

Performance and Best Practices Guide for IBM Storage FlashSystem and IBM SAN Volume Controller

Updated for IBM Storage Virtualize Version 8.6

Bernhard Albrecht
Christian Schroeder
Denis Olshanskiy
Hartmut Lonzer
John Nycz
Jon Herd
Konrad Trojok
Nils Olsson
Patrik Groß
Rene Wuellenweber

Sergey Kubin
Vasfi Gucer
Vineet Sharma



Storage



IBM Redbooks

**Performance and Best Practices Guide for IBM Storage
FlashSystem and IBM SAN Volume Controller**

September 2024

Note: Before using this information and the product it supports, read the information in “Notices” on page xxvii.

First Edition (September 2024)

This edition applies to IBM Storage Virtualize Version 8.6.

© Copyright International Business Machines Corporation 2024. All rights reserved.

Note to U.S. Government Users Restricted Rights -- Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Contents

Figures	xiii
Tables	xxi
Examples	xxv
Notices	xxvii
Trademarks	xxviii
Preface	xxix
Authors	xxix
Now you can become a published author, too!	xxxiii
Comments welcome	xxxiii
Stay connected to IBM Redbooks	xxxiii
Chapter 1. Introduction and system overview	1
1.1 IBM Storage Virtualize	2
1.1.1 Storage virtualization terminology	3
1.2 IBM SAN Volume Controller architectural overview	8
1.2.1 Controller-based approach	9
1.2.2 SAN or fabric-based appliance solution	9
1.2.3 IBM SAN Volume Controller	10
1.2.4 IBM SAN Volume Controller topology	11
1.3 Latest changes and enhancements	13
1.3.1 IBM Storage Virtualize 8.6.0	13
1.3.2 IBM Storage Virtualize 8.5.0	16
1.3.3 IBM Storage Virtualize 8.4.2	18
1.4 IBM SAN Volume Controller family	24
1.4.1 Components	24
1.4.2 Nodes	25
1.4.3 I/O groups	25
1.4.4 System	26
1.4.5 MDisks	26
1.4.6 Cache	27
1.4.7 Quorum disk	30
1.4.8 Storage pool	33
1.4.9 Volumes	34
1.4.10 Easy Tier	36
1.4.11 Hosts	37
1.4.12 Array	37
1.4.13 Encryption	37
1.4.14 iSCSI and iSCSI Extensions over RDMA	39
1.4.15 Data reduction pools	39
1.4.16 IP replication	40
1.4.17 IBM Storage Virtualize copy services	41
1.4.18 Synchronous or asynchronous remote copy	41
1.4.19 Policy-based replication	42
1.4.20 FlashCopy and Transparent Cloud Tiering	43
1.4.21 IBM HyperSwap	44

1.5 IBM SAN Volume Controller models	44
1.5.1 IBM SAN Volume Controller SV3	44
1.5.2 IBM SAN Volume Controller SV2 and SA2	46
1.5.3 IBM SAN Volume Controller model comparisons	48
1.6 IBM FlashSystem family	49
1.6.1 Storage Expert Care	50
1.7 IBM FlashSystem 9500 overview	55
1.7.1 IBM FlashSystem 9500 hardware components	55
1.7.2 IBM FlashSystem 9500 control enclosure	57
1.7.3 IBM FlashSystem 9000 Expansion Enclosure Models AFF and A9F	60
1.8 IBM FlashSystem 9500R Rack Solution overview	61
1.8.1 IBM FlashSystem 9500R Rack Solution diagram	62
1.8.2 FC cabling and clustering	64
1.9 IBM FlashSystem 9000 Expansion Enclosure Models AFF and A9F	64
1.10 IBM FlashSystem 7300 overview	69
1.10.1 IBM FlashSystem 7300 control enclosures	70
1.10.2 IBM FlashSystem 7000 Expansion Enclosure 4657 Models 12G, 24G, and 92G	71
1.11 IBM FlashSystem 5200 overview	74
1.11.1 IBM FlashSystem 5200 expansion enclosures	77
1.12 IBM FlashSystem 5000 family overview	78
1.12.1 IBM FlashSystem 5015	80
1.12.2 IBM FlashSystem 5035	81
1.12.3 IBM FlashSystem 5045	83
1.13 Storage efficiency and data reduction features	88
1.13.1 IBM Easy Tier	88
1.13.2 Data reduction and UNMAP	89
1.13.3 Compression and deduplication	91
1.13.4 Enhanced data security features	92
1.14 Application integration features	94
1.15 Features for manageability	95
1.16 Copy services	100
1.16.1 Policy-based replication	103
1.16.2 HyperSwap	103
1.17 IBM FlashCore Module drives, NVMe SSDs, and SCM drives	105
1.18 Storage virtualization	108
1.19 Business continuity	111
1.19.1 Business continuity with the IBM SAN Volume Controller	111
1.19.2 Business continuity with stretched clusters	112
1.19.3 Business continuity with enhanced stretched cluster	113
1.19.4 Business continuity for FlashSystem and IBM SAN Volume Controller	113
1.19.5 Business continuity with HyperSwap	113
1.19.6 Business continuity with 3-site replication	114
1.19.7 Automatic hot spare nodes (IBM SAN Volume Controller only)	115
1.20 Management and support tools	117
1.20.1 IBM Assist On-site and Remote Support Assistance	117
1.20.2 Event notifications	117
1.21 Licensing	120
1.21.1 Licensing IBM SAN Volume Controller	120
1.21.2 Licensing IBM FlashSystem 9500/R, 9200/R, 7300, 7200, 5200 and 5045	120
1.21.3 Licensing IBM FlashSystem 5035 and 5015	121
Chapter 2. Storage area network guidelines	123
2.1 SAN topology general guidelines	124

2.1.1	SAN performance and scalability	124
2.1.2	ISL considerations	125
2.2	SAN topology-specific guidelines	127
2.2.1	Single-switch IBM Storage Virtualize SANs	127
2.2.2	Basic core-edge topology	128
2.2.3	Edge-core-edge topology	129
2.2.4	Full mesh topology	131
2.2.5	IBM Storage Virtualize as a multi-SAN device	131
2.2.6	Device placement	132
2.2.7	SAN partitioning	134
2.3	IBM Storage Virtualize system ports	134
2.3.1	SAN Volume Controller ports	135
2.3.2	IBM FlashSystem 9200 and 9500 controller ports	136
2.3.3	IBM FlashSystem 7200 and 7300 controller ports	138
2.3.4	IBM FlashSystem 5100 and 5200 and IBM FlashSystem 5015 and 5035 controller ports	140
2.3.5	IBM Storage Virtualize N_Port ID Virtualization, port naming, and distribution.	142
2.3.6	Buffer credits	148
2.4	Zoning	149
2.4.1	Types of zoning	149
2.4.2	Prezoning tips and shortcuts	154
2.4.3	IBM Storage Virtualize internode communications zones	156
2.4.4	IBM Storage Virtualize storage zones	156
2.4.5	IBM Storage Virtualize host zones	173
2.4.6	Hot Spare Node zoning considerations	178
2.4.7	Zoning with multiple IBM Storage Virtualize clustered systems	179
2.4.8	Split storage subsystem configurations	180
2.4.9	IBM Storage Virtualize Ethernet connectivity	180
2.5	Distance extension for remote copy services	181
2.5.1	Optical multiplexors	181
2.5.2	Long-distance SFPs or XFPs	181
2.5.3	Fibre Channel over IP	181
2.5.4	SAN extension with business continuity configurations	183
2.5.5	Native IP replication	186
2.6	Tape and disk traffic that share the SAN	188
2.7	Switch interoperability	189
Chapter 3.	Storage back-end	191
3.1	Internal storage types	192
3.1.1	NVMe storage	192
3.1.2	SAS drives	198
3.1.3	Internal storage considerations	201
3.1.4	IBM SAN Volume Controller internal storage considerations	206
3.2	Arrays	206
3.2.1	Supported RAID types	206
3.2.2	Array considerations	208
3.2.3	Compressed array monitoring	213
3.3	General external storage considerations	215
3.3.1	Storage controller path selection	215
3.3.2	Guidelines for creating an optimal back-end configuration	217
3.3.3	Considerations for compressing and deduplicating back-end controllers	219
3.3.4	Using data reduction at two levels	220
3.3.5	Data reduction pools above a simple RAID	221

3.3.6	Data reduction pools above a data reducing back-end	221
3.4	Controller-specific considerations	221
3.4.1	Considerations for DS8000 series	221
3.4.2	Considerations for XIV Gen3	230
3.4.3	Considerations for IBM FlashSystem A9000 and A9000R	232
3.4.4	Considerations for IBM FlashSystem 5000, 5100, 5200, 7200, 7300, 9100, 9200, and 9500 and IBM SVC SV1, SV2, and SV3.	234
3.4.5	IBM FlashSystem 900 considerations.	238
3.4.6	Path considerations for third-party storage with EMC VMAX, EMC PowerMAX, and Hitachi Data Systems	239
3.5	Quorum disks	239
Chapter 4. Storage pools		243
4.1	Introducing pools	244
4.1.1	Standard pools	244
4.1.2	Data reduction pools	249
4.2	Knowing your data and workload	252
4.2.1	Determining whether your data is compressible	253
4.2.2	Determining whether your data is a deduplication candidate	254
4.2.3	Data reduction estimation tools.	254
4.2.4	Determining the workload and performance requirements	257
4.3	Storage pool planning considerations	259
4.3.1	Selecting pool type	259
4.3.2	Planning for availability	263
4.3.3	Planning for performance	265
4.3.4	Planning for capacity.	266
4.3.5	Extent size considerations	268
4.3.6	External pools	269
4.4	Data reduction pools best practices	270
4.4.1	Data reduction pools with internal storage	271
4.4.2	DRP and external storage considerations.	272
4.4.3	DRP provisioning considerations	273
4.4.4	Standard and data reduction pools coexistence	275
4.4.5	Data migration with data reduction pools	275
4.4.6	Understanding capacity use in a data reduction pool	276
4.5	Operations with storage pools.	279
4.5.1	Adding external MDisks to existing storage pools.	279
4.5.2	Renaming MDisks.	280
4.5.3	Removing MDisks from storage pools	280
4.5.4	Remapping managed MDisks.	285
4.5.5	Controlling the extent allocation order for volume creation.	286
4.6	Considerations when using encryption	286
4.6.1	General considerations.	287
4.6.2	Hardware and software encryption	287
4.6.3	Encryption at rest with USB keys	290
4.6.4	Encryption at rest with key servers	291
4.7	Easy Tier and tiered and balanced storage pools.	296
4.7.1	Easy Tier concepts	297
4.7.2	Easy Tier definitions	299
4.7.3	Easy Tier operating modes.	301
4.7.4	MDisk tier types	304
4.7.5	Changing the tier type of an MDisk.	309
4.7.6	Easy Tier overload protection	309

4.7.7	Easy Tier overallocation protection	310
4.7.8	Removing an MDisk from an Easy Tier pool.	311
4.7.9	Easy Tier implementation considerations	311
4.7.10	Easy Tier settings	317
Chapter 5.	Volumes	319
5.1	Volumes overview	320
5.2	Guidance for creating volumes	321
5.3	Thin-provisioned volumes	323
5.3.1	Compressed volumes	328
5.3.2	Deduplicated volumes.	329
5.3.3	Thin provisioning considerations	330
5.4	Mirrored volumes	333
5.4.1	Write fast failovers	336
5.4.2	Read fast failovers	337
5.4.3	Maintaining data integrity of mirrored volumes	337
5.5	HyperSwap volumes	337
5.6	VMware vSphere Virtual Volumes	339
5.7	Cloud volumes	339
5.7.1	Transparent Cloud Tiering configuration limitations and rules	340
5.7.2	Restoring to the production volume	341
5.7.3	Restoring to a new volume	341
5.8	Volume migration	342
5.8.1	Image-type to striped-type volume migration	343
5.8.2	Migrating to an image-type volume	343
5.8.3	Migrating with volume mirroring	344
5.8.4	Migrating from standard pools to data reduction pools	346
5.8.5	Migrating a volume between systems nondisruptively	347
5.9	Preferred paths to a volume	351
5.10	Moving a volume between I/O groups and nodes.	352
5.10.1	Changing the preferred node of a volume within an I/O group	352
5.10.2	Moving a volume between I/O groups.	352
5.11	Volume throttling	353
5.12	Volume cache mode	356
5.13	Other considerations	359
5.13.1	Volume protection	359
5.13.2	Volume resizing	360
Chapter 6.	Copy services	363
6.1	Introducing copy services	364
6.1.1	FlashCopy	364
6.1.2	Metro Mirror and Global Mirror	364
6.1.3	Volume mirroring.	364
6.2	IBM FlashCopy	365
6.2.1	FlashCopy use cases	365
6.2.2	FlashCopy capabilities overview	367
6.2.3	FlashCopy functional overview	373
6.2.4	FlashCopy planning considerations	381
6.3	Safeguarded Copy	391
6.3.1	Safeguarded Copy use cases	391
6.3.2	Cyber resiliency	392
6.3.3	Safeguarded Copy functional overview.	392
6.3.4	Safeguarded Copy considerations	395

6.4 Remote copy services	397
6.4.1 Remote copy use cases	398
6.4.2 Remote copy functional overview	398
6.4.3 Remote copy network planning	415
6.4.4 Remote copy services planning	430
6.4.5 Multiple site remote copy	441
6.4.6 1920 error	445
6.5 Native IP replication	458
6.5.1 Native IP replication technology	458
6.5.2 IP partnership limitations	459
6.5.3 VLAN support	461
6.5.4 Domain name support for IP replication	462
6.5.5 IP compression	462
6.5.6 Replication portsets	463
6.5.7 Supported configurations examples	465
6.5.8 Native IP replication performance considerations	475
6.6 Volume mirroring	476
6.6.1 Read/write operations	477
6.6.2 Volume mirroring use cases	478
6.6.3 Mirrored volume components	480
6.6.4 Volume mirroring synchronization options	481
6.6.5 Volume mirroring performance considerations	482
6.6.6 Bitmap space for out-of-sync volume copies	484
6.7 Policy-based replication	486
6.7.1 Implementing policy-based replication	486
6.7.2 Limitations and restrictions	491
6.7.3 Supported platforms for 8.5.2 or above:	492
6.7.4 I/O group memory requirement	492
6.7.5 Managing policy-based replication	493
6.7.6 Production overview	494
6.7.7 Recovery overview	496
6.7.8 Synchronization	498
6.7.9 Status and recovery point reporting	500
Chapter 7. Ensuring business continuity	501
7.1 High availability and disaster recovery	502
7.2 Business continuity with a Stretched cluster topology	502
7.2.1 Stretched cluster (SC)	502
7.2.2 Enhanced Stretched cluster (ESC)	503
7.3 Business continuity with HyperSwap	504
7.3.1 HyperSwap overview	505
7.3.2 HyperSwap build steps	509
7.4 Comparing business continuity solutions	510
7.5 Quorum site and the IP quorum application	513
7.5.1 IP quorum overview	513
7.5.2 Quorum modes	514
7.5.3 IP quorum as a service	515
7.6 HyperSwap internals	515
7.7 Other considerations and general recommendations	516
Chapter 8. Hosts	519
8.1 Connection protocols	521
8.2 SCSI FC general configuration guidelines	521

8.2.1	Number of paths	521
8.2.2	Host ports	522
8.2.3	N_Port ID Virtualization.	522
8.2.4	Host to I/O group mapping	522
8.2.5	Volume size versus quantity	522
8.2.6	Host volume mapping	523
8.2.7	Server adapter layout	524
8.2.8	Host status improvements.	524
8.3	SAS host attachment	524
8.4	iSCSI host attachment considerations	525
8.4.1	Connection basics.	525
8.4.2	Performance tuning.	525
8.5	iSCSI Extensions for Remote Direct Memory Access host attachment considerations	527
8.6	NVMe over Fibre Channel host attachment considerations	528
8.7	NVMe over Ethernet host attachment considerations	528
8.7.1	NVMe over Ethernet (RoCE) host attachment considerations	529
8.7.2	NVMe over TCP host attachment considerations	529
8.8	Portsets	530
8.8.1	IP multitenancy	530
8.8.2	Fibre Channel portset	531
8.8.3	Portsets considerations and limitations.	533
8.9	Host pathing	533
8.9.1	Path selection	534
8.10	I/O queues.	534
8.10.1	Queue depths	534
8.11	Host clusters	535
8.11.1	Persistent reservations	536
8.11.2	Clearing reserves	537
8.12	AIX hosts.	538
8.12.1	Multipathing support	538
8.12.2	AIX configuration recommendations	538
8.13	Virtual I/O Server hosts	538
8.13.1	Multipathing support	538
8.13.2	VIOS configuration recommendations	539
8.13.3	Physical and logical volumes	539
8.13.4	Identifying a disk for use as a VSCSI disk	539
8.14	Microsoft Windows hosts	540
8.14.1	Multipathing support	540
8.14.2	Windows and Hyper-V configuration recommendations	540
8.15	Linux hosts	540
8.15.1	Multipathing support	540
8.15.2	Linux configuration recommendations	541
8.16	Oracle Solaris hosts support.	541
8.16.1	Multipathing support	541
8.16.2	Solaris MPxIO configuration recommendations	541
8.16.3	Symantec Veritas DMP configuration recommendations	542
8.17	HP 9000 and HP Integrity hosts	543
8.17.1	Multipathing support	544
8.17.2	HP configuration recommendations	544
8.18	VMware ESXi server hosts	544
8.18.1	Multipathing support	544
8.18.2	VMware configuration recommendations	545
8.19	Container Storage Interface Block Driver	545

8.20	100-Gigabit Ethernet host connectivity	546
8.20.1	Dual port 100 GbE adapter functions	546
8.20.2	Maximum adapter count and slot placement	547
8.20.3	Dual Port 100 GbE adapter cables and connectors	549
Chapter 9. Implementing a storage monitoring system		551
9.1	Generic monitoring	552
9.1.1	Monitoring by using the management GUI	552
9.1.2	Call Home with email notifications	552
9.1.3	Simple Network Management Protocol notification.	553
9.1.4	Syslog notification	555
9.1.5	Monitoring by using quotas and alerts	556
9.2	Performance monitoring	557
9.2.1	Onboard performance monitoring	557
9.2.2	Performance monitoring with IBM Spectrum Control	567
9.2.3	Performance monitoring with IBM Storage Insights	571
9.3	Capacity monitoring	588
9.3.1	Capacity monitoring through the management GUI	589
9.3.2	Capacity monitoring by using IBM Spectrum Control or IBM Storage Insights.	595
9.4	Creating alerts for IBM Storage Control and IBM Storage Insights.	606
9.4.1	Alert examples	607
9.4.2	Alert example to monitor pool capacity: Used Capacity	607
9.5	Health monitoring	613
9.5.1	Health monitoring in the IBM Storage Virtualize GUI	613
9.5.2	Health monitoring through the IBM Storage Virtualize command-line interface (CLI) 615	
9.5.3	Health monitoring in IBM Spectrum Control	616
9.5.4	Health monitoring in IBM Storage Insights	619
9.6	Important performance metrics	620
9.7	Performance diagnostic information	628
9.7.1	Performance diagnostic information included in a support package.	628
9.7.2	Performance diagnostic information exported from IBM Spectrum Control	629
9.7.3	Performance diagnostic information exported from IBM Storage Insights	631
9.8	Metro Mirror and Global Mirror monitoring	632
9.8.1	Monitoring with IBM Copy Services Manager	632
9.8.2	Monitoring MM and GM with scripts	635
9.9	Monitoring Tier 1 SSDs.	635
Chapter 10. Maintaining an IBM Storage Virtualize infrastructure		637
10.1	User interfaces	638
10.1.1	Management GUI	638
10.1.2	Service Assistant Tool GUI	639
10.1.3	Command-line interface	639
10.2	Users and groups	641
10.3	Volumes	643
10.4	Hosts	644
10.5	Software updates	644
10.5.1	Determining the current and target software level	645
10.5.2	Obtaining the software packages	647
10.5.3	Hardware considerations	650
10.5.4	Update sequence	651
10.5.5	SAN fabrics preparation	651
10.5.6	Storage controllers preparation.	652

10.5.7	Host preparation	652
10.5.8	Copy services considerations	652
10.5.9	Running the Upgrade Test Utility	652
10.5.10	Updating the software	656
10.6	Non-disruptive security and software patching	658
10.6.1	Installing a software patch	659
10.6.2	Verifying a software patch.	659
10.6.3	Removing a software patch	660
10.7	Drive firmware updates for IBM FlashSystem.	661
10.8	Remote Code Load	663
10.9	Replacing IBM FlashCore Module in IBM FlashSystem	666
10.10	SAN modifications	667
10.10.1	Cross-referencing WWPNs	667
10.10.2	Cross-referencing LUN IDs	669
10.11	Server HBA replacement	669
10.12	Hardware upgrades	671
10.12.1	Adding control enclosures	671
10.12.2	Adding IBM SVC nodes	673
10.12.3	Upgrading nodes in an existing cluster	675
10.12.4	Upgrading NVMe drives	680
10.12.5	Splitting an IBM Storage Virtualize cluster	681
10.12.6	IBM FlashWatch	681
10.13	I/O throttling.	682
10.13.1	General information about I/O throttling	683
10.13.2	I/O throttling on front-end I/O control	683
10.13.3	I/O throttling on back-end I/O control	683
10.13.4	Overall benefits of using I/O throttling.	684
10.13.5	Considerations for I/O throttling	684
10.13.6	Configuring I/O throttling by using the CLI	685
10.13.7	Creating a throttle by using the GUI	686
10.14	Documenting an IBM Storage Virtualize and SAN environment.	688
10.14.1	Naming conventions	689
10.14.2	SAN fabric documentation	692
10.14.3	IBM Storage Virtualize documentation	694
10.14.4	Storage documentation.	696
10.14.5	Technical support information.	697
10.14.6	Tracking incident and change tickets	697
10.14.7	Automated support data collection	698
10.14.8	Subscribing to IBM Storage Virtualize support	699
Chapter 11.	Storage Virtualize Troubleshooting and diagnostics	701
11.1	Troubleshooting	702
11.1.1	Using the GUI	703
11.1.2	Recommended actions and fix procedure.	705
11.1.3	Using the command-line interface.	709
11.2	Collecting diagnostic data	710
11.2.1	IBM Storage Virtualize systems data collection	710
11.2.2	Host multipath software	715
11.2.3	Drive data collection: drivedumps	721
11.2.4	More data collection	721
11.3	Common problems and isolation techniques	722
11.3.1	Interoperability	723
11.3.2	Host problems.	723

11.3.3	Fibre Channel SAN and IP SAN problems	728
11.3.4	Port issues and transceiver statistics	730
11.3.5	Storage subsystem problems	731
11.3.6	Native IP replication problems	736
11.3.7	Remote Direct Memory Access-based clustering	737
11.3.8	Advanced copy services or data reduction related problems	739
11.3.9	Health status during an upgrade	743
11.3.10	Managing the physical capacity of overprovisioned storage controllers	743
11.3.11	Replacing a failed flash drive	745
11.3.12	Recovering from common events	745
11.4	Remote Support Assistance	746
11.5	Call Home Connect Cloud and Health Checker feature	747
11.5.1	Health Checker	748
11.5.2	Configuring Call Home by using Ansible	748
11.6	IBM Storage Insights	748
11.6.1	IBM Storage Insights customer main dashboard	752
11.6.2	Customized dashboards to monitor your storage	752
11.6.3	Creating a support ticket	752
11.6.4	Updating a support ticket	758
11.6.5	IBM Storage Insights Advisor	763
Appendix A. IBM i considerations		765
IBM i Storage management		767
Single-level storage		768
IBM i response time		770
Planning for IBM i storage capacity		774
Storage connection to IBM i		775
Native attachment		775
VIOS attachment		776
Setting attributes in VIOS		779
FC adapter attributes		779
Disk device attributes		780
Guidelines for Virtual I/O Server resources		780
Disk drives for IBM i		781
Defining LUNs for IBM i		783
Data layout		785
Fibre Channel adapters in IBM i and VIOS		786
Zoning SAN switches		786
IBM i multipath		787
Booting from SAN		788
IBM i mirroring		788
Copy Services considerations		788
Remote replication		788
FlashCopy		790
IBM Lab Services PowerHA Tools for IBM i		791
HyperSwap		794
SAN Volume Controller stretched cluster		797
Db2 mirroring for IBM i		801
Related publications		807
IBM Redbooks		807
Online resources		807
Help from IBM		808

Figures

1-1	Block-level virtualization overview	5
1-2	IBM Storage Virtualize features by product	7
1-3	Overview of block-level virtualization architectures	9
1-4	IBM SAN Volume Controller conceptual and topology overview	12
1-5	IBM Safeguarded Copy Data Resilience	20
1-6	STaaS Tiers comparison	22
1-7	IBM SAN Volume Controller family	24
1-8	Separation of upper and lower cache	29
1-9	Overview of an IBM SAN Volume Controller clustered system with an I/O group	33
1-10	Striped volume	34
1-11	Sequential volume	35
1-12	Image mode volume	35
1-13	Global Mirror with change volumes	42
1-14	IBM SAN Volume Controller SV3 front view	44
1-15	IBM SAN Volume Controller SV3 rear view	45
1-16	IBM SAN Volume Controller SV3 internal hardware components	45
1-17	IBM SAN Volume Controller SV3 internal architecture	46
1-18	IBM SAN Volume Controller SV2 and SA2 front view	46
1-19	IBM SAN Volume Controller SV2 and SA2 rear view	46
1-20	Internal hardware components	47
1-21	IBM SAN Volume Controller SV2 and SA2 internal architecture	47
1-22	IBM FlashSystem family	50
1-23	Storage Expert Care tier levels for IBM FlashSystem 5015 and IBM FlashSystem 5045	52
1-24	IBM FlashSystem 9500 front and rear views	55
1-25	IBM FlashSystem 9500 internal architecture	56
1-26	FS9500 internal hardware components	57
1-27	Logical NVMe drive placement	60
1-28	NVMe drive population with numbering	60
1-29	Legend to figures in this section	63
1-30	IBM FlashSystem 9500R Rack Solution configuration in the rack	63
1-31	IBM FlashSystem 9000 Expansion Enclosure Model AFF	65
1-32	IBM FlashSystem 9000 Expansion Enclosure Model front view	66
1-33	IBM FlashSystem 9200 system that is connected to expansion enclosure	67
1-34	IBM FlashSystem 9500 system that is connected to expansion enclosure	68
1-35	IBM FlashSystem 7300 front and rear views	69
1-36	IBM FlashSystem 7300 internal architecture	70
1-37	Internal hardware components	70
1-38	IBM FlashSystem 7000 LFF Expansion Enclosure Model 12G	72
1-39	Front view of an IBM FlashSystem 7000 SFF expansion enclosure	72
1-40	Rear of an IBM FlashSystem 7000 expansion enclosure	73
1-41	IBM Dense Expansion Drawer	73
1-42	Connecting FS7300 SAS cables while complying with the maximum chain weight	74
1-43	IBM FlashSystem 5200 control enclosure front and 3/4 ISO view	76
1-44	IBM FlashSystem 5015, 5035 and 5045 SFF control enclosure front view	78
1-45	IBM FlashSystem 5015, 5035 and 5045 LFF control enclosure front view	79
1-46	Front view of an IBM FlashSystem 5035	83
1-47	Rear view of an IBM FlashSystem 5035	83

1-48	View of available connectors and LEDs on an IBM FlashSystem 5035 single canister .	83
1-49	Front view of an IBM FlashSystem 5045	86
1-50	Rear view of an IBM FlashSystem 5045.	86
1-51	View of available connectors and LEDs on an IBM FlashSystem 5045 single canister .	86
1-52	Easy Tier concept	88
1-53	IBM Storage Virtualize GUI dashboard.	96
1-54	IBM Storage Insights dashboard.	99
1-55	IBM FlashCore Module (NVMe)	105
1-56	Storage technologies versus latency for Intel drives.	107
1-57	“Star” and “Cascade” modes in a three-site solution.	115
2-1	ISL data flow	126
2-2	Single-switch SAN	128
2-3	Core-edge topology	129
2-4	Edge-core-edge topology	130
2-5	Full mesh topology	131
2-6	IBM Storage Virtualize as a SAN bridge.	132
2-7	Storage and hosts attached to the same SAN switch.	133
2-8	Edge-core-edge segmentation	133
2-9	SAN Volume Controller 2145-SV1 rear port view	135
2-10	SV2/SA2 node layout	136
2-11	SV3 node layout	136
2-12	Port location in the IBM FlashSystem 9200 rear view.	138
2-13	Port location in IBM FlashSystem 9500 rear view.	138
2-14	IBM FlashSystem 7200 rear view	139
2-15	IBM FlashSystem 7300 rear view	139
2-16	IBM FlashSystem 5100 rear view	141
2-17	IBM FlashSystem 5200 rear view	142
2-18	IBM FlashSystem 5015 rear view	142
2-19	IBM FlashSystem 5035 rear view	142
2-20	IBM Storage Virtualize NPIV Port WWPN.	143
2-21	IBM Storage Virtualize NPIV Failover.	143
2-22	IBM Storage Virtualize output of the lstorageportfc command	144
2-23	SAN Volume Controller model 2145-SV1 port distribution	144
2-24	SAN Volume Controller model SV3 port distribution.	145
2-25	IBM FlashSystem 9200 port distribution	145
2-26	Port masking configuration on SVC or IBM FlashSystem with 16 ports	146
2-27	Port masking configuration on IBM FlashSystem or SVC with 24 ports	147
2-28	IBM Storage Virtualize Portsets overview.	151
2-29	Listing the available ports and portsets.	153
2-30	How to assign ports pairs	153
2-31	Output of the IBM Storage Virtualize lstorageportfc command	155
2-32	Back-end storage zoning	157
2-33	V5000 zoning	158
2-34	Dual core zoning schema	159
2-35	ISL traffic overloading	160
2-36	XIV port cabling.	162
2-37	IBM FlashSystem A9000 connectivity.	163
2-38	IBM FlashSystem A9000 grid configuration cabling	165
2-39	Connecting IBM FlashSystem A9000 fully configured as a back-end controller.	166
2-40	V7000 connected as a back-end controller.	167
2-41	IBM FlashSystem 9100 as a back-end controller	168

2-42	IBM FlashSystem 900 connectivity to a SAN Volume Controller cluster	169
2-43	DS8900F I/O adapter layout	171
2-44	DS8886 to SAN Volume Controller connectivity	172
2-45	Host zoning to an IBM Storage Virtualize node	173
2-46	Four-port host zoning	175
2-47	VMware ESX cluster zoning	176
2-48	LPARs SAN connections	177
2-49	Live Partition Mobility	178
2-50	Typical Enhanced Stretched Cluster configuration	184
2-51	Configuration 1: Physical paths shared among the fabrics	185
2-52	Configuration 2: Physical paths not shared among the fabrics	186
2-53	Effect of distance on packet loss	187
2-54	FC frames access methods	188
2-55	Production and backup fabric	189
3-1	FCM capacity monitoring with GUI	202
3-2	Array capacity monitoring with the GUI	214
3-3	IBM System Storage Interoperation Center example	216
3-4	Virtualization concepts of DS8900F for IBM Storage Virtualize	223
3-5	DA pair reduced bandwidth configuration	224
3-6	DA pair correct configuration	226
3-7	The lsarray and lsrank command output	226
3-8	Four DS8900F extent pools as one IBM Storage Virtualize storage pool	229
3-9	XIV rack configuration: 281x-214	231
4-1	Standard and data reduction pool volumes	250
4-2	Garbage-collection principle	251
4-3	Capacity savings analysis	255
4-4	Customized view	256
4-5	Example dashboard capacity view	267
4-6	Compression savings dashboard report	268
4-7	Data reduction pool capacity use example	278
4-8	IBM FlashSystem volume details	283
4-9	IBM FlashSystem volume details for host maps	284
4-10	SAN Volume Controller MDisk details for IBM FlashSystem volumes	285
4-11	Encryption placement in lower layers of the IBM FlashSystem software stack	288
4-12	Mixed encryption in a storage pool	289
4-13	Update Certificate on IBM FlashSystem	292
4-14	Creating a self-signed certificate: IBM Security Key Lifecycle Manager server	292
4-15	Create Device Group for IBM FlashSystem	293
4-16	IBM Security Key Lifecycle Manager replication schedule	294
4-17	Keys that are associated to a device group	295
4-18	Easy Tier single volume with multiple tiers	298
4-19	Easy Tier extent migration types	301
4-20	Single tier storage pool with a striped volume	306
4-21	Multitier storage pool with a striped volume	307
5-1	Thin-provisioned volume	323
5-2	Different kinds of volumes in a DRP	324
5-3	Thin-provisioned volume concepts	325
5-4	Modifying the capacity savings of a volume nondisruptively	327
5-5	Customized view	329
5-6	Mirrored volume creation	335
5-7	Adding a volume copy	335
5-8	Overall HyperSwap diagram	338
5-9	Master and auxiliary volumes	339

5-10	Transparent Cloud Tiering example	340
5-11	Migrating with volume mirroring	346
5-12	Converting volumes with volume mirroring	347
5-13	Write operations from a host through different preferred nodes for each volume. . .	351
5-14	Volume throttling for each volume	353
5-15	Volume throttling	354
5-16	Edit bandwidth and IOPS limit	354
5-17	Cache activated	356
5-18	Cache deactivated	357
5-19	Editing cache mode	358
5-20	Volume Protection.	360
5-21	Expanding a volume	361
5-22	Shrinking volumes.	362
6-1	FlashCopy mapping	367
6-2	Multiple volume mappings in a consistency group	369
6-3	Incremental FlashCopy	370
6-4	Multiple target FlashCopy	371
6-5	Cascaded FlashCopy	371
6-6	FlashCopy mapping states diagram	375
6-7	Cache architecture	378
6-8	Logical placement of the FlashCopy indirection layer.	379
6-9	Interaction between multiple Target FlashCopy mappings	380
6-10	IBM Storage Virtualize presets	384
6-11	IBM Safeguarded Copy provides logical corruption protection to protect sensitive point in time copies of data	393
6-12	Remote copy components and applications	400
6-13	Remote copy partnership	401
6-14	Role and direction changes	402
6-15	Conceptualization of layers.	404
6-16	Supported topologies for remote copy partnerships	405
6-17	Metro Mirror write sequence	407
6-18	Global Mirror relationship write operation	409
6-19	Colliding writes	411
6-20	Global Mirror with Change Volumes	411
6-21	GMCV uses FlashCopy point-in-time copy technology.	413
6-22	Standard SCSI read operation	416
6-23	Standard SCSI write operation	416
6-24	IBM Storage Virtualize remote copy write.	417
6-25	Zoning scheme for >80 ms remote copy partnerships	426
6-26	Typical remote copy network configuration.	427
6-27	Configuration 1: Physical paths shared among the fabrics.	428
6-28	Configuration 2: Physical paths not shared among the fabrics.	429
6-29	Remote copy resources that are not optimized	434
6-30	Optimized Global Mirror resources	435
6-31	Three-site configuration with Enhanced Stretched Cluster.	442
6-32	Using 3-way copy services	443
6-33	Cascading-like infrastructure	444
6-34	Effect of packet size (in bytes) versus the link size.	455
6-35	Typical Ethernet network data flow	458
6-36	Optimized network data flow by using Bridgeworks SANSlide technology.	459
6-37	Portsets.	463
6-38	Only one link on each system and canister with failover ports configured	465
6-39	Clustered or multinode systems with a single inter-site link with only one link	467

6-40	Dual links with two replication portsets on each system configured	469
6-41	Clustered/multinode systems with dual inter-site links between the two systems . .	470
6-42	Multiple IP partnerships with two links and only one I/O group.	472
6-43	Multiple IP partnerships with two links	474
6-44	1-Gbps port throughput trend	476
6-45	Volume mirroring overview	477
6-46	Attributes of a volume and volume mirroring	481
6-47	IOgrp feature example	485
6-48	Possible pool and policy relations.	489
6-49	Region mapping with bitmap.	495
7-1	Typical concept scheme HyperSwap configuration with IBM Storage Virtualize	506
7-2	IBM Storage Virtualize HyperSwap in a storage failure scenario	507
7-3	IBM FlashSystem HyperSwap in a site failure scenario	508
7-4	Initializing the first node of a HyperSwap system	509
7-5	IP quorum network layout.	513
7-6	HyperSwap volume UID	516
8-1	SCSI ID assignment on volume mappings	536
8-2	Dual Port 100 GbE adapter placement on IBM FlashSystem Storage 7300	547
8-3	Dual Port 100 GbE adapter placement on IBM Storage FlashSystem 9500	548
8-4	Dual Port 100 GbE adapter placement on SAN Volume Controller node SV3	548
9-1	Email users showing customizable notifications.	552
9-2	Call Home with cloud services configuration window	553
9-3	SNMP configuration summary	554
9-4	SNMP server configuration window	554
9-5	Syslog servers	555
9-6	Pool threshold.	556
9-7	VDisk threshold.	556
9-8	Monitoring/Performance overview	558
9-9	Workload metrics	559
9-10	Management GUI Dashboard view	559
9-11	Authentication in REST API Explorer: Token displayed in the Response body	560
9-12	The lsnodestats command for node ID 1 (fc_mb) with JSON results in a response body 561	561
9-13	Easy Tier Data Movement window	562
9-14	Easy Tier Movement description window	563
9-15	Easy Tier Composition report window	564
9-16	Easy Tier Composition Description.	564
9-17	Workload skew: Single tier pool	565
9-18	Workload skew: Multitier configuration	566
9-19	IBM Spectrum Control Dashboard	568
9-20	Key Performance Indicators	569
9-21	Write response Time by I/O Group > 5 ms	570
9-22	IBM Storage Insights registration window.	574
9-23	IBM Storage Insights or IBM Storage Insights for IBM Spectrum Control registration options	575
9-24	Registration login window	575
9-25	Creating an IBM account	576
9-26	IBMid account privacy.	577
9-27	IBM Storage Insights registration form	577
9-28	IBM Storage Insights initial setup guide	578
9-29	IBM Storage Insights Deployment Planning	578
9-30	IBM Storage Insights info event after the si_tenant_id was added to Cloud Call Home. 579	579

9-31	Storage Insights - Add Call Home with cloud service device	579
9-32	Select Operating System window	581
9-33	Data collector license agreement	582
9-34	Downloading the data collector in preparation for its installation	582
9-35	Data collector installation on a Linux host.	583
9-36	Adding storage systems to IBM Storage Insights	583
9-37	Operations Dashboard	584
9-38	NOC dashboard	585
9-39	Block Storage Systems table view	586
9-40	Advisor Insights window	587
9-41	Understanding capacity information	588
9-42	Capacity terminology	589
9-43	Usable capacity.	589
9-44	Capacity Savings window	591
9-45	Sidebar > Pools > Properties > Properties for Pool	592
9-46	Sidebar > Pools > MDisks by Pools > Properties > More details	593
9-47	Easy Tier Overallocation Limit GUI support	594
9-48	IBM Spectrum Control overview page	595
9-49	IBM Storage Insights overview page	595
9-50	Block Storage Systems overview	596
9-51	Capacity overview of Storage System	597
9-52	Used Capacity.	597
9-53	Example of Adjusted Used Capacity	598
9-54	Capacity limit example	599
9-55	Capacity-to-Limit	599
9-56	Zero Capacity	603
9-57	IBM Spectrum Control Alert policies	608
9-58	IBM Storage Insights Alert policies	608
9-59	All alert policies in IBM Spectrum Control.	609
9-60	Copying a policy in IBM Spectrum Control	609
9-61	Copy Policy window	610
9-62	New policy with inherited alert definitions	610
9-63	Choosing the required alert definitions	611
9-64	Alert parameters	611
9-65	Setting up the Warning level.	612
9-66	Setting up the informational threshold	612
9-67	IBM Spectrum Control notification settings	612
9-68	System Health state of management GUI Dashboard	613
9-69	Expanded Hardware Components view for a SAN Volume Controller Cluster.	614
9-70	Expanded Hardware Components view for IBM FlashSystem 9100	614
9-71	Prioritizing tiles that need attention	614
9-72	Dashboard entry point drills down to the event log	615
9-73	Events by Priority	615
9-74	IBM Spectrum Control Dashboard summary	616
9-75	IBM Spectrum Control Block Storage Systems.	617
9-76	Detailed Block Storage System view	617
9-77	Offline volumes	618
9-78	Marking the status as acknowledged	618
9-79	Error status cleared.	618
9-80	IBM Storage Insights dashboard showing a volume error	619
9-81	Actions available from the Volume tile	619
9-82	IBM Spectrum Control: Export Performance Data	630
9-83	IBM Spectrum Control: Export Performance Data - Advanced Export	630

9-84 IBM Spectrum Control: Package files example	631
9-85 Selecting Block Storage Systems	631
9-86 Selecting Export Performance Data	632
9-87 CSM sessions preparing	633
9-88 CSM sessions that are prepared and 100% synced	633
9-89 CSM automatic restart is disabled by default	634
9-90 Secondary consistency warning when automatic restart is enabled	635
10-1 Example for restricted view for ownership groups	642
10-2 Current software version	646
10-3 Up-to-date software version	646
10-4 Fix Central download	648
10-5 Upload package manually	648
10-6 Unhide GUI buttons	649
10-7 Transfer update package	650
10-8 Initiating transfer of the update package	650
10-9 IBM Storage Virtualize Upgrade Test Utility by using the GUI	653
10-10 Example result of the IBM Storage Virtualize Upgrade Test Utility	654
10-11 Installing the patch using the GUI	659
10-12 list installed patches in GUI	660
10-13 Drive firmware upgrade	662
10-14 Drive update test result	662
10-15 IBM FlashSystem RCL Schedule Service page	664
10-16 RCL Product type page	664
10-17 Timeframe selection page	665
10-18 RCL Time selection page	665
10-19 RCL booking information page	666
10-20 IBM Storage Virtualize performance statistics (IOPS)	676
10-21 Distribution of controller resources before and after I/O throttling	684
10-22 Creating a volume throttle in the GUI	686
10-23 Creating a host throttle in the GUI	686
10-24 Creating a host cluster throttle in the GUI	687
10-25 Creating a storage pool throttle in the GUI	687
10-26 Creating a system offload throttle in the GUI	688
10-27 Poorly formatted SAN diagram	693
10-28 Brocade SAN Health Options window	693
10-29 Creating a subscription to IBM Storage Virtualize notification	699
11-1 Events icon in the GUI	703
11-2 GUI Dashboard displaying system health events and hardware components	704
11-3 System Health expanded section in the dashboard	705
11-4 Recommended actions	705
11-5 Monitoring → Events window	708
11-6 Properties and Sense Data for an event	708
11-7 <i>Upload Support Package</i> window	712
11-8 Upload Support Package details	713
11-9 PBR replication status	741
11-10 RPO statuses	741
11-11 Remote Support options	746
11-12 IBM Storage Insights versus IBM Storage Insights Pro	750
11-13 Illustration of multiple data collectors	751
11-14 IBM Storage Insights main dashboard	752
11-15 IBM Storage Insights Get Support option	753
11-16 Create Ticket option	753
11-17 Collecting ticket information	754

11-18	Adding a problem description and any other information	755
11-19	Setting the severity level	756
11-20	Final summary before ticket creation	757
11-21	Support ticket summary	757
11-22	IBM Storage Insights Update Ticket	758
11-23	Entering the IBM Support or PMR case number	759
11-24	Log type selection	760
11-25	Collecting new logs	761
11-26	Add a note or attachment	762
11-27	Update of the support ticket is complete	763
11-28	IBM Storage Insights Advisor menu	764
11-29	Advisor detailed summary of recommendations	764
A-1	IBM i storage management spreads objects across logical unit numbers	767
A-2	Virtual address space	768
A-3	IBM i auxiliary storage pools architecture	769
A-4	TIMI atomicity	770
A-5	Disk subsystem	772
A-6	Disk I/O on IBM i	772
A-7	IBM i with different sector sizes	774
A-8	IBM i SAN access by using NPIV	777
A-9	Sizing and modeling for IBM i by using Disk Magic	782
A-10	SAN switch zoning for IBM i with IBM Storage Virtualize storage	786
A-11	IBM i full system replication with IBM Storage Virtualize	789
A-12	IBM i IASP replication with IBM Storage Virtualize	789
A-13	IBM Lab Services PowerHA Tools for IBM i: Full System Replication Manager	792
A-14	IBM Lab Services PowerHA Tools for IBM i: Full System FlashCopy	793
A-15	IBM i HyperSwap SAN fabric connection example	795
A-16	Full system replication that uses SAN Volume Controller volume mirroring	797
A-17	IBM PowerHA System Mirror for i LUN-level switching with SAN Volume Controller stretched cluster	800
A-18	Db2 Mirror environment with one IBM Storage Virtualize storage system	805
A-19	Db2 Mirror environment with two IBM Storage Virtualize storage systems	805
A-20	Db2 Mirror and full system replication	806

Tables

1-1 IBM Storage Virtualize 8.6 supported product list	2
1-2 IBM SAN Volume Controller base models	48
1-3 Historical overview of IBM SAN Volume Controller models	49
1-4 IBM FlashSystems 7200 and 9200 product range	53
1-5 Software PIDs and SWMA feature codes	53
1-6 Billing calculations that are based on customer usage	59
1-7 IBM FlashSystem 9500 Utility Model UG8 billing feature codes	59
1-8 IBM FlashSystem 9500R Rack Solution combinations	62
1-9 Key to rack configuration	62
1-10 IBM FlashSystem 5200 host, drive capacity, and functions summary	76
1-11 Machine type and model comparison for the IBM FlashSystem 5000	79
1-12 IBM FlashSystem 5015 host, drive capacity, and functions summary	80
1-13 2.5-inch supported drives for the IBM FlashSystem 5000 family	80
1-14 3.5-inch supported drives for the IBM FlashSystem 5000 family	81
1-15 IBM FlashSystem 5035 host, drive capacity, and functions summary	81
1-16 IBM FlashSystem 5045 host, drive capacity, and functions summary	84
1-17 Volume types that are available in pools	90
1-18 FCM type capacities	106
1-19 NVMe drive size options	106
1-20 SCM drive options	107
1-21 SCU category definitions	120
2-1 SAN Volume Controller connectivity	135
2-2 IBM FlashSystem 9200 ports connectivity maximum options per enclosure	136
2-3 IBM FlashSystem 9500 ports connectivity maximum options per enclosure	137
2-4 IBM FlashSystem 7200 ports connectivity maximum per enclosure	138
2-5 IBM FlashSystem 7300 ports connectivity maximum per enclosure	139
2-6 IBM FlashSystem 5100 port connectivity maximum per enclosure	140
2-7 IBM FlashSystem 5200 port connectivity maximum per enclosure	141
2-8 IBM FlashSystem 5015	141
2-9 IBM FlashSystem 5035	141
2-10 Alias names examples	154
2-11 Distribution of aliases	155
2-12 XIV connectivity ports as capacity grows	161
2-13 Number of host ports in an IBM FlashSystem A9000R system	163
2-14 Host connections to SAN Volume Controller	164
2-15 DS8900F port configuration	171
3-1 Number of NVMe drive slots per platform	192
3-2 Supported industry-standard NVMe drives on IBM Storage Virtualize	194
3-3 IBM FlashCore Module capacities	197
3-4 Supported SCM drive capacities	198
3-5 Maximum drive slots per SAS expansion chain for IBM FlashSystem	200
3-6 Maximum drive slots per SAS expansion chain for IBM FlashSystem 9500	200
3-7 Maximum drive slots per SAS expansion chain for IBM FlashSystem 7300	200
3-8 Maximum drive slots per SAS expansion chain for IBM FlashSystem 5045	200
3-9 Cumulative writes based on possible DDPD	204
3-10 Supported RAID levels on IBM Storage Virtualize systems	207
3-11 Supported RAID levels with different drive types	207
3-12 XIV minimum volume size and quantity recommendations	231

3-13	Host connections for A9000	233
3-14	Host connections for A9000R	233
3-15	Adapter cages recommendations for IBM FlashSystem 9500	235
3-16	Adapter recommendations for IBM FlashSystem 7300	235
3-17	Adapter cage recommendations for IBM SVC SV3	236
4-1	Compression ratios of common data types	253
4-2	Upper limit of write cache data	266
4-3	Capacity terminology	266
4-4	Pool size by extent size and I/O group number	274
4-5	Minimum recommended pool size by extent size and I/O group number	274
4-6	DRP capacity terms	277
4-7	Easy Tier settings	303
4-8	Recommended 3-tier Easy Tier mapping policy	307
4-9	Four- and five-tier mapping policy	308
4-10	Unsupported temporary four- and five-tier mapping policy	308
4-11	Migration target tier priorities	311
5-1	Migration types and associated commands	342
5-2	Sample syncrate values	345
6-1	Relationship between the rate and data rate per second	368
6-2	Summary table of the FlashCopy indirection layer algorithm	377
6-3	FlashCopy properties and maximum configurations	381
6-4	Relationship of bitmap space to FlashCopy address space for the I/O group	382
6-5	Workload distribution for back-end I/O operations	388
6-6	Maximum round trip	418
6-7	IBM Storage Virtualize inter-system heartbeat traffic (Mbps)	418
6-8	Remote copy maximum limits	430
6-9	Configuration limits for clustering and HyperSwap over FC and Ethernet	431
6-10	Relative Peak I/O Response Time with number of relationships per CG	435
6-11	IP replication limits	461
6-12	Relationship between the rate value and the data copied per second	481
6-13	Relationship of bitmap space to volume mirroring address space	484
6-14	Supported platforms	492
6-15	Bitmap memory per I/O group	493
7-1	Business continuity solutions comparison	510
8-1	100 GbE adapter functions	546
8-2	Maximum 100-GbE adapters per node and PCIe slot placement	547
8-3	Cables	549
9-1	Low space warning percentages for compressed arrays	594
9-2	Event examples for SAN Volume Controller	607
9-3	Alert severities	607
9-4	VDisk and volume metrics (front end)	620
9-5	MDisk and drive metrics	623
9-6	Node metrics	625
9-7	Port metrics	626
9-8	Miscellaneous metrics	627
9-9	Field changes to drive and array devices	636
10-1	UNIX commands available in the CLI	640
10-2	Available memory configurations for one node in a control enclosure	677
10-3	IBM FlashSystem 9500 control enclosure adapter options	678
10-4	IBM FlashSystem 7300 control enclosure adapter options	678
10-5	IBM FlashSystem 5000 family standard configurations	679
10-6	IBM FlashSystem 5000 family adapters	679
10-7	Files that are created by the backup process	695

11-1 Useful AIX lspath commands	716
11-2 Useful AIX lspio commands	716
11-3 Useful Windows mpclaim.exe commands.	717
11-4 Useful Windows PowerShell cmdlets	718
11-5 Selected attributes of the lspostats output	731
A-1 Comparing IBM i native and Virtual I/O Server attachment	775
A-2 Limits increased for maximum disk arms and LUN sizes	784
A-3 Throughput of FC adapters	786
A-4 LUN-level switching IBM PowerHA SystemMirror for i editions	799

Examples

3-1 Manual FCM format	202
3-2 FCM capacity monitoring with CLI	202
3-3 Array capacity monitoring with the CLI	214
3-4 Round-robin enabled storage controller on IBM FlashSystem 9100.	216
3-5 MDisk group balanced path selection (no round-robin enabled) storage controller . .	217
3-6 Command output for the lsarray and lsrank commands	225
3-7 Output of the showvolgrp command	228
3-8 Output for the lshostconnect command	228
3-9 Output of the lscontroller command on IBM FlashSystem	229
3-10 The lsquorum command on IBM FlashSystem 9100.	241
4-1 Results of capacity savings analysis.	255
4-2 DRET command-line interface	257
4-3 Listing of volumes that have extents on an MDisk to be deleted	281
4-4 DS8000 UID example	282
4-5 The lscontroller command.	282
4-6 Command to declare or identify a self-encrypted MDisk from a virtualized external storage	290
4-7 Manually triggered replication.	294
4-8 Verifying the key state.	295
4-9 Commands to enable the key server encryption option on a system upgraded from pre-7.8.0.	296
4-10 Changing an MDisk tier.	309
4-11 Changing Easy Tier settings.	317
5-1 Volume creation without the auto-formatting option	322
5-2 Creating a thin-provisioned volume	327
5-3 Creating a thin-provisioned volume with the deduplication option	329
5-4 Creating a compressed volume with the deduplication option	330
5-5 The migratevdisk command	343
5-6 Monitoring the migration process	343
5-7 Throttle commands example.	355
5-8 Changing the cache mode of a volume	358
6-1 Changing gmlinktolerance to 30 and gmmxhostdelay to 100.	440
6-2 Changing rbuffersize to 64 MB	440
6-3 The maxreplicationdelay and partnershipexclusionthreshold settings	447
6-4 A lsiogrp and chiogrp command example.	484
8-1 The lshostvdiskmap command	523
8-2 Output of using the lshostvdiskmap command	523
8-3 Output of using the lsvdiskhostmap command	524
8-4 The lspportset command	530
8-5 Output of using the lspip command	531
8-6 The lstargerportfc command.	532
8-7 Determining the Symantec Veritas server configuration	542
8-8 Symantec Volume Manager that uses SDD pass-through mode ASL	543
8-9 IBM Storage Virtualize that is configured by using native ASL.	543
9-1 Sample SNMP trap output from snmptrapd	554
9-2 rsyslogd concise output showing audit, login, and authentication events.	555
9-3 Threshold specified as a size	556
9-4 Threshold that is specified as a value.	556

9-5 Latency reported in milliseconds (ms) with microsecond (μ s) granularity	559
9-6 REST API clients	560
9-7 Command to add the SI tenant ID through the CLI.	579
9-8 CMMVC9305E	579
9-9 Getting the value filtered from lssystem	590
9-10 Physical_capacity and physical_free_capacity from lssystem command	590
9-11 Physical_capacity from lssystem command	590
9-12 Deduplication and compression savings and used capacity.	591
9-13 CLI output example for lsportstats command to show the TX & RX power	616
9-14 CLI example to change the interval	629
10-1 The lssystem command	646
10-2 Copying the upgrade test utility to IBM Storage Virtualize	654
10-3 Ensure that files are uploaded	655
10-4 Upgrade test by using the CLI	655
10-5 Verify software version	658
10-6 Installing a patch on all nodes.	659
10-7 Verify patch installation using lsservicestatus.	660
10-8 Removing patch on single nodes	661
10-9 Removing all patches from a node in the system	661
10-10 Listing the firmware level for drives 0, 1, 2, and 3.	663
10-11 Output of the pcmpath query WWPN command.	667
10-12 Output of the lshost <hostname> command.	668
10-13 Cross-referencing information with SAN switches	668
10-14 Mapping the host to I/O groups.	672
10-15 Creating a throttle by using the mkthrottle command in the CLI	685
10-16 Sample svcconfig backup command output	694
10-17 Saving the configuration backup files to your workstation	695
11-1 The svc_livedump command	714
11-2 preplivedump and lslivedump commands.	715
11-3 Output for the multipath -ll command	715
11-4 Output of “esxcli storage core path” list command	718
11-5 Output of esxcli storage core path list -d <naaID>	719
11-6 Output for esxcli storage nmp device list	720
11-7 The triggerdrivedump command	721
11-8 The lshost command.	723
11-9 The lshost <host_id_or_name> command	723
11-10 The lsfabric -host <host_id_or_name> command.	724
11-11 Incorrect WWPN zoning	730
11-12 Correct WWPN zoning	730
11-13 lsportstats command output	730
11-14 Issuing a lsmdisk command	732
11-15 Output of the svcinfo lscontroller command	734
11-16 Determining the ID for the MDisk	736

Notices

This information was developed for products and services offered in the US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing, IBM Corporation, North Castle Drive, MD-NC119, Armonk, NY 10504-1785, US

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you provide in any way it believes appropriate without incurring any obligation to you.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to actual people or business enterprises is entirely coincidental.


COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at “Copyright and trademark information” at <https://www.ibm.com/legal/copytrade.shtml>

The following terms are trademarks or registered trademarks of International Business Machines Corporation, and might also be trademarks or registered trademarks in other countries.

AIX®	IBM®	Interconnect®
Db2®	IBM Cloud®	PowerHA®
DS8000®	IBM FlashCore®	Redbooks®
Easy Tier®	IBM FlashSystem®	Redbooks (logo)  ®
FICON®	IBM Research®	Service Request Manager®
FlashCopy®	IBM Security®	Storwize®
Guardium®	IBM Spectrum®	Tivoli®
HyperSwap®	IBM Z®	XIV®

The following terms are trademarks of other companies:

Intel, Intel Xeon, Intel logo, Intel Inside logo, and Intel Centrino logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

ITIL is a Registered Trade Mark of AXELOS Limited.

The registered trademark Linux® is used pursuant to a sublicense from the Linux Foundation, the exclusive licensee of Linus Torvalds, owner of the mark on a worldwide basis.

Microsoft, Windows, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Java, and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Ansible, OpenShift, Red Hat, are trademarks or registered trademarks of Red Hat, Inc. or its subsidiaries in the United States and other countries.

UNIX is a registered trademark of The Open Group in the United States and other countries.

VMware, VMware vCenter Server, VMware vSphere, and the VMware logo are registered trademarks or trademarks of VMware, Inc. or its subsidiaries in the United States and/or other jurisdictions.

Other company, product, or service names may be trademarks or service marks of others.

Preface

This IBM® Redbooks® publication captures several best practices and describes the performance gains that can be achieved by implementing the IBM Storage FlashSystem and IBM SAN Volume Controller (SVC) products running IBM Storage Virtualize 8.6. These practices are based on field experience.

This book highlights configuration guidelines and best practices for the storage area network (SAN) topology, clustered system, back-end storage, storage pools and managed disks (MDisks), volumes, Remote Copy services, and hosts.

This book is intended for experienced storage, SAN, IBM Storage FlashSystem, and SVC administrators and technicians. Understanding this book requires advanced knowledge of these environments.

Demonstration videos: For a full list of IBM Storage demonstration videos created by the Redbooks authors team, see [Redbooks Demo videos](#).

IBM Storage rebranding: In January 2023, IBM announced that it would be renaming its Spectrum software-defined storage products to IBM Storage products. This change was made to simplify the product portfolio and make it easier for customers to find the products they need. For example, IBM Spectrum Virtualize was renamed to IBM Storage Virtualize. You will likely find documentation under both the Spectrum and Storage names for some time, as the transition to the new names will take place over a period of several months. However, all future documentation will be under the IBM Storage name.

Authors

This book was produced by a team of specialists from around the world.



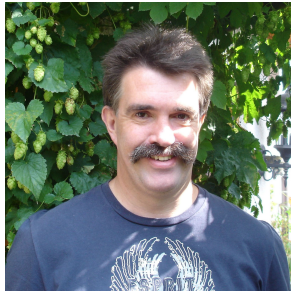
Bernhard Albrecht is a Storage Advisory Partner Technical Specialist for DACH. Bernd has been with IBM in various technical sales roles now for 32 years. He was for example 3 years OEM technical business development storage for Lenovo in IBM Germany. He works at the IBM Germany in Berlin and has over 22 years of storage technical sales experience. His focus is on the IBM FlashSystem® Family and the IBM SAN Volume Controller. His experience in this area goes back to the beginning of these products in 2003. Bernd is a published IBM Redbooks author.



Christian Schroeder provides support with passion to clients worldwide, covering IBM storage products of the Storage Virtualize and FlashSystem products family. He has been with IBM for more than 23 years, working in product support for various platforms as IBM System x servers & BladeCenter, SAN switches, IBM SAN Volume Controller, and IBM FlashSystem. Christian has co-authored IBM Redbooks publications that cover IBM System x servers and SAN Volume Controller V6.3.0 and V8.4.2.



Denis Olshanskiy is a Storage Specialist with a demonstrated history of working in the IT and services industry. His areas of expertise include storage area networks (SANs), data centers, storage solutions, Arduino, and Linux. He has a master's degree in Mechatronics, Robotics, and Automation Engineering from Budapest University of Technology and Economics.



Hartmut Lonzer is Storage Advisory Partner Technical Specialist for DACH. Before this position, he was OEM Alliance Manager for Lenovo in IBM Germany. He works at the IBM Germany headquarter in Ehningen. His main focus is on the IBM FlashSystem Family and the IBM SAN Volume Controller. His experience with the IBM SAN Volume Controller and IBM FlashSystem products goes back to the beginning of these products. Hartmut has been with IBM in various technical and sales roles now for 45 years.



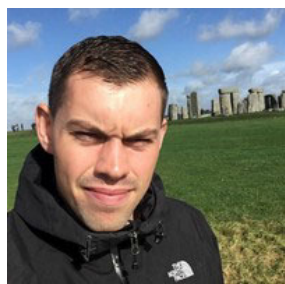
John Nycz is an Advanced Subject Matter Expert for IBM Storage Virtualize and IBM FlashSystem. He has more than 10 years of experience in the areas of systems management, networking hardware, and software. John has been with IBM for more than 20 years and has been a member of numerous development, project management, and support teams. For the last seven years, he has been member of IBM's Storage Virtualization Support Team.



Jon Herd is an IBM Senior Executive Advocate working for the TLS EMEA Remote Technical Support and Client Care team based in IBM Germany. He covers the United Kingdom, Ireland and beyond, advising customers on a portfolio of IBM storage products, including FlashSystem products. He also works as a senior advisor to the TLS EMEA RTS/CC management on new products, strategy and new technologies that might affect the TLS business. Jon has been with IBM for more than 45 years, and has held various technical roles, including Europe, Middle East, and Africa (EMEA) level 2 support on mainframe servers and technical education development. He has written many IBM Redbooks on the FlashSystems products and is a IBM Redbooks Platinum level author. He holds IBM certifications in Product Services profession at a thought leader L3 level, and is a Technical Specialist at an experienced L1 level. He also is a certified Chartered Member of the British Computer Society (MBCS - CITP), a Certified Member of the Institution of Engineering and Technology (MIET), and a Certified Technical Specialist of the Open Group (TOG).



Konrad Trojok has been responsible for the technical team lead for the IBM Storage team at SVA since 2011. The role includes an active part in the daily IBM storage business, including design, implementation, and care of storage solutions. In the partnership with IBM, Konrad also support product development and design. The beginning of his IT career included IBM Power solutions around SP systems and SSA storage. Konrad switched his technical focus with the emerging of SAN and SAN storage.



Nils Olsson is a Subject Matter Expert (previously Product Engineering Professional - Level 2 Support) working in the EMEA Storage Competence Center (ESCC) in Kelsterbach, IBM Germany. He provides remote technical support for IBM Storage Virtualize Products in Europe and worldwide. Nils joined IBM in 2008 and is skilled in SAN, storage, and storage virtualization. He has over 10 years of experience with troubleshooting IBM Storage Virtualize and is certified on the IBM Storwize® portfolio. In his current role, he delivers analysis of complex field issues and provides technical expertise during critical situations, collaborating closely with development



Patrik Groß is a Solution Sales Architect for storage solutions at CANCOM GmbH in Cologne since 2005, having also worked as IT Manager at the customer since 1992. His expertise resides in IBM's Power and Storage products, where he is active in both pre- and post-sales. He supports customers in proof-of-concepts to develop customized solutions and helps troubleshoot and optimize larger storage environments. He manages CANCOM's own IBM demo data center, where he also supports IBM beta programs. His current focus is on IBM Virtualize, Tape, COS and Fusion.



Rene Wuellenweber is an experienced Product Services Professional working in Chemnitz, IBM Germany. Rene has been with IBM for more than 24 years and is working with IBM Storage Virtualize and IBM Flashsystem since 2014. He has experience as a customer engineer and has background in the area of midrange DASD products, archive solution and n-series. In his current job role he is working as Remote Technical Support Engineer for IBM Storage Virtualize and IBM Flashsystem to support customers worldwide.



Sergey Kubin is a Senior Storage Support Engineer working in GBM Qatar. He holds an Electronics Engineer degree from Ural Federal University in Russia and has more than 15 years of experience in IT. In GBM, he provides support and guidance for customers using IBM and multi-vendor storage solutions. His expertise includes file and block storage, and storage area networks.



Vasfi Gucer is a project leader with the IBM Redbooks Team. He has more than 20 years of experience in the areas of systems management, networking hardware, and software. He writes extensively and teaches IBM classes worldwide about IBM products. His focus has been primarily on storage and cloud computing for the last eight years. Vasfi is also an IBM Certified Senior IT Specialist, Project Management Professional (PMP), IT Infrastructure Library (ITIL) V2 Manager, and ITIL V3 Expert.



Vineet Sharma is a Master Certified Technical Specialists in Storage Systems and has expertise several storage system technology areas in IBM Expert Labs, MEA. He works on planning, configuration, and implementation of the IBM Storage Virtualize family and IBM Enterprise Storage(DS8000®). He has delivered many complex solutions and complex data migrations for block storages. He actively works with IBM Business Partners and IBM clients for technology enablement and storage bootcamps. He has over 15 years of experience in IBM storage products such as IBM DS4000 series, IBM Storwize Series, IBM SVC, IBM FlashSystems, IBM DS8000, IBM A9000 and IBM Software products like IBM Spectrum® Control, IBM Copy Services Manager and Storage Insights.

Thanks to the authors of the previous edition of this book: *Performance and Best Practices Guide for IBM Spectrum Virtualize 8.5*, SG24-8521, published on 09 August 2022:

Andy Haerchen, Ashutosh Pathak, Barry Whyte, Cassio Alexandre de Aguiar, Fabio Trevizan de Oliveira, Luis Eduardo Silva Viera, Mahendra S Brahmada, Mangesh M Shirke, Nezih Boyacioglu, Stephen Solewin, Thales Ferreira, Uwe Schreiber and Youssef Largou

Thanks to the following people for their contributions to this project:

Jamie Roszel, Content Editor
IBM Redbooks, RTP

Alex Goodson
IBM UK

Now you can become a published author, too!

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time! Join an IBM Redbooks residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

Find out more about the residency program, browse the residency index, and apply online at:

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

We want our books to be as helpful as possible. Send us your comments about this book or other IBM Redbooks publications in one of the following ways:

- ▶ Use the online **Contact us** review Redbooks form found at:

ibm.com/redbooks

- ▶ Send your comments in an email to:

redbooks@us.ibm.com

- ▶ Mail your comments to:

IBM Corporation, IBM Redbooks
Dept. HYTD Mail Station P099
2455 South Road
Poughkeepsie, NY 12601-5400

Stay connected to IBM Redbooks

- ▶ Find us on LinkedIn:

<https://www.linkedin.com/groups/2130806>

- ▶ Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:

<https://www.redbooks.ibm.com/subscribe>

- ▶ Stay current on recent Redbooks publications with RSS Feeds:

<https://www.redbooks.ibm.com/rss.html>



Introduction and system overview

This chapter defines the concept of storage virtualization and provides an overview of its application in addressing the challenges of the modern storage environment. It also contains an overview of each product that makes up the IBM SAN Volume Controller and IBM Storage FlashSystem families.

This chapter includes the following topics:

- ▶ “IBM Storage Virtualize” on page 2
- ▶ “IBM SAN Volume Controller architectural overview” on page 8
- ▶ “Latest changes and enhancements” on page 13
- ▶ “IBM SAN Volume Controller family” on page 24
- ▶ “IBM SAN Volume Controller models” on page 44
- ▶ “IBM FlashSystem family” on page 49
- ▶ “IBM FlashSystem 9500 overview” on page 55
- ▶ “IBM FlashSystem 9500R Rack Solution overview” on page 61
- ▶ “IBM FlashSystem 9000 Expansion Enclosure Models AFF and A9F” on page 64
- ▶ “IBM FlashSystem 7300 overview” on page 69
- ▶ “IBM FlashSystem 5200 overview” on page 74
- ▶ “IBM FlashSystem 5000 family overview” on page 78
- ▶ “Storage efficiency and data reduction features” on page 88
- ▶ “Application integration features” on page 94
- ▶ “Features for manageability” on page 95
- ▶ “Copy services” on page 100
- ▶ “IBM FlashCore Module drives, NVMe SSDs, and SCM drives” on page 105
- ▶ “Storage virtualization” on page 108
- ▶ “Business continuity” on page 111
- ▶ “Management and support tools” on page 117
- ▶ “Licensing” on page 120

1.1 IBM Storage Virtualize

IBM Storage Virtualize (previously known as IBM Spectrum Virtualize) is a key member of the IBM Storage portfolio. It is a highly flexible storage solution that enables rapid deployment of block storage including storage virtualization, for new and traditional workloads, on-premises, off-premises, or a combination of both.

Note: For more information, see [IBM Storage FlashSystem](#) and [IBM SAN Volume Controller](#).

With the introduction of the IBM Storage family, the *software* that runs on IBM SAN Volume Controller and on IBM Storage FlashSystem (IBM FlashSystem) products is called *IBM Storage Virtualize*. The name of the underlying *hardware* platform remains intact.

IBM FlashSystem storage systems and IBM SAN Volume Controllers are built with award-winning IBM Storage Virtualize software that simplifies infrastructure and eliminates the differences in management, function, and even hybrid multicloud support.

IBM Storage Virtualize is an offering that has been available for years for the IBM SAN Volume Controller and IBM FlashSystem family of storage solutions. The solution offers efficient management and robust protection for massive datasets originating from mobile and social apps, streamlines quick and adaptable cloud service implementations, and ensures peak performance and scalable capabilities essential for deriving valuable insights from the most recent big data analytics tools.

Note: This edition focuses on systems compatible with IBM Storage Virtualize 8.6. Note that certain products mentioned within may have been discontinued by IBM (End of Marketing—EOM), yet they continue to support the 8.6 software. Where this is applicable, it will be mentioned in the text.

Table 1-1 shows the IBM Storage Virtualize 8.6 supported product list and whether the product is still currently sold or has reached EOM.

Table 1-1 IBM Storage Virtualize 8.6 supported product list

Product	Machine Type	Model	Comment
FS9500/R	4666, 4983	AH8, UH8	Current Product
FS7300	4657	924, U7D	Current Product
FS5200	4662	6H2,UH6	Current Product
FS5000 (FS5015, FS5035)	2072, 4680	2N2, 2N4, 3N2, 3N4	Current Product
FS5000 (FS5045)	4680	3P2, 3P4	Current Product
SVC	2145, 2147	SA2, SV3	Current Product
SVC	2145, 2147	SV2	EOM 01/2023
FS9200	9846, 9848, 4666	AG8, UG8	EOM 07/2022
FS7200	2076, 4664	824, U7C	EOM 07/2022
FS9100	9846, 9848	AF8, UF8	EOM 07/2022

Benefits of IBM Storage Virtualize

IBM Storage Virtualize delivers leading benefits that improve storage infrastructure in many ways, including the following examples:

- ▶ Reduces the cost of storing data by increasing the use and accelerating applications to speed business insights. To achieve this goal, the solution:
 - Uses data reduction technologies to increase the amount of data that you can store in the same space.
 - Enables rapid deployment of cloud storage for disaster recovery (DR) along with the ability to store copies of local data.
 - Moves data to the most suitable type of storage based on policies that you define by using IBM Storage Control to optimize storage.
 - Improves storage migration performance so that you can do more with your data.
- ▶ Protects data from theft or inappropriate disclosure while enabling a high availability (HA) strategy that includes protection for data and application mobility and DR. To achieve this goal, the solution:
 - Uses software-based encryption to improve data security.
 - Provides fully duplexed copies of data and automatic switchover across data centers to improve data availability.
 - Eliminates storage downtime with nondisruptive movement of data from one type of storage to another type. Simplifies data by providing a data strategy that is independent of your choice of infrastructure, which delivers tightly integrated functions and consistent management across heterogeneous storage. To achieve this goal, the solution:
 - Integrates with virtualization tools, such as VMware vCenter to improve agility with automated provisioning of storage and easy deployment of new storage technologies.
 - Enables supported storage to be deployed with Kubernetes and Docker container environments, including Red Hat OpenShift.
 - Consolidates storage, regardless of the hardware vendor for simplified management, consistent functions, and greater efficiency.
 - Supports common capabilities across storage types, which provide flexibility in storage acquisition by allowing a mix of vendors in the storage infrastructure.

Note: These benefits are not a complete list of features and functions that are available with IBM Storage Virtualize software.

1.1.1 Storage virtualization terminology

Storage virtualization is a term that is used extensively throughout the storage industry. It can be applied to various technologies and underlying capabilities. In reality, most storage devices technically can claim to be virtualized in one form or another. Therefore, this chapter starts by defining the concept of storage virtualization as it is used in this book.

We describe storage virtualization in the following ways:

- ▶ Storage virtualization is a technology that allows different storage resources with different underlying capabilities to be managed and consumed using a common set of interfaces.

- ▶ Storage virtualization is a logical representation of resources that is not constrained by physical limitations and hides part of the complexity. It also adds or integrates new functions with services and can be nested or applied to multiple layers of a system.

The virtualization model consists of the following layers:

- ▶ Application: The user of the storage domain.
- ▶ Storage domain:
 - File, record, and namespace virtualization and file and record subsystem
 - Block virtualization
 - Block subsystem

Applications typically read and write data as vectors of bytes or records. However, storage presents data as vectors of blocks of a constant size (512 or in the newer devices, 4096 bytes per block).

The *file, record, and namespace virtualization* and *file and record subsystem* layers convert records or files that are required by applications to vectors of blocks, which are the language of the *block virtualization* layer. The block virtualization layer maps requests of the higher layers to physical storage blocks, which are provided by *storage devices* in the *block subsystem*.

Each of the layers in the storage domain abstracts away the complexities of the lower layers and hides them behind an easy to use, standard interface that is presented to upper layers. The resultant decoupling of logical storage space representation and its characteristics that are visible to servers (storage consumers) from underlying complexities and intricacies of storage devices is a key concept of storage virtualization.

The focus of this publication is *block-level virtualization* at the *block virtualization layer*, which is implemented by IBM as IBM Storage Virtualize software that is running on IBM SAN Volume Controller and the IBM FlashSystem family. The IBM SAN Volume Controller is implemented as a clustered appliance in the storage network layer. The IBM FlashSystems are deployed as modular storage systems that can virtualize their internally and externally attached storage.

IBM Storage Virtualize uses the Small Computer System Interface (SCSI) protocol to communicate with its clients. It presents storage space as SCSI logical units (LUs), which are identified by SCSI logical unit numbers (LUNs).

Note: Although LUs and LUNs are different entities, the term *LUN* in practice often is used to refer to a logical disk; that is, an *LU*.

Most applications do not directly access storage; instead, they work with files or records using a file system. However, the operating system of a host must convert these abstractions to the language of storage, which is vectors of storage blocks that are identified by logical block addresses (LBAs) within an LU.

Inside IBM Storage Virtualize, each of the externally visible LUs is internally represented by a volume, which is an amount of storage that is taken out of a storage pool. Storage pools are made of managed disks (MDisks), LUs that are presented to the storage system by external virtualized storage or arrays that consist of internal disks. LUs that are presented to IBM Storage Virtualize by external storage often correspond to RAID arrays that are configured on that storage.

The hierarchy of objects, from a file system block down to a physical block on a physical drive, is shown in Figure 1-1.

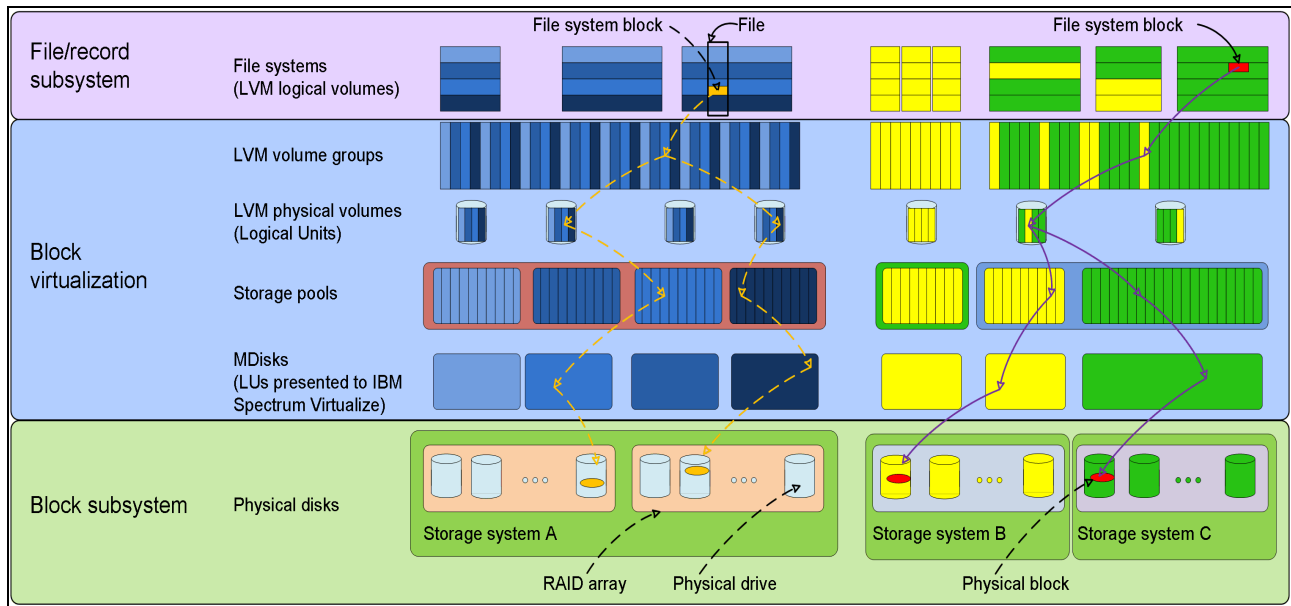


Figure 1-1 Block-level virtualization overview

With storage virtualization, you can manage the mapping between logical blocks within an LU that is presented to a host, and blocks on physical drives. This mapping can be as simple or as complicated as required by a use case. A logical block can be mapped to one physical block or for increased availability, multiple blocks that are physically stored on different physical storage systems, and in different geographical locations.

IBM Storage Virtualize utilizes the concept of an extent, which is a group of physical blocks, as a management construct to allow great flexibility in the utilization of available storage.

Importantly, the mapping can be dynamic: With IBM Easy Tier®, IBM Storage Virtualize can automatically change underlying storage to which an extent is mapped to better match a host's performance requirements with the capabilities of the underlying storage systems.

IBM Storage Virtualize gives a storage administrator a wide range of options to modify volume characteristics: from volume resize to mirroring, creating a point-in-time (PiT) copy with IBM FlashCopy®, and migrating data across physical storage systems.

Importantly, all the functions that are presented to the storage users are independent from the characteristics of the physical devices that are used to store data. This decoupling of the storage feature set from the underlying hardware and ability to present a single, uniform interface to storage users that masks underlying system complexity is a powerful argument for adopting storage virtualization with IBM Storage Virtualize.

Storage virtualization is implemented on many layers. Figure 1-1 shows an example in which a file system block is mirrored by the host's operating system by using features of the logical volume manager (LVM) or the IBM Storage Virtualize system at the storage pool level.

Although the result is similar (the data block is written to two different arrays), the effort that is required for per-host configuration is disproportionately larger than for a centralized solution with organization-wide storage virtualization that is done on a dedicated system and managed from a single GUI.

IBM Storage Virtualize includes the following key features:

- ▶ Simplified storage management by providing a single management interface for multiple storage systems and a consistent user interface for provisioning heterogeneous storage.
- ▶ Online volume migration: IBM Storage Virtualize enables movement of data from one set of physical drives to another in a way that is not apparent to the storage users and without over-straining the storage infrastructure. The migration can be done within a specific storage system (from one set of disks to another set) or across storage systems. Either way, the host that uses the storage is not aware of the operation, and no downtime for applications is needed.
- ▶ Enterprise-level Copy Services functions: Performing Copy Services functions within IBM Storage Virtualize removes dependencies on the capabilities and interoperability of the virtualized storage subsystems. Therefore, it enables the source and target copies to be on any two virtualized storage subsystems.
- ▶ Safeguarded Copy mechanism: This uses FlashCopy functions to take immutable copies and aids in the recovery from ransomware or customer internal actions that might result in loss of data.
- ▶ Inline Data Corruption Detection: A new feature that uses artificial intelligence (AI) and machine learning (ML) services to detect data changes that can be indicative of threats or direct attacks on your data sets in near real-time.
- ▶ Improved storage space usage because of the pooling of resources across virtualized storage systems.
- ▶ Opportunity to improve system performance as a result of volume striping across multiple virtualized arrays or controllers, and the benefits of cache that is provided by IBM Storage Virtualize hardware.
- ▶ Improved data security by using data-at-rest encryption.
- ▶ Data replication, including replication to cloud storage by using advanced copy services for data migration and backup solutions.
- ▶ Data reduction techniques, such as deduplication and compression for space efficiency, which store more data on the storage that is available. IBM Storage Virtualize can enable significant savings that increase the effective capacity of storage systems up to 5x, and decrease the floor space, power, and cooling that are required by the storage system.

Note: IBM Real-time Compression (RtC) is only available for earlier generation engines. The newer IBM SAN Volume Controller engines (SV3, SV2, and SA2) do *not* support RtC; however, they support software compression through Data Reduction Pools (DRP).

Summary

Storage virtualization is a fundamental technology that enables the realization of flexible and reliable storage solutions. It helps enterprises to better align IT architecture with business requirements, simplify their storage administration, and facilitate their IT departments efforts to meet business demands.

IBM Storage Virtualize running on IBM SAN Volume Controller and the IBM FlashSystem family is a mature, 12th-generation virtualization solution that uses open standards and complies with the Storage Networking Industry Association (SNIA) storage model. All products use in-band block virtualization engines that move the control logic (including advanced storage functions) from a multitude of individual storage devices to a centralized entity in the storage network.

IBM Storage Virtualize can improve the use of your storage resources, simplify storage management, and improve the availability of business applications.

Figure 1-2 shows the feature set that is provided by the IBM Storage Virtualize systems.

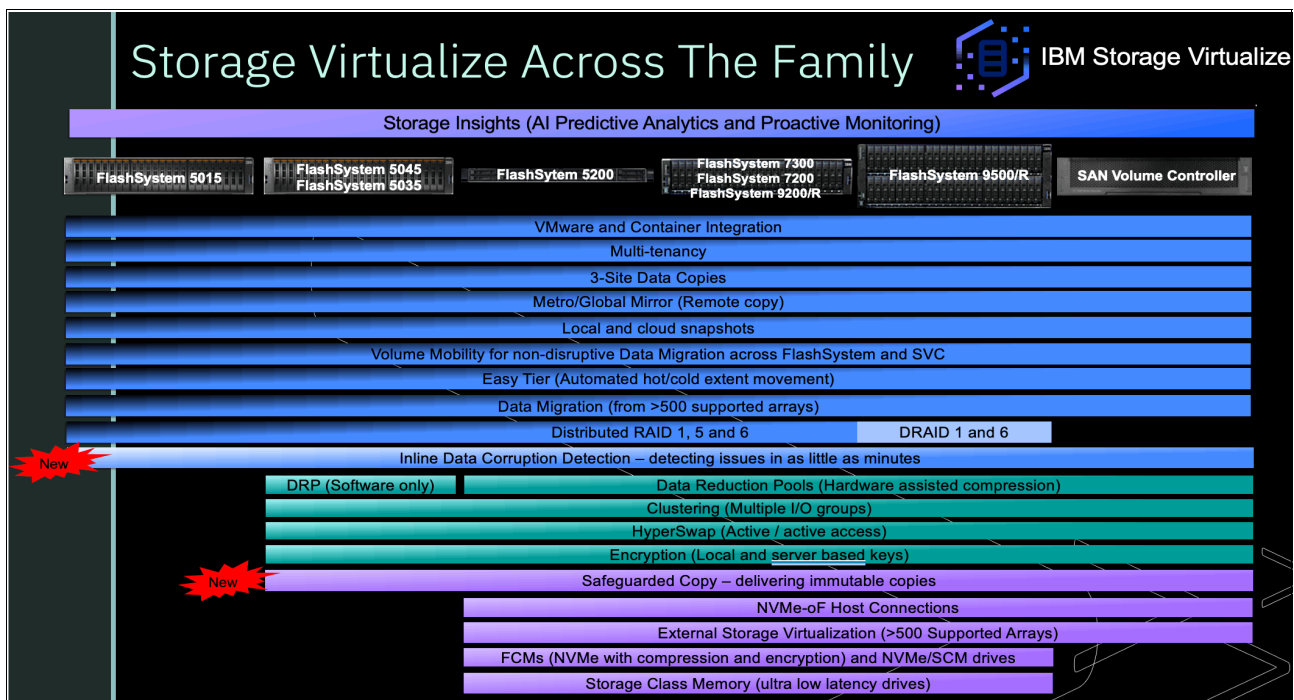


Figure 1-2 IBM Storage Virtualize features by product

1.2 IBM SAN Volume Controller architectural overview

IBM SAN Volume Controller is a SAN block aggregation virtualization appliance that is designed for attachment to various host computer systems.

The following major approaches are used today for the implementation of block-level aggregation and virtualization:

► **Symmetric: In-band appliance**

Virtualization splits the storage that is presented by the storage systems into smaller chunks that are known as *extents*. These extents are then concatenated by using various policies to make virtual disks or *volumes*. With symmetric virtualization, host systems can be isolated from the physical storage. Advanced functions, such as data migration, can run without reconfiguring the host.

With symmetric virtualization, the *virtualization engine* is the central configuration point for the SAN. The virtualization engine directly controls access to the storage and data that is written to the storage. As a result, locking functions that provide data integrity and advanced functions (such as cache and Copy Services) can be run in the virtualization engine. Therefore, the virtualization engine is a central point of control for device and advanced function management.

Symmetric virtualization includes some disadvantages, mainly scalability. Scalability can cause poor performance because all input/output (I/O) must flow through the virtualization engine. To solve this problem, you can use an *n*-way cluster of virtualization engines that includes failover capability.

You can scale the extra processor power, cache memory, and adapter bandwidth to achieve your preferred level of performance. More memory and processing power are needed to run advanced services, such as Copy Services and caching. IBM SAN Volume Controller uses symmetric virtualization. Single virtualization engines, which are known as *nodes*, are combined to create clusters. Each cluster can contain 2 - 8 nodes.

► **Asymmetric: Out-of-band or controller-based**

With asymmetric virtualization, the virtualization engine is outside the data path and performs a metadata-style service. The metadata server contains all of the mapping and the locking tables, and the storage devices contain only data. In asymmetric virtual storage networks, the data flow is separated from the control flow.

A separate network or SAN link is used for control purposes. Because the control flow is separated from the data flow, I/O operations can use the full bandwidth of the SAN. A separate network or SAN link is used for control purposes.

Asymmetric virtualization can have the following disadvantages:

- Data is at risk to increased security exposures, and the control network must be protected with a firewall.
- Metadata can become complicated when files are distributed across several devices.
- Each host that accesses the SAN must know how to access and interpret the metadata. Therefore, specific device drivers or agent software must be running on each of these hosts.
- The metadata server cannot run advanced functions, such as caching or Copy Services, because it only “knows” about the metadata and not the data.

Figure 1-3 shows variations of the two virtualization approaches.

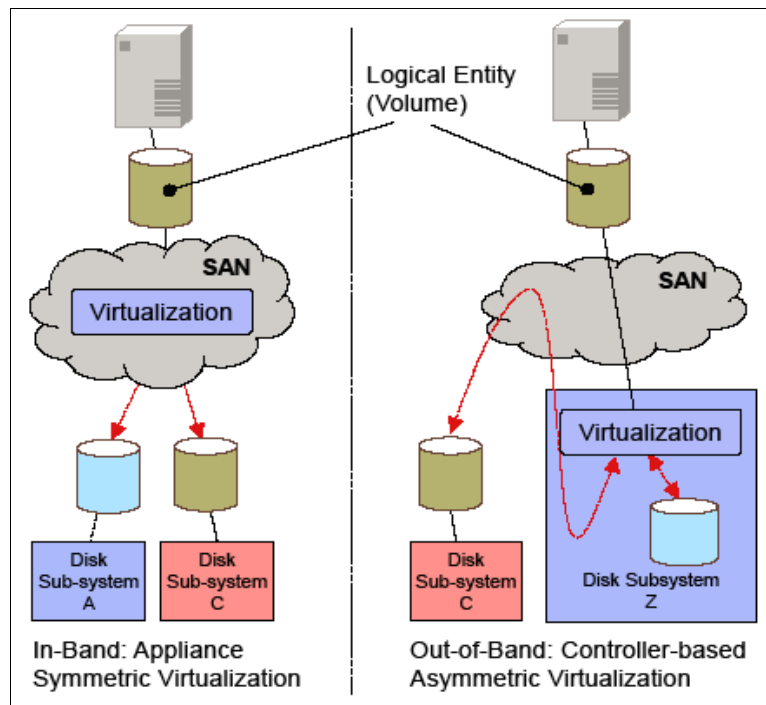


Figure 1-3 Overview of block-level virtualization architectures

Although these approaches provide essentially the same cornerstones of virtualization, interesting side-effects can occur.

1.2.1 Controller-based approach

The out-of-band, controller-based approach includes high functions, but it fails in terms of scalability or upgradeability. Because of the nature of its design, no true decoupling occurs by using this approach, which becomes an issue for the life cycle of this solution, such as with a controller. Data migration issues and questions are challenging, such as how to reconnect the servers to the new controller, and how to reconnect them online without any effect on your applications.

By using this approach, you replace a controller and implicitly replace your entire virtualization solution. In addition to replacing the hardware, other actions (such as updating or repurchasing the licenses for the virtualization feature, and advanced copy functions) might be necessary.

1.2.2 SAN or fabric-based appliance solution

With an in-band, SAN or fabric-based appliance solution that is based on a scale-out cluster architecture, life-cycle management tasks, such as adding or replacing new disk subsystems or migrating data between them, are simple.

Servers and applications remain online, data migration occurs transparently on the virtualization platform, and licenses for virtualization and copy services require no update. No other costs are incurred when disk subsystems are replaced.

Only the fabric-based appliance solution provides an independent and scalable virtualization platform that can provide enterprise-class Copy Services that is open for future interfaces and protocols. By using the fabric-based appliance solution, you can choose the disk subsystems that best fit your requirements, and you are not locked into specific SAN hardware.

For these reasons, IBM chose the SAN-based appliance approach with inline block aggregation for the implementation of storage virtualization with IBM Storage Virtualize.

1.2.3 IBM SAN Volume Controller

IBM SAN Volume Controller includes the following key characteristics:

- ▶ It is highly scalable, which provides an easy growth path from two to eight (grow in a pair of nodes to allow for data protection and reliability).
- ▶ It is SAN interface-independent. It supports connections through Fibre Channel, NVM Express (NVMe) over Fibre Channel (FC-NVMe), NVM Express (NVMe) over RDMA, or a TCP network. Consider the following points:
 - FC-NVMe connections require 4-port 16 GB or 32 GB Fibre Channel adapters that must be purchased.
 - NVMe over RDMA connections support a 25 Gbps RoCE adapter or 100 Gbps RoCE adapter that must be purchased.
 - The IBM SAN Volume Controller 2145-SV3 and 2147-SV3 model support 100 Gbps adapter only.
- ▶ It is host independent for fixed block-based Open Systems environments.
- ▶ It is external storage system independent, which provides a continuous and ongoing process to qualify more types of storage systems.
- ▶ Some nodes can use internal disks (flash drives) or externally direct-attached disks in expansion enclosures.

On the SAN storage that is provided by the disk subsystems, IBM SAN Volume Controller offers the following services:

- Creates a single pool of storage.
- Provides LU virtualization.
- Manages logical volumes.
- Mirrors logical volumes.

IBM SAN Volume Controller running IBM Storage Virtualize 8.6 also provides the following functions:

- ▶ Large scalable cache
- ▶ Copy Services
- ▶ IBM FlashCopy (PiT copy) function, including thin-provisioned FlashCopy to make multiple targets affordable
- ▶ IBM Transparent Cloud Tiering (TCT) function that enables IBM SAN Volume Controller to interact with cloud service providers (CSPs)
- ▶ Metro Mirror (MM), which is a synchronous copy
- ▶ Global Mirror (GM), which is an asynchronous copy
- ▶ Policy-based replication
- ▶ Safeguarded Copy
- ▶ Inline Data Corruption Detection

- ▶ Data migration
- ▶ Storage space efficiency (thin-provisioning, compression, and deduplication)
- ▶ IBM Easy Tier to automatically migrate data between storage types of different performance that is based on disk workload
- ▶ Encryption of external attached storage
- ▶ Supports IBM HyperSwap®
- ▶ Supports VMware vSphere Virtual Volumes (VVOLs) and Microsoft Offloaded Data Transfer (ODX)
- ▶ Direct attachment of hosts
- ▶ Hot spare nodes with a standby function of single or multiple nodes
- ▶ Containerization connectivity with Container Storage Interface (CSI), which enables supported storage to be used as persistent storage in container environments
- ▶ Hybrid Multicloud function with IBM Storage Virtualize for Public Cloud

1.2.4 IBM SAN Volume Controller topology

External storage can be managed by IBM SAN Volume Controller in one or more pairs of hardware nodes. This configuration is referred to as a *clustered system*. These nodes are normally attached to the SAN fabric, with storage and host systems. The SAN fabric is zoned to enable the IBM SAN Volume Controller to communicate with external storage systems and hosts.

Within this software release, IBM SAN Volume Controller also supports iSCSI networks. This feature enables the hosts and storage systems to communicate with IBM SAN Volume Controller to build a storage virtualization solution.

It is important that hosts cannot see or operate on the same volume (LUN) that is mapped to the IBM SAN Volume Controller. Although a set of LUNs can be mapped to IBM SAN Volume Controller, and a separate set of LUNs can be mapped directly to one or more hosts, care must be taken to ensure that a separate set of LUNs is always used.

The zoning capabilities of the SAN switch must be used to create distinct zones to ensure that this rule is enforced. SAN fabrics can include standard FC, FC-NVMe, FCoE, iSCSI over Ethernet, or possible future types.

IBM Storage Virtualize 8.6.0 also supports the following NVMe protocols:

- ▶ NVMe/FC
 - Supported since 8.2.1 [4Q18]
 - Supported with 16/32 Gb FC adapters
 - Supports SLES/RH/ESX/Windows as initiators
 - Will support 64 Gb FC adapters in a future release
- ▶ NVMe/RoCE
 - Supported since 8.5.0 [1Q22]
 - Supports RoCE (Mellanox CX-4/CX-6) adapters with 25Gb/100Gb speeds on storage (target) side
 - Supports SLES/RH/ESX as host initiators OS, with RoCE 25/40/100Gb (Mellanox CX-4/CX-5/CX-6), and Broadcom adapters

- Requires RoCE supported Ethernet infrastructure.
- ▶ NVMe/TCP
 - NVMe/NVMeOF protocol is designed to fully exploit the performance of all-flash storage
 - Ethernet storage connection technology is gaining ground in data centers
 - NVMe/RoCE requires special infrastructure
 - NVMe/TCP is a ubiquitous transport and does not require special infrastructure
 - First release (8.6.0) is controller-based, but will take advantage of hardware offload in the future, making the increased speeds comparable with NVMe/RoCE without the need for special networking hardware

Figure 1-4 shows a conceptual diagram of a storage system that uses IBM SAN Volume Controller. It also shows several hosts that are connected to a SAN fabric or local area network (LAN).

In practical implementations that have HA requirements (most of the target clients for IBM SAN Volume Controller), the SAN fabric cloud represents a redundant SAN. A *redundant SAN* consists of a fault-tolerant arrangement of two or more counterpart SANs, which provides alternative paths for each SAN-attached device.

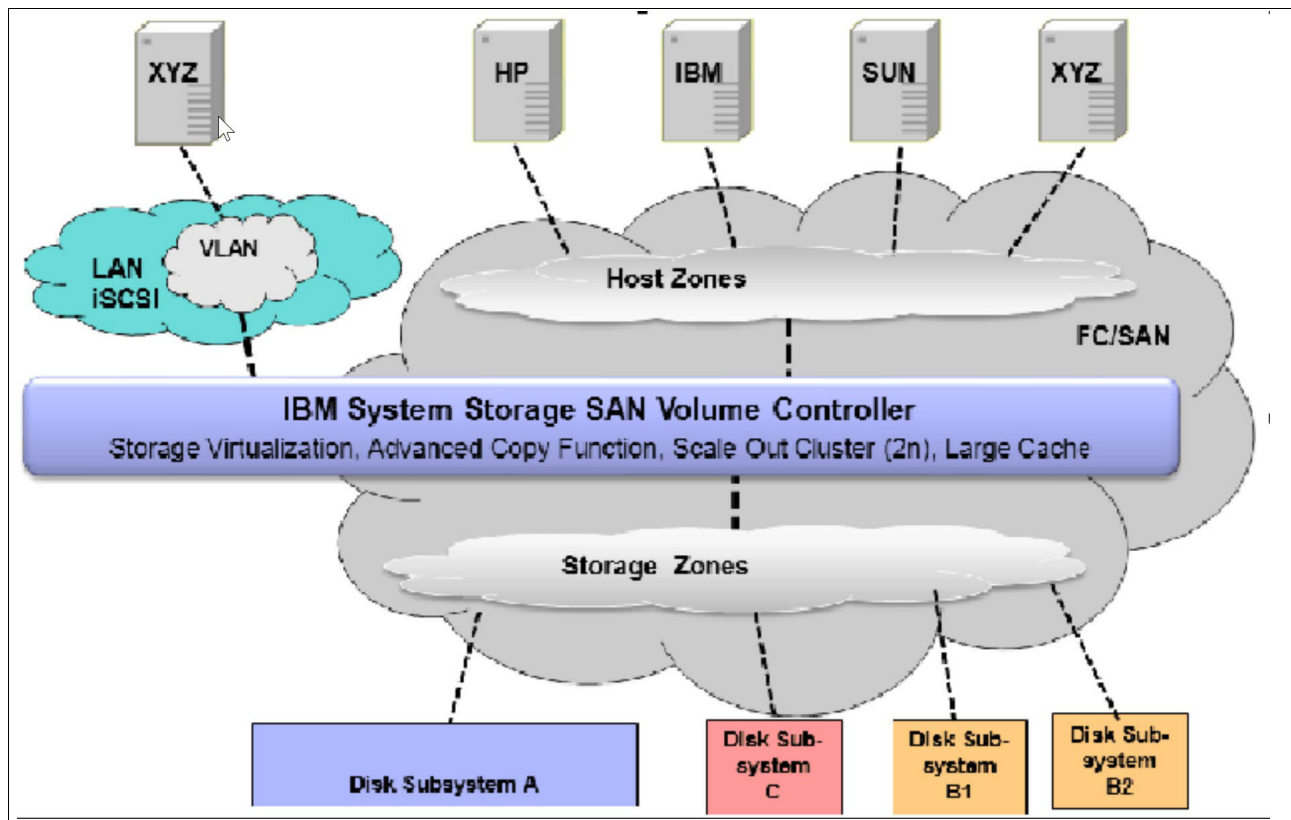


Figure 1-4 IBM SAN Volume Controller conceptual and topology overview

Both scenarios (the use of a single network and the use of two physically separate networks) are supported for iSCSI-based and LAN-based access networks to IBM SAN Volume Controller. Redundant paths to volumes can be provided in both scenarios.

For simplicity, Figure 1-4 on page 12 shows only one SAN fabric and two zones: host and storage. In a real environment, it is a best practice to use two redundant SAN fabrics. IBM SAN Volume Controller can be connected to up to four fabrics.

A clustered system of IBM SAN Volume Controller nodes that are connected to the same fabric presents *logical disks* or volumes to the hosts. These volumes are created from managed LUNs or MDisks that are presented by the storage systems.

The following distinct zones are shown in the fabric:

- ▶ A host zone, in which the hosts can see and address the IBM SAN Volume Controller nodes.
- ▶ A storage zone, in which the IBM SAN Volume Controller nodes can see and address the MDisks or LUNs that are presented by the storage systems.

As explained in 1.4, “IBM SAN Volume Controller family” on page 24, hosts are not permitted to operate on the RAID LUNs directly. All data transfer happens through the IBM SAN Volume Controller nodes. This flow is referred to as *symmetric virtualization*.

For iSCSI-based access, the use of two networks and separating iSCSI traffic within the networks by using a dedicated virtual local area network (VLAN) path for storage traffic prevents any IP interface, switch, or target port failure from compromising the iSCSI connectivity across servers and storage controllers.

1.3 Latest changes and enhancements

In this section, we discuss the latest changes and enhancements in IBM Storage Virtualize 8.6.0. The major functions that were included from 8.5.0 and 8.4.2 releases are also discussed.

1.3.1 IBM Storage Virtualize 8.6.0

IBM Storage Virtualize 8.6.0 provides more features and updates to the IBM Storage Virtualize family of products, including IBM FlashSystems and the IBM SAN Volume Controller.

The following major software and hardware changes are included in 8.6.0:

NVMe over TCP host connectivity

- The NVMe transport protocol provides enhanced performance on high-demand IBM Storage FlashSystem drives.
- NVMe is a logical device interface specification for accessing non-volatile storage media. Host hardware and software use NVMe to fully use the levels of parallelism possible in modern solid-state drives (SSDs).
- Protocols supported (HBA-dependent) are:
 - NVMe over Fibre Channel
 - NVMe over RDMA
 - NVMe over TCP protocol
- For more information, see [IBM System Storage Interoperation Center \(SSIC\)](#).

Support Node CLI to shut down individual Fibre Channel ports

- This command can enable or disable the Fibre Channel port features. The Fibre Channel port supports features like fabric device management interface (FDMI) registration, discovery, and port state.
- You can use the *chfcportfeature* command to enable or disable the Fibre Channel port features.

Support for iSCSI performance improvement

- This procedure provides a solution for Internet Small Computer Systems Interface (iSCSI) host performance problems while connected to a system and its connectivity to the network switch.

Support for 1024 iSCSI hosts per I/O group

- A host can be created with worldwide port names (WWPNs) or iSCSI names. The WWPN name space and the iSCSI name space within the system share the same internal system resources. Each system I/O group can have up to 1024 iSCSI IQNs.

Support for migrating remote copy to policy-based replication

- Policy-based replication uses volume groups and replication policies to automatically deploy and manage replication. Policy-based replication significantly simplifies configuring, managing, and monitoring replication between two systems.
- With policy-based replication, you can replicate data between systems with minimal management, significantly higher throughput and reduced latency compared to the remote-copy function. A replication policy has the following properties:
 - Replication policies cannot be changed after they are created. If changes are required, a new policy can be created and assigned to the associated volume group.
 - Each system supports up to a maximum of 32 replication policies.

Support for Simple Mail Transfer Protocol authentication

- Ability to configure Call Home with email notifications on any of the following email servers:
 - Local email server
 - Emails are sent, received, and archived on physical servers in an on-premises location.
 - Cloud-based email server
 - Emails are sent, received, and archived on cloud-based server providers such as Google Mail and Microsoft Office 365.

Support for TLS 1.3

- Security of all key server communications is governed by TLS 1.2 and TLS 1.3 protocols.
- Encryption keys are distributed between nodes in the system using TLS 1.2 and TLS1.3.
- The system uses AES-256 encryption that uses OpenSSL library interfaces. To establish a connection between the key server and the system, the key server or services must support the configured TLS version.

Support for DNS check to be turned off

- Use the *lsdnserver* command to list information for any Domain Name System (DNS) servers in the system.

Support for FDMI information in Call Home

- The Fabric Device Management Interface (FDMI) enables any storage endpoint to register itself to the Fibre Channel (FC) fabric and query the HBA and port details of the entire fabric.

Version 2 metadata volume for VMware Virtual Volumes (vVols)

- The system provides native support for VMware vSphere APIs for Storage Awareness (VASA) through a VASA Provider (also known as Storage Provider) which sends and receives information about storage that is used by VMware vSphere to the vCenter Server.
- Through VASA, the system also supports VMware Virtual Volumes (also known as vVols), which allows VMware vCenter to automate the creation, deletion and mapping of volumes.

Enhancement to IBM Storage Insights for threat detection

- A key part of monitoring your system includes the detection of potential ransomware attacks. To ensure that you have the latest storage metadata for detecting those types of attacks, compression and cyber resiliency statistics for volumes are collected every five minutes.
- With these statistics, IBM Storage Insights builds a historical model of a storage system and uses its built-in intelligence and formulas to identify when and where ransomware attacks might be occurring.
- For more information on statistics, see, [Storage Insights Overview](#).

Support for non-superuser ability to manage the system

- Each user of the management GUI must provide a username and a password to sign on. Each user also has an associated role, such as monitor or security administrator. These roles are defined at the system level. For example, a user can be the administrator for one system, but the security administrator for another system.
- This feature has allowed non-superuser level, access to monitor and manage certain functions on the system.

Downloading software patches through Call Home using Restful API

- This option enables Call Home with cloud services to connect to support, and directly transfer the latest update available to the storage system.
- To download a software patch file through the Call Home using the Restful API, the user enters the `satask downloadsoftware` command plus the specific patch they require.

For more information, see [What's new in 8.6.0](#).

1.3.2 IBM Storage Virtualize 8.5.0

IBM Storage Virtualize 8.5.0 provides more features and updates to the IBM Storage Virtualize family of products of which IBM FlashSystems and the IBM SAN Volume Controller are part.

The following major software and hardware changes are included in 8.5.0:

- ▶ Support for:
 - The new IBM FlashSystem 9500, 7300, and IBM SAN Volume Controller SV3 systems, including:
 - 100 Gbps Ethernet adapter
 - 48 NVMe drives per distributed RAID-6 array (IBM FlashSystem 9500 only)
 - Secure boot drives
 - Multifactor authentication and single sign-on (SSO)
 - NVMe host attachment over RDMA
 - Fibre Channel port sets
 - I/O stats in microseconds
 - Domain names for IP replication
 - IBM Storage Virtualize 3-Site Orchestrator 4.0
 - Support for increased number of hosts per I/O group
- ▶ Improved distributed RAID array recommendations
- ▶ Improved default time for updates
- ▶ Updates to OpenStack support summary

IBM FlashSystems 9500, 7300 and IBM SAN Volume Controller SV3

The following new hardware platforms were released with the IBM Storage Virtualize 8.5.0. For more information about these new hardware platforms, see 1.5, “IBM SAN Volume Controller models” on page 44, and 1.6, “IBM FlashSystem family” on page 49:

- ▶ IBM FlashSystems 9500
 - The IBM FlashSystems 9500 is a 4U control enclosure and contains up to 48 NVMe-attached IBM FlashCore® Modules or other self-encrypted NVMe-attached SSDs.
 - The NVMe-attached drives in the control enclosures provide significant performance improvements compared to SAS-attached flash drives. The system supports 2U and 5U all-Flash SAS attached expansion enclosure options.
- ▶ IBM FlashSystems 7300
 - The FlashSystem 7300 is a 2U dual controller that contains up to 24 NVMe-attached IBM FlashCore Modules or other self-encrypted NVMe-attached SSDs or Storage Class Module drives.
 - IBM FlashSystem 7300 system also supports 2U and 5U SAS-attached expansion enclosure options.
- ▶ IBM SAN Volume Controller SV3

The IBM SAN Volume Controller SV3 system is a 2U single controller that combines software and hardware into a comprehensive, modular appliance that provides symmetric virtualization. The SV3 is run in pairs from an I/O group, which is the building block for any IBM SAN Volume Controller virtualization setup.

New 100 Gbps Ethernet adapter

Consider the following points:

- ▶ A maximum of six 100 Gbps Ethernet adapters can be installed in the IBM FlashSystems 9500 and 7300 and in a pair of IBM SAN Volume Controller SV3s.
- ▶ In 100 Gbps Ethernet ports, iSCSI performance is equivalent to 25 Gbps iSCSI host attachment.

Support for 48 NVMe drives per distributed RAID-6 array

Enhanced support for 48 NVMe drives in the enclosure by using distributed RAID 6 technology is now available for the IBM FlashSystems 9500.

The following configurations are supported:

- ▶ Distributed RAID 6 arrays of NVMe drive expansion up to 48 member drives.
- ▶ Distributed RAID 6 arrays of FCM NVMe drive support expansion up to 48 member drives.
- ▶ Distributed RAID 6 arrays of extra large (38.4 TB) physical capacity FCM NVMe drives supports up to 24 member drives.

Support for secure boot drives

IBM Storage Virtualize 8.5.0 provides secure boot by pairing each boot drive with a Trusted Platform Module (TPM). TPM provides a secure cryptographic processor that verifies hardware and prevents unauthorized access to hardware and the operating system. On the system, TPM protects secure boot to ensure the code images that are installed are signed, trusted, and unchanged.

Multifactor authentication and SSO

Multifactor authentication requires users to provide multiple pieces of information when they log in to the system to prove their identity. Multifactor authentication uses any combination of two or more methods, called *factors*, to authenticate users to your resources and protect those resources from unauthorized access.

One of the key concepts of multifactor authentication is that each factor comes from a different category: something the user knows, has, or is.

SSO authentication requires users to register their credentials only once when the user signs on to the application for the first time. The user information is stored at the Identity Provider (IdP) that manages the user credentials and determines whether the user is required to authenticate again.

NVMe host attachment over RDMA

The new NVMe/RDMA function uses the RDMA capabilities of the host adapters, as does the existing iSER support. NVMe over RDMA connections support 25 Gbps RoCE adapters or 100 Gbps RoCE adapters.

Fibre Channel port sets

Port sets are groupings of logical addresses that are associated with the specific traffic types. The IBM Storage Virtualize 8.5.0 supports IP and Fibre Channel port sets for host attachment, and IP port sets for backend storage connectivity, and IP replication traffic.

I/O stats in microseconds

Real-time performance statistics provide short-term status information for the system. These statistics summarize the overall performance health of the system and can be used to monitor trends in bandwidth and CPU usage. IBM Storage Virtualize 8.5.0 gives you the ability to see the granularity down at microseconds intervals, which was not available on previous code levels.

Support for fully qualified domain names for IP replication

Metro Mirror and Global Mirror partnerships can be established over Ethernet links that use the IPv4 and IPv6 addresses that are associated with Ethernet ports. Before IBM Storage Virtualize 8.5.0, these addresses had to be physical dotted decimal IP types. IBM Storage Virtualize 8.5.0 allows the user to specify domain names as well when IP partnerships are created. If you specify domain names, a DNS server must be configured on your system.

1.3.3 IBM Storage Virtualize 8.4.2

IBM Storage Virtualize 8.5.0 also provides all of the following previous functions and major enhancements of the 8.4.2 release:

- ▶ Code release schedule
- ▶ Safeguarded Copy
- ▶ Volume mobility
- ▶ Storage as a Service (STaaS)

These updates are described next.

Code release schedule

The new code release schedule includes the following features:

- ▶ Two types of releases: Long Term Releases (LTR) and Continuous Development Releases (CDRs)
- ▶ Four slated updates each year:
 - One LTR:
 - New features are formally announced.
 - Includes all the enhancements of the previous CDRs.
 - Adopts the major numbering scheme (8.4, 8.5, 8.6, and so on).
 - Includes all Program Temporary Fixes (PTFs) of problems that were found with existing features going forward.
 - Three CDRs:
 - No formal announcement.
 - New features are announced, but do not have PTFs.
 - PTFs are generally added to the next CDR (unless critical).
 - The fixes are incorporated into the next CDR.
 - Adopt a minor numbering scheme is used (8.4.1, 8.5.1, and so on).

Safeguarded Copy

Safeguarded Copy is a virtual air gap mechanism that uses FlashCopy functions to take immutable copies. This feature aids in the recovery from ransomware or internal “bad actors” who seek to destroy data.

Note: The Safeguarded Copy function is available with IBM Storage Virtualize software 8.5.0, but is not supported for the FlashSystem 5000, Storwize V5030E, Storwize V7000 Gen2, and Storwize V7000 Gen2+ models.

Safeguarded Copy is a feature or solution with which you can create point-in-time copies (“granularity”) of active production data that cannot be altered or deleted (so that is Immutable or protected). It requires a user with the correct privilege access to modify the Safeguarded Copy expiration settings (Separation of Duties). Lastly, Safeguarded Copy uses copy management software for testing and ease of recovery of copies.

Safeguarded Copy use cases

Safeguarded Copy includes the following use cases:

- ▶ **Validation:** Regular analytics of the copy to provide early detection of a problem or reassurance that the copy is a good copy before further action is taken.
- ▶ **Forensic:** Use a copy of the production system to investigate the problem and determine whether recovery action is necessary.
- ▶ **Surgical:** Extract data from the copy and logically restore back to the production environment.
- ▶ **Catastrophic:** Recover the entire environment back to the point in time of the copy because this option is the only available recovery option.
- ▶ **Offline backup:** Performing an offline backup of data from a consistent point-in-time copy can be used to build a second line of defense, which provides a greater retention period and increased isolation and security.

Safeguarded Copy is the first step toward a cyber resiliency solution. Focusing on the feature set from a customer’s perspective, consider the following three pillars, as shown in Figure 1-5 on page 20:

- ▶ Separation of Duties
- ▶ Protected Copies
- ▶ Automation

IBM Storage Virtualize Safeguarded Copy Data Resilience

IBM FlashSystem Safeguarded Copy feature prevents point-in-time copies of data from being modified or deleted because of user errors, malicious destruction, or ransomware attacks

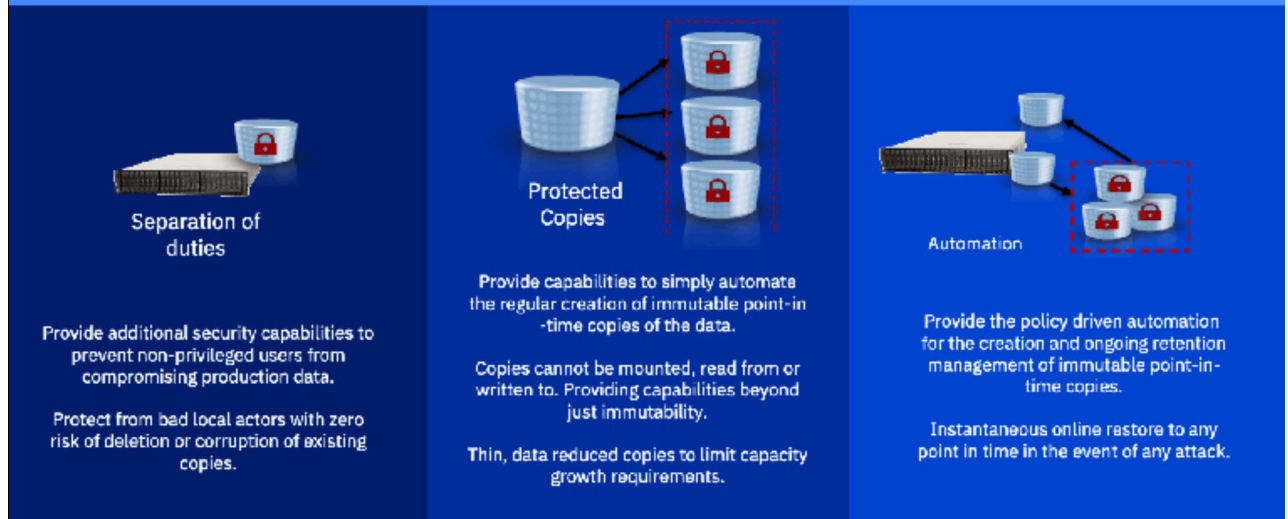


Figure 1-5 IBM Safeguarded Copy Data Resilience

Figure 1-5 shows the following IBM Storage Virtualize Safeguarded Copy Data Resilience examples:

- ▶ Separation of duties
 - Traditional backup and restore capabilities are normally storage administrator controlled and do not protect against intentional (for example, rogue employee) or non-intentional attacks.
 - Primary and backup are treated differently. Protecting current backups, or securing and hardening current backup, does not solve the problem.
- ▶ Protected copies of the data
 - These backups must be immutable; for example, hidden, non-addressable, cannot be altered or deleted, only usable after recovery.
 - Copies must deliver a higher level of security while meeting industry and business regulations.
- ▶ Automation:
 - Initially setting and managing of policies (number of copies, retention period).
 - Automating, managing, and restoring of those copies.

For more information, see [Safeguarded Copy function](#) and *IBM FlashSystem Safeguarded Copy Implementation Guide*, REDP-5654.

Volume mobility

Volume mobility is similar to nondisruptive volume movement between I/O groups, only you are migrating a volume between systems (for example, IBM SAN Volume Controller to FS9500). It allows a user to nondisruptively move data between systems that do not natively cluster. This feature is a major benefit for upgrading and changing Storage Virtualize systems.

Volume migrations also allow:

- ▶ Nondisruptive volume migrations between independent systems or clusters.
- ▶ Nondisruptive migrations between non-clustering platforms.
- ▶ Volume migration away from a cluster that is reaching max capacity limits.

Volume mobility uses enhancements to SCSI (ALUA) path states. The migration is based on Remote Copy (Metro Mirror) functions.

Consider the following restrictions about volume mobility:

- ▶ No 3-site support.
- ▶ Not a DR or HA solution.
- ▶ No support for consistency groups, change volumes, or expanding volumes.
- ▶ Partnership requirements are equivalent to Metro Mirror.
- ▶ Performance considerations apply as with Metro Mirror.
- ▶ Reduced host interoperability support:
 - RHEL
 - SLES
 - ESXi
 - Solaris
 - HP-UX
 - SCSI only (No FC-NVMe)
- ▶ No SCSI persistent reservations or Offloaded Data Transfer (ODX).
- ▶ IBM i, AIX®, Hyper-V, and Windows are *not* supported.
- ▶ Remote Mirror license is not required.

IBM Storage as a Service

The latest member in the IBM Storage Family is the newly announced IBM Storage as a Service (STaaS) offering.

Buyer behavior is shifting from technology-focused to service-level agreement (SLA)-driven, cloud-like simplicity. IT staff is being downsized and transitioning into generalist roles, rather than specializing in specific areas. Enterprise workloads need flexibility where applications are to be deployed. New and better infrastructures, such as 5G, allow this growth outside of the traditional data center.

The IBM STaaS offering is a pure OpEx solution and does not require initial capital.

Consumption-based solutions include the following important features:

- ▶ Flexible scale-up and scale-down model
- ▶ Cloud-like functions in most solutions
- ▶ All the deployment and managed support, optimization, and disposal services included
- ▶ Well-defined upgrade path
- ▶ Clear pricing terms
- ▶ Switch from capital expenditures (CapEx) funding to operating expenses (OpEx) funding
- ▶ Ease of billing and payment terms

With the IBM STaaS offering, the customer makes the following decisions:

- ▶ Which tier level is needed.
- ▶ The amount of storage capacity is needed.
- ▶ For how long the customer wants to use this offering.
- ▶ Connection type that is required.
- ▶ Encryption option that is needed.

Figure 1-6 shows the STaaS tiers available for IBM FlashSystems.

	Extreme (Tier 1)	Premium (Tier 2)	Balanced (Tier 3)
Minimum capacity	25	50	100
Performance (IOPS/TB)*	4,500	2,250	600
Maximum read throughput (GB/s)**	45	45	35
Maximum write throughput (GB/s)**	12	12	10
Availability	99.9999% / optional 100% guarantee***	99.9999% / optional 100% guarantee***	99.9999% / optional 100% guarantee***
Connection type	Ethernet or FC	Ethernet or FC	Ethernet or FC
Encryption	Optional at no cost	Optional at no cost	Optional at no cost
Contract terms	1-5 years	1-5 years	1-5 years
Lifecycle management includes			
Installation	IBM	IBM	IBM
Predictive support	Storage Insights Pro	Storage Insights Pro	Storage Insights Pro
SW S&S	Yes	Yes	Yes
Remote firmware updates	Yes	Yes	Yes
IBM Technical Account Manager	Yes	Yes	Yes
IBM monitoring and capacity growth	Yes	Yes	Yes
HW service	24x7 same day onsite repair IBM performed	24x7 same day onsite repair IBM performed	24x7 same day onsite repair IBM performed
<p><i>*IOPS/TB: This defines the minimum ratio of Input-Output Operations (IOPS) per physical used Terabyte (TB) with usage up to 90% of the usable capacity. IOPS/TB are based on a 70:30 Read / Write mixed workload with 50% cache hit and 16K I/O size using Fibre Channel. Remote replication may impact performance. Actual results will vary by workload, as such the IOPS/TB should be used for the purposes of selecting relative performance tiers.</i></p> <p><i>**Maximum read and write throughput based on 256KB I/O Size using Fibre Channel.</i></p> <p><i>*** 99.9999% objective. 100% guarantee requires HyperSwap configuration installed by IBM Lab Services. Terms and conditions apply.</i></p>			

Figure 1-6 STaaS Tiers comparison

IBM Lifecycle Management provides and includes the following features:

- ▶ Deployment, maintenance, and disposal of the hardware:
 - Offering is based on IBM FlashSystems technology.
 - All hardware is an IBM asset that is installed on-premises at the customer data center.
- ▶ IBM monitoring of capacity growth and IBM installs more if needed (miscellaneous equipment specification [MES]).
- ▶ Periodic technology refresh with 90-day overlap for migration.
- ▶ IBM recycle process of old equipment.

STaaS pseudo machine type

The pseudo machine type for STaaS is 9601, with performance tiers and terms that are differentiated by models (base, medium, and high).

Feature codes to support these models are related to the setup type, annual capacity growth, and options for Ethernet, encryption, and decreasing capacity.

9601 models

The following 9601 models are available:

- ▶ Balanced performance: BT1, BT2, BT3, BT4, and BT5
- ▶ Premium performance: MT1, MT2, MT3, MT4, and MT5
- ▶ Extreme performance: HT1, HT2, HT3, HT4, and HT5

Note: The numeric value in the model number is the duration of the STaaS contract in years.

The IBM STaaS offering also includes the new Storage Expert Care Premium Level of service, which features the resource of a dedicated Technical Account Manager (TAM), enhanced response to severity 1 and 2 issues, and predictive support by way of IBM Storage Insights.

For more information about the STaaS offering, see *IBM Storage as a Service Offering Guide*, [REDP-5644](#).

1.4 IBM SAN Volume Controller family

This section explains the underlying concepts of IBM SAN Volume Controller. It also provides an architectural overview and describes the terminology that is used in a virtualized storage environment. Finally, it introduces the software and hardware components and the other functions that are available with 8.6.

Figure 1-7 shows the complete IBM SAN Volume Controller family that supports the IBM Storage Virtualize 8.6 software.

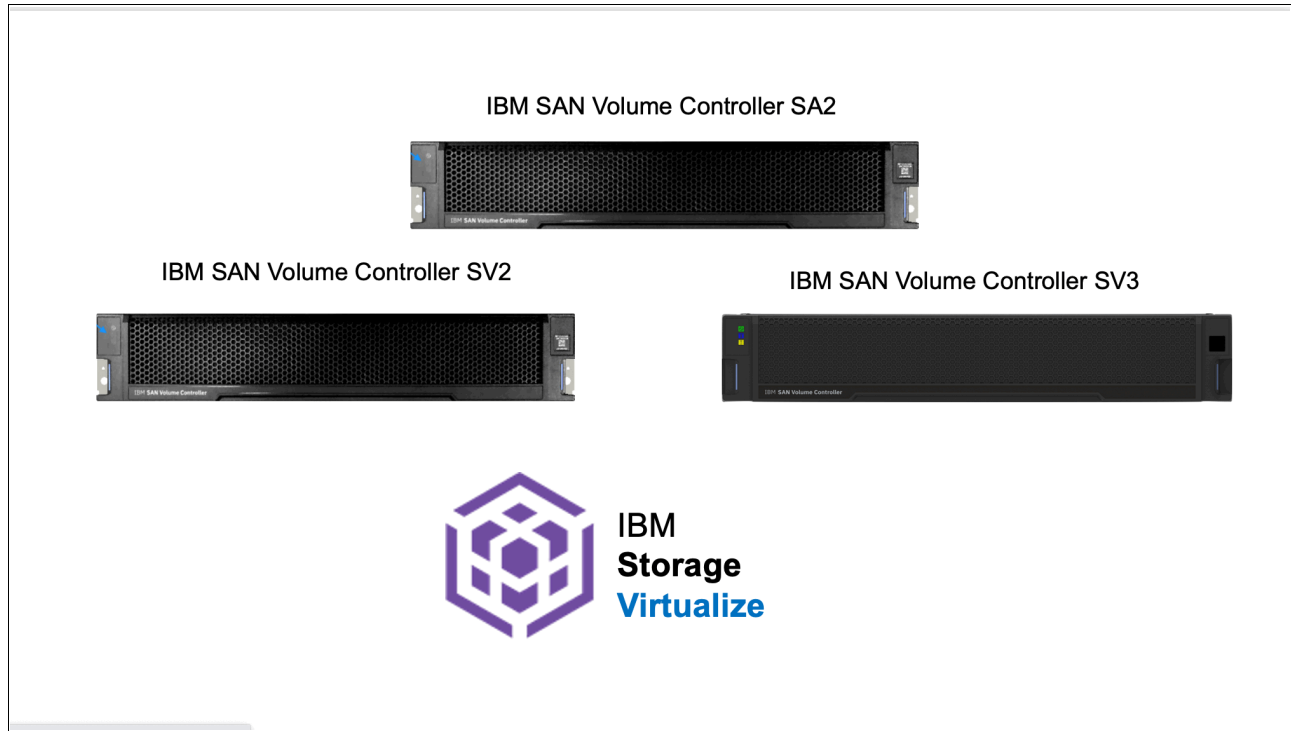


Figure 1-7 IBM SAN Volume Controller family

1.4.1 Components

IBM SAN Volume Controller provides block-level aggregation and volume management for attached disk storage. In simpler terms, the IBM SAN Volume Controller manages several back-end storage controllers or locally attached disks.

IBM SAN Volume Controller maps the physical storage within those controllers or storage systems into logical disk images, or *volumes*, that can be seen by application servers and workstations in the SAN. It logically sits between hosts and storage systems. It presents itself to hosts as the storage provider (*target*) and to storage systems as one large host (*initiator*).

The SAN is zoned such that the application servers cannot “see” the back-end storage or controller. This configuration prevents any possible conflict between IBM SAN Volume Controller and the application servers that are trying to manage the back-end storage.

The IBM SAN Volume Controller is based on the components that are described next.

1.4.2 Nodes

Each IBM SAN Volume Controller hardware unit is called a *node*. Each node is an individual server in an IBM SAN Volume Controller clustered system on which the Storage Virtualize software runs. The node provides the virtualization for a set of volumes, cache, and copy services functions.

The IBM SAN Volume Controller nodes are deployed in pairs (*io_group*), and one or multiple pairs constitute a *clustered system* or *system*. A system can consist of one pair and a maximum of four pairs.

One of the nodes within the system is known as the *configuration node*. The configuration node manages the configuration activity for the system. If this node fails, the system chooses a new node to become the configuration node.

Because the active nodes are installed in pairs, each node provides a failover function to its partner node if a node fails.

1.4.3 I/O groups

Each pair of IBM SAN Volume Controller nodes is also referred to as an *I/O group*. An IBM SAN Volume Controller clustered system can have 1 - 4 I/O groups.

A specific *volume* is always presented to a host server by a single I/O group of the system. The I/O group can be changed.

When a host server performs I/O to one of its volumes, all the I/Os for a specific volume are directed to one specific I/O group in the system. Under normal conditions, the I/Os for that specific volume are always processed by the same node within the I/O group. This node is referred to as the *preferred node* for this specific volume.

Both nodes of an I/O group act as the preferred node for their own specific subset of the total number of volumes that the I/O group presents to the host servers. However, both nodes also act as failover nodes for their respective partner node within the I/O group. Therefore, a node takes over the I/O workload from its partner node when required.

In an IBM SAN Volume Controller-based environment, the I/O handling for a volume can switch between the two nodes of the I/O group. Therefore, it is a best practice that servers are connected to two different fabrics through different FC host bus adapters (HBAs) to use multipath drivers to give redundancy.

The IBM SAN Volume Controller I/O groups are connected to the SAN so that all application servers that are accessing volumes from this I/O group can access this group. Up to 512 host server objects can be per I/O group. The host server objects can access volumes that are provided by this specific I/O group.

If required, host servers can be mapped to more than one I/O group within the IBM SAN Volume Controller system. Therefore, they can access volumes from separate I/O groups. You can move volumes between I/O groups to redistribute the load between the I/O groups. Modifying the I/O group that services the volume can be done concurrently with I/O operations if the host supports nondisruptive volume moves.

It also requires a rescan at the host level to ensure that the multipathing driver is notified that the allocation of the preferred node changed, and the ports (by which the volume is accessed) changed. This modification can be done in the situation where one pair of nodes becomes overused.

1.4.4 System

The system or clustered system consists of 1 - 4 I/O groups. Specific configuration limitations are then set for the entire system. For example, the maximum number of volumes that is supported per system is 10,000, or the maximum capacity of MDisks that is supported is ~28 PiB (32 PB) per system.

All configuration, monitoring, and service tasks are performed at the system level. Configuration settings are replicated to all nodes in the system. To facilitate these tasks, a management IP address is set for the system.

A process is provided to back up the system configuration data on to storage so that it can be restored if a disaster occurs. This method does not back up application data. Only the IBM SAN Volume Controller system configuration information is backed up.

For remote data mirroring, two or more systems must form a partnership before relationships between mirrored volumes are created.

For more information, see [V8.6.0.x Configuration Limits and Restrictions for IBM System Storage SAN Volume Controller](#).

1.4.5 MDisks

The SAN Volume Controller system and its I/O groups view the storage that is presented to them by the back-end storage system as several disks or LUNs, which are known as *MDisks*. Because IBM SAN Volume Controller does not attempt to provide recovery from physical disk failures within the back-end storage system, an MDisk must be provisioned from a RAID array.

These MDisks are placed into storage pools where they are divided into several extents. The application servers do not “see” the MDisks at all. Rather, they see logical disks, which are known as *volumes*. These disks are presented by the IBM SAN Volume Controller I/O groups through the SAN or LAN to the servers.

For information, see [V8.6.0.x Configuration Limits and Restrictions for IBM System Storage SAN Volume Controller](#).

When an MDisk is presented to the IBM SAN Volume Controller, it can be one of the following statuses:

- ▶ Unmanaged MDisk
 - An MDisk is reported as unmanaged when it is not a member of any storage pool. An unmanaged MDisk is not associated with any volumes and has no metadata that is stored on it.
 - IBM SAN Volume Controller does not write to an MDisk that is in unmanaged mode except when it attempts to change the mode of the MDisk to one of the other modes. IBM SAN Volume Controller can see the resource, but the resource is not assigned to a storage pool.
- ▶ Managed MDisk
 - Managed mode MDisks are always members of a storage pool, and they contribute extents to the storage pool. Volumes (if not operated in image mode) are created from these extents. MDisks that are operating in managed mode might have metadata extents that are allocated from them and can be used as *quorum disks*. This mode is the most common and normal mode for an MDisk.

- ▶ Image mode MDisk

Image mode provides a direct block-for-block conversion from the MDisk to the volume by using virtualization. This mode is provided to satisfy the following major usage scenarios:

- Image mode enables the virtualization of MDisks that contain data that was written directly and not through IBM SAN Volume Controller. Rather, it was created by a direct-connected host.
- This mode enables a client to insert IBM SAN Volume Controller into the data path of a storage volume or LUN with minimal downtime.
- Image mode enables a volume that is managed by an IBM SAN Volume Controller to be used with the native copy services function that is provided by the underlying RAID controller. To avoid the loss of data integrity when the IBM SAN Volume Controller is used in this way, it is important that you disable the IBM SAN Volume Controller cache for the volume.
- The IBM SAN Volume Controller can migrate to image mode, which enables the IBM SAN Volume Controller to export volumes and access them directly from a host without the IBM SAN Volume Controller in the path.

Each MDisk that is presented from an external disk controller features an online path count that is the number of nodes that can access that MDisk. The *maximum count* is the maximum number of paths that is detected at any point by the system. The *current count* is what the system sees now. A current value that is less than the maximum can indicate that SAN fabric paths were lost.

Tier

It is likely that the MDisks (LUNs) that are presented to the IBM SAN Volume Controller system have different characteristics because of the disk or technology type on which they are placed. The following tier options are available:

- ▶ tier0_flash
- ▶ tier1_flash
- ▶ tier_enterprise
- ▶ tier_nearline
- ▶ tier_scm

The default value for a newly discovered unmanaged MDisk is enterprise. You can change this value by running the `chmdisk` command.

The tier of external MDisks is not detected automatically and is set to enterprise. If the external MDisk is made up of flash drives or nearline Serial Attached SCSI (SAS) drives and you want to use IBM Easy Tier, you must specify the tier when adding the MDisk to the storage pool or run the `chmdisk` command to modify the tier attribute.

1.4.6 Cache

The primary benefit of storage cache is to improve I/O response time. Reads and writes to a magnetic disk drive experience seek time and latency time at the drive level, which can result in 1 ms - 10 ms of response time (for an enterprise-class disk).

Consider the following points:

- ▶ The IBM SAN Volume Controller Model SA2 and SV2 features 128 GB of memory with options for 768 GB of memory in a 2U 19-inch rack mount enclosure.

- ▶ The IBM SAN Volume Controller Model SV3 features 512 GB of memory with options for 1536 GB of memory in a 2U 19-inch rack mount enclosure.

The IBM SAN Volume Controller provides a flexible cache model as described next.

Cache is allocated in 4 kibibyte (KiB) segments. A *segment* holds part of one track. A *track* is the unit of locking and destaging granularity in the cache. The cache virtual track size is 32 KiB (eight segments).

A track might be only partially populated with valid pages. The IBM SAN Volume Controller combines writes up to a 256 KiB track size if the writes are in the same tracks before destaging. For example, if 4 KiB is written into a track, another 4 KiB is written to another location in the same track.

Therefore, the blocks that are written from the IBM SAN Volume Controller to the disk subsystem can be any size of 512 bytes - 256 KiB. The large cache and advanced cache management algorithms enable it to improve the performance of many types of underlying disk technologies.

The IBM SAN Volume Controller capability to manage in the background the destaging operations that are incurred by writes (in addition to still supporting full data integrity) assists with the IBM SAN Volume Controller capability in achieving good performance.

The cache is separated into two layers: upper cache and lower cache. Figure 1-8 on page 29 shows the separation of the upper and lower cache.

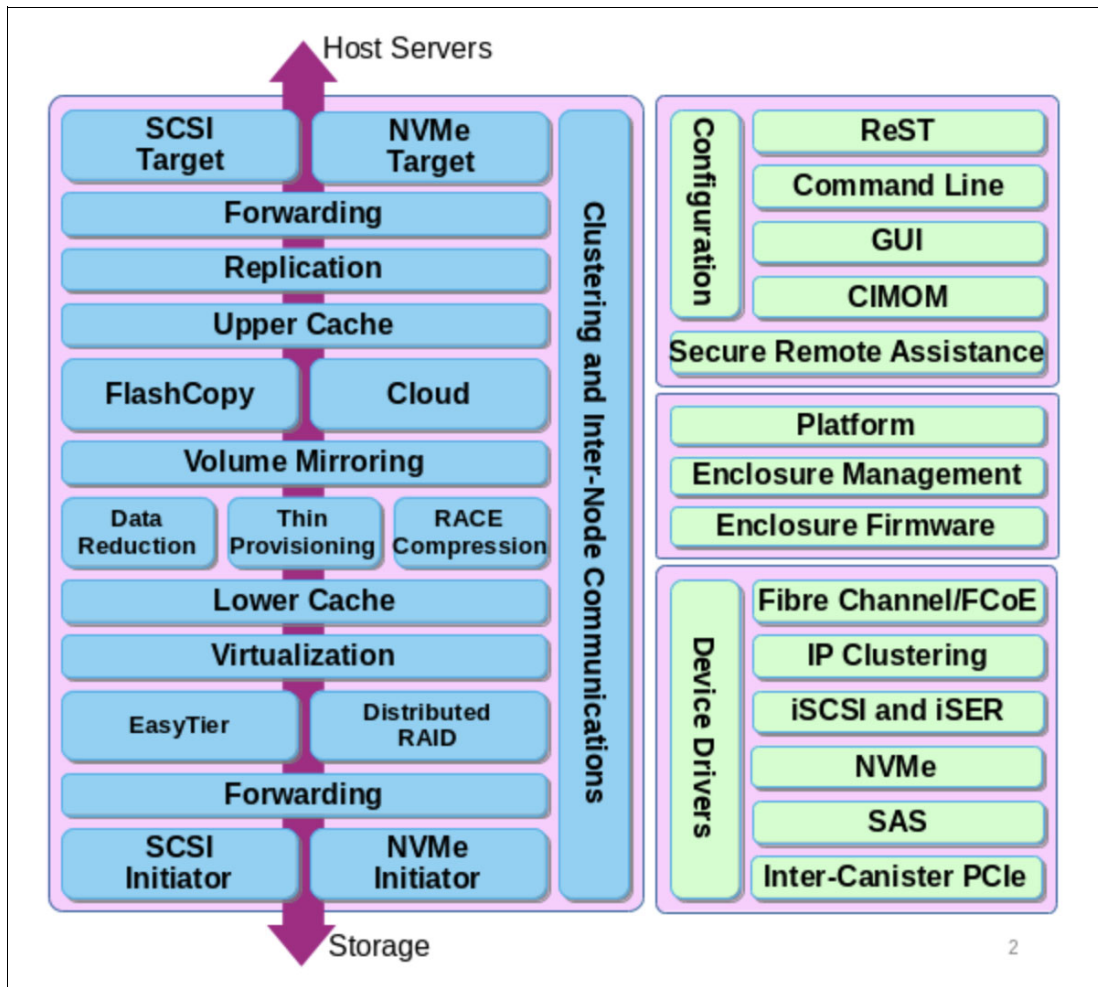


Figure 1-8 Separation of upper and lower cache

The upper cache delivers the following functions, which enable the IBM SAN Volume Controller to streamline data write performance:

- ▶ Fast write response times to the host by being as high up in the I/O stack as possible
- ▶ Partitioning

The lower cache delivers the following functions:

- ▶ Ensures that the write cache between two nodes is in sync.
- ▶ Caches partitioning to ensure that a slow backend cannot use the entire cache.
- ▶ Uses a destaging algorithm that adapts to the amount of data and the back-end performance.
- ▶ Provides read caching and prefetching.

Combined, the two levels of cache also deliver the following functions:

- ▶ Pins data when the LUN goes offline.
- ▶ Provides:
 - Enhanced statistics for IBM Storage Control and IBM Storage Insights
 - Trace for debugging
- ▶ Reports medium errors.

- ▶ Resynchronizes cache correctly and provides the atomic write function.
- ▶ Ensures that other partitions continue operation when one partition becomes 100% full of pinned data.
- ▶ Integrates with T3 recovery procedures.
- ▶ Supports:
 - Fast-write (two-way and one-way), flush-through, and write-through
 - Two-way operation
 - None, read-only, and read/write as user-exposed caching policies
 - Flush-when-idle
 - Expanding cache as more memory becomes available to the platform
 - Credit throttling to avoid I/O skew and offer fairness/balanced I/O between the two nodes of the I/O group
- ▶ Enables switching the preferred node without needing to move volumes between I/O groups.

Depending on the size, age, and technology level of the disk storage system, the total available cache in the IBM SAN Volume Controller nodes can be larger, smaller, or about the same as the cache that is associated with the disk storage.

Because hits to the cache can occur in the IBM SAN Volume Controller or the back-end storage system level of the overall system, the system as a whole can take advantage of the larger amount of cache wherever the cache is available.

In addition, regardless of their relative capacities, both levels of cache tend to play an important role in enabling sequentially organized data to flow smoothly through the system. The IBM SAN Volume Controller cannot increase the throughput potential of the underlying disks in all cases because this increase depends on the underlying storage technology and the degree to which the workload exhibits *hotspots* or sensitivity to cache size or cache algorithms.

However, the write cache is still assigned to a maximum of 12 GB and compression cache to a maximum of 34 GB. The remaining installed cache is used as read cache (including allocation for features, such as IBM FlashCopy, Global Mirror, or Metro Mirror). Data reduction pools share memory with the main I/O process.

1.4.7 Quorum disk

A *quorum disk* is an MDisk or a managed drive that contains a reserved area that is used exclusively for system management. A system automatically assigns quorum disk candidates. Quorum disks are used when a problem exists in the SAN fabric or when nodes are shut down, which leaves half of the nodes remaining in the system. This type of problem causes a loss of communication between the nodes that remain in the system and the nodes that do not remain.

The nodes are split into groups where the remaining nodes in each group can communicate with each other, but not with the other group of nodes that were formerly part of the system. In this situation, some nodes must stop operating and processing I/O requests from hosts to preserve data integrity while maintaining data access. If a group contains less than half the nodes that were active in the system, the nodes in that group stop operating and processing I/O requests from hosts.

It is possible for a system to split into two groups with each group containing half the original number of nodes in the system. A quorum disk determines which group of nodes stops operating and processing I/O requests. In this tiebreaker situation, the first group of nodes that accesses the quorum disk is marked as the owner of the quorum disk. As a result, all of the nodes that belong to the owner group continue to operate as the system and handle all I/O requests.

If the other group of nodes cannot access the quorum disk or discover that the quorum disk is owned by another group of nodes, it stops operating as the system and does not handle I/O requests. A system can have only one active quorum disk that is used for a tiebreaker situation. However, the system uses three quorum disks to record a backup of the system configuration data that is used if a disaster occurs. The system automatically selects one active quorum disk from these three disks.

The other quorum disk candidates provide redundancy if the active quorum disk fails before a system is partitioned. To avoid the possibility of losing all of the quorum disk candidates with a single failure, assign quorum disk candidates on multiple storage systems.

Quorum disk requirements: To be considered eligible as a quorum disk, a LUN must meet the following criteria:

- ▶ It is presented by a storage system that supports IBM SAN Volume Controller quorum disks.
- ▶ It is manually enabled as a quorum disk candidate by running the `chcontroller -allowquorum yes` command.
- ▶ It is in managed mode (no image mode).
- ▶ It includes sufficient free extents to hold the system state information and the stored configuration metadata.
- ▶ It is visible to all of the nodes in the system.

If possible, the IBM SAN Volume Controller places the quorum candidates on separate storage systems. However, after the quorum disk is selected, no attempt is made to ensure that the other quorum candidates are presented through separate storage systems.

Quorum disk placement verification and adjustment to separate storage systems (if possible) reduce the dependency from a single storage system, and can increase the quorum disk availability.

You can list the quorum disk candidates and the active quorum disk in a system by running the `lsquorum` command.

When the set of quorum disk candidates is chosen, it is fixed. However, a new quorum disk candidate can be chosen in one of the following conditions:

- ▶ When the administrator requests that a specific MDisk becomes a quorum disk by running the `chquorum` command.
- ▶ When an MDisk that is a quorum disk is deleted from a storage pool.
- ▶ When an MDisk that is a quorum disk changes to image mode.

An offline MDisk is not replaced as a quorum disk candidate.

For DR purposes, a system must be regarded as a single entity so that the system and the quorum disk can be collocated.

Special considerations are required for the placement of the active quorum disk for a stretched, split cluster or split I/O group configurations. For more information, see [SAN Volume Controller](#).

Important: Running an IBM SAN Volume Controller system without a quorum disk can seriously affect your operation. A lack of available quorum disks for storing metadata prevents any migration operation.

Mirrored volumes can be taken offline if no quorum disk is available. This behavior occurs because the synchronization status for mirrored volumes is recorded on the quorum disk.

During the normal operation of the system, the nodes communicate with each other. If a node is idle for a few seconds, a heartbeat signal is sent to ensure connectivity with the system. If a node fails for any reason, the workload that is intended for the node is taken over by another node until the failed node is restarted and readmitted into the system, which happens automatically.

If the Licensed Internal Code on a node becomes corrupted, which results in a failure, the workload is transferred to another node. The code on the failed node is repaired, and the node is readmitted into the system (which is an automatic process).

IP quorum configuration

In a stretched configuration or IBM HyperSwap configuration, you must use a third, independent site to house quorum devices. To use a quorum disk as the quorum device, this third site must use FC or IP connectivity together with an external storage system. In a local environment, no extra hardware or networking, such as FC or SAS-attached storage, is required beyond what is normally always provisioned within a system.

To use an IP-based quorum application as the quorum device for the third site, no FC connectivity is used. Java applications are run on hosts at the third site. However, strict requirements on the IP network and some disadvantages with the use of IP quorum applications exist.

Unlike quorum disks, all IP quorum applications must be reconfigured and redeployed to hosts when certain aspects of the system configuration change. These aspects include adding or removing a node from the system, or when node service IP addresses are changed.

For stable quorum resolutions, an IP network must provide the following requirements:

- ▶ Connectivity from the hosts to the service IP addresses of all nodes. If IP quorum is configured incorrectly, the network must also deal with the possible security implications of exposing the service IP addresses because this connectivity also can be used to access the service GUI.
- ▶ Port 1260 is used by IP quorum applications to communicate from the hosts to all nodes.
- ▶ The maximum round-trip delay must not exceed 80 ms, which means 40 ms each direction.
- ▶ A minimum bandwidth of 2 MBps for node-to-quorum traffic.

Even with IP quorum applications at the third site, quorum disks at site one and site two are required because they are used to store metadata. To provide quorum resolution, run the `mkquorumapp` command or use the GUI in **Settings** → **Systems** → **IP Quorum** to generate a Java application that is then copied to and run on a host at a third site. A maximum of five applications can be deployed.

1.4.8 Storage pool

A *storage pool* is a collection of MDisks that provides the pool of storage from which volumes are provisioned. A single system can manage up to 1024 storage pools. The size of these pools can be changed (expanded or shrunk) at run time by adding or removing MDisks without taking the storage pool or the volumes offline.

At any point, an MDisk can be a member in one storage pool only, except for image mode volumes.

Figure 1-9 shows the relationships of the IBM SAN Volume Controller entities to each other.

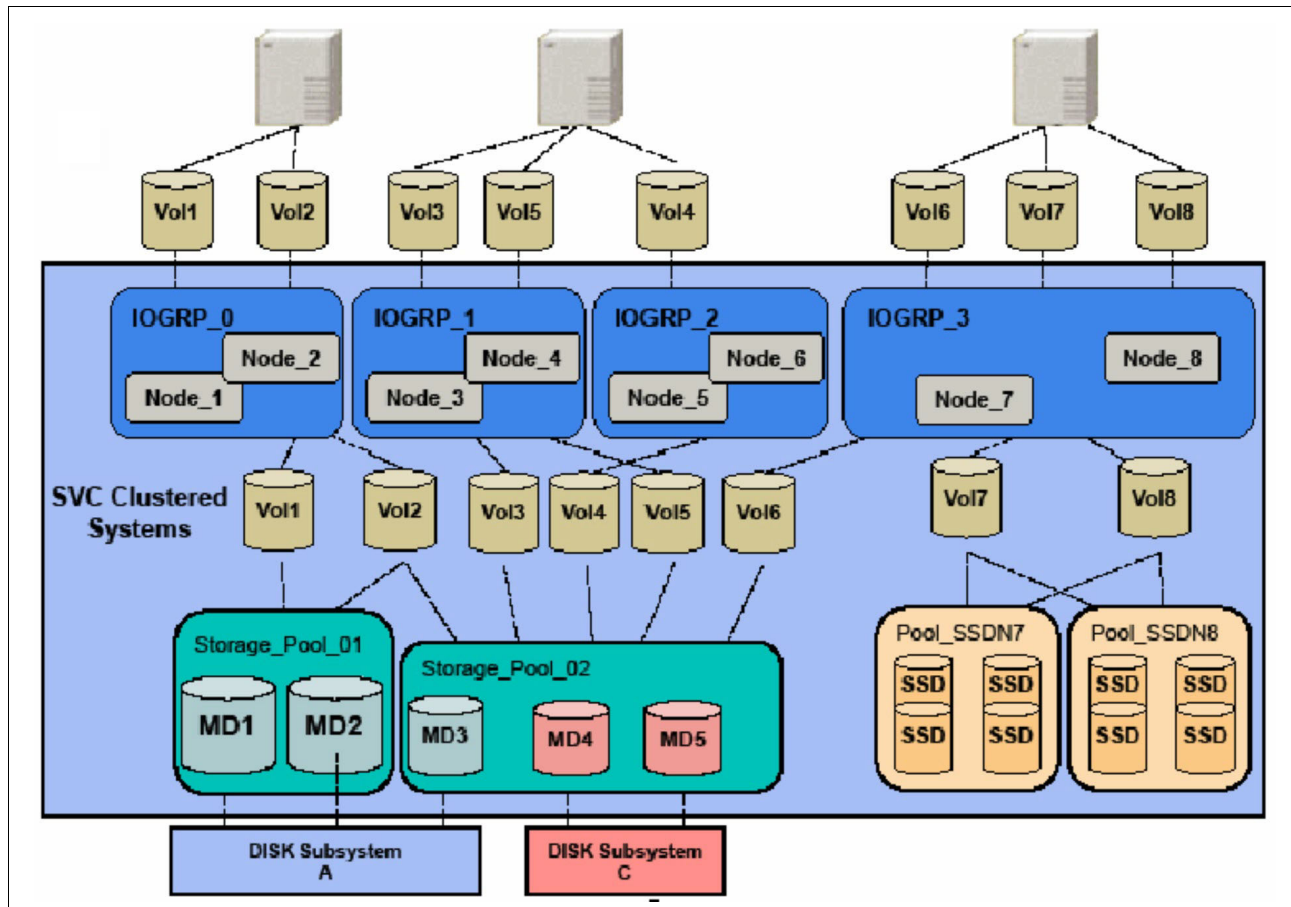


Figure 1-9 Overview of an IBM SAN Volume Controller clustered system with an I/O group

Each MDisk capacity in the storage pool is divided into several extents. The size of the extent is selected by the administrator when the storage pool is created and cannot be changed later. The size of the extent is 16 MiB - 8192 MiB.

It is a best practice to use the same extent size for all storage pools in a system. This approach is a prerequisite for supporting volume migration between two storage pools. If the storage pool extent sizes are not the same, you must use volume mirroring to copy volumes between pools.

The IBM SAN Volume Controller limits the number of extents in a system to $2^{22} \approx 4$ million. Because the number of addressable extents is limited, the total capacity of an IBM SAN Volume Controller system depends on the extent size that is chosen by the IBM SAN Volume Controller administrator.

1.4.9 Volumes

Volumes are logical disks that are presented to the host or application servers by the IBM SAN Volume Controller.

The following types of volumes are available in terms of extents management:

- **Striped**

A striped volume is allocated one extent in turn from each MDisk in the storage pool. This process continues until the space that is required for the volume is satisfied.

It is also possible to supply a list of MDisks to use.

Figure 1-10 shows how a striped volume is allocated (assuming that 10 extents are required).

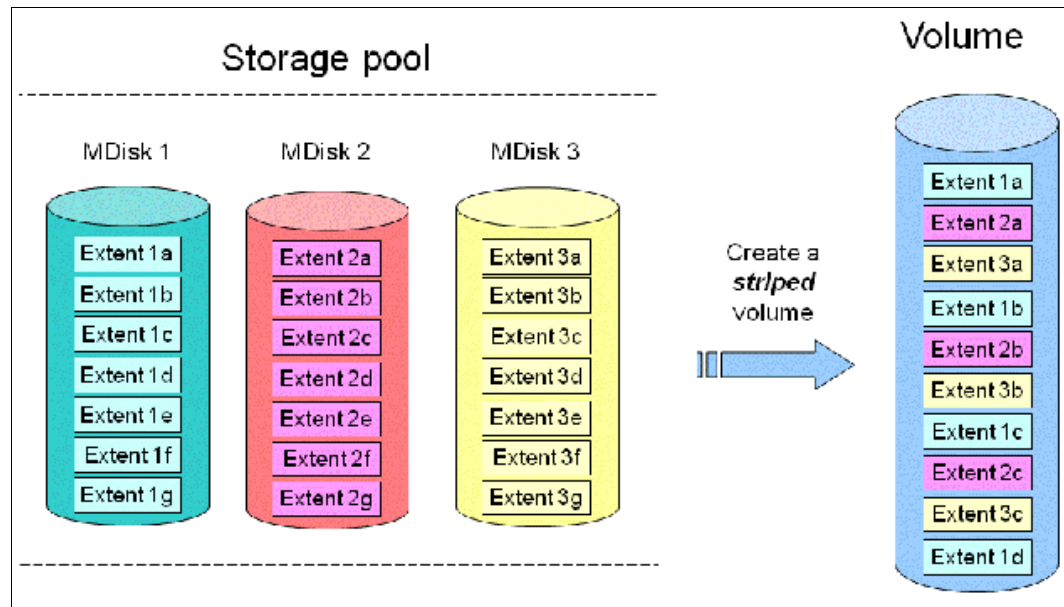


Figure 1-10 Striped volume

► Sequential

A sequential volume is where the extents are allocated sequentially from one MDisk to the next MDisk (see Figure 1-11).

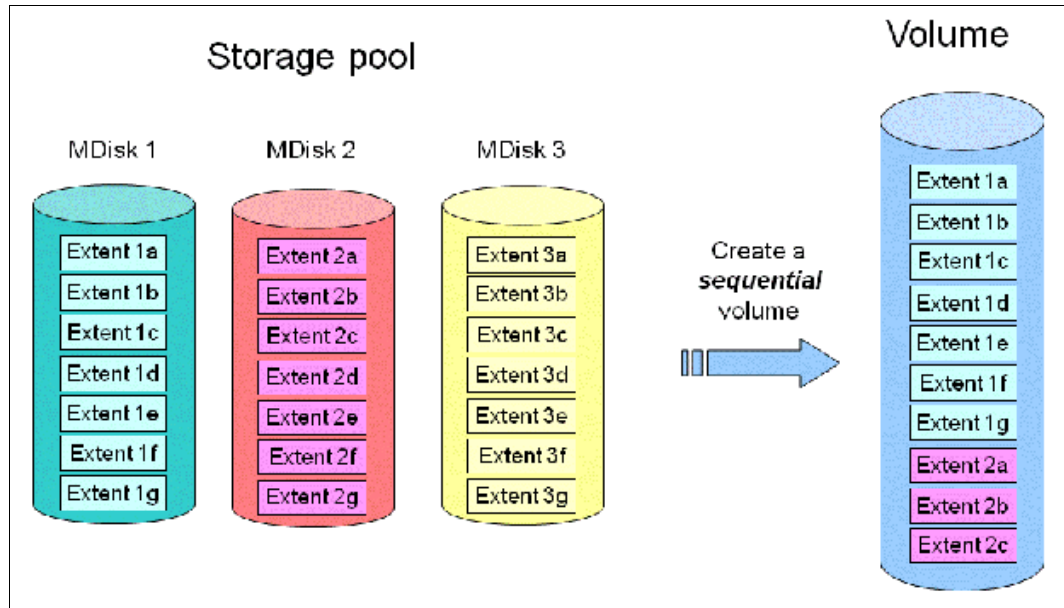


Figure 1-11 Sequential volume

► Image mode

Image mode volumes (see Figure 1-12) are special volumes that have a direct relationship with one MDisk. The most common use case of image volumes is a data migration from your old (typically nonvirtualized) storage to the IBM SAN Volume Controller virtualized infrastructure.

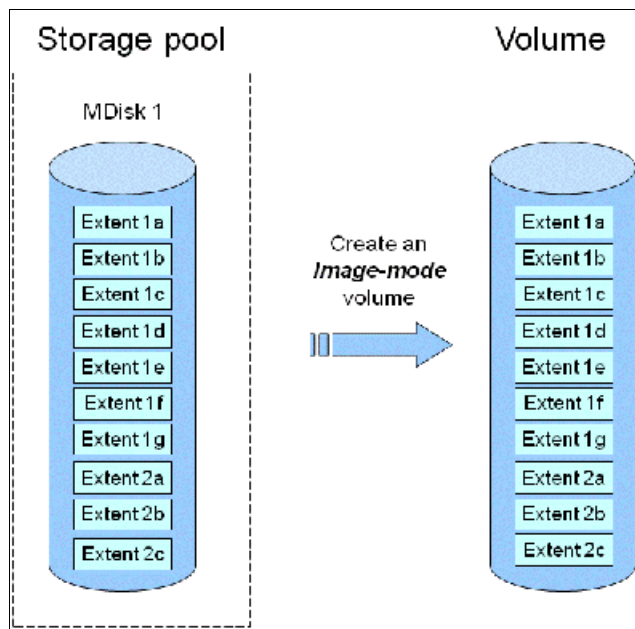


Figure 1-12 Image mode volume

When the image mode volume is created, a direct mapping is made between extents that are on the MDisk and the extents that are on the volume. The LBA x on the MDisk is the same as the LBA x on the volume, which ensures that the data on the MDisk is preserved as it is brought into the clustered system.

Because some virtualization functions are not available for image mode volumes, it is useful to migrate the volume into a new storage pool. After the migration completion, the MDisk becomes a managed MDisk.

If you add an MDisk that contains any historical data to a storage pool, all data on the MDisk is lost. Ensure that you create image mode volumes from MDisks that contain data before adding MDisks to the storage pools.

1.4.10 Easy Tier

Easy Tier is a performance function that automatically migrates extents off a volume to or from one MDisk storage tier to another MDisk storage tier.

Easy Tier monitors the host I/O activity and latency on the extents of all volumes with the Easy Tier function that is turned on in a multitier storage pool over a 24-hour period. Then, it creates an extent migration plan that is based on this activity and dynamically moves high-activity or hot extents to a higher disk tier within the storage pool. It also moves extents whose activity dropped off or cooled down from the high-tier MDisks back to a lower-tiered MDisk.

Easy Tier supports the new SCM drives with a new tier that is called `tier_scm`.

Turning on or off Easy Tier: The Easy Tier function can be turned on or off at the storage pool level and the volume level.

The automatic load-balancing function is enabled by default on each volume and cannot be turned off by using the GUI. This load-balancing feature is not considered an Easy Tier function, although it uses the same principles.

The management GUI supports monitoring Easy Tier data movement in graphical reports. The data in these reports helps you understand how Easy Tier manages data between the different tiers of storage, how tiers within pools are used, and the workloads among the different tiers. Charts for data movement, tier composition, and workload skew comparison can be downloaded as comma-separated value (CSV) files.

You can also offload the statistics file from the IBM SAN Volume Controller nodes and by using the IBM Storage Tier Advisor Tool (STAT) to create a summary report. The STAT can be downloaded for no initial cost from [IBM Storage Tier Advisor Tool](#).

1.4.11 Hosts

A *host* is a logical object that represents a list of worldwide port names (WWPNs), NVMe qualified names (NQN), or iSCSI or iSER names that identify the interfaces that the host system uses to communicate with the IBM SAN Volume Controller. Fibre Channel connections use WWPNs to identify host interfaces to the system. iSCSI or iSER names can be iSCSI qualified names (IQNs) or extended unique identifiers (EUIs). NQNs are used to identify hosts that use FC-NVMe connections.

Volumes can be mapped to a *host* to enable access to a set of volumes.

Node failover can be handled without having a multipath driver that is installed on the iSCSI server. An iSCSI-attached server can reconnect after a node failover to the original target IP address, which is now presented by the partner node. To protect the server against link failures in the network or HBA failures, a multipath driver *must* be used.

N_Port ID Virtualization (NPIV) is a method for virtualizing a physical Fibre Channel port that is used for host I/O. When NPIV is enabled, the partner node takes over the WWPN of the failing node. This takeover allows for rapid recovery of in-flight I/O when a node fails. In addition, path failures that occur because an offline node is masked from host multipathing.

Host cluster

A *host cluster* is a group of logical host objects that can be managed together. For example, you can create a volume mapping that is shared by every host in the host cluster. Host objects that represent hosts can be grouped in a host cluster and share access to volumes. New volumes can also be mapped to a host cluster, which simultaneously maps that volume to all hosts that are defined in the host cluster.

1.4.12 Array

An array is an ordered configuration, or group, of physical devices (drives) that is used to define logical volumes or devices. An array is a type of MDisk that is made up of disk drives (these drives are members of the array). A Redundant Array of Independent Disks (RAID) is a method of configuring member drives to create high availability (HA) and high-performance systems. The system supports *nondistributed* and *distributed* array configurations.

In *nondistributed* arrays, entire drives are defined as “hot-spare” drives. Hot-spare drives are idle and do not process I/O for the system until a drive failure occurs. When a member drive fails, the system automatically replaces the failed drive with a hot-spare drive. The system then resynchronizes the array to restore its redundancy.

However, all member drives within a *distributed* array have a rebuild area that is reserved for drive failures. All the drives in an array can process I/O data and provide faster rebuild times when a drive fails. The RAID level provides different degrees of redundancy and performance; it also determines the number of members in the array.

1.4.13 Encryption

The IBM SAN Volume Controller provides optional encryption of data at rest, which protects against the potential exposure of sensitive user data and user metadata that is stored on discarded, lost, or stolen storage devices. Encryption of system data and system metadata is not required; therefore, system data and metadata are not encrypted.

Planning for encryption involves purchasing a licensed function and then activating and enabling the function on the system.

To encrypt data that is stored on drives, the nodes that are capable of encryption must be licensed and configured to use encryption. When encryption is activated and enabled on the system, valid encryption keys must be present on the system when the system unlocks the drives or the user generates a new key.

Encryption keys can be managed by an external key management system. Supported products are:

- ▶ IBM Security® Guardium® Key Lifecycle Manager (formally IBM Security Key Lifecycle Manager),
- ▶ Thales CipherTrust Manager
- ▶ Gemalto KeySecure key servers

They can also be stored on USB flash drives that are attached to a minimum of one of the nodes. Since 8.1, IBM Storage Virtualize provides a combination of external and USB key repositories.

IBM Security Guardium Key Lifecycle Manager is an IBM solution that provides the infrastructure and processes to locally create, distribute, backup, and manage the lifecycle of encryption keys and certificates. Before activating and enabling encryption, you must determine the method of accessing key information during times when the system requires an encryption key to be present.

When Security Key Lifecycle Manager is used as a key manager for the IBM SAN Volume Controller encryption, you can encounter a deadlock situation if the key servers are running on encrypted storage that is provided by the IBM SAN Volume Controller. To avoid a deadlock situation, ensure that the IBM SAN Volume Controller can communicate with an encryption server to get the unlock key after a power-on or restart scenario. Up to four Security Key Lifecycle Manager servers are supported.

Although both Thales CipherTrust Manager and Gemalto KeySecure key servers support the same type of configurations, you need to ensure that you complete the prerequisites on these key servers before you can enable encryption on the system

Data encryption is protected by the Advanced Encryption Standard (AES) algorithm that uses a 256-bit symmetric encryption key in XTS mode, as defined in the Institute of Electrical and Electronics Engineers (IEEE) 1619-2007 standard as XTS-AES-256.¹ That data encryption key is protected by a 256-bit AES key wrap when it is stored in nonvolatile form.

Another data security enhancement, which is delivered with the Storage Virtualize 8.4.2 code and above, is the new Safeguarded Copy function that can provide protected read-only air gap copies of volumes. This enhancement gives the customer effective data protection against cyber attacks.

For more information, see *IBM FlashSystem Safeguarded Copy Implementation Guide*, [REDP-5654](#).

¹ [1619-2007 - IEEE Standard for Cryptographic Protection of Data on Block-Oriented Storage Devices](#)

1.4.14 iSCSI and iSCSI Extensions over RDMA

iSCSI is an alternative means of attaching hosts and external storage controllers to the IBM SAN Volume Controller.

The iSCSI function is a software function that is provided by the IBM Storage Virtualize software, not hardware. In 7.7, IBM introduced software capabilities to enable the underlying virtualized storage to attach to IBM SAN Volume Controller by using the iSCSI protocol.

The iSCSI protocol enables the transportation of SCSI commands and data over an IP network (TCP/IP), which is based on IP routers and Ethernet switches. iSCSI is a block-level protocol that encapsulates SCSI commands. Therefore, it uses an IP network rather than FC infrastructure.

The major functions of iSCSI include encapsulation and the reliable delivery of CDB transactions between initiators and targets through the IP network, especially over a potentially unreliable IP network.

Every iSCSI node in the network must have the following iSCSI components:

- ▶ An *iSCSI name* is a location-independent, permanent identifier for an iSCSI node. An iSCSI node has one iSCSI name, which stays constant for the life of the node. The terms *initiator name* and *target name* also refer to an iSCSI name.
- ▶ An *iSCSI address* specifies the iSCSI name of an iSCSI node and a location of that node. The address consists of a hostname or IP address, a TCP port number (for the target), and the iSCSI name of the node. An iSCSI node can have any number of addresses, which can change at any time, particularly if they are assigned by way of Dynamic Host Configuration Protocol (DHCP). An IBM SAN Volume Controller node represents an iSCSI node and provides statically allocated IP addresses.

IBM SAN Volume Controller models SV3, SV2, and SA2 support 25 Gbps Ethernet adapters that provide iSCSI and iSCSI Extensions over RDMA (iSER) connections. The IBM SAN Volume Controller models SV3 also supports 100 Gbps Ethernet adapters.

iSER is a network protocol that extends the iSCSI protocol to use RDMA. You can implement RDMA-based connections that use Ethernet networking structures and connections without upgrading hardware. As of this writing, the system supports RDMA-based connections with RDMA over Converged Ethernet (RoCE) or Internet-Wide Area RDMA Protocol (iWARP).

For host attachment, these 25 Gbps adapters support iSCSI and RDMA-based connections; however, for external storage systems, only iSCSI connections are supported through these adapters. When the 25 Gbps adapter is installed on nodes in the system, RDMA technology can be used for node-to-node communications.

Note: The 100 Gbps adapter on the IBM SAN Volume Controller models SV3 supports iSCSI. However, the performance is limited 25 Gbps per port.

1.4.15 Data reduction pools

Data reduction pools (DRPs) represent a significant enhancement to the storage pool concept. The reason is that the virtualization layer is primarily a simple layer that runs the task of lookups between virtual and physical extents.

DRPs are a new type of storage pool that implements various techniques, such as thin-provisioning, compression, and deduplication, to reduce the amount of physical capacity that is required to store data. Savings in storage capacity requirements translate into the reduction of the cost of storing the data.

By using DRPs, you can automatically de-allocate and reclaim the capacity of thin-provisioned volumes that contain deleted data and enable this reclaimed capacity to be reused by other volumes. Data reduction provides more capacity from compressed volumes because of the implementation of the new log-structured array.

Deduplication

Data deduplication is one of the methods of reducing storage needs by eliminating redundant copies of data. Data reduction is a way to decrease the storage disk infrastructure that is required, optimize the usage of storage disks, and improve data recovery infrastructure efficiency.

Existing data or new data is standardized into chunks that are examined for redundancy. If data duplicates are detected, pointers are shifted to reference a single copy of the chunk, and the duplicate data sets are then released.

To estimate potential capacity savings that data reduction can provide on the system, use the Data Reduction Estimation Tool (DRET). This tool scans target workloads on all attached storage arrays, consolidates these results, and generates an estimate of potential data reduction savings for the entire system.

The DRET is available for download at [IBM Data Reduction Estimator Tool \(DRET\) for SVC, Storwize and FlashSystem products](#).

1.4.16 IP replication

IP replication was introduced in 7.2. It enables data replication between IBM Storage Virtualize family members. IP replication uses the IP-based ports of the cluster nodes.

The IP replication function is transparent to servers and applications in the same way as traditional FC-based mirroring. All remote mirroring modes (MM, GM, and GMCV and the new policy-based replication) are supported.

The configuration of the system is straightforward. IBM Storage Virtualize family systems normally “find” each other in the network and can be selected from the GUI.

IP replication includes Bridgeworks SANSlide network optimization technology, and it is available at no additional charge. Remote mirror is a chargeable option, but the price does not change with IP replication. Existing remote mirror users can access the function at no extra charge.

IP connections that are used for replication can have long latency (the time to transmit a signal from one end to the other), which can be caused by distance or by many “hops” between switches and other appliances in the network. Traditional replication solutions transmit data, wait for a response, and then transmit more data, which can result in network utilization as low as 20% (based on IBM measurements). In addition, this scenario worsens the longer the latency.

Bridgeworks SANSlide technology, which is integrated with the IBM Storage Virtualize family, requires no separate appliances and incurs no extra cost and configuration steps. It uses artificial intelligence (AI) technology to transmit multiple data streams in parallel, adjusting automatically to changing network environments and workloads.

SANSlide improves network bandwidth usage up to 3x. Therefore, customers can deploy a less costly network infrastructure, or take advantage of faster data transfer to speed replication cycles, improve remote data currency, and enjoy faster recovery.

1.4.17 IBM Storage Virtualize copy services

IBM Storage Virtualize supports the following copy services functions:

- ▶ Remote copy (synchronous or asynchronous)
- ▶ Policy-based-replication
- ▶ FlashCopy (PiT copy) and Transparent Cloud Tiering (TCT)
- ▶ HyperSwap

Copy services functions are implemented within a single IBM SAN Volume Controller, or between multiple members of the IBM Storage Virtualize family.

The copy services layer sits above and operates independently of the function or characteristics of the underlying disk subsystems that are used to provide storage resources to an IBM SAN Volume Controller.

1.4.18 Synchronous or asynchronous remote copy

The general application of remote copy seeks to maintain two copies of data. Often, the two copies are separated by distance, but not always. The remote copy can be maintained in synchronous or asynchronous modes. IBM Storage Virtualize, Metro Mirror, and Global Mirror are the IBM-branded terms for the functions that are synchronous remote copy and asynchronous remote copy.

Synchronous remote copy ensures that updates are committed at the primary and secondary volumes before the application considers the updates complete. Therefore, the secondary volume is fully dated if it is needed in a failover. However, the application is fully exposed to the latency and bandwidth limitations of the communication link to the secondary volume. In a truly remote situation, this extra latency can have a significant adverse effect on application performance.

Special configuration guidelines exist for SAN fabrics and IP networks that are used for data replication. Consider the distance and available bandwidth of the intersite links.

A function of Global Mirror for low bandwidth was introduced in IBM Storage Virtualize 6.3. It uses change volumes that are associated with the primary and secondary volumes. These points in time copies are used to record changes to the remote copy volume, the FlashCopy map that exists between the secondary volume and the change volume, and between the primary volume and the change volume. This function is called *Global Mirror with change volumes (cycling mode)*.

Figure 1-13 shows an example of this function where you can see the relationship between volumes and change volumes.

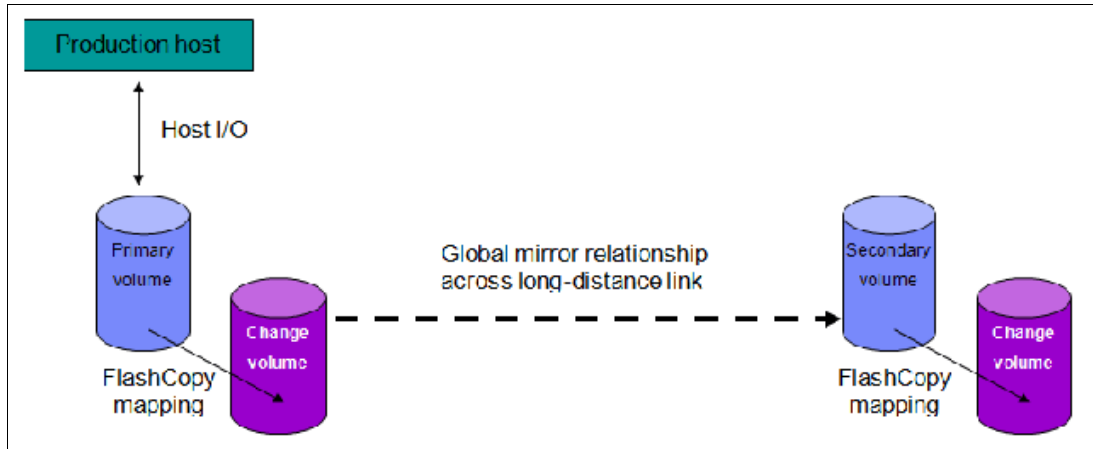


Figure 1-13 Global Mirror with change volumes

In asynchronous remote copy, the application acknowledges that the write is complete before the write is committed at the secondary volume. Therefore, on a failover, specific updates (data) might be missing at the secondary volume.

The application must have an external mechanism for recovering the missing updates, if possible. This mechanism can involve user intervention. Recovery on the secondary site involves starting the application on this recent backup, and then rolling forward or backward to the most recent commit point.

1.4.19 Policy-based replication

Policy-based replication (PBR) is a new feature that provides simplified configuration and management of asynchronous replication between two systems.

Policy-based replication uses volume groups to automatically deploy and manage replication. This feature significantly simplifies configuring, managing, and monitoring replication between two systems. Policy-based replication simplifies asynchronous replication with the following key advantages:

- ▶ Uses volume groups instead of consistency groups. With volume groups, all volumes are replicated based on the assigned policy.
- ▶ Simplifies administration by removing the need to manage relationships and change volumes.
- ▶ Automatically manages provisioning on the remote system.
- ▶ Supports easier visualization of replication during a site failover.
- ▶ Automatically notifies you when the recovery point objective (RPO) is exceeded.
- ▶ Easy-to-understand status and alerts on the overall health of replication.

Policy-based replication is supported in 8.5.2 or later on the following products:

- ▶ IBM SAN Volume Controller
- ▶ IBM Storage FlashSystem 9500
- ▶ IBM Storage FlashSystem 9200
- ▶ IBM Storage FlashSystem 9100
- ▶ IBM Storage FlashSystem 7300

- ▶ IBM Storage FlashSystem 7200
- ▶ IBM Storage FlashSystem 5200
 - Requires a minimum of 128 GiB memory in each node canister
- ▶ IBM Storage Virtualize for Public Cloud

For more information about concepts and objects related to PBR, see [Policy-based replication](#).

For more information about PBR and its implementation, see *Policy-Based Replication with IBM Storage FlashSystem, IBM SAN Volume Controller and IBM Storage Virtualize*, REDP-5704-00.

1.4.20 FlashCopy and Transparent Cloud Tiering

FlashCopy and TCT are used to make a copy of a source volume on a target volume. After the copy operation starts, the original content of the target volume is lost, and the target volume has the contents of the source volume as they existed at a single point in time. Although the copy operation takes time, the resulting data at the target appears as though the copy was made instantaneously.

FlashCopy

FlashCopy is sometimes described as an instance of a time-zero (T0) copy or a point-in-time (PiT) copy technology.

FlashCopy can be performed on multiple source and target volumes. FlashCopy enables management operations to be coordinated so that a common single PiT is chosen for copying target volumes from their respective source volumes.

With IBM Storage Virtualize, multiple target volumes can undergo FlashCopy from the same source volume. This capability can be used to create images from separate PiTs for the source volume, and to create multiple images from a source volume at a common PiT.

Reverse FlashCopy enables target volumes to become restore points for the source volume without breaking the FlashCopy relationship, and without waiting for the original copy operation to complete. IBM Storage Virtualize supports multiple targets and multiple rollback points.

Most customers aim to integrate the FlashCopy feature for PiT copies and quick recovery of their applications and databases. An IBM solution for this goal is provided by IBM Storage Protect and IBM Copy Data Management.

Transparent Cloud Tiering

IBM Storage Virtualize TCT is an alternative solution for data protection, backup, and restores that interfaces to Cloud Services Providers (CSPs), such as IBM Cloud®. The TCT function helps organizations to reduce costs that are related to power and cooling when offsite data protection is required to send sensitive data out of the main site.

TCT uses IBM FlashCopy techniques that provide full and incremental snapshots of several volumes. Snapshots are encrypted and compressed before being uploaded to the cloud. Reverse operations are also supported within that function. When a set of data is transferred out to cloud, the volume snapshot is stored as object storage.

IBM Cloud Object Storage uses an innovative approach and a cost-effective solution to store a large amount of unstructured data. It also delivers mechanisms to provide security services, HA, and reliability.

The management GUI provides an easy- to-use initial setup, advanced security settings, and audit logs that record all backup and restore to cloud.

For more information, see [IBM Cloud Object Storage](#).

1.4.21 IBM HyperSwap

The IBM HyperSwap function is an HA feature that provides dual-site access to a volume. When you configure a system with IBM HyperSwap topology, the system configuration is split between two sites for data recovery, migration, or HA use cases.

When an HyperSwap topology is configured, each node, external storage system, and host in the system configuration must be assigned to one of the sites in the topology. Both nodes of an I/O group must be at the same site. This site must be the same site as the external storage systems that provide the managed disks to that I/O group.

When managed disks are added to storage pools, their site attributes must match. This requirement ensures that each copy in a HyperSwap volume is fully independent and is at a distinct site.

When the system is configured between two sites, HyperSwap volumes have a copy at one site and a copy at another site. Data that is written to the volume is automatically sent to both copies. If one site is no longer available, the other site can provide access to the volume. If ownership groups are used to manage access to HyperSwap volumes, both volume copies and users who access them must be assigned to the same ownership group.

A 2-site HyperSwap configuration can be extended to a third site for DR that uses the IBM Storage Virtualize 3-Site Orchestrator. For more information, see *IBM Storage Virtualize 3-Site Replication*, [SG24-8504](#).

1.5 IBM SAN Volume Controller models

The following IBM SAN Volume Controller models are supported at the IBM Storage Virtualize 8.6 code level:

- ▶ SV3
- ▶ SV2
- ▶ SA2

1.5.1 IBM SAN Volume Controller SV3

Figure 1-14 shows the front view of the IBM SAN Volume Controller SV3.

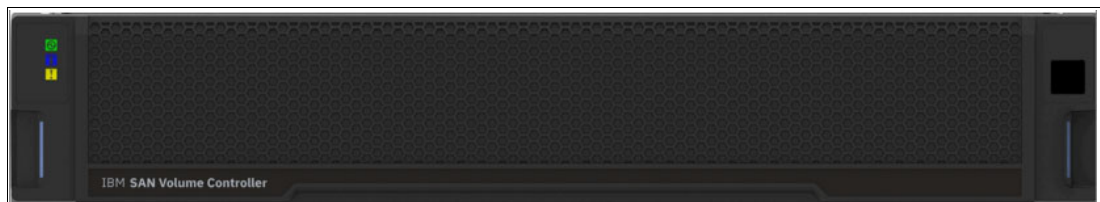


Figure 1-14 IBM SAN Volume Controller SV3 front view

Figure 1-15 shows the rear view of the IBM SAN Volume Controller SV3.

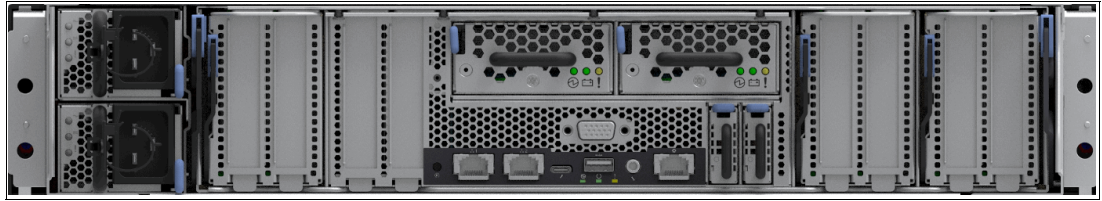


Figure 1-15 IBM SAN Volume Controller SV3 rear view

Figure 1-16 shows the internal hardware components of an IBM SAN Volume Controller SV3 node canister. To the left is the front of the canister where fan modules are located, followed by two Ice Lake CPUs and Dual Inline Memory Module (DIMM) slots. The battery backup units and the PCIe adapter cages are shown on the right side. Each of these adapters cages holds two PCIe adapter cards, except for cage number 2, which is dedicated to the compression card.

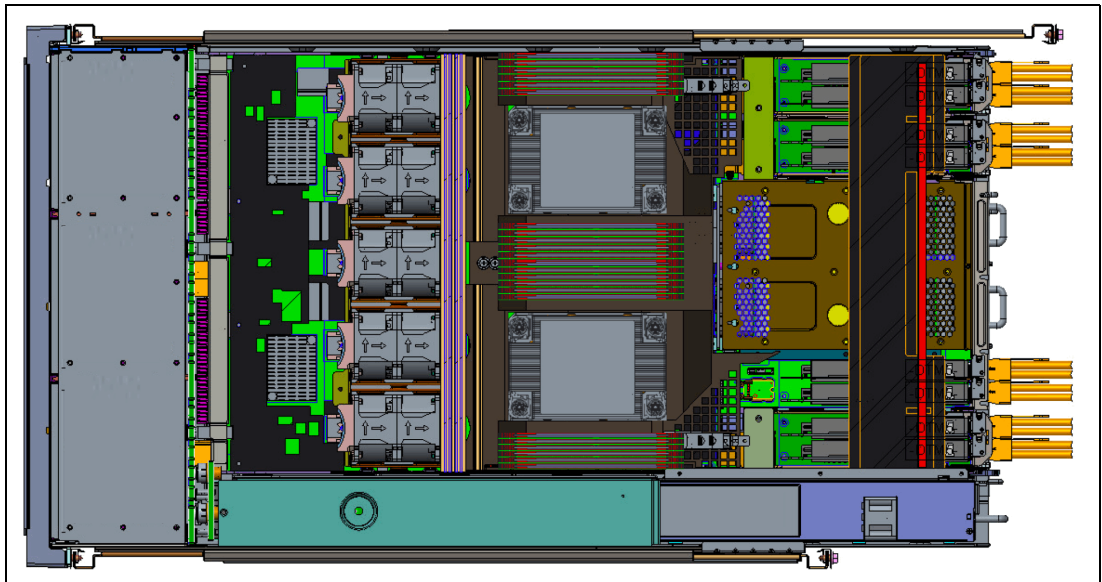


Figure 1-16 IBM SAN Volume Controller SV3 internal hardware components

Figure 1-17 shows the internal architecture of the IBM SAN Volume Controller SV3 model. You can see that the PCIe switch is still present, but has no outbound connections because these models do not support any internal drives. The PCIe switch is used for internal functions and monitoring purposes within the IBM SAN Volume Controller enclosure.

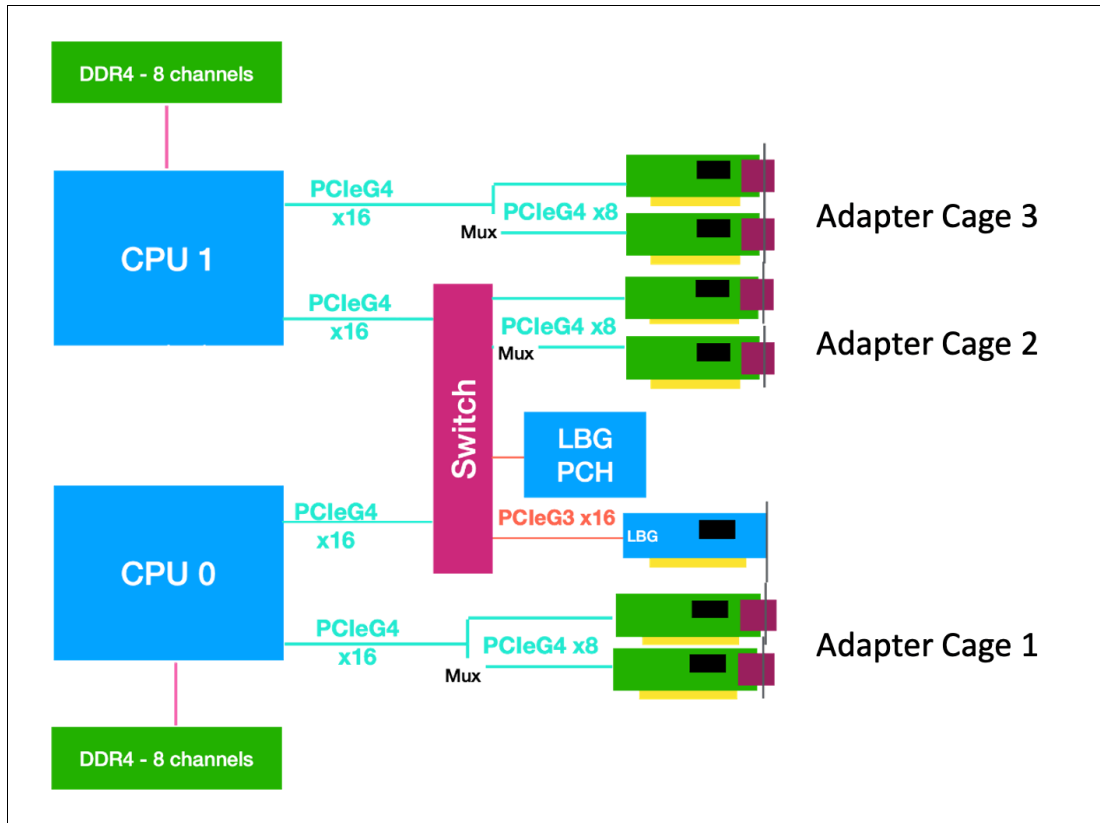


Figure 1-17 IBM SAN Volume Controller SV3 internal architecture

1.5.2 IBM SAN Volume Controller SV2 and SA2

Figure 1-18 shows the front view of the IBM SAN Volume Controller SV2 and SA2.



Figure 1-18 IBM SAN Volume Controller SV2 and SA2 front view

Figure 1-19 shows the rear view of the IBM SAN Volume Controller SV2 and SA2.

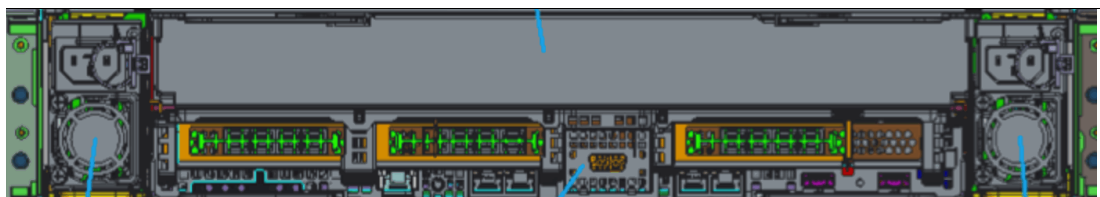


Figure 1-19 IBM SAN Volume Controller SV2 and SA2 rear view

Figure 1-20 shows the internal hardware components of an IBM SAN Volume Controller SV2 and SA2 node canister. To the left is the front of the canister where fan modules and battery backup are located, followed by two Cascade Lake CPUs and Dual Inline Memory Module (DIMM) slots, and PCIe risers for adapters on the right.

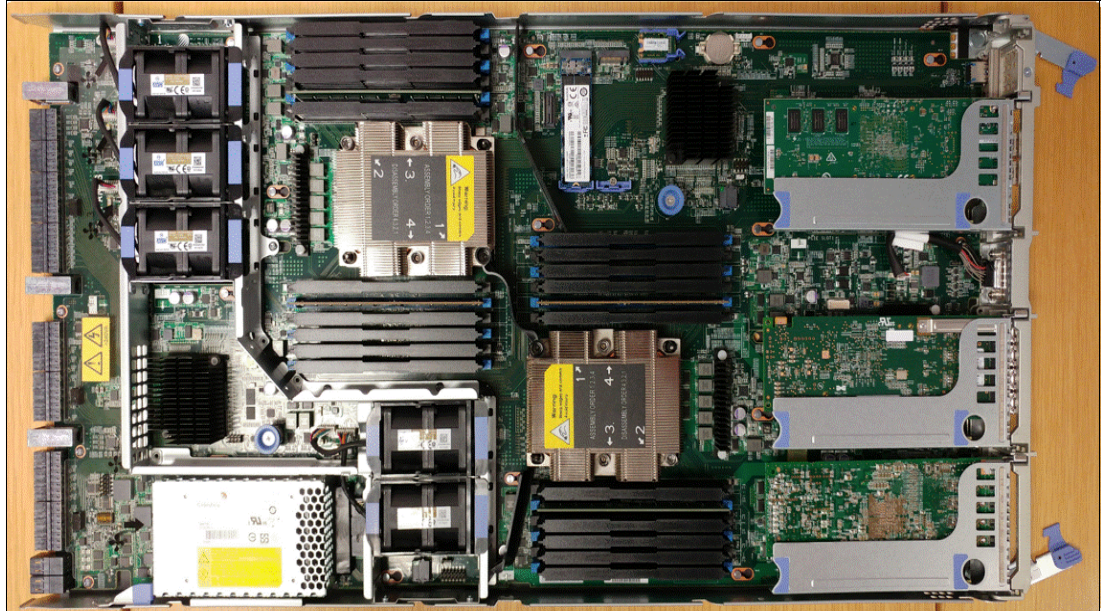


Figure 1-20 Internal hardware components

Figure 1-21 shows the internal architecture of the IBM SAN Volume Controller SV2 and SA2 models. You can see that the PCIe switch is still present, but has no outbound connections because these models do not support any internal drives. The PCIe switch is used for internal monitoring purposes within the IBM SAN Volume Controller enclosure.

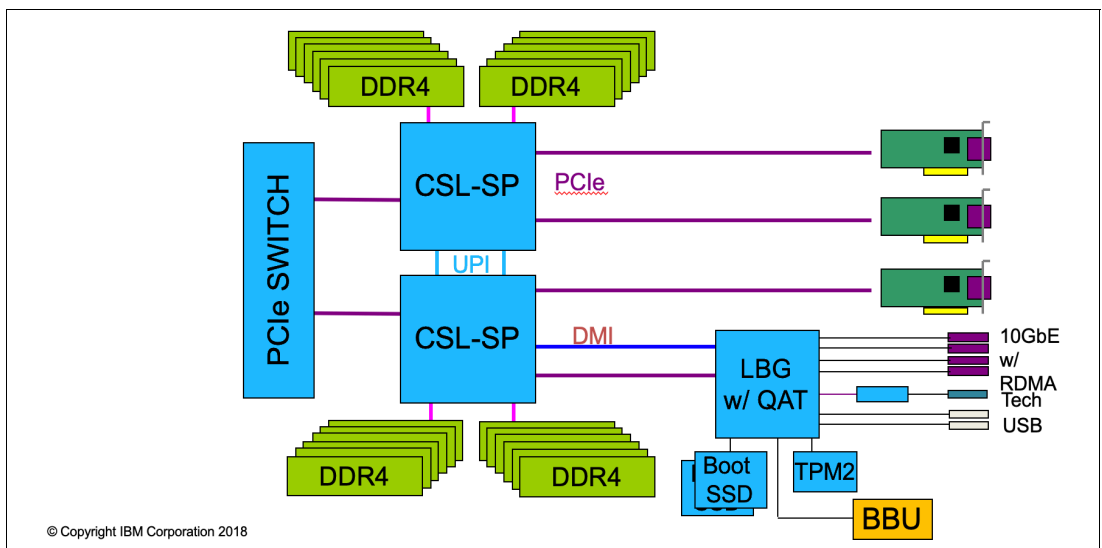


Figure 1-21 IBM SAN Volume Controller SV2 and SA2 internal architecture

Note: IBM SAN Volume Controller SV3, SV2 and SA2 do not support any type of expansion enclosures.

1.5.3 IBM SAN Volume Controller model comparisons

All of the IBM SAN Volume Controller models are delivered in a 2U 19-inch rack-mounted enclosure. At the time of this writing, three models of the IBM SAN Volume Controller are available, as listed in Table 1-2.

More information: For the most up-to-date information about features, benefits, and specifications of the IBM SAN Volume Controller models, see [IBM SAN Volume Controller](#).

The information in this book is valid at the time of this writing and covers IBM Storage Virtualize 8.6. However, as IBM SAN Volume Controller matures, expect to see new features and enhanced specifications.

Table 1-2 IBM SAN Volume Controller base models

Feature	Models		
	2145/2147-SV2	2145/2147-SA2	2145/2147-SV3
Processor	Two Intel Cascade Lake 5218 Series, 16-cores, and 2.30 GHz (Gold)	Two Intel Cascade Lake 4208 Series, 8-cores, and 2.10 GHz (Silver)	Two Intel Ice Lake 4189 Series, 24-cores, and 2.4 GHz (Gold)
Base cache memory	128 GB	128 GB	512 GB
I/O ports and management	Four 10 Gb Ethernet ports for 10 Gb iSCSI connectivity and system management	Four 10 Gb Ethernet ports for 10 Gb iSCSI connectivity and system management	Two 1 Gb Ethernet ports for system management only (non-iSCSI ports)
Technician port	Single 1 Gb Ethernet	Single 1 Gb Ethernet	Single 1 Gb Ethernet
Maximum host interface adapters slots	3	3	6
USB ports	2	2	1
SAS chain	N/A	N/A	N/A
Max number of dense drawers per SAS chain	N/A	N/A	N/A
Integrated battery units	1	1	2
Power supplies and cooling units	2	2	2

The following optional features are available for IBM SAN Volume Controller SV2 and SA2:

- ▶ A 768 GB cache upgrade
- ▶ A 4-port 16 Gb FC/FC over NVMe adapter for 16 Gb FC connectivity
- ▶ A 4-port 32 Gb FC/FC over NVMe adapter for 32 Gb FC connectivity
- ▶ A 2-port 25 Gb iSCSI/iSER/RDMA over Converged Ethernet (RoCE)
- ▶ A 2-port 25 Gb iSCSI/iSER/Internet Wide-area RDMA Protocol (iWARP)

The SV2 and SA2 systems have dual CPU sockets and three adapter slots along with four 10-GbE RJ45 ports on board.

Note: IBM SAN Volume Controller models SA2 and SV2 do not support FCoE.

The following optional features are available for IBM SAN Volume Controller SV3:

- ▶ A 1536 GB cache upgrade
- ▶ A 4-port 32 Gb FC/FC over NVMe adapter for 32 Gb FC connectivity
- ▶ A 2-port 25 Gb iSCSI/iSER/RDMA over Converged Ethernet (RoCE)
- ▶ A 2-port 25 Gb iSCSI/iSER/Internet Wide-area RDMA Protocol (iWARP)
- ▶ A 2-port 100 Gb NVMe/iSCSI/RDMA over Converged Ethernet (RoCE)

Note: The 25 and 100 Gb adapters are NVMe capable; however, to support NVMe, a software dependency exists (at the time of this writing). Therefore, NVMe/NVMeoF is *not* supported on these cards.

All Ethernet cards can be used with the iSCSI protocol. 25 Gb iWARP Ethernet cards can also be used for clustering. 25 Gb and 100 Gb RoCE Ethernet cards can be used for NVMe RDMA.

The comparison of current and previous models of IBM SAN Volume Controller is shown in Table 1-3. Expansion enclosures are not included in the list.

Table 1-3 Historical overview of IBM SAN Volume Controller models

Model	Cache (GB)	FC (Gbps)	iSCSI (Gbps)	Hardware base	Announced
2145-SV2	128 - 768	16 and 32	25, 50, and 100	Intel Xeon Cascade Lake	06 March 2020
2147-SV2	128 - 768	16 and 32	25, 50, and 100	Intel Xeon Cascade Lake	06 March 2020
2145-SA2	128 - 768	16 and 32	25, 50, and 100	Intel Xeon Cascade Lake	06 March 2020
2147-SA2	128 - 768	16 and 32	25, 50, and 100	Intel Xeon Cascade Lake	06 March 2020
2145-SV3	512 - 1536	32	25 and 100 by way of PCIe adapters only	Intel Xeon Ice Lake	08 March 2022
2147-SV3	512 - 1536	32	25 and 100 by way of PCIe adapters only	Intel Xeon Ice Lake	08 March 2022

Note: IBM SAN Volume Controller SV3, SV2, and SA2 do not support any type of SAS expansion enclosures.

1.6 IBM FlashSystem family

The IBM FlashSystem family, running IBM Storage Virtualize software, was simplified with innovations and enterprise-class features for deployments of all sizes, from entry to mid-range to high-end. A one-platform system allows for ease-of-use to manage seamlessly and securely data across your entire IT infrastructure.

IBM FlashSystem 5015, IBM FlashSystem 5035, IBM FlashSystem 5045 and IBM FlashSystem 5200 deliver entry enterprise solutions. IBM FlashSystem 7200 and 7300 provides a midrange enterprise solution. IBM FlashSystem 9200 and 9500 plus the rack-based IBM FlashSystem 9200R and 9500R provide four high-end enterprise solutions.

Although all the IBM FlashSystem family systems are running the same IBM Storage Virtualize software, the feature set that is available with each of the models is different.

Figure 1-22 shows the IBM FlashSystem family.

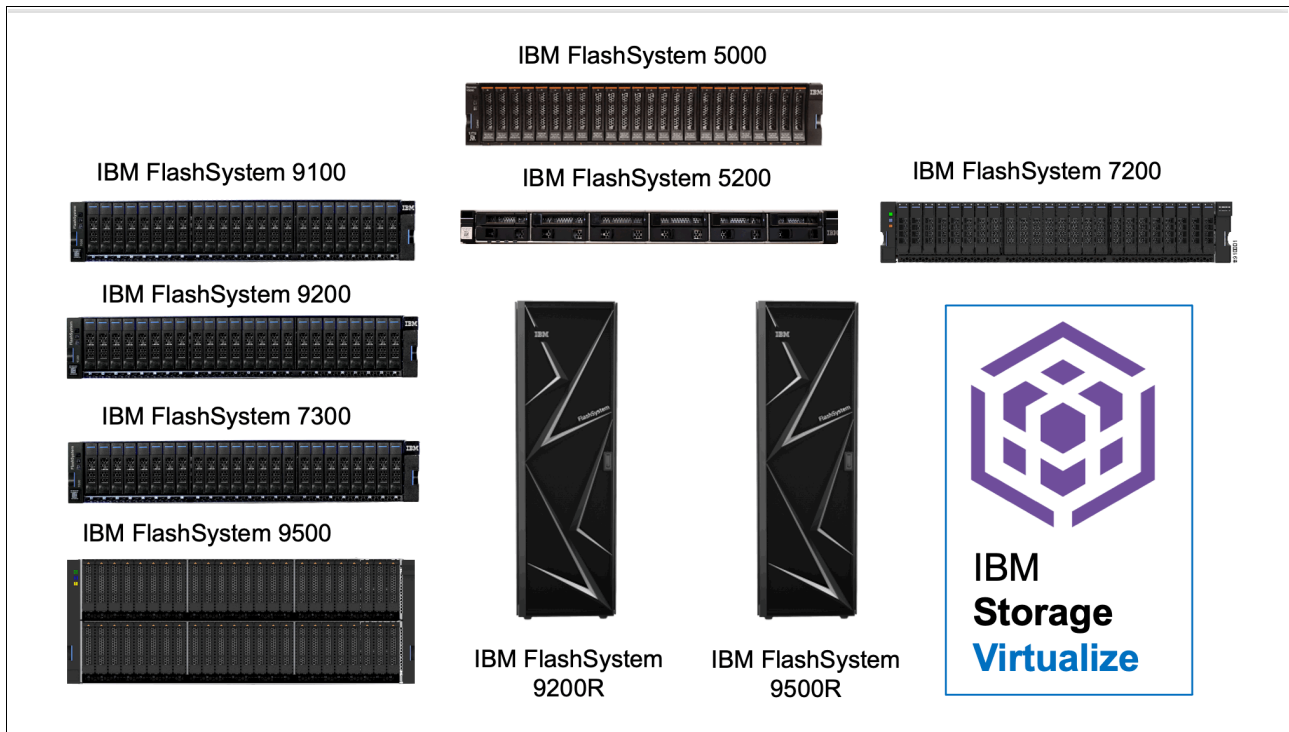


Figure 1-22 IBM FlashSystem family

Note: The IBM FlashSystem 9100 Models AF7, AF8, UF7, and UF8 plus IBM FlashSystem 7200 Model 824 and U7C are no longer sold by IBM, but are included here for completeness because they support IBM Storage Virtualize 8.6 software.

For more information, see the [IBM FlashSystem Family Data Sheet](#).

1.6.1 Storage Expert Care

IBM recently expanded the new service offering that is called *Storage Expert Care*, which allows flexible levels and duration of support contracts to supplement specific machine types and models in the IBM FlashSystems family.

IBM Storage Expert Care is designed to simplify and standardize the support approach on the IBM FlashSystem portfolio to keep customer's systems operating at peak performance.

The Storage Expert Care offering was originally released with the IBM FlashSystem 5200 and now also covers the IBM FlashSystems 7200, 7300, 9200/R, and 9500/R.

Customers can now choose their preferred level of support from up to three tiers (product-dependent), each priced as a simple percentage of the hardware sales price. This feature allows for easy, straightforward quoting from a single system.

These three tiers allow customers to select the best level of required service to support their environment, ranging from base level service, through to premium-enhanced service. This Storage Expert Care offering is designed to improve product resiliency and reliability and reduce the operational costs that are associated with managing and maintaining increasingly complex and integrated IT environments.

The following tier selection and features are available:

- ▶ Basic:
 - Hardware maintenance with next business day onsite response
 - Software support and services
- ▶ Advanced:
 - Hardware maintenance with 24x7 same business day onsite response
 - Software support and services
 - Storage Insights predictive support
- ▶ Premium:
 - Hardware maintenance with 24x7 same business day onsite response
 - Software support and services
 - Machine setup services
 - Predictive support
 - Enhanced response time for defect support
 - Hardware remote code load
 - Access to a dedicated technical account manager
- ▶ Committed Maintenance - CMSL (optional):
 - Enables IBM Hardware Maintenance Services, which is committed maintenance to be included on top of IBM Storage Expert Care Advanced and Premium Tiers.
 - Committed maintenance reduces the cost of downtime by providing a committed time frame to call back, arrive on site, or repair.
 - Reduces the loss of revenue, repair costs, and loss of consumer confidence and shareholder trust by making sure that your products are protected by committed maintenance.

Figure 1-23 shows a summary of the Storage Expert Care Tier Levels.

IBM Storage Expert Care on 5015 & 5045

Extending simplified support offerings

	Warranty	Basic 5015, 5045	Advanced 5015, 5045
IBM SW (Storage Virtualize) fixes, updates and new releases	1 year	Yes	Yes
Guidance on installation, usage and configuration (Support Line)		Yes	Yes
Automated ticket management and alerting		Yes	Yes
Use of Storage Insights for collaborative problem resolution		Yes	Yes
Predictive issue alerting			Yes
IBM Installation		Additional paid service	Additional paid service
Remote code updates (2x year) ***			
Hardware service / parts replacement		9x5 NBD, IBM on-site	24x7 Same day, IBM on-site

Figure 1-23 Storage Expert Care tier levels for IBM FlashSystem 5015 and IBM FlashSystem 5045

Note: Not all geographies and regions offer all the Storage Expert Care levels of support. If the Storage Expert Care is not announced in a specific country, the traditional warranty and maintenance options are still offered.

For more information about countries in which it is applicable, see the following:

- ▶ [FS5200 Announcement Letter](#)
- ▶ [FS7200 Announcement Letter](#)
- ▶ [FS9200 Announcement Letter](#)
- ▶ [FS7300 and FS9500 Announcement Letters](#)

To support the new Storage Expert Care offering on the older IBM FlashSystems 7200 and 9200, new machine types and models were introduced for these products.

Table 1-4 lists the comparison of the old machine types with the traditional warranty and maintenance offering and the new Storage Expert Care offering.

Table 1-4 IBM FlashSystems 7200 and 9200 product range

Product	Previous Machine Type	Expert Care Machine Type	Model	Function
IBMFlashSystem 7200	2076	4664	824, U7C	Control enclosures
	2076	4664	12G, 24G, 92G	Expansion enclosures
IBMFlashSystem 9200	9846 / 9848	4666	AG8, UG8,	Control enclosures
	9846 / 9848	4666	AFF, A9F	Expansion enclosures

Table 1-5 lists the software PIDs and SWMA feature codes that must be added to the order, depending on the required level of cover.

Table 1-5 Software PIDs and SWMA feature codes

Product	Software type	Control enclosure	Expansion enclosure
FlashSystem 7200	New SW PIDs License	5639-F7K (controller)	5639-EC1 (expansion)
FlashSystem 9200	New SW PIDs License	5639-EA4 (controller)	5639-EB2 (expansion)
FlashSystem 7200	3 Month SWMA	5639-EA2 (controller)	5639-EC2 (expansion)
FlashSystem 9200	3 Month SWMA	5639-EA3 (controller)	5639-EB3 (expansion)

When selecting the level of Storage Expert Care, you also must select the duration of the contract, which can be 1 - 5 years. You also can opt for committed maintenance service levels (CMSL).

The contract and duration has its own machine types and models (in addition to the hardware machine type and model that are listed in Table 1-5):

- ▶ FS7200:
 - 4665-P01-05 for Premium
 - 4665-Pxx for Premium with CMSL
- ▶ FS9200:
 - 4673-P01-05 for Premium
 - 4673-Pxx for Premium with CMSL

For example, an FS9200 with Premium Expert care for three years is 4673-PX3, where:

- ▶ P: Premium Level service
- ▶ X: Reserved for committed services (CMSL) if added to the expert care contract
- ▶ 3: Denotes a three-year contract (if 0, no committed services were purchased)

For more information, see the following:

- ▶ [IBM FlashSystem 5200 Storage Expert Care](#)
- ▶ [IBM FlashSystem 7200 Storage Expert Care](#)
- ▶ [IBM FlashSystem 9200 Storage Expert Care](#)
- ▶ [IBM FlashSystem 7300 Storage Expert Care](#)
- ▶ [IBM FlashSystem 9500 Storage Expert Care](#)
- ▶ [IBM FlashSystem 50x5 Storage Expert Care](#)

1.7 IBM FlashSystem 9500 overview

This section describes the IBM FlashSystem 9500 architectural components, available models, and enclosure and software features.

1.7.1 IBM FlashSystem 9500 hardware components

IBM FlashSystem 9500 is an all-flash storage system that consists of a control enclosure that runs the IBM Storage Virtualize Software and manages your storage system, communicates with the hosts, and manages interfaces.

Figure 1-24 shows the IBM FlashSystem 9500 front and rear views.

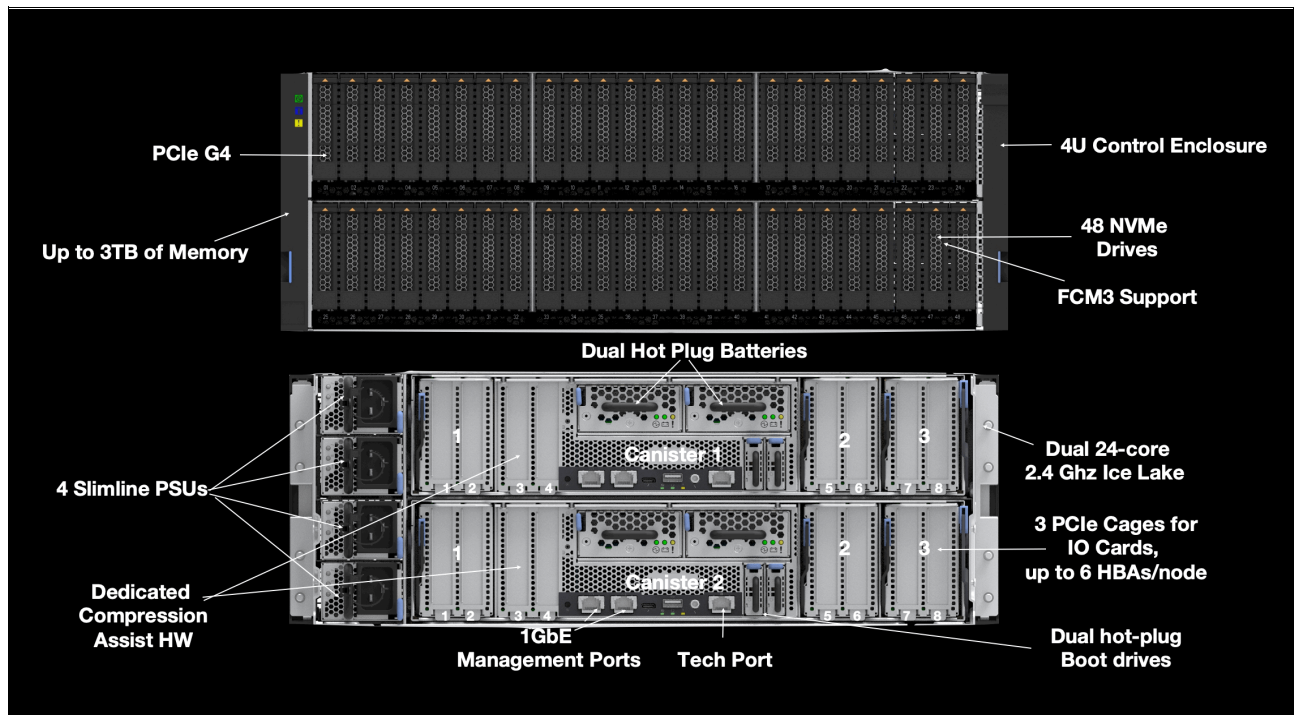


Figure 1-24 IBM FlashSystem 9500 front and rear views

IBM FlashSystem features the following 9500 components:

- ▶ IBM FlashSystem 9500 control enclosure:
 - Node canisters
 - Power supply units (PSUs): Uses C19 connectors not C13 type
 - Battery modules
 - Fan modules
 - Interface cards
 - Ice Lake CPUs and DIMM memory slots
 - USB ports
 - Ethernet ports by way of PCIe adapters
- ▶ Non-Volatile Memory Express (NVMe)-capable flash drives
- ▶ IBM FlashSystem 9000 expansion enclosures (serial-attached Small Computer System Interface [SCSI] [SAS]-attached)

As shown in Figure 1-24 on page 55, the IBM FlashSystem 9500 enclosure consists of redundant PSUs, node canisters, and fan modules to provide redundancy and HA.

Figure 1-25 shows the IBM FlashSystem 9500 internal architecture.

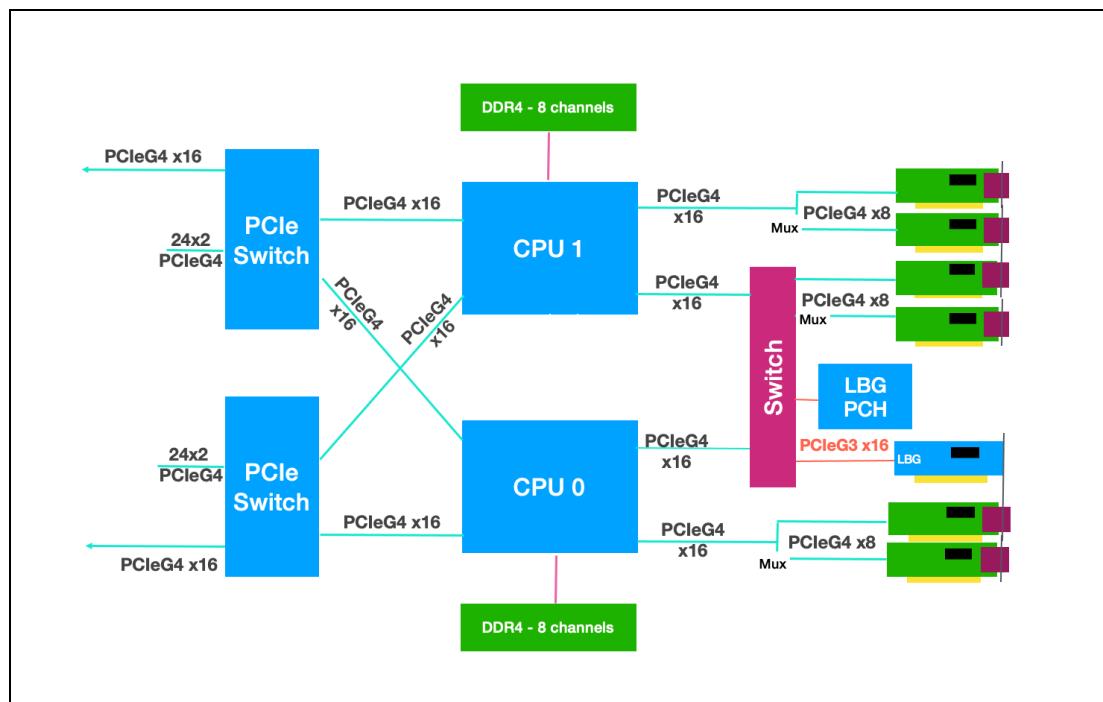


Figure 1-25 IBM FlashSystem 9500 internal architecture

Figure 1-26 shows the internal hardware components of a node canister. On the left is the front of the canister, where the NVMe drives and fan modules are installed, followed by two Ice Lake CPUs and memory DIMM slots, and Peripheral Component Interconnect® Express (PCIe) cages for the adapters on the right. The dual battery backup units are in the center, between the PCIe adapter cages.

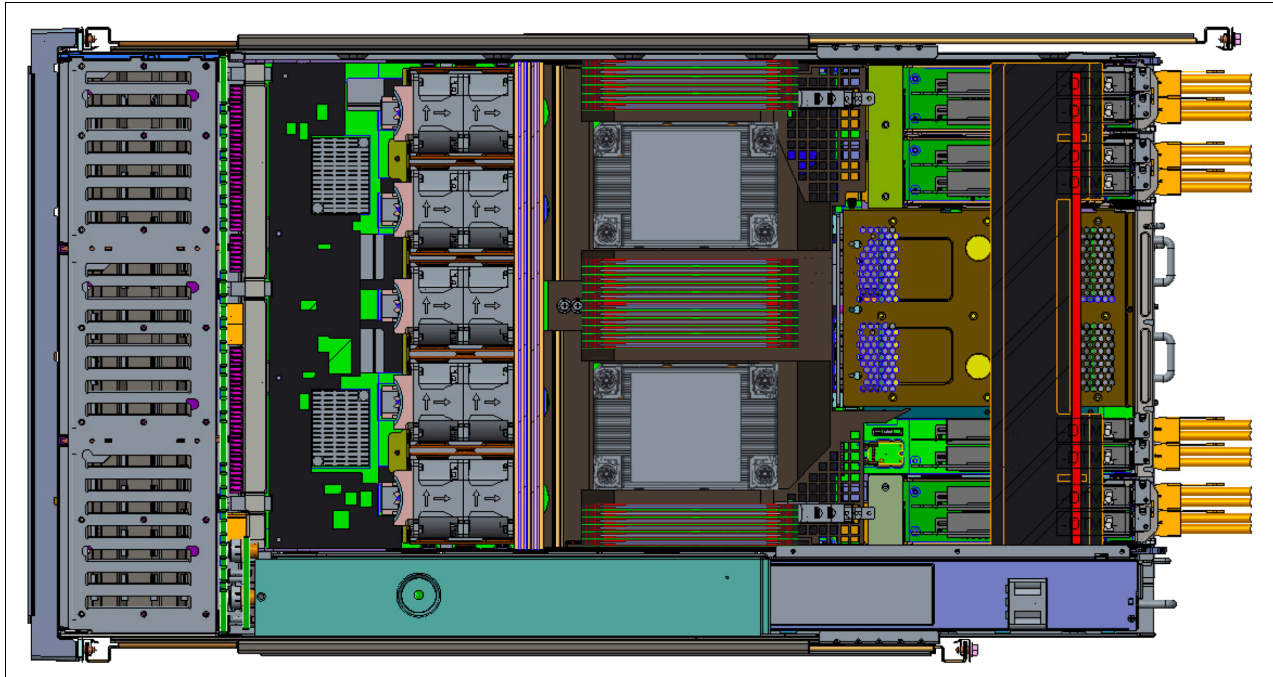


Figure 1-26 FS9500 internal hardware components

1.7.2 IBM FlashSystem 9500 control enclosure

The IBM FlashSystem 9500 system is a 4U model that can house up to 48 NVMe-capable flash drives of various capacities and be configured with up to 3.0 TB of cache.

Note: There are new rules for the plugging of the NVMe drives in the control enclosure. See the “IBM FlashSystem 9500 NVMe drive options” on page 59.

An IBM FlashSystem 9500 clustered system can contain up to two IBM FlashSystem 9500 systems and up to 464 drives in expansion enclosures (Maximum drives per IO Group: 232 (92+92+48)). The following clustering rules must be considered:

- ▶ IBM FlashSystem 9500 systems can be clustered only with another IBM FlashSystem 9500.
- ▶ IBM FlashSystem 9500 systems cannot be clustered with existing IBM FlashSystem 9200 or IBM FlashSystem 7200 or 7300 systems.

The IBM FlashSystem 9500 control enclosure node canisters are configured for active-active redundancy. The node canisters provide a web interface, Secure Shell (SSH) access, and Simple Network Management Protocol (SNMP) connectivity through external Ethernet interfaces. By using the web and SSH interfaces, administrators can monitor system performance and health metrics, configure storage, and collect support data, among other features.

IBM FlashSystem 9500 Control Enclosure Model AH8

IBM FlashSystem 9500 Control Enclosure machine type 4666 Model AG8 features the following components:

- ▶ Two node canisters, each with four 24-core 2.3 GHz Ice Lake CPUs with compression assist up to 100 gigabits per second (Gbps)
- ▶ Cache options from 512 GB (256 GB per canister) to 3.0 TB (1.5 TB per canister)
- ▶ Two 1 Gb Ethernet (GbE) onboard ports for IBM FlashSystem 9500 control enclosure management only
- ▶ Up to 12 (six per canister) I/O adapter features for:
 - Four-port 32 Gb FC-NVMe card
 - Two-port 25 GbE iSCSI/iSCSI Extensions for Remote Direct Memory Access (RDMA)
 - Two-port 25 GbE iSCSI/Hyperswap over iWARP9 (RPQ only) card
 - Two port 100 GbE iSCSI/NVMe RDMA card
 - 12 Gb SAS ports for expansion enclosure attachment
- ▶ 48 slots for 2.5-inch NVMe flash drives: Up to 12 SCM style drives
- ▶ 4U 19-inch rack mount enclosure with AC power supplies
- ▶ Four hot-swappable boot drives
- ▶ Four hot-swappable batteries and AC power supplies

Note: There is a new machine type 4983 models AH8 and UHB being introduced which is identical to the 4666, except it will be sold with Licensed Internal Code (LIC) in line with the other products in the FlashSystems product portfolio. This ensures that all features are included in the product price with the exception of the encryption.

IBM FlashSystem 9500 Utility Model UH8

IBM FlashSystem 9500 Utility Model UH8 provides a variable capacity storage offering. These models offer a fixed capacity with a base subscription of 35% of the total capacity.

IBM Storage Insights is responsible for monitoring the system and reporting the capacity that was used beyond the base 35%, which is then billed on the capacity-used basis. You can grow or shrink usage, and pay only for the configured capacity.

The IBM FlashSystem Utility Model is provided for customers who can benefit from a variable capacity system, where billing is based on actual provisioned space only. The hardware is leased through IBM Global Finance on a three-year lease, which entitles the customer to use approximately 30 - 40% of the total system capacity at no extra cost (depending on the individual customer contract). If storage needs increase beyond that initial capacity, usage is billed based on the average daily provisioned capacity per terabyte per month, on a quarterly basis.

Example: Total system capacity of 115 TB

A customer has an IBM FlashSystem 9500 Utility Model with 4.8 TB NVMe drives for a total system capacity of 115 TB. The base subscription for such a system is 40.25 TB. During the months where the average daily usage is less than 40.25 TB, no extra billing occurs.

The system monitors daily provisioned capacity and averages those daily usage rates over the month term. The result is the average daily usage for the month.

If a customer uses 45 TB, 42.5 TB, and 50 TB in three consecutive months, IBM Storage Insights calculates the overage as listed in Table 1-6, rounding to the nearest terabyte.

Table 1-6 Billing calculations that are based on customer usage

Average daily	Base	Overage	To be billed
45 TB	40.25 TB	4.75 TB	5 TB
42.5 TB	40.25 TB	2.25 TB	2 TB
50 TB	40.25 TB	9.75 TB	10 TB

The total capacity that is billed at the end of the quarter is 17 TB per month in this example.

Flash drive expansions can be ordered with the system in all supported configurations.

Table 1-7 lists the feature codes that are associated with the IBM FlashSystem 9500 Utility Model UH8 billing.

Table 1-7 IBM FlashSystem 9500 Utility Model UG8 billing feature codes

Feature code	Description
#AE00	Variable Usage 1 TB per month
#AE01	Variable Usage 10 TB per month
#AE02	Variable Usage 100 TB per month

These features are used to purchase the variable capacity that is used in the IBM FlashSystem 9500 Utility Models. The features (#AE00, #AE01, and #AE02) provide terabytes of capacity beyond the base subscription on the system. Usage is based on the average capacity that is used per month. The total of the prior three months' usage should be totaled and the corresponding number of #AE00, #AE01, and #AE02 features that are ordered quarterly.

IBM FlashSystem 9500 NVMe drive options

The IBM FlashSystem 9500 control enclosure supports up to 48 NVMe 2.5-inch drives, which can be the IBM FlashCore Module NVMe type drives or the industry-standard NVMe drives.

With partially populated control enclosures, we have some drive slot plugging rules that must be adhered to, ensuring the best possible operating conditions for the drives.

Figure 1-27 on page 60 shows the logical NVMe drive placement, starting from the center of the enclosure (slot 12) on the upper 24 slots. Any slots that do not have an NVMe drive present must have a blank filler installed.

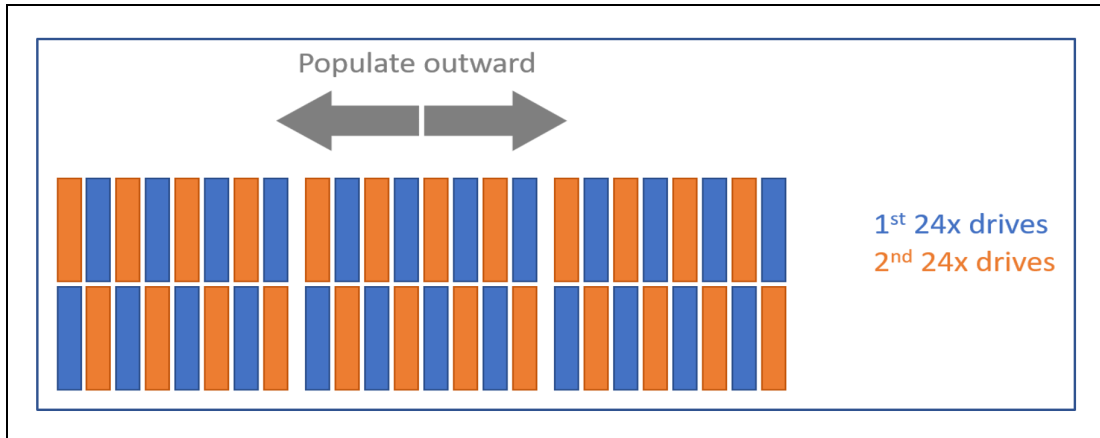


Figure 1-27 Logical NVMe drive placement

Figure 1-28 shows the actual drive population with numbering. This shows how the drives are populated from center out, and then distributing them from top and bottom, as the number of drives increase over time.

Note: The layout in Figure 1-28 has been split at slots 12 and 13 for better clarity on this page; however, slots 1 to 24 and slots 25 to 48 are contiguous.

Slot 1	Slot 2	Slot 3	Slot 4	Slot 5	Slot 6	Slot 7	Slot 8	Slot 9	Slot 10	Slot 11	Slot 12
36	11	34	9	32	7	30	5	28	3	26	1
Slot 25	Slot 26	Slot 27	Slot 28	Slot 29	Slot 30	Slot 31	Slot 32	Slot 33	Slot 34	Slot 35	Slot 36
24	47	22	45	20	43	18	41	16	39	14	37
Slot 13	Slot 14	Slot 15	Slot 16	Slot 17	Slot 18	Slot 19	Slot 20	Slot 21	Slot 22	Slot 23	Slot 24
25	2	27	4	29	6	31	8	33	10	35	12
Slot 37	Slot 38	Slot 39	Slot 40	Slot 41	Slot 42	Slot 43	Slot 44	Slot 45	Slot 46	Slot 47	Slot 48
13	38	15	40	17	42	19	44	21	46	23	48

Figure 1-28 NVMe drive population with numbering

1.7.3 IBM FlashSystem 9000 Expansion Enclosure Models AFF and A9F

IBM FlashSystem 9500 Model AH8 and IBM FlashSystem 9500 Utility Model UH8 support the expansion enclosures IBM FlashSystem 9000 Models AFF and A9F.

For more information, see 1.9, “IBM FlashSystem 9000 Expansion Enclosure Models AFF and A9F” on page 64.

Note: The IBM FlashSystem 9500 Model AH8 and IBM FlashSystem 9500 Model UH8 can support a maximum of one IBM FlashSystem 9000 Model A9F dense expansion or three IBM FlashSystem 9000 Model AFF enclosures per chain.

1.8 IBM FlashSystem 9500R Rack Solution overview

IBM FlashSystem 9500R is a pre-cabled, pre-configured rack solution that contains two IBM FlashSystem 9500 control enclosures. It uses IBM Storage Virtualize to linearly scale performance and capacity through clustering.

The IBM FlashSystem 9500R Rack Solution system features a dedicated FC network for clustering and optional expansion enclosures, which are delivered assembled in a rack. Available with two clustered IBM FlashSystem 9500 systems and up to four expansion enclosures, it can be ordered as an IBM FlashSystem 9502R, with the last number denoting the two AG8 controller enclosures in the rack.

The final configuration occurs on site following the delivery of the systems. More components can be added to the rack after delivery to meet the growing needs of the business.

Note: Other than the IBM FlashSystem 9500 control enclosures and its expansion enclosures, the extra components of this solution are not covered under Storage Expert Care. Instead, they have their own warranty, maintenance terms, and conditions.

Rack rules

The IBM FlashSystem 9500R Rack Solution product represents a limited set of possible configurations. Each IBM FlashSystem 9500R Rack Solution order must include the following components:

- ▶ Two 4666 Model AH8 control enclosures.
- ▶ Two IBM SAN24B-6 or two IBM SAN32C-6 FC switches.
- ▶ The following optional expansion enclosures are available by way of MES only. They cannot be ordered with the machine if ordered as a new build:
 - 0 - 4 4666 Model AFF expansion enclosures, with no more than one expansion enclosure per Model AG8 control enclosure and no mixing with the 9848/4666 Model A9F expansion enclosures.
 - 0 - 2 4666 Model A9F expansion enclosures, with no more than one expansion enclosure per Model AG8 control enclosure and no mixing with 9848/4666 Model AFF expansion enclosures.
- ▶ One 7965-S42 rack with the suitable power distribution units (PDUs) that are required to power components within the rack.
- ▶ All components in the rack must include feature codes #FSRS and #4651.
- ▶ For Model AH8 control enclosures, the first and largest capacity enclosure includes feature code #AL01, and #AL02, in capacity order. The 4666 / 4983 model AH8 control enclosure with #AL01 also must include #AL0R.

Following the initial order, each 4666 Model AH8 control enclosure can be upgraded through a MES.

More components can be ordered separately and added to the rack within the configuration limitations of the IBM FlashSystem 9500 system. Customers must ensure that the space, power, and cooling requirements are met. If assistance is needed with the installation of these additional components beyond the service that is provided by your IBM System Services Representative (IBM SSR), IBM Lab Services are available.

Note: There is a new machine type 4983 model AH8 being introduced that is physically identical to the 4666, except it will be sold with Licensed Internal Code (LIC) in line with the other products in the FlashSystems product line. This ensures all features are included in the product price with the exception of the encryption.

Table 1-8 lists the IBM FlashSystem 9500R Rack Solution combinations, the MTMs, and their associated feature codes.

Table 1-8 IBM FlashSystem 9500R Rack Solution combinations

Machine type and model (MTM)	Description	Quantity
7965-S42	IBM Enterprise Slim Rack	1
8960-F24	IBM SAN24B-6 FC switch (Brocade)	2 ^a
8977-T23	IBM SAN32C-6 FC switch (Cisco)	2 ^b
4666/4983-AH8	IBM FlashSystem 9500 control enclosure with 3-year Storage Expert Care	2
The following expansion enclosures are available by way of MES order only.		
4666-A9F	IBM FlashSystem 9000 5U large form factor (LFF) high-density expansion enclosures with 3-year Storage Expert Care	0 - 2 ^c
4666-AFF	IBM FlashSystem 9000 2U small form factor (SFF) Expansion enclosure with 3-year Storage Expert Care	0 - 4 ^d

- a. For the FC switch, choose two of machine type (MT) 8977 or two of MT 8960.
- b. For the FC switch, choose two of machine type (MT) 8977 or two of MT 8960.
- c. For extra expansion enclosures, choose model AFF, model A9F, or none. You cannot use both.
- d. For extra expansion enclosures, choose model AFF, model A9F, or none. You cannot use both.

1.8.1 IBM FlashSystem 9500R Rack Solution diagram

This section describes the rack diagram (see Figure 1-30, “IBM FlashSystem 9500R Rack Solution configuration in the rack” on page 63) that shows IBM FlashSystem 9500R Rack Solution configuration.

Key to figures

The key to the symbols that are used in the figures in this section are listed in Table 1-9.

Table 1-9 Key to rack configuration

Label	Description
CTL _n	<ul style="list-style-type: none"> ▶ 9848/4666 AH8 control enclosure number <i>n</i> of 2 ▶ CTL1 and CTL2 are required
EXP _n	9848/4666 expansion enclosures number <i>n</i> (optional)
FC SW _n	<ul style="list-style-type: none"> ▶ FC switch <i>n</i> of 2 ▶ These switches are both 8977-T32 or both 8960-F24
PDU A, PDU B	PDUs, both have the same rack feature code: #ECJJ, #ECJL, #ECJN, or #ECJQ.

Figure 1-29 shows the legend that is used to denote the component placement and mandatory gaps for the figures that show the configurations.

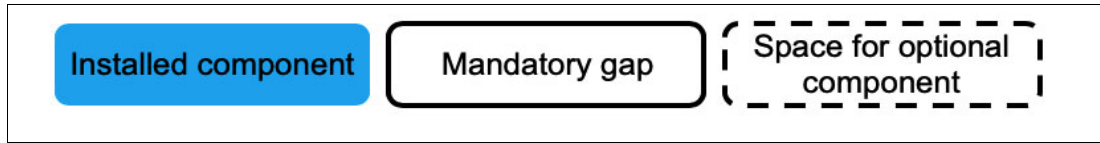


Figure 1-29 Legend to figures in this section

Figure 1-30, “IBM FlashSystem 9500R Rack Solution configuration in the rack” on page 63 shows the standard IBM FlashSystem 9500R Rack Solution configuration in the rack.

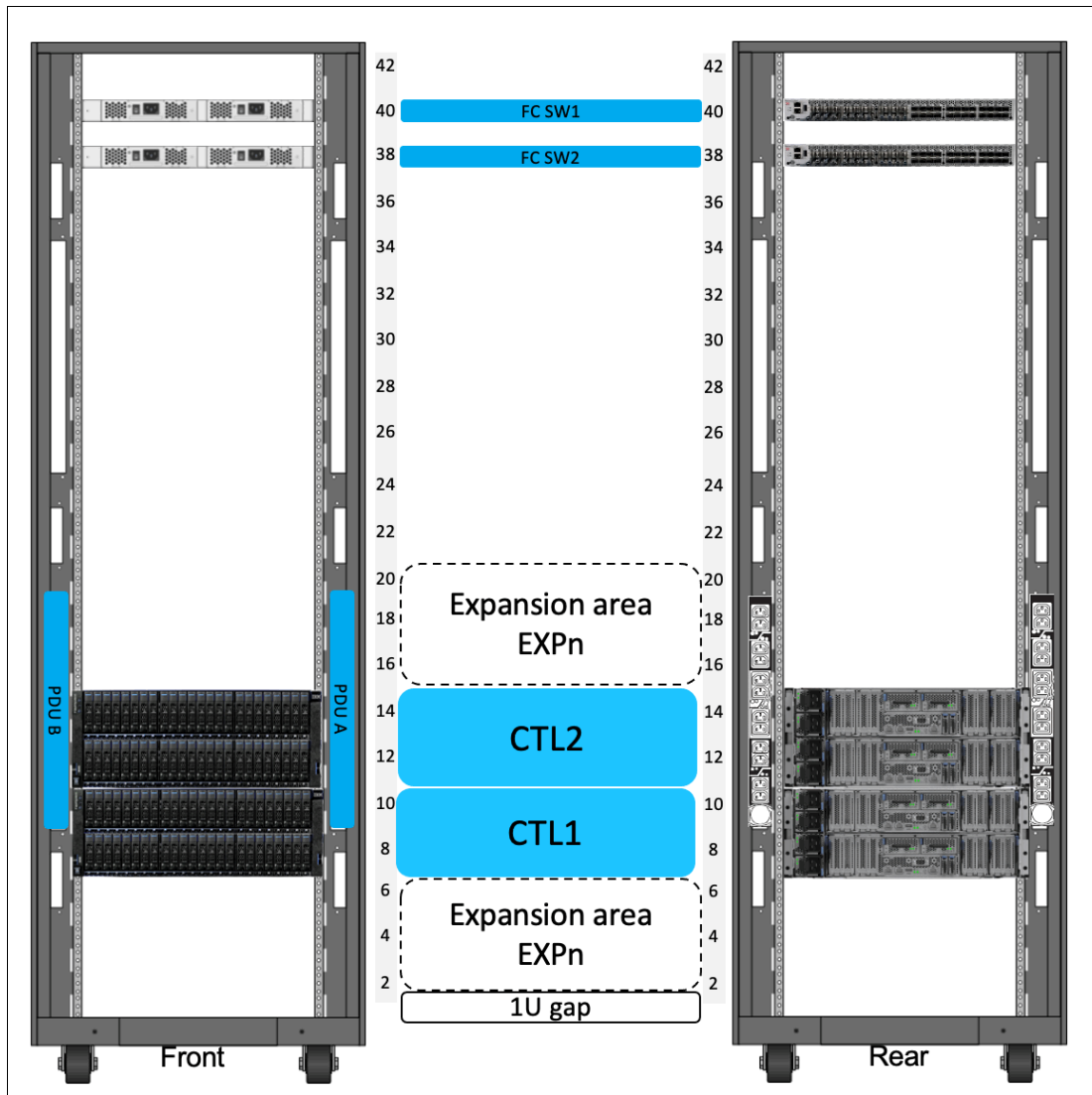


Figure 1-30 IBM FlashSystem 9500R Rack Solution configuration in the rack

Minimum configuration

Consider the following points:

- ▶ Control enclosures (CTL) 1 and 2 are mandatory.
- ▶ The product includes cables that are suitable for inter-system FC connectivity. You must order extra cables for host and Ethernet connectivity.
- ▶ The PDUs and power cabling that are needed depends on what expansion enclosures are ordered.
- ▶ Two PDUs with nine C19 outlets are required. This PDU also has three C13 outlets on the forward-facing side.
- ▶ FC SW1 and FC SW2 are a pair of IBM SAN32C-6 or IBM SAN24B-6 FC switches.
- ▶ You can allocate different amounts of storage (drives) to each CTL component.
- ▶ A gap of 1U is maintained below the expansion area to allow for power cabling routing.

Adding configuration with Model A9F and AFF expansion enclosures by way of MES

Consider the following points:

- ▶ The product includes cables that are suitable for inter-system FC connectivity. You must order extra cables for host and Ethernet connectivity.
- ▶ Any Model A9F or AFF expansion enclosures are installed in U2 - U6 and then, U16-20.
- ▶ The CTLs and EXPs are stacked and cabled to the PDU power, with the highest capacity at the bottom. You go upwards, with EXPn attached to CTLn in a bottom-up order by using an SAS adapter on CTLn and cables.
- ▶ For the 4666/AH8 CTL1 and CTL2, the following adapter cage rules apply:
 - Adapter cage 1 of each AH8 is dedicated to 32 Gb clustering usage.
 - Adapter cage 2 of each AH8 is reserved for the compression adapter.
 - Adapter cage 3 is an SAS adapter if it is required, or one of your choice if it is not.
 - Adapter cage 4 is optional for FC 32 Gb, 25 Gb, or 100 Gb Ethernet.

1.8.2 FC cabling and clustering

The IBM FlashSystem 9500R Rack Solution includes specialized internal cabling that is supplied by the manufacturing plant before the machine is shipped to the customer. Plugging the inter-system internal FC cables is done on site at installation time.

For more information, see [Planning for clustered system connections within the rack](#).

1.9 IBM FlashSystem 9000 Expansion Enclosure Models AFF and A9F

IBM FlashSystem 9000 Expansion Enclosure Models AFF and A9F can be attached to an IBM FlashSystem 9200 or IBM FlashSystem 9500 control enclosure to increase the available capacity. They communicate with the control enclosure through a dual pair of 12 Gbps SAS connections. These expansion enclosures can house many flash (solid-state drive [SSD]) SAS type drives.

IBM FlashSystem 9000 Expansion Enclosure Model AFF

IBM FlashSystem 9000 Expansion Enclosure Model AFF holds up to 24 2.5-inch SAS flash drives in a 2U 19-inch rack mount enclosure.

An intermix of capacity drives can be used in any drive slot. The following attachment rules are applicable for each SAS chain:

- ▶ IBM FlashSystem 9200: Up to 10 IBM FlashSystem 9000 Expansion Enclosure Model AFF enclosures can be attached to the control enclosure to a total of 240 drives maximum.
- ▶ IBM FlashSystem 9500: Up to three IBM FlashSystem 9000 Expansion Enclosure Model AFF enclosures can be attached. This configuration provides extra capacity with a maximum of 72 drives.

Figure 1-31 shows the front view of the IBM FlashSystem 9000 Expansion Enclosure Model AFF.



Figure 1-31 IBM FlashSystem 9000 Expansion Enclosure Model AFF

IBM FlashSystem 9000 Expansion Enclosure Model A9F

The IBM FlashSystem 9000 Expansion Enclosure Model A9F holds up to 92 3.5-inch SAS flash drives (2.5-inch SAS flash drives in carriers) in a 5U 19-inch rack mount enclosure.

An intermix of capacity drives is allowed in any drive slot and the following attachment rules are applicable for each SAS chain:

- ▶ IBM FlashSystem 9200: Up to four IBM FlashSystem 9000 Expansion Enclosure Model A9F enclosures can be attached to the control enclosure to a total of 368 drives maximum.
- ▶ IBM FlashSystem 9500: One IBM FlashSystem 9000 Expansion Enclosure Model A9F can be attached. This configuration provides extra capacity with a maximum of 92 drives.

Figure 1-32 shows the front view of the IBM FlashSystem 9000 Expansion Enclosure Model A9F.

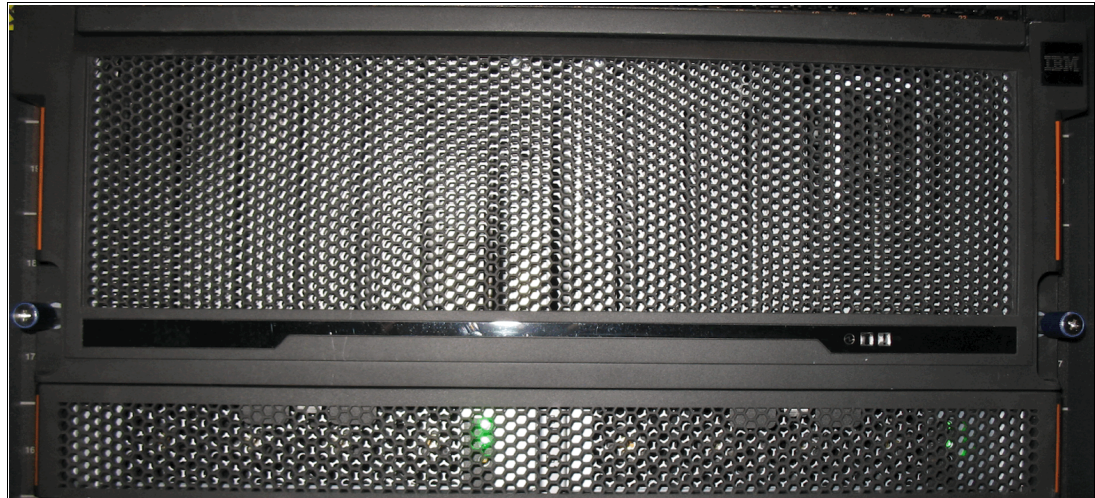


Figure 1-32 IBM FlashSystem 9000 Expansion Enclosure Model front view

SAS chain limitations

When attaching expansion enclosures to the control enclosure, you are not limited by the type of the enclosure. The only limitation for each of the two SAS chains is chain weight. Each type of enclosure has its own chain weight:

- ▶ IBM FlashSystem 9000 Expansion Enclosure Model AFF has a chain weight of 1.
- ▶ IBM FlashSystem 9000 Expansion Enclosure Model A9F has a chain weight of 2.5.
- ▶ For the IBM FlashSystem 9500, the maximum chain weight is 3.

Consider the following example:

- ▶ With the IBM FlashSystem 9500, you can have three IBM FlashSystem 9000 Expansion Enclosure Model AFF or one IBM FlashSystem 9000 Expansion Enclosure Model A9F expansions (3 x 1 or 1 x 2.5).

An example of chain weight 4.5 with two IBM FlashSystem 9000 Expansion Enclosure Model AFF enclosures and one IBM FlashSystem 9000 Expansion Enclosure Model A9F enclosure all correctly cabled is shown in Figure 1-33 on page 67, which shows an IBM FlashSystem 9200 system connecting through SAS cables to the expansion enclosures while complying with the maximum chain weight.

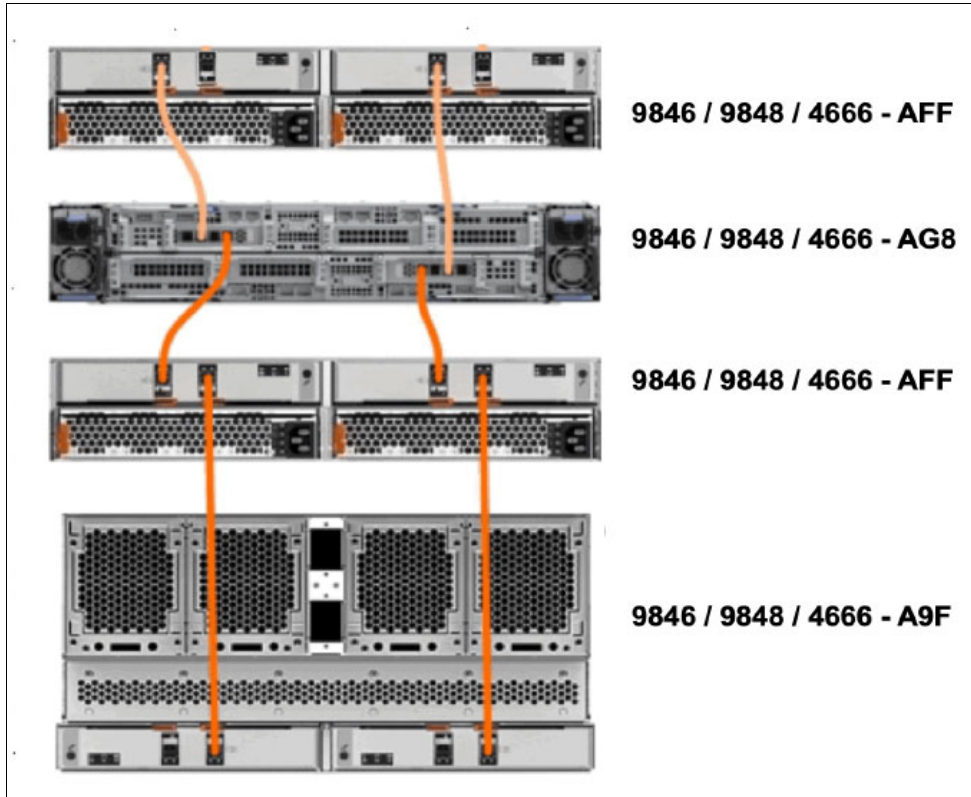


Figure 1-33 IBM FlashSystem 9200 system that is connected to expansion enclosure

An example of chain weights 3 and 2.5 with three IBM FlashSystem 9000 Expansion Enclosure Model AFF enclosures and one IBM FlashSystem 9000 Expansion Enclosure Model A9F enclosure all correctly cabled is shown in Figure 1-34, which shows an IBM FlashSystem 9500 system connecting through SAS cables to the expansion enclosures while complying with the maximum chain weight.

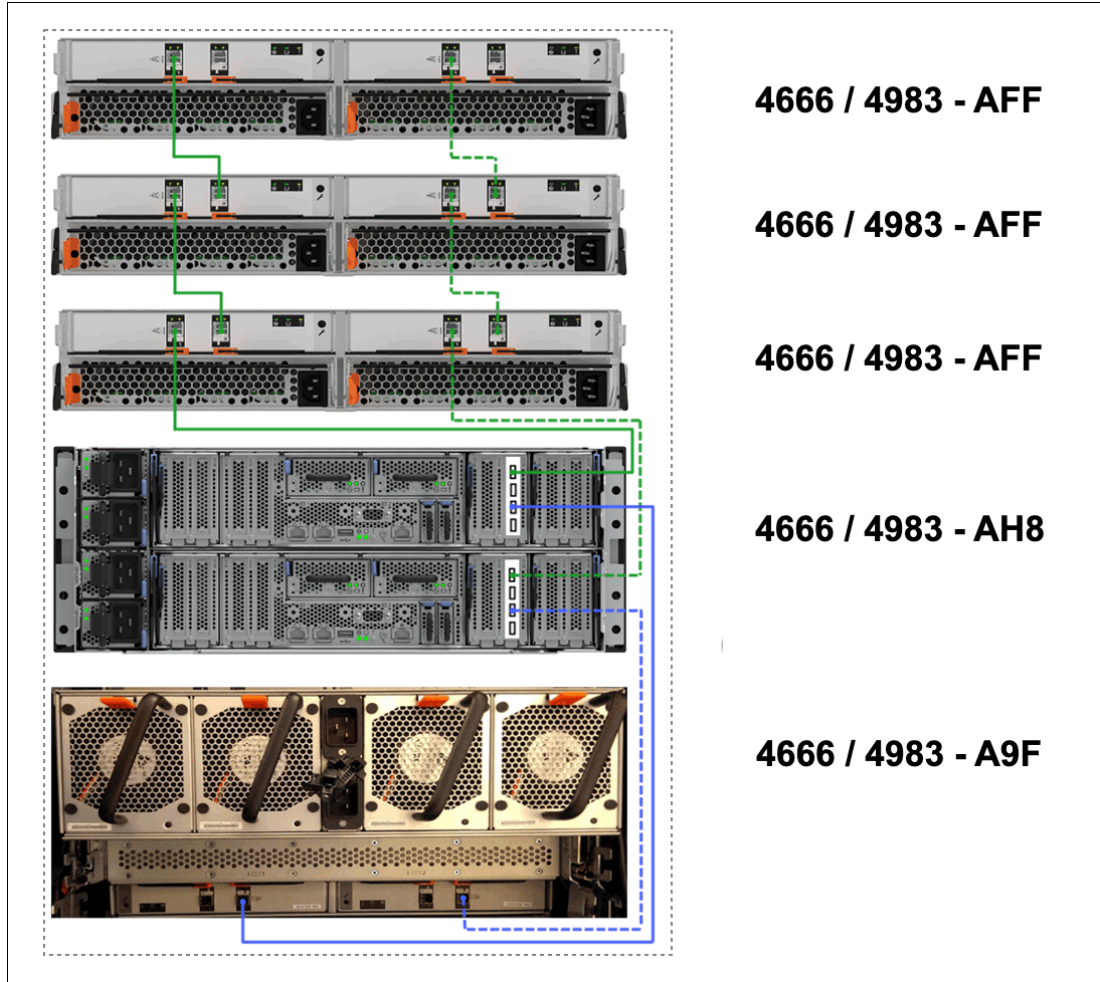


Figure 1-34 IBM FlashSystem 9500 system that is connected to expansion enclosure

Note: The expansion enclosures rules will be the same for the FlashSystem 9500 machine type 4983, which is functionally equivalent to the 4666, except it has Licensed Internal Code (LIC).

For more information, see [IBM FlashSystem 9500](#).

1.10 IBM FlashSystem 7300 overview

Each IBM FlashSystem 7300 system consists of a control enclosure and NVMe-attached flash drives. The control enclosure is the storage server that runs the IBM Storage Virtualize software that controls and provides features to store and manage data.

IBM FlashSystem 7300 has a new machine type of 4657. This new machine type includes the Storage Virtualize component as Licensed Machine Code (LMC); therefore, it does not require the purchase of separate software maintenance (SWMA). IBM FlashSystem 7300 still requires licenses for the external virtualization of storage.

For more information, see section 1.21, “Licensing” on page 120.

Figure 1-35 shows the front and rear views of the IBM FlashSystem 7300 system.

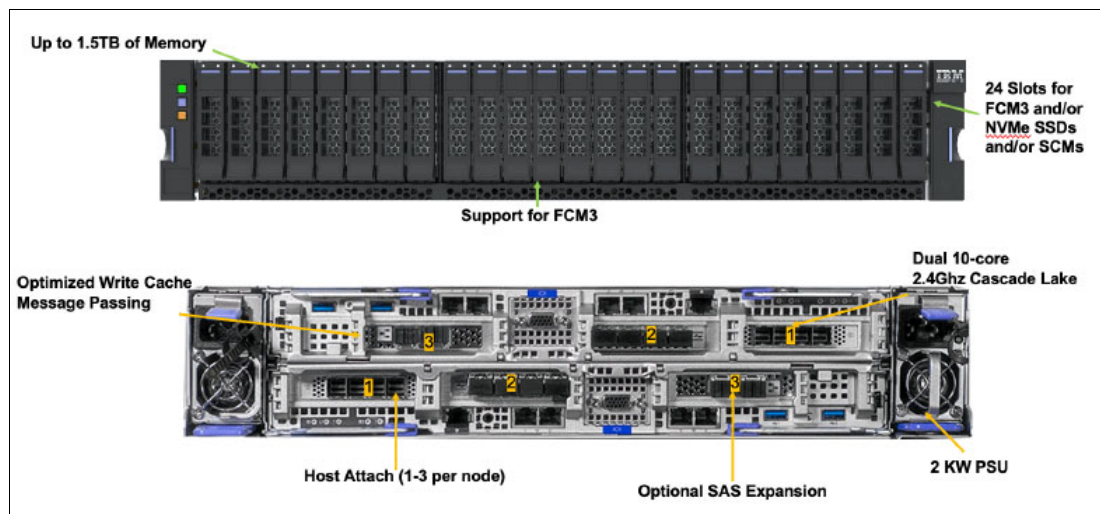


Figure 1-35 IBM FlashSystem 7300 front and rear views

The IBM FlashSystem 7300 features the following core components:

- ▶ IBM FlashSystem 7300 control enclosure:
 - PSUs
 - Node canisters
 - Battery modules
 - Fan modules
 - Interface cards
 - Cascade Lake CPUs and memory slots
- ▶ NVMe drives
- ▶ IBM FlashSystem 7000 expansion enclosures (SAS-attached)

3

Note: Consider the following points:

- ▶ The IBM FlashSystem 7300 (4657-924 and U7D) is available with Storage Expert Care as described in 1.6.1, “Storage Expert Care” on page 50.
- ▶ Machine type 4657 FlashSystem 7300 systems can be clustered only with other FlashSystem 7300 systems. Clustering with machine types 2076, 4664, 4666, 9846, or 9848 is not supported.

As shown in Figure 1-36, the IBM FlashSystem 7300 enclosure consists of redundant PSUs, node canisters, and fan modules to provide redundancy and HA.

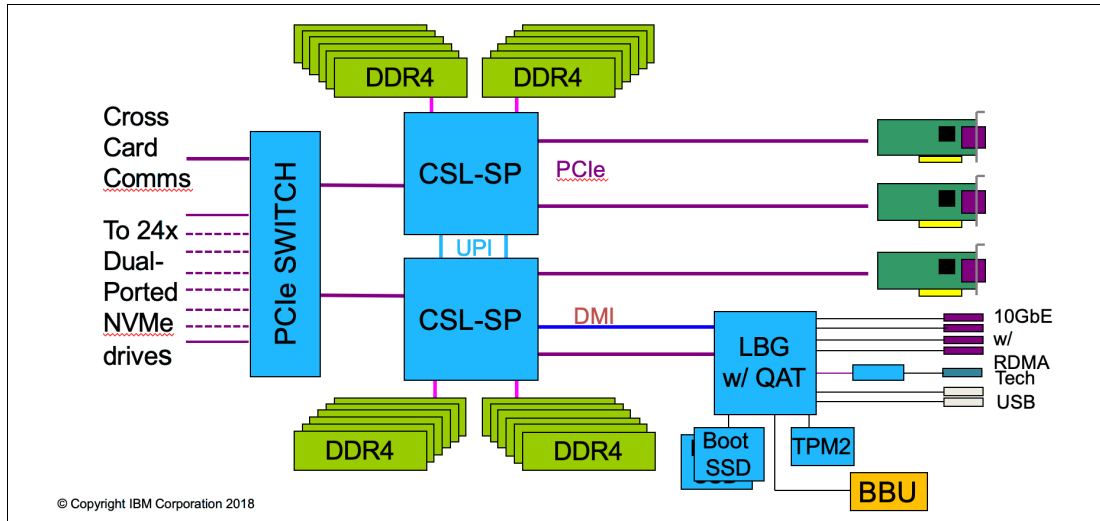


Figure 1-36 IBM FlashSystem 7300 internal architecture

Figure 1-37 shows the internal hardware components of a node canister. On the left side is the front of the canister where fan modules and battery backup are installed, followed by two Cascade Lake CPUs, Dual Inline Memory Module (DIMM) slots, and PCIe risers for adapters on the right side.

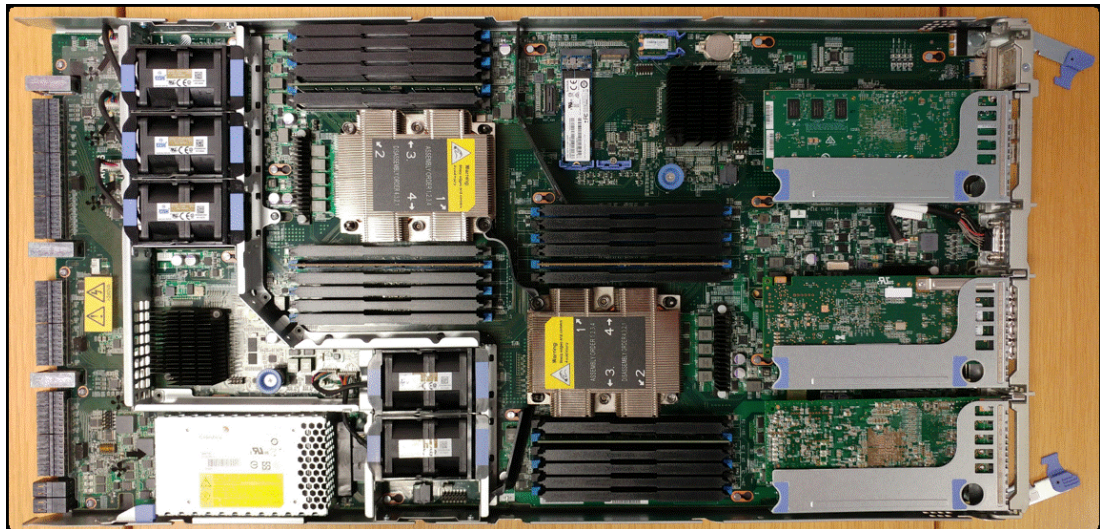


Figure 1-37 Internal hardware components

1.10.1 IBM FlashSystem 7300 control enclosures

IBM FlashSystem 7300 is a 2U model that can support up to 24 NVMe drives (FCM drives with hardware compression and encryption or industry-standard NVMe drives of various capacities or even SCM drives). IBM FlashSystem 7300 can be configured with up to 1.5 TB of cache.

For more information about the supported drive types, see 1.17, “IBM FlashCore Module drives, NVMe SSDs, and SCM drives” on page 105.

An IBM FlashSystem 7300 clustered system can contain up to four IBM FlashSystem 7300 systems and up to 440 drives per IO group. IBM FlashSystem 7300 systems can be added only into clustered systems that include other IBM FlashSystem 7300 systems.

IBM FlashSystem 7300 Model 924 system

The model 924 system offers the following features:

- ▶ Two node canisters, each with two 10-core processors and integrated hardware-assisted compression acceleration
- ▶ 256 GB cache (128 GB per canister) standard with options from 768 GB to 1536 GB (per system)
- ▶ Eight 10 Gb Ethernet ports standard for iSCSI connectivity (copper)
- ▶ 32 Gb FC connectivity options with SCSI and FC-NVMe support (fiber)
- ▶ 10/25 Gb and 100 Gb Ethernet ports for iSCSI and NVMe RDMA connectivity
- ▶ 12 Gb SAS ports for optional expansion enclosure attachment
- ▶ 24 slots for NVMe flash drives, including up to 12 storage-class memory drives
- ▶ 2U, 19-inch rack mount enclosure with AC power supplies
- ▶ Dual boot drives

IBM FlashSystem 7300 Model U7D system

The IBM 4657 Model U7D is the FlashSystem 7300 with a one-year warranty to be used in the Storage Utility Offering space. It is physically and functionally identical to the FlashSystem 7300 Model 924 except for variable capacity billing.

The variable capacity billing uses IBM Storage Insights to monitor the system use, which allows allocated storage use that is above a base subscription rate to be billed per TB, per month.

Allocated storage is identified as storage that is allocated to a specific host (and unusable to other hosts), whether data is written or not written. For thin-provisioning, the data that is written is considered used. For thick provisioning, total allocated volume space is considered used.

1.10.2 IBM FlashSystem 7000 Expansion Enclosure 4657 Models 12G, 24G, and 92G

The following types of expansion enclosures are available:

- ▶ IBM FlashSystem 7000 LFF Expansion Enclosure 4657 Model 12G
- ▶ IBM FlashSystem 7000 SFF Expansion Enclosure 4657 Model 24G
- ▶ IBM FlashSystem 7000 LFF Expansion Enclosure 4657 Model 92G

Note: Attachment and intermixing of machine type 2076/4664 Expansion Enclosure Models 12G, 24G, and 92G with machine type 4657 FlashSystem 7300 controller and expansion models is *not* supported.

Expansion Enclosure 4657 Model 12G

The IBM FlashSystem 7000 LFF 12G Expansion Enclosure includes the following components:

- ▶ Two expansion canisters
- ▶ 12 Gb SAS ports for control enclosure and expansion enclosure attachment
- ▶ A total of 12 slots for 3.5-inch SAS drives
- ▶ 2U 19-inch rack-mounted enclosure with AC power supplies

Figure 1-38 shows the front view of a 4657 Model 12G.



Figure 1-38 IBM FlashSystem 7000 LFF Expansion Enclosure Model 12G

Expansion Enclosure 4657 Model 24G

IBM FlashSystem 7000 SFF Expansion Enclosure 4657 Model 24G includes the following components:

- ▶ Two expansion canisters
- ▶ 12 Gb SAS ports for control enclosure and expansion enclosure attachment
- ▶ A total of 24 slots for 2.5-inch SAS drives
- ▶ 2U 19-inch rack mount enclosure with AC power supplies

The SFF expansion enclosure is a 2U enclosure that includes the following components:

- ▶ A total of 24 2.5-inch drives (hard disk drives [HDDs] or SSDs).
- ▶ Two Storage Bridge Bay (SBB)-compliant Enclosure Services Manager (ESM) canisters.
- ▶ Two fan assemblies, which mount between the drive midplane and the node canisters. Each fan module is removable when the node canister is removed.
- ▶ Two power supplies.
- ▶ An RS232 port on the back panel (3.5 mm stereo jack), which is used for configuration during manufacturing.

Figure 1-39 shows the front of an SFF expansion enclosure.

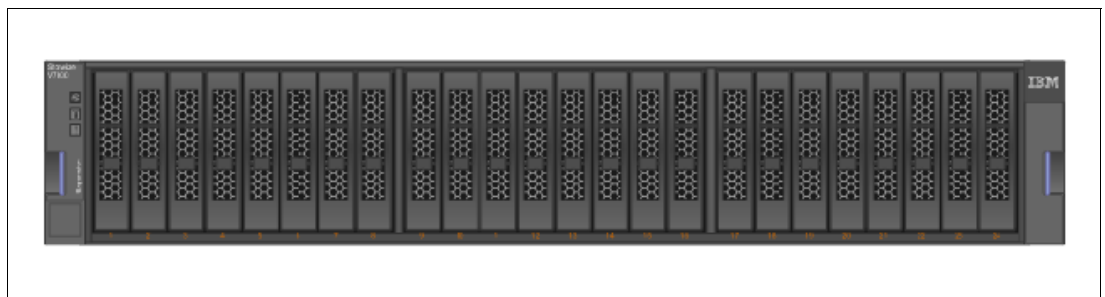


Figure 1-39 Front view of an IBM FlashSystem 7000 SFF expansion enclosure

Figure 1-40 shows the rear view of the expansion enclosure.

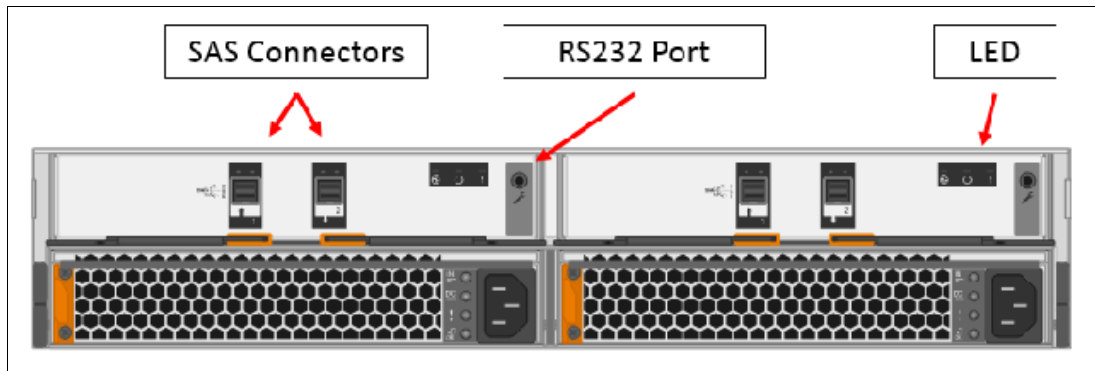


Figure 1-40 Rear of an IBM FlashSystem 7000 expansion enclosure

Dense Expansion Enclosure 92G

Dense expansion drawers, or dense drawers, are disk expansion enclosures that are 5U rack-mounted. Each chassis features two expansion canisters, two power supplies, two expander modules, and a total of four fan modules.

Each dense drawer can hold up to 92 drives that are positioned in four rows of 14 and another three rows of 12 mounted drives assemblies. Two SEMs are centrally located in the chassis: one SEM addresses 54 drive ports, and the other addresses 38 drive ports.

The drive slots are numbered 1 - 14, starting from the left rear slot and working from left to right, back to front.

Each canister in the dense drawer chassis features two SAS ports numbered 1 and 2. The use of SAS port1 is mandatory because the expansion enclosure must be attached to an IBM FlashSystem 7300 node or another expansion enclosure. SAS connector 2 is optional because it is used to attach to more expansion enclosures.

Each IBM FlashSystem 7300 system can support up to four dense drawers per SAS chain.

Figure 1-41 shows a dense expansion drawer.



Figure 1-41 IBM Dense Expansion Drawer

SAS chain limitations

When attaching expansion enclosures to the control enclosure, you are not limited by the type of the enclosure. The only limitation for each of the two SAS chain is its chain weight. Each type of enclosure has its own chain weight:

- ▶ Enclosures 12G and 24G have a chain weight of 1
- ▶ Enclosure 92G has a chain weight of 2.5

The maximum chain weight is 6.

An example of chain weight 4.5 with one 24G, one 12G, and one 92G enclosures, all correctly cabled, is shown in Figure 1-42.

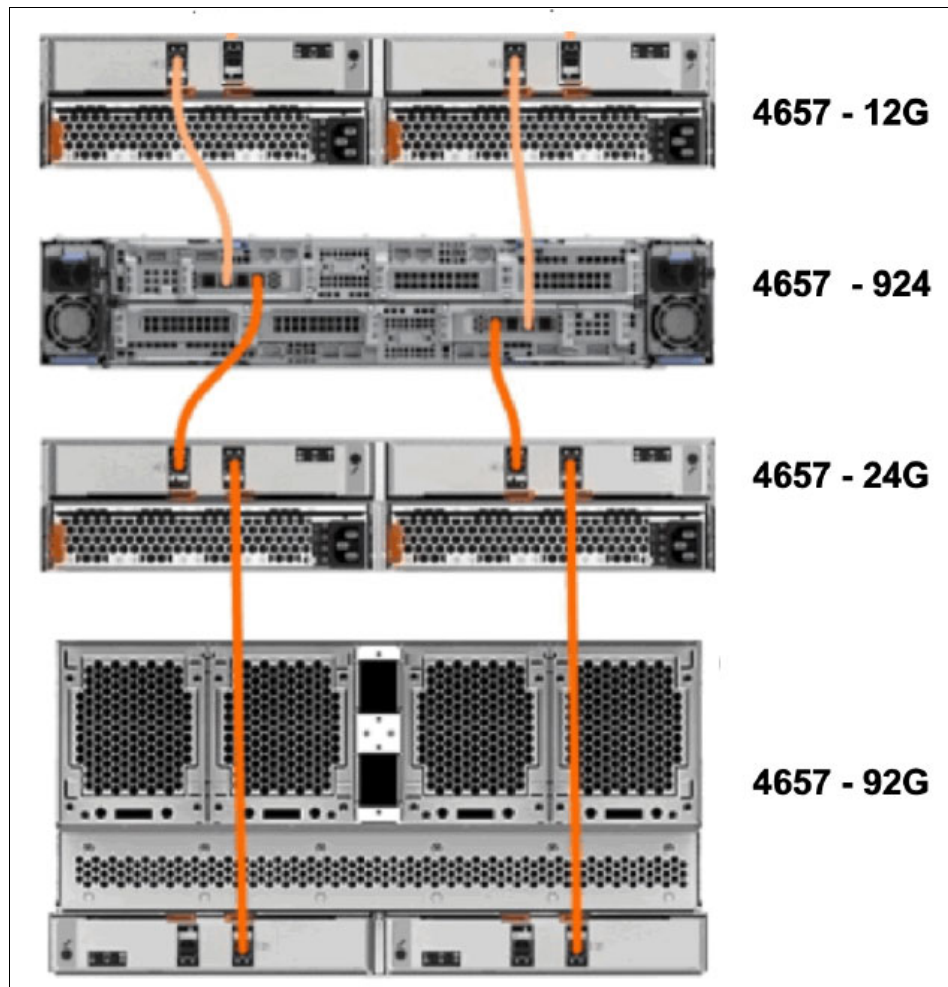


Figure 1-42 Connecting FS7300 SAS cables while complying with the maximum chain weight

1.11 IBM FlashSystem 5200 overview

With IBM FlashSystem 5200, you can be ready for a technology transformation without sacrificing performance, quality, or security while simplifying your data management. This powerful and compact solution is focused on affordability with a wide range of enterprise-grade features of IBM Storage Virtualize that can easily evolve and extend as businesses grows.

This system also has the flexibility and performance of flash and Non-Volatile Memory Express (NVMe) end to end, the innovation of IBM FlashCore technology, and Storage Class Memory (SCM) to help accelerate your business execution.

The innovative IBM FlashSystem family is based on a common storage software platform, IBM Storage Virtualize, that provides powerful all-flash and hybrid-flash solutions that offer feature-rich, cost-effective, and enterprise-grade storage solutions.

Its industry-leading capabilities include a wide range of data services that can be extended to more than 500 heterogeneous storage systems, including the following examples:

- ▶ Automated data movement
- ▶ Synchronous and asynchronous copy services on-premises or to the public cloud
- ▶ HA configurations
- ▶ Storage automated tiering
- ▶ Data reduction technologies, including deduplication

Available on IBM Cloud and Amazon Web Services (AWS), IBM Storage Virtualize for Public Cloud works with IBM FlashSystem 5200 to deliver consistent data management between on-premises storage and public cloud. You can move data and applications between on-premises and public cloud, implement new DevOps strategies, use public cloud for DR without the cost of a second data center, or improve cyber resiliency with “air gap” cloud snapshots.

IBM FlashSystem 5200 offers world-class customer support, product upgrades, and other programs. Consider the following examples:

- ▶ IBM Storage Expert Care service and support IBM Storage Expert Care service and support are simple. You can easily select the level of support and period that best fits your needs with predictable and up-front pricing that is a fixed percentage of the system cost.

Note: For more information, see 1.6.1, “Storage Expert Care” on page 50.

- ▶ The IBM Data Reduction Guarantee helps reduce planning risks and lower storage costs with baseline levels of data compression effectiveness in IBM Storage Virtualize-based offerings.
- ▶ The IBM Controller Upgrade Program enables customers of designated all-flash IBM storage systems to reduce costs while maintaining leading-edge controller technology for essentially the cost of ongoing system maintenance.

The IBM FlashSystem 5200 control enclosure supports up to 12 2.5-inch NVMe-capable flash drives in a 1U high form factor.

Note: The IBM FlashSystem 5200 control enclosure supports the new IBM FCM3 drives if running IBM Storage Virtualize software V8.5 and above. These new drives feature the same capacities as the previous FCM2 drives, but have a higher internal compression ratio of up to 3:1. Therefore, it can effectively store more data, assuming that the data is compressible.

One standard model of IBM FlashSystem 5200 (4662-6H2) and one utility model (4662-UH6) are available.

Figure 1-43 shows the IBM FlashSystem 5200 control enclosure front view with 12 NVMe drives and a 3/4 ISO view.

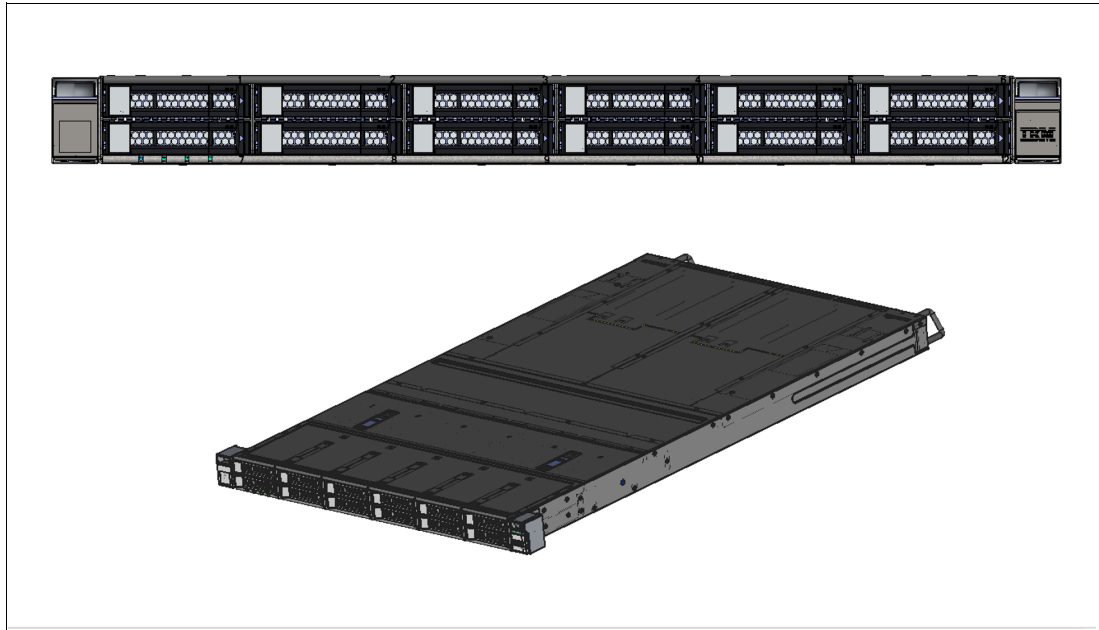


Figure 1-43 IBM FlashSystem 5200 control enclosure front and 3/4 ISO view

Table 1-10 lists the host connections, drive capacities, features, and standard options with IBM Storage Virtualize that are available on IBM FlashSystem 5200.

Table 1-10 IBM FlashSystem 5200 host, drive capacity, and functions summary

Feature or function	Description
Host interface	<ul style="list-style-type: none"> ▶ 10 Gbps Ethernet (iSCSI) ▶ 25 Gbps Ethernet (iSCSI, iSER - iWARP, and RoCE) ▶ 16 Gbps Fibre Channel (FC and FC-NVMe) ▶ 32 Gbps Fibre Channel (FC and FC-NVMe)
Control enclosure supported drives (12 maximum)	<ul style="list-style-type: none"> ▶ 2.5-inch NVMe self-compressing FCMs: 4.8 TB, 9.6 TB, 19.2 TB, and 38.4 TB ▶ NVMe flash drives: 800 GB, 1.92 TB, 3.84 TB, 7.68 TB, and 15.36 TB ▶ NVMe storage-class memory drives: 375 GB, 750 GB, 800 GB, and 1.6 TB
SAS Expansion Enclosures 760 per control enclosure 1,520 per clustered system Model 12G 2U 12 drives Model 24G 2U 24 drives Model 92G 5U 92 drives	<ul style="list-style-type: none"> ▶ 2.5-inch flash drives supported: 800 GB, 1.6 TB, 1.92 TB, 3.84 TB, 7.68 TB, 15.36 TB, and 30.72 TB ▶ 2.5-inch disk drives supported: <ul style="list-style-type: none"> – 600 GB, 900 GB, 1.2 TB, 1.8 TB, and 2.4 TB 10k SAS disk – 2 TB 7.2 K nearline SAS disk ▶ 3.5-inch disk drives supported: 4 TB, 6 TB, 8 TB, 10 TB, 12 TB, 14 TB, 16 TB, and 18 TB 7.2 K nearline SAS disk
RAID levels.	Distributed RAID 5 and 6, TRAIID 1 and 10
Advanced features that are included with each system	<ul style="list-style-type: none"> ▶ Virtualization of internal storage ▶ Data migration ▶ DRPs with thin provisioning ▶ UNMAP ▶ Compression and deduplication ▶ Metro Mirror (synchronous) and Global Mirror (asynchronous)

Feature or function	Description
More available advanced features	<ul style="list-style-type: none"> ▶ Remote mirroring ▶ Policy-based replication ▶ IBM Easy Tier compression ▶ External virtualization ▶ Encryption ▶ FlashCopy ▶ IBM Storage Control ▶ IBM Storage Protect Snapshot

For more information, see [V8.6.0.x Configuration Limits and Restrictions for IBM FlashSystem 50x5 and 5200](#).

1.11.1 IBM FlashSystem 5200 expansion enclosures

The IBM FlashSystem 5200 Model 6H2 and UH6 attach to IBM FlashSystem 5200 Expansion Enclosure Models 4662-12G (2U 12-drive), 4662-24G (2U 24-drive), and 4662-92G (5U 92-drive), which support SAS flash drives and SAS HDD drives.

Note: Attachment and intermixing of IBM Storwize V5100/V5000 Expansion Enclosure Models 12F, 24F, and 92F with IBM FlashSystem 5200 Expansion Enclosure Models 12G, 24G, and 92G is *not* supported by IBM FlashSystem 5200 Model 6H2 and UH6.

The following 2.5-inch SFF flash drives are supported in the expansion enclosures:

- ▶ 400 and 800 GB
- ▶ 1.6, 1.92, 3.2, 3.84, 7.68, 15.36, and 30.72 TB

The following 3.5-inch LFF flash drives are supported in the expansion enclosures:

- ▶ 1.6, 1.92, 3.2, 3.84, 7.68, 15.36, and 30.72 TB
- ▶ 3.5-inch SAS disk drives (Model 12G):
 - 900 GB, 1.2 TB, 1.8 TB, and 2.4 TB 10,000 rpm
 - 4 TB, 6 TB, 8 TB, 10 TB, 12 TB, 14 TB, and 16 TB 7,200 rpm
- ▶ 3.5-inch SAS drives (Model 92G):
 - 1.6 TB, 1.92 TB, 3.2 TB, 3.84 TB, 7.68 TB, 15.36 TB, and 30.72 TB flash drives
 - 1.2 TB, 1.8 TB, and 2.4 TB 10,000 rpm
 - 6 TB, 8 TB, 10 TB, 12 TB, 14 TB, and 16 TB 7,200 rpm
- ▶ 2.5-inch SAS disk drives (Model 24G):
 - 900 GB, 1.2 TB, 1.8 TB, and 2.4 TB 10,000 rpm
 - 2 TB 7,200 rpm
- ▶ 2.5-inch SAS flash drives (Model 24G):
 - 400 and 800 GB
 - 1.6, 1.92, 3.2, 3.84, 7.68, 15.36, and 30.72 TB

1.12 IBM FlashSystem 5000 family overview

IBM FlashSystem 5015, IBM FlashSystem 5035 and IBM FlashSystem 5045 are all-flash and hybrid flash solutions that provide enterprise-grade functions without compromising affordability or performance. They also include the rich features of IBM Storage Virtualize. IBM FlashSystem 5000 helps make modern technologies, such as artificial intelligence (AI), accessible to enterprises of all sizes.

IBM FlashSystem 5000 is a member of the IBM FlashSystem family of storage solutions. It delivers increased performance and new levels of storage efficiency with superior ease of use. This entry storage solution enables organizations to overcome their storage challenges.

The solution includes technology to complement and enhance virtual environments, which delivers a simpler, more scalable, and cost-efficient IT infrastructure. IBM FlashSystem 5000 features two node canisters in a compact, 2U 19-inch rack mount enclosure.

Figure 1-44 shows the IBM FlashSystem 5015, 5035 and 5045 SFF control enclosure front view.



Figure 1-44 IBM FlashSystem 5015, 5035 and 5045 SFF control enclosure front view

Figure 1-45 shows the IBM FlashSystem 5015, 5035 and 5045 LFF control enclosure front view.



Figure 1-45 IBM FlashSystem 5015, 5035 and 5045 LFF control enclosure front view

Table 1-11 lists the model comparison chart for the IBM FlashSystem 5000 family.

Table 1-11 Machine type and model comparison for the IBM FlashSystem 5000

MTM	Full name
2072-2N2	IBM FlashSystem 5015 LFF control enclosure
2072-2N4	IBM FlashSystem 5015 SFF control enclosure
2072-3N2	IBM FlashSystem 5035 LFF control enclosure
2072-3N4	IBM FlashSystem 5035 SFF control enclosure
2072-12G	IBM FlashSystem 5000 LFF expansion enclosure
2072-24G	IBM FlashSystem 5000 SFF expansion enclosure
2072-92G	IBM FlashSystem 5000 High-Density LFF expansion enclosure
4680-2P2	IBM FlashSystem 5015 LFF control enclosure (with Storage Expert Care)
4680-2P4	IBM FlashSystem 5015 SFF control enclosure (with Storage Expert Care)
4680-3P2	IBM FlashSystem 5045 LFF control enclosure (with Storage Expert Care)
4680-3P4	IBM FlashSystem 5045 SFF control enclosure (with Storage Expert Care)
4680-12H	IBM FlashSystem 5000 LFF expansion enclosure (with Storage Expert Care)
4680-24H	IBM FlashSystem 5000 SFF expansion enclosure (with Storage Expert Care)
4680-92H	IBM FlashSystem 5000 High-Density LFF expansion enclosure (with Storage Expert Care)

Note: IBM FlashSystems 5015/5035 (M/T2072) must use the 2072 XXG models of expansion enclosures. Similarly, the 5105/5045 (M/T 4680) must use the 4680 XXH models on expansion enclosures. The two type cannot be intermixed.

1.12.1 IBM FlashSystem 5015

IBM FlashSystem 5015 is an entry-level solution that is focused on affordability and ease of deployment and operation, with powerful scale-up features. It includes many IBM Storage Virtualize features and offers multiple flash and disk drive storage media and expansion options.

Table 1-12 lists the host connections, drive capacities, features, and standard options with IBM Storage Virtualize that are available on IBM FlashSystem 5015.

Table 1-12 IBM FlashSystem 5015 host, drive capacity, and functions summary

Feature / Function	Description
Host interface	<ul style="list-style-type: none"> ▶ 1 Gb iSCSI (on the system board) ▶ 16 Gbps Fibre Channel ▶ 12 Gbps SAS ▶ 25 Gbps iSCSI (iWARP or RoCE) ▶ 10 Gbps iSCSI
Control enclosure and SAS expansion enclosures supported drives	<ul style="list-style-type: none"> ▶ For SFF enclosures, see Table 1-13 on page 80 ▶ For LFF enclosures, see Table 1-14 on page 81
Cache per control enclosure/ clustered system	32 GB or 64 GB
RAID levels	DRAID 1, 5, and 6
Maximum expansion enclosure capacity	<ul style="list-style-type: none"> ▶ Up to 10 standard expansion enclosures per controller ▶ Up to four high-density expansion enclosures per controller
Advanced functions that are included with each system	<ul style="list-style-type: none"> ▶ Virtualization of internal storage ▶ DRPs with thin provisioning and UNMAP ▶ One-way data migration
More available advanced features	<ul style="list-style-type: none"> ▶ Easy Tier ▶ FlashCopy ▶ Remote mirroring

Table 1-13 lists the 2.5-inch supported drives for IBM FlashSystem 5000 family.

Table 1-13 2.5-inch supported drives for the IBM FlashSystem 5000 family

2.5-inch (SFF)	Capacity					
Tier 1 flash	800 GB	1.9 TB	3.84 TB	7.68 TB	15.36 TB	30.72 TB
High-performance enterprise disk drives (10K rpm)	900 GB	1.2 TB	1.8 TB	2.4 TB		
High capacity nearline disk drives (7.2 K rpm)	2 TB					

Table 1-14 on page 81 lists the 3.5-inch supported drives for IBM FlashSystem 5000 family.

Table 1-14 3.5-inch supported drives for the IBM FlashSystem 5000 family

3.5-inch (LFF)	Speed	Capacity							
High-performance, enterprise class disk drives	10,000 RPM	900 GB	1.2 TB	1.8 TB	2.4 TB				
High capacity archival class nearline disk drives	7,200 RPM	4 TB	6 TB	8 TB	10 TB	12 TB	14 TB	16 TB	18 TB

1.12.2 IBM FlashSystem 5035

IBM FlashSystem 5035 provides powerful functions, including powerful encryption capabilities and DRPs with compression, deduplication, thin provisioning, and the ability to cluster for scale-up and scale-out.

There are four models of IBM FlashSystem 5035 as follows:

- ▶ 2072-3N2 (2U 12 Drive Large Form Factor LFF)
- ▶ 2072-3N4 (2U 24 Drive Small Form Factor SFF)

The IBM FlashSystem 5000 expansion enclosures are available in the following form factors:

- ▶ 2U 12 Drive Large Form Factor LFF Model 12H
- ▶ 2U 24 Drive Small Form Factor SFF Model 24H
- ▶ 2U 92 Drive HD Form Factor LFF Model 92H

Available with the IBM FlashSystem 5035 model, DRPs help transform the economics of data storage. When applied to new or existing storage, they can increase usable capacity while maintaining consistent application performance. DRPs can help eliminate or drastically reduce costs for storage acquisition, rack space, power, and cooling, and can extend the useful life of storage assets.

DRPs feature the following capabilities:

- ▶ Block deduplication that works across all the storage in a DRP to minimize the number of identical blocks.
- ▶ New compression technology that ensures consistent 2:1 or better reduction performance across a wide range of application workload patterns.
- ▶ SCSI UNMAP support that de-allocates physical storage when operating systems delete logical storage constructs, such as files in a file system.

Table 1-15 lists the host connections, drive capacities, features, and standard options with IBM Storage Virtualize that are available on IBM FlashSystem 5035.

Table 1-15 IBM FlashSystem 5035 host, drive capacity, and functions summary

Feature / Function	Description
Host interface	<ul style="list-style-type: none"> ▶ 10 Gb iSCSI (on the system board) ▶ 16 Gbps Fibre Channel ▶ 12 Gbps SAS ▶ 25 Gbps iSCSI (iWARP or RoCE) ▶ 10 Gbps iSCSI
Control enclosure and SAS expansion enclosures supported drives	<ul style="list-style-type: none"> ▶ For SFF enclosures, see Table 1-13 on page 80 ▶ For LFF enclosures, see Table 1-14 on page 81

Feature / Function	Description
Cache per control enclosure/ clustered system	32 GB or 64 GB / 64 GB or 128 GB
RAID levels	DRAID 1, 5 (CLI only), and 6
Maximum expansion enclosure capacity	<ul style="list-style-type: none"> ▶ Up to 20 standard expansion enclosures per controller ▶ Up to eight high-density expansion enclosures per controller
Advanced functions that are included with each system	<ul style="list-style-type: none"> ▶ Virtualization of internal storage ▶ DRPs with thin provisioning ▶ UNMAP, compression, and deduplication ▶ One-way data migration ▶ Dual-system clustering
More available advanced features	<ul style="list-style-type: none"> ▶ Easy Tier ▶ FlashCopy ▶ Remote mirroring ▶ Encryption

For more information, see [V8.6.0.x Configuration Limits and Restrictions for IBM FlashSystem 50x5 and 5200](#).

This next section provides hardware information about the IBM FlashSystem 5035 models.

The IBM FlashSystem 5035 control enclosure features the following components:

- ▶ Two node canisters, each with a six-core processor
- ▶ 32 GB cache (16 GB per canister) with optional 64 GB cache (32 GB per canister)
- ▶ 10 Gb iSCSI (copper) connectivity standard with optional 16 Gb FC, 12 Gb SAS, 10 Gb iSCSI (optical), or 25 Gb iSCSI (optical)
- ▶ 12 Gb SAS port for expansion enclosure attachment
- ▶ 12 slots for 3.5-inch LFF SAS drives (Model 3N2) and 24 slots for 2.5-inch SFF SAS drives (Model 3N4)
- ▶ 2U, 19-inch rack mount enclosure with 100 - 240 V AC or -48 V DC power supplies

The IBM FlashSystem 5035 control enclosure models offer the highest level of performance, scalability, and functions and include the following features:

- ▶ Support for 760 drives per system with the attachment of eight IBM FlashSystem 5000 High-Density LFF expansion enclosures and 1,520 drives with a two-way clustered configuration
- ▶ DRPs with deduplication, compression,² and thin provisioning for improved storage efficiency
- ▶ Encryption of data-at-rest that is stored within the IBM FlashSystem 5035 system

² Deduplication and compression require 64 GB of system cache.

Figure 1-46 lists the IBM FlashSystem 5035 SFF control enclosure with 24 drives.



Figure 1-46 Front view of an IBM FlashSystem 5035

Figure 1-47 lists the rear view of an IBM FlashSystem 5035 control enclosure.



Figure 1-47 Rear view of an IBM FlashSystem 5035

Figure 1-48 lists the available connectors and LEDs on a single IBM FlashSystem 5035 canister.

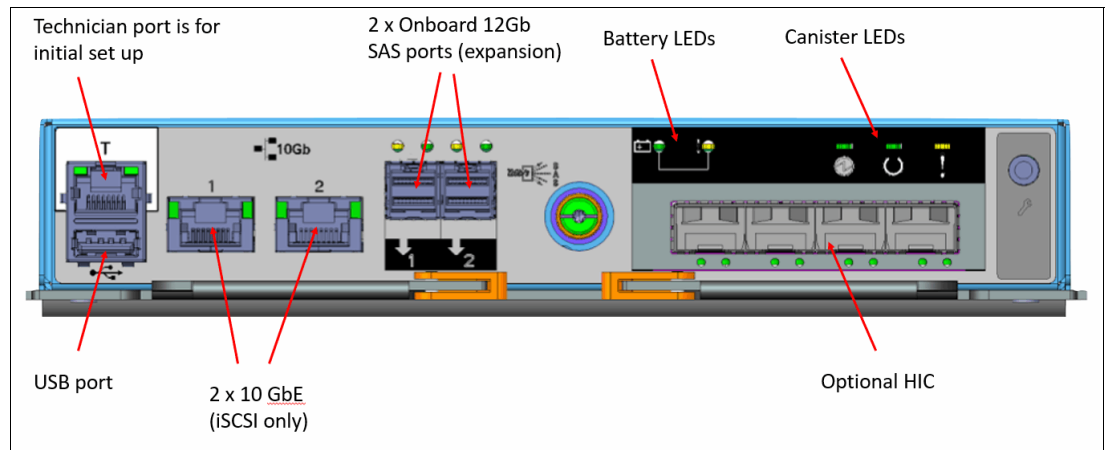


Figure 1-48 View of available connectors and LEDs on an IBM FlashSystem 5035 single canister

For more information, see [V8.4.2.x Configuration Limits and Restrictions for IBM FlashSystem 5x00](#).

1.12.3 IBM FlashSystem 5045

IBM FlashSystem 5045 is a lightweight FlashSystem 5035 replacement, which uses the same base hardware as the FlashSystem 5035 but with software modifications to provide support for software features that are not available on FlashSystem 5035.

Additional software features to be supported:

- ▶ Safeguarded Copy
- ▶ FlashCopy 2 with internal scheduler

IBM FlashSystem 5045 also has a new machine type and model to support the IBM Storage Expert care service offerings:

- ▶ 4680-3P2 (2U 12 Drive Large Form Factor LFF)
- ▶ 4680-3P4 (2U 24 Drive Small Form Factor SFF)

The IBM FlashSystem 5000 expansion enclosures are available in the following form factors:

- ▶ 2U 12 Drive Large Form Factor LFF Model 12H
- ▶ 2U 24 Drive Small Form Factor SFF Model 24H
- ▶ 2U 92 Drive HD Form Factor LFF Model 92H

IBM FlashSystem 5045 provides powerful functions, including powerful encryption capabilities and DRPs with compression, deduplication, thin provisioning, and the ability to cluster for scale-up and scale-out.

Available with the IBM FlashSystem 5045 model, DRPs help transform the economics of data storage. When applied to new or existing storage, they can increase usable capacity while maintaining consistent application performance. DRPs can help eliminate or drastically reduce costs for storage acquisition, rack space, power, and cooling, and can extend the useful life of storage assets.

DRPs feature the following capabilities:

- ▶ Block deduplication that works across all the storage in a DRP to minimize the number of identical blocks.
- ▶ New compression technology that ensures consistent 2:1 or better reduction performance across a wide range of application workload patterns.
- ▶ SCSI UNMAP support that de-allocates physical storage when operating systems delete logical storage constructs, such as files in a file system.

Table 1-16 lists the host connections, drive capacities, features, and standard options with IBM Storage Virtualize that are available on IBM FlashSystem 5045.

Table 1-16 IBM FlashSystem 5045 host, drive capacity, and functions summary

Feature / Function	Description
Host interface	<ul style="list-style-type: none"> ▶ 10 Gb iSCSI (on the system board) ▶ 16 Gbps Fibre Channel ▶ 12 Gbps SAS ▶ 25 Gbps iSCSI (iWARP or RoCE) ▶ 10 Gbps iSCSI
Control enclosure and SAS expansion enclosures supported drives	<ul style="list-style-type: none"> ▶ For SFF enclosures, see Table 1-13 on page 80 ▶ For LFF enclosures, see Table 1-14 on page 81 <ul style="list-style-type: none"> – NOTE Support for Enterprise 15K rpm drives is removed for IBM FlashSystem 5045
Cache per control enclosure/ clustered system	32 GB or 64 GB / 64 GB or 128 GB
RAID levels	DRAID 1, 5 (CLI only), and 6

Feature / Function	Description
Maximum expansion enclosure capacity	<ul style="list-style-type: none"> ▶ Up to 12 standard expansion enclosures per controller ▶ Up to two high-density expansion enclosures per controller
Advanced functions that are included with each system	<ul style="list-style-type: none"> ▶ Virtualization of internal storage ▶ DRPs with thin provisioning ▶ UNMAP, compression, and deduplication ▶ One-way data migration ▶ Dual-system clustering ▶ Volume Mobility ▶ Easy Tier ▶ FlashCopy ▶ Remote mirroring
More available advanced features	<ul style="list-style-type: none"> ▶ Encryption

IBM FlashSystem 5045 does not cluster with IBM FlashSystem 5035 original or any models other than IBM FlashSystem 5045. Otherwise, in common with FlashSystem 5035, a maximum of two I/O groups per cluster is allowed. IBM FlashSystem 5045 also has a reduction in Volume Group extents from 4 million to 1 million.

IBM FlashSystem 5045 uses a License Machine Code (LMC) model similar to that used by FlashSystem 5200. This is a change from FlashSystem 5035 as all feature software (apart from encryption) is included with the base license. Encryption requires a hardware licence key.

In keeping with current trends, expansion chain weight is reduced from 20U in FlashSystem 5035 to 12U for Flash System 5045.

This means that a single chain can contain up to 2x5U92 expansions or up to 6x2U12/24 expansions or a combination of 5U and 2U not exceeding 12U in total. This reduces the maximum number of expansion drive slots per chain, per I/O Group and per system.

- ▶ per chain: $2 \times 92 + 1 \times 24 + 24 = 232$ drives
- ▶ per I/O Group: 2 chains per I/O Group so $4 \times 92 + 2 \times 24 + 24 = 440$ drives
- ▶ per System: 2 I/O Groups so $2 \times 440 = 880$ drives

For more information, see [V8.4.2.x Configuration Limits and Restrictions for IBM FlashSystem 5x00](#).

This next section provides hardware information about the IBM FlashSystem 5045 models.

The IBM FlashSystem 5045 control enclosure features the following components:

- ▶ Two node canisters, each with a six-core processor
- ▶ 32 GB cache (16 GB per canister) with optional 64 GB cache (32 GB per canister)
- ▶ 10 Gb iSCSI (copper) connectivity standard with optional 16 Gb FC, 12 Gb SAS, 10 Gb iSCSI (optical), or 25 Gb iSCSI (optical)
- ▶ 12 Gb SAS port for expansion enclosure attachment
- ▶ 12 slots for 3.5-inch LFF SAS drives (Model 3N2) and 24 slots for 2.5-inch SFF SAS drives (Model 3N4)
- ▶ 2U, 19-inch rack mount enclosure with 100 - 240 V AC power supplies

The IBM FlashSystem 5045 control enclosure models offer the highest level of performance, scalability, and functions and include the following features:

- ▶ Support for 760 drives per system with the attachment of eight IBM FlashSystem 5000 High-Density LFF expansion enclosures and 1,520 drives with a two-way clustered configuration
- ▶ DRPs with deduplication, compression,³ and thin provisioning for improved storage efficiency
- ▶ Encryption of data-at-rest that is stored within the IBM FlashSystem 5045 system

Figure 1-46 on page 83 lists the IBM FlashSystem 5045 SFF control enclosure with 24 drives.



Figure 1-49 Front view of an IBM FlashSystem 5045

Figure 1-47 on page 83 lists the rear view of an IBM FlashSystem 5045 control enclosure.



Figure 1-50 Rear view of an IBM FlashSystem 5045

Figure 1-48 on page 83 lists the available connectors and LEDs on a single IBM FlashSystem 5045 canister.

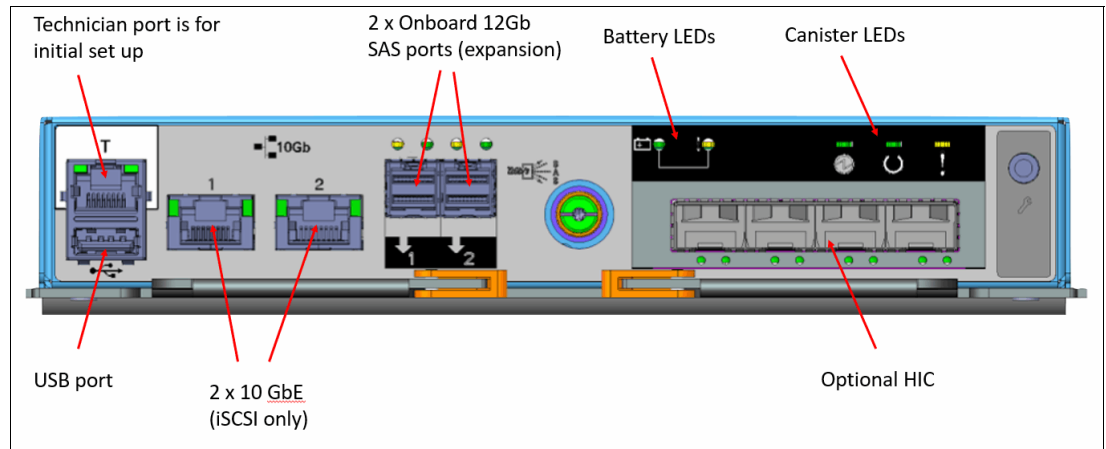


Figure 1-51 View of available connectors and LEDs on an IBM FlashSystem 5045 single canister

³ Deduplication and compression require 64 GB of system cache.

For more information, see [V8.4.2.x Configuration Limits and Restrictions for IBM FlashSystem 5x00](#).

1.13 Storage efficiency and data reduction features

IBM Storage Virtualize software that is running in the IBM FlashSystem storage systems or IBM SAN Volume Controllers offers several functions for storage optimization and efficiency, as described in this section.

1.13.1 IBM Easy Tier

Many applications exhibit a significant skew in the distribution of I/O workload. A small fraction of the storage is responsible for a disproportionately large fraction of the total I/O workload of an environment.

In a tiered storage pool, IBM Easy Tier acts to identify this skew and automatically place data in the suitable tier to take advantage of it. By moving the hottest data onto the fastest tier of storage, the workload on the remainder of the storage is reduced. By servicing most of the application workload from the fastest storage, Easy Tier acts to accelerate application performance.

Easy Tier is a performance optimization function that automatically migrates extents that belong to a volume among different storage tiers based on their I/O load. The movement of the extents is online and unnoticed from a host perspective.

As a result of extent movement, the volume no longer has all its data in one tier, but rather in two or three tiers. Each tier provides optimal performance for the extent, as shown in Figure 1-52.



Figure 1-52 Easy Tier concept

Easy Tier monitors the I/O activity and latency of the extents on all Easy Tier enabled storage pools to create *heat maps*. Based on these maps, Easy Tier creates an extent migration plan and promotes (or moves) high activity or hot extents to a higher disk tier within the same storage pool. It also demotes extents whose activity dropped off, or cooled, by moving them from a higher disk tier managed disk (MDisk) back to a lower tier MDisk.

Storage pools that contain only one tier of storage also can benefit from Easy Tier if they include multiple disk arrays (or MDisks). Easy Tier has a balancing mode: It moves extents from busy disk arrays to less busy arrays of the same tier, which balances I/O load.

All MDisks (disk arrays) belong to one of the tiers. They are classified as SCM, Flash, Enterprise, or NL tier.

For more information, see [Easy Tier function](#).

1.13.2 Data reduction and UNMAP

The UNMAP feature is a set of SCSI primitives that enables hosts to indicate to a SCSI target (storage system) that space that is allocated to a range of blocks on a target storage volume is no longer required. This command enables the storage controller to take measures and optimize the system so that the space can be reused for other purposes.

For example, the most common use case is a host application, such as VMware, which frees storage in a file system. Then, the storage controller can perform functions to optimize the space, such as reorganizing the data on the volume so that space is better used.

When a host allocates storage, the data is placed in a volume. To return the allocated space to the storage pools, the SCSI UNMAP feature is used. UNMAP enables host operating systems to deprovision storage on the storage controller so that the resources can automatically be freed in the storage pools and used for other purposes.

A DRP increases infrastructure capacity use by using new efficiency functions and reducing storage costs. By using the end-to-end SCSI UNMAP function, a DRP can automatically de-allocate and reclaim the capacity of thin-provisioned volumes that contain deleted data so that this reclaimed capacity can be reused by other volumes.

At its core, a DRP uses a Log Structured Array (LSA) to allocate capacity. An LSA enables a tree-like directory to be used to define the physical placement of data blocks that are independent of size and logical location. Each logical block device has a range of logical block addresses (LBAs), starting from 0 and ending with the block address that fills the capacity.

When written, you can use an LSA to allocate data sequentially and provide a directory that provides a lookup to match an LBA with a physical address within the array. Therefore, the volume that you create from the pool to present to a host application consists of a directory that stores the allocation of blocks within the capacity of the pool.

In DRPs, the maintenance of the metadata results in *I/O amplification*. I/O amplification occurs when a single host-generated read or write I/O results in more than one back-end storage I/O request because of advanced functions. A read request from the host results in two I/O requests: a directory lookup and a data read. A write request from the host results in three I/O requests: a directory lookup, a directory update, and a data write. This aspect must be considered when sizing and planning your data-reducing solution.

Standard pools, which make up a classic solution that also is supported by the IBM FlashSystem and IBM SAN Volume Controller system, do not use LSA. A standard pool works as a container that receives its capacity from MDisks (disk arrays), splits it into extents of the same fixed size, and allocates extents to volumes.

Standard pools do not cause I/O amplification and require less processing resource usage compared to DRPs. In exchange, DRPs provide more flexibility and storage efficiency.

Table 1-17 lists the volume capacity saving types that are available with standard pools and DRPs.

Table 1-17 Volume types that are available in pools

Saving type	Standard pool	DRP
Fully allocated	Yes	Yes
Thin-provisioned	Yes	Yes
Thin-provisioned compressed	No	Yes
Thin-provisioned deduplicated	No	Yes
Thin-provisioned compressed and deduplicated	No	Yes

Best practice: If you want to use deduplication, create thin-provisioned compressed and deduplicated volumes.

This book provides only an overview of DRP aspects. For more information, see *Introduction and Implementation of Data Reduction Pools and Deduplication*, [SG24-8430](#).

Fully allocated and thin-provisioned volumes

Volumes can be configured as thin-provisioned or fully allocated. Both versions can be configured in standard pools and DRP pools.

In IBM Storage Virtualized systems, each volume includes virtual capacity and real capacity parameters:

- ▶ *Virtual capacity* is the volume storage capacity that is available to a host. It is used by the host operating system to create a file system.
- ▶ *Real capacity* is the storage capacity that is allocated to a volume from a pool. It shows the amount of space that is used on a physical storage volume.

Fully allocated

Fully allocated volumes are created with the same amount of real capacity and virtual capacity. This type uses no storage efficiency features.

When a fully allocated volume is created on a DRP, it bypasses the LSA structure and works in the same manner as in a standard pool; Therefore, it has no effect on processing and provides no data reduction options at the pool level.

When fully allocated volumes are used on the IBM Storage Virtualized systems with FCM drives (whether a DRP or standard pool is used), capacity savings are achieved by compressing data with hardware compression that runs on the FCM drives. Hardware compression on FCM drives is always on and cannot be turned off. This configuration provides maximum performance in combination with outstanding storage efficiency.

Thin-provisioned

A thin-provisioned volume presents a different capacity to mapped hosts than the capacity that the volume uses in the storage pool. Therefore, real and virtual capacities might not be equal. The virtual capacity of a thin-provisioned volume is typically significantly larger than its real capacity. As more information is written by the host to the volume, more of the real capacity is used. The system identifies read operations to unwritten parts of the virtual capacity, and returns zeros to the server without the use of any real capacity.

In a shared storage environment, thin provisioning is a method for optimizing the use of available storage. Thin provisioning relies on the allocation of blocks of data on demand, versus the traditional method of allocating all of the blocks up front. This method eliminates almost all white space, which helps avoid the poor usage rates that occur in the traditional storage allocation method where large pools of storage capacity are allocated to individual servers but remain unused (not written to).

If a thin-provisioned volume is created in a DRP, the system monitors it for reclaimable capacity from host unmap operations. This capacity can be reclaimed and redistributed into the pool.

Space that is freed from the hosts is a process that is called *UNMAP*. A host can issue **SCSI UNMAP** commands when the user deletes files on a file system, which result in the freeing of all of the capacity that is allocated within that unmapping.

A thin-provisioned volume in a standard pool does not return unused capacity to the pool with **SCSI UNMAP**.

A thin-provisioned volume can be converted nondisruptively to a fully allocated volume, or vice versa, by using the volume mirroring function. For example, you can add a thin-provisioned copy to a fully allocated primary volume and then, remove the fully allocated copy from the volume after they are synchronized.

Note: It is *not* recommended to use noncompressed thin-provisioned volumes on DRPs that contain FCM drives. The system GUI prevents the creation of such types of configurations.

1.13.3 Compression and deduplication

When DRPs are used on IBM Storage Virtualized systems, host data can be compressed or compressed and deduplicated before it is written to the disk drives.

The IBM FlashSystem and IBM SAN Volume Controller family DRP compression is based on the Lempel-Ziv lossless data compression algorithm that operates by using a real-time method. When a host sends a write request, the request is acknowledged by the write cache of the system and then, staged to the DRP.

As part of its staging, the write request passes through the compression engine and is stored in a compressed format. Therefore, writes are acknowledged immediately after they are received by the write cache with compression occurring as part of the staging to internal or external physical storage. This process occurs transparently to host systems, which makes them unaware of the compression.

IBM Comprestimator tool

The IBM Comprestimator tool is available to check whether your data is compressible. It estimates the space savings that are achieved when compressed volumes are used. This utility provides a quick and easy view of showing the benefits of using compression. IBM Comprestimator can be run from the system GUI or command-line interface (CLI), and it checks data that is already stored on the system. In DRPs, IBM Comprestimator is always on starting at code level 8.4, so you can display the compressibility of the data in the GUI and IBM Storage Insights at any time.

IBM Data Reduction Estimation tool

The IBM Data Reduction Estimation tool (DRET) is a host-based command-line utility for estimating the data reduction savings on block storage devices. To help with the profiling and analysis of user workloads that must be migrated to a new system, IBM provides the highly accurate DRET to support both deduplication and compression.

The tool scans target workloads on various earlier storage arrays (from IBM or another company), merges all scan results and then, provides an integrated system-level data reduction estimate.

Both tools are available as stand-alone, host-based utilities that can analyze data on IBM or third-party storage devices. For more information, see [Data Reduction Estimation Tool](#).

Deduplication can be configured with thin-provisioned and compressed volumes in DRPs for added capacity savings. The deduplication process identifies unique chunks of data, or byte patterns, and stores a signature of the chunk for reference when writing new data chunks.

If the new chunk's signature matches a signature, the new chunk is replaced with a small reference that points to the stored chunk. The matches are detected when the data is written. The same byte pattern might occur many times, which greatly reduce the amount of data that must be stored.

Compression and deduplication are not mutually exclusive: One, both, or none of the features can be enabled. If the volume is deduplicated and compressed, data is deduplicated first and then, compressed. Therefore, deduplication references are created on the compressed data that is stored on the physical domain.

1.13.4 Enhanced data security features

To protect data against the potential exposure of sensitive user data and user metadata that is stored on discarded, lost, or stolen storage devices, IBM FlashSystem storage systems and IBM SAN Volume Controller support encryption of data-at-rest.

Encryption is performed by the IBM FlashSystem or IBM SAN Volume Controller controllers for data that is stored:

- ▶ Within the entire system
- ▶ The IBM FlashSystem control enclosure
- ▶ All attached expansion enclosures
- ▶ As externally virtualized by the IBM FlashSystem or IBM SAN Volume Controller storage systems

Encryption is the process of encoding data so that only authorized parties can read it. Data encryption is protected by the Advanced Encryption Standard (AES) algorithm that uses a 256-bit symmetric encryption key in XTS mode, as defined in the IEEE 1619-2007 standard and NIST Special Publication 800-38E as XTS-AES-256.

Two types of encryption are available on devices that are running IBM Storage Virtualize: hardware and software. Which method is used for encryption is chosen automatically by the system based on the placement of the data:

- ▶ **Hardware encryption:** Data is encrypted by IBM FlashCore module (FCM) hardware and SAS hardware for expansion enclosures. It is used only for internal storage (drives).
- ▶ **Software encryption:** Data is encrypted by using the nodes' CPU (the encryption code uses the AES-NI CPU instruction set). It is used only for external storage that is virtualized by the IBM FlashSystem and IBM SAN Volume Controller managed storage systems.

Both methods of encryption use the same encryption algorithm, key management infrastructure, and license.

Note: Only data-at-rest is encrypted. Host-to-storage communication and data that is sent over links that are used for remote mirroring are *not* encrypted.

The IBM FlashSystem also supports self-encrypting drives, in which data encryption is completed in the drive.

Before encryption can be enabled, ensure that a license was purchased and activated.

The system supports two methods of configuring encryption:

- ▶ You can use a centralized external key server that simplifies creating and managing encryption keys on the system. This method of encryption key management is preferred for security and simplification of key management.

A *key server* is a centralized system that generates, stores, and sends encryption keys to the system. Some key server providers support replicating keys among multiple key servers. If multiple key servers are supported, you can specify up to four key servers that connect to the system over a public network or separate private network.

- ▶ The system supports IBM Guardium Key Lifecycle Manager (formerly IBM Security Key Lifecycle Manager) or Gemalto SafeNet KeySecure key servers to handle key management.
- ▶ The system also supports storing encryption keys on USB flash drives. USB flash drive-based encryption requires physical access to the systems and is effective in environments with a minimal number of systems. For organizations that require strict security policies regarding USB flash drives, the system supports disabling a canister's USB ports to prevent unauthorized transfer of system data to portable media devices. If you have such security requirements, use key servers to manage encryption keys.

Another data security enhancement that delivered with the IBM Storage Virtualize code is the new Safeguarded Copy feature that can provide protected read-only logical air gap copies of volumes. This enhancement gives the customer effective data protection against cyber attacks.

For more information, see "Safeguarded Copy" on page 18.

1.14 Application integration features

IBM FlashSystem storage systems include the following features, which enable tight integration with VMware:

- ▶ vCenter plug-in
 - Enables monitoring and self-service provisioning of the system from within VMware vCenter.
- ▶ vStorage application programming interfaces (APIs) for Array Integration (VAAI) support:
 - This function supports hardware-accelerated virtual machine (VM) copy/migration and hardware-accelerated VM initiation, and accelerates VMware Virtual Machine File System (VMFS).
- ▶ Microsoft Windows System Resource Manager (SRM) for VMware Site Recovery Manager
 - Supports automated storage and host failover, failover testing, and failback.
- ▶ VMware vSphere Virtual Volume (VVOL) integration for better usability
 - The migration of space-efficient volumes between storage containers maintains the space efficiency of volumes. Cloning a VM achieves a full independent set of VVOLS, and resiliency is improved for VMs if volumes start running out of space.

VMware vSphere Virtual Volumes

The system supports VVOLS, which allow VMware vCenter to automate the management of system objects, such as volumes and pools. It is an integration and management framework that virtualizes the IBM FlashSystem storage systems, which enables a more efficient operational model that is optimized for virtualized environments and centered on the application instead of the infrastructure.

VVOLS simplify operations through policy-driven automation that enables more agile storage consumption for VMs and dynamic adjustments in real time when they are needed. They also simplify the delivery of storage service levels to individual applications by providing finer control of hardware resources and native array-based data services that can be instantiated with VM granularity.

With VVOLS, VMware offers a paradigm in which an individual VM and its disks, rather than a logical unit number (LUN), becomes a unit of storage management for a storage system. It encapsulates VDisks and other VM files, and natively stores the files on the storage system.

By using a special set of APIs that are called vSphere APIs for Storage Awareness (VASA), the storage system becomes aware of the VVOLS and their associations with the relevant VMs. Through VASA, vSphere and the underlying storage system establish a two-way out-of-band communication to perform data services and offload certain VM operations to the storage system. For example, some operations, such as snapshots and clones, can be offloaded.

For more information about VVOLS and the actions that are required to implement this feature on the host side, see [Understanding Virtual Volumes \(vVols\) in VMware vSphere 6.7/7.0/8.0](#) and [IBM Storage Virtualize and VMware: Integrations, Implementation and Best Practices, SG24-8549](#).

IBM support for VASA is provided by IBM Storage Connect enabling communication between the VMware vSphere infrastructure and the IBM FlashSystem system. The IBM FlashSystem administrator can assign ownership of VVOLs to IBM Storage Connect by creating a user with the VASA Provider security role.

Although the system administrator can complete specific actions on volumes and pools that are owned by the VASA Provider security role, IBM Storage Connect retains management responsibility for VVOLs. For more information, see [IBM Storage Connect documentation](#).

Note: At the time of this writing, VVOLs are *not* supported on DRPs. However, they are still valid with 8.6.0.

1.15 Features for manageability

IBM FlashSystem storage systems and the IBM SAN Volume Controller offer the following manageability and serviceability features:

- ▶ Intuitive GUI
- ▶ IBM Call Home and Remote Support
- ▶ IBM Storage Insights
- ▶ IBM Storage Virtualize RESTful API

IBM Storage Virtualize family system GUI

The IBM storage systems that are running Storage Virtualize include an easy-to-use management GUI that runs on one of the node canisters in the control enclosure. This GUI helps you to monitor, manage, and configure your system. You can access the GUI by opening any supported web browser and entering the management IP addresses.

The IBM storage systems use a GUI with the same look and feel across all platforms for a consistent management experience. The GUI includes an improved overview dashboard that provides all of the information in an easy-to-understand format and enables visualization of effective capacity. By using the GUI, you can quickly deploy storage and manage it efficiently.

Figure 1-53 shows the Storage Virtualize GUI dashboard view. This default view is displayed after the user logs on to the system.

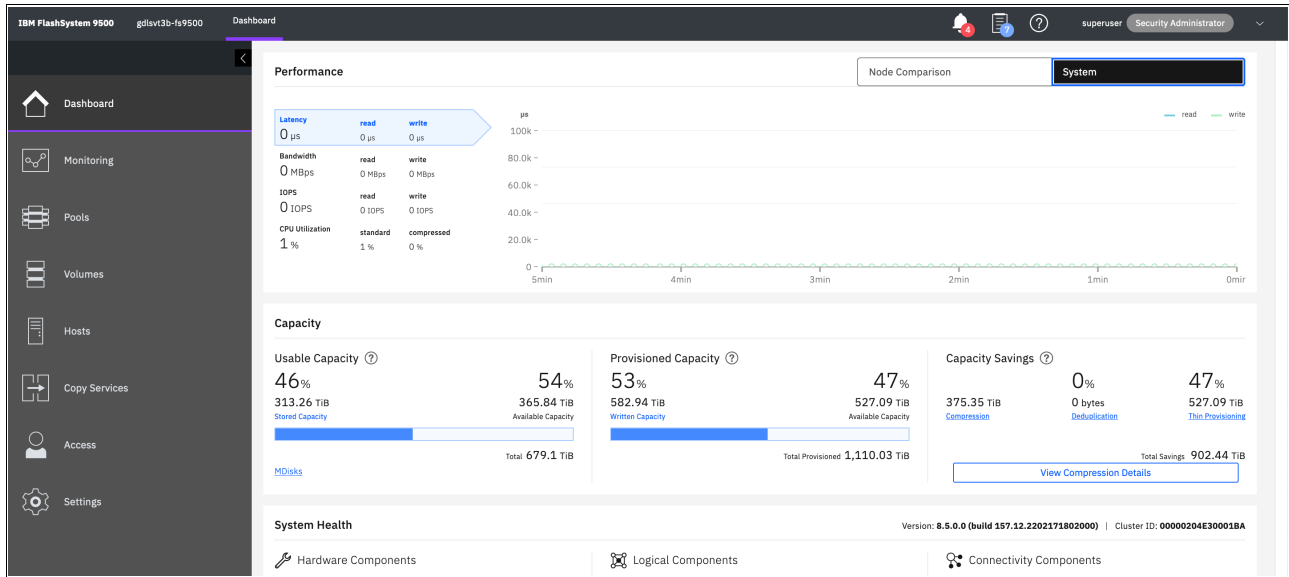


Figure 1-53 IBM Storage Virtualize GUI dashboard

The IBM FlashSystem storage systems and the IBM SAN Volume Controller also provide a CLI, which is useful for advanced configuration and scripting.

The systems support SNMP, email notifications that use Simple Mail Transfer Protocol (SMTP), and syslog redirection for complete enterprise management access.

IBM Call Home and Remote Support

IBM Call Home connects the system to IBM Service Personnel who can monitor and respond to system events to ensure that your system remains running.

The IBM Call Home function opens a service alert if a serious error occurs in the system, which automatically sends the details of the error and contact information to IBM Service Personnel.

If the system is eligible for support, a Cognitive Support Program (CSP) ticket is automatically created and assigned to the suitable IBM Support team. The information that is provided to IBM is an excerpt from the event log that contains the details of the error, and customer contact information from the system.

IBM Service Personnel contact the customer and arrange service on the system, which can greatly improve the speed of resolution by removing the need for the customer to detect the error and raise a support call themselves.

The system supports the following methods to transmit notifications to the support center:

- ▶ Call Home with cloud services

Call Home with cloud services sends notifications directly to a centralized file repository that contains troubleshooting information that is gathered from customers. Support personnel can access this repository and automatically be assigned issues as problem reports.

This method of transmitting notifications from the system to support removes the need for customers to create problem reports manually. Call Home with cloud services also eliminates email filters dropping notifications to and from support, which can delay resolution of problems on the system.

This method sends notifications only to the predefined support center.

► Call Home with email notifications

Call Home with email notification sends notifications through a local email server to support and local users or services that monitor activity on the system. With email notifications, you can send notifications to support and designate internal distribution of notifications, which alerts internal personnel about potential problems. Call Home with email notifications requires configuring at least one email server, and local users.

However, external notifications to the support center can be dropped if filters on the email server are active. To eliminate this problem, Call Home with email notifications is not recommended as the only method to transmit notifications to the support center. Call Home with email notifications can be configured with cloud services.

IBM highly encourages all customers to take advantage of the Call Home feature so that you and IBM can collaborate for your success.

For more information, see [IBM Storage Virtualize Products Call Home and Remote Support Overview](#).

IBM Call Home Connect Cloud

IBM Call Home Connect Cloud (CHCC) is a web application that allows IBM hardware clients to view and monitor key status indicators about their Call Home-enabled IBM hardware assets. Offered at no additional cost to all IBM hardware clients, the website displays information about:

- Critical cases and alerts.
- Warranty and maintenance contract status.
- Last contact status.
- Current software and firmware levels and upgrade recommendations.
- Asset details, such as:
 - Summary data about open and closed cases, with links to IBM Support to view them.
 - Detailed alerts about various Call Home-related events.
 - Links to product-specific support information.

Clients can also:

- Monitor Remote Code Load services on their assets in real time.
- Look up their assets' warranty and maintenance contract information.
- Visualize and export Call Home data in several formats.

Call Home Connect Cloud works offline, too, although real-time updates are not available while offline. For clients who want a more tailored mobile experience, we offer a mobile companion app, Call Home Connect Anywhere, for Android and iOS devices.

For more information about CHCC, see links here, [Introducing IBM Call Home Connect Cloud](#).

IBM Storage Insights

IBM Storage Insights is an IBM Cloud software as a service (SaaS) offering that can help you monitor and optimize the storage resources in the system and across your data center. IBM Storage Insights monitors your storage environment and provides information about the statuses of multiple systems in a single dashboard.

You can view data from the perspectives of the servers, applications, and file systems. Two versions of IBM Storage Insights are available: IBM Storage Insights and IBM Storage Insights Pro.

When you order any IBM FlashSystem storage system or IBM SAN Volume Controller, IBM Storage Insights is available at no extra cost. With this version, you can monitor the basic health, status, and performance of various storage resources.

IBM Storage Insights Pro is a subscription-based product that provides a more comprehensive view of the performance, capacity, and health of your storage resources. In addition to the features that are offered by IBM Storage Insights, IBM Storage Insights Pro provides tools for intelligent capacity planning, storage reclamation, storage tiering, and performance troubleshooting services. Together, these features can help you reduce storage costs and optimize your data center.

Note: With some models of Storage Virtualize systems that offer the Premium Storage Expert Care level of support, Storage Insights Pro is included as part of the offering. For more information, see 1.6.1, “Storage Expert Care” on page 50.

IBM Storage Insights is a part of the monitoring and helps to ensure continued availability of the IBM FlashSystem storage or IBM SAN Volume Controller systems.

The tool provides a single dashboard that gives you a clear view of all your IBM block and file storage and some other storage vendors (the IBM Storage Insights Pro version is required to view other storage vendors’ storage). You can make better decisions by seeing trends in performance and capacity. With storage health information, you can focus on areas that need attention.

When IBM Support is needed, IBM Storage Insights simplifies uploading logs, speeds resolution with online configuration data, and provides an overview of open tickets, all in one place.

The following features are available with IBM Storage Insights:

- ▶ A unified view of IBM systems:
 - Provides a single view to see all your system’s characteristics
 - Displays all of your IBM storage inventory
 - Provides a live event feed so that you know in real time what is going on with your storage so that you can act quickly
- ▶ IBM Storage Insights collects telemetry data and Call Home data and provides real-time system reporting of capacity and performance.
- ▶ Overall storage monitoring by reviewing the following information:
 - The overall health of the system.
 - Monitoring of the configuration to see whether it meets best practices.
 - System resource management determines which system is overtaxed, and provides proactive recommendations to fix it.
- ▶ IBM Storage Insights provides advanced customer service with an event filter that you can use to accomplish the following tasks:
 - You and IBM Support can view, open, and close support tickets, and track trends.

- You can use the autolog collection capability to collect the logs and send them to IBM before IBM Support investigates the problem. This capability can save time in resolving a case.

Figure 1-54 shows a view of the IBM Storage Insights dashboard.

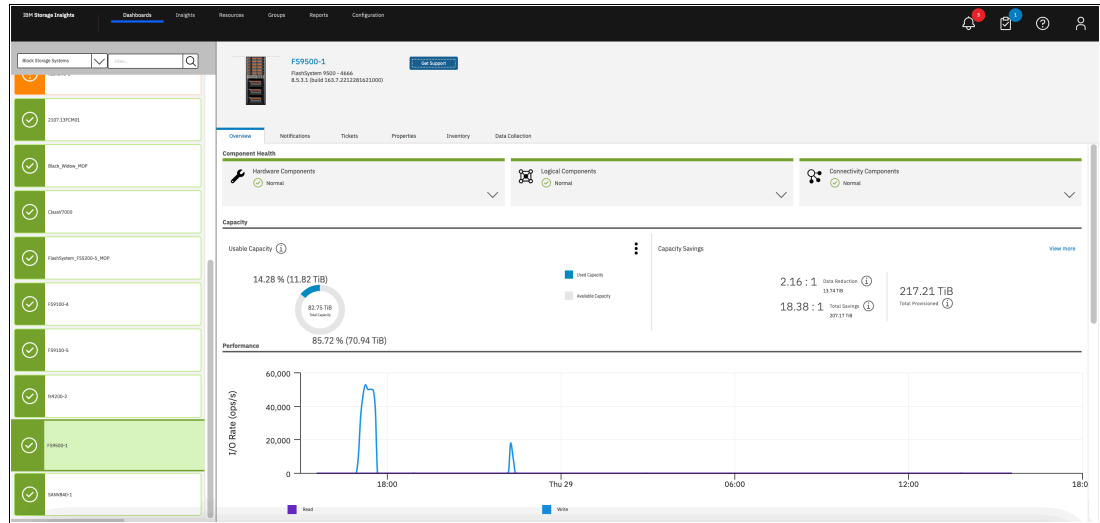


Figure 1-54 IBM Storage Insights dashboard

For IBM Storage Insights to operate, a lightweight data collector must be deployed in your data center to stream only system metadata to your IBM Cloud instance. The metadata flows in one direction: from your data center to IBM Cloud over HTTPS.

The application data that is stored on the storage systems cannot be accessed by the data collector. In IBM Cloud, your metadata is AES256-encrypted and protected by physical, organizational, access, and security controls.

IBM Storage Insights is ISO/IEC 27001 Information Security Management certified.

For more information about IBM Storage Insights, see the following websites:

- ▶ [IBM Storage Insights Fact Sheet](#)
- ▶ [Functional demonstration environment](#) (requires an IBMid)
- ▶ [IBM Storage Insights security information](#)
- ▶ [IBM Storage Insights registration](#)

IBM Storage Virtualize RESTful API

The IBM Storage Virtualize Representational State Transfer (REST) model API consists of command targets that are used to retrieve system information and to create, modify, and delete system resources. These command targets allow command parameters to pass through unedited to the IBM Storage Virtualize CLI, which handles parsing parameter specifications for validity and error reporting. It also uses HTTPS to successfully communicate with the RESTful API server.

The RESTful apiserver does not consider transport security (such as Secure Sockets Layer (SSL)), but instead assumes that requests are started from a local, secured server. The HTTPS protocol provides privacy through data encryption. The RESTful API provides more security by requiring command authentication, which persists for 2 hours of activity or 30 minutes of inactivity, whichever occurs first.

Uniform Resource Locators (URLs) target different node objects on the system. The **HTTPS POST** method acts on command targets that are specified in the URL. To make changes or view information about different objects on the system, you must create and send a request to the system. You must provide specific elements for the RESTful API server to receive and transform the request into a command.

To interact with the system by using the RESTful API, make an HTTPS command request with a valid configuration node URL destination. Open TCP port 7443 and include the keyword **rest**. Then, use the following URL format for all requests:

```
https://system_node_ip:7443/rest/command
```

Note the following values:

- ▶ *system_node_ip* is the system IP address, which is the address that is taken by the configuration node of the system.
- ▶ The port number is always 7443 for the IBM Storage Virtualize RESTful API.
- ▶ **rest** is a keyword.
- ▶ *command* is the target command object (such as `auth` or `lseventlog` with any parameters). The command specification uses the following format:

```
command_name,method="POST",headers={'parameter_name': 'parameter_value',  
'parameter_name': 'parameter_value',...}
```

1.16 Copy services

IBM Storage Systems running IBM Storage Virtualize provide copy services functions that can be used to improve availability and support DR.

Volume mirroring

By using volume mirroring, a volume can have two physical copies in one IBM Storage System. Each volume copy can belong to a different pool and use a different set of capacity saving features.

When a host writes to a mirrored volume, the system writes the data to both copies. When a host reads a mirrored volume, the system picks one of the copies to read. If one of the mirrored volume copies is temporarily unavailable, the volume remains accessible to servers. The system remembers which areas of the volume are written, and resynchronizes these areas when both copies are available.

You can create a volume with one or two copies, and you can convert a non-mirrored volume into a mirrored volume by adding a copy. When a copy is added in this way, the system synchronizes the new copy so that the new copy is the same as the existing volume. Servers can access the volume during this synchronization process.

Volume mirroring can be used to migrate data to or from an IBM Storage Systems running IBM Storage Virtualize. For example, you can start with a non mirrored image mode volume in the migration pool and then, add a copy to that volume in the destination pool on internal storage. After the volume is synchronized, you can delete the original copy that is in the source pool. During the synchronization process, the volume remains available.

Volume mirroring is also used to convert fully allocated volumes to use data reduction technologies, such as thin-provisioning, compression, or deduplication, or to migrate volumes between storage pools.

FlashCopy

The FlashCopy or snapshot function creates a point-in-time (PiT) copy of data that is stored on a source volume to a target volume. FlashCopy is sometimes described as an instance of a time-zero (T0) copy. Although the copy operation takes some time to complete, the resulting data on the target volume is presented so that the copy appears to occur immediately, and all data is available immediately. Advanced functions of FlashCopy allow operations to occur on multiple source and target volumes.

Management operations are coordinated to provide a common, single PiT for copying target volumes from their respective source volumes to create a consistent copy of data that spans multiple volumes.

The function also supports multiple target volumes to be copied from each source volume, which can be used to create images from different PiTs for each source volume.

FlashCopy is used to create consistent backups of dynamic data and test applications, and to create copies for auditing purposes and for data mining. It can be used to capture the data at a specific time to create consistent backups of dynamic data. The resulting image of the data can be backed up; for example, to a tape device object storage or another disk/flash based storage technology. When the copied data is on tape, the data on the FlashCopy target disks becomes redundant and can be discarded.

Another possible FlashCopy application is creating test environments. FlashCopy can be used to test an application with real business data before the production version of the application is updated or replaced. With FlashCopy, a fully functional and space-efficient clone of a volume that contains real data can be created. It enables read and write access for the test environment while keeping the real production environment data safe and untouched. After testing is complete, the clone volume can be discarded or retained for future use.

FlashCopy can perform a restore from any FlashCopy mapping. Therefore, you can restore (or copy) from the target to the source of your regular FlashCopy relationships. When restoring data from FlashCopy, this method can be qualified as reversing the direction of the FlashCopy mappings. This approach can be used for various applications, such as recovering a production database application after an errant batch process that caused extensive damage.

Remote mirroring

You can use remote mirroring (also referred as Remote Copy [RC]) function to set up a relationship between two volumes, where updates made to one volume are mirrored on the other volume. The volumes can be on two different systems (intersystem) or on the same system (intrasystem).

For an RC relationship, one volume is designated as the primary and the other volume is designated as the secondary. Host applications write data to the primary volume, and updates to the primary volume are copied to the secondary volume. Normally, host applications do not run I/O operations to the secondary volume.

The following types of remote mirroring are available:

- ▶ **Metro Mirror (MM)**

Provides a consistent copy of a source volume on a target volume. Data is written to the target volume synchronously after it is written to the source volume so that the copy is continuously updated.

With synchronous copies, host applications write to the primary volume but do not receive a confirmation that the write operation completed until the data is written to the secondary

volume, which ensures that both volumes have identical data when the copy operation completes. After the initial copy operation completes, the MM function always maintains a fully synchronized copy of the source data at the target site. The MM function supports copy operations between volumes that are separated by distances up to 300 km (186.4 miles).

For DR purposes, MM provides the simplest way to maintain an identical copy on the primary and secondary volumes. However, as with all synchronous copies over remote distances, host application performance can be affected. This performance effect is related to the distance between primary and secondary volumes and depending on application requirements, its use might be limited based on the distance between sites.

► Global Mirror (GM)

Provides a consistent copy of a source volume on a target volume. The data is written to the target volume asynchronously and the copy is continuously updated. When a host writes to the primary volume, a confirmation of I/O completion is received before the write operation completes for the copy on the secondary volume. Because of this situation, the copy might not contain the most recent updates when a DR operation is completed.

If a failover operation is started, the application must recover and apply any updates that were not committed to the secondary volume. If I/O operations on the primary volume are paused for a short period, the secondary volume can become a match of the primary volume. This function is comparable to a continuous backup process in which the last few updates are always missing. When you use GM for DR, you must consider how you want to handle these missing updates.

The secondary volume is generally less than one second behind the primary volume, which minimizes the amount of data that must be recovered if a failover occurs. However, a high-bandwidth link must be provisioned between the two sites.

► Global Mirror with Change Volumes (GMCV)

Enables support for GM with a higher recovery point objective (RPO) by using change volumes. This function is for use in environments where the available bandwidth between the sites is smaller than the update rate of the replicated workload.

With GMCV, or GM with cycling, change volumes must be configured for the primary and secondary volumes in each relationship. A copy is taken of the primary volume in the relationship to the change volume. The background copy process reads data from the stable and consistent change volume and copies the data to the secondary volume in the relationship.

CoW technology is used to maintain the consistent image of the primary volume for the background copy process to read. The changes that occurred while the background copy process was active also are tracked. The change volume for the secondary volume also can be used to maintain a consistent image of the secondary volume while the background copy process is active.

GMCV provides fewer requirements to inter-site link bandwidth than other RC types. It is mostly used when link parameters are insufficient to maintain the RC relationship without affecting host performance.

Intersystem replication is possible over an FC or IP link. The native IP replication feature enables replication between any family systems running IBM Storage Virtualize by using the built-in networking ports of the system nodes.

Note: Although all three types of RC are supported to work over an IP link, the recommended type is GMCV.

1.16.1 Policy-based replication

Policy-based replication (PBR) uses volume groups and replication policies to automatically deploy and manage replication. Policy-based replication significantly simplifies configuring, managing, and monitoring replication between two systems.

Note: IBM FlashSystems 5015, 5035 and 5045 do not support policy-based replication.

With PBR, you can replicate data between systems with minimal management, significantly higher throughput and reduced latency compared to the remote-copy function. A replication policy has following properties:

- ▶ A replication policy can be assigned to one or more volume groups.
- ▶ Replication policies cannot be changed after they are created. If changes are required, a new policy can be created and assigned to the associated volume group.
- ▶ Each system supports up to a maximum of 32 replication policies.
 - Replication policies
 - Replication policies define the replication settings that are assigned to the volume groups. Replication policies replicate the volume groups and ensure that consistent data is available on the production and recovery system.
 - Volume groups for PBR
 - PBR uses volume groups and replication policies to automatically deploy and manage replication. PBR significantly simplifies configuring, managing, and monitoring replication between two systems.
 - Partnerships
 - Two-site partnerships replicate volume data that is on one system to a remote system. Two-site partnerships are required for policy-based replication. Partnerships can be used for migration, 3-site replication, and DR situations.

For more information, see [Getting Started with policy-based-replication](#).

1.16.2 HyperSwap

The IBM HyperSwap function is an HA feature that provides dual-site, active-active access to a volume. It is available on systems that can support more than one I/O group.

With HyperSwap, a fully independent copy of the data is maintained at each site. When data is written by hosts at either site, both copies are synchronously updated before the write operation is completed. The HyperSwap function automatically optimizes to minimize data that is transmitted between two sites, and to minimize host read and write latency.

If the system or the storage at either site goes offline and an online and accessible up-to-date copy is left, the HyperSwap function can automatically fail over access to the online copy. The HyperSwap function also automatically resynchronizes the two copies when possible.

To construct HyperSwap volumes, active-active replication relationships are made between the copies at each site. These relationships automatically run and switch direction according to which copy or copies are online and up to date.

The relationships provide access to whichever copy is up to date through a single volume, which has a unique ID. This volume is visible as a single object across both sites (I/O groups), and is mounted to a host system.

A 2-site HyperSwap configuration can be extended to a third site for DR that uses the IBM Storage Virtualize 3-Site Orchestrator.

IBM Storage Virtualize 3-Site Orchestrator coordinates replication of data for disaster recovery and high availability scenarios between systems that are on three geographically dispersed sites. IBM Storage Virtualize 3-Site Orchestrator is a command-line based application that runs on a separate Linux host that configures and manages supported replication configurations on IBM Storage Virtualize products.

To extend a HyperSwap system or to create a HyperSwap configuration with IBM Storage Virtualize 3-Site Orchestrator requires more planning, installing, and configuring steps.

The HyperSwap function works with the standard multipathing drivers that are available on various host types, with no extra host support that is required to access the highly available volume. Where multipathing drivers support Asymmetric Logical Unit Access (ALUA), the storage system tells the multipathing driver which nodes are closest to it and must be used to minimize I/O latency. You tell the storage system which site a host is connected to, and it configures host pathing optimally.

1.17 IBM FlashCore Module drives, NVMe SSDs, and SCM drives

This section describes the three types of flash drives that can be installed in the control enclosures:

- ▶ FCM drives
- ▶ NVMe SSDs
- ▶ Storage-class memory (SCM) drives

Note: The SCM drives and XL FCM drives require IBM Storage Virtualize V8.3.1 or later to be installed on the IBM FlashSystem control enclosure.

The following IBM FlashSystem products can support all three versions of these drives:

- ▶ 9500
- ▶ 9500R Rack Solution
- ▶ 9200
- ▶ 9200R Rack Solution
- ▶ 7300
- ▶ 7200
- ▶ 5200
- ▶ 5100

They are not supported in any of the expansion enclosures.

Figure 1-55 shows an FCM (NVMe) with a capacity of 19.2 TB. The first generation of FCM drives were built by using 64-layer Triple Level Cell (TLC) flash memory and an Everspin MRAM cache into a U.2 form factor. FCM generations 2 and 3 are built using 96-layer Quad Level Cell (QLC) flash memory. These later generations also reformat some of the flash capacity to a pseudo-SLC mode (pSLC) to improve performance, reduce latency, and provide a dynamic read cache on the device.



Figure 1-55 IBM FlashCore Module (NVMe)

FCM drives are designed for high parallelism and optimized for 3D QLC and updated FPGAs. IBM also enhanced the FCM drives by adding read cache to reduce latency on highly compressed pages. Also added was four-plane programming to lower the overall power during writes. FCM drives offer hardware-assisted compression up to 3:1 and are FIPS 140-2 compliant.

FCM drives carry IBM Variable Stripe RAID (VSR) at the FCM level and use DRAID to protect data at the system level. VSR and DRAID together optimize RAID rebuilds by off-loading rebuilds to DRAID, and they offer protection against FCM failures.

Table 1-18 lists the capacities of the FCM type drives.

Table 1-18 FCM type capacities

FCM module size	Physical size (TBu)
Small	4.8 TBu
Medium	9.6 TBu
Large	19.2 TBu
XLarge	38.4 TBu

Industry-standard SSD NVMe drives

All the IBM FlashSystem models that are described in this book provide an option to use industry-standard SSD NVMe drives, which are sourced from Samsung and Toshiba and available in the several capacity variations, as listed in Table 1-19.

Table 1-19 NVMe drive size options

Drive type	Physical size (TBu)
NVMe Flash Drive	1.92 TB
NVMe Flash Drive	3.84 TB
NVMe Flash Drive	7.68 TB
NVMe Flash Drive	15.36 TB
NVMe Flash Drive	30.72 TB

NVMe and adapter support

NVMe is a NUMA-optimized, high-performance, and highly scalable storage protocol that is designed to access nonvolatile storage media by using a host PCIe bus. NVMe uses low-latency and available parallelism, and reduces I/O impact.

NVMe supports multiple I/O queues up to 64 K queues, and each queue can support up to 64 K entries. Earlier generations of SAS and Serial Advanced Technology Attachment (SATA) support a single queue with only 254 and 32 entries and use many more CPU cycles to access data. NVMe handles more workload for the same infrastructure footprint.

NVMe-oF is a technology specification that is designed to enable NVMe message-based commands to transfer data between a host computer and a target SSD or system. Data is transferred over a network, such as Ethernet, FC, or InfiniBand.

Storage-class memory

SCM drives use persistent memory technologies that improve endurance and reduce the latency of flash storage device technologies. All SCM drives use the NVMe architecture. IBM Research® is actively engaged in researching these new technologies.

For more information, see the [SNIA Educational Library](#).

These technologies fundamentally change the architecture of today's storage infrastructures. Figure 1-56 shows the different types of storage technologies versus the latency for Intel drives.

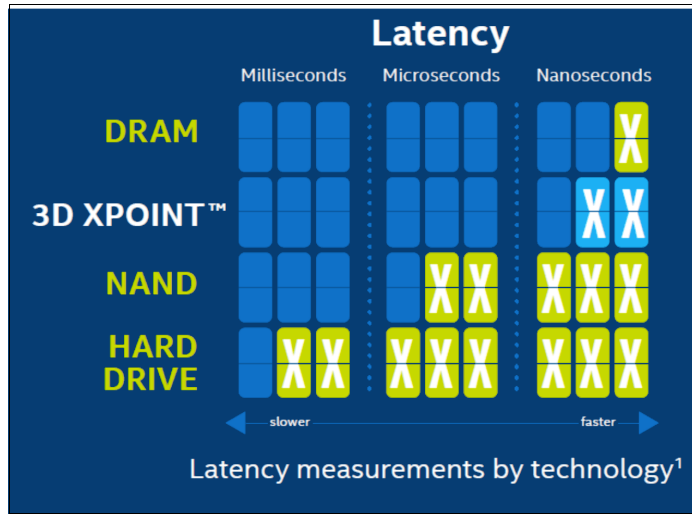


Figure 1-56 Storage technologies versus latency for Intel drives

IBM supports SCM class drives. Table 1-20 lists the SCM drive size options.

Table 1-20 SCM drive options

Drive type	Physical size (TBu)	Supported in
NVMe SCM Drive	375 GB	FS5200, FS7200 *
NVMe SCM Drive	750 GB	FS5200, FS7200 *
NVMe SCM Drive	800 GB	FS5200, FS7200 *
NVMe SCM Drive	1.6 TB	FS5200, FS7200, FS7300, FS9500/R

* These drives are not sold anymore, but they are still supported.

Easy Tier supports the SCM drives with a new tier that is called `tier_scm`.

Note: The SCM drive type supports only DRAID 6, DRAID 5, DRAID 1, and TRAIT 1 or 10.

1.18 Storage virtualization

Storage virtualization is a term that is used extensively throughout the storage industry. It can be applied to various technologies and underlying capabilities. In reality, most storage devices technically can claim to be virtualized in one form or another. Therefore, this section starts by defining the concept of storage virtualization as it is used in this book.

We describe storage virtualization in the following ways:

- ▶ Storage virtualization is a technology that makes one set of resources resemble another set of resources (preferably with more wanted characteristics).
- ▶ Storage virtualization is a logical representation of resources that is not constrained by physical limitations and hides part of the complexity. It also adds or integrates new functions with services, and can be nested or applied to multiple layers of a system.

The virtualization model consists of the following layers:

- ▶ Application: The user of the storage domain.
- ▶ Storage domain:
 - File, record, and namespace virtualization, and file and record subsystem
 - Block virtualization
 - Block subsystem

Applications typically read and write data as vectors of bytes or records. However, storage presents data as vectors of blocks of a constant size (512, or in the newer devices, 4096 bytes per block).

The file, record, and namespace virtualization and file and record subsystem layers convert records or files that are required by applications to vectors of blocks, which are the language of the block virtualization layer. The block virtualization layer maps requests of the higher layers to physical storage blocks, which are provided by storage devices in the block subsystem.

Each of the layers in the storage domain abstracts away complexities of the lower layers and hides them behind an easy to use, standard interface that is presented to upper layers. The resultant decoupling of logical storage space representation and its characteristics that are visible to servers (storage consumers) from underlying complexities and intricacies of storage devices is a key concept of storage virtualization.

The focus of this publication is block-level virtualization at the block virtualization layer, which is implemented by IBM as IBM Storage Virtualize software that is running on an IBM SAN Volume Controller and the IBM FlashSystem family. The IBM SAN Volume Controller is implemented as a clustered appliance in the storage network layer. The IBM FlashSystem storage systems are deployed as modular systems that can virtualize their internally and externally attached storage.

IBM Storage Virtualize uses the SCSI protocol to communicate with its clients and presents storage space as SCSI logical units (LUs), which are identified by SCSI LUNs.

Note: Although LUs and LUNs are different entities, the term *LUN* in practice is often used to refer to a logical disk, that is, an LU.

Although most applications do not directly access storage but work with files or records, the operating system of a host must convert these abstractions to the language of storage; that is, vectors of storage blocks that are identified by LBAs within an LU.

Inside IBM Storage Virtualize, each of the externally visible LUs is internally represented by a volume, which is an amount of storage that is taken out of a storage pool. Storage pools are made of MDisks; that is, they are LUs that are presented to the storage system by external virtualized storage or arrays that consist of internal disks. LUs that are presented to IBM Storage Virtualize by external storage usually correspond to RAID arrays that are configured on that storage.

With storage virtualization, you can manage the mapping between logical blocks within an LU that is presented to a host and blocks on physical drives. This mapping can be as simple or as complicated as required. A logical block can be mapped to one physical block, or for increased availability, multiple blocks that are physically stored on different physical storage systems, and in different geographical locations.

Importantly, the mapping can be dynamic: With Easy Tier, IBM Storage Virtualize can automatically change underlying storage to which groups of blocks (extent) are mapped to better match a host's performance requirements with the capabilities of the underlying storage systems.

IBM Storage Virtualize gives a storage administrator various options to modify volume characteristics, from volume resize to mirroring, creating a PiT copy with FlashCopy and migrating data across physical storage systems.

Importantly, all the functions that are presented to the storage users are independent from the characteristics of the physical devices that are used to store data. This decoupling of the storage feature set from the underlying hardware and ability to present a single, uniform interface to storage users that masks underlying system complexity is a powerful argument for adopting storage virtualization with IBM Storage Virtualize.

IBM Storage Virtualize includes the following key features:

- ▶ Simplified storage management by providing a single management interface for multiple storage systems, and a consistent user interface for provisioning heterogeneous storage.
- ▶ IBM Storage Virtualize enables online volume migration (moving the data from one set of physical drives to another set) in a way that is not apparent to the storage users and without over-straining the storage infrastructure. The migration can be done within a specific storage system (from one set of disks to another set) or across storage systems. Either way, the host that uses the storage is not aware of the operation, and no downtime for applications is needed.
- ▶ Performing copy services functions within IBM Storage Virtualize removes dependencies on the capabilities and interoperability of the virtualized storage subsystems. Therefore, it enables the source and target copies to be on any two virtualized storage subsystems.
- ▶ Improved storage space usage because of the pooling of resources across virtualized storage systems.
- ▶ Opportunity to improve system performance as a result of volume striping across multiple virtualized arrays or controllers, and the benefits of cache that is provided by IBM Storage Virtualize hardware.
- ▶ Improved data security by using data-at-rest encryption.
- ▶ Data replication, including replication to cloud storage by using advanced copy services for data migration and backup solutions.
- ▶ Today, open systems typically use less than 50% of the provisioned storage capacity. IBM Storage Virtualize can enable savings, increase the effective capacity of storage systems up to five times, and decrease the floor space, power, and cooling that are required by the storage system.

IBM FlashSystem and IBM SAN Volume Controller families are scalable solutions running on a HA platform that can use diverse back-end storage systems to provide all the benefits to various attached hosts.

External storage virtualization

You can use IBM Storage Systems running IBM Storage Virtualize to manage the capacity of other storage systems with external storage virtualization. When IBM Storage Virtualize virtualizes a storage system, its capacity is managed similarly to internal disk drives or flash modules. Capacity in external storage systems inherits all of the rich functions and ease of use.

You can use IBM Storage Systems running IBM Storage Virtualize to preserve your investments in storage, centralize management, and make storage migrations easier with storage virtualization and Easy Tier. Virtualization helps insulate applications from changes that are made to the physical storage infrastructure.

To verify whether your storage can be virtualized by IBM FlashSystem or IBM SAN Volume Controller, see the [IBM System Storage Interoperation Center \(SSIC\)](#).

All the IBM Storage Systems that are running IBM Storage Virtualize can migrate data from external storage controllers, including migrating from any other IBM or third-party storage systems. IBM Storage Virtualize uses the functions that are provided by its external virtualization capability to perform the migration. This capability places external LUs under the control of an IBM FlashSystem or IBM SAN Volume Controller system. Then, hosts continue to access them through the IBM FlashSystem system or IBM SAN Volume Controller system, which acts as a proxy.

The migration process typically consists of the following steps:

1. Input/output (I/O) to the LUs that are on the external storage system must be stopped, and the mapping of the storage system must be changed so that the original LUs are presented directly to the IBM FlashSystem or IBM SAN Volume Controller Family machine and not to the hosts. IBM FlashSystem or IBM SAN Volume Controller discovers the external LUs and recognizes them as *unmanaged* external storage back-end devices (MDisks).
2. The unmanaged MDisks are imported to the IBM FlashSystem or IBM SAN Volume Controller image mode volumes and placed in a migration storage pool. This storage pool is now a logical container for the externally attached LUs. Each volume has a one-to-one mapping with an external LU. From a data perspective, the image mode volume represents the SAN-attached LUs exactly as they were before the import operation. The image mode volumes are on the same physical drives of the storage system, and the data remains unchanged.
3. Your hosts are configured for an IBM FlashSystem or IBM SAN Volume Controller attachment, and image-mode volumes are mapped to them. After the volumes are mapped, the hosts discover their volumes and are ready to continue working with them so that I/O can be resumed.
4. Image-mode volumes are migrated to the internal storage of IBM FlashSystem by using the volume mirroring feature. Mirrored copies are created online so that a host can still access and use the volumes during the mirror synchronization process.
5. After the mirror operations are complete, the image mode volumes are removed (deleted), and external storage system can be disconnected and decommissioned or reused elsewhere.

The GUI of the IBM Storage Systems that are running IBM Storage Virtualize provides a storage migration wizard, which simplifies the migration task. The wizard features intuitive steps that guide users through the entire process.

Note: The IBM FlashSystem 5015, 5035 and 5045 systems do not support external virtualization for any other purpose other than data migration.

Summary

Storage virtualization is a fundamental technology that enables the realization of flexible and reliable storage solutions. It helps enterprises to better align their IT architecture with business requirements, simplify their storage administration, and facilitate their IT departments efforts to meet business demands.

IBM Storage Virtualize that is running on the IBM FlashSystem family is a mature, 11th-generation virtualization solution that uses open standards and complies with the SNIA storage model. All the products are appliance-based storage, and use in-band block virtualization engines that move the control logic (including advanced storage functions) from many individual storage devices to a centralized entity in the storage network.

IBM Storage Virtualize can improve the use of your storage resources, simplify storage management, and improve the availability of business applications.

1.19 Business continuity

In today's online, highly connected, and fast-paced world, we expect that today's IT systems provide HA and continuous operations, and that they can be quickly recovered if a disaster occurs. Yet, today's IT environment also features an ever-growing time to market pressure, with more projects to complete, more IT problems to solve, and a steep rise in time and resource limitations.

Thankfully, today's IT technology also features unprecedented levels of functions, features, and lowered cost. In many ways, it is easier than ever before to find IT technology that can address today's business concerns. This section describes some IBM FlashSystem storage solutions that can be applied to today's business continuity requirements.

1.19.1 Business continuity with the IBM SAN Volume Controller

In simple terms, a *clustered system* or *system* is a collection of servers that together provide a set of resources to a client. The key point is that the client has no knowledge of the underlying physical hardware of the system. The client is isolated and protected from changes to the physical hardware. This arrangement offers many benefits including, most significantly, HA.

Resources on the clustered system act as HA versions of unclustered resources. If a node (an individual computer) in the system is unavailable or too busy to respond to a request for a resource, the request is passed transparently to another node that can process the request. The clients are "unaware" of the exact locations of the resources that they use.

The IBM SAN Volume Controller is a collection of up to eight nodes, which are added in pairs that are known as *I/O groups*. These nodes are managed as a set (system), and they present a single point of control to the administrator for configuration and service activity.

The eight-node limit for an IBM SAN Volume Controller system is a limitation that is imposed by the Licensed Internal Code, and not a limit of the underlying architecture. Larger system configurations might be available in the future.

Although the IBM SAN Volume Controller code is based on a purpose-optimized Linux kernel, the clustered system feature is not based on Linux clustering code. The clustered system software within the IBM SAN Volume Controller (that is, the event manager cluster framework) is based on the outcome of the COMPASS research project. It is the key element that isolates the IBM SAN Volume Controller application from the underlying hardware nodes.

The clustered system software makes the code portable. It provides the means to keep the single instances of the IBM SAN Volume Controller code that are running on separate systems' nodes in sync. Therefore, restarting nodes during a code upgrade, adding nodes, removing nodes from a system, or failing nodes cannot affect IBM SAN Volume Controller availability.

All active nodes of a system must know that they are members of the system. This knowledge is especially important in situations where it is key to have a solid mechanism to decide which nodes form the active system, such as the split-brain scenario where single nodes lose contact with other nodes. A worst case scenario is a system that splits into two separate systems.

Within an IBM SAN Volume Controller system, the *voting set* and a quorum disk are responsible for the integrity of the system. If nodes are added to a system, they are added to the voting set. If nodes are removed, they are removed quickly from the voting set. Over time, the voting set and the nodes in the system can change so that the system migrates onto a separate set of nodes from the set on which it started.

The IBM SAN Volume Controller clustered system implements a dynamic quorum. Following a loss of nodes, if the system can continue to operate, it adjusts the quorum requirement so that further node failure can be tolerated.

The lowest Node Unique ID in a system becomes the boss node for the group of nodes. It determines (from the quorum rules) whether the nodes can operate as the system. This node also presents the maximum two-cluster IP addresses on one or both of its nodes' Ethernet ports to enable access for system management.

1.19.2 Business continuity with stretched clusters

Within standard implementations of the IBM SAN Volume Controller, all the I/O group nodes are physically installed in the same location. To supply the different HA needs that customers have, the *stretched system configuration* was introduced. In this configuration, each node (from the same I/O group) on the system is physically on a different site. When implemented with mirroring technologies, such as volume mirroring or copy services, these configurations can be used to maintain access to data on the system if power failures or site-wide outages occur.

Stretched clusters are considered HA solutions because both sites work as instances of the production environment (no standby location exists). Combined with application and infrastructure layers of redundancy, stretched clusters can provide enough protection for data that requires availability and resiliency.

In a stretched cluster configuration, nodes within an I/O group can be separated by a distance of up to 10 km (6.2 miles) by using specific configurations. You can use FC inter-switch links (ISLs) in paths between nodes of the same I/O group. In this case, nodes can be separated by a distance of up to 300 km (186.4 miles); however, potential performance impacts can result.

1.19.3 Business continuity with enhanced stretched cluster

Enhanced stretched cluster (ESC) further improves stretched cluster configurations with the *site awareness* concept for nodes, hosts, and external storage systems. It also provides a feature that enables you to manage effectively rolling disaster scenarios.

The site awareness concept enables more efficiency for host I/O traffic through the SAN, and an easier host path management.

The use of an IP-based quorum application as the quorum device for the third site does not require FC connectivity. Java applications run on hosts at the third site.

Note: Stretched cluster and ESC features are supported for IBM SAN Volume Controller only. They are not supported for the IBM FlashSystem family of products.

For more information and implementation guidelines about deploying stretched cluster or ESC, see *IBM Spectrum Virtualize and SAN Volume Controller Enhanced Stretched Cluster with VMware*, SG24-8211.

1.19.4 Business continuity for FlashSystem and IBM SAN Volume Controller

In this section, we discuss business continuity for FlashSystem and IBM SAN Volume Controller.

1.19.5 Business continuity with HyperSwap

The HyperSwap HA feature in the IBM Storage Virtualize software enables business continuity during hardware, power, or connectivity failures, or disasters, such as fire or flooding. The HyperSwap feature is available on the IBM SAN Volume Controller and IBM FlashSystem products that are running IBM Storage Virtualize software.

The HyperSwap feature provides HA volumes that are accessible through two sites at up to 300 km (186.4 miles) apart. A fully independent copy of the data is maintained at each site.

When data is written by hosts at either site, both copies are synchronously updated before the write operation is completed. The HyperSwap feature automatically optimizes to minimize data that is transmitted between sites and to minimize host read and write latency.

HyperSwap includes the following key features:

- ▶ Works with IBM SAN Volume Controller and IBM FlashSystem products that are running IBM Storage Virtualize software.
- ▶ Uses intra-cluster synchronous RC (MM) capabilities along with change volume and access I/O group technologies.
- ▶ Makes a host's volumes accessible across two I/O groups in a clustered system by using the MM relationship in the background. They look like a single volume to the host.

- ▶ Works with the standard multipathing drivers that are available on various host types, with no other host support that is required to access the HA volume.

For more information about HyperSwap implementation use cases and guidelines, see the following publications:

- ▶ *IBM Storwize V7000, Spectrum Virtualize, HyperSwap, and VMware Implementation*, SG24-8317
- ▶ *High Availability for Oracle Database with IBM PowerHA SystemMirror and IBM Spectrum Virtualize HyperSwap*, REDP-5459
- ▶ *IBM Spectrum Virtualize HyperSwap SAN Implementation and Design Best Practices*, REDP-5597

1.19.6 Business continuity with 3-site replication

A three-site replication solution was made available in limited deployments for 8.3.1, where data is replicated from the primary site to two alternative sites, and the remaining two sites are aware of the difference between themselves. This solution ensures that if a disaster occurs at any one of the sites, the remaining two sites can establish a consistent_synchronized RC relationship among themselves with minimal data transfer; that is, within the expected RPO.

IBM Storage Virtualize 8.4 and above expands the three-site replication model to include HyperSwap, which improves data availability options in three-site implementations. Systems that are configured in a three-site topology have high DR capabilities, but a disaster might take the data offline until the system can be failed over to an alternative site.

HyperSwap allows active-active configurations to maintain data availability, which eliminates the need to failover if communications are disrupted. This solution provides a more robust environment, which allows up to 100% uptime for data, and recovery options that are inherent to DR solutions.

To better assist with 3-site replication solutions, IBM Storage Virtualize 3-Site Orchestrator coordinates replication of data for DR and HA scenarios between systems.

IBM Storage Virtualize 3-Site Orchestrator is a command-line based application that runs on a separate Linux host that configures and manages supported replication configurations on IBM Storage Virtualize products.

Figure 1-57 shows the two supported topologies for the 3-site, replication-coordinated solutions.

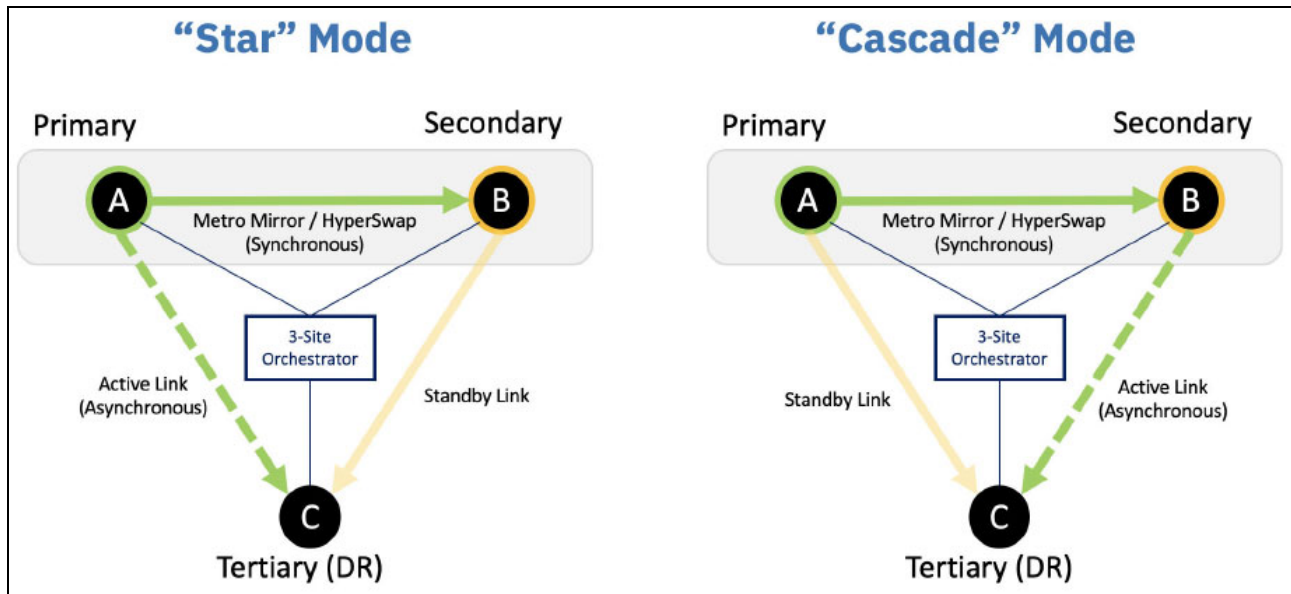


Figure 1-57 “Star” and “Cascade” modes in a three-site solution

For more information about this type of implementation, see *Spectrum Virtualize 3-Site Replication*, SG24-8474.

1.19.7 Automatic hot spare nodes (IBM SAN Volume Controller only)

In previous stages of IBM SAN Volume Controller development, the scripted *warm standby* procedure enables administrators to configure spare nodes in a cluster by using the concurrent hardware upgrade capability of transferring WWPNs between nodes. The system can automatically take on the spare node to replace a failed node in a cluster or to keep the entire system under maintenance tasks, such as software upgrades. These extra nodes are called *hot spare nodes*.

Up to four nodes can be added to a single cluster. When the hot-spare node is used to replace a node, the system attempts to find a spare node that matches the configuration of the replaced node perfectly.

However, if a perfect match does not exist, the system continues the configuration check until a matching criteria is found. The following criteria is used by the system to determine suitable hot-spare nodes:

- ▶ Requires an exact match:
 - Memory capacity
 - Fibre Channel port ID
 - Compression support
 - Site
- ▶ Recommended to match, but can be different:
 - Hardware type
 - CPU count
 - Number of Fibre Channel ports

If the criteria are not the same for both, the system uses lower criteria until the minimal configuration is found. For example, if the Fibre Channel ports do not match exactly but all the other required criteria match, the hot-spare node can still be used. The minimal configuration that the system can use as a hot-spare node includes identical memory, site, Fibre Channel port ID, and, if applicable, compression settings.

If the nodes on the system support and are licensed to use encryption, the hot-spare node must also support and be licensed to use encryption.

The hot spare node essentially becomes another node in the cluster, but is not doing anything under normal conditions. Only when it is needed does it use the N_Port ID Virtualization (NPIV) feature of the Storage Virtualize virtualized storage ports to take over the job of the failed node. It performs this takeover by moving the NPIV WWPNs from the failed node first to the surviving partner node in the I/O group and then, over to the hot spare node.

Approximately 1 minute passes intentionally before a cluster swaps in a node to avoid any thrashing when a node fails. In addition, the system must be sure that the node definitely failed, and is not (for example) restarting. The cache flushes while only one node is in the I/O group, the full cache is returned when the spare swaps in.

This entire process is transparent to the applications; however, the host systems notice a momentary path lost for each transition. The persistence of the NPIV WWPNs lessens the multipathing effort on the host considerably during path recovery.

Note: A warm start of active node (code assert or restart) does not cause the hot spare to swap in because the restarted node becomes available within 1 minute.

The other use case for hot spare nodes is during a software upgrade. Normally, the only impact during an upgrade is slightly degraded performance. While the node that is upgrading is down, the partner in the I/O group writes through cache and handles both nodes' workload. Therefore, to work around this issue, the cluster uses a spare in place of the node that is upgrading. The cache does not need to go into write-through mode and the period of degraded performance from running off a single node in the I/O group is significantly reduced.

After the upgraded node returns, it is swapped back so that you roll through the nodes as normal, but without any failover and failback at the multipathing layer. This process is handled by the NPIV ports; therefore, the upgrades must be seamless for administrators who are working in large enterprise IBM SAN Volume Controller deployments.

Note: After the cluster commits new code, it also automatically upgrades hot spares to match the cluster code level.

This feature is available to IBM SAN Volume Controller only. Although IBM FlashSystem systems can use NPIV and realize the general failover benefits, no hot spare canister or split I/O group option is available for the enclosure-based systems.

1.20 Management and support tools

The IBM Storage Virtualize system can be managed by using the included management software that runs on the controller hardware.

1.20.1 IBM Assist On-site and Remote Support Assistance

IBM Assist On-site and Remote Support Assistance are two tools that you can use to manage and support an IBM Storage Virtualize system.

IBM Assist On-site

With the IBM Assist On-site tool, a member of the IBM Support team can view your desktop and share control of your server to provide you with a solution. This tool is a remote desktop-sharing solution that is offered through the IBM website. With it, the IBM System Services Representative (IBM SSR) can remotely view your system to troubleshoot a problem.

You can maintain a chat session with the IBM SSR so that you can monitor this activity and understand how to fix the problem yourself or enable them to fix it for you.

For more information, see [IBM remote assistance: Assist On-site](#).

When you access the website, you sign in and enter a code that the IBM SSR provides to you. This code is unique to each IBM Assist On-site session. A plug-in is downloaded to connect you and your IBM SSR to the remote service session. The IBM Assist On-site tool contains several layers of security to protect your applications and your computers. The plug-in is removed after the next restart.

You also can use security features to restrict access by the IBM SSR. Your IBM SSR can provide you with more information about the use of the tool.

Remote Support Assistance

The embedded part of the IBM Storage Virtualize 8.6 code is a software toolset that is called *Remote Support Client*. It establishes a network connection over a secured channel with Remote Support Server in the IBM network.

The Remote Support Server provides predictive analysis of the IBM Storage Controller running IBM Storage Virtualize software status and assists administrators with troubleshooting and fix activities. Remote Support Assistance is available at no extra charge, and no extra license is needed.

For more information, see [Configuring support assistance](#).

1.20.2 Event notifications

IBM Storage Virtualize system can use SNMP traps, syslog messages, and a Call Home email or cloud based notification, to notify you and the IBM Support Center when significant events are detected. Any combination of these notification methods can be used simultaneously.

Notifications are normally sent immediately after an event is raised. Each event that IBM Storage Controller detects is assigned a notification type of Error, Warning, or Information. You can configure the IBM Storage Controller to send each type of notification to specific recipients.

Simple Network Management Protocol traps

SNMP is a standard protocol for managing networks and exchanging messages. IBM Storage Virtualize can send SNMP messages that notify personnel about an event.

You can use an SNMP manager to view the SNMP messages that IBM Storage Virtualize sends. You can use the management GUI or the CLI to configure and modify your SNMP settings.

The MIB file for SNMP can be used to configure a network management program to receive SNMP messages that are sent by the IBM Storage Virtualize.

Syslog messages

The syslog protocol is a standard protocol for forwarding log messages from a sender to a receiver on an IP network. The IP network can be Internet Protocol Version 4 (IPv4) or Internet Protocol Version 6 (IPv6).

IBM SAN Volume Controller and IBM FlashSystems can send syslog messages that notify personnel about an event. The event messages can be sent in expanded or concise format. You can use a syslog manager to view the syslog messages that IBM SAN Volume Controller or IBM FlashSystems sends.

IBM Storage Virtualize uses UDP to transmit the syslog message. You can use the management GUI or the CLI to configure and modify your syslog settings.

Call Home notification

Call Home notification improves the response time for issues on the system. Call Home notifications send diagnostic data to IBM Support personnel who can quickly determine solutions for these problems, which can disrupt operations on the system.

Call Home email

This feature transmits operational and error-related data to you and IBM through a Simple Mail Transfer Protocol (SMTP) server connection in the form of an event notification email. You can use the Call Home function if you have a maintenance contract with IBM or if IBM SAN Volume Controller or IBM FlashSystems are within the warranty period.

To send email, you must configure at least one SMTP server. You can specify as many as five other SMTP servers for backup purposes. The SMTP server must accept the relaying of email from the IBM SAN Volume Controller clustered system IP address. Then, you can use the management GUI or the CLI to configure the email settings, including contact information and email recipients. Set the reply address to a valid email address.

Send a test email to check that all connections and infrastructure are set up correctly. You can disable the Call Home function at any time by using the management GUI or CLI.

Cloud Call Home

Call Home with cloud services sends notifications directly to a centralized file repository that contains troubleshooting information that is gathered from customers. Support personnel can access this repository and be assigned issues automatically as problem reports. This method of transmitting notifications from the system to support removes the need for customers to create problem reports manually.

Call Home with cloud services also eliminates email filters dropping notifications to and from support, which can delay resolution of problems on the system. Call Home with cloud services uses Representational State Transfer (RESTful) APIs, which are a standard for transmitting data through web services.

For new system installations, Call Home with cloud services is configured as the default method to transmit notifications to support. When you update the system software, Call Home with cloud services is also set up automatically. You must ensure that network settings are configured to allow connections to the support center. This method sends notifications to only the predefined support center.

To use Call Home with cloud services, ensure that all of the nodes on the system have internet access, and that a valid service IP is configured on each node on the system. In addition to these network requirements, you must configure suitable routing to the support center through a domain name service (DNS) or by updating your firewall configuration so it includes connections to the support center.

After a DNS server is configured, update your network firewall settings to allow outbound traffic to `esupport.ibm.com` on port 443.

If not using DNS but you have a firewall to protect your internal network from outside traffic, you must enable specific IP addresses and ports to establish a connection to the support center. Ensure that your network firewall allows outbound traffic to the following IP addresses on port 443:

- ▶ 129.42.21.70 (New)
- ▶ 129.42.56.189 **
- ▶ 129.42.60.189 **

Note: ** For the latest updates on changes to IP addresses, see [Preparing customer firewalls and proxies for the upcoming infrastructure changes – Call Home, Electronic Fix Distribution](#).

You can configure either of these methods or configure both for redundancy. DNS is the preferred method because it ensures that the system can still connect to the IBM Support center if the underlying IP addresses to the support center change.

IBM highly encourages all customers to take advantage of the Call Home feature so that you and IBM can collaborate for your success.

For more information about the features and functions of Call Home methods, see [IBM Storage Virtualize Products Call Home and Remote Support Overview](#).

1.21 Licensing

All IBM FlashSystem functional capabilities are provided through IBM Storage Virtualize software. Each platform is licensed as described in the following sections.

1.21.1 Licensing IBM SAN Volume Controller

The IBM SAN Volume Controller uses perpetual licenses. A perpetual license is the “traditional” model that is used to purchase software. You pay for your software license up-front and can use it indefinitely. For more information about the following storage capacity for external virtualization features, based on storage classification, see Table 1-21.

- ▶ FlashCopy/TiB based
- ▶ Remote Mirroring/TiB based
- ▶ Encryption/Key license

1.21.2 Licensing IBM FlashSystem 9500/R, 9200/R, 7300, 7200, 5200 and 5045

The IBM FlashSystems 9500, 9500R, 9200, 9200R, 7300, 7200, 5200 and 5045 include all-inclusive licensing for all functions except encryption (a country-limited feature code) and external virtualization.

Note: All internal enclosures in the FS9xx0, 7300, 7200, and 5200 require a license. However, the new FS9500 (4983 only), the FS5200, FS5045 and FS7300 software is License Machine Code.

Any externally virtualized storage requires the External Virtualization license per storage capacity unit (SCU) that is based on the tier of storage that is available on the external storage system. In addition, if you use FlashCopy and Remote Mirroring on an external storage system, you must purchase a per-tebibyte license to use these functions.

The SCU is defined in terms of the category of the storage capacity, as listed in Table 1-21.

Table 1-21 SCU category definitions

License	Drive class	SCU ratio
SCM	SCM devices	SCU equates to 1.00 TiB usable of Category 1 storage.
Flash	All flash devices, other than SCM drives	SCU equates to 1.18 TiB usable of Category 2 storage.
Enterprise	10 K or 15 K RPM drives	SCU equates to 2 TiB usable of Category 3 storage.
NL	NL SATA drives	SCU equates to 4.00 TiB usable of Category 4 storage.

Encryption license (key-based)

Encryption is enabled on IBM FlashSystem systems by obtaining the Encryption Enablement feature. This feature enables encryption at the system level and externally virtualized storage subsystems.

The encryption feature uses a key-based license that is activated by using an authorization code. The authorization code is sent with the IBM FlashSystem Licensed Function Authorization documents that you receive after purchasing the license.

The Encryption USB Flash Drives (Four Pack) feature or an external key manager, such as the IBM Security Key Lifecycle Manager, are required for encryption keys management.

1.21.3 Licensing IBM FlashSystem 5035 and 5015

The base license that is provided with the system includes its basic functions. However, extra licenses can be purchased to expand the capabilities of the system. Administrators are responsible for purchasing extra licenses and configuring the systems within the license agreement, which includes configuring the settings of each licensed function.

IBM FlashSystem 5000 licenses (key-based)

The IBM FlashSystem 5015 and 5035 systems use key-based licensing in which an authorization code is used to activate licensed functions on the system. The authorization code is sent with the IBM FlashSystem 5000 Licensed Function Authorization documents that you receive after purchasing the license. These documents contain the authorization codes that are required to obtain keys (also known as *DFSA license keys*) for each licensed function that you purchased for your system. For each license that you purchase, a separate document with an authorization code is sent to you.

Each function is licensed to an IBM FlashSystem 5000 control enclosure. It covers the entire system (control enclosure and all attached expansion enclosures) if it consists of one I/O group. If the IBM FlashSystem 5030 / 5035 systems consists of two I/O groups, two keys are required.

The following functions require a license key before they can be activated on the system:

- ▶ Easy Tier

Easy Tier automatically and dynamically moves frequently accessed data to flash (solid-state) drives in the system, which results in flash drive performance without manually creating and managing storage tier policies. Easy Tier makes it easy and economical to deploy flash drives in the environment. In this dynamically tiered environment, data movement is seamless to the host application, regardless of the storage tier in which the data is stored.

- ▶ Remote Mirroring

The Remote Mirroring (also known as remote copy [RC]) function enables you to set up a relationship between two volumes so that updates that are made by an application to one volume are mirrored on the other volume.

The license settings apply to only the system on which you are configuring license settings. For RC partnerships, a license also is required on any remote systems that are in the partnership.

- ▶ FlashCopy upgrade

The FlashCopy upgrade extends the base FlashCopy function that is included with the product. The base version of FlashCopy limits the system to 64 target volumes. With the FlashCopy upgrade license activated on the system, this limit is removed. If you reach the limit that is imposed by the base function before activating the upgrade license, you cannot create more FlashCopy mappings.

To help evaluate the benefits of these new capabilities, Easy Tier and RC licensed functions can be enabled at no extra charge for a 90-day trial. Trials are started from the

IBM FlashSystem management GUI and do not require any IBM intervention. When the trial expires, the function is automatically disabled unless a license key for that function is installed onto the machine.

If you use a trial license, the system warns you at regular intervals when the trial is about to expire. If you do not purchase and activate the license on the system before the trial license expires, all configurations that use the trial licenses are suspended.

Encryption license (key-based)

Encryption is enabled on IBM FlashSystem 5035 through the acquisition of the Encryption Enablement feature. This feature enables encryption on the entire IBM FlashSystem family system and externally virtualized storage subsystems.

Note: Encryption hardware feature is available on the IBM FlashSystem 5035 and 5045 only.

This encryption feature uses a key-based license and is activated with an authorization code. The authorization code is sent with the IBM FlashSystem 5000 Licensed Function Authorization documents that you receive after purchasing the license.

The Encryption USB flash drives (Four Pack) feature or IBM Security Key Lifecycle Manager are required for encryption keys management.



Storage area network guidelines

A storage area network (SAN) is one of the most important aspects when implementing and configuring an IBM Storage Virtualize system. Because of its unique behavior and the interaction with other storage, specific SAN design and zoning recommendations differ from classic storage practices.

This chapter provides guidance to connect IBM Storage Virtualize in a SAN to achieve a stable, redundant, resilient, scalable, and performance-likely environment. Although this chapter does *not* describe how to design and build a flawless SAN from the beginning, you can consider the principles that are presented here when building your SAN.

This chapter includes the following topics:

- ▶ “SAN topology general guidelines” on page 124
- ▶ “SAN topology-specific guidelines” on page 127
- ▶ “IBM Storage Virtualize system ports” on page 134
- ▶ “Zoning” on page 149
- ▶ “Distance extension for remote copy services” on page 181
- ▶ “Tape and disk traffic that share the SAN” on page 188
- ▶ “Switch interoperability” on page 189

2.1 SAN topology general guidelines

The SAN topology requirements for IBM Storage Virtualize do not differ much from any other SAN. A correctly sized and designed SAN enables you to build a redundant and failure-proof environment and minimize performance issues and bottlenecks. Therefore, before installing any of the products that are covered by this book, ensure that your environment follows an actual SAN design and architecture with vendor-recommended SAN devices and code levels.

For more information about SAN design and best practices, see [SAN Fabric Administration Best Practices Guide](#).

A topology is described in terms of how the switches are interconnected. There are several different SAN topologies, such as core-edge, edge-core-edge, or full mesh. Each topology has its utility, scalability, and cost, so one topology is a better fit for some SAN demands than others. Independent of the environment demands, there are a few best practices that must be followed to keep your SAN working correctly, performing correctly, and be redundant and resilient.

IBM Storage Virtualize systems support end-to-end NVMe connectivity along with Small Computer System Interface (SCSI). NVMe is a high-speed transfer protocol to leverage the parallelism of solid-state drives (SSDs) / FlashCore Modules (FCMs) in IBM Storage Virtualize systems.

IBM Storage Virtualize supports NVMe on three types of fabric transports:

- ▶ NVMe over Fabrics (NVMe-oF) by using Remote Direct Access Memory (RDMA)
 - RDMA over Converged Ethernet (RoCE) (Ethernet)
- ▶ NVMe over Fabrics by using NVMe/TCP
- ▶ NVMe over Fabrics (NVMe-oF) by using Fibre Channel Protocol (FCP): Fibre Channel-Nonvolatile Memory Express (FC-NVMe)

IBM Storage Virtualize systems support Ethernet connectivity by using 25 Gb and 100 Gb adapter options for internet Small Computer Systems Interface (iSCSI), iSCSI Extensions for RDMA (iSER) on RoCE or iWARP, or NVMe over RDMA. Each adapter supports a different use case. With 25 Gb and 100 Gb, both adapters support host attachment by using iSCSI and NVMe over RDMA (RoCE). iSER (RoCE or iWARP), clustering or HyperSwap, native IP replication, and external virtualization are supported only on a 25 Gb adapter.

2.1.1 SAN performance and scalability

Regardless of the storage and the environment, planning and sizing the SAN makes a difference when growing your environment and troubleshooting problems.

Because most SAN installations continue to grow over the years, the main SAN industry-lead companies design their products to support a certain type of growth. Your SAN must be designed to accommodate both short-term and medium-term growth.

From the performance standpoint, the following topics must be evaluated and considered:

- ▶ Host-to-storage fan-in fan-out ratios
- ▶ Host to Inter-Switch Link (ISL) oversubscription ratio
- ▶ Edge switch to core switch oversubscription ratio
- ▶ Storage to ISL oversubscription ratio
- ▶ Size of the trunks
- ▶ Monitoring for slow drain device issues

From a scalability standpoint, ensure that your SAN can support the new storage and host traffic. Make sure that the chosen topology can support growth in performance and port density.

If new ports must be added to the SAN, you might need to drastically modify the SAN to accommodate a larger-than-expected number of hosts or storage. Sometimes these changes increase the number of hops on the SAN and so cause performance and ISL congestion issues. For more information, see 2.1.2, “ISL considerations” on page 125.

Consider using SAN director-class switches. They reduce the number of switches in a SAN and provide the best scalability that is available. Most of the SAN equipment vendors provide high port density switching devices. With MDS 9718 Multilayer Director, Cisco offers the industry’s highest port density single chassis with up to seven hundred and sixty-eight 16 or 32 Gb ports. The Brocade UltraScale Inter-Chassis Links (ICL) technology enables you to create multichassis configurations with up to nine directors, or four thousand six hundred and eight 16 or 32 Gb ports.

IBM offers the IBM Storage Networking SAN128B-7 high-density 64G switch that provides an opportunity for high scale fabrics in a less rack space.

Therefore, if possible, plan for the maximum size configuration that you expect your IBM Storage Virtualize installation to reach. Planning for the maximum size does not mean that you must purchase all the SAN hardware initially; it requires you to design only the SAN to reach the expected maximum size.

2.1.2 ISL considerations

ISLs are responsible for interconnecting the SAN switches, creating SAN flexibility and scalability. For this reason, they can be considered the core of a SAN topology. Therefore, they are sometimes the main cause of issues that can affect a SAN. For this reason, it is important to take extra caution when planning and sizing the ISL in your SAN.

Regardless of your SAN size, topology, or the size of your IBM Storage Virtualize installation, consider applying the following best practices to your SAN ISL design:

- ▶ Be aware of the ISL oversubscription ratio.

The standard recommendation is up to 7:1 (seven hosts that use a single ISL). However, it can vary according to your SAN behavior. Most successful SAN designs are planned with an oversubscription ratio of 7:1, and some extra ports are reserved to support a 3:1 ratio. However, high-performance SANs start at a 3:1 ratio.

Exceeding the standard 7:1 oversubscription ratio requires you to implement fabric bandwidth threshold alerts. If your ISLs exceed 70%, the schedule fabric changes to distribute the load.

- ▶ Avoid unnecessary ISL traffic.

Connect all IBM Storage Virtualize node ports in a clustered system to the same SAN switches or directors as all the storage devices with which the clustered system of IBM Storage Virtualize is expected to communicate. Conversely, storage traffic and internode traffic should be avoided over an ISL (except during migration scenarios). For certain configurations, such as HyperSwap or Stretched Cluster topology it is a good practice to organize private SANs (for inter-node communication) and if possible separate ISL communication.

Keep high-bandwidth use servers and I/O-intensive applications on the same SAN switches as the IBM Storage Virtualize host ports. Placing these servers on a separate switch can cause unexpected ISL congestion problems. Also, placing a high-bandwidth server on an edge switch wastes ISL capacity.

- ▶ Properly size the ISLs on your SAN. They must have adequate bandwidth and buffer credits to avoid traffic or frames congestion. A congested ISL can affect the overall fabric performance.
- ▶ Always deploy redundant ISLs on your SAN. Using an extra ISL avoids congestion if an ISL fails because of certain issues, such as a SAN switch line card or port blade failure.
- ▶ Use the link aggregation features, such as Brocade Trunking or Cisco Port Channel to obtain better performance and resiliency.
- ▶ Avoid exceeding two hops between IBM Storage Virtualize and the hosts. More than two hops are supported. However, when ISLs are not sized properly, more than two hops can lead to ISL performance issues and buffer credit starvation (SAN congestion).

When sizing over two hops, consider that all the ISLs that go to the switch where IBM Storage Virtualize is connected also handle the traffic that is coming from the switches on the edges, as shown in Figure 2-1.

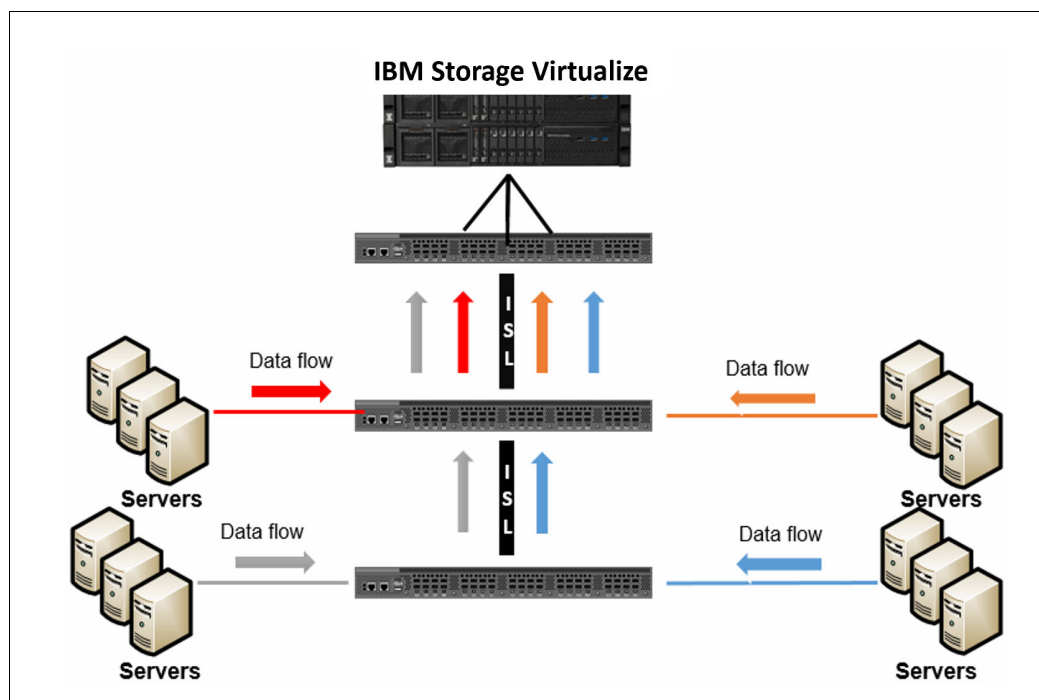


Figure 2-1 ISL data flow

Consider the following points:

- ▶ If possible, use SAN directors to avoid many ISL connections. Problems that are related to oversubscription or congestion are much less likely to occur within SAN director fabrics.
- ▶ When interconnecting SAN directors through an ISL, spread the ISL cables across different director blades. In a situation where an entire blade fails, the ISL still is redundant through the links that are connected to other blades.
- ▶ Plan for the peak load, not for the average load.
- ▶ Lower hop counts maximize the performance by reducing the fabric latency.

2.2 SAN topology-specific guidelines

Some best practices (see 2.1, “SAN topology general guidelines” on page 124) apply to all SANs. However, specific best practices requirements exist for each available SAN topology. In this section, we describe the differences between the types of topology and highlight the specific considerations for each one.

This section covers the following topologies:

- ▶ Single-switch fabric
- ▶ Core-edge fabric
- ▶ Edge-core-edge
- ▶ Full mesh

2.2.1 Single-switch IBM Storage Virtualize SANs

The most basic IBM Storage Virtualize topology consists of a single switch per SAN fabric. This switch can range from a 24-port 1U switch for a small installation of a few hosts and storage devices to a director with hundreds of ports. This configuration is a low-cost design solution that has the advantage of simplicity and a sufficient architecture for small-to-medium IBM Storage Virtualize installations.

One of the advantages of a single-switch SAN is that no hop exists when all servers and storages are connected to the same switches.

Note: To meet redundancy and resiliency requirements, a single-switch solution needs at least two SAN switches or SAN directors (one per different fabric).

A best practice is to use a multislot director-class single switch over setting up a core-edge fabric that is made up solely of lower-end switches, as described in 2.1.1, “SAN performance and scalability” on page 124.

The single switch topology, as shown in Figure 2-2, has only two switches, so the IBM Storage Virtualize ports must be equally distributed on both fabrics.

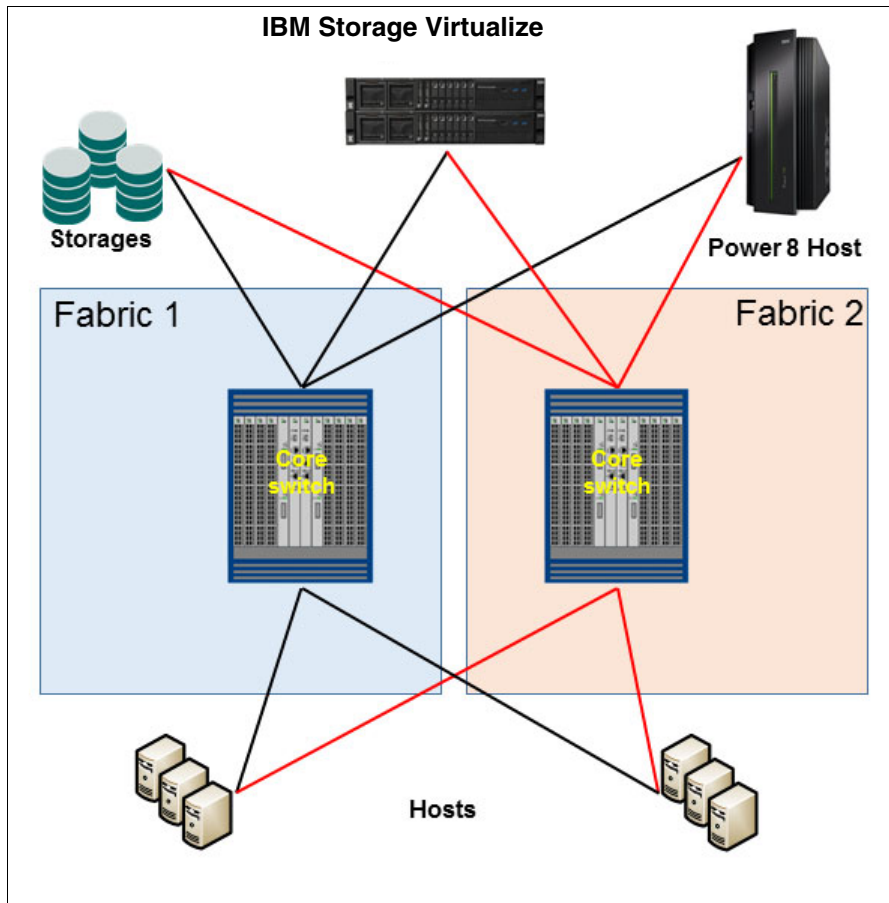


Figure 2-2 Single-switch SAN

Note: To correctly size your network, always calculate the short-term and mid-term growth to avoid lack of ports. On this topology, the limit of ports is based on the switch size. If other switches are added to the network, the topology type is changed automatically.

2.2.2 Basic core-edge topology

The core-edge topology (as shown in Figure 2-3 on page 129) is easily recognized by most SAN architects. This topology consists of a switch in the center (usually, a director-class switch), which is surrounded by other switches. The *core switch* contains all IBM Storage Virtualize ports, storage ports, and high-bandwidth hosts. It is connected by using ISLs to the edge switches. The edge switches can be of any size from 24 port switches up to multi-slot directors.

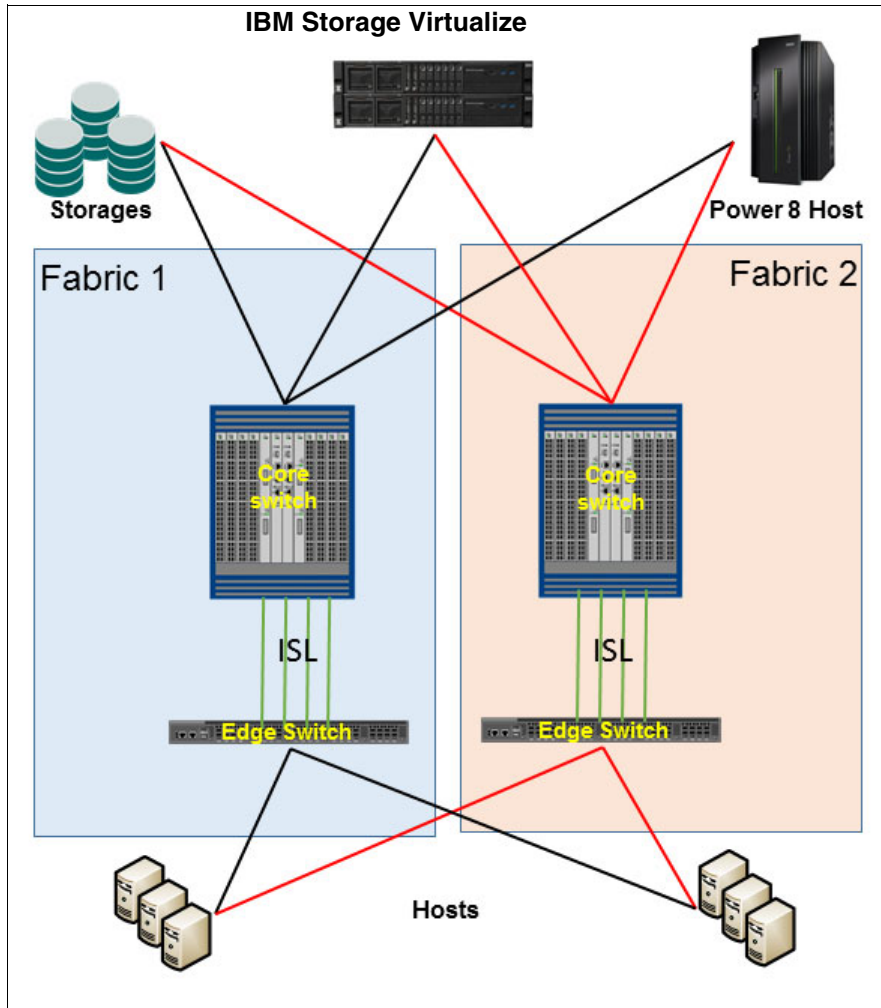


Figure 2-3 Core-edge topology

When IBM Storage Virtualize and the servers are connected to different switches, the hop count for this topology is one.

Note: This topology is commonly used to easily grow your SAN network by adding edge switches to the core. Consider the ISL ratio and usage of physical ports from the core switch when adding edge switches to your network.

2.2.3 Edge-core-edge topology

Edge-core-edge is the most scalable topology. It is used for installations where a core-edge fabric that is made up of multislot director-class SAN switches is insufficient. This design is useful for large, multiclustered system installations. Like a regular core-edge, the edge switches can be of any size, and multiple ISLs must be installed per switch.

Figure 2-4 shows an edge-core-edge topology with two different edges, one of which is exclusive for the IBM Storage Virtualize system and high-bandwidth servers. The other pair is exclusively for servers.

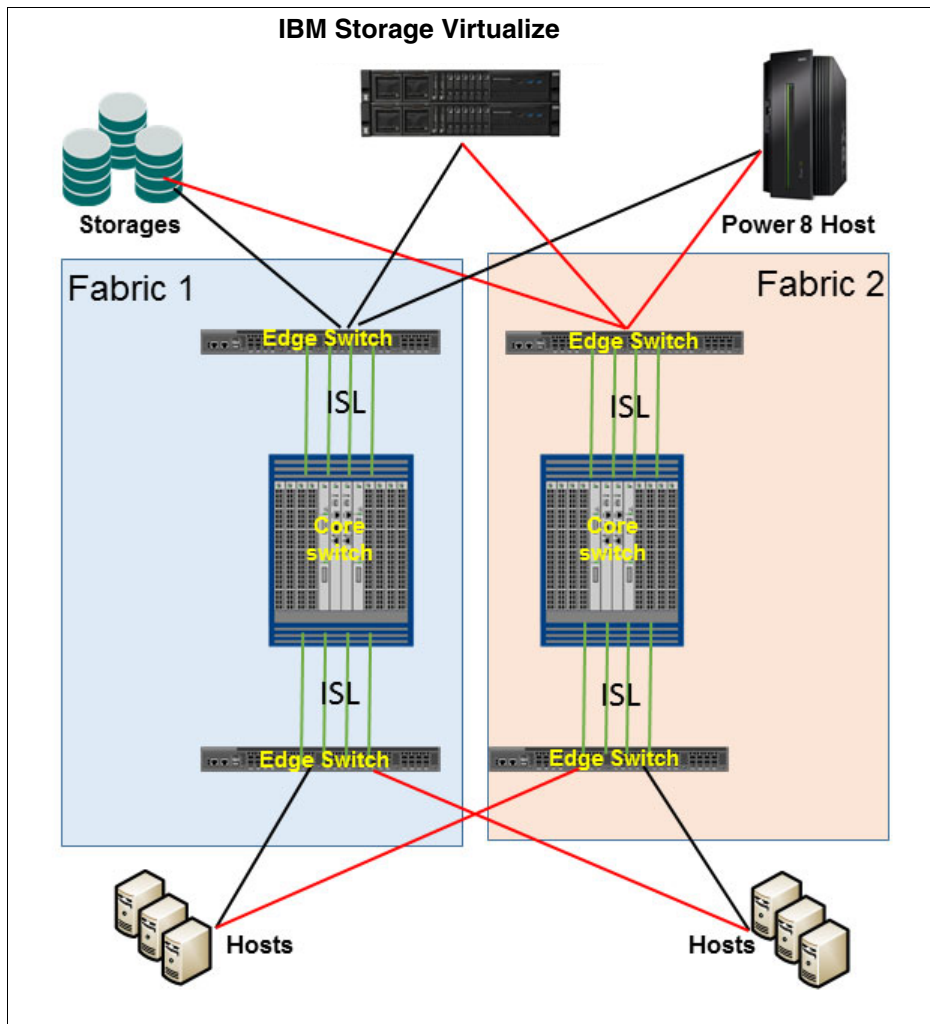


Figure 2-4 Edge-core-edge topology

Performance can be slightly affected if the number of hops increases, which depends on the total number of switches and the distance between the host and the IBM Storage Virtualize system.

Edge-core-edge fabrics allow better isolation between tiers. For more information, see 2.2.6, “Device placement” on page 132.

2.2.4 Full mesh topology

In a full mesh topology, all switches are interconnected to all other switches on the same fabric. Therefore, the server and storage placement is not a concern after the number of hops is no more than one hop. A full mesh topology is shown in Figure 2-5.

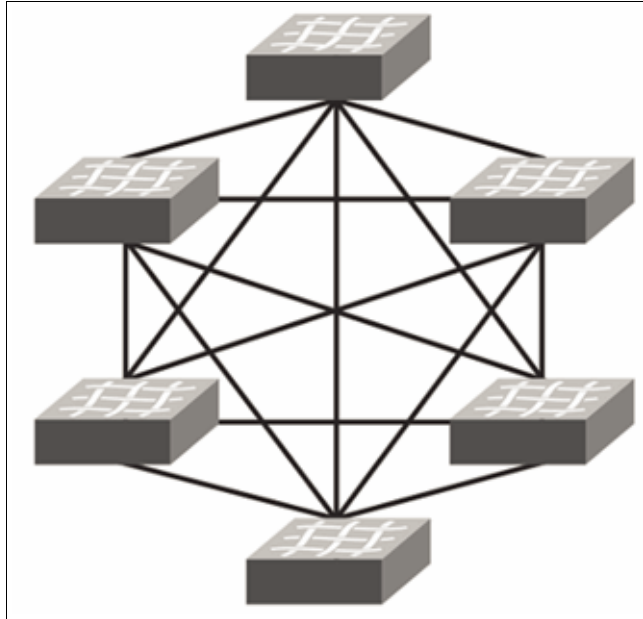


Figure 2-5 Full mesh topology

Note: Each ISL uses one physical port. Depending on the total number of ports that each switch has and the total number of switches, this topology uses several ports from your infrastructure to be set up.

2.2.5 IBM Storage Virtualize as a multi-SAN device

IBM Storage Virtualize system supports different port configurations in each model. In addition to the increased throughput capacity, the number of ports enables new possibilities and allows different kinds of topologies and migration scenarios.

One of these topologies is the usage of an IBM Storage Virtualize system as a multi-SAN device between two isolated SANs. This configuration is useful for storage migration or sharing resources between SAN environments without merging them.

To use an external storage with an IBM Storage Virtualize system, this external storage must be attached to the IBM Storage Virtualize system through the zoning configuration and set up as virtualized storage. This feature can be used for storage migration and decommissioning processes and to speed up host migration. In some cases, based on the external storage configuration, virtualizing external storage with an IBM Storage Virtualize system can increase performance based on the cache capacity and processing.

Figure 2-6 shows an example of IBM Storage Virtualize as a multi-SAN device.

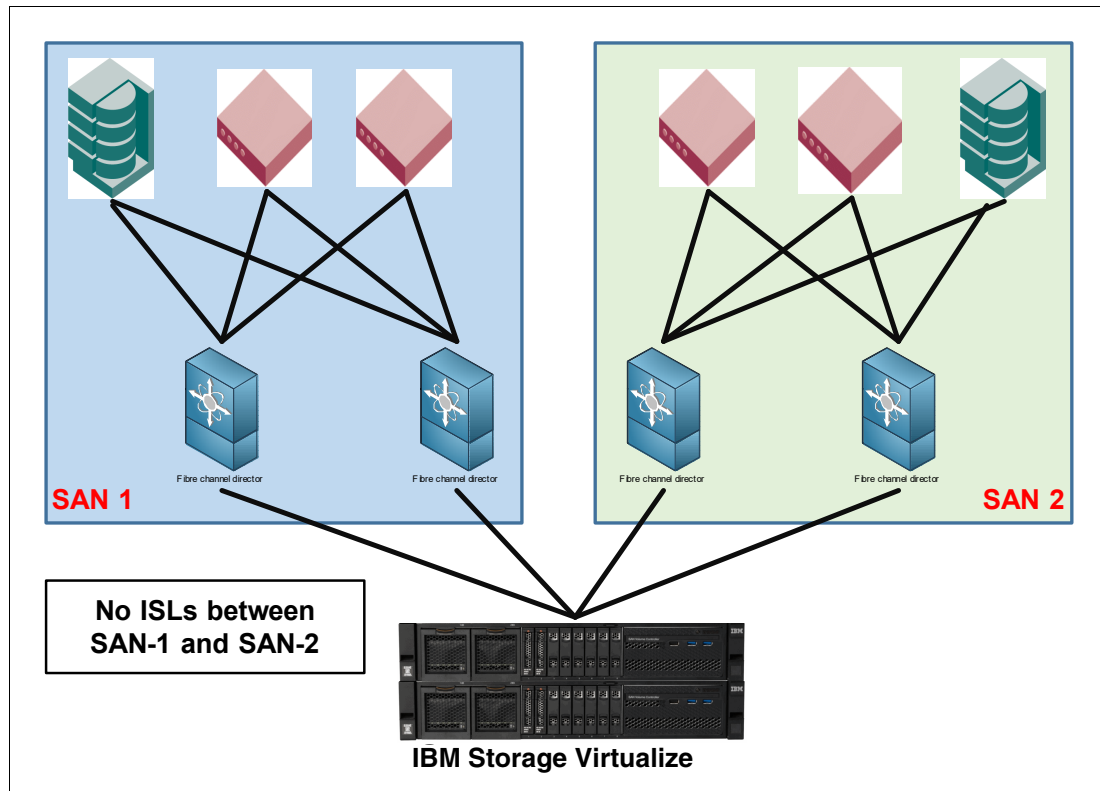


Figure 2-6 IBM Storage Virtualize as a SAN bridge

In Figure 2-6, both SANs are isolated. When connected to both SAN networks, the IBM Storage Virtualize system can allocate storage to hosts on both SAN networks. It is also possible to virtualize storage from each SAN network. This way, you can have established storage on SAN2 (SAN 2 in Figure 2-6) that is attached to the IBM Storage Virtualize system and provide disks to hosts on SAN1 (SAN 1 in Figure 2-6). This configuration is commonly used for migration purposes or in cases where the established storage has a lower performance compared to the IBM Storage Virtualize system.

2.2.6 Device placement

In a correctly sized environment, it is not usual to experience frame congestion on the fabric. Device placement seeks to balance the traffic across the fabric to ensure that the traffic is flowing in a specific way to avoid congestion and performance issues. The ways to balance the traffic consist of isolating traffic by using zoning, virtual switches, or traffic isolation zoning.

Keeping the traffic local to the fabric is a strategy to minimize the traffic between switches (and ISLs) by keeping storages and hosts attached to the same SAN switch, as shown in Figure 2-7.

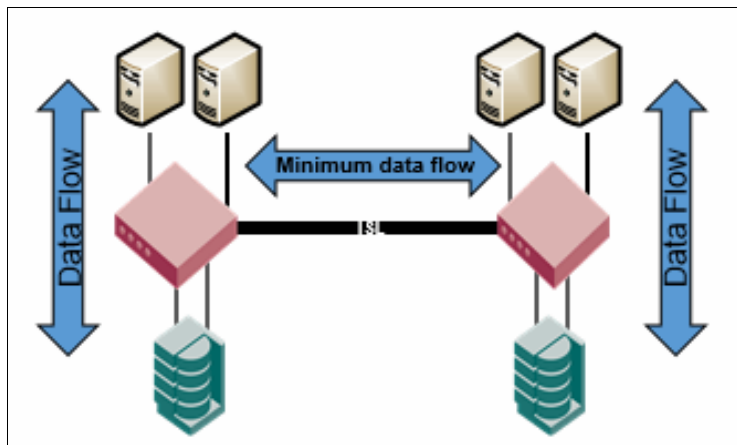


Figure 2-7 Storage and hosts attached to the same SAN switch

This solution can work well in small- and medium-sized SANs. However, it is not as scalable as other topologies that are available. The most scalable SAN topology is the edge-core-edge, which is described in 2.2, “SAN topology-specific guidelines” on page 127.

In addition to scalability, this topology provides different resources to isolate the traffic and reduce possible SAN bottlenecks. Figure 2-8 shows an example of traffic segmentation on the SAN by using edge-core-edge topology.

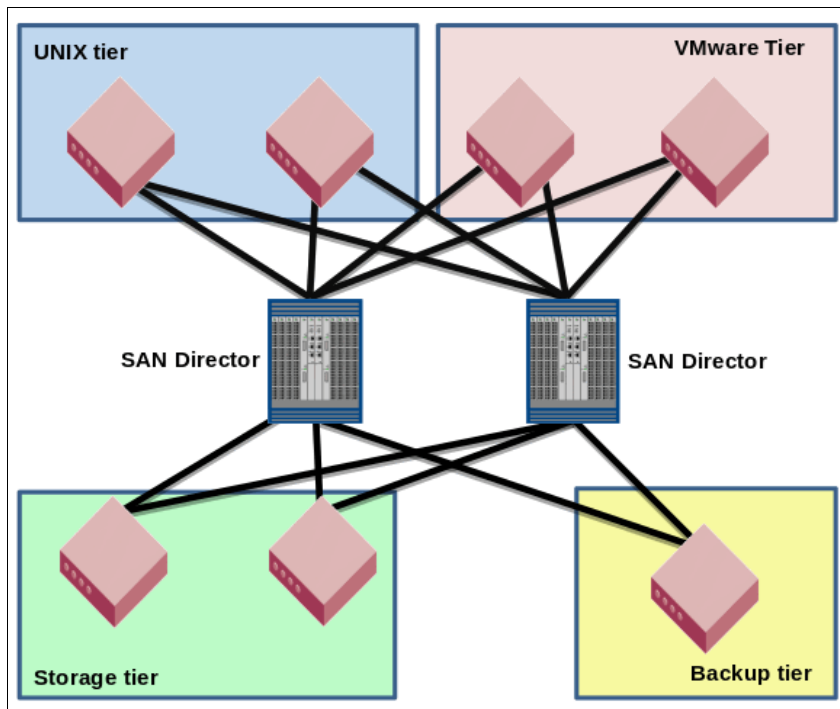


Figure 2-8 Edge-core-edge segmentation

Even when sharing core switches, it is possible to use virtual switches (see 2.2.7, “SAN partitioning” on page 134) to isolate one tier from another one. This configuration helps avoid traffic congestion that is caused by slow drain devices that are connected to the backup tier switch.

2.2.7 SAN partitioning

SAN partitioning is a hardware-level feature that allows SAN switches to share hardware resources by partitioning its hardware into different and isolated virtual switches. Both Brocade and Cisco provide SAN partitioning features called *Virtual Fabric* (Brocade) and *virtual storage area network (VSAN)* (Cisco).

Hardware-level fabric isolation is accomplished through the concept of switch virtualization, which allows you to partition physical switch ports into one or more “virtual switches.” Then, virtual switches are connected to form virtual fabrics.

As the number of available ports on a switch continues to grow, partitioning switches allow storage administrators to take advantage of high port density switches by dividing physical switches into different virtual switches. From a device perspective, SAN partitioning is transparent, so the same guidelines and practices that apply to physical switches apply also to virtual ones.

Although the main purposes of SAN partitioning are port consolidation and environment isolation, this feature is also instrumental in the design of a business continuity solution that is based on IBM Storage Virtualize.

SAN partitioning can be used to dedicate the IBM Storage Virtualize Fibre Channel (FC) ports for internode communication, replication communication, and host to storage communication along with IBM Storage Virtualize port masking.

Note: When director-class switches are used, use ports from different blades to seek load balance and avoid a single point of failure.

For more information about IBM Storage Virtualize business continuity solutions, see Chapter 7, “Ensuring business continuity” on page 501.

2.3 IBM Storage Virtualize system ports

IBM Storage Virtualize adds a common operating environment for all integrated arrays. The IBM Storage Virtualize family is composed of the following products:

- ▶ IBM SAN Volume Controller (SVC)
- ▶ IBM FlashSystem:
 - IBM FlashSystem 9500
 - IBM FlashSystem 9500R
 - IBM FlashSystem 9200
 - IBM FlashSystem 7300 and 7200
 - IBM FlashSystem 5015 and 5035
 - IBM FlashSystem 5100 and 5200

2.3.1 SAN Volume Controller ports

Port connectivity options are significantly changed in SVC hardware for model 2145-SV3 with 8.6.0 code release depending on chosen configuration, as shown in Table 2-1.

Table 2-1 SAN Volume Controller connectivity

Feature	2145-SV1	2145-SV2	2145-SA2	2145-SV3
FC Host Bus Adapters (HBAs)	Four Quad 16 Gb	Three Quad 16 or 32 Gb (FC-NVMe supported)	Three Quad 16 or 32 Gb (FC-NVMe supported)	Six Quad 32 Gb (FC-NVMe supported) OR Three Quad 64 Gb
Ethernet I/O	Four Quad 10 Gb iSCSI and FCoE	One Dual 25 Gb (available up to three 25 Gb)	One Dual 25 Gb (available up to three 25 Gb)	Six Dual 25 Gb OR Six Dual 100 Gb for iSCSI, iSER, RoCE, RoCEv2, and iWARP
Built-in ports	Four 10 Gb	Four 10 Gb	Four 10 Gb	Two 1 Gb
Serial-attached SCSI (SAS) expansion ports	Four 12 Gb SAS	N/A	N/A	N/A

Note: Ethernet adapters support RoCE or iWARP.

This new port density expands the connectivity options and provides new ways to connect the SVC to the SAN. This section describes some best practices and use cases that show how to connect an SVC on the SAN to use this increased capacity.

Slots and ports identification

The SVC can have up to four quad FC HBA cards (16 FC ports) per node. Figure 2-9 shows the port location in the rear view of the 2145-SV1 node.

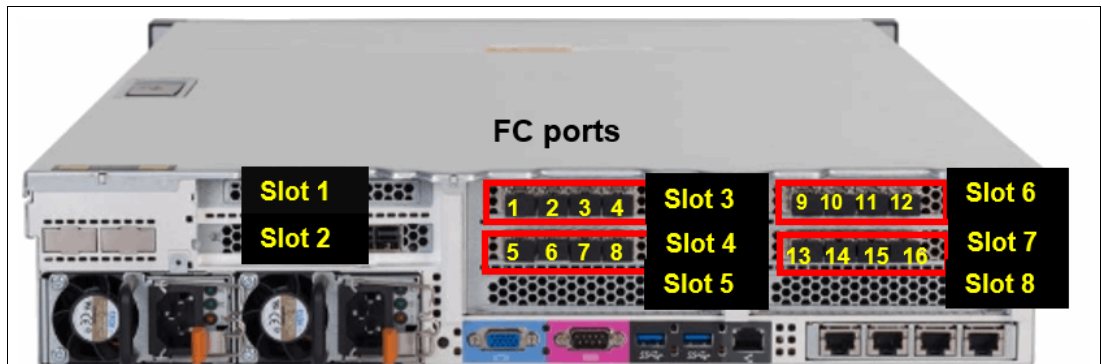


Figure 2-9 SAN Volume Controller 2145-SV1 rear port view

Figure 2-10 shows the port locations for the SV2/SA2 nodes.



Figure 2-10 SV2/SA2 node layout

The IBM SAN Volume Controller SV3 can have up to six quad FC HBA cards (24 FC ports) per node. Figure 2-11 shows the port location in the rear view of the 2145-SV3 node.

SVC SV3 supports eight Peripheral Component Interconnect Express (PCIe) slots that are organized by cages, where three cages (cage1, cage2, cage3) can be used for host/storage connectivity adapter installation:

- ▶ Slots 1 - 2 (cage1), 5 - 6 (cage2), and 7 - 8 (cage3)
- ▶ Slot 3 is for compression offload.
- ▶ Slot 4 is empty



Figure 2-11 SV3 node layout

For maximum redundancy and resiliency, spread the ports across different fabrics. Because the port count varies according to the number of cards that is included in the solution, try to keep the port count equal on each fabric.

2.3.2 IBM FlashSystem 9200 and 9500 controller ports

Port connectivity options are significantly increased with IBM FlashSystem 9200 hardware. Models 9846-AG8 and 9848-AG8 can deliver up to twelve 16 Gb or twelve 32 Gb FC ports per node canister. Ports connectivity per control enclosure is listed in Table 2-2.

Table 2-2 IBM FlashSystem 9200 ports connectivity maximum options per enclosure

Connectivity type	IBM FlashSystem 9200 ports per enclosure
Fibre Channel	24 x 16 Gb ports(3 x Quad port FC HBA per node canister) or 24 x 32 Gb ports(3 x Quad port FC HBA per node canister)

Connectivity type	IBM FlashSystem 9200 ports per enclosure
Ethernet I/O	4 x 25 Gb ports iWARP or RoCE for iSCSI or iSER (2 x Dual Ethernet HBAs per node canister)
Built-in ports	8 x 10 Gb ports (Four 10 Gb ports per canister node) for internet Small Computer Systems Interface (iSCSI) and 2 x 1Gb technician ports
SAS expansion ports	Two Quad 12 Gb SAS (two ports active)(1x SAS expansion adapter per node canister)

Port connectivity options are significantly increased with IBM FlashSystem 9500 compared to previous models. IBM FlashSystem 9500 Model AH8 at code level 8.6.0 number of connectivity ports per enclosure (two node canisters) are listed in Table 2-3.

Table 2-3 IBM FlashSystem 9500 ports connectivity maximum options per enclosure

Connectivity type	IBM FlashSystem 9500 ports per enclosure
Fibre Channel	24 x 64 Gb ports (3xQuad port FC HBA per node canister) OR 48 x 32 Gb ports (6xQuad port FC HBA per node canister)
Ethernet I/O	24 x 10/25 Gb ports (iWARP or ROCE, iSCSI) OR 24x 100 Gb ports (NVME/RDMA ROCEv2, iSCSI) that is 6x Dual port Ethernet HBA per node canister
Built-in ports	4 x 10 Gb for management (iSCSI) and 2 x 1 Gb technician ports.
SAS expansion ports	Two Quad 12 Gb SAS (two ports active) (1x SAS expansion adapter per node canister)

Note:

- ▶ IBM FlashSystem 9200 node canisters feature three PCIe slots where you can combine the adapters as needed. If expansions are used, one of the slots must have the SAS expansion card. Then, two ports are left for FC HBA cards, whether iWARP or RoCE Ethernet adapters.
- ▶ The IBM FlashSystem 9500 node canister features eight PCIe slots where you can combine the cards as needed. And the information in the table above shows the maximum combination of specific one type connectivity HBA. For example if 24 Fibre Channel ports 64Gb are desired for configuration. This configuration occupies all three cages in both node canisters, each cage with one 64Gb card installed, thus eliminating possibility to install Ethernet HBAs.

For more information, see [IBM FlashSystem 9200](#).

Slots and ports identification

IBM FlashSystem 9200 can have up to three quad FC HBA cards (12 FC ports) per node canister. Figure 2-12 on page 138 shows the port location in the rear view of the IBM FlashSystem 9200 node canister.

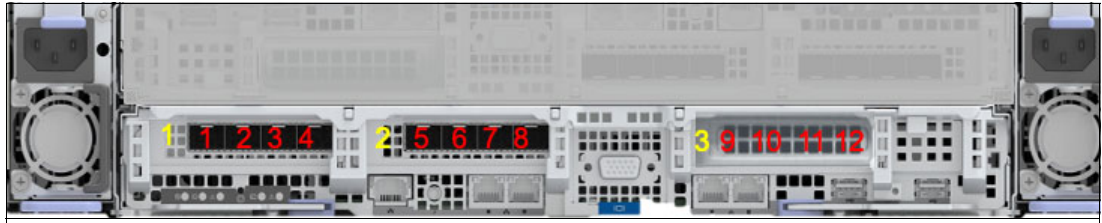


Figure 2-12 Port location in the IBM FlashSystem 9200 rear view

For maximum redundancy and resiliency, spread the ports across different fabrics. Because the port count varies according to the number of cards that is included in the solution, try to keep the port count equal on each fabric.

IBM FlashSystem 9500 has eight PCIe slots where six of them (dedicated for cage1, cage2, cage3) can be used for host/storage connectivity adapters:

- ▶ Slots 1 - 2 (cage1), 5 - 6 (cage2), and 7 - 8 (cage3)
- ▶ Slot 3 for compression offload
- ▶ Slot 4 empty

Figure 2-13 shows the port location in the rear view of the IBM FlashSystem 9200 node canister.



Figure 2-13 Port location in IBM FlashSystem 9500 rear view

2.3.3 IBM FlashSystem 7200 and 7300 controller ports

The port connectivity options in IBM FlashSystem 7200 support 16 Gb and 32 Gb FC and four onboard 10-gigabit Ethernet (GbE) ports, as listed in Table 2-4.

Table 2-4 IBM FlashSystem 7200 ports connectivity maximum per enclosure

Feature	IBM FlashSystem 7200 per enclosure
FC HBA	24 x 16 Gb FC/NVMeoF OR 24 x 32 Gb FC/NVMeoF that is 3 x Quad port FC HBAs per node canister
Ethernet I/O	12 x 25 Gb (iSCSI, iSER, or iWARP or RoCE) that is 3 x Dual port Ethernet cards per node canister
Built-in ports	8 x 10 Gb ports (4 per each node canister) and 2 x 1 Gb technician ports (1 per each node canister)

Feature	IBM FlashSystem 7200 per enclosure
SAS expansion ports	4 x Quad 12 Gb SAS (1 card per each node canister, 2 ports active only)

Table 2-5 lists the port connectivity options for IBM FlashSystem 7300.

Table 2-5 IBM FlashSystem 7300 ports connectivity maximum per enclosure

Feature	IBM FlashSystem 7300
FC HBA	24 x 16/32 Gb (FC or NVMeoF)
Ethernet I/O	12 x 25 Gb (iSCSI, iSER, or iWARP or RoCE) 12 x 100 Gb (iSCSI or NVMe over RDMA)
Built-in ports	8 x 10 Gb and 2 x 1 Gb technician port
SAS expansion ports	4 x Quad 12 Gb SAS (1 card per each node canister, 2 ports active only)

Slots and ports identification

The IBM FlashSystem 7200 can have up to three quad FC HBA cards (12 FC ports) per node canister. Figure 2-14 shows the port location in the rear view of the IBM FlashSystem 7200 node canister.

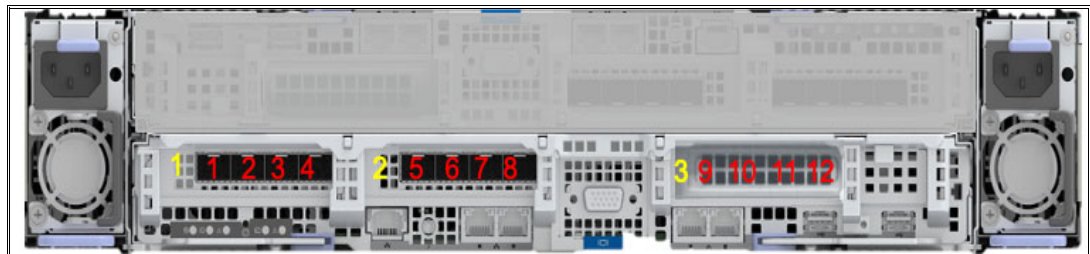


Figure 2-14 IBM FlashSystem 7200 rear view

IBM FlashSystem 7300 can have up to three quad FC HBA cards (12 FC ports) per node canister. Figure 2-15 shows the port location in the rear view of the IBM FlashSystem 7300 node canister.

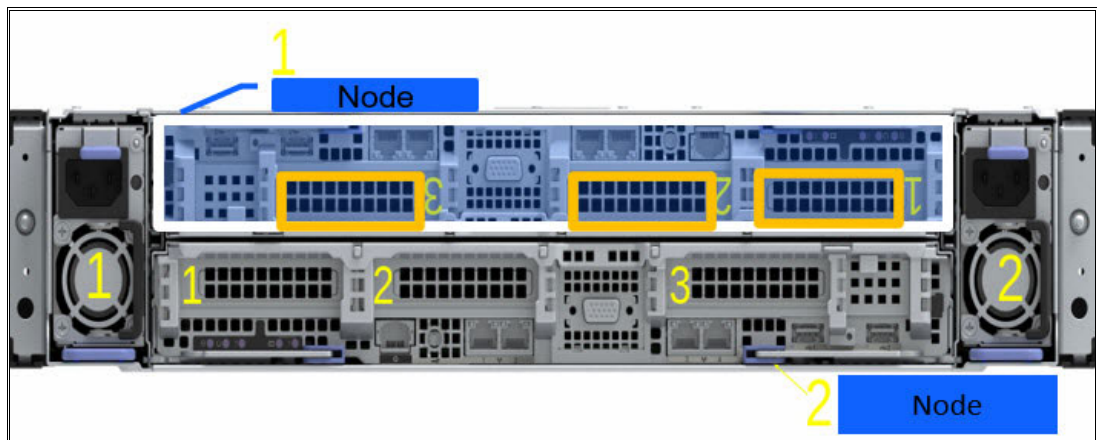


Figure 2-15 IBM FlashSystem 7300 rear view

For maximum redundancy and resiliency, spread the ports across different fabrics. Because the port count varies according to the number of cards that is included in the solution, try to keep the port count equal on each fabric.

2.3.4 IBM FlashSystem 5100 and 5200 and IBM FlashSystem 5015 and 5035 controller ports

IBM FlashSystem 5xxx systems bring the simplicity and innovation of other family members. The following tables list the port connectivity options for IBM FlashSystem 5100 (Table 2-6), IBM FlashSystem 5200 (Table 2-7 on page 141), IBM FlashSystem 5015 (Table 2-8 on page 141), and IBM FlashSystem 5035 (Table 2-9 on page 141).

Table 2-6 IBM FlashSystem 5100 port connectivity maximum per enclosure

Feature	IBM FlashSystem 5100
FC HBA	16 x 16 Gb (FC or NVMeoF) 16 x 32 Gb (FC or NVMeoF)
Ethernet I/O	4 x 25 Gb (iSCSI, iSER, or iWARP or RoCE)
Built-in ports	8 x 10 Gb and two 1 Gb technician ports
SAS expansion ports	4 x 12 Gb SAS

Table 2-7 IBM FlashSystem 5200 port connectivity maximum per enclosure

Feature	IBM FlashSystem 5200
FC HBA	16 x 16 Gb (FC or NVMeoF) 8 x 32 Gb (FC or NVMeoF)
Ethernet I/O	2 x 25 Gb (iSCSI, iSER, or iWARP or RoCE)
Built-in ports	8 x 10 Gb and 2 x 1 Gb technician ports
SAS expansion ports	2 x 12 Gb SAS

Table 2-8 IBM FlashSystem 5015

Feature	IBM FlashSystem 5015
FC HBA	8 x 16 Gb (FC or NVMeoF)
Ethernet I/O	4 x 25 Gb (iSCSI)
Built-in ports	8 x 10 Gb, four 1 Gb (iSCSI), and 2 x 1 Gb technician ports
SAS expansion ports	2 x 12 Gb SAS

Table 2-9 IBM FlashSystem 5035

Feature	IBM FlashSystem 5035
FC HBA	8 x 16 Gb (FC or NVMeoF) 8 x 32 Gb (FC or NVMeoF)
Ethernet I/O	4 x 25 Gb (iSCSI, iSER, or iWARP or RoCE)
Built-in ports	8 x 10 Gb and 2 x 1 Gb technician ports
SAS expansion ports	4 x 12 Gb SAS

Slots and ports identification

IBM FlashSystem 5100 can have up to two quad FC HBA cards (eight 16 Gb FC ports) per node canister. Figure 2-16 shows the port location in the rear view of the IBM FlashSystem 5100 node canister.

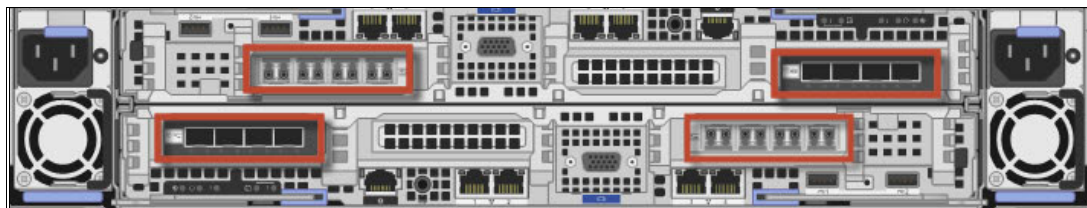


Figure 2-16 IBM FlashSystem 5100 rear view

IBM FlashSystem 5200 can have up to two dual FC HBA cards (four 32 Gb FC ports) per node canister. Figure 2-17 on page 142 shows the port location in the rear view of the IBM FlashSystem 5200 node canister.

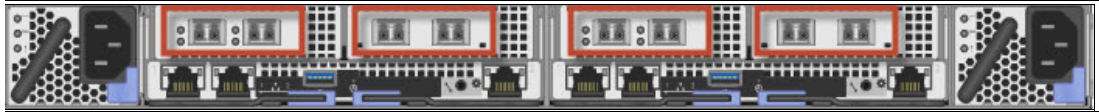


Figure 2-17 IBM FlashSystem 5200 rear view

IBM FlashSystem 5015 can have up to one quad FC HBA cards (four 16 Gb FC ports) per node canister. Figure 2-18 shows the port location in the rear view of the IBM FlashSystem 5015 node canister.

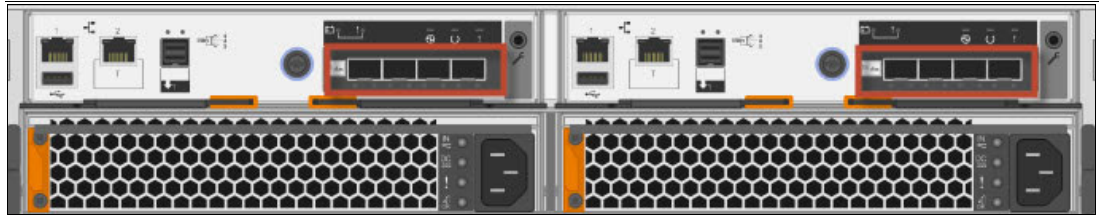


Figure 2-18 IBM FlashSystem 5015 rear view

IBM FlashSystem 5035 can have up to one quad FC HBA card (four 16 Gb or 32 Gb FC ports) per node canister. Figure 2-19 shows the port location in the rear view of the IBM FlashSystem 5035 node canister.



Figure 2-19 IBM FlashSystem 5035 rear view

2.3.5 IBM Storage Virtualize N_Port ID Virtualization, port naming, and distribution

In this section, we describe IBM Storage Virtualize N_Port ID Virtualization (NPIV), port naming, and distribution.

IBM Storage Virtualize NPIV

IBM Storage Virtualize uses NPIV by default, which reduces failover time and allows for features such as Hot Spare Nodes (HSNs). NPIV creates multiple vFC ports per physical FC port and removes the dependence on multipathing software during failover. It allows a partner node to take over the worldwide port names (WWPNs) of a failed node and improves the availability from application perspective.

Transitional mode on an IBM Storage Virtualize system can be used to change or update your zoning from traditional physical WWPNs to NPIV WWPNs that are based on zoning.

Figure 2-20 and Figure 2-21 represent the IBM Storage Virtualize NPIV port WWPN and failover.

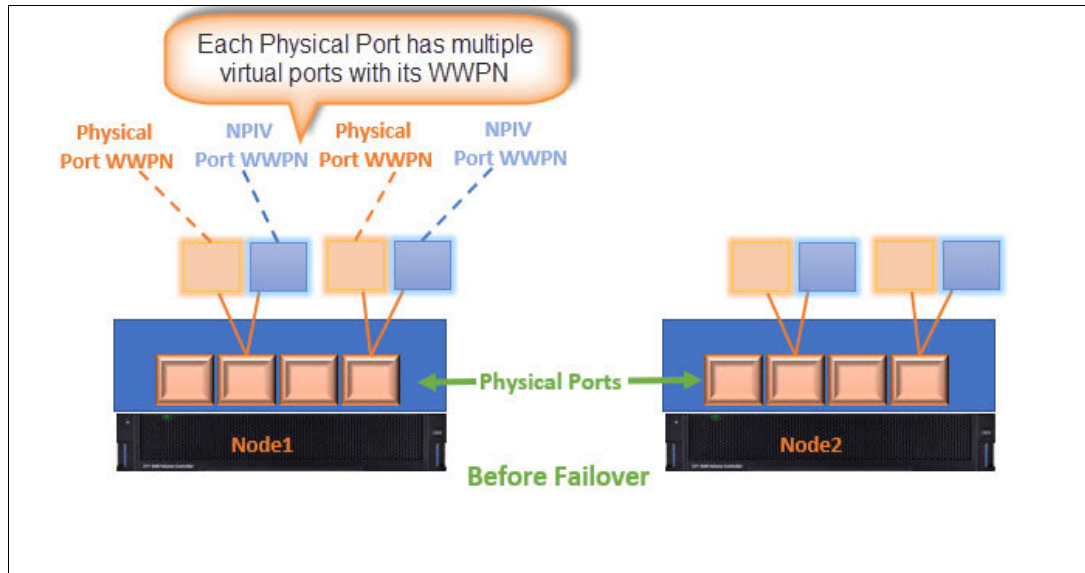


Figure 2-20 IBM Storage Virtualize NPIV Port WWPN

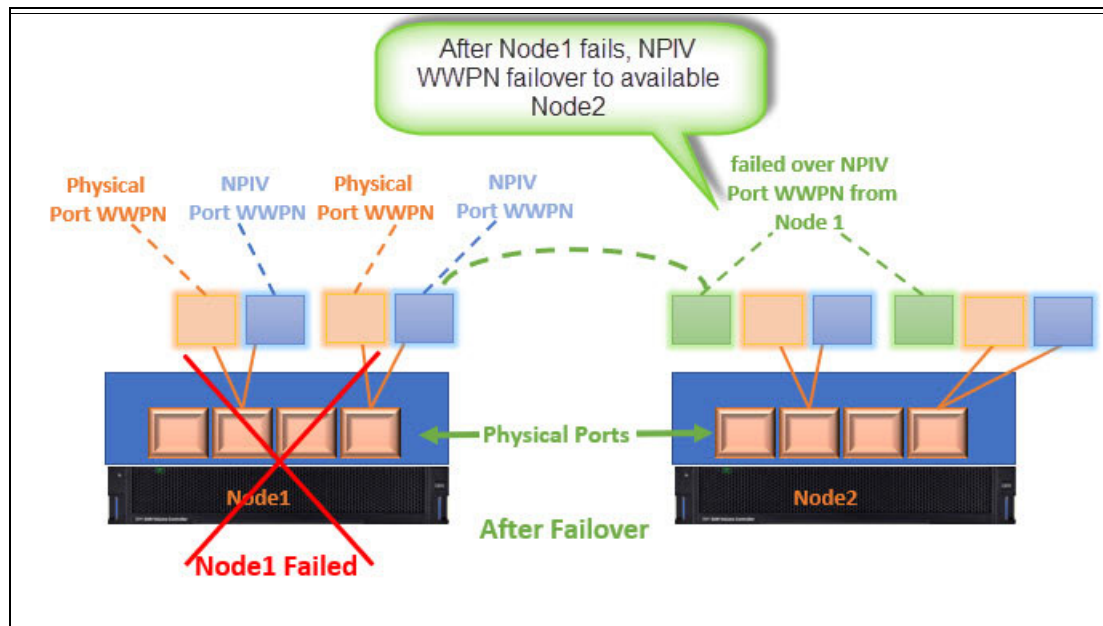


Figure 2-21 IBM Storage Virtualize NPIV Failover

The `lstargerportfc` command output (see Figure 2-22 on page 144) from the IBM Storage Virtualize command-line interface (CLI) shows that each port has three WWPNs: One is a physical WWPN for SCSI connectivity; the second is an NPIV WWPN for SCSI connectivity; and the third is an NPIV WWPN for NVMe connectivity.

```

ruser>l1stargetportfc
port_id owning_node_id current_node_id nportid host_io_permitted virtualized protocol fc_io_port_id portset_count host_count active_login_coun
810000214 1 5 5 080100 no no scsi 1 0 0 2
810000214 1 5 5 080101 yes yes scsi 1 1 0 0
810000214 1 5 5 080102 yes yes nvme 1 1 0 0
810000214 2 5 5 080000 no no scsi 2 0 0 2
810000214 2 5 5 080001 yes yes scsi 2 1 0 0
810000214 2 5 5 080002 yes yes nvme 2 1 0 0
810000214 3 5 5 120100 no no scsi 3 0 0 2
810000214 3 5 5 120101 yes yes scsi 3 1 0 0
810000214 3 5 5 120102 yes yes nvme 3 1 0 0
810000214 4 5 5 120000 no no scsi 4 0 0 2
810000214 4 5 5 120001 yes yes scsi 4 1 0 0
810000214 4 5 5 120002 yes yes nvme 4 1 0 0
810000216 1 4 4 080200 no no scsi 1 0 0 2
810000216 1 4 4 080201 yes yes scsi 1 1 0 0
810000216 1 4 4 080202 yes yes nvme 1 1 0 0
810000216 2 4 4 080300 no no scsi 2 0 0 2
810000216 2 4 4 080301 yes yes scsi 2 1 0 0
810000216 2 4 4 080302 yes yes nvme 2 1 0 0
810000216 3 4 4 120200 no no scsi 3 0 0 2
810000216 3 4 4 120201 yes yes scsi 3 1 0 0
810000216 3 4 4 120202 yes yes nvme 3 1 0 0
810000216 4 4 4 120300 no no scsi 4 0 0 2
810000216 4 4 4 120301 yes yes scsi 4 1 0 0
810000216 4 4 4 120302 yes yes nvme 4 1 0 0

```

Figure 2-22 IBM Storage Virtualize output of the `l1stargetportfc` command

Note: NPIV is not supported for Ethernet connectivity, such as FCOE, iSCSI, and iSER.

The same ports (port IDs) must be in the same fabric to fail over in a hardware failure. For example:

- ▶ Port IDs 1 and 3 of all nodes and spare nodes are part of the odd fabric.
- ▶ Port IDs 2 and 4 of all nodes and spare nodes are part of the even fabric.

Use the Transitional mode to convert your host zoning from physical WWPN to virtual WWPN.

IBM Storage Virtualize Port naming

In the field, fabric naming conventions vary. However, it is common to find fabrics that are named, for example, PROD_SAN_1 and PROD_SAN_2, or PROD_SAN_A and PROD_SAN_B. This type of naming convention is used to simplify the IBM Storage Virtualize systems with their denomination followed by *1* and *2* or *A* and *B*, which specifies that the devices that are connected to those fabrics contains the redundant paths of the same servers and SAN devices.

To simplify the SAN connection identification and troubleshooting, keep all odd ports on the odd fabrics or “A” fabrics, and the even ports on the even fabric or “B” fabrics, as shown in Figure 2-23, which shows the port arrangement for IBM Storage Virtualize different models.

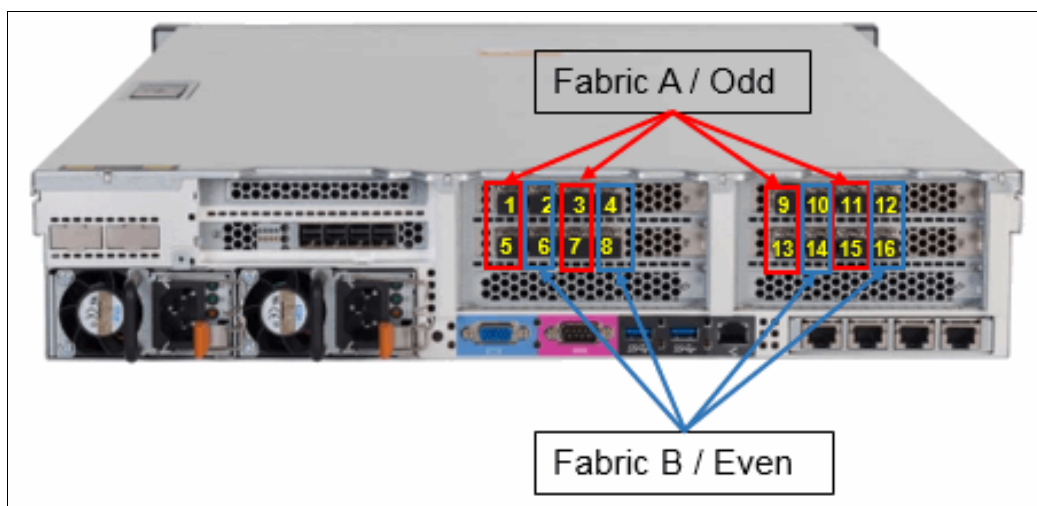


Figure 2-23 SAN Volume Controller model 2145-SV1 port distribution

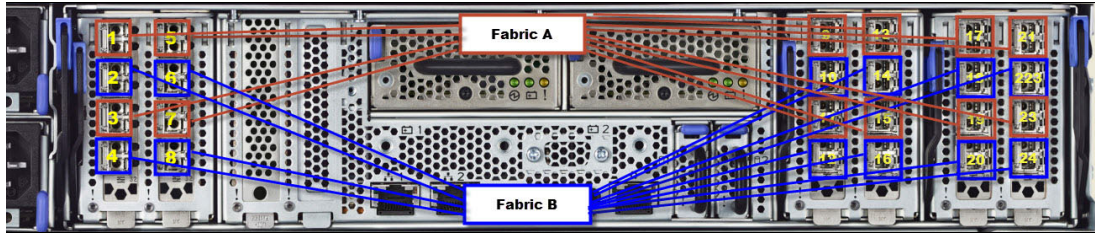


Figure 2-24 SAN Volume Controller model SV3 port distribution

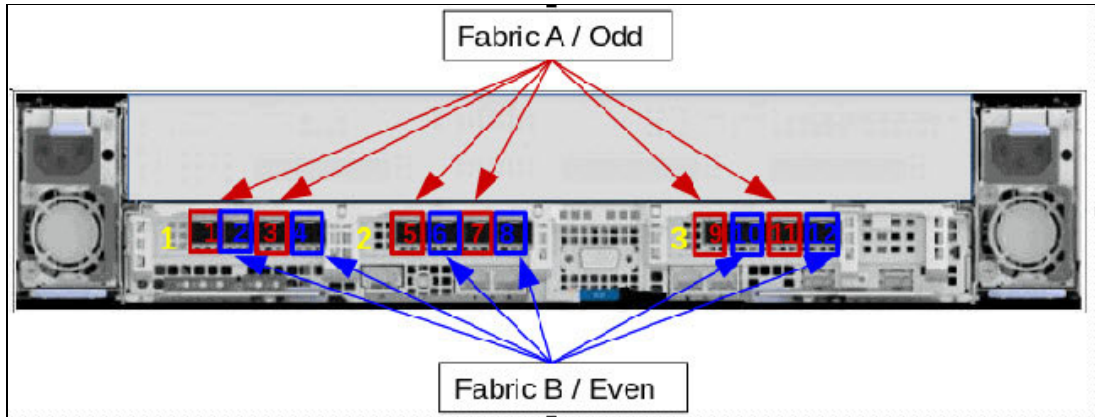


Figure 2-25 IBM FlashSystem 9200 port distribution

The same port distribution can be used for other IBM Storage Virtualize model, for example, IBM FlashSystem 9500 or 9500R, IBM FlashSystem 7300, and IBM FlashSystem 5015 or 5035.

IBM Storage Virtualize port distribution

As a best practice, assign specific uses to specific IBM FlashSystem or IBM SVC ports. This technique helps to optimize the port utilization in a matter of dedicated purpose and in a way prevents failures in one network traffic impacting the other.

Because of the increased port availability on IBM FlashSystem or SVC clusters and the increased bandwidth with the 32 Gb ports and 64Gb ports, it is not only possible to separate the port utilization among hosts and storage, node-to-node (intracluster) and replication traffic but also very sensible.

The example port utilization designation that is shown in the following figures can be used for up to 16 ports (Figure 2-26 on page 146) (32 ports total per cluster) and 24 ports (48 ports total per cluster) (Figure 2-27 on page 147) IBM Storage Virtualize systems.

Design your port designation as required by your workload and to achieve maximum availability in case of a failure as it separates failure domains, and helps in better analysis and system management.

Intracluster (inter-node) communication must be separated in multi-I/O groups, SVC, Enhanced Stretched Cluster (ESC), and HyperSwap scenarios.

Fibre channel port masking provides tool for such port communication segregation. There are two masks available, one that allows to designate ports for local inter-node communication that is recommended to have in case of multi-IO group configuration, and could be checked by the `lsystem` command using the `localfcportmask` parameter.

This mask should be read from right to left, meaning that the far right bit in the mask corresponds with `fc_io_port_id 1`, and so on. Setting corresponding bits of this mask to "1" enables the designated ports for intra-cluster communication exclusively. While ports that have "0" in the mask are restricted to only host/storage traffic and are not used for intracluster communication.

The second mask enables assigning particular node ports exclusively to the replication functionality (inter-node communication with remote cluster nodes). It is mentioned in the output of the `lssystem` command, specifically in the `remotefcportmask` parameter. This value behaves similarly to `localfcportmask`, with each bit representing a specific port. Setting a bit to "1" enables just those particular ports for data replication, while "0" disables communication for all other ports except for hosts or storage systems (based on SAN zoning). Figure 2-26 shows a 16 port configuration, while Figure 2-27 on page 147 shows a 24 port configuration.

Cage (applicable for FS9500/FS9200)	Adapter/fc_io_port_id	4 ports	8 ports	12 ports	16 port	SAN fabric
1	Adapter1/1	Host+Storage	Host+Storage	Host+Storage	Host+Storage	A
	Adapter1/2	Host+Storage	Host+Storage	Host+Storage	Host+Storage	B
	Adapter1/3	Intracluster+Replication	Intracluster or Replication	Intracluster or Replication	Intracluster	A
	Adapter1/4	Intracluster+Replication	Intracluster	Intracluster	Intracluster	B
	Adapter2/5		Host+Storage	Host+Storage	Host+Storage	A
	Adapter2/6		Host+Storage	Host+Storage	Host+Storage	B
	Adapter2/7		Intracluster	Replication or Host+Storage	Replication or Host+Storage	A
	Adapter2/8		Intracluster or Replication	Replication or Host+Storage	Replication or Host+Storage	B
2	Adapter3/9			Host+Storage	Host+Storage	A
	Adapter3/10			Host+Storage	Host+Storage	B
	Adapter3/11			Intracluster	Intracluster	A
	Adapter3/12			Intracluster or Replication	Intracluster	B
	Adapter4/13				Host+Storage	A
	Adapter4/14				Host+Storage	B
	Adapter4/15				Replication or Host+Storage	A
	Adapter4/16				Replication or Host+Storage	B
3	Adapter5/17					A
	Adapter5/18					B
	Adapter5/19					A
	Adapter5/20					B
	Adapter6/21					A
	Adapter6/22					B
	Adapter6/23					A
	Adapter6/24					B
port masking	localfcportmask	1100	11001100 or 01001000	110000001100 or 010000001000	0000110000001100	
port masking	remotefcportmask	1100	00000000 or 10000100	000011000000 or 100000000100	1100000011000000 or all 0	

Host refers to the host objects defined in the system that needs access to the data.
Storage refers to the external storage servers, also known as controllers that are virtualized under/behind IBM Spectrum Virtualize.
Intracluster refers to node-to-node communication inside IBM Spectrum Virtualize cluster in case of multi-IO group configuration.
Replication refers to the ports that are used for data replication between other/external clusters nodes.
"+" means that port is recommended to use for the types of communication.
"or" means that one of options can be used depending on the design considerations.

Figure 2-26 Port masking configuration on SVC or IBM FlashSystem with 16 ports

Cage (applicable for FS9500/FS9200)	Adapter/fc_io_port_id	24 ports	SAN fabric
1	Adapter1/1	Host	A
	Adapter1/2	Host	B
	Adapter1/3	Intracuster	A
	Adapter1/4	Intracuster	B
	Adapter2/5	Host or Storage	A
	Adapter2/6	Host or Storage	B
	Adapter2/7	Replication or Host/Storage	A
	Adapter2/8	Replication or Host/Storage	B
2	Adapter3/9	Host or Storage	A
	Adapter3/10	Host or Storage	B
	Adapter3/11	Intracuster	A
	Adapter3/12	Intracuster	B
	Adapter4/13	Host	A
	Adapter4/14	Host	B
	Adapter4/15	Host or Storage	A
	Adapter4/16	Host or Storage	B
3	Adapter5/17	Host or Storage	A
	Adapter5/18	Host or Storage	B
	Adapter5/19	Host	A
	Adapter5/20	Host	B
	Adapter6/21	Host or Storage	A
	Adapter6/22	Host or Storage	B
	Adapter6/23	Replication or Host/Storage	A
	Adapter6/24	Replication or Host/Storage	B
port masking	localfcportmask	000000000000110000001100	
port masking	remotefcportmask	110000000000000011000000 or all 0	

Figure 2-27 Port masking configuration on IBM FlashSystem or SVC with 24 ports

Note: Consider the following points:

- ▶ On SVC clusters with 12 ports or more per node, if you use advanced copy services (such as volume mirroring, FlashCopy, or remote copy) on an ESC, it is a best practice that you use four ports per node for inter-node communication. The SVC uses the internode ports for all of these functions, and the usage of these features greatly increases the data rates that are sent across these ports. If you are implementing a new cluster and plan to use any copy services, plan on having more than 12 ports per cluster node. It is also applicable for IBM FlashSystem clusters with multi-IO group configuration
- ▶ In case of IBM FlashSystem system with a single I/O group, the system keeps `localfcportmask` set to default all “1” and does not allow you to change this setting. However, this is not an issue because the inter-node traffic occurs on the internal PCI mid-plane link inside one enclosure. The port-masking recommendations that are shown in Figure 2-27 on page 147 and Figure 2-26 on page 146 apply to systems with more than one I/O group.
- ▶ Depending on the workload or number of I/O groups, you can reserve ports 3, 4 and 9, 10 for inter-node communication traffic. In this case, port masking is used and *other ports are disabled from inter-node traffic communication*, and thus `localfcportmask` is set to 001100001100.

Host and storage ports have different traffic behavior, so keeping host and storage ports together produces maximum port performance and utilization by benefiting from its full duplex bandwidth. For this reason, sharing host and storage traffic in the same ports is a best practice. However, traffic segmentation can also provide some benefits in terms of troubleshooting and host zoning management, see Figure 2-27 on page 147 for example of traffic segmentation. Consider, for example, SAN congestion conditions due to a slow draining device.

In this case, segregating the ports simplifies the identification of the device that is causing the problem while limiting the effects of the congestion to the hosts or back-end ports only. Furthermore, dedicating ports for host traffic reduces the possible combinations of host zoning and simplifies SAN management. It is best practice to implement port traffic segmentation with configurations with more ports only.

2.3.6 Buffer credits

IBM Storage Virtualize systems FC ports have a predefined number of buffer credits. The number of buffer credits determines the available throughput over distances:

- ▶ All 8 Gbps adapters have 41 credits available per port, saturating links at up to 10 km (6.2 miles) at 8 Gbps.
- ▶ Two-port 16 Gbps (DH8 only nodes) adapters have 80 credits available per port, saturating links at up to 10 km (6.2 miles) at 16 Gbps.
- ▶ Four-port 16 Gbps adapters have 40 credits available per port, saturating links at up to 5 km at (3.1 miles) 16 Gbps.
- ▶ Four-port 32 Gbps adapters have 40 credits available per port, saturating links at up to 2.5 km at (1.5 miles) 32 Gbps.

Switch port buffer credit: For Stretched cluster and IBM HyperSwap configurations that do not use ISLs for the internode communication, it is best practice to set the switch port buffer credits to match the IBM Storage Virtualize port.

2.4 Zoning

This section describes the zoning recommendations for IBM Storage Virtualize systems. Zoning an IBM Storage Virtualize cluster into a SAN fabric requires planning and following specific guidelines.

Important: Errors that are caused by improper IBM Storage Virtualize system zoning are often difficult to isolate, and the steps to fix them can affect the SAN environment. Therefore, create your zoning configuration carefully.

The initial configuration for IBM Storage Virtualize requires the following separate zones:

- ▶ Internode and intra-cluster zones
- ▶ Replication zones (if replication is used)
- ▶ Back-end storage to IBM Storage Virtualize zoning for external virtualization
- ▶ Host to IBM Storage Virtualize zoning

Different guidelines must be followed for each zoning type, as described in 2.4.1, “Types of zoning” on page 149.

Note: Although internode and intra-cluster zone is not necessary for non-clustered IBM Storage Virtualize systems except SVC, it is generally preferred to use one of these zones.

2.4.1 Types of zoning

Modern SAN switches feature two types of zoning: port zoning, and WWPN zoning. The preferred method is to use only WWPN zoning. A common misconception is that WWPN zoning provides poorer security than port zoning, which is not the case. Modern SAN switches enforce the zoning configuration directly in the switch hardware. Also, you can use port binding functions to enforce a WWPN to be connected to a specific SAN switch port.

When switch-port based zoning is used, the ability to allow only specific hosts to connect to an IBM Storage Virtualize cluster is lost.

Consider an NPV device, such as an access gateway that is connected to a fabric. If 14 hosts are attached to that NPV device, and switch port-based zoning is used to zone the switch port for the NPV device to IBM Storage Virtualize system node ports, all 14 hosts can potentially connect to the IBM Storage Virtualize cluster, even if IBM Storage Virtualize is providing storage for only four or five of those hosts.

However, the problem is exacerbated when the IBM Storage Virtualize NPIV feature is used in transitional mode. In this mode, a host can connect to the physical and virtual WWPNs on the cluster. With switch port zoning, this configuration doubles the connection count for each host that is attached to the IBM Storage Virtualize cluster. This issue can affect the function of path failover on the hosts by resulting in too many paths, and the IBM Storage Virtualize Cluster can exceed the maximum host connection count on a large fabric.

If you have the NPIV feature enabled on your IBM Storage Virtualize system, you must use WWPN-based zoning.

Zoning types: Avoid using a zoning configuration that includes a mix of port and WWPN zoning. For NPIV configurations, host zoning must use the WWPN zoning type.

A best practice for traditional zone design calls for *single initiator* zoning, that is, a zone can consist of many target devices, but only one initiator because target devices often wait for an initiator device to connect to them, and initiators actively attempt to connect to each device to which they are zoned. The single initiator approach removes the possibility of a misbehaving initiator affecting other initiators.

The drawback to single initiator zoning is that on a large SAN that features many zones, the SAN administrator's job can be more difficult, and the number of zones on a large SAN can exceed the zone database size limits.

Cisco and Brocade developed features that can reduce the number of zones by allowing the SAN administrator to control which devices in a zone can communicate with other devices in the zone. The features are called Cisco Smart Zoning and Brocade Peer Zoning, which are supported by IBM Storage Virtualize systems.

A brief overview of these features is provided next.

Note: Brocade Traffic Isolation (TI) zoning is deprecated in Brocade Fabric OS 9.0. You can still use TI zoning if you have existing zones, but you must keep at least one switch running a pre-9.0 version of FOS in the fabric to make changes to the TI zones.

Cisco Smart Zoning

Cisco Smart Zoning is a feature that, when enabled, restricts the initiators in a zone to communicate only with target devices in the same zone. For our cluster example, this feature allows a SAN administrator to zone all the host ports for a VMware cluster in the same zone with the storage ports to which all of the hosts need access. Smart Zoning configures the access control lists in the fabric routing table to allow only the initiator (host) ports to communicate with target ports.

For more information about Smart Zoning, see [Smart Zoning](#).

For more information about implementation, see [Implementing Smart Zoning on IBM c-Type and Cisco Switches](#).

Brocade Peer Zoning

Brocade Peer Zoning is a feature that provides a function to restrict what devices can see other devices within the same zone. However, Peer Zoning is implemented such that some devices in the zone are designated as principal devices. The non-principal devices can communicate only with the principal device, and not with each other. As with Cisco, the communication is enforced in the fabric routing table.

For more information, see the section "Peer zoning" in *Modernizing Your IT Infrastructure with IBM b-type Gen 6 Storage Networking and IBM Spectrum Storage Products*, SG24-8415.

Note: Use Smart and Peer Zoning for the host zoning only. Use traditional zoning for intracluster, back-end, and intercluster zoning.

A simple zone for small environments

As an option for small environments, IBM Storage Virtualize based systems support a simple set of zoning rules that enable a small set of host zones to be created for different environments.

For systems with fewer than 64 hosts that are attached, zones that contain host HBAs must contain no more than 40 initiators, including the ports that acts as initiators, such as the IBM Storage Virtualize based system ports that are target + initiator.

Therefore, a valid zone can be 32 host ports plus eight IBM Storage Virtualize based system ports. Include only one port from each node in the I/O groups that are associated with this host.

Note: Do not place more than one HBA port from the same host in the same zone. Also, do not place dissimilar hosts in the same zone. Dissimilar hosts are hosts that are running different operating systems or are different hardware products.

IBM Storage Virtualize Portsets

IBM Storage Virtualize Portsets is a feature to provide effective port management and host access on storage. A group of storage ports can be created that you associate with a specific traffic type, for example, host access, replication, and back-end storage connectivity. IBM Storage Virtualize Portsets provides a way to limit the host logins per I/O port, which improves the resource utilization for performance on storage by load balancing across FC ports on nodes, which removes the skewed I/O workload problem.

IBM Storage Virtualize supports Ethernet and FC portsets for host attachment. Back-end and external storage connectivity, host attachment, and IP replication can be configured by using IP portsets, and host attachment can be configured on FC portsets.

A maximum of 72 (FC + Ethernet) portsets can be created per system.

A host can access the storage only from those IP addresses or FC ports that are configured on a portset and associated with that host.

Every portset is identified by a name. The default portsets are portset0 and portset64 portsets for host attachment. For Ethernet ports connectivity - portset0. And portset64 is for FC ports connectivity host attachment. The default portset - portset3 is for the storage port type by using Ethernet connectivity, as shown in Figure 2-28.

portset0 (Default) Host Attachment	PORT TYPE Ethernet	PORT COUNT 0	MEMBER COUNT 0
portset64 (Default) Host Attachment	PORT TYPE Fibre Channel	PORT COUNT 4	MEMBER COUNT 1
portset1 Remote Copy	PORT TYPE Ethernet	PORT COUNT 0	MEMBER COUNT 0
portset2 Remote Copy	PORT TYPE Ethernet	PORT COUNT 0	MEMBER COUNT 0
portset3 (Default) Storage	PORT TYPE Ethernet	PORT COUNT 0	MEMBER COUNT -

Figure 2-28 IBM Storage Virtualize Portsets overview

A host should always be associated with a correct portset on the storage. If you have a zoned host with different ports that are not part of the same portset, the storage generates an event of a wrong port login (Event ID 064002).

For more information about resolving a blocked FC login event, see [Resolving a problem with a blocked Fibre Channel login](#).

Note:

- ▶ A host can be part of only one portset, but an FC port can be part of multiple portsets.
- ▶ A correct zoning configuration along with a portset is recommended to achieve the recommended number of paths on a host.

Here is an example, say we have two hosts and FlashSystem storage with 8 total ports for fibre channel connectivity and 4 Ethernet ports. Two hosts have fibre channel connectivity and two have Ethernet connectivity and we would like to make sure that those hosts are separated for some reason, due to difference in communication speed, for example.

Therefore, it is feasible to establish two portsets for fiber channel-connected hosts. Initially, it is advisable to create the portsets prior to defining the host objects on the storage system. Afterward, allocate the fiber channel ports to the designated portsets. Note that all eight FC ports can be assigned to a portset; however, if you wish to differentiate communications based on factors such as communication speed or specific fault domains, you can divide the four FC ports into separate portsets. Furthermore, ensure that these ports reside within distinct fabrics for failover purposes. Consequently, each portset should consist of four ports capable of communicating with the host via redundant fabric.

Ethernet ports can also be similarly assigned to their designated portsets.

After portsets are created and ports are assigned, host objects can be defined on the storage, and each host object should be assigned to one portset. Thus, they would be able to communicate with the storage only through the dedicated ports and if communication speeds are different that can improve the reliability of communication and if at some point there is any congestion or other communication issues, this segregation enhances the ability to investigate and find the culprit faster.

IBM Storage Virtualize Portsets can be used for effective workload distribution, proper failure domain segmentation and host-based functional grouping (FC_SCSI, NVMeoF, and performance). Figure 2-29 and Figure 2-30 show an example of configuration using the portsets. Figure 2-29 on page 153 shows initial configuration with fibre channel ports and list of portsets on the system. Notice the portset its0V860 that was added to the list.

```

IBM_FlashSystem:FS9xx0:superuser>lsportfc
id fc_io_port_id port_id type port_speed node_id node_name WWPN nportid status attachment cluster_use
adapter_location adapter_port_id
0 1 1 fc 16Gb 5 node1 5005076810110214 080100 active switch
local_partner 1 1
1 2 2 fc 16Gb 5 node1 5005076810120214 080000 active switch
local_partner 1 2
2 3 3 fc 16Gb 5 node1 5005076810130214 120100 active switch
local_partner 1 3
3 4 4 fc 16Gb 5 node1 5005076810140214 120000 active switch
local_partner 1 4
24 1 1 fc 16Gb 2 node2 5005076810110216 080300 active switch
local_partner 1 1
25 2 2 fc 16Gb 2 node2 5005076810120216 120200 active switch
local_partner 1 2
26 3 3 fc 16Gb 2 node2 5005076810130216 120300 active switch
local_partner 1 3
27 4 4 fc 16Gb 2 node2 5005076810140216 080200 active switch
local_partner 1 4
IBM_FlashSystem:FS9xx0:superuser>lsportset
id name type port_count host_count lossless owner_id owner_name port_type is_default
0 portset0 host 0 0 ethernet yes
1 portset1 replication 0 0 ethernet no
2 portset2 replication 0 0 ethernet no
3 portset3 storage 0 0 ethernet no
4 itsoV860 host 0 0 yes fc no
64 portset64 host 4 2 yes fc yes

```

Figure 2-29 Listing the available ports and portsets.

Figure 2-30 shows how to assign ports pairs with `fc_io_port_id 1` and `fc_io_port_id 3` to `itsoV840` portset group. It also lists those specific pairs.

```

IBM_FlashSystem:FS9xx0:superuser>addfcportsetmember -portset itsoV860 -fcioportid 1&&addfcportsetmember -portset
itsoV860 -fcioportid 3
IBM_FlashSystem:FS9xx0:superuser>lsportset
id name type port_count host_count lossless owner_id owner_name port_type is_default
0 portset0 host 0 0 ethernet yes
1 portset1 replication 0 0 ethernet no
2 portset2 replication 0 0 ethernet no
3 portset3 storage 0 0 ethernet no
4 itsoV860 host 2 0 yes fc no
64 portset64 host 4 2 yes fc yes
IBM_FlashSystem:FS9xx0:superuser>lstargetportfc -filtervalue "port_id=1"&&lstargetportfc -filtervalue "port_id=3"
id WWPN WWNN port_id owning_node_id current_node_id nportid host_io_permitted virtualized
protocol fc_io_port_id portset_count host_count active_login_count
1 5005076810110214 5005076810000214 1 5 5 080100 no no
scsi 1 0 0 1
2 5005076810150214 5005076810000214 1 5 5 080101 yes yes
scsi 1 2 2
3 5005076810190214 5005076810000214 1 5 5 080102 yes yes
nvme 1 2 0
73 5005076810110216 5005076810000216 1 2 2 080300 no no
scsi 1 0 0 2
74 5005076810150216 5005076810000216 1 2 2 080301 yes yes
scsi 1 2 2 0
75 5005076810190216 5005076810000216 1 2 2 080302 yes yes
nvme 1 2 0
id WWPN WWNN port_id owning_node_id current_node_id nportid host_io_permitted virtualized
protocol fc_io_port_id portset_count host_count active_login_count
7 5005076810130214 5005076810000214 3 5 5 120100 no no
scsi 3 0 0 1
8 5005076810170214 5005076810000214 3 5 5 120101 yes yes
scsi 3 2 2
9 5005076810180214 5005076810000214 3 5 5 120102 yes yes
nvme 3 2 0
79 5005076810130216 5005076810000216 3 2 2 120300 no no
scsi 3 0 0 2
80 5005076810170216 5005076810000216 3 2 2 120301 yes yes
scsi 3 2 2
81 5005076810180216 5005076810000216 3 2 2 120302 yes yes
nvme 3 2 0

```

Figure 2-30 How to assign ports pairs

While the portsets provide additional layer for keeping the configuration clean and ordered as well as improves conditions for maintenance it is still important to plan it carefully taking into account the overall zoning.

2.4.2 Prezoning tips and shortcuts

In this section, we describe several tips and shortcuts that are available for IBM Storage Virtualize systems zoning.

Naming convention and zoning scheme

When you create and maintaining an IBM Storage Virtualize system zoning configuration, you must have a defined naming convention and zoning scheme. If you do not define a naming convention and zoning scheme, your zoning configuration can be difficult to understand and maintain.

Environments have different requirements, which means that the level of detail in the zoning scheme varies among environments of various sizes. Therefore, ensure that you have an easily understandable scheme with an appropriate level of detail. Then, make sure that you use it consistently and adhere to it whenever you change the environment.

For more information about IBM Storage Virtualize system naming conventions, see 10.14.1, “Naming conventions” on page 689.

Aliases

Use zoning aliases when you create your IBM Storage Virtualize system zones if they are available on your specific type of SAN switch. Zoning aliases makes your zoning easier to configure and understand, and causes fewer possibilities for errors. Table 2-10 shows some alias name examples.

Table 2-10 Alias names examples

Port or WWPN	Use	Alias
Card 1 Port 1 physical WWPN	External Storage back-end	FSx_N1P1_STORAGE
Card 1 Port 1 NPIV WWPN	Host attachment	FSx_N1P1_HOST_NPIV
Card 1 Port 2 physical WWPN	External Storage back-end	FSx_N1P2_STORAGE
Card 1 Port 2 NPIV WWPN	Host attachment	FSx_N1P2_HOST_NPIV
Card 1 Port 3 physical WWPN	Inter-node traffic	FSx_N1P3_CLUSTER
Card 1 Port 3 NPIV WWPN	No use	No alias
Card 1 Port 4 physical WWPN	Inter-node traffic	FSx_N1P4_CLUSTER
Card 1 Port 4 NPIV WWPN	No use	No alias
Card 2 Port 3 physical WWPN	Replication traffic	FSx_N1P7_REPLICATION
Card 2 Port 3 NPIV WWPN	No use	No alias
Card 2 Port 4 physical WWPN	Replication traffic	FSx_N1P8_REPLICATION
Card 2 Port 4 NPIV WWPN	No use	No alias

Note: In Table 2-10, not all ports have an example for aliases. NPIV ports can be used for host attachment only, as shown in Figure 2-31. If you are using external virtualized back-ends, use the physical port WWPN. For replication and inter-node, use the physical WWPN. In the alias examples that are listed in Table 2-10, the *N* is for node, and all examples are from node 1. An N2 example is FSx_N2P4_CLUSTER. The x represents the model of your IBM Storage Virtualize system, for example, SVC or IBM FlashSystem 9200 or 9500, IBM FlashSystem 7300, or IBM FlashSystem 5015 or 5035.

```

IBM FlashSystem:FS9110:superuser>lstargetportfc
id  WWPN          WWNN          port_id  owning_node_id  current_node_id  nportid  host_io_permitted  virtualized  protocol
1  5005076810110214  5005076810000214  1        5                5           080100  no             no         scsi
2  5005076810150214  5005076810000214  1        5                5           080101  yes            yes        scsi
3  5005076810190214  5005076810000214  1        5                5           080102  yes            yes        nvme
4  5005076810120214  5005076810000214  2        5                5           080000  no             no         scsi
5  5005076810160214  5005076810000214  2        5                5           080001  yes            yes        scsi
6  50050768101A0214  5005076810000214  2        5                5           080002  yes            yes        nvme
7  5005076810130214  5005076810000214  3        5                5           120100  no             no         scsi
8  50050768101C0214  5005076810000214  3        5                5           120101  yes            yes        scsi
9  50050768101B0214  5005076810000214  3        5                5           120102  yes            yes        nvme
10 50050768101D0214  5005076810000214  4        5                5           120000  no             no         scsi
11 50050768101E0214  5005076810000214  4        5                5           120001  yes            yes        scsi
12 50050768101C0214  5005076810000214  4        5                5           120002  yes            yes        nvme
13 5005076810110216  5005076810000216  1        4                4           080200  no             no         scsi

```

Figure 2-31 Output of the IBM Storage Virtualize `lstargetportfc` command

One approach is to include multiple members in one alias because zoning aliases can normally contain multiple members (similar to zones). This approach can help avoid some common issues that are related to zoning and make it easier to maintain the port balance in a SAN.

Create the following zone aliases:

- ▶ One zone alias for each IBM Storage Virtualize system port.
- ▶ One alias for each host initiator.
- ▶ One host initiator alias to IBM Storage Virtualize port 1 from node 1, and to port 1 from node 2. Then, name this zone HOST1_HBA1_T1_FSx.
- ▶ Zone an alias group for each storage subsystem port pair (the IBM Storage Virtualize system must reach the same storage ports on both I/O group nodes).

By creating template zones, you keep the number of paths on the host side to four for each volume and a good workload balance among the IBM Storage virtualize ports. Table 2-11 shows how the aliases are distributed if you create template zones as described in the example.

Table 2-11 Distribution of aliases

Template	IBM FlashSystem 9200 ports on Fabric A	IBM FlashSystem 9200 ports on Fabric B
T1	Node 1 port 1 Node 2 port 1	Node 1 port 2 Node 2 port 2
T2	Node 1 port 3 Node 2 port 3	Node 1 port 4 Node 2 port 4
T3	Node 1 port 5 Node 2 port 5	Node 1 port 6 Node 2 port 6
T4	Node 1 port 7 Node 2 port 7	Node 1 port 8 Node 2 port 8

2.4.3 IBM Storage Virtualize internode communications zones

Internode (or intra-cluster) communication is critical to the stable operation of the cluster. The ports that carry internode traffic are used for mirroring write cache and metadata exchange between nodes and canisters.

To establish efficient, redundant, and resilient intracluster communication, the intracluster zone must contain at least two ports from each node or canister. For IBM Storage Virtualize nodes with eight ports or more, isolate the intracluster traffic by dedicating node ports specifically to internode communication.

The ports to be used for intracluster communication varies according to the machine type and model number and port count. For more information about port assignment recommendations, see Figure 2-26 on page 146. Use the port from different adapters of each node to achieve maximum redundancy while creating zones for node-to-node communication.

NPIV configurations: On NPIV-enabled configurations, use the physical WWPN for the intracluster zoning.

Only 16-port logins are allowed from one node to any other node in a SAN fabric. Ensure that you apply the correct port masking to restrict the number of port logins. Without port masking, any IBM Storage Virtualize system port and any member of the same zone can be used for intracluster communication, even the port members of an IBM Storage Virtualize system to a host and an IBM Storage Virtualize system to storage zoning.

Inter-node communication is supported on 25 GbE by using iSCSI (RoCE or iWARP). Check the network stability (congestion, packet drop, latency, and maximum transmission unit (MTU)) before configuring node-to-node communication on Ethernet.

As a best practice, use FlashCopy mapping between a source and target on the same node or I/O group to minimize cross-I/O-group communication. The same best practice applies to change volumes (CVs) in Global Mirror with Change Volumes (GMCV) replication.

High availability (HA) solutions such as ESC and HyperSwap rely on node-to-node communication between I/O group or nodes. In such scenarios, dedicate enough ports for node-to-node and intracluster communication, which is used for metadata exchange and mirroring write cache.

Configure a private SAN along with port masking over dark fibre links for HA (ESC or HyperSwap) configurations to achieve maximum performance.

If a link or ISL between sites is not stable in a HA solution (for example, ESC or HyperSwap), this instability makes the node-to-node communication and cluster unstable.

Note: To check whether the login limit is exceeded, count the number of distinct ways by which a port on node X can log in to a port on node Y. This number must not exceed 16. For more information about port masking, see Chapter 8, “Hosts” on page 519.

2.4.4 IBM Storage Virtualize storage zones

The zoning between an IBM Storage Virtualize system and other storage is necessary to allow the virtualization of any storage space under the IBM Storage Virtualize system. This storage is referred to as *back-end storage* or external storage.

A zone for each back-end storage to each IBM Storage Virtualize system node or canister must be created in both fabrics, as shown in Figure 2-32. Doing so reduces the overhead that is associated with many logins. The ports from the storage subsystem must be split evenly across the dual fabrics.

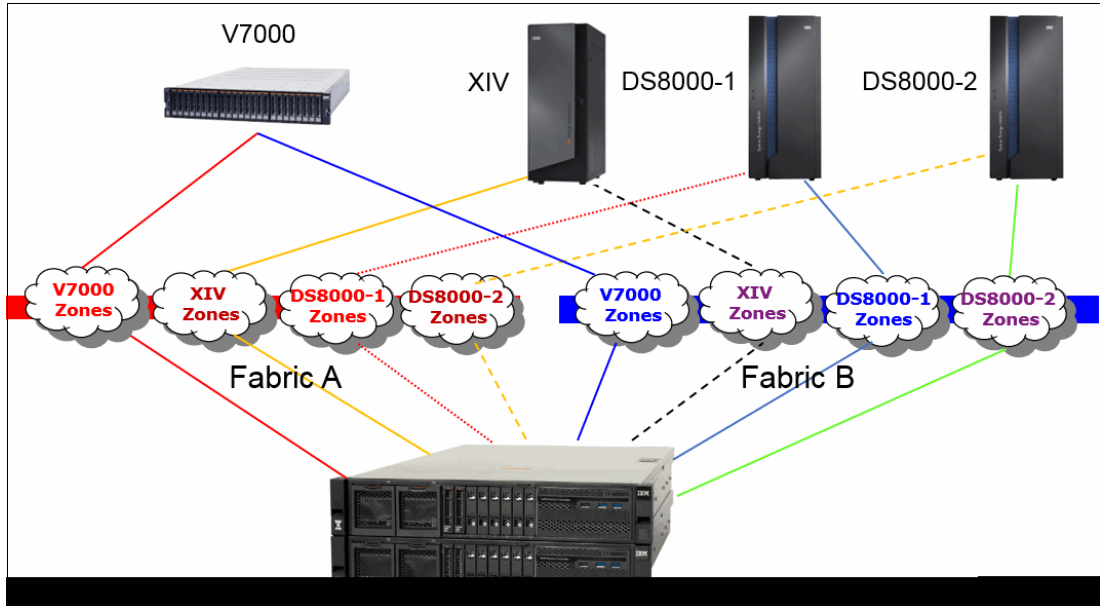


Figure 2-32 Back-end storage zoning

Often, all nodes or canisters in an IBM Storage Virtualize system should be zoned to the same ports on each back-end storage system, with the following exceptions:

- ▶ When implementing ESC or HyperSwap configurations where the back-end zoning can be different for the nodes or canisters according to the site definition (for more information, see *IBM Spectrum Virtualize and SAN Volume Controller Enhanced Stretched Cluster with VMware*, SG24-8211 and *IBM Storwize V7000, Spectrum Virtualize, HyperSwap, and VMware Implementation*, SG24-8317).
- ▶ When the SAN has a multi-core design that requires special zoning considerations, as described in “Zoning to storage best practice” on page 158.

Important: Consider the following points:

- ▶ On NPIV-enabled systems, use the physical WWPN for the zoning to the back-end controller.
- ▶ If a back-end controller is an IBM Storage Virtualize appliance, such as an IBM FlashSystem 7200, 9200, 9500, or 7300 (storage layer) that is used by SVC (replication layer), then replication layer primary ports are zoned to NPIV target ports on the storage layer device.

When two nodes or canisters are zoned to different set of ports for the same storage system, the IBM Storage Virtualize operation mode is considered degraded. Then, the system logs errors that request a repair action. This situation can occur if incorrect zoning is applied to the fabric.

Figure 2-33 shows a zoning example (that uses generic aliases) between a 2-node SVC and an IBM Storwize V5000. Both SVC nodes can access the same set of Storwize V5000 ports.

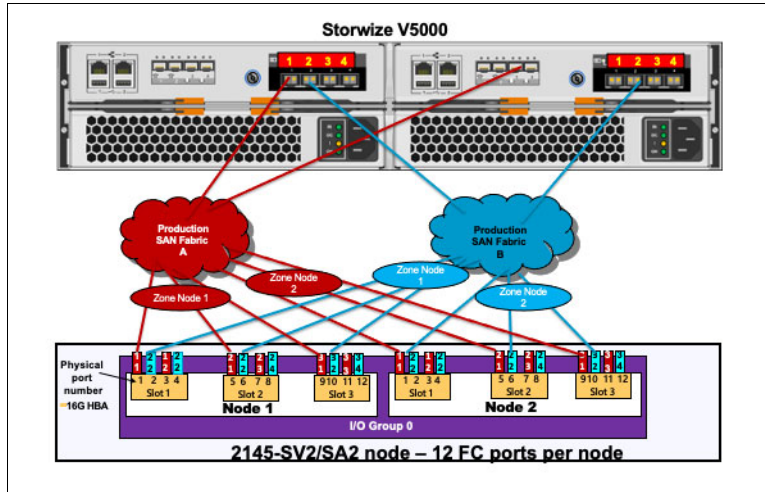


Figure 2-33 V5000 zoning

Each storage controller or model has its own zoning and port placement best practices. The generic guideline for all storage is to use the ports that are distributed between the redundant storage components, such as nodes, controllers, canisters, and FC adapters (respecting the port count limit, as described in “Back-end storage port count” on page 161).

The following sections describe the IBM Storage specific zoning guidelines. Storage vendors other than IBM might have similar best practices. For more information, contact your vendor.

Zoning to storage best practice

In 2.1.2, “ISL considerations” on page 125, we describe ISL considerations for ensuring that the IBM Storage Virtualize system is connected to the same physical switches as the back-end storage ports.

For more information about SAN design options, see 2.2, “SAN topology-specific guidelines” on page 127.

This section describes best practices for zoning IBM Storage Virtualize system ports to controller ports on each of the different SAN designs.

The high-level best practice is to configure zoning such that the SVC and IBM FlashSystem ports are zoned only to the controller ports that are attached to the same switch. For single-core designed fabrics, this practice is not an issue because only one switch is used on each fabric to which the SVC, IBM FlashSystem, and controller ports are connected. For the mesh and dual-core and other designs in which the IBM Storage Virtualize system is connected to multiple switches in the same fabric, zoning might become an issue.

Figure 2-34 shows the best practice zoning on a dual-core fabric. You can see that two zones are used:

- ▶ Zone 1 includes only the IBM Storage Virtualize system and back-end ports that are attached to the core switch on the left.
- ▶ Zone 2 includes only the IBM Storage Virtualize system and back-end ports that are attached to the core switch on the right.

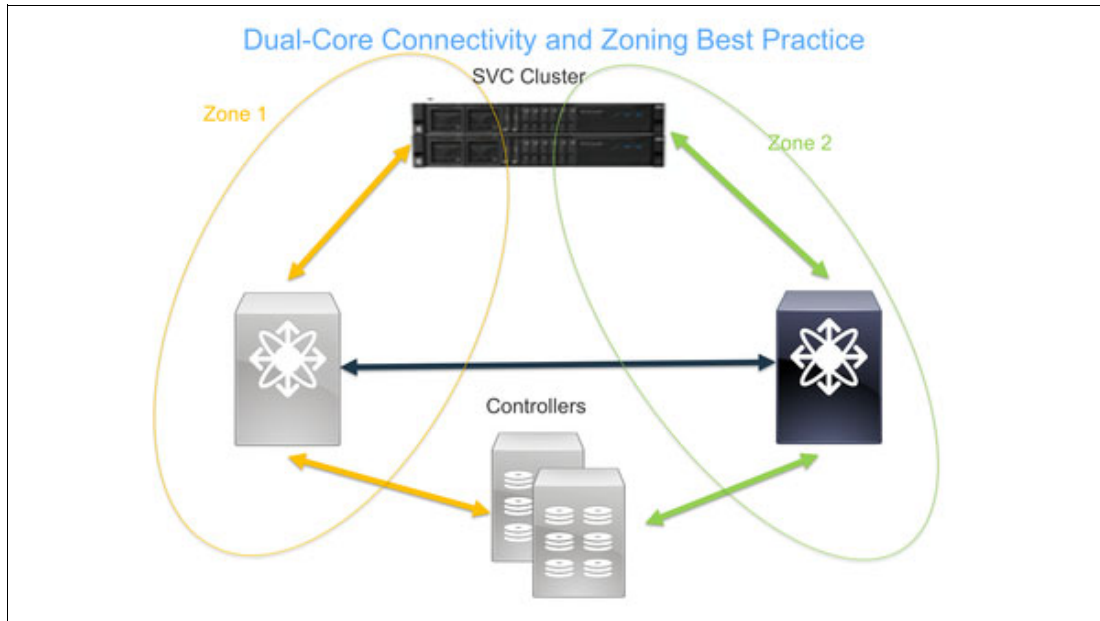


Figure 2-34 Dual core zoning schema

Mesh fabric designs that have the IBM Storage Virtualize and controller ports that are connected to multiple switches follow the same general guidelines. Failure to follow this best practice might result in IBM Storage Virtualize system performance impacts to the fabric.

Real-life potential effect of deviation from best practice zoning

Figure 2-35 shows a design that consists of a dual-core Brocade fabric with the SVC cluster that is attached to one switch and controllers that are attached to the other switch. An IBM General Parallel File System (GPFS) cluster is attached to the same switch as the controllers. This real-world design was used for a customer that was experiencing extreme performance problems on its SAN. The customer had dual fabrics, and each fabric had this same flawed design.

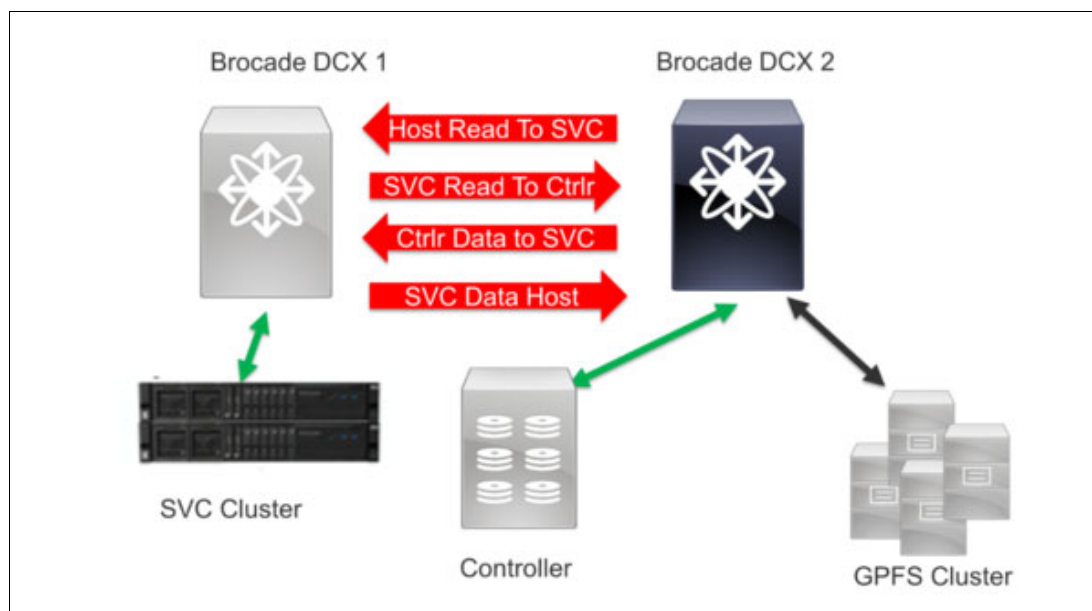


Figure 2-35 ISL traffic overloading

The design violates the best practices of ensuring that the IBM Storage Virtualize system and storage ports are connected to the same switches, and zoning the ports, as shown in Figure 2-34 on page 159. It also violates the best practice of connecting the host ports (the GPFS cluster) to the same switches as the IBM Storage Virtualize system where possible.

This design creates an issue with traffic that is traversing the ISL unnecessarily, as shown in Figure 2-35. I/O requests from the GPFS cluster must traverse the ISL four times. This design must be corrected such that the IBM Storage Virtualize system, controller, and GPFS cluster ports are all connected to both core switches, and zoning is updated to be in accordance with the example that is shown in Figure 2-34 on page 159.

Figure 2-35 shows a real-world customer SAN design. The effect of the extra traffic on the ISL between the core switches from this design caused significant delays in command response time from the GPFS cluster to the SVC or IBM FlashSystem and from the SVC to the controller.

The SVC or IBM FlashSystem cluster also logged nearly constant errors against the controller, including disconnecting from controller ports. The SAN switches logged frequent link timeouts and frame drops on the ISL between the switches. Finally, the customer had other devices sharing the ISL that were not zoned to the SVC or IBM FlashSystem. These devices also were affected.

Back-end storage port count

The current firmware that is available (version 8.6 at the time of writing) sets limit of 1024 worldwide node names (WWNNs) per SVC or IBM FlashSystem cluster and up to 1024 WWPNS. The rule is that each port represents a WWPN count on the IBM Storage Virtualize cluster or I/O group. However, the WWNN count differs based on the type of storage.

For example, at the time of writing, EMC DMX/Symmetrix, all Hitachi Data Systems (HDS) storage, and SUN/HP use one WWNN per port. This configuration means that each port appears as a separate controller to the SVC or IBM FlashSystem. Therefore, each port that is connected to the IBM Storage Virtualize system means one WWPN and a WWNN increment.

IBM storage and EMC Clariion/VNX use one WWNN per storage subsystem, so each appears as a single controller with multiple port WWPNS.

A best practice is to assign up to 16 ports from each back-end storage to the SVC or IBM FlashSystem cluster. The reason for this limitation is that since version 8.5, the maximum number of ports that is recognized by the IBM Storage Virtualize system per each WWNN is 16. The more ports that are assigned, the more throughput is obtained.

In a situation where the back-end storage has hosts direct-attached, do not mix the host ports with the IBM Storage Virtualize system ports. The back-end storage ports must be dedicated to the IBM Storage Virtualize system. Therefore, sharing storage ports are functional only during migration and for a limited time. However, if you intend to have some hosts that are permanently directly attached to the back-end storage, you must separate the IBM Storage Virtualize system ports from the host ports.

IBM XIV storage subsystem

IBM XIV® storage is modular storage and available as fully or partially populated configurations. An XIV hardware configuration can include 6 - 15 modules. Each extra module that is added to the configuration increases the XIV capacity, CPU, memory, and connectivity.

From a connectivity standpoint, four FC ports are available in each interface module for a total of 24 FC ports in a fully configured XIV system. The XIV modules with FC interfaces are present on modules 4 - 9. Partial rack configurations do not use all ports, even though they might be physically present.

Table 2-12 lists the XIV port connectivity according to the number of installed modules.

Table 2-12 XIV connectivity ports as capacity grows

XIV modules	Total ports	Port interfaces	Active port modules
6	8	2	4 and 5
9	16	4	4, 5, 7, and 8
10	16	4	4, 5, 7, and 8
11	20	5	4, 5, 7, 8, and 9
12	20	5	4, 5, 7, 8, and 9
13	24	6	4, 5, 6, 7, 8, and 9
14	24	6	4, 5, 6, 7, 8, and 9
15	24	6	4, 5, 6, 7, 8, and 9

Note: If the XIV includes the capacity on demand (CoD) feature, all active FC interface ports are usable at the time of installation regardless of how much usable capacity you purchased. For example, if a 9-module system is delivered with six modules active, you can use the interface ports in modules 4, 5, 7, and 8, although effectively three of the nine modules are not yet activated through CoD.

To use the combined capabilities of SVC, IBM FlashSystem, and XIV, you must connect two ports (one per fabric) from each interface module with the SVC or IBM FlashSystem ports.

For redundancy and resiliency purposes, select one port from each HBA that is present on the interface modules. Use port 1 and 3 because both ports are on different HBAs. By default, port 4 is set as a SCSI initiator and dedicated to XIV replication.

Therefore, if you decide to use port 4 to connect to an SVC or IBM FlashSystem, you must change its configuration from initiator to target. For more information, see *IBM XIV Storage System Architecture and Implementation*, SG24-7659.

Figure 2-36 shows how to connect an XIV frame to an SVC storage controller.

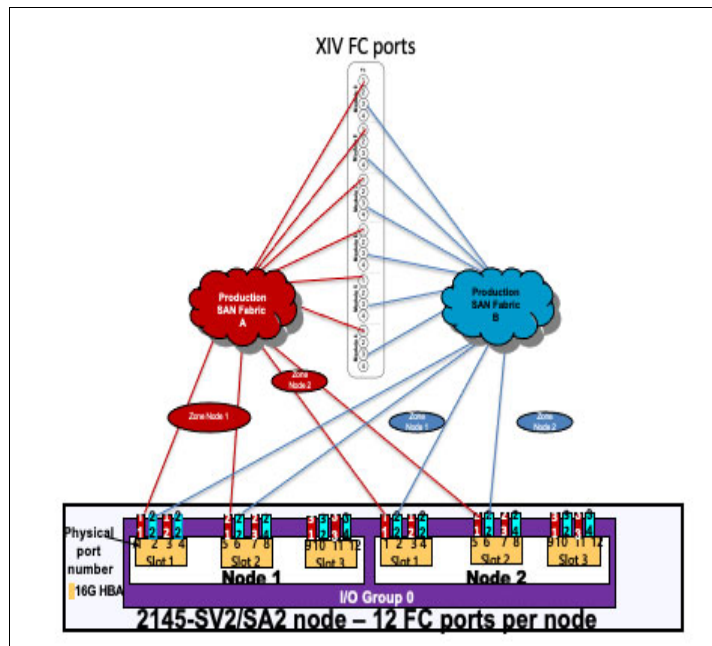


Figure 2-36 XIV port cabling

A best practice for zoning is to create a single zoning to each SVC or IBM FlashSystem node on each SAN fabric. This zone must contain all ports from a single XIV and the SVC or IBM FlashSystem node ports that are destined to connect host and back-end storage. All nodes in an SVC or IBM FlashSystem cluster must see the same set of XIV host ports.

Figure 2-36 shows that a single zone is used for each XIV to SVC or IBM FlashSystem node. For this example, the following zones are used:

- ▶ Fabric A, XIV → SVC Node 1: All XIV fabric A ports to SVC node 1
- ▶ Fabric A, XIV → SVC Node 2: All XIV fabric A ports to SVC node 2
- ▶ Fabric B, XIV → SVC Node 1: All XIV fabric B ports to SVC node 1
- ▶ Fabric B, XIV → SVC Node 2: All XIV fabric B ports to SVC node 2

For more information about other best practices and XIV considerations, see Chapter 3, “Storage back-end” on page 191.

IBM FlashSystem A9000 and A9000R storage systems

An IBM FlashSystem A9000 system has a fixed configuration with three grid elements, with a total of 12 FC ports. A best practice is to restrict ports 2 and 4 of each grid controller for replication and migration use, and use ports 1 and 3 for host access.

However, considering that any replication or migration is done through IBM Storage Virtualize, ports 2 and 4 also can be used for IBM Storage Virtualize connectivity. Port 4 must be set to target mode for replication or migration to work.

Assuming a dual fabric configuration for redundancy and resiliency purposes, select one port from each HBA that is present on the grid controller. Therefore, a total of six ports (three per fabric) are used.

Figure 2-37 shows a possible connectivity scheme for IBM SAN Volume Controller 2145-SV2/SA2 nodes and IBM FlashSystem A9000 systems.

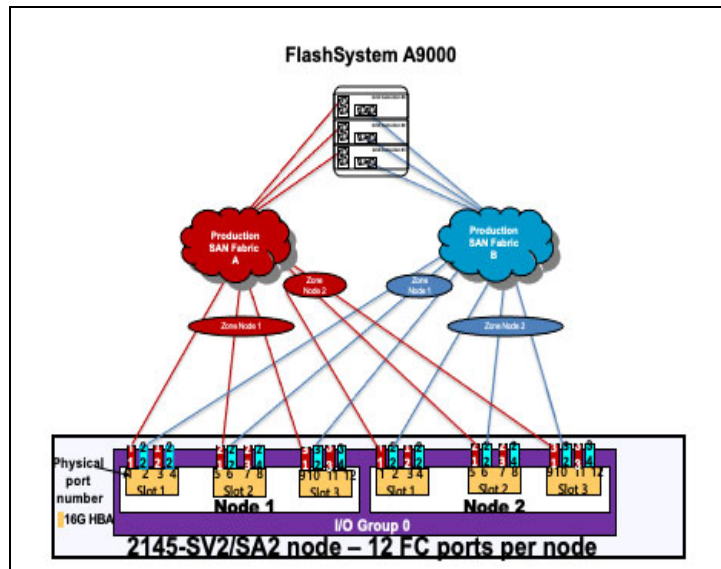


Figure 2-37 IBM FlashSystem A9000 connectivity

The IBM FlashSystem A9000R system has more choices because many configurations are available, as listed in Table 2-13.

Table 2-13 Number of host ports in an IBM FlashSystem A9000R system

Grid elements	Total available host ports
2	8
3	12
4	16
5	20
6	24

However, IBM Storage Virtualize can support only 16 WWPN from any single WWNN. The IBM FlashSystem A9000 or IBM FlashSystem A9000R system has only one WWNN, so you are limited to 16 ports to any IBM FlashSystem A9000R system.

Table 2-14 shows the same table, but with columns added to show how many and which ports can be used for connectivity. The assumption is a dual fabric, with ports 1 in one fabric, and ports 3 in the other.

Table 2-14 Host connections to SAN Volume Controller

Grid elements	Total host ports available	Total ports that are connected to IBM Storage Virtualize	Total ports that are connected to IBM Storage Virtualize
2	8	8	All controllers, ports 1 and 3
3	12	12	All controllers, ports 1 and 3
4	16	8	Odd controllers, port 1 Even controllers, port 3
5	20	10	Odd controllers, port 1 Even controllers, port 3
6	24	12	Odd controllers, port 1 Even controllers, port 3

For the 4-grid element system, it is possible to attach 16 ports because that is the maximum that IBM Storage Virtualize allows. For 5- and 6-grid element systems, it is possible to use more ports up to the 16 maximum; however, that configuration is *not* recommended because it might create unbalanced workloads to the grid controllers with two ports attached.

Figure 2-38 shows a possible connectivity scheme for SVC 2145-SV2/SA2 nodes and IBM FlashSystem A9000R systems with up to three grid elements.

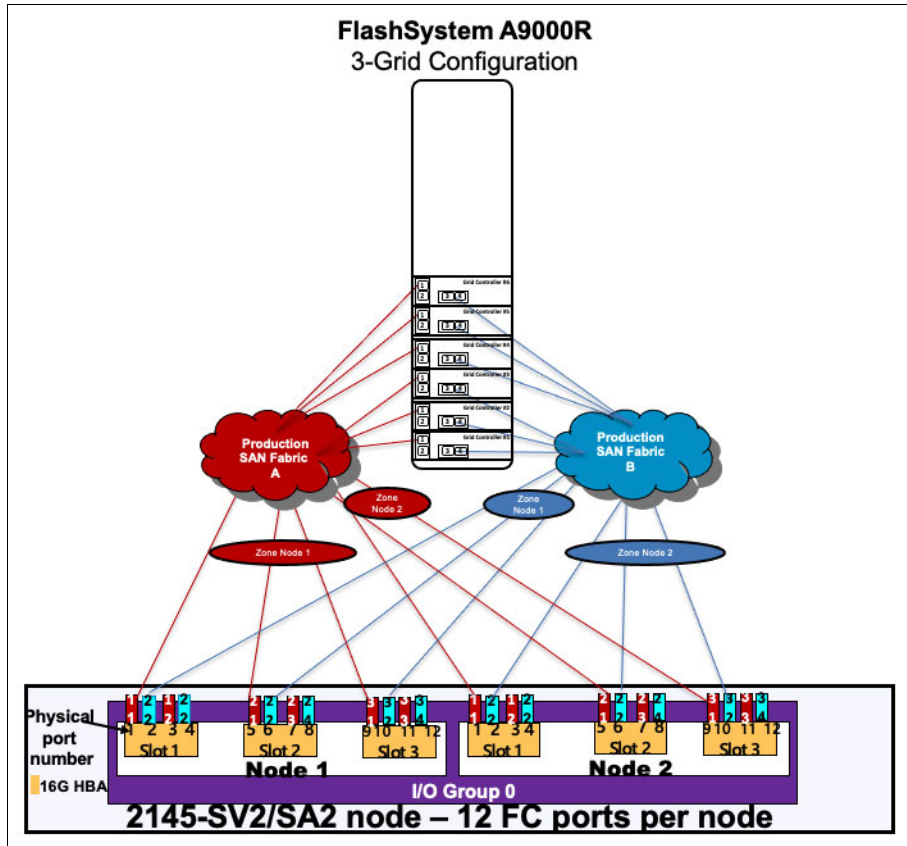


Figure 2-38 IBM FlashSystem A9000 grid configuration cabling

Figure 2-39 shows a possible connectivity schema for SVC 2145-SV2/SA2 nodes and IBM FlashSystem A9000R systems fully configured.

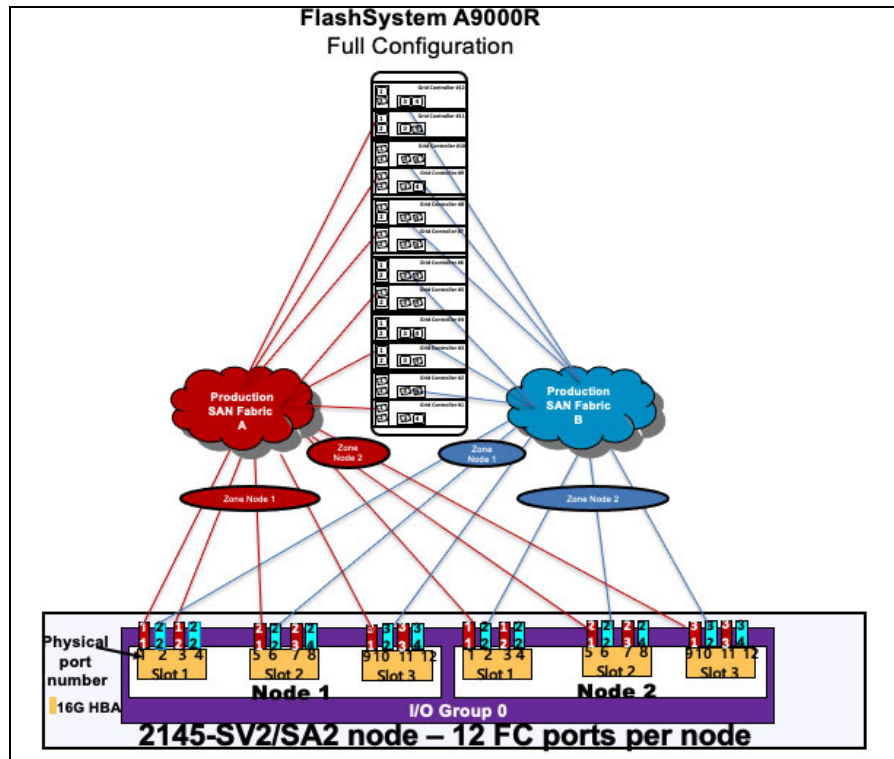


Figure 2-39 Connecting IBM FlashSystem A9000 fully configured as a back-end controller

For more information about IBM FlashSystem A9000 and A9000R implementation, see *IBM FlashSystem A9000 and A9000R Architecture and Implementation (Version 12.3.2)*, SG24-8345.

IBM Storage Virtualize storage subsystem

IBM Storage Virtualize external storage systems can present volumes to another IBM Storage Virtualize system. If you want to virtualize one IBM Storage Virtualize system by using another IBM Storage Virtualize system, change the *layer* of the IBM Storage Virtualize system to be used as the virtualizer. By default, SVC includes the layer of *replication*; IBM Storage Virtualize includes the layer of *storage*.

Volumes that form the storage layer can be presented to the replication layer and are seen on the replication layer as managed disks (MDisks), but not vice versa. That is, the storage layer cannot see a replication layer's MDisks.

The SVC layer of replication cannot be changed, so you cannot virtualize SVC behind IBM Storage Virtualize. However, IBM FlashSystem can be changed from storage to replication and from replication to storage layer.

If you want to virtualize one IBM FlashSystem behind another one, the IBM FlashSystem that is used as external storage must have a layer of storage; the IBM FlashSystem that is performing virtualization must have a layer of replication.

The storage layer and replication layer have the following differences:

- ▶ In the *storage layer*, an IBM Storage Virtualize family system has the following characteristics and requirements:
 - The system can complete Metro Mirror (MM) and Global Mirror (GM) replication with other storage layer systems.
 - The system can provide external storage for replication layer systems or SVC.
 - The system cannot use another IBM Storage Virtualize family system that is configured with the storage layer as external storage.
- ▶ In the *replication layer*, an IBM Storage Virtualize family system has the following characteristics and requirements:
 - The system can complete MM and GM replication with other replication layer systems or SVC.
 - The system cannot provide external storage for a replication layer system or SVC.
 - The system can use another IBM FlashSystem family system that is configured with a storage layer as external storage.

Note: To change the layer, you must disable the visibility of every other IBM FlashSystem or SVC on all fabrics. This process involves deleting partnerships, remote copy relationships, and zoning between IBM FlashSystem and other IBM FlashSystem or SVC. Then, run the `chsystem -layer` command to set the layer of the system.

For more information, see [System layers](#).

To zone the IBM FlashSystem as a back-end storage controller of SVC, every SVC node must access the same IBM FlashSystem ports as a minimum requirement. Create one zone per SVC node per fabric to the same ports from an IBM FlashSystem storage.

Figure 2-40 shows a zone between a 16-port IBM Storwize or IBM FlashSystem and an SVC.

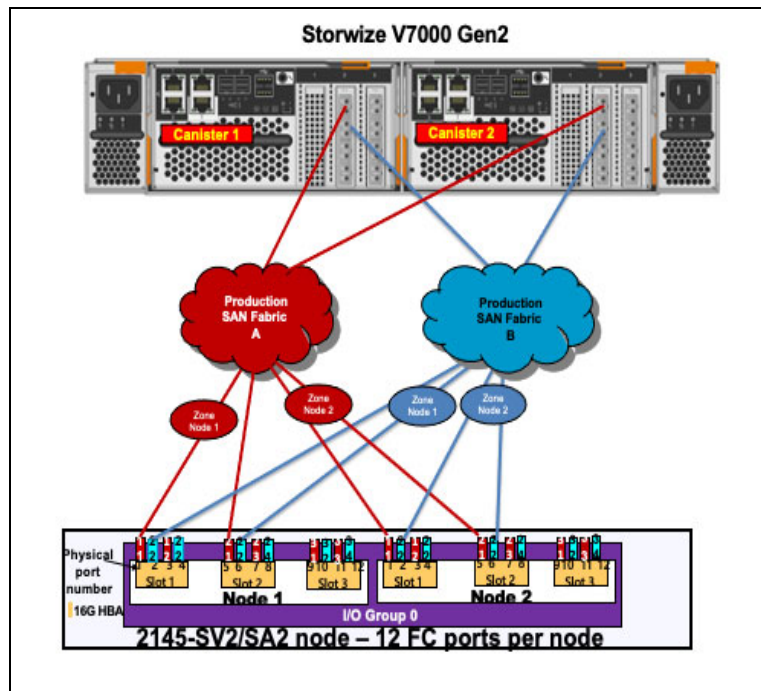


Figure 2-40 V7000 connected as a back-end controller

The ports from the Storwize V7000 system in Figure 2-40 on page 167 are split between both fabrics. The odd ports are connected to Fabric A and the even ports are connected to Fabric B. You also can spread the traffic across the IBM Storwize V7000 FC adapters on the same canister.

However, it does not significantly increase the availability of the solution because the mean time between failures (MTBF) of the adapters is not significantly less than that of the non-redundant canister components.

Note: If you use an NPIV-enabled IBM FlashSystem system as back-end storage, only the NPIV ports on the IBM FlashSystem system must be used for the storage back-end zoning.

Connect as many ports as necessary to service your workload to the SVC. For more information about back-end port limitations and best practices, see “Back-end storage port count” on page 161.

Considering the IBM Storage Virtualize family configuration, the configuration is the same for new IBM FlashSystem systems (see Figure 2-41, which shows an IBM FlashSystem 9100 as an SVC back-end zone example).

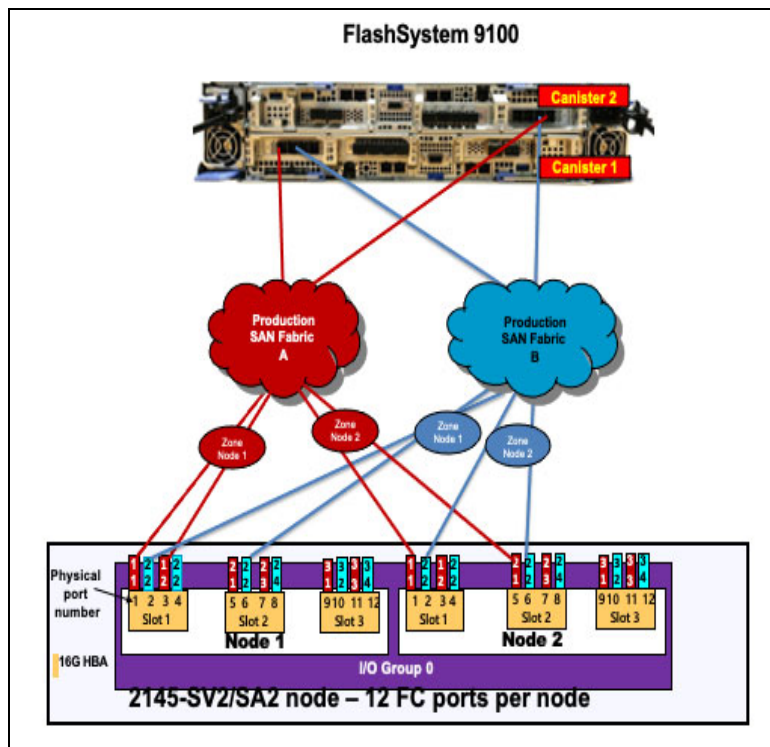


Figure 2-41 IBM FlashSystem 9100 as a back-end controller

Connect as many ports as necessary to service your workload to the SVC. For more information about back-end port limitations and best practices, see “Back-end storage port count” on page 161.

IBM FlashSystem 900

IBM FlashSystem 900 is an all-flash storage array that provides extreme performance and can sustain highly demanding throughput and low latency across its FC interfaces. It includes up to 16 ports of 8 Gbps or eight ports of 16 Gbps FC. It also provides enterprise-class reliability, large capacity, and green data center power and cooling requirements.

The main advantage of integrating IBM FlashSystem 900 with IBM Storage Virtualize is to combine the extreme performance of IBM FlashSystem with the IBM Storage Virtualize enterprise-class solution with such features as tiering, mirroring, IBM FlashCopy, thin provisioning, IBM Real-time Compression (RtC), and copy services.

Before starting, work closely with your IBM Sales, pre-sales, and IT architect to correctly size the solution by defining the suitable number of IBM FlashSystem I/O groups or clusters and FC ports that are necessary according to your servers and application workload demands.

To maximize the performance that you can achieve when deploying the IBM FlashSystem 900 with IBM Storage Virtualize, carefully consider the assignment and usage of the FC HBA ports on IBM Storage Virtualize, as described in 2.3.2, “IBM FlashSystem 9200 and 9500 controller ports” on page 136. The IBM FlashSystem 900 ports must be dedicated to the IBM Storage Virtualize workload, so do *not* mix direct-attached hosts on IBM FlashSystem 900 with IBM Storage Virtualize ports.

Connect IBM FlashSystem 900 to the SAN network by completing the following steps:

1. Connect the IBM FlashSystem 900 odd-numbered ports to the odd-numbered SAN fabric (or SAN fabric A) and the even-numbered ports to the even-numbered SAN fabric (or SAN fabric B).
2. Create one zone for each IBM Storage Virtualize node with all IBM FlashSystem 900 ports on each fabric.

Figure 2-42 shows a 16-port IBM FlashSystem 900 zoning to an SVC.

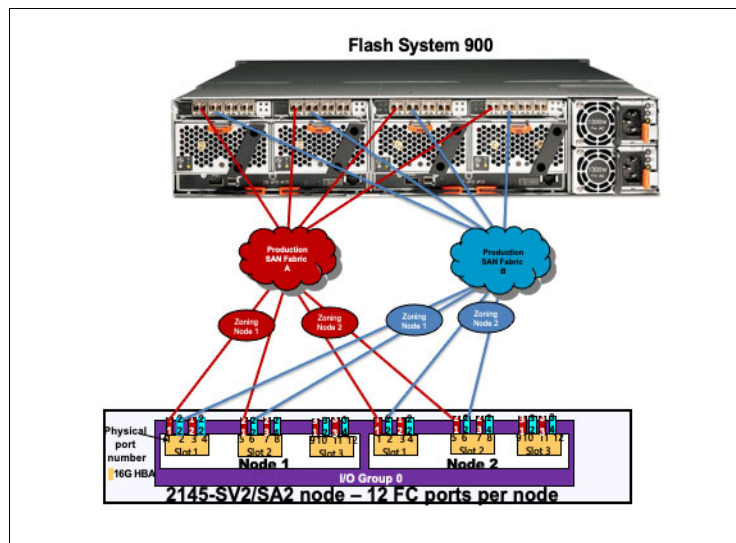


Figure 2-42 IBM FlashSystem 900 connectivity to a SAN Volume Controller cluster

After the IBM FlashSystem 900 is zoned to two SVC nodes, four zones exist with one zone per node and two zones per fabric.

You can decide to share or not the IBM Storage Virtualize ports with other back-end storage. However, it is important to monitor the buffer credit usage on IBM Storage Virtualize switch ports and, if necessary, modify the buffer credit parameters to properly accommodate the traffic to avoid congestion issues.

For more information about IBM FlashSystem 900 best practices, see Chapter 3, “Storage back-end” on page 191.

IBM DS8900F

The IBM DS8000 family is a high-performance, high-capacity, highly secure, and resilient series of disk storage systems. The DS8900F family is the latest and most advanced of the DS8000 series offerings to date. The HA, multiplatform support, including IBM Z®, and simplified management tools help provide a cost-effective path to an on-demand world.

From a connectivity perspective, the DS8900F family is scalable. Two different types of host adapters are available: 16 gigabit Fibre Channel (GFC) and 32 GFC. Both can auto-negotiate their data transfer rate down to an 8 Gbps full-duplex data transfer. The 16 GFC and 32 GFC host adapters are all 4-port adapters.

Both adapters contain a high-performance application-specific integrated circuit (ASIC). To ensure maximum data integrity, the ASIC supports metadata creation and checking. Each FC port supports a maximum of 509 host login IDs and 1,280 paths. This configuration enables the creation of large SANs.

Tip: As a best practice for using 16 GFC or 32 GFC technology in DS8900F and IBM Storage Virtualize, consider using the IBM Storage Virtualize maximum of 16 ports for the DS8900F. Also, ensuring that more ranks can be assigned to the IBM Storage Virtualize system than the number of slots that are available on that host ensures that the ports are not oversubscribed.

A single 16 or 32 GFC host adapter does not provide full line rate bandwidth with all ports active:

- ▶ 16 GFC host adapter: 3300 MBps read and 1730 MBps write
- ▶ 32 GFC host adapter: 6500 MBps read and 3500 MBps write

The DS8910F model 993 configuration supports a maximum of eight host adapters. The DS8910F model 994 configurations support a maximum of 16 host adapters in the base frame. The DS8950F model 996 configurations support a maximum of 16 host adapters in the base frame and an extra 16 host adapters in the DS8950F model E96.

Host adapters are installed in slots 1, 2, 4, and 5 of the I/O bay. Figure 2-43 shows the locations for the host adapters in the DS8900F I/O bay.

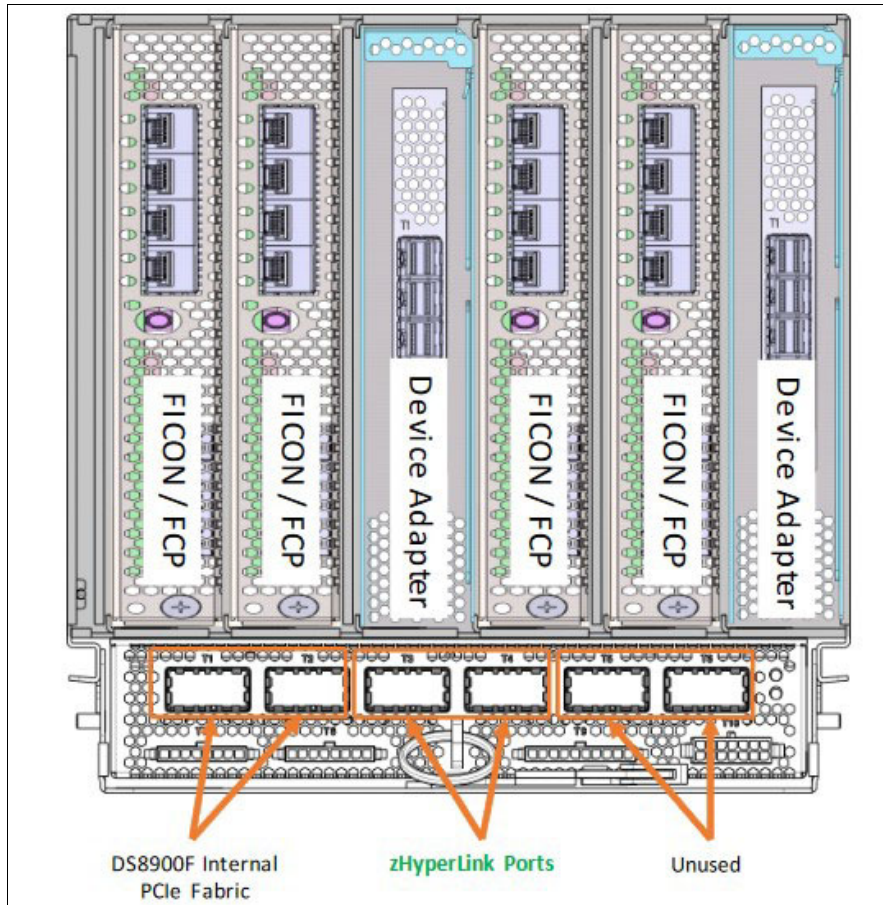


Figure 2-43 DS8900F I/O adapter layout

The system supports an intermix of both adapter types up to the maximum number of ports, as listed in Table 2-15.

Table 2-15 DS8900F port configuration

Model	Minimum and maximum host adapters	Minimum and maximum host adapters ports
994	2 and 16	8 and 64
996	2 and 16	8 and 64
996 + E96	2 and 32	8 and 128

Important: Each of the ports on a DS8900F host adapter can be independently configured for FCP or IBM FICON®. The type of port can be changed through the DS8900F Data Storage Graphical User Interface (DS GUI) or by using Data Storage Command-Line Interface (DS CLI) commands. To work with SAN and IBM Storage Virtualize, use the Small Computer System Interface- Fibre Channel Protocol (SCSI- FCP) FC-switched fabric. FICON is for IBM Z only.

For more information about DS8900F hardware, port, and connectivity, see *IBM Storage DS8900F Architecture and Implementation: Updated for Release 9.3.2*, SG24-8456.

Despite the wide DS8900F port availability, you attach a DS8900F series to an IBM Storage Virtualize system by using Disk Magic to know how many host adapters are required according to your workload, and you spread the ports across different HBAs for redundancy and resiliency proposes. However, consider the following points as a place to start for a single IBM Storage Virtualize cluster configuration:

- ▶ For 16 or fewer arrays, use two host adapters - 8 FC ports.

Note: For redundancy, use four host adapters as a minimum.

- ▶ For 17 - 48 arrays, use four host adapters (16 FC ports).
- ▶ For 48 or more arrays, use eight host adapters (16 FC ports). This configuration matches the most high-performance and demanding environments.

Note: To check the current code MAX limitation, search for the term “and restrictions” for your code level and IBM Storage Virtualize 8.6 at [Support Information for SAN Volume Controller](#).

Figure 2-44 shows the connectivity between an SVC and a DS8886.

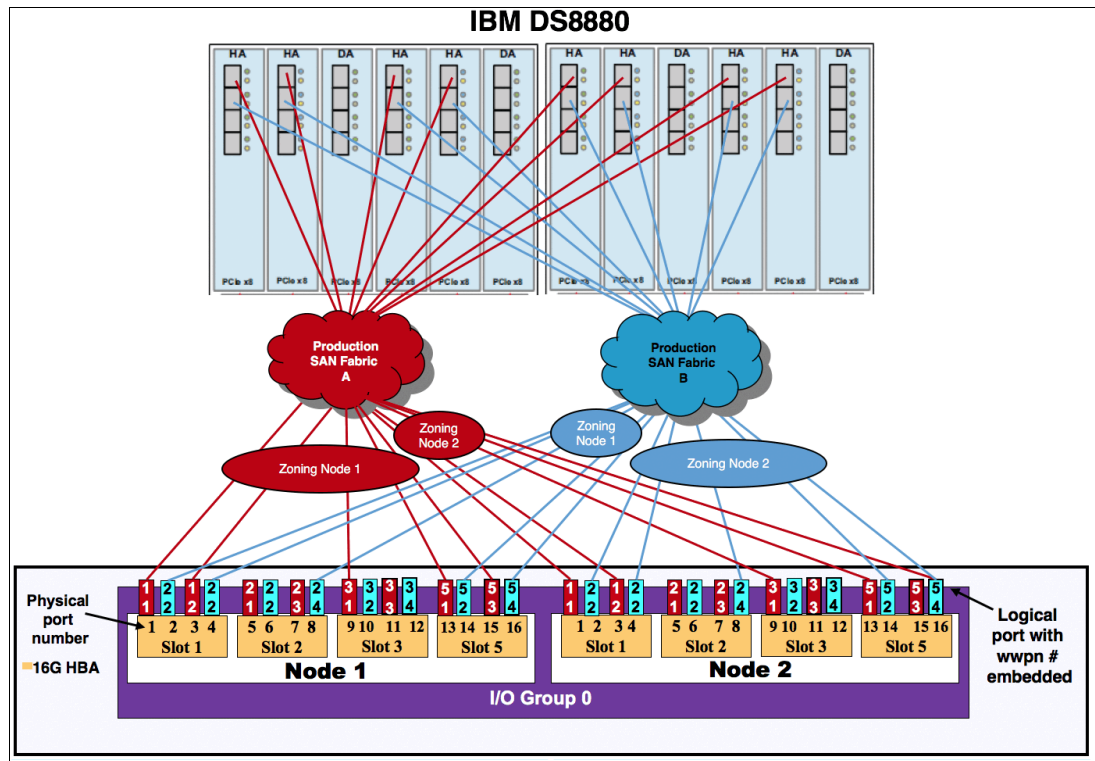


Figure 2-44 DS8886 to SAN Volume Controller connectivity

Note: Figure 2-44 also is valid example that can be used for DS8900F to SVC connectivity.

In Figure 2-44 on page 172, 16 ports are zoned to IBM Storage Virtualize, and the ports are spread across the different HBAs that are available on the storage.

To maximize performance, the DS8900F ports must be dedicated to the IBM Storage Virtualize connections. However, the IBM Storage Virtualize ports must be shared with hosts so that you can obtain the maximum full duplex performance from these ports.

For more information about port usage and assignments, see 2.3.2, “IBM FlashSystem 9200 and 9500 controller ports” on page 136.

Create one zone per IBM Storage Virtualize system node per fabric. IBM Storage Virtualize must access the same storage ports on all nodes. Otherwise, the DS8900F operation status is set to Degraded on the IBM Storage Virtualize system.

After the zoning steps, you must configure the *host connections* by using the DS8900F D DS GUI or DS CLI commands for all IBM Storage Virtualize nodes WWPNs. This configuration creates a single volume group that adds all IBM Storage Virtualize cluster ports within this volume group.

For more information about volume group, host connection, and DS8000 administration, see *IBM Storage DS8900F Architecture and Implementation: Updated for Release 9.3.2*, SG24-8456.

The specific best practices to present DS8880 logical unit numbers (LUNs) as back-end storage to the SVC are described in Chapter 3, “Storage back-end” on page 191.

2.4.5 IBM Storage Virtualize host zones

A best practice to connect a host into an IBM Storage Virtualize system is to create a single zone for each host port. This zone must contain the host port and *one* port from each IBM Storage Virtualize system node that the host must access. Although two ports from each node per SAN fabric are in the usual dual-fabric configuration, ensure that the host accesses only one of them, as shown in Figure 2-45.

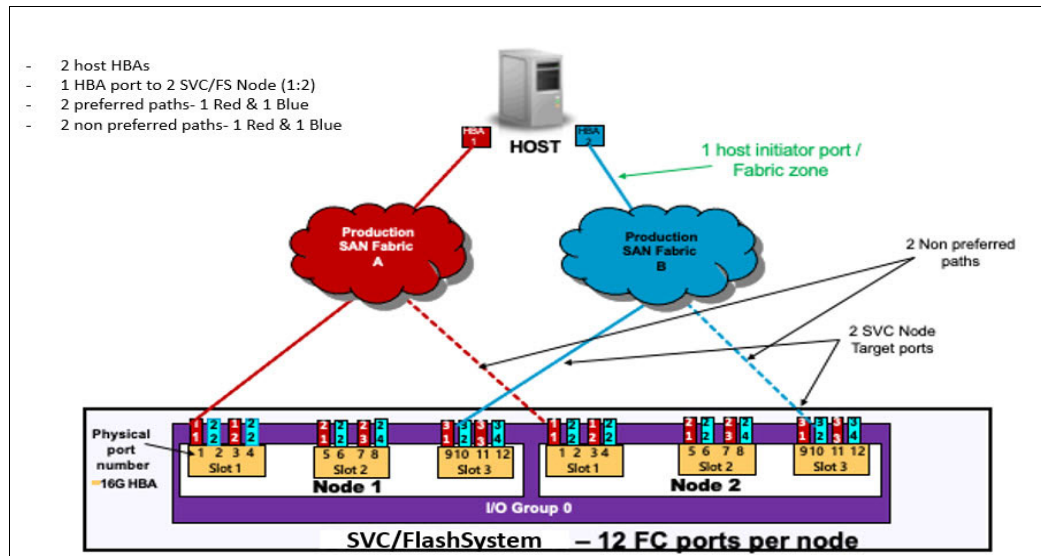


Figure 2-45 Host zoning to an IBM Storage Virtualize node

This configuration provides four paths to each volume, with two preferred paths (one per fabric) and two non-preferred paths. Four paths is the number of paths (per volume) that is optimal for multipathing software, such as AIX Path Control Module (AIXPCM), Linux device mapper, and the IBM Storage Virtualize system.

IBM Storage virtualize supports 32 Gb FC and 64 Gb FC ports. Cisco and Brocade new generation switches also support 32 Gb FC ports connectivity.

A multipath design balances the usage of physical resources. The application should be able to continue to work at the required performance even if the redundant hardware fails. More paths do not equate to better performance or HA. For example, if there are eight paths to a volume and 250 volumes that are mapped to a single host, this configuration would generate 2000 active paths for a multipath driver to manage. Too many paths to a volume can cause excessive I/O waits, resulting in application failures and, under certain circumstances, it can reduce performance.

When the recommended number of paths to a volume is exceeded, sometimes path failures are not recovered in the required amount of time.

It is a best practice to know the workload and application requirements and plan for the number of paths. The number of paths to a host must not exceed eight. *Using end-to-end 32 Gb FC ports, four paths for a volume is the suggested configuration.*

A best practice is to keep paths to all available nodes in an I/O group to achieve maximum availability.

NPIV consideration: All the recommendations in this section also apply to NPIV-enabled configurations. For more information about the systems that are supported by NPIV, see [IBM SAN Volume Controller Configuration Limits](#).

Eight paths per volume are supported. However, this design provides no performance benefits and in some circumstances can reduce performance. Also, it does not significantly improve reliability or availability. However, fewer than four paths do not satisfy the minimum redundancy, resiliency, and performance requirements.

To obtain the best overall performance of the system and to prevent overloading, the workload on each IBM Storage Virtualize system's ports must be equal. Having the same amount of workload typically involves zoning approximately the same number of host FC ports to each IBM Storage Virtualize system node FC port.

Use an IBM Storage Virtualize FC Portsets configuration along with zoning to distribute the load equally on storage FC ports.

Hosts with four or more Host Bus Adapters

If you have four HBAs in your host instead of two HBAs, more planning is required. Because eight paths is not an optimum number, configure your IBM Storage Virtualize host definitions (and zoning) along with IBM Storage Virtualize Portsets as though the single host is two separate hosts. During the volume assignment, you can alternate which volume was assigned to one of the *pseudo-hosts*.

The reason for not assigning one HBA to each path is because the IBM Storage Virtualize I/O group works as a cluster. When a volume is created, one node is assigned as preferred and the other node serves solely as a backup node for that specific volume, which means that using one HBA to each path will never balance the workload for that particular volume.

Therefore, it is better to balance the load by I/O group instead so that the volume is assigned to nodes automatically.

Figure 2-46 shows an example of a 4-port host zoning.

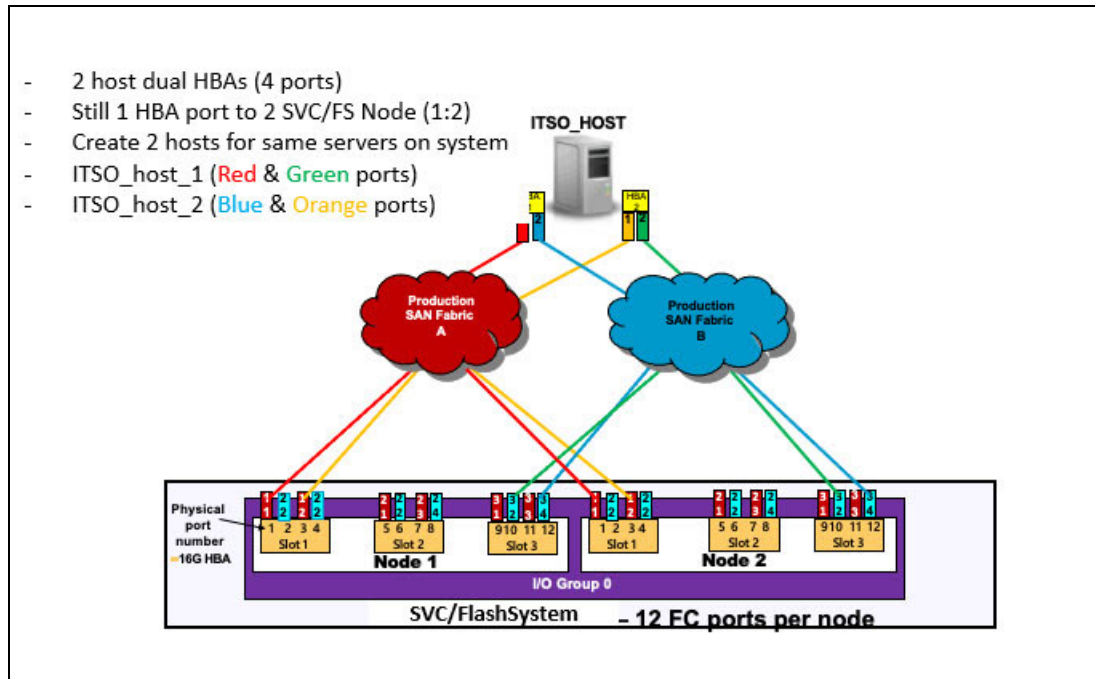


Figure 2-46 Four-port host zoning

Because the optimal number of volume paths is four, you must create two or more hosts on the IBM Storage Virtualize system. During the volume assignment, alternate which volume is assigned to one of the pseudo-hosts in a round-robin fashion.

Note: Pseudo-hosts is not a defined function or feature of IBM Storage Virtualize. To create a pseudo-host, you must add another host ID to the IBM Storage Virtualize host configuration. Instead of creating one host ID with four WWPNs, you define two hosts with two WWPNs, so you must pay extra attention to the SCSI IDs that are assigned to each of the pseudo hosts to avoid having two different volumes from the same storage subsystem with the same SCSI ID.

VMware ESX cluster zoning

For VMware ESX clusters, you must create separate zones for each host node in the VMware ESX cluster, as shown in Figure 2-47.

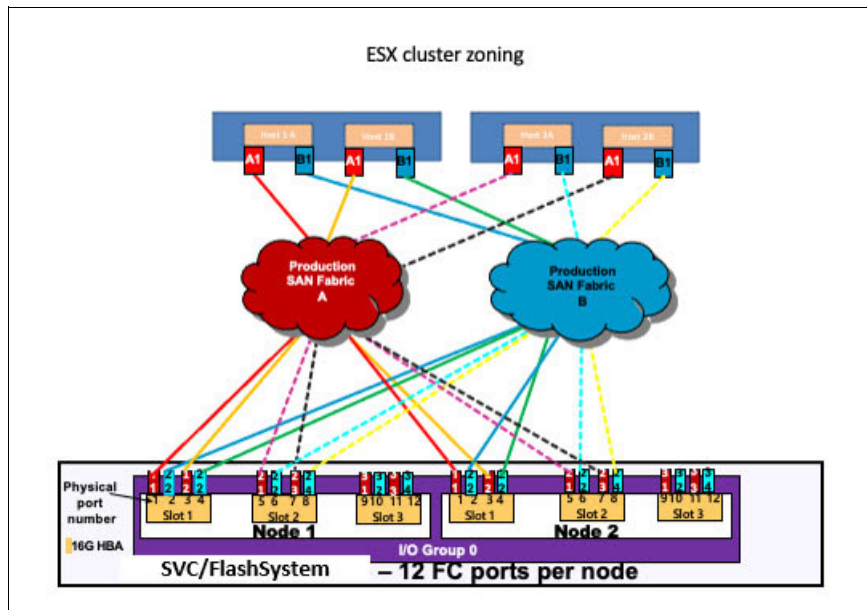


Figure 2-47 VMware ESX cluster zoning

Ensure that you follow these best practices when configuring your VMware ESX clustered hosts:

- ▶ Zone a single VMware ESX cluster in a manner that avoids ISL I/O traversing.
- ▶ Spread multiple host clusters evenly across the IBM Storage Virtualize system node ports and I/O groups.
- ▶ Map LUNs and volumes evenly across zoned ports, alternating the preferred node paths evenly for optimal I/O spread and balance.
- ▶ Create separate zones for each host node in IBM Storage Virtualize and on the VMware ESX cluster.

AIX Virtual I/O Server: Live Partition Mobility zoning

When zoning AIX Virtual I/O Server (VIOS) to IBM Storage Virtualize, you must plan carefully. Because of the complexity of this solution, it is common to create more than four paths to each volume, and MDisk or not provide for proper redundancy. The following best practices can help you avoid a non-degraded path error on IBM Storage Virtualize with four paths per volume:

- ▶ Create two separate and isolated zones on each fabric for each logical partition (LPAR).
- ▶ Create a portset with two IBM Storage Virtualize FC ports and two host FC ports.
- ▶ Do not put both the active and inactive LPAR WWPNs in either the same zone or same IBM Storage Virtualize host definition.
- ▶ Map LUNs to the virtual host FC HBA port WWPNs, not the physical host FCA adapter WWPN.
- ▶ When using NPIV, do not have a ratio of more than one physical adapter to eight virtual ports. This configuration avoids I/O bandwidth oversubscription to the physical adapters.

- ▶ Create a pseudo-host in IBM Storage Virtualize host definitions that contain only two virtual WWPNs (one from each fabric), as shown in Figure 2-48.
- ▶ Map the LUNs or volumes to the pseudo-LPARs (active and inactive) in a round-robin fashion.

Figure 2-48 shows the correct SAN connection and zoning for LPARs.

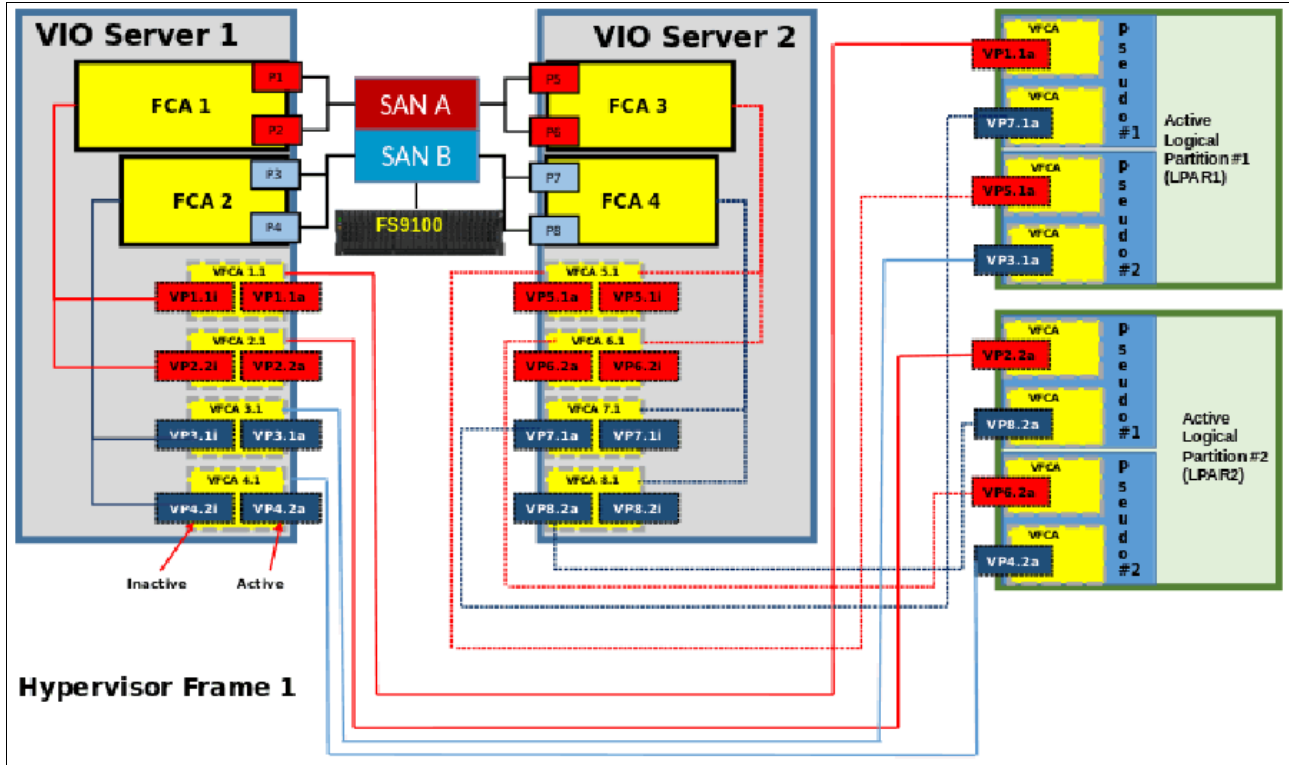


Figure 2-48 LPARs SAN connections

During Live Partition Mobility (LPM), both inactive and active ports are active. When LPM is complete, the previously active ports show as inactive, and the previously inactive ports show as active.

Figure 2-49 shows LPM from the hypervisor frame to another frame.

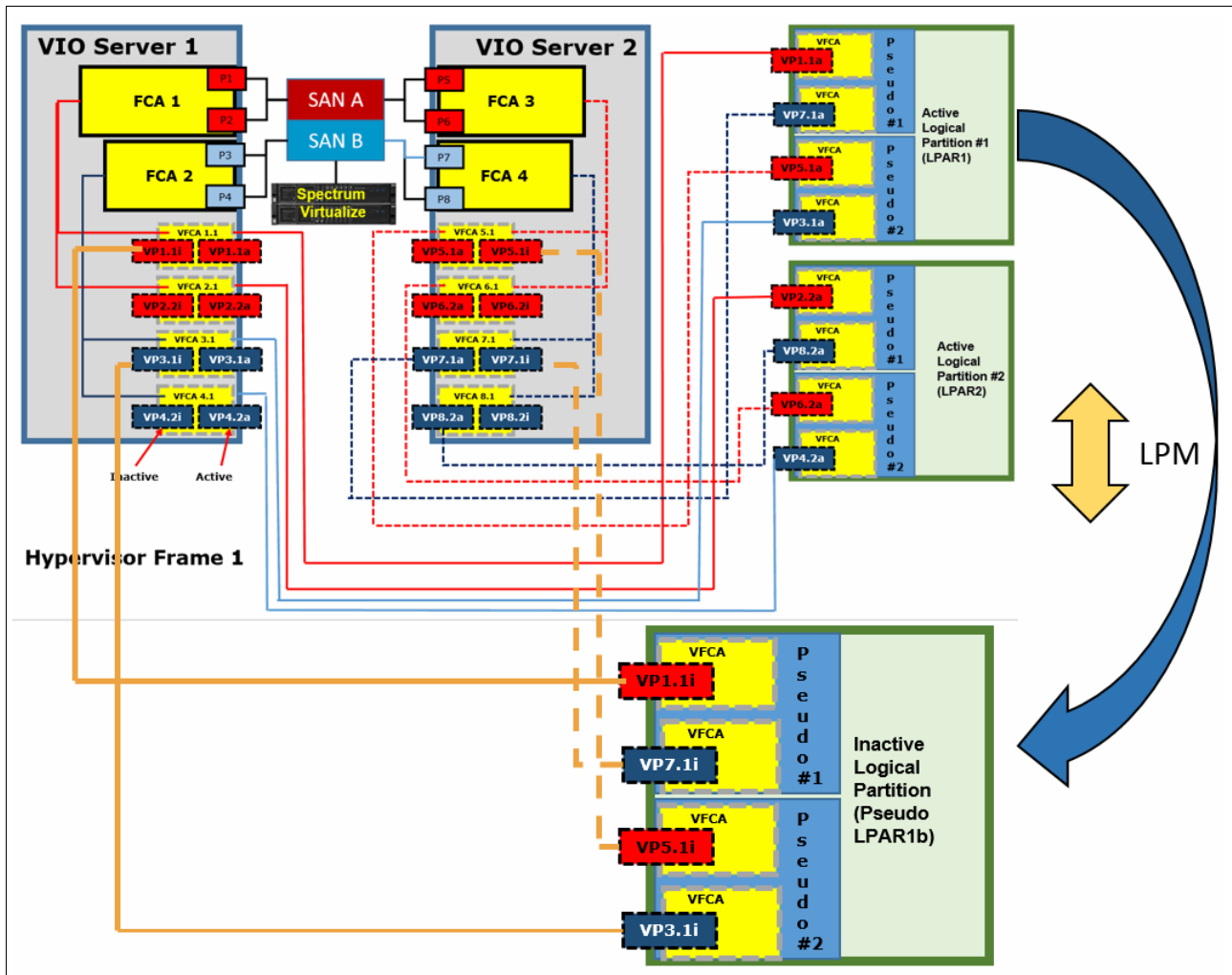


Figure 2-49 Live Partition Mobility

Note: During LPM, the number of paths double from four to eight. Starting with eight paths per LUN or volume results in 16 unsupported paths during LPM, which can lead to I/O interruption.

2.4.6 Hot Spare Node zoning considerations

IBM Storage Virtualize 8.1 introduced the HSN feature, which provides HA for SVC clusters by automatically swapping a spare node into the cluster if the cluster detects a failing node. Also, the maintenance procedures, like code updates and hardware upgrades, benefit from this feature by helping you avoid prolonged loss of redundancy during node maintenance.

For more information about HSNs, see *IBM Spectrum Virtualize: Hot-Spare Node and NPIV Target Ports*, REDP-5477.

For the HSN feature to be fully effective, you must enable the NPIV feature. In an NPIV-enabled cluster, each physical port is associated with two WWPNs. When the port initially logs in to the SAN, it uses the normal WWPN (*primary port*), which does not change from previous releases or from NPIV-disabled mode. When the node completes its startup and is ready to begin processing I/O, the *NPIV target ports* log on to the fabric with the second WWPN.

Special zoning requirements must be considered when implementing the HSN function.

Host zoning with HSN

Hosts should be zoned with NPIV target ports only. Spare node ports must not be included in the host zoning.

Intercluster and intracluster zoning with HSN

Communications between IBM Storage Virtualize nodes, including between different clusters, take place over primary ports. Spare node ports must be included in the intracluster zoning likewise the other nodes.

Similarly, when a spare node comes online, its primary ports are used for remote copy relationships, so the spare node must be zoned with the remote cluster.

Back-end controllers zoning with HSN

Back-end controllers must be zoned to the primary ports on IBM Storage Virtualize nodes. When a spare node is in use, that nodes ports must be included in the back-end zoning, as with the other nodes.

Note: Currently, the zoning configuration for spare nodes is not policed while the spare is inactive, and no errors are logged if the zoning or back-end configuration is incorrect.

Back-end controller configuration with HSN

IBM Storage Virtualize uses the primary ports to communicate with the back-end controller, including the spare. Therefore, all MDisks must be mapped to all IBM Storage Virtualize nodes, including spares.

For IBM Storage Virtualize based back-end controllers, such as IBM Storwize V7000, it is a best practice that the host clusters function is used, with each node forming one host within this cluster. This configuration ensures that each volume is mapped identically to each IBM Storage Virtualize node.

2.4.7 Zoning with multiple IBM Storage Virtualize clustered systems

Unless two separate IBM Storage Virtualize systems participate in a mirroring relationship, configure all zoning so that the two systems do not share a zone. If a single host requires access to two different clustered systems, create two zones with each zone to a separate system.

The back-end storage zones must be separate, even if the two clustered systems share a storage subsystem. You must zone separate I/O groups if you want to connect them in one clustered system. Up to four I/O groups can be connected to form one clustered system.

2.4.8 Split storage subsystem configurations

In some situations, a storage subsystem might be used for IBM Storage Virtualize attachment and direct-attach hosts. In this case, pay attention during the LUN masking process on the storage subsystem. Assigning the same storage subsystem LUN to a host and the IBM Storage Virtualize system can result in swift data corruption.

If you perform a migration into or out of the IBM Storage Virtualize system, make sure that the LUN is removed from one place *before* it is added to another place.

2.4.9 IBM Storage Virtualize Ethernet connectivity

IBM Storage Virtualize systems support Ethernet connectivity by using iSCSI, iSER (RoCE or iWARP), NVMe with ROCEv2 and NVMe/TCP.

The iSCSI protocol is based on TCP/IP, and the iSER protocol is an extension of iSCSI that uses RDMA technology (RoCE or iWARP).

IBM Storage Virtualize provides three adapter options for Ethernet connectivity: 10 GbE, 25 GbE, and 100 GbE adapters.

A 100 GbE adapter supports only host attachment by using iSCSI and NVMe over RDMA and NVMe/TCP on IBM FlashSystem 9500 and 9500R, IBM FlashSystem 7300, and SVC SV3. These models do not support iSER (RoCE or iWARP) for host attachment.

A 25 GbE adapter can be used for clustering, HyperSwap, IP replication, and host attachment by using iSCSI or iSER (RoCE or iWARP) on all IBM Storage Virtualize models.

Consider the following items:

- ▶ A 3260 port must be opened on the network to enable communication between IBM Storage Virtualize and the host.
- ▶ The MTU size must be same on IBM Storage Virtualize and the customer network. MTU sizes of 1500 and 9000 (Jumbo frames) are two standard settings on networks.
- ▶ Four logins are allowed from each port to a single IBM Storage virtualize node.
- ▶ Configure the Ethernet portset for optimum logins.
- ▶ An IP address can be part of multiple portsets.
- ▶ Each portset can contain a maximum of four IP addresses from each node.
- ▶ IBM Storage Virtualize supports a maximum 64 Ethernet portsets.
- ▶ Portset3 is reserved for back-end and storage connectivity.
- ▶ TCP delayed acknowledgment (ACK) should be disabled on initiators.

For more information about configuring Linux and Windows hosts by using iSER connectivity, see the following resources:

- ▶ [Windows host](#)
- ▶ [Linux host](#)

2.5 Distance extension for remote copy services

To implement remote copy services over distance, the following choices are available:

- ▶ Optical multiplexors, such as Dense Wavelength Division Multiplexing (DWDM) or Coarse Wavelength Division Multiplexing (CWDM) devices.
- ▶ Long-distance small form-factor pluggables (SFPs) and 10-Gb small form factor pluggables (XFPs).
- ▶ FC-to-IP address conversion boxes.
- ▶ Native IP address-based replication with IBM Storage Virtualize code.

Of these options, the optical varieties of distance extension are preferred. IP address distance extension introduces more complexity, is less reliable, and has performance limitations. However, in many cases optical distance extension is impractical because of cost or unavailability.

2.5.1 Optical multiplexors

Optical multiplexors can extend your SAN up to hundreds of kilometers at high speeds. For this reason, they are the preferred method for long-distance expansion. When you are deploying optical multiplexing, make sure that the optical multiplexor is certified to work with your SAN switch model. IBM Storage Virtualize model-neutral regarding optical multiplexors.

If you use multiplexor-based distance extension, closely monitor your physical link error counts in your switches. Optical communication devices are high-precision units. When they shift out of calibration, you start to see errors in your frames.

An HA solution such as ESC and HyperSwap optimal performance depends on the configuration of a long-distance link that is based on a multiplexor because node-to-node communication happens on this link along with mirroring of production data.

2.5.2 Long-distance SFPs or XFPs

Long-distance optical transceivers have the advantage of extreme simplicity. Although no expensive equipment is required, a few configuration steps are necessary. Ensure that you use transceivers that are designed for your particular SAN switch *only*. Because each switch vendor supports only a specific set of Cisco SFP or XFP transceivers, it is unlikely that Cisco SFPs work in a Brocade switch.

2.5.3 Fibre Channel over IP

Fibre Channel over IP (FCIP) conversion is by far the most common and least expensive form of distance extension. FCIP is a technology that allows FC routing to be implemented over long distances by using the TCP/IP protocol. In most cases, FCIP is implemented in disaster recovery (DR) scenarios with some kind of data replication between the primary and secondary sites.

FCIP is a tunneling technology, which means FC frames are encapsulated in the TCP/IP packets. As such, it is not apparent to the devices that are connected through the FCIP link. To use FCIP, you need some kind of tunneling device on both sides of the TCP/IP link that integrates FC and Ethernet connectivity. Most of the SAN vendors offer FCIP capability through stand-alone devices (multiprotocol routers) or by using blades that are integrated in the director class product. IBM Storage Virtualize systems support FCIP connection.

An important aspect of the FCIP scenario is the IP link quality. With IP-based distance extension, you must dedicate bandwidth to your FC to IP traffic if the link is shared with other IP traffic. Because the link between two sites is low-traffic or used only for email, do not assume that this type of traffic is always the case. The design of FC is sensitive to congestion, and you do not want a spyware problem or a DDOS attack on an IP network to disrupt IBM Storage Virtualize.

Also, when you communicate with your organization's networking architects, distinguish between megabytes per second (MBps) and megabits per second (Mbps). In the storage world, bandwidth often is specified in MBps, but network engineers specify bandwidth in Mbps. If you fail to specify MB, you can end up with an impressive-sounding 155 Mbps OC-3 link, which supplies only 15 MBps or so to IBM Storage Virtualize. If you include the safety margins, this link is not as fast as you might hope, so ensure that the terminology is correct.

Consider the following points when you are planning for your FCIP TCP/IP links:

- ▶ For redundancy purposes, use as many TCP/IP links between sites as you have fabrics in each site to which you want to connect. In most cases, there are two SAN FC fabrics in each site, so you need two TCP/IP connections between sites.
- ▶ Try to dedicate TCP/IP links only for storage interconnection. Separate them from other LAN or wide area network (WAN) traffic.
- ▶ Make sure that you have a service-level agreement (SLA) with your TCP/IP link vendor that meets your needs and expectations.
- ▶ If you do not use GMCV, make sure that you have sized your TCP/IP link to sustain peak workloads.
- ▶ Using IBM Storage Virtualize internal GM simulation options can help you test your applications before production implementation. You can simulate the GM environment within one SVC system without partnership with another one. To perform GM testing, run the **chsystem** command with the following parameters:
 - `gminterdelaysimulation`
 - `gmintradelaysimulation`

For more information about GM planning, see Chapter 6, "Copy services" on page 363.

- ▶ If you are not sure about your TCP/IP link security, enable Internet Protocol Security (IPsec) on the all FCIP devices. IPsec is enabled on the Fabric OS level, so you do not need any external IPsec appliances.

In addition to planning your TCP/IP link, consider adhering to the following best practices:

- ▶ Set the link bandwidth and background copy rate of partnership between your replicating IBM Storage Virtualize systems to a value *lower* than your TCP/IP link capacity. Failing to set this rate can cause an unstable TCP/IP tunnel, which can lead to stopping all your remote copy relations that use that tunnel.
- ▶ The best case is to use GMCV when replication is done over long distances and bandwidth is limited or the link is shared among multiple workloads.
- ▶ Use compression on corresponding FCIP devices.

- ▶ Use at least two ISLs from your local FC switch to local FCIP router.
- ▶ On a Brocade SAN, use the Integrated Routing feature to avoid merging fabrics from both sites.

For more information about FCIP, see the following publications:

- ▶ *IBM System Storage b-type Multiprotocol Routing: An Introduction and Implementation*, SG24-7544
- ▶ *IBM/Cisco Multiprotocol Routing: An Introduction and Implementation*, SG24-7543

2.5.4 SAN extension with business continuity configurations

IBM Storage Virtualize ESC and HyperSwap technologies provide business continuity solutions over metropolitan areas with distances up to 300 km (186.4 miles). These business continuity solutions over metropolitan areas are achieved by using a SAN extension over WDM technology.

Furthermore, to avoid single points of failure, multiple WDMs and physical links are implemented. When implementing these solutions, particular attention must be paid in the intercluster connectivity setup.

Important: HyperSwap and ESC clusters require implementing dedicated private fabrics for the internode communication between the sites. For more information about the requirements, see *IBM Spectrum Virtualize HyperSwap SAN Implementation and Design Best Practices*, REDP-5597.

Consider a typical implementation of an ESC that uses ISLs, as shown in Figure 2-50.

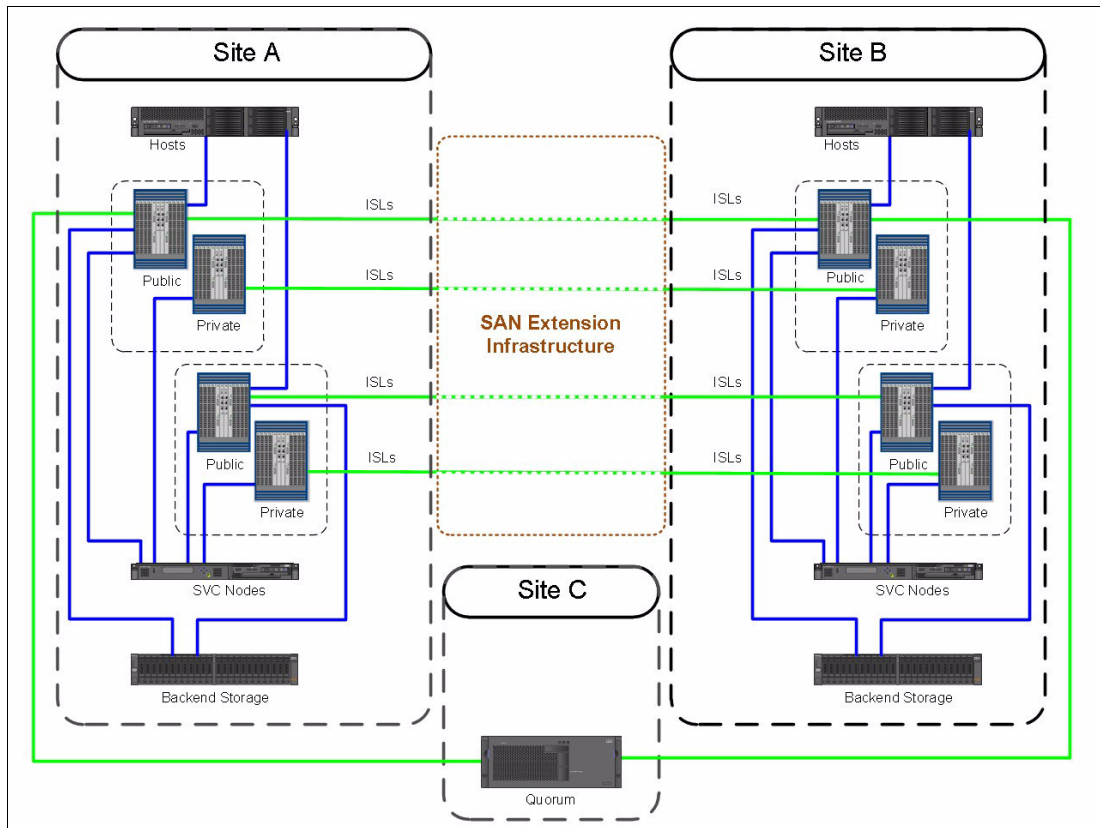


Figure 2-50 Typical Enhanced Stretched Cluster configuration

In this configuration, the intercluster communication is isolated in a private SAN that interconnects Site A and Site B through a SAN extension infrastructure that consists of two DWDMs. For redundancy reasons, assume that two ISLs are used for each fabric for the private SAN extension.

Two possible configurations are available to interconnect the private SANs. In Configuration 1 (see Figure 2-51), one ISL per fabric is attached to each DWDM. In this case, the physical paths Path A and Path B are used to extend both fabrics.

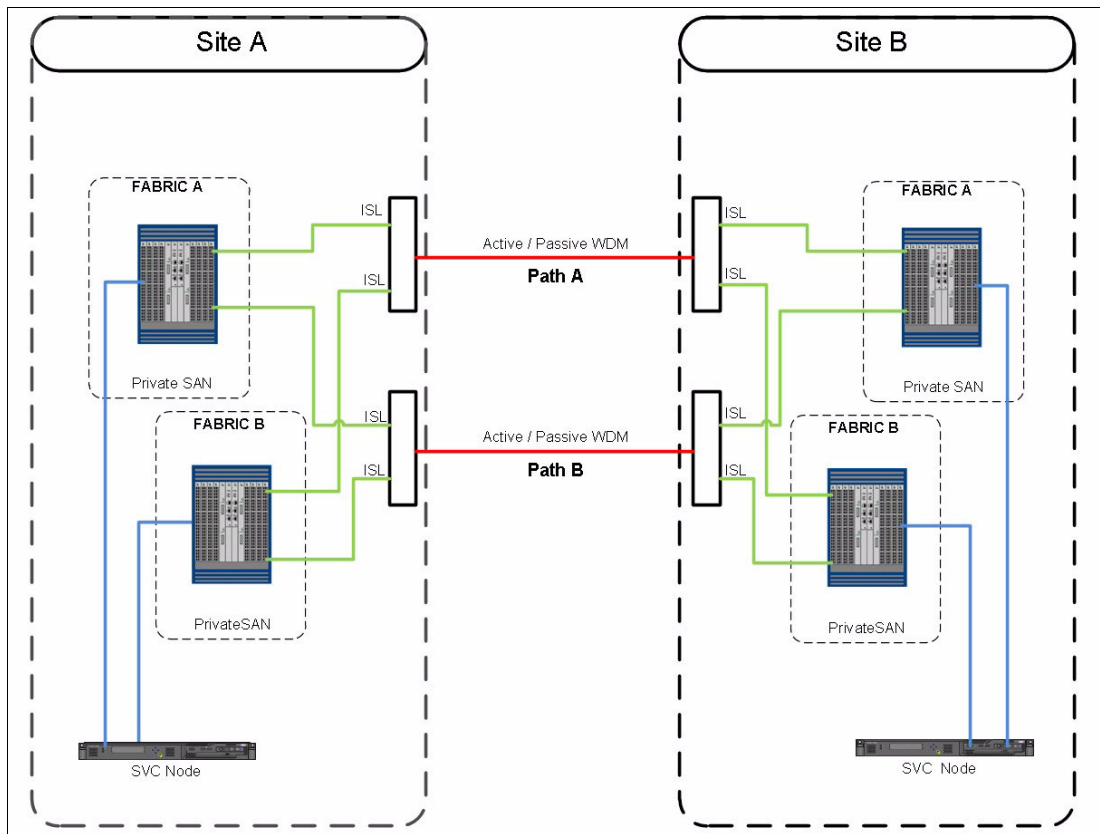


Figure 2-51 Configuration 1: Physical paths shared among the fabrics

In Configuration 2 (see Figure 2-52), ISLs of fabric A are attached only to Path A, while ISLs of fabric B are attached only to Path B. In this case, the physical paths are not shared between the fabrics.

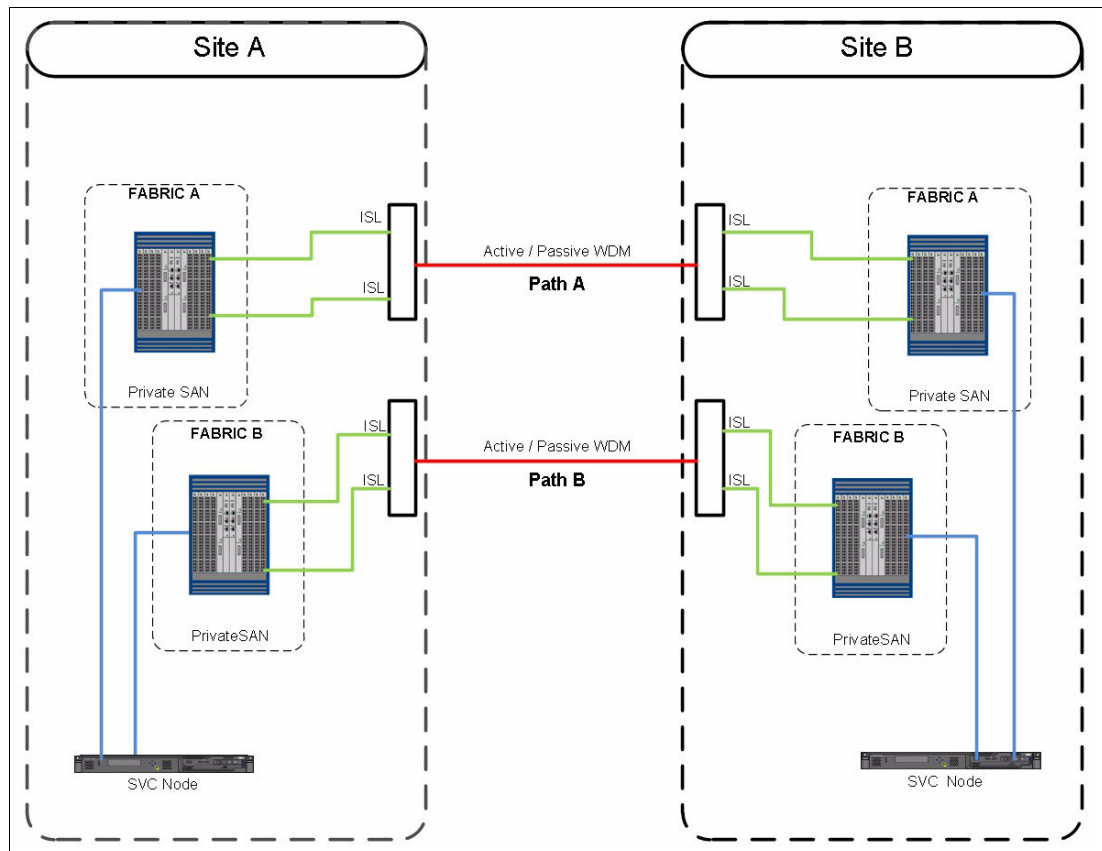


Figure 2-52 Configuration 2: Physical paths not shared among the fabrics

With Configuration 1, in a failure of one of the physical paths, both fabrics are simultaneously affected, and a fabric reconfiguration occurs because of an ISL loss. This situation can lead to a temporary disruption of the intracenter communication and in the worst case to a split-brain condition. To mitigate this situation, link aggregation features such as Brocade ISL trunking can be implemented.

With Configuration 2, a physical path failure leads to a fabric segmentation of one of the two fabrics, leaving the other fabric unaffected. In this case, the intracenter communication would be ensured through the unaffected fabric.

In summary, the recommendation is to fully understand the implication of a physical path or DWDM loss in the SAN extension infrastructure and implement a suitable architecture to avoid a simultaneous impact.

2.5.5 Native IP replication

It is possible to implement native IP-based replication. *Native* means that the IBM Storage Virtualize system does not need any FCIP routers to create a partnership. This partnership is based on the IP network and not on the FC network. Native IP replication is supported on 1 Gb, 10 Gb, and 25 Gb adapters and ports but not on 100 Gb adapters and ports. For more information about native IP replication, see Chapter 6, “Copy services” on page 363.

To enable native IP replication, IBM Storage Virtualize implements the Bridgeworks SANSlide network optimization technology. For more information about this solution, see *IBM SAN Volume Controller and Storwize Family Native IP Replication*, REDP-5103.

The main design point for the initial SANSlide implementation and subsequent enhancements, including the addition of replication compression is to reduce link utilization to allow the links to run closer to their respective line speed at distance and over poor quality links. IP replication compression does not significantly increase the effective bandwidth of the links beyond the physical line speed of the links.

If bandwidths are required that exceed the line speed of the physical links, alternative technologies should be considered (such as FCIP), where compression is done in the tunnel and often yields an increase in effective bandwidth of 2:1 or more.

The effective bandwidth of an IP link highly depends on latency and the quality of the link in terms of the rate of packet loss. Even a small amount of packet loss and the resulting retransmits will significantly degrade the bandwidth of the link.

Figure 2-53 shows the effects that distance and packet loss have on the effective bandwidth of the links in MBps. Numbers reflect a pre-compression data rate with compression on and 50% compressible data. These numbers are as tested and can vary depending on specific link and data characteristics.

1G						10G					
	0ms	20ms	40ms	60ms	80ms		0ms	1ms	2ms	5ms	10ms
0%	122	108	61	41	30	0%	632	683	712	459	266
0.1%	80	66	41	29	25	0.1%	120	120	115	117	87
0.2%	59	44	30	24	20	0.2%	76	81	72	74	59
0.5%	42	28	23	18	14	0.5%	49	48	41	41	35
1%	31	21	17	14	13	1%	33	33	31	28	26

Figure 2-53 Effect of distance on packet loss

Notes: Consider the following points:

- ▶ The maximum bandwidth for a typical IP replication configuration that consists of two 1 Gb links is approximately 244 MBps at zero latency and zero packet loss.
- ▶ When two links are used, replication performs at twice the speed of the lower performing link. For example, the maximum combined data rate for two 1 Gb IP links at 0 latency and 0% packet loss on link A and 0.1% packet loss on link B is 160 MBps.
- ▶ 10 Gb links should not be used with latencies beyond 10 ms. Beyond 10 ms, a 1 Gb link outperforms a 10 Gb link.
- ▶ IBM Storage Virtualize supports volume compression. However, replication runs above volume compression in the IBM Storage Virtualize software stack, which means volumes are replicated at their full uncompressed capacity. This behavior differs from some storage systems, such as the IBM XIV, where replication runs below volume compression and therefore replicates the compressed capacity of the volumes. This difference must be considered when sizing workloads that are moving from one storage system technology to another one.

2.6 Tape and disk traffic that share the SAN

If free ports are available on your core switch, you can place tape devices (and their associated backup servers) on the IBM Storage Virtualize system SAN. However, do not put tape and disk traffic on the same host FC HBA.

To avoid any effects on ISL links and congestion on your SAN, do not put tape ports and backup servers on different switches. Modern tape devices have high-bandwidth requirements.

During your backup SAN configuration, use the switch virtualization to separate the traffic type. The backup process has different frames than production and can affect performance.

Backup requests tend to use all network resources that are available to finish writing on its destination target. Until the request is finished, the bandwidth is occupied and does not allow other frames to access the network.

The difference between these two types of frames is shown in Figure 2-54.

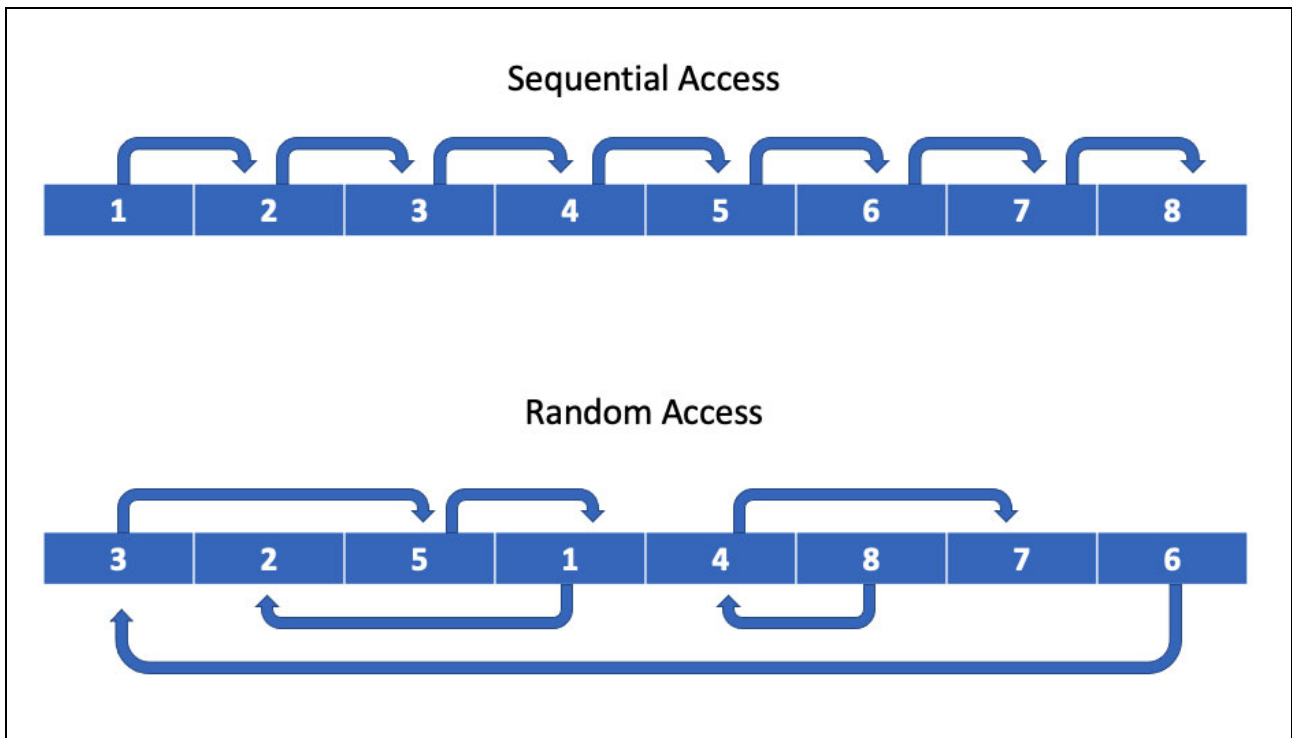


Figure 2-54 FC frames access methods

Backup frames use the sequential method to write data. It releases only the path after it is done writing, and production frames write and read data randomly. Writing and reading is constantly occurring with the same physical path. If backup and production are set up on the same environment, production frames (read/write) can run tasks only when backup frames are complete, which causes latency to your production SAN network.

Figure 2-55 shows one example of a backup and production SAN configuration to avoid congestion because of high-bandwidth usage by the backup process.

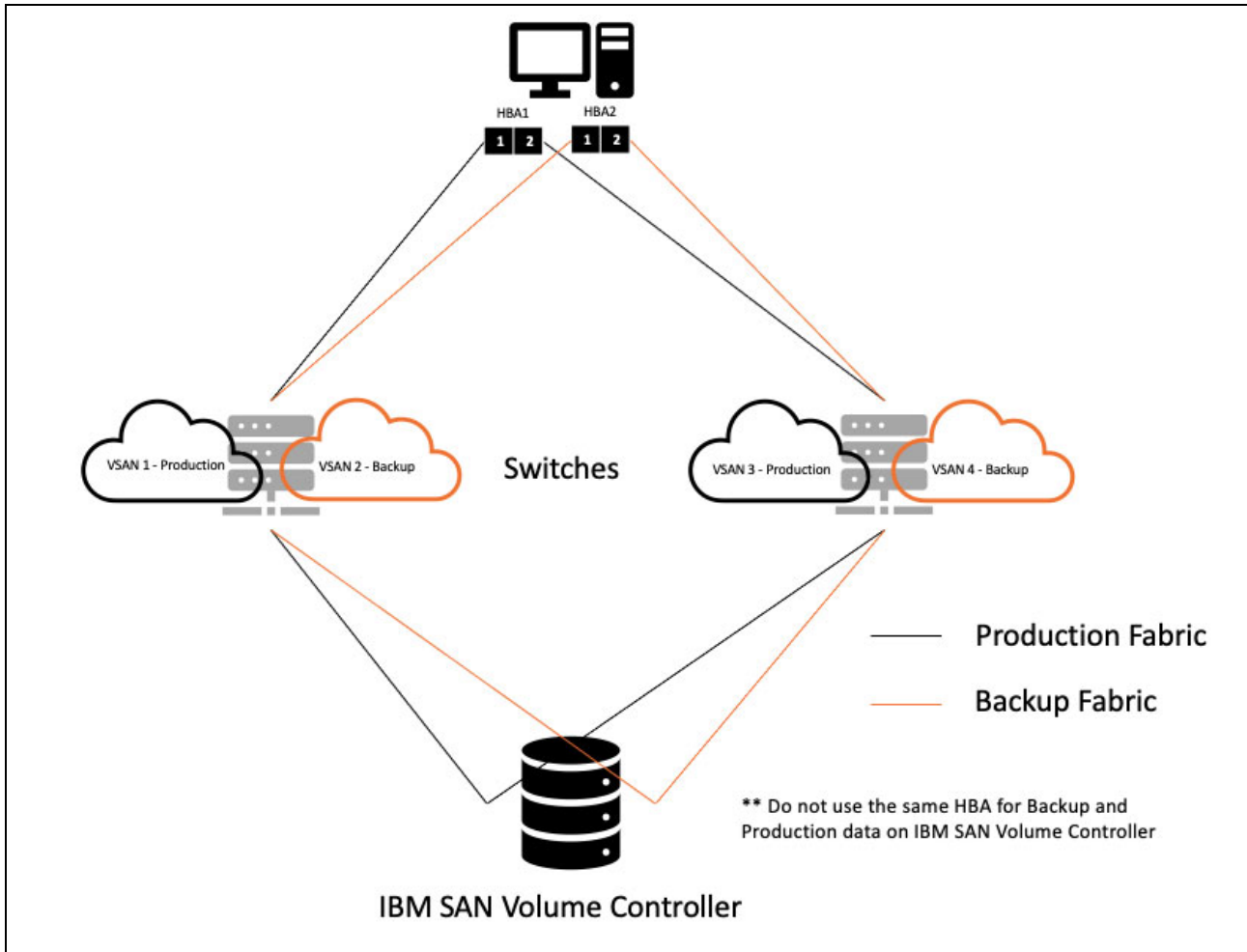


Figure 2-55 Production and backup fabric

2.7 Switch interoperability

Note: For more information about interoperability, see [IBM System Storage Interoperation Center \(SSIC\)](#).

IBM Storage Virtualize is flexible as far as switch vendors are concerned. All the node connections on an IBM Storage Virtualize clustered system must go to the switches of a single vendor, that is, you must not have several nodes or node ports plugged into vendor A and several nodes or node ports plugged into vendor B.

IBM Storage Virtualize supports some combinations of SANs that are made up of switches from multiple vendors in the same SAN. However, this approach is not preferred in practice. Despite years of effort, interoperability among switch vendors is less than ideal because FC standards are not rigorously enforced. Interoperability problems between switch vendors are notoriously difficult and disruptive to isolate. Also, it can take a long time to obtain a fix. For these reasons, run only multiple switch vendors in the same SAN long enough to migrate from one vendor to another vendor, if this setup is possible with your hardware.

You can run a mixed-vendor SAN if you have agreement from both switch vendors that they fully support attachment with each other. However, Brocade does *not* support interoperability with any other vendors.

Interoperability between Cisco switches and Brocade switches is not recommended, except during fabric migrations, and then only if you have a back-out plan in place. Also, when connecting BladeCenter switches to a core switch, consider the use of the NPIV technology.

When you have SAN fabrics with multiple vendors, pay special attention to any particular requirements. For example, observe from which switch in the fabric the zoning must be performed.



Storage back-end

This chapter describes the aspects and best practices to consider when the internal and external back-end storage for a system is planned, configured, and managed:

- ▶ Internal storage consists of flash and disk drives that are installed in the control and expansion enclosures of the system.
- ▶ External storage is acquired by IBM Storage Virtualize by virtualizing a separate IBM or third-party storage system, which is attached with Fibre Channel (FC) or internet Small Computer Systems Interface (iSCSI).

At time of writing, XIV, A9000, A9000R and FlashSystem 900 are still supported.

This chapter also provides information about traditional quorum disks. For information about IP quorum, see Chapter 7, “Ensuring business continuity” on page 501.

This chapter includes the following topics:

- ▶ “Internal storage types” on page 192
- ▶ “Arrays” on page 206
- ▶ “General external storage considerations” on page 215
- ▶ “Controller-specific considerations” on page 221
- ▶ “Quorum disks” on page 239

3.1 Internal storage types

IBM Storage Virtualize Storage supports the following three types of devices that are attached by using the Non-Volatile Memory Express (NVMe) protocol:

- ▶ Storage-class memory (SCM) drives
- ▶ Industry-standard NVMe flash drives
- ▶ IBM FlashCore Module (FCM)

With a serial-attached Small Computer System Interface (SCSI) (SAS) attachment, flash (solid-state drives (SSDs)), and hard disk drives (HDDs) are supported. The set of supported drives depends on the platform.

3.1.1 NVMe storage

IBM FlashSystem offers end-to-end NVMe and provides feature-rich, enterprise-grade storage solutions that help support crucial workloads and applications.

The number of NVMe drive slots per platform is listed in Table 3-1.

Table 3-1 Number of NVMe drive slots per platform

IBM FlashSystem Family	NVMe slots
IBM FlashSystem 50x0	Not supported
IBM FlashSystem 5200	Twelve 2.5-inch slots
IBM FlashSystem 5100	Twenty-four 2.5-inch slots
IBM FlashSystem 7200	Twenty-four 2.5-inch slots
IBM FlashSystem 7300	Twenty-four 2.5-inch slots
IBM FlashSystem 9100	Twenty-four 2.5-inch slots
IBM FlashSystem 9200	Twenty-four 2.5-inch slots
IBM FlashSystem 9500	Forty-eight 2.5-inch slots

The NVMe protocol

NVMe is an optimized, high-performance scalable host controller interface that addresses the needs of systems that use Peripheral Component Interconnect Express (PCIe)-based solid-state storage. The NVMe protocol is an interface specification for communicating with storage devices. It is functionally analogous to other protocols, such as SAS. However, the NVMe interface was designed for fast storage media, such as flash-based SSDs and low-latency non-volatile storage technologies.

NVMe storage devices are typically directly attached to a host system over a PCIe bus, that is, the NVMe controller is contained in the storage device itself, alleviating the need for an extra I/O controller between the CPU and the storage device. This architecture results in lower latency, throughput scalability, and simpler system designs.

The NVMe protocol supports multiple I/O queues versus older SAS and Serial Advanced Technology Attachment (SATA) protocols, which use only a single queue.

NVMe as a protocol like SCSI. It allows for discovery, error recovery, and read/write operations. However, NVMe uses Remote Direct Memory Access (RDMA) over new or existing physical transport layers, such as PCIe, FC, or Ethernet. The major advantage of an NVMe-drive attachment is that it usually uses PCIe connectivity, so the drives are physically connected to the CPU by a high-bandwidth PCIe connection rather than by using a “middle man”, such as a SAS controller chip, which limits total bandwidth to what is available to the PCIe connection into the SAS controller. Where a SAS controller might use 8 or 16 PCIe lanes in total, each NVMe drive has its own dedicated pair of PCIe lanes, which means that a single drive can achieve data rates in excess of multiple GiBps rather than hundreds of MiBps when compared with SAS.

Overall latency can be improved by the adoption of larger parallelism and the modern device drivers that are used to control NVMe interfaces. For example, NVMe over FC versus SCSI over FC are both bound by the same FC network speeds and bandwidths. However, the overhead on older SCSI device drivers (for example, reliance on kernel-based interrupt drivers) means that the software functions in the device driver might limit its capability when compared with an NVMe driver because an NVMe driver typically uses a polling loop interface rather than an interrupt driven interface.

A polling interface is more efficient because the device itself looks for work to do and typically runs in user space (rather than kernel space). Therefore, the interface has direct access to the hardware. An interrupt-driven interface is less efficient because the hardware tells the software when it work must be done by pulling an interrupt line, which the kernel must process and then hand control of the hardware to the software. Therefore, interrupt-driven kernel drivers waste time switching between kernel and user space. As a result, all useful work that is done is bound by the work that a single CPU core can handle. Typically, a single hardware interrupt is owned by just one core.

All IBM Storage Virtualize FC and SAS drivers are implemented as polling drivers. Thus, on the storage side, almost no latency is saved when you switch from SCSI to NVMe as a protocol. However, the bandwidth increases are seen when a SAS controller is switched to a PCIe-attached drive.

Most of the advantages of using an end-to-end NVMe solution when attaching to an IBM Storage Virtualize system are seen as a reduction in the CPU cycles that are needed to handle the interrupts on the host server where the FC Host Bus Adapter (HBA) resides. Most SCSI device drivers remain interrupt-driven, so switching to NVMe over FC results in the same latency reduction. CPU cycle reduction and general parallelism improvements have been part of IBM Storage Virtualize products since 2003.

Industry-standard NVMe drives

IBM Storage Virtualize system control enclosures provide an option to use self-encrypting industry-standard NVMe flash drives, which are available with 800 GB - 30.72 TB capacity.

Table 3-2 lists the supported industry-standard NVMe drives on an IBM Storage Virtualize system.

Table 3-2 Supported industry-standard NVMe drives on IBM Storage Virtualize

IBM FlashSystem	800 GB	1.92 TB	3.84 TB	7.68 TB	15.36 TB	30.72 TB
IBM FlashSystem 5100	Yes	Yes	Yes	Yes	Yes	-
IBM FlashSystem 5200	Yes	Yes	Yes	Yes	Yes	-
IBM FlashSystem 7200	Yes	Yes	Yes	Yes	Yes	-
IBM FlashSystem 7300	-	Yes	Yes	Yes	Yes	Yes
IBM FlashSystem 9100	-	Yes	Yes	Yes	Yes	-
IBM FlashSystem 9200 and 9200R	-	Yes	Yes	Yes	Yes	-
IBM FlashSystem 9500 and 9500R	-	Yes	Yes	Yes	Yes	Yes

Industry-standard NVMe drives start at a smaller capacity point than FCM drives, which allows for a smaller system.

NVMe FlashCore Module

At the heart of the IBM Storage Virtualize Storage system is IBM FlashCore technology. FCM is a family of high-performance flash drives that provides performance-neutral, hardware-based data compression and self-encryption.

FCM introduces the following features:

- ▶ A hardware-accelerated architecture that is engineered for flash, with a hardware-only data path.
- ▶ A modified dynamic compression algorithm for data compression and decompression, implemented completely in drive hardware.
- ▶ Dynamic SLC (single-level cell) cache memory for reduced latency.
- ▶ Cognitive algorithms for wear leveling and heat separation.

A variable stripe redundant array of independent disks (RAID) (VSR) stripes data across more granular, subchip levels, which allows for failing areas of a chip to be identified and isolated without failing the entire chip. Asymmetric wear-leveling monitors the health of blocks within the chips and tries to place “hot” data within the healthiest blocks to prevent the weaker blocks from wearing out prematurely.

Note: At the time of writing, IBM is the only vendor to deliver VSR for multiple dimensions of RAID protection while maintaining peak performance.

The multiple dimensions come from factoring in system-level RAID protection. The advantage is that many of the things that would normally require intervention by system-level RAID are not a problem for IBM solutions because they are dealt with at the module level.

Bit errors that are caused by electrical interference are continually scanned for, and if any errors are found, they are corrected by an enhanced Error Correcting Code (ECC) algorithm. If an error cannot be corrected, then the IBM Storage Virtualize Storage system distributed RAID (DRAID) layer is used to rebuild the data.

Note: NVMe FCMs use inline hardware compression to reduce the amount of physical space that is required. Compression cannot be disabled. If the written data cannot be compressed further or compressing the data causes it to grow in size, the uncompressed data is written. In either case, because the FCM compression is done in the hardware, there is no performance impact.

IBM Storage Virtualize FCMs are not interchangeable with the flash modules that are used in IBM FlashSystem 900 storage enclosures because they have a different form factor and interface.

Modules that are used in IBM FlashSystem 5100, 5200, 7200, 7300, 9100, 9200, and 9500 are a built-in 2.5-inch U2 dual-port form factor.

FCMs are available with a physical or usable capacity of 4.8, 9.6, 19.2, and 38.4 TB. The *usable capacity* is a factor of how many bytes the flash chips can hold.

FCMs have a maximum effective capacity (or virtual capacity) beyond which they cannot be filled. *Effective capacity* is the total amount of user data that can be stored on a module, assuming that the compression ratio (CR) of the data is at least equal to (or higher than) the ratio of effective capacity to usable capacity. Each FCM contains a fixed amount of space for metadata, and the maximum effective capacity is the amount of data it takes to fill the metadata space.

IBM FlashCore Gen3 modules

IBM offers the third generation of FCMs, FCM3, which provides better performance and lower latency than FCM Gen1 and Gen2.

IBM Storage Virtualize 8.5 adds support for FCM3 with increased effective capacity and compression.

Here are the benefits of the new FCM3:

- ▶ Logical to Physical Table (LPT) paging enables increased effective capacity. Increased effective capacity stems from the FCM storing the LPT more efficiently in DRAM. In prior iterations of FCMs, IBM allocated the entire LPT in DRAM, increasing the cost and limiting the amount of effective capacity. LPT paging enables the new FCM3 to increase the effective capacity.
- ▶ IBM Storage Virtualize and FCM3 hints improve performance and reduce latency:
 - FCM3 has some significant performance and latency improvements.
 - Large and extra-large performance is significantly increased:
 - PCIe Gen 4 technology.
 - In comparison with FCM2, read throughput increases from 2.25 GBps to 3.52 GBps, and write throughput doubles.
- ▶ New hints between IBM Storage Virtualize and FCM reduce system-level latency and improve performance. The drive can be sent hints regarding the access frequency of an I/O that it is receiving so that hot data can be placed in faster pages while colder data can be placed in slower pages. This process is also referred to as *smart data placement*.
- ▶ Data reduction pool (DRP) metadata hints to prioritize DRP metadata.
- ▶ RAID parity hints to prioritize RAID parity.
- ▶ Parity scrub hints to de-prioritize background scrubber reads.
- ▶ RAID rebuild enhancements.

Notes:

- ▶ FCM3 is available only on IBM FlashSystem 5200, 7300, and 9500 with IBM Storage Virtualize 8.5.0. and higher.
- ▶ IBM FlashSystem 9500 with FCM3 19.2 TB and 38.4 TB drives use 7 nm (nanometer) technology and PCI Gen 4 to increase the throughput.
- ▶ With 8.5.2 the FS9500 can now support 48 FCM3 XL (38.4 TB) drives.
- ▶ The RAID bitmap space on the FS9500 is increased to 800MB.
- ▶ DRAID expansion can now be from 1 to 42 drives at a time (max was 12 previously).
- ▶ IBM FlashSystem 5200, 7300, and 9500 do not support mixed FCM1, FCM2, and FCM3 in an array.
- ▶ An array with intermixed FCM1 and FCM2 drives performs like an FCM1 array.

FCM capacities are listed in Table 3-3.

Table 3-3 IBM FlashCore Module capacities

Usable capacity	Compression ratio at maximum effective capacity	Maximum effective capacity
4.8 TB	4.5: 1	21.99 TB
9.6 TB	2.3: 1	21.99 TB
19.2 TB	2.3: 1 for FCM2 3: 1 for FCM3	43.98 TB for FCM2 57.6 TB for FCM3
38.4 TB	2.3: 1 for FCM2 3: 1 for FCM3	87.96 TB for FCM2 115.2 TB for FCM3

A 4.8 TB FCM has a higher CR because it has the same amount of metadata space as the 9.6 TB.

For more information about usable and effective capacities, see 3.1.3, “Internal storage considerations” on page 201.

Storage-class memory drives

SCM is a term that is used to describe non-volatile memory devices that perform faster (~10 μs) than traditional NAND SSDs (100 ns).

IBM FlashSystem supports SCM drives that are built on two different technologies:

- ▶ 3D XPoint technology from Intel, which is developed by Intel and Micron (Intel Optane drives)
- ▶ zNAND technology from Samsung (Samsung zSSD)

Available SCM drive capacities are listed in Table 3-4.

Table 3-4 Supported SCM drive capacities

Technology	Small capacity	Large capacity
3D XPoint	350 GB	750 GB
zNAND	800 GB	1.6 TB ^a

a. IBM FlashSystem 7300 supports only a 1.6 TB NVMe SCM drive.

Note: IBM FlashSystem 9500 will not support SCM drives

SCM drives have their own technology type and drive class in an IBM Storage Virtualize configuration. They cannot intermix in the same array with standard NVMe or SAS drives.

Due to their speed, SCM drives are placed in a new top tier, which is ranked higher than existing tier0_flash that is used for NVMe NAND drives.

A maximum of 12 SCM drives can be installed per control enclosure.

When using Easy Tier, think about maximizing the capacity to get the most benefits (unless the working set is small).

SCM with Easy Tier reduces latency, and in some cases improves input/output operations per second (IOPS). If you want the benefits of SCM across all your capacity, then Easy Tier will continually automatically move the hottest data onto the SCM tier and leave the rest of the data on the lower tiers. This action can benefit DRPs when the metadata is moved to the SCM drives.

If you have a particular workload that requires the best performance and lowest latency and it fits into the limited SCM capacity that is available, then use SCM as a separate pool and pick which workloads use that pool.

3.1.2 SAS drives

IBM FlashSystem 5100, 5200, 7200, 7300, 9100, 9200, and 9500 control enclosures have only NVMe drive slots, but systems can be scaled up by attaching SAS expansion enclosures with SAS drives.

IBM FlashSystem 5015 and 5035 control enclosures have twelve 3.5-inch large form factor (LFF) or twenty-four 2.5-inch small form factor (SFF) SAS drive slots. They can be scaled up by connecting SAS expansion enclosures.

A single IBM FlashSystem 5100, 5200, 7200, 9100, or 9200 control enclosure supports the attachment of up to 20 expansion enclosures with a maximum of 760 drives (748 drives for IBM FlashSystem 5200), including NVMe drives in the control enclosure. By clustering control enclosures, the size of the system can be increased to a maximum of 1520 drives for IBM FlashSystem 5100, 2992 drives for IBM FlashSystem 5200, and 3040 drives for IBM FlashSystem 7200, 9100, and 9200.

A single IBM FlashSystem 9500 control enclosure can support up to three IBM FlashSystem 9000 SFF expansion enclosures or one IBM FlashSystem 9000 LFF HD expansion enclosure for a combined maximum of 232 NVMe and SAS drives per system. Intermixing of expansion enclosures in a system is supported.

A single IBM FlashSystem 7300 Model 924 control enclosure can support up to 10 IBM FlashSystem 7000 expansion enclosures with a maximum of 440 drives per system. IBM FlashSystem 7300 systems can be clustered to help deliver greater performance, bandwidth, and scalability. A clustered IBM FlashSystem 7300 system can contain up to four IBM FlashSystem 7300 systems and up to 1,760 drives. Maximum drives per control enclosure:

- ▶ 24 drives in the control enclosure
- ▶ 4 x 92G expansion enclosures
- ▶ 2 x 24G expansion enclosures

IBM FlashSystem 5045 control enclosure supports up to 6x 2Uxx or 2x 5U92 + 1x 2Uxxx expansion enclosures with a maximum of 440 drives (including drives in the control enclosure). With two-way clustering, which is available for IBM FlashSystem 5045, up to 880 drives per system are allowed.

IBM FlashSystem 5035 control enclosure supports up to 20 expansion enclosures with a maximum of 504 drives (including drives in the control enclosure). With two-way clustering, which is available for IBM FlashSystem 5035, up to 1008 drives per system are allowed.

IBM FlashSystem 5015 control enclosure supports up to 10 expansions and 392 drives maximum.

Expansion enclosures are dynamically added without downtime, helping to quickly and seamlessly respond to growing capacity demands.

The following types of SAS-attached expansion enclosures are available for the IBM FlashSystem family:

- ▶ 2U 19-inch rack mount SFF expansion with 24 slots for 2.5-inch drives
- ▶ 2U 19-inch rack mount LFF expansion with 12 slots for 3.5-inch drives (not available for IBM FlashSystem 9x00)
- ▶ 5U 19-inch rack mount LFF high-density expansion enclosure with 92 slots for 3.5-inch drives

Different expansion enclosure types can be attached to a single control enclosure and intermixed with each other.

Note: Intermixing expansion enclosures with machine type (MT) 2077 and MT 2078 is not allowed.

IBM FlashSystem 5035, 5045, 5100, 5200, 7200, 7300, 9100, 9200, and 9500 control enclosures have two SAS chains for attaching expansion enclosures. Keep both SAS chains equally loaded. For example, when attaching ten 2U enclosures, connect half of them to chain 1 and the other half to chain 2.

IBM FlashSystem 5015 has only a single SAS chain.

The number of drive slots per SAS chain is limited to 368. To achieve this goal, you need four 5U high-density enclosures. Table 3-5 on page 200, Table 3-6 on page 200 and Table 3-7 on page 200 list the maximum number of drives that are allowed when different enclosures are attached and intermixed. For example, if three 5U enclosures of an IBM FlashSystem 9200 system are attached to a chain, you cannot connect more than two 2U enclosures to the same chain, and you get 324 drive slots as the result.

Table 3-5 Maximum drive slots per SAS expansion chain for IBM FlashSystem¹

5U expansions	2U expansions										
	0	1	2	3	4	5	6	7	8	9	10
0	0	24	48	72	96	120	144	168	192	216	240
1	92	116	140	164	188	212	236	260	--	--	--
2	184	208	232	256	280	304	--	--	--	--	--
3	276	300	324	--	--	--	--	--	--	--	--
4	368	--	--	--	--	--	--	--	--	--	--

Table 3-6 Maximum drive slots per SAS expansion chain for IBM FlashSystem 9500

5U expansions	2U expansions			
	0	1	2	3
0	0	24	48	72
1	92	-	-	-

Table 3-7 Maximum drive slots per SAS expansion chain for IBM FlashSystem 7300

5U expansions	2U expansions					
	0	1	2	3	4	5
0	0	24	48	72	96	120
1	92	116	140	-	-	-
2	184	-	-	-	-	-

Table 3-8 Maximum drive slots per SAS expansion chain for IBM FlashSystem 5045

5U expansions	2U expansions						
	0	1	2	3	4	5	6
0	0	24	48	72	96	120	144
1	92	116	--	--	--	--	--
2	184	208	--	--	--	--	--

IBM FlashSystem 5015, 5035 and 5045 node canisters have on-board SAS ports for expansions. IBM FlashSystem 5100, 5200, 7200, 9100, and 9200 need a 12 GB SAS interface card to be installed in both nodes of a control enclosure to attach SAS expansions.

Expansion enclosures can be populated with HDDs (high-performance enterprise-class disk drives or high-capacity nearline disk drives) or with SSDs.

A set of allowed drive types depends on the system:

- ▶ IBM FlashSystem 9x00 is all flash.
- ▶ Other members of the family can be configured as all flash or hybrid. In hybrid configurations, different drive types can be intermixed inside a single enclosure.

¹ Except for IBM FlashSystem 7300 and 9500.

Drive capacities vary from less than 1 TB to more than 30 TB.

3.1.3 Internal storage considerations

In this section, we describe best practices for planning and managing IBM FlashSystem internal storage.

Planning for performance and capacity

With IBM FlashSystem 50x0, 5100, 5200, 7200, 7300, 9100, 9200, and 9500, SAS enclosures are used to scale capacity within the performance envelope of a single controller enclosure. Clustering multiple control enclosures scales performance with the extra NVMe storage.

For best performance results, plan to operate your storage system with ~85% or less physical capacity used. Flash drives depend on free pages being available to process new write operations and to quickly process garbage collection.

Without some level of free space, the internal operations to maintain drive health and host requests might over-work the drive, which causes the software to proactively fail the drive, or a hard failure might occur in the form of the drive becoming write-protected (zero free space left). The free space helps in rebuild scenarios where the drives have plenty of room to get background pre-erase workloads done as data is written to the drives and general write amplification occurs.

If you are using data reduction, then regardless of the technology that you choose, it is a best practice to keep the system below ~85% to allow it to respond to sudden changes in the rate of data reduction (such as host encryption being enabled). Also, as you run the system close to full, the garbage-collection function is working harder while new writes are processed, which might slow down the system and increase latency to the host.

Note: For more information about physical flash provisioning, see [Do not provision 100% of the physical flash](#).

Intermix rules

Drives of the same form factor and connector type can be intermixed within an expansion enclosure.

For systems that support NVMe drives, NVMe and SAS drives can be intermixed in the same system. However, NVMe drives can exist only in the control enclosure, and SAS drives can exist only in SAS expansion enclosures.

Within a NVMe control enclosure, NVMe drives of different types and capacities can be intermixed. Industry-standard NVMe drives and SCMs can be intermixed with FCMs.

For more information about rules for mixing different drives in a single DRAID array, see “Drive intermix rules” on page 211.

Formatting

Drives and FCMs must be formatted before they can be used. The format that you use is important because when an array is created, its members must have zero used capacity. Drives automatically are formatted when they become a candidate.

An FCM is expected to format in under 70 seconds. Formatting an SCM drive takes much longer than an FCM or industry-standard NVMe drive. On Intel Optane, drive formatting can take 15 minutes.

While a drive is formatting, it appears as an offline candidate. If you attempt to create an array before formatting is complete, the create command is delayed until all formatting is done. After formatting is done, the command completes.

If a drive fails to format, it goes offline. If so, a manual format is required to bring it back online. The command-line interface (CLI) scenario is shown in Example 3-1.

Example 3-1 Manual FCM format

```
IBM_IBM FlashSystem:IBM FlashSystem 9100-ITS0:superuser>lsdrive | grep offline
id status error_sequence_number use tech_type ...
13 offline 118 candidate tier0_flash ...
IBM_IBM FlashSystem:IBM FlashSystem 9100-ITS0:superuser>chdrive -task format 13
```

Securely erasing

All SCM, FCM, and industry-standard NVMe drives that are used in the system are self-encrypting. For SAS drives, encryption is performed by an SAS chip in the control enclosure.

For industry-standard NVMe drives, SCMs, and FCMs, formatting the drive completes a cryptographic erase of the drive. After the erasure, the original data on that device becomes inaccessible and cannot be reconstructed.

To securely erase SAS or NVMe drive, use the **chdrive -task erase <drive_id>** command.

The methods and commands that are used to securely delete data from drives enable the system to be used in compliance with European Regulation EU2019/424.

Monitoring FCM capacity

The IBM FlashSystem GUI (as shown in Figure 3-1) and CLI (as shown in Example 3-2) allow you to monitor effective and physical capacity for each FCM.

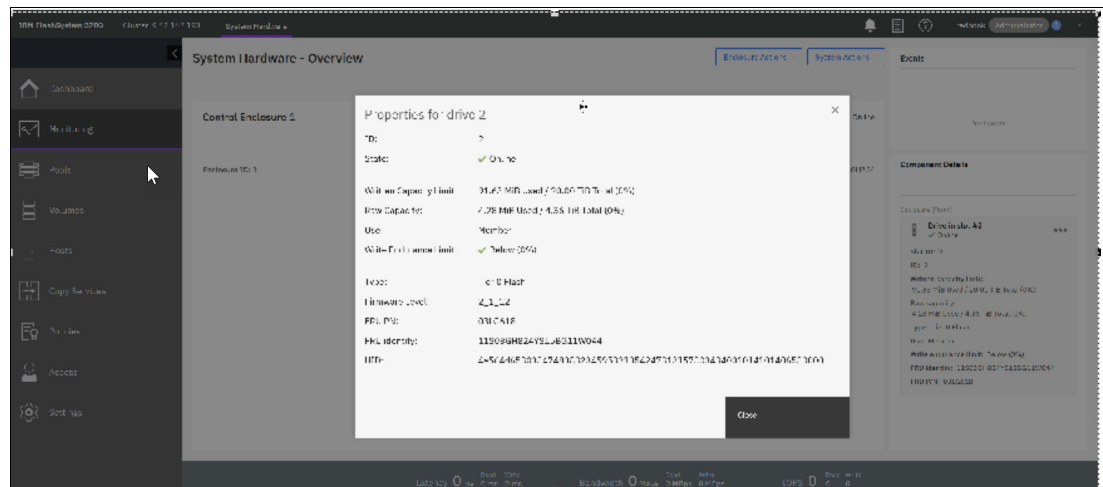


Figure 3-1 FCM capacity monitoring with GUI

Example 3-2 FCM capacity monitoring with CLI

```
IBM_IBM FlashSystem:IBM FlashSystem 9100-ITS0:superuser>lsdrive 2
id 2
...
tech_type tier0_flash
```

```
capacity 20.0TB
...
write_endurance_used 0
write_endurance_usage_rate
replacement_date
transport_protocol nvme
compressed yes
physical_capacity 4.36TB
physical_used_capacity 138.22MB
effective_used_capacity 3.60GB
```

Both examples show the same 4.8 TB FCM with a maximum effective capacity of 20 TiB (or 21.99 TB).

To calculate the actual CR, divide the effective used capacity by the physical used capacity. Here, we have $3.60/0.134 = 26.7$, so written data is compressed 26.7:1 (highly compressible).

Physical used capacity is expected to be nearly the same on all modules in one array.

When FCMs are used, data CRs should be thoroughly planned and monitored.

If highly compressible data is written to an FCM, it still becomes full when it reaches the maximum effective capacity. Any spare data space remaining is used to improve the performance of the module and extend the wear-leveling.

Example: A total of 20 TiB of data that is compressible 10:1 is written to a 4.8 TB module.

The maximum effective capacity of the module is 21.99 TB, which equals 20 TiB.

The usable capacity of the module is 4.8 TB = 4.36 TiB.

After 20 TiB of data is written, the module is 100% full for the array because it has no free effective (logical) capacity. At the same time, the data uses only 2 TiB of the physical capacity. The remaining 2.36 TiB cannot be used for host writes, only for drive internal tasks and to improve the module's performance.

If non-compressible or low-compressible data is written, the module fills until the maximum physical capacity is reached.

Example: A total of 20 TiB of data that is compressible 1.2:1 is written to a 19.2 TB module.

The module's maximum effective capacity is 43.99 TB, which equals 40 TiB. The module's usable capacity is 19.2 TB = 17.46 TiB.

After 20 TiB is written, only 50% of effective capacity is used. With 1.2:1 compression, it occupies 16.7 TiB of physical capacity, which makes the module physically 95% full, and potentially affects the module's performance.

Pool-level and array-level warnings can be set to alert and prevent compressed drive overflow.

Drive writes per day

Drive writes per day (DWPD) is a term that is used to express the number of times that the total capacity of a drive can be written per day within its warranty period. This metric shows drive write endurance.

If the drive write workload is continuously higher than the specified DWPD, the system alerts that the drive is wearing faster than expected. Because DWPD is account for during system sizing, this alert usually means that workload differs from what was expected on the array and it must be revised.

DWPD numbers are important with SSD drives of smaller sizes. With drive capacities below 1 TB, it is possible to write the total capacity of a drive several times a day. When a single SSD provides tens of terabytes, it is unlikely that you can overrun the DWPD measurement.

Consider the following points:

- ▶ SAS-attached Tier1 flash drives support up to 1 DWPD, which means that full drive capacity can be written on it every day and it lasts the 5-year lifecycle.

Example: A total of 3.84 TB read-intensive (RI) SAS drive is rated for 1 DWPD, which means 3,840,000 MB of data can be written on it each day. Each day has $24 \times 60 \times 60 = 86400$ seconds, so $3840000/86400 = 44.4$ MBps of average daily write workload is required to reach 1 DWPD.

Total cumulative writes over a 5-year period are $3.84 \times 1 \text{ DWPD} \times 365 \times 5 = 6.8 \text{ PB}$.

- ▶ FCM2/FCM3 drives are rated with two DWPD over five years, which is measured in usable capacity. Therefore, if the data is compressible (for example, 2:1), the DWPD doubles.

Example: A total of 19.2 TB FCM is rated for 2 DWPD. Its effective capacity is nearly 44 TB = 40 TiB, so if you use 2.3:1 compression, to reach the DWPD limit, the average daily workload over 5 years must be around 1 GBps. Total cumulative writes over a 5-year period are more than 140 PB.

- ▶ SCM drives are rated with 30 DWPD over five years.

The system monitors the number of writes for each drive that supports the DWPD parameter, and it logs a warning event if this amount is above then DWPD for the specific drive type.

It is acceptable that the write endurance usage rate has high warnings, which indicate that the write data rate exceeds the expected threshold for the drive type during the initial phase of system implementation or stress testing. Afterward, when system's workload is stabilized, the system recalculates usage rate and removes the warnings. Calculation is based on a long-run average, so it can take up to 1 month for them to be automatically cleared.

Cumulative writes that are based on possible DWPD numbers are listed in Table 3-9. It provides an overview of the total cumulative writes over a 5-year period with various DWPD numbers.

Table 3-9 Cumulative writes based on possible DWPD

Capacity of flash drive	One DWPD	Two DWPD	Five DWPD
3.84 TB SAS or NVMe SSD	7.0 PB	14.0 PB	35.0 PB
7.68 TB SAS or NVMe SSD	14.0 PB	28.0 PB	70.0 PB

Capacity of flash drive	One DWPD	Two DWPD	Five DWPD
15.36 TB SAS or NVMe SSD	28.0 PB	56.0 PB	140.0 PB
4.8 TB IBM FCM3	26.3 PB	52.6 PB	131.5 PB
9.6 TB IBM FCM3	52.5 PB	105.0 PB	262.5 PB
19.2 TB IBM FCM3	105.1 PB	210.2 PB	525.6 PB
38.4 TB IBM FCM3	210.2 PB	420.4 PB	1.05 EB

To understand the value behind FCMs, consider the following comparison of three drives that share similar-sized physical capacities:

- ▶ A 3.84 TB NVMe SSD and a 4.8 TB FCM. Because of the no-penalty compression that is built into the FCM3, it delivers up to 3.75X the cumulative capacity.
- ▶ A 7.68 TB NVMe SSD and a 9.6 TB FCM. Once again, the built-in compression means that the FCM3 delivers 3.75X the cumulative capacity.
- ▶ A 15.36 TB NVMe SSD and a 19.2 TB. With FCM3, we achieve about 3.75X the cumulative capacity of NVMe SSDs.

So, the DWPD measurement is largely irrelevant for FCMs and large SSDs.

3.1.4 IBM SAN Volume Controller internal storage considerations

IBM SAN Volume Controller (SVC), which is built on DH8 and SV1 nodes, supports SAS-attached expansions with SSDs and HDDs.

For best practices and DRAID configuration, see 3.2.2, “Array considerations” on page 208.

Note: IBM SV2, SA2, and SV3 nodes do not support internal storage.

3.2 Arrays

To use internal IBM Storage Virtualize drives in storage pools and provision their capacity to hosts, the drives must be joined in RAID arrays to form array-type managed disks (MDisks).

3.2.1 Supported RAID types

RAID provides the following key design goals:

- ▶ Increased data reliability
- ▶ Increased I/O performance

IBM Storage Virtualize supports the following RAID types:

- ▶ Traditional RAID (TRAIID)

In a TRAIID approach, data is spread among drives in an array. However, the spare space is constituted by spare drives, which sit outside of the array. Spare drives are idling and do not share I/O load that comes to an array.

When one of the drives within the array fails, all data is read from the mirrored copy (for RAID 10) or is calculated from remaining data stripes and parity (for RAID 5 or RAID 6), and then written to a single spare drive.

- ▶ DRAID

With DRAID, spare capacity is used instead of the idle spare drives from a TRAIID. The spare capacity is spread across the disk drives. Because no drives are idling, all drives contribute to array performance.

If a drive fails, the rebuild load is distributed across multiple drives. By spreading this load, DRAID addresses two main disadvantages of a TRAIID approach: It reduces rebuild times by eliminating the bottleneck of one drive, and it increases array performance by increasing the number of drives that are sharing the workload.

An IBM Storage Virtualize implementation of DRAID allows an effectively spread workload across multiple node canister CPU cores, which provides significant performance improvement over single-threaded TRAIID arrays.

Table 3-10 lists the RAID type and level support on different IBM Storage Virtualize systems.

Table 3-10 Supported RAID levels on IBM Storage Virtualize systems

RAID types	Non-distributed arrays (traditional RAID)				Distributed arrays (DRAID)		
	RAID 0	RAID 1 / 10	RAID 5	RAID 6	DRAID 1 / 10	DRAID 5	DRAID 6
IBM FlashSystem Family	Yes	Yes	-	-	-	Yes	Yes
IBM FlashSystem 5010 and 5030	Yes	Yes	-	-	-	Yes	Yes
IBM FlashSystem 5015 5035 and 5045	-	-	-	-	Yes	Yes	Yes
IBM FlashSystem 5100	Yes	Yes	-	-	-	Yes	Yes
IBM FlashSystem 5200	-	-	-	-	Yes	Yes	Yes
IBM FlashSystem 7200	Yes	Yes	-	-	Yes	Yes	Yes
IBM FlashSystem 7300	-	-	-	-	-	Yes	Yes
IBM FlashSystem 9100	Yes	Yes	-	-	-	Yes	Yes
IBM FlashSystem 9200 and 9200R	Yes	Yes	-	-	Yes	Yes	Yes
IBM FlashSystem 9500 and 9500R	-	-	-	-	Yes	-	Yes

NVMe FCMs that are installed in an IBM Storage Virtualize system can be aggregated into DRAID 6, DRAID 5, or DRAID 1. All TRAIID levels are not supported on FCMs.

SCM drives support DRAID levels 1, 5, and 6, and TRAIID 0.

Some limited RAID configurations do not allow large drives. For example, DRAID 5 cannot be created with any drive type if drives capacities are equal or above 8 TB. Creating such arrays is blocked intentionally to prevent long rebuilding times.

As drive capacities increase, the rebuild time that is required after a drive failure increases significantly. Together with the fact that with larger capacities the chance for a previously unreported (and uncorrectable) medium error increases, customers that configure DRAID 5 arrays on newer platforms or products or with newer drive models are more likely to have a second drive failure or a medium error that is found during rebuild, which would result in an unwanted Customer Impact Event (CIE), and potentially a data loss.

Table 3-11 lists the supported drives, array types, and RAID levels.

Table 3-11 Supported RAID levels with different drive types

Supported drives	Non-distributed arrays (traditional RAIDs)				Distributed arrays (DRAIDs)		
	RAID 0	RAID 1 or 10	RAID 5	RAID 6	DRAID 1 or 10	DRAID 5	DRAID 6
SAS HDDs	Yes	Yes	-	-	Yes ^a	Yes ^b	Yes
SAS flash drives	Yes	Yes	-	-	Yes	Yes ^b	Yes
NVMe drives	Yes	Yes	-	-	Yes	Yes ^b	Yes

Supported drives	Non-distributed arrays (traditional RAIDs)				Distributed arrays (DRAIDs)		
FCMs	-	-	-	-	Yes ^c	Yes	Yes
SCM drives	Yes	Yes	-	-	Yes	Yes	Yes

- a. Three or more HDDs are required for DRAID 1. An array with two members cannot be created with an HDD type, although it can be created with other types. HDDs larger than 8 TiB are not supported for DRAID.
- b. Drives with capacities of 8 TiB or more are not supported for DRAID 5.
- c. Extra-large (38.4 TB) FCMs are not supported by DRAID 1.

3.2.2 Array considerations

In this section, we describe practices that must be considered when planning and managing drive arrays in an IBM FlashSystem environment.

RAID level

Consider the following points when determining which RAID level to use:

- ▶ DRAID 6 is recommended for all arrays with more than six drives.

TRAID levels 5 and 6 are not supported on the current generation of IBM FlashSystem because DRAID is superior to the TRAID levels.

For most use cases, DRAID 5 has no performance advantage compared to DRAID 6. At the same time, DRAID 6 offers protection from the second drive failure, which is vital because rebuild times are increasing together with the drive size. Because DRAID 6 offers the same performance level but provides more data protection, it is the top recommendation.
- ▶ On platforms that support DRAID 1, DRAID 1 is the recommended RAID level for arrays that consist of two or three drives.

DRAID 1 has a mirrored geometry. It consists of mirrors of two strips, which are exact copies of each other. These mirrors are distributed across all array members.
- ▶ For arrays with four or five members, it is a best practice to use DRAID 1 or DRAID 5, with preference to DRAID 1 where it is available.

DRAID 5 provides a capacity advantage over DRAID 1 with same number of drives, at the cost of performance. Particularly during rebuild, the performance of a DRAID 5 array is worse than a DRAID 1 array with the same number of drives.
- ▶ For arrays with six members, the choice is between DRAID 1 and DRAID 6.
- ▶ On platforms that support DRAID 1, do not use TRAID 1 or RAID 10 because they do not perform as well as the DRAID type.
- ▶ On platforms that do not support DRAID 1, the recommended RAID level for NVMe SCM drives is TRAID 10 for arrays of two drives, and DRAID 5 for arrays of four or five drives.
- ▶ RAID configurations that differ from the recommendations that are listed here are not available with the system GUI. If the wanted configuration is supported but differs from these recommendations, arrays of required RAID levels can be created by using the system CLI.

Notes:

- ▶ DRAID 1 arrays are supported only for pools with extent sizes of 1024 MiB or greater.
- ▶ IBM Storage Virtualize 8.5 does not allow more than a single DRAID array that is made of compressing drives (for example, FCM) in the same storage pool (MDisk group).
- ▶ Starting with IBM Storage Virtualize 8.5, shared pool bitmap memory for DRAID arrays is adjusted automatically on creation or deletion of arrays. You do not need to use the **chiogrp** command before running **mkdistributedarray** to create a distributed array and add it to a storage pool.
- ▶ Support for 48 NVMe drives per DRAID 6 array: For IBM FlashSystem 9500, there is now enhanced support for 48 NVMe drives in the enclosure by using DRAID 6 technology. The system must be upgraded to 8.5.2 or a later release to create distributed RAID arrays of more than 24 x 38.4 TB FlashCore Modules (for both one or more arrays) in the same control enclosure. The following configurations are supported:
 - DRAID 6 arrays of NVMe drives support expansion up to 48 member drives, including up to four distributed rebuild areas.
 - DRAID 6 arrays of FCM NVMe drives support expansion up to 48 member drives, including one distributed rebuild area.
 - At the time of writing, DRAID 6 arrays of extra large (38.4 TB) physical capacity FCM NVMe drives support up to 24 member drives, including one distributed rebuild area.

RAID geometry

Consider the following points when determining your RAID geometry:

- ▶ Data, parity, and spare space must be striped across the number of devices that is available. The higher the number of devices, the lower the percentage of overall capacity the spare and parity devices consume, and the more bandwidth that is available during rebuild operations.

Fewer devices are acceptable for smaller capacity systems that do not have a high-performance requirement, but solutions with a few large drives should be avoided. Sizing tools must be used to understand performance and capacity requirements.

- ▶ DRAID code makes full use of the multi-core environment, so splitting the same number of drives into multiple DRAID arrays does not bring performance benefits compared to a single DRAID array with the same number of drives. Maximum system performance can be achieved from a single DRAID array. Recommendations that were given for TRAIT, for example, to create four or eight arrays to spread load across multiple CPU threads, do not apply to DRAID.
- ▶ Consider the following guidelines to achieve the best rebuild performance in a DRAID array:
 - For FCMs and industry-standard NVMe drives, the optimal number of drives in an array is 16 - 24. This limit ensures a balance between performance, rebuild times, and usable capacity. An array of NVMe drives cannot have more than 24 members.²

² Except for IBM FlashSystem 9500.

Notes:

- ▶ The optimal number of drives in an IBM FlashSystem 9500 is 16 - 48.
- ▶ IBM FlashSystem 9500 machine type 4666 delivers 48 drives in a single 4U form factor.

- For SAS HDDs, configure at least 40 drives to the array rather than create many DRAID arrays with much fewer drives in each one to achieve the best rebuild times. A typical best benefit is approximately 48 - 64 HDD drives in a single DRAID 6.
- For SAS SSD drives, the optimal array size is 24 - 36 drives per DRAID 6 array.
- For SCM, the maximum number of drives in an array is 12.
- ▶ Distributed spare capacity, or rebuild areas, are configured with the following guidelines:
 - DRAID 1 with two members: The only DRAID type that is allowed to not have spare capacity (zero rebuild areas).
 - DRAID 1 with 3 - 16 members: The array must have only one rebuild area.
 - DRAID 5 or 6: The minimum recommendation is one rebuild area every 36 drives. One rebuild area per 24 drives is optimal.
 - Arrays with FCM drives cannot have more than one rebuild area per array.

- ▶ The DRAID stripe width is set during array creation and indicates the width of a single unit of redundancy within a distributed set of drives. Reducing the stripe width does not enable the array to tolerate more failed drives. DRAID 6 does not get more redundancy than is determined for DRAID 6 regardless of the width of a single redundancy unit.

A reduced width increases capacity overhead, but also increases rebuild speed because there is a smaller amount of data that the RAID must read to reconstruct the missing data. For example, a rebuild on a DRAID with a 14+P+Q geometry (width = 16) would be slower or have a higher write penalty than a rebuild on a DRAID with the same number of drives but with a 3+P+Q geometry (width = 5). In return, usable capacity for an array with a width = 5 is smaller than for an array with a width = 16.

The default stripe width settings (12 for DRAID 6) provide an optimal balance between those parameters.

- ▶ The array strip size must be 256 KiB. With IBM Storage Virtualize code releases before 8.4.x, it was possible to choose 128 - 256 KiB if the DRAID member drive size was below 4 TB. From 8.4.x and later, you can create arrays with only a 256 KiB strip size.

Arrays that were created on previous code levels with a strip size 128 KiB are fully supported.

- ▶ The stripe width and strip size (both) determine the Full Stride Write (FSW) size. With FSW, data does not need to be read in a stride, so the RAID I/O penalty is greatly reduced.

For better performance, it is often said that you should set the host file system block size to the same value as the FSW size or a multiple of the FSW stripe size. However, the IBM Storage Virtualize cache is designed to perform FSW whenever possible, so no difference is noticed in the performance of the host in most scenarios.

For fine-tuning for maximum performance, adjust the stripe width or host file system block size to match each other. For example, for a 2 MiB host file system block size, the best performance is achieved with an 8+P+Q DRAID6 array (eight data disks x 256 KiB stripe size, with an array stripe width = 10).

Drive intermix rules

Consider the following points when intermixing drives:

- ▶ Compressing drives (FCMs) and non-compressing drives (SAS or NVMe) cannot be mixed in an array.
- ▶ SCM drives cannot be mixed in the same array with other types of NVMe or SAS devices.
- ▶ Physical and logical capacity:
 - For all types of NVMe drives: Members of an array must have the same physical and logical capacity. It is not possible to replace an NVMe drive with an NVMe drive with greater capacity.
 - For SAS drives: Members of an array do not need to have the same physical capacity. When creating an array on SAS drives, you can allow “superior” drives (drives that have better characteristics than the selected drive type). However, array capacity and performance in this case is determined by the base drive type.

For example, if you have four 1.6 TB SSD drives and two 3.2 TB SSD drives, you still can create a DRAID 6 out of those six drives. Although 3.2 TB drives are members of this array, they use only half of their capacity (1.6 TB); the remaining capacity is never used.
- ▶ Mixing devices from different enclosures:
 - For NVMe devices, you cannot mix NVMe devices from different control enclosures in a system into one array.
 - For SAS drives, you can mix SAS drives from different control or expansion enclosures in a system into one array. One DRAID 6 can span across multiple enclosures.

Drive failure and replacement

When a drive fails in a DRAID, arrays recover redundancy by rebuilding to spare capacity, which is distributed between all array members. After a failed drive is replaced, the array performs copyback, which uses the replaced drive and frees the rebuild area.

A DRAID distinguishes non-critical and critical rebuilds. If a single drive fails in a DRAID 6 array, the array still has redundancy, and a rebuild is performed with limited throughput to minimize the effect of the rebuild workload to an array’s performance.

If an array has no more redundancy (which resulted from a single drive failure in DRAID 5 or double drive failure in DRAID 6), a critical rebuild is performed. The goal of a critical rebuild is to recover redundancy as fast as possible. A critical rebuild is expected to perform nearly twice as fast as a non-critical one.

When a failed drive that was an array member is replaced, the system includes it back to an array. For this process, the drive must be formatted first, which might take some time for an FCM or SCM.

If the drive was encrypted in another array, it comes up as failed because this system does not have the required keys. The drive must be manually formatted to make it a candidate.

Note: An FCM drive that is a member of a RAID array must not be reseated unless you are directly advised to do so by IBM Support. Reseating FCM drives that are still in use by an array can cause unwanted consequences.

RAID expansion

Consider the following points for RAID expansion:

- ▶ You can expand distributed arrays to increase the available capacity. As part of the expansion, the system automatically migrates data for optimal performance for the new expanded configuration. Expansion is non-disruptive and compatible with other functions, such as IBM Easy Tier and data migrations.
- ▶ New drives are integrated and data is restriped to maintain the algorithm placement of stripes across the existing and new components. Each stripe is handled in turn, that is, the data in the existing stripe is redistributed to ensure the DRAID protection across the new larger set of component drives.
- ▶ Only the number of member drives and rebuild areas can be increased. RAID level and RAID stripe width stay as they were set during array creation. If you want to change the stripe width for better capacity efficiency, you must create an array, migrate the data, and then expand the array after deleting the original array.
- ▶ The RAID-member count cannot be decreased. It is not possible to shrink an array.
- ▶ DRAID 5, DRAID 6, and DRAID 1 can be expanded. TRAIID arrays do not support expansion.
- ▶ Only one expansion process can run on array at a time. During a single expansion, up to 12 drives can be added.

Only one expansion per storage pool is allowed, with a maximum of four per system.

- ▶ Once expansion is started, it cannot be canceled. You can only wait for it to complete or delete an array.
- ▶ As the array capacity increases, it becomes available to the pool as expansion progresses. There is no need to wait for expansion to be 100% complete because added capacity can be used while expansion is still in progress.

When you expand an FCM array, the physical capacity is not immediately available, and the availability of new physical capacity does not track with logical expansion progress.

- ▶ Array expansion is a process that is designed to run in the background. It can take a significant amount of time.

Array expansion can affect host performance and latency, especially when expanding an array of HDDs. Do not expand an array when the array has over 50% load. If you do not reduce the host I/O load, the amount of time that is needed to complete the expansion increases greatly.

- ▶ Array expansion is not possible when an array is in write-protected mode because it is full (out of physical capacity). Any capacity issues must be resolved first.
- ▶ Creating a separate array can be an alternative for DRAID expansion.

For example, if you have a DRAID 6 array of 40 NL-SAS drives and you have 24 new drives of the same type, the following options are available:

- Perform two DRAID expansions by adding 12 drives in one turn. With this approach, the configuration is one array of 64 drives; however, the expansion process might take a few weeks for large capacity drives. During that time, the host workload must be limited, which can be unacceptable.
- Create a separate 24-drive DRAID 6 array and add it to the same pool as a 40-drive array. The result is that you get two DRAID 6 arrays with different performance capabilities, which is suboptimal. However, the back-end performance-aware cache and Easy Tier balancing can compensate for this flaw.

RAID capacity

Consider the following points when determining RAID capacity:

- ▶ If you are planning only your configuration, use the [IBM Storage Modeler](#) tool, which is available for IBM Business Partners.
- ▶ If your system is deployed, you can use the `lspotentialarraysize` CLI command to determine the capacity of a potential array for a specified drive count, drive class, and RAID level in the specified pool.
- ▶ To get the approximate amount of available space in a DRAID 6 array, use the following formula:

$$\text{Array Capacity} = D / ((W * 256) + 16) * ((N - S) * (W - 2) * 256)$$

D	Drive capacity
N	Drive count
S	Rebuild areas (spare count)
W	Stripe width

Example: For the capacity of a DRAID 6 array of sixteen 9.6 TB FCMs, use the following values:

- ▶ D = 9.6 TB = 8.7 TiB
- ▶ N = 16
- ▶ S = 1
- ▶ W = 12

$$\text{Array capacity} = 8.7 \text{ TiB} / ((12 * 256) + 16) * ((16 - 1) * (12 - 2) * 256) = 8.7 \text{ TiB} / 3088 * 38400 = 108.2 \text{ TiB}$$

3.2.3 Compressed array monitoring

DRAID arrays on FCMs must be carefully monitored and planned because they are over-provisioned, which means that they are susceptible to an out-of-space condition.

To minimize the risk of an out-of-space condition, ensure that the following tasks are done:

- ▶ The data CR is known and account for when planning for an array's physical and effective capacity.
- ▶ Monitor the array's free space and avoid filling it up with more than 85% of physical capacity.

To monitor arrays, use IBM Spectrum Control or IBM Storage Insights with configurable alerts. For more information, see Chapter 9, "Implementing a storage monitoring system" on page 551.

The IBM Storage Virtualize Storage GUI and CLI displays the used and available effective and physical capacities. For examples, see Figure 3-2 and Example 3-3.

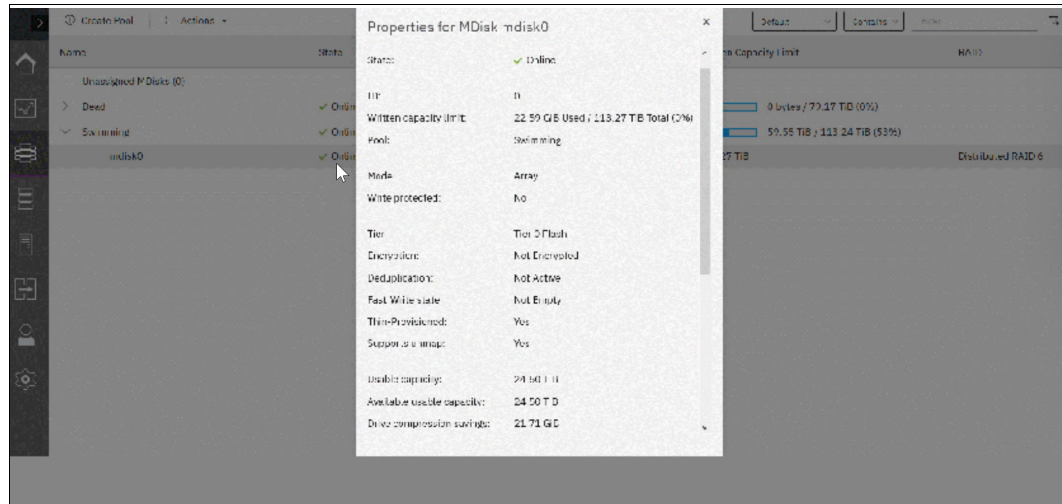


Figure 3-2 Array capacity monitoring with the GUI

Example 3-3 Array capacity monitoring with the CLI

```
IBM_IBM FlashSystem:IBM FlashSystem 9100-ITS0:superuser>lsarray 0
mdisk_id 0
mdisk_name mdisk0
capacity 113.3TB
...
physical_capacity 24.50TB
physical_free_capacity 24.50TB
write_protected no
allocated_capacity 58.57TB
effective_used_capacity 22.59GB
```

- ▶ If the used physical capacity of the array reaches 99%, IBM Storage Virtualize raises event ID 1241: 1% physical space left for compressed array. This event is a call for immediate action.

To prevent running out of space, one or a combination of the following corrective actions must be taken:

- Add storage to the pool and wait while data is balanced between the arrays by Easy Tier.
- Migrate volumes with extents on the MDisk that is running low on physical space to another storage pool or migrate extents from the array that is running low on physical space to other MDisks that have sufficient extents.
- Delete or migrate data from the volumes by using a host that supports UNMAP commands. IBM Storage Virtualize Storage system issues UNMAP to the array and space is released.

For more information about out-of-space recovery, see [Handling out of physical space conditions](#).

- ▶ Arrays are most in danger of running out of space during a rebuild or when they are degraded. DRAID spare capacity, which is distributed across array drives, remains free during normal DRAID operation, thus reducing overall drive fullness. If the array capacity is 85% full, each array FCM is used for less than that because of the spare space reserve. When a DRAID is rebuilding, this space is used.

After the rebuild completes, the extra space is filled and the drives might be truly full, resulting in high levels of write amplification and degraded performance. In the worst case (for example, if the array is more than 99% full before rebuild starts), there is a chance that the rebuild might cause a physical out-of-space condition.

3.3 General external storage considerations

IBM Storage Virtualize can virtualize external storage and make it available to the system. External back-end storage systems (or *controllers* in IBM Storage Virtualize terminology) provide their logical volumes (LVs), which are detected by IBM FlashSystem as MDisks and can be used in storage pools.

This section covers aspects of planning and managing external storage that is virtualized by IBM Storage Virtualize.

External back-end storage can be connected to IBM Storage Virtualize with FC (SCSI) or iSCSI. NVMe-FC back-end attachment is *not* supported because it provides no performance benefits for IBM Storage Virtualize. For more information, see “The NVMe protocol” on page 192.

On IBM FlashSystem 5010 and 5030 and IBM FlashSystem 5015, 5035 and 5045, virtualization is allowed only for data migration. Therefore, these systems can be used to externally virtualize storage as an image mode device for the purposes of data migration, *not* for long-term virtualization.

3.3.1 Storage controller path selection

When a MDisk logical unit (LU) is accessible through multiple storage system ports, the system ensures that all nodes that access this LU coordinate their activity and access the LU through the same storage system port.

An MDisk path that is presented to the storage system for all system nodes must meet the following criteria:

- ▶ The system node is a member of a storage system.
- ▶ The system node has FC or iSCSI connections to the storage system port.
- ▶ The system node has successfully discovered the LU.
- ▶ The port selection process has not caused the system node to exclude access to the MDisk through the storage system port.

When the IBM Storage Virtualize node canisters select a set of ports to access the storage system, the two types of path selection that are described in the next sections are supported to access the MDisks. A type of path selection is determined by external system type and cannot be changed.

To determine which algorithm is used for a specific back-end system, see [IBM System Storage Interoperation Center \(SSIC\)](#), as shown in Figure 3-3.

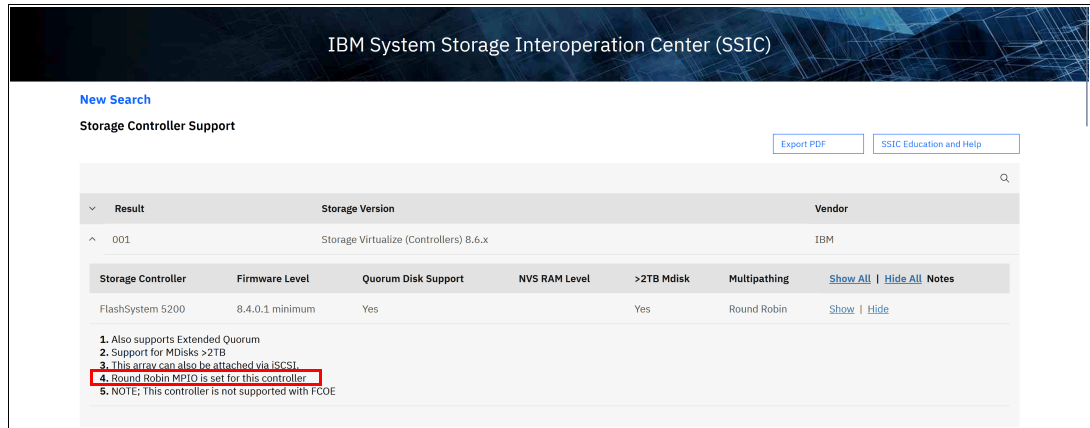


Figure 3-3 IBM System Storage Interoperation Center example

Round-robin path algorithm

With the round-robin path algorithm, each MDisk uses one path per target port per IBM Storage Virtualize node, which means that in cases of a storage system without a preferred controller, such as XIV or DS8000, each MDisk uses all the available FC ports of that storage controller.

With a round-robin compatible storage controller, there is no need to create as many volumes as there are storage FC ports anymore. Every volume and MDisk uses all the available IBM Storage Virtualize ports.

This configuration results in a significant increase in performance because the MDisk is no longer bound to one back-end FC port. Instead, it can issue I/Os to many back-end FC ports in parallel. Particularly, the sequential I/O within a single extent can benefit from this feature.

Additionally, the round-robin path selection improves resilience to certain storage system failures. For example, if one of the back-end storage system FC ports has performance problems, the I/O to MDisk is sent through other ports. Moreover, because I/Os to MDisk are sent through all back-end storage FC ports, the port failure can be detected more quickly.

Best practice: If you have a storage system that supports the round-robin path algorithm, you should zone as many FC ports as possible from the back-end storage controller. IBM Storage Virtualize supports up to 16 FC ports per storage controller. For FC port connection and zoning guidelines, see your storage system documentation.

Example 3-4 shows a storage controller that supports round-robin path selection.

Example 3-4 Round-robin enabled storage controller on IBM FlashSystem 9100

```
IBM_IBM FlashSystem:IBM FlashSystem 9100-ITS0:superuser>lsmdisk 4
id 4
name mdisk4
...
preferred_WWPN 20010002AA0244DA
active_WWPN many                                     <<< Round Robin Enabled
```

MDisk group balanced and controller balanced

Although round-robin path selection provides optimized and balanced performance with minimum configuration required, there are storage systems that still require manual intervention to achieve the same goal.

With storage subsystems that use active-passive type systems, IBM Storage Virtualize accesses an MDisk LU through one of the ports on the preferred controller. To best use the back-end storage, make sure that the number of LUs that is created is a multiple of the connected FC ports and aggregate all LUs to a single MDisk group.

Example 3-5 shows a storage controller that supports the MDisk group balanced path selection.

Example 3-5 MDisk group balanced path selection (no round-robin enabled) storage controller

```
IBM_IBM FlashSystem:IBM FlashSystem 9100-ITS0:superuser>lsmdisk 5
id 5
name mdisk5
...
preferred_WWPN
active_WWPN 20110002AC00C202 <<< indicates MDisk group balancing
```

3.3.2 Guidelines for creating an optimal back-end configuration

Most of the back-end controllers aggregate HDDs or SSDs into RAID arrays, and then join arrays into pools. Logical volumes are created on those pools and provided to hosts. When connected to external back-end storage, IBM Storage Virtualize acts as a host. It is important to create a back-end controller configuration that provides performance and resiliency because IBM Storage Virtualize relies on back-end storage when serving I/O to attached host systems.

If your back-end system has homogeneous storage, create the required number of RAID arrays (RAID 6 or RAID 10 are recommended) with an equal number of drives. The type and geometry of an array depends on the back-end controller vendor's recommendations. If your back-end controller can spread the load stripe across multiple arrays in a resource pool (for example, by striping), create a single pool and add all the arrays there.

On back-end systems with mixed drives, create a separate resource pool for each type of drive (HDD or SSD). Keep the drive type in mind because you must assign the correct tier for an MDisk when it is used by IBM Storage Virtualize.

Create a set of fully allocated logical volumes from the back-end system storage pool (or pools). Each volume is detected as an MDisk on IBM Storage Virtualize. The number of logical volumes to create depends the type of drives that are used by your back-end controller.

Back-end controller with HDDs

If your back-end is using HDDs, the volume number calculation must be based on a queue depth. *Queue depth* is the number of outstanding I/O requests of a device.

For optimal performance, HDDs need 8 - 10 concurrent I/O at the device, which does not change with drive rotation speed. Make sure that in a highly loaded system that any given IBM Storage Virtualize MDisk can queue up approximately eight I/O per back-end system drives.

The IBM Storage Virtualize queue depth per MDisk is approximately 60. The exact maximum on a real system might vary depending on the circumstances. However, for this calculation, it does not matter.

The queue depth per MDisk number leads to the *HDD Rule of 8*. According to this rule, to achieve eight I/Os per drive with a queue depth 60 per MDisk from IBM Storage Virtualize, a back-end array with $60/8 = 7.5$ that is approximately equal to eight physical drives is optimal, or one logical volume per every eight drives in an array.

Example: The back-end controller to be virtualized is IBM Storwize V5030 with 64 nearline serial-attached SCSI (NL-SAS) 8 TB drives.

The system is homogeneous. According to recommendations that are described in 3.2.2, “Array considerations” on page 208, create a single DRAID 6 array in the IBM Storwize and include it in a storage pool. By using the HDD Rule of 8, you want $64/8 = 8$ MDisks, so create eight volumes from a pool to present to IBM Storage Virtualize and assign them to the nearline tier.

All-flash back-end controllers

For all-flash controllers, the considerations are more about I/O distribution across IBM Storage Virtualize ports and processing threads than about queue depth per drive. Because most all-flash arrays that are put behind a virtualizer have high I/O capabilities, make sure that IBM Storage Virtualize is given the optimal chance to spread the load and evenly use its internal resources. Queue depths are less of a concern here (because of the lower latency per I/O).

For all-flash back-end arrays, a best practice is to create at least 32 logical volumes from the array capacity to keep the queue depths high enough and spread the work across the virtualizer resources.

For IBM FlashSystem 9500 with the capacity and performance that it provides, you should consider creating 64 logical volumes from the array capacity.

For smaller setups with a low number of SSDs, this number can be reduced to 16 logical volumes (which results in 16 MDisks) or even eight volumes.

Example: As an example, the back-end controllers that are virtualized are IBM FlashSystem 5035 with 24 Tier1 7.6 TB drives and IBM FlashSystem 900. The virtualizer needs a pool with two storage tiers.

- ▶ On IBM FlashSystem 5035, create a single DRAID 6 array and add it to a storage pool. Using the all-flash rule, you must create 32 volumes to present as MDisks. However, because it is small setup, you can reduce the number of volumes to 16.
- ▶ On the virtualizer, add 16 MDisks from IBM FlashSystem 5035 as Tier1 flash and 32 MDisks as Tier0 flash to a single multitier pool.
- ▶ On IBM FlashSystem 900, join all IBM MicroLatency modules into a RAID 5 array and add it to a storage pool. Because IBM FlashSystem 900 is a Tier0 solution, use the all-flash rule and add 32 MDisks as Tier0 flash to a single multitier pool.

Large setup considerations

For controllers like IBM DS8000 and XIV, you can use an all-flash rule of 32. However, with installations involving this type of back-end controllers, it might be necessary to consider a maximum queue depth per back-end controller port, which is set to 1000 for most supported high-end storage systems.

With high-end controllers, queue depth per MDisk can be calculated by using the following formula:

$$Q = ((P \times C) / N) / M$$

- Q** Calculated queue depth for each MDisk.
- P** Number of back-end controller host ports (unique worldwide port names (WWPNs)) that are zoned to IBM FlashSystem (minimum is 2 and maximum is 16).
- C** Maximum queue depth per WWPN, which is 1000 for controllers, such as XIV Gen3 or DS8000.
- N** Number of nodes in the IBM Storage Virtualize cluster (2, 4, 6, or 8).
- M** Number of volumes that are presented by a back-end controller and detected as MDisks.

For a result of $Q = 60$, calculate the number of volumes that is needed to create as $M = (P \times C) / (N \times Q)$, which can be simplified to $M = (16 \times P) / N$.

Example: A 4-node IBM FlashSystem 9200 is used with 12 host ports on the IBM XIV Gen3 System.

By using the previous formula, we must create $M = (16 \times 12) / 4 = 48$ volumes on the IBM XIV Gen3 to obtain a balanced high-performing configuration.

3.3.3 Considerations for compressing and deduplicating back-end controllers

IBM Storage Virtualize supports over-provisioning on selected back-end controllers, which means that if back-end storage performs data deduplication or data compression on LUs that are provisioned from it, the LUs still can be used as external MDisks on IBM Storage Virtualize.

The implementation steps for thin-provisioned MDisks are the same as for fully allocated storage controllers. Extreme caution should be used when planning capacity for such configurations.

IBM Storage Virtualize detects the following items:

- ▶ If the MDisk is thin-provisioned.
- ▶ The total physical capacity of the MDisk.
- ▶ The used and remaining physical capacity of the MDisk.
- ▶ Whether **unmap** commands are supported by the back-end. By sending SCSI **unmap** commands to thin-provisioned MDisks, the system marks data that is no longer in use. Then, the garbage-collection processes on the back-end can free unused capacity and reallocate it to free space.

Using an appropriate compression or data deduplication ratio is key to achieving a stable environment. If you are not sure about the real compression or data deduplication ratio, contact your IBM technical sales representative to obtain more information.

The nominal capacity from a compression- and deduplication-enabled storage system is not fixed; it varies based on the nature of the data. Always use a conservative data reduction ratio for the initial configuration.

Using a suitable ratio for capacity assignment can cause an out-of-space situation. If the MDisks do not provide enough capacity, IBM Storage Virtualize disables access to all the volumes in the storage pool.

Example: This example includes the following assumptions:

- ▶ Assumption 1: Sizing is performed with an optimistic 5:1 rate.
- ▶ Assumption 2: The real rate is 3:1.

Therefore:

- ▶ Physical Capacity: 20 TB.
- ▶ Calculated capacity: 20 TB x 5 = 100 TB.
- ▶ The volume that is assigned from the compression- or deduplication-enabled storage subsystem to IBM Storage Virtualize or IBM FlashSystem is 100 TB.
- ▶ Real usable capacity: 20 TB x 3 = 60 TB.

If the hosts attempt to write more than 60 TB data to the storage pool, the storage subsystem cannot provide any more capacity. Also, all volumes that are used as IBM Storage Virtualize or FlashSystem MDisks and all related pools go offline.

Thin-provisioned back-end storage must be carefully monitored. It is necessary to set up capacity alerts to be aware of the real remaining physical capacity.

A best practice is to have an emergency plan and know the steps to recover from an “Out Of Physical Space” situation on the back-end controller. The plan must be prepared during the initial implementation phase.

3.3.4 Using data reduction at two levels

If you create a solution where data reduction technologies are applied at both the storage and the virtualization appliance levels (this situation should be avoided if possible), then here are the rules that you should follow:

- ▶ All DRP volumes should run with compression turned on (performance bottlenecks come with DRP metadata, not compression).
- ▶ Start conservative when using FCMs, so assuming that the SVC is not doing data reduction, create 1:1 volumes for the FCM capacity to present to the SVC. Over time, monitor your savings, and then add more volumes on the back-end as needed.
- ▶ Fully allocated volumes should be in their own pool.
- ▶ If you want to use DRPs with an existing overallocated back-end, you must reclaim storage and configure it according to best practices.

IBM FlashSystem A9000 and A9000R considerations

IBM FlashSystem A9000 and A9000R use the IBM industry-leading data reduction technology that combines inline, real-time pattern matching and removal, data deduplication, and compression. Compression uses hardware cards inside each grid controller.

IBM FlashSystem A9000 and A9000R systems always have data reduction on, and because of the grid architecture, they can use all the resources of the grid for the active I/Os. Data reduction should be done at the IBM FlashSystem A9000 or A9000R system, and not at the SVC.

IBM XIV Gen3 considerations

It is a best practice that compression is done in the SVC when attaching XIV Gen3 models 114 and 214 unless the SVC is older hardware and cannot add CPU cores and memory. In this case, depending on workload, it might be better to use XIV compression.

If the XIV Gen3 is a Model 314, it is preferable to do the compression in the XIV system because there are more resources in the grid that are assigned to the compression task. However, if operational efficiency is more important, you can choose to enable compression in the SVC.

3.3.5 Data reduction pools above a simple RAID

In this scenario, An SVC DRP is deployed over IBM Storwize V7000 Gen3 or other fully allocated volumes. The following rules and recommendations apply:

- ▶ Use DRP at the top level to plan for deduplication and snapshot optimizations.
- ▶ DRP at the top level provides best application capacity reporting (volume-written capacity).
- ▶ Always use compression in DRP to get the best performance.
- ▶ Bottlenecks in compression performance come from metadata overhead, not compression processing.

3.3.6 Data reduction pools above a data reducing back-end

In this scenario, an SVC DRP with compression is deployed over IBM Storwize with FCMs or IBM FlashSystem 9200 with FCM. The following rules and recommendations apply:

- ▶ You should assume 1:1 compression in back-end storage.
- ▶ Using DRP with an over-allocated back-end might lead to the DRP garbage causing an out-of-space condition.
- ▶ A small extra savings can be realized from compressing metadata.

For existing systems, you should evaluate whether you need to move to DRP to get the benefits of deduplication or that hardware compression can meet your needs.

3.4 Controller-specific considerations

This section describes implementation-specific information that is related to different supported back-end systems.

3.4.1 Considerations for DS8000 series

In this section, we describe considerations for the DS8000 series.

Interaction between DS8000 and IBM Storage Virtualize

It is important to understand the DS8000 drive virtualization process, which is the process of preparing physical drives for storing data that belongs to a volume that is used by a host. In this case, the host is IBM Storage Virtualize.

In this regard, the basis for virtualization begins with the physical drives of DS8000, which are mounted in storage enclosures. Virtualization builds on the physical drives as a series of layers:

- ▶ Array sites
- ▶ Arrays
- ▶ Ranks
- ▶ Extent pools
- ▶ Logical volumes
- ▶ Logical subsystems

Array sites are the building blocks that are used to define arrays, which are data storage systems for block-based, file-based, or object based storage. Instead of storing data on a server, storage arrays use multiple drives that are managed by a central management and can store a huge amount of data.

In general terms, eight identical drives that have the same capacity, speed, and drive class comprise the array site. When an array is created, the RAID level, array type, and array configuration are defined. RAID 5, RAID 6, and RAID 10 levels are supported.

Important: Normally, RAID 6 is highly preferred and the default while using the Data Storage Graphical Interface (DS GUI). As with large drives in particular, the RAID rebuild times (after one drive failure) become larger. Using RAID 6 reduces the danger of data loss due to a double-RAID failure. For more information, see [RAID implementation](#).

A *rank*, which is a logical representation for the physical array, is relevant for IBM Storage Virtualize because of the creation of a fixed-block (FB) pool for each array that you want to virtualize. Ranks in DS8000 are defined in a one-to-one relationship to arrays. It is for this reason that a rank is defined as using only one array.

An FB rank features one of the following extent sizes:

- ▶ 1 GiB, which is a large extent.
- ▶ 16 MiB, which is a small extent.

An *extent pool* or storage pool in DS8000 is a logical construct to add the extents from a set of ranks, forming a domain for extent allocation to a logical volume.

In synthesis, a *logical volume* consists of a set of extents from one extent pool or storage pool. DS8900F supports up to 65,280 logical volumes.

A logical volume that is composed of fix block extents is called logical unit number (LUN). An FB LUN consists of one or more 1 GiB (large) extents, or one or more 16 MiB (small) extents from one FB extent pool. A LUN is not allowed to cross extent pools. However, a LUN can have extents from multiple ranks within the same extent pool.

Important: DS8000 Copy Services does not support FB logical volumes larger than 2 TiB. Therefore, you cannot create a LUN that is larger than 2 TiB if you want to use Copy Services for the LUN, unless the LUN is integrated as MDisks in an IBM FlashSystem. Use IBM Storage Virtualize Copy Services instead. Based on the considerations, the maximum LUN sizes to create for a DS8900F and present to IBM FlashSystem are as follows:

- ▶ 16 TB LUN with large extents (1 GiB)
- ▶ 16 TB LUN with small extents (16 MiB) for DS8880F with R8.5 or later and for DS8900F R9.0 or later

Logical subsystems (LSSs) are another logical construct, and they are mostly used with FB volumes. Thus, a maximum of 255 LSSs can exist on DS8900F. For more information, see [Actions on the volumes page](#).

The concepts of virtualization of DS8900F for IBM Storage Virtualize are shown in Figure 3-4.

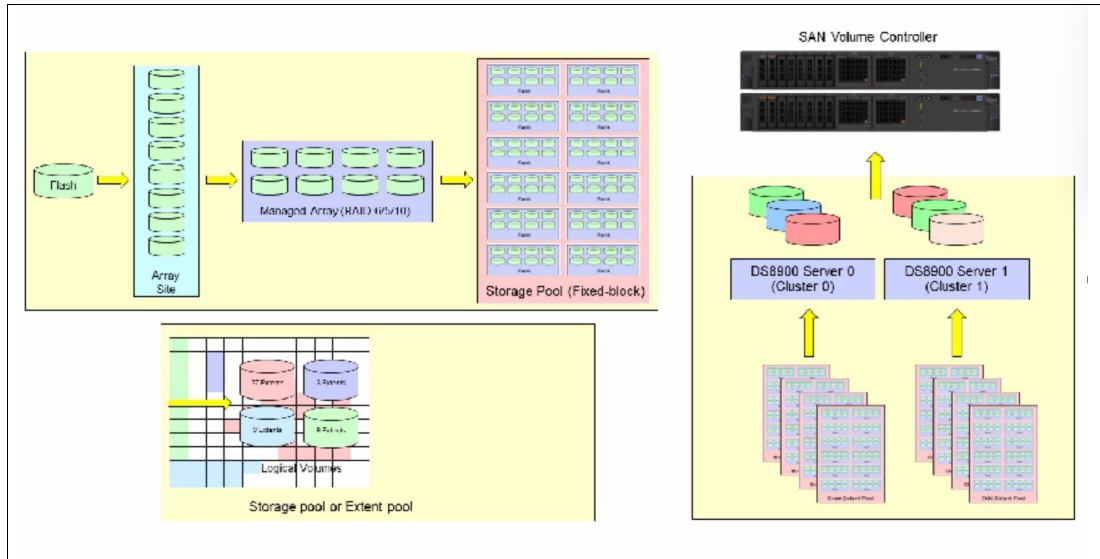


Figure 3-4 Virtualization concepts of DS8900F for IBM Storage Virtualize

Connectivity considerations

The number of DS8000 ports that are used is at least eight. With large and workload-intensive configurations, consider using more ports, up to 16, which is the maximum that is supported by IBM FlashSystem.

Generally, use ports from different host adapters and, if possible, from different I/O enclosures. This configuration is also important because during a DS8000 Licensed Internal Code (LIC) update, a host adapter port might need to be taken offline. This configuration allows the IBM Storage Virtualize I/O to survive a hardware failure on any component on the storage area network (SAN) path.

For more information about SAN best practices and connectivity, see Chapter 2, “Storage area network guidelines” on page 123.

Defining storage

To optimize DS8000 resource utilization, use the following guidelines:

- ▶ Distribute capacity and workload across device adapter (DA) pairs.
- ▶ Balance the ranks and extent pools between the two DS8000 internal servers to support the corresponding workloads on them.
- ▶ Spread the logical volume workload across the DS8000 internal servers by allocating the volumes equally on rank groups 0 and 1.
- ▶ Use as many disks as possible. Avoid idle disks, even if all storage capacity is not to be used initially.
- ▶ Consider using multi-rank extent pools.
- ▶ Stripe your logical volume across several ranks, which is the default for multi-rank extent pools.

Balancing workload across DS8000 series controllers

When you configure storage on the DS8000 series disk storage subsystem, ensure that the ranks on a DA pair are evenly balanced between odd and even extent pools. If you do not ensure that the ranks are balanced, uneven DA loading can cause a considerable performance degradation.

The DS8000 series controllers assign server (controller) affinity to ranks when they are added to an extent pool. Ranks that belong to an even-numbered extent pool have an affinity to Server 0, and ranks that belong to an odd-numbered extent pool have an affinity to Server 1.

Figure 3-5 shows an example of a configuration that results in a 50% reduction in available bandwidth.

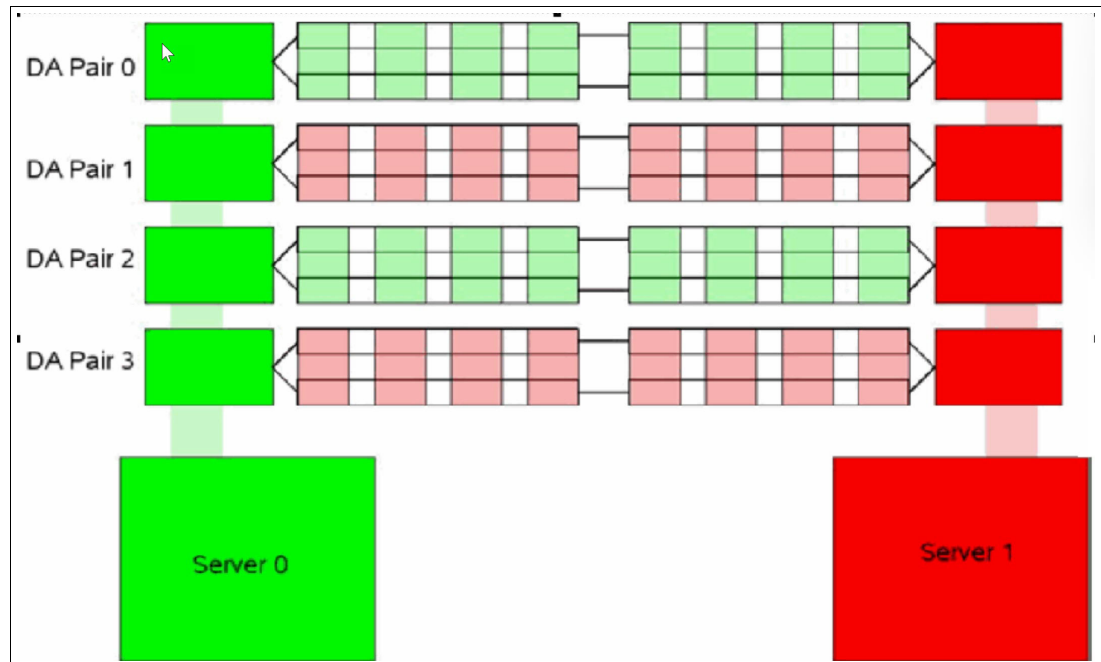


Figure 3-5 DA pair reduced bandwidth configuration

Arrays on each of the DA pairs are accessed only by one of the adapters. In this case, all ranks on DA pair 0 are added to even-numbered extent pools, which means that they all have an affinity to Server 0. Therefore, the adapter in Server 1 is sitting idle. Because this condition is true for all four DA pairs, only half of the adapters are actively performing work. This condition can also occur on a subset of the configured DA pairs.

Example 3-6 shows the invalid configuration, as depicted in the CLI output of the **lsarray** and **lsrank** commands. The arrays that are on the same DA pair contain the same group number (0 or 1), meaning that they have affinity to the same DS8000 series server. Here, Server 0 is represented by Group 0, and server1 is represented by group1.

As an example of this situation, consider arrays A0 and A4, which are attached to DA pair 0. In this example, both arrays are added to an even-numbered extent pool (P0 and P4) so that both ranks have affinity to Server 0 (represented by Group 0), which leaves the DA in Server 1 idle.

Example 3-6 Command output for the lsarray and lsrank commands

```
dscli> lsarray -l
Date/Time: Oct 20, 2016 12:20:23 AM CEST IBM DSCLI Version: 7.8.1.62 DS: IBM.2107-75L2321
Array State Data RAID type arsite Rank DA Pair DDMcap(10^9B) diskclass
=====
A0 Assign Normal 5 (6+P+S) S1 R0 0 146.0 ENT
A1 Assign Normal 5 (6+P+S) S9 R1 1 146.0 ENT
A2 Assign Normal 5 (6+P+S) S17 R2 2 146.0 ENT
A3 Assign Normal 5 (6+P+S) S25 R3 3 146.0 ENT
A4 Assign Normal 5 (6+P+S) S2 R4 0 146.0 ENT
A5 Assign Normal 5 (6+P+S) S10 R5 1 146.0 ENT
A6 Assign Normal 5 (6+P+S) S18 R6 2 146.0 ENT
A7 Assign Normal 5 (6+P+S) S26 R7 3 146.0 ENT
```

```
dscli> lsrank -l
Date/Time: Oct 20, 2016 12:22:05 AM CEST IBM DSCLI Version: 7.8.1.62 DS: IBM.2107-75L2321
ID Group State datastate Array RAIDtype extpoolID extpoolnam stgtype exts usedexts
=====
R0 0 Normal Normal A0 5 P0 extpool0 fb 779 779
R1 1 Normal Normal A1 5 P1 extpool1 fb 779 779
R2 0 Normal Normal A2 5 P2 extpool2 fb 779 779
R3 1 Normal Normal A3 5 P3 extpool3 fb 779 779
R4 0 Normal Normal A4 5 P4 extpool4 fb 779 779
R5 1 Normal Normal A5 5 P5 extpool5 fb 779 779
R6 0 Normal Normal A6 5 P6 extpool6 fb 779 779
R7 1 Normal Normal A7 5 P7 extpool7 fb 779 779
```

Figure 3-6 shows a configuration that balances the workload across all four DA pairs.

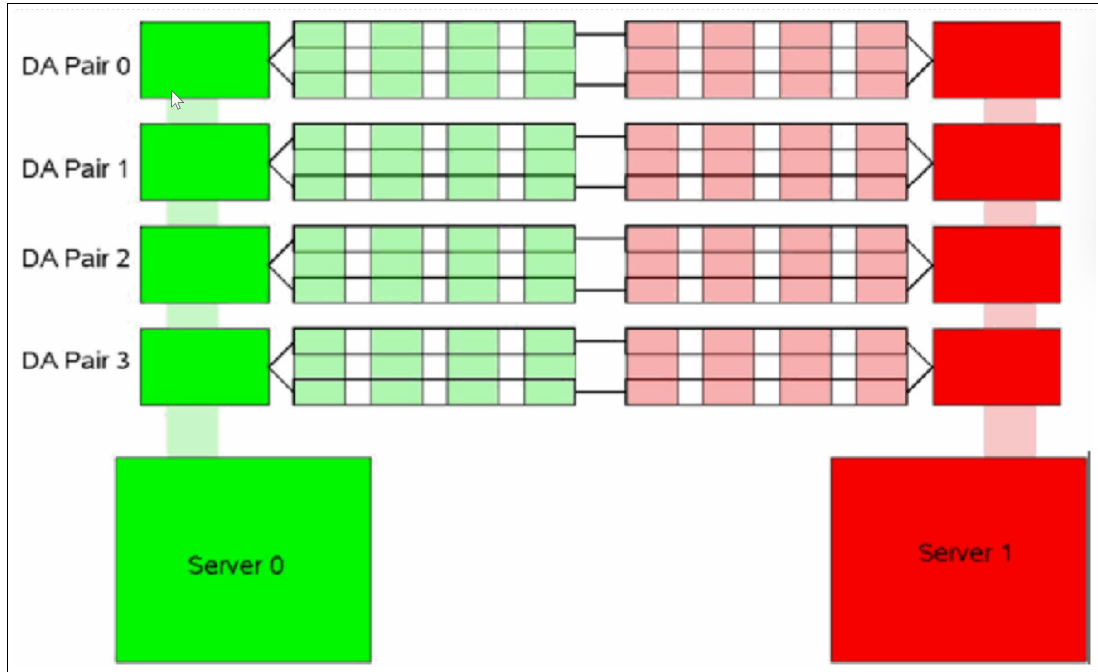


Figure 3-6 DA pair correct configuration

Figure 3-7 shows a correct configuration, as depicted in the CLI output of the `lsarray` and `lsrank` commands. The output shows that this configuration balances the workload across all four DA pairs with an even balance between odd and even extent pools. The arrays that are on the same DA pair are split between groups 0 and 1.

```

dscli> lsarray -l
Date/Time: Oct 20, 2016 10:15:43 AM CEST IBM DSCLI Version: 7.8.1.62 DS: IBM.2107-75L2321
Array State Data RAID type arsite Rank DA Pair DDMcap(10^9B) diskclass
-----
A0 Assign Normal 5 (6+P+S) S1 R0 0 1200.0 ENT
A1 Assign Normal 5 (6+P+S) S2 R1 1 1200.0 ENT
A2 Assign Normal 5 (6+P+S) S3 R2 2 1200.0 ENT
A3 Assign Normal 5 (6+P+S) S4 R3 3 1200.0 ENT
A4 Assign Normal 5 (6+P+S) S5 R4 0 1200.0 ENT
A5 Assign Normal 5 (6+P+S) S6 R5 1 1200.0 ENT
A6 Assign Normal 5 (6+P+S) S7 R6 2 1200.0 ENT
A7 Assign Normal 5 (6+P+S) S8 R7 3 1200.0 ENT

dscli> lsrank -l
Date/Time: Oct 20, 2016 10:20:23 AM CEST IBM DSCLI Version: 7.8.1.62 DS: IBM.2107-75L2321
ID Group State datastate Array RAIDtype extpoolID extpoolnam stgtype exts usedexts encryptgrp marray
-----
R0 0 Normal Normal A0 5 P0 extpool10 fb 6348 6348 - MA1
R1 1 Normal Normal A1 5 P1 extpool11 fb 6348 6348 - MA2
R2 0 Normal Normal A2 5 P2 extpool12 fb 6348 6348 - MA3
R3 1 Normal Normal A3 5 P3 extpool13 fb 6348 6348 - MA4
R4 1 Normal Normal A4 5 P5 extpool15 fb 6348 6348 - MA5
R5 0 Normal Normal A5 5 P4 extpool14 fb 6348 6348 - MA6
R6 1 Normal Normal A6 5 P7 extpool17 fb 6348 6348 - MA7
R7 0 Normal Normal A7 5 P6 extpool16 fb 6348 6348 - MA8
  
```

Figure 3-7 The `lsarray` and `lsrank` command output

DS8000 series ranks to extent pools mapping

In the DS8000 architecture, extent pools are used to manage one or more ranks. An extent pool is visible to both processor complexes in the DS8000 storage system, but it is directly managed by only one of them. You must define a minimum of two extent pools with one extent pool that is created for each processor complex to fully use the resources. You can use the following approaches:

- ▶ One-to-one approach: One rank per extent pool configuration.

With the one-to-one approach, DS8000 is formatted in 1:1 assignment between ranks and extent pools. This configuration disables any DS8000 storage-pool striping or auto-rebalancing activity, if they were enabled. You can create one or two volumes in each extent pool exclusively on one rank only and put all of those volumes into one IBM Storage Virtualize storage pool. IBM Storage Virtualize stripes across all these volumes and balances the load across the RAID ranks by that method. No more than two volumes per rank are needed with this approach. So, the rank size determines the volume size.

Often, systems are configured with at least two storage pools:

- One (or two) that contain MDisks of all the 6+P RAID 5 ranks of the DS8000 storage system.
- One (or more) that contain the slightly larger 7+P RAID 5 ranks.

This approach maintains equal load balancing across all ranks when the IBM Storage Virtualize striping occurs because each MDisk in a storage pool is the same size.

The IBM Storage Virtualize extent size is the stripe size that is used to stripe across all these single-rank MDisks.

This approach delivered good performance and has its advantages. However, it also has a few minor drawbacks:

- A natural skew, such as a small file of a few hundred KiB that is heavily accessed.
- When you have more than two volumes from one rank, but not as many IBM Storage Virtualize storage pools, the system might start striping across many entities that are effectively in the same rank, depending on the storage pool layout. Such striping should be avoided.

An advantage of this approach is that it delivers more options for fault isolation and control over where a certain volume and extent are located.

- ▶ Many-to-one approach: Multi-rank extent pool configuration.

A more modern approach is to create a few DS8000 extent pools, for example, two DS8000 extent pools. Use DS8000 storage pool striping or automated Easy Tier rebalancing to help prevent overloading individual ranks.

Create at least two extent pools for each tier to balance the extent pools by tier and controller affinity. Mixing different tiers on the same extent pool is effective only when Easy Tier is activated on the DS8000 pools. However, virtualized tier management has more advantages when handled by IBM Storage Virtualize.

For more information about choosing the level on which to run Easy Tier, see 4.7, “Easy Tier and tiered and balanced storage pools” on page 296.

You need only one volume size with this multi-rank approach because plenty of space is available in each large DS8000 extent pool. The maximum number of back-end storage ports to be presented to IBM Storage Virtualize is 16. Each port represents a path to IBM Storage Virtualize.

Therefore, when sizing the number of LUNs or MDisks to present to the IBM Storage Virtualize, the suggestion is to present at least 2 - 4 volumes per path. So, using the maximum of 16 paths, create 32, 48, or 64 DS8000 volumes. For this configuration, IBM Storage Virtualize maintains a good queue depth.

To maintain the highest flexibility and for easier management, large DS8000 extent pools are beneficial. However, if the DS8000 installation is dedicated to shared-nothing environments, such as Oracle ASM, IBM Db2® warehouses, or IBM General Parallel File System (GPFS), use the single-rank extent pools.

LUN masking

For a storage controller, all IBM Storage Virtualize nodes must detect the same set of LUs from all target ports that are logged in. If the target ports are visible to the nodes or canisters that do not have the same set of LUs assigned, IBM Storage Virtualize treats this situation as an error condition and generates error code 1625.

You must validate the LUN masking from the storage controller and then confirm the correct path count from within IBM Storage Virtualize.

The DS8000 series controllers perform LUN masking that is based on the volume group. Example 3-7 shows the output of the `showvolgrp` command for volume group V0, which contains 16 LUNs that are presented to a 2-node IBM Storage Virtualize cluster.

Example 3-7 Output of the showvolgrp command

```

dscli> showvolgrp V0
Date/Time: Oct 20, 2016 10:33:23 AM BRT IBM DCLI Version: 7.8.1.62 DS: IBM.2107-75FPX81
Name ITS0_SVC
ID V0
Type SCSI Mask
Vols 1001 1002 1003 1004 1005 1006 1007 1008 1101 1102 1103 1104 1105 1106 1107 1108

```

Example 3-8 shows output for the `lshostconnect` command from the DS8000 series. In this example, four ports of the two-node cluster are assigned to the same volume group (V0), so they are assigned to the same four LUNs.

Example 3-8 Output for the lshostconnect command

```

dscli> lshostconnect -volgrp v0
Date/Time: Oct 22, 2016 10:45:23 AM BRT IBM DCLI Version: 7.8.1.62 DS: IBM.2107-75FPX81
Name          ID   WWPN          HostType Profile          portgrp volgrpID ESSIOport
=====
ITS0_SVC_N1C1P4 0001 500507680C145232 SVC     SAN Volume Controller    1 V0     all
ITS0_SVC_N1C2P3 0002 500507680C235232 SVC     SAN Volume Controller    1 V0     all
ITS0_SVC_N2C1P4 0003 500507680C145231 SVC     SAN Volume Controller    1 V0     all
ITS0_SVC_N2C2P3 0004 500507680C235231 SVC     SAN Volume Controller    1 V0     all

```

In Example 3-8, you can see that only the IBM Storage Virtualize WWPNS are assigned to V0.

Attention: Data corruption can occur if the same LUN is assigned to IBM Storage Virtualize nodes and other devices, such as hosts that are attached to DS8000.

Next, you see how IBM Storage Virtualize detects these LUNs if the zoning is properly configured. The MDisk Link Count (`mdisk_link_count`) represents the total number of MDisks that are presented to the IBM Storage Virtualize cluster by that specific controller.

Example 3-9 shows the general details of the output storage controller by using the system CLI.

Example 3-9 Output of the lscontroller command on IBM FlashSystem

```

IBM_IBM FlashSystem:IBM FlashSystem 9100-ITS0:superuser>svcinfolcontroller
DS8K75FPX81
id 1
controller_name DS8K75FPX81
WWNN 5005076305FFC74C
mdisk_link_count 16
max_mdisk_link_count 16
degraded no
vendor_id IBM
product_id_low 2107900
...
WWPN 500507630500C74C
path_count 16
max_path_count 16
WWPN 500507630508C74C
path_count 16
max_path_count 16
  
```

IBM Storage Virtualize MDisks and storage pool considerations

A best practice is to create a single IBM Storage Virtualize storage pool per DS8900F system, which provides simplicity of management and best overall performance.

An example of the preferred configuration is shown in Figure 3-8. Four storage pools or extent pools (one even and one odd) of DS8900F are joined into one IBM Storage Virtualize storage pool.

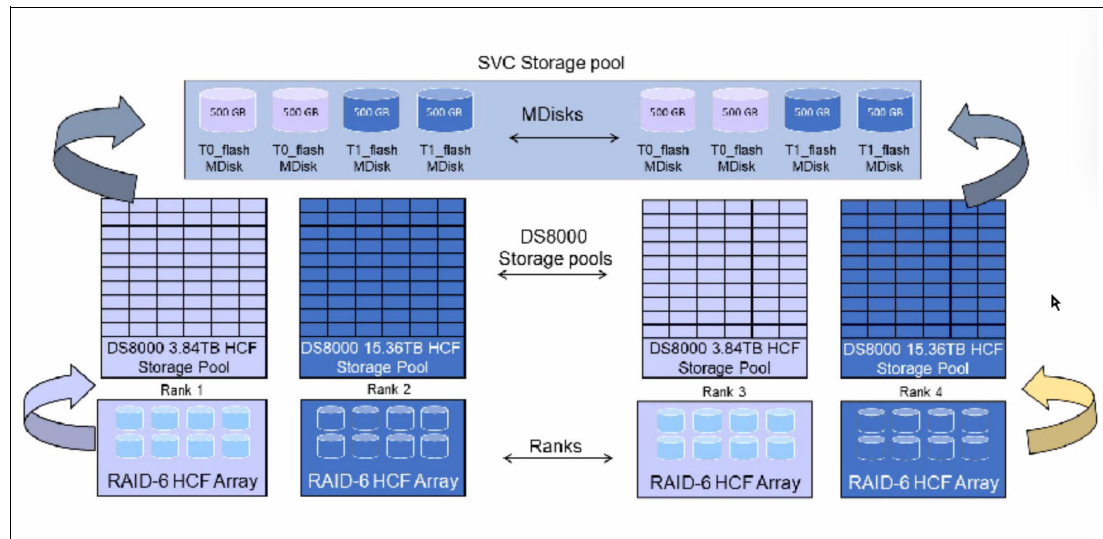


Figure 3-8 Four DS8900F extent pools as one IBM Storage Virtualize storage pool

To determine how many logical volumes must be created to present to IBM Storage Virtualize as MDisks, see 3.3.2, “Guidelines for creating an optimal back-end configuration” on page 217.

3.4.2 Considerations for XIV Gen3

The XIV Gen3 volumes can be provisioned to IBM Storage Virtualize by using iSCSI and FC. However, it is preferable that you implement FC attachment for performance and stability considerations unless a dedicated IP infrastructure for storage is available.

Host options and settings for XIV systems

You must use specific settings to identify IBM Storage Virtualize as a host to XIV systems. An XIV node within an XIV system is a single WWPN. An XIV node is considered to be a single SCSI target. Each host object that is created within the XIV must be associated with the same LUN map.

From an IBM Storage Virtualize perspective, an XIV type 281x controller can consist of more than one WWPN. However, all are placed under one WWNN that identifies the entire XIV system.

Creating a host object for IBM FlashSystem for an XIV

A single host object with all WWPNs of IBM FlashSystem nodes can be created when implementing XIV. This technique makes the host configuration easier to configure. However, the ideal host definition is to consider each node of IBM Storage Virtualize as a host object and create a cluster object to include all nodes or canisters.

When implemented in this manner, statistical metrics are more effective because performance can be collected and analyzed on the IBM Storage Virtualize node level.

A detailed procedure to create a host on XIV is available in *IBM XIV Gen3 with IBM System Storage SAN Volume Controller and Storwize V7000*, REDP-5063.

Volume considerations

As modular storage, XIV storage can be available in a minimum of six modules and up to a maximum of 15 modules in a configuration. Each additional module added to the configuration increases the XIV capacity, CPU, memory, and connectivity.

The XIV system supports the following configurations:

- ▶ 28 - 81 TB when 1-TB drives are used.
- ▶ 55 - 161 TB when 2-TB disks are used.
- ▶ 84 - 243 TB when 3-TB disks are used.
- ▶ 112 - 325 TB when 4-TB disks are used.
- ▶ 169 - 489 TB when 6-TB disks are used.

Figure 3-9 on page 231 shows how XIV configurations vary according to the number of modules that are on the system.

Rack Configuration								
Total number of modules (Configuration type)	6 partial	9 partial	10 partial	11 partial	12 partial	13 partial	14 partial	15 full
Total number of data modules	3	3	4	5	6	7	8	9
Total number of interface modules	3	6	6	6	6	6	6	6
Number of active interface modules	2	4	4	5	5	6	6	6
Interface module 9 state		Disabled	Disabled	Enabled	Enabled	Enabled	Enabled	Enabled
Interface module 8 state		Enabled	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled
Interface module 7 state		Enabled	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled
Interface module 6 state		Disabled	Disabled	Disabled	Disabled	Disabled	Enabled	Enabled
Interface module 5 state		Enabled	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled
Interface module 4 state		Enabled	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled
FC ports	8	16	16	20	20	24	24	24
iSCSI ports (1 Gbps – mod 114)	6	14	14	16	16	22	22	22
iSCSI ports (10 Gbps – mod 214)	4	8	8	10	10	12	12	12
Number of disks	72	108	120	132	144	156	168	180
Usable capacity (1 / 2 / 3 / 4 / 6 TB)	28 TB	44 TB	51 TB	56 TB	63 TB	67 TB	75 TB	81 TB
	55 TB	88 TB	102 TB	111 TB	125 TB	134 TB	149 TB	161 TB
	84 TB	132 TB	154 TB	168 TB	190 TB	203 TB	225 TB	243 TB
	112 TB	177 TB	207 TB	225 TB	254 TB	272 TB	301 TB	325 TB
	169 TB	267 TB	311 TB	338 TB	382 TB	409 TB	453 TB	489 TB
# of CPUs (one per Module)	6	9	10	11	12	13	14	15
Memory (24 GB per module w 1/2/3 TB)	144 GB	216 GB	240 GB	264 GB	288 GB	312 GB	336 GB	360 GB
Memory (48 GB per module w 4/6 TB)	288 GB	432 GB	480 GB	480 GB	528 GB	576 GB	624 GB	720 GB
(Optional for 1, 2, 3, 4, 6 TB XIVs) 400 GB Flash Cache	2.4 TB	3.6 TB	4.0 TB	4.4 TB	4.8 TB	5.2 TB	5.6 TB	6.0 TB
(Optional for 4, 6 TB XIVs) 800 GB Flash Cache	4.8 TB	7.2 TB	8.0 TB	8.8 TB	9.2 TB	10.4 TB	11.2 TB	12.0 TB
Power (kVA) - Model 281x-214 / with SSD	2.5 / 2.6	3.6 / 3.9	4.0 / 4.3	4.3 / 4.6	4.7 / 5.09	5.0 / 5.4	5.5 / 5.8	5.8 / 6.2

Figure 3-9 XIV rack configuration: 281x-214

Although XIV has its own queue depth characteristics for direct host attachment, the best practices that are described in 3.3.2, “Guidelines for creating an optimal back-end configuration” on page 217 are preferred when you virtualize XIV with IBM Storage Virtualize.

Table 3-12 lists the suggested volume sizes and quantities for IBM Storage Virtualize on the XIV systems with different drive capacities.

Table 3-12 XIV minimum volume size and quantity recommendations

Modules	XIV host ports	Volume size (GB) 1 TB drives	Volume size (GB) 2 TB drives	Volume size (GB) 3 TB drives	Volume size (GB) 4 TB drives	Volume size (GB) 6 TB drives	Volume quantity	Volumes to XIV host ports
6	4	1600	3201	4852	6401	9791	17	4.3
9	8	1600	3201	4852	6401	9791	27	3.4
10	8	1600	3201	4852	6401	9791	31	3.9
11	10	1600	3201	4852	6401	9791	34	3.4
12	10	1600	3201	4852	6401	9791	39	3.9
13	12	1600	3201	4852	6401	9791	41	3.4
14	12	1600	3201	4852	6401	9791	46	3.8
15	12	1600	3201	4852	6401	9791	50	4.2

Other considerations

Consider the following restrictions when using the XIV system as back-end storage for IBM Storage Virtualize:

- ▶ Volume mapping

When mapping a volume, you must use the same LUN ID to all IBM Storage Virtualize nodes. Therefore, map the volumes to the cluster, not to individual nodes.

- ▶ XIV storage pools

When creating an XIV storage pool, define the Snapshot Size as 0. Snapshot space does not need to be reserved because it is not recommended that you use XIV snapshots on LUNs mapped as MDisks. The snapshot functions should be used on the IBM Storage Virtualize level.

Because all LUNs on a single XIV system share performance and capacity characteristics, use a single IBM Storage Virtualize storage pool for a single XIV system.

- ▶ Thin provisioning

XIV thin provisioning pools are not supported by IBM Storage Virtualize. Instead, you must use a regular pool.

- ▶ Copy functions for XIV models

You cannot use advanced copy functions, such as taking a snapshot and remote mirroring, for XIV models with disks that are managed by IBM Storage Virtualize.

For more information about the configuration of XIV behind IBM FlashSystem, see *IBM XIV Gen3 with IBM System Storage SAN Volume Controller and Storwize V7000*, REDP-5063.

3.4.3 Considerations for IBM FlashSystem A9000 and A9000R

IBM FlashSystem A9000 and IBM FlashSystem A9000R use industry-leading data-reduction technology that combines inline, real-time pattern matching and removal, data deduplication, and compression. Compression also uses hardware cards inside each grid controller. Compression can easily provide a 2:1 data reduction saving rate on its own, effectively doubling the system storage capacity. Combined with pattern removal and data deduplication services, IBM FlashSystem A9000 and A9000R can easily yield an effective data capacity of five times the original usable physical capacity.

Deduplication can be implemented on IBM Storage Virtualize by attaching an IBM FlashSystem A9000 or A9000R as external storage instead of using IBM Storage Virtualize DRP-level deduplication.

There are several considerations when you are attaching an IBM FlashSystem A9000 or A9000R system as a back-end controller.

Volume considerations

IBM FlashSystem A9000 and A9000R designate resources to data reduction, and because this designation is always on, it is advised that data reduction be done only in IBM FlashSystem A9000 or A9000R and not in the IBM Storage Virtualize cluster. Otherwise, when IBM FlashSystem A9000 or A9000R tries to reduce the data, unnecessary additional latency occurs.

Estimated data reduction is important because it helps determine volume size. Always try to use a conservative data-reduction ratio when attaching IBM FlashSystem A9000 or A9000R because the storage pool goes offline if the back-end storage runs out of capacity.

To determine the controller volume size, complete the following tasks:

- ▶ Calculate effective capacity: Reduce the measured-data reduction ratio (for example, if the Data Reduction Estimator Tool (DRET) provides a ratio of 4:1, use 3.5:1 for calculations) and multiply it to determine physical capacity.
- ▶ Determine the number of connected FC ports by using Table 3-13 and Table 3-14.
- ▶ Volume size is equal to effective capacity that is divided by the number of ports taken twice (effective capacity/path*2).

The remaining usable capacity can be added to the storage pool after the system reaches a stable data reduction ratio.

Table 3-13 Host connections for A9000

Number of controllers	Total FC ports available	Total ports that are connected to SAN Volume Controller	Connected ports
3	12	6	All controllers, ports 1 and 3

Table 3-14 Host connections for A9000R

Grid element	Number of controllers	Total FC ports available	Total ports that are connected to SAN Volume Controller	Connected ports
2	4	16	8	All controllers, ports 1 and 3
3	6	24	12	All controllers, ports 1 and 3
4	8	32	8	Controllers 1 - 4, port 1 Controllers 5 - 8, port 3
5	10	40	10	Controllers 1 - 5, port 1 Controllers 6 - 10, port 3
6	12	48	12	Controllers 1 - 6, port 1 Controllers 7 - 12, port 3

It is important not to run out of hard capacity on the back-end storage because doing so takes the storage pool offline. It is important to closely monitor the IBM FlashSystem A9000 or A9000R. If you start to run out of space, you can use the migration functions of IBM Storage Virtualize to move data to another storage system.

Examples: Consider the following examples:

- ▶ An IBM FlashSystem A9000 with 57 TB of usable capacity, or 300 TB of effective capacity, at the standard 5.26:1 data efficiency ratio.

We ran the data reduction tool on a good representative sample of the volumes that we are virtualizing. We know that we have a data reduction ratio of 4.2:1, and for extra safety, used 4:1 for further calculations. Using 4 x 57 gives you 228 TB. Divide this result by 12 (six paths x 2), and you get 19 TB per volume.

- ▶ A five-grid element IBM FlashSystem A9000R that uses 29 TB flash enclosures has a total usable capacity of 145 TB.

We used 10 paths and did not run any of the estimation tools on the data. However, we know that the host was not compressing the data. We assume a CR of 2:1, so 2 x 145 gives 290, and divided by 20 gives 14.5 TB per volume. In this case, if we see that we are getting a much better data reduction ratio than we planned for, we can always create volumes and make them available to IBM Storage Virtualize.

The biggest concern about the number of volumes is to ensure that there is adequate queue depth. Given that the maximum volume size on the IBM FlashSystem A9000 or A9000R is 1 PB and you are ensuring two volumes per path, you should be able to create a few larger volumes and still have good queue depth and not have numerous volumes to manage.

Other considerations

IBM Storage Virtualize can detect that the IBM FlashSystem A9000 controller is using deduplication technology and show that the Deduplication attribute of the MDisk is Active.

Deduplication status is important because it allows IBM Storage Virtualize to enforce the following restrictions:

- ▶ Storage pools with deduplicated MDisks should contain only MDisks from the same IBM FlashSystem A9000 or IBM FlashSystem A9000R storage controller.
- ▶ Deduplicated MDisks cannot be mixed in an Easy Tier enabled storage pool.

3.4.4 Considerations for IBM FlashSystem 5000, 5100, 5200, 7200, 7300, 9100, 9200, and 9500 and IBM SVC SV1, SV2, and SV3

Recommendations that are described in this section apply to a solution with the IBM FlashSystem family or IBM Storwize family system that is virtualized by IBM Storage Virtualize.

Connectivity considerations

It is expected that N_Port ID Virtualization (NPIV) is enabled on both systems: the system that is virtualizing storage, and the system that works as a back-end. Zone “host” or “virtual” WWPNs of the back-end system to physical WWPNs of the front-end or virtualizing system.

For more information about SAN and zoning best practices, see Chapter 2, “Storage area network guidelines” on page 123.

System layers

IBM Storage Virtualize systems have a concept of system layers. There are two layers: *storage* and *replication*. Systems that are configured into a storage layer can work as back-end storage. Systems that are configured into replication layer can virtualize other IBM Storage Virtualize clusters and use them as back-end controllers.

Systems that are configured with the same layer can be replication partners. Systems in the different layers cannot.

By default, IBM FlashSystem is configured to storage layer. The system layer on IBM FlashSystem can be switched. SVC is configured to the replication layer, and it cannot be changed.

For more information and instructions and limitations, see [System layers](#).

Adapter cages recommendations for IBM FlashSystem 9500

IBM FlashSystem 9500 supports forty-eight 32 gigabit Fibre Channel (GFC), twenty 10 or 25-gigabit Ethernet (GbE), or twelve 100 GbE per enclosure. For best practices, see Table 3-15 on page 235.

Table 3-15 Adapter cages recommendations for IBM FlashSystem 9500

Adapter cages per enclosure	Min or max ports per enclosure	System bandwidth	Recommendations
Two adapter cages One per controller	2/8	33%	Recommended for fewer than 16 drives and modest IOPS requirements.
Four adapter cages Two per controller	4/32	66%	Recommended for high IOPS workloads and systems with more than 16 drives, or remote copy or HyperSwap with a low host port count. An extra memory upgrade should also be used.
Six adapter cages Three per controller	6/48	100%	Recommended for higher port count host attach and high-bandwidth workloads. An extra memory upgrade should also be used.

Adapter recommendations for IBM FlashSystem 7300

For more information about best practices when configuring IBM FlashSystem 7300 adapters, see Table 3-16.

Table 3-16 Adapter recommendations for IBM FlashSystem 7300

Slot # and maximum cards per Enclosure	Ports per enclosure	System bandwidth	Recommendations
1/2	8	50%	Recommended for fewer than 12 drives and modest IOPS requirements.
2/4	16	100%	Recommended for high IOPS workloads and systems with more than 12 drives, or remote copy or HyperSwap with a low host port count. Should be used with memory upgrade 1 or 2.
3/6	24	100%	Recommended for remote copy or HyperSwap configurations with a higher port count host attach. Should be used with memory upgrade 1 or 2.

Adapter cage recommendations for IBM SAN Volume Controller SV3

For best practices when configuring IBM SAN Volume Controller SV3 adapter cages, see Table 3-17.

Table 3-17 Adapter cage recommendations for IBM SVC SV3

Adapter cages per enclosure	Min and max ports per I/O group	System bandwidth	Recommendations
Two adapter cages	2/8	33%	Recommended for modest IOPS requirements and a small number of host ports.
Four adapter cages	4/32	66%	Recommended for high IOPS workloads and systems, or remote copy or HyperSwap with a low host port count. An extra memory upgrade should be used.
Six adapter cages	6/48	100%	Recommended for high host fan-in attach and high-bandwidth workloads. An extra memory upgrade should also be used.

Automatic configuration

IBM FlashSystem family systems can be automatically configured for optimal performance as a back-end storage behind SVC.

An automatic configuration wizard must be used on a system that has no volumes, pools, and host objects that are configured. An available wizard configures internal storage devices, creates volumes, and maps to the host object, which represents the SVC.

Array and disk pool considerations

A back-end IBM FlashSystem family system can have a hybrid configuration containing FCMs and SSD drives, or SSDs and HDDs.

Internal storage that is attached to the back-end system must be joined into RAID arrays. You might need one or more DRAID 6 arrays, depending on the number and the type of available drives. For RAID recommendations, see 3.2.2, “Array considerations” on page 208.

Consider creating a separate disk pool for each type (tier) of storage and use the Easy Tier function on a front-end system. Front-end IBM FlashSystem family systems cannot monitor the Easy Tier activity of the back-end storage.

If Easy Tier is enabled on front- and back-end systems, they independently rebalance the hot areas according to their own heat map. This process causes a rebalance over a rebalance. Such a situation can eliminate the performance benefits of extent reallocation. For this reason, Easy Tier must be enabled only on one level (preferably the front end).

For more information about recommendations about Easy Tier with external storage, see 4.2, “Knowing your data and workload” on page 252.

For most use cases, standard pools are preferred to data-reduction pools on the back-end storage. If planned, the front end performs reduction. Data reduction on both levels is not recommended because it adds processing overhead and does not result in capacity savings.

If Easy Tier is disabled on the back-end as advised here, the back-end IBM FlashSystem pool extent size is not a performance concern.

SCSI UNMAP considerations

When virtualized, IBM Storage Virtualize Storage treats the “virtualizer” system as a host. Back-end UNMAP is enabled by default on all IBM Storage Virtualize systems, and it is a best practice to keep UNMAP turned on for most use cases. Host UNMAP is off by default.

Consider enabling host UNMAP support to achieve better capacity management if the system that is going to be virtualized meets the following qualifications:

- ▶ Contains FCMs.
- ▶ Contains flash only (no HDDs).

Consider leaving host UNMAP disabled to protect a virtualized system from being over-loaded if you are going to virtualize a hybrid system and the storage that will be virtualized uses HDDs.

To turn on or off host UNMAP support, use the `chsystem` CLI command. For more information, see [chsystem](#).

Volume considerations

Volumes in IBM FlashSystem can be created as *striped* or *sequential*. The general rule is to create striped volumes. Volumes on a back-end system must be fully allocated.

To determine the number of volumes to create on a back-end IBM FlashSystem to provide to IBM Storage Virtualize as MDisks, see the general rules that are provided in 3.3.2, “Guidelines for creating an optimal back-end configuration” on page 217. When virtualizing a back-end with HDDs, perform queue depth calculations.

For all-flash solutions, create 32 volumes from the available pool capacity, which can be reduced to 16 or even 8 for small arrays (for example, if you have 16 or fewer flash drives in a back-end pool). For FCM arrays, the number of volumes is also governed by load distribution. 32 volumes out of a pool with an FCM array is recommended.

When choosing volume size, consider which system (front-end or back-end) performs the compression. If data is compressed and deduplicated on the front-end IBM Storage Virtualize system, FCMs cannot compress it further, which results in a 1:1 CR.

Therefore, the back-end volume size is calculated from the pool physical capacity that is divided by the number of volumes (16 or more).

Example: Assume that you have an IBM FlashSystem 9200 with twenty-four 19.2 TB modules. This configuration provides a raw disk capacity of 460 TB, with 10+P+Q DRAID 6 and one distributed spare, and the physical array capacity is 365 TB or 332 TiB.

Because it is not recommended to provision more than 85% of physical flash, we have 282 TiB. Because we do not expect any compression on FCM (the back-end is getting data that is compressed by upper levels), we provision storage to an upper level and assume 1:1 compression, which means we create 32 volumes of $282 \text{ TiB} / 32 = 8.8 \text{ TiB}$ each.

If IBM Storage Virtualize is not compressing data, space savings are achieved with FCM hardware compression. Use compression-estimation tools to determine the expected CR and use a smaller ratio for further calculations (for example, if you expect 4.5:1 compression, use 4.3:1). Determine the volume size by using the calculated effective pool capacity.

Example: Assume that you have an IBM FlashSystem 7300 with twelve 9.6 TB modules. This configuration provides raw disk capacity of 115 TB, with 9+P+Q DRAID 6 and one distributed spare. The physical capacity is 85 TB or 78 TiB.

Because it is not recommended to provision more than 85% of a physical flash, we have 66 TiB. The Compressimator shows that we can achieve a 3.2:1 CR; decreasing in and assuming 3:1, we have 66 TiB x 3 = 198 TiB of effective capacity.

Create 16 volumes of 198 TiB / 16 = 12.4 TiB each. If the CR is higher than expected, we can create and provision more volumes to the front end.

3.4.5 IBM FlashSystem 900 considerations

The main advantage of integrating IBM FlashSystem 900 with IBM Storage Virtualize is to combine the extreme performance of IBM FlashSystem 900 with the IBM Storage Virtualize enterprise-class solution, such as tiering, volume mirroring, deduplication, and copy services.

When you configure IBM FlashSystem 900 as a back-end for IBM Storage Virtualize systems, remember the considerations that are described in this section.

Defining storage

IBM FlashSystem 900 supports up to 12 IBM MicroLatency modules. IBM MicroLatency modules are installed in IBM FlashSystem 900 based on the following configuration guidelines:

- ▶ A minimum of four MicroLatency modules must be installed in the system. RAID 5 is the only supported configuration for IBM FlashSystem 900.
- ▶ The system supports configurations of 4, 6, 8, 10, and 12 MicroLatency modules in RAID 5.
- ▶ All MicroLatency modules that are installed in the enclosure must be identical in capacity and type.
- ▶ For optimal airflow and cooling, if fewer than 12 MicroLatency modules are installed in the enclosure, populate the module bays beginning in the center of the slots and adding the modules on either side until all 12 slots are populated.

The array configuration is performed during system setup. The system automatically creates MDisks or arrays and defines the RAID settings based on the number of flash modules in the system. The default supported RAID level is RAID 5.

Volume considerations

To fully use all IBM Storage Virtualize system resources, create 32 volumes (or 16 volumes if IBM FlashSystem 900 is not fully populated). This way, all CPU cores, nodes, and FC ports of the virtualizer are fully used.

However, one important factor must be considered when volumes are created from a pure IBM FlashSystem 900 MDisks storage pool. IBM FlashSystem 900 can process I/Os much faster than traditional storage. Sometimes, it is even faster than cache operations because with cache, all I/Os to the volume must be mirrored to another node in I/O group.

This operation can take as much as 1 millisecond while I/Os that are issued directly (which means without cache) to IBM FlashSystem 900 can take 100 - 200 microseconds. So, in some rare use cases, it might be recommended to disable the IBM Storage Virtualize cache to optimize for maximum IOPS.

You must keep the cache *enabled* in the following situations:

- ▶ If volumes from the IBM FlashSystem 900 pool are compressed.
- ▶ If volumes from the IBM FlashSystem 900 pool are in a Metro Mirror (MM) and Global Mirror (GM) relationship.
- ▶ If volumes from the IBM FlashSystem 900 pool are in a FlashCopy relationship (either source or target).
- ▶ If the same pool that has MDisks from IBM FlashSystem 900 also contains MDisks from other back-end controllers.

For more information, see *Implementing IBM FlashSystem 900*, SG24-8271.

3.4.6 Path considerations for third-party storage with EMC VMAX, EMC PowerMAX, and Hitachi Data Systems

Many third-party storage options are available and supported. This section describes the multipathing considerations for EMC VMAX, EMC PowerMax, and Hitachi Data Systems (HDS).

Most storage controllers, when presented to IBM Storage Virtualize, are recognized as a single worldwide node name (WWNN) per controller. However, for some EMC VMAX, EMC PowerMAX, and HDS storage controller types, the system recognizes each port as a different WWNN. For this reason, each storage port, when zoned to IBM Storage Virtualize, appears as a different external storage controller.

IBM Storage Virtualize supports a maximum of 16 WWNNs per storage system, so it is preferred to connect up to 16 storage ports.

To determine the number of logical volumes or LUNs to be configured on third-party storage, see 3.3.2, “Guidelines for creating an optimal back-end configuration” on page 217.

3.5 Quorum disks

Note: This section does not cover IP-attached quorum disks. For more information about IP-attached quorum disks, see Chapter 7, “Ensuring business continuity” on page 501.

A system uses a quorum disk for two purposes:

- ▶ To break a tie when a SAN fault occurs and exactly half the nodes that were previously a member of the system are present.
- ▶ To hold a copy of important system configuration data.

After internal drives are prepared to be added to an array or external MDisks become managed, a small portion of the capacity is reserved for quorum data. The size is less than 0.5 GiB for a drive and not less than one pool extent for an MDisk.

Three devices from all available internal drives and managed MDisks are selected for the *quorum disk* role. They store system metadata, which is used for cluster recovery after a disaster. Despite only three devices that are designated as quorum disks, capacity for quorum data is reserved on each of them because the designation might change (for example, if a quorum disk has a physical failure).

Only one of those disks is selected as the active quorum disk. It is used as a tie-breaker. If as a result of a failure the cluster is split in half and both parts lose sight of each other (for example, the inter-site link failed in a HyperSwap cluster with two I/O groups), they appeal to the tie-breaker active quorum device. The half of the cluster nodes that can reach and reserve the quorum disk after the split occurs lock the disk and continue to operate. The other half stops its operation. This design prevents both sides from becoming inconsistent with each other.

The storage device must match following criteria to be considered a quorum candidate:

- ▶ The internal drive or module must follow these rules:
 - It must be a member of an array or be a candidate.
 - Not be in the “Unused” state.
 - The MDisk must be in the “Managed” state. MDisks that are in the “Unmanaged” or “Image” states cannot be quorum disks.
- ▶ External MDisks can be provisioned over only FC.
- ▶ An MDisk must be presented by a disk subsystem (LUNs) that are supported as quorum disks.

The system uses the following rules when selecting quorum devices:

- ▶ Fully connected candidates are preferred over partially connected candidates.
 - In a multiple enclosure environment, MDisks are preferred over drives.
- ▶ Drives are preferred over MDisks.
 - If there is only one control enclosure and no external storage in the cluster, drives are considered first.
- ▶ Drives from a different control enclosure are preferred over a second drive from the same enclosure.
 - If the IBM Storage Virtualize system contains more than one I/O group, at least one of the candidates from each group is selected.
- ▶ NVMe drives are preferred over SAS drives.
 - NVMe drives in a control enclosure are chosen rather than an SAS expansion drive.

To become an active quorum device (tie-breaker device), the storage must be visible to all nodes in a cluster.

In practice, these rules have the following meanings:

- ▶ For an IBM Storage Virtualize system with a single control enclosure, quorums that include active quorum disks are assigned automatically outside the internal drives. No action is required.
- ▶ For an IBM Storage Virtualize system with two or more I/O groups and external storage virtualized, the active quorum is assigned to an external MDisk. None of the internal drives can become the active quorum because they are connected to a single control enclosure and visible only by one pair of nodes.
- ▶ For an IBM Storage Virtualize system with two or more I/O groups and without external storage, no active quorum disk is selected automatically. However, a standard topology cluster in most use cases operates without any issues. For a HyperSwap topology, an IP quorum or FC-attached quorum must be deployed on the third site.

To list the IBM Storage Virtualize Storage quorum devices, run the **lsquorum** command, as shown in Example 3-10.

Example 3-10 The lsquorum command on IBM FlashSystem 9100

```
IBM_IBM FlashSystem:IBM FlashSystem 9100-ITS0:superuser>lsquorum
quorum_index status id name controller_id controller_name active object_type
0           online 4                               no      drive
1           online 1                               yes     drive
2           online 2                               no      drive
```

To move the quorum assignment, use the **chquorum** command. The command is not supported on NVMe drives, so you can move it only *from* the NVMe drive, but not *to* the NVMe drive.



Storage pools

This chapter describes considerations for planning storage pools for an IBM Storage Virtualize system implementation. It explains various pool configuration options, including Easy Tier and data reduction pools (DRPs). It provides best practices on implementation and an overview of some typical operations with managed disks (MDisks).

This chapter includes the following topics:

- ▶ “Introducing pools” on page 244
- ▶ “Knowing your data and workload” on page 252
- ▶ “Storage pool planning considerations” on page 259
- ▶ “Data reduction pools best practices” on page 270
- ▶ “Operations with storage pools” on page 279
- ▶ “Considerations when using encryption” on page 286
- ▶ “Easy Tier and tiered and balanced storage pools” on page 296

4.1 Introducing pools

A *storage pool* or *pool*, sometimes referred to as an *MDisk Group*, is a grouping of storage capacity that is used to provision volumes and logical units (LUs) that can then be made visible to hosts.

IBM Storage Virtualize system supports the following types of pools:

- ▶ Standard pools
- ▶ Data reduction pools (DRP)

Standard pools have been available since the initial release of IBM Storage Virtualize in 2003 and can include fully allocated or thin-provisioned volumes.

Note: With the current hardware and software generation, support for a volume-level compression with standard pools is withdrawn. Standard pools cannot be configured to use IBM Real-time Compression (RtC). Only IBM FlashCore Module (FCM) level compression is available.

With DRP, volume-level compression and deduplication is available too.

Data reduction pools were introduced with Storage Virtualize release 8.1.0. DRPs increase infrastructure capacity usage by employing new efficiency functions and reducing storage costs. The pools enable you to automatically de-allocate (unmap) and reclaim the capacity of thin-provisioned and compressed volumes that contain deleted data. In addition, the pools enable this reclaimed capacity to be reused by other volumes. Data reduction pools allow volume-level compression and cross-volume deduplication.

Either pool type can be made up of different tiers. A tier defines a performance characteristic of that subset of capacity in the pool. Every pool supports three tier types (fastest, average, and slowest). The tiers and their usage are managed automatically by the Easy Tier function inside the pool.

Either pool type can have child pools. With standard pools, child pools allow to dedicate (reserve) a part of pool capacity to a subset of volumes. With DRP, child pools are quotaless. With the both pool types, child pools can have an independent from a parent pool throttle setting (performance limit) and an independent provisioning policy.

4.1.1 Standard pools

Standard pools (also referred to as traditional storage pools) virtualize assigned to it back-end storage as a set of fixed-size allocation units, which are called extents. Subsets of extents are be joined into volumes that can be assigned (mapped) to hosts as LUNs. For detailed information about guidelines for implementing standard pools, see 4.2, “Knowing your data and workload” on page 252.

You can create child pools within a standard pool. A pool, which contains child pools is referred to as a *parent pool*. A parent pool has all the capabilities and functions of a regular pool, but a part of its capacity is reserved for a child. A *child* pool is a logical subdivision of a storage pool or MDisk group. Like a parent pool, a child pool supports volume creation and migration.

When you create a child pool in a standard parent pool, you must specify a capacity limit for the child pool. This limit allows for a quota of capacity to be allocated to the child pool. This

capacity is reserved for the child pool and subtracts from the available capacity in the parent pool.

A child pool inherits its tier setting from the parent pool. Changes to a parent's tier setting are inherited by child pools. The child pool also inherits Easy Tier status, pool status, capacity information, and back-end storage information. The I/O activity of a parent pool is the sum of the I/O activity of itself and the child pools.

Parent pools

Parent pools receive their capacity from MDisks. To track the space that is available on an MDisk, the system divides each MDisk into chunks of equal size. These chunks are called *extents* and they are indexed internally. The choice of extent size affects the total amount of storage that is managed by the system. The extent size remains constant throughout the lifetime of the parent pool.

All MDisks in a pool are split into extents of the same size. Volumes are created from the extents that are available in the pool. You can add MDisks to a pool at any time to increase the number of extents that are available for new volume copies or to expand volume copies. The system automatically balances volume extents between the MDisks to provide the best performance to the volumes by using Easy Tier function.

Choose your extent size wisely according to your future needs. A small extent size limits your overall usable capacity, but a larger extent size can waste storage. For example, if you select an extent size of 8 GiB but then create only a 6 GiB volume, one entire extent is allocated to this volume (8 GiB) and 2 GiB is unused.

When you create or manage a standard pool, consider the following general guidelines:

- ▶ An MDisk can be associated with only one pool.
- ▶ You can add only MDisks that are in unmanaged mode to a parent pool. When MDisks are added to a parent pool, their mode changes from unmanaged to managed.
- ▶ Ensure that all MDisks that are allocated to the same tier of a parent pool have the same redundant array of independent disks (RAID) level. This configuration ensures that the same resiliency is maintained across that tier. Similarly, for performance reasons, do not mix RAID types within a tier. The performance of all volumes is reduced to the lowest achiever in the tier, and a mismatch of tier members can result in I/O convoying effects where everything is waiting on the slowest member.
- ▶ You can delete MDisks from a parent pool under the following conditions:
 - The volumes are not using any of the extents that are on the MDisk.
 - Enough free extents are available elsewhere in the pool to move extents that are in use from this MDisk.
- ▶ If the parent pool is deleted, you cannot recover the mapping that existed between extents that are in the pool or the extents that the volumes use. If the parent pool includes associated child pools, you must delete the child pools first and return its extents to the parent pool. After the child pools are deleted, you can delete the parent pool. The MDisks that were in the parent pool are returned to unmanaged mode and can be added to other parent pools. Because the deletion of a parent pool can cause a loss of data, you must force the deletion if volumes are associated with it.

Note: Deleting a child or parent pool is unrecoverable.

If you force-delete a pool, all volumes in that pool are deleted, even if they are mapped to a host and are still in use. Use extreme caution when force-deleting pool objects because volume-to-extent mapping cannot be recovered after the delete is processed.

Force-deleting a storage pool is possible only with command-line interface (CLI) tools. For more information, see the man page for the `rmmdiskgrp` command.

- ▶ You should specify a warning capacity for a pool. A warning event is generated when the amount of space that is used in the pool exceeds the warning capacity. The warning threshold is especially useful with thin-provisioned volumes that are configured to automatically use space from the pool.
- ▶ Volumes are associated with just one pool, except during any migration between parent pools. However, volume copies of the same volume can be in different pools.
- ▶ Volumes that are allocated from a parent pool are by default striped across all the storage assigned into that parent pool. Wide striping can provide performance benefits.
- ▶ You cannot use the volume migration functions to migrate volumes between parent pools that feature different extent sizes. However, you can use volume mirroring to move data to a parent pool that has a different extent size.
- ▶ When you delete a pool with mirrored volumes, consider the following points:
 - If the volume is mirrored and the synchronized copies of the volume are all in the same pool, the mirrored volume is destroyed when the storage pool is deleted.
 - If the volume is mirrored and a synchronized copy exists in a different pool, the volume copy remains after the pool is deleted.

You might not be able to delete a pool or child pool if Volume Delete Protection is enabled. In version 8.3.1 and later, Volume Delete Protection is enabled by default. However, the granularity of protection is improved: You can now specify Volume Delete Protection to be enabled or disabled on a per-pool basis rather than on a system basis as was previously the case.

Child pools

Instead of being created directly from MDisks, child pools are created from existing capacity that is allocated to a parent pool. As with parent pools, volumes can be created that specifically use the capacity that is allocated to the child pool. Child pools are like parent pools with similar properties and can be used for volume copy operation.

Child pools are created with fully allocated physical capacity, that is, the physical capacity that is applied to the child pool is reserved from the parent pool, as though you created a fully allocated volume of the same size in the parent pool.

The allocated capacity of the child pool must be smaller than the free capacity that is available to the parent pool. The allocated capacity of the child pool is no longer reported as the *free* space of its parent pool. Instead, the parent pool reports the entire child pool as *used* capacity. You must monitor the used capacity (instead of the free capacity) of the child pool instead.

When you create or work with a child pool, consider the following information:

- ▶ As with parent pools, you can specify a warning threshold that alerts you when the capacity of the child pool is reaching its upper limit. Use this threshold to ensure that access is not lost when the capacity of the child pool is close to its allocated capacity.
- ▶ Ensure that any child pools that are associated with a parent pool have enough capacity for the volumes that are in the child pool before removing MDisks from a parent pool. The system automatically migrates all extents that are used by volumes to other MDisks in the parent pool to ensure that data is not lost.
- ▶ You cannot shrink the capacity of a child pool to less than its real capacity. The system also resets the warning level when the child pool is shrunk, and issues a warning if the level is reached when the capacity is shrunk.
- ▶ On systems with encryption enabled, child pools can be created to migrate existing volumes in a non-encrypted pool to encrypted child pools. When you create a child pool after encryption is enabled, an encryption key is created for the child pool even when the parent pool is not encrypted. Then, you can use volume mirroring to migrate the volumes from the non-encrypted parent pool to the encrypted child pool.
- ▶ The system supports migrating a copy of volumes between child pools within the same parent pool or migrating a copy of a volume between a child pool and its parent pool. Migrations between a source and target child pool with different parent pools are not supported. However, you can migrate a copy of the volume from the source child pool to its parent pool. Then, the volume copy can be migrated from the parent pool to the parent pool of the target child pool. Finally, the volume copy can be migrated from the target parent pool to the target child pool.
- ▶ Migrating a volume between a parent pool and a child pool (with the same encryption key or no encryption) results in a *nocopy migration*. The data does not move. Instead, the extents are reallocated to the child or parent pool and the accounting of the used space is corrected.
- ▶ Child pools are created automatically by an IBM Spectrum Connect vSphere API for Storage Awareness (VASA) client to implement VMware vSphere Virtual Volumes (VVOLs).
- ▶ A throttle can be assigned to a child pool to limit its IO rate or data rate. Child pool throttle can be used to restrict a set of volumes in it from using more performance resources than it is desired.
- ▶ Child pool provisioning policy can be different from a policy that is assigned to its parent.
- ▶ Child pool can be assigned to an ownership group.

Thin-provisioned volumes in a standard pool

A thin-provisioned volume presents a different capacity to mapped hosts than the capacity that the volume uses in the storage pool. Storage Virtualize system supports thin-provisioned volumes in standard pools.

In standard pools, thin-provisioned volumes are created as a specific volume type, that is, based on capacity-savings criteria. These properties are managed at the volume level. The virtual capacity of a thin-provisioned volume is typically larger than its real capacity. Each system uses the real capacity to store data that is written to the volume, and metadata that describes the thin-provisioned configuration of the volume. As more information is written to the volume, more of the real capacity is used.

The system identifies read operations to unwritten parts of the virtual capacity and returns zeros to the server without the usage of any real capacity. For more information about storage system, pool, and volume capacity metrics, see Chapter 9, “Implementing a storage monitoring system” on page 551.

Thin-provisioned volumes can also help simplify server administration. Instead of assigning a volume with some capacity to an application and increasing that capacity as the needs of the application change, you can configure a volume with a large virtual capacity for the application. Then, you can increase or shrink the real capacity as the application needs change, without disrupting the application or server.

It is important to monitor physical capacity if you want to provide more space to your hosts than is physically available in the pool and pool’s back-end storage. For more information about monitoring the physical capacity of your storage and an explanation of the difference between thin provisioning and over-allocation, see 9.4, “Creating alerts for IBM Storage Control and IBM Storage Insights” on page 606.

Standard pools with IBM FlashCore Modules

There is no pool layer compression available with the standard pools. Volumes can be either fully allocated or thin provisioned. Data compression will happen only on FCM layer, independently of a type of a pool or selected volume capacity savings method.

If you use the compression functions that are provided by the FCM modules in your system as a mechanism to add data reduction to a standard pool while maintaining the maximum performance, take care to understand the capacity reporting, in particular if you want to thin provision on top of the FCMs.

The FCM RAID array reports its written capacity limit, which can be as large as 4:1 to its physical capacity. This capacity is the maximum that can be stored on the FCM array. However, it might not reflect the compression savings that you achieve with your data.

You must first understand your expected compression ratio (CR). In an initial deployment, allocate approximately 50% fewer savings.

For example, you have 100 TiB of physical usable capacity in an FCM RAID array before compression. Your compression results show savings of approximately 2:1, which suggests that you can write 200 TiB of volume data to this RAID array.

Start at 150 TiB of volumes that are mapped to hosts. Monitor the real compression rates and usage and over time add in the other 50 TiB of volume capacity. Be sure to leave spare space for unexpected growth, and consider the guidelines that are outlined in 3.2, “Arrays” on page 206.

If you often over-provision your hosts at much higher rates, you can use a standard pool and create thin-provisioned volumes in that pool. However, be careful that you do not run out of physical capacity. You now must monitor the back-end array used capacity. In essence, you are double accounting with the thin provisioning, that is, expecting 2:1 on the FCM compression, and then whatever level you over-provision at the volumes.

If you know that your hosts rarely grow to use the provisioned capacity, this process can be safely done. However, the risk comes from run-away applications (writing large amounts of capacity) or an administrator suddenly enabling application encryption and writing to fill the entire capacity of the thin-provisioned volume.

4.1.2 Data reduction pools

IBM Storage Virtualize uses data reduction pools that incorporate deduplication and hardware-accelerated compression technology, plus volume capacity reclamation support. DRPs also use all the thin-provisioning and data-efficiency features that you expect from IBM Storage Virtualize storage to potentially reduce your capital expenditure (CapEx) and operational expenditure (OpEx). All these benefits extend to over 500 heterogeneous storage arrays from multiple vendors.

DRPs were designed with space reclamation being a fundamental consideration. DRPs provide the following benefits:

- ▶ Log Structured Array (LSA) allocation (redirect on all overwrites)
- ▶ Garbage collection to free whole extents
- ▶ Fine-grained (8 KB) chunk allocation/de-allocation within an extent
- ▶ Unmap commands end-to-end support with automatic space reclamation
- ▶ Support for compression
- ▶ Support for deduplication
- ▶ Support for traditional fully allocated volumes

Data reduction increase storage efficiency and reduce storage costs, especially for flash storage. Data reduction reduces the amount of data that is stored on external storage systems and internal drives by compressing and deduplicating capacity and reclaiming capacity that is no longer in use.

Quotaless data reduction child pool

The concepts and high-level description of parent-child pools are the same as for standard pools, described in “Child pools” on page 246, with a few exceptions:

- ▶ You cannot define capacity or a quota for a DRP child pool.
- ▶ A DRP child pool shares the same encryption key as its parent.
- ▶ Capacity warning levels cannot be set on a DRP child pool. Instead, you must rely on the warning levels of the DRP parent pool.
- ▶ A DRP child pool consumes space from the DRP parent pool as volumes are written to it.
- ▶ VVOL for DRP is not supported at the time of writing.

DRP shares capacity between volumes (when deduplication is used), it is impossible to attribute capacity ownership of a specific grain to a specific volume because it might be used by two more volumes, which the is value proposition of deduplication. This process results in the differences between standard and DRP child pools.

Object-based access control (OBAC) or multi-tenancy can be applied to DRP child pools or volumes because OBAC requires a child pool to function.

DRP internal details

DRP consists of several logical containers (volumes), that store data and metadata. It is important to understand how these metadata volumes are used and mapped to user (host) volumes.

The internal layout of a DRP is different from a standard pool. A standard pool creates volume objects within the pool. Some fine-grained internal metadata is stored within a thin-provisioned or real-time-compressed volume in a standard pool. Overall, the pool contains volume objects.

A DRP reports volumes to the user in the same way as a standard pool. All volumes in a single DRP use the same Customer Data Volume to store their data. Therefore, deduplication is possible across volumes in a single DRP. There is a Directory Volume for each user volume that is created within the pool. The directory points to grains of data that is stored in the Customer Data Volume.

Other internal volumes are created, one per DRP. There is one Journal Volume per I/O group that can be used for recovery purposes and to replay metadata updates if needed. There is one Reverse Lookup Volume per I/O group that is used by garbage collection.

Figure 4-1 shows the difference between DRP volumes and volumes in standard pools.

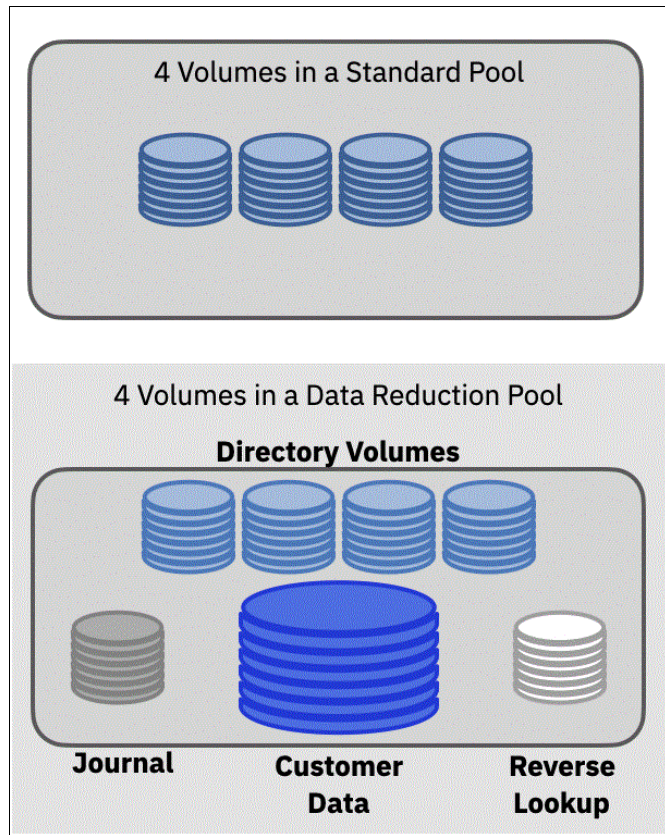


Figure 4-1 Standard and data reduction pool volumes

The Customer Data Volume normally uses greater than 97% of pool capacity. The I/O pattern is a large sequential write pattern (256 KB) that is coalesced into full stride writes, and you typically see a short, random read pattern.

Directory Volumes occupy approximately 1% of pool capacity. They typically have a short 4 KB random read/write I/O. The Journal Volume occupies approximately 1% of pool capacity, and shows large sequential write I/O (256 KB typically).

Journal Volumes are only read for recovery scenarios (for example, T3 recovery). Reverse Lookup Volumes are used by the garbage-collection process and occupy less than 1% of pool capacity. Reverse Lookup Volumes have a short, semi-random read/write pattern.

The primary task of garbage collection (see Figure 4-2) is to reclaim space, that is, to track all the regions that were invalidated and make this capacity usable for new writes. As a result of compression and deduplication, when you overwrite a host-write, the new data does not always use the same amount of space that the previous data used. This issue leads to the writes always occupying new space on back-end storage while the old data is still in its original location.

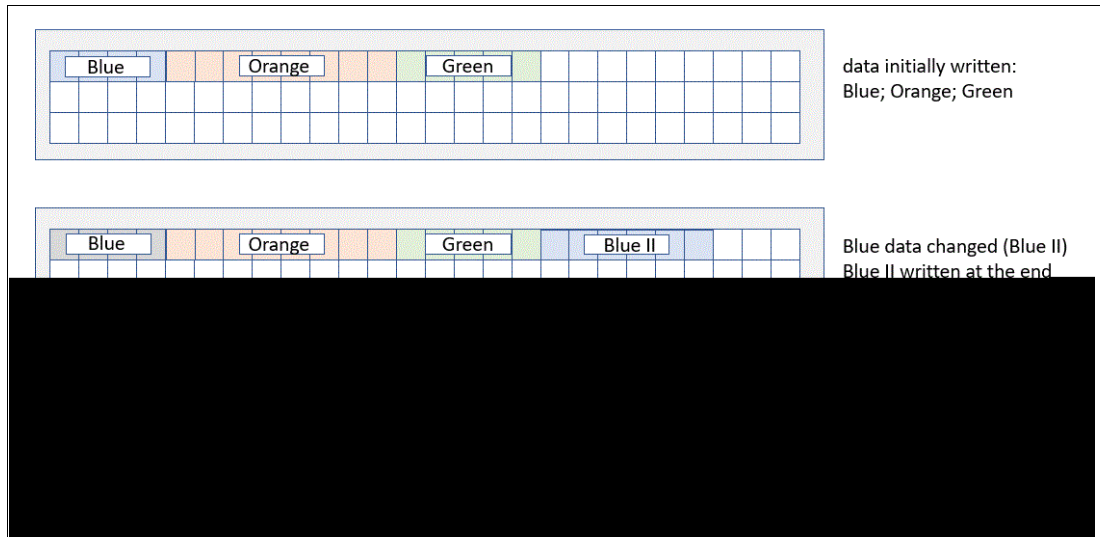


Figure 4-2 Garbage-collection principle

Stored data is divided into regions. As data is overwritten, a record is kept of which areas of those regions were invalidated. Regions that have many invalidated parts are potential candidates for garbage collection. When most of a region has invalidated data, it is inexpensive to move the remaining data to another location, which frees the whole region.

DRPs include built-in services to enable garbage collection of unused blocks. Therefore, many smaller unmaps end up enabling a much larger chunk (extent) to be freed back to the pool. Trying to fill small holes is inefficient because too many I/Os are needed to keep reading and rewriting the directory. Therefore, garbage collection waits until an extent has many small holes and moves the remaining data into the extent, compacts the data, and rewrites the data. When there is an empty extent, it can be freed back to the virtualization layer (and back-end with unmap) or start writing into the extent with new data (or rewrites).

The reverse lookup metadata volume tracks the extent usage, or more importantly the holes that are created by overwrites or unmaps. Garbage collection looks for extents with the most unused space. After a whole extent has all valid data moved elsewhere, it can be freed back to the set of unused extents in that pool or reused for new written data.

Garbage collection needs free regions to move data during its operation, so it is suggested that you size pools to keep a specific amount of free capacity available. This best practice ensures that there is some free space for garbage collection. For more information, see 4.4.6, “Understanding capacity use in a data reduction pool” on page 276.

Fully allocated volumes in a DRP

It is possible to create fully allocated volumes in a DRP.

A fully allocated volume uses the entire capacity of the volume. When the volume is created, the space is reserved (used) from the DRP and not available for other volumes in the DRP.

Data is not deduplicated or compressed in a fully allocated volume. Similarly, because the volume does not use the internal fine-grained allocation functions, technically it operates in the same way as a fully allocated volume in a standard pool.

Compressed and deduplicated volumes in a DRP

It is possible to create compressed-only volumes in a DRP. A compressed volume is thin-provisioned by its nature. A compressed volume uses only its compressed data size in the pool. The volume grows only as you write data to it.

Deduplication in DRP runs on a pool basis, so deduplication is performed across all the volumes in a single data reduction pool. The DRP first looks for deduplication matches, and then it compresses the data before writing to the storage.

It is possible to create a deduplicated-only volume in a DRP, however it is not the best practice. A deduplicated volume is thin-provisioned in nature. The extra processing that is required to compress the deduplicated on block is minimal because of the hardware compression acceleration, so it is recommended that you create a compressed and deduplicated volume rather than only a deduplicated volume.

Pool level compression is independent from a backend-level (FCM) compression. Compression can occur on both levels. In this case FCMs will get a data that is already compressed. Only small compression savings, achieved by metadata compressions, will occur on the FCM level.

Thin-provisioned only volumes in a DRP

Thin-provisioned volumes use the fine-grained allocation functions of a DRP. They provide capacity savings by allocating space only for data that was written by a host, and will de-allocate (unmap) capacity when host deletes it data.

It is not recommended to use thin-only volumes in DRP on an FCM backend. All hardware platforms that support FCMs, have compression acceleration hardware, therefore performance penalty for compression will be minimal. At the same time, using pool level compression on top of FCMs simplifies capacity monitoring. So, it is not recommended to use thin-only volumes in such configurations, as there is no practical benefit in that.

Note: When the back-end storage is thin-provisioned or data-reduced, the GUI will not offer the option to create only thin-provisioned volumes in a DRP. GUI aims to comply with the best practice, and thin-only volume on FCM backend can cause capacity monitoring difficulties. You can still use CLI to create volumes with this capacity savings type if required.

4.2 Knowing your data and workload

Pool type and pool settings selection depend on characteristics of the data that will be written on it. It is also important to know in advance how the data on a pool is accessed and what performance is expected.

4.2.1 Determining whether your data is compressible

Compression is performed on FCM level if your platform has them, also compression can be configured to run on DRP level on all platforms except FlashSystem 5015. The compression algorithm is used to reduce the on-disk footprint of data that is written to back-end by thin provisioning.

In Storage Virtualize, both FCM compression and DRP compression are inline, which means that it happens when the data is being written, rather than an attempt to compress data as a background task.

In DRP, compression can be enabled on a per-volume basis, and thin provisioning is a prerequisite. The input I/O is split into fixed 8 KiB blocks for internal handling, and compression is performed on each block. Then, these compressed blocks are consolidated into 256 K chunks of compressed data for consistent write performance by allowing the cache to build full stride writes, which enable the most efficient RAID throughput.

Data compression techniques depend on the type of data that must be compressed and on the needed performance. Effective compression savings generally rely on the accuracy of your planning and understanding whether the specific data is compressible or not.

The compression is lossless, that is, data is compressed without losing any of the data. The original data can be recovered after the compress or expand cycle. Good compression savings might be achieved in the data types such as:

- ▶ Virtualized Infrastructure
- ▶ Database and data warehouse
- ▶ Home directory, shares, and shared project data
- ▶ CAD/CAM
- ▶ Oil and gas data
- ▶ Log data
- ▶ Software development
- ▶ Text and some picture files

However, if the data is compressed in some cases, the savings are less, or even negative. Pictures (for example, GIF, JPG, and PNG), audio (MP3 and WMA) and video or audio (AVI and MPG), and even compressed databases data might not be good candidates for compression.

Table 4-1 lists the compression ratio that is expected for the common data types.

Table 4-1 Compression ratios of common data types

Data types/applications	Compression ratio
Databases	Up to 80%
Server or desktop virtualization	Up to 75%
Engineering data	Up to 70%
Email	Up to 80%

If the data is encrypted by host or application before it is written to the Storage Virtualize system, it cannot be compressed. Compressing already encrypted data does not result in much savings because it contains pseudo-random data. The compression algorithm relies on patterns to gain efficient size reduction. Because encryption destroys such patterns, the compression algorithm would be unable to provide much data reduction.

Note: Saving assumptions that are based on the type of data are imprecise. Therefore, you should determine compression savings with the proper tools. See 4.2.3, “Data reduction estimation tools” on page 254 for tools information.

4.2.2 Determining whether your data is a deduplication candidate

Deduplication can be performed only on DRP level. It runs cross-volume, and process all volumes that are configured to be deduplicated in a single data reduction pool.

Deduplication is done by using hash tables to identify previously written copies of data. If duplicates are found, instead of writing the data to disk, the algorithm references the previously found data.

- ▶ Deduplication uses 8 KiB deduplication grains and an SHA-1 hashing algorithm.
- ▶ DRPs build 256 KiB chunks of data consisting of multiple de-duplicated and compressed 8 KiB grains.
- ▶ DRPs write contiguous 256 KiB chunks for efficient write streaming with the capability for cache and RAID to operate on full stride writes.
- ▶ DRPs provide deduplication and then compress capability.
- ▶ The scope of deduplication is within a DRP within an I/O Group.

Some environments have data with high deduplication savings, and are therefore candidates for deduplication.

Good deduplication savings can be achieved in several environments, such as virtual desktop and some virtual machine (VM) environments. Therefore, these environments might be good candidates for deduplication.

IBM provides the Data Reduction Estimator Tool (DRET) to help determine the deduplication capacity-saving benefits.

4.2.3 Data reduction estimation tools

IBM provides two tools to estimate the savings when you use data reduction technologies:

- ▶ Comprestimator

This tool is built into Storage Virtualize code. It reports the expected compression savings on a per-volume basis in the GUI and CLI. It also exists as a separate executable.

- ▶ DRET

The DRET tool must be installed on a host and used to scan the volumes that are mapped to a host. It is primarily used to assess the deduplication savings. The DRET tool is the most accurate way to determine the estimated savings. However, it must scan all your volumes to provide an accurate summary.

Comprestimator

Comprestimator is available in the following ways:

- ▶ As a stand-alone, host-based CLI utility. It can be used to estimate the expected compression for block volumes where you do not have an IBM Storage Virtualize product providing those volumes.
- ▶ Integrated into IBM Storage Virtualize code.

Host-based Comprestimator

The Comprestimator is a CLI and host-based utility that can be used to estimate an expected compression rate for block devices. The tool and instructions can be downloaded from [IBM FlashSystem Comprestimator](#).

Integrated Comprestimator

IBM Storage Virtualize also features an integrated Comprestimator tool that is available through the management GUI and CLI. If you want to apply capacity efficiency features to volumes that already exist in the system, you can use this tool to evaluate whether compression will generate capacity savings.

To access the Comprestimator tool in the management GUI, select **Volumes** → **Volumes**.

If you want to analyze all the volumes in the system, select **Actions** → **Capacity Savings** → **Estimate Compression Savings**.

If you want to evaluate only the capacity savings of selected volumes, select a list of volumes and select **Actions** → **Capacity Savings** → **Analyze**, as shown in Figure 4-3.

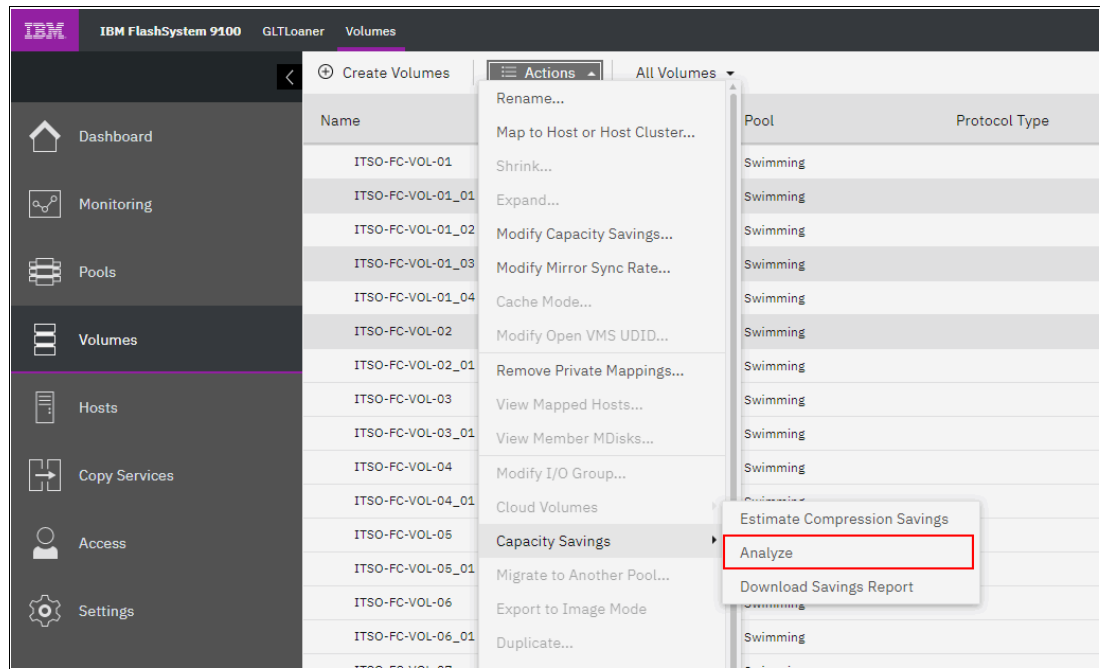


Figure 4-3 Capacity savings analysis

To display the results of the capacity savings analysis, select **Actions** → **Capacity Savings** → **Download Savings Report**, as shown in Figure 4-3, or enter the command `lsvdiskanalysis` in the CLI, as shown in Example 4-1.

Example 4-1 Results of capacity savings analysis

```
IBM_IBM FlashSystem:superuser>lsvdiskanalysis TESTVOL01
id 64
name TESTVOL01
state estimated
started_time 201127094952
analysis_time 201127094952
capacity 600.00GB
thin_size 47.20GB
```

```

thin_savings 552.80GB
thin_savings_ratio 92.13
compressed_size 21.96GB
compression_savings 25.24GB
compression_savings_ratio 53.47
total_savings 578.04GB
total_savings_ratio 96.33
margin_of_error 4.97

```

You can customize the Volume view to view estimation results in a convenient way to help make your decision, as shown in Figure 4-4.

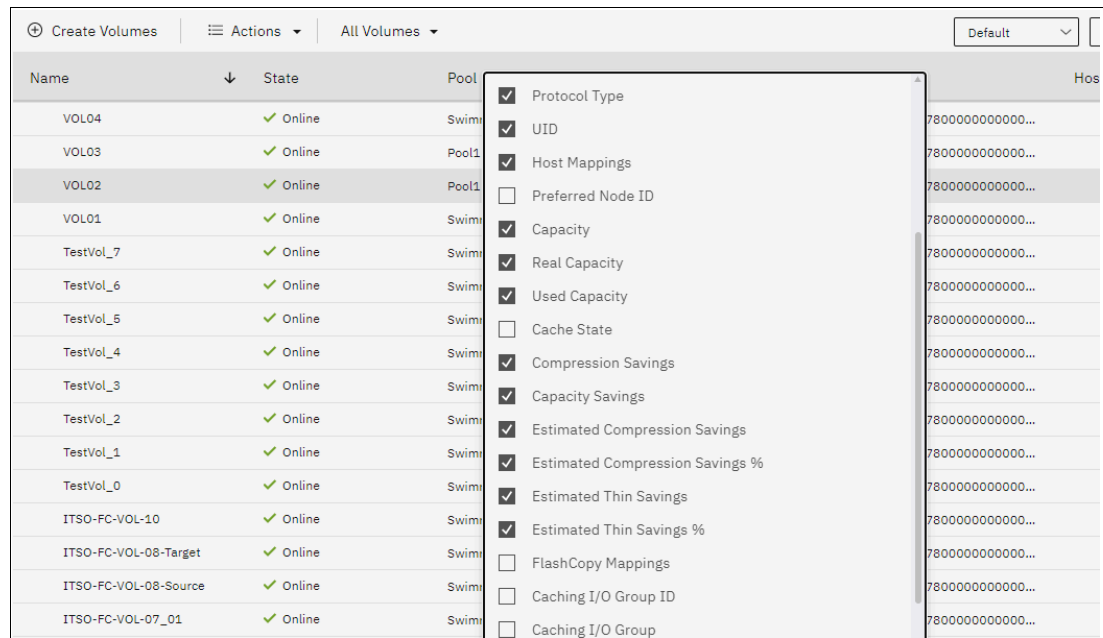


Figure 4-4 Customized view

Comprestimator is always enabled and running in background, so you can view the expected capacity savings in the main dashboard view, pool views, and volume views. However, on older codes it needs to be started manually by triggering the “estimate” or “analyze” tasks.

Data Reduction Estimator Tool

IBM provides DRET to support both deduplication and compression. The host-based CLI tool scans target workloads on various older storage arrays (from IBM or another company), merges all scan results, and then provides an integrated system-level data reduction estimate for your system planning.

DRET uses advanced mathematical and statistical algorithms to perform an analysis with a low memory “footprint”. The utility runs on a host that can access the devices to be analyzed. It performs only read operations, so it has no effect on the data that is stored on the device. Depending on the configuration of the environment, in many cases the DRET is used on more than one host to analyze more data types.

It is important to understand block device behavior when analyzing traditional (fully allocated) volumes. Traditional volumes that were created without initially zeroing the device might contain traces of old data on the block device level. Such data is not accessible or viewable on the file system level. When the DRET is used to analyze such volumes, the expected reduction results reflect the savings rate to be achieved for all the data on the block device level, including traces of old data.

Regardless of the block device type being scanned, it is also important to understand a few principles of common file system space management. When files are deleted from a file system, the space they occupied before the deletion becomes free and available to the file system. The freeing of space occurs even though the data on disk was not removed, but the file system index and pointers were updated to reflect this change.

When DRET is used to analyze a block device that is used by a file system, all underlying data in the device is analyzed regardless of whether this data belongs to files that were already deleted from the file system. For example, you can fill a 100 GB file system and use 100% of the file system, and then delete all the files in the file system, which makes it 0% used. When scanning the block device that is used for storing the file system in this example, DRET (or any other utility) can access the data that belongs to the files that are deleted.

To reduce the impact of the block device and file system behavior, it is recommended that you use DRET to analyze volumes that contain as much active data as possible rather than volumes that are mostly empty of data. This usage increases the accuracy level and reduces the risk of analyzing old data that is deleted, but might still have traces on the device.

DRET can be downloaded from [IBM Data Reduction Estimator Tool \(DRET\) for SVC, Storwize and FlashSystem products](#).

Example 4-2 shows the DRET CLI.

Example 4-2 DRET command-line interface

```
Data-Reduction-Estimator -d <device> [-x Max MBps] [-o result data filename] [-s Update interval] [--command scan|merge|load|partialscan] [--mergefiles Files to merge] [--loglevel Log Level] [--batchfile batch file to process] [-h]
```

Note: According to the results of the DRET, use DRPs to use the available data deduplication savings unless performance requirements exceed what DRP can deliver.

Do not enable deduplication if the data set is not expected to provide deduplication savings.

4.2.4 Determining the workload and performance requirements

An important factor of sizing and planning for an IBM Storage Virtualize environment is the knowledge of the workload characteristics of that specific environment.

Sizing and performance is affected by the following workloads (among others):

► Read/write ratio

Read/write (%) ratio affects performance because higher writes cause more IOPS to the DRP. To effectively size an environment, the read/write ratio should be considered. During a write I/O, when data is written to the DRP, it is stored on the data disk, the forward lookup structure is updated, and the I/O is completed.

DRPs use metadata. Even when volumes are not in the pool, some of the space in the pool is used to store the metadata. The space that is allocated to metadata is relatively small. Regardless of the type of volumes that the pool contains, metadata is always stored separately from customer data.

In DRPs, the maintenance of the metadata results in I/O amplification. I/O amplification occurs when a single host-generated read or write I/O results in more than one back-end storage I/O request because of advanced functions. A read request from the host results in two I/O requests: a directory lookup and a data read. A write request from the host results in three I/O requests: a directory lookup, a directory update, and a data write. Therefore, DRPs create *more IOPS* on the FCMs or drives.

- ▶ Block size

The concept of a block size is simple and the impact on storage performance might be distinct. Block size effects might have an impact on overall performance. Therefore, consider that larger blocks affect performance more than smaller blocks. Understanding and considering block sizes in the design, optimization, and operation of the storage system-sizing leads to more predictable behavior of the entire environment.

Note: Where possible, limit the maximum transfer size that is sent to IBM FlashSystem to no more than 256 KiB. This limitation is a best practice and not specific to only DRP.

- ▶ IOPS, MBps, and response time

Storage constraints are IOPS, throughput, and latency, and it is crucial to correctly design the solution or plan for a setup for speed and bandwidth. Suitable sizing requires knowledge about the expected requirements.

- ▶ Capacity

During the planning of an IBM Storage Virtualize environment, both written and physical capacity must be sized. FlashCore Modules have a maximum written capacity limit, specifying the maximum amount of uncompressed raw data that can be written onto it, and physical capacity limit, which shows how much data it can store after compression. Both those limits need to be taken into consideration. For optimal performance and safety, the recommendation is to use the DRP to a maximum of 85%.

Before planning a new environment, consider monitoring the storage infrastructure requirements with monitoring or management software (such as IBM Spectrum Control or IBM Storage Insights). At busy times, the peak workload (such as IOPS or MBps) and peak response time provide you with an understanding of the required workload plus expected growth. Also, consider allowing enough room for the performance that is required during planned and unplanned events (such as upgrades and possible defects or failures).

It is important to understand the relevance of application response time rather than internal response time with required IOPS or throughput. Typical online transaction processing (OLTP) applications require IOPS and low latency.

Do not place capacity over performance while designing or planning a storage solution. Even if capacity might be sufficient, the environment can suffer from low performance. Deduplication and compression might satisfy capacity needs, but aim for performance and robust application performance.

To size an IBM Storage Virtualize environment, your IBM account team or IBM Business Partner must access IBM Storage Modeller (StorM). The tool can be used to determine whether the system can provide suitable bandwidth and latency with the requested data efficiency features enabled.

4.3 Storage pool planning considerations

The implementation of storage pools in a Storage Virtualize system requires a holistic approach that involves application availability and performance considerations. Usually, a tradeoff between these two aspects must be accounted for.

The main best practices in the storage pool planning activity are described in this section. Most of these practices apply to both standard and DRP pools, except where otherwise specified. For more specific best practices for DRPs, see 4.4, “Data reduction pools best practices” on page 270.

4.3.1 Selecting pool type

Each of two available pool types, standard pools and data reduction pools, has its own traits that are specific only to it. It is important to choose right pool type at the system planning phase because if it selected wrong, the system might not fulfill your business needs.

Depending on the requirements and configuration, you can also configure multiple pools, and pools of a different type can coexist on a single system.

Consider information provided below to select a proper pool type for your data.

Capacity savings methods

The main difference between pool types is an availability of various capacity efficiency features. Pool-level compression and deduplication are available only with data reduction pools.

Note: With the previous generation of IBM Storage Virtualize hardware, there was an RTC compression method supported for standard pools. However, with the current hardware generation, it is no longer available.

Capacity saving methods available with data reduction pools:

- ▶ None (Fully Allocated)
- ▶ Thin-provisioned
- ▶ Thin-provisioned and compressed
- ▶ Thin-provisioned and deduplicated
- ▶ Thin-provisioned, compressed, and deduplicated

Capacity saving methods available with standard pools:

- ▶ None (Fully Allocated)
- ▶ Thin-provisioned

Capacity savings method is assigned per-volume basis in the pool. Volumes with the different methods can be mixed in a single pool without any limitations.

If the system is equipped with FlashCore Modules, compression on FCM level is always enabled and works independently of the pool type. This means that even with standard pools, data will be compressed on FCMs.

Performance of Fully Allocated (FA) volumes in both pool types is the same with both pool types. Any type of the pool can be used.

The way how thin provisioning is implemented is different: in DRP, thin provisioning is provided by a use of DRP's LSA data structures, and carries significant computation and I/O overhead comparing to standard pools. Consider the following if you plan to use only thin provisioning without pool-level compression or deduplication:

- ▶ On entry-level systems that do not support compression: use thin provisioning in standard pools.
- ▶ On entry-level systems that do not have compression acceleration hardware, which is configured with all-flash drives and where a flash-tier latency is expected: use thin provisioning in standard pools.
- ▶ On systems with FCMs it is preferred not use thin-provisioned volumes without compression. Recommended types are either fully allocated in any type of the pool, or thin-provisioned and compressed (and optionally deduplicated).
- ▶ On platforms with compression acceleration hardware, in most of the use cases compressed volumes are preferred over to thin-provisioned non-compressed, as most of the IO and processing overhead in DRP is generated by thin-provisioning processing. Latency added by compression processing after that overhead is in most configurations neglectable.

To automate volume creation with the required capacity savings method, system has a concept of provisioning policies. A provisioning policy can be assigned to a pool of any type or its child pool, and it enforces configured with it capacity savings method on all volumes that are newly created on the pool.

Features and limitations unique to pool type

Most of the system's features are supported by both types of the pools. However, some of them can be used only with one of the other pool type.

Features that can be used only with standard pools are:

- ▶ VMware VVOLs, you have to use standard pools. VVOLs implementation in IBM Storage Virtualize uses child pool as a container for VMware virtual volumes. That child pool must belong to standard pool type parent pool.
- ▶ Child Pools with capacity quotas. When you create a child pool from a standard pool, you have to specify capacity that is reserved for the child pool and a parent. Volumes and other child pools in the parent pool can't use that capacity, as it is guaranteed for the child pool. In DRP, child pools don't have any capacity guarantee. Child pool in DRP occupies only capacity that is in use by volumes in it.
- ▶ Thin-provisioned vdisks excluded from Easy Tier processing or striped to a subset of MDisks in the pool. This might be useful if you want to keep a single volume in the upper tier of a multi-tier pool.
- ▶ Image-mode vdisks.
- ▶ Cache-disabled thin-provisioned vdisks.

Features that are unique to DRP are:

- ▶ Pool-level compression and deduplication (see "Capacity savings methods" on page 259).
- ▶ Volume capacity reclamation with host unmap (see "Capacity deallocation (unmap)" on page 261).
- ▶ Redirect-on-Write flashcopy (snapshots). IBM Storage Virtualize uses Copy-on-Write snapshots. Redirect-on-Write flashcopy offers better performance, better deduplication ratio for snapshots and less I/O amplification. It is automatically enabled if all the requirements for both source and target volumes are true:

- they are in the same DRP
- they are in the same IO group
- they both are configured to use deduplication
- they both are not mirrored (have only a single copy).

If you plan to use DRP, you need to be aware of extra limitations that they imply:

- ▶ Minimum extent size is 1024 MB. For standard pools it can be smaller. (Except Flashsystem 9500)
- ▶ There is a limit for total thin-provisioned and compressed capacity for all volumes in a single data reduction pool. Standard pools only have a limit for a single volume. For more information, see [Configuration Limits and Restrictions](#).

Capacity deallocation (unmap)

By using **SCSI UNMAP** or **NVMe dataset management deallocate** commands, hosts can inform storage system that data blocks indicated in a command are no longer needed, and can deallocate and reclaim as a free capacity.

There is a significant difference in deallocation command handling between Standard pools and DRPs. They both support and process those commands, but only DRP can reclaim volume capacity.

When a user deletes a file on a host, the operating system issues an unmap command for the blocks that made up the file. A large amount of capacity can be freed if the user deletes (or Vmotions) a volume on a data store on a host. This process might result in many contiguous blocks being freed. Each of these contiguous blocks results in an unmap command that is sent to the LUN provisioned on a storage device.

Data reduction pools support end-to-end unmap functions. Space that is deallocated by the hosts with an unmap command results in the reduction of the used space in the volume and the pool.

In a DRP, when a system receives an unmap command, the result is that the capacity that is allocated within that contiguous chunk is freed. After the capacity is marked as free, it is accounted as a pool's *reclaimable capacity*. Similarly, deleting a volume at the DRP level marks all the volume's capacity as reclaimable. Garbage collection process, which runs at the background, works with reclaimable capacity and issues unmap commands to the system's backend (array or externally virtualized system) to free up unused space.

If a volume that receives unmap command belongs to a *standard pool*, the command is accepted, but the thin-provisioned volume's used capacity will not decrease, and free capacity will not be released to a pool free space.

In the back-end, unmap processing is same for both pool types.

If a pool's back-end storage supports unmap, then an unmap command to a corresponding LBA of a back-end MDisk is generated. For FCM storage, this process results in the device freeing the used physical capacity back to its available space.

On a regular, non space-efficient, solid-state drive (SSD)-based MDisk, unmap processing improves the efficiency of garbage collection on the device, leading to improved performance.

If back-end storage does not support unmap, commands that are sent by a host are also accepted and processed. Internally, the system issues a `write_same` command of zeros to the corresponding back-end storage LBAs. This behavior is required per the standard because storage system must ensure that a subsequent read to an unmapped region returns zeros and not any other data. For particular storage devices, such as nearline serial-attached SCSI (NL-SAS) drives, the `write_same` commands can create workload which is higher than the device can sustain, which results in performance problems.

Back-end unmap ensures that physical capacity on the FCM arrays is freed when a host deletes its data. Also it helps to ensure a good MDisk performance: flash drives can reuse the space for wear-leveling and to maintain a healthy capacity of “pre-erased” (ready to be used) blocks.

In virtualization scenarios, for example in configurations with an IBM SVC virtualizing IBM FlashSystem in the back-end, SVC will issue unmap or `write_same` commands to the virtualized system if the system announces unmap support during SCSI discovery.

Important: A standard pool will not shrink its used space as the result of a host unmap command processing. However, the back-end used capacity will shrink, if back-end is space-efficient (for example, if backend is an array of FCMs).

For this reason, even with a standard pool it is still beneficial to have host unmap support that is enabled for systems that use FCMs. If a system contains not only FCMs but a mix of storage, thorough planning is required.

Enabling, monitoring, throttling, and disabling SCSI unmap

There are two settings that control unmap:

- ▶ Host unmap support controls if unmap capabilities are announced to the hosts that are mounted to system’s volumes
- ▶ Back-end unmap support controls if unmap commands are issued to back-end storage, internal (arrays) and external.

By default, host-based unmap support is disabled on all product other than the IBM FlashSystem 9000 series, as those systems are all-flash. On systems that can be ordered in hybrid configurations, you might need to enable host unmap support.

Back-end unmap is enabled by default on all products, and best practice is to leave it enabled.

To enable or disable host unmap support, run the following command:

```
chsystem -hostunmap on|off
```

To enable or disable backend unmap commands, run the following command:

```
chsystem -backendunmap on|off
```

You can check how much unmap processing is occurring on a per volume or per-pool basis by using the performance statistics. This information can be viewed with IBM Spectrum Control or IBM Storage Insights.

Note: Processing unmap can add workload to the back-end storage.

Performance monitoring helps you notice possible effects, and if the unmap workload is affecting performance, consider taking the necessary steps and consider the data rates that are observed. It might be expected to see GiBps of unmap if you deleted many volumes.

You can throttle the amount of “host offload” operations (such as the SCSI UNMAP) by using the per-node settings for offload throttle. For example:

```
mkthrottle -type offload -bandwidth 500
```

This setting limits each node to 500 MiBps of offload commands.

You can also stop the system from announcing unmap support to one or more host systems, if you are not able to stop any host from issuing unmap commands by using host-side settings. To modify unmap support for a particular host, change the host type:

```
chhost -type generic_no_unmap <host_id_or_name>
```

If you experience severe performance problems as a result of unmap operations, you can disable unmap on the entire system or on per-host basis.

Flexibility for the future

It is not possible to change pool type, so when selecting pool type it is important to consider not only current requirements, but also requirements that can appear later during the system’s life cycle.

If other important factors do not lead you to choose standard pools, then DRPs are the right choice. Using DRPs can increase storage efficiency and reduce costs because it reduces the amount of data that is stored on hardware and reclaims previously used storage resources that are no longer needed by host systems.

Also, DRPs provide great flexibility for future use because they add the ability of compression and deduplication of data at the volume level in a specific pool, even if these features are initially not used at creation time.

Note: We recommend the use of DRP pools with fully allocated volumes if the restrictions on capacity do not affect your environment.

4.3.2 Planning for availability

Availability concepts need to be considered when planning for the number of pools in the system and the number of MDisks (array) in each.

By design, IBM Storage Virtualize systems take the entire storage pool offline if a single MDisk in that storage pool goes offline, which means that the storage pool’s MDisk quantity and size define the failure domain. Reducing the hardware failure domain for back-end storage is only part of your considerations. When you are determining the storage pool layout, you must also consider application boundaries and dependencies to identify any availability benefits that one configuration might have over another one.

Sometimes, reducing the hardware failure domain, such as placing the volumes of an application into a single storage pool, is not always an advantage from the application perspective.

Alternatively, splitting the volumes of an application across multiple storage pools increases the chances of having an application outage if one of the storage pools that is associated with that application goes offline.

Finally, increasing the number of pools to reduce the failure domain is not always a viable option. An example of such a situation is a system with a low number of physical drives: creating multiple arrays to place them into different pools reduces the usable space because of spare and protection capacity.

When virtualizing external storage, the failure domain is defined by the external storage itself rather than by the pool definition on the front-end system. For example, if you provide 20 MDisks from external storage and all of these MDisks are using the same physical arrays, the failure domain becomes the total capacity of these MDisks, no matter how many pools you have distributed them across.

The following actions are the starting best practices when planning storage pools for availability:

- ▶ Create a separate pool for arrays that are created from drives connected to a single control enclosure (IO group). A pool can join MDisks from different IO groups in enclosure-based system, but this approach increases failure domain size and dramatically increases inter-node traffic traversing the SAN. In SVC environment, storage is normally connected to all IO groups, so this is not a concern.
- ▶ Create a separate pool for each FCM array. Only a single FCM array is allowed in a pool.
- ▶ Create separate pools for internal storage and external storage, unless you are creating a hybrid pool that is managed by Easy Tier (see 4.3.6, “External pools” on page 269).
- ▶ Create a storage pool for each external virtualized storage subsystem, unless you are creating a hybrid pool that is managed by Easy Tier (see 4.3.6, “External pools” on page 269).

Note: If capacity from different external storage is shared across multiple pools, provisioning groups are created.

IBM Storage Virtualize detects that a resource (MDisk) shares its physical storage with other MDisks and monitors provisioning group capacity. MDisks in a single provisioning group should not be shared between storage pools because capacity consumption on one pool can affect free capacity on other pools. System detects this condition and shows that the pool contains shared resources.

- ▶ Use dedicated pools for image mode volumes.

Limitation: Image mode volumes are not supported in DRPs.

- ▶ For Easy Tier enabled storage pools, always allow free capacity for Easy Tier to deliver better performance.
- ▶ Consider implementing child pools when you must have a logical division of your volumes for each application set. There are cases where you want to subdivide a storage pool but maintain many MDisks in that pool. Child pools allow logical subdivision of a pool. Throttles and provisioning policies can be set independently per child pool. In standard pools, capacity thresholds for child pools can also be set.

When you select storage subsystems, the decision often comes down to the ability of the storage subsystem to be more reliable and resilient, and meet application requirements.

Although IBM Storage Virtualize can provide any physical level-data redundancy for virtualized external storages by using advanced features, with the availability characteristics of the storage subsystem's controllers have the most impact on the overall availability of the data that is virtualized.

4.3.3 Planning for performance

IBM Storage Virtualize will stripe each volumes across all MDisks in the pool, which will set volume performance limits much higher than the limit for a single MDisk because multiple read and write operations will run in parallel.

However, IBM Storage Virtualize DRAID can also utilize multiple threads. This means that full system performance can be achieved with a single DRAID array that joins, for example, all FCM drives in FlashSystem 9200. There is no need to create multiple DRAID arrays, until the arrays conform to best practices for arrays on your drive type, as listed in 3.2.2, "Array considerations" on page 208.

To implement performance-oriented pools with externally virtualized storage, create large pools with multiple arrays rather than more pools with few arrays. This approach usually works better for performance than spreading the application workload across many smaller pools because typically the workload is not evenly distributed across the volumes and then across the pools. For more details, see 4.3.6, "External pools" on page 269.

The following actions are the starting best practices when planning storage pools for performance:

- ▶ Create a dedicated storage pool with dedicated resources if there is a specific performance application request.
- ▶ When using external storage in an Easy Tier enabled pool, do not intermix MDisks in the same tier with different performance characteristics.
- ▶ In an IBM FlashSystem clustered environment, create storage pools with IOgrp or Control Enclosure affinity. You use only arrays or MDisks that are supplied by the internal storage that is directly connected to one IOgrp SAS chain only. This configuration avoids unnecessary IOgrp-to-IOgrp communication traversing the storage area network (SAN) and consuming Fibre Channel (FC) bandwidth.

Note: In IBM FlashSystem setup with multiple IO groups, do not mix MDisks from different control enclosures (I/O groups) in a single pool.

In an SVC cluster environment, this recommendation does not apply.

- ▶ Use dedicated pools for image mode volumes.

Limitation: Image mode volumes are not supported by DRPs.

- ▶ For Easy Tier enabled storage pools, always allow free capacity for Easy Tier to deliver better performance.
- ▶ Consider implementing child pools when you must have a logical division of your volumes for each application set. Cases often exist where you want to subdivide a storage pool but maintain many f MDisks in that pool. Child pools are logically like storage pools, but allow you to specify one or more subdivided child pools. Thresholds and throttles can be set independently per child pool.

Cache partitioning

System's write cache is partitioned, and each pool is assigned to a single cache partition. The system automatically defines a logical cache partition per storage pool. Child pools do not count toward cache partitioning.

A cache partition is a logical threshold that stops a single partition from consuming the entire cache resource. This partition is provided as a protection mechanism and does not affect performance in normal operations. Only when a storage pool becomes overloaded does the partitioning activate and essentially slow down write operations in the pool to the same speed that the back-end can handle. *Overloaded* means that the front-end write throughput is greater than the back-end storage can sustain. This situation should be avoided.

Table 4-2 shows the upper limit of write cache data that any one partition, or storage pool, can occupy.

Table 4-2 Upper limit of write cache data

Number of storage pools	Upper limit
1	100%
2	66%
3	40%
4	30%
5 or more	25%

You can think of the rule as no single partition occupies more than its upper limit of cache capacity with write data.

In recent versions of IBM Spectrum Control, the fullness of the cache partition is reported and can be monitored. You should not see partitions reaching 100% full. If you do, then it suggests the corresponding storage pool is in an overload situation, and the workload should be moved from that pool or extra storage capability should be added to that pool.

4.3.4 Planning for capacity

Capacity planning is never an easy task. Capacity monitoring has become more complex with the advent of data reduction. It is important to understand the terminology that is used to report usable, used, and free capacity.

The terminology and its reporting in the GUI changed in recent versions, and they are listed in Table 4-3.

Table 4-3 Capacity terminology

Old term	New term	Meaning
Physical capacity	Usable capacity	The amount of capacity that is available for storing data on a system, pool, array, or MDisk after formatting and RAID techniques are applied.
Volume capacity	Provisioned capacity	The total capacity of all volumes in the system.

Old term	New term	Meaning
N/A	Written capacity	The total capacity that is written to the volumes in the system, which is shown as a percentage of the provisioned capacity and reported before any data reduction.

The *usable capacity* describes the amount of capacity that can be written to on the system and includes any back-end data reduction (that is, the “virtual” capacity is reported to the system).

An example of the dashboard capacity view is shown in Figure 4-5.

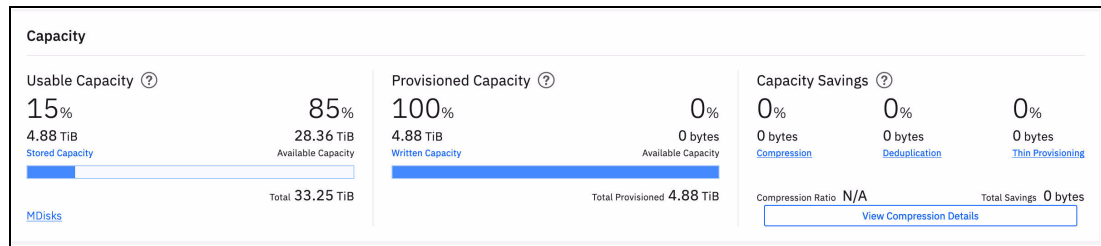


Figure 4-5 Example dashboard capacity view

Note: With DRP, capacity reporting is more complex than with standard pools. For details, refer to 4.4, “Data reduction pools best practices” on page 270.

For FCMs, the usable capacity is the maximum capacity that can be written to the system. However, for the smaller capacity drives (4.8 TB), the reported usable capacity is 20 TiB. The actual usable capacity might be lower because of the actual data reduction that is achieved from the FCM compression.

Plan to achieve the default 2:1 compression, which is approximately an average of 10 TiB of usable space. Careful monitoring of the actual data reduction should be considered if you plan to provision to the maximum stated usable capacity when the small capacity FCMs are used.

The larger FCMs, 9.6 TB and above, report just over 2:1 usable capacity. Therefore, 22, 44, and 88 for the 9.6, 19.2, and 38.4 TB modules.

The *provisioned capacity* shows the total provisioned capacity in terms of the volume allocations. This capacity is the “virtual” capacity that is allocated to fully allocated and thin-provisioned volumes. Therefore, it is in theory that the capacity can be written if all volumes were filled 100% by the system.

The *written capacity* is the actual amount of data that is written into the provisioned capacity:

- ▶ For fully allocated volumes, the written capacity is always 100% of the provisioned capacity.
- ▶ For thin-provisioned volumes (including data reduced volumes), the written capacity is the actual amount of data that the host writes to the volumes.

The final set of capacity numbers relates to the data reduction, which is reported in two ways:

- ▶ As the savings from DRP (compression and deduplication) that is provided at the DRP level, as shown in Figure 4-6 on page 268.

- ▶ As the FCM compression.

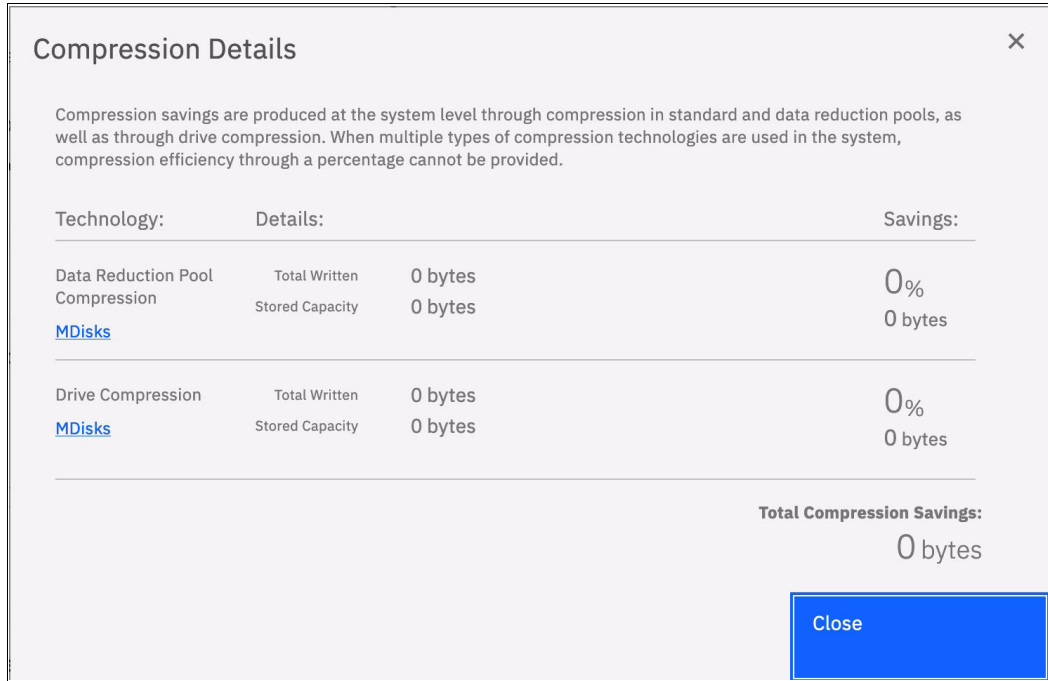


Figure 4-6 Compression savings dashboard report

4.3.5 Extent size considerations

When adding MDisks to a pool, they are logically divided into chunks of equal size. These chunks are called *extents* and they are indexed internally. IBM Storage Virtualize can manage 2^{22} extents in system, independently of the extent size. This means that the choice of extent size affects the total amount of storage that can be addressed.

Extent sizes can be 16, 32, 64, 128, 256, 512, 1024, 2048, 4096, or 8192 MB. However, the choice of sizes can be limited by a pool type and by hardware platform. For example, DRP supports only extent sizes not smaller than 1024 MB on any platform, and FlashSystem 9500 allows only to choose between 4096 and 8192 MB. For the exact capacity limits, see [Configuration Limits and Restrictions](#).

Extent size cannot be changed after a pool is created, it must remain constant through the lifetime of the pool.

For pool-extent size planning, consider the following recommendations:

- ▶ Aim to keep extent size minimal, while keeping in mind system's maximum managed capacity, including all potential future growth.
- ▶ For DRP, consider a limit for all thin-provisioned volumes in the pool, including possible future growth. Avoid small extent sizes with DRP.
- ▶ With Easy Tier enabled hybrid pools, consider smaller extent sizes to better use the higher tier resources and provide better performance.
- ▶ Aim to keep the same extent size for all pools in the system. The extent-based migration function between pools is not possible with different extent sizes. However, you can still migrate volumes between pools using volume mirroring feature.

Limitation: Extent-based migrations from standard pools to DRPs are not supported unless the volume is fully allocated.

4.3.6 External pools

Both IBM FlashSystem and SVC systems can virtualize external storage systems. This section describes special considerations when configuring storage pools with external storage.

Availability considerations

The external storage virtualization feature provides many advantages through consolidation of storage. You must understand the availability implications that storage component failures can have on availability domains within a system.

IBM Storage Virtualize offers significant performance benefits through its ability to stripe across back-end storage volumes. However, consider the effects that various configurations have on availability:

- ▶ When you select MDisks for a storage pool, performance is often the primary consideration. However, in many cases, the availability of the configuration is traded for little or no performance gain.
- ▶ System takes the entire storage pool offline if a single MDisk in that storage pool goes offline. Consider an example where you have 40 external arrays of 1 TB each for a total capacity of 40 TB with all 40 arrays in the same storage pool. In this case, you place the entire 40 TB of capacity at risk if one of the 40 arrays fails (which causes the storage pool to go offline). If you then spread the 40 arrays out over some of the storage pools, the effect of an array failure (an offline MDisk) affects less storage capacity, which limits the failure domain.

To ensure optimum availability to well-designed storage pools, consider the following best practices:

- ▶ It is recommended that each storage pool contains only MDisks from a single storage subsystem. An exception exists when you are working with Easy Tier hybrid pools. For more information, see 4.7, “Easy Tier and tiered and balanced storage pools” on page 296.
- ▶ It is suggested that each storage pool contains only MDisks from a single storage tier (SSD or flash, enterprise, or NL-SAS) unless you are working with Easy Tier hybrid pools. For more information, see 4.7, “Easy Tier and tiered and balanced storage pools” on page 296.

IBM Storage Virtualize does not provide any physical-level data redundancy for virtualized external storage. The availability characteristics of the storage subsystems’ controllers have the most impact on the overall availability of the data that is virtualized by IBM Storage Virtualize.

Performance considerations

Using the following external virtualization capabilities, you can boost the performance of the back-end storage systems:

- ▶ Using wide-striping across multiple arrays
- ▶ Adding more read/write cache capability

Wide-striping can add approximately 10% extra input/output processor (IOP) performance to the back-end system by using these mechanisms.

Another factor is that virtualized-storage subsystems can be scaled up or scaled out. For example, IBM System Storage DS8000 series is a scale-up architecture that delivers the best performance per unit, and the IBM FlashSystem series can be scaled out with enough units to deliver the same performance.

With a virtualized system, there is debate whether to scale out back-end systems or add them as individual systems behind IBM Storage Virtualize. Either case is valid. However, adding individual controllers is likely to allow IBM Storage Virtualize to generate more I/O based on queuing and port-usage algorithms. It is recommended that you add each controller enclosure (I/O group) of an IBM FlashSystem back-end as its own controller, that is, do not cluster IBM FlashSystem when it acts as an external storage controller behind another IBM Storage Virtualize product, such as SVC.

Adding each controller (I/O group) of an IBM FlashSystem back-end as its own controller adds more management IP addresses and configurations. However, this action provides the best scalability in terms of overall solution performance.

A significant consideration when you compare native performance characteristics between storage subsystem types is the amount of scaling that is required to meet the performance objectives. Although lower-performing subsystems can typically be scaled to meet performance objectives, the extra hardware that is required lowers the availability characteristics of the IBM Storage Virtualize system.

All storage subsystems possess an inherent failure rate. Therefore, the failure rate of a storage pool becomes the failure rate of the storage subsystem times the number of units.

Number of MDisks per pool

The number of MDisks per pool also will affect availability and performance. Adding more arrays to a pool, rather than creating a separate pool, can be a way to improve the overall pool performance because of data striping across the MDisks and I/O parallelism.

The back-end storage access is controlled through MDisks where the IBM Storage Virtualize systems act like a host to the back-end controller. Just as you must consider volume queue depths when accessing storage from a host, these systems must calculate queue depths to maintain high throughput capability while ensuring the lowest possible latency.

For more information about the queue depth algorithm and the rules about how many MDisks to present for an external pool, see “Volume considerations” on page 230, which describes how many volumes to create on the back-end controller (that are seen as MDisks by the virtualizing controller) based on the type and number of drives or flash modules.

4.4 Data reduction pools best practices

This section describes DRP planning and implementation best practices.

4.4.1 Data reduction pools with internal storage

Important: If you plan to use DRP with deduplication and compression that is enabled with FCM storage, assume zero extra compression from the FCMs, that is, use the reported physical or usable capacity from the RAID array as the usable capacity in the pool and ignore the maximum effective capacity.

The reason for assuming zero extra compression from the FCMs is because the DRP function is sending compressed data to the FCMs, which cannot be further compressed. Therefore, the data reduction (effective) capacity savings are reported at the front-end pool level and the back-end pool capacity is almost 1:1 for the physical capacity.

Some small amount of other compression savings might be seen because of the compression of the DRP metadata on the FCMs.

The main point to consider is whether the data is deduplicable. Tools are available to provide estimation of the deduplication ratio. For more information, see 4.2.2, “Determining whether your data is a deduplication candidate” on page 254.

Consider DRP configurations with FCM drives:

- ▶ Data is highly deduplicatable. In this case, the recommendation is to use compressed and deduplicated volume type. The fact that FCMs will get data that is already compressed, will not impact FCM performance and capacity savings.
- ▶ Data is not deduplicatable. In this case, you might use fully allocated volumes in DRP or in standard pools, and let the FCM hardware do the compression because the overall achievable throughput will be higher.

With the industry standard NVMe drives, which do not perform inline compression, similar considerations apply:

- ▶ Data is deduplicable. In this case, the recommendation is to use a compressed and deduplicated volume type. The DRP compression technology has more than enough compression bandwidth for these purposes, so compression should always be done.
- ▶ Data is not deduplicable. In this case, the recommendation is to use only a compressed volume type. The internal compression technology provides enough compression bandwidth.

Note: In general, avoid creating DRP volumes that are only thin-provisioned and deduplicated. When using DRP volumes, they should be either fully allocated, or deduplicated and compressed.

Various configuration items affect the performance of compression on the system. To attain high compression ratios and performance on your system, ensure that the following guidelines are met:

- ▶ For performance-optimized configurations, aim to use FCM compression as the only capacity efficiency method.

- ▶ Using a small amount (1 - 3%) of storage-class memory (SCM) capacity in a DRP significantly improves DRP metadata performance. As the directory data is the most frequently accessed data in the DRP and the design of DRP maintains directory data on the same extents, Easy Tier quickly promotes the metadata extents to the fastest available tier.
- ▶ Never create a DRP with only NL-SAS capacity. If you want to use predominantly NL-SAS drives, ensure that you have a small amount of flash or SCM capacity for the metadata.
- ▶ Do not compress encrypted data. If the application or operating system provides encryption, do not attempt to use DRP compression, instead use only FCM compression and account for no capacity savings (1:1 ratio of physical to effective capacity). Data at rest encryption, which is provided by the system, is still possible because the encryption is performed after the data is reduced or by the FCM itself. If host-based encryption is unavoidable, assume that data reduction is not possible.
- ▶ Although DRP and FCM do not have performance penalties if data cannot be compressed (that is, you can attempt to compress all data), the extra overhead of managing DRP volumes can be avoided by using fully allocated volumes if no data reduction benefits are realized.
- ▶ You can use tools that estimate the compressible data or use commonly known ratios for common applications and data types. Storing these data types on compressed volumes saves disk capacity and improves the benefit of using compression on your system. For more information, see “Knowing your data and workload” on page 252.
- ▶ Avoid the usage of any client, file system, or application based-compression with the system compression. If this approach is not possible, use same considerations as for host-encrypted data.

4.4.2 DRP and external storage considerations

Avoid configurations that attempt to perform data reduction at two levels.

Never use DRP on both front-end (virtualizing) and back-end (virtualized) systems concurrently (DRP over DRP). It is recommended to use DRP at the virtualizer level rather than the back-end storage because this approach simplifies capacity management and reporting.

For storage behind the virtualizer, best practice is to provision fully allocated volumes. By running in this configuration, you ensure the following items:

- ▶ The virtualizer understands the real physical capacity that is available and can warn you about and avoid out-of-space situations (where access is lost due to no space).
- ▶ Capacity monitoring can be performed on the virtualizer level because it sees the true physical and effective capacity usage.
- ▶ The virtualizer performs efficient data reduction on previously unreduced data. Generally, the virtualizer has offload hardware and more CPU resources than the back-end storage because it does not need to deal with RAID and other such considerations.

If you cannot avoid back-end data reduction (for example, the back-end storage controller cannot disable its data reduction features), ensure that you follow these best practices:

- ▶ Do not excessively overprovision the physical capacity on the back-end:
 - For example, assume that you have 100 TiB of real capacity. Start by presenting only 100 TiB of volumes to IBM Storage Virtualize. Monitor the actual data reduction on the back-end controller. If your data is reducing well over time, increase the capacity that is provisioned to the virtualizer.

- This approach ensures that you can monitor and validate your data reduction rates and avoids a panic if you do not achieve the expected rates and presented too much capacity to the system.
- ▶ Ensure that capacity usage is carefully and constantly monitored on both levels. Set up alerting and have an emergency plan for the situation when back-end device is close to being out of physical space.

Important: Never run DRP on top of DRP. This approach is wasteful and causes performance problems without extra capacity savings.

4.4.3 DRP provisioning considerations

This section describes best practices to consider during DRP provisioning.

DRP restrictions

Consider the following important restrictions when planning for a DRP implementation:

- ▶ The maximum number of supported DRPs in the system is four.
- ▶ VVOLs are not currently supported in DRP.
- ▶ Volume shrinking is not supported in DRP with thin or compressed volumes.
- ▶ Non-Disruptive Volume Movement (NDVM) is not supported by DRP volumes.
- ▶ The volume copy split of a volume mirror in a different I/O group is not supported for DRP thin-provisioned or compressed volumes.
- ▶ Image and sequential mode virtual disks (VDisks) are not supported in DRP.
- ▶ Extent level migration is not allowed between DRPs (or between a DRP and a standard pool) unless volumes are fully allocated.
- ▶ Volume migration for any volume type is permitted between a quotaless child and its parent DRP pool.
- ▶ There is a limit of maximum of 128 K extents per Customer Data Volume per I/O group:
 - Therefore, the pool extent size dictates the maximum physical capacity in a pool after data reduction.
 - Use a 4 GB extent size or greater.
- ▶ The recommended pool size is at least 20 TB.
- ▶ Use less than 1 PB per I/O group.
- ▶ Your pool should be no more than 85% occupied.

In addition, the following considerations apply to DRP:

- ▶ The real, used, free, and tier capacities are not reported per volume for DRP volumes. Instead, only information on a pool level is available.
- ▶ Cache mode is always read/write on compressed or deduplicated volumes.
- ▶ Autoexpand is always on.
- ▶ No ability to place specific volume capacity on specific MDisks.

Extent size considerations

With DRP, the number of extents available per pool is limited by the internal structure of the pool and specifically by the size of the data volume. For more information, see 4.1.2, “Data reduction pools” on page 249.

At the time of writing, the maximum number of extents that are supported for a data volume is 128 K. As shown in Figure 4-1 on page 250, one data volume is available per pool.

Table 4-4 lists the maximum size per pool, by extent size and I/O group number.

Table 4-4 Pool size by extent size and I/O group number

Extent size	Max size with one I/O group	Max size with two I/O groups	Max size with three I/O group	Max size with four I/O group
1024	128 TB	256 TB	384 TB	512 TB
2048	256 TB	512 TB	768 TB	1024 TB
4096	512 TB	1024 TB	1536 TB	2048 TB
8192	1024 TB	2048 TB	3072 TB	4096 TB

Considering that the extent size cannot be changed after the pool is created, it is recommended that you carefully plan the extent size according to the environment capacity requirements. For most of the configurations, an extent size of 4 GB is recommended for DRP.

Pool capacity requirements

A minimum capacity must be provisioned in a DRP to provide capacity for the internal metadata structures. This capacity is reserved on a DRP at the creation time. Table 4-5 shows the minimum capacity that is required by extent size and I/O group number.

Table 4-5 Minimum recommended pool size by extent size and I/O group number

Extent size	Min size with one I/O group	Min size with two I/O group	Min size with three I/O group	Min size with four I/O group
1024	255 GB	516 GB	780 GB	1052 GB
2048	510 GB	1032 GB	1560 GB	2104 GB
4096	1020 GB	2064 GB	3120 GB	4208 GB
8192	2040 GB	4128 GB	6240 GB	8416 GB

The values that are reported in Table 4-5 represent the minimum required capacity for a DRP to create a single volume.

Important: When sizing a DRP, the garbage-collection process is constantly running to reclaim the unused space, which optimizes the extents usage. For more information about the garbage-collection process, see “DRP internal details” on page 249.

This garbage-collection process requires a certain amount of free space to work efficiently. For this reason, it is recommended to keep at least *15% free space* in a DRP pool.

4.4.4 Standard and data reduction pools coexistence

Although homogeneous configurations for the pool type are preferable, there is no technical reason to avoid using standard pools and DRPs in the same system. In some circumstances, this coexistence is unavoidable. Consider the following scenarios:

- ▶ Systems that require VVOLs support and data reduction capabilities for other environments. This scenario requires the definition of both standard pools and DRPs because of the restriction of DRPs regarding VVOLs. In this case, the standard pool is used for VVOL environments only, and the DRP is used for the other environments. Some data-reduction capability can be achieved for the VVOLs standard pool by using the inline data compression that is provided by the FCMs on IBM FlashSystem.
- ▶ Systems that require an external pool for image mode volumes and data reduction capabilities for other environments require the definition of standard pools and DRPs because of the restriction of DRPs regarding the image mode volumes. In this case, the standard pool is used for image mode volumes only, and optionally with the write cache disabled if needed for the back-end native copy services usage. For more information, see Chapter 6, “Copy services” on page 363. DRP is used for all the other environments.
- ▶ Installations where an IBM FlashSystem provides capacity for the hosts and for the virtualizer front-end system. A pool providing back-end capacity for a virtualizer must be a standard pool, as recommended in 4.4.2, “DRP and external storage considerations” on page 272. Host capacity can be provided as volumes with capacity efficiency features in DRP, if needed.
- ▶ An installation that requires more than four pools.

4.4.5 Data migration with data reduction pools

As mentioned in “DRP restrictions” on page 273, extent-level migration to and from a DRP (such as migrate-volume or migrate-extent functions) is not supported. For an existing configuration, where you plan to move data to or from a DRP and use data-reduced volumes, there are two options: host-based migration and volume mirroring-based migration.

Host-based migration

Host-based migration uses operating system features or software tools that run on the hosts to concurrently move data to the normal host operations. VMware vMotion and AIX Logical Volume Mirroring (LVM) are two examples of these features. When you use this approach, a specific amount of capacity on the target pool is required to provide the migration target volumes.

The process includes the following steps:

1. Create the target volumes of the migration in the target pool. Depending on the migration technique, the size and the number of the volumes can be different from the original ones. For example, you can migrate two 2 TB VMware datastore volumes in a single 4 TB data store volume.
2. Map the target volumes to the host.
3. Rescan the Host Bus Adapters (HBAs) to attach the new volumes to the host.
4. Activate the data move or mirroring feature from the old volumes to the new ones.
5. Wait until the copy is complete.
6. Detach the old volumes from the host.
7. Unmap and remove the old volumes from the system.

When migrating data to a DRP, consider the following options:

- ▶ Migrate directly to compressed or deduplicated volumes. With this option, the migration duration mainly depends on the host-migration throughput capabilities. The target volumes are subject to high write-workload, which can use many resources because of the compression and deduplication tasks.

To avoid a potential effect on the performance of the workload, try to limit the migration throughput at the host level. If this limit cannot be used, implement the throttling function at the volume level.

- ▶ Migrate first to fully allocated volumes, and then convert them to compressed or deduplicated volumes. Also with this option, the migration duration mainly depends on the host capabilities, but usually more throughput can be sustained because there is no overhead for compression and deduplication. The space-saving conversion can be done by using the volume mirroring feature.

Volume mirroring-based migration

The volume mirroring feature can be used to migrate data from a pool to another pool while changing the space saving characteristics of a volume. Like host-based migration, volume mirroring-based migration requires free capacity on the target pool, but it is not needed to create volumes manually.

Volume mirroring migration is a three-step process:

1. Add a volume copy on the DRP and specify the data reduction features.
2. Wait until the copies are synchronized.
3. Remove the original copy.

With volume mirroring, the throughput of the migration activity can be adjusted at a volume level by specifying the Mirror Sync Rate parameter. Therefore, if performance is affected, the migration speed can be lowered or even suspended.

Note: Volume mirroring supports only two copies of a volume. If a configuration uses both copies, one of the copies must be removed first before you start the migration.

The volume copy split of a volume mirror in a different I/O group is not supported for a thin-provisioned or compressed volume in a DRP.

4.4.6 Understanding capacity use in a data reduction pool

Capacity monitoring in DRP is a complex task and requires a knowledge of DRP capacity terms. They are shown in Table 4-6 on page 277.

Data reduction pools use Log Structured Array (LSA) mechanism to store data of thin-provisioned and compressed volumes. When host overwrites a block of the data on its volume, LSA will not write new data to the same place as the old block. Instead, it will write data to the new area, and mark an old block as garbage. Same process occurs when a sends unmap (deallocate) commands after it has deleted a part of his data: those blocks are also marked as a garbage.

When you delete a thin-provisioned or compressed volume in DRP, its capacity does not immediately return as free. In DRP, volume deletion is asynchronous process, which includes the following phases:

- The grain must be inspected to determine whether this volume was the last one that referenced this grain (deduplicated):

- If so, the grains can be freed.
- If not, the grain references must be updated, and the grain might need to be rehomed to belong to one of the remaining volumes that still require this grain.
- When all grains that are to be deleted are identified, these grains are returned to the “reclaimable” capacity (marked as a garbage).

All the blocks marked as garbage data are accounted as a pool’s *Reclaimable Capacity*.

Table 4-6 *DRP capacity terms*

Use	Description
Reduced customer data	The data that is written to the DRP, in compressed and de-duplicated form.
Fully allocated data	The amount of capacity that is allocated to fully allocated volumes (assumed to be 100% written).
Free	The amount of free space that is not in use by any volume.
Reclaimable data	The amount of data in the pool marked as a garbage, but not yet accounted as free. It is either old (overwritten) data, or data that is unmapped by host, or associated with recently deleted volumes. It will become free after it is collected by a Garbage Collector process.
Metadata	Approximately 1 - 3% overhead for DRP metadata volumes.

To turn reclaimable capacity into free capacity, it needs to be processed by a garbage collection process. It runs in the background and scans DRP capacity trying to find a full extent of garbage. If a full extent can’t be found, it will move blocks of data and garbage to accumulate garbage blocks consequentially on extent boundary, and then reclaim full extent.

After a reasonable usage period, the DRP can show up to only 15 - 20% of overall free space, and a significant part of the capacity as reclaimable. The garbage collection algorithm must balance the need to free space with the overhead of performing garbage collection. Therefore, the incoming write/overwrite rates and any unmap operations dictate how much “reclaimable space” is present at any time.

Balancing how much garbage collection is done versus how much free space is available dictates how much reclaimable space is present at any time. The system dynamically adjusts the target rate of garbage collection to maintain a suitable amount of free space.

Data reduction pool attempts to maintain a sensible amount of free space. If there is little free space and you delete many volumes, the garbage-collection code might trigger a large amount of back-end data movement and might result in performance issues.

An example steady state DRP is shown in Figure 4-7 on page 278.

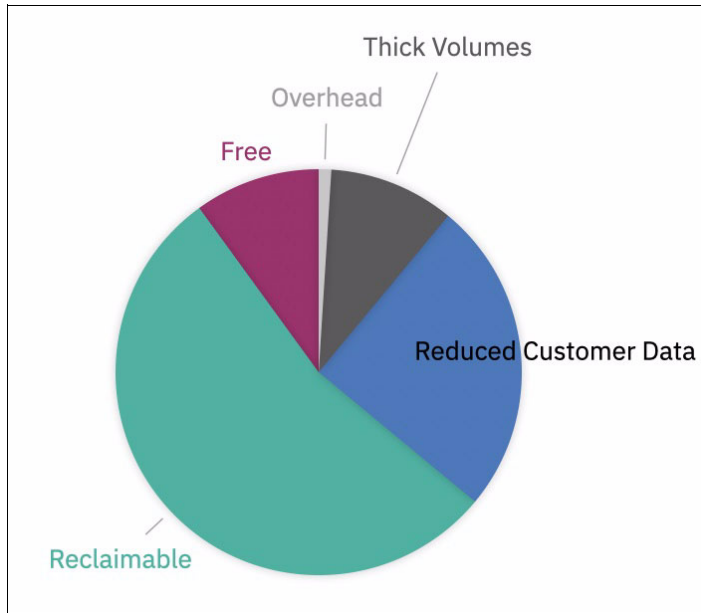


Figure 4-7 Data reduction pool capacity use example

Consider the following points:

- ▶ If you create a large capacity of fully allocated volumes in a DRP, you are taking this capacity directly from free space only, which might result in triggering heavy garbage collection if there is little free space remaining and a large amount of reclaimable space remains.
- ▶ If you create many fully allocated volumes and experience degraded performance due to garbage collection, you can reduce the required work by temporarily deleting unused fully allocated volumes.
- ▶ When deleting a fully allocated volume, the capacity is returned directly to free space.
- ▶ When deleting a thin-provisioned volume, it does not immediately create free space.
- ▶ If you are at risk of running out of space, but much reclaimable space exists, you can force garbage collection to work harder by creating a temporary fully allocated volume to reduce the amount of real free space and trigger more garbage collection.

Important: Use caution when using up all or most of the free space with fully allocated volumes. Garbage collection requires free space to coalesce data blocks into whole extents and free capacity. If little free space is available, the garbage collector must work harder to free space.

- ▶ It might be worth creating some “airbag” fully allocated volumes in a DRP. This type of volume reserves some space that you can quickly return to the free space resources if you reach a point where you are almost out of space, or when garbage collection is struggling to free capacity efficiently.

Consider these points:

- This type of volume should not be mapped to hosts.
- This type of volume should be labeled, for example, “RESERVED_CAPACITY_DO_NOT_USE”.

Note: In DRP, volume's used capacity and tier used capacity are not reported per volume. These items are reported only at the parent pool level because of the complexities of deduplication capacity reporting.

4.5 Operations with storage pools

In the following section, we describe some guidelines for the typical operation with pools, which apply both to standard and DRP pool types. Only highlights and best practices are provided. All operations are described in details in *Implementation Guide for IBM Storage FlashSystem and IBM SAN Volume Controller: Updated for IBM Storage Virtualize Version 8.6*, SG24-8542.

4.5.1 Adding external MDisks to existing storage pools

If MDisks are being added to an IBM Storage Virtualize system, it is likely because you want to provide more capacity or performance. In Easy Tier enabled pools, the storage-pool balancing feature ensures that the newly added MDisks are automatically populated with extents that come from the other MDisks. Therefore, manual intervention is not required to rebalance the capacity across the available MDisks.

Important: When adding external MDisks, the system does not know to which tier the MDisk belongs. Ensure that you specify or change the tier type to match the tier type of the MDisk.

This specification is vital to ensure that Easy Tier keeps a pool as a single tier pool and balances across all MDisks, or Easy Tier adds the MDisk to the correct tier in a multitier pool.

Failure to set the correct tier type creates a performance problem that might be difficult to diagnose in the future.

Adding MDisks to storage pools is a simple task, but it is suggested that you perform some checks in advance, especially when adding external MDisks.

Checking access to new MDisks

Be careful when you add external MDisks to existing storage pools to ensure that the availability of the storage pool is not compromised by adding a faulty MDisk. The reason is that loss of access to a single MDisk causes the entire storage pool to go offline.

In IBM Storage Virtualize, there is a feature that tests an MDisk automatically for reliable read/write access before it is added to a storage pool. Therefore, user action is not required. The test fails under the following conditions:

- ▶ One or more nodes cannot access the MDisk through the chosen controller port.
- ▶ I/O to the disk does not complete within a reasonable time.
- ▶ The SCSI inquiry data that is provided for the disk is incorrect or incomplete.
- ▶ The IBM Storage Virtualize cluster suffers a software error during the MDisk test.

Image-mode MDisks are not tested before they are added to a storage pool because an offline image-mode MDisk does not take the storage pool offline. Therefore, the suggestion here is to use a dedicated storage pool for each image mode MDisk.

This best practice makes it easier to discover what the MDisk is going to be virtualized as, and reduces the chance of human error.

Persistent reserve

A common condition where external MDisks can be configured by IBM FlashSystem and SVC but cannot perform read/write is when a persistent reserve is left on a LUN from a previously attached host.

In this condition, rezone the back-end storage and map it back to the host that is holding the reserve. Alternatively, map the back-end storage to another host that can remove the reserve by using a utility such as the Microsoft Windows SDD Persistent Reserve Tool.

4.5.2 Renaming MDisks

After you discover MDisks, rename them from their default name. This action can help during problem isolation and avoid confusion that can lead to an administrative error by using a naming convention for MDisks that associates the MDisk with the controller and array.

When multiple tiers of storage are on the same system, you might also want to indicate the storage tier in the name. For example, you can use R5 and R10 to differentiate RAID levels, or you can use T1, T2, and so on, to indicate the defined tiers.

Best practice: For MDisks, use a naming convention that associates the MDisk with its corresponding controller and array within the controller, such as DS8K_<extent pool name or id>_<volume id>.

4.5.3 Removing MDisks from storage pools

You might want to remove MDisks from a storage pool (for example, when you decommission a storage controller or drive shelf). When you remove MDisks from a storage pool, consider whether to manually migrate extents from the MDisks. It is also necessary to make sure that you remove the correct MDisks.

Note: The removal of MDisks occurs only if sufficient space is available to migrate the volume data to other extents on other MDisks that remain in the storage pool. After you remove the MDisk from the storage pool, it takes time to change the mode from managed to unmanaged, depending on the size of the MDisk that you are removing.

When you remove the MDisk made of internal disk drives from the storage pool, the MDisk is deleted. This process also deletes the array on which this MDisk was built, and converts all drives that were included in this array to a candidate state. You can now use those disk drives to create another array of a different size and RAID type, or you can use them as hot spares.

When external MDisk is deleted from a pool, it turns to *unmanaged* state. After that you can use back-end controller management interface to remove the MDisk.

Migrating extents from the MDisk to be deleted

If an MDisk contains volume extents, you must move these extents to the remaining MDisks in the storage pool. Example 4-3 shows how to list the volumes that have extents on an MDisk by using the CLI.

Example 4-3 Listing of volumes that have extents on an MDisk to be deleted

```
IBM_2145:itsosvcc11:admin>lsmdiskextent mdisk14
id          number_of_extents  copy_id
5           16                 0
3           16                 0
6           16                 0
8           13                 1
9           23                 0
8           25                 0
```

DRP restriction: The `lsmdiskextent` command does not provide accurate extent usage for thin-provisioned or compressed volumes on DRPs.

Specify the `-force` flag on the `rmmdisk` command, or select the corresponding option in the GUI. Both actions cause the system to automatically move all used extents on the MDisk to the remaining MDisks in the storage pool. Even if you use `-force` flag, system will still perform appropriate checks to make sure that you can safely remove the MDisks from a pool.

Alternatively, you might want to manually perform the extent migrations. Otherwise, the automatic migration randomly allocates extents to MDisks (and areas of MDisks). After all the extents are manually migrated, the MDisk removal can proceed without the `-force` flag.

Verifying the identity of an MDisk before removal

External MDisks must appear to the system as unmanaged before their controller LUN mapping is removed. Unmapping LUNs that are still part of a storage pool results in the storage pool going offline, which affects all hosts with mappings to volumes in that storage pool.

If the MDisk was named by using the best practices, the correct LUNs are easier to identify. However, ensure that the identification of LUNs that are being unmapped from the controller match the associated MDisk on IBM FlashSystem and SVC by using the Controller LUN Number field and the unique identifier (UID) field.

The UID is unique across all MDisks on all controllers. However, the controller LUN is unique only within a specified controller and for a certain host. Therefore, when you use the controller LUN, check that you are managing the correct storage controller and that you are looking at the mappings for the correct IBM Storage Virtualize host object.

Tip: Renaming your back-end storage controllers also helps you with MDisk identification.

Correlating the back-end volume with the MDisk

The correct correlation between the back-end volume (LUN) with the external MDisk is crucial to avoid mistakes and possible outages. You can correlate the back-end volume with MDisk for DS8000 series, XIV, and IBM FlashSystem storage controllers.

DS8000 LUN

The LUN ID uniquely identifies LUNs only within the same storage controller. If multiple storage devices are attached to the same IBM Storage Virtualize cluster, the LUN ID must be combined with the worldwide node name (WWNN) attribute to uniquely identify LUNs within the IBM Storage Virtualize.

To get the WWNN of the DS8000 controller, take the first 16 digits of the MDisk UID and change the first digit from 6 to 5, such as 6005076305ffc74c to 5005076305ffc74c. When detected as IBM FlashSystem and SVC ctrl_LUN_#, the DS8000 LUN is decoded as 40XX40YY00000000, where XX is the logical subsystem (LSS) and YY is the LUN within the LSS. As detected by the DS8000, the LUN ID is the four digits starting from the 29th digit, as shown in Example 4-4.

Example 4-4 DS8000 UID example

```
6005076305ffc74c000000000000001007000000000000000000000000000000000000000000000
```

In Example 4-4, you can identify the MDisk that is supplied by the DS8000, which is LUN ID 1007.

XIV system volumes

Identify the XIV volumes by using the volume serial number and the LUN that is associated with the host mapping. The example in this section uses the following values:

- ▶ Serial number: 897
- ▶ LUN: 2

Complete the following steps:

1. To identify the volume serial number, right-click a volume and select **Properties**. Example 4-5 shows the Volume Properties dialog box that opens.

Example 4-5 The lscontroller command

```
IBM_2145:tpcsvc62:admin>lscontroller 10
id 10
controller_name controller10
WWNN 5001738002860000
...
```

2. To identify your LUN, in the volumes by Hosts view, expand your IBM FlashSystem and SVC host group and then review the LUN column.
3. The MDisk UID field consists of part of the controller WWNN from bits 2 - 13. You might check those bits by using the **lscontroller** command.
4. The correlation can now be performed by taking the first 16 bits from the MDisk UID field:
 - Bits 1 - 13 refer to the controller WWNN.
 - Bits 14 - 16 are the XIV volume serial number (897) in hexadecimal format (resulting in 381 hex).
 - The conversion is
001738000286038100.

Where:

- The controller WWNN (bits 2 - 13) is 0017380002860.
- The XIV volume serial number that is converted to hex is 381.

5. To correlate the IBM Storage Virtualize `ctr1_LUN_#`:
 - a. Convert the XIV volume number to hexadecimal format.
 - b. Check the last 3 bits from the IBM Storage Virtualize `ctr1_LUN_#`.
 In this example, the number is 0000000000000002, as shown in Figure 4-8.

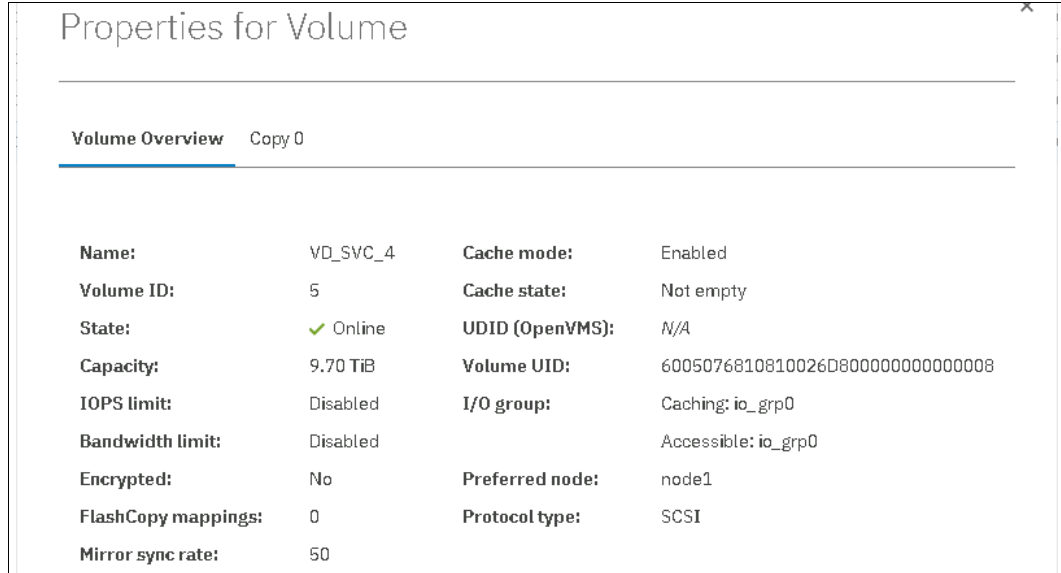


Figure 4-8 IBM FlashSystem volume details

IBM FlashSystem volumes

To correlate the IBM FlashSystem volumes with the external MDisks that are seen by the virtualizer, complete the following steps:

1. From the back-end IBM FlashSystem side, check the Volume UID field for the volume that was presented to the virtualizer, as shown in Figure 4-8.
2. On the **Host Maps** tab, check the SCSI ID number for the specific volume, as shown in Figure 4-9 on page 284. This value is used to match the virtualizer `ctr1_LUN_#` (in hexadecimal format).

Host Details: SVC

Overview Mapped Volumes Port Definitions

Volumes Mapped to the Host

Default Contains Filter

SCSI ID	Name ↑	UID	Caching I/O...	
5	VD_SVC_4	6005076810810026D800000000000008	0	
2	VD_SVC...	6005076810810026D800000000000005	0	
3	VD_SVC...	6005076810810026D800000000000006	0	
4	VD_SVC...	6005076810810026D800000000000007	0	
0	VD_SVC...	6005076810810026D800000000000003	0	
1	VD_SVC...	6005076810810026D800000000000004	0	

Showing 6 mappings | Selecting 1 mapping

Figure 4-9 IBM FlashSystem volume details for host maps

- At the virtualizer, review the MDisk details and compare the MDisk UID field with the IBM FlashSystem Volume UID, as shown in Figure 4-10 on page 285. The first 32 bits should be the same.

FlashSystem and IBM SAN Volume Controller: Updated for IBM Storage Virtualize Version 8.6, SG24-8542) or in the back-end system logs.

4.5.5 Controlling the extent allocation order for volume creation

When creating a volume on a standard pool, the allocation of extents is performed by using a round-robin algorithm, taking one extent from each MDisk in the pool in turn.

The first MDisk to allocate an extent from is chosen in a pseudo-random way rather than always starting from the same MDisk. The pseudo-random algorithm avoids the situation where the “striping effect” inherent in a round-robin algorithm places the first extent for many volumes on the same MDisk.

Placing the first extent of a number of volumes on the same MDisk might lead to poor performance for workloads that place a large I/O load on the first extent of each volume or that create multiple sequential streams.

However, this allocation pattern is unlikely to remain for long because Easy Tier balancing moves the extents to balance the load evenly across all MDisk in the tier. The hot and cold extents also are moved between tiers.

In a multitier pool, the middle tier is used by default for new volume creation. If free space is not available in the middle tier, the cold tier is used if it exists. If the cold tier does not exist, the hot tier is used. For more information about Easy Tier, see 4.7, “Easy Tier and tiered and balanced storage pools” on page 296.

DRP restriction: With compressed and deduplicated volumes on DRP, the extent distribution cannot be checked across the MDisks. Initially, only a minimal number of extents are allocated to the volume, based on the `rsize` parameter.

4.6 Considerations when using encryption

IBM Storage Virtualize supports optional encryption of data at rest. This support protects against the potential exposure of sensitive user data and user metadata that is stored on discarded, lost, or stolen storage devices. To use encryption on the system, an encryption license is required for each cluster I/O group.

Note: Hot Spare Nodes (HSNs) also need encryption licenses if they are to be used to replace the failed nodes that support encryption.

Two ways to storage encryption keys are supported:

- ▶ Encryption key servers. System supports IBM Security Key Lifecycle Manager, Gemalto SafeNet KeySecure and Thales CipherTrust Manager. For more information, see [IBM Storage Virtualize Supported Key Servers](#).
- ▶ Storing encryption keys on USB flash drives. USB flash drive-based encryption requires physical access to the systems and is effective in environments with a minimal number of systems.

4.6.1 General considerations

When encryption is activated and enabled, valid encryption keys must be present on the system when the system unlocks the drives or the user generates a new key. If USB encryption is enabled on the system, the encryption key must be stored on USB flash drives that contain a copy of the key that was generated when encryption was enabled. If key server encryption is enabled on the system, the key is retrieved from the key server.

It is not possible to convert the existing data to an encrypted copy in-place. You can use the volume migration function to migrate the data to an encrypted storage pool or encrypted child pool. Alternatively, you can also use the volume mirroring function to add a copy to an encrypted storage pool or encrypted child pool and delete the unencrypted copy after the migration.

Before you activate and enable encryption, you must determine the method of accessing key information during times when the system requires an encryption key to be present. The system requires an encryption key to be present during the following operations:

- ▶ System power-on
- ▶ System restart
- ▶ User-initiated rekey operations
- ▶ System recovery

Several factors must be considered when planning for encryption:

- ▶ Physical security of the system.
- ▶ Need and benefit of manually accessing encryption keys when the system requires them.
- ▶ Availability of key data.
- ▶ Encryption license is purchased, activated, and enabled on the system.

For more information about configuration details about IBM Storage Virtualize encryption, see *Implementation Guide for IBM Storage FlashSystem and IBM SAN Volume Controller: Updated for IBM Storage Virtualize Version 8.6*, SG24-8542.

4.6.2 Hardware and software encryption

Encryption can be performed by using hardware encryption or software encryption.

Both methods protect against the potential exposure of sensitive user data that is stored on discarded, lost, or stolen media. Both methods can facilitate the warranty return or disposal of hardware. The method that is used for encryption is chosen automatically by the system based on the placement of the data.

Figure 4-11 shows encryption placement in the lower layers of the IBM Storage Virtualize software stack.

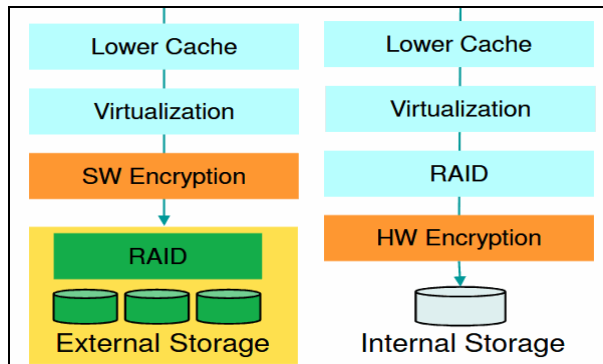


Figure 4-11 Encryption placement in lower layers of the IBM FlashSystem software stack

Hardware encryption

Hardware encryption has the following characteristics:

- ▶ The algorithm is a built-in SAS chip for all SAS-attached drives, or built in to the drive itself for NVMe-attached drives (FCM, industry-standard NVMe, and SCM).
- ▶ No system overhead.
- ▶ Only available to direct-attached SAS or NVMe disks.
- ▶ Can be enabled only when you create internal arrays.
- ▶ Child pools cannot be encrypted if the parent storage pool is not encrypted.
- ▶ Child pools are automatically encrypted if the parent storage pool is encrypted, but can have different encryption keys.
- ▶ DRP child pools can use only the same encryption key as their parent.

Software encryption-only storage pool

Software encryption has the following characteristics:

- ▶ The algorithm is running at the interface device driver.
- ▶ Uses a special CPU instruction set and engines (AES_NI).
- ▶ Allows encryption for virtualized external storage controllers, which are not capable of self-encryption.
- ▶ Less than 1% system overhead.
- ▶ Available only to virtualized external storage.
- ▶ Is enabled only when you create storage pools and child pools that are made up of virtualized external storage.
- ▶ Child pools can be encrypted even if the parent storage pool is not encrypted.

Mixed encryption in a storage pool

It is possible to mix hardware and software encryption in a storage pool, as shown in Figure 4-12.

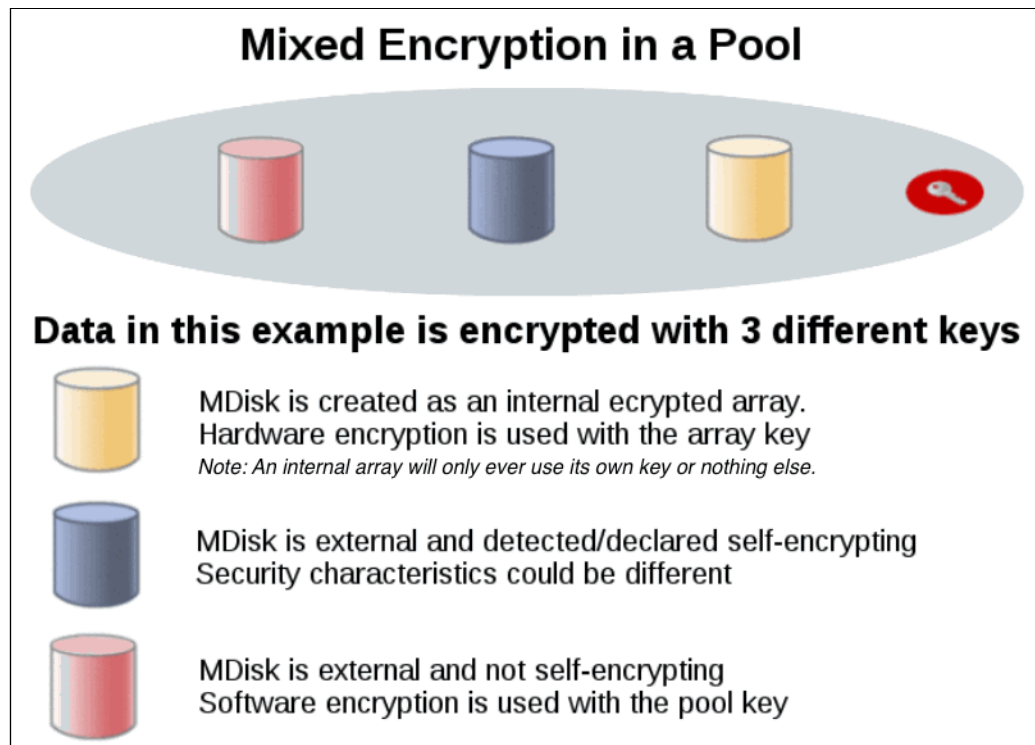


Figure 4-12 Mixed encryption in a storage pool

However, if you want to create encrypted child pools from an unencrypted storage pool containing a mix of internal arrays and external MDisks, the following restrictions apply:

- ▶ The parent pool must not contain any unencrypted internal arrays.
- ▶ All system nodes must have an activated encryption license.

Note: An encrypted child pool that is created from an unencrypted parent storage pool reports as unencrypted if the parent pool contains unencrypted internal arrays. Remove these arrays to ensure that the child pool is fully encrypted.

The general rule is to not mix different types of MDisks in a storage pool unless you intend to use the Easy Tier tiering function. In this scenario, the internal arrays must be encrypted if you want to create encrypted child pools from an unencrypted parent storage pool. All methods of encryption use the same encryption algorithm, the same key management infrastructure, and the same license.

Note: In a virtualized storage environment, always implement encryption on the self-encryption capable back-end storage, such as IBM FlashSystem.

Declare or identify the self-encrypted virtualized external MDisk as encrypted on by setting the **-encrypt** option to **yes** in the **chmdisk** command, as shown in Example 4-6. This configuration is important to prevent back-end storage from trying to encrypt them again.

Example 4-6 Command to declare or identify a self-encrypted MDisk from a virtualized external storage

```
IBM_2145:ITSO_A:superuser>chmdisk -encrypt yes mdisk0
```

Note: It is important to declare or identify the self-encrypted MDisks from a virtualized external storage before creating an encrypted storage pool or child pool on IBM Storage Virtualize.

4.6.3 Encryption at rest with USB keys

The following section describes the characteristics of using USB flash drives for encryption and the available options to access the key information.

USB flash drives have the following characteristics:

- ▶ Physical access to the system is required to process a rekeying operation.
- ▶ No mechanical components to maintain with almost no read/write operations to the USB flash drive.
- ▶ Inexpensive to maintain and use.
- ▶ Convenient and easy to have multiple identical USB flash drives available as backups.

Two options are available for accessing key information on USB flash drives:

- ▶ USB flash drives are left inserted in the system always.

If you want the system to restart automatically, a USB flash drive must be left inserted in all the nodes on the system. When you power on, then all nodes have access to the encryption key. This method requires that the physical environment where the system is located is secure. If the location is secure, it prevents an unauthorized person from making copies of the encryption keys, stealing the system, or accessing data that is stored on the system.

- ▶ USB flash drives are not left inserted into the system except as required.

For the most secure operation, do not keep the USB flash drives inserted into the nodes on the system. However, this method requires that you manually insert the USB flash drives that contain copies of the encryption key into the nodes during operations that the system requires an encryption key to be present. USB flash drives that contain the keys must be stored securely to prevent theft or loss.

4.6.4 Encryption at rest with key servers

The following section describes the characteristics of using key servers for encryption and essential recommendations for key server configuration with IBM FlashSystem and SVC.

Key servers

Key servers have the following characteristics:

- ▶ Physical access to the system is not required to process a rekeying operation.
- ▶ Support for businesses that have security requirements to not use USB ports.
- ▶ Strong key generation.
- ▶ Key self-replication and automatic backups.
- ▶ Implementations follow an open standard that aids in interoperability.
- ▶ Audit details.
- ▶ Ability to administer access to data separately from storage devices.

Encryption key servers create and manage encryption keys that are used by the system. In environments with many systems, key servers distribute keys remotely without requiring physical access to the systems. A key server is a centralized system that generates, stores, and sends encryption keys to the system. If the key server provider supports replication of keys among multiple key servers, you can specify up to four key servers (one master and three clones) that connect to the system over both a public network or a separate private network.

Recommendations for IBM GKLM key server configuration

The following section provides some essential recommendations for key server configuration with IBM Storage Virtualize.

Self-signed certificate type and validity period

The default certificate type on IBM Security Guardian Key Lifecycle Manager and IBM Storage Virtualize is RSA. If you use a different certificate type, make sure that you match the certificate type on both ends. The default certificate validity period is 1095 days on IBM Security Key Lifecycle Manager server and 5475 days on IBM Storage Virtualize.

You can adjust the validity period to comply with specific security policies and always match the certificate validity period on IBM FlashSystem, SVC, and IBM Security Guardian Key Lifecycle Manager servers. A mismatch causes a certificate authorization error and leads to unnecessary certificate exchange.

In the management GUI, select **Settings** → **Security** → **Secure Communications** → **Update Certificate**.

Figure 4-13 on page 292 shows the default certificate type and validity period on IBM Storage Virtualize.

Update Certificate

Certificate type: Self-signed certificate
 Signed certificate

Key type: 2048-bit RSA

Validity days: 5,475

Figure 4-13 Update Certificate on IBM FlashSystem

Figure 4-14 shows the default certificate type and validity period on an IBM Security Key Lifecycle Manager server.

Self-signed Certificate

*Certificate label in keystore:

*Certificate description (common name):

*Validity period of new certificate (in days; for example, 3 years is 365 x 3 = 1095 days):
 The interval in days ranges from 1 to 9000

*Algorithm:

Figure 4-14 Creating a self-signed certificate: IBM Security Key Lifecycle Manager server

Device group configuration

The IBM Spectrum_VIRT device group is not predefined on IBM Security Key Lifecycle Manager. It must be created based on an IBM General Parallel File System (GPFS) device, as shown in Figure 4-15.

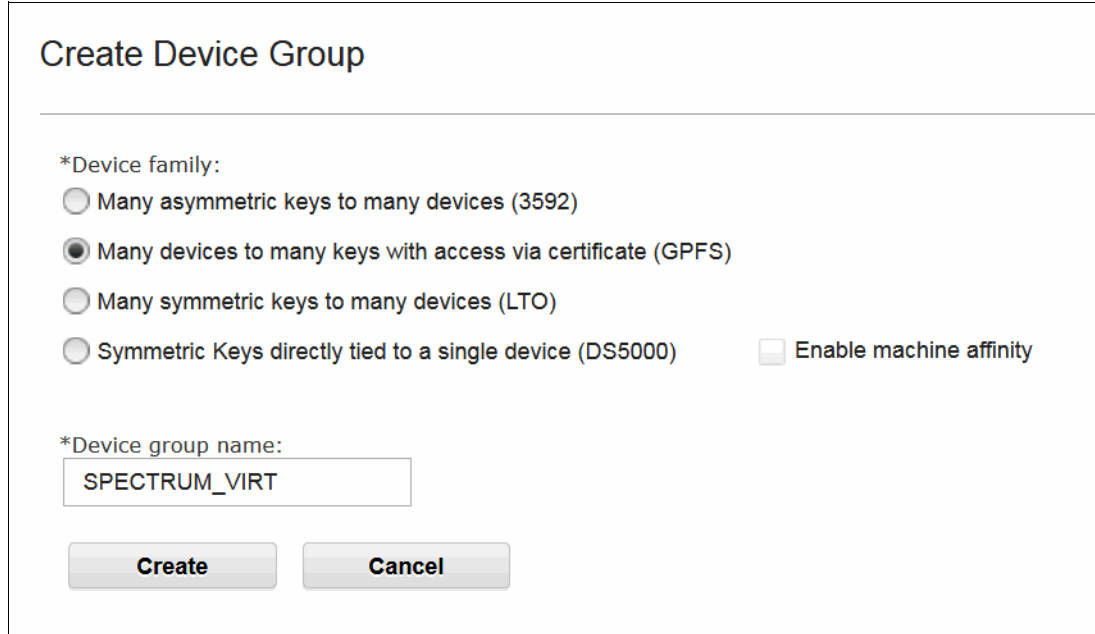


Figure 4-15 Create Device Group for IBM FlashSystem

By default, in IBM Storage Virtualize, the *IBM Spectrum_VIRT* group name is predefined in the encryption configuration wizard. *IBM Spectrum_VIRT* contains all the keys for the managed IBM FlashSystem and SVC. However, it is possible to use different device groups if they are GPFS device-based, for example, one device group for each environment (production or disaster recovery (DR)). Each device group maintains its own key database, and this approach allows more granular key management.

Clone servers configuration management

The minimum replication interval on IBM Security Guardian Key Lifecycle Manager is 1 hour, as shown in Figure 4-16 on page 294. It is more practical to perform backup and restore or manual replication for the initial configuration to speed up the configuration synchronization.

Also, the rekey process creates a configuration on the IBM Security Guardian Key Lifecycle Manager server, and it is important not to wait for the next replication window but to manually synchronize the configuration to the extra key servers (clones). Otherwise, an error message is generated by the IBM Storage Virtualize system, which indicates that the key is missing on the clones.

Figure 4-16 on page 294 shows the replication interval.

Basic Properties **Advance Properties**

Replication backup destination directory:

Maximum number of replication files to keep before rollover:

▼ Replication Scheduler Section

Replication frequency (in hours):

Daily replication time (in HH:MM format):

▼ Replication Log Section

Replication log file name:

Maximum log file size (in KB):

Maximum number of log files to keep:

Figure 4-16 IBM Security Key Lifecycle Manager replication schedule

Example 4-7 shows an example of manually triggered replication.

Example 4-7 Manually triggered replication

```
/opt/IBM/WebSphere/AppServer/bin/wsadmin.sh -username SKLMAdmin -password
<password> -lang jython -c "print AdminTask.tklmReplicationNow()"
```

Encryption key management

There is always only one active key for each encryption-enabled system. The previously used key is deactivated after the rekey process. It is possible to delete the deactivated keys to keep the key database tidy and up to date.

Figure 4-17 on page 295 shows the keys that are associated with a device group. In this example, the SG247933_BOOK device group contains one encryption-enabled system, and it has three associated keys. Only one of the keys is activated, and the other two were deactivated after the rekey process.

SG247933_BOOK

The screen below allows you to add or delete certificate and their associated node name. As well as, modify the node name associated with a certificate. New keys can be added and a name associated with that key(s).

Home Refresh Add Modify Delete

No filter applied		
Certificate UUID	Name	Endpoint Count
CERTIFICATE-8a89d57-70cf447-adda-4b29-9b1c-89c200fd1745	sg247933_redbook	2

Total: 1 Selected: 0

No filter applied	
Key UUID	Name
KEY-8a89d57-15bf8f41-cea6-4df3-8f4e-be0c36318615	mmm008a89d57000000870
KEY-8a89d57-74edaef9-b6d9-4766-9b39-7e21d9911011	mmm008a89d5700000086e
KEY-8a89d57-ebe5d5a1-8987-4aff-ab58-5f808a078269	mmm008a89d5700000086f

Total: 3 Selected: 0

Figure 4-17 Keys that are associated to a device group

Example 4-8 shows an example to check the state of the keys.

Example 4-8 Verifying the key state

```
/opt/IBM/WebSphere/AppServer/bin/wsadmin.sh -username SKLMAdmin -password
<password> -lang jython
wsadmin>print AdminTask.tklmKeyList('[-uuid
KEY-8a89d57-15bf8f41-cea6-4df3-8f4e-be0c36318615]')
CTGKM0001I Command succeeded.
```

```
uuid = KEY-8a89d57-15bf8f41-cea6-4df3-8f4e-be0c36318615
alias = mmm008a89d57000000870
key algorithm = AES
key store name = defaultKeyStore
key state = ACTIVE
creation date = 18/11/2017, 01:43:27 Greenwich Mean Time
expiration date = null
```

```
wsadmin>print AdminTask.tklmKeyList('[-uuid
KEY-8a89d57-74edaef9-b6d9-4766-9b39-7e21d9911011]')
CTGKM0001I Command succeeded.
```

```
uuid = KEY-8a89d57-74edaef9-b6d9-4766-9b39-7e21d9911011
alias = mmm008a89d5700000086e
key algorithm = AES
key store name = defaultKeyStore
key state = DEACTIVATED
creation date = 17/11/2017, 20:07:19 Greenwich Mean Time
expiration date = 17/11/2017, 23:18:37 Greenwich Mean Time
```

```
wsadmin>print AdminTask.tklmKeyList('[-uuid
KEY-8a89d57-ebe5d5a1-8987-4aff-ab58-5f808a078269]')
```

CTGKM0001I Command succeeded.

```
uuid = KEY-8a89d57-ebe5d5a1-8987-4aff-ab58-5f808a078269
alias = mmm008a89d5700000086f
key algorithm = AES
key store name = defaultKeyStore
key state = DEACTIVATED
creation date = 17/11/2017, 23:18:34 Greenwich Mean Time
expiration date = 18/11/2017, 01:43:32 Greenwich Mean Time
```

Note: The initial configuration, such as certificate exchange and TLS configuration, is required only on the master IBM Security Key Lifecycle Manager server. The restore or replication process duplicates all the required configurations to the clone servers.

If encryption was enabled on a pre-7.8.0 system and the system is updated to later releases, you must run a USB rekey operation to enable key server encryption. Run the **chencryption** command before you enable key server encryption. To perform a rekey operation, run the commands that are shown in Example 4-9.

Example 4-9 Commands to enable the key server encryption option on a system upgraded from pre-7.8.0

```
chencryption -usb newkey -key prepare
chencryption -usb newkey -key commit
```

4.7 Easy Tier and tiered and balanced storage pools

Easy Tier was originally developed to provide the maximum performance benefit of a few SSDs or flash drives. Because of their low response times, high throughput, and IOPS characteristics, SSDs and flash arrays were a welcome addition to the storage system, but initially their acquisition cost per gigabyte was more than for HDDs.

By implementing an evolving artificial intelligence (AI) like algorithm, Easy Tier moved the most frequently accessed blocks of data to the lowest latency device. Therefore, it provides an exponential improvement in performance when compared to a small investment in SSD and flash capacity.

The industry moved on in the more than 10 years since Easy Tier was first introduced. The current cost of SSD and flash-based technology meant that more users can deploy all-flash environments.

HDD-based large capacity NL-SAS drives are still the most cost-effective online storage devices. Although SSD and flash ended the 15K RPM drive market, it has yet to reach a price point that competes with NL-SAS for lower performing workloads. The use cases for Easy Tier changed, and most deployments now use either all-flash or “flash and trash” approaches, with 10% or more flash capacity and the remainder using NL-SAS.

Easy Tier also provides balancing within a tier. This configuration ensures that no one single component within a tier of the same capabilities is more heavily loaded than another one. It does so to maintain an even latency across the tier and help to provide consistent and predictable performance.

As the industry strives to develop technologies that can enable higher throughput and lower latency than even flash, Easy Tier continues to provide user benefits. For example, SCM technologies, which were introduced to IBM FlashSystem in 2020, now provide lower latency than even flash, but as with flash when it was first introduced, at a considerably higher cost of acquisition per gigabyte.

Choosing the correct mix of drives and data placement is critical to achieve optimal performance at the lowest cost. Maximum value can be derived by placing “hot” data with high I/O density and low response time requirements on the highest tier, while targeting lower tiers for “cooler” data, which is accessed more sequentially and at lower rates.

Easy Tier dynamically automates the ongoing placement of data among different storage tiers. It can be enabled for internal and external storage to achieve optimal performance.

The Easy Tier feature that is called *storage pool balancing* automatically moves extents within the same storage tier from overloaded to less loaded MDisks. Storage pool balancing ensures that your data is optimally placed among all disks within storage pools.

Storage pool balancing is designed to balance extents between tiers in the same pool to improve overall system performance and avoid overloading a single MDisk in the pool.

However, Easy Tier considers only performance, and it does *not* consider capacity. Therefore, if two FCM arrays are in a pool and one of them is nearly out of space and the other is empty, Easy Tier does not attempt to move extents between the arrays.

For this reason, it is recommended that if you must increase the capacity on an MDisk, increase the size of the array rather than add an FCM array.

4.7.1 Easy Tier concepts

IBM Storage Virtualize products implement Easy Tier enterprise storage functions, which were originally designed with the development of Easy Tier on IBM DS8000 enterprise-class storage systems. Easy Tier enables automated sub-volume data placement throughout different or within the same storage tiers. This feature intelligently aligns the system with current workload requirements and optimizes the usage of high-performance storage, such as SSD, FCM, and SCM.

Easy Tier reduces the I/O latency for hot spots, but it does not replace storage cache. Both Easy Tier and storage cache solve a similar access latency workload problem. However, these two methods weigh differently in the algorithmic construction that is based on *locality of reference*, recency, and frequency. Easy Tier monitors I/O performance from the device end (after cache), it can pick up the performance issues that cache cannot solve, and complements the overall storage system performance.

The primary benefit of Easy Tier is to reduce latency for hot spots, but this feature includes an added benefit where the remaining “medium” (that is, not cold) data has less contention for its resources and performs better as a result (that is, lower latency).

Easy Tier can be used in a single tier pool to balance the workload across storage MDisks. It ensures an even load on all MDisks in a tier or pool. Therefore, bottlenecks and convoying effects are removed when striped volumes are used. In a multitier pool, each tier is balanced.

In general, the storage environment’s I/O is monitored at a volume level, and the entire volume is always placed inside one suitable storage tier. Determining the amount of I/O, moving part of the underlying volume to an appropriate storage tier, and reacting to workload changes is too complex for manual operation. It is in this situation that the Easy Tier feature can be used.

Easy Tier is a performance optimization function that automatically migrates extents that belong to a volume between different storage tiers (see Figure 4-18) or the same storage tier. Because this migration works at the extent level, it is often referred to as sublogical unit number (LUN) migration. Movement of the extents is dynamic, nondisruptive, and not visible from the host perspective. As a result of extent movement, the volume no longer has all its data in one tier; rather, it is in two or three tiers, or balanced between MDisks in the same tier.

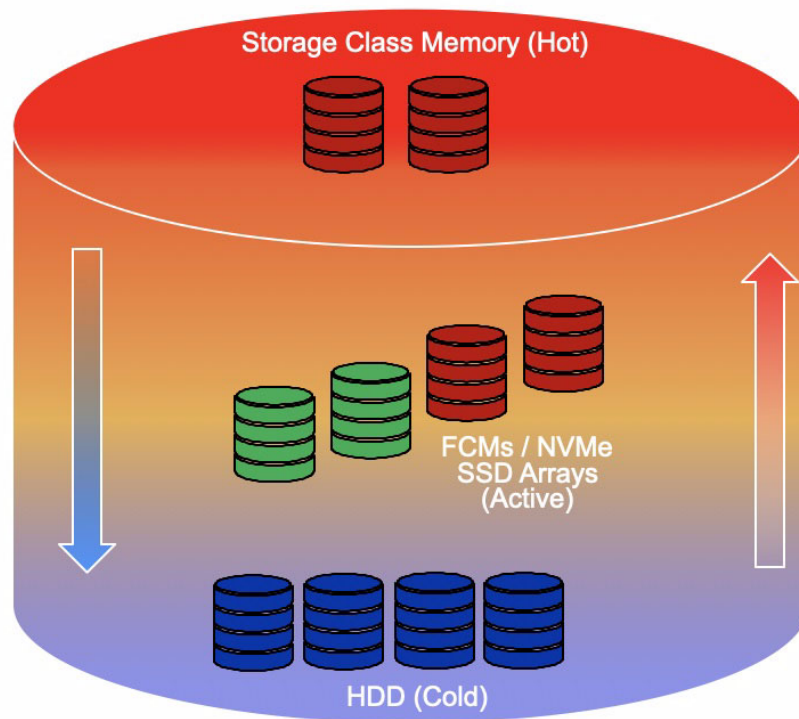


Figure 4-18 Easy Tier single volume with multiple tiers

You can enable Easy Tier on a per volume basis, except for non-fully allocated volumes in a DRP where Easy Tier is always enabled. It monitors the I/O activity and latency of the extents on all Easy Tier enabled volumes.

Based on the performance characteristics, Easy Tier creates an extent migration plan and dynamically moves (promotes) high activity or hot extents to a higher disk tier within the same storage pool. Generally, a new migration plan is generated on a stable system once every 24 hours. Instances might occur when Easy Tier reacts within 5 minutes, for example, when detecting an overload situation.

It also moves (demotes) extents whose activity dropped off, or cooled, from higher disk tier MDisks back to a lower tier MDisk. When Easy Tier runs in a storage pool rebalance mode, it moves extents from busy MDisks to less busy MDisks of the same type.

Note: Image mode and sequential volumes are not candidates for Easy Tier automatic data placement because all extents for those types of volumes must be on one specific MDisk, and they cannot be moved.

4.7.2 Easy Tier definitions

Easy Tier measures and classifies each extent into one of its three tiers. It performs this classification process by looking for extents that are the outliers in any system:

1. It looks for the hottest extents in the pool. These extents contain the most frequently accessed data of a suitable workload type (less than 64 KiB I/O). Easy Tier plans to migrate these extents into whatever set of extents that comes from MDisks that are designated as the hot tier.
2. It looks for coldest extents in the pool, which are classed as having done < 1 I/O in the measurement period. These extents are planned to be migrated onto extents that come from the MDisks that are designated as the cold tier.
3. It is not necessary for Easy Tier to look for extents to place in the middle tier. By definition, if something is not designated as “hot” or “cold”, it stays or is moved to extents that come from MDisks in the middle tier.

With these three tier classifications, an Easy Tier pool can be optimized.

Internal processing

The Easy Tier function includes the following four main processes:

► I/O monitoring

This process operates continuously and monitors volumes for host I/O activity. It collects performance statistics for each extent, and derives averages for a rolling 24-hour period of I/O activity.

Easy Tier makes allowances for large block I/Os; therefore, it considers only I/Os of up to 64 kilobytes (KiB) as migration candidates.

This process is efficient and consumes negligible processing resources of the system's nodes.

► Data Placement Advisor (DPA)

The DPA uses workload statistics to make a cost-benefit decision about which extents will be candidates for migration to a higher performance tier.

This process identifies extents that can be migrated back to a lower tier.

► Data Migration Planner (DMP)

By using the extents that were previously identified, the DMP builds the extent migration plans for the storage pool. The DMP builds two plans:

- The Automatic Data Relocation (ADR) mode plan to migrate extents across adjacent tiers.
- The Rebalance (RB) mode plan to migrate extents within the same tier.

► Data migrator

This process involves the actual movement or migration of the volume's extents up to, or down from, the higher disk tier. The extent migration rate is capped so that a maximum of up to 12 GiB every 5 minutes is migrated, which equates to approximately 3.4 TiB per day that is migrated between disk tiers.

Note: You can increase the target migration rate to 48 GiB every 5 minutes by temporarily enabling accelerated mode. For more information, see “Easy Tier acceleration” on page 317.

When active, Easy Tier performs the following actions across the tiers:

► Promote

Moves the hotter extents to a higher performance tier with available capacity. Promote occurs within adjacent tiers.

► Demote

Demotes colder extents from a higher tier to a lower tier. Demote occurs within adjacent tiers.

► Swap

Exchanges a cold extent in an upper tier with a hot extent in a lower tier.

► Warm demote

Prevents performance overload of a tier by demoting a warm extent to a lower tier. This process is triggered when the bandwidth or IOPS exceeds a predefined threshold. If you see these operations, it is a trigger to suggest that you should add more capacity to the higher tier.

► Warm promote

This feature addresses the situation where a lower tier suddenly becomes active. Instead of waiting for the next migration plan, Easy Tier can react immediately. Warm promote acts in a similar way to warm demote. If the 5-minute average performance shows that a layer is overloaded, Easy Tier immediately starts to promote extents until the condition is relieved. This action is often referred to as “overload protection”.

► Cold demote

Demotes inactive (or cold) extents that are on a higher performance tier to its adjacent lower-cost tier. Therefore, Easy Tier automatically frees extents on the higher storage tier before the extents on the lower tier become hot. Only supported between HDD tiers.

► Expanded cold demote

Demotes appropriate sequential workloads to the lowest tier to better use nearline disk bandwidth.

- ▶ Auto rebalance

Redistributes extents within a tier to balance usage across MDisks for maximum performance. This process moves hot extents from high-use MDisks to low-use MDisks, and exchanges extents between high-use MDisks and low-use MDisks.
- ▶ Space reservation demote

Introduced in version 8.4.0 to prevent out-of-space conditions on thin-provisioned (compressed) back-ends, Easy Tier stops the migration of new data into a tier, and if necessary, migrates extents to a lower tier.

Easy Tier attempts to migrate the most active volume extents up to an SSD first.

If a new migration plan is generated before the completion of the previous plan, the previous migration plan and queued extents that are not yet relocated are abandoned. However, migrations that are still applicable are included in the new plan.

Note: Extent migration occurs only between adjacent tiers. For example, in a three-tiered storage pool, Easy Tier does not move extents from the flash tier directly to the nearline tier and vice versa without moving them first to the enterprise tier.

The Easy Tier extent migration types are shown in Figure 4-19.

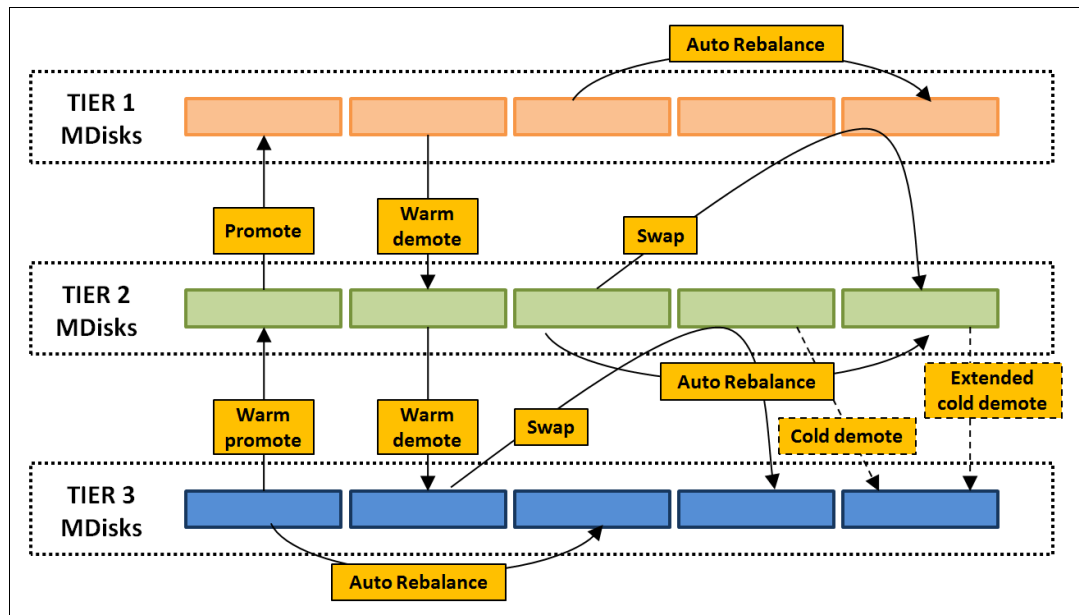


Figure 4-19 Easy Tier extent migration types

4.7.3 Easy Tier operating modes

Easy Tier includes the following main operating modes:

- ▶ Off
- ▶ On
- ▶ Automatic
- ▶ Measure

Easy Tier is a licensed feature on some IBM Storage Virtualize hardware platforms. If the license is not present and Easy Tier is set to Auto or On, the system runs in Measure mode.

Options: The Easy Tier function can be turned on or off at the storage pool level *and* at the volume level, except for non-fully allocated volumes in a DRP where Easy Tier is always enabled.

Easy Tier off mode

With Easy Tier turned off, statistics are not recorded, and cross-tier extent-migration does not occur.

Measure mode

Easy Tier can be run in an evaluation or measurement-only mode, and it collects usage statistics for each extent in a storage pool where the Easy Tier value is set to Measure.

This collection is typically done for a single-tier pool so that the benefits of adding more performance tiers to the pool can be evaluated before any major hardware acquisition.

The heat and activity of each extent can be viewed in the GUI by selecting **Monitoring** → **Easy Tier Reports**.

Automatic mode

In Automatic mode, the storage pool parameter `-easytier auto` must be set, and the volumes in the pool must have `-easytier` set to `on`.

The behavior of Easy Tier depends on the pool configuration. Consider the following points:

- ▶ If the pool contains only MDisks with a single tier type, the pool is in balancing mode.
- ▶ If the pool contains MDisks with more than one tier type, the pool runs automatic data placement and migration in addition to balancing within each tier.

Dynamic data movement is transparent to the host server and application users of the data, other than providing improved performance. Extents are automatically migrated, as explained in “Implementation rules” on page 312.

There might be situations where the Easy Tier setting is “auto” but the system is running in monitoring mode only, for example, with unsupported tier types or if you have not enabled the Easy Tier license. For more information, see Table 4-9 on page 308.

The GUI provides the same reports as available in measuring mode, and in addition provide the data movement report that shows the breakdown of the actual migration events that are triggered by Easy Tier. These migrations are reported in terms of the migration types, as described in “Internal processing” on page 299.

Easy Tier on mode

This mode forces Easy Tier to perform the tasks as in Automatic mode.

For example, when Easy Tier detects an unsupported set of tier types in a pool, as outlined in Table 4-9 on page 308, using the “on” mode forces Easy Tier to the active state, and it performs to the best of its ability. The system raises an alert. There is an associated Directed Maintenance Procedure that guides you to fix the unsupported tier types.

Important: Avoid creating a pool with more than three tiers. Although the system attempts to create generic hot, medium, and cold “buckets”, you might end up with Easy Tier running in measure mode only.

These configurations are unsupported because they can cause performance problems in the long term, for example, disparate performance within a single tier.

The ability to override Automatic mode is provided to enable temporary migration from an older set of tiers to new tiers, which must be rectified as soon as possible.

Storage pool balancing

This feature assesses the extents that are written in a pool and balances them automatically across all MDisks within the pool. This process works with Easy Tier when multiple classes of disks exist in a single pool. In this case, Easy Tier moves extents between the different tiers, and storage pool balancing moves extents within the same tier to enable a balance in terms of workload across all MDisks that belong to a tier.

Balancing is when you maintain equivalent latency across all MDisks in a tier, which can result in different capacity usage across the MDisks. However, performance balancing is preferred over capacity-balancing in most cases.

The process automatically balances existing data when new MDisks are added into an existing pool, even if the pool contains only a single type of drive.

Balancing is automatically active on all storage pools no matter what the Easy Tier setting is. For a single tier pool, the Easy Tier state reports as Balancing.

Note: Storage pool balancing can be used to balance extents when mixing different-sized disks of the same performance tier. For example, when adding larger capacity drives to a pool with smaller capacity drives of the same class, storage pool balancing redistributes the extents to leverage the extra performance of the new MDisks.

Easy Tier mode settings

The Easy Tier setting can be changed on a storage pool and volume level. Depending on the Easy Tier setting and the number of tiers in the storage pool, Easy Tier services might function in a different way. Table 4-7 lists possible combinations of Easy Tier settings.

Table 4-7 Easy Tier settings

Storage pool Easy Tier setting	Number of tiers in the storage pool	Volume copy Easy Tier setting ^a	Volume copy Easy Tier status
Off	One or more	Off	Inactive ^b
		On	Inactive ^b
Measure	One or More	Off	Measured ^c
		On	Measured ^c

Storage pool Easy Tier setting	Number of tiers in the storage pool	Volume copy Easy Tier setting ^a	Volume copy Easy Tier status
Auto	One	Off	Measured ^c
		On	Balanced ^d
	Two - four	Off	Measured ^c
		On	Active ^{e f}
	Five	Any	Measured ^c
On	One	Off	Measured ^c
		On	Balanced ^d
	Two - four	Off	Measured ^c
		On	Active ^e
	Five	Off	Measured ^c
		On	Active ^f

- a. If the volume copy is in image or sequential mode or being migrated, the volume copy Easy Tier status is Measured rather than Active.
- b. When the volume copy status is Inactive, no Easy Tier functions are enabled for that volume copy.
- c. When the volume copy status is Measured, the Easy Tier function collects usage statistics for the volume, but automatic data placement is not active.
- d. When the volume copy status is Balanced, the Easy Tier function enables performance-based pool balancing for that volume copy.
- e. When the volume copy status is Active, the Easy Tier function operates in automatic data placement mode for that volume.
- f. When five-tier (or some four-tier) configurations are used and Easy Tier is in the On state, Easy Tier is forced to operate but might not behave exactly as expected. For more information, see Table 4-9 on page 308.

Note: The default Easy Tier setting for a storage pool is Auto, and the default Easy Tier setting for a volume copy is On. Therefore, Easy Tier functions, except for pool performance balancing, are disabled for storage pools with a single tier. Automatic data placement mode is enabled by default for all striped volume copies in a storage pool with two or more tiers.

4.7.4 MDisk tier types

The three Easy Tier tier types (“hot”, “medium”, and “cold”) are generic “buckets” that Easy Tier uses to build a set of extents that belong to each tier. You must tell Easy Tier which MDisks belong to which bucket.

The type of disk and RAID geometry that is used by internal or external MDisks define their expected performance characteristics. These characteristics are used to help define a *tier type* for each MDisk in the system.

Five tier types can be assigned. The tables in this section use the numbers from this list as a shorthand for the tier name:

tier_scm	Represents SCM MDisks.
tier0_flash	Represents enterprise flash technology, including FCM.
tier1_flash	Represents lower performing Tier1 flash technology (lower drive writes per day (DWPD)).
tier_enterprise	Represents enterprise HDD technology (both 10 K and 15 K RPM).
tier_nearline	Represents nearline HDD technology (7.2 K RPM).

Consider the following points:

- ▶ Easy Tier is designed to operate with up to three tiers of storage: “hot”, “medium”, and “cold”.
- ▶ An MDisk can belong only to one tier type.
- ▶ At the time of writing, five MDisk tier types exist.
- ▶ Internal MDisks have their tier type set automatically.
- ▶ External MDisks default to the “enterprise” tier and might need to be changed by the user.
- ▶ The number of MDisk tier types that is found in a pool determines whether the pool is a single-tier pool or a multitier pool.

Attention: As described in 4.7.5, “Changing the tier type of an MDisk” on page 309, system does not automatically detect the type of external MDisks. Instead, all external MDisks are initially put into the enterprise tier by default. The administrator must then manually change the MDisks tier and add them to storage pools.

Single-tier storage pools

Figure 4-20 on page 306 shows a scenario in which a single storage pool is populated with MDisks that are presented by an external storage controller. In this solution, the striped volumes can be measured by Easy Tier and benefit from *storage pool balancing* mode, which moves extents between MDisks of the same type.

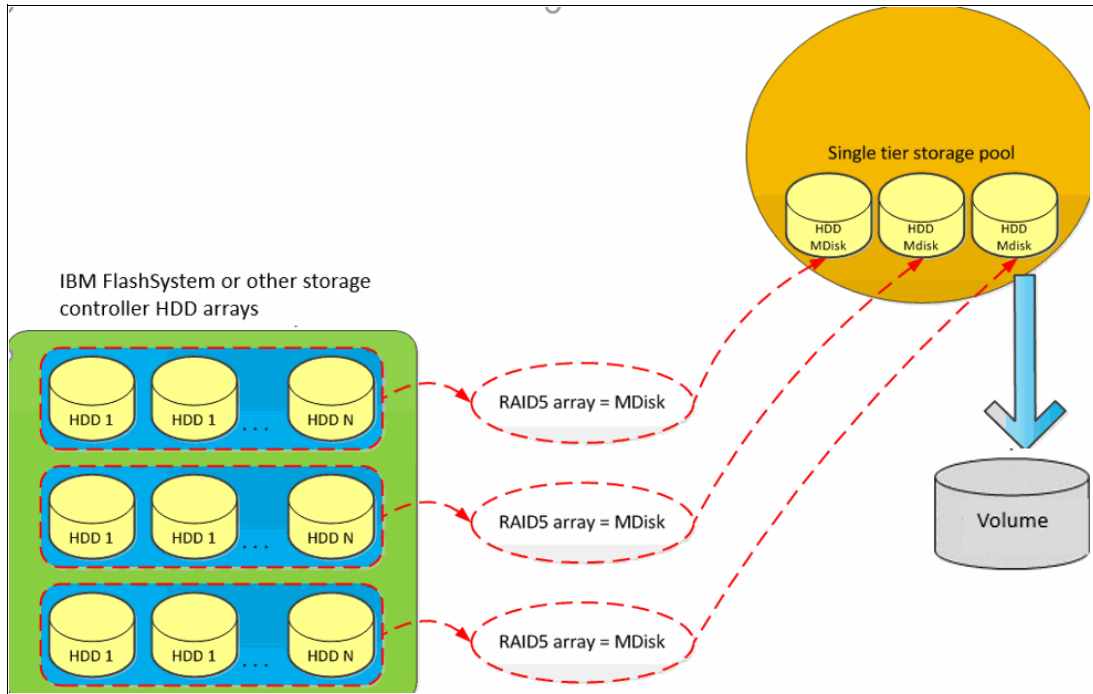


Figure 4-20 Single tier storage pool with a striped volume

MDisks that are used in a single-tier storage pool should have the same hardware characteristics. These characteristics include the same RAID type, RAID array size, disk type, disk RPM, and controller performance characteristics.

For external MDisks, attempt to create all MDisks with the same RAID geometry (number of disks). If this approach is not possible, you can modify the Easy Tier load setting to manually balance the workload, but you must be careful. For more information, see “MDisk Easy Tier load” on page 318.

For internal MDisks, the system can cope with different geometries because the number of drives is reported to Easy Tier, which then uses the Overload Protection information to balance the workload, as described in 4.7.6, “Easy Tier overload protection” on page 309.

Multitier storage pools

A multitier storage pool has a mix of MDisks with more than one type of MDisk tier attribute. This pool can be, for example, a storage pool that contains a mix of enterprise and SSD MDisks or enterprise and NL-SAS MDisks.

Figure 4-21 on page 307 shows a scenario in which a storage pool is populated with three different MDisk types:

- ▶ One belonging to an SSD array
- ▶ One belonging to an SAS HDD array
- ▶ One belonging to an NL-SAS HDD array)

Although Figure 4-21 on page 307 shows RAID 5 arrays other RAID types can be used.

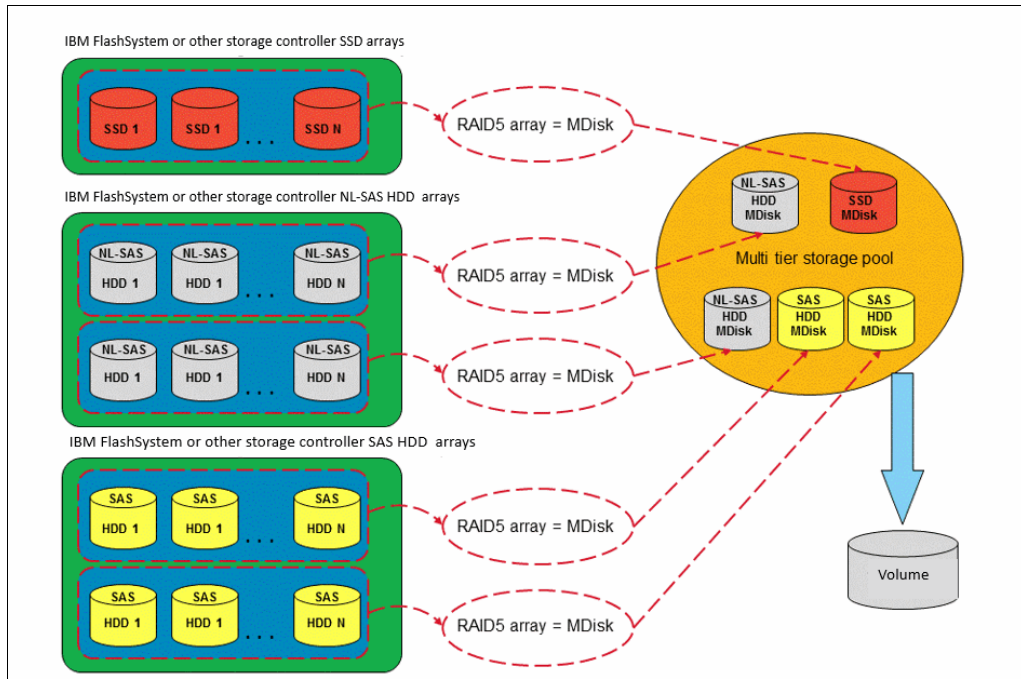


Figure 4-21 Multitier storage pool with a striped volume

Note: If you add MDisk to a pool and they have (or you assign) more than three tier types, Easy Tier tries to group two or more of the tier types into a single “bucket” and use them both as either the “middle” or “cold” tier. The groupings are described in Table 4-9 on page 308.

However, overload protection and pool balancing might result in a bias on the load being placed on those MDisk despite them being in the same “bucket”.

Easy Tier mapping to MDisk tier types

The five MDisk tier types are mapped to the three Easy Tier tiers depending on the pool configuration, as shown in Table 4-8.

Table 4-8 Recommended 3-tier Easy Tier mapping policy

Tier mix	1+2 1+3 1+4 1+5	2+3 2+4 2+5	3+4 3+5	4+5	1+2+3 1+2+4 1+2+5	1+3+4 1+3+5	1+4+5 2+4+5 3+4+5	2+3+4 2+3+5
Hot tier	1	2			1	1	1 or 2 or 3	2
Middle tier	2 or 3 or 4 or	3 or 4 or	3	4	2	3	4	3
Cold tier	5	5	4 or 5	5	3 or 4 or 5	4 or 5	5	4 or 5

For more information about the tier descriptions, see 4.7.4, “MDisk tier types” on page 304.

Four- and five-tier pools

In general, Easy Tier tries to place tier_enterprise (4) and tier1_flash (3) based tiers into one bucket to reduce the number of tiers that are defined in a pool to 3, as shown in Table 4-9.

Table 4-9 Four- and five-tier mapping policy

Tier mix	1+2+3+4, 1+2+3+5, or 1+2+4+5	1+3+4+5, 2+3+4+5	1+2+3+4+5
Hot tier	Not supported. Measure only.	1 or 2	Not supported. Measure only.
Middle tier		3 and 4	
Cold tier		5	

If you create a pool with all five tiers or one of the unsupported four-tier pools and Easy Tier is set to “auto” mode, Easy Tier enters “measure” mode and measures the statistics but does not move any extents. To return to a supported tier configuration, remove one or more MDisk.

Important: Avoid creating a pool with more than three tiers. Although the system attempts to create “buckets”, the result might be that Easy Tier runs in Measure mode only.

Temporary unsupported four- or five-tier mapping

If you must temporarily define four- or five-tier mapping in a pool and you end up with one of the unsupported configurations, you can force Easy Tier to migrate data by setting the Easy Tier mode to “on”.

Attention: Before you force Easy Tier to run in this mode, you must have a *full understanding* of the implications of doing so. Be *very cautious* about using this mode.

The on setting is provided to allow temporary migrations where you cannot avoid creating one of these unsupported configurations. The implications are that long-term use in this mode can cause performance issues due to the grouping of unlike MDisk within a single Easy Tier tier.

For these configurations, Easy Tier uses the mappings that are shown in Table 4-10.

Table 4-10 Unsupported temporary four- and five-tier mapping policy

Tier mix	1+2+3+4 or 1+2+3+5 ^a	1+2+4+5 ^b	1+2+3+4+5 ^{a b}
Hot tier	1	1	1
Middle tier	2 & 3	2	2 & 3
Cold tier	4 or 5	4 & 5	4 & 5

- a. In these configurations, enterprise HDDs and nearline HDDs are placed into the cold tier. These two drive types have different latency characteristics, and the difference can skew the metrics that are measured by Easy Tier for the cold tier.
- b. In these configurations, Tier 0 and Tier 1 flash devices are placed in the middle tier. The different DWPD does not make the most efficient use of the Tier 0 flash.

4.7.5 Changing the tier type of an MDisk

IBM Storage Virtualize cannot determine the technology type of the external MDisk, and by default assigns it with the enterprise tier type.

Attention: When adding external MDisks to a pool, validate that the `tier_type` setting is correct. Incorrect `tier_type` settings can cause performance problems, for example, if you inadvertently create a multitier pool.

Internal MDisks are automatically created with the correct `tier_type` because the system is aware of the drives that are used to create the RAID array and so can set the correct `tier_type` automatically.

The `tier_type` can be set when adding an MDisk to a pool, or you can later change the tier of an MDisk by using the CLI, as shown in Example 4-10.

Note: Changing the tier type is possible only for external MDisks, not for arrays.

Example 4-10 Changing an MDisk tier

```
IBM_2145:SVC_ESC:superuser>lsmdisk -delim " "
id name status mode mdisk_grp_id mdisk_grp_name capacity ctrl_LUN_#
controller_name UID tier encrypt site_id site_name distributed dedupe
1 mdisk1 online managed 1 POOL_V7K_SITEB 250.0GB 0000000000000001 V7K_SITEB_C2
6005076802880102c0000000000000200000000000000000000000000 tier_enterprise
no 2 SITE_B no no
2 mdisk2 online managed 1 POOL_V7K_SITEB 250.0GB 0000000000000002 V7K_SITEB_C2
6005076802880102c0000000000000210000000000000000000000000 tier_enterprise
no 2 SITE_B no no

IBM_2145:SVC_ESC:superuser>chmdisk -tier tier_nearline 1

IBM_2145:SVC_ESC:superuser>lsmdisk -delim " "
id name status mode mdisk_grp_id mdisk_grp_name capacity ctrl_LUN_#
controller_name UID tier encrypt site_id site_name distributed dedupe
1 mdisk1 online managed 1 POOL_V7K_SITEB 250.0GB 0000000000000001 V7K_SITEB_C2
6005076802880102c0000000000000200000000000000000000000000 tier_nearline no
2 SITE_B no no
2 mdisk2 online managed 1 POOL_V7K_SITEB 250.0GB 0000000000000002 V7K_SITEB_C2
6005076802880102c0000000000000210000000000000000000000000 tier_enterprise
no 2 SITE_B no no
```

It is also possible to change the external MDisk tier by using the GUI.

This change happens online and has no effect on hosts or the availability of the volumes.

4.7.6 Easy Tier overload protection

Easy Tier is defined as a “greedy” algorithm. If overload protection is not used, Easy Tier attempts to use every extent on the hot tier. In some cases, this issue leads to overloading the hot tier MDisks and creates a performance problem.

Therefore, Easy Tier implements overload protection to ensure that it does not move too much workload onto the hot tier. If this protection is triggered, no other extents are moved onto that tier while the overload is detected. Extents can still be swapped, so if one extent becomes colder and another hotter, they can be swapped.

To implement overload protection, Easy Tier must understand the capabilities of an MDisk. For internal MDisks, this understanding is handled automatically because the system can instruct Easy Tier about the type of drive and RAID geometry (for example, 8+P+Q); therefore, the system can calculate the expected performance ceiling for any internal MDisk.

With external MDisks, the only measure or details Easy Tier knows are the storage controller type. We know whether the controller is an enterprise, midrange, or entry level system and can make some assumptions about the load that it can handle.

However, external MDisks cannot automatically have their MDisk tier type or “Easy Tier Load” defined. Set the tier type manually and (if needed), modify the load setting. For more information about Easy Tier loads, see “MDisk Easy Tier load” on page 318.

Overload protection is also used by the “warm promote” function. If Easy Tier detects a sudden change on a cold tier in which a workload is causing overloading of the cold tier MDisks, it can quickly react and recommend migration of the extents to the middle tier. This feature is useful when provisioning new volumes that overrun the capacity of the middle tier or when no middle tier is present, for example, with flash and nearline only configurations.

4.7.7 Easy Tier overallocation protection

On systems with compressing backend, Easy Tier operations can potentially cause an “Out of physical capacity” situation. To prevent this, Easy Tier overallocation limit feature is implemented.

Arrays that are created from self-compressing drives have a written capacity limit (virtual capacity before compression) that is higher than the array’s usable capacity (physical capacity). Writing highly compressible data to the array means that the written capacity limit can be reached without running out of usable capacity. However, if data is not compressible or the compression ratio is low, it is possible to run out of usable capacity before reaching the written capacity limit of the array, which means the amount of data that is written to a self-compressing array must be controlled to prevent the array from running out of space.

Without a maximum overallocation limit, Easy Tier scales the usable capacity of the array based on the actual compression ratio of the data that is stored on the array at a point in time. Easy Tier migrates data to the array and might use a large percentage of the usable capacity in doing so, but it stops migrating to the array when the array comes close to running out of usable capacity. Then, it might start migrating data away from the array again to free space.

However, Easy Tier migrates storage only at a slow rate, which might not keep up with changes to the compression ratio within the tier. When Easy Tier swaps extents or data is overwritten by hosts, compressible data might be replaced with data that is less compressible, which increases the amount of usable capacity that is consumed by extents and might result in self-compressing arrays running out of space, which can cause a loss of access to data until the condition is resolved.

You can specify the maximum overallocation ratio for pools that contain FCM arrays to prevent out-of-space scenarios. The value acts as a multiplier of the physically available space in self-compressing arrays. The allowed values are a percentage in the range of 100% (default) to 400% or off. Setting the value to off disables this feature.

When the limit is set, Easy Tier scales the available usable capacity of self-compressing arrays by using the specified overallocation limit and adjusts the migration plan to make sure the fullness of these arrays stays below the maximum overallocation. Specify the maximum overallocation limit based on the estimated lowest compression ratio of the data that is written to the pool.

The default setting of 100% allows no Easy Tier overallocation on new pools, which is the most conservative and safe option. By increasing the limit, you allow Easy Tier to put more data into the FCM tier and benefit from its performance, but at the same time chances to get into “out of physical space” (OOPS) situation if compression ratio suddenly drops is much higher.

4.7.8 Removing an MDisk from an Easy Tier pool

When you remove an MDisk from a pool that still includes defined volumes and that pool is an Easy Tier pool, the extents that are still in use on the MDisk that you are removing are migrated to other free extents in the pool.

Easy Tier attempts to migrate the extents to another extent within the same tier. However, if there is not enough space in the same tier, Easy Tier picks the highest-priority tier with free capacity. Table 4-11 describes the migration target-tier priorities.

Table 4-11 Migration target tier priorities

Tier of MDisk being removed	Target tier priority (pick highest with free capacity)				
	1	2	3	4	5
tier_scm	tier_scm	tier0_flash	tier1_flash	tier_enterprise	tier_nearline
tier0_flash	tier0_flash	tier_scm	tier1_flash	tier_enterprise	tier_nearline
tier1_flash	tier1_flash	tier0_flash	tier_scm	tier_enterprise	tier_nearline
tier_enterprise	tier_enterprise	tier1_flash	tier_nearline	tier0_flash	tier_scm
tier_nearline	tier_nearline	tier_enterprise	tier1_flash	tier0_flash	tier_scm

The tiers are chosen to optimize the typical migration cases, for example, replacing the enterprise HDD tier with Tier 1 flash arrays or replacing nearline HDDs with Tier 1 flash arrays.

4.7.9 Easy Tier implementation considerations

Easy Tier is part of the IBM Storage Virtualize code. For Easy Tier to migrate extents between different tier disks, storage that offers different tiers must be available (for example, a mix of flash and HDDs). With single tier (homogeneous) pools, Easy Tier uses storage pool balancing only.

Important: Easy Tier uses the extent migration capabilities of IBM Storage Virtualize. These migrations require free capacity because an extent is first cloned to a new extent before the old extent is returned to the free capacity in the relevant tier.

It is recommended that a minimum of 16 extents are needed for Easy Tier to operate. However, if only 16 extents are available, Easy Tier can move at most 16 extents at a time.

Easy Tier and storage pool balancing do not function if you allocate 100% of the storage pool to volumes.

Implementation rules

Remember the following implementation and operational rules when you use Easy Tier:

- ▶ Easy Tier automatic data placement is not supported on image mode or sequential volumes. I/O monitoring for such volumes is supported, but you cannot migrate extents on these volumes unless you convert image or sequential volume copies to striped volumes.
- ▶ Automatic data placement and extent I/O activity monitors are supported on each copy of a mirrored volume. Easy Tier works with each copy independently of the other copy.

Volume mirroring considerations: Volume mirroring can have different workload characteristics on each copy of the data because reads are normally directed to the primary copy and writes occur to both copies. Therefore, the number of extents that Easy Tier migrates between the tiers might be different for each copy.

- ▶ If possible, the system creates volumes or expands volumes by using extents from MDisks from the enterprise tier. However, if necessary, it uses extents from MDisks from the flash tier.
- ▶ Do not provision 100% of Easy Tier enabled pool capacity. Reserve at least 16 extents for each tier for the Easy Tier movement operations.

When a volume is migrated out of a storage pool that is managed with Easy Tier, Easy Tier automatic data placement mode is no longer active on that volume. Automatic data placement is turned off while a volume is being migrated, even when it is between pools that both have Easy Tier automatic data placement enabled. Automatic data placement for the volume is reenabled when the migration is complete.

Limitations

When you use Easy Tier, consider the following limitations:

- ▶ Removing an MDisk by using the **-force** parameter.
When an MDisk is deleted from a storage pool with the **-force** parameter, extents in use are migrated to MDisks in the same tier as the MDisk that is being removed, if possible. If insufficient extents exist in that tier, extents from another tier are used.
- ▶ Migrating extents.
When Easy Tier automatic data placement is enabled for a volume, you cannot use the **migrateexts** CLI command on that volume.
- ▶ Migrating a volume to another storage pool.
When a system migrates a volume to a new storage pool, Easy Tier automatic data placement between the two tiers is temporarily suspended. After the volume is migrated to its new storage pool, Easy Tier automatic data placement resumes for the moved volume, if appropriate.

When the system migrates a volume from one storage pool to another one, Easy Tier attempts to migrate each extent to an extent in the new storage pool from the same tier as the original extent. In several cases, such as where a target tier is unavailable, another tier is used based on the priority rules that are outlined in 4.7.8, “Removing an MDisk from an Easy Tier pool” on page 311.

- ▶ Migrating a volume to an image mode copy.

Easy Tier automatic data placement does not support image mode. When a volume with active Easy Tier automatic data placement mode is migrated to an image mode volume, Easy Tier automatic data placement mode is no longer active on that volume.

- ▶ Image mode and sequential volumes cannot be candidates for automatic data placement. However, Easy Tier supports evaluation mode for image mode volumes.

Extent size considerations

The *extent size* determines the granularity level at which Easy Tier operates, which is the size of the chunk of data that Easy Tier moves across the tiers. By definition, a *hot extent* refers to an extent that has more I/O workload compared to other extents in the same pool and in the same tier.

It is unlikely that all the data that is contained in an extent has the same I/O workload, and as a result, the same temperature. Therefore, moving a hot extent likely also moves data that is not hot. The overall Easy Tier efficiency to put hot data in the correct tier is then inversely proportional to the extent size.

Consider the following points:

- ▶ Easy Tier efficiency affects the storage solution cost-benefit ratio. It is more effective for Easy Tier to place hot data in the top tier. In this case, less capacity can be provided for the relatively more expensive Easy Tier top tier.
- ▶ The extent size determines the bandwidth requirements for the Easy Tier background process. The smaller the extent size, the lower that the bandwidth consumption.

However, Easy Tier efficiency is not the only factor that is considered when choosing the extent size. Manageability and capacity requirement considerations also must be accounted for.

Generally, use the default 1 GB (standard pool) or 4 GB (DRP) extent size for Easy Tier enabled configurations.

External controller tiering considerations

IBM Easy Tier is an algorithm that was developed by IBM Almaden Research and made available to many members of the IBM storage family, such as the DS8000, SVC, and IBM FlashSystem products. The DS8000 is the most advanced in Easy Tier implementation and provides features that are not yet available for IBM FlashSystem technology, such as Easy Tier Application, Easy Tier Heat Map Transfer, and Easy Tier Control.

In general, using Easy Tier at the highest level, that is, the virtualizer, is recommended. So, it is a best practice to disable Easy Tier on back-end systems, but to leave it enabled on the virtualizer.

Important: Never run tiering at two levels. Doing so causes thrashing and unexpected heat and cold jumps at both levels.

Consider the following two options:

- ▶ Easy Tier is done at the virtualizer level.

In this case, complete the following steps at the back-end level:

- Set up homogeneous pools according to the tier technology that is available.
- Create volumes to present to the virtualizer from the homogeneous pool.
- Disable tiering functions.

At the virtualizer level, complete the following steps:

- Discover the MDisks that are provided by the back-end storage and set the tier properly.
- Create hybrid pools that aggregate the MDisks.
- Enable the Easy Tier function.

- ▶ Easy Tier is done at the backend level.

In this case, complete the following steps at the back-end level:

- Set up hybrid pools according to the tier technology that is available.
- Create volumes to present to the virtualizer from the hybrid pools.
- Enable the tiering functions.

At the virtualizer level, complete the following actions:

- Discover the MDisks that are provided by the back-end storage and set the same tier for all.
- Create standard pools that aggregate the MDisks.
- Disable the Easy Tier function.

Although both of these options provide benefits in term of performance, they have different characteristics.

Option 1 provides the following advantages compared to Option 2:

- ▶ With option 1, Easy Tier can be enabled or disabled at the volume level. This feature allows users to decide which volumes benefit from Easy Tier and which do not.

With option 2, this goal cannot be achieved.

- ▶ With option 1, the volume heat map matches directly to the host workload profile by using the volumes. This option also allows you to use Easy Tier across different storage controllers, which use lower performance and cost systems to implement the middle or cold tiers.

With option 2, the volume heat map on the back-end storage is based on the system workload. Therefore, it does not represent the hosts workload profile because of the effects of the system's caching.

- ▶ With option 1, the extent size can be changed to improve the overall Easy Tier efficiency.

Option 2, especially with DS8000 as the back-end, offers some advantages compared to option 1. For example, when external storage is used, the virtualizer uses generic performance profiles to evaluate the workload that can be placed on a specific MDisk, as described in "MDisk Easy Tier load" on page 318. These profiles might not match the back-end capabilities, which can lead to a resource usage that is not optimized.

However, this problem rarely occurs with option 2 because the performance profiles are based on the real back-end configuration.

Easy Tier and thin-provisioned back-end considerations

When a data reduction-capable back-end is used in Easy Tier enabled pools, the data-reduction ratio on the physical back-end might vary over time because of Easy Tier data-moving.

Easy Tier continuously moves extents across the tiers (and within the same tier) and attempts to optimize performance. As result, the amount of data that is written to the back-end (and therefore the compression ratio) can unpredictably fluctuate over time, even though the data is not modified by the user.

Note: It is not recommended to intermix data reduction-capable and non-data reduction-capable storage in the same tier of a pool with Easy Tier enabled.

Easy Tier and remote copy considerations

When Easy Tier is enabled, the workloads that are monitored on the primary and secondary systems can differ. Easy Tier at the primary system sees a normal workload; it sees only the write workloads at the secondary system.

This situation means that the optimized extent distribution on the primary system can differ considerably from the one that is on the secondary system. The optimized extent reallocation that is based on the workload learning on the primary system is not sent to the secondary system now to allow the same extent optimization on both systems based on the primary workload pattern.

In a DR situation with a failover from the primary site to a secondary site, the extent distribution of the volumes on the secondary system is not optimized to match the primary workload. Easy Tier relearns the production I/O profile and builds a new extent migration plan on the secondary system to adapt to the new production workload.

The secondary site eventually achieves the same optimization and level of performance as on the primary system. This task takes a little time, so the production workload on the secondary system might not run at its optimum performance during that period. The Easy Tier acceleration feature can be used to mitigate this situation. For more information, see “Easy Tier acceleration” on page 317.

IBM Storage Virtualize remote copy configurations that use the nearline tier at the secondary system must be carefully planned, especially when practicing DR by using FlashCopy. In these scenarios, FlashCopy often starts just before the beginning of the DR test. It is likely that the FlashCopy target volumes are in the nearline tier because of prolonged inactivity.

When the FlashCopy starts, an intensive workload often is added to the FlashCopy target volumes because of both the background and foreground I/Os. This situation can easily lead to overloading, and then possibly performance degradation of the nearline storage tier if it is not correctly sized in terms of resources.

Easy Tier on DRP and interaction with garbage collection

DRPs use LSA structures that need garbage-collection activity to be done regularly. An LSA always appends new writes to the end of the allocated space. For more information, see “DRP internal details” on page 249.

Even if data exists and the write is an overwrite, the new data is not written in that place. Instead, the new write is appended at the end and the old data is marked as needing garbage collection. This process provides the following advantages:

- ▶ Writes to a DRP volume always are treated as sequential. Therefore, all the 8 KB chunks can be built into a larger 256 KB chunk and destage the writes from cache as full stripe writes or as large as a 256 KB sequential stream of smaller writes.
- ▶ Easy Tier with DRP gives the best performance both in terms of RAID on back-end systems and on flash, where it becomes easier for the flash device to perform its internal garbage collection on a larger boundary.

To improve the Easy Tier efficiency with this write workload profile, you can start to record metadata about how frequently certain areas of a volume are overwritten. The Easy Tier algorithm was modified so that you can then bin-sort the chunks into a heat map in terms of rewrite activity, and then group commonly rewritten data onto a single extent. This method ensures that Easy Tier operates correctly for read/write data when data reduction is used.

Before DRP, write operations to compressed volumes had a lower value to the Easy Tier algorithms because writes were always to a new extent, so the previous heat was lost. Now, we can maintain the heat over time and ensure that frequently rewritten data is grouped. This process also aids the garbage-collection process where it is likely that large contiguous areas are garbage that is collected together.

Tier sizing considerations

Tier sizing is a complex task that always requires an environment workload analysis to match the performance and costs expectations.

Consider the following sample multi-tier configurations that address some of most common requirements. The same benefits can be achieved by adding SCM to the configuration. In these examples, the top flash tier can be replaced with an SCM tier, or SCM can be added as the hot tier and the corresponding medium and cold tiers are shifted down to drop the coldest tier:

- ▶ 5% SCM and 95% FCM
This configuration provides FCM capacity efficiency and performance, with the performance boost because of DRP and other frequently accessed metadata on SCM tier.
- ▶ 20-50% flash and 80-50% nearline
This configuration provides a mix of storage for latency-sensitive and capacity-driven workloads, it provides reduced costs and performance comparable to a single-tier flash solution.
- ▶ 10 - 20% flash and 80 - 90% enterprise
This configuration provides flash-like performance with reduced costs.
- ▶ 5% Tier 0 flash, 15% Tier 1 flash, and 80% nearline
This configuration provides flash-like performance with reduced costs.
- ▶ 3 - 5% flash and 95 - 97% enterprise
This configuration provides improved performance compared to a single-tier solution. All data is ensured to have at least enterprise performance. It also removes the requirement for overprovisioning for high-access density environments.
- ▶ 3 - 5% flash, 25 - 50% enterprise, and 40 - 70% nearline
This configuration provides improved performance and density compared to a single-tier solution. It also provides significant reduction in environmental costs.

4.7.10 Easy Tier settings

The Easy Tier settings for storage pools and volumes can be changed only from the CLI. All the changes are done online without any effect on the host or data availability.

Turning on and off Easy Tier

Use the `chvdisk` command to turn on or off Easy Tier on selected volumes. Use the `chmdiskgrp` command to change the status of Easy Tier on selected storage pools, as shown in Example 4-11.

Example 4-11 Changing Easy Tier settings

```
IBM_IBM FlashSystem:ITS0:superuser>chvdisk -easytier on test_vol_2
IBM_IBM FlashSystem:ITS0:superuser>chmdiskgrp -easytier auto test_pool_1
```

Tuning Easy Tier

It is also possible to change more advanced parameters of Easy Tier. These parameters should be used with caution because changing the default values can affect system performance.

Easy Tier acceleration

The first setting is called *Easy Tier acceleration*. This setting is a system-wide one that is disabled by default. Turning on this setting makes Easy Tier move extents up to four times faster than when in the default setting. In accelerate mode, Easy Tier can move up to 48 GiB every 5 minutes, while in normal mode it moves up to 12 GiB. Enabling Easy Tier acceleration is advised only during periods of low system activity. The following use cases for acceleration are the most likely ones:

- ▶ When installing a new system, accelerating Easy Tier quickly produces a steady state and reduces the time that is needed to reach an optimal configuration. This approach applies to single-tier and multitier pools. In a single-tier pool, this approach allows balancing to spread the workload quickly, and in a multitier pool it allows both inter-tier movement and balancing within each tier.
- ▶ When adding capacity to the pool, accelerating Easy Tier can quickly spread existing volumes onto the new MDisk by using pool balancing. This approach can help if you added more capacity to stop warm demote operations. In this case, Easy Tier knows that certain extents are hot and were demoted only due to lack of space or because Overload Protection was triggered.
- ▶ When migrating the volumes between the storage pools in cases where the target storage pool has more tiers than the source storage pool, accelerating Easy Tier can quickly promote or demote extents in the target pool.

This setting can be changed online without affecting host or data availability. To turn on or off Easy Tier acceleration mode, run the following command:

```
chsystem -easytieracceleration <on/off>
```

Important: Do not leave accelerated mode on indefinitely. It is a best practice to run in accelerated mode only for a few days to weeks to enable Easy Tier to reach a steady state quickly. After the system is performing fewer migration operations, disable accelerated mode to ensure that Easy Tier does not affect system performance.

MDisk Easy Tier load

The second setting is called *MDisk Easy Tier load*. This setting is set on an individual MDisk basis, and it indicates how much load Easy Tier can put on that MDisk. This setting was introduced to handle situations where Easy Tier is either underutilizing or overutilizing an external MDisk.

This setting cannot be changed for internal MDisks (an array) because the system can determine the exact load that an internal MDisk can handle based on the type of drive (HDD or SSD), the number of drives, and type of RAID in use per MDisk.

For an external MDisk, Easy Tier uses specific performance profiles based on the characteristics of the external controller and on the tier that is assigned to the MDisk. These performance profiles are generic, which means that they do not account for the actual back-end configuration. For example, the same performance profile is used for a DS8000 with 300 GB 15 K RPM and 1.8 TB 10 K RPM.

This feature is provided for advanced users to change the Easy Tier load setting to better align it with a specific external controller configuration.

Note: The load setting is used with the MDisk tier type setting to calculate the number of concurrent I/Os and expected latency from the MDisk. Setting this value incorrectly or by using the wrong MDisk tier type can have a detrimental effect on overall pool performance.

The following values can be set to each MDisk for the Easy Tier load:

- ▶ Default
- ▶ Low
- ▶ Medium
- ▶ High
- ▶ Very high

The system uses a default setting based on the controller performance profile and the MDisk tier setting of the presented MDisks.

Change the default setting to any other value only when you are certain that a MDisk is underutilized and can handle more load or that the MDisk is overutilized and the load should be lowered. Change this setting to “very high” only for SDDs and flash MDisks.

This setting can be changed online without affecting the host or data availability.

To change this setting, run the following command:

```
chmdisk -easytierload high mdisk0
```

Note: After changing the load setting, note the old and new settings and record the date and time of the change. Use IBM Storage Insights to review the performance of the pool in the coming days to ensure that you have not inadvertently degraded the performance of the pool.

You can also gradually increase the load setting and validate that with each change you are seeing an increase in throughput without a corresponding detrimental increase in latency (and vice versa if you are decreasing the load setting).



Volumes

In an IBM Storage Virtualize system, a *volume* is a logical disk that the system presents to attached hosts. This chapter describes the various types of volumes and provides guidance about managing their properties.

This chapter includes the following topics:

- ▶ “Volumes overview” on page 320
- ▶ “Guidance for creating volumes” on page 321
- ▶ “Thin-provisioned volumes” on page 323
- ▶ “Mirrored volumes” on page 333
- ▶ “HyperSwap volumes” on page 337
- ▶ “VMware vSphere Virtual Volumes” on page 339
- ▶ “Cloud volumes” on page 339
- ▶ “Volume migration” on page 342
- ▶ “Preferred paths to a volume” on page 351
- ▶ “Moving a volume between I/O groups and nodes” on page 352
- ▶ “Volume throttling” on page 353
- ▶ “Volume cache mode” on page 356
- ▶ “Other considerations” on page 359

5.1 Volumes overview

A volume can have one or two volume copies on the local storage system. A volume also can be replicated to a remote storage system. A *basic volume* has one local copy. A *mirrored volume* has two local copies. Each volume copy can be in different pools and have different capacity reduction attributes.

For best performance, spread the host workload over multiple volumes.

Volumes can be created with the following attributes:

- ▶ Standard-provisioned volumes

Volumes with no special attributes, which are also referred to as *fully allocated volumes*. A standard provisioned volumes completely uses the allocated capacity at time of creation.

- ▶ Thin-provisioned volumes

Volumes that present a larger capacity to the host than their real capacity. A thin volume presents the provisioned size to the host. For example, such a volume provisioned for 4 TiB stores only the amount data that is actually written. A host to which the volume is mapped reports the size as 4 TiB.

- ▶ Compressed volumes

A compressed volume can save capacity. The compression occurs when data is written to disk. Conversely, when reading the data is decompressed.

- ▶ Deduplicated volumes

Volumes whose data is deduplicated with other volumes in a data reduction pool (DRP). Deduplication is a type of data reduction that eliminates duplicate copies of data. Deduplication of user data occurs within a data reduction pool and only between volumes or volume copies that are marked as deduplicated. Some models or software versions require specific hardware or software to use this function.

- ▶ Mirrored volumes

A mirrored volume can have two exact copies. Each volume copy can belong to a different pool, and each copy has the same provisioned capacity as the volume. In the management GUI, an asterisk (*) indicates the primary copy of the mirrored volume. The primary copy indicates the preferred volume for read requests. A mirrored volume provides higher reliability because a host can access either copy of the volume.

- ▶ HyperSwap volumes

HyperSwap volumes create copies on separate sites for systems that are configured with HyperSwap topology. Data that is written to a HyperSwap volume is automatically sent to both copies so that either site can provide access to the volume if the other site becomes unavailable. HyperSwap volumes are supported on systems that contain more than one I/O group.

- ▶ VMware vSphere Virtual Volumes (VVOLs)

Volumes that are managed remotely by VMware vCenter. IBM Storage Virtualize provides native support for VMware vSphere APIs for Storage Awareness (VASA) through a VASA Provider (also known as Storage Provider) which sends and receives information about storage that is used by VMware vCenter Server.

- ▶ Cloud volumes

Volumes that are enabled for IBM Transparent Cloud Tiering (TCT).

Volumes in standard pools of an IBM SAN Volume Controller (SVC) or IBM FlashSystem can feature the following attributes that affect where the extents are allocated:

- ▶ **Striped**

A volume that is striped at the extent level. The extents are allocated from each managed disk (MDisk) that is in the storage pool. This volume type is the most frequently used because each I/O to the volume is spread across external storage MDisks.

- ▶ **Sequential**

A volume on which extents are allocated sequentially from one MDisk. This type of volume is rarely used because a striped volume is better suited to most cases.

- ▶ **Image**

Image-mode volumes are special volumes that have a direct relationship with one MDisk. If you have an MDisk that contains data that you want to merge into the clustered system, you can create an image-mode volume. When you create an image-mode volume, a direct mapping is made between extents that are on the MDisk and extents that are on the volume. The MDisk is not virtualized. The logical block address (LBA) *x* on the MDisk is the same as LBA *x* on the volume.

When you create an image-mode volume copy, you must assign it to a storage pool. An image-mode volume copy must be at least one extent in size. The minimum size of an image-mode volume copy is the extent size of the storage pool to which it is assigned.

The extents are managed in the same way as other volume copies. When the extents are created, you can move the data onto other MDisks that are in the storage pool without losing access to the data. After you move one or more extents, the volume copy becomes a virtualized disk, and the mode of the MDisk changes from image to managed.

5.2 Guidance for creating volumes

When creating volumes, consider the following guidelines:

- ▶ Consider the naming rules before you create volumes. Consistent volume-naming conventions help you avoid confusion and possible misuse of volumes. Proper naming at the time of creation also avoids the need to go back after the fact and rename volumes.
- ▶ Choose which type of volume that you want to create. First, decide whether fully allocated (standard volumes) or thin-provisioned volumes are going to be created. If you decide to create a thin-provisioned volume, analyze whether you need compression and deduplication enabled. Volume capacity reduction options in the IBM Storage Virtualize system are independent of any reduction done that is by the back-end controller.
- ▶ A fully allocated volume is automatically formatted, which can be a time-consuming process. However, this background process does not impede the immediate usage of the volume. During the format, extents are overwritten with zeros and Small Computer System Interface (SCSI) **unmap** commands are sent to the back-end storage, if required and supported.

Actions, such as moving, expanding, shrinking, or adding a volume copy, are unavailable when the specified volume is formatting.

You also can create volumes by using the command-line interface (CLI). Example 5-1 on page 322 shows the command to disable the auto-formatting option with the **-nofmtdisk** parameter.

Example 5-1 Volume creation without the auto-formatting option

```
superuser>mkvdisk -name VOL01 -mdiskgrp 0 -size 1 -unit gb -vtype striped
-iogrp io_grp0 -nofmtdisk
Virtual Disk, id [52], successfully created
superuser>lsvdisk VOL01
id 52
name VOL01
IO_group_id 0
IO_group_name io_grp0
status online
mdisk_grp_id 0
mdisk_grp_name Swimming
capacity 1.00GB
type striped
formatted no
formatting no
.
lines removed for brevity
```

When you create a volume, it takes some time to completely format it completely (depending on the volume size). The **syncrate** parameter of the volume specifies the volume copy synchronization rate, and it can be modified to accelerate the completion of the format process.

For example, the initialization of a 1 TB volume can take more than 120 hours to complete with the default syncrate value 50, or approximately 4 hours if you manually set the syncrate to 100. If you increase the syncrate to accelerate the volume initialization, remember to reduce it again to avoid issues the next time you use volume mirroring to perform a data migration of that volume.

For more information about creating a thin-provisioned volume, see 5.3, “Thin-provisioned volumes” on page 323.

- ▶ Each volume is associated with an I/O group and has a preferred node inside that I/O group. When creating a volume on an SVC, consider balancing volumes across the I/O groups to balance the load across the cluster. When creating a volume on a clustered IBM FlashSystem, ensure each MDisk group completely resides in one IBM FlashSystem.

If a host can access only one I/O group, the volume must be created in the I/O group to which the host has access.

Also, it is possible to define a list of I/O groups in which a volume can be accessible to hosts. It is a best practice that a volume is accessible to hosts by the caching only I/O group. You can have more than one I/O group in the access list of a volume in some scenarios with specific requirements, such as when a volume is migrated to another I/O group.

Tip: Migrating volumes across I/O groups can be a disruptive action. Therefore, specify the correct I/O group at the time the volume is created.

- ▶ By default, the *preferred node*, which owns a volume within an I/O group, is selected in a round-robin basis. Although it is not easy to estimate the workload when the volume is created, distribute the workload evenly on each node within an I/O group.
- ▶ Except in a few cases, the cache mode of a volume is set to read/write. For more information, see 5.12, “Volume cache mode” on page 356.

- ▶ A volume occupies an integer number of extents, but its length does not need to be an integer multiple of the extent size. However, the length does need to be an integer multiple of the block size. Any space that is left over between the last logical block in the volume and the end of the last extent in the volume is unused.
- ▶ The maximum number of volumes per I/O group and system is listed in the “Configurations Limits and Restrictions” section for your system’s code level at the following IBM Support web pages:
 - [IBM SAN Volume Controller](#)
 - [IBM FlashSystem 9500](#)
 - [IBM FlashSystem 9100 and 9200](#)
 - [IBM FlashSystem 7200 and 7300](#)
 - [IBM FlashSystem 50x5 and 5200](#)

5.3 Thin-provisioned volumes

A thin-provisioned volume presents a different capacity to mapped hosts than the capacity that the volume uses in the storage pool. The system supports thin-provisioned volumes in standard pools and DRPs.

Note: We do not recommend using thin-provisioned volumes in a DRP with IBM FlashCore Module (FCM).

Figure 5-1 shows the basic concept of a thin-provisioned volume.

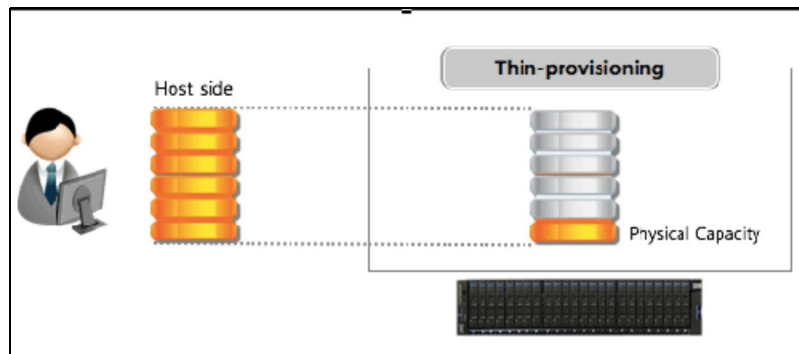


Figure 5-1 Thin-provisioned volume

The different types of volumes in a DRP are shown in Figure 5-2.

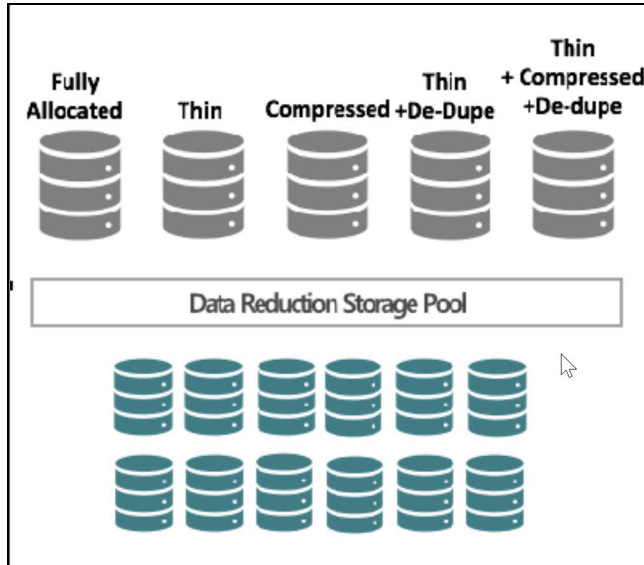


Figure 5-2 Different kinds of volumes in a DRP

In standard pools, thin-provisioned volumes are created based on capacity savings criteria. These properties are managed at the volume level. However, in DRPs, all the benefits of thin-provisioning are available to all the volumes that are assigned to the pool. For the thin-provisioned volumes in DRPs, you can configure compression and data deduplication on these volumes, which increases the capacity savings for the entire pool.

You can enhance capacity efficiency for thin-provisioned volumes by monitoring the hosts' usage of capacity. When the host indicates that the capacity is no longer needed, the space is released and can be reclaimed by the DRP. Standard pools do not have these functions.

Figure 5-3 shows the concepts of thin-provisioned volumes.

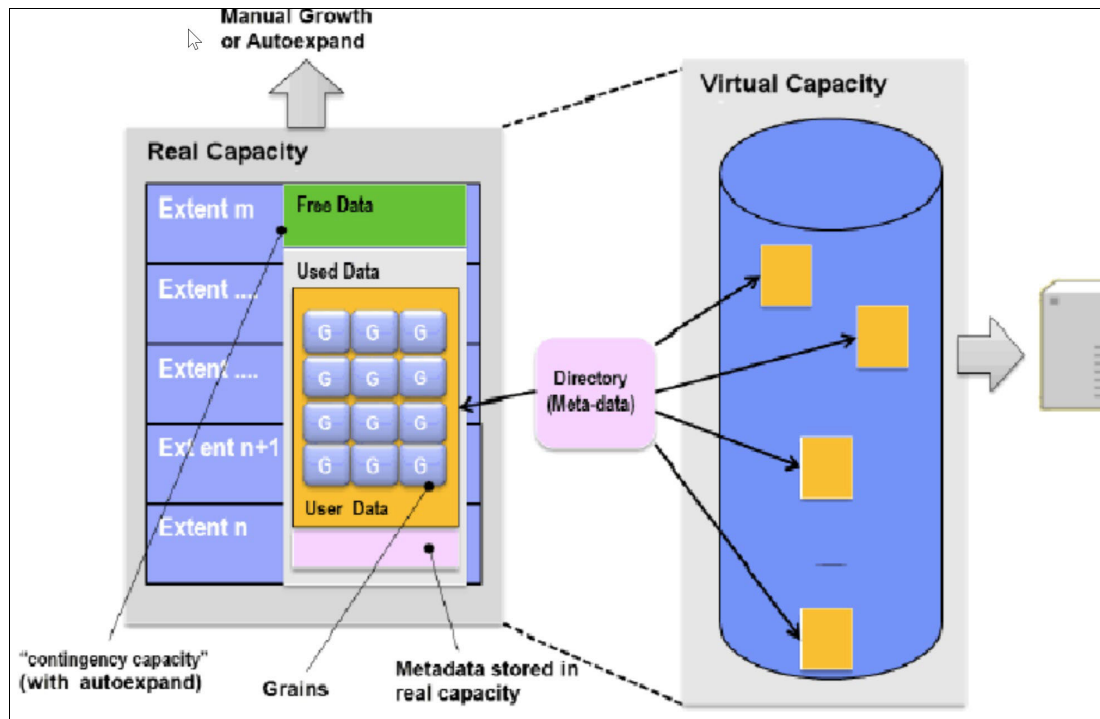


Figure 5-3 Thin-provisioned volume concepts

Real capacity defines how much disk space from a pool is allocated to a volume. *Virtual capacity* is the capacity of the volume that is reported to the hosts. A volume's virtual capacity is typically larger than its real capacity. However, as data continues to be written to the volume, that difference diminishes.

Each system uses the real capacity to store data that is written to the volume and metadata that describes the thin-provisioned configuration of the volume. As more information is written to the volume, more of the real capacity is used. The system identifies read operations to unwritten parts of the virtual capacity and returns zeros to the server without using any real capacity.

Thin-provisioned volumes in a standard pool are available in two operating modes: autoexpand and noautoexpand. You can switch the mode at any time. Thin-provisioned volumes in a DRP always have autoexpand enabled.

If you select the autoexpand feature, the IBM Storage Virtualize system automatically adds a fixed amount of real capacity to the thin volume as required. Therefore, the autoexpand feature attempts to maintain a fixed amount of unused real capacity for the volume. We recommend the usage of autoexpand by default to avoid volume-offline issues.

This amount of extra real capacity is known as the *contingency capacity*. The contingency capacity is initially set to the real capacity that is assigned when the volume is created. If the user modifies the real capacity, the contingency capacity is reset to be the difference between the used capacity and real capacity.

A volume that is created *without* the autoexpand feature (and therefore has a zero contingency capacity) goes offline when the real capacity is used. In this case, it must be expanded.

When creating a thin-provisioned volume with compression and deduplication enabled, you must be careful about out-of-space issues in the volume and pool where the volume is created. Set the warning threshold notification in the pools that contain thin-provisioned volumes, and in the volume.

Warning threshold: When you are working with thin-provisioned volumes, enable the warning threshold (by using email or a Simple Network Management Protocol (SNMP) trap) in the storage pool. If the autoexpand feature is not used, you also must enable the warning threshold on the volume level. If the pool or volume runs out of space, the volume goes offline, which results in a loss of access.

If you do not want to be concerned with monitoring volume capacity, it is highly recommended that the autoexpand option is enabled. Also, when you create a thin-provisioned volume, you must specify the space that is initially allocated to it (by using the **-rsize** option in the CLI) and the grain size.

By default, **-rsize** (or real capacity) is set to 2% of the volume virtual capacity, and grain size is 256 KiB. These default values, with the autoexpand enabled and warning disabled options, work in most scenarios. Some instances exist in which you might consider using different values to suit your environment.

Example 5-2 shows the command to create a volume with the suitable parameters.

Example 5-2 Creating a thin-provisioned volume

```
superuser>mkvdisk -name VOL02 -mdiskgrp Pool1 -size 100 -unit gb -vtype striped
-iogrp io_grp0 -rsize 2% -autoexpand -warning 0 -grainsize 256
Virtual Disk, id [53], successfully created
superuser>lsvdisk VOL02
id 53
name VOL02
.
lines removed for brevity
.
capacity 100.00GB
.
lines removed for brevity
.
used_capacity 0.75MB
real_capacity 2.02GB
free_capacity 2.01GB
overallocation 4961
autoexpand on
warning 0
grainsize 256
se_copy yes
.
lines removed for brevity
```

A thin-provisioned volume can be converted nondisruptively to a fully allocated volume or vice versa. Figure 5-4 shows how to modify the capacity savings of a volume. You can right-click the volume and select **Modify Capacity Savings**.

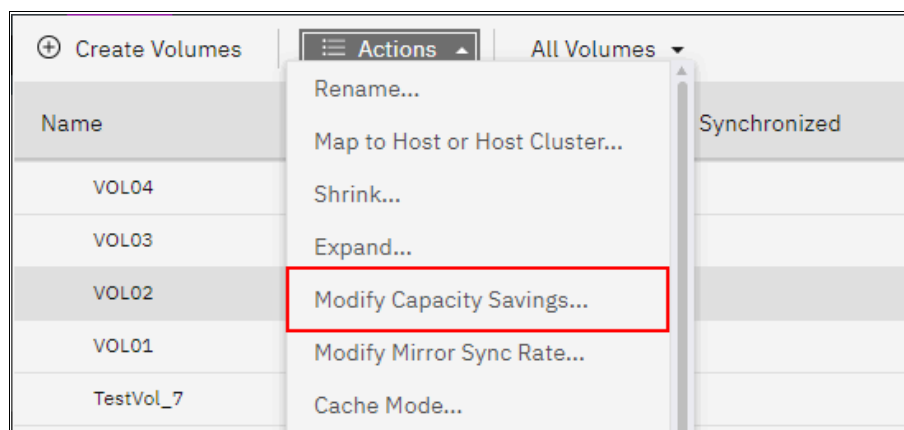


Figure 5-4 Modifying the capacity savings of a volume nondisruptively

The fully allocated to thin-provisioned migration procedure uses a zero-detection algorithm so that grains that contain all zeros do not cause any real capacity to be used.

5.3.1 Compressed volumes

When you create volumes, you can specify compression as a method to save capacity for the volume. With compressed volumes, data is compressed as it is written to disk, which saves more space. When data is read to hosts, the data is decompressed.

Note: The volume compression attribute is independent of any compression that is performed by FCMs or any other compressing back-end.

IBM Storage Virtualize systems support compressed volumes in a DRP only. A DRP also reclaims capacity that is not used by hosts if the host supports SCSI **unmap** commands. When these hosts issue SCSI **unmap** commands, a DRP reclaims the released capacity.

Compressed volumes in DRPs do not display their individual compression ratio (CR). The pool's used capacity *before* reduction indicates the total amount of data that is written to volume copies in the storage pool before data reduction occurs. The pool's used capacity *after* reduction is the space that is used after thin provisioning, compression, and deduplication. This compression solution provides nondisruptive conversion between compressed and decompressed volumes.

If you are planning to virtualize volumes that are connected to your hosts directly from any storage subsystems and you want an estimate of the space saving that likely is to be achieved, run the IBM Data Reduction Estimator Tool (DRET).

DRET is a CLI- and host-based utility that can be used to estimate an expected compression rate for block devices. This tool also can evaluate capacity savings by using deduplication. For more information, see [IBM Data Reduction Estimator Tool for SVC, Storwize and FlashSystems products](#).

IBM Storage Virtualize systems also include an integrated Comprestimator tool, which is available through the management GUI and CLI. If you are considering applying compression on noncompressed volumes in an IBM FlashSystem, you can use this tool to evaluate whether compression generates enough capacity savings.

For more information, see 4.2.3, "Data reduction estimation tools" on page 254.

As shown in Figure 5-5, customize the Volume view to see the compression savings for a compressed volume and estimated compression savings for a noncompressed volume that you are planning to migrate.

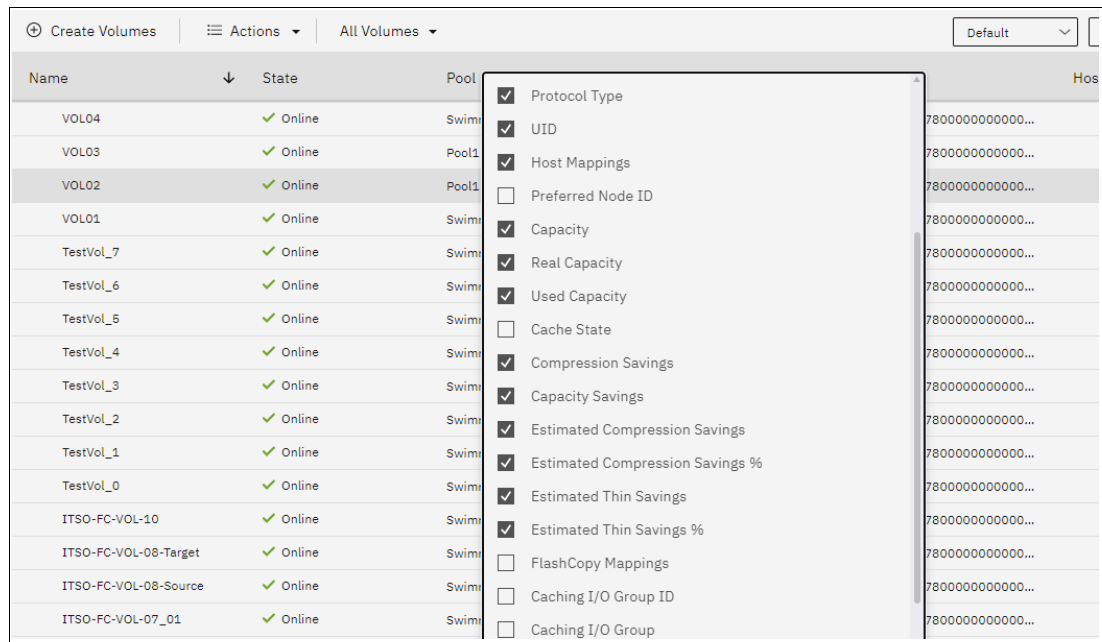


Figure 5-5 Customized view

5.3.2 Deduplicated volumes

Deduplication is a data reduction technique for eliminating duplicate copies of data. It can be configured with thin-provisioned and compressed volumes in a DRP.

The deduplication process identifies unique chunks of data (or byte patterns) and stores a signature of the chunk for reference when writing new data chunks. If the new chunk's signature matches an existing signature, the new chunk is replaced with a small reference that points to the stored chunk. The same byte pattern can occur many times, which result in the amount of data that must be stored being greatly reduced.

If a volume is configured with deduplication and compression, data is deduplicated first and then compressed. Therefore, deduplication references are created on the compressed data that is stored on the physical domain.

The scope of deduplication is all deduplicated volumes in the same pool, regardless of the volume's preferred node or I/O group.

To create a thin-provisioned volume that uses deduplication and does not require a provisioning policy, enter the command into the CLI that is shown in Example 5-3.

Example 5-3 Creating a thin-provisioned volume with the deduplication option

```
superuser>mkvolume -name dedup_test_01 -size 10 -unit gb -pool 0 -thin
-deduplicated
Volume, id [55], successfully created
```

To create a compressed volume that uses deduplication and does not required a provisioning policy, enter the command that is shown in Example 5-4 on page 330.

Example 5-4 Creating a compressed volume with the deduplication option

```
superuser>mkvolume -name dedup_test_02 -size 10 -unit gb -pool 0 -compressed  
-deduplicated  
Volume, id [56], successfully created
```

To maximize the space that is available for the deduplication database, the system distributes it between all nodes in the I/O groups that contain deduplicated volumes. Each node holds a distinct portion of the records that are stored in the database.

Depending on the data type that is stored on the volume, the capacity savings can be significant. Examples of use cases that typically benefit from deduplication are backup servers and virtual environments with multiple virtual machines (VMs) running the same operating system.

In both cases, it is expected that multiple copies of identical files exist, such as components of the standard operating system or applications that are used in the organization. Data that is encrypted or compressed at the file-system level does not benefit from deduplication. Deduplication works by finding patterns, and encryption essentially works by obfuscating whatever patterns might exist in the data.

If you want to evaluate whether savings are realized by migrating a set of volumes to deduplicated volumes, you can use DRET. For more information about DRET, see 4.2.3, “Data reduction estimation tools” on page 254.

5.3.3 Thin provisioning considerations

Most modern file systems and operating systems can benefit from thin provisioning. However, if the host performs a low-level format of the entire volume or the file system writes to a large percentage of the volume, then it is possible that no advantage is gained by using a thin-provisioning volume over a fully allocated volume.

Consider the following properties of thin-provisioned volumes that are useful to understand:

- ▶ When the used capacity first exceeds the volume *warning threshold*, an event is raised, which indicates that real capacity is required. The default warning threshold value is 80% of the volume capacity. To disable warnings, specify 0%.
- ▶ Compressed volumes include an attribute called *decompressed used capacity* (for standard pools) and *used capacity before reduction* (for a DRP). These volumes are the used capacities before compression or data reduction. They are used to calculate the CR.

Thin provisioning and overallocation

Because thin-provisioned volumes do not store the zero blocks, a storage pool is overallocated only after the sum of all volume capacities exceeds the size of the storage pool.

Storage administrators must be concerned about the out-of-space problem. If enough capacity exists on disk to store fully allocated volumes and you convert them to thin-provisioned volumes, enough space exists to store data (even if the hosts write to every byte of virtual capacity). Therefore, this issue is not going to be a problem for the short term, and you have time to monitor your system and understand how your capacity grows.

Monitoring capacity with thin-provisioned volumes

Note: It is critical that capacity is monitored when thin-provisioned or compressed volumes are used. Be sure to add capacity *before* running out of space.

If you run out of space on a volume or storage pool, the host that uses the affected volumes cannot perform new write operations to these volumes. Therefore, an application or database that is running on this host becomes unavailable.

In a storage pool with only fully allocated volumes, the storage administrator can easily manage the used and available capacity in the storage pool as the used capacity grows when volumes are created or expanded.

However, in a pool with thin-provisioned volumes, the used capacity increases any time that the host writes data. For this reason, the storage administrator must consider capacity planning carefully. It is critical to put in place volume and pool capacity monitoring.

Tools, such as IBM Spectrum Control and IBM Storage Insights, can display the capacity of a storage pool in real time and graph how it is growing over time. These tools are important because they are used to predict when the pool runs out of space.

IBM Storage Virtualize also alerts you by including an event in the event log when the storage pool reaches the configured threshold, which is called the *warning level*. The GUI sets this threshold to 80% of the capacity of the storage pool by default.

By using enhanced Call Home and IBM Storage Insights, IBM now can monitor and flag systems that have low capacity. This ability can result in a support ticket being generated and the client being contacted.

What to do if you run out of space in a storage pool

You can use one or a combination of the following options that are available if a storage pool runs out of space:

- ▶ Contingency capacity on thin-provisioned volumes.

If the storage pool runs out of space, each volume has its own contingency capacity, which is an amount of storage that is reserved by the volume. This capacity is sizable. Contingency capacity is defined by the *real capacity* parameter that is specified when the volume is created, which has a default value of 2%.

The contingency capacity protects the volume from going offline when its storage pool runs out of space by having the storage pool use this reserved space first. Therefore, you have some time to repair things before everything starts going offline.

If you want more safety, you might implement a policy of creating volumes with 10% of *real capacity*. Also, you do not need to have the same contingency capacity for every volume.

Note: This protection likely solves most immediate problems. However, after you are informed that you ran out of space, a limited amount of time exists to react. You need a plan in place and the next steps must be understood.

- ▶ Have unallocated storage on standby.

You can always have spare drives or MDisks ready to be added within only a few minutes to whichever storage pool runs out of space. This capacity provides some breathing room while you take other actions. The more drives or MDisks you have, the more times you must perform the same set of steps to solve the problem.

- ▶ **Sacrificial emergency space volume.**

Consider using a fully allocated sacrificial emergency space volume in each pool. If the storage pool is running out of space, you can delete or shrink this volume to quickly provide more available space in the pool.

- ▶ **Move volumes.**

You can migrate volumes to other pools to free space. However, data migration on IBM FlashSystem is designed to move slowly to avoid performance problems. Therefore, it might be impossible to complete this migration before your applications go offline.

- ▶ **Policy-based solutions.**

No policy is going to solve a problem if you run out of space, but you can use policies to reduce the likelihood of that ever happening to the point where you feel comfortable using fewer of the other options.

You can use these types of policies for thin-provisioning.

Note: The following policies use arbitrary numbers. These numbers are designed to make the suggested policies more readable. We do not provide any recommended numbers to insert into these policies because they are determined by business risk, and this consideration is different for every client.

- Manage free space such that enough free capacity always is available for your 10 largest volumes to reach 100% full without running out of free space.
- Never overallocate more than 200%. For example, if you have 100 TB of capacity in the storage pool, the sum of the volume capacities in the same pool must not exceed 200 TB.
- Always start the process of adding capacity when the storage pool reaches 70% full.

Grain size

The *grain size* is defined when the thin-provisioned volume is created. The grain size can be set to 32 KB, 64 KB, 128 KB, or 256 KB (default). The grain size cannot be changed after the thin-provisioned volume is created.

Smaller grain sizes can save more space, but they have larger directories. For example, if you select 32 KB for the grain size, the volume size cannot exceed 260,000 GB. Therefore, if you are not going to use the thin-provisioned volume as a FlashCopy source or target volume, use 256 KB by default to maximize performance.

Thin-provisioned volume copies in DRPs have a grain size of 8 KB. This predefined value cannot be set or changed.

If you are planning to use thin-provisioning with FlashCopy, the grain size for FlashCopy volumes can be only 64 KB or 256 KB. In addition, to achieve best performance, the grain size for the thin-provisioned volume and FlashCopy mapping must be same. For this reason, it is not recommended to use thin-provisioned volume in DRPs as a FlashCopy source or target volume.

Note: Using thin-provisioned volumes in a DRP for FlashCopy is not recommended.

5.4 Mirrored volumes

By using volume mirroring, a volume can have two copies. Each copy of the volume can belong to a different pool and have different capacity reduction attributes. Both copies contain the same virtual capacity. In the management GUI, an asterisk (*) indicates the primary copy of the mirrored volume. The primary copy indicates the preferred volume for read requests.

When a server writes to a mirrored volume, the system writes the data to both copies. When a server reads a mirrored volume, the system picks one of the copies to read. If one of the mirrored volume copies is temporarily unavailable (for example, because the storage system that provides the pool is unavailable), the volume remains accessible to servers. The system remembers which areas of the volume are written and resynchronizes these areas when both copies are available.

You can create a volume with one or two copies, and you can convert a non-mirrored volume into a mirrored volume by adding a copy. When a copy is added in this way, the system synchronizes the new copy so that it is the same as the existing volume. Servers can access the volume during this synchronization process.

You can convert a mirrored volume into a nonmirrored volume by deleting one copy or by splitting one copy to create a non-mirrored volume.

The volume copy can be any type: image, striped, or sequential. The volume copy can use thin-provisioning or compression to save capacity. If the copies are in DRPs, you also can use deduplication to the volume copies to increase the capacity savings.

If you are creating a volume, the two copies can use different capacity reduction attributes. You can add a deduplicated volume copy in a DRP to a volume in a standard pool. You can use this method to migrate volumes to DRPs.

You can use mirrored volumes for the following reasons:

- ▶ Improving the availability of volumes by protecting them from a single storage system failure.
- ▶ Providing concurrent maintenance of a storage system that does not natively support concurrent maintenance.
- ▶ Providing an alternative method of data migration with better availability characteristics. While a volume is migrated by using the data migration feature, it is vulnerable to failures on the source and target pool. Volume mirroring provides an alternative because you can start with a non-mirrored volume in the source pool, and then add a copy to that volume in the destination pool.

When the volume is synchronized, you can delete the original copy that is in the source pool. During the synchronization process, the volume remains available, even if a problem occurs with the destination pool.

- ▶ Converting fully allocated volumes to use data reduction technologies, such as thin-provisioning, compression, or deduplication.
- ▶ Converting compressed or thin-provisioned volumes in standard pools to DRPs to improve capacity savings.

After a volume mirror is synchronized, a mirrored copy can become unsynchronized if it goes offline and write I/O requests must be processed, or if a mirror fast failover occurs. The fast failover isolates the host systems from temporarily slow-performing mirrored copies, which affect the system with a short interruption to redundancy.

Note: In standard volumes, the primary volume formats before synchronizing to the volume copies. The `-syncrate` parameter for the `mkvdisk` command controls the format and synchronization speed.

You can create a mirrored volume by using the **Mirrored** option in the Create Volume window, as showing in Figure 5-6. To display the Volume copy type selection ensure the Advanced settings mode is enabled.

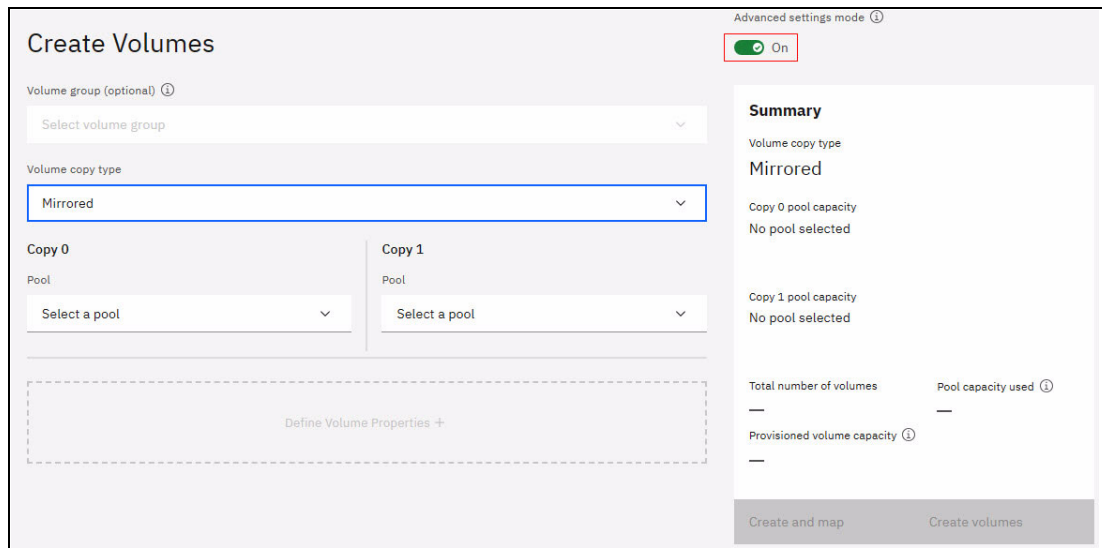


Figure 5-6 Mirrored volume creation

You can convert a non-mirrored volume into a mirrored volume by adding a copy, as shown in Figure 5-7.

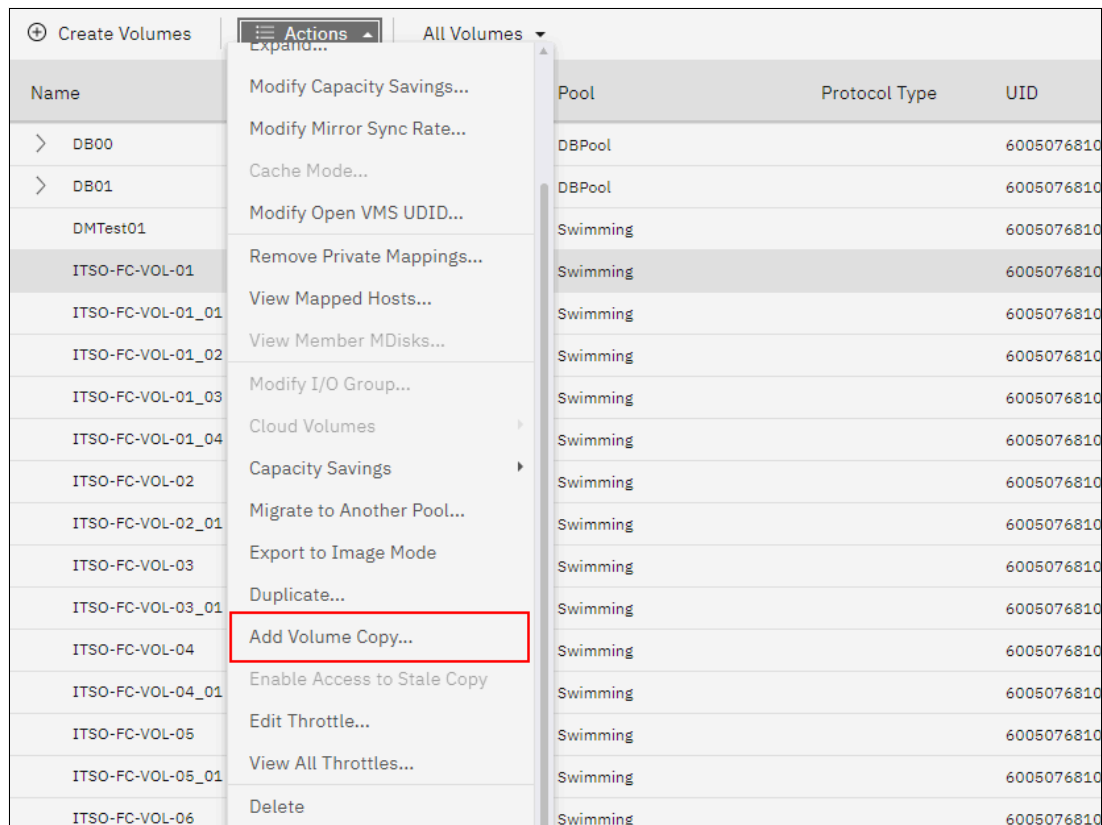


Figure 5-7 Adding a volume copy

5.4.1 Write fast failovers

With write fast failovers, the system submits writes to both copies during processing of host write I/O. If one write succeeds and the other write takes longer than 10 seconds, the slower request times out and ends. The duration of the ending sequence for the slow copy I/O depends on the back-end from which the mirror copy is configured. For example, if the I/O occurs over the Fibre Channel (FC) network, the I/O ending sequence typically completes in 10 - 20 seconds.

However, in rare cases, the sequence can take more than 20 seconds to complete. When the I/O ending sequence completes, the volume mirror configuration is updated to record that the slow copy is now no longer synchronized. When the configuration updates finish, the write I/O can be completed on the host system.

The volume mirror stops by using the slow copy for 4 - 6 minutes; subsequent I/O requests are satisfied by the remaining synchronized copy. During this time, synchronization is suspended. Also, the volume's synchronization progress shows less than 100% and decreases if the volume receives more host writes. After the copy suspension completes, volume mirroring synchronization resumes and the slow copy starts synchronizing.

If another I/O request times out on the unsynchronized copy during the synchronization, volume mirroring again stops by using that copy for 4 - 6 minutes. If a copy is always slow, volume mirroring attempts to synchronize the copy again every 4 - 6 minutes and another I/O timeout occurs.

The copy is not used for another 4 - 6 minutes and becomes progressively unsynchronized. Synchronization progress gradually decreases as more regions of the volume are written.

If write fast failovers occur regularly, an underlying performance problem might exist within the storage system that is processing I/O data for the mirrored copy that became unsynchronized. If one copy is slow because of storage system performance, multiple copies on different volumes are affected. The copies might be configured from the storage pool that is associated with one or more storage systems. This situation indicates possible overloading or other back-end performance problems.

When you run the `mkvdisk` command to create a volume, the `mirror_write_priority` parameter is set to `latency` by default. Fast failover is enabled. However, fast failover can be controlled by changing the value of the `mirror_write_priority` parameter on the `chvdisk` command. If the `mirror_write_priority` is set to `redundancy`, fast failover is disabled.

The system applies a full SCSI initiator-layer error recovery procedure (ERP) for all mirrored write I/O. If one copy is slow, the ERP can take up to 5 minutes. If the write operation is still unsuccessful, the copy is taken offline. Carefully consider whether maintaining redundancy or fast failover and host response time (at the expense of a temporary loss of redundancy) is more important.

Note: Mirrored volumes can be taken offline if no quorum disk is available. This behavior occurs because the synchronization status for mirrored volumes is recorded on the quorum disk. To protect against mirrored volumes being taken offline, follow the guidelines for setting up quorum disks.

5.4.2 Read fast failovers

Read fast failovers affect how the system processes read I/O requests. A read fast failover determines which copy of a volume the system tries first for a read operation. The primary-for-read copy is the copy that the system tries first for read I/O.

The system submits a host read I/O request to one copy of a volume at a time. If that request succeeds, the system returns the data. If it is not successful, the system retries the request to the other copy volume.

With read fast failovers, when the primary-for-read copy goes slow for read I/O, the system fails over to the other copy. Therefore, the system tries the other copy first for read I/O during the following 4 - 6 minutes. After that attempt, the system reverts to read the original primary-for-read copy.

During this period, if read I/O to the other copy also is slow, the system reverts immediately. Also, if the primary-for-read copy changes, the system reverts to try the new primary-for-read copy. This issue can occur when the system topology changes or when the primary or local copy changes. For example, in a standard topology, the system normally tries to read the primary copy first. If you change the volume's primary copy during a read fast failover period, the system reverts to read the newly set primary copy immediately.

The read fast failover function is always enabled on the system. During this process, the system does not suspend the volumes or make the copies out of sync.

5.4.3 Maintaining data integrity of mirrored volumes

Volume mirroring improves data availability by allowing hosts to continue I/O to a volume, even if one of the back-end storage systems fails. However, this mirroring does not enhance data integrity. If either of the back-end storage systems corrupts the data, the host is at risk of reading that corrupted data in the same way as for any other volume.

Therefore, before you perform maintenance on a storage system that might affect the data integrity of one copy, it is important to check that both volume copies are synchronized. Then, remove that volume copy before you begin the maintenance.

5.5 HyperSwap volumes

HyperSwap volumes create copies on two separate sites for systems that are configured with HyperSwap topology. Data that is written to a HyperSwap volume is automatically sent to both copies so that either site can provide access to the volume if the other site becomes unavailable.

HyperSwap is a system topology that enables high availability (HA) and disaster recovery (DR) (HADR) between I/O groups at different locations. Before you configure HyperSwap volumes, the system topology must be configured for HyperSwap and sites must be defined. Figure 5-8 on page 338 shows an overall view of HyperSwap that is configured with two sites.

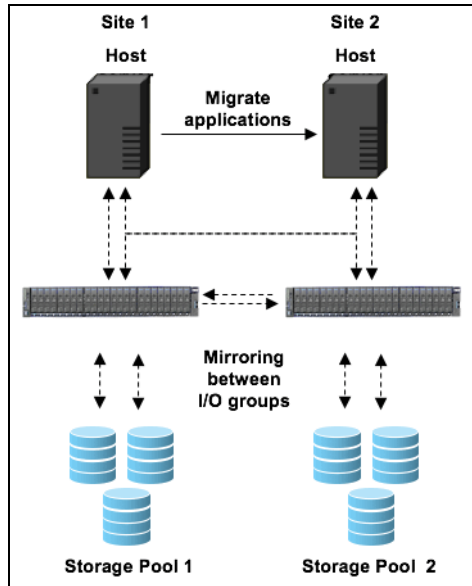


Figure 5-8 Overall HyperSwap diagram

In the management GUI, HyperSwap volumes are configured by specifying volume details, such as quantity, capacity, name, and the method for saving capacity. As with basic volumes, you can choose compression or thin-provisioning to save capacity on volumes.

For thin-provisioning or compression, you can also select to use deduplication for the volume that you create. For example, you can create a compressed volume that also uses deduplication to remove duplicated data.

The method for capacity savings applies to all HyperSwap volumes and copies that are created. The volume location displays the site where copies are located, based on the configured sites for the HyperSwap system topology. For each site, specify a pool and I/O group that are used by the volume copies that are created on each site. If you select to deduplicate volume data, the volume copies must be in DRPs on both sites.

The management GUI creates an HyperSwap relationship and change volumes (CVs) automatically. HyperSwap relationships manage the synchronous replication of data between HyperSwap volume copies at the two sites.

If your HyperSwap system supports self-compressing FCMs and the base volume is fully allocated in a DRP, the corresponding CV is created with compression enabled. If the base volume is in a standard pool, the CV is created as a thin-provisioned volume.

You can specify a consistency group (CG) that contains multiple active-active relationships to simplify management of replication and provide consistency across multiple volumes. A CG is commonly used when an application spans multiple volumes. CVs maintain a consistent copy of data during resynchronization. CVs allow an older copy to be used for DR if a failure occurred on the up-to-date copy before resynchronization completes.

You can also use the `mkvolume` CLI to create a HyperSwap volume. The command also defines pools and sites for HyperSwap volume copies and creates the active-active relationship and CVs automatically.

You can see the relationship between the master and auxiliary volume in a 2-site HyperSwap topology in Figure 5-9 on page 339.

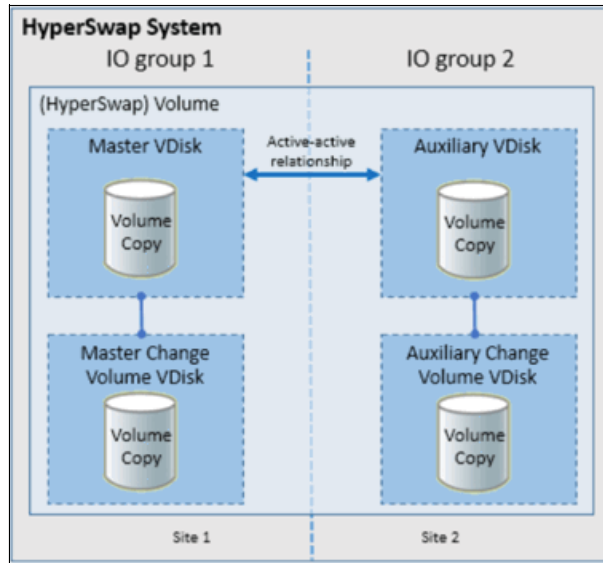


Figure 5-9 Master and auxiliary volumes

For more information about HyperSwap volumes, see 7.6, “HyperSwap internals” on page 515.

5.6 VMware vSphere Virtual Volumes

IBM Storage Virtualize supports vSphere Virtual Volumes (vVols), which allow VMware vCenter to automate the management of system objects, such as volumes and pools. Refer to the IBM Redbooks *IBM Storage Virtualize and VMware: Integrations, Implementation and Best Practices*, SG24-8549-01 for more details on vVols.

5.7 Cloud volumes

A *cloud volume* is any volume that is enabled for TCT. After TCT is enabled on a volume, point-in-time copies or snapshots can be created and copied to cloud storage that is provided by a cloud service provider (CSP). These snapshots can be restored to the system for DR purposes. Before you create cloud volumes, a valid connection to a supported CSP must be configured.

With TCT, the system supports connections to CSPs and the creation of cloud snapshots of any volume or volume group on the system. Cloud snapshots are point-in-time copies of volumes that are created and transferred to cloud storage that is managed by a CSP.

A cloud account defines the connection between the system and a supported CSP. It also must be configured before data can be transferred to or restored from the cloud storage. After a cloud account is configured with the CSP, you determine which volumes you want to create cloud snapshots of and enable TCT on those volumes.

Figure 5-10 shows an example of TCT.

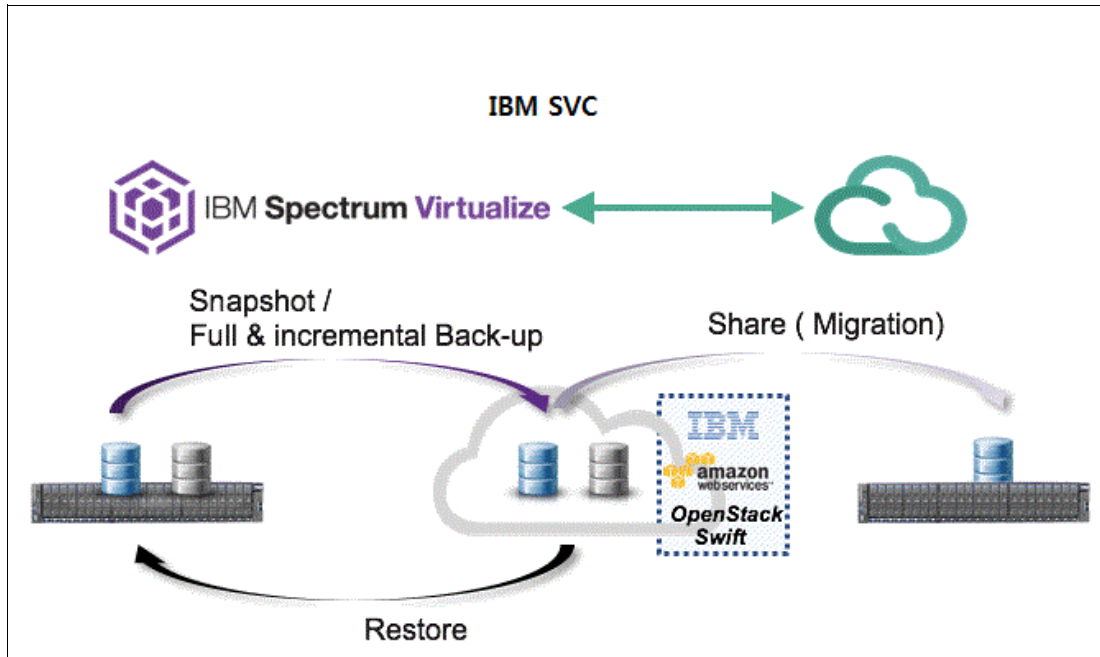


Figure 5-10 Transparent Cloud Tiering example

A cloud account is an object on the system that represents a connection to a CSP by using a particular set of credentials. These credentials differ depending on the type of CSP that is being specified. Most CSPs require the hostname of the CSP and an associated password. Some CSPs also require certificates to authenticate users of the cloud storage. Public clouds use certificates that are signed by well-known certificate authorities.

Private CSPs can use a self-signed certificate or a certificate that is signed by a trusted certificate authority. These credentials are defined on the CSP and passed to the system through the administrators of the CSP.

A cloud account defines whether the system can successfully communicate and authenticate with the CSP by using the account credentials.

If the system is authenticated, it can access cloud storage to copy data to the cloud storage or restore data that is copied to cloud storage back to the system. The system supports one cloud account to a single CSP. Migration between providers is *not* supported.

5.7.1 Transparent Cloud Tiering configuration limitations and rules

Consider the following limitations and rules regarding TCT:

- ▶ One cloud account per system.
- ▶ A maximum of 1024 volumes can have cloud snapshot-enabled volumes.
- ▶ The maximum number of active snapshots per volume is 256.
- ▶ The maximum number of volume groups is 512.
- ▶ Cloud volumes cannot be resized, either larger or smaller.

- ▶ A volume cannot be configured for a cloud snapshot if any of the following conditions exist:
 - The volume is part of a remote copy relationship (Metro Mirror, Global Mirror, active-active) master, auxiliary, or change volume. This restriction prevents the cloud snapshot from being used with HyperSwap volumes.
 - The volume is a VMware Virtual Volume, including FlashCopy owned volumes that are used internally for VMware Virtual Volume restoration functions.
 - The volume is:
 - A file system volume.
 - Associated with any user-owned FlashCopy maps.
 - A mirrored volume with copies in different storage pools.
 - Being migrated between storage pools.
- ▶ A volume cannot be enabled for cloud snapshots if the cloud storage is set to import mode.
- ▶ A volume cannot be enabled for cloud snapshots if the maximum number of cloud volumes exists. The maximum number of cloud volumes on the system is 1024. If the system exceeds this limit, you can disable cloud snapshots on a cloud volume and delete its associated snapshots from the cloud storage to accommodate snapshots on new cloud volumes.
- ▶ A volume cannot be used for a restore operation if it meets any of the following criteria:
 - A VMware Virtual Volume, including FlashCopy volumes that are used internally for VMware Virtual Volume restoration functions.
 - A file system volume.
 - Part of a remote copy relationship (MM, GM, or active-active) master, auxiliary, or CV.
- ▶ A volume that is configured for backup or is being used for restoration cannot be moved between I/O groups.
- ▶ Only one operation (cloud snapshot, restore, or snapshot deletion) is allowed at a time on a cloud volume.
- ▶ Cloud volume traffic is allowed only through management interfaces (1 G or 10 G).

5.7.2 Restoring to the production volume

This process is used to restore a snapshot version to the production volume, which is the original volume from which the snapshots were created. After the restore operation completes, the snapshot version completely replaces the current data that is on the production volume. During the restore operation, the production volume goes offline until it completes. Data is not fully restored to the production volume until the changes are committed.

5.7.3 Restoring to a new volume

If you do not want to have the production volume offline for the restore, you can restore a cloud snapshot to a new volume. The production volume remains online and host operations are not disrupted.

When the snapshot version is restored to a new volume, you can use the restored data independently of the original volume from which the snapshot was created. If the new volume exists on the system, the restore operation uses the unique identifier (UID) of the new volume.

If the new volume does not exist on the system, you must choose whether to use the UID from the original volume or create a UID. If you plan to use the new volume on the same system, use the UID that is associated with the snapshot version that is being restored.

5.8 Volume migration

Migrating an image mode volume to managed mode volume or vice versa is done by migrating a volume from one storage pool to another one. A non-image mode volume also can be migrated to a different storage pool.

When migrating from image to managed or vice versa, the command varies, as shown in Table 5-1.

Table 5-1 Migration types and associated commands

Storage pool-to-storage pool type	Command
Managed-to-managed or Image-to-managed	migratevdisk
Managed-to-image or Image-to-image	migratetoimage

Migrating a volume from one storage pool to another one is nondisruptive to the host application that uses the volume. Depending on the workload of the IBM Storage Virtualize system, performance might be slightly affected.

The migration of a volume from one storage pool to another storage pool by using the **migratevdisk** command is allowed only if both storage pools feature the same extent size. Volume mirroring can be used if a volume must be migrated from one storage pool to another storage pool with different extent sizes. Also, you may use the **migratevdisk** command to or from DRP only if a volume is fully allocated.

5.8.1 Image-type to striped-type volume migration

When you are migrating storage into IBM FlashSystem, the storage is brought in as *image-type volumes*, which mean that the volume is based on a single MDisk. The CLI command that you can use is **migratevdisk**.

Example 5-5 shows the **migratevdisk** command that can be used to migrate an image-type volume to a striped-type volume. The command also can be used to migrate a striped-type volume to a striped-type volume.

Example 5-5 The migratevdisk command

```
superuser> migratevdisk -mdiskgrp MDG1DS4K -threads 4 -vdisk Migrate_sample
```

This command migrates the volume `Migrate_sample` to the storage pool `MDG1DS4K`, and uses four threads when migrating. Instead of using the volume name, you can use its ID number.

You can monitor the migration process by using the **lsmigrate** command, as shown in Example 5-6.

Example 5-6 Monitoring the migration process

```
superuser> lsmigrate
migrate_type MDisk_Group_Migration
progress 0
migrate_source_vdisk_index 3
migrate_target_mdisk_grp 2
max_thread_count 4
migrate_source_vdisk_copy_id 0
```

5.8.2 Migrating to an image-type volume

An *image-type volume* is a direct, “straight-through” mapping to one image mode MDisk. If a volume is migrated to another MDisk, the volume is represented as being in managed mode during the migration (because it is striped on two MDisks).

A volume is represented only as an image-type volume after it reaches the state where it is a straight-through mapping. An image-type volume cannot be expanded.

Image-type disks are used to migrate data to IBM FlashSystem and migrate data out of virtualization. In general, the reason for migrating a volume to an image-type volume is to move the data on the disk to a non-virtualized environment.

If the migration is interrupted by a cluster recovery, the migration resumes after the recovery completes.

The **migratetoimage** command migrates the data of a user-specified volume by consolidating its extents (which might be on one or more MDisks) onto the extents of the target MDisk that you specify. After migration is complete, the volume is classified as an image-type volume, and the corresponding MDisk is classified as an image-mode MDisk.

The MDisk that is specified as the target must be in an *unmanaged* state at the time that the command is run. Running this command results in the inclusion of the MDisk into the user-specified storage pool.

Remember: This command cannot be used if the source volume copy is in a child pool or if the target MDisk group that is specified is a child pool. This command does not work if the volume is fast formatting.

The **migratetoimage** command fails if the target or source volume is offline. Correct the offline condition before attempting to migrate the volume.

If the volume (or volume copy) is a target of a FlashCopy mapping with a source volume in an active-active relationship, the new MDisk group must be in the same site as the source volume. If the volume is in an active-active relationship, the new MDisk group must be in the same site as the source volume. Also, the site information for the MDisk being added must be defined and match the site information for other MDisks in the storage pool.

Note: You cannot migrate a volume or volume image between storage pools if cloud snapshot is enabled on the volume.

An encryption key cannot be used when migrating an image-mode MDisk. To use encryption (when the MDisk has an encryption key), the MDisk must be self-encrypting before configuring the storage pool.

The **migratetoimage** command is useful when you want to use your system as a data mover. For more information, see [migratetoimage](#).

5.8.3 Migrating with volume mirroring

With volume mirroring, you can migrate volumes between storage pools with different extent sizes.

To migrate volumes between storage pools, complete the following steps:

1. Add a copy to the target storage pool.
2. Wait until the synchronization is complete.
3. Remove the copy from the source storage pool.

To migrate from a thin-provisioned volume to a fully allocated volume, the process is similar:

1. Add a target fully allocated copy.
2. Wait for synchronization to complete.
3. Remove the source thin-provisioned copy.

In both cases, if you set the **autodelete** option to yes when creating the volume copy, the source copy is automatically deleted, and you can skip the third step in both processes. The best practice for this type of migration is to try not to overload the systems with a high syncrate or with too many migrations at the same time.

The **syncrate** parameter specifies the copy synchronization rate. A value of zero prevents synchronization. The default value is 50. The supported **-syncrate** values and their corresponding rates are listed in Table 5-2.

Table 5-2 Sample syncrate values

User-specified syncrate attribute value	Data copied per second
1 - 10	128 KB
11 - 20	256 KB
21 - 30	512 KB
31 - 40	1 MB
41 - 50	2 MB
51 - 60	4 MB
61 - 70	8 MB
71 - 80	16 MB
81 - 90	32 MB
91 - 100	64 MB
101 - 110	128 MB
111 - 120	256 MB
121 - 130	512 MB
131 - 140	1 GB
141 - 150	2 GB

We recommend modifying syncrate after monitoring overall bandwidth and latency. Then, if the performance is not affected for migration, increase the syncrate to complete within the allotted time.

You also can use volume mirroring when you migrate a volume from a non-virtualized storage device to an IBM Storage Virtualize system. As you can see in Figure 5-11 on page 346, you first must attach the storage to the IBM Storage Virtualize system (in this instance, an SVC), which requires some downtime because the hosts need to stop I/O, rediscover the volume through the IBM Storage Virtualize system, and then resume access.

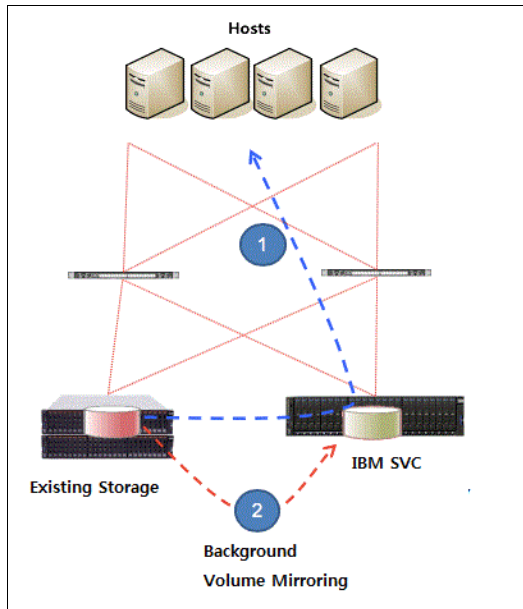


Figure 5-11 Migrating with volume mirroring

After the storage is correctly attached to the IBM Storage Virtualize system, map the image-type volumes to the hosts so that the host recognizes volumes as though they were accessed through the non-virtualized storage device. Then, you can restart applications.

After that process completes, you can use volume mirroring to migrate the volumes to a storage pool with managed MDisks, which creates striped-type copies of each volume in this target pool. Data synchronization in the volume copies and then starts in the background.

For more information, see [Migrating volumes between pools using the CLI](#).

5.8.4 Migrating from standard pools to data reduction pools

If you want to move volumes to a DRP, you can move them by using volume mirroring between a standard pool and DRP. Host I/O operations are not disrupted during migration.

Figure 5-12 shows two examples of how you can use volume mirroring to convert volumes to a different type or migrate volumes to a different type of pool.

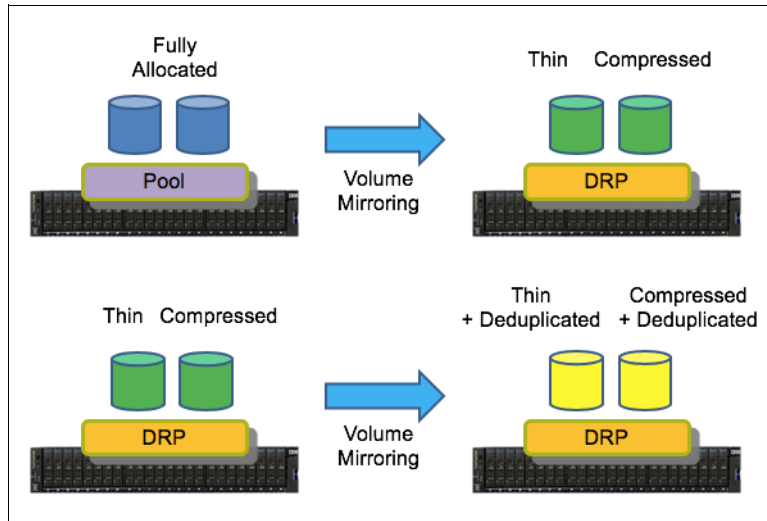


Figure 5-12 Converting volumes with volume mirroring

You also can move compressed or thin-provisioned volumes in standard pools to DRPs to simplify management of reclaimed capacity. The DRP tracks the unmap operations of the hosts and reallocates capacity automatically. The system supports volume mirroring to create a copy of the volume in a new DRP. This method creates a copy of the volume in a new DRP and does not disrupt host operations.

Deleting a volume copy in a DRP is a background task and can take a significant amount of time. During the deletion process, the deleting copy is still associated with the volume and a new volume mirror cannot be created until the deletion is complete. If you want to use volume mirroring again on the same volume without waiting for the delete, split the copy to be deleted to a new volume before deleting it.

For more information, see [Moving volumes to a data reduction pool](#).

5.8.5 Migrating a volume between systems nondisruptively

With nondisruptive system migration, storage administrators can migrate volumes from one IBM Storage Virtualize system to another without any application downtime. This function supports several use cases. For example, you can use this function to balance the load between multiple systems or to update and decommission hardware.

You also can migrate data between node-based systems and enclosure-based systems. Unlike replication remote copy types, nondisruptive system migration does not require a remote mirroring license before you can configure a remote copy relationship that is used for migration.

There are some configuration and host operating system restrictions that are documented in the *Configuration Limits and Restrictions* document under [Volume Mobility](#).

Prerequisites

The following prerequisites must be met for nondisruptive system migration:

- ▶ Both systems are running version 8.4.2 or later.
- ▶ An FC or IP partnership exists between the two systems that you want to migrate volumes between. The maximum supported round-trip time (RTT) between the two systems is 3 milliseconds. Ensure that the partnership has sufficient bandwidth to support the write throughput for all the volumes you are migrating. For more information, see the `mkfcpartnership` command for creating FC partnerships and `mkippartnership` command for creating IP partnerships.
- ▶ Any hosts that are mapped to volumes that you are migrating are correctly zoned to both systems. Hosts must appear in an online state on both systems.

Using the management GUI

To configure volume migration by using the GUI, complete the following steps:

1. On the source system, select **Volumes** → **Volumes**. On the Volumes page, identify the volumes that you want to migrate and record the volume name and capacity.
2. On the target system, select **Volumes** → **Volumes** and select **Create Volume**. Create the target volume within the appropriate storage tier with the same capacity as the source volume.
3. On the source system, select **Copy Services** → **Remote Copy**.
4. Click **Independent Relationship**.
5. Click **Create Relationship**.
6. On the Create Relationship page, click **Non-disruptive System Migration**.
7. Ensure that the auxiliary volume location specifies the system that you want to migrate to, and click **Next**.
8. Select the **Master** and **Auxiliary** volumes to use in the relationship.

Note: The volumes must be the same size. If the GUI window does not show the expected auxiliary volume, check the size by running the `lsvdisk -unit b <volume name or id>` command.

9. Select **Yes** to start the copy. Click **Finish**.
10. In the management GUI, select **Copy Services** → **Remote Copy** → **Independent Relationship**. Wait until the migration relationship that you created displays the Consistent Synchronized state.

Note: Data is copied to the target system at the lowest of partnership background copy rate or relationship bandwidth. The relationship bandwidth default is 25 MBps per relationship and can be increased by running the `chsystem -relationshipbandwidthlimit <new value in MBps>` command if needed.

11. Create host mappings to the auxiliary volumes on the remote system. Ensure that all auxiliary volumes are mapped to the same hosts that were previously mapped to the master volumes on the older system.
12. Ensure that the Host Bus Adapters (HBAs) in all hosts that are mapped to the volume are rescanned to ensure that all new paths are detected to the auxiliary volumes. Record the current path states on any connected hosts. Identify the worldwide port names (WWPNs) that are used for the active and standby (ghost) paths.

13. In the management GUI, select **Copy Services** → **Remote Copy** → **Independent Relationship**. Right-click the migration relationship and select **Switch Direction**. This action reverses the copy direction of the relationship and switches the active and standby paths, which result in all host I/O being directed to the new volume.
14. Validate that all hosts use the new paths to the volume by verifying that the paths that were reporting as standby (or ghost) are now reporting active. Verify that all previously active paths are now reporting standby (or ghost).

Note: Do not proceed if the added standby paths are not visible on the host. Standby paths might be listed under a different name on the host, such as “ghost” paths. Data access can be affected if all standby paths are not visible to the hosts when the direction is switched on the relationship.

15. Validate that all hosts use the target volume for I/O and verify that no issues exist.
16. On the original source system that was used in the migration, select **Hosts** → **Hosts**. Right-click the hosts and select **Unmap Volumes**. Verify the number of volumes that are being unmapped, and then click **Unmap**.
17. On the original source system, select **Volumes** → **Volumes**. Right-click the volumes and select **Delete**. Verify the number of volumes that are being deleted and click **Continue**.
The volume migration process is complete.

Using the command-line interface

To configure volume migration by using the CLI, complete the following steps:

1. On the source system, enter the `lsvdisk` command to determine all the volumes that you want to migrate to the target system.
In the results, record the name, ID, and capacity for each volume that you want to migrate to the target system.
2. On the target system, create volumes for each volume that you want to migrate, ensuring that you create the volume with the same capacity as the source volume, for example, `mkvolume -pool 0 -size 1000 -unit gb`.
3. On the source system, enter the following command to create a relationship for migration:
`mkrcrelationship -master sourcevolume -aux targetvolume -cluster system2 -migration -name migrationrc`
sourcevolume is the name or ID of the master volume on the source system, and targetvolume is the name or ID of the auxiliary volume that you created on the target system.
The `-migration` flag indicates that the remote copy relationship can be used to migrate data between the two systems that are defined in the partnership only.
Optionally, you can specify a name with the `-name` parameter. In this example, `migrationrc` is the name of the relationship. If no name is specified, an ID is assigned to the relationship.
4. On the source system, start the relationship by entering the running the `startcrelationship migrationrc` command, where `migrationrc` is the name of the relationship.
5. Verify that the state of the relationship is `consistent_synchronized` by entering the `lsrcrelationship migrationrc` command, where `migrationrc` is the name of the relationship. In the results that display, ensure that the state is `consistent_synchronized`.

Attention: Do not proceed until the relationship is in the `consistent_synchronized` state.

Depending on the amount of data that is being migrated, the process can take some time.

Note: Data is copied to the target system at the lowest of the partnership background copy rate or the relationship bandwidth. The relationship bandwidth default is 25 MBps per relationship, which can be increased by running the `chsystem -relationshipbandwidthlimit <new value in MBps>` command.

6. After the relationship is in the `consistent_synchronized` state, create host mappings to the auxiliary volumes on the target system by entering the `mkvdiskhostmap -host host1 targetvolume` command, where *targetvolume* is the name of the auxiliary volume on the target system. Ensure that all auxiliary volumes are mapped to the same hosts that were previously mapped to the master volumes on the source system.
7. On all hosts, ensure that the HBAs are mapped to the volume are rescanned to ensure that all new paths are detected to the auxiliary volumes. Record the current path states on any connected hosts. Identify the WWPNs that are used for the active and standby (ghost) paths.

Attention: Do not proceed if the added standby paths are not visible on the host. Standby paths might be listed under a different name on the host, such as “ghost” paths. Data access can be affected if all standby paths are not visible to the hosts when the direction is switched on the relationship.

8. Switch the direction of the relationship so that the auxiliary volume on the target system becomes the primary source for host I/O operations by running the `switchrelationship -primary aux migrationrc` command, where *migrationrc* indicates the name of the relationship. This command reverses the copy direction of the relationship and switches the active and standby paths, which result in all host I/O being directed to the auxiliary volume.
9. Validate that all hosts use the new paths to the volume by verifying that the paths previously reporting as standby (or ghost) are now reporting active.
10. Verify that all previously active paths are now reporting standby (or ghost).
11. Validate that all hosts use the target volume for I/O and verify that no issues exist.
12. On the original source system, unmap hosts from the original volumes by entering the `rmvdiskhostmap -host host1 sourcevolume` command, where *sourcevolume* is the name of the original volume that was migrated.
13. On the original source system, delete the original source volumes by entering the `rmvolume sourcevolume` command, where *sourcevolume* is the name of the original volume that was migrated.

The migration process is now complete.

5.9 Preferred paths to a volume

When a volume is created, it is assigned to an I/O group and assigned a preferred node. The *preferred node* is the node that normally processes I/Os for the volume. The *primary purposes* of a preferred node are load balancing and determining which node destages writes.

Preferred node assignment is normally automatic. The system selects the node in the I/O group that includes the fewest volumes. However, the preferred node can be specified or changed, if needed.

All modern multipathing drivers support Asymmetric Logical Unit Access (ALUA). This access allows the storage to mark certain paths as preferred (paths to the preferred node). ALUA multipathing drivers acknowledge preferred pathing and send I/O to only the other node if the preferred node is not accessible.

Figure 5-13 shows write operations from a host to two volumes with different preferred nodes.

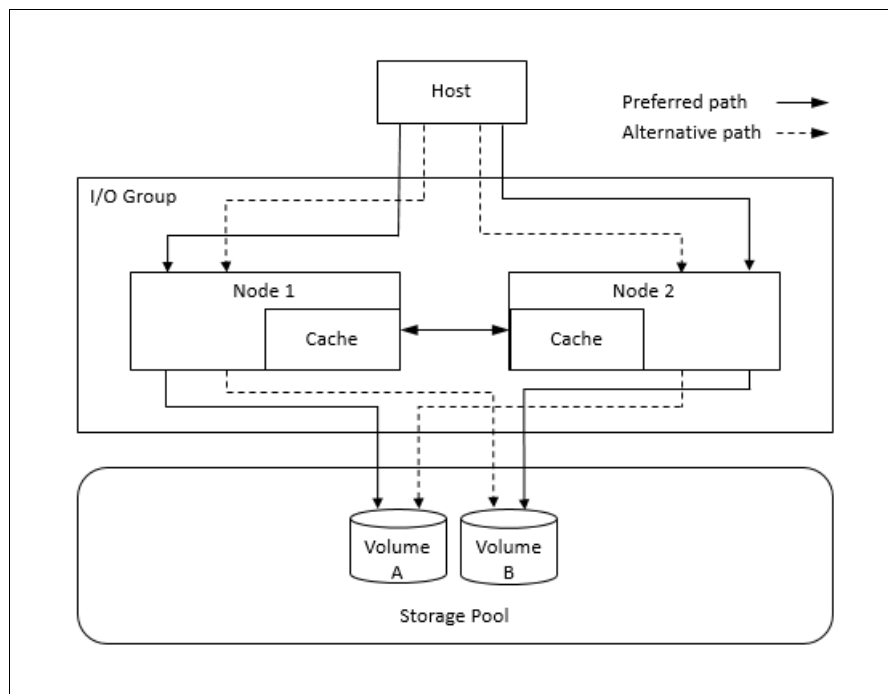


Figure 5-13 Write operations from a host through different preferred nodes for each volume

When debugging performance problems, it can be useful to review the Non-Preferred Node Usage Percentage metric in IBM Spectrum Control or IBM Storage Insights. I/O to the non-preferred node might cause performance problems for the I/O group, which can be identified by these tools.

For more information about this performance metric and IBM Spectrum Control, see [Performance metrics for resources that run IBM Spectrum Virtualize](#).

5.10 Moving a volume between I/O groups and nodes

To balance the workload across I/O groups and nodes, you can move volumes between I/O groups and nodes.

Changing the preferred node of a volume within an I/O group or to another I/O group is a nondisruptive process.

5.10.1 Changing the preferred node of a volume within an I/O group

Changing the preferred node within an I/O group can be done with concurrent I/O. However, it can lead to some delay in performance, and in some specific operating systems or applications, they might detect some timeouts.

This operation can be done by using the CLI and GUI, but if you have only one I/O group, this operation is not possible by using the GUI. To change the preferred node within an I/O group by using the CLI, run the following command:

```
movevdisk -node <node_id or node_name> <vdisk_id or vdisk_name>
```

5.10.2 Moving a volume between I/O groups

Moving a volume between I/O groups should be done on only an SVC. When using IBM FlashSystem, a volume should be migrated so that the contents of the volume are on the internal storage of the owning IBM FlashSystem. When moving a volume between I/O groups, it is recommended that the system chooses the volume preferred node in the new I/O group. However, it is possible to manually set the preferred node during this operation by using the GUI and CLI.

Some limitations exist in moving a volume across I/O groups, which is called Non-Disruptive Volume Movement (NDVM). These limitations are mostly in host cluster environments. You can check their compatibility at [IBM System Storage Interoperation Center \(SSIC\)](#).

Note: These migration tasks can be nondisruptive if performed correctly and the hosts that are mapped to the volume support NDVM. The cached data that is held within the system first must be written to disk before the allocation of the volume can be changed.

Modifying the I/O group that services the volume can be done concurrently with I/O operations if the host supports nondisruptive volume move. It also requires a rescan at the host level to ensure that the multipathing driver is notified that the allocation of the preferred node changed and the ports by which the volume is accessed changed. This rescan can be done in a situation where one pair of nodes becomes over-used.

If any host mappings are available for the volume, the hosts must be members of the target I/O group or the migration fails.

Verify that you created paths to I/O groups on the host system. After the system successfully adds the new I/O group to the volume's access set and you moved the selected volumes to another I/O group, detect the new paths to the volumes on the host.

The commands and actions on the host vary depending on the type of host and the connection method that is used. These steps must be completed on all hosts to which the selected volumes are currently mapped.

Note: If the selected volume is performing quick initialization, this wizard is unavailable until quick initialization completes.

5.11 Volume throttling

Volume throttling effectively throttles the number of input/output operations per second (IOPS) or bandwidth (MBps) that can be achieved to and from a specific volume. You might want to use I/O throttling if you have a volume that has an access pattern that adversely affects the performance of other volumes.

For example, volumes that are used for backup or archive operations can have I/O-intensive workloads, potentially taking bandwidth from production volumes. A volume throttle can be used to limit I/Os for these types of volumes so that I/O operations for production volumes are not affected.

Figure 5-14 shows an example of volume throttling.

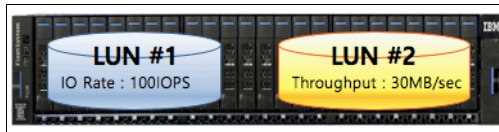


Figure 5-14 Volume throttling for each volume

When deciding between using IOPS or bandwidth as the I/O governing throttle, consider the disk access pattern of the application. Database applications often issue much I/O, but they transfer only a relatively small amount of data. In this case, setting an I/O governing throttle that is based on MBps does not achieve the expected result. Therefore, it is better to set an IOPS limit.

However, a streaming video application often issues a small amount of I/O but transfers much data. In contrast to the database example, defining an I/O throttle that is based on IOPS does not achieve a good result. For a streaming video application, it is better to set an MBps limit.

You can edit the throttling value in the menu, as shown in Figure 5-15.

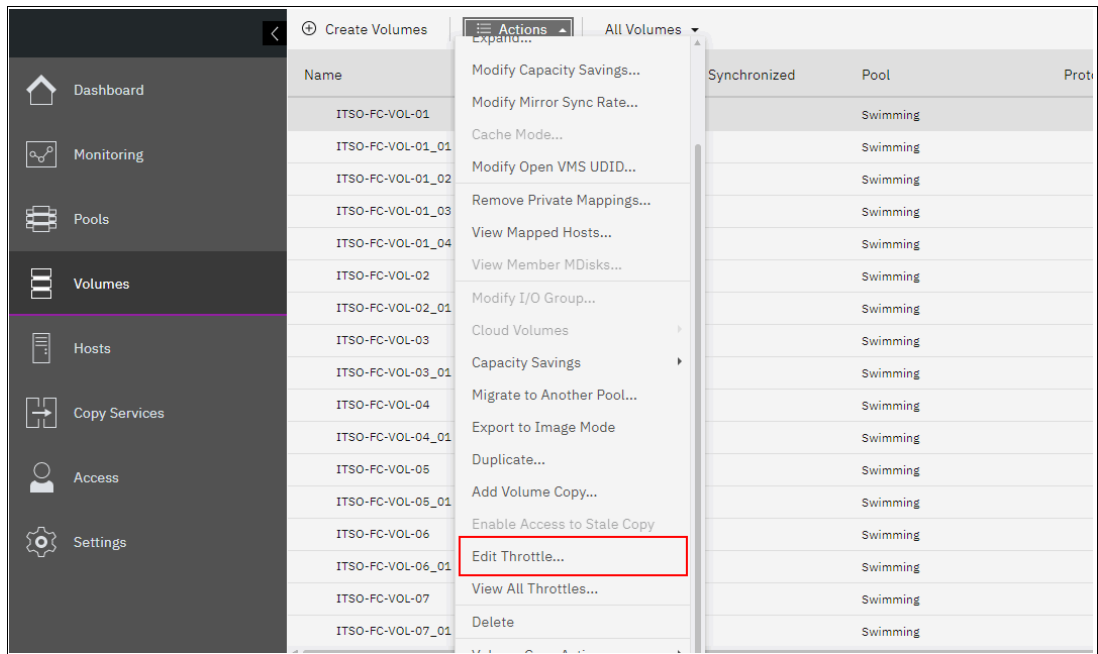


Figure 5-15 Volume throttling

Figure 5-16 shows both the bandwidth and IOPS parameters that can be set.

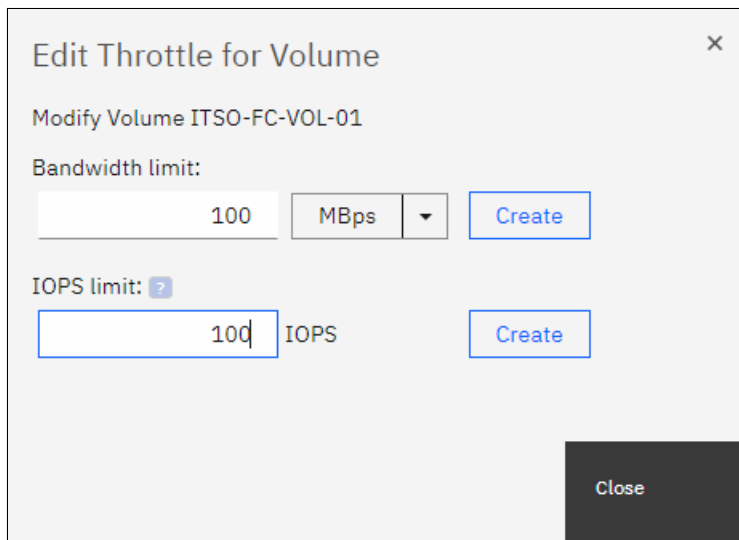


Figure 5-16 Edit bandwidth and IOPS limit

Throttling at a volume level can be set by using the following commands:

► **mkthrottle**

This command is used to set I/O throttles for volumes that use this command. The command must be used with the **-type vdisk** parameter, followed by **-bandwidth bandwidth_limit_in_mbdisk** or **-iops iops_limit** to define MBps and IOPS limits.

► **chvdisk**

When used with **-rate throttle_rate** parameter, this command specifies the IOPS and MBps limits. The default **throttle_rate** units are I/Os. To change the **throttle_rate** units to Mbps, specify the **-unitmb** parameter. If **throttle_rate** value is zero, the throttle rate is disabled. By default, the **throttle_rate** parameter is disabled.

Note: The **mkthrottle** command can be used to create throttles for volumes, hosts, host clusters, pools, or system offload commands.

When the IOPS limit is configured on a volume and it is smaller than 100 IOPS, the throttling logic rounds it to 100 IOPS. Even if the throttle is set to a value smaller than 100 IOPS, the throttling occurs at 100 IOPS.

After any of the commands that were described thus far are used to set volume throttling, a throttle object is created. Then, you can list your created throttle objects by using the **lsthrottle** command and change their parameters with the **chthrottle** command. Example 5-7 shows some command examples.

Example 5-7 Throttle commands example

```
superuser>mkthrottle -type vdisk -bandwidth 100 -vdisk Vo101
Throttle, id [0], successfully created.
superuser>lsthrottle
throttle_id throttle_name object_id object_name throttle_type IOPs_limit
bandwidth_limit_MB
0          throttle0      52          Vo101          vdisk          100

superuser>chthrottle -iops 1000 throttle0
superuser>lsthrottle
throttle_id throttle_name object_id object_name throttle_type IOPs_limit
bandwidth_limit_MB
0          throttle0      52          Vo101          vdisk          1000          100

superuser>lsthrottle throttle0
id 0
throttle_name throttle0
object_id 52
object_name Vo101
throttle_type vdisk
IOPs_limit 1000
bandwidth_limit_MB 100
```

Note: The throttle reduces IOPS or bandwidth by adding latency as a resource approaches a defined throttle. This increased response time is observable in performance monitoring tools.

For more information, see [Managing throttles for volumes](#).

5.12 Volume cache mode

Cache mode in IBM Storage Virtualize systems determines whether read/write operations are stored in cache. For each volume, one of the following cache modes can be used:

- ▶ readwrite (enabled)

All read/write I/O operations that are performed by the volume are stored in cache. This default cache mode is used for all volumes. A volume or volume copy that is created from a DRP must have a cache mode of readwrite.

When you create a thin-provisioned volume, set the cache mode to readwrite to maximize performance. If you set the mode to none, the system cannot cache the thin-provisioned metadata, and performance is decreased. In a DRP, a thin-provisioned or compressed volume copy setting cannot be created for a cache mode that is different than readwrite.

- ▶ readonly

All read I/O operations that are performed by the volume are stored in cache.

- ▶ none (disabled)

All read/write I/O operations that are performed by the volume are not stored in cache.

By default, when a volume is created, the cache mode is set to readwrite. Disabling cache can affect performance and increase read/write response time.

Figure 5-17 shows write operation behavior when a volume cache is activated (readwrite).

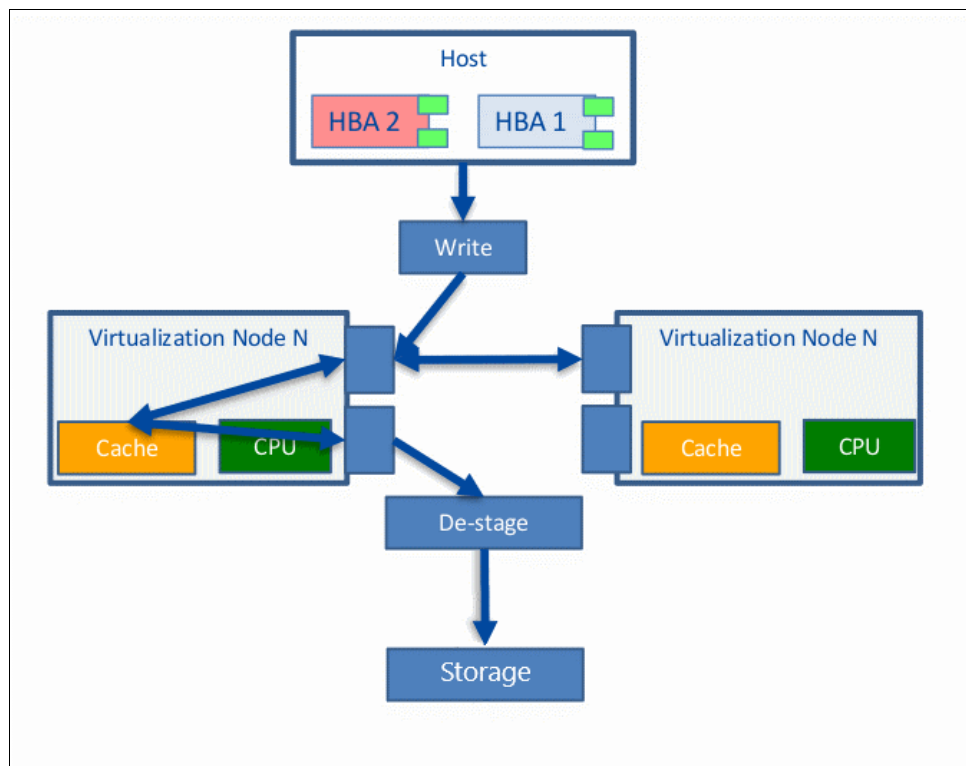


Figure 5-17 Cache activated

Figure 5-18 shows a write operation behavior when volume cache is deactivated (none).

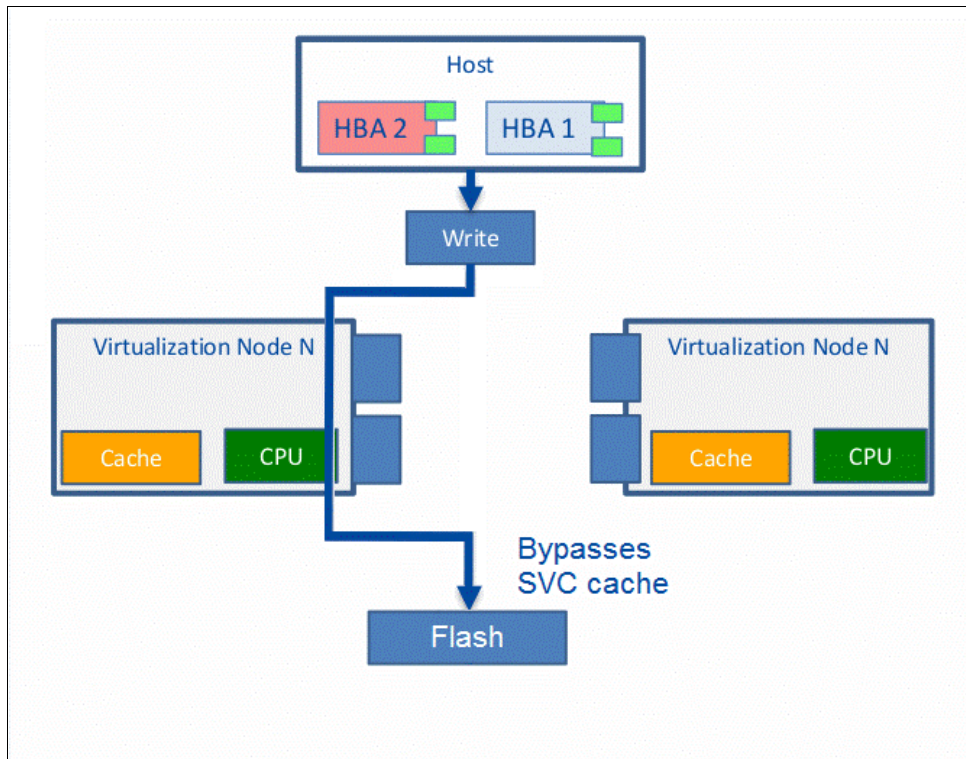


Figure 5-18 Cache deactivated

In most cases, the volume with readwrite cache mode is recommended because disabling cache for a volume can result in performance issues to the host. However, some specific scenarios exist in which it is recommended to disable the readwrite cache.

You might use cache-disabled (none) volumes when you have remote copy or FlashCopy in a back-end storage controller, and these volumes are virtualized in IBM FlashSystem devices as image virtual disks (VDisks). Another possible usage of a cache-disabled volume is when intellectual capital is in Copy Services automation scripts. Keep the usage of cache-disabled volumes to a minimum for normal workloads.

You can also use cache-disabled volumes to control the allocation of cache resources. By disabling the cache for specific volumes, more cache resources are available to cache I/Os to other volumes in the same I/O group. An example of this usage is a non-critical application that uses volumes in MDisks from all-flash storage.

Note: Volumes with readwrite cache enabled are recommended.

By default, volumes are created with cache mode enabled (read/write); however, you can specify the cache mode when the volume is created by using the `-cache` option.

The cache mode of a volume can be concurrently changed (with I/O) by using the `chvdisk` command or using the GUI by selecting **Volumes** → **Volumes** → **Actions** → **Cache Mode**. Figure 5-19 on page 358 shows the editing cache mode for a volume.

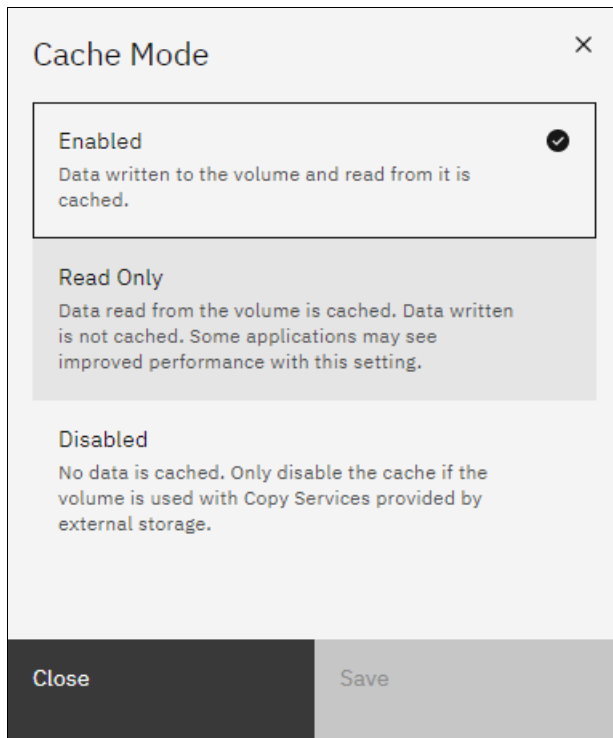


Figure 5-19 Editing cache mode

The CLI does not fail I/O to the user, and the command must be allowed to run on any volume. If used correctly without the **-force** flag, the command does not result in a corrupted volume. Therefore, the cache must be flushed and you must discard cache data if the user disables cache on a volume.

Example 5-8 shows an image volume `VDISK_IMAGE_1` that changed the cache parameter after it was created.

Example 5-8 Changing the cache mode of a volume

```

superuser>mkvdisk -name VDISK_IMAGE_1 -iogrp 0 -mdiskgrp IMAGE_Test -vtype image
-mdisk D8K_L3331_1108
Virtual Disk, id [9], successfully created
superuser>lsvdisk VDISK_IMAGE_1
id 9
.
lines removed for brevity
.
fast_write_state empty
cache readwrite
.
lines removed for brevity

superuser>chvdisk -cache none VDISK_IMAGE_1
superuser>lsvdisk VDISK_IMAGE_1
id 9
.
lines removed for brevity
.

```

cache none

lines removed for brevity

In an environment with copy services (FlashCopy, MM, GM, and volume mirroring) and typical workloads, disabling IBM FlashSystem cache is detrimental to overall performance.

Attention: Carefully evaluate the effect to the entire system with quantitative analysis before and after making this change.

5.13 Other considerations

This section describes other considerations regarding volumes.

5.13.1 Volume protection

You can protect volumes to prevent active volumes or host mappings from being deleted. IBM Storage Virtualize systems feature a global setting that is enabled by default that prevents these objects from being deleted if the system detects recent I/O activity. You can set this value to apply to all volumes that are configured on your system, or control whether the system-level volume protection is enabled or disabled on specific pools.

To prevent an active volume from being deleted unintentionally, administrators must enable volume protection. They also can specify a period that the volume must be idle before it can be deleted. If volume protection is enabled and the period is not expired, the volume deletion fails, even if the **-force** parameter is used.

When you delete a volume, the system verifies whether it is a part of a host mapping, FlashCopy mapping, or remote copy relationship. In these cases, the system fails to delete the volume unless the **-force** parameter is specified. However, if volume protection is enabled, the **-force** parameter does not delete a volume if it has I/O activity in the last minutes that are defined in the protection duration time in volume protection.

Note: The **-force** parameter overrides the volume dependencies, not the volume protection setting. Volume protection must be disabled to permit a volume or host-mapping deletion if the volume had recent I/O activity.

Consider enabling volume protection by running the following command:

```
chsystem vdiskprotectionenabled yes -vdiskprotectiontime <value_in_minutes>
```

If you want volume protection enabled in your system but disabled in a specific storage pool, run the following command:

```
chmdiskgrp -vdiskprotectionenabled no <pool_name_or_ID>
```

You can also manage volume protection in the GUI by selecting **Settings** → **System** → **Volume Protection**, as shown in Figure 5-20.

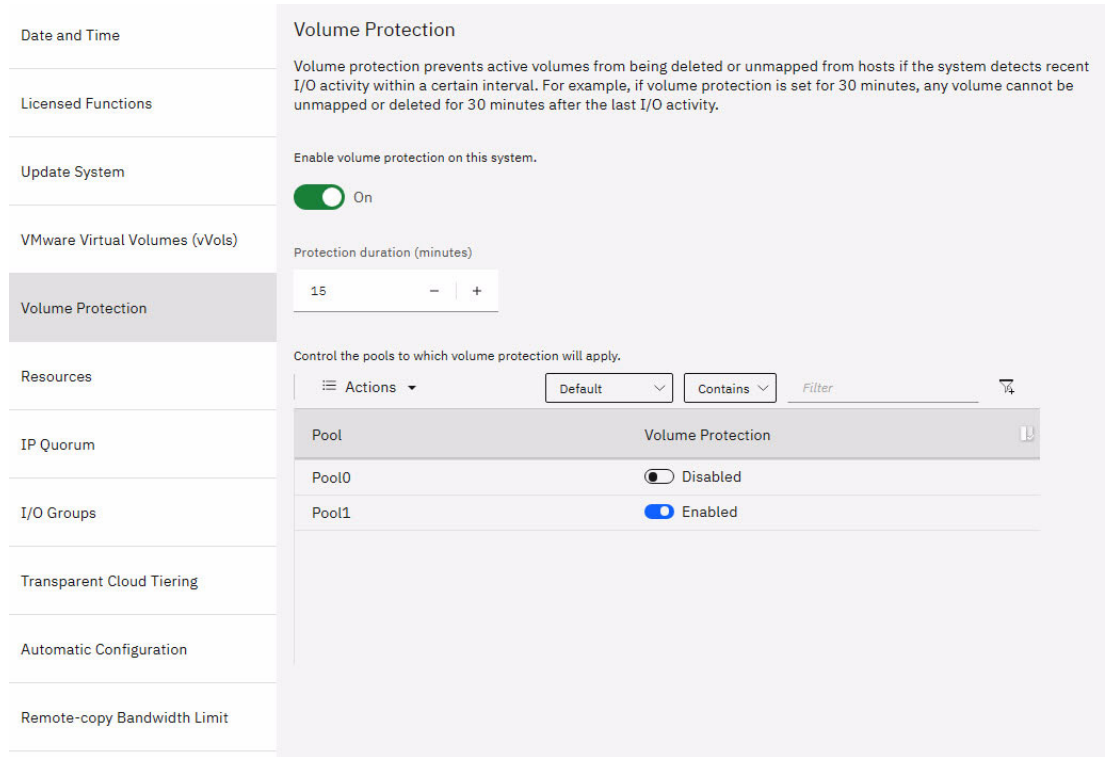


Figure 5-20 Volume Protection

5.13.2 Volume resizing

You can increase and decrease the sizes of fully allocated and thin-provisioned volumes. A volume can be expanded with concurrent I/Os for some operating systems. However, *never* attempt to shrink a volume that is in use that contains data because volume capacity is removed from the end of the disk, whether that capacity is in use by a server. A volume cannot be expanded or shrunk during its quick initialization process.

Expanding a volume

You can expand volumes for the following reasons:

- ▶ To increase the available capacity on a specific volume that is mapped to a host.
- ▶ To increase the size of a volume to make it match the size of the source or master volume so that it can be used in a FlashCopy mapping or MM relationship.

Figure 5-21 shows the Modify Volume Capacity window.

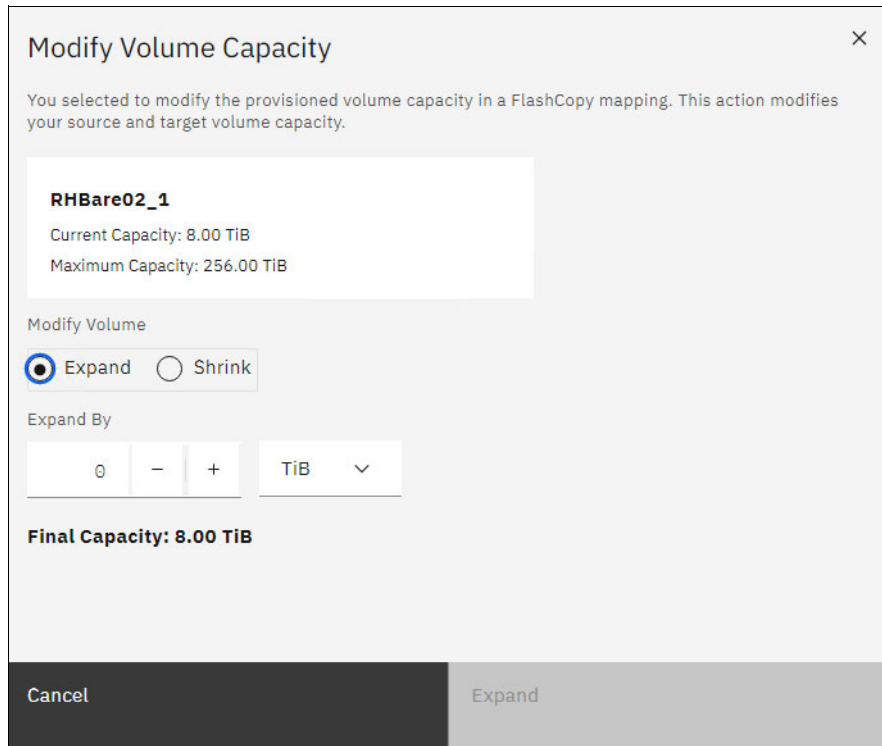


Figure 5-21 Expanding a volume

Shrinking a volume

Volumes can be reduced in size if necessary. If a volume does not contain any data, it is unlikely that you will encounter any issues when shrinking its size. However, if a volume is in use and contains data, do not shrink its size because IBM Storage Virtualize is unaware of whether it is removing used or non-used capacity.

Attention: When you shrink a volume, capacity is removed from the end of the disk, whether that capacity is in use. Even if a volume includes free capacity, do not assume that only unused capacity is removed when you shrink a volume.

Figure 5-22 shows the Modify Volume Capacity window, which you can use to shrink volumes.

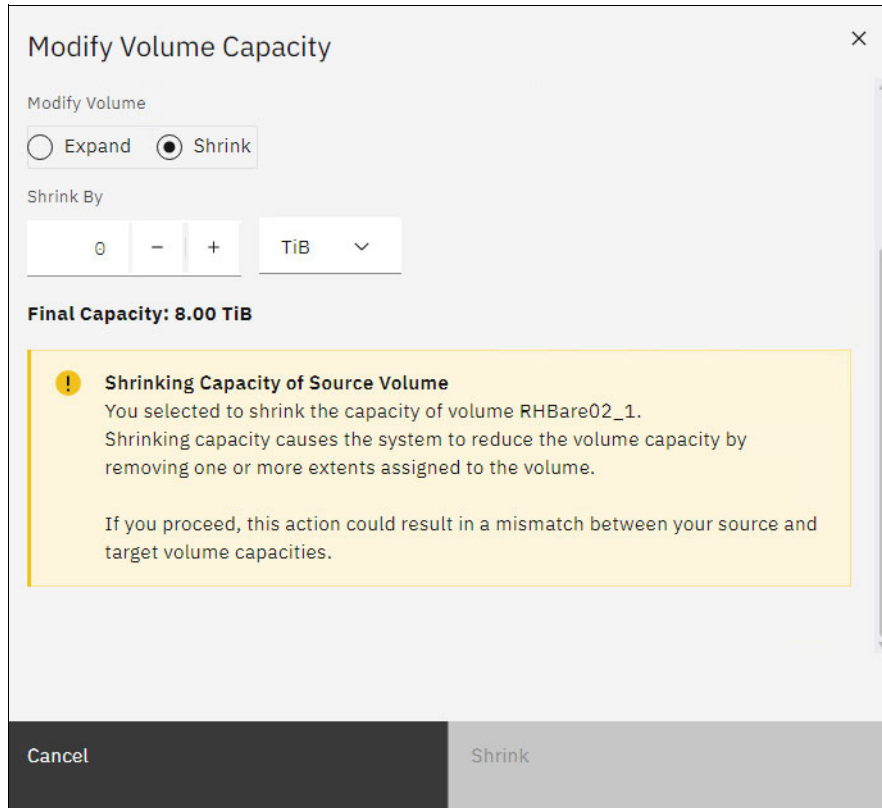


Figure 5-22 Shrinking volumes



Copy services

Copy services are a collection of functions that provide capabilities for disaster recovery (DR), data migration, and data duplication solutions.

This chapter provides an overview and best practices of IBM Storage Virtualize Copy Services capabilities, including FlashCopy, Metro Mirror (MM) and Global Mirror (GM), and volume mirroring.

This chapter includes the following topics:

- ▶ “Introducing copy services” on page 364
- ▶ “IBM FlashCopy” on page 365
- ▶ “Safeguarded Copy” on page 391
- ▶ “Remote copy services” on page 397
- ▶ “Native IP replication” on page 458
- ▶ “Volume mirroring” on page 476
- ▶ “Policy-based replication” on page 486

6.1 Introducing copy services

IBM Storage Virtualize based systems, including IBM SAN Volume Controller (SVC) and the IBM FlashSystem family, offer a complete set of copy services functions that provide capabilities for DR, business continuity, data movement, and data duplication solutions.

6.1.1 FlashCopy

FlashCopy is a function that you can use to create a point-in-time copy of one of your volumes. This function is helpful when performing backups or application testing. These copies can be cascaded on one another, read from, written to, and even reversed. These copies can conserve storage (if needed) by being space-efficient copies that record only items that changed from the originals instead of full copies.

6.1.2 Metro Mirror and Global Mirror

MM and GM are technologies that you use to keep a real-time copy of a volume at a remote site that contains another IBM Storage Virtualize system. Consider the following points:

- ▶ MM makes *synchronous* copies of volumes. The write operation on the source storage is not considered complete unless the write operation on the target storage is complete. The distance between two sites is usually determined by the amount of latency that the applications can handle.
- ▶ GM makes *asynchronous* copies of volumes. The write operation is considered complete immediately after write completion on local storage. GM does not wait for a write to complete on target storage like MM does. This requirement greatly reduces the latency that is experienced by applications if the other system is far away. However, it also means that during a failure, the data on the remote copy might not have the most recent changes that are committed to the local volume.

IBM Storage Virtualize provides two types of asynchronous mirroring technology:

- Standard GM
- Global Mirror with Change Volumes (GMCV)

6.1.3 Volume mirroring

Volume mirroring is a function that can increase the high availability (HA) of the storage infrastructure. It can create up to two local copies of a volume. Volume mirroring can use space from two storage pools, and preferably from two separate back-end disk subsystems.

Primarily, you use this function to insulate hosts from the failure of a storage pool and from the failure of a back-end disk subsystem. During a storage pool failure, the system continues to provide service for the volume from the other copy on the other storage pool, with no disruption to the host.

You can also use volume mirroring to change the capacity savings of a volume and migrate data between storage pools of different extent sizes and characteristics.

6.2 IBM FlashCopy

By using the IBM FlashCopy function of IBM Spectrum Virtualize, you can perform a *point-in-time copy* of one or more volumes. This section describes the inner workings of FlashCopy, and provides some best practices for its use.

You can use FlashCopy to solve critical and challenging business needs that require duplication of data of your source volume. Volumes can remain online and active while you create consistent copies of the data sets. Because the copy is performed at the block level, it operates below the host operating system and its cache. Therefore, the copy is not apparent to the host.

Important: Because FlashCopy operates at the block level below the host operating system and cache, those levels do need to be flushed for consistent FlashCopy copies.

While the FlashCopy operation is performed, the source volume is stopped briefly to initialize the FlashCopy bitmap, and then input/output (I/O) can resume. Although several FlashCopy options require the data to be copied from the source to the target in the background, which can take time to complete, the resulting data on the target volume is presented so that the copy appears to complete immediately.

This process is performed by using a bitmap (or bit array) that tracks changes to the data after the FlashCopy is started, and an indirection layer that enables data to be read from the source volume transparently.

6.2.1 FlashCopy use cases

When you are deciding whether FlashCopy addresses your needs, you must adopt a combined business and technical view of the problems that you want to solve. First, determine the needs from a business perspective. Then, determine whether FlashCopy can address the technical needs of those business requirements.

The business applications for FlashCopy are wide-ranging. In the following sections, a short description of the most common use cases is provided.

Backup improvements with FlashCopy

FlashCopy does not reduce the time that it takes to perform a backup to traditional backup infrastructure. However, it can be used to minimize and, under certain conditions, eliminate application downtime that is associated with performing backups.

After the FlashCopy is performed, the resulting image of the data can be backed up to tape as though it were the source system. After the copy to tape is complete, the image data is redundant, and the target volumes can be discarded. For time-limited applications, such as these examples, “no copy” or incremental FlashCopy is used most often. The usage of these methods puts less load on your infrastructure.

When FlashCopy is used for backup purposes, the target data is usually managed as read-only at the operating system level. This approach provides extra security by ensuring that your target data was not modified and remains true to the source.

Restoring with FlashCopy

FlashCopy can perform a restore from any existing FlashCopy mapping. Therefore, you can restore (or copy) from the target to the source of your regular FlashCopy relationships. It might be easier to think of this method as reversing the direction of the FlashCopy mappings. This capability has the following benefits:

- ▶ There is no need to worry about pairing mistakes because you trigger a restore.
- ▶ The process appears instantaneous.
- ▶ You can maintain a pristine image of your data while you are restoring what was the primary data.

This approach can be used for various applications, such as recovering your production database application after an errant batch process that caused extensive damage.

Best practices: Although restoring from FlashCopy is quicker than a traditional tape media restore, do not use restoring from FlashCopy as a substitute for good archiving practices. Instead, keep one to several iterations of your FlashCopy copies so that you can near-instantly recover your data from the most recent history. Keep your long-term archive for your business.

In addition to the restore option, which copies the original blocks from the target volume to modified blocks on the source volume, the target can be used to perform a restore of individual files. To do that task, you must make the target available on a host. Do not make the target available to the source host because seeing duplicates of disks causes problems for most host operating systems. Copy the files to the source by using the normal host data copy methods for your environment.

Moving and migrating data with FlashCopy

FlashCopy can be used to facilitate the movement or migration of data between hosts while minimizing downtime for applications. By using FlashCopy, application data can be copied from source volumes to new target volumes while applications remain online. After the volumes are fully copied and synchronized, the application can be brought down and then immediately brought back up on the new server that is accessing the new FlashCopy target volumes.

Use case: FlashCopy can be used to migrate volumes from and to data reduction pools (DRPs), which do not support extent-based migrations.

This method differs from the other migration methods, which are described later in this chapter. Common uses for this capability are host and back-end storage hardware refreshes.

Application testing with FlashCopy

It is often important to test a new version of an application or operating system that is using actual production data. This testing ensures the highest quality possible for your environment. FlashCopy makes this type of testing easy to accomplish without putting the production data at risk or requiring downtime to create a constant copy.

Create a FlashCopy of your source and use it for your testing. This copy is a duplicate of your production data down to the block level so that even physical disk IDs are copied. Therefore, it is impossible for your applications to tell the difference.

Cyber resiliency

FlashCopy is the foundation of the IBM Storage Virtualize *Safeguarded Copy* function that can create cyber-resilient point-in-time copies of volumes that cannot be changed or deleted through user errors, malicious actions, or ransomware attacks. For more information, see 6.3, “Safeguarded Copy” on page 391.

6.2.2 FlashCopy capabilities overview

FlashCopy occurs between a source volume and a target volume in the same storage system. The minimum granularity that IBM Storage Virtualize supports for FlashCopy is an entire volume. It is not possible to use FlashCopy to copy only part of a volume.

To start a FlashCopy operation, a relationship between the source and the target volume must be defined. This relationship is called *FlashCopy mapping*.

FlashCopy mappings can be stand-alone or a member of a consistency group (CG). You can perform the actions of preparing, starting, or stopping FlashCopy on either a stand-alone mapping or a CG.

Note: Starting from IBM Storage Virtualize 8.4.2, the maximum number of FlashCopy mappings per system is 15 864.^a

a. Applies only to IBM Storwize V7000 Gen3, IBM FlashSystem 7200, 9100, and 9200, and IBM SAN Volume Controller DH8, SV1, SA2, SV2, and SV3.

Table 6-1 on page 368 shows the concept of FlashCopy mapping.

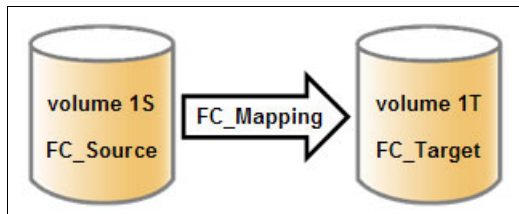


Figure 6-1 FlashCopy mapping

A FlashCopy mapping has a set of attributes and settings that define the characteristics and the capabilities of the FlashCopy.

These characteristics are explained in more detail in the following sections.

Background copy

The *background copy rate* is a property of a FlashCopy mapping that you use to specify whether a background physical copy of the source volume to the corresponding target volume occurs. A value of 0 disables the background copy. If the FlashCopy background copy is disabled, only data that changed on the source volume is copied to the target volume. A FlashCopy with background copy disabled is also known as *no-copy* FlashCopy.

The benefit of using a FlashCopy mapping with background copy enabled is that the target volume becomes a real clone (independent from the source volume) of the FlashCopy mapping source volume after the copy completes. When the background copy function is not performed, the target volume remains a valid copy of the source data while the FlashCopy mapping remains in place.

Valid values for the background copy rate are 0 - 150. The background copy rate can be defined and changed dynamically for individual FlashCopy mappings.

Table 6-1 lists the relationship of the background copy rate value to the attempted amount of data to be copied per second.

Table 6-1 Relationship between the rate and data rate per second

Value	Data copied per second
1 - 10	128 KB
11 - 20	256 KB
21 - 30	512 KB
31 - 40	1 MB
41 - 50	2 MB
51 - 60	4 MB
61 - 70	8 MB
71 - 80	16 MB
81 - 90	32 MB
91 - 100	64 MB
101-110	128 MB
111-120	256 MB
121-130	512 MB
131-140	1024 MB
141-150	2048 MB

Note: To ensure optimal performance of all IBM Storage Virtualize features, it is a best practice to not exceed a copyrate value of 130.

FlashCopy consistency groups

CGs can be used to help create a consistent point-in-time copy across multiple volumes. They are used to manage the consistency of dependent writes that are run in the application following the correct sequence.

When CGs are used, the FlashCopy commands are issued to the CGs. The groups simultaneously perform the operation on all FlashCopy mappings that are contained within the CGs.

Table 6-2 on page 377 shows a CG that consists of two volume mappings.

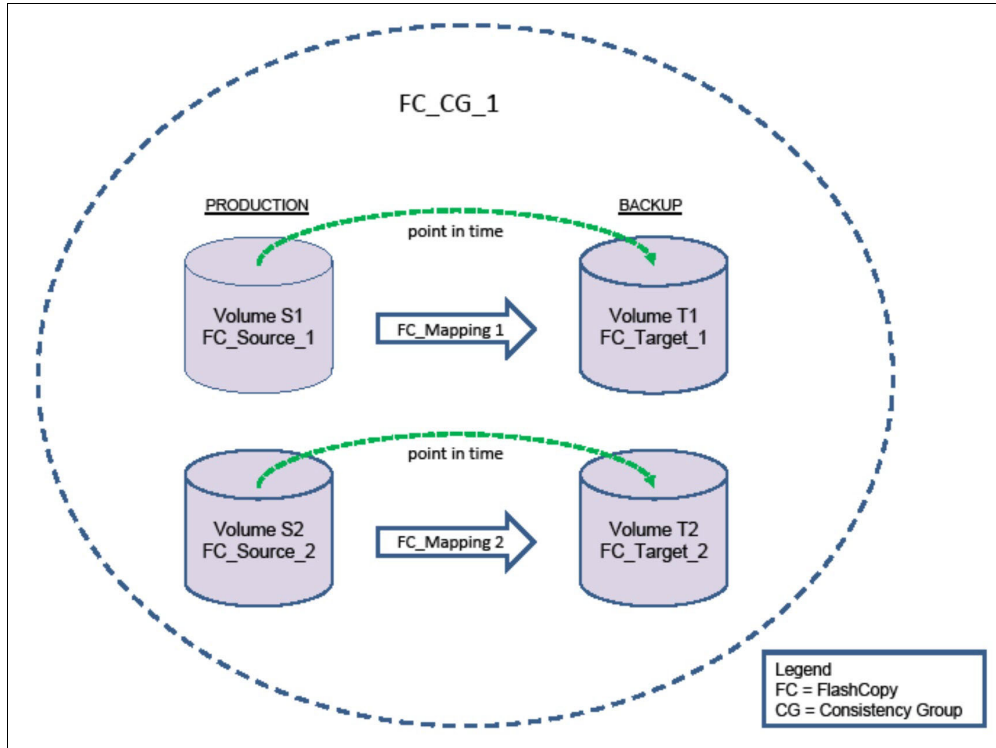


Figure 6-2 Multiple volume mappings in a consistency group

FlashCopy mapping considerations: If the FlashCopy mapping is added to a CG, it can be managed only as part of the group. This limitation means that FlashCopy operations are no longer allowed on the individual FlashCopy mappings.

Incremental FlashCopy

By using Incremental FlashCopy, you can reduce the required time to copy. Also, because less data must be copied, the workload that is put on the system and the back-end storage is reduced.

Incremental FlashCopy does not require that you copy an entire disk source volume whenever the FlashCopy mapping is started. Instead, only the changed regions on source volumes are copied to target volumes, as shown in Table 6-3 on page 381.

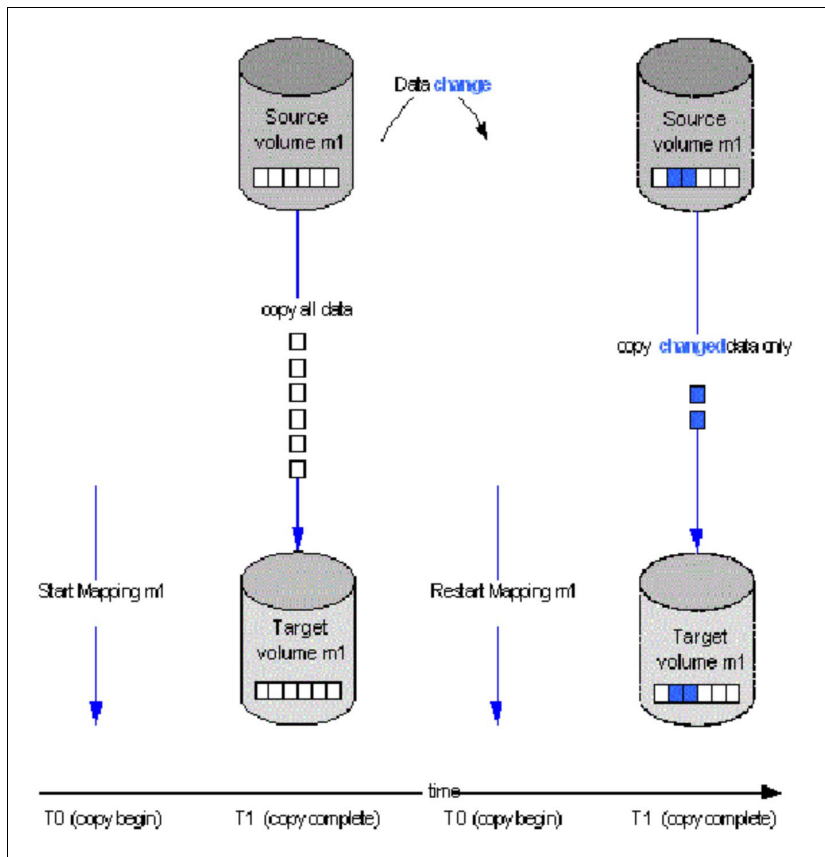


Figure 6-3 Incremental FlashCopy

If the FlashCopy mapping was stopped before the background copy completed, then when the mapping is restarted, the data that was copied before the mapping stopped will not be copied again. For example, if an incremental mapping reaches 10 percent progress when it is stopped and then it is restarted, that 10 percent of data will not be recopied when the mapping is restarted, assuming that it was not changed.

Stopping an incremental FlashCopy mapping: If you are planning to stop an incremental FlashCopy mapping, make sure that the copied data on the source volume will not be changed, if possible. Otherwise, you might have an inconsistent point-in-time copy.

A *difference* value is provided in the query of a mapping, which makes it possible to know how much data changed. This data must be copied when the Incremental FlashCopy mapping is restarted. The difference value is the percentage (0 - 100 percent) of data that changed. This data must be copied to the target volume to get a fully independent copy of the source volume.

An incremental FlashCopy can be defined by setting the *incremental* attribute in the FlashCopy mapping.

Multiple Target FlashCopy

In Multiple Target FlashCopy, a source volume can be used in multiple FlashCopy mappings while the target is a different volume, as shown in Figure 6-4.

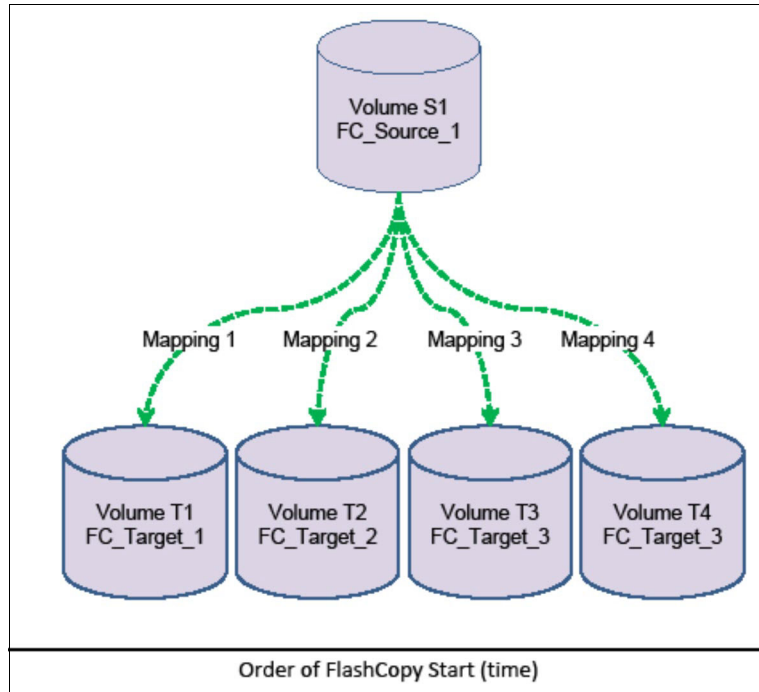


Figure 6-4 Multiple target FlashCopy

Up to 256 different mappings are possible for each source volume. These mappings are independently controllable from each other. Multiple target FlashCopy mappings can be members of the same or different CGs. In cases where all the mappings are in the same CG, the result of starting the CG will be to FlashCopy to multiple identical target volumes.

Cascaded FlashCopy

With Cascaded FlashCopy, you can have a source volume for one FlashCopy mapping and as the target for another FlashCopy mapping, which is referred to as a *Cascaded FlashCopy*. This function is illustrated in Figure 6-5.

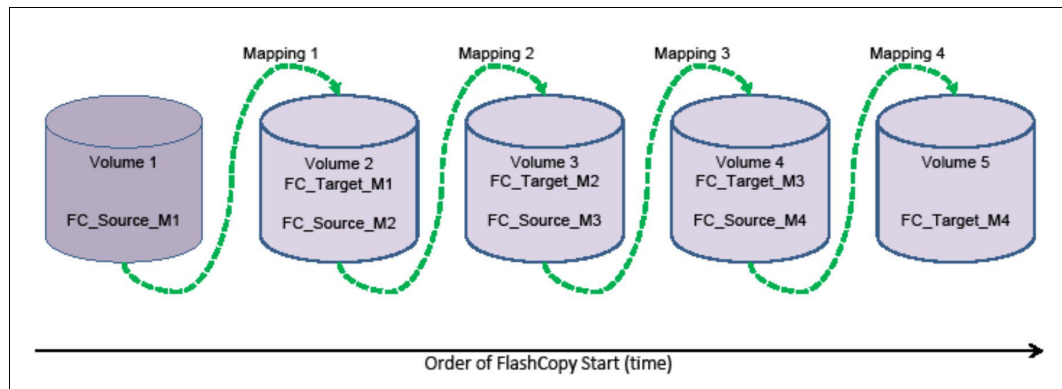


Figure 6-5 Cascaded FlashCopy

A total of 255 mappings are possible for each cascade.

Reverse FlashCopy

Reverse FlashCopy enables FlashCopy targets to become restore points for the source without breaking the FlashCopy relationship, and without having to wait for the original copy operation to complete. It can be used with the Multiple Target FlashCopy to create multiple rollback points.

A key advantage of the Multiple Target reverse FlashCopy function is that the reverse FlashCopy does not destroy the original target. This feature enables processes that are using the target, such as a tape backup, to continue uninterrupted. IBM Storage Virtualize also allows you to create an optional copy of the source volume to be made before the reverse copy operation starts. This ability to restore back to the original source data can be useful for diagnostic purposes.

Thin-provisioned FlashCopy

When a new volume is created, you can designate it as a *thin-provisioned volume*, and it has a virtual capacity and a real capacity.

Virtual capacity is the volume storage capacity that is available to a host. *Real capacity* is the storage capacity that is allocated to a volume copy from a storage pool. In a fully allocated volume, the virtual capacity and real capacity are the same. However, in a thin-provisioned volume, the virtual capacity can be much larger than the real capacity.

The virtual capacity of a thin-provisioned volume is typically larger than its real capacity. On IBM Storage Virtualize systems, the real capacity is used to store data that is written to the volume, and metadata that describes the thin-provisioned configuration of the volume. As more information is written to the volume, more of the real capacity is used.

Thin-provisioned volumes can also help to simplify server administration. Instead of assigning a volume with some capacity to an application and increasing that capacity following the needs of the application if those needs change, you can configure a volume with a large virtual capacity for the application. Then, you can increase or shrink the real capacity as the application's needs change, without disrupting the application or server.

When you configure a thin-provisioned volume, you can use the warning level attribute to generate a warning event when the used real capacity exceeds a specified amount or percentage of the total real capacity. For example, if you have a volume with 10 GB of total capacity and you set the warning to 80 percent, an event is registered in the event log when you use 80 percent of the total capacity. This technique is useful when you need to control how much of the volume is used.

If a thin-provisioned volume does not have enough real capacity for a write operation, the volume is taken offline and an error is logged (error code 1865, event ID 060001). Access to the thin-provisioned volume is restored by either increasing the real capacity of the volume or increasing the size of the storage pool on which it is allocated.

You can use thin volumes for Cascaded FlashCopy and Multiple Target FlashCopy. It is also possible to mix thin-provisioned with normal volumes. Thin-provisioning can be used for incremental FlashCopy too, but using thin-provisioned volumes for incremental FlashCopy makes sense only if the source and target are thin-provisioned.

When using thin-provisioned volumes on DRPs, consider implementing compression because it provides several benefits:

- ▶ Reduced amount of I/O operation to the back end as the amount of data to be written to the back-end reduces with compressed data. This situation is particularly relevant with a poorly performing back end, but less of an issue with the high-performing back end on IBM Spectrum Virtualize.
- ▶ Space efficiency because the compressed data provides more capacity savings.
- ▶ Better back-end capacity monitoring because DRP pools with thin-provisioned uncompressed volumes do not provide physical allocation information.

Therefore, the recommendation is to always enable compression on DRP thin-provisioned volumes.

Thin-provisioned incremental FlashCopy

The implementation of thin-provisioned volumes does not preclude the usage of incremental FlashCopy on the same volumes. It does not make sense to have a fully allocated source volume and then use incremental FlashCopy, which is always a full copy at first, to copy this fully allocated source volume to a thin-provisioned target volume. However, this action is not prohibited.

Consider this optional configuration:

- ▶ A thin-provisioned source volume can be copied incrementally by using FlashCopy to a thin-provisioned target volume. Whenever FlashCopy is performed, only data that is modified is recopied to the target. If space is allocated on the target because of I/O to the target volume, this space will not be reclaimed with subsequent FlashCopy operations.
- ▶ A fully allocated source volume can be copied incrementally by using FlashCopy to another fully allocated volume while it is being copied to multiple thin-provisioned targets (taken at separate points in time). This combination allows a single full backup to be kept for recovery purposes, and separates the backup workload from the production workload. Concurrently, it allows older thin-provisioned backups to be retained.

6.2.3 FlashCopy functional overview

Understanding how FlashCopy works internally helps you to configure it and enables you to obtain more benefits from it.

FlashCopy mapping states

A FlashCopy mapping defines the relationship that copies data between a source volume and a target volume. FlashCopy mappings can be either stand-alone or a member of a CG. You can perform the actions of preparing, starting, or stopping FlashCopy on either a stand-alone mapping or a CG.

A FlashCopy mapping has an attribute that represents the state of the mapping. The FlashCopy states are as follows:

- ▶ Idle_or_copied
- ▶ Copying
- ▶ Stopped
- ▶ Stopping
- ▶ Suspended
- ▶ Preparing
- ▶ Prepared

Idle_or_copied

Read/write caching is enabled for both the source and the target. A FlashCopy mapping exists between the source and target, but the source and target behave as independent volumes in this state.

Copying

The FlashCopy indirection layer (see “Indirection layer” on page 376) governs all I/O to the source and target volumes while the background copy is running. The background copy process is copying *grains* from the source to the target. Reads/writes run on the target as though the contents of the source were instantaneously copied to the target during the running of the `startfcmaporstartfcconsistgrp` command. The source and target can be independently updated. Internally, the target depends on the source for certain tracks. Read/write caching is enabled on the source and the target.

Stopped

The FlashCopy was stopped either by a user command or by an I/O error. When a FlashCopy mapping is stopped, the integrity of the data on the target volume is lost. Therefore, while the FlashCopy mapping is in this state, the target volume is in the Offline state. To regain access to the target, the mapping must be started again (the previous point-in-time is lost) or the FlashCopy mapping must be deleted. The source volume is accessible, and read/write caching is enabled for the source. In the Stopped state, a mapping can either be prepared again or deleted.

Stopping

The mapping is transferring data to a dependent mapping. The behavior of the target volume depends on whether the background copy process completed while the mapping was in the Copying state. If the copy process completed, the target volume remains online while the stopping copy process completes. If the copy process did not complete, data in the cache is discarded for the target volume. The target volume is taken offline, and the stopping copy process runs. After the data is copied, a `stop complete` asynchronous event notification is issued. The mapping moves to the Idle/Copied state if the background copy completed or to the Stopped state if the background copy did not complete. The source volume remains accessible for I/O.

Suspended

The FlashCopy was in the Copying or Stopping state when access to the metadata was lost. As a result, both the source and target volumes are offline and the background copy process halted. When the metadata becomes available again, the FlashCopy mapping returns to the Copying or Stopping state. Access to the source and target volumes is restored, and the background copy or stopping process resumes. Unflushed data that was written to the source or target before the FlashCopy was suspended is pinned in cache until the FlashCopy mapping leaves the Suspended state.

Preparing

The FlashCopy is preparing the mapping. While in this state, data from cache is destaged to disk and a consistent copy of the source exists on disk. Now, cache is operating in write-through mode and writes to the source volume experience more latency. The target volume is reported as online, but it does not perform reads or writes. These reads/writes are failed by the Small Computer System Interface (SCSI) front end.

Before starting the FlashCopy mapping, it is important that any cache at the host level, for example, buffers on the host operating system or application, are instructed to flush any outstanding writes to the source volume. Performing the cache flush that is required as part of the `startfcmap` or `startfcconsistgrp` command causes I/Os to be delayed while waiting for the cache flush to complete. To overcome this problem, FlashCopy supports the `prestartfcmap` or `prestartfcconsistgrp` commands. These commands prepare for a FlashCopy start while still allowing I/Os to continue to the source volume.

In the Preparing state, the FlashCopy mapping is prepared by the following steps:

1. Flush any modified write data that is associated with the source volume from the cache. Read data for the source is left in the cache.
2. Place the cache for the source volume into write-through mode so that subsequent writes wait until data is written to disk before completing the write command that is received from the host.
3. Discard any read or write data that is associated with the target volume from the cache.

Prepared

While in the Prepared state, the FlashCopy mapping is ready to perform a start. While the FlashCopy mapping is in this state, the target volume is in the Offline state. In the Prepared state, writes to the source volume experience more latency because the cache is operating in write-through mode.

Figure 6-6 represent the FlashCopy mapping state diagram. It illustrates the states in which a mapping can exist, and which events are responsible for a state change.

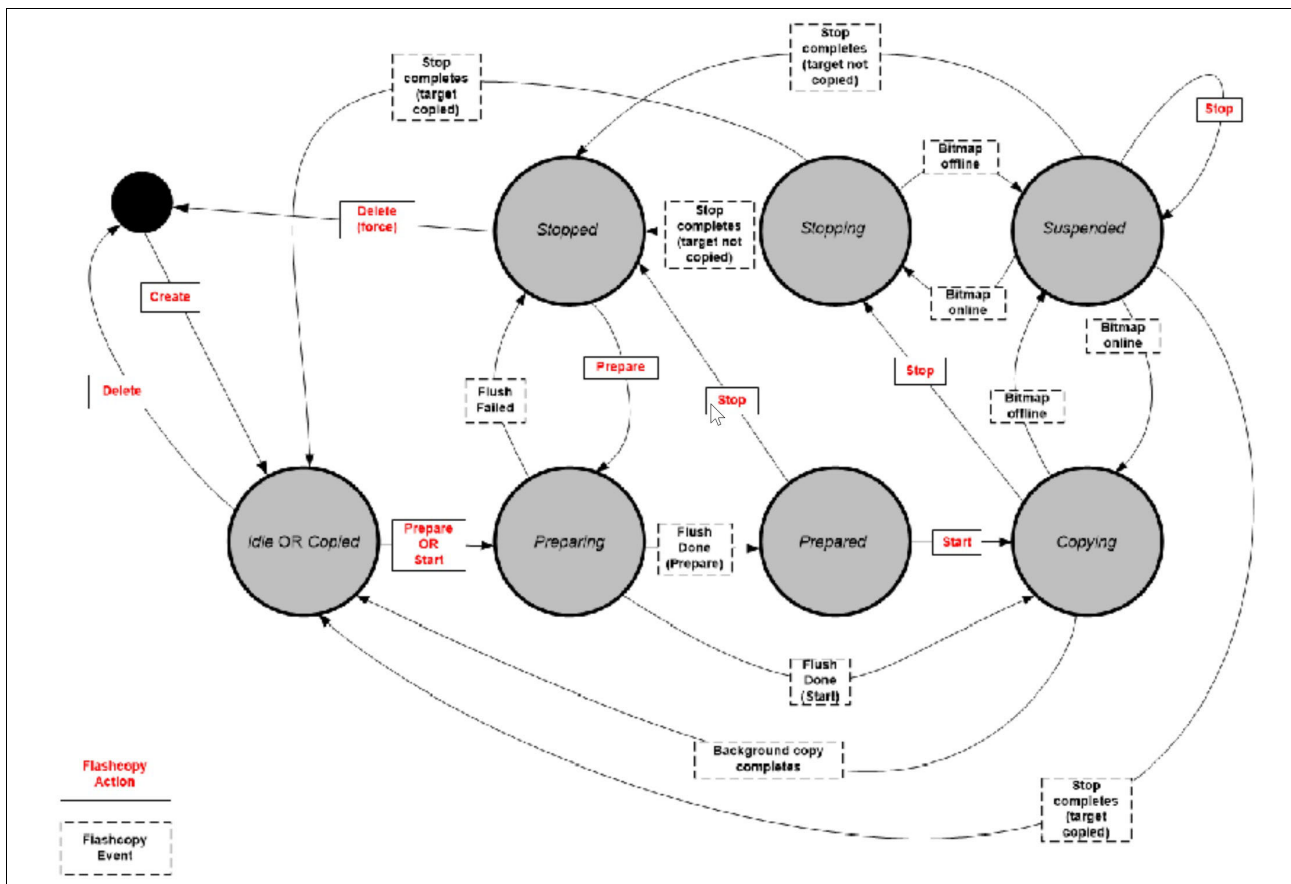


Figure 6-6 FlashCopy mapping states diagram

FlashCopy bitmaps and grains

A *bitmap* is an internal data structure that is stored in a particular I/O group that is used to track which data in FlashCopy mappings was copied from the source volume to the target volume. *Grains* are units of data that are grouped to optimize the usage of the bitmap. One bit in each bitmap represents the state of one grain. A FlashCopy grain can be either 64 KB or 256 KB.

A FlashCopy bitmap takes up the bitmap space in the memory of the I/O group that must be shared with bitmaps of other features (such as remote copy bitmaps, volume mirroring bitmaps, and redundant array of independent disks (RAID) bitmaps).

Indirection layer

The *FlashCopy indirection layer* governs the I/O to the source and target volumes when a FlashCopy mapping is started. This process is done by using a FlashCopy bitmap. The purpose of the FlashCopy indirection layer is to enable both the source and target volumes for read/write I/O immediately after FlashCopy starts.

The following description illustrates how the FlashCopy indirection layer works when a FlashCopy mapping is prepared and then started.

When a FlashCopy mapping is prepared and started, the following sequence is applied:

1. Flush the write cache to the source volume or volumes that are part of a CG.
2. Put the cache into write-through mode on the source volumes.
3. Discard the cache for the target volumes.
4. Establish a sync point on all of the source volumes in the CG (creating the FlashCopy bitmap).
5. Ensure that the indirection layer governs all the I/O to the source volumes and target.
6. Enable the cache on source volumes and target volumes.

FlashCopy provides the semantics of a point-in-time copy that uses the indirection layer, which intercepts I/O that is directed at either the source or target volumes. The act of starting a FlashCopy mapping causes this indirection layer to become active in the I/O path, which occurs automatically across all FlashCopy mappings in the CG. Then, the indirection layer determines how each I/O is routed based on the following factors:

- ▶ The volume and the logical block address (LBA) to which the I/O is addressed.
- ▶ Its direction (read or write).

The indirection layer allows an I/O to go through the underlying volume, which preserves the point-in-time copy. To do this, the IBM Storage Virtualize code uses two mechanisms:

- ▶ **Copy-on-write (CoW):** With this mechanism, when a write operation occurs in the source volume, a portion of data (grain) containing the data to be modified is copied to the target volume before the operation completion.
- ▶ **Redirect-on-write (RoW):** With this mechanism, when a write operation occurs in the source volume, the data to be modified is written in another area, leaving the original data unmodified to be used by the target volume.

IBM Storage Virtualize implements CoW and RoW logics transparently to the user with the aim to optimize the performance and capacity. By using the RoW mechanism, the performance can improve by reducing the number of physical I/Os for the write operations while a significant capacity savings can be achieved by improving the overall deduplication ratio.

The RoW was introduced with IBM Storage Virtualize 8.4 and is used in the following conditions:

- ▶ Source and target volumes in the same pool.
- ▶ Source and target volumes in the same I/O group.
- ▶ The pool that contains the source and target volumes must be a DRP.
- ▶ Source and target volumes do not participate in a volume mirroring relationship.
- ▶ Source and target volumes are not fully allocated.

In all the cases in which the RoW is not applicable, the CoW is used.

Table 6-2 lists the indirection layer algorithm for CoW.

Table 6-2 Summary table of the FlashCopy indirection layer algorithm

Volume being accessed	Has the grain been copied?	Host I/O operation	
		Read	Write
Source	No	Read from the source volume.	Copy the grain to the most recently started target for this source, and then write to the source.
	Yes	Read from the source volume.	Write to the source volume.
Target	No	If any newer targets exist for this source in which this grain already was copied, read from the oldest of these targets. Otherwise, read from the source.	Hold the write. Check the dependency target volumes to see whether the grain was copied. If the grain is not already copied to the next oldest target for this source, copy the grain to the next oldest target. Then, write to the target.
	Yes	Read from the target volume.	Write to the target volume.

Interacting with cache

IBM Storage Virtualize provides a two-layer cache, as follows:

- ▶ The *upper cache* serves mostly as write cache and hides the write latency from the hosts and application.
- ▶ The *lower cache* is a read/write cache and optimizes I/O to and from disks.

Figure 6-7 shows the IBM Storage Virtualize cache architecture.

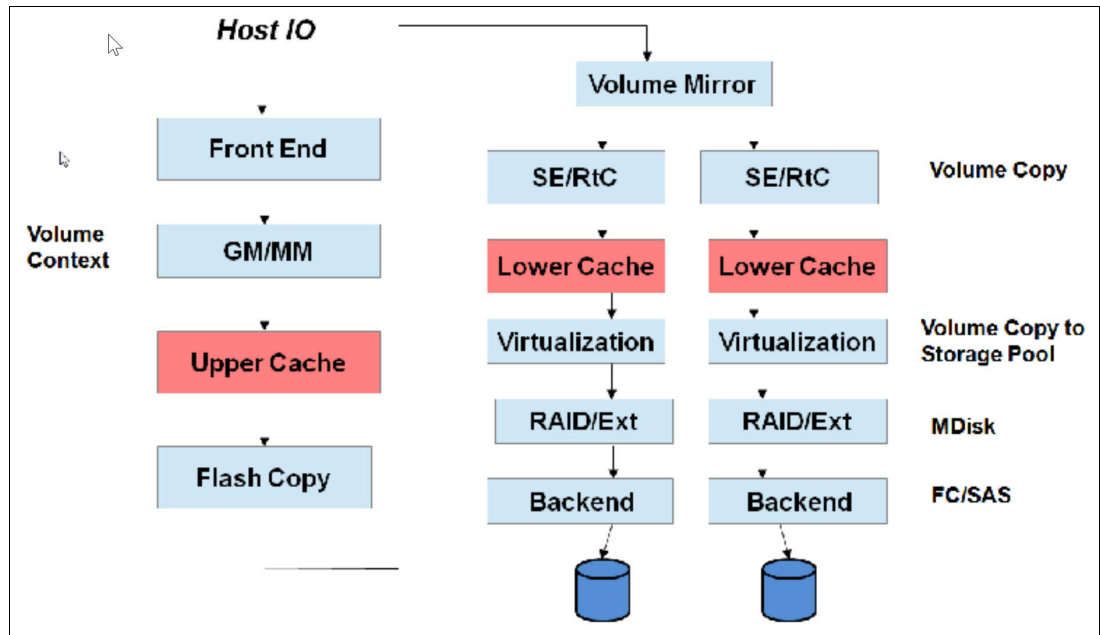


Figure 6-7 Cache architecture

The CoW process might introduce significant latency into write operations. To isolate the active application from this additional latency, the FlashCopy indirection layer is placed logically between the upper and lower cache. Therefore, the additional latency that is introduced by the CoW process is encountered only by the internal cache operations, and not by the application.

The logical placement of the FlashCopy indirection layer is shown in Figure 6-8.

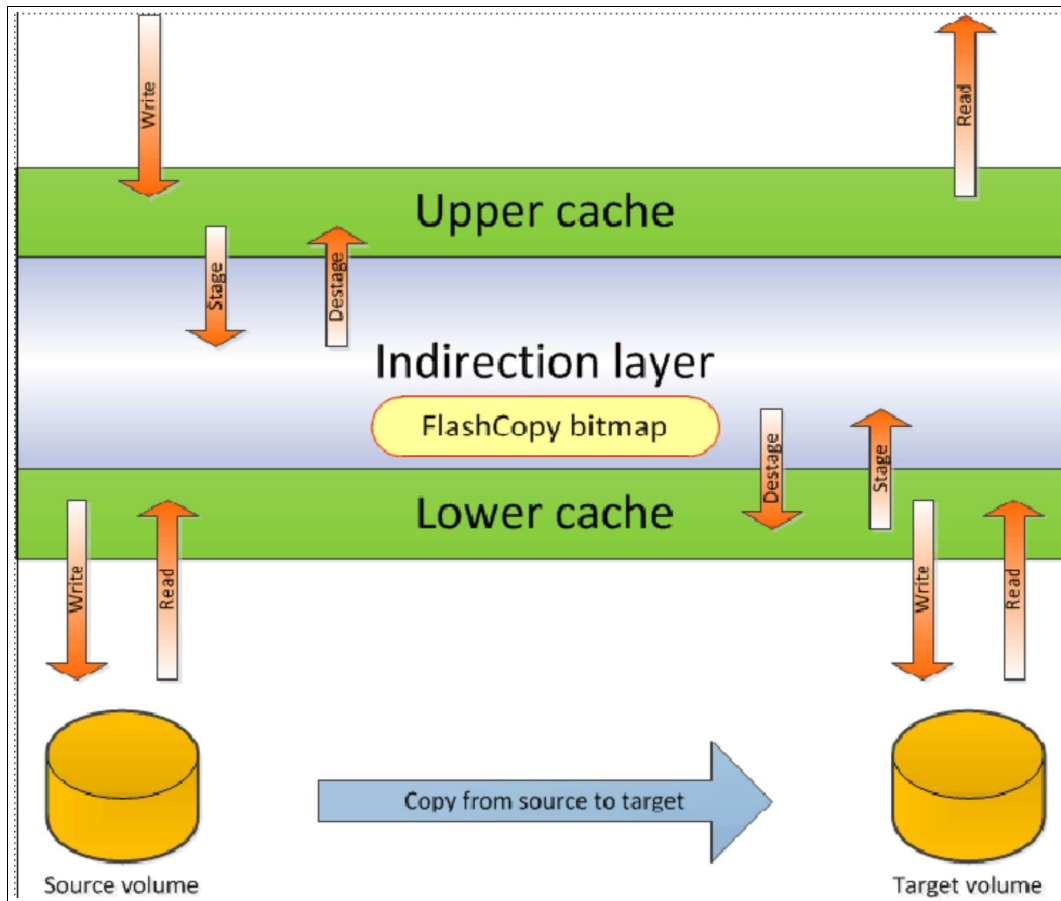


Figure 6-8 Logical placement of the FlashCopy indirection layer

The two-level cache architecture provides performance benefits to the FlashCopy mechanism. Because the FlashCopy layer is above the lower cache in the IBM Storage Virtualize software stack, it can benefit from read prefetching and coalescing writes to back-end storage.

Also, preparing FlashCopy is fast because the upper cache write data does not have to go directly to back-end storage, but to the lower cache layer only.

Interaction and dependency between multiple Target FlashCopy mappings

Figure 6-9 shows a set of three FlashCopy mappings that share a common source. The FlashCopy mappings target target volumes Target 1, Target 2, and Target 3.

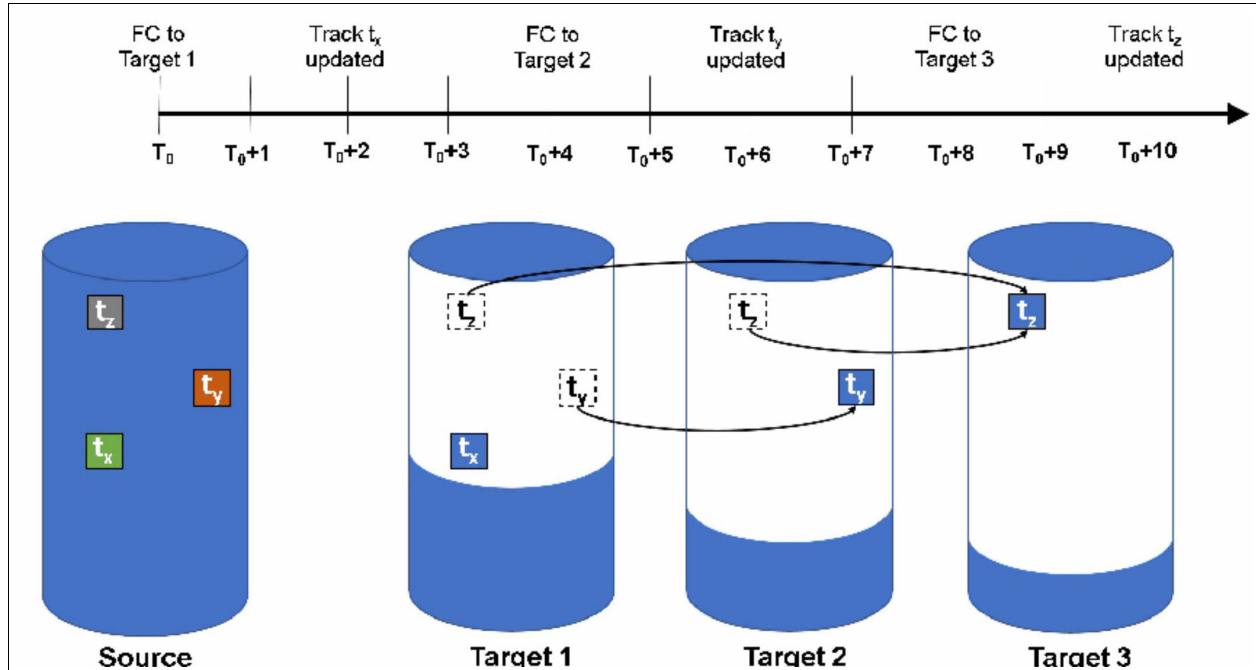


Figure 6-9 Interaction between multiple Target FlashCopy mappings

Consider the following events timeline:

- ▶ At time T_0 , a FlashCopy mapping is started between the source and Target 1.
- ▶ At time T_0+2 , track t_x is updated in the source. Because this track has not yet been copied in the background on Target 1, the CoW process copies this track to Target 1 before being updated on the source.
- ▶ At time T_0+4 , a FlashCopy mapping starts between the source and Target 2.
- ▶ At time T_0+6 , track t_y is updated in the source. Because this track has not yet been copied in the background on Target 2, the CoW process copies this track to Target 2 only before being updated on the source.
- ▶ At time T_0+8 , a FlashCopy mapping starts between the source and Target 3.
- ▶ At time T_0+10 , track t_z is updated in the source. Because this track has not yet been copied in the background on Target 3, the CoW process copies this track to Target 3 only before being updated on the source.

As a result of this sequence of events, the configuration in Figure 6-9 has the following characteristics:

- ▶ Target 1 depends on Target 2 and Target 3. It remains dependent until all of Target 1 is copied. No target depends on Target 1, so the mapping can be stopped without needing to copy any data to maintain the consistency in the other targets.
- ▶ Target 2 depends on Target 3, and remains dependent until all of Target 2 is copied. Target 1 depends on Target 2, so if this mapping is stopped, the cleanup process is started to copy all data that is uniquely held on this mapping (that is t_y) to Target 1.

- ▶ Target 3 does not depend on any target, but it has Target 1 and Target 2 depending on it, so if this mapping is stopped, the cleanup process is started to copy all data that is uniquely held on this mapping (that is t_z) to Target 2.

Target writes with Multiple Target FlashCopy

A write to an intermediate or the newest target volume must consider the state of the grain within its own mapping and the state of the grain of the next oldest mapping:

- ▶ If the grain of the next oldest mapping has not been copied yet, it must be copied before the write is allowed to proceed to preserve the contents of the next oldest mapping. The data that is written to the next oldest mapping comes from a target or source.
- ▶ If the grain in the target being written has not been copied yet, the grain is copied from the oldest already copied grain in the mappings that are newer than the target, or the source if none are already copied. After this copy is done, the write can be applied to the target.

Target reads with Multiple Target FlashCopy

If the grain that is being read already was copied from the source to the target, the read simply returns data from the target being read. If the grain was not copied, each of the newer mappings is examined in turn, and the read is performed from the first copy that is found. If none are found, the read is performed from the source.

6.2.4 FlashCopy planning considerations

The FlashCopy function, like all the advanced IBM Storage Virtualize features, offers useful capabilities. However, some basic planning considerations should be followed for a successful implementation.

FlashCopy configurations limits

To plan for and implement FlashCopy, you must check the configuration limits and adhere to them. Table 6-3 lists the system limits that apply to the latest version as of the time of writing.

Table 6-3 *FlashCopy properties and maximum configurations*

FlashCopy property	Maximum	Comment
FlashCopy targets per source	256	This maximum is the maximum number of FlashCopy mappings that can exist with the same source volume.
FlashCopy mappings per system	15864	This maximum is the maximum number of FlashCopy mappings per system.
FlashCopy CGs per system	500	This maximum is an arbitrary limit that is policed by the software.
FlashCopy volume space per I/O group	4096 TB	This maximum is a limit on the quantity of FlashCopy mappings by using bitmap space from one I/O group.
FlashCopy mappings per CG	512	This limit is due to the time that is taken to prepare a CG with many mappings.

Configuration limits: The configuration limits always change with the introduction of new hardware and software capabilities. For more information, see [V8.4.0.x Configuration Limits and Restrictions for IBM FlashSystem 9200](#).

The total amount of cache memory that is reserved for the FlashCopy bitmaps limits the amount of capacity that can be used as a FlashCopy target. Table 6-4 shows the relationship of bitmap space to FlashCopy address space, depending on the size of the grain and the kind of FlashCopy service being used.

Table 6-4 Relationship of bitmap space to FlashCopy address space for the I/O group

Copy service	Grain size (KB)	1 MB of memory provides the following volume capacity for the specified I/O group
FlashCopy	256	2 TB of target volume capacity
FlashCopy	64	512 GB of target volume capacity
Incremental FlashCopy	256	1 TB of target volume capacity
Incremental FlashCopy	64	256 GB of target volume capacity

Mapping consideration: For multiple FlashCopy targets, you must consider the number of mappings. For example, for a mapping with a 256 KB grain size, 8 KB of memory allows one mapping between a 16 GB source volume and a 16 GB target volume. Alternatively, the same amount of memory and the same grainsize allows two mappings, between one 8 GB source volume and two 8 GB target volumes. You need to consider the total target capacity when doing bitmap calculations, not the source.

When you create a FlashCopy mapping, if you specify an I/O group other than the I/O group of the source volume, the memory accounting goes toward the specified I/O group, not toward the I/O group of the source volume.

The default amount of memory for FlashCopy is 20 MB. This value can be increased or decreased by using the **chiogrp** command or through the GUI. The maximum amount of memory that can be specified for FlashCopy is 2048 MB (512 MB for 32-bit systems). The maximum combined amount of memory across all copy services features is 2600 MB (552 MB for 32-bit systems).

Bitmap allocation: When creating a FlashCopy mapping, you can optionally specify the I/O group where the bitmap is allocated. If you specify an I/O group other than the I/O group of the source volume, the memory accounting goes toward the specified I/O group, not toward the I/O group of the source volume. This option can be useful when an I/O group is exhausting the memory that is allocated to the FlashCopy bitmaps and no more free memory is available in the I/O group.

FlashCopy general restrictions

The following implementation restrictions apply to FlashCopy:

- ▶ The size of source and target volumes must be the same when creating a FlashCopy mapping.
- ▶ Multiple FlashCopy mappings that use the same target volume can be defined, but only one of these mappings can be started at a time. This limitation means that multiple FlashCopy mappings cannot be active to the same target volume.

- ▶ The following restrictions apply when expanding or shrinking volumes that are defined in a FlashCopy mapping:
 - Target volumes cannot be shrunk.
 - A source volume can be shrunk, but only to the largest starting size of a target volume (in a multiple target or cascading mappings) when in the copying or stopping state.
 - Source and target volumes must be same size when the mapping is prepared or started.
 - Source and target volumes can be expanded in any order except for incremental FlashCopy, where the target volume must be expanded before the source volume can be expanded.

Note: Expanding or shrinking volumes that are participating in a FlashCopy map is allowed with code level 8.4.2 or later.

- ▶ In a cascading FlashCopy, the grain size of all the FlashCopy mappings that participate must be the same.
- ▶ In a multi-target FlashCopy, the grain size of all the FlashCopy mappings that participate must be the same.
- ▶ In a reverse FlashCopy, the grain size of all the FlashCopy mappings that participate must be the same.
- ▶ No FlashCopy mapping can be added to a CG while the FlashCopy mapping status is Copying.
- ▶ No FlashCopy mapping can be added to a CG while the CG status is Copying.
- ▶ Using CGs is restricted when using cascading FlashCopy. A CG starts FlashCopy mappings at the same point in time. Within the *same* CG, it is not possible to have mappings with these conditions:
 - The source volume of one mapping is the target of another mapping.
 - The target volume of one mapping is the source volume for another mapping.

These combinations are not useful because within a CG, mappings cannot be established in a certain order. This limitation renders the content of the target volume undefined. For example, it is not possible to determine whether the first mapping was established before the target volume of the first mapping that acts as a source volume for the second mapping.

Even if it were possible to ensure the order in which the mappings are established within a CG, the result is equal to multiple target FlashCopy (two volumes holding the same target data for one source volume). In other words, a cascade is useful for copying volumes in a certain order (and copying the changed content targets of FlashCopy copies) rather than at the same time in an undefined order (from within one single CG).

- ▶ Source and target volumes can be used as primary in a remote copy relationship. For more information about the FlashCopy and the remote copy possible interactions, see “Interaction between remote copy and FlashCopy” on page 432.

FlashCopy presets

The IBM Storage Virtualize GUI provides three FlashCopy presets (Snapshot, Clone, and Backup) to simplify the more common FlashCopy operations. Figure 6-10 shows the preset selection window in the GUI.

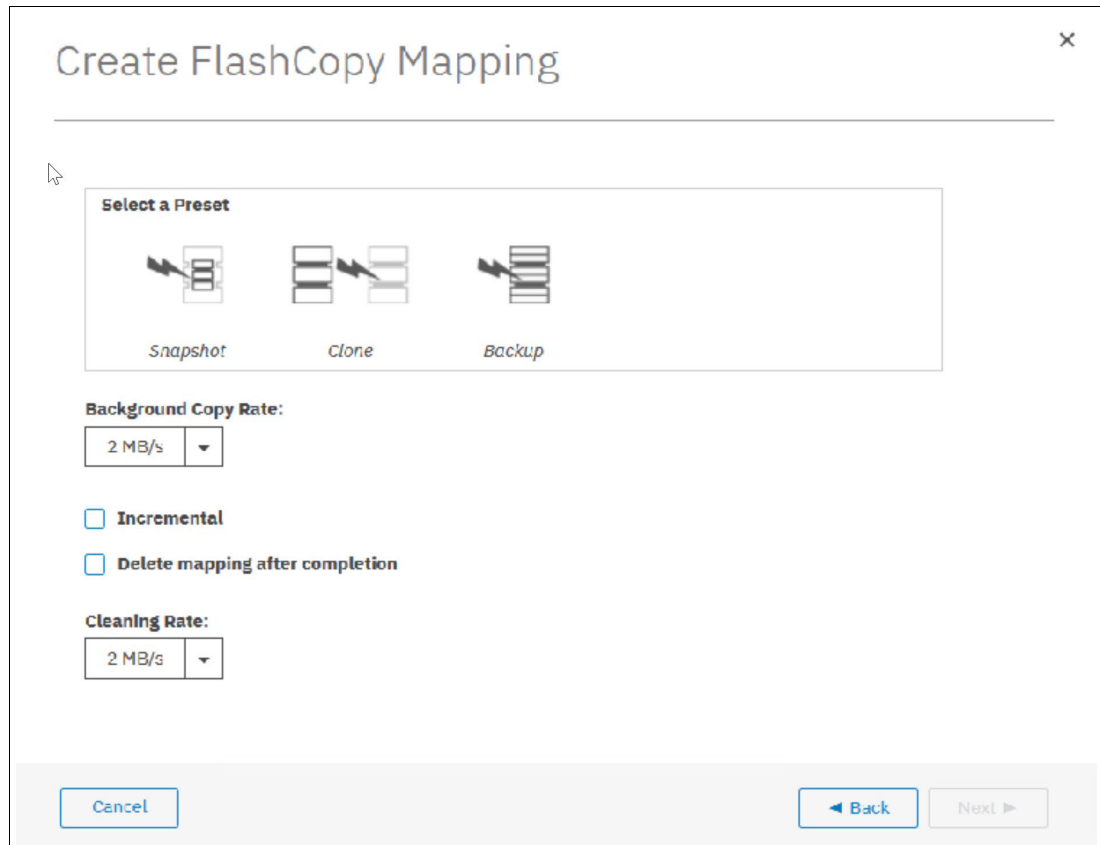


Figure 6-10 IBM Storage Virtualize presets

Although these presets meet most FlashCopy requirements, they do not support all possible FlashCopy options. If more specialized options are required that are not supported by the presets, the options must be performed by using command-line interface (CLI) commands.

This section describes the three preset options and their use cases.

Snapshot

This preset creates a CoW or RoW point-in-time copy. For more information about RoW prerequisites, see “Indirection layer” on page 376.

The snapshot is not intended to be an independent copy. Instead, the copy is used to maintain a view of the production data at the time that the snapshot is created. Therefore, the snapshot holds only the data from regions of the production volume that have changed since the snapshot was created. Because the snapshot preset uses thin-provisioning, only the capacity that is required for the changes is used.

Snapshot uses the following preset parameters:

- ▶ Background copy: None.
- ▶ Incremental: No.
- ▶ Delete after completion: No.

- ▶ Cleaning rate: No.
- ▶ The target pool is the primary copy source pool.

A typical use case for the snapshot is when the user wants to produce a copy of a volume without affecting the availability of the volume. The user does not anticipate many changes to be made to the source or target volume. A significant proportion of the volumes remains unchanged.

By ensuring that only changes require a copy of data to be made, the total amount of disk space that is required for the copy is reduced. Therefore, many snapshot copies can be used in the environment.

Snapshots are useful for providing protection against corruption or similar issues with the validity of the data. However, they do not provide protection from physical controller failures. Snapshots can also provide a vehicle for performing repeatable testing (including “what-if” modeling that is based on production data) without requiring a full copy of the data to be provisioned.

Clone

The clone preset creates a replica of the volume, which can then be changed without affecting the original volume. After the copy completes, the mapping that was created by the preset is automatically deleted.

Clone uses the following preset parameters:

- ▶ Background copy rate: 50.
- ▶ Incremental: No.
- ▶ Delete after completion: Yes.
- ▶ Cleaning rate: 50.
- ▶ The target pool is the primary copy source pool.

A typical use case for the snapshot is when users want a copy of the volume that they can modify without affecting the original volume. After the clone is established, there is no expectation that it is refreshed, or that there is any further need to reference the original production data again. If the source is thin-provisioned, the target is thin-provisioned for the auto-create target.

Backup

The backup preset creates a point-in-time replica of the production data. After the copy completes, the backup view can be refreshed from the production data, with minimal copying of data from the production volume to the backup volume.

Backup uses the following preset parameters:

- ▶ Background Copy rate: 50.
- ▶ Incremental: Yes.
- ▶ Delete after completion: No.
- ▶ Cleaning rate: 50.
- ▶ The target pool is the primary copy source pool.

The backup preset can be used when the user wants to create a copy of the volume that can be used as a backup if the source becomes unavailable. This unavailability can happen during loss of the underlying physical controller. The user plans to periodically update the secondary copy, and does not want to suffer from the resource demands of creating a copy each time.

Incremental FlashCopy times are faster than full copy, which helps to reduce the window where the new backup is not yet fully effective. If the source is thin-provisioned, the target is also thin-provisioned in this option for the auto-create target.

Another use case, which is not supported by the name, is to create and maintain (periodically refresh) an independent image. This image can be subjected to intensive I/O (for example, data mining) without affecting the source volume's performance.

Thin-provisioning considerations

When creating FlashCopy with thin-provisioned target volumes, the no-copy option is often used. The real size of a thin-provisioned volume is an attribute that defines how much physical capacity is reserved for the volume. The real size can vary 0 - 100% of the virtual capacity.

When thin-provisioned volumes are used as FlashCopy targets, it is important to provide a nonzero real size. This size is required because when the FlashCopy is started, the CoW process requires that you allocate capacity on the target volumes. If some capacity is not yet allocated, the write I/O can be delayed until the capacity is made available (as with thin-provisioned volumes with zero real size). Often, the write caching hides this effect, but in heavy write workloads, the performance might be affected.

Sizing consideration

When thin-provisioned FlashCopy is used, an estimation of the physical capacity consumption is required. Consider that while a FlashCopy is active, the thin-provisioned target volume allocates physical capacity whenever a grain is modified for the first time on source or target volume.

The following factors must be considered that so that an accurate sizing can be completed:

- ▶ The FlashCopy duration in terms of seconds (D).
- ▶ The write operation per second (W).
- ▶ The grain size in terms of KB (G).
- ▶ The rewrite factor. This factor represents the average chance that a write operation reoccurs in the same grain (R) in percentage.

Although the first three factors are easy to assess, the rewrite factor can only be roughly estimated because it depends on the workload type and the FlashCopy duration. The used capacity (CC) of a thin-provisioned target volume of C size while the FlashCopy is active can be estimated by using the following equation:

$$CC = \min\{(W - W \times R) \times G \times D, C\}$$

For example, consider a 100 GB volume that has FlashCopy active for 3 hours (10.800 seconds) with a grain size of 64 K. Consider also a write workload of 100 input/output operations per second (IOPS) with a rewrite factor of 85% (85% of writes occur on the same grains). In this case, the estimation of the used capacity is:

$$CC = (100 - 85) \times 64 \times 10.800 = 10.368.000 \text{ KB} = 9,88 \text{ GB}$$

Important: Consider the following points:

- ▶ The recommendation with thin-provisioned target volumes is to assign at least 2 GB of real capacity.
- ▶ Thin-provisioned FlashCopy can greatly benefit from the RoW capability that was introduced with IBM Storage Virtualize 8.4. For more information, see "Indirection layer" on page 376.

Grain size considerations

When you create a mapping, a grain size of 64 KB can be specified instead of the default 256 KB. This smaller grain size was introduced specifically for incremental FlashCopy, even though its usage is not restricted to incremental mappings.

In incremental FlashCopy, the modified data is identified by using the bitmaps. The amount of data to be copied when refreshing the mapping depends on the grain size. If the grain size is 64 KB, as compared to 256 KB, there might be less data to copy to get a fully independent copy of the source again.

Here are the preferred settings for thin-provisioned FlashCopy:

- ▶ The thin-provisioned volume grain size should be equal to the FlashCopy grain size. If the 256 KB thin-provisioned volume grain size is chosen, it is still beneficial to limit the FlashCopy grain size to 64 KB. It is possible to minimize the performance impact to the source volume, even though this size increases the I/O workload on the target volume.
- ▶ The thin-provisioned volume grain size must be 64 KB for the best performance and the best space efficiency.

The exception is where the thin-provisioned target volume is going to become a production volume (and likely to be subjected to ongoing heavy I/O). In this case, the 256 KB thin-provisioned grain size is preferable because it provides better long-term I/O performance at the expense of a slower initial copy.

FlashCopy limitation: Configurations with many FlashCopy or remote copy relationships might be forced to choose a 256 KB grain size for FlashCopy to avoid constraints on the amount of bitmap memory.

Cascading FlashCopy and Multiple Target FlashCopy require that all the mappings that are participating in the FlashCopy chain feature the same grain size. For more information, see “FlashCopy general restrictions” on page 382.

Volume placement considerations

The source and target volumes placement among the pools and the I/O groups must be planned to minimize the effect of the underlying FlashCopy processes. In normal conditions (that is, with all the canisters fully operative), the FlashCopy background copy workload distribution follows this schema:

- ▶ The preferred node of the *source* volume is responsible for the background copy *read* operations.
- ▶ The preferred node of the *target* volume is responsible for the background copy *write* operations.

Table 6-5 shows how the back-end I/O operations are distributed across the nodes.

Table 6-5 Workload distribution for back-end I/O operations

Node	Read from source	Read from target	Write to source	Write to target
Node that performs the back-end I/O if the grain is copied.	Preferred node in the source volume's I/O group.	Preferred node in the target volume's I/O group.	Preferred node in the source volume's I/O group.	Preferred node in the target volume's I/O group.
Node that performs the back-end I/O if the grain is not yet copied.	Preferred node in the source volume's I/O group.	Preferred node in the source volume's I/O group.	The preferred node in the source volume's I/O group will read/write, and the preferred node in target volume's I/O group will write.	The preferred node in the source volume's I/O group will read, and the preferred node in target volume's I/O group will write.

The data transfer among the source and the target volume's preferred nodes occurs through the node-to-node connectivity. Consider the following volume placement alternatives:

- ▶ Source and target volumes use the same preferred node.
In this scenario, the node that is acting as preferred node for the source and target volumes manages all the read/write FlashCopy operations. Only resources from this node are used for the FlashCopy operations, and no node-to-node bandwidth is used.
- ▶ Source and target volumes use the different preferred node.
In this scenario, both nodes that are acting as preferred nodes manage read/write FlashCopy operations according to the previously described scenarios. The data that is transferred between the two preferred nodes goes through the node-to-node network.

Both alternatives that are described have advantages and disadvantages, but in general option 1 (source and target volumes use the same preferred node) is preferred. Consider the following exceptions:

- ▶ A clustered IBM FlashSystem system with multiple I/O groups in HyperSwap, where the source volumes are evenly spread across all the nodes.
In this case, the preferred node placement should follow the location on site B, and then the target volumes preferred node must be in site B. Placing the target volumes preferred node in site A causes the redirection of the FlashCopy write operation through the node-to-node network.
- ▶ A clustered IBM FlashSystem system with multiple control enclosures, where the source volumes are evenly spread across all the canisters.
In this case, the preferred node placement should follow the location of source and target volumes on the internal storage. For example, if the source volume is on the internal storage that is attached to control enclosure A and the target volume is on internal storage that is attached to control enclosure B, then the target volumes preferred node must be in one canister of control enclosure B. Placing the target volumes preferred node on control enclosure A causes the redirection of the FlashCopy write operation through the node-to-node network.

Placement on the back-end storage is mainly driven by the availability requirements. Generally, use different back-end storage controllers or arrays for the source and target volumes.

DRP-optimized snapshots: To use the RoW capability that was introduced with IBM Storage Virtualize 8.4, check the volume placement restrictions that are described in “Indirection layer” on page 376.

Background copy considerations

The background copy process uses internal resources, such as CPU, memory, and bandwidth. This copy process tries to reach the target copy data rate for every volume according to the background copy rate parameter setting (see Table 6-1 on page 368).

If the copy process cannot achieve these goals, it starts contending resources that go to the foreground I/O (that is, the I/O that is coming from the hosts). As a result, both background copy and foreground I/O tend to see an increase in latency and a reduction in throughput compared to the situation where the bandwidth is not limited. Degradation is graceful. Both background copy and foreground I/O continue to progress, and do not stop, hang, or cause the node to fail.

To avoid any impact on the foreground I/O, that is, in the hosts' response time, carefully plan the background copy activity by accounting for the overall workload running in the systems. The background copy basically reads/writes data to managed disks (MDisks). Usually, the most affected component is the back-end storage. CPU and memory are not normally significantly affected by the copy activity.

The theoretical added workload due to the background copy is easily estimable. For example, starting 20 FlashCopy copies, each with a background copy rate of 70, adds a maximum throughput of 160 MBps for the reads and 160 MBps for the writes.

The source and target volumes distribution on the back-end storage determines where this workload is going to be added. The duration of the background copy depends on the amount of data to be copied. This amount is the total size of volumes for full background copy or the amount of data that is modified for incremental copy refresh.

Performance monitoring tools like IBM Spectrum Control can be used to evaluate the existing workload on the back-end storage in a specific time window. By adding this workload to the foreseen background copy workload, you can estimate the overall workload running toward the back-end storage. Disk performance simulation tools, like Disk Magic or IBM Storage Modeller (StorM), can be used to estimate the effect, if any, of the added back-end workload to the host service time during the background copy window. The outcomes of this analysis can provide useful hints for the background copy rate settings.

When performance monitoring and simulation tools are not available, use a conservative and progressive approach. Consider that the background copy setting can be modified at any time, even when the FlashCopy already started. The background copy process can be stopped by setting the background copy rate to 0.

Initially set the background copy rate value to add a limited workload to the back end (for example, less than 100 MBps). If no effects on hosts are noticed, the background copy rate value can be increased. Do this process until you see negative effects. The background copy rate setting follows an exponential scale, so changing, for example, from 50 to 60 doubles the data rate goal from 2 MBps to 4 MBps.

Cleaning process and cleaning rate

The cleaning rate is the rate at which the data is copied among dependent FlashCopy copies, such as Cascaded and Multiple Target FlashCopy. The cleaning process aims to release the dependency of a mapping in such a way that it can be stopped immediately (without going to the stopping state). The typical use case for setting the cleaning rate is when it is required to stop a Cascaded or Multiple Target FlashCopy that is not the oldest in the FlashCopy chain. To avoid the stopping state lasting for a long time in this case, the cleaning rate can be adjusted.

An interaction occurs between the background copy rate and the cleaning rate settings:

- ▶ Background copy = 0 and cleaning rate = 0
No background copy or cleaning take place. When the mapping is stopped, it goes into the stopping state and a cleaning process starts with the default cleaning rate, which is 50 or 2 MBps.
- ▶ Background copy > 0 and cleaning rate = 0
The background copy takes place at the background copy rate, but no cleaning process starts. When the mapping is stopped, it goes into the stopping state, and a cleaning process starts with the default cleaning rate (50 or 2 MBps).
- ▶ Background copy = 0 and cleaning rate > 0
No background copy takes place, but the cleaning process runs at the cleaning rate. When the mapping is stopped, the cleaning completes (if not yet completed) at the cleaning rate.
- ▶ Background copy > 0 and cleaning rate > 0
The background copy takes place at the background copy rate, but no cleaning process starts. When the mapping is stopped, it goes into the stopping state, and a cleaning process starts with the specified cleaning rate.

Regarding the workload considerations for the cleaning process, the same guidelines as for background copy apply.

Host and application considerations to ensure FlashCopy integrity

Because FlashCopy works at the block level, it is necessary to understand the interaction between your application and the host operating system. From a logical standpoint, it is easiest to think of these objects as “layers” that sit on top of one another. The application is the topmost layer, and beneath it is the operating system layer.

Both of these layers have various levels and methods of caching data to provide better speed. Because IBM Storage Virtualize and FlashCopy sit below these layers, they are unaware of the cache at the application or operating system layers.

To ensure the integrity of the copy that is made, it is necessary to flush the host operating system and application cache for any outstanding reads or writes before the FlashCopy operation is performed. Failing to flush the host operating system and application cache produces what is referred to as a *crash consistent* copy.

The resulting copy requires the same type of recovery procedure, such as log replay and file system checks, that is required following a host crash. FlashCopy copies that are crash-consistent often can be used following file system and application recovery procedures.

Note: Although a best practice to perform FlashCopy is to flush the host cache first, some companies, such as Oracle, support using snapshots without it, as described in [Very Large Database \(VLDB\) Backup & Recovery Best Practices](#).

Various operating systems and applications provide facilities to stop I/O operations and ensure that all data is flushed from the host cache. If these facilities are available, they can be used to prepare for a FlashCopy operation. When this type of facility is not available, the host cache must be flushed manually by quiescing the application and unmounting the file system or drives.

Best practice: From a practical standpoint, when you have an application that is backed by a database and you want to make a FlashCopy of that application's data, it is sufficient in most cases to use the write-suspend method that is available in most modern databases. You can use this method because the database maintains strict control over I/O.

This method is as opposed to flushing data from both the application and the backing database, which is always the suggested method because it is safer. However, this method can be used when facilities do not exist or your environment includes time sensitivity.

6.3 Safeguarded Copy

The Safeguarded Copy function supports creating cyber-resilient copies of your important data by implementing the following features:

- ▶ Separation of duties provides more security capabilities to prevent nonprivileged users from compromising production data. Operations that are related to Safeguarded backups are restricted to only a subset of users with specific roles on the system (Administrator, Security Administrator, and Superuser).
- ▶ Protected Copies provides capabilities to regularly create Safeguarded backups. Safeguarded backups cannot be mapped directly to hosts to prevent any application from changing these copies.

Automation manages safeguarded backups and restores and recovers data with the integration of IBM Copy Services Manager (IBM CSM). IBM CSM automates the creation of Safeguarded backups according to the schedule that is defined in a Safeguarded policy. IBM CSM supports testing, restoring, and recovering operations with Safeguarded backups.

6.3.1 Safeguarded Copy use cases

Data recovery and restoration can span from more minor, or contained, situations to catastrophic. Safeguarded Copy can be used across the recovery range. Here are some use cases:

- ▶ Validation: Regular analytics on the copy provide early detection of a problem or reassurance that the copy is a valid copy before further actions.
- ▶ Forensic: You can start a copy of the production systems from the copy to investigate the problem and determine the required recovery actions.
- ▶ Surgical: You can recover a subset of volumes or logical unit numbers (LUNs) or extract data (that is, specific corrupted tables in a database by using database utilities) from a recovered copy and restore it back to the production environment.

- ▶ Catastrophic: You can recover the entire environment back to the point in time of the copy because it is the only recovery option.
- ▶ Offline backup: You can perform an offline backup of data from a consistent point-in-time copy to build a second line of defense, which provides a greater retention period and increased isolation and security.

6.3.2 Cyber resiliency

IBM Storage Virtualize Safeguarded Copy provides cyber resiliency against a cyberattack by providing two main capabilities:

- ▶ Immutability: Use Safeguarded Copy for immutable point-in-time copies of production data.
- ▶ Isolation: Use an air gap for “offline by design”.

Immutability

Immutability is defined by how easy it is to change, corrupt, or destroy data. Protection against all forms of corruption becomes more critical because in addition to hardware or software failures, corruption can be caused by inadvertent user error, malicious intent, or cyberattack.

To keep data safe, IBM Storage Virtualize end-to-end enterprise cyber resiliency features provide many options for protection systems and data from user errors, malicious destruction, and ransomware attacks.

Because an unrelenting tide of data breaches is driving increased interest in providing secure authentication across multicloud environments, IBM Storage Virtualize offers the powerful data security function of IBM Safeguarded Copy. With this new technology, businesses can prevent data tampering or deletion for any reason by enabling the creation of immutable point-in-time copies of data for a production volume.

Isolation

Isolation is a term that means that the protected copies of data are isolated from the active production data so that they cannot be corrupted by a compromised host system.

Safeguarded Backups are invisible to hackers, and are hidden and protected from being modified or deleted by user error, malicious destruction, or ransomware attacks.

The data can be used only after a Safeguarded Backup is recovered to a separate recovery volume. Recovery volumes can be accessed by using a recovery system that you use to restore production data. Safeguarded Backups are a trusted and secure source of data that can be used for forensic analysis or a surgical or catastrophic recovery.

6.3.3 Safeguarded Copy functional overview

Safeguarded Copy on IBM Storage Virtualize is implemented by using a new way of working with the FlashCopy function. The new method uses volume groups, which are similar to CGs. However, in addition to ensuring that a group of volumes is preserved at the same point in time, the volume group also enables the simplification of restoration or recovery to that point in time. It achieves this goal through the association of a group of volumes with a snapshot policy that determines frequency and retention duration. For more information, see *IBM FlashSystem Safeguarded Copy Implementation Guide*, REDP-5654.

Figure 6-11 shows a Safeguarded Copy functional overview.

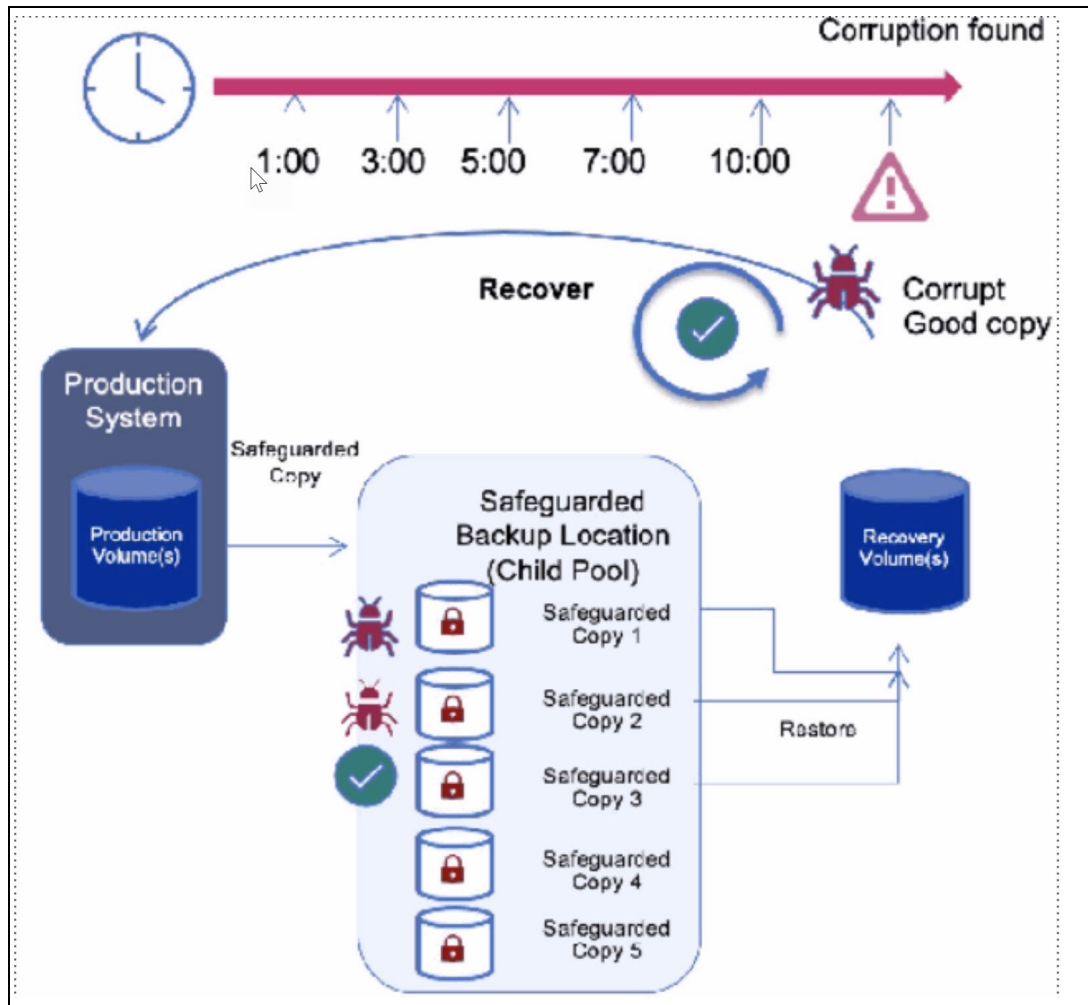


Figure 6-11 IBM Safeguarded Copy provides logical corruption protection to protect sensitive point in time copies of data

In this scenario, we have a Safeguarded Copy configuration with the production volumes, the recovery volumes, and five Safeguarded Backups (SGC1 to SGC5), with SGC5 as the most recent one, representing five recovery points. The recovery process (to point in time SGC2) consists of the following steps:

1. IBM Storage Virtualize establishes a FlashCopy from the production volumes to the recovery volumes, which make the recovery volumes identical to the production volumes.
2. IBM Storage Virtualize creates a recovery bitmap that indicates all data that was changed since SGC2 and must be referenced from the logs SGC5, SGC4, SGC3, and SGC2, rather than from the production volumes.
3. If the recovery system reads data from recovery volume, IBM Storage Virtualize examines the recovery bitmap and decides whether it must fetch the requested data from production volumes or from one of the CG logs.

Safeguarded Copy enables a user to create up to 15,864 immutable copies of a source or production environment. These backups are immutable and cannot be directly accessed by any server after creation.

Note: Safeguarded Copy is not a direct replacement for FlashCopy, and both can be used as part of a cyber resilience solution. Unlike with FlashCopy, the recovery data is not stored in separate regular volumes, but in a storage space that is called Safeguarded Backup Capacity.

Recovering a backup

The following considerations apply to recovering a backup:

- ▶ *Recover* is used when you want to test or run forensics against a backup.
- ▶ When you issue the Recover Backup command, IBM CSM creates a set of new R1 recovery volumes that are used to create an image of the data that is contained in the backup.
- ▶ IBM CSM allows customers to attach hosts to the R1 volumes through IBM CSM, or they can be attached through IBM Spectrum Virtualize.
- ▶ IBM CSM creates the R1 volumes with a name that is the concatenation of the source volume name and the time of the backup. With this approach, you can do quick filtering in the IBM Storage Virtualize GUI.
- ▶ By default, the recovery volumes are created in the Source pool for the H1 volumes, but on the **Recover Options** tab customers can select an alternative pool where the R1 volumes will be created.
- ▶ When creating R1 volumes, a provisioning policy can be used on the pool to define the default characteristics of the volume. By default, the volumes are created as thin volumes.

Restoring a backup

The following considerations apply to restoring a backup:

- ▶ *Restore* is used when you want to restore your source production volumes (H1) to the image of the data in the backup.
- ▶ Although not required, it is a best practice that you recover a backup, and test the contents before issuing the restore backup command.
- ▶ The restore backup replaces all content on the H1 source volumes with the content in the backup.
- ▶ For added protection, you can enable Dual Control in IBM CSM to ensure that two users approve a restore before it occurs.
- ▶ After a restore is issued, the session details panel indicates the time to which the H1s were restored.

6.3.4 Safeguarded Copy considerations

The following prerequisites and recommendations must be considered before implementing IBM Safeguarded Copy.

System requirements for Safeguarded Copy

Here are the system requirements for Safeguarded Copy:

- ▶ All supported IBM FlashSystem and SVC systems must be upgraded to 8.4.2.
- ▶ You must size extra capacity for SGC and Recovery Volume and data validation capacity (you can do sizing by using the IBM Storage Modeler).
- ▶ FlashCopy license.
- ▶ IBM CSM license:
 - Must be at 6.3.0.1.
 - Licensed for the source capacity of SGC.
 - The same CSM instance can be used for other IBM Storage Virtualize or DS8000 Copy Services.
 - Active and standby servers, and virtual machines (VMs), are required.

Note: IBM CSM is required for automating taking, maintaining, restoring, recovering, and deleting the Safeguarded copies. IBM CSM must be licensed.

Planning a Safeguarded Copy implementation

Implementing a Safeguarded Copy solution requires meeting defined business requirements and planning considerations similar to implementing a HA, DR, or high availability and disaster recovery (HADR) solution, such as for MM, GM, and Metro Global Mirror.

Whether building out a new solution or adding to an existing HADR solution, consider the following points if you must plan and prepare for the use cases that are described in 6.3.1, “Safeguarded Copy use cases” on page 391:

- ▶ Identify the potential problem or exposure that you must solve, for example, protect against inadvertent deletion, malicious destruction, selective manipulation, or ransomware attack.
- ▶ Identify the data that you need to protect, for example, mission- or business-critical only or the entire environment or a subset.
- ▶ Specify the frequency that you need to take Safeguarded backup copies, for example, take one every 10 or 30 minutes or every 1, 3, 5, or 24 hours.
- ▶ Identify how long do you need to keep the Safeguarded backup copies, for example, keep them for 1 day, 2 days, 5 days, 1 week, or 1 month.
- ▶ Determine how to perform validation and forensics for logical corruption detection.
- ▶ Decide on the servers that you will use for validation and forensics, and which ones you intend to use for recovery and run production after the data is recovered.
- ▶ Plan and prepare for surgical recovery and catastrophic recovery.
- ▶ Determine the frequency of the offline backups, retaining period, backups location, and required speed of recovery requirements from offline backups.
- ▶ Determine the current or planned HADR solution and how is it managed, for example, by using IBM CSM or a script.

- ▶ Decide on how should Safeguarded backup copies be implemented, for example, at the production data center, the DR data center, both, or at another site.
- ▶ Decide whether virtual or physical isolation is required.
- ▶ Be aware of the IBM Storage Virtualize maximum object and FlashCopy limits (256 FlashCopy mappings per source volume).

If a Safeguarded Copy Backup Policy creates a Safeguarded backup copy every hour and keeps the copies safe for 12 days, you would hit the 256 maximum FlashCopy mappings in 256 hours from the scheduled start time (1 backup every hour and 24 backups per day, for 12 days is (24 x 12) 360 backups, which exceeds the 256 limit).

- ▶ Determine the current or planned HADR solution. The type of isolation determines the topology, for example, 2-site, 3-site, and so on.

Limitations and restrictions

Here are the limitations and restrictions of Safeguarded Copy:

- ▶ Starting with IBM Storage Virtualize 8.4.2, the number of virtual disks (VDisks) and Fibre Channel (FC) mappings increased from 10,000 to 15,864.
- ▶ Safeguarded Copy supports a maximum of the following items:
 - 256 volume groups.
 - 512 volumes per volume group.
 - 256 Safeguarded copies per source volume.
 - 32 Safeguarded policies:
 - There are three default IBM supplied policies.
 - To create user-defined backup policies, use the `mksafeguardedpolicy` command.
 - The GUI does not currently support creating user-defined Safeguarded backup policies (but can display both).
- ▶ Safeguarded Copy backup location:
 - Defined as an immutable target location for Safeguarded Copy backups.
 - Resides in the same parent pool as any source volumes that are defined as requiring a Safeguarded Copy (parent pools can be traditional pools or DRPs).
- ▶ The source volume of a Safeguarded Copy cannot be any of the following items:
 - A volume that uses volume (VDisk) mirroring.
 - A source volume in an ownership group.
 - A change volume (CV) that is used in either HyperSwap or GM relationships.
 - A source volume that is used for cloud backups with the IBM Transparent Cloud Tiering (TCT) function.

Capacity planning and performance considerations

Here are the capacity planning and performance considerations for Safeguarded Copy:

- ▶ The extra space that is required for Safeguarded Copy depends on three factors:
 - Data write change rate.
 - Number and frequency of snapshots.
 - Duration and retention of snapshots.

- ▶ Consider not using Safeguarded Copy for every use case:
 - Use it strategically for mission- and business-critical workloads that need to be protected.
 - Backup solutions like IBM Spectrum Protect Plus are still required for long-term recovery and applications that can tolerate longer recovery times from the backup media (for example, off-premises object storage).
- ▶ Consider the performance and system overhead considerations of many FlashCopy copies.
- ▶ Leverage IBM Storage Modeler to help size the capacity requirements.

6.4 Remote copy services

IBM Storage Virtualize offers various remote copy services functions that address DR and business continuity needs.

MM is designed for metropolitan distances with a zero recovery point objective (RPO) to achieve zero data loss. This objective is achieved with a synchronous copy of volumes. Writes are not acknowledged to host until they are committed to both storage systems. By definition, any vendors' synchronous replication makes the host wait for write I/Os to complete at both the local and remote storage systems. Round-trip replication network latencies are added to source volume response time.

MM has the following characteristics:

- ▶ Zero RPO
- ▶ Synchronous
- ▶ Production application performance that is affected by round-trip latency

GM technologies are designed to minimize the effect of network latency on source volume by replicating data asynchronously. IBM Storage Virtualize provides two types of asynchronous mirroring technology:

- ▶ Standard GM
- ▶ GMCV

With GM, writes are acknowledged as soon as they can be committed to the local storage system. At the same time, they are sequence-tagged and passed on to the replication network. This technique allows GM to be used over longer distances. By definition, any vendors' asynchronous replication results in an RPO greater than zero. However, for GM, the RPO is small, typically anywhere from several milliseconds to some number of seconds.

Although GM is asynchronous, it tries to achieve near-zero RPO. Hence, the network and the remote storage system must be able to cope with peaks in traffic.

GM has the following characteristics:

- ▶ Near-zero RPO
- ▶ Asynchronous
- ▶ Production application performance that is affected by I/O sequencing preparation time

GMCV can replicate point-in-time copies of volumes. This option generally requires lower bandwidth because it is the average rather than the peak throughput that must be accommodated. The RPO for GMCV is higher than traditional GM.

GMCV has the following characteristics:

- ▶ Larger RPO (a user can define RPO that is based on available resources and workload).
- ▶ Point-in-time copies.
- ▶ Asynchronous.
- ▶ A possible system performance effect because point-in-time copies are created locally.
- ▶ Eliminates the replication link latency impact on production volumes.

Successful implementation of remote copy depends on taking a holistic approach in which you consider all components and their associated properties. The components and properties include host application sensitivity, local and remote storage area network (SAN) configurations, local and remote system and storage configuration, and the inter-system network.

6.4.1 Remote copy use cases

Data replication techniques are the foundations of DR and business continuity solutions. In addition to these common use cases, remote copy technologies can be used in other data movement scenarios that are described in the following sections.

Storage systems renewal

Remote copy functions can be used to facilitate the migration of data between storage systems while minimizing downtime for applications. By using remote copy, application data can be copied from an IBM Storage Virtualize system to another one without application downtime. After the volumes are fully copied and synchronized, the application can be stopped and then immediately started on the new storage system.

Starting with IBM Storage Virtualize 8.4.2, the Nondisruptive Volume Migration capability was introduced. This feature uses the remote copy capabilities to transparently move host volumes between IBM Storage Virtualize based systems.

For more information, see 5.8, “Volume migration” on page 342.

Data center moving

Remote copy functions can be used to move data between IBM Storage Virtualize based systems to facilitate data center moving operations. By using remote copy, application data can be copied from volumes in a source data center to volumes in another data center while applications remain online. After the volumes are fully copied and synchronized, the applications can be stopped and then immediately started in the target data center.

6.4.2 Remote copy functional overview

This section presents the terminology and the basic functional aspects of the remote copy services.

Common terminology and definitions

When such a breadth of technology areas is covered, the same technology component can have multiple terms and definitions. This document uses the following definitions:

- ▶ *Local system or master system or source system*
The system on which the foreground applications run.
- ▶ *Local hosts*
Hosts that run on the foreground applications.

▶ *Master volume or source volume*

The local volume that is being mirrored. The volume has non-restricted access. Mapped hosts can read/write to the volume.

▶ *Inter-system link or inter-system network*

The network that provides connectivity between the local and the remote site. It can be an FC network (SAN), an IP network, or a combination of the two.

▶ *Remote system or auxiliary system or target System*

The system that holds the remote mirrored copy.

▶ *Auxiliary volume or target volume*

The remote volume that holds the mirrored copy. It is read-access only.

▶ *Remote copy*

A generic term that is used to describe an MM or GM relationship in which data on the source volume is mirrored to an identical copy on a target volume. Often, the two copies are separated by some distance, which is why the term *remote* is used to describe the copies.

A remote copy relationship includes the following states:

– Consistent relationship

A remote copy relationship where the data set on the target volume represents a data set on the source volumes at a certain point.

– Synchronized relationship

A relationship is *synchronized* if it is consistent *and* the point that the target volume represents is the current point. The target volume contains identical data as the source volume.

- ▶ *Synchronous remote copy*
Writes to the source and target volumes that are committed in the foreground before confirmation is sent about completion to the local host application. MM is a synchronous remote copy type.
- ▶ *Asynchronous remote copy*
A foreground write I/O is acknowledged as complete to the local host application before the mirrored foreground write I/O is cached at the remote system. Mirrored foreground writes are processed asynchronously at the remote system, but in way that a consistent copy is always present in the remote system. GM and GMCV are asynchronous remote copy types.
- ▶ The *background copy* process manages the initial synchronization or resynchronization processes between source volumes to target mirrored volumes on a remote system.
- ▶ *Foreground I/O* reads and writes I/O on a local SAN, which generates a mirrored foreground write I/O that is across the inter-system network and remote SAN.

Figure 6-12 shows some of the concepts of remote copy.

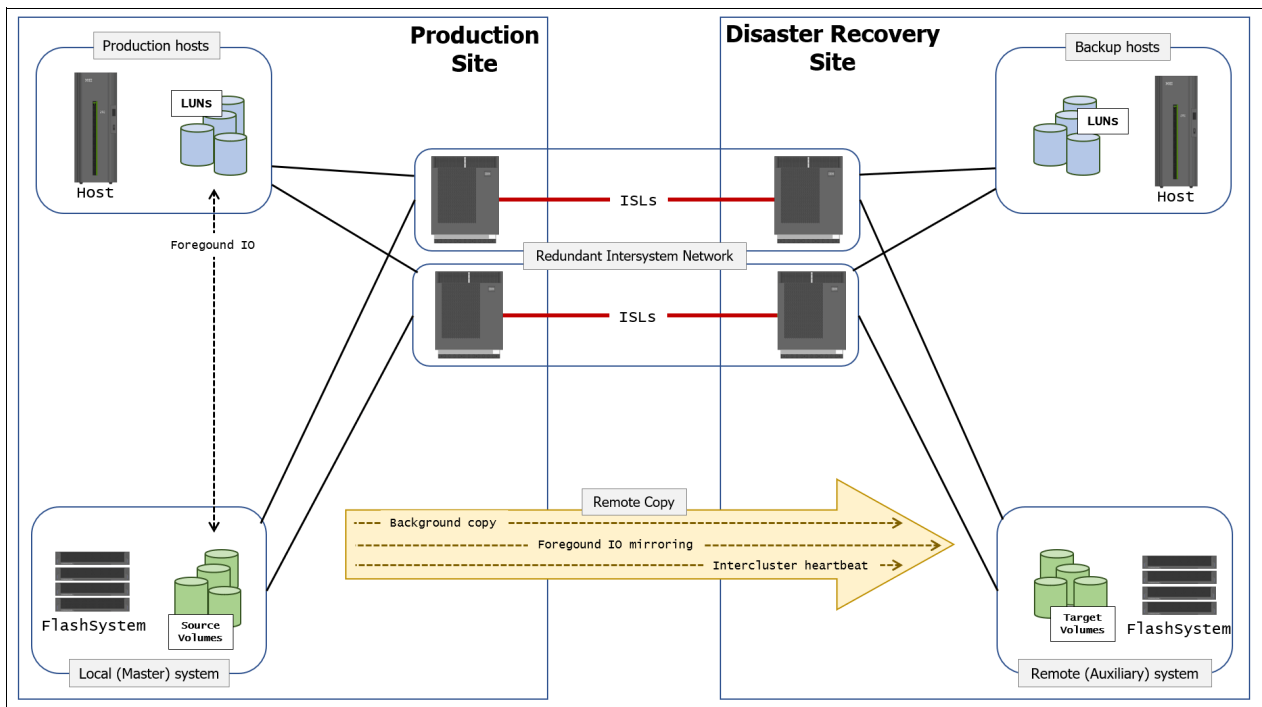


Figure 6-12 Remote copy components and applications

A successful implementation of inter-system remote copy services significantly depends on the quality and configuration of the inter-system network.

Remote copy partnerships and relationships

A remote copy *partnership* is a partnership that is established between a master (local) system and an auxiliary (remote) system, as shown in Figure 6-13.

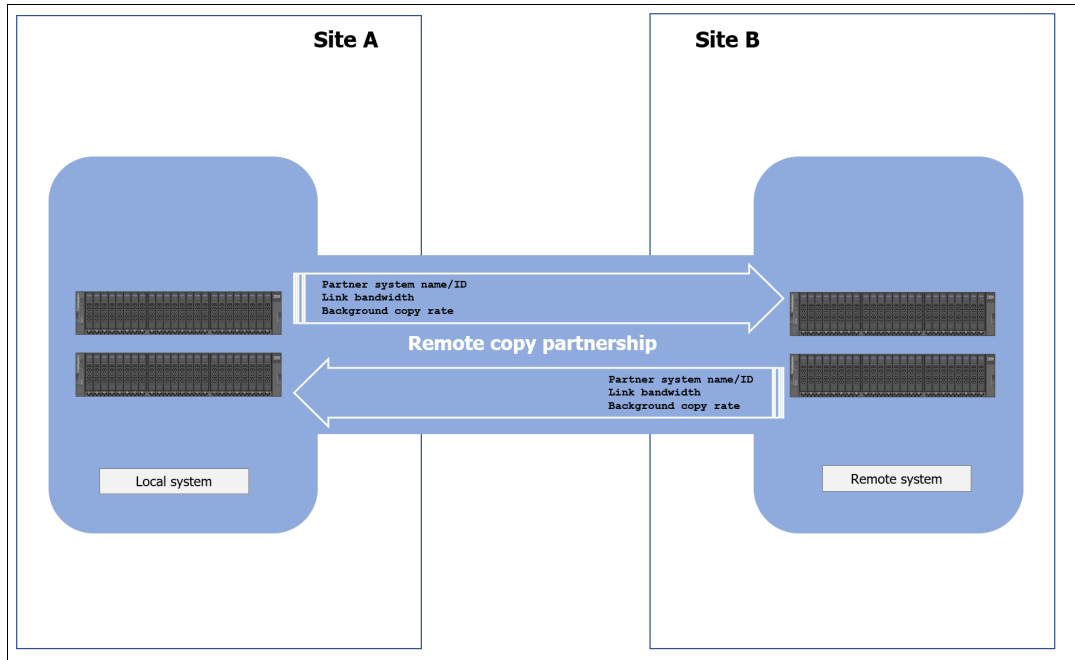


Figure 6-13 Remote copy partnership

Partnerships are established between two systems by issuing the **mkfcpartnership** (for an FC-based partnership) or **mkippartnership** (for an IP-based partnership) command once from each end of the partnership. The following parameters must be specified:

- ▶ Management IP is required. Both IPv4 and IPv6 supported.
- ▶ The remote system name (or ID).
- ▶ The link bandwidth (in Mbps).
- ▶ The background copy rate as a percentage of the link bandwidth.
- ▶ The background copy parameter that determines the maximum speed of the initial synchronization and resynchronization of the relationships.

Tip: To establish a fully functional MM or GM partnership, issue the **mkfcpartnership** (for an FC-based partnership) or **mkippartnership** (for a IP-based partnership) command from both systems.

In addition to the background copy rate setting, the initial synchronization can be adjusted at the relationship level with the **relationship_bandwidth_limit** parameter. The **relationship_bandwidth_limit** command is a system-wide parameter that sets the maximum bandwidth that can be used to initially synchronize a single relationship.

After background synchronization or resynchronization is complete, a remote copy relationship provides and maintains a consistent mirrored copy of a source volume to a target volume.

Copy directions and default roles

When a remote copy relationship is created, the source volume is assigned the role of the *master*, and the target volume is assigned the role of the *auxiliary*. This design implies that the initial copy direction of mirrored foreground writes and background resynchronization writes (if applicable) is from master to auxiliary. When a remote copy relationship is initially started, the master volume assumes the role of *primary* volume while the auxiliary volume becomes the *secondary* volume.

After the initial synchronization is complete, you can change the copy direction (see Figure 6-14) by switching the roles of the primary and secondary. The ability to change roles is used to facilitate DR.

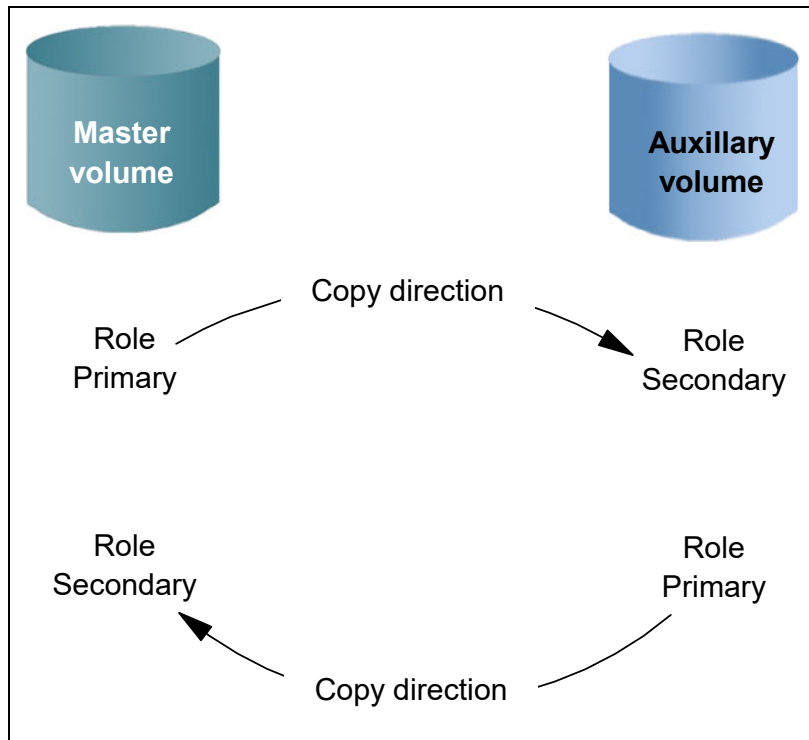


Figure 6-14 Role and direction changes

Attention: When the direction of the relationship is changed, the primary and secondary roles of the volumes are altered, which changes the read/write properties of the volume. The master volume takes on a secondary role and becomes read-only, and the auxiliary volume takes on the primary role and facilitates read/write access.

Consistency groups

A CG is a collection of relationships that can be treated as one entity. This technique is used to preserve write order consistency across a group of volumes that pertain to one application, for example, a database volume and a database log file volume.

After a remote copy relationship is added into a CG, you cannot manage the relationship in isolation from the CG. Issuing any command that can change the state of the relationship fails if it is run on an individual relationship that is already part of a CG. For example, a **stoprelationship** command on the stand-alone volume would fail because the system knows that the relationship is part of a CG.

Like remote copy relationships, remote copy CG also assigns the role of *master* to the source storage system and *auxiliary* to the target storage system.

Consider the following points regarding CGs:

- ▶ Each volume relationship can belong to only one CG.
- ▶ Volume relationships can also be stand-alone, that is, not in any CG.
- ▶ CGs can be created and left empty or can contain one or many relationships.
- ▶ You can create up to 256 CGs on a system.
- ▶ All volume relationships in a CG must have matching primary and secondary systems, but they do not need to share I/O groups.
- ▶ All relationships in a CG have the same copy direction and state.
- ▶ Each CG is either for MM or for GM relationships, but not both. This choice is determined by the first volume relationship that is added to the CG.

Consistency group consideration: A CG relationship does not have to directly match the I/O group number at each site. A CG that is owned by I/O group 1 at the local site does not have to be owned by I/O group 1 at the remote site. If you have more than one I/O group at either site, you can create the relationship between any two I/O groups. This technique spreads the workload, for example, from local I/O group 1 to remote I/O group 2.

Streams

CGs can be used as a way to spread replication workload across multiple streams within a partnership.

The GM partnership architecture allocates traffic from each CG in a round-robin fashion across 16 streams, that is, cg0 traffic goes into stream0, and cg1 traffic goes into stream1.

Any volume that is *not* in a CG also goes into stream0. You might want to consider creating an empty CG 0 so that stand-alone volumes do not share a stream with active CG volumes.

You can optimize your streams by creating more CGs. Within each stream, each batch of writes must be processed in tag sequence order and any delays in processing any particular write also delays the writes behind it in the stream. Having more streams (up to 16) reduces this kind of potential congestion.

Each stream is sequence-tag-processed by one node, so generally you want to create at least as many CGs as you have IBM Storage Virtualize nodes and controllers, and ideally perfect multiples of the node count.

Note: Streams are implemented only in GM.

Layer concept

The *layer* is an attribute of IBM Storage Virtualize based systems that allow you to create partnerships among different IBM Storage Virtualize products. The key points concerning layers are listed here:

- ▶ SVC is always in the *Replication* layer.
- ▶ By default, IBM FlashSystem products are in the *Storage* layer. A user can change it to the *Replication* layer.
- ▶ A system can form partnerships only with systems in the same layer.

- ▶ An SVC can virtualize an IBM FlashSystem system only if the IBM FlashSystem is in the Storage layer.
- ▶ An IBM FlashSystem system in the Replication layer can virtualize an IBM FlashSystem system in the Storage layer.

Figure 6-15 shows the concept of layers.

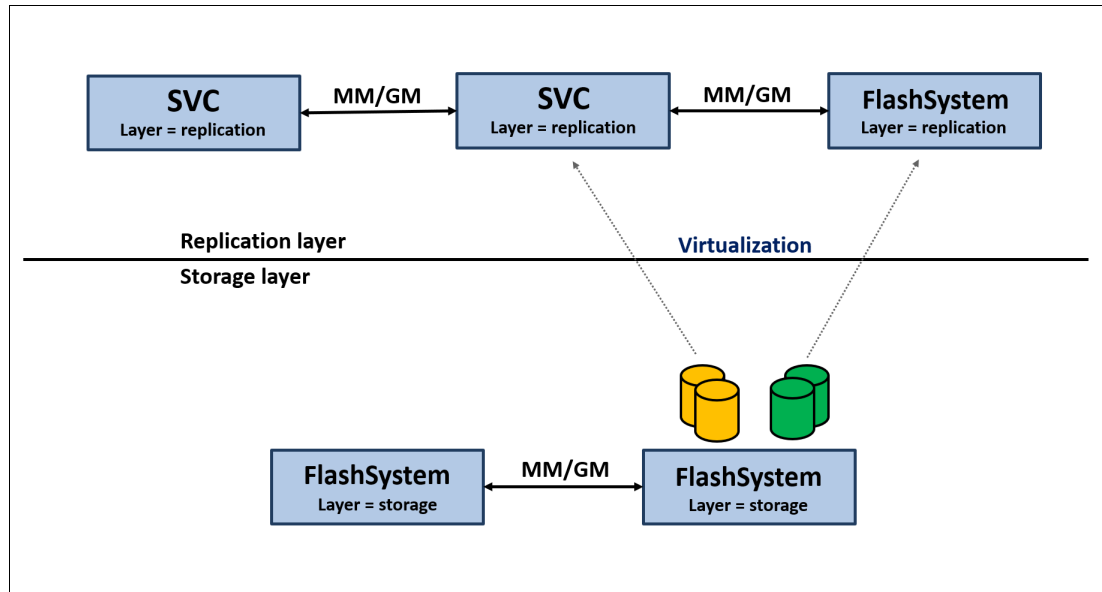


Figure 6-15 Conceptualization of layers

Generally, changing the layer is performed only at initial setup time or as part of a major reconfiguration. To change the layer of an IBM Storage Virtualize system, the system must meet the following preconditions:

- ▶ The IBM Storage Virtualize system must not have IBM Spectrum Virtualize, Storwize, or IBM FlashSystem host objects that are defined, and it must not be virtualizing any other IBM Storage Virtualize system.
- ▶ The IBM Storage Virtualize system must not be visible to any other IBM Storage Virtualize system in the SAN fabric.
- ▶ The IBM Storage Virtualize system must not have any system partnerships defined. If it is already using MM or GM, the existing partnerships and relationships must be removed first.

Changing an IBM Storage Virtualize system from the Storage layer to the Replication layer can be performed only by using the CLI. After you are certain that all the preconditions are met, issue the following command to change the layer from *Storage* to *Replication*:

```
chsystem -layer replication
```


Partnership topologies

IBM Storage Virtualize allows various partnership topologies, as shown in Figure 6-16. Each box represents an IBM Storage Virtualize based system.

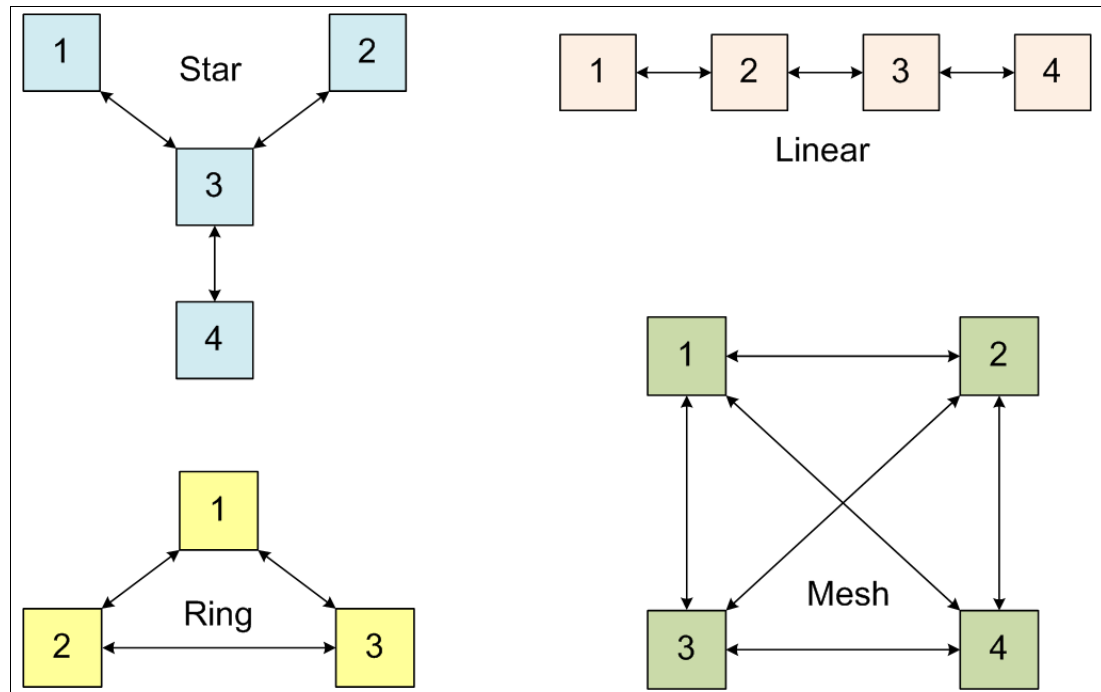


Figure 6-16 Supported topologies for remote copy partnerships

The set of systems that are directly or indirectly connected form the *connected set*. A system can be connected to up to three remote systems. No more than four systems can be in the same connected set.

Star topology

A star topology can be used to share a centralized DR system (three in this example) with up to three other systems, for example, replicating $1 \rightarrow 3$, $2 \rightarrow 3$, and $4 \rightarrow 3$.

Ring topology

A ring topology (three or more systems) can be used to establish a one-in, one-out implementation. For example, the implementation can be $1 \rightarrow 2$, $2 \rightarrow 3$, $3 \rightarrow 1$ to spread replication loads evenly among three systems.

Linear topology

A linear topology of two or more sites is also possible. However, it is simpler to create partnerships between system 1 and system 2, and separately between system 3 and system 4.

Mesh topology

A fully connected mesh topology is where every system has a partnership to each of the three other systems. This topology allows flexibility in that volumes can be replicated between any two systems.

Topology considerations:

- ▶ Although systems can have up to three partnerships, any one volume can be part of only a single relationship. You cannot establish a multi-target remote copy relationship for a specific volume. However, three-site replication is possible with IBM Storage Virtualize 3-site replication. For more information, see *IBM Spectrum Virtualize 3-Site Replication*, SG24-8504.
- ▶ Although various topologies are supported, it is advisable to keep your partnerships as simple as possible, which in most cases mean system pairs or a star.

Intrasystem remote copy

The intrasystem remote copy feature allows remote copy relationships to be created within the same IBM Storage Virtualize system. A preconfigured *local partnership* is created by default in the system for the intrasystem remote copy.

Considering that within a single system a remote copy does not protect data in a disaster scenario, this capability has no practical use except for functional testing. For this reason, intrasystem remote copy is not officially supported for production data.

Metro Mirror functional overview

MM provides synchronous replication. It is designed to ensure that updates are committed to both the primary and secondary volumes before sending an acknowledgment (ACK) of the completion to the server.

If the primary volume fails completely for any reason, MM is designed to ensure that the secondary volume holds the same data as the primary did at the time of failure.

MM provides the simplest way to maintain an identical copy on both the primary and secondary volumes. However, as with any synchronous copy over long distance, there can be a performance impact to host applications due to network latency.

MM supports relationships between volumes that are up to 300 kilometers (km) apart. Latency is an important consideration for any MM network. With typical fiber optic round-trip latencies of 1 millisecond (ms) per 100 km, you can expect a minimum of 3 ms extra latency due to the network alone on each I/O if you are running across the 300 km separation.

Figure 6-17 shows the order of MM write operations.

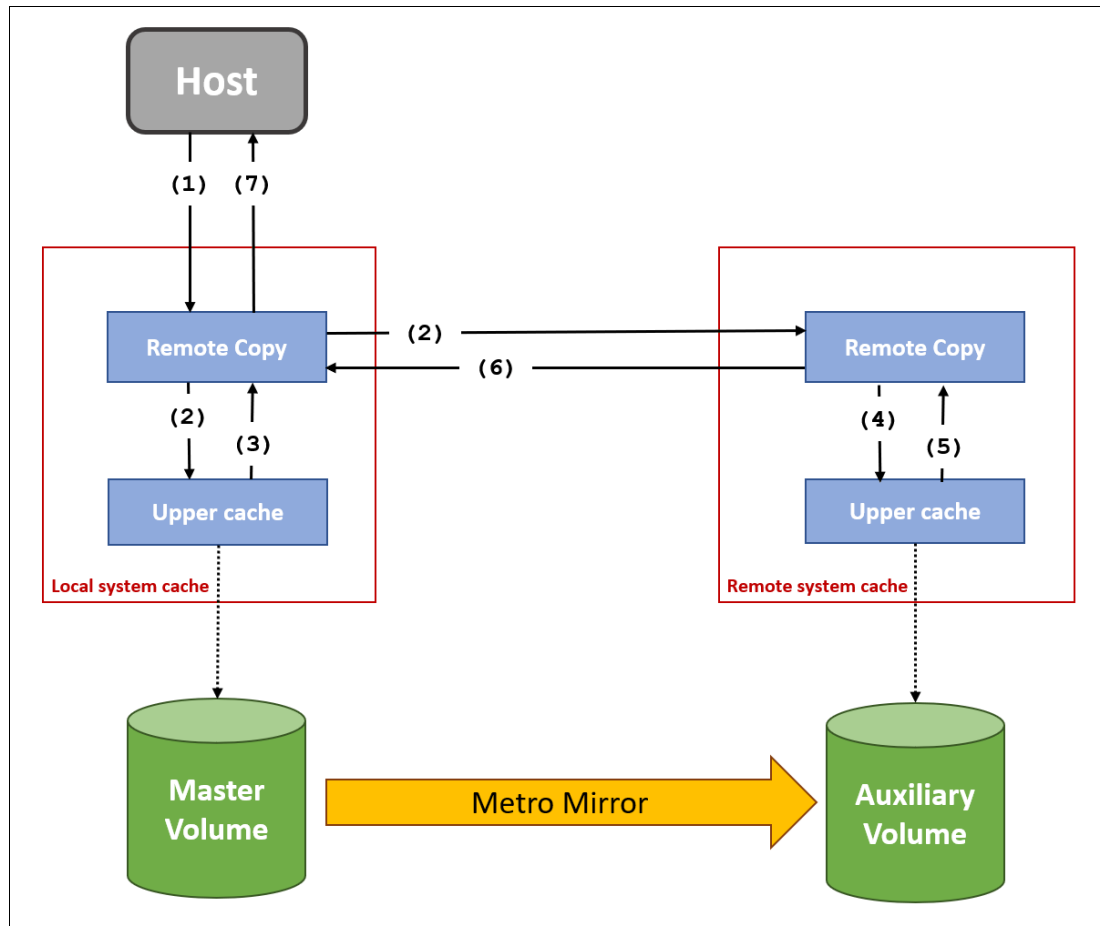


Figure 6-17 Metro Mirror write sequence

The write operation sequence includes the following steps:

1. The write operation is initiated by the host and intercepted by the remote copy component of the local system cache.
2. The write operation is simultaneously written in the upper cache component and sent to the remote system.
3. The write operation on local system upper cache is acknowledged back to the remote copy component on the local system.
4. The write operation is written in the upper cache component of the remote system. This operation is initiated when the data arrives from the local system, and it does not depend on an ongoing operation in the local system.
5. The write operation on the remote system upper cache is acknowledged to the remote copy component on remote system.
6. The remote write operation is acknowledged to the remote copy component on the local system.
7. The write operation is acknowledged to the host.

For a write to be considered as committed, the data must be written in both local and remote systems cache. De-staging to disk is a natural part of I/O management, but it is not generally in the critical path for an MM write acknowledgment.

Global Mirror functional overview

GM provides *asynchronous* replication. It is designed to reduce the dependency on round-trip network latency by acknowledging the primary write in parallel with sending the write to the secondary volume.

If the primary volume fails for any reason, GM ensures that the secondary volume holds the same data as the primary did at a point a short time before the failure. That short period of data loss is typically 10 ms - 10 seconds, but varies according to individual circumstances.

GM provides a way to maintain a write-order-consistent copy of data at a secondary site only slightly behind the primary. GM has minimal impact on the performance of the primary volume.

GM is an asynchronous remote copy technique, where foreground writes at the local system and mirrored foreground writes at the remote system are not wholly independent of one another. The IBM Storage Virtualize implementation of GM uses algorithms to maintain a consistent image at the target volume always.

This consistent image is achieved by identifying sets of I/Os that are active concurrently at the source, assigning an order to those sets, and applying these sets of I/Os in the assigned order at the target. The multiple I/Os within a single set are applied concurrently.

The process that marshals the sequential sets of I/Os operates at the remote system, and therefore is not subject to the latency of the long-distance link.

Figure 6-18 on page 409 shows that a write operation to the master volume is acknowledged to the host that issues the write before the write operation is mirrored to the cache for the auxiliary volume.

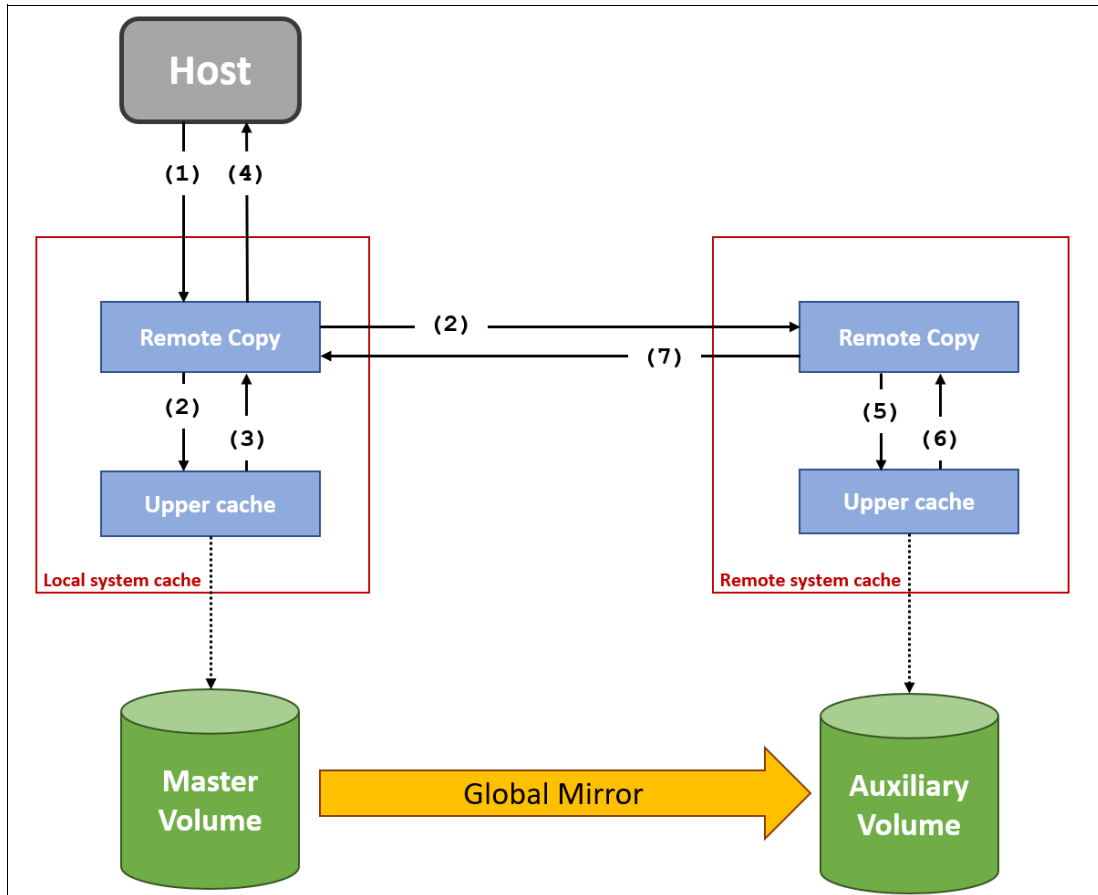


Figure 6-18 Global Mirror relationship write operation

The write operation sequence includes the following steps:

1. The write operation is initiated by the host and intercepted by the remote copy component of the local system cache.
2. The remote copy component on the local system completes the sequence tagging, and the write operation is simultaneously written in the upper cache component and sent to the remote system (along with the sequence number).
3. The write operation on the local system upper cache is acknowledged to the remote copy component on the local system.
4. The write operation is acknowledged to the host.
5. The remote copy component on the remote system initiates the write operation to the upper cache component according to the sequence number. This operation is initiated when the data arrives from the local system, and it does not depend on an ongoing operation in the local system.
6. The write operation on the remote system upper cache is acknowledged to the remote copy component on the remote system.
7. The remote write operation is acknowledged to the remote copy component on the local system.

With GM, a confirmation is sent to the host server before the host receives a confirmation of the completion at the auxiliary volume. The GM function identifies sets of write I/Os that are active concurrently at the primary volume. Then, it assigns an order to those sets and applies these sets of I/Os in the assigned order at the auxiliary volume.

Further writes might be received from a host when the secondary write is still active for the same block. In this case, although the primary write might complete, the new host write on the auxiliary volume is delayed until the previous write is completed. Any delay is reflected in the write-delay on the primary volume.

Write ordering

Many applications that use block storage must survive failures, such as a loss of power or a software crash. They also must not lose data that existed before the failure. Because many applications must perform many update operations in parallel to that storage block, maintaining write ordering is key to ensuring the correct operation of applications after a disruption.

An application that performs a high volume of database updates is often designed with the concept of dependent writes. Dependent writes ensure that an earlier write completes before a later write starts. Reversing the order of dependent writes can undermine the algorithms of the application and lead to problems, such as detected or undetected data corruption.

To handle this situation, IBM Storage Virtualize uses a write ordering algorithm while sending data to remote site by using remote copy. Each write gets tagged in the primary storage cache for its ordering. By using this order, data is sent to the remote site and committed on the target or remote storage.

Colliding writes

Colliding writes are defined as new write I/Os that overlap existing active write I/Os.

The original GM algorithm required only a single write to be active on any 512-byte LBA of a volume. If another write was received from a host while the auxiliary write was still active, the new host write was delayed until the auxiliary write was complete (although the master write might complete). This restriction was needed if a series of writes to the auxiliary had to be retried (which is known as *reconstruction*). Conceptually, the data for reconstruction comes from the master volume.

If multiple writes might be applied to the master for a sector, only the most recent write had the correct data during reconstruction. If reconstruction was interrupted for any reason, the intermediate state of the auxiliary became inconsistent.

Applications that deliver such write activity do not achieve the performance that GM is intended to support. A volume statistic is maintained about the frequency of these collisions. The original GM implementation was modified to allow multiple writes to a single location to be outstanding in the GM algorithm.

A need still exists for master writes to be serialized. The intermediate states of the master data must be kept in a nonvolatile journal while the writes are outstanding to maintain the correct write ordering during reconstruction. Reconstruction must never overwrite data on the auxiliary with an earlier version. The colliding writes of volume statistic monitoring are now limited to those writes that are not affected by this change.

Figure 6-19 on page 411 shows a colliding write sequence.

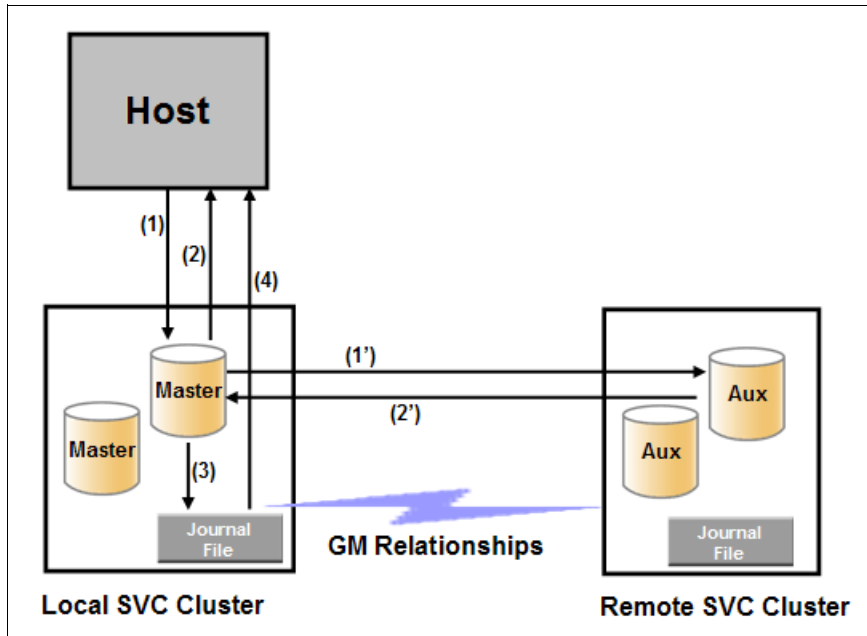


Figure 6-19 Colliding writes

The following numbers correspond to the numbers that are shown in Figure 6-19:

1. A first write is performed from the host to LBA X.
 2. A host receives acknowledgment that the write is complete, even though the mirrored write to the auxiliary volume is not yet complete.
- The first two actions (1 and 2) occur asynchronously with the first write.
3. A second write is performed from the host to LBA X. If this write occurs before the host receives acknowledgment (2), the write is written to the journal file.
 4. A host receives acknowledgment that the second write is complete.

Global Mirror with Change Volumes functional overview

GMCV provides asynchronous replication based on point-in-time copies of data. It is designed to allow for effective replication over lower bandwidth networks and reduce any impact on production hosts.

MM and GM both require the bandwidth to be sized to meet the peak workload. GMCV must be sized to meet only the average workload across a cycle period.

Figure 6-20 shows a high-level conceptual view of GMCV. GMCV uses FlashCopy to maintain image consistency and isolate host volumes from the replication process.

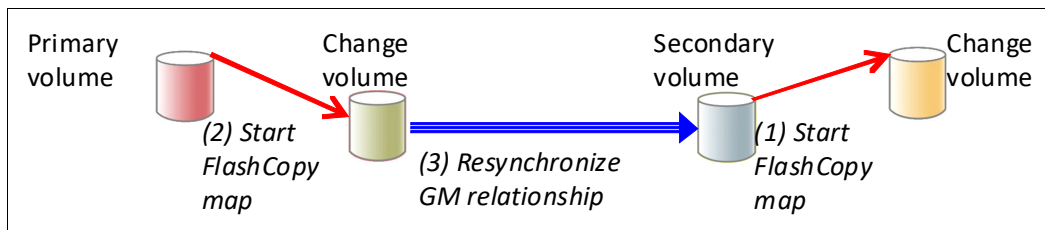


Figure 6-20 Global Mirror with Change Volumes

GMCV sends only one copy of a changed grain that might have been rewritten many times within the cycle period.

If the primary volume fails completely for any reason, GMCV ensures that the secondary volume holds the same data as the primary volume did at a specific point in time. That period of data loss is typically 5 minutes - 24 hours, but varies according to the design choices that you make.

CVs hold point-in-time copies of 256 KB grains. If any of the disk blocks in a grain change, that grain is copied to the CV to preserve its contents. CVs are also maintained at the secondary site so that a consistent copy of the volume is always available even when the secondary volume is being updated.

Change volumes considerations

Here are the CVs considerations:

- ▶ Primary volumes and CVs are always in the same I/O group.
- ▶ CVs are always thin-provisioned.
- ▶ CVs cannot be mapped to hosts and used for host I/O.
- ▶ CVs cannot be used as a source for any other FlashCopy or GM operations.

Figure 6-21 on page 413 shows how a CV is used to preserve a point-in-time data set, which is then replicated to a secondary site. The data at the secondary site is in turn preserved by a CV until the next replication cycle completes.

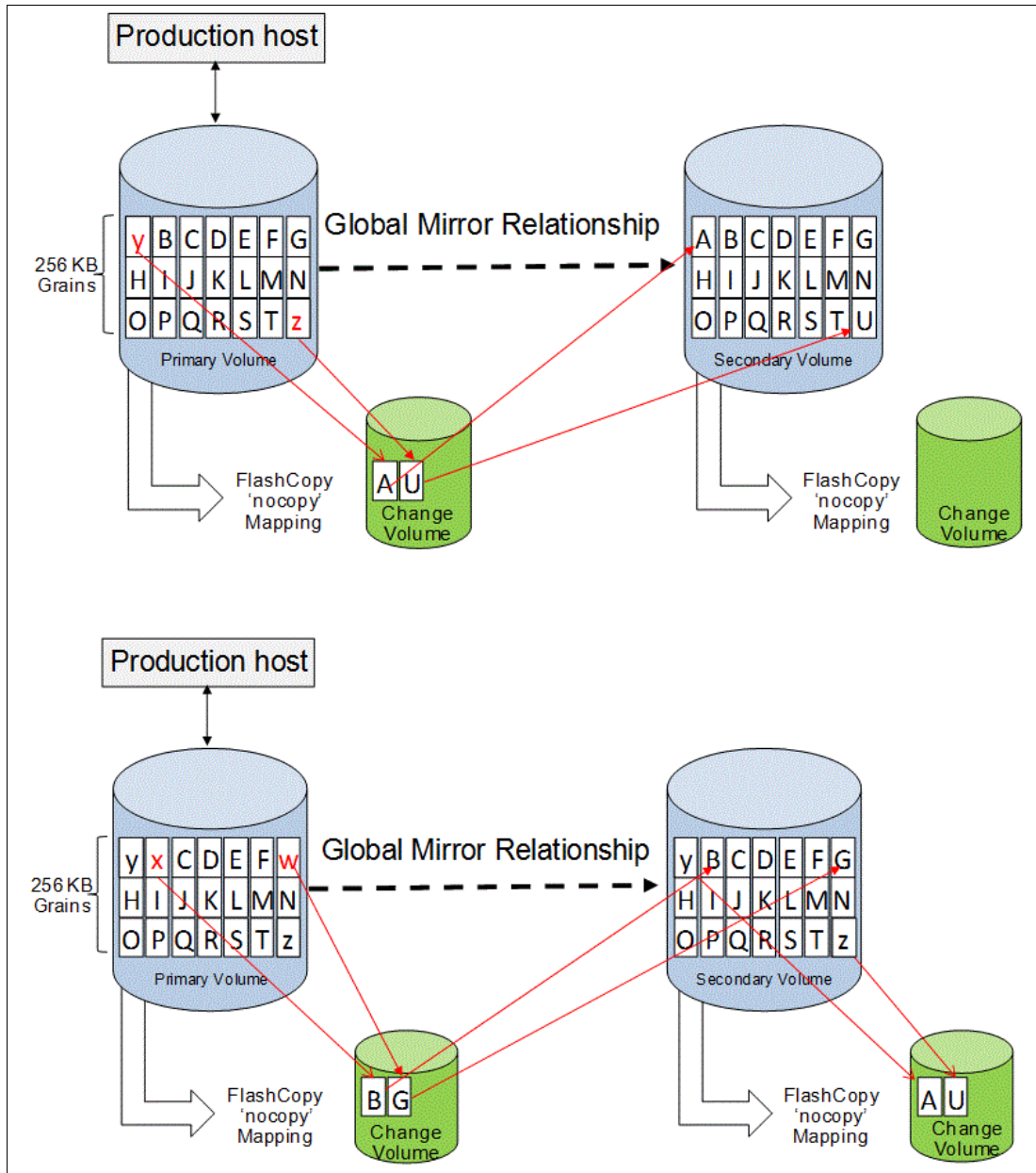


Figure 6-21 GMCV uses FlashCopy point-in-time copy technology

GMCV FlashCopy mapping note: GMCV FlashCopy mappings are not standard FlashCopy volumes and are not accessible for general use. They are internal structures that are dedicated to supporting GMCV.

The options for `-cyclingmode` are `none` and `multi`.

Specifying or taking the default `none` means that GM acts in its traditional mode without CVs.

Specifying `multi` means that GM starts cycling based on the cycle period, which defaults to 300 seconds. The valid range is 60 - 24*60*60 seconds (86,400 seconds = one day).

If all the changed grains cannot be copied to the secondary site within the specified time, then the replication takes as long as it needs, and it starts the next replication when the earlier one completes. You can choose to implement this approach by deliberately setting the cycle period to a short period, which is a perfectly valid approach. However, the shorter the cycle period, the less opportunity there is for peak write I/O smoothing, and the more bandwidth you need.

The `-cyclingmode` setting can be changed only when the GM relationship is in a stopped state.

Recovery point objective by using change volumes

RPO is the maximum tolerable period in which data might be lost if you switch over to your secondary volume.

If a cycle completes within the specified cycle period, then the RPO is not more than 2x cycle long. However, if it does not complete within the cycle period, then the RPO is not more than the sum of the last two cycle times.

The current RPO can be determined by looking at the `lsrcrelationship` freeze time attribute. The freeze time is the timestamp of the last primary CV that completed copying to the secondary site. Note the following example:

1. The cycle period is the default of 5 minutes. The cycle is triggered at 6:00 AM and completes at 6:03 AM. The freeze time would be 6:00 AM, and the RPO is 3 minutes.
2. The cycle starts again at 6:05 AM. The RPO now is 5 minutes. The cycle is still running at 6:12 AM, and the RPO is now up to 12 minutes because 6:00 AM is still the freeze time of the last complete cycle.
3. At 6:13 AM, the cycle completes and the RPO now is 8 minutes because 6:05 AM is the freeze time of the last complete cycle.
4. Because the cycle period was exceeded, the cycle immediately starts again.

Consistency protection overview

Consistency protection improves data availability and recovery if the system loses connectivity to volumes or CGs. When data is copied from a primary volume (master) to a secondary volume (auxiliary), consistency protection ensures that the secondary volume contains a write-order consistent set of data. The consistency of the secondary volume can be protected during resynchronization. For more information, see [Consistency protection](#).

The MM, GM, GMCV, and HyperSwap Copy Services functions create remote copy or remote replication relationships between volumes or CGs. If the secondary volume in a remote copy relationship becomes unavailable to the primary volume, the system maintains the relationship. However, the data might become out of sync when the secondary volume becomes available.

CVs can be used to maintain a consistent image of the secondary volume. HyperSwap relationships and GM relationships with cycling mode set to Multiple must always be configured with CVs. MM and GM with cycling mode set to None can optionally be configured with CVs.

When a secondary CV is configured, the relationship between the primary and secondary volumes does not stop if the link goes down or the secondary volume is offline. The relationship does not go in to the Consistent stopped status. Instead, the system uses the secondary CV to automatically copy the previous consistent state of the secondary volume.

The relationship automatically moves to the Consistent copying status as the system resynchronizes and protects the consistency of the data. The relationship status changes to Consistent synchronized when the resynchronization process completes. The relationship automatically resumes replication after the temporary loss of connectivity.

You are not required to configure a secondary CV on an MM or GM (without cycling) relationship. However, if the link goes down or the secondary volume is offline, the relationship goes in to the Consistent stopped status. If write operations take place on either the primary or secondary volume, the data is no longer synchronized (out of sync).

Consistency protection must be enabled on all relationships in a CG. Every relationship in a CG must be configured with a secondary CV.

6.4.3 Remote copy network planning

Remote copy partnerships and relationships do not work reliably if the connectivity on which they are running is configured incorrectly. This section focuses on the inter-system network, giving an overview of the remote system connectivity options.

Terminology

The inter-system network is specified in terms of *latency* and *bandwidth*. These parameters define the capabilities of the link regarding the traffic that it can carry. They must be chosen so that they support all forms of traffic, including mirrored foreground writes, background copy writes, and inter-system heartbeat messaging (node-to-node communication).

Link latency is the time that is taken by data to move across a network from one location to another one. It is measured in milliseconds. The latency measures the time that is spent to send the data and to receive the acknowledgment back (round-trip time (RTT)).

Link bandwidth is the network capacity to move data as measured in millions of bits per second or megabits per second (Mbps) or billions of bits per second or gigabits per second (Gbps).

The term *bandwidth* is also used in the following context:

- ▶ Storage bandwidth: The ability of the back-end storage to process I/O. Measures the amount of data (in bytes) that can be sent in a specified amount of time.
- ▶ Remote copy partnership bandwidth (parameter): The rate at which background write synchronization is attempted (unit of Mbps).

Inter-system connectivity supports mirrored foreground and background I/O. A portion of the link is also used to carry traffic that is associated with the exchange of low-level messaging between the nodes of the local and remote systems. A *dedicated amount* of the link bandwidth is required for the exchange of heartbeat messages and the initial configuration of inter-system partnerships.

FC connectivity is the standard connectivity that is used for the remote copy inter-system networks. It uses the FC protocol and SAN infrastructures to interconnect the systems.

Native IP connectivity is a connectivity option that is based on standard TPC/IP infrastructures that are provided by IBM Storage Virtualize technology.

Standard SCSI operations and latency

A single SCSI read operation over an FC network is shown in Figure 6-22.

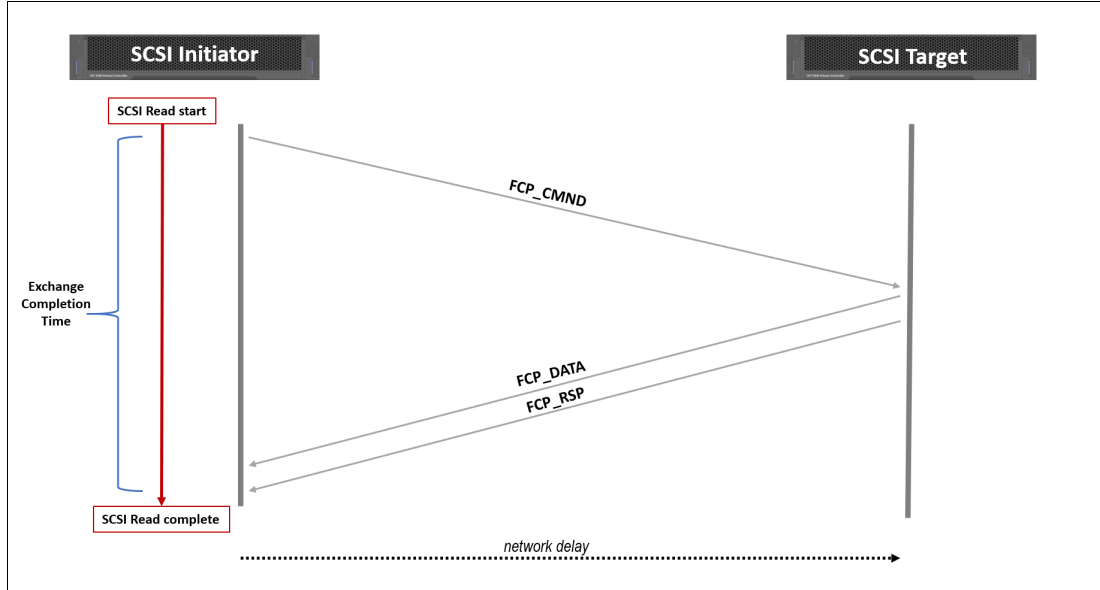


Figure 6-22 Standard SCSI read operation

The initiator starts by sending a read command (**FCP_CMND**) across the network to the target. The target is responsible for retrieving the data and responding by sending the data (**FCP_DATA_OUT**) to the initiator. Finally, the target completes the operation by sending the command completed response (**FCP_RSP**). **FCP_DATA_OUT** and **FCP_RSP** are sent to the initiator in sequence. Overall, one round trip is required to complete the read, so the read takes at least one RTT plus the time for the data out.

A typical SCSI behavior for a write is shown in Figure 6-23.

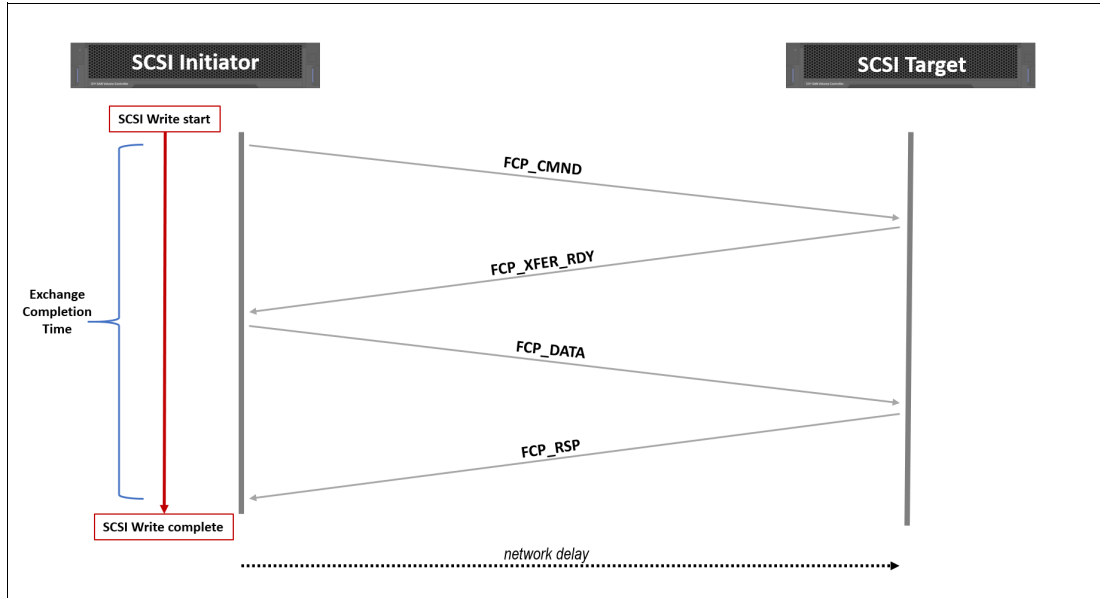


Figure 6-23 Standard SCSI write operation

A standard-based SCSI write is a two-step process:

1. The write command (**FCP_CMND**) is sent across the network to the target. The first round trip is essentially asking transfer permission from the target. The target responds with an acceptance (**FCP_XFR_RDY**).

The initiator waits until it receives a response from the target before starting the second step, that is, sending the data (**FCP_DATA_OUT**).

2. The target completes the operation by sending the command completed response (**FCP_RSP**). Overall, two round trips are required to complete the write, so the write takes at least $2 \times \text{RTT}$, plus the time for the data out.

Within the confines of a data center, where the latencies are measured in microseconds (μsec), no issues exist. However, across a geographical network where the latencies are measured in milliseconds (ms), the overall service time can be significantly affected.

Considering that the network delay over fiber optics per kilometer (km) is approximately $5 \mu\text{sec}$ ($10 \mu\text{sec}$ RTT), the resulting minimum service time per every km of distance for a SCSI operation is $10 \mu\text{sec}$ (reads) and $20 \mu\text{sec}$ (writes), for example, a SCSI write over 50 km has a minimum service time of $1000 \mu\text{sec}$ (that is, 1 ms).

IBM Storage Virtualize remote write operations

With standard SCSI operations, writes are especially affected by the latency. IBM Storage Virtualize implements a proprietary protocol to mitigate the effects of the latency in the write operations over an FC network.

Figure 6-24 shows how a remote copy write operation is performed over an FC network.

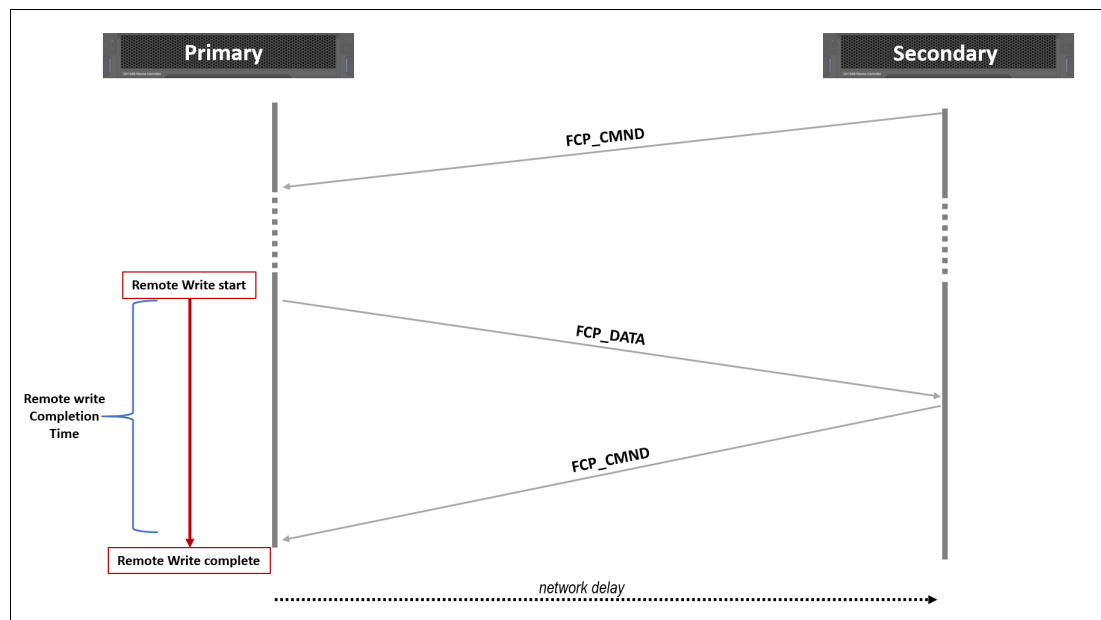


Figure 6-24 IBM Storage Virtualize remote copy write

When the remote copy is initialized, the target system (secondary system) sends a dummy read command (**FCP_CMND**) to the initiator (primary system). This command waits on the initiator until a write operation is requested.

When a write operation is started, the data is sent to the target as response of the dummy read command (**FCP_DATA_OUT**). Finally, the target completes the operation by sending a new dummy read command (**FCP_CMND**).

Overall, one round trip is required to complete the remote write by using this protocol, so replicating a write takes at least one RTT plus the time for the data out.

Network latency considerations

The maximum supported round-trip latency between sites depends on the type of partnership between systems. Table 6-6 lists the maximum round-trip latency. This restriction applies to all variants of remote mirroring.

Table 6-6 Maximum round trip

Partnership		
FC	1 Gbps IP	10 Gbps IP
250 ms	80 ms	10 ms

More configuration requirements and guidelines apply to systems that perform remote mirroring over extended distances, where the RTT is greater than 80 ms. If you use remote mirroring between systems with 80 - 250 ms round-trip latency, you must meet the following extra requirements:

- ▶ The RC buffer size setting must be 512 MB on each system in the partnership. This setting can be accomplished by running the `chsystem -rcbuffersize 512` command on each system.

Important: Changing this setting is disruptive to MM and GM operations. Use this command only before partnerships are created between systems, or when all partnerships with the system are stopped.

- ▶ Two FC ports on each node that will be used for replication must be dedicated for replication traffic. This configuration can be achieved by using SAN zoning and port masking. Starting with IBM Storage Virtualize 8.5, a user can configure a remote copy portset to achieve remote copy traffic isolation.
- ▶ SAN zoning should be applied to provide separate intrasystem zones for each local-remote I/O group pair that is used for replication. For more information about zoning guidelines, see “Remote system ports and zoning considerations” on page 425.

Link bandwidth that is used by internode communication

IBM Storage Virtualize uses part of the bandwidth for its internal inter-system heartbeat. The amount of traffic depends on how many nodes are in each of the local and remote systems. Table 6-7 shows the amount of traffic (in Mbps) that is generated by different sizes of systems.

Table 6-7 IBM Storage Virtualize inter-system heartbeat traffic (Mbps)

Local or remote system	Two nodes	Four nodes	Six nodes	Eight nodes
Two nodes	5	6	6	6
Four nodes	6	10	11	12
Six nodes	6	11	16	17
Eight nodes	6	12	17	21

These numbers represent the total traffic between the two systems when *no* I/O is occurring to a mirrored volume on the remote system. Half of the data is sent by one system, and half of the data is sent by the other system. The traffic is divided evenly over all available connections. Therefore, if you have two redundant links, half of this traffic is sent over each link during a fault-free operation.

If the link between the sites is configured with redundancy to tolerate single failures, size the link so that the bandwidth and latency statements continue to be accurate even during single failure conditions.

Network sizing considerations

Proper network sizing is essential for the remote copy services operations. Failing to estimate the network sizing requirements can lead to poor performance in remote copy services and the production workload.

Consider that inter-system bandwidth should support the combined traffic of the following items:

- ▶ Mirrored foreground writes, as generated by your server applications at peak times
- ▶ Background write synchronization, as defined by the GM bandwidth parameter
- ▶ Inter-system communication (*heartbeat messaging*)

Calculating the required bandwidth is essentially a question of mathematics based on your current workloads, so you should start by assessing your current workloads.

Metro Mirror and Global Mirror network sizing

Because MM is synchronous replication, the amount of replication bandwidth that is required to mirror a foreground write-data throughput is more than or equal to the foreground write-data throughput.

GM, which does not have write buffering resources, tends to mirror the foreground write when it is committed in cache, so the bandwidth requirements are similar to MM.

For a proper bandwidth sizing with MM or GM, you must know your peak write workload to at least a 5-minute interval. This information can be easily gained from tools like IBM Spectrum Control. Finally, you must allow for background copy, intercluster communication traffic, and a safe margin for unexpected peaks and workload growth.

Recommendation: Do not compromise on bandwidth or network quality when planning an MM or GM deployment. If bandwidth is likely to be an issue in your environment, consider GMCV.

As an example, consider a business with the following I/O profile:

- ▶ The average write size is 8 KB ($8 \times 8 \text{ bits}/1024 = 0.0625 \text{ Mb}$).
- ▶ For most of the day between 8 AM and 8 PM, the write activity is approximately 1500 writes per second.
- ▶ Twice a day (once in the morning and once in the afternoon), the system bursts up to 4500 writes per second for up to 10 minutes.

This example represents a general traffic pattern that might be common in many medium-sized sites. Furthermore, 20% of bandwidth must be left available for the background synchronization.

MM or GM require bandwidth on the instantaneous peak of 4500 writes per second as follows:

$$4500 \times 0.0625 = 282 \text{ Mbps} + 20\% \text{ resync allowance} + 5 \text{ Mbps heartbeat} = 343 \text{ Mbps}$$

This calculation provides the dedicated bandwidth that is required. The user should consider a safety margin plus growth requirement for the environment.

$$(\text{Peak Write I/O} \times \text{Write I/O Size} = [\text{Total Data Rate}] + 20\% \text{ resync allowance} + 5 \text{ Mbps heartbeat} = [\text{Total Bandwidth Required}])$$

GMCV network sizing

The GMCV is typically less demanding in terms of bandwidth requirements for many reasons:

- ▶ By using its journaling capabilities, the GMCV provides a way to maintain point-in-time copies of data at a secondary site where insufficient bandwidth is available to replicate the peak workloads in real time.
- ▶ It sends only one copy of a changed grain, which might have been rewritten many times within the cycle period.

The GMCV network sizing is basically a tradeoff between RPO, journal capacity, and network bandwidth. A direct relationship exists between the RPO and the physical occupancy of the CVs. The lower the RPO, the less capacity is used by CVs. However, higher RPO requires usually less network bandwidth.

For a proper bandwidth sizing with GMCV, you must know your average write workload during the cycle time. This information can be obtained easily from tools like IBM Spectrum Control. Finally, you must consider the background resync workload, intercluster communication traffic, and a safe margin for unexpected peaks and workload growth.

As an example, consider a business with the following I/O profile:

- ▶ Average write size 8 KB ($8 \times 8 \text{ bits}/1024 = 0.0625 \text{ Mb}$).
- ▶ For most of the day from 8 AM - 8 PM, the write activity is approximately 1500 writes per second.
- ▶ Twice a day (once in the morning and once in the afternoon), the system bursts up to 4500 writes per second for up to 10 minutes.
- ▶ Outside of the 8 AM - 8 PM window, there is little or no I/O write activity.

This example is intended to represent a general traffic pattern that might be common in many medium-sized sites. Furthermore, 20% of bandwidth must be left available for the background synchronization.

Consider the following sizing exercises:

- ▶ GMCV peak 30-minute cycle time

If we look at this time as broken into 10-minute periods, the peak 30-minute period is made up of one 10-minute period of 4500 writes per second, and two 10-minute periods of 1500 writes per second. The average write rate for the 30-minute cycle period can be expressed mathematically as follows:

$$(4500 + 1500 + 1500) / 3 = 2500 \text{ writes/sec for a 30-minute cycle period}$$

The minimum bandwidth that is required for the cycle period of 30 minutes is as follows:

$$2500 \times 0.0625 = 157 \text{ Mbps} + 20\% \text{ resync allowance} + 5 \text{ Mbps heartbeat} = 195 \text{ Mbps}$$

dedicated plus any safety margin plus growth

► GMCV peak 60-minute cycle time

For a cycle period of 60 minutes, the peak 60-minute period is made up of one 10-minute period of 4500 writes per second, and five 10-minute periods of 1500 writes per second. The average write for the 60-minute cycle period can be expressed as follows:

$$(4500 + 5 \times 1500)$$

Replicating until 8 AM at the latest would probably require at least the following bandwidth:

$$(9000 + 70 \times 1500) / 72 = 1584 \times 0.0625 = 99 \text{ Mbps} + 100\% + 5 \text{ Mbps heartbeat} = 203 \text{ Mbps}$$

at night, plus any safety margin and growth, and nondedicated and time-shared with daytime traffic.

► GMCV with daily cycle time

Suppose that the business does not have aggressive RPO requirements and does not want to provide dedicated bandwidth for GM. However, the network is available and unused at night, so GM can use that network. An element of risk exists here, that is, if the network is unavailable for any reason, GMCV cannot keep running during the day until it catches up. Therefore, you must allow a much higher resync allowance in your replication window. For example, 100 percent.

A GMCV replication that is based on daily point-in-time copies at 8 PM each night and replicating until 8 AM at the latest likely requires at least the following bandwidth:

$$(9000 + 70 \times 1500) / 72 = 1584 \times 0.0625 = 99 \text{ Mbps} + 100\% + 5 \text{ Mbps heartbeat} = 203 \text{ Mbps}$$

at night, plus any safety margin and growth, and nondedicated and time-shared with daytime traffic.

The central principle of sizing is that you need to know your write workload:

- For MM and GM, you need to know the peak write workload.
- For GMCV, you need to know the average write workload.

GMCV bandwidth: In the above examples, the bandwidth estimation for GMCV is based on the assumption that the write operations occur in such a way that a CV grain (that has a size of 256 KB) is completely changed before it is transferred to the remote site. In the real life, this situation is unlikely to occur.

Usually, only a portion of a grain is changed during a GMCV cycle, but the transfer process always copies the whole grain to the remote site. This behavior can lead to an unforeseen processor burden in the transfer bandwidth that, in an edge case, can be even higher than the one that is required for standard GM.

Global Mirror and GMCV coexistence considerations

GM and GMCV relationships can be defined in the same system. With these configurations, particular attention must be paid to bandwidth sizing and the partnership settings.

The two GM technologies, as previously described, use the available bandwidth in different ways:

- GM uses the amount of bandwidth that is needed to sustain the write workload of the replication set.
- GMCV uses the fixed amount of bandwidth as defined in the partnership as background copy.

For this reason, during GMCV cycle-creation, a fixed part of the bandwidth is allocated for the background copy and only the remaining part of the bandwidth is available for GM.

To avoid bandwidth contention, which can lead to a 1920 error (see 6.5.6, “Replication portsets” on page 463) or delayed GMCV cycle creation, the bandwidth must be sized to consider both requirements.

Note: GM can use bandwidth that is reserved for background copy if it is not used by a background copy workload.

Ideally, in these cases the bandwidth should be enough to accommodate the peak write workload for the GM replication set plus the estimated bandwidth that is needed to fulfill the RPO of GMCV. If these requirements cannot be met due to bandwidth restrictions, the option with the least impact is to increase the GMCV cycle period, and then reduce the background copy rate to minimize the chance of a 1920 error.

These considerations also apply to configurations where multiple IBM Storage Virtualize based systems are sharing bandwidth resources.

Fibre Channel connectivity

When you use FC technology for the inter-system network, consider the following items:

- ▶ Redundancy
- ▶ Basic topology and problems
- ▶ Distance extensions options
- ▶ Hops
- ▶ Buffer credits
- ▶ Remote system ports and zoning considerations

Redundancy

The inter-system network must adopt the same policy toward redundancy as for the local and remote systems to which it is connecting. The Inter-Switch Links (ISLs) must have redundancy, and the individual ISLs must provide the necessary bandwidth in isolation.

Basic topology and problems

Because of the nature of FC, you must avoid ISL congestion whether within individual SANs or across the inter-system network. Although FC (and IBM Spectrum Virtualize) can handle an overloaded host or storage array, the mechanisms in FC are ineffective for dealing with congestion in the fabric in most circumstances. The problems that are caused by fabric congestion can range from dramatically slow response time to storage access loss. These issues are common with all high-bandwidth SAN devices and inherent to FC. They are not unique to IBM Storage Virtualize products.

When an FC network becomes congested, the FC switches stop accepting more frames until the congestion clears. They also can drop frames. Congestion can quickly move upstream in the fabric and clog the end devices from communicating.

This behavior is referred to as *head-of-line blocking*. Although modern SAN switches internally have a nonblocking architecture, head-of-line-blocking still exists as a SAN fabric problem. Head-of-line blocking can result in an IBM Storage Virtualize node that cannot mirror its write caches because you have a single congested link that leads to an edge switch.

Distance extensions options

To implement remote mirroring over a distance by using the FC, you have the following choices:

- ▶ *Optical multiplexors*, such as Dense Wavelength Division Multiplexing (DWDM) or Coarse Wavelength Division Multiplexing (CWDM) devices.

Optical multiplexors can extend a SAN up to hundreds of kilometers (or miles) at high speeds. For this reason, they are the preferred method for long-distance expansion. If you use multiplexor-based distance extensions, closely monitor your physical link error counts in your switches. Optical communication devices are high-precision units. When they shift out of calibration, you start to see errors in your frames.

- ▶ Long-distance small form-factor pluggable (SFP) transceivers and 10-Gb small form factor pluggables (XFPs).

Long-distance optical transceivers have the advantage of extreme simplicity. You do not need expensive equipment, and only a few configuration steps need to be performed. However, ensure that you use only transceivers that are designed for your particular SAN switch.

- ▶ FC-to-IP conversion boxes. Fibre Channel over IP (FCIP) is, by far, the most common and least expensive form of distance extension. It also is complicated to configure. Relatively subtle errors can have severe performance implications.

With IP-based distance extension, you must dedicate bandwidth to your FCIP traffic if the link is shared with other IP traffic. Do not assume that because the link between two sites has low traffic or is used only for email that this type of traffic is always the case. FC is far more sensitive to congestion than most IP applications.

Also, when you are communicating with the networking architects for your organization, make sure to distinguish between *megabytes per second* (MBps) as opposed to *megabits per second* (Mbps). In the storage world, bandwidth is often specified in MBps, and network engineers specify bandwidth in Mbps.

Of these options, the optical distance extension is the preferred method. IP distance extension introduces more complexity, is less reliable, and has performance limitations. However, optical distance extension can be impractical in many cases because of cost or unavailability.

For more information, see [SAN Volume Controller Supported environment](#).

Hops

The hop count is not increased by the intersite connection architecture. For example, if you have a SAN extension that is based on DWDM, the DWDM components are not apparent to the number of hops. The hop count limit within a fabric is set by the fabric devices (switch or director) operating system. It is used to derive a frame hold time value for each fabric device.

This hold time value is the maximum amount of time that a frame can be held in a switch before it is dropped or the fabric busy condition is returned. For example, a frame might be held if its destination port is unavailable. The hold time is derived from a formula that uses the error detect timeout value and the resource allocation timeout value. Every extra hop adds about 1.2 microseconds of latency to the transmission.

Currently, IBM Storage Virtualize Copy Services support three hops when protocol conversion exists. Therefore, if you have DWDM extended between primary and secondary sites, three SAN directors or switches can exist between the primary and secondary systems.

Buffer credits

SAN device ports need memory to temporarily store frames as they arrive, assemble them in sequence, and deliver them to the upper layer protocol. The number of frames that a port can hold is called its *buffer credit*. The FC architecture is based on a flow control that ensures a constant stream of data to fill the available pipe.

When two FC ports begin a conversation, they exchange information about their buffer capacities. An FC port sends only the number of buffer frames for which the receiving port gives credit. This method avoids overruns and provides a way to maintain performance over distance by filling the pipe with in-flight frames or buffers.

The following types of transmission credits are available:

▶ **Buffer_to_Buffer Credit (BB_Credit)**

During login, N_Ports and F_Ports at both ends of a link establish its BB_Credit.

▶ **End_to_End Credit (EE_Credit)**

In the same way during login, all N_Ports establish EE_Credit with each other. During data transmission, a port must not send more frames than the buffer of the receiving port can handle before you receive an indication from the receiving port that it processed a previously sent frame. Two counters are used: BB_Credit_CNT and EE_Credit_CNT. Both counters are initialized to zero during login.

FC Flow Control: Each time that a port sends a frame, it increments BB_Credit_CNT and EE_Credit_CNT by one. When it receives R_RDY from the adjacent port, it decrements BB_Credit_CNT by one. When it receives ACK from the destination port, it decrements EE_Credit_CNT by one.

At any time, if BB_Credit_CNT becomes equal to the BB_Credit, or EE_Credit_CNT becomes equal to the EE_Credit of the receiving port, the transmitting port stops sending frames until the respective count is decremented.

The previous statements are true for a Class 2 service. Class 1 is a dedicated connection. Therefore, BB_Credit is not important, and only EE_Credit is used (EE Flow Control). However, Class 3 is an unacknowledged service, so it uses only BB_Credit (BB Flow Control), but the mechanism is the same in all cases.

The number of buffers is an important factor in overall performance. You need enough buffers to ensure that the transmitting port can continue to send frames without stopping to use the full bandwidth, which is true with distance. The total amount of buffer credit that is needed to optimize the throughput depends on the link speed and the average frame size.

For example, consider an 8 Gbps link connecting two switches that are 100 km apart. At 8 Gbps, a full frame (2148 bytes) occupies about 0.51 km of fiber. In a 100 km link, you can send 198 frames before the first one reaches its destination. You need an ACK to go back to the start to fill EE_Credit again. You can send another 198 frames before you receive the first ACK.

You need at least 396 buffers to allow for nonstop transmission at 100 km distance. The maximum distance that can be achieved at full performance depends on the capabilities of the FC node that is attached at either end of the link extenders, which are vendor-specific. A match should occur between the buffer credit capability of the nodes at either end of the extenders.

Remote system ports and zoning considerations

Ports and zoning requirements for the remote system partnership changed over time.

The current preferred configuration is based on information that is available at [IBM Support](#).

The preferred practice for IBM Storage Virtualize is to provision dedicated node ports for local node-to-node traffic (by using port masking) and isolate GM node-to-node traffic between the local nodes from other local SAN traffic.

Remote port masking: To isolate the node-to-node traffic from the remote copy traffic, the local and remote port masking implementation is preferable.

This configuration of local node port masking is less of a requirement on nonclustered IBM FlashSystem systems, where traffic between node canisters in an I/O group is serviced by the dedicated PCI inter-canister link in the enclosure. The following guidelines apply to the remote system connectivity:

- ▶ The minimum requirement to establish a remote copy partnership is to connect at least one node per system. When remote connectivity among all the nodes of both systems is not available, the nodes of the local system that are not participating in the remote partnership use the node or nodes that are defined in the partnership as a bridge to transfer the replication data to the remote system.

This replication data transfer occurs through the node-to-node connectivity. This configuration, even though it is supported, allows the replication traffic to go through the node-to-node connectivity, which is not recommended.

- ▶ Partnered systems should use the same number of nodes in each system for replication.
- ▶ For maximum throughput, all nodes in each system should be used for replication, both in terms of balancing the preferred node assignment for volumes and for providing inter-system FC connectivity.
- ▶ Where possible, use the minimum number of partnerships between systems. For example, assume site A contains systems A1 and A2, and site B contains systems B1 and B2. In this scenario, creating separate partnerships between pairs of systems (such as A1-B1 and A2-B2) offers greater performance for GM replication between sites than a configuration with partnerships defined between all four systems.

For zoning, the following rules for the remote system partnership apply:

- ▶ For remote copy configurations where the round-trip latency between systems is less than 80 milliseconds, zone two FC ports on each node in the local system to two FC ports on each node in the remote system.
- ▶ For remote copy configurations where the round-trip latency between systems is more than 80 milliseconds, apply SAN zoning to provide separate intrasystem zones for each local-remote I/O group pair that is used for replication, as shown in Figure 6-25.

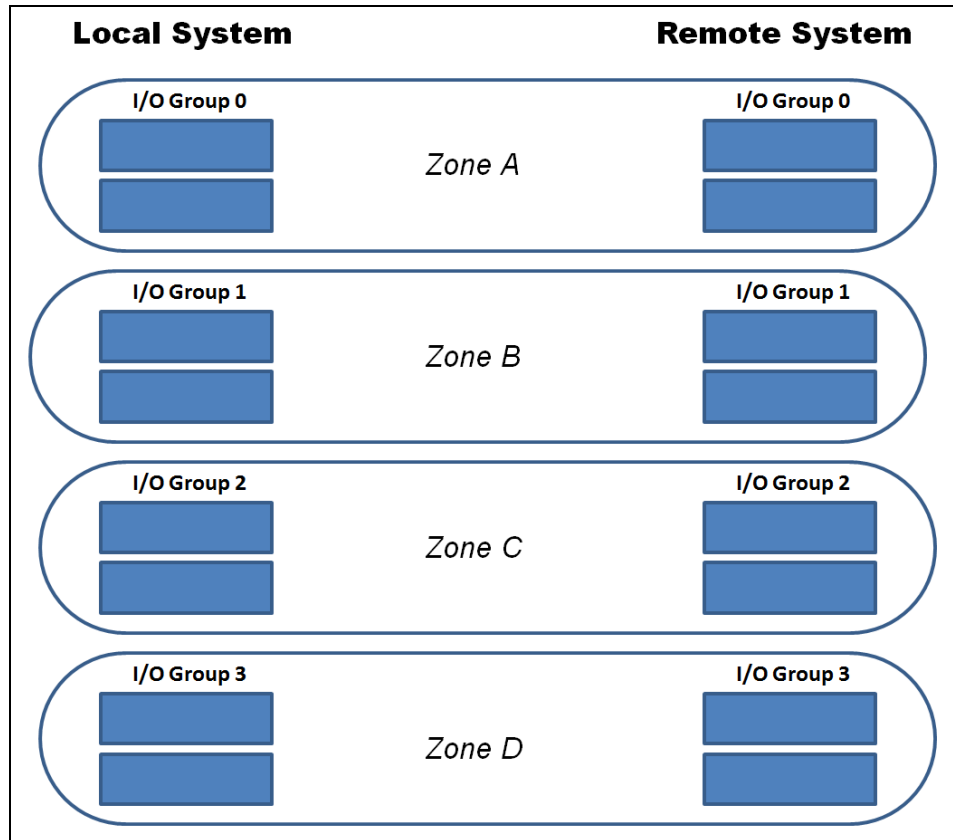


Figure 6-25 Zoning scheme for >80 ms remote copy partnerships

N_Port ID Virtualization (NPIV): IBM Storage Virtualize systems with the NPIV feature enabled provide virtual worldwide port names (WWPNs) for the host zoning. These WWPNs are intended for host zoning only, and they cannot be used for the remote copy partnership.

SAN extension design considerations

DR solutions based on remote copy technologies require reliable SAN extensions over geographical links. To avoid single points of failure, multiple physical links are usually implemented. When implementing these solutions, particular attention must be paid to the remote copy network connectivity setup.

Consider a typical implementation of a remote copy connectivity by using ISLs, as shown in Figure 6-26.

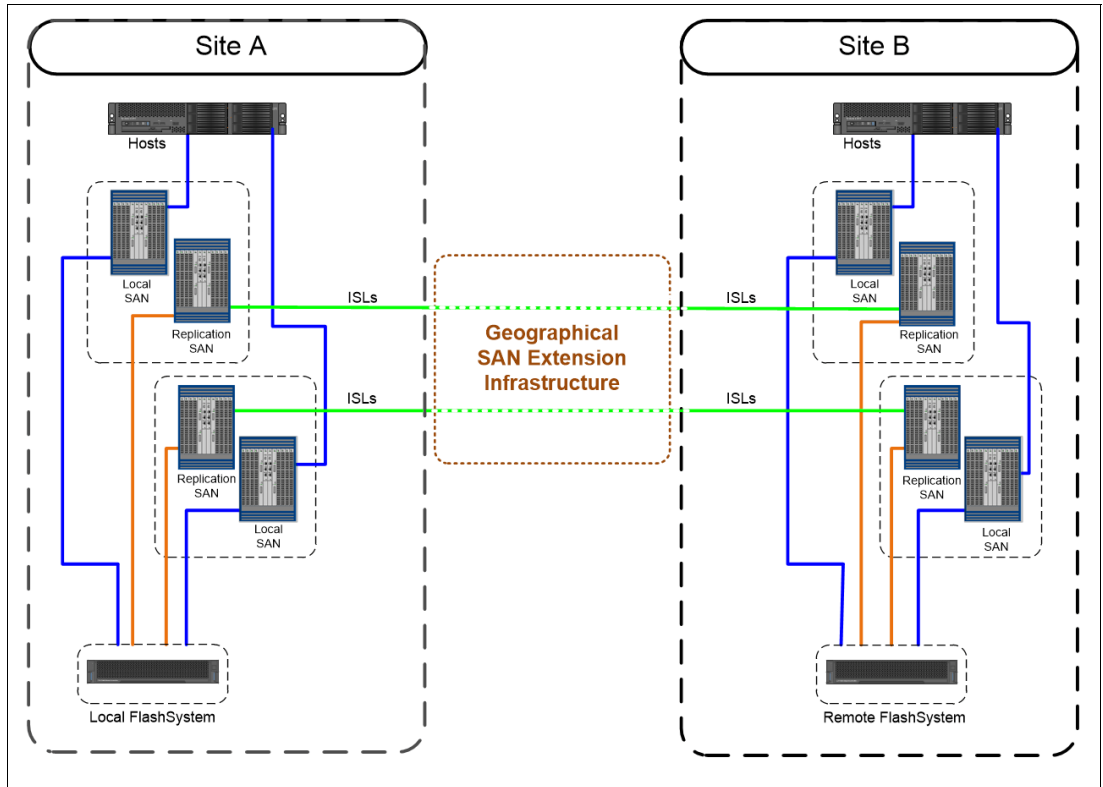


Figure 6-26 Typical remote copy network configuration

In the configuration that is shown in Figure 6-26 on page 427, the remote copy network is isolated in a replication SAN that interconnects Site A and Site B through a SAN extension infrastructure through two physical links. Assume that, for redundancy reasons, two ISLs are used for each fabric for the replication SAN extension.

There are two possible configurations to interconnect the replication SANs. In configuration 1, as shown in Figure 6-27, one ISL per fabric is attached to each physical link through xWDM or FCIP routers. In this case, the physical paths Path A and Path B are used to extend both fabrics.

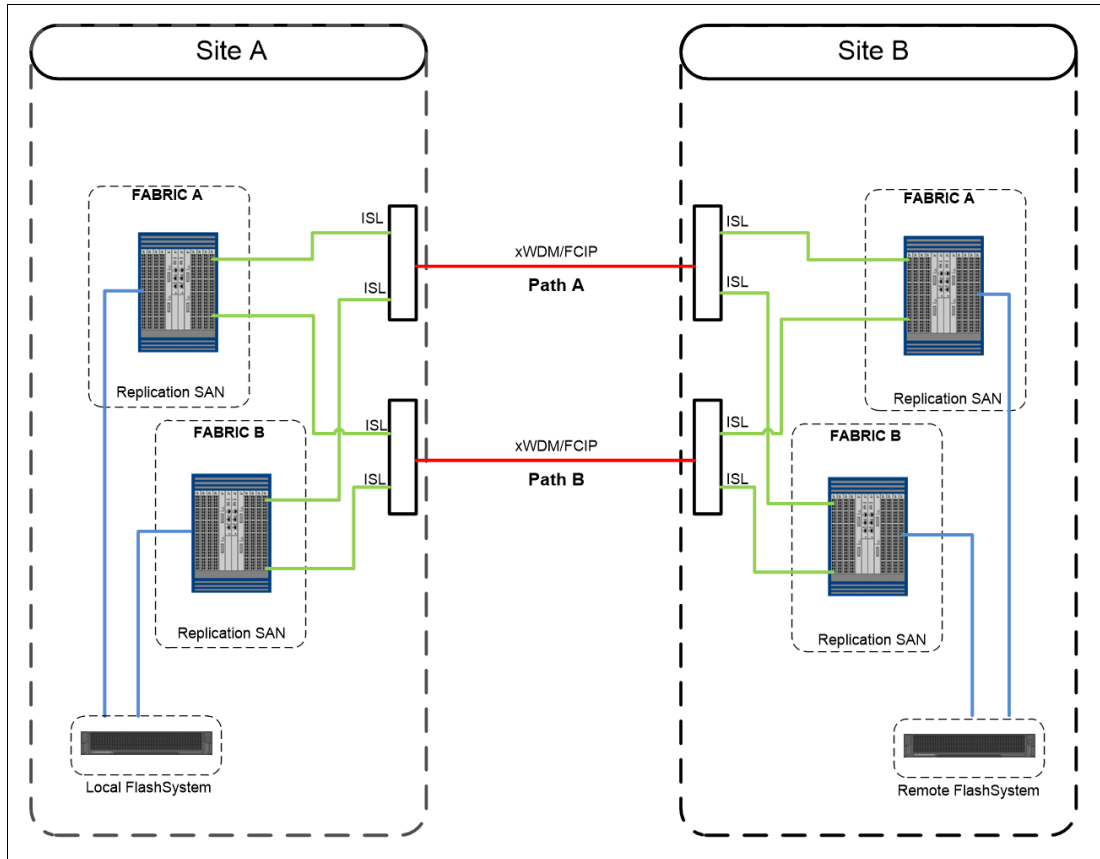


Figure 6-27 Configuration 1: Physical paths shared among the fabrics

In configuration 2, the physical paths are not shared between the fabrics, as shown in Figure 6-28.

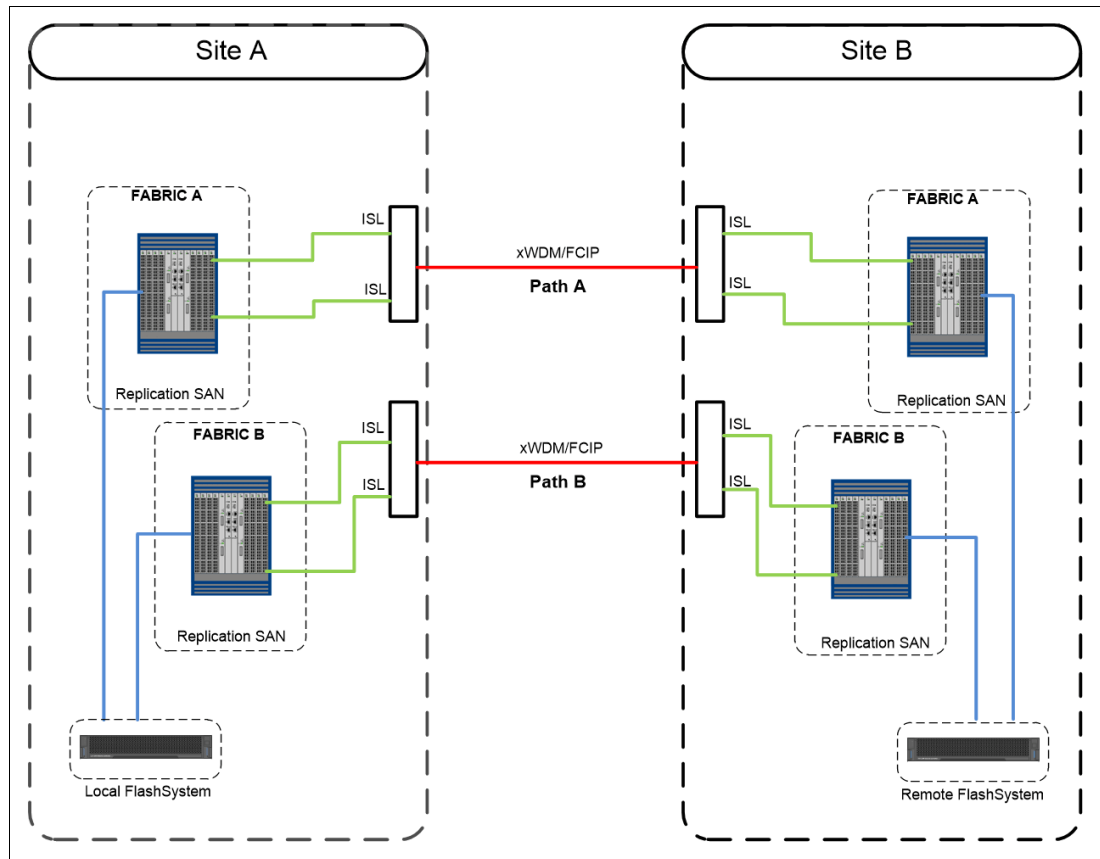


Figure 6-28 Configuration 2: Physical paths not shared among the fabrics

With configuration 1, in a failure of one of the physical paths, both fabrics are simultaneously affected, and a fabric reconfiguration occurs because of an ISL loss. This situation might lead to a temporary disruption of the remote copy communication, and in the worst case to partnership loss condition. To mitigate this situation, link aggregation features like Brocade ISL trunking can be implemented.

With configuration 2, a physical path failure leads to a fabric segmentation of one of the two fabrics, leaving the other fabric unaffected. In this case, the remote copy communication is ensured through the unaffected fabric.

You should fully understand the implication of a physical path or xWDM or FCIP router loss in the SAN extension infrastructure and implement the appropriate architecture to avoid a simultaneous impact.

6.4.4 Remote copy services planning

When you plan for remote copy services, you must keep in mind the considerations that are outlined in the following sections.

Remote copy configurations limits

To plan for and implement remote copy services, you must check the configuration limits and adhere to them. Table 6-8 shows the limits for a system that currently apply to IBM FlashSystem 8.5. Check the online documentation because these limits might change.

Table 6-8 Remote copy maximum limits

Remote copy property	Maximum	Comment
Remote copy (MM and GM) relationships per system	10000	This configuration can be any mix of MM and GM relationships.
Active-active relationships	2000	The limit for the number of HyperSwap volumes in a system.
Remote copy relationships per CG (<= 256 GMCV relationships are configured.)	None	No limit is imposed beyond the remote copy relationships per system limit. Apply to GM and MM.
Remote copy relationships per CG (> 256 GMCV relationships are configured.)	200	
Remote copy CGs per system	256	
Total MM and GM volume capacity per I/O group	2048 TB	The total capacity for all master and auxiliary volumes in the I/O group.
Total number of GMCV relationships per system	256	60 s cycle time.
	2500	300 s cycle time.
3-site remote copy (MM) relationships per CG	256	
3-site remote copy (MM) CGs per system	16	
3-site remote copy (MM) relationships per system	1250	
Inter-cluster IP partnerships per system	3	A system may be connected to up to three remote systems.
Inter-site links per IP partnership	2	A maximum of two inter-site links can be used between two IP partnership sites.
Ports per node	1	A maximum of one port per node can be used for IP partnership.

Like FlashCopy, remote copy services require memory to allocate the bitmap structures that are used to track the updates while volumes are suspended or synchronizing. The default amount of memory for remote copy services is 20 MB. This value can be increased or decreased by using the `chiogrp` command. The maximum amount of memory that can be specified for remote copy services is 512 MB. The grain size for the remote copy services is 256 KB.

Partnerships between systems for MM or GM replication can be used with both FC and native Ethernet connectivity. Distances greater than 300 meters are supported only when using an FCIP link or FC between source and target.

Table 6-9 shows the configuration limits for clustering and HyperSwap over FC and Ethernet.

Table 6-9 Configuration limits for clustering and HyperSwap over FC and Ethernet

Clustering over Fibre Channel	Clustering over 25-gigabit Ethernet (GbE)	HyperSwap over Fibre Channel	HyperSwap over Ethernet (25 Gb only)	Metro/Global Mirror replication over Fibre Channel	Metro/Global Mirror replication over Ethernet (10 Gb or 25 GB)
Yes (up to two I/O groups)	Yes (up to two I/O groups)	Yes (up to 2 I/O groups)	Yes (up to 2 I/O groups)	Yes	Yes

Remote copy general restrictions

To use MM and GM, you must adhere to the following rules:

- ▶ You must have the same size for the source and target volumes when defining a remote copy relationship. However, the target volume can be a different type (image, striped, or sequential mode) or have different cache settings (cache-enabled or cache-disabled).
- ▶ You cannot move remote copy source or target volumes to different I/O groups.
- ▶ Remote copy volumes can be resized with the following restrictions:
 - Resizing applies to MM and GM only. GMCV is not supported.
 - The Remote Copy Consistency Protection feature is not allowed and must be removed before resizing the volumes.
 - The remote copy relationship must be in the synchronized status.
 - The resize order must ensure that the target volume is always larger than the source volume.

Note: Volume expansion for MM and GM volumes was introduced with IBM Storage Virtualize 7.8.1 with some restrictions:

- ▶ In the first implementation (up to 8.2.1), only thin-provisioned or compressed volumes were supported.
- ▶ With 8.2.1, nonmirrored fully allocated volumes also were supported.
- ▶ With 8.4, all the restrictions on volume type were removed.

- ▶ You can mirror intrasystem MM or GM only between volumes in the same I/O group.

Intrasystem remote copy: Intrasystem GM is not supported on IBM Storage Virtualize based systems running version 6 or later.

- ▶ GM is not recommended for cache-disabled volumes that are participating in a GM relationship.

Changing the remote copy type

Changing the remote copy type for an existing relationship is an easy task. It is enough to stop the relationship, if it is active, and change the properties to set the new remote copy type. Remember to create the CVs in case of a change from MM or GM to GMCV.

Interaction between remote copy and FlashCopy

Remote copy functions can be used with the FlashCopy function so that you can have both functions operating concurrently on the same volume. The possible combinations between remote copy and FlashCopy are as follows:

- ▶ Remote copy source:
 - A remote copy source can be a FlashCopy source.
 - A remote copy source can be a FlashCopy target with the following restrictions:
 - A FlashCopy target volume cannot be updated while it is the source volume of an MM or GM relationship that is actively mirroring. A FlashCopy mapping cannot be started while the target volume is in an active remote copy relationship.
 - The I/O group for the FlashCopy mappings must be the same as the I/O group for the FlashCopy target volume (that is the I/O group of the remote copy source).
- ▶ Remote copy target:
 - A remote copy target can be a FlashCopy source.
 - A remote copy target can be a FlashCopy target if the FlashCopy mapping is in the `idle_copied` state when its target volume is the target volume of an active MM or GM relationship.

When implementing FlashCopy functions for volumes in GMCV relationships, FlashCopy multi-target mappings are created. As described in “Interaction and dependency between multiple Target FlashCopy mappings” on page 380, this creating results in dependent mappings that can affect the cycle formation due to the cleaning process. For more information, see “**Cleaning process and cleaning rate**” on page 390.

With such configurations, it is a best practice to set the cleaning rate as needed. This best practice also applies to Consistency Protection volumes and HyperSwap configurations.

Native back-end controller copy functions considerations

IBM Storage Virtualize provides a widespread set of copy services functions that cover most client requirements.

However, some storage controllers can provide specific copy services capabilities that are not available with the current version of IBM Spectrum Virtualize. IBM Storage Virtualize addresses these situations by using cache-disabled image mode volumes that virtualize LUNs that participate in the native back-end controller’s copy services relationships.

Keeping the cache disabled ensures data consistency throughout the I/O stack, from the host to the back-end controller. Otherwise, by leaving the cache enabled on a volume, the underlying controller does not receive any write I/Os as the host writes them. IBM Storage Virtualize caches them and processes them later. This process can have more ramifications if a target host depends on the write I/Os from the source host as they are written.

Note: Native copy services are not supported on all storage controllers. For more information about the known limitations, see [Using Native Controller Copy Services](#).

As part of its copy services function, the storage controller might take a LUN offline or suspend reads or writes. IBM Storage Virtualize does not recognize why this process happens. Therefore, it might log errors when these events occur. For this reason, if IBM Storage Virtualize must detect the LUN, ensure that the LUN remains in the unmanaged state until full access is granted.

Native back-end controller copy services can also be used for LUNs that are not managed by IBM Spectrum Virtualize. Accidental incorrect configurations of the back-end controller copy services involving IBM Storage Virtualize attached LUNs can produce unpredictable results.

For example, if you accidentally use a LUN with IBM Storage Virtualize data on it as a point-in-time target LUN, you can corrupt that data. Moreover, if that LUN was a MDisk in a managed-disk group with striped or sequential volumes on it, the MDisk group might be brought offline. This situation makes all the volumes that belong to that group go offline, leading to a widespread host access disruption.

Remote copy and code upgrade considerations

When you upgrade system software where the system participates in one or more inter-system relationships, upgrade only one cluster at a time. Do *not* upgrade the systems concurrently.

Attention: Upgrading both systems concurrently is not monitored by the software upgrade process.

Allow the software upgrade to complete on one system before you start it on the other system. Upgrading both systems concurrently can lead to a loss of synchronization. In stress situations, it can further lead to a loss of availability.

Usually, pre-existing remote copy relationships are unaffected by a software upgrade that is performed correctly. However, always check in the target code release notes for special considerations on the copy services.

Although it is not a best practice, a remote copy partnership can be established with some restrictions among systems with different IBM Storage Virtualize versions. For more information, see [IBM Storage Virtualize Family of Products Inter-System Metro Mirror and Global Mirror Compatibility Cross Reference](#).

Volume placement considerations

You can optimize the distribution of volumes within I/O groups at the local and remote systems to maximize performance.

Although defined at a system level, the partnership bandwidth and the background copy rate are evenly divided among the cluster's I/O groups. The available bandwidth for the background copy can be used by either nodes or shared by both nodes within the I/O group.

This bandwidth allocation is independent from the number of volumes for which a node is responsible. Each node divides its bandwidth evenly between the (multiple) remote copy relationships with which it associates volumes that are performing a background copy.

Volume preferred node

Conceptually, a connection (path) goes between each node on the primary system to each node on the remote system. Write I/O, which is associated with remote copying, travels along this path. Each node-to-node connection is assigned a finite amount of remote copy resources and can sustain only in-flight write I/O to this limit.

The node-to-node in-flight write limit is determined by the number of nodes in the remote system. The more nodes that exist at the remote system, the lower the limit is for the in-flight write I/Os from a local node to a remote node. Less data can be outstanding from any one local node to any other remote node. To optimize performance, GM volumes must have their preferred nodes distributed evenly between the nodes of the systems.

The preferred node property of a volume helps to balance the I/O load between nodes in that I/O group. This property is also used by remote copy to route I/O between systems.

The IBM Storage Virtualize node that receives a write for a volume is normally the preferred node of the volume. For volumes in a remote copy relationship, that node is responsible for sending that write to the preferred node of the target volume. The primary preferred node is responsible for sending any writes that relate to the background copy. Again, these writes are sent to the preferred node of the target volume.

Each node of the remote system has a fixed pool of remote copy system resources for *each* node of the primary system. Each remote node has a separate queue for I/O from each of the primary nodes. This queue is a fixed size and is the same size for every node. If preferred nodes for the volumes of the remote system are set so that every combination of primary node and secondary node is used, remote copy performance is maximized.

Figure 6-29 shows an example of remote copy resources that are not optimized. Volumes from the local system are replicated to the remote system. All volumes with a preferred node of Node 1 are replicated to the remote system, where the target volumes also have a preferred node of Node 1.

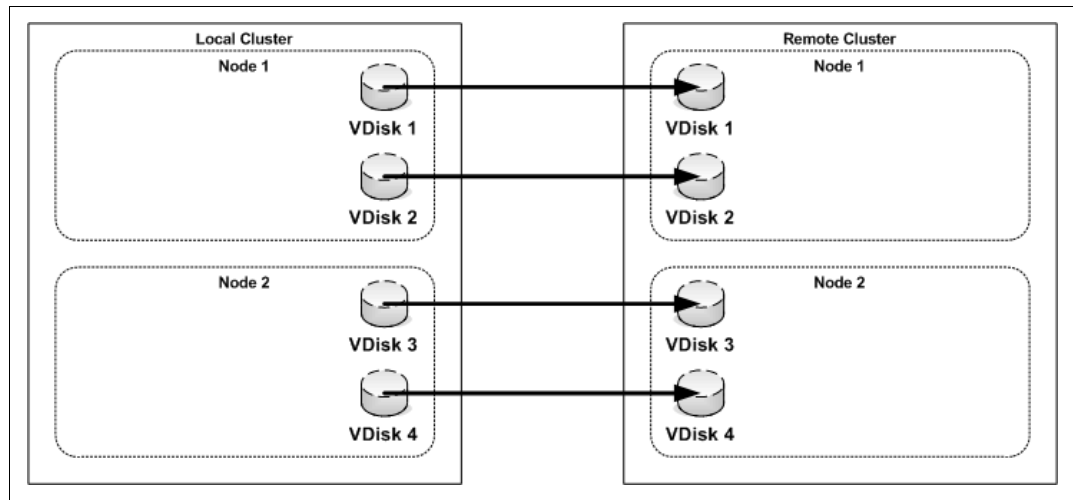


Figure 6-29 Remote copy resources that are not optimized

With the configuration that is shown in Figure 6-29, the resources for remote system Node 1 that are reserved for local system Node 2 are not used. Also, the resources for local system Node 1 that are reserved for remote system Node 2 are not used.

If the configuration that is shown in Figure 6-29 changes to the configuration that is shown in Figure 6-30 on page 435, all remote copy resources for each node are used, and remote copy operates with better performance.

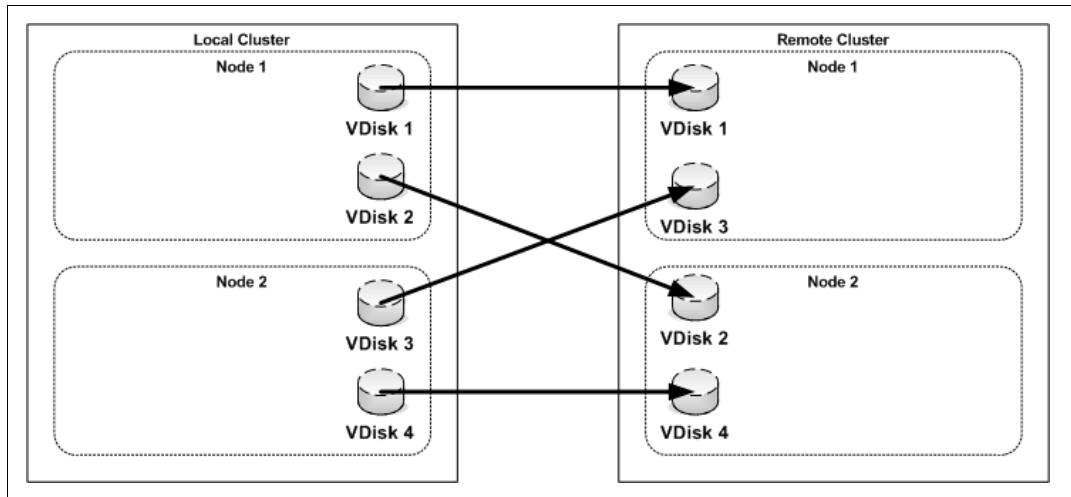


Figure 6-30 Optimized Global Mirror resources

GMCV number of volume per consistency group considerations

GMCV uses FlashCopy technology to have a consistent point in time copy of data, which is replicated periodically to DR site. While taking this periodic FlashCopy, the system quiesces I/O operations for volumes (which are part of this replication CG), which can lead to an increase in Peak I/O Response time. However, it does not affect average I/O Response Time. Nonreplicated volumes are not affected in this process.

Note: The pause period is short. I/O is paused only while FlashCopy mapping is being prepared.

Peak I/O Response Time varies based on the number of relationships in a CG. Lower the number of relationships in a CG to make Peak I/O Response Time better. It is a best practice to have fewer volumes per CG wherever possible. Table 6-10 shows the relative Peak I/O Response Time with the number of relationships per CG.

Table 6-10 Relative Peak I/O Response Time with number of relationships per CG

Relationships per CG	Peak I/O Response Time (Approximate)
1	1.0x
25	1.2x
50	2.0x
150	3.0x
256	5.0x

GMCV change volumes placement considerations

The CVs in a GMCV configuration are thin-provisioned volumes that are used as FlashCopy targets. For this reason, the same considerations apply that are described in “Volume placement considerations” on page 387. The CVs can be compressed to reduce the amount of space that is used, but note that the CVs might be subject to a heavy write workload both in the primary and secondary system.

Therefore, the placement on the back end is critical to provide adequate performance. Consider using DRP for the CVs only if it is beneficial in terms of space savings.

Tip: The internal FlashCopy that is used by the GMCV uses a 256 KB grain size. However, it is possible to force a 64 KB grain size by creating a FlashCopy with a 64 KB grain size from the GMCV volume and a dummy target volume before assigning the CV to the relationship. You can do this procedure for both the source and target volumes. After the CV assignment is done, the dummy FlashCopy can be deleted.

Background copy considerations

The remote copy partnership bandwidth parameter *explicitly* defines the rate at which the background copy is attempted, but also *implicitly* affects foreground I/O. Background copy bandwidth can affect foreground I/O latency in one of the following ways:

- ▶ Increasing latency of foreground I/O

If the remote copy partnership bandwidth parameter is set too high for the actual inter-system network capability, the background copy resynchronization writes use too much of the inter-system network. The writes starve the link of the ability to service synchronous or asynchronous mirrored foreground writes. Delays in processing the mirrored foreground writes increase the latency of the foreground I/O as perceived by the applications.

- ▶ Read I/O overload of primary storage

If the remote copy partnership background copy rate is set too high, the added read I/Os that are associated with background copy writes can overload the storage at the primary site and delay foreground (read/write) I/Os.

- ▶ Write I/O overload of auxiliary storage

If the remote copy partnership background copy rate is set too high for the storage at the secondary site, the background copy writes overload the auxiliary storage. Again, they delay the synchronous and asynchronous mirrored foreground write I/Os.

Important: An increase in the peak foreground workload can have a detrimental effect on foreground I/O by pushing more mirrored foreground write traffic along the inter-system network, which might not have the bandwidth to sustain it. It can also overload the primary storage.

To set the background copy bandwidth optimally, consider all aspects of your environments, starting with the following biggest contributing resources:

- ▶ Primary storage
- ▶ Inter-system network bandwidth
- ▶ Auxiliary storage

Provision the most restrictive of these three resources between the background copy bandwidth and the peak foreground I/O workload. Perform this provisioning by calculation or by determining experimentally how much background copy can be allowed before the foreground I/O latency becomes unacceptable.

Then, reduce the background copy to accommodate peaks in workload. In cases where the available network bandwidth cannot sustain an acceptable background copy rate, consider alternatives to the initial copy, as described in “Initial synchronization options and offline synchronization” on page 437.

Changes in the environment or increasing its workload can affect the foreground I/O. IBM Storage Virtualize provides a means to monitor and a parameter to control how foreground I/O is affected by running remote copy processes. IBM Storage Virtualize monitors the delivery of the mirrored foreground writes.

If latency or performance of these writes extends beyond a (predefined or client-defined) limit for a period, the remote copy relationship is suspended. For more information, see “1920 error” 6.5.6, “Replication portsets” on page 463.

Finally, with GMCV, the cycling process that transfers the data from the local to the remote system is a background copy task. For more information, see “Global Mirror and GMCV coexistence considerations” on page 421. For this reason, the background copy rate and the `relationship_bandwidth_limit` setting affect the available bandwidth during the initial synchronization and the normal cycling process.

Background copy bandwidth allocation: As described in “Volume placement considerations” on page 433 the available bandwidth of a remote copy partnership is evenly divided among the cluster’s I/O groups. In a case of unbalanced distribution of the remote copies among the I/O groups, the partnership bandwidth should be adjusted to reach the needed background copy rate.

Consider, for example, a 4-I/O group cluster that has a partnership bandwidth of 4,000 Mbps and a background copy percentage of 50%. The expected maximum background copy rate for this partnership is 250 MBps.

Because the available bandwidth is evenly divided among the I/O groups, every I/O group in this cluster can theoretically synchronize data at a maximum rate of approximately 62 MBps (50% of 1,000 Mbps). Now, in an edge case where only volumes from one I/O group are replicated, the partnership bandwidth should be adjusted to 16,000 Mbps to reach the full background copy rate (250 MBps).

Initial synchronization options and offline synchronization

When creating a remote copy relationship, two options regarding the initial synchronization process are available:

- ▶ The **not synchronized** option is the default. With this option, when a remote copy relationship starts, a full data synchronization at the background copy rate occurs between the source and target volumes. It is the simplest approach because apart from issuing the necessary IBM Storage Virtualize commands, other administrative activity is not required. However, in some environments, the available bandwidth makes this option unsuitable.
- ▶ The **already synchronized** option does not force any data synchronization when the relationship starts. The administrator must ensure that the source and target volumes contain identical data before a relationship is created. The administrator can perform this check in one of the following ways:
 - Format both volumes to change all data to zero.
 - Copy a complete tape image (or other method of moving data) from one disk to the other.

In either technique, write I/O must not take place to the source and target volume before the relationship is established. Then, the administrator must complete the following actions:

- Create the relationship with the synchronized settings (`-sync` option).
- Start the relationship.

Attention: If you do not perform these steps correctly, the remote copy reports the relationship as being *consistent* when it is not. This setting is likely to cause auxiliary volumes to be useless.

By understanding the methods to start an MM and GM relationship, you can use one of them as a means to implement the remote copy relationship saving bandwidth.

Consider a situation where you have a large source volume (or many source volumes) containing already active data that you want to replicate to a remote site. Your planning shows that the mirror initial-sync time takes too long (or is too costly if you pay for the traffic that you use). In this case, you can set up the sync by using another medium that is less expensive. This synchronization method is called *offline synchronization*.

This example uses tape media as the source for the initial sync for the MM relationship or the GM relationship target before it uses remote copy services to maintain the MM or GM. This example does not require downtime for the hosts that use the source volumes.

Before you set up GM relationships and save bandwidth, complete the following steps:

1. Ensure that the hosts are running and using their volumes normally. The MM relationship or GM relationship is not yet defined.

Identify all volumes that become the source volumes in an MM relationship or in a GM relationship.

2. Establish the remote copy partnership with the target IBM Storage Virtualize system.

To set up GM or MM relationships and save bandwidth, complete the following steps:

1. Define an MM relationship or a GM relationship for each source disk. When you define the relationship, ensure that you use the **-sync** option, which stops the system from performing an initial sync.

Attention: If you do not use the **-sync** option, all these steps are redundant because the IBM Storage Virtualize system performs a full initial synchronization.

2. Stop each mirror relationship by using the **-access** option, which enables write access to the target volumes. You need write access later.
3. Copy the source volume to the alternative media by using the **dd** command to copy the contents of the volume to tape. Another option is to use your backup tool (for example, IBM Spectrum Protect) to make an image backup of the volume.

Change tracking: Although the source is modified while you copy the image, the IBM Storage Virtualize software is tracking those changes. The image that you create might have some of the changes and is likely to miss some of the changes.

When the relationship is restarted, IBM Storage Virtualize applies all the changes that occurred since the relationship stopped in step 2. After all the changes are applied, you have a consistent target image.

4. Ship your media to the remote site and apply the contents to the targets of the MM or GM relationship. You can mount the MM and GM target volumes to a UNIX server and use the **dd** command to copy the contents of the tape to the target volume.

If you used your backup tool to make an image of the volume, follow the instructions for your tool to restore the image to the target volume. Remember to remove the mount if the host is temporary.

Tip: It does not matter how long it takes to get your media to the remote site to perform this step. However, the faster that you can get the media to the remote site and load it, the quicker that the IBM Storage Virtualize system starts running and maintaining the MM and GM.

5. Unmount the target volumes from your host. When you start the MM and GM relationships later, IBM Storage Virtualize stops write-access to the volume while the mirror relationship is running.
6. Start your MM and GM relationships. The relationships must be started with the **-clean** parameter. This way, changes that are made on the secondary volume are ignored. Only changes that are made on the clean primary volume are considered when synchronizing the primary and secondary volumes.
7. While the mirror relationship catches up, the target volume is not usable at all. When it reaches the `ConsistentSynchnonized` status, your remote volume is ready for use in a disaster.

Back-end storage considerations

To reduce the overall solution costs, it is a common practice to provide the remote systems with lower performance characteristics compared to the local system, especially when using asynchronous remote copy technologies. This approach can be risky, especially when using the GM or MM technology where the application performances at the primary system can be limited by the performance of the remote system.

The best practice is to perform an accurate back-end resource sizing for the remote system to fulfill the following capabilities:

- ▶ The peak application workload to the GM or MM volumes
- ▶ The defined level of background copy
- ▶ Any other I/O that is performed at the remote site

Remote copy tunable parameters

Several commands and parameters help to control remote copy and its default settings. You can display the properties and features of the systems by using the **lssystem** command. Also, you can change the features of systems by using the **chsystem** command.

relationshipbandwidthlimit

The **relationshipbandwidthlimit** parameter is an optional parameter that specifies the new background copy bandwidth of 1 - 1000 MBps. The default is 25 MBps. This parameter operates system-wide, and defines the maximum background copy bandwidth that any relationship can adopt. The existing background copy bandwidth settings that are defined on a partnership continue to operate with the lower of the partnership and volume rates attempted.

Important: Do not set this value higher than the default without establishing that the higher bandwidth can be sustained.

The **relationshipbandwidthlimit** parameter also applies to MM relationships.

gmlinktolerance and gmmaxhostdelay

The **gmlinktolerance** and **gmmaxhostdelay** parameters are critical for deciding internally whether to terminate a relationship due to a performance problem. In most cases, these two parameters must be considered in tandem. The defaults do not normally change unless you have a specific reason to do so.

The **gmlinktolerance** parameter can be thought of as how long you allow the host delay to go on being significant before you decide to terminate a GM volume relationship. This parameter accepts values of 20 - 86,400 seconds in increments of 10 seconds. The default is 300 seconds. You can disable the link tolerance by entering a value of zero for this parameter.

The **gmmaxhostdelay** parameter can be thought of as the maximum host I/O impact that is due to GM, that is, how long that local I/O would take with GM turned off, and how long does it take with GM turned on. The difference is the host delay due to the GM tag and forward processing.

Although the default settings are adequate for most situations, increasing one parameter while reducing another one might deliver a tuned performance environment for a particular circumstance.

Example 6-1 shows how to change the **gmlinktolerance** and **gmmaxhostdelay** parameters by using the **chsystem** command.

Example 6-1 Changing gmlinktolerance to 30 and gmmaxhostdelay to 100

```
chsystem -gmlinktolerance 30
chsystem -gmmaxhostdelay 100
```

Test and monitor: To reiterate, thoroughly test and carefully monitor the host impact of any changes before putting them into a live production environment.

For more information about settings considerations for the **gmlinktolerance** and **gmmaxhostdelay** parameters, see 6.5.6, “Replication portsets” on page 463.

rcbuffersize

The **rcbuffersize** parameter was introduced to manage workloads with intense and bursty write I/O that do not fill the internal buffer while GM writes are undergoing sequence tagging.

Important: Do not change the **rcbuffersize** parameter except under the direction of IBM Support.

Example 6-2 shows how to change **rcbuffersize** to 64 MB by using the **chsystem** command. The default value for **rcbuffersize** is 48 MB, and the maximum value is 512 MB.

Example 6-2 Changing rcbuffersize to 64 MB

```
chsystem -rcbuffersize 64
```

Any extra buffers that you allocate are taken away from the general cache.

maxreplicationdelay and partnershipexclusionthreshold

maxreplicationdelay is a system-wide parameter that defines a maximum latency (in seconds) for individual writes that pass through the GM logic. If a write is hung for the specified amount of time, for example, due to a rebuilding array on the secondary system, GM stops the relationship (and any containing CG), which triggers a 1920 error.

The **partnershipexclusionthreshold** parameter was introduced so that users can set the timeout for an I/O that triggers a temporarily dropping of the link to the remote cluster. The value must be 30 - 315.

Important: Do not change the **partnershipexclusionthreshold** parameter except under the direction of IBM Support.

For more information about settings considerations for the **maxreplicationdelay** parameter, see 6.5.6, “Replication portsets” on page 463.

Link delay simulation parameters

Even though GM is an asynchronous replication method, there can be an impact to server applications due to GM managing transactions and maintaining write order consistency over a network. To mitigate this impact, as a testing and planning feature, you can use GM to simulate the effect of the round-trip delay between sites by using the following parameters:

► **gminterclusterdelaysimulation**

This optional parameter specifies the inter-system delay simulation, which simulates the GM round-trip delay between two systems in milliseconds. The default is 0. The valid range is 0 - 100 milliseconds.

► **gmintraclusterdelaysimulation**

This optional parameter specifies the intrasystem delay simulation, which simulates the GM round-trip delay in milliseconds. The default is 0. The valid range is 0 - 100 milliseconds.

6.4.5 Multiple site remote copy

The most common use cases for the remote copy functions are obviously DR solutions. Code level 8.3.1 introduced further DR capabilities such as *3-site replication*, which provides a solution for coordinated DR across three sites in various topologies. A complete discussion about the DR solutions that are based on IBM Storage Virtualize technology is beyond the intended scope of this book. For an overview of the DR solutions with IBM Storage Virtualize Copy Services, see *IBM System Storage SAN Volume Controller and Storwize V7000 Replication Family Services*, SG24-7574. For a deepening of the 3-site replication, see *IBM Spectrum Virtualize 3-Site Replication*, SG24-8504.

Another typical remote copy use case is data movement among distant locations as required, for example, for data center relocation and consolidation projects. In these scenarios, the IBM Storage Virtualize remote copy technology is particularly effective when combined with the image copy feature that allows data movement among storage systems of different technologies or vendors.

Mirroring scenarios that involve multiple sites can be implemented by using a combination of IBM Storage Virtualize capabilities.

Enhanced Stretched Cluster 3-site mirroring (only applicable to SAN Volume Controller)

With SVC Enhanced Stretched Cluster (ESC), remote copy services can be combined with volume mirroring to implement 3-site solutions, as shown in Figure 6-31.

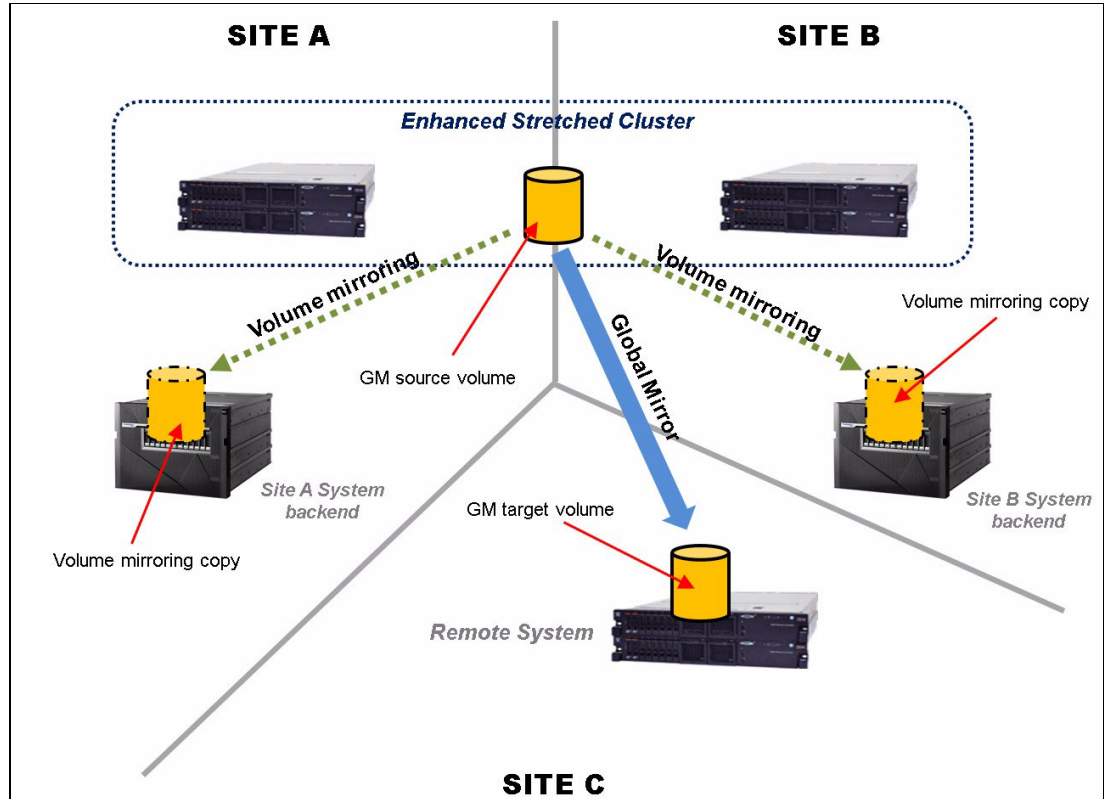


Figure 6-31 Three-site configuration with Enhanced Stretched Cluster

Three-site configurations also can be implemented by using special cascading configurations.

Performing cascading copy service functions

Cascading copy service functions that use IBM Storage Virtualize are not directly supported. However, you might require a three-way (or more) replication by using copy service functions (synchronous or asynchronous mirroring). You can address this requirement both by using IBM Storage Virtualize Copy Services and by combining IBM Storage Virtualize Copy Services (with image mode cache-disabled volumes) and native storage controller copy services.

DRP limitation: Currently, the image mode VDisk is not supported by DRP.

Cascading with native storage controller copy services

Figure 6-32 describes the configuration for 3-site cascading by using the native storage controller copy services in combination with IBM Storage Virtualize remote copy functions.

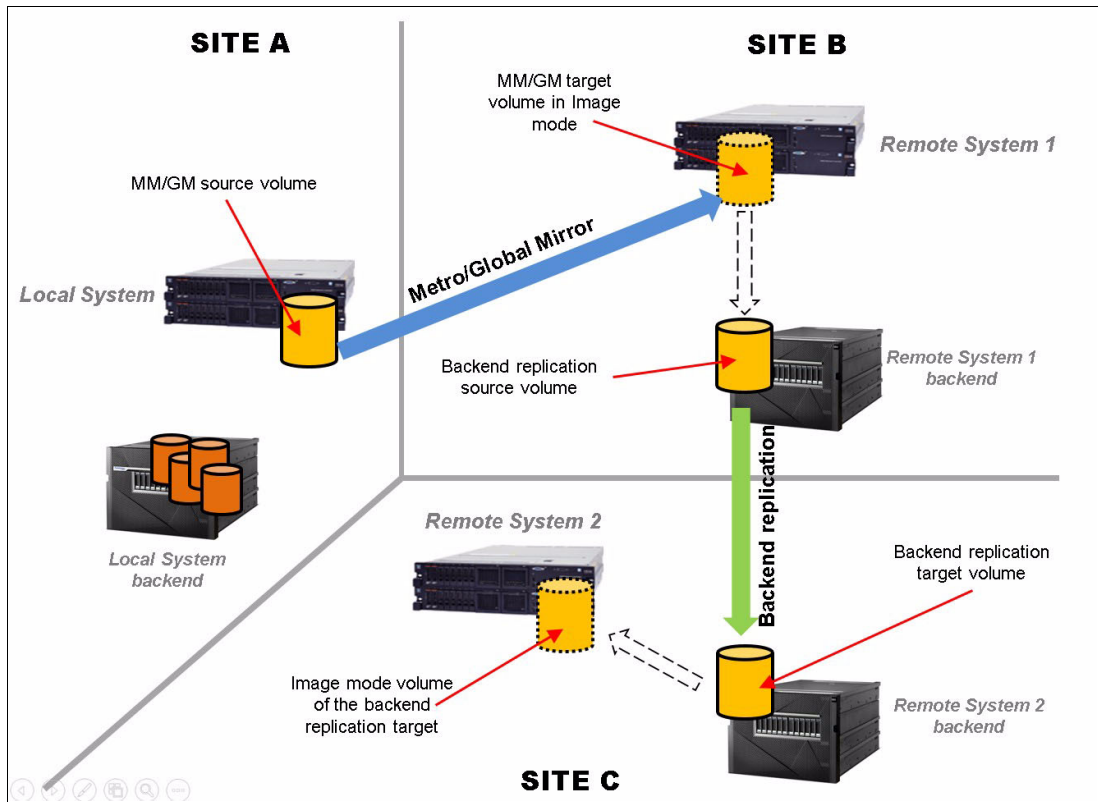


Figure 6-32 Using 3-way copy services

In Figure 6-32, the primary site uses IBM Storage Virtualize remote copy functions (GM or MM) at the secondary site. Therefore, if a disaster occurs at the primary site, the storage administrator enables access to the target volume (from the secondary site) and the business application continues processing.

While the business continues processing at the secondary site, the storage controller copy services replicate to the third site. This configuration is allowed under the following conditions:

- ▶ The back-end controller native copy services must be supported by IBM Spectrum Virtualize. For more information, see “Native back-end controller copy functions considerations” on page 432.
- ▶ The source and target volumes that are used by the back-end controller native copy services must be imported to the IBM Storage Virtualize system as image-mode volumes with the cache disabled.

Cascading with IBM Storage Virtualize Copy Services

Remote copy services cascading is allowed with the IBM Storage Virtualize 3-site replication capability. For more information, see *IBM Spectrum Virtualize 3-Site Replication*, SG24-8504. However, a cascading-like solution is also possible by combining IBM Storage Virtualize Copy Services. These remote copy services implementations are useful in 3-site DR solutions and data center moving scenarios.

In the configuration that is described in Figure 6-33, a GM (MM also can be used) solution is implemented between the Local System at Site A, which is the production site, and the Remote System 1 at Site B, which is the primary DR site. A third system, Remote System 2, is at Site C, which is the secondary DR site. Connectivity is provided between Site A and Site B, between Site B and Site C, and optionally between Site A and Site C.

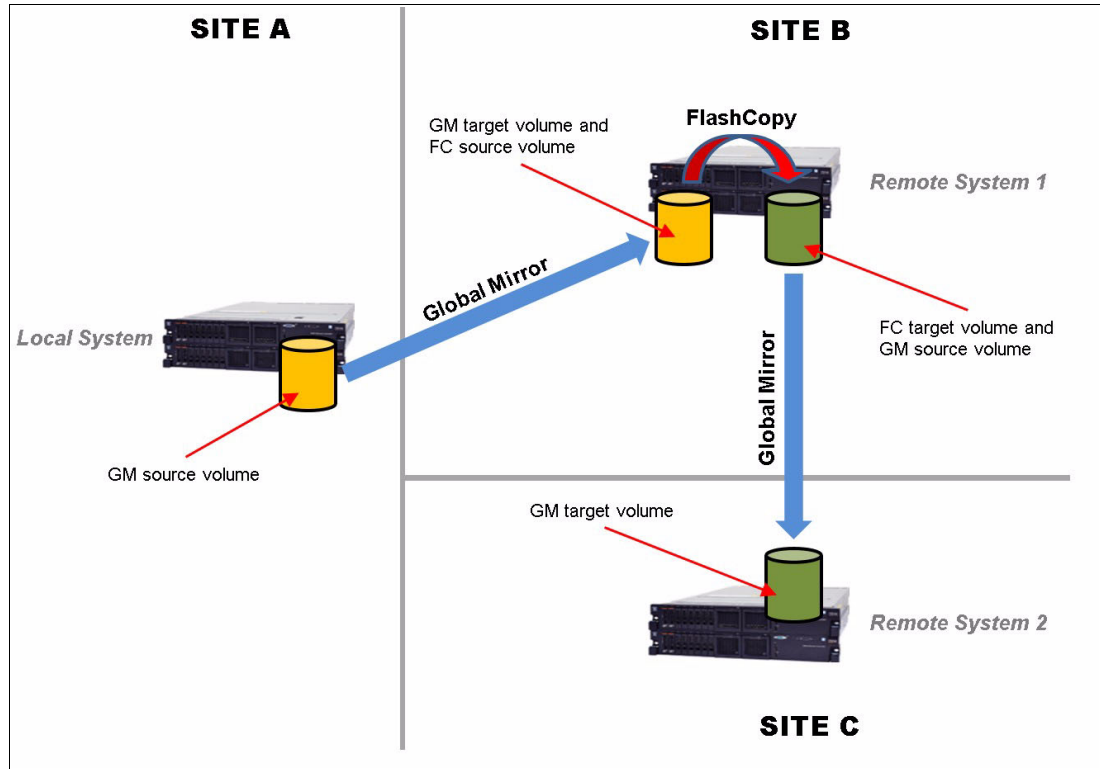


Figure 6-33 Cascading-like infrastructure

To implement a cascading-like solution, the following steps must be completed:

1. Setup phase. To initially set up the environment, complete the following steps:
 - a. Create the GM relationships between the Local System and Remote System 1.
 - b. Create the FlashCopy mappings in the Remote System 1 by using the target GM volumes as FlashCopy source volumes. The FlashCopy must be incremental.
 - c. Create the GM relationships between Remote System 1 and Remote System 2 by using the FlashCopy target volumes as GM source volumes.
 - d. Start the GM from Local System to Remote System 1.

After the GM is in the ConsistentSynchronized state, you are ready to create the cascading-like solution.
2. Consistency point creation phase. The following actions must be performed every time a consistency point creation in the Site C is required:
 - a. Check whether the GM between Remote System 1 and Remote System 2 is in the stopped or idle status, and if it is not, stop the GM.
 - b. Stop the GM between the Local System to Remote System 1.
 - c. Start the FlashCopy in Remote Site 1.
 - d. Resume the GM between the Local System and Remote System 1.
 - e. Start or resume the GM between Remote System 1 and Remote System 2.

The first time that these operations are performed, a full copy between Remote System 1 and Remote System 2 occurs. Later runs of these operations perform incremental resynchronization instead. After the GM between Remote System 1 and Remote System 2 is in the Consistent Synchronized state, the consistency point in Site C is created. The GM between Remote System 1 and Remote System 2 can now be stopped to be ready for the next consistency point creation.

6.4.6 1920 error

An IBM Storage Virtualize based system generates a 1920 error message whenever an MM or GM relationship stops because of adverse conditions. If this condition is left unresolved, it might affect the performance of the foreground I/O.

A 1920 error can occur for many reasons. The condition might be the result of a temporary interruption, such as maintenance on the inter-system connectivity, an unexpectedly higher foreground host I/O workload, or a permanent error because of a hardware failure. It is also possible that not all relationships are affected and that multiple 1920 errors can be posted.

The 1920 error can be triggered for MM *and* GM relationships. However, in MM configurations, the 1920 error is associated with only a permanent I/O error condition. For this reason, the main focus of this section is 1920 errors in a GM configuration.

Internal Global Mirror control policy and raising 1920 errors

Although GM is an asynchronous remote copy service, the local and remote sites have some interplay. When data comes into a local volume, work must be done to ensure that the remote copies are consistent. This work can add a delay to the local write. Normally, this delay is low.

To mitigate the effects of the GM to the foreground I/Os, the IBM Storage Virtualize code implements different control mechanisms for Slow I/O and Hung I/O conditions. The *Slow I/O condition* is a persistent performance degradation on write operations that are introduced by the remote copy logic. The *Hung I/O condition* is a long delay (seconds) on write operations.

Slow I/O condition: gmmaxhostdelay and gmlinktolerance

The `gmmaxhostdelay` and `gmlinktolerance` parameters help to ensure that hosts do not perceive the latency of the long-distance link, regardless of the bandwidth of the hardware that maintains the link or the storage at the secondary site.

In terms of nodes and back-end characteristics, the system configuration must be provisioned so that when combined, they can support the maximum throughput that is delivered by the applications at the primary that uses GM.

If the capabilities of the system configuration are exceeded, the system becomes backlogged, and the hosts receive higher latencies on their write I/O. Remote copy in GM implements a protection mechanism to detect this condition and halts mirrored foreground write and background copy I/O. Suspension of this type of I/O traffic ensures that misconfiguration or hardware problems (or both) do not affect host application availability.

GM attempts to detect and differentiate between backlogs that occur because of the operation of the GM protocol. It does not examine the general delays in the system when it is heavily loaded, where a host might see high latency even if GM were disabled.

GM uses the **gmmaxhostdelay** and **gmlinktolerance** parameters to monitor GM protocol backlogs in the following ways:

- ▶ Users set the **gmmaxhostdelay** and **gmlinktolerance** parameters to control how software responds to these delays. The **gmmaxhostdelay** parameter is a value in milliseconds with a maximum value of 100.
- ▶ Every 10 seconds, GM samples all the GM writes and determines how much of a delay they added. If the delay added by at least a third of these writes is greater than the **gmmaxhostdelay** setting, that sample period is marked as *bad*.
- ▶ Software keeps a running count of *bad periods*. Each time that a bad period occurs, this count goes up by one. Each time a good period occurs, this count goes down by 1, to a minimum value of 0. As an example, 10 bad periods that are followed by five good periods and then followed by 10 bad periods result in a bad period count of 15.
- ▶ The **gmlinktolerance** parameter is defined in seconds. Because bad periods are assessed at intervals of 10 seconds, the maximum bad period count is the **gmlinktolerance** parameter value that is divided by 10. For example, with a **gmlinktolerance** value of 300, the maximum bad period count is 30. When maximum bad period count is reached, a 1920 error is reported.

In this case, the 1920 error is identified with the specific event ID 985003 that is associated to the GM relationship, which in the last 10-second period had the greatest accumulated time spent on delays. This event ID is generated with the text `Remote Copy retry timeout`.

An edge case is achieved by setting the **gmmaxhostdelay** and **gmlinktolerance** parameters to their minimum settings (1 ms and 20 s). With these settings, you need only two consecutive bad sample periods before a 1920 error condition is reported. Consider a foreground write I/O that has a light I/O load, for example, a single I/O happens in 20 s. With unlucky timing, a single bad I/O result (that is, a write I/O that took over 1 ms in remote copy) spans the boundary of two, 10-second sample periods. This single bad I/O theoretically can be counted as twice the number of bad periods and triggers a 1920 error.

A higher **gmlinktolerance** value, **gmmaxhostdelay** setting, or I/O load, might reduce the risk of encountering this edge case.

Hung I/O condition: maxreplicationdelay and partnershipexclusionthreshold

The system-wide **maxreplicationdelay** attribute configures how long a single write can be outstanding from the host before the relationship is stopped, which triggers a 1920 error. It can protect the hosts from seeing timeouts because of secondary hung I/Os.

This parameter is mainly intended to protect from secondary system issues. It does not help with ongoing performance issues, but can be used to limit the exposure of hosts to long write response times that can cause application errors. For example, setting **maxreplicationdelay** to 30 means that if a write operation for a volume in a remote copy relationship does not complete within 30 seconds, the relationship is stopped, which triggers a 1920 error. This error happens even if the cause of the write delay is not related to the remote copy. For this reason, the **maxreplicationdelay** settings can lead to false positive 1920 error triggering.

In addition to the 1920 error, the specific event ID 985004 is generated with the text `Maximum replication delay exceeded`.

The **maxreplicationdelay** values can be 0 - 360 seconds. Setting **maxreplicationdelay** to 0 disables the feature.

partnershipexclusionthreshold is a system-wide parameter that sets the timeout for an I/O that triggers a temporary dropping of the link to the remote system. Like **maxreplicationdelay**, the **partnershipexclusionthreshold** attribute provides some flexibility in a part of replication that tries to shield a production system from hung I/Os on a secondary system.

To better understand the **partnershipexclusionthreshold** parameter, consider the following scenario. By default in IBM Spectrum Virtualize, a node assert (a restart with a 2030 error) occurs if any individual I/O takes longer than 6 minutes. To avoid this situation, some actions are attempted to clean up anything that might be hanging I/O before the I/O reaches 6 minutes.

One of these actions is temporarily dropping (for 15 minutes) the link between systems if any I/O takes longer than 5 minutes and 15 seconds (315 seconds). This action often removes hang conditions that are caused by replication problems. The **partnershipexclusionthreshold** parameter introduced the ability to set this value to a time lower than 315 seconds to respond to hung I/O more swiftly. The **partnershipexclusionthreshold** value must be a number 30 - 315.

If an I/O takes longer than the **partnershipexclusionthreshold** value, a 1720 error is triggered (with an event ID 987301), and any regular GM or MM relationships stop on the next write to the primary volume.

Important: Do not change the **partnershipexclusionthreshold** parameter except under the direction of IBM Support.

To set the **maxreplicationdelay** and **partnershipexclusionthreshold** parameters, run the **chssystem** command, as shown in Example 6-3.

Example 6-3 The maxreplicationdelay and partnershipexclusionthreshold settings

```
IBM_IBM FlashSystem:ITS0:superuser>chssystem -maxreplicationdelay 30
IBM_IBM FlashSystem:ITS0:superuser>chssystem -partnershipexclusionthreshold 180
```

The **maxreplicationdelay** and **partnershipexclusionthreshold** parameters do not interact with the **gmlinktolerance** and **gmmxhostdelay** parameters.

Troubleshooting 1920 errors

When you are troubleshooting 1920 errors that are posted across multiple relationships, you must diagnose the cause of the earliest error first. You must consider whether other higher priority system errors exist and fix these errors because they might be the underlying cause of the 1920 error.

The diagnosis of a 1920 error is assisted by SAN performance statistics. To gather this information, you can use IBM Spectrum Control with a statistics monitoring interval of 1 or 5 minutes. Also, turn on the internal statistics gathering function **I0stats** in IBM Spectrum Virtualize. Although not as powerful as IBM Spectrum Control, **I0stats** can provide valuable debug information if the **snap** command gathers system configuration data close to the time of failure.

The following main performance statistics must be investigated for the 1920 error:

► Write I/O Rate and Write Data Rate

For volumes that are primary volumes in relationships, these statistics are the total amount of write operations that are submitted per second by hosts on average over the sample period, and the bandwidth of those writes. For secondary volumes in relationships, these statistics are the average number of replicated writes that are received per second, and the bandwidth that these writes consume. Summing the rate over the volumes that you intend to replicate gives a coarse estimate of the replication link bandwidth that is required.

► Write Response Time and Peak Write Response Time

On primary volumes, these items are the average time (in milliseconds) and the peak time between a write request being received from a host and the completion message being returned. The Write Response Time is the best way to show what kind of write performance that the host is seeing.

If a user complains that an application is slow, and the stats show that the Write Response Time leaps from 1 ms to 20 ms, the two are most likely linked. However, some applications with high queue depths and low to moderate workloads are not affected by increased response times. This high queue depth is an effect of some other problem. The Peak Write Response Time is less useful because it is sensitive to individual glitches in performance, but it can show more details about the distribution of write response times.

On secondary volumes, these statistics describe the time for the write to be submitted from the replication feature into the system cache, and should normally be of a similar magnitude to the ones on the primary volume. Generally, the Write Response Time should be below 1 ms for a fast-performing system.

► Global Mirror Write I/O Rate

This statistic shows the number of writes per second that the (regular) replication feature is processing for this volume. It applies to both types of GM and to MM, but only for the secondary volume. Because writes are always separated into 32 KB or smaller tracks before replication, this setting might be different from the Write I/O Rate on the primary volume (magnified further because the samples on the two systems will not be aligned, so they capture a different set of writes).

► Global Mirror Overlapping Write I/O Rate

This statistic monitors the amount of overlapping I/O that the GM feature is handling for regular GM relationships, which is where an LBA is written again after the primary volume is updated, but before the secondary volume is updated for an earlier write to that LBA. To mitigate the effects of the overlapping I/Os, a journaling feature was implemented, as described in “Colliding writes” on page 410.

► Global Mirror secondary write lag

This statistic is valid for regular GM primary and secondary volumes. For primary volumes, it tracks the length of time in milliseconds that replication writes are outstanding from the primary system. This amount includes the time to send the data to the remote system, consistently apply it to the secondary nonvolatile cache, and send an acknowledgment back to the primary system.

For secondary volumes, this statistic records only the time that is taken to consistently apply it to the system cache, which is normally up to 20 ms. Most of that time is spent coordinating consistency across many nodes and volumes. Primary and secondary volumes for a relationship tend to record times that differ by the RTT between systems. If this statistic is high on the secondary system, look for congestion on the secondary system’s fabrics, saturated auxiliary storage, or high CPU utilization on the secondary system.

- ▶ Write-cache Delay I/O Rate

These statistics show how many writes might not be instantly accepted into the system cache because the cache was full. These statistics are a good indication that the write rate is faster than the storage can cope with. If this amount starts to increase on auxiliary storage while primary volumes suffer from increased Write Response Time, it is possible that the auxiliary storage is not fast enough for the replicated workload.

- ▶ Port to Local Node Send Response Time

The time in milliseconds that it takes this node to send a message to other nodes in the same system (which will mainly be the other node in the same I/O group) and get an acknowledgment back. This amount should be well below 1 ms, with values below 0.3 ms being essential for regular GM to provide a Write Response Time below 1 ms.

This requirement is necessary because up to three round-trip messages within the local system happen before a write completes to the host. If this number is higher than you want, look at fabric congestion (Zero Buffer Credit Percentage) and CPU Utilization of all the nodes in the system.

- ▶ Port to Remote Node Send Response Time

This value is the time in milliseconds that it takes to send a message to nodes in other systems and get an acknowledgment back. This amount is not separated out by remote system, but for environments that have replication to only one remote system. This amount should be close to the low-level ping time between your sites. If this amount goes significantly higher, it is likely that the link between your systems is saturated, which usually causes high Zero Buffer Credit Percentage as well.

- ▶ Sum of Port-to-local node send response time and Port-to-local node send queue time

The time must be less than 1 ms for the primary system. A number in excess of 1 ms might indicate that an I/O group is reaching its I/O throughput limit, which can limit performance.

- ▶ System CPU Utilization

These values show how heavily loaded the nodes in the system are. If any core has high utilization (for example, over 90%) and there is an increase in write response time, it is possible that the workload is being CPU limited. You can resolve this situation by upgrading to faster hardware, or spreading out some of the workload to other nodes and systems.

- ▶ Zero Buffer Credit Percentage or Port Send Delay I/O Percentage

These statistics are the fraction of messages that this node attempted to send through FC ports that had to be delayed because the port ran out of buffer credits. If you have a long link from the node to the switch to which it is attached, it might be beneficial to get the switch to grant more buffer credits on its port.

It is more likely to be the result of congestion on the fabric because running out of buffer credits is how FC performs flow control. Normally, this value is well under 1%. 1 - 10% is a concerning level of congestion, but you might find the performance acceptable. Over 10% indicates severe congestion. This amount is called out on a port-by-port basis in the port-level statistics, which gives finer granularity about where any congestion might be.

When looking at the port-level statistics, high values on ports that are used for messages to nodes in the same system are much more concerning than the ones on ports that are used for messages to nodes in other systems.

► **Back-end Write Response Time**

This value is the average response time in milliseconds for write operations to the back-end storage. This time might include several physical I/O operations, depending on the type of RAID architecture.

Poor back-end performance on a secondary system is a frequent cause of 1920 errors, while it is not so common for primary systems. Exact values to watch out for depend on the storage technology, but usually the response time should be less than 50 ms. A longer response time can indicate that the storage controller is overloaded. If the response time for a specific storage controller is outside of its specified operating range, investigate for the same reason.

Focus areas for 1920 errors

The causes of 1920 errors might be numerous. To fully understand the underlying reasons for posting this error, consider the following components that are related to the remote copy relationship:

- The inter-system connectivity network
- Primary storage and remote storage
- IBM Storage Virtualize node
- SAN

Data collection for diagnostic purposes

A successful diagnosis depends on the collection of the following data at both systems:

- The **snap** command with **livedump** (triggered at the point of failure)
- I/O stats that are running at the operating-system level (if possible)
- IBM Spectrum Control performance statistics data (if possible)
- The following information and logs from other components:
 - Inter-system network and switch details:
 - Technology.
 - Bandwidth.
 - Typical measured latency on the inter-system network.
 - Distance on all links (which can take multiple paths for redundancy).
 - Whether trunking is enabled.
 - How the link interfaces with the two SANs.
 - Whether compression is enabled on the link.
 - Whether the link is dedicated or shared; if so, the resource and amount of those resources that they use.
 - Switch Write Acceleration to check with IBM for compatibility or known limitations.
 - Switch Compression, which should be transparent, but complicates the ability to predict bandwidth.
 - Storage and application:
 - Specific workloads at the time of 1920 errors, which might not be relevant, depending upon the occurrence of the 1920 errors and the volumes that are involved.
 - RAID rebuilds.
 - Whether 1920 errors are associated with Workload Peaks or Scheduled Backup.

Inter-system network

For diagnostic purposes, ask the following questions about the inter-system network:

- ▶ Was network maintenance being performed?

Consider the hardware or software maintenance that is associated with the inter-system network, such as updating firmware or adding more capacity.

- ▶ Is the inter-system network overloaded?

You can find indications of this situation by using statistical analysis with the help of I/O stats, IBM Spectrum Control, or both. Examine the internode communications, storage controller performance, or both. By using IBM Spectrum Control, you can check that the storage metrics for the GM relationships were stopped, which can be tens of minutes depending on the `gmLinkTolerance` and `maxReplicationDelay` parameters.

Diagnose the overloaded link by using the following methods:

- Look at the statistics that are generated by the routers or switches near your most bandwidth-constrained link between the systems.

Exactly what is provided and how to analyze it varies depending on the equipment that is used.

- Look at the port statistics for high response time in internode communication.

An overloaded long-distance link causes high response times in the internode messages (the *Port to Remote Node Send Response Time* statistic) that are sent by IBM Spectrum Virtualize. If delays persist, the messaging protocols exhaust their tolerance elasticity, and the GM protocol is forced to delay handling new foreground writes while waiting for resources to free up.

- Look at the port statistics for buffer credit starvation.

The *Zero Buffer Credit Percentage* and *Port Send Delay I/O Percentage* statistic can be useful here because you normally have a high value here as the link saturates. Only look at ports that are replicating to the remote system.

- Look at the volume statistics (before the 1920 error is posted):

- Target volume write throughput approaches the link bandwidth.

If the write throughput on the target volume is equal to your link bandwidth, your link is likely overloaded. Check what is driving this situation. For example, does the peak foreground write activity exceed the bandwidth, or does a combination of this peak I/O and the background copy exceed the link capacity?

- Source volume write throughput approaches the link bandwidth.

This write throughput represents only the I/O that is performed by the application hosts. If this number approaches the link bandwidth, you might need to upgrade the link's bandwidth. Alternatively, reduce the foreground write I/O that the application is attempting to perform, or reduce the number of remote copy relationships.

- Target volume write throughput is greater than the source volume write throughput.

If this condition exists, the situation suggests a high level of background copy and mirrored foreground write I/O. In these circumstances, decrease the background copy rate parameter of the GM partnership to bring back the combined mirrored foreground I/O and background copy I/O rates within the remote links bandwidth.

- Look at the volume statistics (after the 1920 error is posted).

Source volume write throughput after the GM relationships were stopped.

If write throughput increases greatly (by 30% or more) after the GM relationships are stopped, the application host was attempting to perform more I/O than the remote link can sustain.

When the GM relationships are active, the overloaded remote link causes higher response times to the application host. This overload decreases the throughput of application host I/O at the source volume. After the GM relationships stop, the application host I/O sees a lower response time, and the true write throughput returns.

To resolve this issue, increase the remote link bandwidth, reduce the application host I/O, or reduce the number of GM relationships.

Storage controllers

Investigate the primary and remote storage controllers, starting at the remote site. If the back-end storage at the secondary system is overloaded, or another problem is affecting the cache there, the GM protocol fails to keep up. Similarly, the problem exhausts the (**gmLinktolerance**) elasticity and has a similar effect at the primary system.

In this situation, ask the following questions:

- ▶ Are the storage controllers at the remote system overloaded (performing slowly)?

Use IBM Spectrum Control to obtain the back-end write response time for each MDisk at the remote system. A response time for any individual MDisk that exhibits a sudden increase of 50 ms or more or that is higher than 100 ms generally indicates a problem with the back end. For a 1920 error that is triggered by the “max replication delay exceeded” condition, check the peak back-end write response time to see whether it exceeded the **maxreplicatiodelay** value around the 1920 occurrence.

Check whether an error condition is on the internal storage controller, for example, media errors, a failed physical disk, or a recovery activity, such as RAID array rebuilding that uses more bandwidth.

If an error occurs, fix the problem, and then restart the GM relationships.

If no error occurs, consider whether the secondary controller can process the required level of application host I/O. You might improve the performance of the controller in the following ways:

- Adding more or faster physical disks to a RAID array.
- Changing the cache settings of the controller and checking that the cache batteries are healthy, if applicable.

- ▶ Are the storage controllers at the primary site overloaded?

Analyze the performance of the primary back-end storage by using the same steps that you use for the remote back-end storage. The main effect of bad performance is to limit the amount of I/O that can be performed by application hosts. Therefore, you must monitor the back-end storage at the primary site regardless of GM. For a 1920 error that is triggered by the “max replication delay exceeded” condition, check the peak back-end write response time to see whether it exceeded the **maxreplicatiodelay** value around the 1920 occurrence.

However, if bad performance continues for a prolonged period, a false 1920 error might be flagged.

Node

For IBM Storage Virtualize node hardware, the possible cause of the 1920 errors might be from a heavily loaded secondary or primary system. If this condition persists, a 1920 error might be posted.

GM must synchronize its I/O processing across all nodes in the system to ensure data consistency. If any node is running out of memory, it can affect all relationships. So, check the CPU cores usage statistic. If CPU usage looks higher when there is a performance problem, then running out of CPU bandwidth might be causing the problem. Of course, CPU usage goes up when the IOPS going through a node goes up, so if the workload increases, you would expect to see CPU usage increase.

If there is an increase in CPU usage on the secondary system but no increase in IOPS and volume write latency increases too, it is likely that the increase in CPU usage caused the increased volume write latency. In that case, try to work out what might have caused the increase in CPU usage (for example, starting many FlashCopy mappings). Consider moving that activity to a time with less workload. If there is an increase in both CPU usage and IOPS, and the CPU usage is close to 100%, then that node might be overloaded. A *Port-to-local node send queue time* value higher than 0.2 ms often denotes overloaded CPU cores.

In a primary system, if it is sufficiently busy, the write ordering detection in GM can delay writes enough to reach a latency of `gmmxhostdelay` and cause a 1920 error. Stopping replication potentially lowers CPU usage, and also lowers the opportunities for each I/O to be delayed by slow scheduling on a busy system.

Solve overloaded nodes by upgrading them to newer, faster hardware if possible, or by adding more I/O groups or control enclosures (or systems) to spread the workload over more resources.

Storage area network

Issues and congestions both in local and remote SANs can lead to 1920 errors. The *Port to local node send response time* is the key statistic to investigate. It captures the RTT between nodes in the same system. Anything over 1.0 ms is high and causes high secondary volume write response time. Values greater than 1 ms on primary system cause an impact on write latency to GM primary volumes of 3 ms or more.

If you checked CPU core utilization on all the nodes and it has not gotten near 100%, a high *Port to local node send response time* means that there is fabric congestion or a slow-draining FC device.

A good indicator of SAN congestion is the Zero Buffer Credit Percentage and Port Send Delay I/O Percentage on the port statistics. For more information about buffer credit, see “Buffer credits” on page 424.

If a port has more than 10% zero buffer credits, that situation definitely causes a problem for all I/O, not just GM writes. Values 1 - 10% are moderately high and might contribute to performance issues.

For both primary and secondary systems, congestion on the fabric from other slow-draining devices becomes much less of an issue when only dedicated ports are used for node-to-node traffic within the system. However, these ports become an option only on systems with more than four ports per node. Use port masking to segment your ports.

FlashCopy considerations

Check that FlashCopy mappings are in the *prepared* state. Check whether the GM target volumes are the sources of a FlashCopy mapping and whether that mapping was in the *prepared* state for an extended time.

Volumes in the prepared state are cache-disabled, so their performance is impacted. To resolve this problem, start the FlashCopy mapping, which re-enables the cache and improves the performance of the volume and of the GM relationship.

FlashCopy can add significant workload to the back-end storage, especially when the background copy is active (see “Background copy” on page 367). In cases where the remote system is used to create golden or practice copies for DR testing, the workload that is added by the FlashCopy background processes can overload the system. This overload can lead to poor remote copy performances and then to a 1920 error, even though with IBM FlashSystem this issue is not much of one because of a high-performing flash back end.

Careful planning of the back-end resources is important with these kinds of scenarios. Reducing the FlashCopy background copy rate can also help to mitigate this situation. Furthermore, the FlashCopy CoW process adds some latency by delaying the write operations on the primary volumes until the data is written to the FlashCopy target.

This process does not directly affect the remote copy operations because it is logically placed below the remote copy processing in the I/O stack, as shown in Figure 6-7 on page 378. Nevertheless, in some circumstances, especially with write-intensive environments, the CoW process tends to stress some of the internal resources of the system, such as CPU and memory. This condition also can affect the remote copy that competes for the same resources, eventually leading to 1920 errors.

FCIP considerations

When you get a 1920 error, always check the latency first. The FCIP routing layer can introduce latency if it is not properly configured. If your network provider reports a much lower latency, you might have a problem at your FCIP routing layer. Most FCIP routing devices have built-in tools to enable you to check the RTT. When you are checking latency, remember that TCP/IP routing devices (including FCIP routers) report RTT by using standard 64-byte ping packets.

In Figure 6-34 on page 455, you can see why the effective transit time must be measured only by using packets that are large enough to hold an FC frame, or 2148 bytes (2112 bytes of payload and 36 bytes of header). Set estimated resource requirements to be a safe amount because various switch vendors have optional features that might increase this size. After you verify your latency by using the proper packet size, proceed with normal hardware troubleshooting.

Look at the second largest component of your RTT, which is *serialization delay*. Serialization delay is the amount of time that is required to move a packet of data of a specific size across a network link of a certain bandwidth. The required time to move a specific amount of data decreases as the data transmission rate increases.

Figure 6-34 on page 455 shows the orders of magnitude of difference between the link bandwidths. It is easy to see how 1920 errors can arise when your bandwidth is insufficient. Never use a TCP/IP ping to measure RTT for FCIP traffic.

Packet Size	Link Size	Serialization Delay (Time Required to Send Data)	Unit
64	256 Kbps	2.0E+03	microseconds
64	1.5 Mbps	3.4E+02	microseconds
64	100 Mbps	5.1E+00	microseconds
64	155 Mbps	3.3E+00	microseconds
64	622 Mbps	8.2E-01	microseconds
64	1 Gbps	5.1E-04	microseconds
64	10 Gbps	5.1E-05	microseconds
1500	256 Kbps	4.7E+04	microseconds
1500	1.5 Mbps	8.0E+03	microseconds
1500	100 Mbps	1.2E+02	microseconds
1500	155 Mbps	7.7E+01	microseconds
1500	622 Mbps	1.9E+01	microseconds
1500	1 Gbps	1.2E+01	microseconds
1500	10 Gbps	1.2E+00	microseconds
2148	256 Kbps	6.7E+04	microseconds
2148	1.5 Mbps	1.1E+04	microseconds
2148	100 Mbps	1.7E+02	microseconds
2148	155 Mbps	1.1E+02	microseconds
2148	622 Mbps	2.8E+01	microseconds
2148	1 Gbps	1.7E+01	microseconds
2148	10 Gbps	1.7E-03	microseconds

Figure 6-34 Effect of packet size (in bytes) versus the link size

In Figure 6-34, the amount of time in microseconds that is required to transmit a packet across network links of varying bandwidth capacity is compared. The following packet sizes are used:

- ▶ 64 bytes: The size of the common ping packet
- ▶ 1500 bytes: The size of the standard TCP/IP packet
- ▶ 2148 bytes: The size of an FC frame

Finally, your path maximum transmission unit (MTU) affects the delay that is incurred to get a packet from one location to another location. An MTU might cause fragmentation or be too large and cause too many retransmits when a packet is lost.

Hung I/O

A hung I/O condition is reached when a write operation is delayed in the IBM Storage Virtualize stack for a significant time (typically seconds). This condition is monitored by IBM Spectrum Virtualize, which eventually leads to a 1920 error if the delay is higher than **maxreplicationdelay** settings.

Hung I/Os can be caused by many factors, such as back-end performance, cache fullness, internal resource starvation, and remote copy issues. When the **maxreplicationdelay** setting triggers a 1920 error, the following areas must be investigated:

- ▶ Inter-site network disconnections: This kind of event generates partnership instability, which leads to delayed mirrored write operations until the condition is resolved.
- ▶ Secondary system poor performance: In the case of bad performance, the secondary system can become unresponsive, which delays the replica of the write operations.
- ▶ Primary or secondary system node warmstarts: During a node warmstart, the system freezes all the I/Os for few seconds to get a consistent state of the cluster resources. These events often are not directly related to the remote copy operations.

Note: The `maxreplicationdelay` trigger can occur even if the cause of the write delay is not related to the remote copy. In this case, the replication suspension does not resolve the hung I/O condition.

To exclude the remote copy as the cause of the hung I/O, the duration of the delay (peak write response time) can be checked by using tools, such as IBM Spectrum Control. If the measured delay is greater than the `maxreplicationdelay` settings, it is unlikely that the remote copy is responsible.

Recovering after 1920 errors

After a 1920 error occurs, the GM auxiliary volumes are no longer in a Consistent Synchronized state. You must establish the cause of the problem and fix it before you restart the relationship.

When the relationship is restarted, you must resynchronize it. During this period, the data on the MM or GM auxiliary volumes on the secondary system is inconsistent, and your applications cannot use the volumes as backup disks. To address this data consistency exposure on the secondary system, a FlashCopy of the auxiliary volumes can be created to maintain a consistent image until the GM (or the MM) relationships are synchronized again and back in a consistent state.

IBM Storage Virtualize provides the *Remote Copy Consistency Protection* feature that automates this process. When Consistency Protection is configured, the relationship between the primary and secondary volumes does not go in to the Inconsistent copying status after it is restarted. Instead, the system uses a secondary CV to automatically copy the previous consistent state of the secondary volume.

The relationship automatically moves to the Consistent copying status as the system resynchronizes and protects the consistency of the data. The relationship status changes to Consistent synchronized when the resynchronization process completes.

For more information about the Consistency Protection feature, see *Implementation Guide for IBM Spectrum Virtualize Version 8.5*, SG24-8520.

To ensure that the system can handle the background copy load, delay restarting the MM or GM relationship until a quiet period occurs. If the required link capacity is unavailable, you might experience another 1920 error, and the MM or GM relationship might stop in an inconsistent state.

Copy services tools, like IBM CSM, or manual scripts can be used to automatize the relationships to restart after a 1920 error. CSM implements a logic to avoid recurring restart operations in the case of a persistent problem. CSM attempts an automatic restart for every occurrence of a 1720 or 1920 error of a certain number of times (determined by the `gmLinktolerance` value) within a 30-minute period.

If the number of allowable automatic restarts is exceeded within the period, CSM does not automatically restart GM on the next 1720 or 1920 error. Furthermore, with CSM it is possible to specify the amount of time, in seconds, in which the tool waits after a 1720 or 1920 error before automatically restarting the GM. For more information, see [IBM Copy Services Manager welcome page](#).

Tip: When implementing automatic restart functions, it is a best practice to preserve data consistency on GM target volumes during the resynchronization by using features such as FlashCopy or Consistency Protection.

Adjusting the Global Mirror settings

Although the default values are valid in most configurations, the settings of **gmlinktolerance** and **gmmaxhostdelay** can be adjusted to accommodate particular environment or workload conditions.

For example, GM can look at average delays. However, some hosts, such as VMware ESX, might not tolerate a single I/O getting old, for example, 45 seconds, before it decides to restart. Because it is better to terminate a GM relationship than it is to restart a host, you might want to set **gmlinktolerance** to something like 30 seconds and then compensate so that you do not get too many relationship terminations by setting **gmmaxhostdelay** to something larger, such as 100 ms.

If you compare the two approaches, the default (**gmlinktolerance 300, gmmaxhostdelay 5**) is a rule that means “If more than one third of the I/Os are slow and that happens repeatedly for 5 minutes, then terminate the busiest relationship in that stream.” In contrast, the example of **gmlinktolerance 30, gmmaxhostdelay 100** is a rule that means “If more than one third of the I/Os are slow and that happens repeatedly for 30 seconds, then terminate the busiest relationship in the stream.”

So the first approach picks up general slowness, and the other approach picks up shorter bursts of extreme slowness that might disrupt your server environment. The general recommendation is to change the **gmlinktolerance** and **gmmaxhostdelay** values progressively and evaluate the overall impact to find an acceptable compromise between performances and GM stability.

You can even disable the **gmlinktolerance** feature by setting the **gmlinktolerance** value to 0. However, the **gmlinktolerance** parameter cannot protect applications from extended response times if it is disabled. You might consider disabling the **gmlinktolerance** feature in the following circumstances:

- ▶ During SAN maintenance windows, where degraded performance is expected from SAN components and application hosts can withstand extended response times from GM volumes.
- ▶ During periods when application hosts can tolerate extended response times and it is expected that the **gmlinktolerance** feature might stop the GM relationships. For example, you are testing the usage of an I/O generator that is configured to stress the back-end storage. Then, the **gmlinktolerance** feature might detect high latency and stop the GM relationships. Disabling the **gmlinktolerance** parameter stops the GM relationships at the risk of exposing the test host to extended response times.

Another tunable parameter that interacts with the GM is the **maxreplicationdelay**. The **maxreplicationdelay** settings do not mitigate the 1920 error occurrence because the parameter adds a trigger to the 1920 error. However, **maxreplicationdelay** provides users with a fine granularity mechanism to manage the hung I/Os condition and it can be used in combination with **gmlinktolerance** and **gmmaxhostdelay** settings to better address particular environment conditions.

In the VMware example, an alternative option is to set the **maxreplicationdelay** to 30 seconds and leave the **gmlinktolerance** and **gmmaxhostdelay** settings at their default. With these settings, the **maxreplicationdelay** timeout effectively handles the hung I/Os conditions, and the **gmlinktolerance** and **gmmaxhostdelay** settings still provide an adequate mechanism to protect from ongoing performance issues.

6.5 Native IP replication

The native IP replication feature enables replication between any IBM Storage Virtualize products by using the built-in networking ports or optional 1-, 10-, or 25-Gb adapter.

Native IP replication uses SANslide technology, which was developed by Bridgeworks Limited of Christchurch, UK. They specialize in products that can bridge storage protocols and accelerate data transfer over long distances. Adding this technology at each end of a wide area network (WAN) TCP/IP link improves the utilization of the link.

This technology improves the link utilization by applying patented artificial intelligence (AI) to hide latency that is normally associated with WANs. Doing so can greatly improve the performance of mirroring services, in particular GMCV over long distances.

6.5.1 Native IP replication technology

Remote mirroring over IP communication is supported on IBM Storage Virtualize by using Ethernet communication links. IBM Storage Virtualize Software IP replication uses the innovative *Bridgeworks SANSlide* technology to optimize network bandwidth and utilization. With this new function, you can use a lower-speed and lower-cost networking infrastructure for data replication.

Bridgeworks SANSlide technology, which is integrated into IBM Spectrum Virtualize, uses AI to help optimize network bandwidth usage and adapt to changing workload and network conditions. This technology can improve remote mirroring network bandwidth usage up to three times. It can enable clients to deploy a less costly network infrastructure, or speed up remote replication cycles to enhance DR effectiveness.

With an Ethernet network data flow, the data transfer can slow down over time. This condition occurs because of the latency that is caused by waiting for the acknowledgment of each set of packets that are sent. The next packet set cannot be sent until the previous packet is acknowledged, as shown in Figure 6-35.

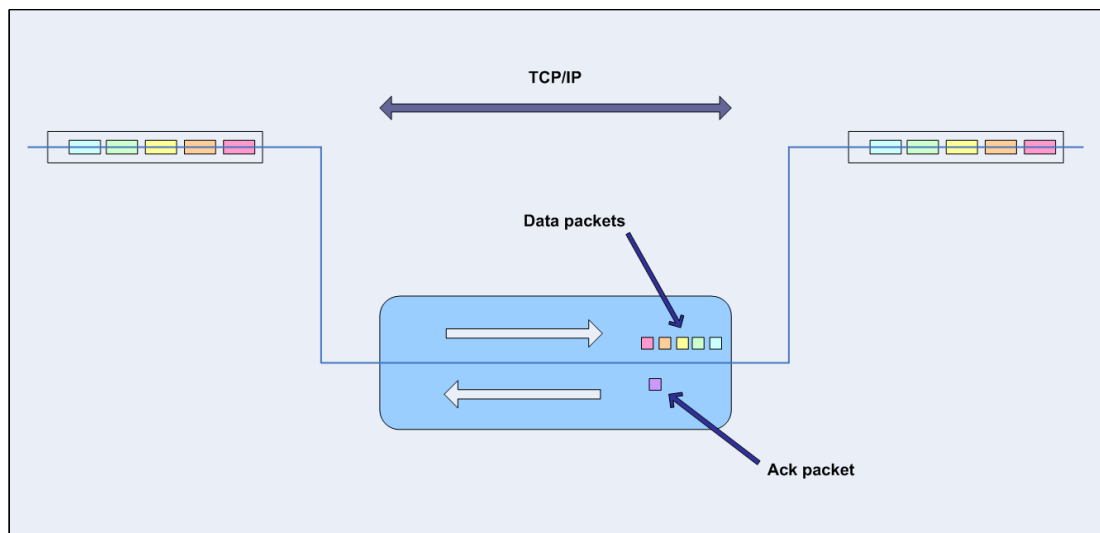


Figure 6-35 Typical Ethernet network data flow

However, by using the embedded IP replication, this behavior can be eliminated with the enhanced parallelism of the data flow. This parallelism uses multiple virtual connections (VCs) that share IP links and addresses.

The AI engine can dynamically adjust the number of VCs, receive window size, and packet size to maintain optimum performance. While the engine is waiting for one VC's ACK, it sends more packets across other VCs. If packets are lost from any VC, data is automatically retransmitted, as shown in Figure 6-36.

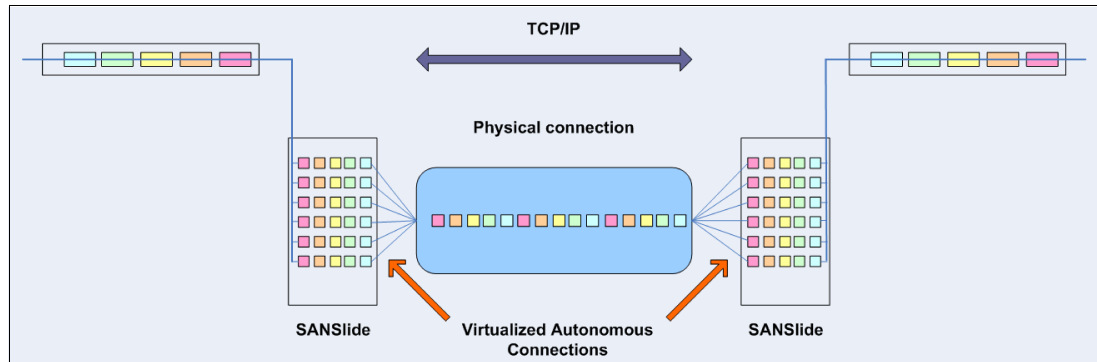


Figure 6-36 Optimized network data flow by using Bridgeworks SANSlide technology

For more information about this technology, see *IBM SAN Volume Controller and Storwize Family Native IP Replication*, REDP-5103.

MM, GM, and GMCV are supported by native IP partnership.

6.5.2 IP partnership limitations

The following prerequisites and assumptions must be considered before IP partnership between two IBM Storage Virtualize based systems can be established:

- ▶ The systems have version 7.2 or later code levels.
- ▶ The systems have the necessary licenses that enable remote copy partnerships to be configured between two systems. A separate license is not required to enable IP partnership.
- ▶ The storage SANs are configured correctly and the correct infrastructure to support the systems in remote copy partnerships over IP links are in place.
- ▶ The two systems must be able to ping each other and perform the discovery.
- ▶ The maximum number of partnerships between the local and remote systems, including IP *and* FC partnerships, is limited to the current maximum that is supported, which is three partnerships (four systems total).

Note: With code versions earlier than 8.4.2, only a single partnership over IP is supported.

- ▶ A system can have simultaneous partnerships over FC and IP, but with separate systems. The FC zones between two systems must be removed before an IP partnership is configured.
- ▶ The use of WAN-optimization devices, such as Riverbed, is not supported in IP partnership configurations containing SVC.

- ▶ IP partnerships are supported by 25-, 10-, and 1-Gbps links. However, the intermix on a single link is not supported.
- ▶ The maximum supported RTT is 80 ms for 1-Gbps links.
- ▶ The maximum supported RTT is 10 ms for 25- and 10-Gbps links.
- ▶ The minimum supported link bandwidth is 10 Mbps.
- ▶ The inter-cluster heartbeat traffic uses 1 Mbps per link.
- ▶ Migrations of remote copy relationships directly from FC-based partnerships to IP partnerships are not supported.
- ▶ IP partnerships between the two systems can be over either IPv4 or IPv6, but not both.
- ▶ Virtual local area network (VLAN) tagging of the IP addresses that are configured for remote copy is supported.
- ▶ Management IP addresses and internet Small Computer Systems Interface (iSCSI) IP addresses on the same port can be in a different network.
- ▶ An added layer of security is provided by using Challenge Handshake Authentication Protocol (CHAP) authentication.
- ▶ Direct-attached systems configurations are supported by the following restrictions:
 - Only two direct-attach links are allowed.
 - The direct-attach links must be on the same I/O group.
 - Use two portsets, where a portset contains only the two ports that are directly linked.
- ▶ TCP ports 3260 and 3265 are used for IP partnership communications. Therefore, these ports must be open in firewalls between the systems.
- ▶ Network address translation (NAT) between systems that are being configured in an IP partnership group is not supported.
- ▶ Only one remote copy data session per portset can be established. It is intended that only one connection (for sending or receiving remote copy data) is made for each independent physical link between the systems.

Note: A physical link is the physical IP link between the two sites: A (local) and B (remote). Multiple IP addresses on local system A can be connected (by Ethernet switches) to this physical link. Similarly, multiple IP addresses on remote system B can be connected (by Ethernet switches) to the same physical link. At any point, only a single IP address on cluster A can form a remote copy data session with an IP address on cluster B.

- ▶ The maximum throughput is restricted based on the usage of 1-Gbps or 10-Gbps Ethernet ports. The output varies based on distance (for example, round-trip latency) and quality of the communication link (for example, packet loss). The following maximum throughputs are achievable:
 - One 1-Gbps port can transfer up to 120 MB.
 - One 10-Gbps port can transfer up to 600 MB.

Table 6-11 lists the IP replication limits.

Table 6-11 IP replication limits

Remote copy property	Maximum	Apply to	Comment
Inter-system IP partnerships per system	Three systems	All models	A system can be connected with up to three remote systems.
Inter-site links per IP partnership	Two links	All models	A maximum of two inter-site links can be used between two IP partnership sites.
Ports per node	One port	All models	A maximum of one port per node can be used for an IP partnership.
IP partnership software compression limit	140 MBps	All models	Not the total limit of IP replication, but only of compression.

6.5.3 VLAN support

VLAN tagging is supported for iSCSI host attachment and IP replication. Hosts and remote copy operations can connect to the system through Ethernet ports. Each traffic type has different bandwidth requirements, which can interfere with each other if they share a port.

VLAN tagging creates two separate connections on the same IP network for different types of traffic. The system supports VLAN configuration on both IPv4 and IPv6 connections.

When the VLAN ID is configured for the IP addresses that are used for iSCSI host attachment or IP replication, the suitable VLAN settings on the Ethernet network and servers must be configured correctly to avoid connectivity issues. After the VLANs are configured, changes to the VLAN settings disrupt iSCSI and IP replication traffic to and from the partnerships.

During the VLAN configuration for each IP address, the VLAN settings for the local and failover ports on two nodes of an I/O group can differ. To avoid any service disruption, switches must be configured so that the failover VLANs are configured on the local switch ports, and the failover of IP addresses from a failing node to a surviving node succeeds.

If failover VLANs are not configured on the local switch ports, no paths are available to IBM Storage Virtualize during a node failure, and the replication fails.

Consider the following requirements and procedures when implementing VLAN tagging:

- ▶ VLAN tagging is supported for IP partnership traffic between two systems.
- ▶ VLAN provides network traffic separation at layer 2 for Ethernet transport.
- ▶ VLAN tagging by default is disabled for any IP address of a node port. You can use the CLI or GUI to set the VLAN ID for port IP addresses on both systems in the IP partnership.
- ▶ When a VLAN ID is configured for the port IP addresses that are used in remote copy port groups, the VLAN settings on the Ethernet network must also be properly configured to prevent connectivity issues.

Setting VLAN tags for a port is disruptive. Therefore, VLAN tagging requires that you stop the partnership first before you configure VLAN tags. Then, restart when the configuration is complete.

6.5.4 Domain name support for IP replication

Starting with IBM Storage Virtualize 8.5, you can specify domain names when creating IP partnerships. If you specify domain names, a DNS server must be configured on your system. To configure a DNS server for the system, select **Settings** → **Network** → **DNS**. You can also use the `mkdnserver` command to configure DNS servers.

Using the IBM Storage Virtualize management GUI, you can use DNS of the partner system by using the following procedure:

1. Select **Copy Services** → **Partnerships** and select **Create Partnership**.
2. Select **2-Site Partnership** and click **Continue**.
3. On the Create Partnership page, select **IP**.
4. To configure the partnership, enter either the partner system IP address or domain name, and then select the IP address or domain name of the partner system.

6.5.5 IP compression

IBM Storage Virtualize can use the IP compression capability to speed up replication cycles or to reduce bandwidth utilization.

This feature reduces the volume of data that must be transmitted during remote copy operations by using compression capabilities like the ones with existing IBM Real-time Compression (RtC) implementations.

No license: The IP compression feature does not require a RtC software license.

The data compression is made within the IP replication component of the IBM Storage Virtualize code. This feature can be used with MM, GM, and GMCV. The IP compression feature provides two kinds of compression mechanisms: hardware compression and software compression.

IP compression can be enabled on hardware configurations that support hardware-assisted compression acceleration engines. The hardware compression is active when compression accelerator engines are available. Otherwise, software compression is used.

Hardware compression uses currently underused compression resources. The internal resources are shared between data and IP compression. Software compression uses the system CPU and might have an impact on heavily used systems.

To evaluate the benefits of the IP compression, use the Comprestimator tool to estimate the compression ratio (CR) of the data to be replicated. The IP compression can be enabled and disabled without stopping the remote copy relationship by using the `mkippartnership` and `chpartnership` commands with the `-compress` parameter. Furthermore, in systems with replication that is enabled in both directions, the IP compression can be enabled in only one direction. IP compression is supported for IPv4 and IPv6 partnerships.

6.5.6 Replication portsets

This section describes the replication portsets and different ways to configure the links between the two remote systems. Two systems can be connected over one link or at most two links. For the systems to know about the physical links between the two sites, you use portsets.

Portsets are groupings of logical addresses that are associated with the specific traffic types. IBM Storage Virtualize supports IP (iSCSI or iSCSI Extensions for Remote Direct Memory Access (RDMA) (iSER)) or FC portsets for host attachment, IP portsets for back-end storage connectivity (iSCSI only), and IP replication traffic. Each physical Ethernet port can have a maximum of 64 IP addresses with each IP address on a unique portset.

A *portset object* is a system-wide object that might contain IP addresses from every I/O group. Figure 6-37 shows a sample of a portsets definition across the canister ports in a 2-I/O group IBM Storage Virtualize Storage system cluster.

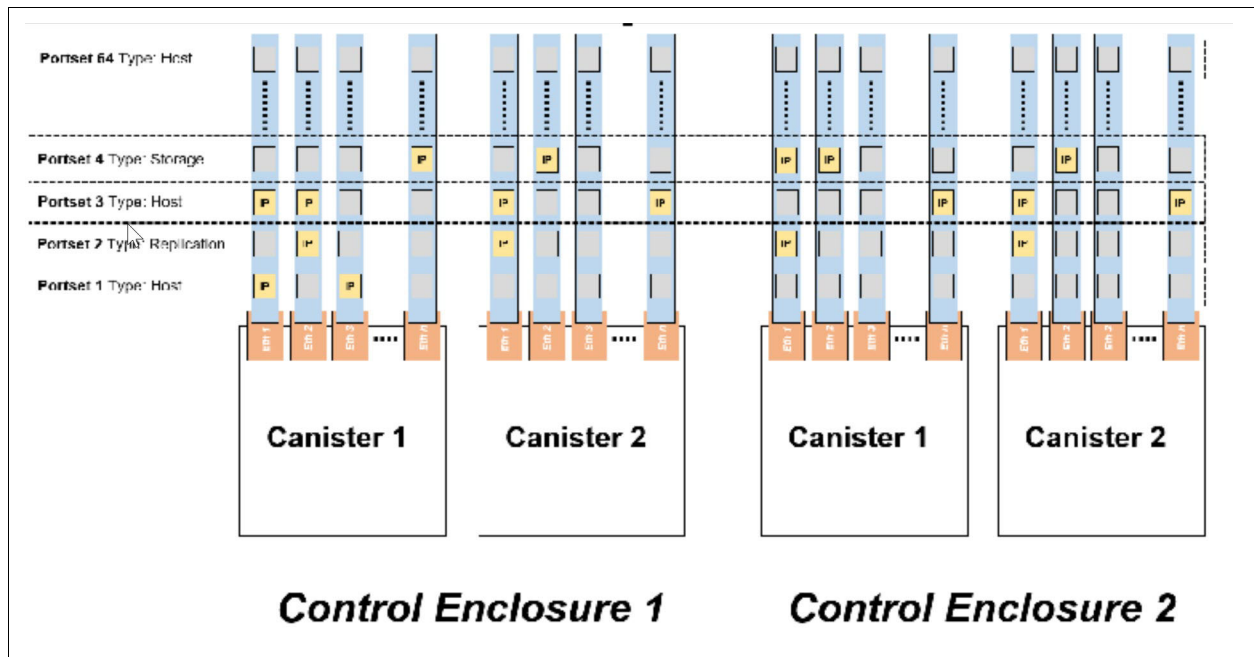


Figure 6-37 Portsets

To establish an IP partnership between two systems, complete the following steps:

1. Identify the Ethernet ports to be used for IP replication.
2. Define a replication type portset.
3. Set the IP addresses to the identified ports and add them to the portset.
4. Create the IP partnership from both systems specifying the portset to be used.

Multiple IBM Storage Virtualize canisters or nodes can be connected to the same physical long-distance link by setting IP addresses in the same portset. Samples of supported configurations are described in 6.5.7, “Supported configurations examples” on page 465.

In scenarios with two physical links between the local and remote clusters, two separate replication portsets must be used to designate which IP addresses are connected to which physical link. The relationship between the physical links and the replication portsets is not monitored by the IBM Storage Virtualize code. Therefore, two different replication portsets can be used with a single physical link and vice versa.

All IP addresses in a replication portset must be IPv4 *or* IPv6 addresses (IP types cannot be mixed). IP addresses can be shared among replication and host type portsets, although it is not recommended.

Note: The concept of a portset was introduced in IBM Storage Virtualize 8.4.2 and the IP Multi-tenancy feature. Versions before 8.4.2 use the remote copy port groups concept to tag the IP addresses to associate with an IP partnership. For more information, see [Remote-copy port groups](#).

When upgrading to 8.4.2, an automatic process occurs to convert the remote copy port groups configuration to an equivalent replication portset configuration.

Failover operations within and between portsets

Within one portset, only one IP address from each system is selected for sending and receiving remote copy data at any one time. Therefore, on each system, at most one IP address for each portset group is reported as used.

If the IP partnership cannot continue over an IP address, the system fails over to another IP address within that portset. Some reasons this issue might occur include the switch to which it is connected fails, the node goes offline, or the cable that is connected to the port is unplugged.

For the IP partnership to continue during a failover, multiple ports must be configured within the portset. If only one link is configured between the two systems, configure at least two IP addresses (one per node) within the portset. You can configure these two IP addresses on two nodes within the same I/O group or within separate I/O groups.

While failover is in progress, no connections in that portset exist between the two systems in the IP partnership for a short time. Typically, failover completes within 30 seconds to 1 minute. If the systems are configured with two portsets, the failover process within each portset continues independently of each other.

The disadvantage of configuring only one link between two systems is that during a failover, a discovery is initiated. When the discovery succeeds, the IP partnership is reestablished. As a result, the relationships might stop, in which case a manual restart is required. To configure two inter-system links, you must configure two replication type portsets.

When a node fails in this scenario, the IP partnership can continue over the other link until the node failure is rectified. Then, failback occurs when both links are again active and available to the IP partnership. The discovery is triggered so that the active IP partnership data path is made available from the new IP address.

In a two-node system or when more than one I/O group exists and the node in the other I/O group has IP addresses within the replication portset, the discovery is triggered. The discovery makes the active IP partnership data path available from the new IP address.

6.5.7 Supported configurations examples

Different IP replication topologies are available depending on the number of physical links, the number of nodes, and the number of IP partnerships. In the following sections, some typical configurations are described.

Single partnership configurations

In this section, some single partnership configurations are described.

Single inter-site link configurations

Consider single control enclosure systems in IP partnership over a single inter-site link (with failover ports configured), as shown in Figure 6-38.

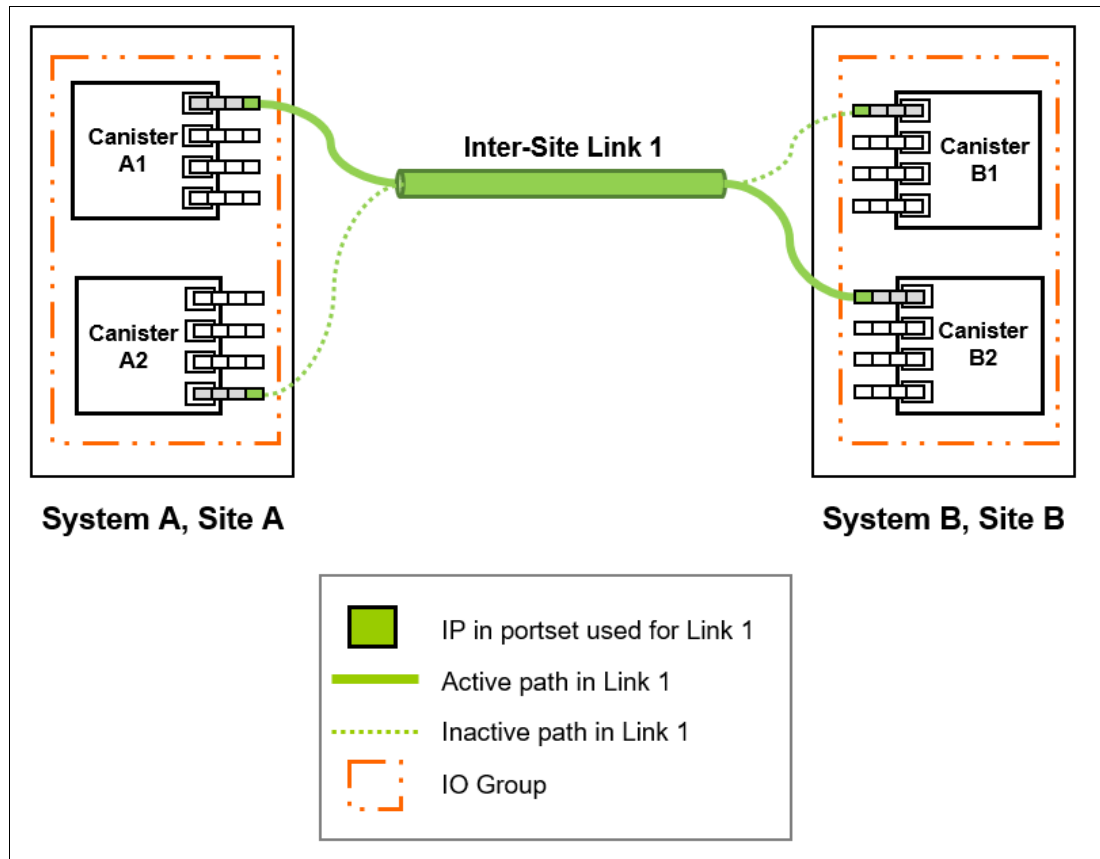


Figure 6-38 Only one link on each system and canister with failover ports configured

Figure 6-38 shows two systems: System A and System B. A single portset is used with IP addresses on two Ethernet ports, one each on Canister A1 and Canister A2 on System A. Similarly, a single portset is configured on two Ethernet ports on Canister B1 and Canister B2 on System B.

Although two ports on each system are configured in the portset, only one Ethernet port in each system actively participates in the IP partnership process. This selection is determined by a path configuration algorithm that is designed to choose data paths between the two systems to optimize performance.

The other port on the partner canister or node in the control enclosure behaves as a standby port that is used during a canister or node failure. If Canister or Node A1 fails in System A, IP partnership continues servicing replication I/O from Ethernet Port 2 because a failover port is configured on Canister or Node A2 on Ethernet Port 2.

However, it might take some time for discovery and path configuration logic to reestablish paths post-failover. This delay can cause partnerships to change to `Not_Present` for that time. The details of the particular IP port that is actively participating in IP partnership is provided in the `1spartnership` output (reported as `link1_ip_id` and `link2_ip_id`).

This configuration has the following characteristics:

- ▶ Each canister in the control enclosure or node in the I/O group has ports with IP addresses that are defined in the same replication type portset. However, only one path is active at any time at each system.
- ▶ If Canister or Node A1 in System A or Canister or Node B2 in System B fails in the respective systems, IP partnerships rediscovery is triggered and continues servicing the I/O from the failover port.
- ▶ The discovery mechanism that is triggered because of failover might introduce a delay where the partnerships momentarily change to the `Not_Present` state and recover.

A 4-control enclosure system or 8-node system in IP partnership with a 2-control enclosure system or 5-node system over a single inter-site link is shown in Figure 6-39 on page 467.

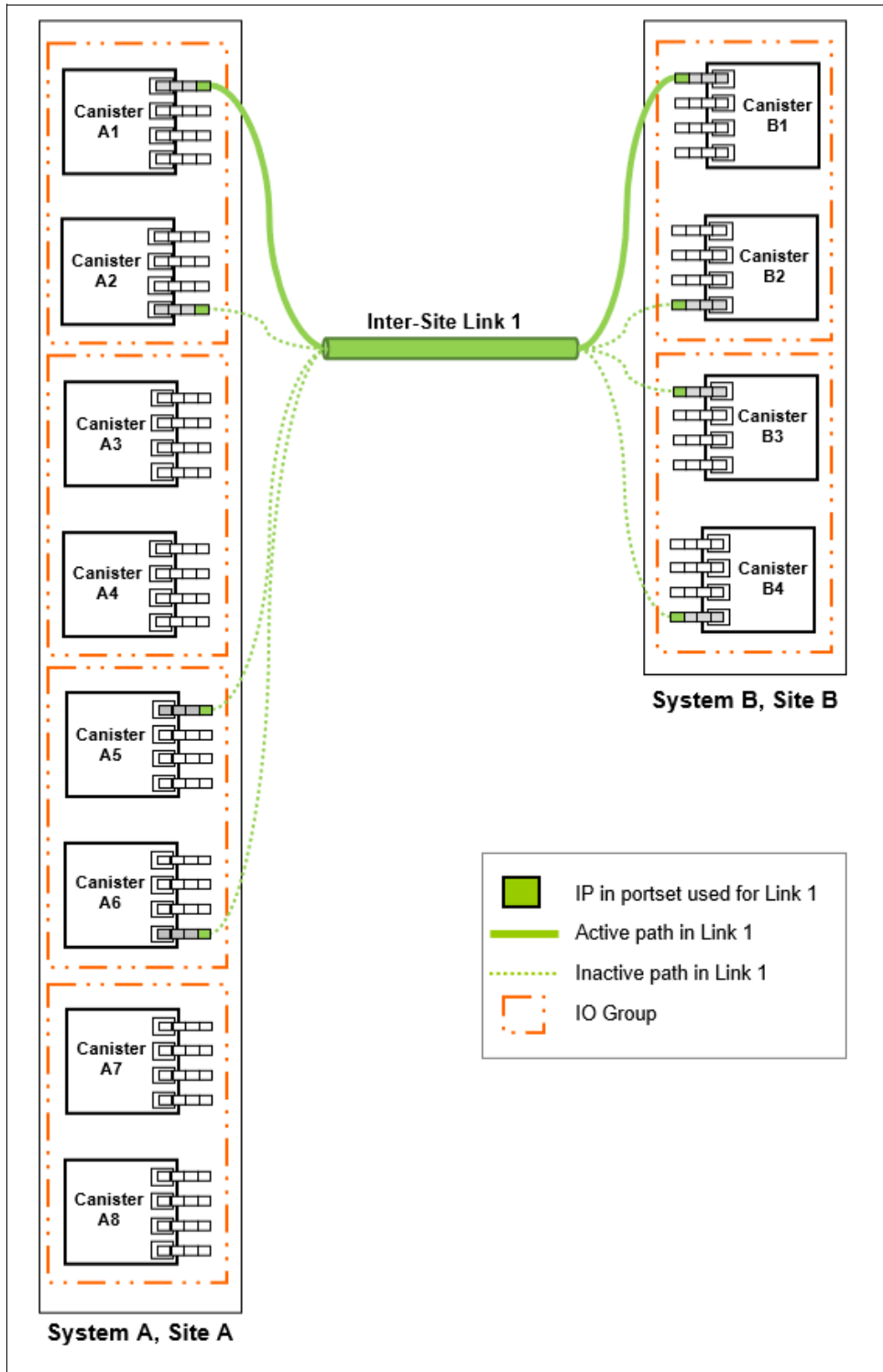


Figure 6-39 Clustered or multinode systems with a single inter-site link with only one link

Figure 6-39 shows a 4-control enclosure system or an 8-node system (System A in Site A) and a 2-control enclosure system or a 4-node system (System B in Site B). A single replication portset is used on canisters or nodes A1, A2, A5, and A6 on System A at Site A. Similarly, a single portset is used on canisters or nodes B1, B2, B3, and B4 on System B.

Although four control enclosures or four I/O groups (eight nodes) are in System A, only two control enclosures or I/O groups are configured for IP partnerships. Port selection is determined by a path configuration algorithm. The other ports play the role of standby ports.

If Canister or Node A1 fails in System A, IP partnership continues to use one of the ports that is configured in the portset from any of the canisters or nodes from either of the two control enclosures in System A.

However, it might take some time for discovery and path configuration logic to reestablish paths post-failover. This delay might cause partnerships to change to the `Not_Present` state. This process can lead to remote copy relationships stopping. The administrator must manually start them if the relationships do not auto-recover.

The details of which particular IP port is actively participating in IP partnership process is provided in the `1spartnership` output (reported as `link1_ip_id` and `link2_ip_id`).

This configuration includes the following characteristics:

- ▶ The replication portset that is used contains IP addresses from canisters of all the control enclosures. However, only one path is active at any time at each system.
- ▶ If Canister or Node A1 in System A or Canister or Node B2 in System B fails in the system, the IP partnerships trigger discovery and continue servicing the I/O from the failover ports.
- ▶ The discovery mechanism that is triggered because of failover might introduce a delay where the partnerships momentarily change to the `Not_Present` state and then recover.
- ▶ The bandwidth of the single link is used completely.

Two inter-site link configurations

A single control enclosure or two 2-node systems with a 2-inter-site link configuration is shown in Figure 6-40 on page 469.

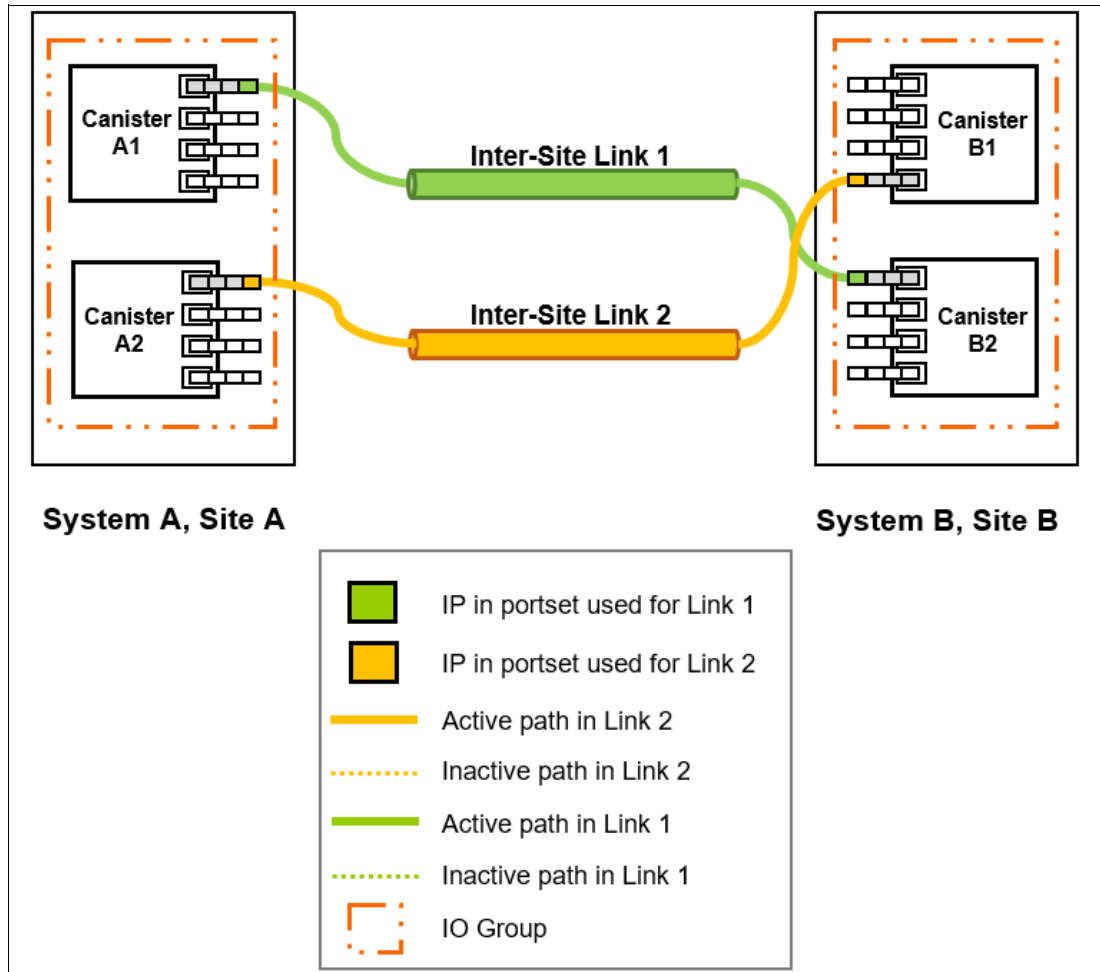


Figure 6-40 Dual links with two replication portsets on each system configured

As shown in Figure 6-40, two replication portsets are configured on System A and System B because two inter-site links are available. In this configuration, the failover ports are not configured on partner canisters or nodes in the control enclosure or I/O group. Rather, the ports are maintained in different portsets on both of the canisters or nodes. They can remain active and participate in an IP partnership by using both of the links. Failover ports cannot be used with this configuration because only one active path per canister per partnership is allowed.

However, if either of the canisters or nodes in the control enclosure or I/O group fail (that is, if Canister or Node A1 on System A fails), the IP partnership continues from only the available IP that is configured in portset that is associated to link 2. Therefore, the effective bandwidth of the two links is reduced to 50% because only the bandwidth of a single link is available until the failure is resolved.

This configuration includes the following characteristics:

- ▶ Two inter-site links exist, and two replication portset are used.
- ▶ Each node has only one IP address in each replication portset.
- ▶ Both IP addresses in the two portsets participate simultaneously in IP partnerships. Therefore, both the links are used.

- ▶ During a canister or node failure or link failure, the IP partnership traffic continues from the other available link. Therefore, if two links of 10 Mbps each are available and you have 20 Mbps of effective link bandwidth, bandwidth is reduced to 10 Mbps only during a failure.
- ▶ After the canister or node failure or link failure is resolved and failback happens, the entire bandwidth of both of the links is available as before.

A 4-control enclosure clustered system or an 8-node system in IP partnership with a 2-control enclosure clustered system or a 4-node system over dual inter-site links is shown in Figure 6-41.

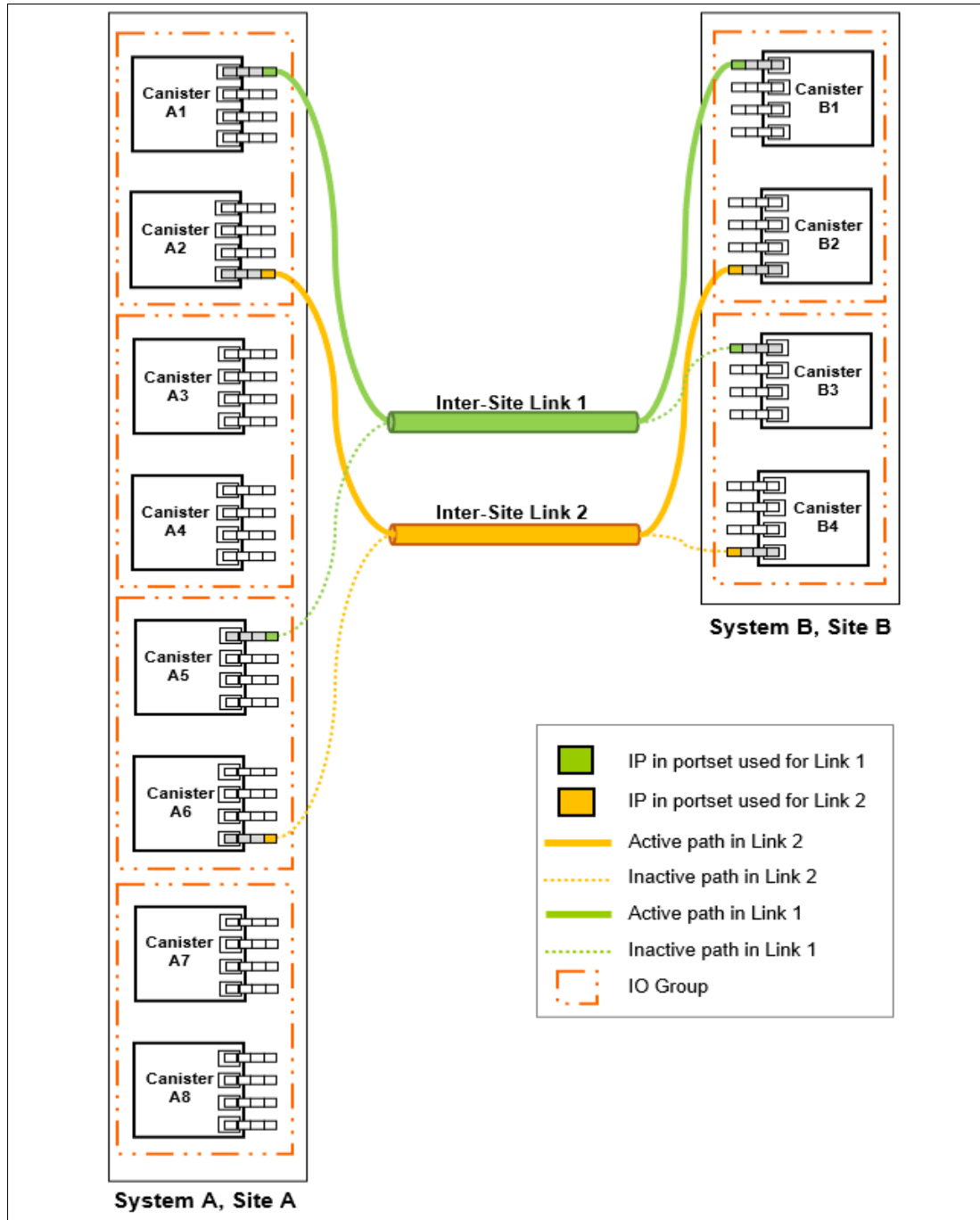


Figure 6-41 Clustered/multinode systems with dual inter-site links between the two systems

Figure 6-41 on page 470 shows a 4-control enclosure or an 8-node System A in Site A and a 2-control enclosure or a 4-node System B in Site B. Canisters or nodes from only two control enclosures or two I/O groups are configured with replication portsets in System A.

In this configuration, two links and two control enclosures or two I/O groups are configured with replication portsets. However, path selection logic is managed by an internal algorithm. Therefore, this configuration depends on the pathing algorithm to decide which of the canisters or nodes actively participate in IP partnership. Even if Canister or Node A5 and Canister or Node A6 have IP addresses that are configured within replication portsets properly, active IP partnership traffic on both of the links can be driven from Canister or Node A1 and Canister or Node A2 only.

If Canister or Node A1 fails in System A, IP partnership traffic continues from Canister or Node A2 (that is, link 2). The failover also causes IP partnership traffic to continue from Canister or Node A5 on which a portset that is associated to link 1 is configured. The details of the specific IP port actively participating in the IP partnership process is provided in the **Ispartnership** output (reported as `link1_ip_id` and `link2_ip_id`).

This configuration includes the following characteristics:

- ▶ Two control enclosures or two I/O groups have IP addresses that are configured in two replication portsets because two inter-site links for participating in IP partnership are used. However, only one IP per system in a particular portset remains active and participates in an IP partnership.
- ▶ One IP address per system from each replication portset participates in an IP partnership simultaneously. Therefore, both of the links are used.
- ▶ If a canister or node or port on the canister that is actively participating in the IP partnership fails, the remote copy data path is established from that port because another IP address is available on an alternative canister or node in the system within the replication portset.
- ▶ The path selection algorithm starts discovery of available IP addresses in the affected portset in the alternative I/O groups and paths are reestablished. This process restores the total bandwidth across both links.

Multiple partnerships configurations

In this section, some multiple partnerships configurations are described.

Figure 6-42 shows a 2-control enclosure or a 4-node System A in Site A, a 2-control enclosure or a 4-node System B in Site B, and a 2-control enclosure or 4-node System C in Site C.

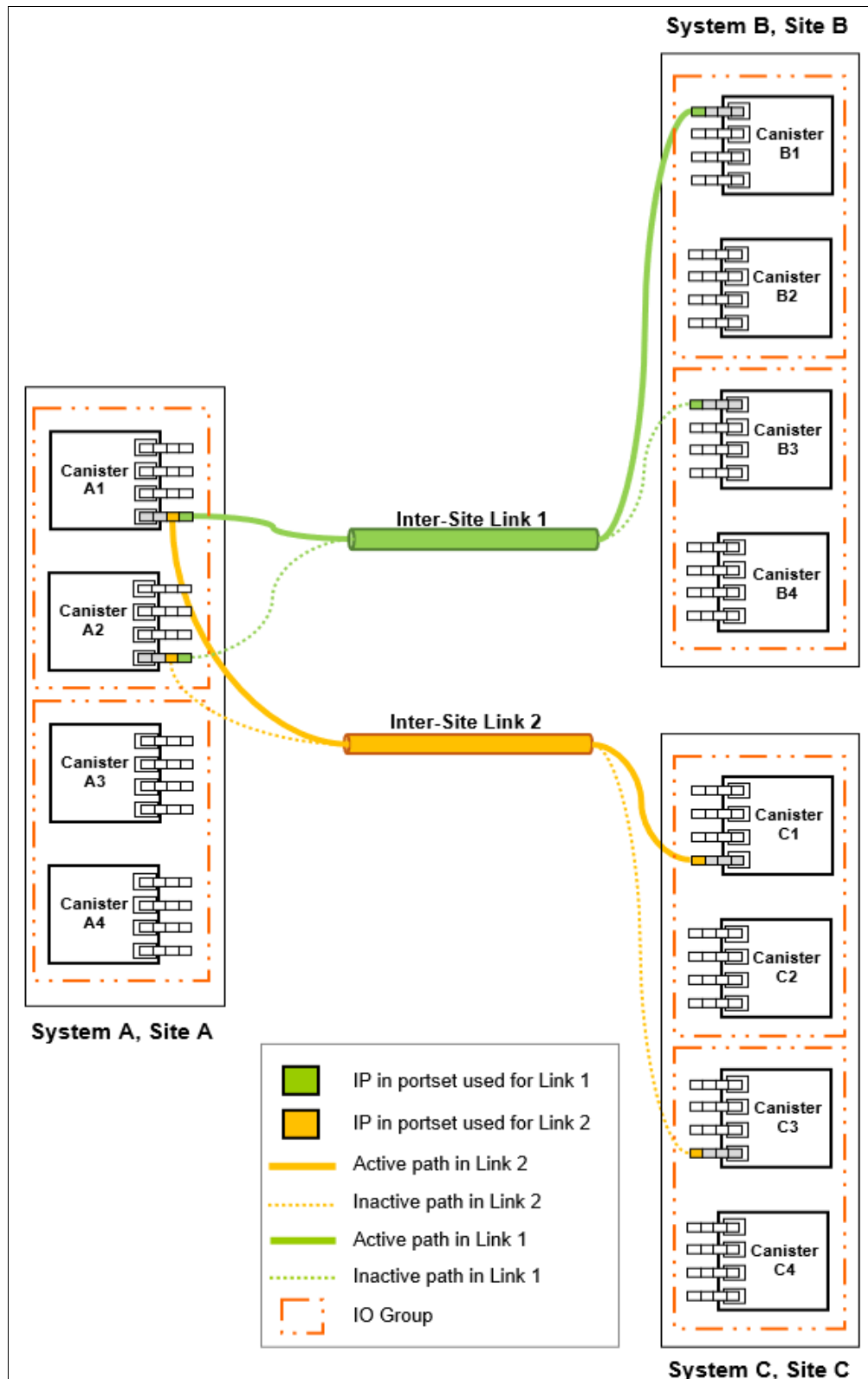


Figure 6-42 Multiple IP partnerships with two links and only one I/O group

In this configuration, two links and only one control enclosure or one I/O group are configured with replication portsets in System A. Both replication portsets use the same Ethernet ports in Canister or Node A1 and A2. System B uses a replication portset that is associated to link 1, and System C uses a replication portset that is associated to link 2. System B and System C have configured portsets across both control enclosures.

However, path selection logic is managed by an internal algorithm. Therefore, this configuration depends on the pathing algorithm to decide which of the canisters or nodes actively participate in IP partnerships. In this example, the active paths go from Canister or Node A1 to Canister or Node B1 and Canister or Node A1 to Canister or Node C1. In this configuration, multiple paths are allowed for a single canister because they are used for different IP partnerships.

If Canister or Node A1 fails in System A, IP partnerships continue servicing replication I/O from Canister or Node A2 because a fail-over port is configured on that node.

However, it might take some time for discovery and path configuration logic to reestablish paths post-failover. This delay can cause partnerships to change to Not_Present for that time, which can lead to a replication stopping. The details of the specific IP port that is actively participating in then IP partnership is provided in the `lspartnership` output (reported as `link1_ip_id` and `link2_ip_id`).

This configuration includes the following characteristics:

- ▶ One IP per system from each replication portset participates in the IP partnership simultaneously. Therefore, both links are used.
- ▶ Replication portsets on System A for both links are defined in the same physical ports.
- ▶ If a canister or node or port on the canister or node that is actively participating in an IP partnership fails, the remote copy data path is established from that port because another IP address is available on an alternative canister in the system within the replication portset.
- ▶ The path selection algorithm starts the discovery of available IP addresses in the affected portset in the alternative control enclosures, and paths are reestablished. This process restores the total bandwidth across both links.

An alternative partnership layout configuration is shown in Figure 6-43.

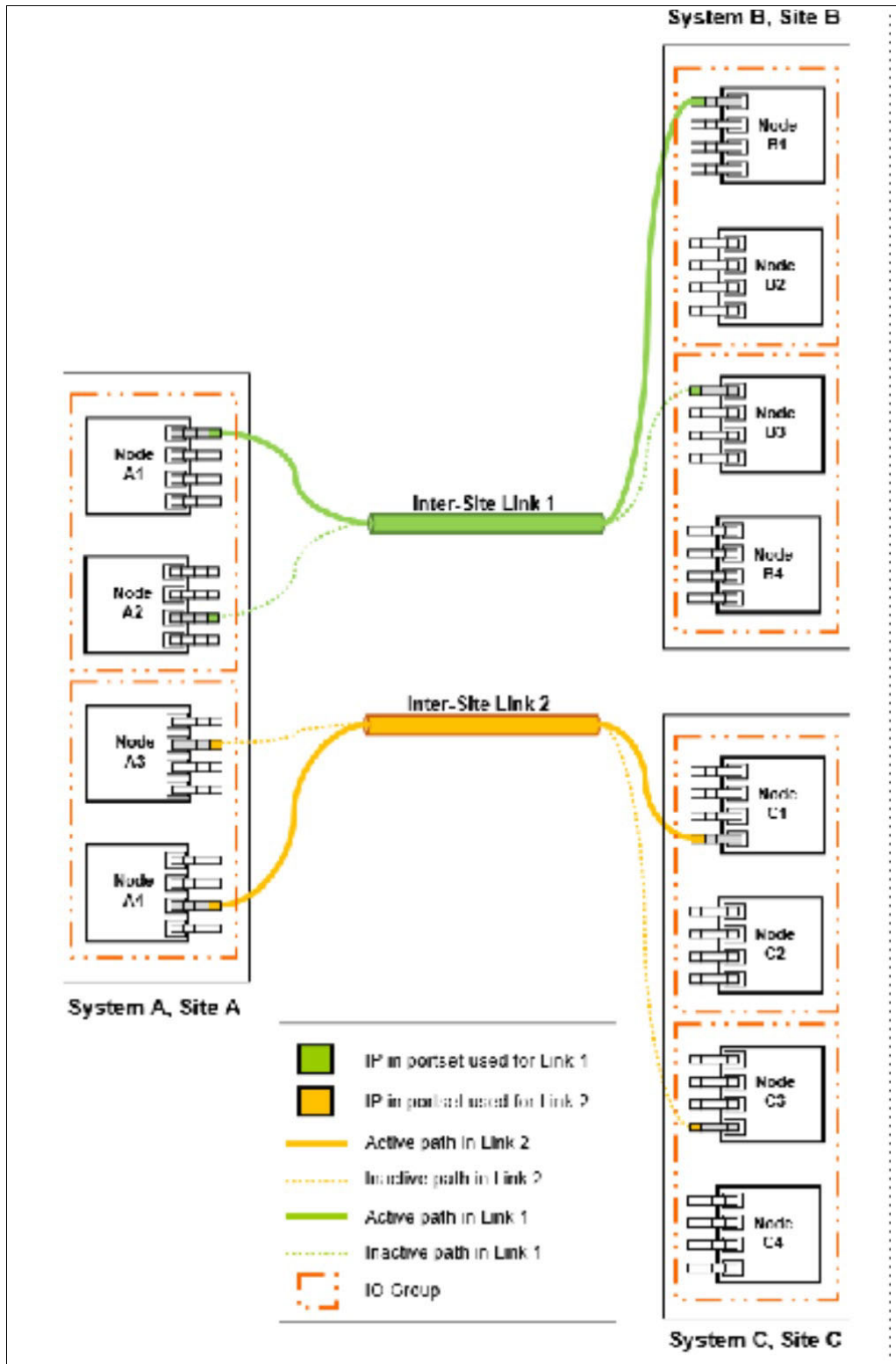


Figure 6-43 Multiple IP partnerships with two links

In this configuration, two links and two control enclosures or two I/O groups are configured with replication portsets in System A. System A control enclosure 0 or I/O Group 0 (Canister or Node A1 and Canister or Node A2) use IP addresses on the replication portset that is associated to link 1, and control enclosure 1 or I/O Group 1 (Canister or Node A3 and Canister or Node A4) use IP addresses on the replication portset that is associated to link 2. System B uses a replication portset that is associated to link 1, and System C uses a replication portset that is associated to link 2. System B and System C have configured portsets across both control enclosures or I/O groups.

However, path selection logic is managed by an internal algorithm. Therefore, this configuration depends on the pathing algorithm to decide which of the nodes actively participate in IP partnerships. In this example, the active paths go from Canister or Node A1 to Canister or Node B1 and Canister or Node A4 to Canister or Node C1 for System A to System B and System A to System C.

If Canister or Node A1 fails in System A, the IP partnership for System A to System B continues servicing replication I/O from Canister or Node A2 because a failover port is configured on that node.

However, it might take some time for the discovery and path configuration logic to reestablish paths post-failover. This delay can cause partnerships to change to Not_Present for that time, which can lead to a replication stopping. The partnership for System A to System C remains unaffected. The details of the specific IP port that is actively participating in the IP partnership are provided in the `lspartnership` output (reported as `link1_ip_id` and `link2_ip_id`).

This configuration includes the following characteristics:

- ▶ One IP address per system from each replication portset participates in the IP partnership simultaneously. Therefore, both of the links are used.
- ▶ Replication portsets on System A for the two links are defined in different physical ports.
- ▶ If a canister or node or port on the canister or node that is actively participating in the IP partnership fails, the remote copy data path is established from that port because another IP address is available on an alternative canister or node in the system within the replication portset.
- ▶ The path selection algorithm starts discovery of the available IP addresses in the affected portset in the alternative control enclosures or I/O groups, and paths are reestablished. This process restores the total bandwidth across both links.
- ▶ If a canister or node or link failure occurs, only one partnership is affected.

Replication portsets: Configuring two replication portsets provides more bandwidth and resilient configurations in a link failure. Two replication portsets also can be configured with a single physical link. This configuration makes sense only if the total link bandwidth exceeds the aggregate bandwidth of two replication portsets together. The usage of two portsets when the link bandwidth does not provide the aggregate throughput can lead to network resources contention and bad link performance.

6.5.8 Native IP replication performance considerations

Many factors affect the performance of an IP partnership. Some of these factors are latency, link speed, number of intersite links, host I/O, MDisk latency, and hardware. Since the introduction of IP partnerships, many improvements were made to make IP replication better performing and more reliable.

Nevertheless, with poor quality networks that have significant packet loss and high latency, the actual usable bandwidth might decrease considerably.

Figure 6-44 shows the throughput trend for a 1-Gbps port regarding the packet loss ratio and the latency.

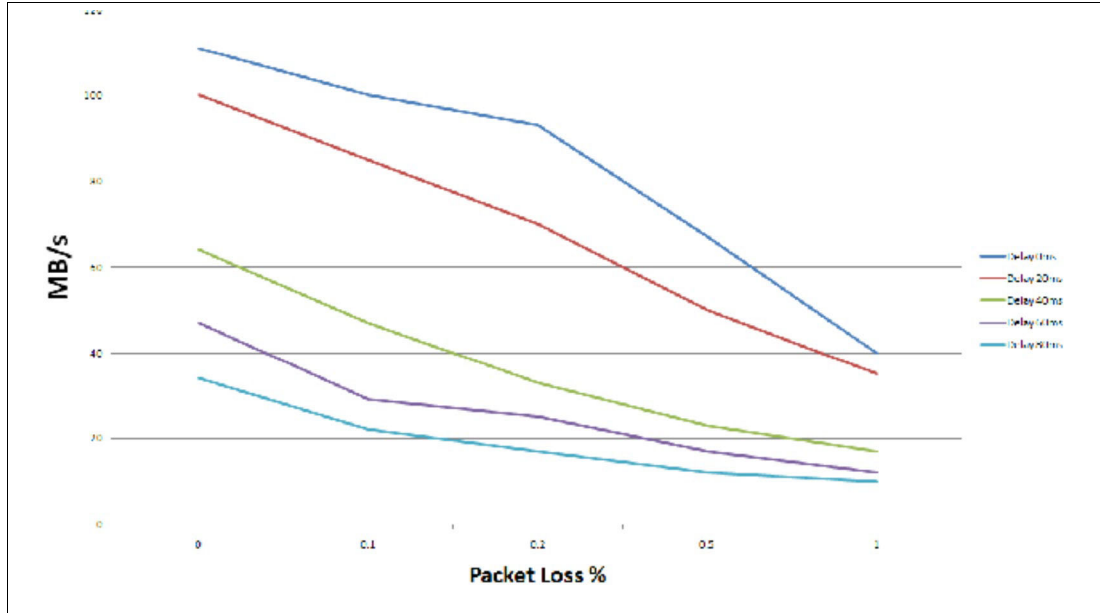


Figure 6-44 1-Gbps port throughput trend

Figure 6-44 shows how the combined effect of packet loss and latency can lead to a throughput reduction of more than 85%. For these reasons, the IP replication option should be considered only for replication configurations that are not affected by poor quality and poor performing networks. Due to its characteristic of low-bandwidth requirement, GMCV is the preferred solution with IP replication.

To improve performance when using compression and an IP partnership in the same system, use a different port for iSCSI host I/O and IP partnership traffic. Also, use a different VLAN ID for iSCSI host I/O and IP partnership traffic.

6.6 Volume mirroring

By using volume mirroring, you can have two physical copies of a volume that provide a basic RAID 1 function. These copies can be in the same storage pool or in different storage pools, with different extent sizes of the storage pool. Typically, the two copies are allocated in different storage pools.

The first storage pool contains the original (primary volume copy). If one storage controller or storage pool fails, a volume copy is not affected if it was placed on a different storage controller or in a different storage pool.

If a volume is created with two copies, both copies use the same virtualization policy. However, you can have two copies of a volume with different virtualization policies. In combination with *thin-provisioning*, each mirror of a volume can be thin-provisioned, compressed or fully allocated, and in striped, sequential, or image mode.

A mirrored (secondary) volume has all the capabilities of the primary volume copy. It also has the same restrictions (for example, a mirrored volume is owned by an I/O group, as with any other volume). This feature also provides a *point-in-time copy* function that is achieved by “splitting” a copy from the volume. However, the mirrored volume does not address other forms of mirroring that are based on remote copy (GM or MM functions), which mirrors volumes across I/O groups or clustered systems.

One copy is the primary copy, and the other copy is the secondary copy. Initially, the first volume copy is the primary copy. You can change the primary copy to the secondary copy if required.

Figure 6-45 shows an overview of volume mirroring.

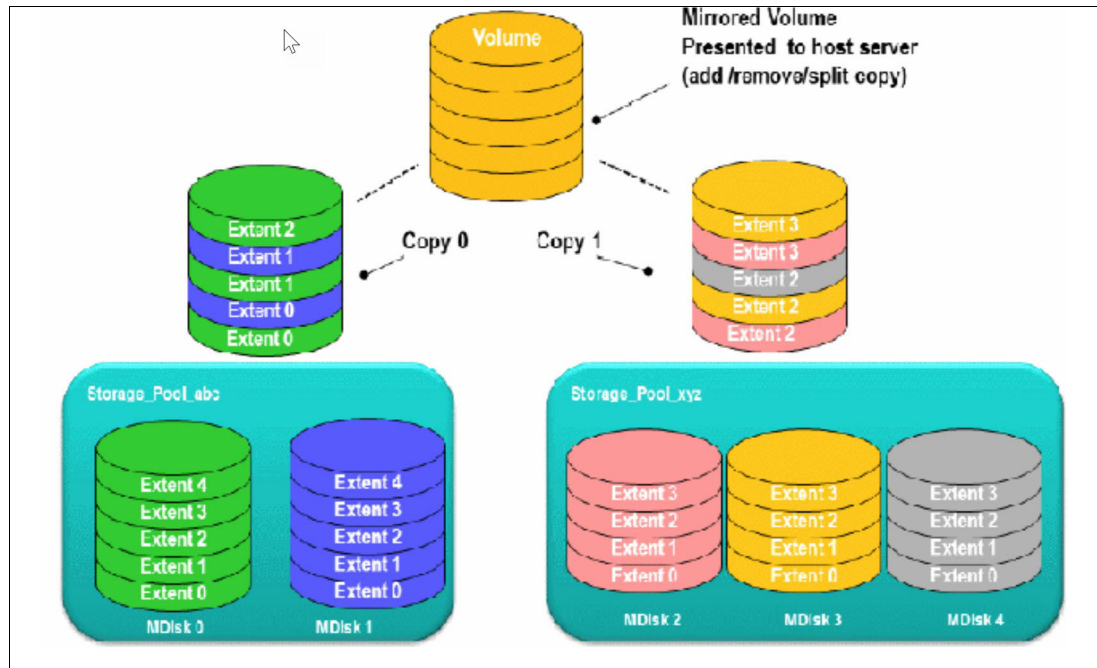


Figure 6-45 Volume mirroring overview

6.6.1 Read/write operations

Read/write operations behavior depends on the status of the copies and on other environment settings. During the initial synchronization or a resynchronization, only one of the copies is in synchronized status, and all the reads are directed to this copy. The write operations are directed to both copies.

When both copies are synchronized, the write operations are again directed to both copies. The read operations usually are directed to the primary copy unless the system is configured in an ESC topology, which applies to an SVC system only. With this system topology and the enablement of site awareness capability, the concept of primary copy still exists, but is not more relevant. The read operation follows the site affinity.

For example, consider an ESC configuration with mirrored volumes with one copy in Site A and the other in Site B. If a host I/O read is attempted to a mirrored disk through an IBM Storage Virtualize node in Site A, then the I/O read is directed to the copy in Site A, if it is available. Similarly, a host I/O read that is attempted through a node in Site B goes to the Site B copy.

Important: With an SVC ESC, keep consistency between the hosts, nodes, and storage controller site affinity as long as possible to ensure the best performance.

During back-end storage failure, note the following points:

- ▶ If one of the mirrored volume copies is temporarily unavailable, the volume remains accessible to servers.
- ▶ The system tracks the volume blocks that are written or changed, and resynchronizes these areas when both copies are available.
- ▶ The remaining copy service can read I/O without user intervention when the failing one is offline.

6.6.2 Volume mirroring use cases

Volume mirroring can provide extra copies of the data that can be used for HA solutions and data migration scenarios. You can convert a nonmirrored volume into a mirrored volume by adding a copy. When a copy is added by using this method, the cluster system synchronizes the new copy so that it is the same as the existing volume. You can convert a mirrored volume into a nonmirrored volume by deleting one copy or by splitting one copy to create a nonmirrored volume.

Access: Servers can access the volume during the synchronization processes that are described.

You can use mirrored volumes to provide extra protection for your environment or perform a migration. This solution offers several options:

- ▶ Stretched cluster configurations (only applicable to SVC)
Standard and ESC SVC configuration uses the volume mirroring feature to implement the data availability across the sites.
- ▶ Export to Image mode
With this option, you can move storage from *managed mode* to *image mode*. This option is useful if you are using IBM Storage Virtualize as a migration device. For example, suppose vendor A's product cannot communicate with vendor B's product, but you need to migrate existing data from vendor A to vendor B.
By using *Export to image mode*, you can migrate data by using copy services functions and then return control to the native array while maintaining access to the hosts.
- ▶ Import to Image mode
With this option, you can import an existing storage MDisk or LUN with its existing data from an external storage system without putting metadata on it. The existing data remains intact. After you import it, you can use the volume mirroring function to migrate the storage to the other locations while the data remains accessible to your hosts.
- ▶ Volume cloning by using volume mirroring and then by using the Split into New Volume option
With this option, any volume can be cloned without any interruption to host access. You must create two mirrored copies of the data and then break the mirroring with the split option to make two independent copies of data. This option does not apply to already mirrored volumes.

- ▶ Volume pool migration by using the volume mirroring option

With this option, any volume can be moved between storage pools without any interruption to host access. You might use this option to move volumes as an alternative to the Migrate to Another Pool function.

Compared to the Migrate to Another Pool function, volume mirroring provides more manageability because it can be suspended and resumed anytime, and you can move volumes among pools with different extent sizes. This option does not apply to already mirrored volumes.

Use case: Volume mirroring can be used to migrate volumes from and to DRPs, which do not support extent-based migrations. For more information, see 4.3.6, “External pools” on page 269.

- ▶ Volume capacity savings change

With this option, you can modify the capacity savings characteristics of any volume from standard to thin-provisioned or compressed and vice versa without any interruption to host access. This option works the same as the volume pool migration but specifies a different capacity savings for the newly created copy. This option does not apply to already mirrored volumes.

When you use volume mirroring, consider how quorum candidate disks are allocated. Volume mirroring maintains some state data on the quorum disks. If a quorum disk is not accessible and volume mirroring cannot update the state information, a mirrored volume might need to be taken offline to maintain data integrity. To ensure the HA of the system, ensure that multiple quorum candidate disks, which are allocated on different storage systems, are configured.

Quorum disk consideration: Mirrored volumes can be taken offline if there is no quorum disk that is available. This behavior occurs because the synchronization status for mirrored volumes is recorded on the quorum disk. To protect against mirrored volumes being taken offline, follow the guidelines that are described in the previous paragraph.

Here are other volume mirroring use cases and characteristics:

- ▶ Creating a mirrored volume:

- A user can create a maximum of two copies per volume.
- Both copies are created with the same virtualization policy, by default.

To have a volume that is mirrored by using different policies, you must add a volume copy with a different policy to a volume that has only one copy.

- Both copies can be in different storage pools. The first storage pool that is specified contains the primary copy.
- It is not possible to create a volume with two copies when specifying a set of MDisks.
- ▶ Add a volume copy to an existing volume:
 - The volume copy that is added can have a different space allocation policy.
 - Two existing volumes with one copy each cannot be merged into a single mirrored volume with two copies.
- ▶ Remove a volume copy from a mirrored volume:
 - The volume remains with only one copy.
 - It is not possible to remove the last copy from a volume.

- ▶ Split a volume copy from a mirrored volume and create a volume with the split copy:
 - This function is allowed only when the volume copies are synchronized. Otherwise, use the **-force** command.
 - It is not possible to recombine the two volumes after they are split.
 - Adding and splitting in one workflow enables migrations that are not currently allowed.
 - The split volume copy can be used as a means for creating a point-in-time copy (clone).
- ▶ Repair or validate volume copies by comparing them and performing the following three functions:
 - Report the first difference found. The function can iterate by starting at a specific LBA by using the **-startlba** parameter.
 - Create virtual medium errors where there are differences. This function is useful if there is back-end data corruption.
 - Correct the differences that are found (reads from primary copy and writes to secondary copy).
- ▶ View volumes that are affected by a back-end disk subsystem being offline:
 - Assume that a standard usage is for a mirror between disk subsystems.
 - Verify that mirrored volumes remain accessible if a disk system is being shut down.
 - Report an error in case a quorum disk is on the back-end disk subsystem.
- ▶ Expand or shrink a volume:
 - This function works on both of the volume copies at once.
 - All volume copies always have the same size.
 - All copies must be synchronized before expanding or shrinking them.

DRP limitation: DRPs do not support thin or compressed volumes shrinking.

- ▶ Delete a volume. When a volume is deleted, all copies are deleted for that volume.
- ▶ Migration commands apply to a specific volume copy.
- ▶ Out-of-sync bitmaps share the bitmap space with FlashCopy and MM and GM. Creating, expanding, and changing I/O groups might fail if there is insufficient memory.
- ▶ GUI views contain volume copy IDs.

6.6.3 Mirrored volume components

Note the following points regarding mirrored volume components:

- ▶ A mirrored volume is always composed of two copies (copy 0 and copy1).
- ▶ A volume that is not mirrored consists of a single copy (which, for reference, might be copy 0 or copy 1).

A mirrored volume looks the same to upper-layer clients as a nonmirrored volume. Upper layers within the cluster software, such as FlashCopy and MM and GM, and storage clients, do not know whether a volume is mirrored. They all continue to handle the volume as they did before without being aware of whether the volume is mirrored.

Figure 6-46 shows the attributes of a volume and volume mirroring.

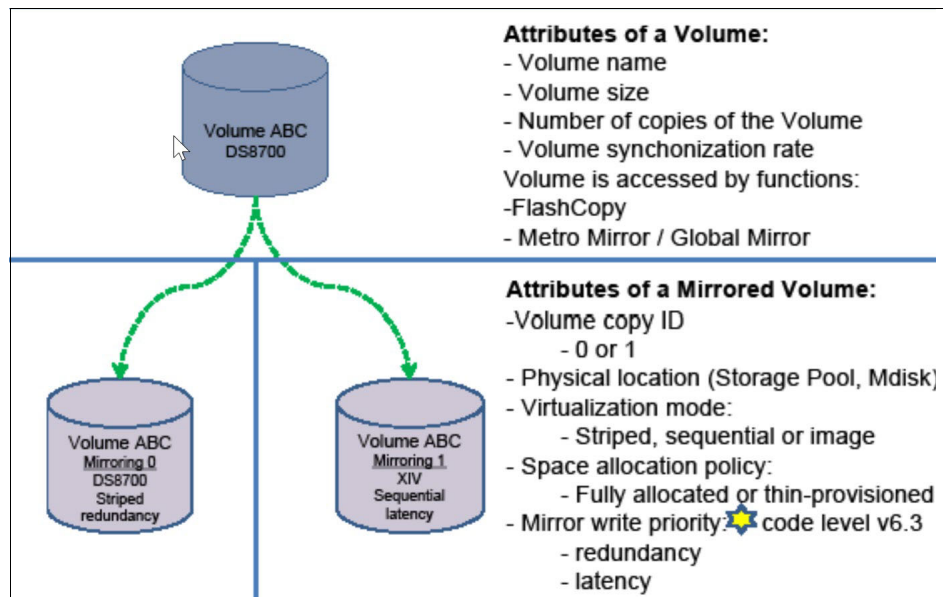


Figure 6-46 Attributes of a volume and volume mirroring

In Figure 6-46, IBM XIV and IBM DS8700 show that a mirrored volume can use different storage devices.

6.6.4 Volume mirroring synchronization options

When a volume is created with two copies, initially the copies are in the *out-of-sync* status. The primary volume copy (located in the first specified storage pool) is defined as in sync, and the secondary volume copy is defined as out of sync. The secondary copy is synchronized through the synchronization process.

This process runs at the default synchronization rate of 50 (as shown in Table 6-12), or at the defined rate while creating or modifying the volume. For more information about the effect of the copy rate setting, see 6.7.5, “Managing policy-based replication” on page 493. When the synchronization process completes, the volume mirroring copies are in the *in-sync* state.

Table 6-12 Relationship between the rate value and the data copied per second

User-specified rate attribute value per volume	Data copied per second
0	Synchronization is disabled.
1 - 10	128 KB
11 - 20	256 KB
21 - 30	512 KB
31 - 40	1 MB
41 - 50	2 MB (50% is the default value.)
51 - 60	4 MB
61 - 70	8 MB
71 - 80	6 MB

User-specified rate attribute value per volume	Data copied per second
81 - 90	32 MB
91 - 100	64 MB
101 - 110	128 MB
111 - 120	256 MB
121 - 130	512 MB
131 - 140	1024 MB
141 - 150	2048 MB

By default, when a mirrored volume is created, a format process also is initiated. This process ensures that the volume data is zeroed to prevent access to data that is still present on the reused extents.

This format process runs in the background at the defined synchronization rate, as shown in Table 6-12 on page 481. Before IBM Storage Virtualize 8.4, the format processing overwrites, with zeros, only Copy 0, and then synchronizes Copy 1. With version 8.4 or later, the format process is initiated concurrently to both volume mirroring copies, which eliminates the second synchronization step.

You can specify that a volume is synchronized (the `-createsync` parameter), even if it is not. Using this parameter can cause data corruption if the primary copy fails and leaves an unsynchronized secondary copy to provide data. Using this parameter can cause loss of read stability in unwritten areas if the primary copy fails, data is read from the primary copy, and then different data is read from the secondary copy. To avoid data loss or read stability loss, use this parameter only for a primary copy that was formatted and not written to. When using the `-createsync` setting, the initial formatting is skipped.

Another example use case for `-createsync` is for a newly created mirrored volume where both copies are thin-provisioned or compressed because no data is written to disk and unwritten areas return zeros (0). If the synchronization between the volume copies was lost, the resynchronization process is incremental, which means that only grains that were written to must be copied, which then receive synchronized volume copies again.

The progress of volume mirror synchronization can be obtained from the GUI or by using the `lsvdisksyncprogress` command.

6.6.5 Volume mirroring performance considerations

The write operation of mirrored volumes always occurs to both copies, which causes mirrored volumes to generate more workload on the cluster, back-end disk subsystems, and the connectivity infrastructure. The mirroring is symmetrical, and writes are acknowledged only when the write to the last copy completes. The result is that if the volumes copies are on storage pools with different performance characteristics, the slowest storage pool determines the performance of writes to the volume. This performance applies when writes must be destaged to disk.

Tip: Place volume copies of one volume on storage pools with the same or similar characteristics. Usually, if only good read performance is required, you can place the primary copy of a volume in a storage pool with better performance. Because the data is read always and only from one volume copy, reads are not faster than without volume mirroring.

However, this situation is only true when both copies are synchronized. If the primary is out of sync, then reads are submitted to the other copy.

Synchronization between volume copies has a similar impact on the cluster and the back-end disk subsystems as FlashCopy or data migration. The synchronization rate is a property of a volume that is expressed as a value of 0 - 150. A value of 0 disables synchronization.

Table 6-12 on page 481 shows the relationship between the rate value and the data that is copied per second.

Rate attribute value: The rate attribute is configured on each volume that you want to mirror. The default value of a new volume mirror is 50%.

In large IBM Storage Virtualize configurations, the settings of the copy rate can considerably affect the performance in scenarios where a back-end storage failure occurs. For example, consider a scenario where a failure of a back-end storage controller is affecting one copy of 300 mirrored volumes. The host continues the operations by using the remaining copy.

When the failed controller comes back online, the resynchronization process for all 300 mirrored volumes starts concurrently. With a copy rate of 100 for each volume, this process can add a theoretical workload of 18.75 GBps, which overloads the system.

Then, the general suggestion for the copy rate settings is to evaluate the impact of massive resynchronization and set the parameters. Consider setting the copy rate to high values for initial synchronization only, and with a few volumes at a time. Alternatively, consider defining a volume provisioning process that allows the safe creation of already synchronized mirrored volumes, as described in 6.7.4, “I/O group memory requirement” on page 492.

Volume mirroring I/O timeout configuration

A mirrored volume has pointers to the two copies of data, usually in different storage pools. Each write completes on both copies before the host receives the I/O completion status.

Synchronized mirrored volume copy is taken offline and goes out of sync if the following conditions occur. The volume remains online and continues to service I/O requests from the remaining copy.

- ▶ If a write I/O to a copy failed or a long timeout expired.
- ▶ The system completed all available controller level error recovery procedures (ERPs).

The *fast failover* feature isolates hosts from temporarily and poorly performing back-end storage of one copy at the expense of a short interruption to redundancy. The fast failover feature behavior is that during normal processing of host write I/O, the system submits writes to both copies with a timeout of 10 seconds (20 seconds for stretched volumes). If one write succeeds and the other write takes longer than 5 seconds, then the slow write is stopped. The FC abort sequence can take around 25 seconds.

When the stop completes, one copy is marked as out of sync, and the host write I/O completed. The overall fast failover ERP aims to complete the host I/O in approximately 30 seconds (or 40 seconds for stretched volumes).

The fast failover can be set for *each* mirrored volume by using the **chvdisk** command and the **mirror_write_priority** attribute settings:

- ▶ *Latency* (default value): A short timeout prioritizing low host latency. This option enables the fast failover feature.
- ▶ *Redundancy*: A long timeout prioritizing redundancy. This option indicates that a copy that is slow to respond to a write I/O can use the full ERP time. The response to the I/O is delayed until it completes to keep the copy in sync if possible. This option disables the fast failover feature.

Volume mirroring ceases to use slow copy for 4 - 6 minutes, and subsequent I/O data is not affected by a slow copy. Synchronization is suspended during this period. After the copy suspension completes, volume mirroring resumes, which allows I/O data and synchronization operations to the slow copy, which often quickly completes the synchronization.

If another I/O times out during the synchronization, then the system stops using that copy again for 4 - 6 minutes. If one copy is always slow, then the system tries it every 4 - 6 minutes and the copy gets progressively more out of sync as more grains are written. If fast failovers are occurring regularly, there is probably an underlying performance problem with the copy's back-end storage.

The preferred **mirror_write_priority** setting for the ESC configurations is *latency*.

6.6.6 Bitmap space for out-of-sync volume copies

The grain size for the synchronization of volume copies is 256 KB. One grain takes up 1 bit of bitmap space. 20 MB of bitmap space supports 40 TB of mirrored volumes. This relationship is the same as the relationship for copy services (Global and MM) and standard FlashCopy with a grain size of 256 KB (see Table 6-13).

Table 6-13 Relationship of bitmap space to volume mirroring address space

Function	Grain size in KB	1 byte of bitmap space gives a total of	4 KB of bitmap space gives a total of	1 MB of bitmap space gives a total of	20 MB of bitmap space gives a total of	512 MB of bitmap space gives a total of
Volume mirroring	256	2 MB of volume capacity	8 GB of volume capacity	2 TB of volume capacity	40 TB of volume capacity	1024 TB of volume capacity

Shared bitmap space: This bitmap space on one I/O group is shared between MM, GM, FlashCopy, and volume mirroring.

The command to create mirrored volumes can fail if there is not enough space to allocate bitmaps in the target I/O Group. To verify and change the space that is allocated and available on each I/O group by using the CLI, see Example 6-4.

Example 6-4 A *lsiogrp* and *chiogrp* command example

```
IBM_IBM FlashSystem:ITS0:superuser>lsiogrp
id name          node_count vdisk_count host_count site_id site_name
0  io_grp0        2          9          0          0
1  io_grp1        0          0          0          0
```



```

2 io_grp2      0      0      0
3 io_grp3      0      0      0
4 recovery_io_grp 0      0      0
IBM_IBM FlashSystem:ITS0:superuser>lsiogrp io_grp0|grep _memory
flash_copy_total_memory 20.0MB
flash_copy_free_memory 20.0MB
remote_copy_total_memory 20.0MB
remote_copy_free_memory 20.0MB
mirroring_total_memory 20.0MB
mirroring_free_memory 20.0MB
raid_total_memory 40.0MB
raid_free_memory 40.0MB
flash_copy_maximum_memory 2048.0MB
compression_total_memory 0.0MB
.
IBM_IBM FlashSystem:ITS0:superuser>chiogrp -feature mirror -size 64 io_grp0
IBM_IBM FlashSystem:ITS0:superuser>lsiogrp io_grp0|grep _memory
flash_copy_total_memory 20.0MB
flash_copy_free_memory 20.0MB
remote_copy_total_memory 20.0MB
remote_copy_free_memory 20.0MB
mirroring_total_memory 64.0MB
mirroring_free_memory 64.0MB
raid_total_memory 40.0MB
raid_free_memory 40.0MB
flash_copy_maximum_memory 2048.0MB
compression_total_memory 0.0MB

```

To verify and change the space that is allocated and available on each I/O group by using the GUI, see Figure 6-47.

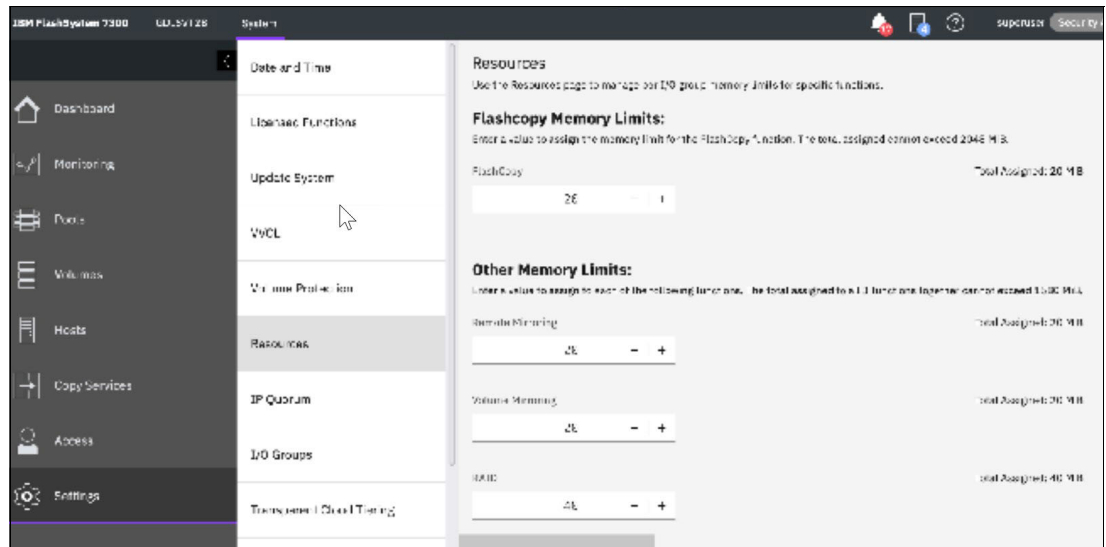


Figure 6-47 IOgrp feature example

6.7 Policy-based replication

Starting with firmware 8.5.2, a new feature called policy-based replication (PBR) has been introduced with enhanced replication implementation solution. The purpose of the PBR is to simplify configuration, management, and monitoring of the replication through the use of policies and volume groups.

Policy-based replication addresses the challenges of replication performance, usability, and scalability with Remote Copy. In the current release, it supports only Asynchronous replication over FC or IP partnerships. One can still continue using remote copy after firmware upgrade.

6.7.1 Implementing policy-based replication

You need to consider the following points when implementing policy-based replication:

- ▶ **Defining Replication Configuration:** Very first task in setting up PBR is the configuration of the policies. Normally, a *storage architect* is the person who defines the provisioning policies of the replicated volumes, policies for replication and bandwidth of the replication in partnership settings. One has to be a superuser in order to define the required configuration settings.
- ▶ **Daily Operations:** Managing the replicated volumes are carried out by a *storage administrator*. The tasks required for the management of replication are adding the volumes in the volume group, removing the volumes from the volume group, stopping/starting the replication, performing disaster recovery operations.

Note: The new method uses volume groups, which are similar to CGs. However, in addition to ensuring that a group of volumes is preserved at the same point in time, the volume group also enables the simplification of restoration or recovery to that point in time. It achieves this goal through the association of a group of volumes with a snapshot policy that determines frequency and retention duration.

Note: Adding a volume to a group automatically configures replication for it. It also permits configuration changes to be performed while the partnership is disconnected. The system automatically reconfigures the DR system once the partnership is reconnected. The same volume group might be used for multiple features (for example, with or without replication, Flash Copy snapshots and Safeguarded Copy), which allows for a single definition of the volumes that are required by an application

Establishing the partnership

Creating the partnership between Primary system and Secondary system is still the same and requires FC or IP connectivity as explained in Figure 6-13 on page 401.

A new requirement for partnerships that are used for policy-based replication is the installation of Transport Layer Security (TLS) certificates between partnered systems. The certificates are required to provide secure communication between systems for managing replication between the systems. Management traffic is authenticated by a certificate and is encrypted using SSL. However, system takes care of exchanging the certificates and no external certificate is required to be arranged and to be uploaded. (Self-signed and CA-signed certificates are supported too.)

It is recommended that you create partnerships by using the GUI as this automates the certificate exchange and setup. Both self-signed and signed certificates are supported. During this process, both system exchange certificates to give each system configuration access using REST API.

The maximum amount of bandwidth that is consumed by policy-based replication is controlled by the settings on the partnership. Differences between policy-based replication and Remote Copy exist in how the partnership settings are used. Bandwidth is not divided between Remote Copy and policy-based replication.

The partnership bandwidth defines the amount of bandwidth that is dedicated for replication between each of the I/O groups in megabits per second.

The background copy rate defines the amount of the bandwidth, expressed as a percentage that can be used for synchronization. For example, a link bandwidth of 1000 Mbps and a background copy rate of 50 allows for 500 Mbps of synchronization traffic (62.5 MBps). This control is a separate parameter because when Remote Copy is used, this control must be managed by the user to balance foreground host writes and background synchronization activity. Policy-based replication does not distinguish between the two cases and automatically manages the bandwidth for each type.

Best practices:

- ▶ If you are using only policy-based replication, then the background copy rate should be set to 100% to allow all available bandwidth to be used.
- ▶ If you are using both Remote Copy and Policy-based replication on the same partnership, then the background copy rate can be used to balance the bandwidth between policy-based replication and Remote Copy.

All traffic for policy-based replication is treated as background copy, so the rate can be managed to guarantee bandwidth to Remote Copy, which is more sensitive to changes in the available bandwidth.

Remote Copy has additional system settings for the following items:

- ▶ To control the synchronization rate of an individual relationship.

This setting is not used for policy-based replication; the system automatically manages the synchronization bandwidth.

- ▶ To control the amount of memory used for replication.

This setting is not used for policy-based replication.

IP partnership is supported up to 80ms RTT and FC partnership is supported up to 250ms with I/O group to I/O group zoning.

License requirement: The feature uses the existing remote license and the existing rules apply. PBR does not require or consume a FlashCopy license.

Creating storage pools

Storage pool creation and assigning the capacity on each system is the next major task in the replication configuration. Optionally, child pools inside the parent pool can also be used for the volume placement. Volumes which are part of the replication can reside in either parent pool or in the child pool.

Assigning provisioning policies

Provisioning policy is the feature which helps define how the new volume capacity should be provisioned in the pool. For example, full-allocated, thin provisioned, compressed+deduplicated and so on.

A set of default provisioning policies are created when the first parent pool is created. These are

- ▶ capacity_optimized (thin-provisioned, compressed)
- ▶ performance_optimized (creates fully-allocated volumes)

These policies can be renamed, deleted or user can create more similar custom policies. If IBM FlashCore modules are used in the storage pool, then fully-allocated volumes still benefit from compression performed by the FlashCore module.

Notes on provisioning policy:

- ▶ A provisioning policy specifying thin-provisioning can only be used with standard storage pools.
- ▶ A provisioning policy specifying compression can only be used with data reduction storage pools.
- ▶ A provisioning policy specifying deduplication can only be used with data reduction pools.

A provisioning policy can be associated with multiple pools. A pool can be associated with, at most, one provisioning policy. Provisioning policies are supported on parent pools and child pools. A provisioning policy exists within a single system (unlike replication policies, which are replicated between systems). Intentionally, few configurable options exist in the provisioning policy. The defaults for the other attributes are chosen to match the most common usage and best practices.

The association of the provisioning policy to a pool automatically causes it to be used for new volumes that are created in the pool. This means that the provisioning policy is used simply by specifying a pool that has a policy that is associated when creating a volume by using the GUI or the `mkvolume` command. Provisioning policies can also be used for non-replicated volumes to simplify volume creation.

Different pools can use different policies and they do not need to be symmetric.

Note: Provisioning policies are not used when change volumes are created for policy-based replication as these are always created by the system by using best practices.

Pool link

Storage pool linking is a mechanism which defines the target pool on the recovery system for the replication. Unlike partnership that defines the remote system to be used during disaster recovery, storage pool linking selects a specific pool on the recovery system for the volume copies, also referred as secondary volumes. A pool can be linked to only one pool per partnership. A pool link is required to be configured for a partnership to create replicated volumes in that pool. The pools can be parent pools or child pools.

For a standard topology system, if a volume is created locally as a mirrored volume between different pools, the system uses the pool link of the primary volume copy to identify the remote pool to use.

If the system topology is stretched, then the pool in site 1 is used to identify the linked pool; pools in site 2 are not required to be linked to replicate stretched volumes. The system does not provide a way to automatically create mirrored volumes on a remote system, but mirror copies can be added by the user.

The system does not provide a way to automatically create mirrored volumes on a remote system, but mirror copies can be added by the user.

Figure 6-48 shows the various options for pool linking for replication.

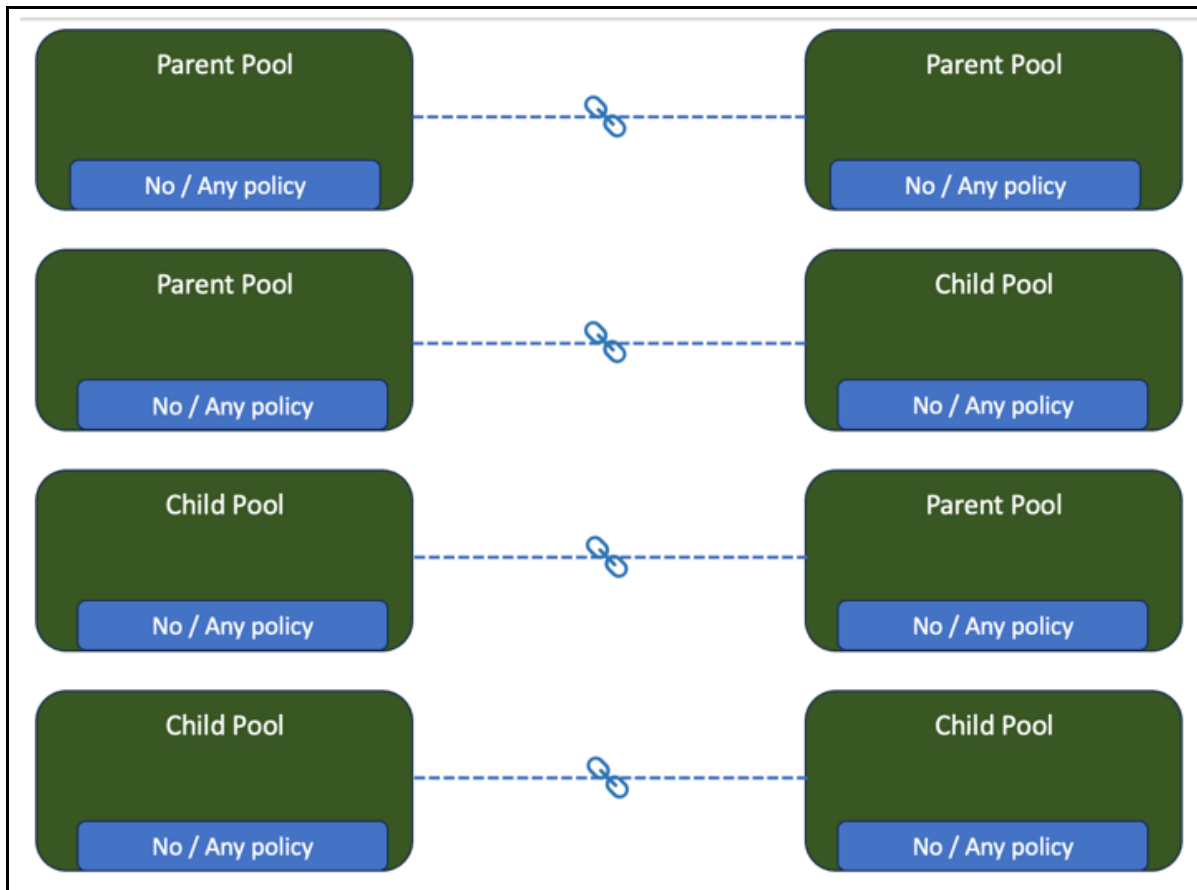


Figure 6-48 Possible pool and policy relations

Best practices: However, not all combinations will make sense for the pool linking. The best practice is to maintain the symmetry in assigning the policies across both locations while linking the pools.

Some best practice configuration options

The following are some scenarios and best practice configuration options:

Scenario-1

Client needs to create the pool configuration and want all volumes to be thick, no capacity savings.

We aim to create the simplest configuration with all volumes set to thick provisioning, without any capacity savings.

Link the parent pools. Provisioning policies are not required.

Scenario-2

We want all the volumes to be created with a particular capacity saving, e.g. thin/compressed/deduplicated.

Create and assign a provisioning policy on the local and remote pools. The policy on each system would specify the same capacity savings so that the production and recovery volumes are created the uniformly.

Link the parent pools.

Scenario-3

We need some volumes to be created thick and others to be created thin (or some other capacity saving), such as thin/compressed/deduplicated.

To achieve this, we will use child pools with different provisioning policies for each pool. Using a data reduction pool as the parent pool will allow the child pools to be quotaless, simplifying capacity management and ensuring efficient storage utilization.

Scenario-4

Client wants to allocate a fixed amount of storage for different groups of users/applications by utilizing child pools.

By setting a standard pool as the parent pool, we can enforce quotas on the child pools, ensuring that each group gets the allocated storage. Optionally, we can assign provisioning policies to the volumes within the child pools, based on your preferences, if capacity savings are desired. This approach allows for efficient storage allocation and management across various user/application groups.

Replication policies

Replication policies are a key concept for policy-based replication as they define how replication should be configured between partnered systems. A replication policy specifies how replication should be applied to a volume group and therefore to all volumes within that group. A replication policy can be associated with any number of volume groups. A volume group can have at most one replication policy associated with it.

A replication policy defines three key attributes:

- ▶ **A set of locations** defining which I/O groups on which of the partnered systems should contain a copy of the volume group, where the data should be replicated.
- ▶ **A topology** representing how the systems are organized and the type of replication that should be performed between each location, how the data should be replicated between the locations.
- ▶ **A name** to uniquely identify the replication policy on both systems.

Replication policy cannot be renamed and modified once they are in use and are associated with the replication. This is important as it guarantees that both systems always have the same definition of how replication should be configured. If changes are required to a replication policy, a new policy can be created with the desired changes, and the associated volume groups can be reassigned with the new policy.

Note: Each system supports up to 32 replication policies to be created, which allows for various locations, topologies, and RPOs to be defined.

The creation of a replication policy results in the system creating the policy on all systems that are defined in the replication policy. A replication policy can only be created when the systems are connected. A replication policy can be deleted only if there are no volume groups that are associated with the policy and can be deleted while the systems are disconnected.

Three replication modes exist for volume groups:

- ▶ **Production:** Volumes in the volume group are accessible for host I/O and configuration changes are allowed. This copy acts as the source for replication to any recovery copies. By definition, a production volume group always contains an up-to-date copy of application data for host access.
- ▶ **Recovery:** Volumes in the group are able to receive only replication I/O; volumes act as a target for replication and cannot be used for host I/O. Most configuration changes are not permitted and must be performed on the production copy.
- ▶ **Independent:** If independent access is enabled for the volume group, such as during a disaster, each copy of the volume group is accessible for host I/O and permits configuration changes. Replication is suspended in this state.

Replication policies do not define the direction in which the replication is performed. Instead, the direction is determined when a replication policy is associated with a volume group, or a volume group is created by specifying a replication policy. The system where this action was performed is configured as the production copy of the volume group.

6.7.2 Limitations and restrictions

The following are limitations and restrictions:

- ▶ **Policy-based replication:**
 - Only 32 replication policies are allowed per system.
 - Maximum of two I/O groups per system using policy-based replication. Use volumes in the volume group from the same I/O group.
 - Only 1024 volume groups per system are allowed.
 - 512 volumes per volume group (reminder: as with snapshots, the pause time is proportionate to the volume count per volume group).
 - Policy-based replication capacity per I/O group.
 - Policy-based replication cannot be used with systems in HyperSwap topology.
- ▶ **Volume groups:**
 - Volumes must be in a volume group to be replicated.
 - Volumes cannot be moved between replicated volume groups.
 - Volume groups cannot be renamed once a replication policy is associated.
 - Volumes in replicated volume groups cannot be renamed.
 - Volume groups must have a name that is unique on both systems as the recovery copy is created with the same name.
 - Volume names must be unique only on the system they are created on.
 - If the volume name is not available on the recovery system, it will be appended with `_<number>` until a unique value is found.
- ▶ **Volume related considerations:**
 - Recovery copies of a volume are never created as mirrored volumes.

- A replicated volume group cannot use Transparent Cloud Tiering.
 - Remote Copy and policy-based replication can be configured on the same volume only if the volume is operating as the production volume for both types of replication, for the purpose of migrating from Remote Copy to policy-based replication.
 - Restore-in-place using FlashCopy or using replicated volumes as the target of a legacy FlashCopy map is not supported.
 - Replicated volumes cannot be expanded or shrunk.
 - Workaround is to remove from the volume group, expand the volume, add to volume group.
 - Replicated volumes must have cache enabled.
 - No support for VVol replication and TCT (Transparent Cloud Tiering) Volumes.
 - Image-mode volumes are not supported.
- ▶ **General:**
- 3-site capabilities are not supported by PBR.
 - Ownership groups are not supported by PBR.

6.7.3 Supported platforms for 8.5.2 or above:

Table 6-14 provides the supported platforms for policy-based replication with firmware 8.5.2 or above.

Table 6-14 Supported platforms

Model	Replicated Volumes	Replicated Capacity
FlashSystem 50x5	Not supported	Not supported
FlashSystem 5200*	7932	1024 TiB
FlashSystem 7200	7932	2048 TiB
FlashSystem 7300	7932	2048 TiB
FlashSystem 91x0	7932	2048 TiB
FlashSystem 9200	7932	2048 TiB
FlashSystem 9500	10000	2048 TiB**
SAN Volume Controller	7932	2048 TiB
SV Public Cloud(Azure)*	7932	1024 TiB
SV Public Cloud (Others)	Not supported	Not supported

* Requires minimum of 64GB of memory per node.

** In the Future releases

6.7.4 I/O group memory requirement

Table 6-15 on page 493 shows the bitmap memory requirement per I/O group for policy-based replication.

Table 6-15 Bitmap memory per I/O group

Model	Total I/O group memory (excluding FlashCopy)	Required I/O group memory for policy-based replication	Available for other features (RAID, Mirroring, Remote Copy)
FlashSystem 5200	2088 MiB	1537 MiB	551 MiB
FlashSystem 7200	3628 MiB	2561 MiB	1067 MiB
FlashSystem 7300	3628 MiB	2561 MiB	1067 MiB
FlashSystem 91x0	3328 MiB	2561 MiB	1067 MiB
FlashSystem 9200	3628 MiB	2561 MiB	1067 MiB
FlashSystem 9500	5920 MiB	4609 MiB	1311 MiB
SAN Volume Controller	3628 MiB	2561 MiB	1067 MiB
SV Public Cloud (Azure)	2088 MiB	1537 MiB	551 MiB

6.7.5 Managing policy-based replication

Replication is configured on the volume group by assigning a replication policy to the volume group. The system then automatically performs actions to configure replication according to the replication policy for all volumes in the volume group. This simplifies the configuration of replication as it is only configured once. Adding a volume to a group automatically configures replication for it. It also permits configuration changes to be performed while the partnership is disconnected. The system automatically reconfigures the DR system once the partnership is reconnected.

To change the direction of replication, the volume group must first be made independent. After the volume group is independent, the direction can be selected by choosing which location should be the production copy of the volume group and then restarting replication from that system. If you wish to perform a planned failover, it is important to suspend application I/O and ensure that the recovery point is more recent than when the application stopped performing I/O. Failure to do this can result in data missing after the change of direction.

Note: Creating volumes in a replicated volume groups skips the need for a full synchronization.

An important concept is that the configuration and the data on the volumes are coupled and the recovery point is formed from both the configuration and volume data. Adding, creating, removing, or deleting volumes from a volume group occurs on the recovery system at the equivalent point-in-time as they did on the production system. This means that if independent access is enabled on a volume group during a disaster, then the data and volumes that are presented are as they were from a previous point in time on the production system. This might result in partially synchronized volumes being removed from the recovery system when independent access is enabled, if those volumes are not in the recovery point.

When a volume is deleted from the production volume group, the copy of the volume from the recovery volume group will not be immediately deleted. This volume is deleted when the recovery volume group has a recovery point that does not include that volume.

This coupling requires that all configuration changes to the volume group are made from the system that has the production copy of the volume group and the changes are reflected to the recovery system. The exception to this requirement is the action of enabling independent access on a recovery copy of a volume group as this action can be performed only on the system that includes the recovery copy of the volume group.

When a replication policy is removed from a volume group, the recovery volume group and its volumes are deleted. If you need to keep the recovery copy, independent access should be enabled on the recovery copy first. Then, the replication policy can be removed from the volume group.

Note: The replication policy that is associated with a volume group might be changed to a compatible policy without requiring a full resynchronization. A compatible policy is defined as having the same locations and allows for a replication policy to be replaced by another policy with a different recovery-point objective without interrupting replication. If you attempt to change the policy to an incompatible policy, it is rejected by the system. The replication policy needs to be removed and the new policy must be associated with the volume group.

6.7.6 Production overview

The production system has two main responsibilities: processing host I/O requests and replicating the contents of the volume to the remote system. The production system can be identified by the location that is shown as production for the volume group.

The production copy can operate in the following three modes:

- ▶ **Change recording mode:** where new host writes are tracked, written locally but are not replicated.
- ▶ **Journaling mode:** where new host writes are tracked, written locally, and replicated in-order. This mode has a lower recovery point than cycling mode, but requires more bandwidth.
- ▶ **Cycling mode:** where new host writes are tracked, written locally and periodically replicated from a snapshot of the production volume recorded on a change volume.

Change recording mode is the default mode and is selected whenever the volume group cannot replicate; this could be because of offline volumes, errors, pending configuration changes, or if the partnership is disconnected. In this mode, the system records which regions of the volume have been written to and need to be synchronized when replication resumes.

When replication becomes possible, the production copy automatically selects either journaling or cycling mode to perform replication based on which is best for the current conditions. This auto-tuning allows for the system to dynamically switch between the two modes, attempting to achieve the lowest possible recovery point while also avoiding latency problems for the production volumes.

Write tracking and overlapping writes

Each replicated volume is allocated non-volatile memory to track unsynchronized volume regions, known as a bitmap. Both nodes in an I/O group include a copy of the bitmap and updates to the bitmap are mirrored between the nodes. Every volume is split into regions that are called grains; each grain is a contiguous 128 KiB range and each bit in the bitmap represents one grain.

Figure 6-49 depicts the sample of bitmap and region mapping.

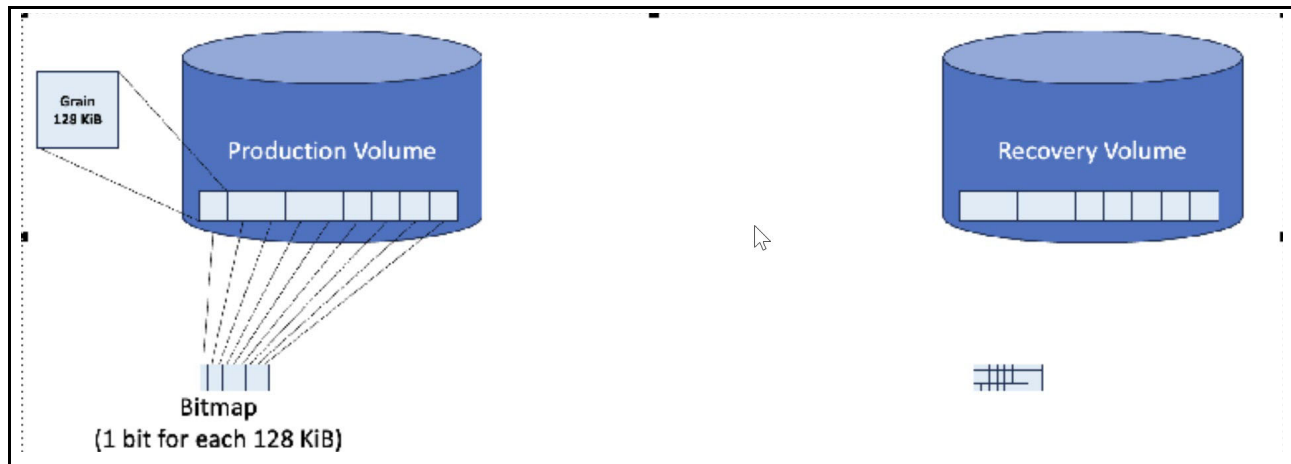


Figure 6-49 Region mapping with bitmap

Whenever a write is received from an application, the grain is recorded as unsynchronized. The bitmap is set to synchronized for a grain once the recovery volume contains identical data. The bitmap is primarily used for to identify regions of the volume that need to be synchronized if replication is suspended. The bitmap also allows for overlapping writes to be detected and resolved.

If one or more writes to the same region of the volume are active at the same time, it is known as an *overlapping write*. It might sometimes be referred to as a colliding write. The storage system needs to guarantee that the writes are applied in the same order on both systems to maintain data consistency. In theory there should never be two active writes to the same logical block; this behavior is described as invalid by storage specifications, but in practice it can happen and must be handled by the storage array. To guarantee the ordering of overlapping writes, the system selects one of the writes to process first. This write must be completed before the processing of the second write is started. For best possible performance, application workloads should be aligned to 8 KiB boundary or multiples thereof (16 KiB, 32 KiB, and so on).

Sequencing

In journaling mode, every write is tagged with a sequence number that defines the order in which the writes must be replayed at the recovery system to maintain consistency. Sequencing is across all volumes in a volume group to maintain mutual consistency. To achieve this, one node in the production I/O group is selected to generate sequence numbers for the volume group. As a write is written into the write cache, a parallel process requests a sequence number for the write. Once the sequence number is obtained and the local write is complete, the host write is eligible for replication.

For performance reasons, sequencing is performed only within an I/O group and all volumes in a volume group that uses policy-based replication must be in the same I/O group. A common cause of performance problems with Remote Copy Global Mirror was the requesting sequence numbers as the sequence number generator could be any node in the system. In this new model, it is either the node that received the write or its I/O group partner node that generates the sequence numbers. The system has significantly better performance characteristics when it sends messages to its I/O group partner node than any other node in the system, which results in better performance and a more consistent solution. Each volume group operates independently and there is no coordination of sequence numbers between volume groups.

Synchronization operations are also tagged with a sequence number, so they are correctly interleaved into the stream of writes.

Journal

Each node maintains a large, volatile memory buffer for journaling host writes and synchronization reads. The exact amount of journal capacity varies between models but (at the time of writing) it can be between 1 GiB and 32 GiB per node. The journal size is related to the number of CPU cores and memory in the node. The purpose of the journal is twofold.

- ▶ Providing temporary in-memory storage for host writes or synchronization reads until they can be replicated to the recovery copy and written to the local cache.
- ▶ Buffering host writes in journaling mode so that replication and the local host write are decoupled, which means replication problems cause the recovery point to extend instead of delaying I/O for the production application.

Writes are replicated from the journal to the recovery system in sequence number order. On a per-recovery system I/O group basis, writes are replicated in a first-in-first-out basis to ensure that the recovery system can always make progress. Writes are replicated to the preferred node of the volume in the remote system. If a direct connection is not present between the two nodes, it is routed through local or remote nodes. If the preferred node is not available in the remote system, the write is replicated to the online node in the caching I/O group for the recovery volume.

To ensure that replication does not cause application performance problems, the journal usage is constantly monitored by the system so it can proactively ensure that there is free capacity in the journal for new host writes. The journal is divided by CPU core and maintains lists of writes per volume group. However, resources are not divided between volume groups as this could lead to unfairly penalizing volume groups with higher write workloads.

If the throughput of replication is less than the throughput of the local writes, then the journal starts to fill up. The journal size is determined so that it can accommodate temporary bursts in write throughput. However, if it is a sustained increase in the write throughput versus what can be replicated, a proactive journal purge is triggered to prevent exhausting the resources. All volume groups that are replicating in the I/O group have their journals evaluated, within the context of the RPO defined on the replication policy, for peak and average usage and evaluated. Volume groups that are associated with higher RPO replication policies and those consuming the most journal resources are contenders to be purged first.

Purging a journal involves discarding the journal resources recording host writes that have completed locally, but not yet replicated for the volume group. The volume group then uses the bitmap to synchronize any grains that were affected by writes that were in the journal.

Volume groups have their journals purged if the volume group encounters an error that prevents replication, such as offline volumes or changes in connectivity between the two systems.

6.7.7 Recovery overview

The recovery system runs a process that replays the writes in the order that they were processed at the production location and maintains mutual consistency for the volumes in the volume group.

As writes are received from the production location, the recovery system orders them based on their sequence number. The writes are then mirrored between the two nodes in the recovery system I/O group and written to the volume once all earlier writes are mirrored and processed. Non-overlapping writes can be submitted in parallel to optimize performance but overlapping writes are replayed in sequence number order to ensure consistency. Only one node submits a write, usually the preferred node, but both nodes track its progress through the process.

Writes are recorded in non-volatile memory on both nodes in the I/O group so that they can be recovered if the nodes restart. After the write is stored in non-volatile memory on both nodes, it is marked as committed. Committing a write guarantees that it will be written in the future to the volume and that it forms part of the recovery point.

This is a simplified overview of the process that provides a high throughput method of replaying the writes, which is always able to establish a mutually consistent recovery point for all volumes in a volume group.

Note: Volumes in the recovery system are always provisioned by the system; existing volumes will not be used.

Change volumes

Every replicated volume has a change volume associated; these are managed automatically by the system and are hidden from the CLI and GUI views as there are limited user-interactions permitted to them.

A change volume is a thin-provisioned volume or a thin-provisioned and compressed volume if the volume it is associated with is in a data-reduction pool. If all copies of the volume are deduplicated then the change volume is also created as deduplicated, but not available as deduplication source. Change volumes are in the same I/O group with the same preferred node as the host-accessible volume and are created in the same pool; if a mirrored volume is created, then a copy of the change volume is created in both pools. They are always created as cache-enabled volumes with default Easy Tier settings.

Two FlashCopy maps are created between the volume and the change volume so that these maps can be used to either take or restore a snapshot of the volume. Change volumes do not require or contribute towards the FlashCopy license.

It is crucial to consider the capacity of change volumes when provisioning a system that uses replication. The used capacity of change volumes changes significantly depending on the amount of data that is required to preserve the snapshot. In most cases the data consumes a small percentage of the virtual capacity of the volume. However, in extreme cases it is possible for the data to consume an equal capacity to the volume. A general rule is to size the system with an additional 10-20% capacity of the replicated volumes for change volumes. However, this can vary by system. If there is low bandwidth between sites, but a high write-throughput at the production location, then anticipate the need for more capacity for change volumes. Similarly, if you are planning for the systems to be disconnected for an extended time, the synchronization when they reconnect might need to copy a significant amount of data. The change-volume capacity grows according to the amount of synchronization that is needed.

Note: For example, if a system has 25 volumes that are each 40 GiB, the virtual capacity of that system is 1000 GiB. The amount of usable capacity that is used depends on the data that is written and the data reduction features that are used. Twenty-five change volumes are created in each system that, by default, consume negligible usable capacity. If data reduction features are not used, it would be a best practice to have between 1.1 and 1.2 TiB of usable capacity in this example.

If a change volume runs out of space, it might prevent replication.

The FlashCopy maps that are used by the change volumes are automatically started and stopped, as required, to achieve the desired behavior for replication.

In general, the change volume is automatically maintained by the system in line with the volume it protects. However, special cases exist for mirroring and migration when you use the CLI.

- ▶ If a volume is migrated to a different pool, the change volume must also be migrated. When you use the GUI, this migration happens automatically.

However, a CLI user is required to migrate the change volume in the same way as the user volume. The change-volume vdisk ID can be seen in the **lsvdisk** view of the user volume in the `changevolume1_id` field.

Note: A CLI user can also specify the `-showhidden` parameter on the CLI views to display the hidden change volume vdisks and FlashCopy maps.

- ▶ For mirroring operations that add or remove a volume copy, the GUI automatically applies the same operation to the change volume, system, which keeps it aligned with the user volume.

A CLI user must use the **addvdiskcopy** and **rmvdiskcopy** commands to manage mirrored copies of the change volume when the user volume is modified. When the **addvdiskcopy** command is used on a change volume, the only optional parameters that can be specified are those relating to the mirroring process itself (for example, `autodelete` and `mirrorwritepriority`). This ensures that the change volume is created according to best practice. All other parameters for the new copy of the change volume are defined by the system.

6.7.8 Synchronization

The system automatically performs a synchronization whenever the volume group transitions from change recording mode to either journaling or cycling mode. Before any synchronization starts, the FlashCopy maps for the change volumes at the recovery system are started to preserve the consistency of the volume group. The recovery point is frozen at the time of the last write before the synchronization during the synchronization. This is required because synchronization does not replay writes in order. Therefore, it does not maintain consistency while it is in progress.

Synchronization uses the bitmap to identify the grains that need to be read from the production volumes and written to the recovery volumes. Once the change volumes are maintaining a snapshot of the recovery volumes, the synchronization process starts with the recovery copy controlling the requests for grains to read from the production copy. Synchronization reads are interleaved with host writes (if in journaling mode) and added to the journal. The production system automatically manages the amount of synchronization

based on the usage of resources available in the journal. This ensures that synchronization activity does not consume too many journal resources, thus avoiding journal purges that are caused by synchronization.

If synchronization encounters a read error, such as a medium error, the volume group stops replicating and an error is logged against the volume group. After the problem that caused the read error is resolved, replication restarts when the error is marked as fixed. If synchronization is interrupted, such as by the partnership disconnecting, it restarts automatically when it is able and the change volume continues to protect the original snapshot. The snapshot for the change volume is discarded only after all volumes in the volume group complete synchronization or the recovery copy of the volume group is deleted.

If the recovery copy of a volume group is made independent during a synchronization, the change volume is used to restore the snapshot onto the host-accessible volumes. Once the FlashCopy maps are reversed, the volumes are made accessible to the host. This triggers a background process to undo any writes to the volumes that were done as part of the partial synchronization.

Journaling mode

Journaling mode provides a high-throughput, low recovery point asynchronous replication mode. Typically, the recovery point is expected to be below one second, but it can vary based on the RTT (round-trip time), the available bandwidth between systems, and the performance of the recovery system.

In this mode, every host write is tagged with a sequence number and added to the journal for replicating. This is similar in characteristics to Remote Copy Global Mirror, but journaling mode has a significantly greater throughput and the ability to extend the RPO to avoid performance problems.

Cycling mode

Cycling mode uses synchronization and change volumes to periodically replicate to the recovery system in a way that requires less bandwidth and consumes fewer journal resources. Periodically, a snapshot is captured (using FlashCopy) of all volumes in the volume group. The snapshot is stored on the change volume at the production system and is maintained by using either Copy-on-Write or Redirect-on-Write, depending on the storage-pool type and volume-capacity savings. The background synchronization process copies only the changes to the remote system. After the synchronization is complete, the snapshot at the production system is discarded and the process might repeat.

Cycling mode results in a higher recovery point than journaling mode, but has the distinct advantage that it can coalesce writes to the same region of the volume, thus reducing the bandwidth required between systems.

Mode switching

Asynchronous replication automatically adapts to the conditions by switching volume groups between journaling and cycling mode. If a volume group experiences too many journal-purges in a short period of time, it starts replicating by using cycling mode. This usually happens if the replication throughput is not high enough to sustain the write throughput for this volume group. Unlike Remote Copy Global Mirror with Change Volumes, a defined cycle period does not exist. Rather, replication will cycle frequently enough to ensure that the RPO that is defined by the replication policy is achieved (in the absence of errors). The system aims to ensure that all volume groups meet their defined RPO.

Conversely, if a volume group is found to be cycling too frequently it starts to replicate by using journaling mode. Journaling mode can be more resource-efficient, so it is the preferable option if the conditions can support it.

The transitions between the modes are transparent and are managed by the system. The mode that a volume group is currently using cannot be easily identified; the important aspect is the current recovery point, which is visible against the volume group. It is this metric that should be monitored. If the volume group exceeds the RPO that is defined on the replication policy, an alert is raised.

6.7.9 Status and recovery point reporting

The status of replication for a volume group is visible in the GUI on the Policies tab of the volume group view, or on the CLI using the `ls volumegroupreplication` command.

The status includes information such as:

- ▶ Which system is currently production and which system is recovery, and therefore which direction replication is operating.
- ▶ Whether the systems are connected or disconnected.
- ▶ Whether replication is running or suspended.
- ▶ Information about the time of the recovery point that exists on the recovery copy, expressed either as a time behind production (if replication is running) or a fixed timestamp (if replication is not running).
- ▶ Whether the recovery point available for the volume group is within the recovery point objective (RPO) that is defined in the replication policy.

The recovery point can be tracked by using the statistics that are produced by the production system. It is available on the GUI, CLI, or REST API on both systems. This value is updated periodically and is rounded up to the nearest second. More granular reporting is only available within the XML statistics that can be retrieved by external monitoring tools.

Note: A running recovery point of zero does not guarantee that the copies are identical as the value is updated periodically.

- ▶ For detailed implementation of the policy-based replication, refer to *Policy-Based Replication with IBM Storage FlashSystem, IBM SAN Volume Controller and IBM Storage Virtualize*, REDP-5704.



Ensuring business continuity

Business continuity and continuous application availability are among the most important requirements for many organizations. Advances in virtualization, storage, and networking made enhanced business continuity possible.

Information technology solutions now can manage planned and unplanned outages, and provide the flexibility and cost efficiencies that are available from cloud-computing models.

This chapter briefly describes the Stretched Cluster, Enhanced Stretched Cluster (ESC), and HyperSwap solutions for IBM Storage Virtualize systems. Technical details or implementation guidelines are not presented in this chapter because they are described in separate publications.

This chapter includes the following topics:

- ▶ “High availability and disaster recovery” on page 502
- ▶ “Business continuity with a Stretched cluster topology” on page 502
- ▶ “Business continuity with HyperSwap” on page 504
- ▶ “Comparing business continuity solutions” on page 510
- ▶ “Quorum site and the IP quorum application” on page 513
- ▶ “HyperSwap internals” on page 515
- ▶ “Other considerations and general recommendations” on page 516

7.1 High availability and disaster recovery

IBM Storage Virtualize 8.6 provides a choice of high availability (HA) and disaster recovery (DR) (HADR) solutions, but before selecting the solution that fulfills your business needs, it is necessary to understand how those solutions work and what is the difference between them:

► DR

Can resume access to data as some point after a specific disaster impacting a component or a whole site:

- Intended to handle problems after a system fails.
- Requires admin interaction or extra automation to perform failover.
- Associated with nonzero recovery point objective (RPO) and recovery time objective (RTO).

► HA

Have continuous access to data during and after a disaster impacting a system's components or a whole site:

- Intended to handle problems while a system is running.
- Does not require any admin interaction.
- RPO and RTO are zero.

Both HA and DR can be used simultaneously. For example, a solution can implement HyperSwap as HA with a replica on a third site as DR.

Note: This book does not cover 3-site replication solutions. For more information, see *IBM Spectrum Virtualize 3-Site Replication*, SG24-8504.

IBM Storage Virtualize systems support three different cluster topologies: standard, Stretched cluster, and HyperSwap. Standard topology is supposed to provide redundancy with 99,9999 for a single Flash System but does not protect from a complete site failure. The other two topologies are designed to have two production sites, and provide HA by keeping data access in a failure at one of the sites. Stretched cluster and HyperSwap are described in this chapter.

7.2 Business continuity with a Stretched cluster topology

IBM SAN Volume Controller (SVC) can be configured as a Stretched cluster and ESC. This topology is not available for enclosure-based systems, such as all members of the IBM FlashSystem family. As well currently the Ethernet clustering without opening the SCORE with IBM for iWARP (TCP protocol) since IBM Storage Virtualize 8.5.2. FlashSystem 50XX do not support iSER clustering.

7.2.1 Stretched cluster (SC)

Within standard implementations of IBM Storage Virtualize, all nodes are physically installed in the same location and configured into redundant cluster. Cluster contains IO groups, depending on the IBM Storage Virtualize solution there could be up to 4 IO groups. Each IO group consists of two nodes for redundancy. And each IO group services IO to specific set of volumes configured for that IO group. To fulfill the different HA needs of customers, the Stretched cluster (or split-cluster) configuration was introduced, in which each node from the same I/O group is physically installed at a different site, and the storage capacities (that are virtualized but SVC) are also installed on each site to provide fail proof domains.

Stretched cluster uses the volume mirroring feature to maintain synchronized independent copies of user data on each site.

When implemented, you can use this configuration to maintain access to data on the system, even if failures occur at different levels, such as the storage area network (SAN), back-end storage, IBM Storage Virtualize node, or data center power provider.

Stretched cluster is considered a HA solution because both sites work as instances of the production environment (no standby location is used). Combined with application and infrastructure layers of redundancy, Stretched clusters can provide enough protection for data that requires availability and resiliency.

When IBM Storage Virtualize was first introduced, the maximum supported distance between nodes within an I/O group was 100 meters (328 feet). With the evolution of code and the introduction of new features, Stretched cluster configurations were enhanced to support distances up to 300 km (186.4 miles). These geographically dispersed solutions use specific configurations that use Fibre Channel (FC) or Fibre Channel over IP (FC/IP) switches, or Multiprotocol Router (MPR) Inter-Switch Links (ISLs) between different locations. While as it was discussed above the clustered configurations were enhanced to support bigger distances, it is very important to take into account that each 100km of distance introduces 1ms delay, thus at maximum distance introduced communication delay might raise three times.

A Stretched cluster solution can still be configured and used in small setups where both production sites are in the same data center and it is recommended to take into account the account different fire protection zones. However, for most use cases, ESC is a preferred option.

Note: In this case, the Stretched cluster is organized by the SAN fabric configuration.

7.2.2 Enhanced Stretched cluster (ESC)

IBM Storage Virtualize 7.2 introduced the ESC feature, which further improved and enhanced the standard Stretched cluster configurations. ESC introduced the *site awareness* concept for nodes and external storages that are virtualized behind SVC, as well as communication bandwidth optimization.

With IBM Storage Virtualize 7.5, site awareness was extended to hosts. This extension enables more efficient distribution of host I/O traffic through the SAN, and easier host path management.

Stretched cluster and ESC solutions can be combined with DR features, such as Metro Mirror (MM) or Global Mirror (GM), which make it possible to keep three independent data copies, and enable you to effectively manage rolling disaster scenarios.

In an ESC configuration, each site is defined as an independent failure domain. If one site experiences a failure, the other site can continue to operate without disruption. Sites can be in the same room, across rooms in the data center, in different buildings at the same campus, or in different cities. Different types of sites protect against different types of failures.

In addition to two sites storing copies of data, in SC and ESC environments you must configure a third site, which must be independent of data sites. The third site hosts a quorum device that provides an automatic tie-breaker in case of communication failure between the two main sites (for more information, see 7.4, “Comparing business continuity solutions” on page 510) or any failure that prevents nodes on both sites from communication with each other (for example one site is totally down).

Communication loss condition, while nodes on both sites are still active but inter-site communication is not possible, is also known as *split-brain scenario*. Therefore, it is necessary to decide which site should continue to service IO and which site should go to standby in order to keep data consistent. This is one of the roles of the quorum site to help with this decision.

Two sites within a single location

If each site has different power source or supply within a single location or data center, the system can survive the failure of any single power domain. For example, one node can be placed in one rack installation and the other node can be in another rack. Each rack is considered a separate site with its own power phase. In this case, if power was lost to one of the racks, the partner node in the other rack can be configured to process requests and effectively provide availability to data, even when the other node is offline because of a power disruption.

Two sites at separate locations

If each site is a different physical location, the system can survive the failure of any single location. These sites can span shorter distances (for example, two sites in the same city), or they can be spread farther geographically, such as two sites in separate cities. If one site experiences a site-wide disaster, the other site can remain available to process requests.

If configured correctly, the system continues to operate after the loss of one site. The key prerequisite is that each site contains only one node from each I/O group. However, placing one node from each I/O group in different sites for a stretched system configuration does *not* provide HA. You must also configure the suitable mirroring technology and ensure that all configuration requirements for those technologies are correctly configured.

Best practices:

- ▶ As a best practice, configure an ESC system to include at least two I/O groups (four nodes).
- ▶ Communication between nodes, so called inter-site communication between nodes should be separated from other IO, therefore dedicated private redundant fabric should be configured.
- ▶ The inter-site communication fabrics between sites have to be totally independent from each other and should not have any common elements that can influence the fabrics communication capabilities. It does not matter how many physical links there are between sites (ISLs) as inter-site communication between nodes depends on the fabric end-to-end communication. And if one fabric is compromised even having single physical link down (and for example frames were lost), cluster considers fabric non reliable and would switch to another fabric. If both fabrics are compromised, that can happen. If, for example, one has two ISL providers and one provider has problems and providers are shared between fabrics and no single provider is dedicated to specific fabric, cluster considers communication loss between two sites (as frames are lost on both fabrics, even for short period) and one of the sites is set to standby.

7.3 Business continuity with HyperSwap

The *HyperSwap* HA feature can be used with SVC and IBM FlashSystem family products. The feature enables business continuity during a hardware failure, power outage, connectivity problem, or other disasters, such as fire or flooding.

It provides HA volumes that are accessible through two sites that are up to 300 kilometers (km) apart. A fully independent copy of the data is maintained at each site. When data is written by hosts at either site, both copies are synchronously updated before the write operation completes. HyperSwap automatically optimizes itself to minimize data that is transmitted between sites, and to minimize host read/write latency. For more information about the optimization algorithm, see 7.6, “HyperSwap internals” on page 515.

Note: For more technical information about HyperSwap, see [IBM HyperSwap: An automated disaster recovery solution](#).

7.3.1 HyperSwap overview

HyperSwap includes the following key features:

- ▶ Works with all IBM Storage Virtualize products except for IBM FlashSystem 5010 and 5015.
- ▶ Uses the intra-cluster synchronous non-breakable remote copy (active-active MM) capability, with change volumes (CVs) and access I/O group technologies.
- ▶ It makes a host’s volumes accessible across two IBM Storage Virtualize I/O groups in a clustered system by using the active-active MM relationship. The volumes are presented as a single logical unit number (LUN) to the host.
- ▶ Works with the standard multipathing drivers that are available on various host types. Extra host support is not required to access the HA volumes.

While SC or ESC actually stretched the IO groups between two sites, and it was possible because SVC has standalone nodes that can be organized into IO groups and physically moved around, IBM FlashSystem storages contains of enclosures with storage capabilities and each enclosure already has two nodes on board that cannot be split physically and that are already organized into the IO group. So they cannot be physically moved and the IO groups cannot be stretched. To resolve this condition and provide an HA solution the IBM Storage Virtualize HyperSwap configuration was introduced and this solution requires that at least one control enclosure is implemented in each location.

Therefore, a minimum of two control enclosures (two IO groups) cluster is needed to implement HyperSwap on IBM Flash System storages. Configurations with three or four control enclosures are also supported for the HyperSwap, though it is recommended to have the same number of controllers on both sites.

In addition to the active-active MM feature that is used to mirror (replicate) data between sites, the HyperSwap feature also introduced the *site awareness* concept for node canisters, internal and external storage, and hosts.

The typical simplified scheme of IBM FlashSystem HyperSwap implementation is shown in Figure 7-1.

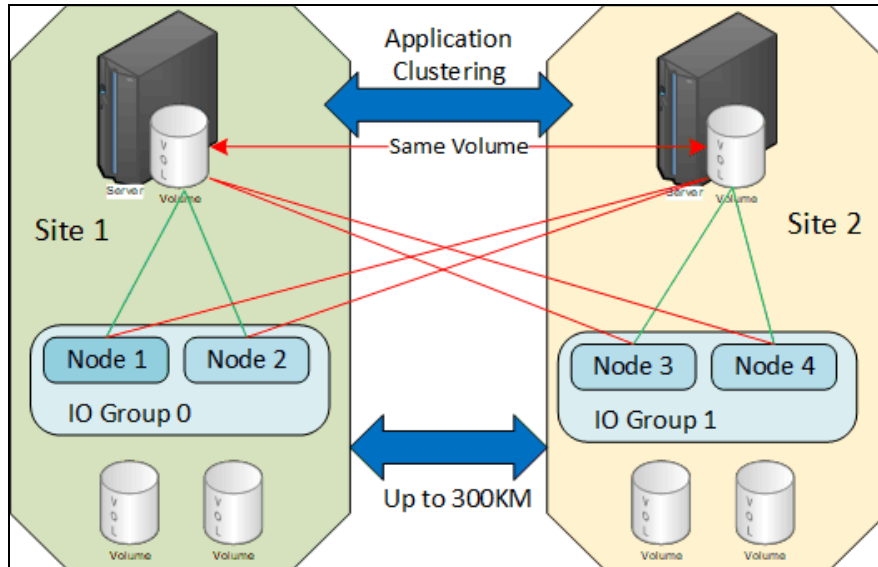


Figure 7-1 Typical concept scheme HyperSwap configuration with IBM Storage Virtualize

With a copy of the data that is stored at each location, HyperSwap configurations can handle different failure scenarios.

The Small Computer System Interface (SCSI) protocol allows storage devices to indicate the preferred ports for hosts to use when they submit I/O requests. By using the Asymmetric Logical Unit Access (ALUA) state for a volume, a storage controller can inform the host about what paths are active and which ones are preferred. In a HyperSwap system topology, the system advertises the host paths to “local” nodes (nodes on the same site as the host) as *Active Optimized*. The path to remote nodes (nodes on a different site) is advertised as *Active Unoptimized*. If after a failure there are no *Optimized* paths for a host, it starts by using *Unoptimized*.

Figure 7-2 shows how HyperSwap operates in an I/O group failure at one of the sites. The host on Site 1 detects that there are no paths to local nodes, and starts using paths to LUN through Site 2.

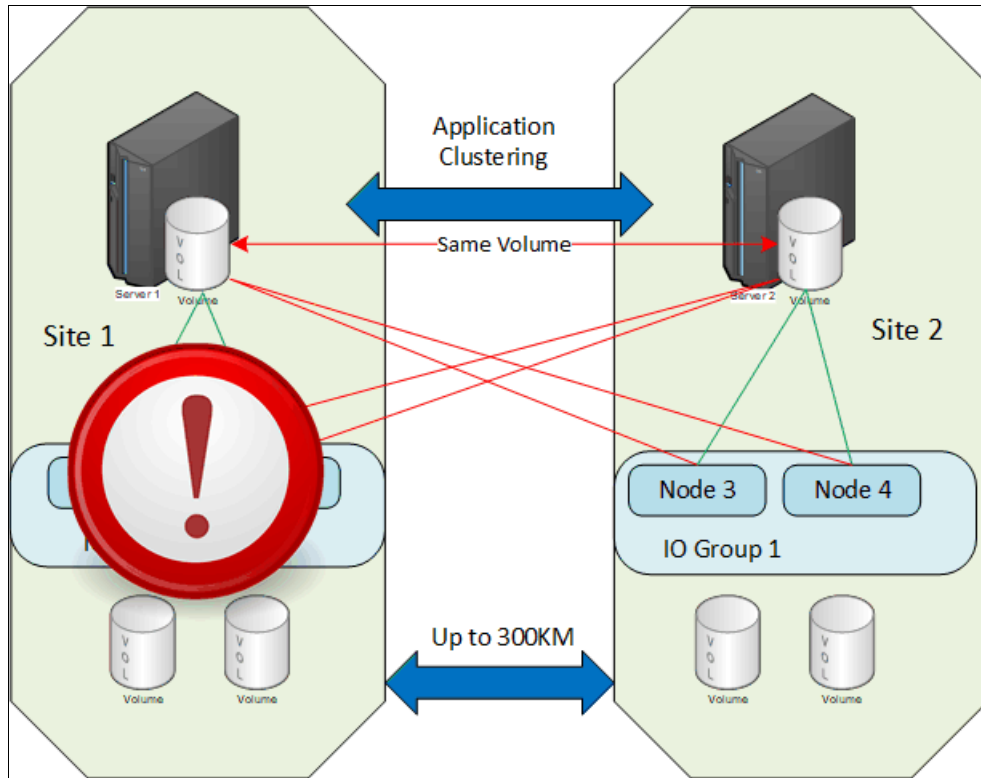


Figure 7-2 IBM Storage Virtualize HyperSwap in a storage failure scenario

Even if the access to an entire location (site) is lost, as shown in Figure 7-3 on page 508, access to the disks remains available at the alternative location. This behavior requires clustering software at the application and server layer to fail over to a server at the alternative location and resume access to the disks.

The active-active synchronous mirroring feature, keeps both copies of the storage in synchronization. Therefore, the loss of one location causes no disruption to the alternative location.

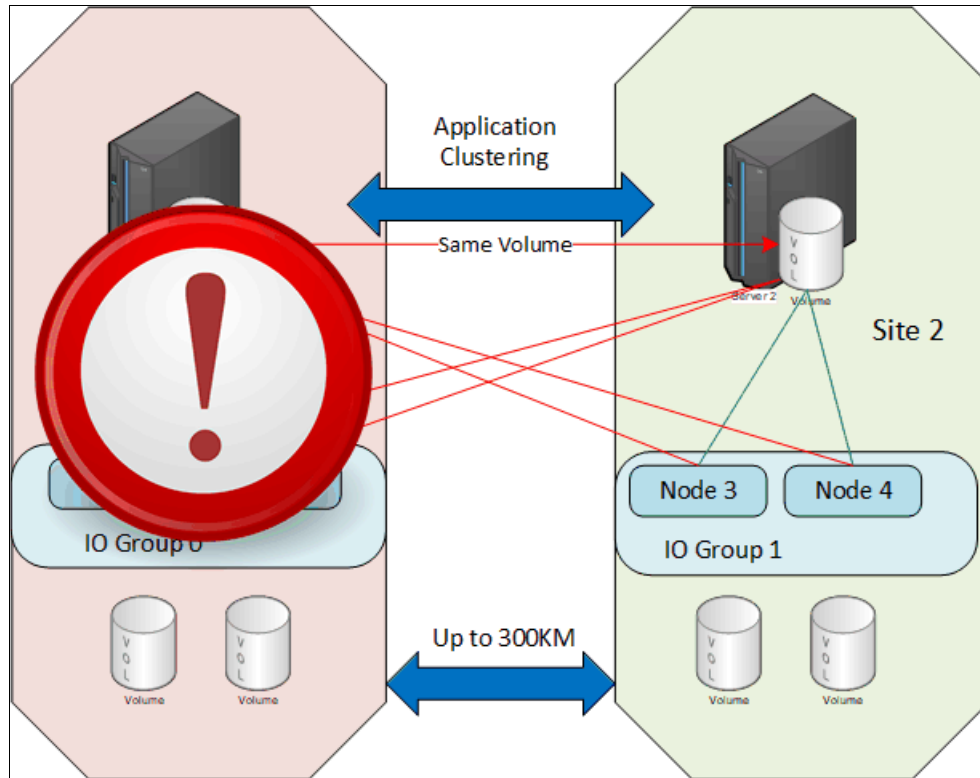


Figure 7-3 IBM FlashSystem HyperSwap in a site failure scenario

The HyperSwap system depends on a quality and stability of an inter-site link. A best practice requires isolation of the traffic between IBM Storage Virtualize cluster nodes on a dedicated private SAN from other types of traffic traversing through the inter-site link, such as host and back-end controller traffic. This task can be performed by using dedicated hardware, with a Cisco virtual storage area network (VSAN), or Brocade Virtual Fabric when using the same SAN switches.

Note: SAN24B-6 can now also be used with virtual fabric using F.O.S. 9.1.1 or higher.

The isolation is necessary because all host writes to a HyperSwap volume are mirrored to a remote site by using an internode link. If these writes are delayed, then host performance deteriorates because its writes are acknowledged only after they complete at both sites.

Note: For more information about traffic isolation and a recommended fabric configuration, see *IBM Spectrum Virtualize HyperSwap SAN Implementation and Design Best Practices*, REDP-5597.

7.3.2 HyperSwap build steps

A HyperSwap solution consists of a single system (cluster) that contains two or more I/O groups (control enclosures). To configure HyperSwap, complete the following steps:

1. If all control enclosures have no system configuration (for example, if they are new):
 - a. Perform hardware installation, cabling, and SAN zoning. On Ethernet it is necessary to add IP addresses to the other IO group. It is important to make sure that SAN zoning and Ethernet communication (if Ethernet clustering is used) are configured correctly.
 - b. Initialize the first node of the first control enclosure by selecting **As the first node in a new system**, as shown in Figure 7-4.

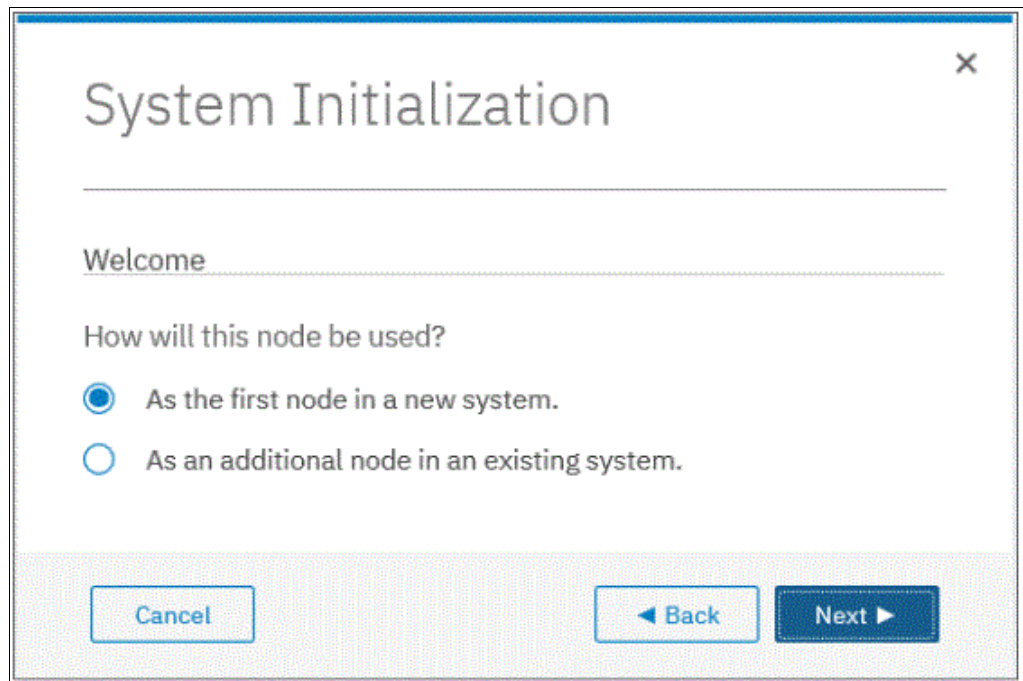


Figure 7-4 Initializing the first node of a HyperSwap system

- c. After the initialization wizard completes and a cluster on a first control enclosure is set up, initialize the remaining control enclosures by selecting **As an additional node in an existing system**.
 - d. Complete the HyperSwap configuration by following the system setup wizard.
 - e. Change the system topology to HyperSwap by using the GUI topology wizard.
2. If control enclosures already are configured and running (for example, you want to join two separate IBM FlashSystem clusters into one HyperSwap system):
 - a. Select one of the control enclosures that will retain its configuration and keep the data.
 - b. Migrate all the data away from the other control enclosures (systems) that need to become a part of HyperSwap configuration.
 - c. After data is migrated, delete the system configuration and the data from those systems to reset them to the “factory” state. For more information, see [Removing a control enclosure from a system](#).
 - d. Adjust the zoning to the HyperSwap configuration that you want so that all control enclosures can see each other over the FC SAN.

- e. Use the GUI of the first control enclosure (the one that was identified in step a on page 509) to add control enclosures to the cluster.
- f. Use the GUI topology wizard to change the topology to HyperSwap.

Note: For more information about system initialization, adding control enclosures, and changing the topology, see *Implementation Guide for IBM Storage FlashSystem and IBM SAN Volume Controller: Updated for IBM Storage Virtualize Version 8.6*, SG24-8542.

7.4 Comparing business continuity solutions

The business continuity solutions that are described in this chapter feature different characteristics in terms of implementation and features. Table 7-1 provides a comparison of these solutions that can help you to identify the most fitting solution for a specific environment and needs.

Table 7-1 Business continuity solutions comparison

Solution feature	Standard Stretched Cluster	Enhanced Stretched Cluster	HyperSwap
The function is available on these products.	SVC only.	SVC only.	All IBM Storage Virtualize based products that support two or more I/O groups.
Complexity of the configuration.	Command-line interface (CLI) or GUI on a single system, and simple object creation.	CLI or GUI on a single system, and simple object creation.	CLI or GUI on a single system, and simple object creation.
The number of sites on which data is stored.	Two.	Two.	Two.
Distance between sites.	Up to 300 km (186.4 miles).	Up to 300 km (186.4 miles).	Up to 300 km (186.4 miles).
Maintains independent copies of data.	Two.	Two.	Two (four if you use more volume mirroring to two pools in each site).
Technology for host to access multiple copies and automatically fail over.	Standard host multipathing driver.	Standard host multipathing driver.	Standard host multipathing driver.
Cache is retained if only one site is online?	Yes, if a spare node is used. Otherwise, no.	Yes, if a spare node is used. Otherwise, no.	Yes.
Host-to-storage-system path optimization.	Manual configuration of the preferred node.	Automatic configuration that is based on the host site settings. Uses Asymmetric Logical Unit Access (ALUA) or ANA (NVME) and Target Port Group Support (TPGS).	Automatic configuration that is based on the host site settings. Uses ALUA or ANA (NVME) and TPGS.
Synchronization and resynchronization of copies.	Automatic.	Automatic.	Automatic.

Solution feature	Standard Stretched Cluster	Enhanced Stretched Cluster	HyperSwap
Stale consistent data is retained during the resynchronization for DR?	No.	No.	Yes.
Scope of failure and resynchronization.	Single volume.	Single volume.	One or more volumes. The scope is user-configurable.
Ability to use FlashCopy with an HA solution.	Yes (there is no awareness of the site locality of the data).	Yes (there is no awareness of the site locality of the data).	Limited. You can use FlashCopy maps with a HyperSwap volume as a source to avoid sending data across the link between sites.
Ability to use MM, GM, or Global Mirror with Change Volumes (GMCV) with an HA solution.	One remote copy. You can maintain current copies on up to four sites.	One remote copy. You can maintain current copies on up to four sites.	Support for 3-site solutions is available with IBM Storage Virtualize 8.4 or later.
Maximum number of HA volumes.	5,000.	5,000.	1,250 in the IBM FlashSystem 5000/5100 products, or any other product running code level 8.3.1 or earlier. 2,000 in other IBM Storage Virtualize products with code level 8.4 or later.
Minimum required paths for each logical unit (LUN) for each host port.	Two.	Two.	Four.
Minimum number of I/O groups.	One.	One I/O group is supported, but it is a best practice to have two or more I/O groups.	Two.
Rolling disaster support.	No.	Yes.	Yes.
Licensing.	Included in base product.	Included in base product.	Requires a Remote Mirroring license for volumes. The exact license requirements might vary by product but it is included in the most of the actual FlashSystems.

Solution feature	Standard Stretched Cluster	Enhanced Stretched Cluster	HyperSwap
Capacity limits.	<p>Total mirrored volume capacity per I/O group 1PiB Depending on the hardware model involved in configuration this limit can be higher. In case of SV3 model bitmap capacity increased to 20GiB shared pool between Volume Group Snapshots and Volume Mirroring. While, that giving 40 PiB theoretical capacity for VGM without VGS/SGC, as a best practice do not plan to use more than 20PiB on SV3 for VDM</p>	<p>Depending on the hardware involved in to configuration Total mirrored volume capacity per I/O group is 1 PiB. In case of SV3 model bitmap capacity increased to 20GiB shared pool between Volume Group Snapshots and Volume Mirroring. While that giving 40 PiB theoretical capacity for VGM without VGS/SGC as a best practice do use more than 20PiB on SV3 for VDM</p>	<p>1 PB or 2 PB limit depending on the code level and system type (IBM Storage (Spectrum) Virtualize 8.4.2 added the 2 PiB limit for 7200 upwards and SVC SA2/SV2 upwards.)</p>

7.5 Quorum site and the IP quorum application

With the Stretched cluster or HyperSwap configurations, you must use a third, independent site to house a quorum device to act as the tie-breaker in split-brain scenarios. The quorum device can also hold a backup copy of the cluster metadata to be used in certain situations that might require a full cluster recovery.

7.5.1 IP quorum overview

To use a quorum disk as the quorum device, the third site must have FC connectivity between an external storage system and the IBM Storage Virtualize cluster. Sometimes, this third-site quorum disk requirement turns out to be expensive in terms of infrastructure and network costs. For this reason, a less demanding solution that is based on a Java application, which is known as the IP quorum application, was introduced with the 7.6 release.

Initially, IP quorum was used only as a tie-breaker solution. However, with the 8.2.1 release, it was expanded to store cluster configuration metadata, fully serving as an alternative for quorum disk devices.

Note: IP quorum with metadata demands higher link bandwidth between system's nodes and the IP quorum host than IP quorum without metadata support. Also, enabling metadata storage requires lower link latency.

Consider deploying IP quorum with metadata storage only if there are no other ways to store metadata (for example, all the system's back-end storage is internet Small Computer Systems Interface (iSCSI)), or if you have an ensured high-quality network link between nodes and the IP quorum host.

FC connectivity is not needed, to use an IP quorum application as the quorum device for the third site (up to 5 IP quorums can be deployed, but only one quorum can be set to be active). An IP quorum application can be run on any host at the third site, as shown in Figure 7-5.

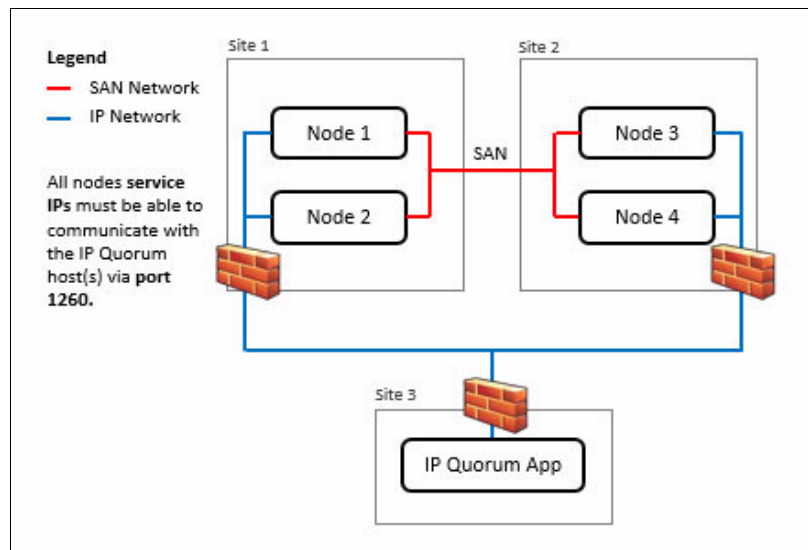


Figure 7-5 IP quorum network layout

However, the following strict requirements must be met on the IP network when an IP quorum application is used:

- ▶ Connectivity from the servers that are running an IP quorum application to the service IP addresses of all nodes or node canisters. The network also must handle the possible security implications of exposing the service IP addresses because this connectivity also can be used to access the service assistant interface if the IP network security is configured incorrectly.
- ▶ On each server that runs an IP quorum application, ensure that only authorized users can access the directory that contains the IP quorum application. Metadata is stored in the directory in a readable format, so ensure access to the IP quorum application and the metadata is restricted to only authorized users.
- ▶ The gateway should not be susceptible to failure if one site goes down.
- ▶ Port 1260 is used by the IP quorum application to communicate from the hosts to all nodes or enclosures.
- ▶ The maximum round-trip delay must not exceed 80 milliseconds (ms), which means 40 ms each direction.
- ▶ If you are configuring the IP quorum application without a quorum disk for metadata, a minimum bandwidth of 2 megabytes per second (MBps) is ensured for traffic between the system and the quorum application. If your system is using an IP quorum application with quorum disk for metadata, a minimum bandwidth of 64 MBps is ensured for traffic between the system and the quorum application.
- ▶ Ensure that the directory that stores an IP quorum application with metadata contains at least 250 MB of available capacity.

Quorum devices are also required at Site 1 and Site 2, and can be either disk-based quorum devices or IP quorum applications. A maximum number of five IP quorum applications can be deployed.

Important: *Do not* host the quorum disk devices or IP quorum applications on storage that is provided by the system it is protecting, because during a tie-break situation, this storage pauses I/O.

For more information about IP quorum requirements and installation, including supported operating systems and Java runtime environments (JREs), see [Configuring quorum](#).

For more information about quorum disk devices, see 3.5, “Quorum disks” on page 239.

Note: The IP quorum configuration process is integrated into the IBM Storage Virtualize GUI and can be found by selecting **Settings** → **Systems** → **IP Quorum**.

7.5.2 Quorum modes

Quorum mode is a configuration option that was added to the IP quorum function with IBM Storage Virtualize 8.4. By default, the IP quorum mode is set to **Standard**. In HyperSwap clusters, this mode can be changed to **Preferred** or **Winner**.

With this configuration, you can specify which site will resume I/O after a disruption based on the applications that run on each site or other factors. For example, you can specify whether a selected site is the preferred for resuming I/O, or if the site automatically “wins” in tie-breaker scenarios.

Preferred mode

If only one site runs critical applications, you can configure this site as *preferred*. During a split-brain situation, the system delays processing tie-breaker operations on other sites that are not specified as “preferred”. The designated preferred site has a timed advantage when a split-brain situation is detected, and starts racing for the quorum device a few seconds before the nonpreferred sites.

Therefore, the likelihood of reaching the quorum device first is higher. If the preferred site is damaged or cannot reach the quorum device, the other sites have the chance to win the tie-breaker and continue I/O.

Winner mode

This configuration is recommended for use when a third site is not available for a quorum device to be installed. In this case, when a split-brain situation is detected, the site that is configured as the winner is always the one to continue processing I/O regardless of the failure and its condition. The nodes at the nonwinner site always lose the tie-breaker and stop processing I/O requests until the fault is fixed.

7.5.3 IP quorum as a service

IP quorum can be implemented as a Linux service, and it also can be implemented by using Ansible. For more information, see the following publications, which are not directly managed by IBM:

- ▶ [How to configure a systemd service for the IBM Storage Virtualize IP quorum app](#)
- ▶ [Automated IBM Storage Virtualize IP-Quorum installation with Ansible](#)

7.6 HyperSwap internals

Each *HyperSwap volume* is internally represented as four *VDisks* and one *remote copy relationship*. It consists of a master volume and a master CV (Change Volume) in one system site, and an auxiliary volume and auxiliary CV (Change Volume) in the other system site. An active-active synchronous mirroring relationship exists between the two sites. Similar to a regular MM relationship, the active-active relationship keeps the master volume and auxiliary volume synchronized but it is non-breakable and the directions cannot be manually controlled.

The relationship employs the CVs as journals throughout any resynchronization procedure. Both the master CV and its counterpart, the auxiliary CV, should ideally reside within the same I/O Group as their respective master and auxiliary volumes, respectively. This guideline also pertains to Volume Groups, where applicable, requiring available space within them.

For more information, see 5.5, “HyperSwap volumes” on page 337.

The HyperSwap volume always uses the unique identifier (UID) of the master volume. The HyperSwap volume is assigned to the host by mapping only the master volume, even though access to the auxiliary volume is ensured by the HyperSwap function.

Figure 7-6 shows how a HyperSwap volume handles the UID relationship. Note that both volumes in this active-active relationship is presented under one UID on the system.

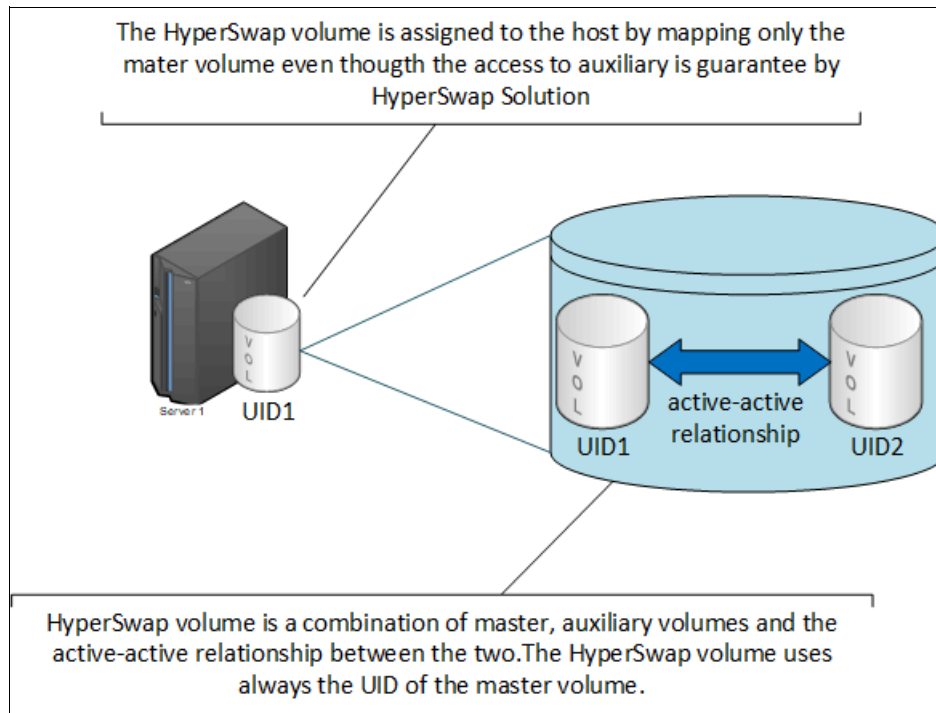


Figure 7-6 HyperSwap volume UID

In HyperSwap, host write operations can be submitted by hosts at both sites, but they are always routed to the volume, which is the primary copy of the HyperSwap relationship. Then, data is mirrored to the secondary copy over the inter-site link.

HyperSwap can automatically switch replication direction between sites. If a sustained write workload (that is, more than 75% of write I/O operations for at least 20 minutes) is submitted to a site with the secondary volume, the HyperSwap function switches the direction of the active-active relationships, swapping the secondary volume to primary, and vice versa.

Replication direction can be switched on any HyperSwap volume independently, unless the volumes are added to a single consistency group (CG). You can have the primary on Site 1 for one HyperSwap volume, and the primary on Site 2 for another HyperSwap volume at the same time.

Host read operations that are submitted on any site always were directed to a primary copy before IBM Storage Virtualize 8.3.1. Starting with 8.3.1, reads always are a processed local copy of the volume.

7.7 Other considerations and general recommendations

A business continuity solution implementation requires special consideration about the infrastructure and network setup. In HyperSwap and Stretched cluster or ESC topologies, the communication between the IBM Storage Virtualize controllers must be optimal and free of errors for best performance because the internode messaging is done across the sites. Have a dedicated private SAN for internode communication so that the communication is not impacted by regular SAN activities.

The HyperSwap and Stretched cluster or ESC features require implementing the storage network to ensure that the inter-node communication on the FC ports on the control enclosures (nodes) between the sites is on dedicated fabrics. No other traffic (hosts or back-end controllers) or traffic that is unrelated to the distributed cluster can be allowed on this fabric. Two fabrics are used: one private for the inter-node communication, and one public for all other data.

A few SAN designs are available that can achieve this separation and some incorrect SAN designs can result in some potential problems that can occur with incorrect SAN design and implementation.

One other important consideration is to review the site attribute of all the components to make sure that they are accurate. With the site awareness algorithm that is present in the IBM Storage Virtualize code, optimizations are done to reduce the cross-site workload. If this attribute is missing or not accurate, there might be unnecessary increased cross-site traffic, which might lead to higher response time to the applications.

For more information about design options and some common problems, see the following resources:

- ▶ *SAN and Fabric Resiliency Best Practices for IBM b-type Products*, REDP-4722
- ▶ [Designing a Resilient SAN for IBM HyperSwap SVC and IBM Storage Virtualize](#)

For step-by-step configuration instructions, see [HyperSwap system configuration details](#).



Hosts

This chapter provides general guidelines and best practices for configuring host systems for IBM Storage Virtualize based storage systems.

Before attaching a new host, confirm that the host is supported by IBM Storage Virtualize. For more information about a detailed compatibility matrix, see [IBM System Storage Interoperation Center \(SSIC\)](#).

The host configuration guidelines apply equally to all IBM Storage Virtualize systems. Therefore, the product name often is referred to as an *IBM Storage FlashSystem / SAN Volume Controller*.

For more information about host attachment, see the [Host attachment](#).

For more information about hosts that are connected by using Fibre Channel (FC), see Chapter 2, “Storage area network guidelines” on page 123. Host connectivity is a key consideration in overall storage area network (SAN) design.

This chapter includes the following topics:

- ▶ “Connection protocols” on page 521
- ▶ “SCSI FC general configuration guidelines” on page 521
- ▶ “SAS host attachment” on page 524
- ▶ “iSCSI host attachment considerations” on page 525
- ▶ “iSCSI Extensions for Remote Direct Memory Access host attachment considerations” on page 527
- ▶ “NVMe over Fibre Channel host attachment considerations” on page 528
- ▶ “NVMe over Ethernet host attachment considerations” on page 528
- ▶ “Portsets” on page 530
- ▶ “Host pathing” on page 533
- ▶ “I/O queues” on page 534
- ▶ “Host clusters” on page 535
- ▶ “AIX hosts” on page 538
- ▶ “Virtual I/O Server hosts” on page 538
- ▶ “Microsoft Windows hosts” on page 540
- ▶ “Linux hosts” on page 540
- ▶ “Oracle Solaris hosts support” on page 541
- ▶ “HP 9000 and HP Integrity hosts” on page 543

- ▶ “VMware ESXi server hosts” on page 544
- ▶ “Container Storage Interface Block Driver” on page 545
- ▶ “100-Gigabit Ethernet host connectivity” on page 546

8.1 Connection protocols

There are two main protocols types for accessing volumes on IBM Storage Virtualize:

- ▶ **SCSI** - standard protocol since 1980s, built for conventional hard disk drives
- ▶ **NVME** - newer protocol with fewer commands and better performance

A volume in IBM Storage Virtualize can only be accessed using a single protocol, either SCSI or NVMe. This is because the storage controller in the system can only present a volume to a host in one way.

However, a volume can be accessed using different transport protocols (iSCSI, FC, and so on) from different hosts. For example, a volume could be accessed using iSCSI from one host and FC from another host. This is because the transport protocol is the way that the host communicates with the storage system, and the storage system can present a volume to different hosts using different transport protocols.

IBM Storage Virtualize does not support the same host accessing a volume using different transport protocols via the same host bus adapter. This is because the host bus adapter can only communicate with the storage system using one transport protocol at a time.

Here is an example of how this would work:

- ▶ Host A could access a volume using iSCSI.
- ▶ Host B could access the same volume using FC.
- ▶ Host C could not access the same volume using both iSCSI and FC.

All hosts and cluster connections need to use the same protocol type for a volume. Different hosts can have different protocols working with the same storage port. So iSCSI, iWarp, iSER, NVMe/RDMA and NVMe/TCP can coexist on the same port with different host mapping types. This includes the ports that are used for replication.

Note: At the time of writing, systems with ROCE adapters cannot have MTU greater than 1500 on 8.6.0.0 or later. The workaround is to reduce MTU to 1500.

8.2 SCSI FC general configuration guidelines

In this section, we describe some general configuration guidelines. The information that is presented here complements the content in Chapter 2, “Storage area network guidelines” on page 123.

8.2.1 Number of paths

As a best practice, do not use more than four FC paths per volume. For HyperSwap and stretched cluster configurations, eight paths per volume are recommended. Adding paths does not significantly increase redundancy, and tends to bog down the host with path management. Also, too many paths might increase the failover time.

8.2.2 Host ports

Each host uses two ports from two different host bus adapters (HBAs). These ports should go to separate SAN fabrics and be zoned to one target port of each node or node canister. When the volumes are created, they are assigned to an I/O group, and the resulting path count between the volume and the host should be four.

Best practice: Keep FC tape (including virtual tape libraries) and FC disks on separate HBAs. These devices have two different data patterns when operating in their optimum mode. Switching between them can cause unwanted processor usage and performance slowdown for the applications.

8.2.3 N_Port ID Virtualization

IBM Storage Virtualize uses N_Port ID Virtualization (NPIV) by default. It reduces failover time and allows for features such as Hot Spare Nodes (HSNs).

For more information about the IBM Storage Virtualize 8.6 NPIV configuration and details, see 7.5, “N_Port ID Virtualization support”, in *Implementation Guide for IBM Storage FlashSystem and IBM SAN Volume Controller: Updated for IBM Storage Virtualize Version 8.6*, SG24-8542.

For more information about configuring NPIV, see Chapter 2, “Storage area network guidelines” on page 123.

8.2.4 Host to I/O group mapping

An *I/O group* consists of two nodes or node canisters that share the management of volumes within the cluster. Use a single I/O group (iogrp) for all volumes that are allocated to a specific host. This guideline results in the following benefits:

- ▶ Minimizes port fan-outs within the SAN fabric.
- ▶ Maximizes the potential host attachments to IBM Storage Virtualize because the maximums are based on I/O groups.
- ▶ Reduces the number of target ports that must be managed within the host.

8.2.5 Volume size versus quantity

In general, host resources, such as memory and processing time, are used up by each storage logical unit number (LUN) that is mapped to the host. For each extra path, more memory is used, and a portion of more processing time is also required.

Each LUN has their own IO queues, so especially in SCSI environments where you have only one queue per LUN, you can get more throughput with LUNs working in parallel.

Best practice is minimum number on LUNs on the IO group should be equal or higher than the number of cores in the IO group.

If you use a high number of volumes and build volume groups or consistency groups for Snapshots or FlashCopy, it makes sense to have not too many volumes in there, definitely less than 128. All volumes in a volume groups or consistency groups will be stopped at the same time.

The user can control this effect by using fewer larger LUNs rather than many small LUNs. However, you might need to tune queue depths and I/O buffers to support controlling the memory and processing time efficiently.

For more information about queue depth, see, 8.10.1, “Queue depths” on page 534.

Note: Larger volume sizes can also help to reduce the number of remote copy relationships in an RC consistency group (CG), which can lead to a performance benefit for Global Mirror with Change Volumes (GMCV) in large environments. For more information about the remote copy configuration limits, see Table 6-8 on page 430.

8.2.6 Host volume mapping

Host mapping is the process of controlling which hosts can access specific volumes within the system. IBM Storage Virtualize always presents a specific volume with the same Small Computer System Interface (SCSI) ID on all ports of the I/O group. When a volume is mapped, IBM Storage Virtualize software automatically assigns the next available SCSI ID if none is specified. In addition, each volume has a unique identifier (UID).

You can allocate the operating system volume of the SAN boot as the lowest SCSI ID (zero for most hosts), and then allocate the various data disks. If you share a volume among multiple hosts, consider controlling the SCSI ID so that the IDs are identical across the hosts. This consistency ensures ease of management at the host level and prevents potential issues during IBM Storage Virtualize updates and even node restarts, mostly for VMware ESX operating systems.

If you are using image mode to migrate a host to IBM Storage Virtualize, allocate the volumes in the same order that they were originally assigned on the host from the back-end storage.

The `lshostvdiskmap` command displays a list of virtual disks (VDisks) (volumes) that are mapped to a host. These volumes are recognized by the specified host.

Example 8-1 shows the syntax of the `lshostvdiskmap` command that is used to determine the SCSI ID and the UID of volumes.

Example 8-1 The `lshostvdiskmap` command

```
svcinfo lshostvdiskmap -delim : 0 (host id)
id:name:SCSI_id:host_id:host_name:vdisk_UID:IO_group_id:IO_group_name:mapping_type
:host_cluster_id:host_cluster_name:protocol
0:ESXVOL_01:0:0:HG-ESX6:600507681280000C70000000000004B2:0:io_grp0:private:::scsi
```

Example 8-2 shows the results of using the `lshostvdiskmap` command.

Example 8-2 Output of using the `lshostvdiskmap` command

```
svcinfo lshostvdiskmap -delim : HG-ESX6 (host name)
id:name:SCSI_id:vdisk_id:vdisk_name:vdisk_UID:IO_group_id:IO_group_name:mapping_type:host_cluste
r_id:host_cluster_name:protocol
3:HG-ESX6:0:5:DB_Volume:60050768108104A2F000000000000037:0:io_grp0:private:::scsi
3:HG-ESX6:1:15:Infra_Volume:60050768108104A2F000000000000041:0:io_grp0:private:::scsi
3:HG-ESX6:2:43:onprem_volume_Ansible:60050768108104A2F000000000000081:0:io_grp0:private:::scsi
3:HG-ESX6:3:14:Volume IP Replication:60050768108104A2F000000000000040:0:io_grp0:private:::scsi
3:HG-ESX6:4:48:ansible:60050768108104A2F000000000000086:0:io_grp0:private:::scsi
3:HG-ESX6:5:49:ansible2:60050768108104A2F000000000000087:0:io_grp0:private:::scsi
3:HG-ESX6:6:34:Onprem_Demo_Ansible_Vol:60050768108104A2F00000000000009F:0:io_grp0:private:::scsi
```

```
3:HG-ESX6:7:50:vol_HG-ESX6_1:60050768108104A2F000000000000A5:0:io_grp0:private::scsi
3:HG-ESX6:8:51:vol_HG-ESX6_10:60050768108104A2F000000000000A8:0:io_grp0:private::scsi
```

Example 8-3 shows the results of using the `lsvdiskhostmap` command.

Example 8-3 Output of using the lsvdiskhostmap command

```
svcinfo lsvdiskhostmap -delim : EEXCLS_HBin01
id:name:SCSI_id:host_id:host_name:vdisk_UID:IO_group_id:IO_group_name:mapping_type:host_cluster_
id:host_cluster_name:protocol
950:EEXCLS_HBin01:14:109:HDMCENTEX1N1:600507680191011D480000000000466:0:io_grp0:private::scsi
950:EEXCLS_HBin01:13:110:HDMCENTEX1N2:600507680191011D480000000000466:0:io_grp0:private::scsi
950:EEXCLS_HBin01:14:111:HDMCENTEX1N3:600507680191011D480000000000466:0:io_grp0:private::scsi
950:EEXCLS_HBin01:14:112:HDMCENTEX1N4:600507680191011D480000000000466:0:io_grp0:private::scsi
950:EEXCLS_HBin01:14:113:HDMCENTEX1N5:600507680191011D480000000000466:0:io_grp0:private::scsi
```

Note: Example 8-3 shows the same volume that is mapped to five different hosts, but host 110 features a different SCSI ID than the other four hosts. *This example is not a recommended practice that can lead to loss of access in some situations because of SCSI ID mismatch.*

8.2.7 Server adapter layout

If your host system includes multiple internal I/O buses, place the two adapters that are used for IBM Storage Virtualize cluster access on two different I/O buses to maximize the availability and performance. When purchasing a server, always have two cards instead of one. For example, two dual-port HBA cards are preferred over one quad-port HBA card because you can spread the I/O and add redundancy.

8.2.8 Host status improvements

IBM Storage Virtualize provides an alternative for reporting host status.

Previously, a host was marked as *degraded* if one of the host ports logged off the fabric. However, examples exist in which this marking might be normal and may cause confusion.

At the host level, a new `status_policy` setting is available that includes the following settings:

- ▶ The `complete` setting uses the original host status definitions.
- ▶ By using the `redundant` setting, a host is not reported as degraded unless not enough ports are available for redundancy.

8.3 SAS host attachment

SAS Attachment is used in the very entry level for attachment of 1-4 Servers. This attachment is only available on entry FlashSystem products.

Check the supported server adapter and operating systems at [IBM System Storage Interoperation Center \(SSIC\)](#).

8.4 iSCSI host attachment considerations

It is recommended to separate Storage traffic from other network traffic.

Internet Small Computer Systems Interface (iSCSI) is a common protocol usable on any Ethernet network. With IBM Storage Virtualize 8.6 the performance was improved, especially on the midrange and high-end systems. IBM Storage Virtualize 8.6 now supports up to 1024 iSCSI Hosts per IO Group, depending on System.

Priority Flow Control for iSCSI / iSER is supported on Emulex and Chelsio adapters (SVC supported) with all DCBX enabled switches.

8.4.1 Connection basics

The cabling needs to be symmetrical. Port x of all node of an IO group needs to be connected to the same network.

Maximum 4 ports in a port group per host. All ports need to be on the same speed.

8.4.2 Performance tuning

Some of the attributes and host parameters that might affect iSCSI performance:

- ▶ Transmission Control Protocol (TCP) Delayed ACK
- ▶ Ethernet jumbo frame
- ▶ Network bottleneck or oversubscription
- ▶ iSCSI session login balance
- ▶ Priority flow control (PFC) setting and bandwidth allocation for iSCSI on the network

Disable the TCP delayed acknowledgment feature.

To disable this feature, refer to OS/platform documentation.

- ▶ VMware: <https://kb.vmware.com/s/article/1002598>.
- ▶ Windows: <http://support.microsoft.com/kb/823764>.
- ▶ Linux (RHEL): <https://access.redhat.com/solutions/1529183>.
- ▶ Linux (Other): Use IP utility to enable quickack on host side. For example, `ip route show; ip route change 192.168.140.0/24 dev eth7 proto kernel scope link src 192.168.140.55 quickack 1`.

The primary signature of this issue: read performance is significantly lower than write performance. Transmission Control Protocol (TCP) delayed acknowledgment is a technique that is used by some implementations of the TCP to improve network performance. However, in this scenario where the number of outstanding I/O is 1, the technique can significantly reduce I/O performance.

In essence, several ACK responses can be combined into a single response, reducing protocol overhead. As described in RFC 1122, a host can delay sending an ACK response by up to 500 ms. Additionally, with a stream of full-sized incoming segments, ACK responses must be sent for every second segment.

Enable jumbo frame for iSCSI.

Jumbo frames are Ethernet frames with a size in excess of 1,500 bytes. The maximum transmission unit (MTU) parameter is used to measure the size of jumbo frames.

Note: At the moment systems with RoCE Adapters cannot have MTU greater than 1500 on 8.6.0.0 or later. The workaround is to reduce MTU to 1500.

The network must support jumbo frames end-to-end to be effective. To verify, send a ping packet to be delivered without fragmentation to verify that the network supports jumbo frames. For example:

► Windows:

```
ping -t <iscsi target ip> -S <iscsi initiator ip> -f -l <new mtu size - packet overhead (usually 36, might differ)>
```

The following command is an example of a command that is used to check whether a 9000-bytes MTU is set correctly on a Windows system:

```
ping -t -S 192.168.1.117 192.168.1.217 -f -l 8964
```

The following output is an example of a successful reply:

```
192.168.1.217: bytes=8964 time=1ms TTL=52
```

► Linux:

```
ping -l <source iscsi initiator ip> -s <new mtu size> -M do <iscsi target ip>
```

► ESXi:

```
ping <iscsi target ip> -I <source iscsi initiator ip> -s <new mtu size - 28> -d
```

Verify the switch's port statistic where initiator/target ports are connected to make sure that packet drops are not high. Review network architecture to avoid any bottlenecks and oversubscription. The network needs to be balanced to avoid any packet drop; packet drop significantly reduces storage performance. Involve networking support to fix any such issues.

Optimize and utilize all iSCSI ports.

To optimize system resource utilization, all iSCSI ports must be used. Each port is assigned to one CPU, and by balancing the login, one can maximize CPU utilization and achieve better performance. Ideally, configure subnets equal to the number of iSCSI ports on the system node. Configure each port of a node with an IP on a different subnet and keep it the same for other nodes. The following example displays an ideal configuration:

```
Node 1
Port 1: 192.168.1.11
Port 2: 192.168.2.21
Port 3: 192.168.3.31
```

```
Node 2:
Port 1: 192.168.1.12
Port 2: 192.168.2.22
Port 3: 192.168.3.33
```

Avoid situations where 50 hosts are logged in to port 1 and only five hosts are logged in to port 2.

Use proper subnetting to achieve a balance between the number of sessions and redundancy.

Ensure that proper bandwidth is given to iSCSI on the network

You can divide the bandwidth among the various types of traffic. It is important to assign proper bandwidth for good performance. To assign bandwidth for iSCSI traffic, you need to first enable the priority flow control for iSCSI.

IBM FlashSystem systemwide CHAP is a two-way authentication method that uses CHAP to validate both the initiator and target. Changing or removing this systemwide CHAP can be disruptive, as it requires changes to the host configuration. If CHAP is changed without updating the host configuration, it can result in volume or LUN outages.

For any changes to systemwide CHAP, the host `iscsi` configuration requires an update to ensure it should only use a valid target (FlashSystem) CHAP, which is being updated.

Important: The `node.session.auth.username_in` needs to be assigned as “FlashSystem clustername”, which can be seen with the `lshost` command.

The configuration steps can be seen at [Setting up authentication for Linux hosts](#).

It is possible to have unique target authentication CHAP for individual host initiators. Using the `chhost` command allows individual host objects to offer unique username and chapsecret. However, the target system(FlashSystem) will have a common chapsecret and username (using CLI `ls` or `chsystem` can help managing chapsecret, while username will be name of the cluster) to offer for initiator authentication.

When using the SendTarget discovery method with FlashSystem, the IP addresses that are part of the same node and belong to the same portsets are returned in response. This means that if you have 8 nodes, you will need to perform iSCSI SendTarget discovery 8 times to get all of the IP addresses for the target.

For more information, see [iSCSI performance analysis and tuning](#).

8.5 iSCSI Extensions for Remote Direct Memory Access host attachment considerations

For host access, iSCSI Extensions for RDMA (iSER) is not supported on IBM Storage FlashSystem 7300, 9500, and IBM SAN Volume Controller (SVC) SV3 (both RDMA over Converged Ethernet (RoCE) and internet Wide-area Remote Direct Memory Access (RDMA) Protocol (iWARP)). If you are using iSER host attachment on existing IBM Storage Virtualize storage, these systems continue to support both RoCE and iWARP on IBM Storage Virtualize 8.6. For more information, see [IBM Support: V8.6.0.x Configuration Limits and Restrictions for IBM Storage FlashSystem 9500](#).

Note: Although iSER is still supported on the IBM Storage FlashSystem 5200, you should consider migrating to RoCE.

Priority Flow Control for iSCSI or iSER is supported on Emulex and Chelsio adapters (SVC supported) with all DCBX enabled switches.

8.6 NVMe over Fibre Channel host attachment considerations

IBM Storage Virtualize supports a single host initiator port that uses SCSI and Fibre Channel-Nonvolatile Memory Express (FC-NVMe) connections to the storage. NPIV needs to be enabled on the FlashSystem or SVC and the NVMe FC ports need to be zoned. NVMe/FC will not be used for external virtualization. Use different Zones for SCSI/FC and NVMe/FC for an easier management and overview. A volume can only be accessed using a single protocol (SCSI or NVMe), not both.

You need to create two host definitions for the host, one for SCSI and one for NVMe, if you want to use both type of volumes from one host, for example for migration.

- ▶ Host1scsi-LUN1-FC(SCSI)
- ▶ Host1nvme-LUN2-FC(NVMe)

Best practice is using only one protocol from one host.

Asymmetric Namespace Access was added to the FC-NVMe protocol standard, which gives it functions that are similar to Asymmetric Logical Unit Access (ALUA). As a result, FC-NVMe can now be used in stretched clusters. Consider NVMe over Fibre Channel target limits when you plan and configure the hosts.

Include the following points in your plan:

An NVMe host can connect to four NVMe controllers on each system node. The maximum per node is four with an extra four in failover.

1. Zone up to four ports in a single host to detect up to four ports on a node. To allow failover and avoid outages, zone the same or additional host ports to detect an extra four ports on the second node in the I/O group.
2. A single I/O group can contain up to 256 FC-NVMe I/O controllers. The maximum number of I/O controllers per node is 128 plus an extra 128 in failover. Following the recommendation in step 1, zone a total maximum of 16 hosts to detect a single I/O group. Also, consider that a single system target port allows up to 16 NVMe I/O controllers.

IBM Storage Virtualize 8.6 allows a maximum of 32/64 NVMe hosts per system and 16 hosts per I/O group, if no other types of hosts are attached. IBM Storage Virtualize code does not monitor or enforce these limits.

For more information about using NVMe hosts with IBM FlashSystem, see [NVMe over Fibre Channel Host Properties](#).

Note: Do not map the same volumes to SCSI and NVMe hosts concurrently, even one after the other. Also, take care not to add NVMe hosts and SCSI hosts to the same host cluster.

8.7 NVMe over Ethernet host attachment considerations

It is recommended to separate Storage traffic from other network traffic.

Asymmetric Namespace Access was added to the NVMe over RDMA and NVMe over TCP protocol standard, which gives it functions that are similar to Asymmetric Logical Unit Access (ALUA). As a result, FC-NVMe can now be used in stretched clusters.

iSCSI, iWarp, iSER, NVMe/RDMA and NVMe/TCP can coexist on the same port with different host mapping types. This includes the ports that are used for replication. A single host working with multiple protocols is not supported. Different hosts can have different protocols working with the same storage port.

Note: Do not map the same volumes to SCSI and NVMe hosts concurrently, even one after the other. Also, take care not to add NVMe hosts and SCSI hosts to the same host cluster.

Tip: IBM Storage Virtualize uses discovery port 4420 for all RDMA protocol instead of 8009.

8.7.1 NVMe over Ethernet (RoCE) host attachment considerations

IBM Storage Virtualize 8.6 allows a maximum of 512 NVMe hosts per system and 256 hosts per I/O group, if no other types of hosts are attached. IBM Storage Virtualize code does not monitor or enforce these limits. RoCE v2 is a very performing network protocol based on UDP with RDMA.

RDMA can reduce the CPU load of the server, compared with other Ethernet protocols. One limitation of existing NVMe/RoCE is that it requires special Ethernet infrastructure (lossless Ethernet / DCB). How to configure it depends on the switches you use, but you will need network skills. RoCE capable host adapters are also needed. Check [SSIC](#) to understand what is supported. RoCE v2 is routable, but prefer layer 2 networks.

Note: At the moment systems with ROCE Adapters cannot have MTU greater than 1500 on 8.6.0.0 or later. The workaround is to reduce MTU to 1500.

Check the release notes for future releases to see if the restriction is lifted.

8.7.2 NVMe over TCP host attachment considerations

Starting with Version 8.6.0.x NVMe/TCP is supported. Further improvements are planned. There are still some limits, check if these are lifted.

NVMe/TCP needs more CPU resources than protocols using RDMA. NVMe/TCP is a ubiquitous transport allowing NVMe performance without any constraint to the data center infrastructure. Each NVMe/TCP port on FlashSystem supports multiple IPs and multiple VLANs. NVMe/TCP will be supported on FlashSystem platforms that are installed with Mellanox CX-4 or CX-6 adapters.

NVMe-TCP is generally switch agnostic and routable. For Operating system support and Multipathing check at [SSIC](#).

For more information about using NVMe hosts with IBM FlashSystem, see [V8.6.0.x Configuration Limits and Restrictions for IBM FlashSystem 9500](#).

Note: At the moment systems with RoCE Adapters cannot have MTU greater than 1500 on 8.6.0.0 or later. The workaround is to reduce MTU to 1500.

Check on the release notes to see if the restriction is lifted.

8.8 Portsets

Portsets are groupings of logical addresses that are associated with specific traffic types. IBM Storage Virtualize 8.6.0 systems support both IP and FC portsets for host attachment, back-end storage connectivity, and replication traffic.

A system can have maximum of 72 portsets, which is a collective maximum limit for FC and Ethernet portsets. A portset can be of the host attach, remote copy, or storage type. The default portset is the host attach type. A portset of a specific type can be used only for that function, for example, a host attach type portset cannot be used for a remote copy partnership.

To see the FC and IP portsets, run the **lspportset** command.

Example 8-4 The lspportset command

```
svcinfolspportset
```

id	name	type	port_count	host_count	lossless	owner_id	owner_name	port_type	is_default
0	portset0	host	0	0				ethernet	yes
1	portset1	replication	0	0				ethernet	no
2	portset2	replication	0	0				ethernet	no
3	portset3	storage	0	0				ethernet	no
4	PortSet16	host	1	1	yes	0	Bank_Gr_Owngrp	fc	no
5	portset32	host	1	1	yes	1	Health_Gr_Owngrp	fc	no
64	portset64	host	6	4	yes			fc	yes

8.8.1 IP multitenancy

IP support for all IBM Storage Virtualize products previously allowed only a single IPv4 and IPv6 address per port for use with Ethernet connectivity protocols (internet Small Computer Systems Interface (iSCSI) and iSER).

As of version 8.4.2, IBM Storage Virtualize removed that limitation and supports an increased per port limit to 64 IP addresses (IPv4, IPv6, or both). The scaling of the IP definition also scaled the virtual local area network (VLAN) limitation, which can be done per IP address or as needed.

The object-based access control (OBAC) model (that is, OBAC-based per tenant administration and partitioned for multitenant cloud environments) also was added to the Ethernet configuration management.

The IBM Storage Virtualize new IP object model introduced a new feature that is named the *portset*. The *portset object* is a group of logical addresses that represents a typical IP function and traffic type. Portsets can be used for multiple traffic types, such as host attachment, back-end storage connectivity (iSCSI only), or IP replication.

The following commands can be used to manage an IP or Ethernet configuration:

- ▶ **lspportset**
- ▶ **mkportset**
- ▶ **chportset**
- ▶ **rmportset**
- ▶ **lsip** (**lspportip** deprecated)
- ▶ **mkip** (**cfgportip** deprecated)
- ▶ **rmkip** (**rmpportip** deprecated)
- ▶ **lspportethernet** (**lspportip** deprecated)

- ▶ **chportethernet** (**cfgport ip** deprecated)
- ▶ **mkhost** (with the parameter **-portset** to bind the host to the portset)
- ▶ **chost** (with the parameter **-portset** to bind the host to the portset)

A host can access storage through the IP addresses that are included in the portset that is mapped to the host. The process to bind a host to a portset includes the following steps:

1. Create the portset.
2. Configure the IP addresses with the portset.
3. Create a host object.
4. Bind the host to the portset.
5. Discover and log in from the host.

IP portsets can be added by using the management GUI or the command-line interface (CLI). You can configure portsets by using the GUI and selecting **Settings** → **Network** → **Portsets**.

After the portsets are created, IP addresses can be assigned by using the management GUI or the CLI. You can configure portsets by using the GUI and selecting **Settings** → **Network** → **Ethernet Ports**.

Example 8-5 shows the results of the usage of the **lsip** command.

Example 8-5 Output of using the lsip command

```

svcinfolsip
id node_id node_name port_id portset_id portset_name IP_address prefix vlan gateway
owner_id owner_name
0 1 node1 1 0 portset0 10.0.240.110 24 10.0.240.9
1 1 node1 1 1 portset1 10.0.240.110 24 10.0.240.9
2 2 node2 1 0 portset0 10.0.240.111 24 10.0.240.9
3 2 node2 1 1 portset1 10.0.240.111 24 10.0.240.9

```

8.8.2 Fibre Channel portset

FC portsets were introduced in IBM Storage Virtualize 8.5 for effective port management and host access. An FC portset is a group of FC I/O ports. A host must be associated to a portset, and can access storage through the associated portset only. This configuration is recommended for larger configurations (256 and greater host counts). Conventionally, host access is managed by FC zoning, which might be complex and result in unbalanced resource utilization. When used with good zoning practices, portsets provide better management and resource utilization in configurations with large host counts (256 and greater).

Two modes of portsets are available:

- ▶ **Legacy mode:** On new installation or when upgrading to version 8.6.0.x, all FC ports are added to default portset. All existing host objects automatically are associated with the default portset. This mode is useful for smaller configurations, and does not require portset knowledge.
- ▶ **User-defined portsets:** Recommended for larger configurations. Can be defined by a user as part of the initial configuration. All host objects are limited by their associated portset for storage access.

While using the FC portset feature, each FC I/O port can be added to multiple FC portsets; however, a host can be added to only one FC portset. Every portset can support up to four FC I/O ports.

Each portset is identified by a unique name. Portset 0 is an Ethernet default portset, and portset 64 is a default FC portset that is configured when the system is created or updated. Portsets 1 and 2 are replication portsets. Portset 3 is a storage portset.

To see the host login for each FC port, run the `lstorageportfc` command.

Example 8-6 The lstorageportfc command

```

svcinfo lstorageportfc
id WWPN          WWNN          port_id owning_node_id current_node_id nportid host_io_permitted virtualized protocol fc_io_port_id
portset_count host_count active_login_count
1 5005076812110967 5005076812000967 1 1 1 060600 no no scsi 1 0
0 3
2 5005076812150967 5005076812000967 1 1 1 060601 yes yes scsi 1 1
6 1
3 5005076812190967 5005076812000967 1 1 1 060602 yes yes nvme 1 1
0 0
4 5005076812120967 5005076812000967 2 1 1 070600 no no scsi 2 0
0 3
5 5005076812160967 5005076812000967 2 1 1 070601 yes yes scsi 2 1
6 1
6 50050768121A0967 5005076812000967 2 1 1 070602 yes yes nvme 2 1
0 0
7 5005076812130967 5005076812000967 3 1 1 080600 no no scsi 3 0
0 1
8 5005076812170967 5005076812000967 3 1 1 080601 yes yes scsi 3 1
6 0
9 5005076812180967 5005076812000967 3 1 1 080602 yes yes nvme 3 1
0 0
10 5005076812140967 5005076812000967 4 1 1 090600 no no scsi 4 0
0 1
11 5005076812180967 5005076812000967 4 1 1 090601 yes yes scsi 4 1
6 0
12 50050768121C0967 5005076812000967 4 1 1 090602 yes yes nvme 4 1
0 0

```

The output shows details about each FC port, associated portset count, host count, and active login counts. If some FC ports have a higher number of active logins, they can cause an unbalanced performance.

Note: Event ID 088007 “Fibre Channel I/O port has more than recommended active login” is logged when there are more than 256 active logins on all or any of the worldwide port names (WWPNs) on the FC I/O port on any node. This event informs the customer that an FC I/O port is serving login more than the recommended limit, and the system is being under-utilized because the load is not distributed uniformly across the FC I/O ports and nodes. The event is cleared when an administrator fixes the zoning such that the total active login count of the FC I/O port becomes less than or equal to 256.

Generally for FC, the host to storage connection is controlled by SAN zoning. A portset helps to set a rule on storage layer to avoid many FC logins to same port while other ports remain idle. Misconfiguring portsets and hosts or wrong ports that are used in zones result in the Event ID 064002 being logged in the IBM Storage Virtualize 8.6 event log.

Note: Event ID 064002 “Host and Fibre Channel port must be in same portset.” is logged when the host tries to log in to a port that is associated with a different portset. A login is detected from the host to a storage port that is not part of the same portset of which the host is a part. For an FC host, check “Fibre Channel Connectivity” for the host with a state of Blocked, and for a NVMe host, check “NVMe Connectivity” for a host with a state of Invalid. Add storage port WWPNs that are assigned to the correct portset on the host to the storage zone and remove the wrong ones. This action automatically clears the event.

For multi-tenant configuration portsets, separate the FC ports to the tenants. Example 8-4 on page 530 shows two portsets that are assigned to different ownership groups.

8.8.3 Portsets considerations and limitations

Consider the following points about portsets:

- ▶ Multiple hosts can be mapped to a single portset.
- ▶ A single host cannot be mapped to multiple portsets.
- ▶ A system can have maximum of 72 portsets.
- ▶ IP addresses and FC ports can belong to multiple portsets.
- ▶ Port masking is used to enable or disable each port per feature for specific traffic types (host, storage, and replication).
- ▶ Portset 0, Portset 3, Portset 64, and the replication portset are predefined.
- ▶ When an IP address or host is configured, a portset must be specified.
- ▶ Portset 0 is the default portset for iSCSI, and Portset 64 is the default portset for FC that is automatically configured when the system is updated or created and cannot be deleted.
- ▶ Portset 0 and Portset 64 allow administrators to continue to use an original configuration that does not require multi-tenancy.
- ▶ Portset 3 is used for iSCSI back-end storage virtualization.
- ▶ After an update to the 8.5.x code or later, all configured FC host objects are automatically mapped to Portset 64, and iSCSI host objects are mapped to Portset 0.
- ▶ Unconfigured logins are rejected upon discovery.
- ▶ The internet Storage Name Service (iSNS) function registers IP addresses in Portset 0 only with the iSNS server.
- ▶ Each port can be configured with only one unique routable IP address (gateway-specified).

8.9 Host pathing

Each host mapping associates a volume with a host object and allows all HBA ports in the host object to access the volume. You can map a volume to multiple host objects.

When a mapping is created, multiple paths normally exist across the SAN fabric from the hosts to the IBM Storage Virtualize system. Most operating systems present each path as a separate storage device. Therefore, multipathing software is required on the host. The multipathing software manages the paths that are available to the volume, presents a single storage device to the operating system, and provides failover if a path is lost.

If your IBM Storage Virtualize system uses NPIV, path failures that occur because of an offline node are masked from host multipathing.

Note: When using NPIV and an SVC cluster with site awareness, the remaining node in the same I/O group take over the addresses. Be aware of the potential impact on inter-data center traffic here. This does not apply if Hot Spare nodes are used per site. We recommend to zone the host to both sides.

8.9.1 Path selection

I/O for a specific volume is handled exclusively by the nodes in a single I/O group. Although both nodes in the I/O group can service the I/O for the volume, the system prefers to use a consistent node, which is called the *preferred node*. The primary reasons for using a preferred node are load-balancing and to determine which node destages writes to the back-end storage.

When a volume is created, an I/O group and preferred node are defined, and optionally can be set by the administrator. The owner node for a volume is the preferred node when both nodes are available.

With HyperSwap, configuration nodes in multiple I/O groups can potentially service I/O for the same volume, and site IDs are used to optimize the I/O routing.

IBM Storage Virtualize uses Asymmetric Logical Unit Access (ALUA), as do most multipathing drivers. Therefore, the multipathing driver gives preference to paths to the preferred node. Most modern storage systems use ALUA.

Note: Some competitors claim that ALUA means that IBM Storage Virtualize is effectively an active-passive cluster. This claim is not true. Both nodes in IBM Storage Virtualize can and do service I/O concurrently.

In the small chance that an I/O goes to the non-preferred node, that node services the I/O without issue.

8.10 I/O queues

Host operating system and HBA software must have a way to fairly prioritize I/O to the storage. The host bus might run faster than the I/O bus or external storage. Therefore, you must have a way to queue I/O to the devices. Each operating system and host adapter use unique methods to control the I/O queue.

The I/O queue can be controlled by using one of the following unique methods:

- ▶ Host adapter-based
- ▶ Memory and thread resources-based
- ▶ Based on the number of commands that are outstanding for a device

8.10.1 Queue depths

Queue depth is used to control the number of concurrent operations that occur on different storage resources. Queue depth is the number of I/O operations that can be run in parallel on a device.

Queue depths apply at various levels of the system:

- ▶ Disk or flash
- ▶ Storage controller
- ▶ Per volume and HBA on the host

For example, each IBM Storage Virtualize node has a queue depth of 10,000. A typical disk drive operates efficiently at a queue depth of 8. Most host volume queue depth defaults are approximately around 32.

Guidance for limiting queue depths in large SANs that was described in previous documentation was replaced with calculations for overall I/O group-based queue depth considerations.

No set rule is available for setting a queue-depth value per host HBA or per volume. The requirements for your environment are driven by the intensity of each workload.

Ensure that one application or host cannot use the entire controller queue. However, if you have a specific host application that requires the lowest latency and highest throughput, consider giving it a proportionally larger share than others.

Consider the following points:

- ▶ A single IBM Storage Virtualize FC port accepts a maximum concurrent queue depth of 2048.
- ▶ A single IBM Storage Virtualize node accepts a maximum concurrent queue depth of 10,000. After this depth is reached, it reports a full status for the queue.
- ▶ Host HBA queue depths must be set to the maximum (typically 1024).
- ▶ The host queue depth must be controlled through the per volume value:
 - A typical random workload volume must use a value of approximately 32.
 - To limit the workload of a volume, use a value of 4 or less.
 - To maximize throughput and give a higher share to a volume, use a value of 64.

The total workload capability can be calculated by multiplying the number of volumes by their respective queue depths and summing. With low latency storage, a workload of over 1 million input/output operations per second (IOPS) can be achieved with concurrency on a single I/O group of 1000.

For more information about queue depths, see the following IBM Documentation:

- ▶ [FC hosts](#)
- ▶ [iSCSI hosts](#)
- ▶ [iSER hosts](#)

8.11 Host clusters

IBM Storage Virtualize supports host clusters. This feature allows multiple hosts to access the same set of volumes.

Volumes that are mapped to that host cluster are assigned to all members of the host cluster with the same SCSI ID. A typical use case is to define a host cluster that contains all the WWPNs that belong to the hosts that are participating in a host operating system-based cluster, such as IBM PowerHA®, Microsoft Cluster Server (MSCS), or VMware ESXi clusters.

The following commands can be used to manage host clusters:

- ▶ `lshostcluster`
- ▶ `lshostclustermember`
- ▶ `lshostclustervolumemap`
- ▶ `addhostclustermember`
- ▶ `chostcluster`
- ▶ `mkhost` (with parameter `-hostcluster` to create the host in one cluster)
- ▶ `mkhostcluster`
- ▶ `mkvolumehostclustermap`
- ▶ `rmhostclustermember`

- ▶ **rmhostcluster**
- ▶ **rmvolumehostclustermap**

Host clusters can be added by using the GUI. By using the GUI, the system assigns the SCSI IDs for the volumes (you can also manually assign them). For ease of management purposes, use separate ranges of SCSI IDs for hosts and host clusters.

For example, you can use SCSI IDs 0 - 99 for non-cluster host volumes, and greater than 100 for the cluster host volumes. When you choose the **System Assign** option, the system automatically assigns the SCSI IDs starting from the first available in the sequence.

If you choose **Self Assign**, the system enables you to select the SCSI IDs manually for each volume. On the right side of the window, the SCSI IDs that are used by the selected host or host cluster are shown (see Figure 8-1).

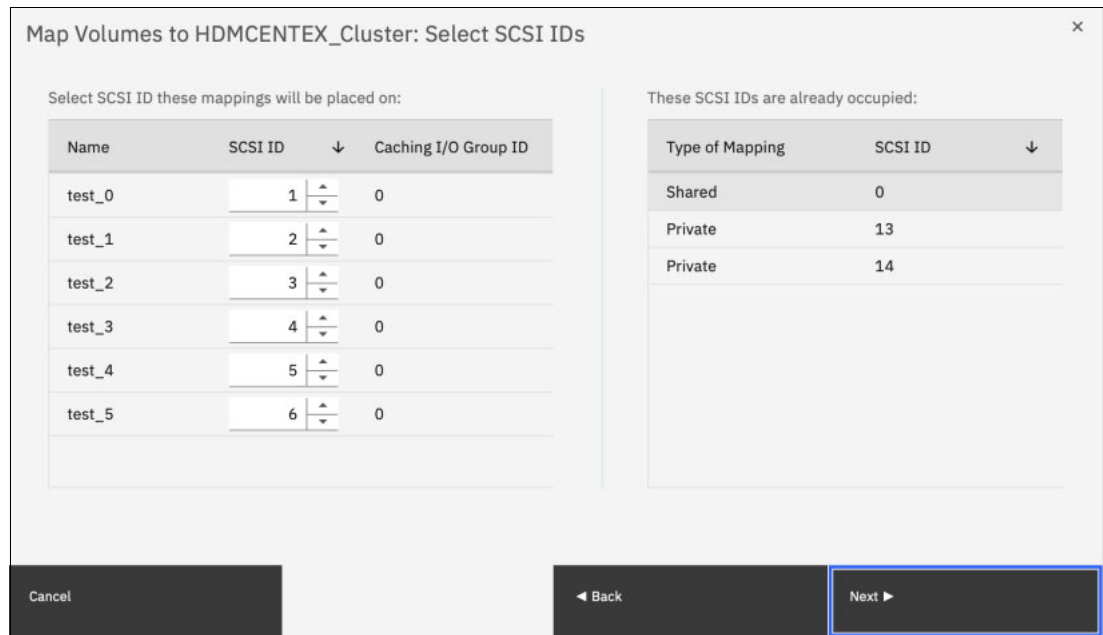


Figure 8-1 SCSI ID assignment on volume mappings

Note: Although extra care is always recommended when dealing with hosts, IBM Storage Virtualize does not allow you to join a host into a host cluster if it includes a volume mapping with a SCSI ID that also exists in the host cluster:

```
IBM_2145:ITS0-SVCLab:superuser>addhostclustermember -host ITS0_HOST3
ITS0_CLUSTER1
CMMVC9068E Hosts in the host cluster have conflicting SCSI IDs for their
private mappings.
IBM_2145:ITS0-SVCLab:superuser>
```

8.11.1 Persistent reservations

To prevent hosts from sharing storage inadvertently, they must follow a storage reservation mechanism. The mechanisms for restricting access to IBM Storage Virtualize volumes use the SCSI-3 persistent reserve commands or the SCSI-2 reserve and release commands.

The host software uses several methods to implement host clusters. These methods require sharing the volumes on IBM Storage Virtualize between hosts. To share storage between hosts, the cluster must maintain control over accessing the volumes. Some clustering software uses software locking methods.

You can choose other methods of control by directing the clustering software or the device drivers to use the SCSI architecture reserve or release mechanisms. The multipathing software can change the type of reserve that is used from an earlier reserve to persistent reserve, or remove the reserve.

Persistent reserve refers to a set of SCSI-3 standard commands and command options that provide SCSI initiators with the ability to establish, preempt, query, and reset a reservation policy with a specified target device. The functions that are provided by the persistent reserve commands are a superset of the original reserve or release commands.

The persistent reserve commands are incompatible with the earlier reserve or release mechanism. Also, target devices can support only reservations from the earlier mechanism or the new mechanism. Attempting to mix persistent reserve commands with earlier reserve or release commands results in the target device returning a reservation conflict error.

Earlier reserve and release mechanisms (SCSI-2) reserved the entire LUN (volume) for exclusive use down a single path. This approach prevents access from any other host or even access from the same host that uses a different host adapter. The persistent reserve design establishes a method and interface through a reserve policy attribute for SCSI disks. This design specifies the type of reservation (if any) that the operating system device driver establishes before it accesses data on the disk.

The following possible values are supported for the reserve policy:

- ▶ `No_reserve`: No reservations are used on the disk.
- ▶ `Single_path`: Earlier reserve or release commands are used on the disk.
- ▶ `PR_exclusive`: Persistent reservation is used to establish *exclusive host access* to the disk.
- ▶ `PR_shared`: Persistent reservation is used to establish *shared host access* to the disk.

When a device is opened (for example, when the AIX `varyonvg` command opens the underlying hdisks), the device driver checks the object data manager (ODM) for a `reserve_policy` and a `PR_key_value`. Then, the driver opens the device. For persistent reserve, each host that is attached to the shared disk must use a unique registration key value.

8.11.2 Clearing reserves

It is possible to accidentally leave a reserve on the IBM Storage Virtualize volume or on the IBM Storage Virtualize managed disk (MDisk) during migration into IBM Storage Virtualize, or when disks are reused for another purpose. Several tools are available from the hosts to clear these reserves.

Instances exist in which a host image mode migration appears to succeed; however, problems occur when the volume is opened for read/write I/O. The problems can result from not removing the reserve on the MDisk before image mode migration is used in IBM Storage Virtualize.

You cannot clear a leftover reserve on an IBM Storage Virtualize MDisk from IBM Storage Virtualize. You must clear the reserve by mapping the MDisk back to the owning host and clearing it through host commands, or through back-end storage commands as advised by IBM technical support.

8.12 AIX hosts

This section describes support and considerations for AIX hosts.

For more information about configuring AIX hosts, see [IBM Power Systems AIX hosts](#).

8.12.1 Multipathing support

Subsystem Device Driver Path Control Module (SDDPCM) is no longer supported. Use the default AIX PCM. For more information, see [The Recommended Multi-path Driver to use on IBM AIX and VIOS When Attached to SVC and Storwize storage](#).

8.12.2 AIX configuration recommendations

These device settings can be changed by using the **chdev** AIX command:

```
reserve_policy=no_reserve
```

The default reserve policy is **single_path** (SCSI-2 reserve). Unless a specific need exists for reservations, use **no_reserve**.

```
algorithm=shortest_queue
```

If coming from SDD PCM, AIX defaults to **fail_over**. You cannot set the algorithm to **shortest_queue** unless the reservation policy is **no_reserve**:

```
queue_depth=32
```

The default queue depth is 20. IBM recommends 32:

```
rw_timeout=30
```

The default for SDD PCM is 60. For AIX PCM, the default is 30. IBM recommends 30.

For more information about configuration best practices, see [AIX Multi Path Best Practices](#).

8.13 Virtual I/O Server hosts

This section describes support and considerations for Virtual I/O Server (VIOS) hosts.

For more information about configuring VIOS hosts, see [IBM Power Systems with Virtual I/O Server](#).

8.13.1 Multipathing support

Subsystem Device Driver Path Control Module (SDDPCM) is no longer supported. Use the default AIX PCM.

For more information, see [The Recommended Multi-path Driver to use on IBM AIX and VIOS When Attached to SVC and Storwize storage](#). Where VIOS SAN Boot or dual VIOS configurations are required, see [SSIC](#).

For more information about VIOS, see this [IBM Virtual I/O Server overview](#).

8.13.2 VIOS configuration recommendations

These device settings can be changed by using the **chdev** AIX command:

```
reserve_policy=single_path
```

The default reserve policy is **single_path** (SCSI-2 reserve):

```
algorithm=fail_over
```

If coming from SDD PCM, AIX defaults to **fail_over**:

```
queue_depth=32
```

The default queue depth is 20. IBM recommends 32.

```
rw_timeout=30
```

The default for SDD PCM is 60. For AIX PCM, the default is 30. IBM recommends 30.

8.13.3 Physical and logical volumes

Virtual Small Computer System Interface (VSCSI) is based on a client/server relationship. The VIOS owns the physical resources and acts as the server or target device.

Physical storage with attached disks (in this case, volumes on IBM Storage Virtualize) on the VIOS partition can be shared by one or more client logical partitions (LPARs). These client LPARs contain a VSCSI client adapter (SCSI initiator) that detects these virtual devices (VSCSI targets) as standard SCSI-compliant devices and LUNs.

You can create the following types of volumes on a VIOS:

- ▶ Physical volume (PV) VSCSI hdisks
- ▶ Logical volume (LV) VSCSI hdisks

PV VSCSI hdisks are entire LUNs from the VIOS perspective. If you are concerned about the failure of a VIOS and configured redundant VIOSs for that reason, you must use PV VSCSI hdisks. An LV VSCSI hdisk cannot be served up from multiple VIOSs.

LV VSCSI hdisks are in Logical Volume Mirroring (LVM) volume groups on the VIOS and must not span PVs in that volume group or be striped LVs. Because of these restrictions, use PV VSCSI hdisks.

8.13.4 Identifying a disk for use as a VSCSI disk

The VIOS uses the following methods to uniquely identify a disk for use as a VSCSI disk:

- ▶ Unique device identifier (UDID)
- ▶ Physical volume identifier (PVID)
- ▶ IEEE volume ID

Each of these methods can result in different data formats on the disk. The preferred disk identification method for volumes is to use UDIDs. For more information about how to determine your disk IDs, see [Identifying exportable disks](#).

8.14 Microsoft Windows hosts

This section describes support and considerations for Microsoft Windows hosts, including Microsoft Hyper-V.

For more information about configuring Windows hosts, see [Hosts that run the Microsoft Windows Server operating system](#).

8.14.1 Multipathing support

For multi-pathing support, use Microsoft multipath I/O (MPIO) with Microsoft Device Specific Module (MS DSM), which is included in the Windows Server operating system. The older Subsystem Device Driver Device Specific Module (SDDDSM) is no longer supported. For more information, see [IBM Storage Virtualize Multipathing Support for AIX and Windows Hosts](#).

The Windows multipathing software supports the following maximum configuration:

- ▶ Up to eight paths to each volume
- ▶ Up to 2048 volumes per windows server or host
- ▶ Up to 512 volumes per Hyper-V host

8.14.2 Windows and Hyper-V configuration recommendations

Ensure that the following components are configured:

- ▶ Operating system service packs and patches and clustered-system software
- ▶ HBAs and HBA device drivers
- ▶ Multipathing drivers (Microsoft Device Specific Module (MSDSM))

Regarding disk timeout for Windows servers, change the disk I/O timeout value to 60 in the Windows registry.

8.15 Linux hosts

This section describes support and considerations for Linux hosts.

For more information about configuring Linux hosts, see [Hosts that run the Linux operating system](#).

8.15.1 Multipathing support

IBM Storage Virtualize supports Linux hosts that use native Device Mapper-Multipathing (DM-MP) and native multipathing support.

Note: Occasionally, we see that storage administrators modify parameters in the `multipath.conf` file to address some perceived shortcoming in the DM-MP configuration. These modifications can create unintended and unexpected behaviors. The recommendations that are provided in IBM Documentation are optimal for most configurations.

8.15.2 Linux configuration recommendations

Consider the following points about configuration settings for Linux:

- ▶ Settings and udev rules can be edited in `/etc/multipath.conf`.
- ▶ Some Linux levels require `polling_interval` to be under the defaults section. If `polling_interval` is under the device section, comment it out by using the `#` key, as shown in the following example:

```
# polling_interval
```
- ▶ Use the default values as described in [Settings for Linux hosts](#).
- ▶ The `dev_loss_tmo` settings control how long to wait for devices or paths to be pruned. If the inquiry is too short, it might time out before the paths are available. IBM recommends 120 seconds for this setting.
- ▶ For better performance use LVM with multiple disks instead of filesystems running on a single disk.

Best practice: The `scsi_mod.inq_timeout` should be set to 70. If this timeout is set incorrectly, it can cause paths to not be rediscovered after a node is restarted.

For more information about this setting and other attachment requirements, see [Attachment requirements for hosts that are running the Linux operating system](#).

8.16 Oracle Solaris hosts support

This section describes support and considerations for Oracle hosts. SAN boot and clustering support are available for Oracle hosts.

For more information about configuring Solaris hosts, see [Oracle hosts](#).

8.16.1 Multipathing support

IBM Storage Virtualize supports multipathing for Oracle Solaris hosts through Oracle Solaris MPxIO, Symantec Veritas Volume Manager Dynamic Multi-Pathing (DMP), and the Native Multipathing Plug-in (NMP). Specific configurations depend on file system requirements, HBA, and operating system level.

Note: The NMP does not support the Solaris operating system in a clustered-system environment. For more information about your supported configuration, see [SSIC](#).

8.16.2 Solaris MPxIO configuration recommendations

IBM Storage Virtualize software supports load balancing of the MPxIO software, using a command like:

```
svctask mkhost -name new_name_arg -hbawpn wwpn_list
```

Note: Starting with IBM Storage Virtualize version 8.1 a host type of `tpgs` is no longer mandatory. The behavior is now the same as with a host type of `generic`.

To complete your configuration, complete the following steps:

1. Install the latest Solaris host patches.
2. Copy the `/kernel/drv/scsi_vhci.conf` file to the `/etc/driver/drv/scsi_vhci.conf` file.
3. Set the `load-balance="round-robin"` parameter.
4. Set the `auto-failback="enable"` parameter.
5. Comment out the `device-type-scsi-options-list = "IBM 2145", "symmetric-option"` parameter.
6. Comment out the `symmetric-option = 0x1000000` parameter.
7. Restart hosts or run `stmsboot -u` based on the host level.
8. Verify changes by running `luxadm display /dev/rdisk/cXtYdZs2`, where `cXtYdZs2` is your storage device.
9. Check that the preferred node paths are primary and online, and that the non-preferred node paths are secondary and online.

8.16.3 Symantec Veritas DMP configuration recommendations

When you are managing IBM Storage Virtualize storage in Symantec volume manager products, you must install an Array Support Library (ASL) on the host so that the volume manager is aware of the storage subsystem properties (active/active or active/passive).

If a suitable ASL is not installed, the volume manager does not claim the LUNs. Using ASL is required to enable the special failover or failback multipathing that IBM Storage Virtualize requires for error recovery.

To determine the basic configuration of a Symantec Veritas server, run the commands that are shown in Example 8-7.

Example 8-7 Determining the Symantec Veritas server configuration

```
pkginfo -l (lists all installed packages)
showrev -p |grep vxvm (to obtain version of volume manager)
vxddladm listsupport (to see which ASLs are configured)
vxdisk list
vxdmpadm listctrl all (shows all attached subsystems, and provides a type where
possible)
vxdmpadm getsubpaths ctrl=cX (lists paths by controller)
vxdmpadm getsubpaths dmpnodename=cxtxdxs2' (lists paths by LUN)
```

The commands that are shown in Example 8-8 and Example 8-9 determine whether IBM Storage Virtualize is correctly connected. They also show which ASL is used: native Dynamic Multi-Pathing (DMP), ASL, or SDD ASL.

Example 8-8 shows what you see when Symantec Volume Manager correctly accesses IBM Storage Virtualize by using the SDD pass-through mode ASL.

Example 8-8 Symantec Volume Manager that uses SDD pass-through mode ASL

```
# vxdmpadm list enclosure all
ENCLR_NAME ENCLR_TYPE ENCLR_SNO STATUS
=====
OTHER_DISKS OTHER_DISKS OTHER_DISKS CONNECTED
VPATH_SANVCO VPATH_SANVC 0200628002faXX00 CONNECTED
```

Example 8-9 shows what you see when IBM Storage Virtualize is configured by using native DMP ASL.

Example 8-9 IBM Storage Virtualize that is configured by using native ASL

```
# vxdmpadm list enclosure all
ENCLR_NAME ENCLR_TYPE ENCLR_SNO STATUS
=====
OTHER_DISKS OTHER_DSKSI OTHER_DISKS CONNECTED
SAN_VCO SAN_VC 0200628002faXX00 CONNECTED
```

For more information about the latest ASL levels to use native DMP, see the array-specific module table that is available at [Veritas](#).

To check the installed Symantec Veritas version, run the following command:

```
showrev -p |grep vxvm
```

To check which IBM ASLs are configured in the volume manager, run the following command:

```
vxddladm listsupport |grep -i ibm
```

After you install a new ASL by using the `pkgadd` command, restart your system or run the `vxctl enable` command. To list the ASLs that are active, run the following command:

```
vxddladm listsupport
```

8.17 HP 9000 and HP Integrity hosts

This section describes support and considerations for Linux hosts. SAN boot is supported for all HP-UX 11.3x releases on both HP 9000 and HP Integrity servers.

For more information about configuring Linux hosts, see [HP 9000 and HP Integrity](#).

8.17.1 Multipathing support

IBM Storage Virtualize supports multipathing for HP-UX hosts through HP PVLlinks and the NMP. Dynamic multipathing is available when you add paths to a volume or when you present a new volume to a host.

To use PVLlinks while NMP is installed, ensure that NMP did not configure a vpath for the specified volume.

For more information about a list of configuration maximums, see [Multipathing configuration maximums for HP 9000 and HP Integrity servers](#).

8.17.2 HP configuration recommendations

Consider the following configuration recommendations for HP:

- ▶ HP-UX 11.31 September 2007 and later 0803 releases are supported.
- ▶ HP-UX 11.31 contains native multipathing as part of the mass storage stack feature.
- ▶ NMP supports only HP-UX 11iv1 and HP-UX 11iv2 operating systems in a clustered-system environment.
- ▶ SCSI targets that use more than eight LUNs must have type attribute **hpux** set to host object.
- ▶ Ensure that the host object is configured with the type attribute set to **hpux** as shown in the following example:

```
svctask mkhost -name new_name_arg -hbawwpn wwpn_list -type hpux
```
- ▶ Configure the Physical Volume timeout for NMP for 90 seconds.
- ▶ Configure the Physical Volume timeout for PVLlinks for 60 seconds (the default is 4 minutes).

8.18 VMware ESXi server hosts

This section describes considerations for VMware hosts.

For more information about configuring VMware hosts, see [Hosts that run the VMware ESXi operating system](#).

To determine the various VMware ESXi levels that are supported, see the [SSIC](#).

We recommend using VMware Tools on the virtual machines. Using VMware-Plugin the new native or via IBM Spectrum Connect can make the management easier, like the same volume name on VMware as on IBM FlashSystem.

There are some limitations by VMware on NVMe volumes.

- ▶ [Requirements and Limitations of VMware NVMe Storage VMware 7](#)
- ▶ [Requirements and Limitations of VMware NVMe Storage VMware 8](#)

8.18.1 Multipathing support

VMware features a built-in multipathing driver that supports IBM Storage Virtualize ALUA-preferred path algorithms.

For information about limits and a complete list of maximums, see [VMware Configuration Maximums](#).

8.18.2 VMware configuration recommendations

For more information about specific configuration best practices for VMware, see [Configuring the ESXi operating system](#).

Consider and verify the following settings:

- ▶ The storage array type plug-in should be ALUA (VMW_SATP_ALUA).
- ▶ Path selection policy should be RoundRobin (VMW_PSP_RR).
- ▶ Set the XCopy transfer size to 4096.
- ▶ The recommended number of paths per volume is four. It is recommended not to exceed 8. In a HyperSwap configuration, you can use eight, up to 16 paths per volume is supported.
- ▶ The RoundRobin IOPS should be changed from 1000 to 1 so that I/Os are evenly distributed across as many ports on the system as possible. For more information about how to change this setting, see [Adjusting Round Robin IOPS limit on VMware knowledge base](#).
- ▶ If preferred, all VMware I/O paths (active optimized and non-optimized) can be used by running the following `esxcli` command:

```
esxcli storage nmp psp roundrobin deviceconfig set --useano=1 -d <naa of the device>
```
- ▶ For Ethernet attachment, like NVMe over TCP it should be considered to use a dedicated pair of ports, configured using dedicated vSwitches and VMkernel ports.

For more information about active optimized and active non-optimized paths, see [Active - active capability](#).

Note: You check whether your volumes are seen as flash on the ESXi server. In some cases, VMware marks IBM FlashSystem volumes as hard disk drives (HDDs). As a best practice, mark volumes as Flash before creating a datastore on them.

8.19 Container Storage Interface Block Driver

The Container Storage Interface (CSI) enables the Container Orchestrators Platform to perform actions on storage systems. The CSI Block Driver connects Kubernetes and Red Hat OpenShift Container Platform (OCP) to IBM Block storage devices (IBM Storage Virtualize, IBM FlashSystem, and IBM DS8000). This process is done by using persistent volumes (PVs) to dynamically provision block storage with stateful containers. Provisioning can be fully automated to scale, deploy, and manage containerized applications. The CSI driver allows IBM Storage Virtualize to participate in hybrid multicloud environments for modern infrastructures.

To use the IBM Block storage CSI driver, complete the following steps:

1. Create an array secret.
2. Create a storage class.
3. Create a Persistent Volume Claim (PVC) that is 1 Gb.
4. Display the PVC and the created PV.
5. Create a StatefulSet.

For more information about installing, configuring, and using CSI Block Driver, see [IBM block storage CSI driver](#).

8.20 100-Gigabit Ethernet host connectivity

IBM Storage FlashSystem 9500, FlashSystem 7300, and SVC SV3 support new 100-gigabit Ethernet (GbE) Ethernet host connectivity options, which provide end-to-end NVMe capability in the I/O path from hosts to the back-end.

With Storage Virtualize 8.6 the Ethernet restrictions have been removed on the FS9500 and IBM SAN Volume Controller SV3.

All cages can be fully populated with two port 25 Gb/100 Gb Ethernet cards

For 100Gb adapters, fully populating a cage means the *bandwidth will be oversubscribed*. Each adapter will get 128 Gb of its theoretical 200 Gb capability (as in the FS7300).

Note: In some situations you need maximum bandwidth, other situations need maximum ports. By easing the restrictions on the Ethernet adapters we are allowing the customer to make that choice. If maximum ports are selected, the bandwidth will be oversubscribed

8.20.1 Dual port 100 GbE adapter functions

Table 8-1 lists the supported and unsupported adapter functions and their caveats.

Table 8-1 100 GbE adapter functions

Function	Supported	Not supported
NVMe over TCP host attachment	✓	
NVMe over RDMA (RoCE V2) connectivity for host attachment ^a	✓	
iSCSI connectivity for host attachment ^b	✓	
Clustering / HyperSwap		✓
Replication connectivity		✓
iSCSI-based external storage connectivity		✓
iSER connectivity for host attachment		✓

a. The primary use case for optimal performance.

b. A secondary use case to allow boot from SAN through 100 GbE. Performance is equivalent to 25-GbE iSCSI host attachment.

8.20.2 Maximum adapter count and slot placement

Table 8-2 lists the maximum number of Dual Port 100 GbE adapters that are supported per node on IBM Storage FlashSystem 7300, FlashSystem 9500, and SVC SV3.

Table 8-2 Maximum 100-GbE adapters per node and PCIe slot placement

System	Maximum Dual Port 100 GbE adapter count	Adapter slot placement per node for maximum port performance
IBM FlashSystem 7300	3	1, 2, and 3
IBM FlashSystem 9500	6	1, 5, and 7
SVC SV3	6	1, 5, and 7

Note: The 100 Gbps Ethernet adapter is limited to Peripheral Component Interconnect Express (PCIe) adapters on this hardware. As a best practice, attempt to balance I/O over all the ports as evenly as possible, especially for NVMe over RDMA host attachment on IBM FlashSystem because the PCIe slots are oversubscribed. Performance should be calculated for use on a primary or failover model to avoid PCIe slot oversubscription.

Figure 8-2 depicts the slots on an IBM Storage FlashSystem 7300 that can contain Dual Port 100 GbE adapters.

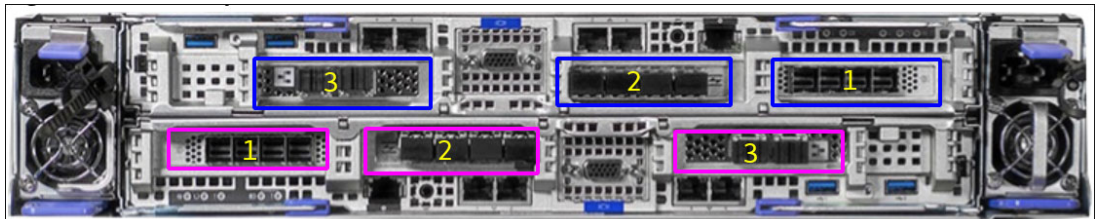


Figure 8-2 Dual Port 100 GbE adapter placement on IBM FlashSystem Storage 7300

Note: On IBM FlashSystem 7300 when both ports of an adapter are used for NVMe over RDMA, or NVMe over TCP, or both, I/O operation types are limited to 100 Gbps per adapter (not per port).

Figure 8-3 on page 548 depicts the slots on an IBM Storage FlashSystem 9500 that can contain Dual Port 100 GbE adapters.



Figure 8-3 Dual Port 100 GbE adapter placement on IBM Storage FlashSystem 9500

Figure 8-4 shows the slots on an SVC SV3 node that can contain Dual Port 100 GbE adapters.



Figure 8-4 Dual Port 100 GbE adapter placement on SAN Volume Controller node SV3

Note: When one or more Dual Port 100 GbE adapters are installed in IBM Storage FlashSystem 9500 or SVC SV3 nodes, they should always occupy the lower numbered slot in the adapter cage, and the other slot must not contain an adapter. For adapter cage 2, slots 3 and 4 are for internal use only.

8.20.3 Dual Port 100 GbE adapter cables and connectors


Table 8-3 describes the type of connectors and cables and their minimum standards for use with the Dual Port 100 GbE adapters. These cable and transceiver options are orderable from IBM or third-party providers, but they must meet the minimum requirements.

Table 8-3 Cables

Connector	Cable type	Minimum standard
QSFP28 (SR4) – Transceiver (IBM feature)	Multi-fiber push on connectors (MPO) fabric	<ul style="list-style-type: none"> ▶ OM3 (up to 70 m) ▶ OM4 (up to 100 m) ▶ SFF-8436 transceiver specification: <ul style="list-style-type: none"> – IEC60825-1 product safety specification. – Operational temperature should be 0 °C - 70 °C. However, in some cases the temperature can reach to 85 °C.
QSFP28 (SRBD) – Transceiver	LC fabric optical: OM3/OM4	
Active Optical Cable (AOC)	AOC: Ribbon optical fiber cable	
Direct Attached Copper (DAC)	DAC: Twin-ax copper cable (up to 2 m)	

Planning for 100 Gbps Ethernet adapters:

- ▶ [FlashSystem 7300](#)
- ▶ [FlashSystem 9500](#)
- ▶ [SAN Volume Controller SV3](#)



Implementing a storage monitoring system

Monitoring in a storage environment is crucial, and is part of what often is called *storage governance*.

With a robust and reliable storage monitoring system, you can realize significant financial savings and minimize pain in your operation by monitoring and predicting usage bottlenecks in your virtualized storage environment.

It is also possible to use the data that is collected from monitoring to create strategies and apply configurations to improve performance, tuning connections, and tools usability.

This chapter provides suggestions and the basic concepts about how to implement a storage monitoring system for IBM Storage Virtualize by using their specific functions or external IBM tools.

This chapter includes the following topics:

- ▶ “Generic monitoring” on page 552
- ▶ “Performance monitoring” on page 557
- ▶ “Capacity monitoring” on page 588
- ▶ “Creating alerts for IBM Storage Control and IBM Storage Insights” on page 606
- ▶ “Health monitoring” on page 613
- ▶ “Important performance metrics” on page 620
- ▶ “Performance diagnostic information” on page 628
- ▶ “Metro Mirror and Global Mirror monitoring” on page 632

9.1 Generic monitoring

With IBM Storage Virtualize, you can implement generic monitoring by using specific functions that are integrated with the product itself without adding any external tools or cost.

9.1.1 Monitoring by using the management GUI

The management GUI is the primary tool that is used to service your system. You can regularly monitor the status of the system by using the management GUI. If you suspect a problem, use the management GUI first to diagnose and resolve the problem.

Use the views that are available in the management GUI to verify the status of the system, the hardware devices, the physical storage, and the available volumes. Selecting **Monitoring** → **Events** provides access to all problems that exist on the system. Select the **Recommended Actions** filter to display the most important events that must be resolved.

If a service error code exists for the alert, you can run a fix procedure that helps you resolve the problem. These fix procedures analyze the system and provide more information about the problem. These actions also ensure that the required changes do not cause volumes to be inaccessible to the hosts and automatically perform configuration changes that are required to return the system to its optimum state.

If any interaction is required, fix procedures suggest actions to take and guide you through those actions that automatically manage the system where necessary. If the problem is fixed, the alert is excluded.

9.1.2 Call Home with email notifications

Call Home connects your system to service representatives who can monitor issues and respond to problems efficiently and quickly to keep your system running. The Call Home feature transmits operational and event-related data to you and IBM through a Simple Mail Transfer Protocol (SMTP) server or cloud services connection through Representational State Transfer (RESTful) APIs.

The system either sends notifications through an SMTP email server to IBM Support or through a RESTful application programming interface (API). If Call Home with cloud services is enabled, as depicted in Figure 9-2 on page 553. Multiple email recipients can be added to receive notifications from the storage system. You can also customize the type of information that is sent to each recipient, as shown in Figure 9-1.

Server IP or Domain	Server Port	Status	Username	Password
smtp.com	587	Active	ibmstoragevirtuali

Email Address	Error	Warning	Info	Inventory
support@company.com	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
user@company.com	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Figure 9-1 Email users showing customizable notifications

Note: With 8.6, a cloud based E-Mail SMTP server is supported. For security reasons, an app-specific password may need to be set on the provider site for authorization.

Call Home with cloud services is the industry standards for transmitting data through web services. With the introduction of this cloud based delivery mechanism, IBM has been implemented a method to messages to the IBM call home servers, which is not affected by spam filters or other technologies preventing IBM from receiving the call home messages.

Cloud Call Home is a key building block which IBM will continue to enhance by providing more predictive support. These features will not be available to clients using email call home. Call Home with cloud services is the preferred call home method for the IBM Storage Virtualize products, to ensure the most optimal end-to-end reliable delivery mechanism. Call Home with an active instance of E-Mail and Cloud services are shown in Figure 9-2.

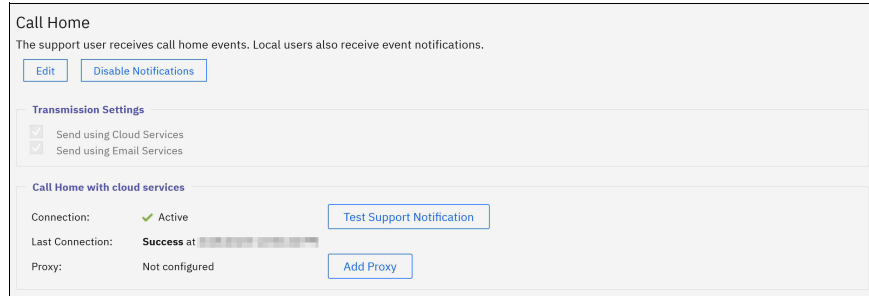


Figure 9-2 Call Home with cloud services configuration window

Note: Web proxy servers can be either configured in the GUI or using the CLI.

More details about Call Home can be found at [IBM Storage Virtualize Products Call Home and Remote Support Overview](#).

Starting with 8.6, the IBM Storage Virtualize Call Home functionality offers the possibility to add a cloud based e-mail provider.

9.1.3 Simple Network Management Protocol notification

Simple Network Management Protocol (SNMP) is a standard protocol for managing networks and exchanging messages. The system can send SNMP messages that notify personnel about an event. You can use an SNMP manager to view the SNMP messages that are sent by the IBM Storage Virtualize system.

The management information base (MIB) file describes the format of the SNMP messages that are sent by the system. Use this MIB file to configure a network management program to receive SNMP event notifications that are sent from an IBM Storage Virtualize system. This MIB file is suitable for use with SNMP messages from all versions of IBM Storage Virtualize.

For more information about the MIB file for the IBM Storage Virtualize system, see [Management Information Base file for SNMP](#).

Note: Since 8.6, the MIB file can be downloaded from the GUI.


```
Apr 12 09:20:01 ITS0-cluster sshd[5098]: Accepted keyboard-interactive/pam for
superuser from a.b.c.d port 39842 ssh2
Apr 12 09:20:02 ITS0-cluster sshd[5098]: pam_unix(sshd:session): session opened
for user superuser by (uid=0)
```

9.1.5 Monitoring by using quotas and alerts

In an IBM Storage Virtualize system, the space usage of storage pools and thin-provisioned or compressed virtual disks (VDisks) can be monitored by setting some specific quota alerts. These alerts can be defined in the management GUI and by using the command-line interface (CLI).

Storage pool

On a storage pool level, an integer defines a threshold at which a warning is generated. The warning is generated the first time that the threshold is exceeded by the used-disk capacity in the storage pool. The threshold can be specified with a percentage (see Figure 9-6) or size (see Example 9-3) value.

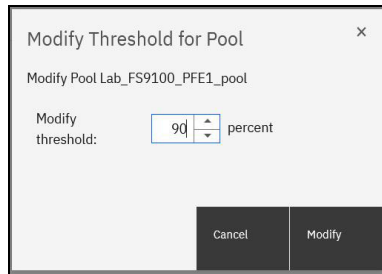


Figure 9-6 Pool threshold

Example 9-3 Threshold specified as a size

```
IBM_FlashSystem:FSxxx:superuser>chmdiskgrp -warning 1 -unit tb ITS0
```

VDisk

At the VDisk level, a warning is generated when the used disk capacity on the thin-provisioned or compressed copy first exceeds the specified threshold. The threshold can be specified with a percentage (see Figure 9-7) or size (see Example 9-4) value.

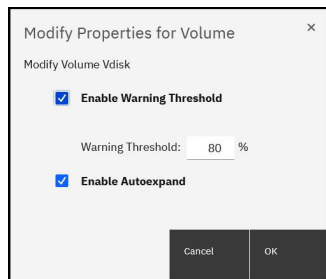


Figure 9-7 VDisk threshold

Example 9-4 Threshold that is specified as a value

```
IBM_FlashSystem:FSxxx:superuser>svctask chvdisk -copy 0 -warning 1 -unit gb ITS0
```

Note: You can specify a `disk_size` integer, which defaults to megabytes (MB) unless the `-unit` parameter is specified, or you can specify `disk_size%`, which is a percentage of the volume size. If both copies are thin-provisioned and the `-copy` parameter is not specified, the specified `-warning` parameter is set on both copies. To disable warnings, specify 0 or 0%. The default value is 0. This option is not valid for thin or compressed volumes in a data reduction pool (DRP).

9.2 Performance monitoring

The ability to collect historical performance metrics is as essential to properly monitoring and managing storage subsystems as it is for IBM Storage Virtualize systems. During troubleshooting and performance tuning, the historical data can be used as a baseline to identify unwanted behavior.

The next sections describe the performance analysis tools that are integrated with IBM Storage Virtualize systems. Also described are the IBM external tools that are available to collect performance statistics to allow historical retention.

Performance statistics are useful to debug or prevent some potential bottlenecks, and to make capacity planning for future growth easier.

9.2.1 Onboard performance monitoring

In this section we cover onboard performance monitoring.

Monitoring and Dashboard views

In IBM Storage Virtualize, real-time performance statistics provide short-term status information for your systems. The statistics are shown as graphs in the management GUI.

You can use system statistics to monitor the aggregate workload of all the volumes, interfaces, and managed disks (MDisks) that are used on your system. The workload can be displayed in megabytes per second (MBps) or input/output operations per second (IOPS). Additionally, read/write latency metrics can be displayed for volumes and MDisks.

You can also monitor the overall CPU usage for the system. These statistics also summarize the recent performance health of the system in almost real time.

You can monitor changes to stable values or differences between related statistics, such as the latency between volumes and MDisks. Then, these differences can be further evaluated by performance diagnostic tools.

With system-level statistics, you can also view quickly the aggregate bandwidth of volumes, interfaces, and MDisks. Each of these graphs displays the current bandwidth in megabytes per second and a view of bandwidth over time.

Each data point can be accessed to determine its individual bandwidth usage and evaluate whether a specific data point might represent performance impacts. For example, you can monitor the interfaces such as for Fibre Channel (FC), internet Small Computer System Interface (iSCSI), Serial Attached SCSI (SAS) or IP Replication to determine whether the interface rate is different from the expected rate.

You can also select node-level statistics, which can help you determine the performance impact of a specific node. As with system statistics, node statistics help you to evaluate whether the node is operating within a normal range of performance metrics.

The CPU utilization graph shows the current percentage of CPU usage and specific data points on the graph that show peaks in utilization. If compression is being used, you can monitor the amount of CPU resources that are being used for compression and the amount that is available to the rest of the system. The Compression CPU utilization chart is not relevant for DRP compression.

The Volumes and MDisks graphs on the Performance window show four metrics: Read, Write, Read latency, and Write latency. You can use these metrics to help determine the overall performance health of the volumes and MDisks on your system. Consistent unexpected results can indicate errors in configuration, system faults, connectivity issues, or workload-specific behavior.

Each graph represents 5 minutes of collected statistics, which are updated every 5 seconds. They also provide a means of assessing the overall performance of your system, as shown in Figure 9-8.



Figure 9-8 Monitoring/Performance overview

Note: Starting with 8.5, the latency metrics in the Monitoring view and the Dashboard view switch dynamically between milliseconds (ms) and microseconds (µs) and the graph scales as needed. This function aids in monitoring submillisecond response times of highly performant systems.

You can select the workload metric that you want to be displayed, as shown in Figure 9-9 on page 559.

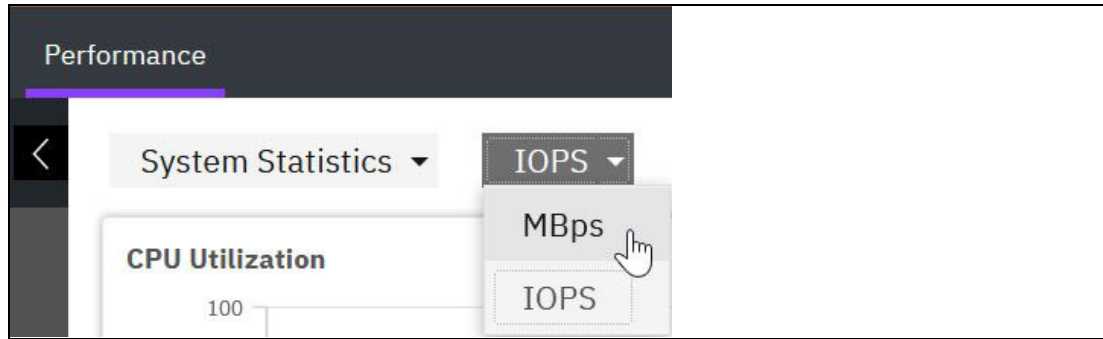


Figure 9-9 Workload metrics

You can also obtain a quick overview of the performance in the GUI dashboard: **System** → **Dashboard**, as shown in Figure 9-10.

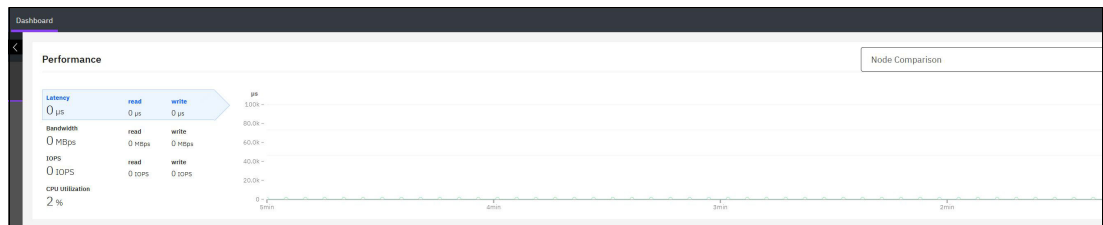


Figure 9-10 Management GUI Dashboard view

Command-line interface

The `lsnodestats`, `lsnodecanisterstats`, and `lssystemstats` commands continue to report latency in milliseconds only. The latency was reported as an integer value before code level 8.5, but now it is reported with three decimal places of granularity. This format makes it possible to monitor variations in response time that are less than 1 ms, as shown in Example 9-5.

Example 9-5 Latency reported in milliseconds (ms) with microsecond (μ s) granularity

```
IBM_FlashSystem:FS9xxx:superuser>lssystemstats -history drive_w_ms|grep -E
2306220704
230622070402 drive_w_ms 0.000
230622070407 drive_w_ms 0.000
230622070412 drive_w_ms 0.000
230622070417 drive_w_ms 0.000
230622070422 drive_w_ms 0.066
230622070427 drive_w_ms 0.000
230622070432 drive_w_ms 0.000
230622070437 drive_w_ms 0.000
230622070442 drive_w_ms 0.000
230622070447 drive_w_ms 0.000
230622070452 drive_w_ms 0.000
230622070457 drive_w_ms 0.000
```

REST API explorer

The IBM Storage Virtualize Representational State Transfer (REST) model application programming interface (API) consists of command targets that are used to create, read, update, and delete system resources. These command targets allow command parameters to pass through unedited to the Storage Virtualize command-line interface, which handles parsing parameter specifications for validity and error reporting. Use Hypertext Transfer Protocol Secure (HTTPS) to successfully communicate with the RESTful API server.

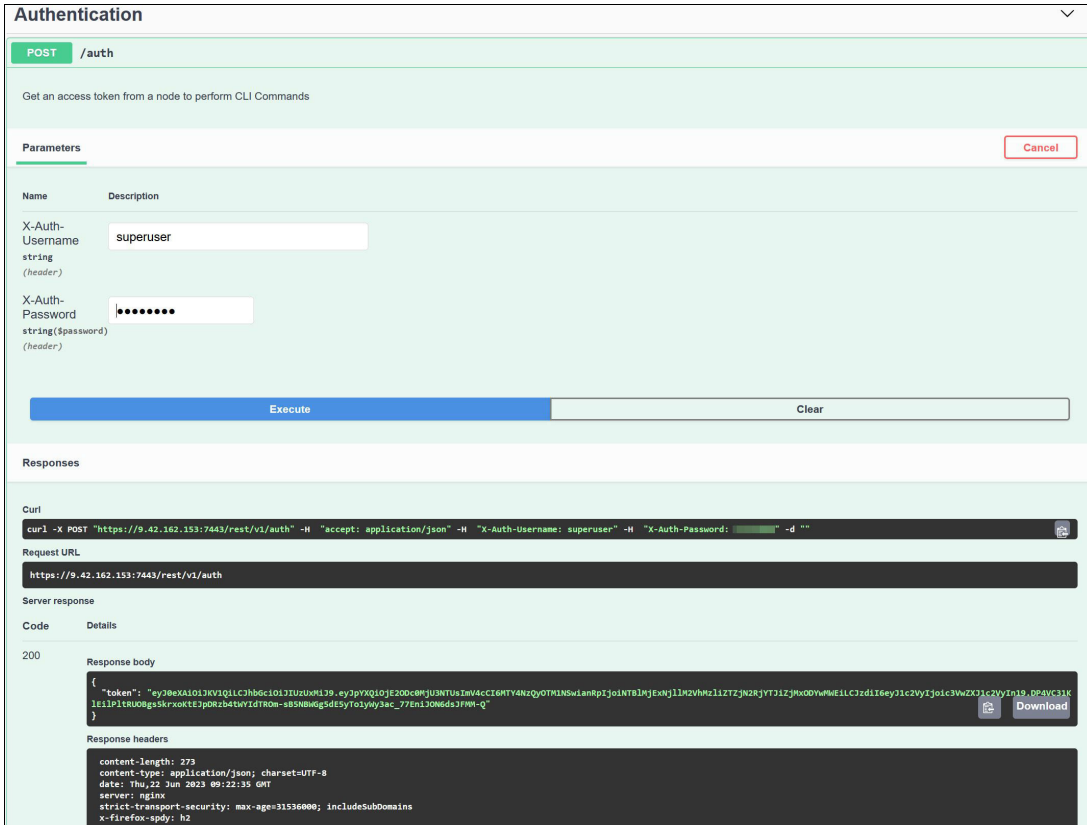
REST API client is used to perform a query of the REST API. The results would be returned to the console environment of the REST API client, as shown in Example 9-6.

Example 9-6 REST API clients

```
curl
perl
powershell + InvokeRestMethod cmdlet
python + requests library
```

An OpenAPI explorer guide is embedded within the IBM Storage Virtualize product and can be reached by using the following address: <https://<system-ip>:7443/rest/explorer/>.

The REST API explorer can retrieve the X-Auth-Token for the provided credentials (see Figure 9-11) and run a REST API query and display the results in a browser (see Figure 9-12 on page 561).



The screenshot displays the 'Authentication' interface in the REST API Explorer. It shows a POST request to the endpoint '/auth' with the following parameters:

- X-Auth-Username:** superuser
- X-Auth-Password:** [REDACTED]

The response is a 200 status code with a JSON body containing an X-Auth-Token. The token is a long alphanumeric string. The response headers include:

- content-length: 273
- content-type: application/json; charset=UTF-8
- date: Thu, 22 Jun 2023 09:22:35 GMT
- server: nginx
- strict-transport-security: max-age=31536000; includeSubdomains
- x-firefox-spy: h2

Figure 9-11 Authentication in REST API Explorer: Token displayed in the Response body

Monitoring Easy Tier by using the GUI

Since version 8.3.1, the GUI includes various reports and statistical analysis that can be used to understand what Easy Tier movement, activity, and skew is present in a storage pool. These reports replace the old IBM Storage Tier Advisor Tool (STAT) and STAT Charting Tool.

Unlike previous versions, where you were required to download the necessary log files from the system and upload them to the STAT tool, from version 8.3.1 onwards, the system continually reports the Easy Tier information, so the GUI always displays the most up-to-date information.

Accessing Easy Tier reports

To show the Easy Tier Reports window, select **Monitoring** → **Easy Tier Reports**.

Note: If the system or Easy Tier was running for less than 24 hours, no data might be available to display.

The Reports window features the following views that can be accessed by using the tabs at the top of the window, which are described next:

- ▶ Data Movement
- ▶ Tier Composition
- ▶ Workload Skew Comparison

Data Movement report

The Data Movement report shows the amount of data that was moved in a specific period. You can change the period by using the drop-down selection at the right side (see Figure 9-13).



Figure 9-13 Easy Tier Data Movement window

The report breaks down the type of movement, which is described in terms of the internal Easy Tier extent movement types (see 4.7.2, “Easy Tier definitions” on page 299).

To aid your understanding and remind you of the definitions, click **Movement Description** to view the information window (see Figure 9-14).

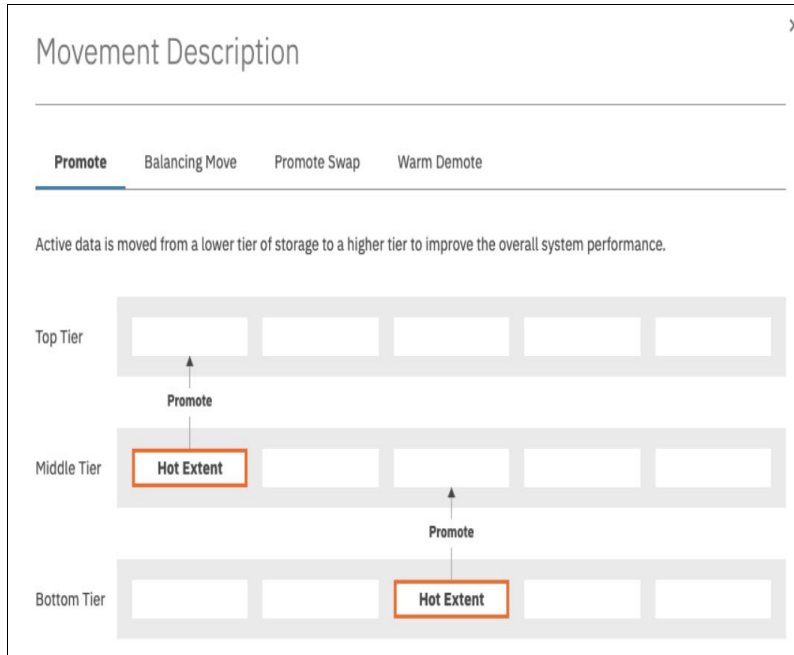


Figure 9-14 Easy Tier Movement description window

Tier Composition report

Important: If you are regularly seeing “warm demote” in the movement data, consider increasing the amount of hot tier that is available. A warm demote suggests that an extent is hot, but not enough capacity or Overload Protection was triggered in the hot tier.

The Tier Composition window (see Figure 9-15 on page 564) shows how much data in each tier is active versus inactive. In an ideal case, most of your active data is in the hot tier alone. In most cases, the active data set cannot fit in only the hot tier; therefore, expect to also see active data in the middle tier. Here we can see that most of the data in the middle tier is inactive or the workload does meet the criteria for Easy Tier optimization.

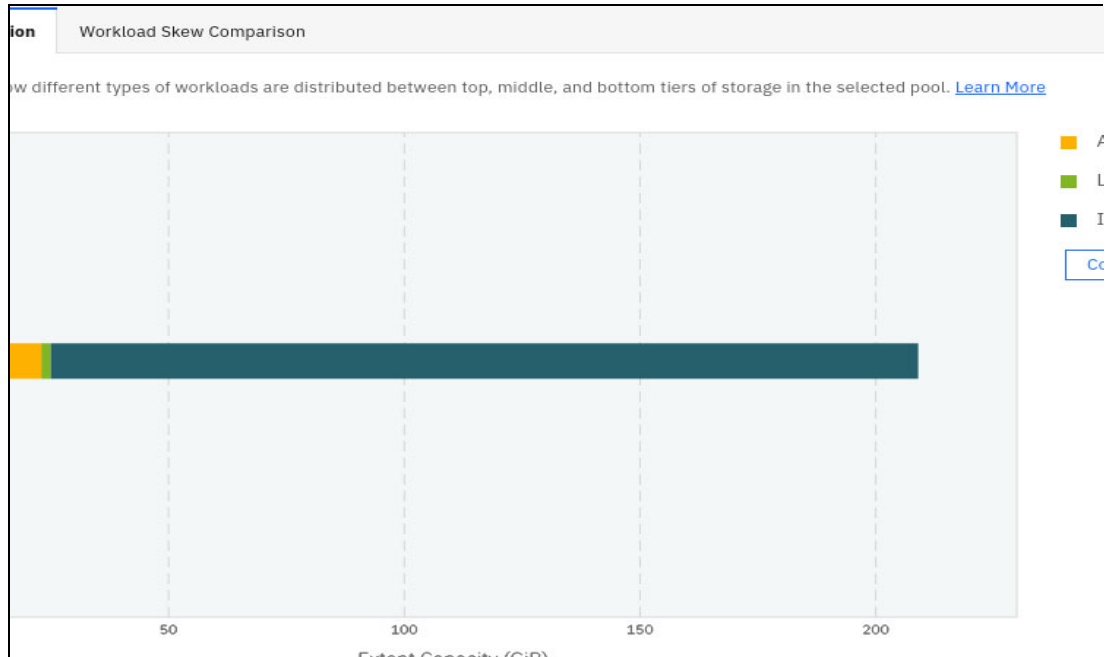


Figure 9-15 Easy Tier Composition report window

If all active data can fit in the hot tier, you see the best possible performance from the system. *Active large* is data that is active but is being accessed at block sizes larger than the 64 KiB for which Easy Tier is optimized. This data is still monitored and can contribute to “expanded cold demote” operations.

The presence of any active data in the cold tier (regularly) suggests that you must increase the capacity or performance in the hot or middle tiers.

In the same way as with the Data Movement window, you can click **Composition Description** to view the information for each composition type (see Figure 9-16).

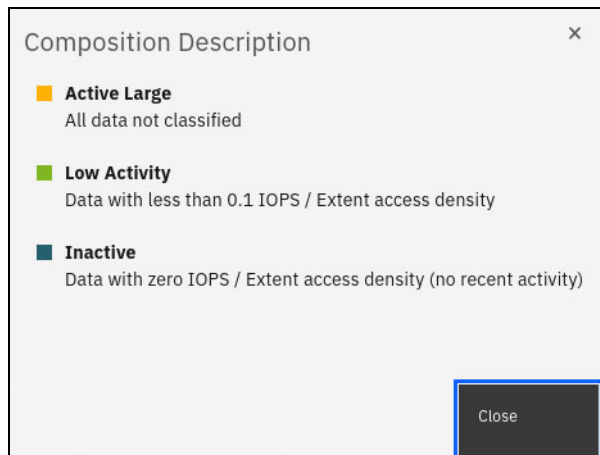


Figure 9-16 Easy Tier Composition Description

Workload Skew Comparison report

The Workload Skew Comparison report plots the percentage of the workload against the percentage of capacity. The skew shows a good estimate for how much capacity is required in the top tier to realize the most optimal configuration that is based on your workload.

Tip: The skew can be viewed when the system is in measuring mode with a single tier pool to help guide the recommended capacity to purchase that can be added to the pool in a hot tier.

A highly skewed workload (the line on the graph rises sharply within the first percentage of capacity) means that a smaller proportional capacity of hot tier is required. A low skewed workload (the line on the graph rises slowly and covers a large percentage of the capacity) requires more hot tier capacity, which you should consider as a good performing middle tier when you cannot configure enough hot tier capacity (see Figure 9-17).



Figure 9-17 Workload skew: Single tier pool

In the first example that is shown in Figure 9-17, you can clearly see that this workload is highly skewed. This single-tier pool uses less than 5% of the capacity, but is performing 99% of the workload in terms of IOPS and MBps.

This result is a prime example for adding a small amount of faster storage to create a “hot” tier and improve overall pool performance (see Figure 9-18).

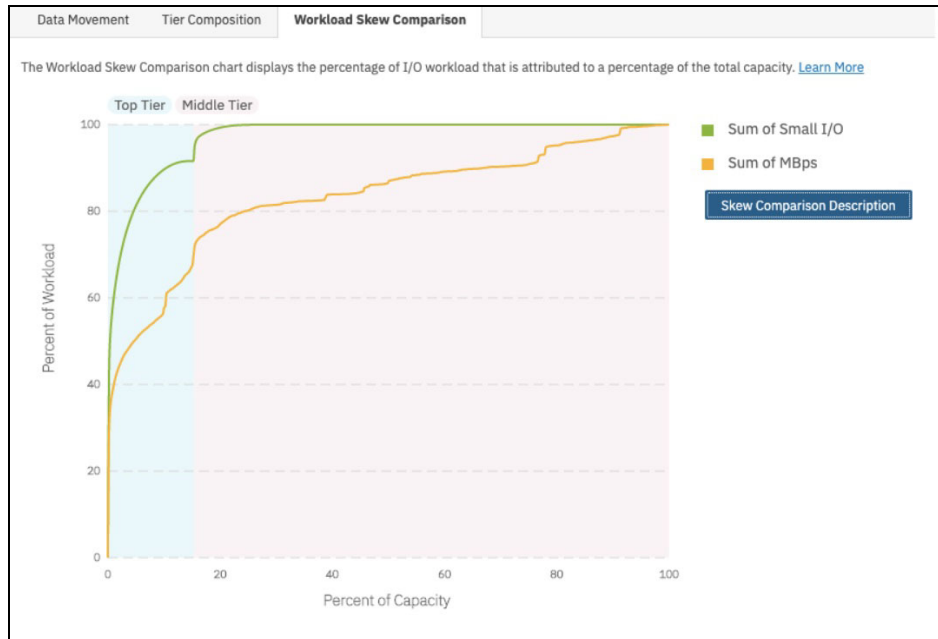


Figure 9-18 Workload skew: Multitier configuration

In this second example that is shown in Figure 9-18, the system is configured as a multitier pool, and Easy Tier optimized the data placement for some time. This workload is less skewed than in the first example, with almost 20% of the capacity performing up to 99% of the workload.

Here again, it might be worth considering increasing the amount of capacity in the top tier because approximately 10% of the IOPS workload is coming from the middle tier and can be optimized to reduce latency.

The graph that is shown in Figure 9-18 also shows the split between IOPS and MBps. Although the middle tier is not handling much of the IOPS workload, it is providing a reasonably large proportion of the MBps workload.

In these cases, ensure that the middle tier can manage good large block throughput. A case might be made for further improving performance by adding some higher throughput devices as a new middle tier, and demoting the current middle tier to the cold tier; however, this change depends on the types of storage that is used to provide the existing tiers.

Any new configuration with three tiers must comply with the configuration rules regarding the different types of storage that is supported in three-tier configurations (see “Easy Tier mapping to MDisk tier types” on page 307).

If you implemented a new system and see that most of the workload is coming from a middle or cold tier, it might take only a day or two for Easy Tier to complete the migrations after it initially analyzes the system.

If after a few days a distinct bias still exists to the lower tiers, you might want to consider enabling “Accelerated Mode” for a week or so; however, disable this mode after the system reaches a steady state. For more information, see “Easy Tier acceleration” on page 317.

9.2.2 Performance monitoring with IBM Spectrum Control

IBM Spectrum Control is an on-premises storage management, monitoring, and reporting solution. It uses the metadata that it collects about vendors' storage devices to provide services, such as custom alerting, analytics, and replication management. Both IBM Spectrum Control and IBM Storage Insights monitor storage systems fabrics and switches, but IBM Spectrum Control also monitors hypervisors to provide you with unique analytics and insights into the topology of your storage network.

IBM Spectrum Control also provides more granular collection of performance data with 1-minute intervals rather than the 5-minute intervals in IBM Storage Insights or IBM Storage Insights Pro. For more information about IBM Storage Insights, see 9.2.3, "Performance monitoring with IBM Storage Insights" on page 571.

Because IBM Spectrum Control is an on-premises tool, it does not send the metadata about monitored devices off-site, which is ideal for dark shops and sites that do not want to open ports to the cloud.

For more information about the capabilities of IBM Spectrum Control, see this [Product overview](#).

For more information about pricing and other purchasing information, see [IBM Spectrum Control](#).

Note: If you use IBM Spectrum Control or manage IBM block storage systems, you can access the no-charge version of IBM Storage Insights. For more information, see [Getting Started with IBM Storage Insights](#).

IBM Spectrum Control offers several reports that you can use to monitor IBM Storage Virtualize 8.6 systems to identify performance problems. IBM Spectrum Control provides improvements to the web-based user interface that is designed to offer easy access to your storage environment.

IBM Spectrum Control provides a large amount of detailed information about IBM Storage Virtualize 8.6 systems. This sections provide some basic suggestions about what metrics must be monitored and analyzed to debug potential bottleneck problems. It also covers alerting profiles and their thresholds that are considered important for detecting and resolving performance issues.

For more information about the installation, configuration, and administration of IBM Spectrum Control (including how to add a storage system), see the following web pages:

- ▶ [Supported Storage Products in IBM Spectrum Control 5.4.x](#)
- ▶ [IBM Spectrum Control - Installing](#)

Note IBM Spectrum Control 5.3.x has reached end of support. 5.4.0 or higher is recommended for monitoring IBM Storage Virtualize systems.

IBM Spectrum Control Dashboard

The IBM Spectrum Control Dashboard gives you a status overview of all monitored resources and identifies potential problem areas in a storage environment. It represents the following information:

- ▶ Condition and usage of resources
- ▶ Entities that use storage on those resources
- ▶ Number and status of unacknowledged alert conditions that are detected on the monitored resources
- ▶ Most active storage systems in your environment.

Figure 9-19 shows the IBM Spectrum Control Dashboard.

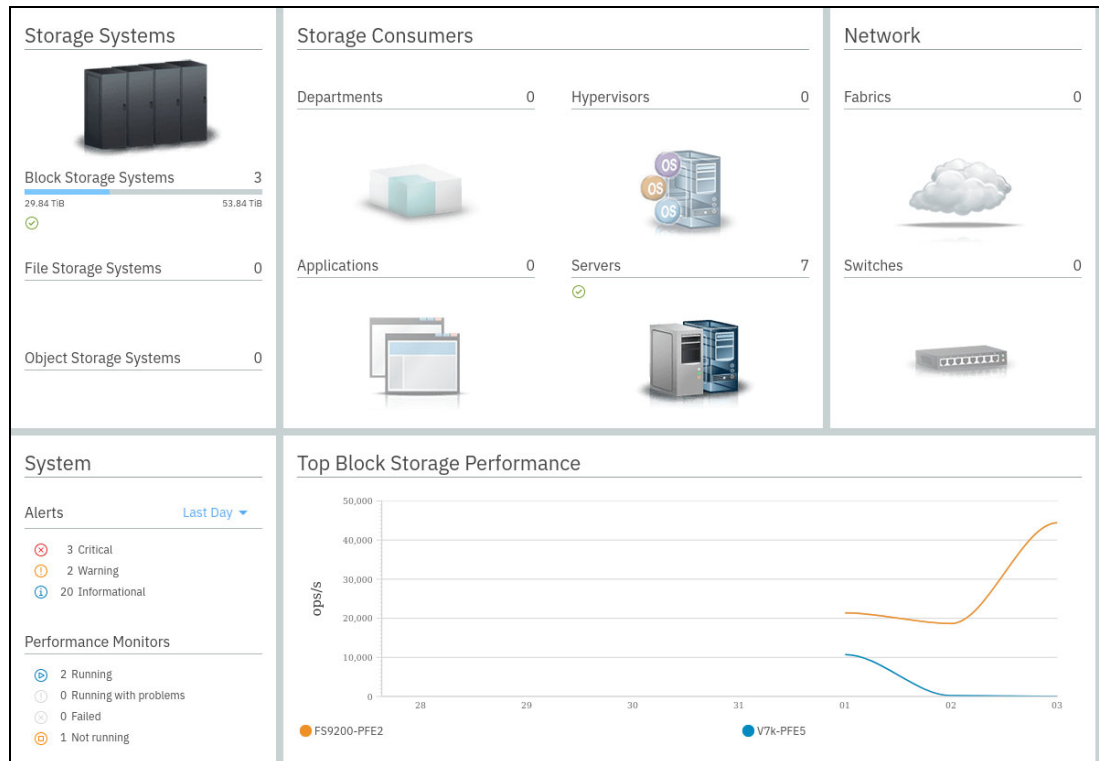


Figure 9-19 IBM Spectrum Control Dashboard

Key performance indicators

IBM Spectrum Control provides *key performance indicators* (in earlier releases, *Best Practice Performance Guidelines*) for the critical monitoring metrics. These guidelines do *not* represent the maximum operating limits of the related components. Instead, they suggest limits that are selected with an emphasis on maintaining a stable and predictable performance profile.

The key performance indicators GUI of IBM Spectrum Control (see Figure 9-20 on page 569) displays by default the last 24 hours from the active viewing time and date. Selecting an individual element from the chart overlays the corresponding 24 hours for the previous day and seven days before. This display allows for an immediate historical comparison of the respective metric. The day of reference can also be changed to allow historical comparison of previous days.

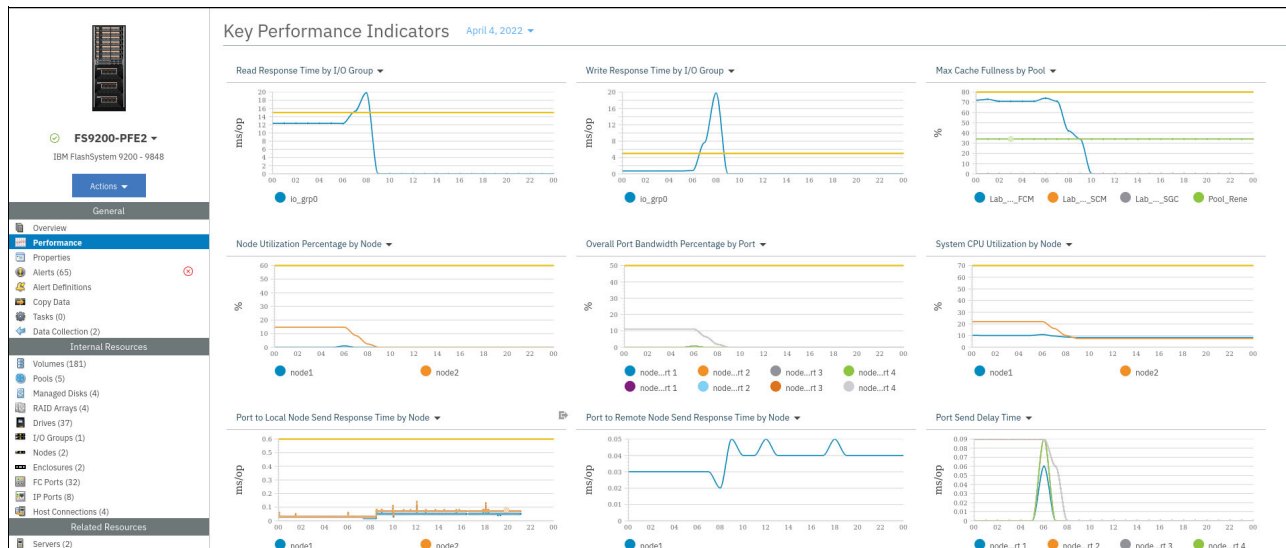


Figure 9-20 Key Performance Indicators

The yellow lines that are shown in Figure 9-20 represent guidelines that were established at the levels that allow for a diverse set of workload characteristics while maintaining a stable performance profile. The other lines on each chart represent the measured values for the metric for the resources on your storage system: I/O groups, ports, or nodes.

You can use the lines to compare how close your resources are to potentially becoming overloaded. If your storage system is responding poorly and the charts indicate overloaded resources, you might need to better balance the workload. You can balance the workload between the hardware of the cluster by adding hardware to the cluster or moving some workload to other storage systems.

The charts that are shown in Figure 9-20 show the hourly performance data that is measured for each resource on the selected day. Use the following charts to compare the workloads on your storage system with the following key performance indicators:

- ▶ **Node Utilization Percentage by Node**

The average of the bandwidth percentages of those ports in the node that are actively used for host and MDisk send and receive operations. The average is weighted by port speed and adjusted according to the technology limitations of the node hardware. This chart is empty for clusters without FC ports (or when no host I/O is going on). Compare the guideline value for this metric, for example, 60% utilization, with the measured value from your system.

- ▶ **Overall Port Bandwidth Percentage by Port**

The percentage of the port bandwidth that is used for receive and send operations. This value is an indicator of port bandwidth usage that is based on the speed of the port. The guideline value is 50%. Compare the guideline value for this metric with the values that are measured for the switch ports. A cluster can have many ports. The chart shows only the eight ports with the highest average bandwidth.

- ▶ **Port-to-Local Node Send Response Time by Node**

The average number of milliseconds to complete a send operation to another node that is in the local cluster. This value represents the external response time of the transfers. Compare the guideline value for this metric, for example, 0.6 ms/op, with the measured value from your system.

- ▶ **Port-to-Remote Node Send Response Time by Node**
 The average number of milliseconds it takes to complete a send operation to a node in the remote cluster, and the average number of milliseconds it takes to complete a receive operation from a node in the remote cluster. This value represents the external response time of the transfers. A guideline value is not available for this metric because response times for copy-service operations can vary widely. You can correlate the response times to identify discrepancies between the response times for the different nodes.
- ▶ **Read Response Time by I/O Group**
 The average number of milliseconds to complete a read operation. Compare the guideline value for this metric, for example, 15 ms/op, with the measured value from your system.
- ▶ **Max Cache Fullness by Pool**
 The maximum amount of the lower cache that the write cache partitions on the nodes that manage the pool are using for write operations. If the value is 100%, one or more cache partitions on one or more pools is full. The operations that pass through the pools with full cache partitions are queued and I/O response times increase for the volumes in the affected pools. Available in IBM Storage Virtualize 7.3 or later.
- ▶ **Write Response Time by I/O Group**
 The average number of milliseconds to complete a write operation. Compare the guideline value for this metric, for example, 5 ms/op, with the measured value from your system.
- ▶ **Zero Buffer Credit Percentage by Node**
 The amount of time as a percentage that the port was not able to send frames between ports because of insufficient buffer-to-buffer credit. The amount of time value is measured from the last time that the node was reset. In FC technology, buffer-to-buffer credit is used to control the flow of frames between ports. Information about zero buffer credit is collected and analyzed only for 8 Gbps FC ports. The guideline value is 20%.
- ▶ **Port Send Delay Time**
 The average number of milliseconds of delay that occurs on the port for each send operation. The reason for these delays might be a lack of buffer credits. A guideline value is not available for this metric because delay times can vary significantly depending on configuration and usage. Correlate the delay times to identify discrepancies between the ports' delay times and any spikes that might correlate with the time of any reported performance problems.

Note: *System CPU Utilization by Node* was removed from this view and replaced by *Max Cache Fullness by Pool*. Additionally, either *Zero Buffer Credit Percentage by Node* or *Port Send Delay Time* (not both) are shown depending on the model of the system.

Figure 9-21 shows an example of the Write Response Time by I/O Group, which exceeded the best practice limit (yellow line). The dropdown menu provides further options.

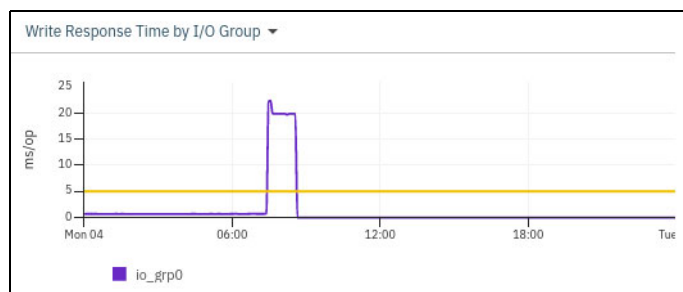


Figure 9-21 Write response Time by I/O Group > 5 ms

Note: The guidelines are not strict thresholds. They are derived from real field experience for many configurations and workloads. When appropriate, these guidelines can be adopted as alert thresholds within an alert policy.

9.2.3 Performance monitoring with IBM Storage Insights

IBM Storage Insights is an off-premises IBM Cloud service that provides cognitive support capabilities, monitoring, and reporting for storage systems and switches and fabrics. It is an IBM Cloud service, so getting started is simple, and upgrades are handled automatically.

By using the IBM Cloud infrastructure, IBM Support can monitor your storage environment to help minimize the time to resolution of problems and collect diagnostic packages without requiring you to manually upload them. This support experience, from environment to instance, is unique to IBM Storage Insights and transforms how and when you get help.

IBM Storage Insights is a software as a service (SaaS) offering with its core running over IBM Cloud. IBM Storage Insights provides an unparalleled level of visibility across your storage environment to help you manage complex storage infrastructures and make cost-saving decisions. IBM Storage Insights combines proven IBM data management leadership with IBM analytics leadership from IBM Research and a rich history of storage management expertise with a cloud delivery model, enabling you to take control of your storage environment.

As a cloud-based service, IBM Storage Insights enables you to deploy quickly and save storage administration time while optimizing your storage. IBM Storage Insights also helps automate aspects of the support process to enable faster resolution of issues. IBM Storage Insights optimizes storage infrastructure by using cloud-based storage management and a support platform with predictive analytics.

With IBM Storage Insights, you can optimize performance and tier your data and storage systems for the right combination of speed, capacity, and economy. IBM Storage Insights provides comprehensive storage management and helps to keep costs low, and might prevent downtime and loss of data or revenue.

Here are the key features for IBM Storage Insights:

- ▶ Rapid results when you need them.
- ▶ A single-pane view across your storage environment.
- ▶ Performance analyses at your fingertips.
- ▶ Valuable insights from predictive analytics.
- ▶ Two editions that meet your needs.
- ▶ Simplified, comprehensive, and proactive product support.

Note: As a best practice, use IBM Storage Insights or IBM Spectrum Control for a better user experience.

IBM Storage Insights versus IBM Storage Insights Pro

The free version is called *IBM Storage Insights* and provides a unified view of a storage environment with diagnostic event information, an integrated support experience, and key capacity and performance metrics. This includes component health overviews for switches and fabrics with key performance metrics. IBM Storage Insights is available at no cost to IBM Storage Insights Pro subscribers and owners of IBM block storage systems who sign up.

The capacity-based, subscription version is called *IBM Storage Insights Pro* and includes all the features of IBM Storage Insights plus a more comprehensive view of the performance, capacity, and health of storage resources. This includes key health, performance, and diagnostic information for switches and fabrics. It also helps you reduce storage costs and optimize your data center by providing features like intelligent capacity planning, storage reclamation, storage tiering, and advanced performance metrics. The storage systems that you can monitor are expanded to include IBM file, object, software-defined storage (SDS) systems, and non-IBM block and file storage systems, such as Dell EMC storage systems.

In both versions, when problems occur on your storage, you can get help to identify and resolve those problems and minimize potential downtime, where and when you need it.

A table about the difference of all features is documented in [IBM documentation - Storage Insights vs Storage Insights Pro](#).

IBM Storage Insights for IBM Spectrum Control

IBM Storage Insights for IBM Spectrum Control is an IBM Cloud service that can help you predict and prevent storage problems before they impact your business. It is complementary to IBM Spectrum Control and available at no additional cost if you have an active license with a current subscription and a support agreement for IBM Virtual Storage Center, IBM Spectrum Storage Suite, or any edition of IBM Spectrum Control.

As an on-premises application, IBM Spectrum Control does not send the metadata about monitored devices off-site, which is ideal for dark shops and sites that do not want to open ports to the cloud. However, if your organization allows for communication between its network and the cloud, you can use IBM Storage Insights for IBM Spectrum Control to transform your support experience for IBM block storage.

IBM Storage Insights for IBM Spectrum Control and IBM Spectrum Control work together to monitor your storage environment. Here is how IBM Storage Insights for IBM Spectrum Control can transform your monitoring and support experience:

- ▶ Open, update, and track IBM Support tickets easily for your IBM block storage devices.
- ▶ Get hassle-free log collection by allowing IBM Support to collect diagnostic packages for devices so that you do not have to.
- ▶ Use Call Home to monitor devices, get best practice recommendations, and filter events to quickly isolate trouble spots.
- ▶ Use IBM Support to view the current and historical performance of your storage systems and help reduce the time-to-resolution of problems.

You can use IBM Storage Insights for IBM Spectrum Control if you have an active license with a current subscription and a support agreement for an IBM Spectrum Control license. If your subscription and support lapses, you are no longer eligible for IBM Storage Insights for IBM Spectrum Control. To continue using IBM Storage Insights for IBM Spectrum Control, renew your IBM Spectrum Control license. You can also choose to subscribe to IBM Storage Insights Pro.

To compare the features of IBM Spectrum Control and IBM Storage Insights for IBM Spectrum Control, check out the [Feature comparison](#).

You can upgrade IBM Storage Insights to IBM Storage Insights for IBM Spectrum Control if you have an active license for IBM Spectrum Control. For more information, see [IBM Storage Insights Registration](#), choose the option for IBM Spectrum Control, and follow the prompts.

IBM Storage Insights for IBM Spectrum Control does not include the service-level agreement (SLA) for IBM Storage Insights Pro. Terms and conditions for IBM Storage Insights for IBM Spectrum Control are available at [Cloud Services Terms](#).

IBM Storage Insights, IBM Storage Insights Pro, and IBM Storage Insights for IBM Spectrum Control show some similarities, but the following differences exist:

- ▶ IBM Storage Insights is an off-premises IBM Cloud service that is available at no extra charge if you own IBM block storage systems. It provides a unified dashboard for IBM block storage systems and switches and fabrics with a diagnostic events feed, a streamlined support experience, and key capacity and performance information.
- ▶ IBM Storage Insights Pro is an off-premises IBM Cloud service that is available on subscription and expands the capabilities of IBM Storage Insights. You can monitor IBM file, object, and SDS systems, and non-IBM block and file storage systems, such as Dell or EMC storage systems and IBM and non-IBM switches or fabrics.

IBM Storage Insights Pro also includes configurable alerts and predictive analytics that help you to reduce costs, plan capacity, and detect and investigate performance issues. You get recommendations for reclaiming unused storage, recommendations for optimizing the placement of tiered data, capacity planning analytics, and performance troubleshooting tools.

- ▶ IBM Storage Insights for IBM Spectrum Control is similar to IBM Storage Insights Pro in capability, and it is available for no additional cost if you have an active license with a current subscription and support agreement for IBM Virtual Storage Center, IBM Spectrum Storage Suite, or any edition of IBM Spectrum Control.

IBM Spectrum Storage Suite

IBM Spectrum Storage Suite provides unlimited access to the IBM Spectrum Storage software family and IBM Cloud Object Storage software with licensing on a flat, cost-per-TB basis to make pricing easy to understand and predictable as capacity grows. Structured specifically to meet changing storage needs, the suite is ideal for organizations starting out with SDS, and for those organizations with established infrastructures who must expand their capabilities.

The suite includes the following components:

- ▶ IBM Spectrum Control: Integrated data and storage management software that provides monitoring, automation, and analytics for storage.
- ▶ IBM Spectrum Protect: Scalable hybrid cloud data protection for physical file servers, applications, and virtual environments.
- ▶ IBM Spectrum Protect Plus: Complete VM protection and availability that is easy to set up and manage yet scalable for the enterprise.
- ▶ IBM Spectrum Archive: Direct, intuitive, and graphical access to data that is stored in IBM tape drives and libraries.
- ▶ IBM Storage Virtualize: Virtualization of mixed block storage environments to improve efficiency, reduce cost, and simplify management.

- ▶ IBM Spectrum Scale: Advanced storage management of unstructured data for cloud, big data, analytics, objects, and more.
- ▶ IBM Cloud Object Storage: Flexible, scalable, and simple object storage with geo-dispersed enterprise availability and security for hybrid cloud workloads.
- ▶ IBM Spectrum Discover: Modern artificial intelligence (AI) workflow and metadata management software for exabyte-scale file and object storage with hybrid multicloud support.

Because IBM Spectrum Storage Suite contains IBM Spectrum Control, you can deploy IBM Storage Insights for IBM Spectrum Control.

Note: Alerts are a good way to be notified about conditions and potential problems that are detected in your storage. If you use IBM Spectrum Control and IBM Storage Insights for IBM Spectrum Control together to enhance your monitoring capabilities. As a best practice, define alerts in one of the offerings and not both.

By defining all your alerts in one offering, you can avoid receiving duplicate or conflicting notifications when alert conditions are detected.

Implementation and setup of IBM Storage Insights

To use IBM Storage Insights with IBM Storage Virtualize, you must sign up at [IBM Storage Insights Registration](#).

Sign-up process

Consider the following points about the sign-up process:

- ▶ For the sign-up process, you need an IBMid. If you do not have an IBMid, create your IBM account and complete the short form.
- ▶ When you register, specify an owner for IBM Storage Insights. The owner manages access for other users and acts as the main contact.
- ▶ You receive a Welcome email when IBM Storage Insights is ready. The email contains a direct link to your dashboard.

Figure 9-22 shows the IBM Storage Insights registration window.

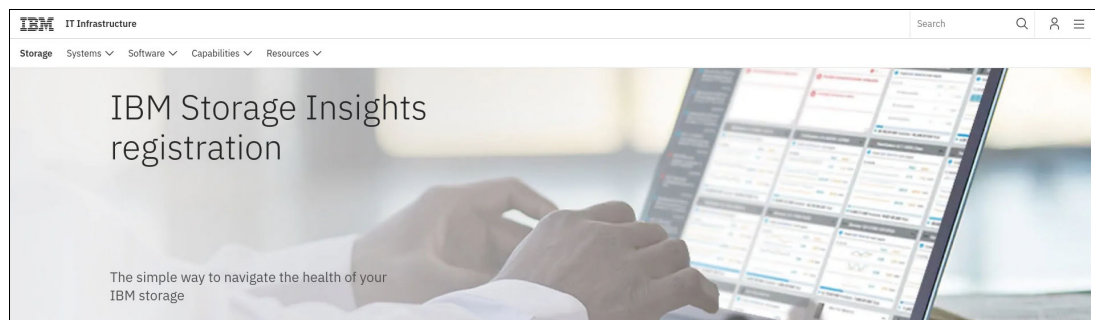


Figure 9-22 IBM Storage Insights registration window

At the dashboard, complete the following steps:

1. Figure 9-23 on page 575 shows the registration options. Select whether you want to register for IBM Storage Insights or IBM Storage Insights for IBM Spectrum Control. For more information about the differences of the IBM Storage Insights software, see "" on page 571.

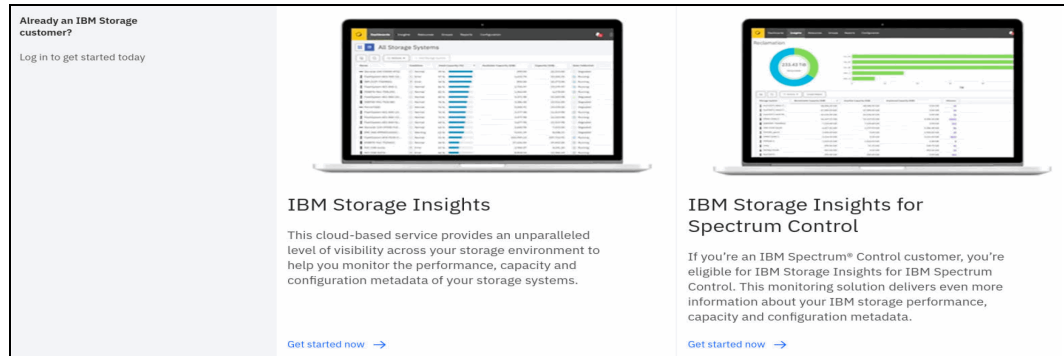


Figure 9-23 IBM Storage Insights or IBM Storage Insights for IBM Spectrum Control registration options

- Figure 9-24 shows the log-in window in the registration process. If you have your credentials, enter your IBMid and proceed to the next window by clicking **Continue**. If you do not have an IBMid, click **Create an IBMid**.

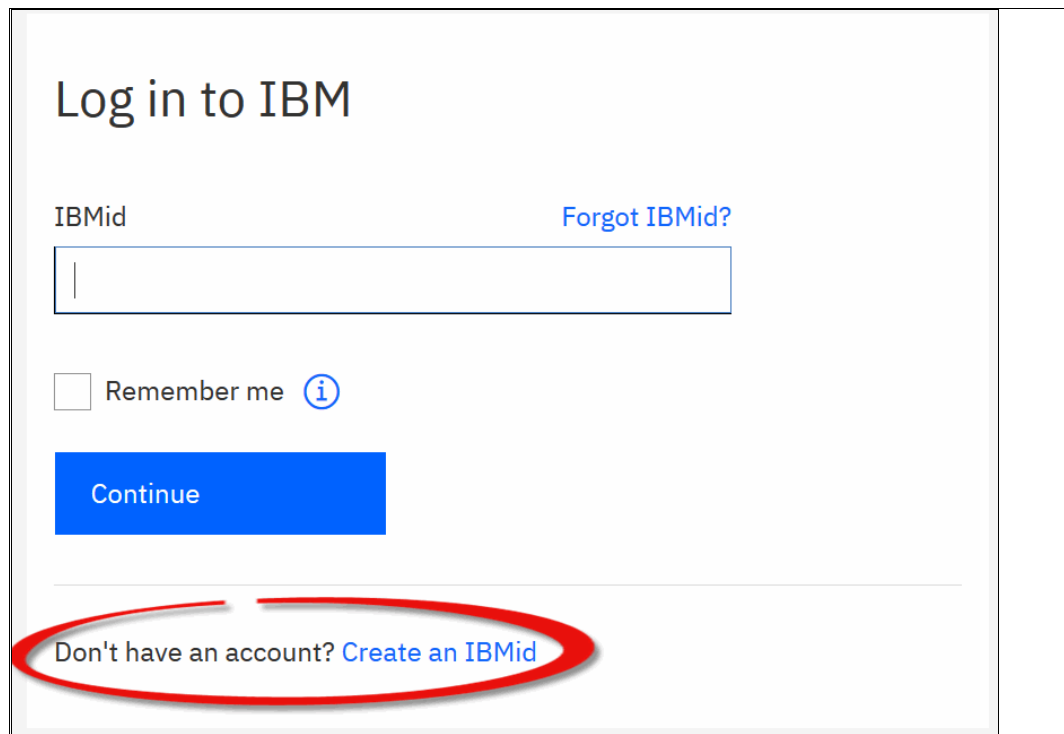


Figure 9-24 Registration login window

3. If you want to create an IBMid, see Figure 9-25 for reference. Provide the following information and click **Next**:

- Email
- First name
- Last name
- Country or region
- Password

Enter the one-time code that was sent to your email address.

Select **by email** checkbox if you want to receive information from IBM to keep you informed about products, services, and offerings. You can withdraw your marketing consent at any time by sending an email to netsupp@us.ibm.com. Also, you can unsubscribe from receiving marketing emails by clicking the unsubscribe link in any email.

For more information about our processing, see the [IBM Privacy Statement](#).

Click **Create account**.

The screenshot shows the IBM account creation interface. On the left, there's a decorative graphic of colored squares. The main content area is titled 'Create your IBM account' and 'Sign up for an IBMid'. It has two sections: '1. Account information' with fields for E-mail, First name, Last name, and Country; and '2. Verify email' with a '7 digit code' field and a 'by email' checkbox. A 'Create account' button is at the bottom.

Figure 9-25 Creating an IBM account

4. In the next window, sign in with your IBM Account and password and review the summary information about your IBMid account privacy, as shown in Figure 9-26 on page 577.

About your IBMid Account Privacy ×

This notice provides information about accessing your IBMid user account (Account). If you have previously been presented with a version of this notice, please refer to "Changes since the previous version of this notice" below for information about the new updates. Updates to the [IBM Privacy Statement](#) since this notice was originally published provide additional information about how your personal information is processed by IBM.

- Changes since the previous version of this notice ▼
- What data does IBM collect? ▼
- Why IBM needs your data ▼
- How your data was obtained ▼
- How IBM uses your data ▼
- How IBM protects your data ▼
- How long we keep your data ▼
- Your rights ▼

I acknowledge that I understand how IBM is using my Basic Personal Data and I am at least 16 years of age.

This document was last updated on 2020-03-05

Cancel
Proceed

Figure 9-26 IBMid account privacy

5. Complete the following information in the IBM Storage Insights registration form (see Figure 9-27). The following items are mandatory:
 - IBM Storage Insights service name (must be unique)
 - IBMid
 - First name and last name

Register

Name your IBM Storage Insights service

Your Storage Insights service name must be unique. We recommend using your company name and another identifying feature, such as a location or department — for example, "Bank ABC North America," or "Bank ABC Europe."

●

This field is required.

Owner details

Register here to view your company's storage trends, stay informed about the health of your storage, and easily access support via the IBM Storage Insights dashboard.

Email address

You may already have an IBMid. If so, enter the email address associated with your ID. If you do not have one, this process will create an IBMid for you.

First name Last name

Privacy and Terms

This data, at any time revocable by you, may be stored by IBM or an affiliate on an international server and used by IBM or an affiliate. IBM may use and store your business contact information wherever we do business, including in the U.S. and other countries outside of Europe. IBM may use this contact information in furtherance of your business relationship with IBM. The IBM Storage Insights Cloud Service is governed by the terms of the IBM Cloud Service Agreement (<https://www.ibm.com/terms>) and the IBM Storage Insights Service Description (<http://www.ibm.com/software/sla/laab.nslsla/sla/sd-8055-02>). Your use of the IBM Storage Insights Cloud Service indicates your acceptance of the terms referenced in these documents.

By clicking "Submit" you agree that IBM may process your data in the manner indicated above and as described in Privacy and Terms.

Submit

Figure 9-27 IBM Storage Insights registration form

IBM Storage Insights initial setup guide

After your registration for IBM Storage Insights is complete, the first login starts the IBM Storage Insights initial setup guide automatically in the browser.

The Welcome window guides you on the next steps, as shown in Figure 9-28.

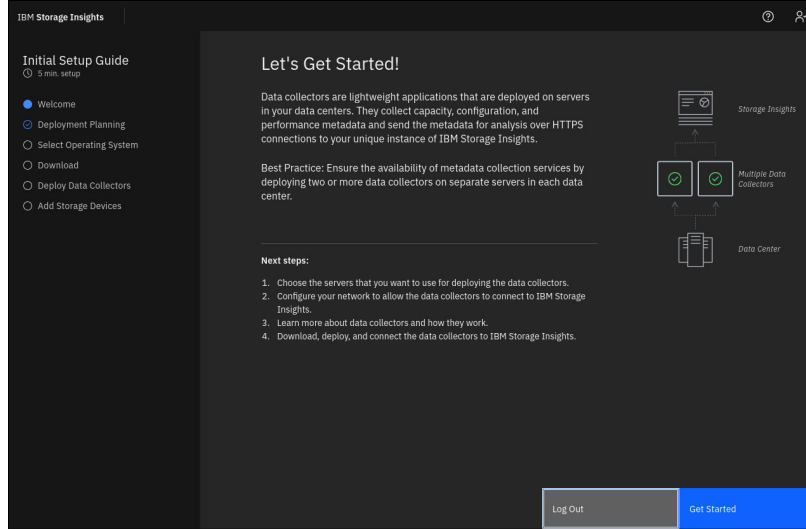


Figure 9-28 IBM Storage Insights initial setup guide

Note: The IBM Storage Insights URL is contained in the welcome email. If you have not received the welcome email, log in to [IBM Support](#) and open a ticket.

The Deployment Planning window provides guidance about the list of supported operating systems for the data collectors. Data collectors are lightweight applications that are deployed on servers or virtual machines in your data centers. Data collectors collect capacity, configuration, and performance metadata about your monitored devices and send the metadata for analysis over HTTPS connections to your IBM Storage Insights service. The network security requirements, and requirements for proxy configuration. Figure 9-29 shows the Deployment Planning window.

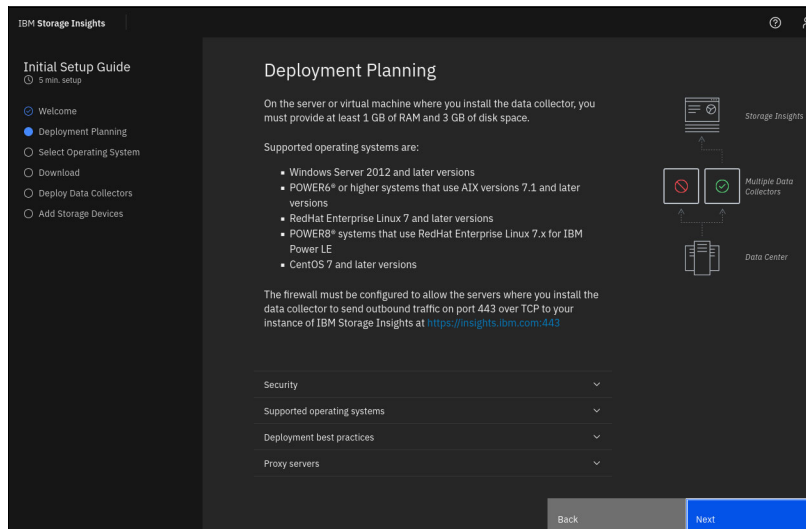


Figure 9-29 IBM Storage Insights Deployment Planning

Embedded data collector

Starting with version 8.5.3 all IBM Storage Virtualize products with a configured memory of 128GB and supports the embedded version of a data collector. This provides the possibility, that a data collector does not need to be installed on a separate operating system.

Example 9-7 shows the command to add the IBM Storage Insights service to the Cloud Call Home function.

Example 9-7 Command to add the SI tenant ID through the CLI

```
IBM_FlashSystem:FSxxx:superuser>chcloudcallhome -sitenantid
xxxxxxxx-xxxx-xxxx-xxxx-xxxxxxxxxxxxx
IBM_FlashSystem:FSxxx:superuser>lsccloudcallhome
status enabled
connection active
error_sequence_number
last_success 230626091945
last_failure
si_tenant_id xxxxxxxx-xxxx-xxxx-xxxx-xxxxxxxxxxxxx
```

After the connection of the embedded data collector was established successfully, an informational event will pop up in IBM Storage Insights, as shown in Figure 9-30.

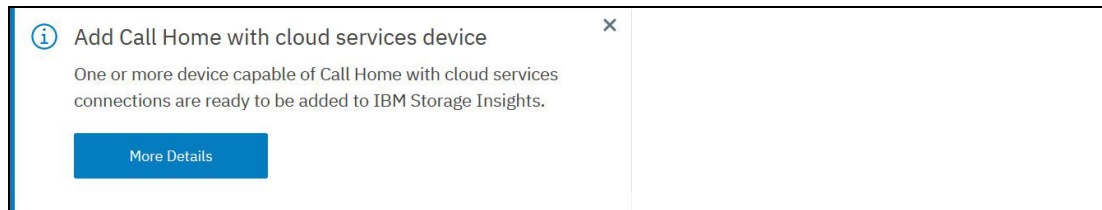


Figure 9-30 IBM Storage Insights info event after the si_tenant_id was added to Cloud Call Home

To complete to process the cloud service device, which triggered the request to send data to the IBM Storage Insights instance needs to be approved, as shown in Figure 9-31.

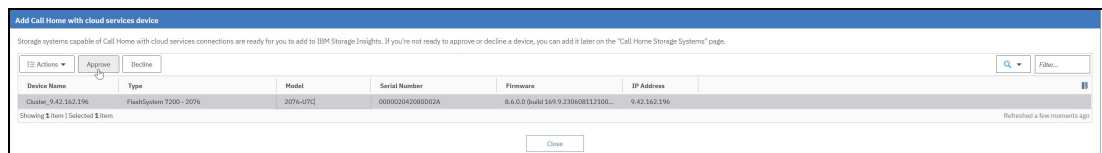


Figure 9-31 Storage Insights - Add Call Home with cloud service device

Example 9-8 shows the error message, which will be logged in case the IBM Storage Insights instance id will be added to the Cloud Call Home function on a 64GB memory system.

Example 9-8 CMMVC9305E

```
IBM_FlashSystem:FS9xxx:superuser>chcloudcallhome -sitenantid
xxxxxxxx-xxxx-xxxx-xxxx-xxxxxxxxxxxxx
CMMVC9305E The Cloud Call Home feature is not supported on this platform.
IBM_FlashSystem:FS9xxx:superuser>sainfo lshardware | grep memory_configured
memory_configured 64
IBM_FlashSystem:FS9xxx:superuser>
```

Deployable data collector

This section covers the security prerequisites and supported operating systems of the insatiable version of the data collector.

► Security

The data collector initiates outbound-only connections over HTTPS to transmit metadata to your unique instance of IBM Storage Insights in the IBM Cloud data center.

The data collector collects metadata about your storage and does not access application, personal, or identity data.

Learn more about how data is collected, transmitted, and protected in the [IBM Storage Insights: Security Guide](#).

► Supported operating systems

Data collectors can be installed on Windows, AIX, and Linux servers. You must provide at least 1 GB of RAM and 3 GB of disk space on each server, in addition to the disk space that is required to install the data collector. The server or VM where you install data collectors must be available 24x7. For more information, see “Monitoring large environments” on page 581.

– Windows

The Windows data collector runs on Windows Server 2012 and later versions. To install and run the data collector on Windows servers, you need to be logged in as Administrator.

– AIX

The AIX data collector runs on POWER6 or later systems that use AIX 7.1 or later. The AIX data collector can run on a physical AIX installation or a logical partition (LPAR). To install and run the data collector, log in to the server as root.

– Linux

The Linux data collector runs on Linux x86-64 operating systems. The supported Linux operating systems are Red Hat Enterprise Linux 7 and later and CentOS 7 and later. To install and run the data collector, log in to the server as root.

– Linux for IBM Power LE

The Linux for IBM Power LE data collector runs on Linux ppc64le operating systems. The supported Linux operating systems are Red Hat Enterprise Linux 7.x for IBM Power LE.

Restriction: You cannot monitor third-party, IBM FlashSystem A9000, IBM XIV, and IBM Spectrum Accelerate devices. To install and run the data collector, log in to the server as root.

Deployment best practices

Avoid high network latency and avoid interruptions in metadata collection services by installing two or more data collectors on separate servers in each of your data centers:

► Redundancy

To make your data collection services more robust, install two or more data collectors on separate servers in your data center.

When you add storage devices, the data collectors that you deployed are tested to see whether they can communicate with your storage devices. If multiple data collectors can communicate with a storage device, then the data collector with the best response time gets the metadata collection job.

If the collection of metadata is interrupted, the data collectors are tested again, and the data collectors with the best response times take over the collection of metadata.

► **Monitoring storage devices in multiple data centers**

To avoid high network latency and avoid interruptions in the collection of metadata when you monitor storage devices in data centers in different locations, install two or more data collectors on separate servers in each data center.

Example: You install data collectors in your Washington and Chicago data centers and both data centers are connected over the network. If the data collectors in your Washington data center go offline, then the data collectors in your Chicago data center take over the collection of your metadata for both data centers.

► **Monitoring large environments**

A best practice is to deploy one data collector for every 25 storage devices that you want to monitor. The number of volumes that your storage devices manage also determines the number of data collectors that you must deploy. For example, if you add 10 storage devices that manage 50,000 volumes, you must deploy more data collectors to manage the collection of metadata.

Note: Make your metadata collection more robust by installing more data collectors. IBM Storage Insights can automatically switch the collection of metadata to other data collectors if the metadata collection is interrupted.

IBM Support monitors issues with the collection of metadata to make you aware of issues with collecting metadata. Recurring issues with collecting metadata might indicate that you must deploy more data collectors to share the collection of metadata collection workload.

In large environments, ensure that the servers that are used to install and run the data collectors have 4 GB of available RAM and 4 GB of available drive capacity.

Proxy servers

When you install the data collector, you can connect the data collector through a proxy server.

To connect to the proxy server, the hostname and port number is needed. The connection through a secure proxy server, username and password credentials are needed as well.

Deploying data collectors

To deploy data collectors, complete the following steps:

1. Select the candidate operating system of the host that will be used to install the data collector, as shown in Figure 9-32.

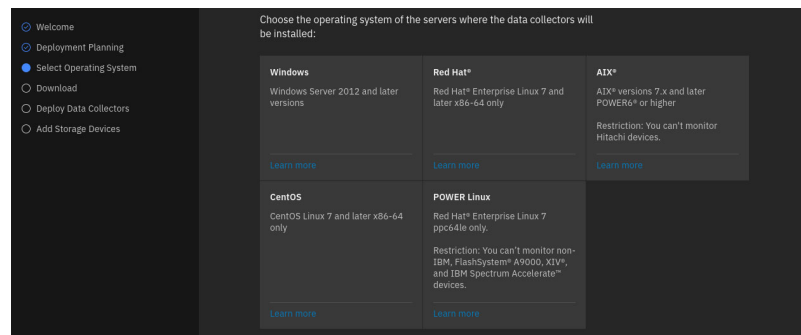


Figure 9-32 Select Operating System window

2. Accept the license agreement for the data collector, as shown in Figure 9-33.

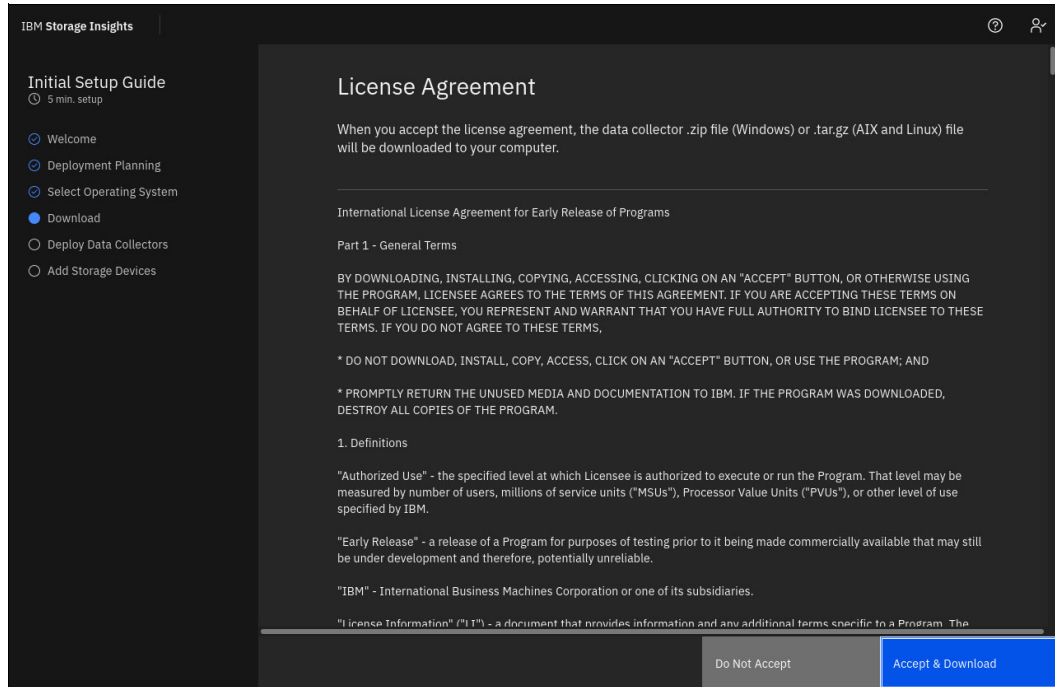


Figure 9-33 Data collector license agreement

3. Follow the guidance in the Deploy Data Collectors window to download and deploy the data collector, as shown in Figure 9-34.

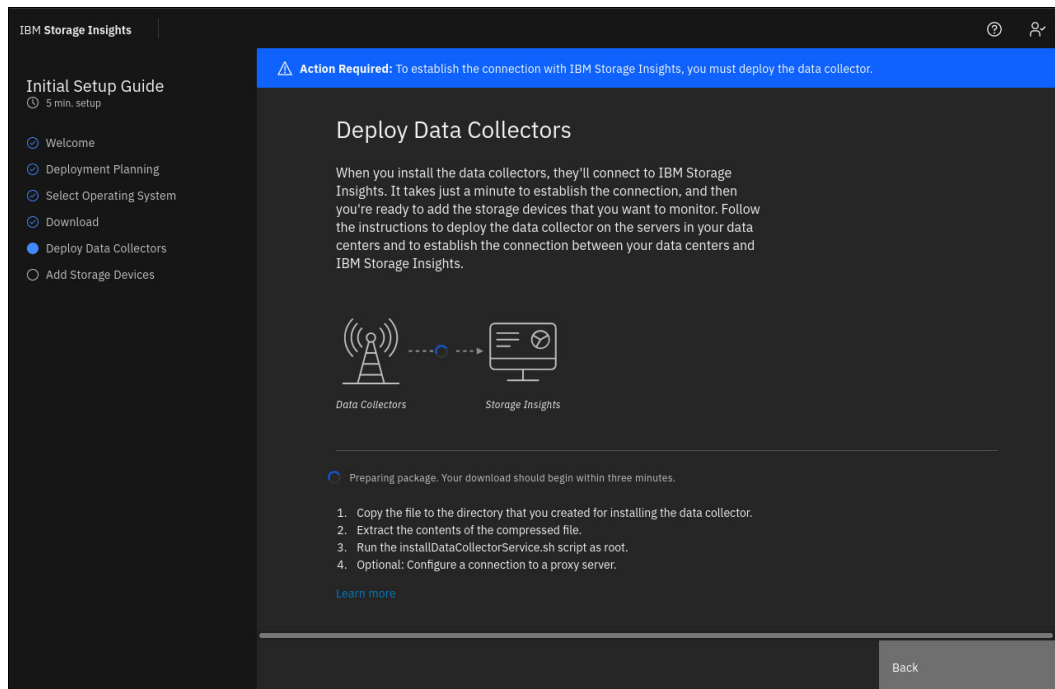


Figure 9-34 Downloading the data collector in preparation for its installation

4. Install the data collector according to the provided instructions. Figure 9-35 shows the installation of the data collector on a Linux host.

```
root:Cloud_DataCollector_linux# ./installDataCollectorService.sh

Warning no default label for /tmp/DC/Cloud_DataCollector_linux
Data Collector service will be registered in /etc/systemd/system.

Does your data collector require a proxy server? [yes/no] (default = yes): no
You made no configuration modification.

Verifying the connection to the storage management service...
The connection verification succeeded.

The data collector was installed as a service: "dataCollector_170d51fa354ed31212
3a2b1a77550dc5".
Base directory: /tmp/DC/Cloud_DataCollector_linux.

Starting the service:...
root:Cloud_DataCollector_linux# █
```

Figure 9-35 Data collector installation on a Linux host

5. After the data collector is installed and communication is established, the IBM Storage Insights Dashboard view starts with an Add Storage System prompt, as shown in Figure 9-36.

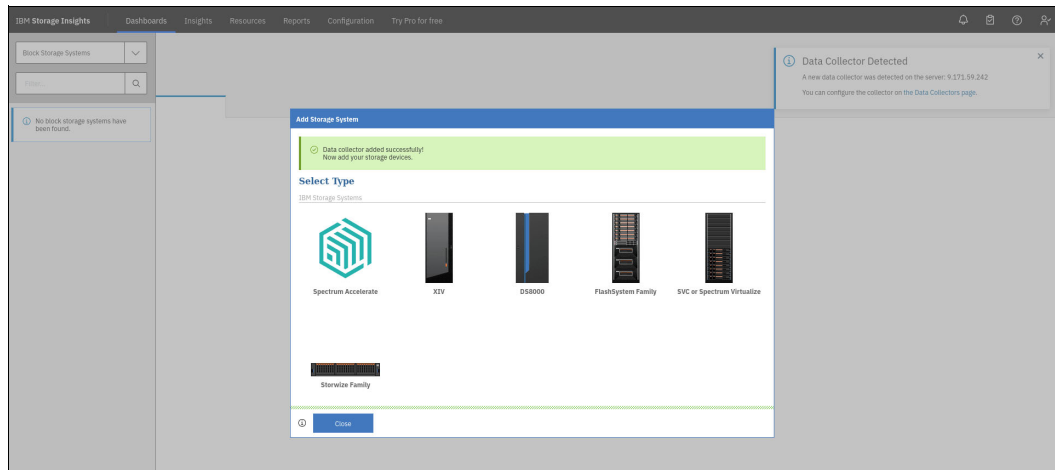


Figure 9-36 Adding storage systems to IBM Storage Insights

Note: If you have multiple geographically dispersed data centers and plan to install data collectors in each data center, the association between storage systems and data collector can be modified after multiple data collectors are available.

IBM Storage Insights dashboards

You can create customized dashboards to monitor specific storage systems or help troubleshoot issues. For example, you might have dashboards to monitor storage by data center location or storage usage by production platform.

Operations dashboard

You can use the Operations dashboard to identify which block storage systems or fabrics in your inventory need attention, such as the ones with error or warning conditions. You can manage your storage and fabrics in the Operations dashboard by using key insights and analysis about health, capacity, and performance.

To view the Operations dashboard, select **Dashboards** → **Operations**. With IBM Storage Insights, you get the information that you need to monitor the health of your block storage environment and fabrics on the Operations dashboard.

You can click a storage system in the list to get an overview of the health of the storage system components or resources, key capacity metrics, including compression savings, and key performance metrics. You can open the GUI for the storage system from the Component Health overview.

You can view more details about the storage system and components from the overview (IBM Storage Insights Pro only):

- ▶ *Notifications details and actions* that you can take to manage events.
- ▶ *Tickets details and actions* that you can take to manage tickets.
- ▶ *Properties details*, including editable name, location, and custom tag fields, and support information.
- ▶ *Inventory of nodes and enclosures for an SVC storage system*, including support information. (IBM Storage Insights Pro only).
- ▶ *Data collection details*, such as the status of the data collection, when the most recent data collection occurred, and a list of the available data collectors.

Figure 9-37 shows an example of the Operations dashboard.

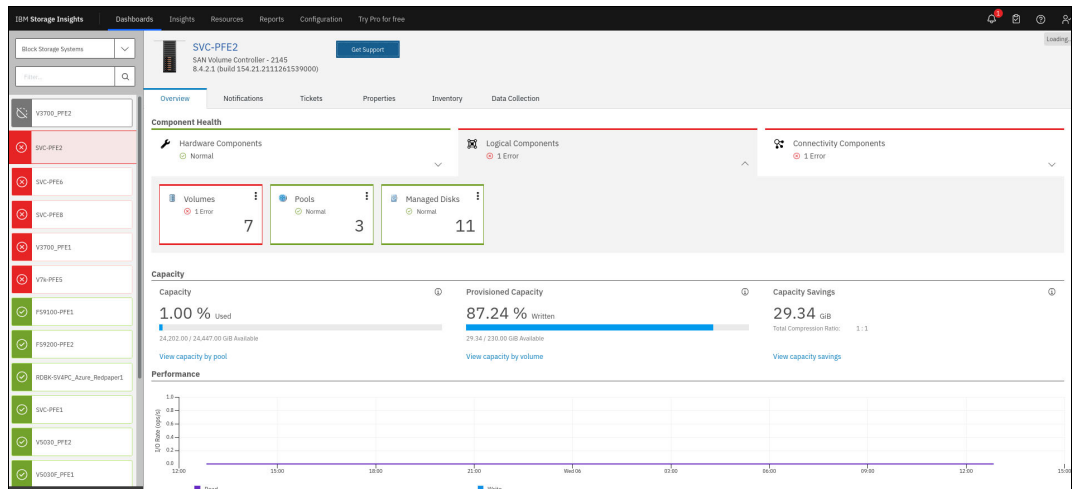


Figure 9-37 Operations Dashboard

Network Operations Center dashboard

Monitor events and storage systems in your inventory at a glance. The Network Operations Center (NOC) dashboard provides key insights and analysis about the health, capacity, and performance of your block storage in one, central location. You can create customized dashboards to selectively monitor particular storage systems.

To view the NOC dashboard, select **Dashboards** → **NOC**. You can display it on a dedicated monitor in your network operations center so that you can monitor storage system changes at a glance.

The block storage systems that are being monitored are displayed in tiles or rows on the dashboard. Call Home must be enabled on the storage systems that are monitored.

Use the Tile view to quickly access essential information about your storage systems, including the overall condition. The overall condition is determined by the most critical status that was detected for the storage system's internal resources. Storage systems with error conditions are displayed at the top of the dashboard, followed by storage systems with warning conditions.

On each tile, a snapshot of performance and capacity is displayed. Click the tile to view the following information:

- ▶ Overview of the health of the storage system components or resources, key capacity metrics including compression savings, and key performance metrics. You can open the GUI for the storage system from the Component Health overview.
- ▶ (IBM Storage Insights Pro only) You can view more details about the storage system and components from the overview.
- ▶ Notifications details and actions that you can take to manage events.
- ▶ Tickets details and actions that you can take to manage tickets.
- ▶ Properties details, including editable name, location, and custom tag fields, and support information.
- ▶ (IBM Storage Insights Pro only) Inventory of nodes and enclosures for IBM Storage Virtualize systems, including support information, if available.

Figure 9-38 shows an example of the NOC dashboard.

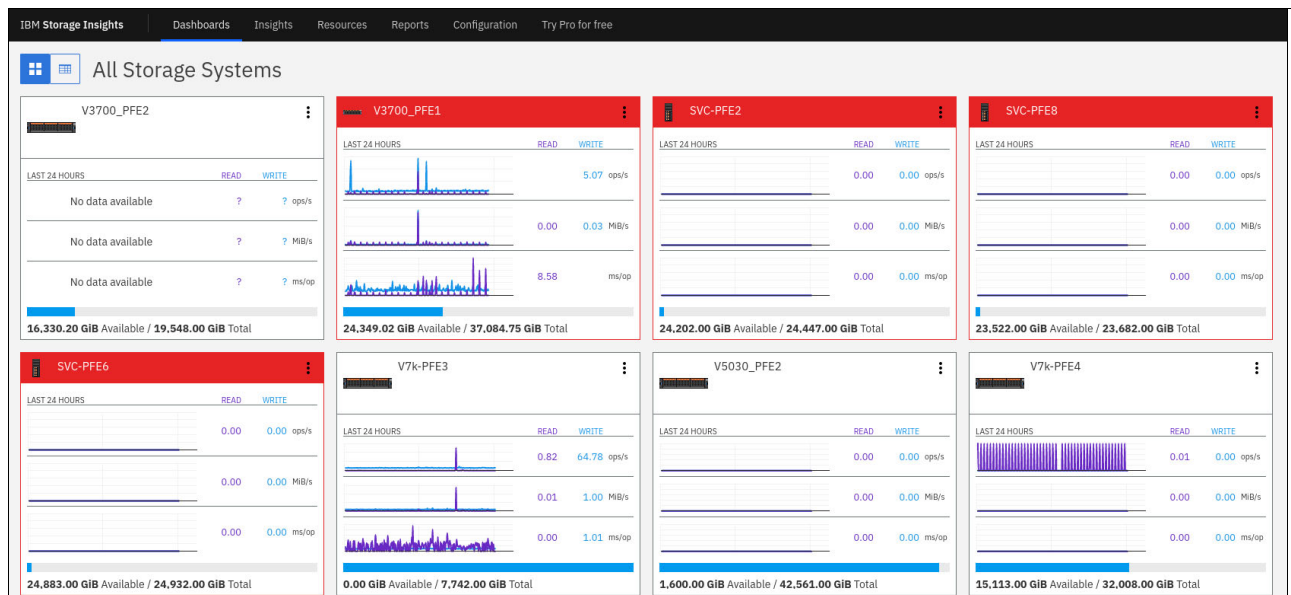


Figure 9-38 NOC dashboard

Resources tabular view

These tabular views are available for each type of resource that IBM Storage Insights supports. They are Block Storage Systems, Switches, Fabrics, and Hosts.

This view is useful because it can be sorted and filtered based on the selected column. For example, Figure 9-39 shows Block Storage Systems that are sorted in ascending order by IBM Storage Virtualize code level.

Block Storage Systems

11 Normal
0 Warning
1 Unreachable
4 Error

Name	Condition	Firmware	Events	Data Collection	Capacity (GiB)	Used Capacity (%)
V3700_PFE1	Error	7.5.0.14 (build 115.54.1804161640000)	Informational	Running	37,084.75	34 %
V3700_PFE2	Device Un...	7.8.1.12 (build 135.9.2006121041000)		Failed	19,548.00	16 %
V7k-PFE3	Normal	8.1.3.6 (build 143.13.1904231708000)	Informational	Running	7,742.00	100 %
V5030_PFE2	Normal	8.2.1.13 (build 147.21.2104210349000)	Informational	Running	42,561.00	96 %
V5030_PFE1	Normal	8.3.1.5 (build 150.27.2104221539000)	Informational	Running	23,941.00	32 %
SVC-PFE6	Error	8.3.1.6 (build 150.28.2110050819000)		Running	24,932.00	0 %
SVC-PFE8	Error	8.4.0.6 (build 152.24.2201050004000)		Running	23,682.00	1 %
V7k-PFE4	Normal	8.4.0.6 (build 152.24.2201050004000)	Informational	Running	32,008.00	53 %
v7000-ctr-61	Normal	8.4.0.6 (build 152.24.2201050004000)	Informational	Running	36,123.00	15 %
v7000-ctr-60	Normal	8.4.0.6 (build 152.24.2201050004000)	Informational	Running	36,008.00	14 %
FS9100-PFE1	Normal	8.4.0.6 (build 152.24.2201050004005)		Running	36,135.46	7 %
SVC-PFE1	Normal	8.4.2.1 (build 154.21.0000000000000)	Informational	Running	5,187.00	6 %
SVC-PFE2	Error	8.4.2.1 (build 154.21.2111261539000)		Running	24,447.00	1 %
FS9200-PFE2	Normal	8.4.2.1 (build 154.21.2111261539000)	Informational	Running	35,702.34	20 %
V7k-PFE5	Normal	8.4.2.1 (build 154.21.2111261539008)	Informational	Running	17,944.00	40 %
RDBK-SV4PC_Azure_Red...	Normal	8.4.3.0 (build 155.0.2112040719000)	Informational - Ackn...	Task expir...	1,024.00	3 %

Figure 9-39 Block Storage Systems table view

For more information, see [Dashboards in IBM Storage Insights](#).

Adding users to your dashboard

Optional: Add users, such as other storage administrators, IBM Technical Advisors, and IBM Business Partners, at any time so that they can access your IBM Storage Insights dashboard, by completing the following steps:

1. In IBM Storage Insights, click your username in the upper right of the dashboard.
2. Click **Manage Users**.
3. On your My IBM page, ensure that **IBM Storage Insights** is selected.
4. Click **Add new user**.

For more information, see [Adding and removing users](#).

Enabling Call Home

Get the most out of IBM Storage Insights by enabling Call Home on your IBM block storage systems. With Call Home, your dashboard includes a diagnostic feed of events and notifications about their health and status.

Stay informed so that you can act quickly to resolve incidents before they affect critical storage operations.

For more information, see [Monitoring resources with Call Home](#).

An extra benefit is that the Call Home data is screened against a set of rules to identify misconfiguration, deviations from best practices, and IBM Support Flashes that are applicable. The results are displayed in IBM Storage Insights.

To see these best practices, select **Insights** → **Advisor**, as shown in In Figure 9-40.

Keep your infrastructure healthy with these recommendations.

Unacknowledged Recommendations: 24 24 Informational | 25 Acknowledged

Actions Acknowledge

Event	Severity	Time	Device Name	More Information
Hosts have more than the recommended number of paths to volumes	Informational	Dec 13, 2021, 21:58:54	v7000-ctr-60	Based on our general experience, it is best to limit the total number of paths fr
Consistency Protection can be configured for Global Mirror and Metro Mirror relationships	Informational	Dec 13, 2021, 21:58:54	v7000-ctr-60	IBM Spectrum Virtualize V7.8.1 introduced the Remote Copy Consistency Prot
The system has not been configured to use Network Time Protocol (NTP)	Informational	Dec 17, 2021, 13:54:40	V7k-PFE5	The system is not currently configured to use NTP to keep the system clock sh
Performance statistics configuration outside of recommended parameters	Informational	Dec 17, 2021, 13:54:40	V7k-PFE5	Collection of performance statistics is either disabled or the frequency is set to
Defined hosts show as offline status.	Informational	Dec 17, 2021, 13:58:24	V7k-PFE5	There are one or most hosts which are configured in the system that show an o
Unused volumes identified in the system	Informational	Dec 17, 2021, 13:58:24	V7k-PFE5	Volumes were found in storage pools that appear to be unused. These may hav
Consider enabling volume protection	Informational	Dec 17, 2021, 13:58:24	V7k-PFE5	To prevent an active volume from being deleted unintentionally, administrators
Preferred node optimization for FlashCopies	Informational	Jan 6, 2022, 10:10:29	V7k-PFE4	FlashCopy mappings were identified that have source and target volumes in th
Unused volumes identified in the system	Informational	Jan 6, 2022, 10:10:29	V7k-PFE4	Volumes were found in storage pools that appear to be unused. These may hav
The system has not been configured to use Network Time Protocol (NTP)	Informational	Jan 12, 2022, 10:16:43	SVC-PFE1	The system is not currently configured to use NTP to keep the system clock sh
Performance statistics configuration outside of recommended parameters	Informational	Jan 12, 2022, 10:16:43	SVC-PFE1	Collection of performance statistics is either disabled or the frequency is set to
Preferred node optimization for FlashCopies	Informational	Jan 27, 2022, 21:36:42	FS9200-PF...	FlashCopy mappings were identified that have source and target volumes in th
Software Update Recommendation	Informational	Feb 1, 2022, 06:08:56	V3700_PFE1	
Software Update Recommendation	Informational	Feb 1, 2022, 06:09:35	V3700_PFE1	
Software Update Recommendation	Informational	Feb 3, 2022, 06:07:26	V5030F_PF...	
Software Update Recommendation	Informational	Feb 3, 2022, 06:07:51	V5030F_PF...	
Software Update Recommendation	Informational	Feb 4, 2022, 06:10:07	V3700_PFE1	
Enable Configuration Reporting for automated system analysis	Informational	Mar 15, 2022, 21:00:35	FS9200-PF...	IBM systems automatically check your configuration for known issues and bes

Figure 9-40 Advisor Insights window

More benefits of the Advisor can be found in IBM Documentation: [Monitoring recommended actions in the Advisor](#).

9.3 Capacity monitoring

Effective and exact capacity management is based on fundamental knowledge of capacity metrics in IBM Storage Virtualize. DRPs, thin provisioning, compression, and deduplication add many metrics to the IBM Storage Virtualize management GUI, IBM Spectrum Control, and IBM Storage Insights.

This section is divided into three sections to describe capacity monitoring by using any of the following interfaces:

- ▶ The management GUI
- ▶ IBM Spectrum Control
- ▶ IBM Storage Insights

This section describes the key capacity metrics of the IBM Storage Virtualize management GUI, IBM Spectrum Control (based on 5.4.6), and IBM Storage Insights.

Figure 9-41 shows how to interpret the capacity and savings in a storage environment.

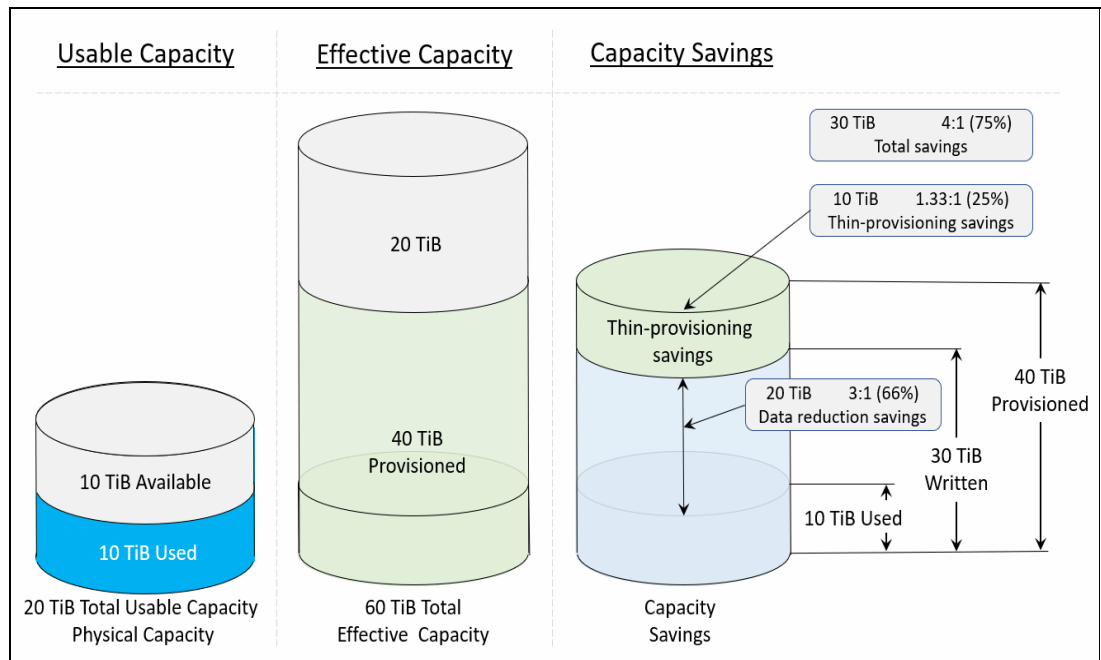


Figure 9-41 Understanding capacity information

9.3.1 Capacity monitoring through the management GUI

Different capacity reports are implemented to the management GUI of the IBM Storage Virtualize products. Figure 9-42 shows the definition of these values for a better understanding of the capacity terminologies.

Previous term or new term description	New term	Definition
free capacity or free space (definition updated)	available capacity	The amount of usable capacity that is not yet used in a system, pool, array, or MDisk.
(definition updated)	capacity	The amount of data that can be contained on a storage medium.
(definition updated)	data reduction	A set of techniques that can be used to reduce the amount of usable capacity that is required to store data. Examples of data reduction include data deduplication and compression.
(new term)	data reduction savings	The total amount of usable capacity that is saved in a system, pool, or volume through the application of an algorithm such as compression or deduplication on the written data. This saved capacity is the difference between the written capacity and the used capacity.
projected capacity	effective capacity	The amount of provisioned capacity that can be created in a system or pool without running out of usable capacity given the current data reduction savings being achieved. This capacity equals the usable capacity divided by the data reduction savings percentage.
reserved capacity	overhead capacity	An amount of usable capacity that is occupied by metadata in a system or pool and other data that is used for system operations.
(new term)	overprovisioned ratio	The ratio of provisioned capacity to usable capacity in a system or pool.
(new term)	overprovisioning	The result of creating more provisioned capacity in a storage system or pool than there is usable capacity. Overprovisioning occurs when thin provisioning or data reduction techniques ensure that the used capacity of the provisioned volumes is less than their provisioned capacity.
virtual capacity or allocated capacity (term refers to capacity of volume)	provisioned capacity	The total capacity of all volumes and volume copies in a system or pool.
physical capacity (when term refers to individual drives)	raw capacity	The reported capacity of the drives in the system before formatting or RAID (Redundant Array of Independent Disks) is applied.
physical capacity	usable capacity	The amount of capacity that is provided for storing data on a system, pool, array, or MDisk after formatting and RAID techniques are applied.
fully allocated volume	standard-provisioned volume	A volume that completely uses storage at creation.
(new term)	standard provisioning	The ability to completely use a volume's capacity for that specific volume.
(new term)	thin-provisioning savings	The total amount of usable capacity that is saved in a system, pool, or volume by consuming usable capacity only when needed as a result of write operations. The capacity that is saved is the difference between the provisioned capacity minus the written capacity.
(new term)	total capacity savings	The total amount of usable capacity that is saved in a system, pool, or volume through thin provisioning and data reduction techniques. This saved capacity is the difference between the used usable capacity and the provisioned capacity.
(new term)	used capacity	The amount of usable capacity that is taken up by data or overhead capacity in a system, pool, array, or MDisk after data reduction techniques have been applied.
(definition updated)	written capacity	The amount of usable capacity that would have been used to store written data in a system or pool if data reduction was not applied.
(new term)	written capacity limit	The largest amount of capacity that can be written to a drive, array, or MDisk. The limit can be reached even when usable capacity is still available.

Figure 9-42 Capacity terminology

Dashboard

The **Capacity** section on the Dashboard provides an overall view of system capacity. This section displays usable capacity, used/stored capacity, and capacity savings.

Usable capacity (see Figure 9-43) visualized the amount of physical capacity that is available for storing data on a system, pool, array, or MDisk after formatting and RAID techniques are applied. The pie graph and divided into three categories: Used/Stored Capacity, Available Capacity, and Total Capacity.



Figure 9-43 Usable capacity

Note: If data reduction is used at two layers (Data reduction pool with compressed volumes and FlashCore Modules), it is recommended to allocate the physical storage 1:1 (physical capacity instead of the effective capacity), to prevent an out of space scenario.

The compression rate of FlashCore-Modules in a two-layer data reduction configuration is minimal and only affects the data reduction pool metadata

Used/Stored Capacity is the amount of usable capacity that is taken up by data or overhead capacity in a system, pool, array, or MDisk after data reduction techniques have been applied.

The **Available Capacity** is the amount of usable capacity that is not yet used in a system, pool, array, or MDisk.

Total capacity is the amount of total physical capacity of all standard-provisioned and thin-provisioned storage that is managed by the storage system. The value is rounded to two decimal places.

To determine the usable capacity on your system through the command-line interface, several parameter values are used from the **lssystem** command to calculate Used/Stored Capacity, Available Capacity, and Total Capacity.

To get the Used/Stored Capacity value through the command-line interface, the `physical_capacity` have to be subtracted by `physical_free_capacity` and `total_reclaimable_capacity`. Example 9-9 shows how to get these value filtered from **lssystem**.

Example 9-9 Getting the value filtered from lssystem

```
IBM_FlashSystem:FS9xxx:superuser>lssystem|grep physical_capacity&&lssystem|grep
physical_free&&lssystem|grep total_reclaimable_capacity
physical_capacity 17.09TB
physical_free_capacity 17.09TB
total_reclaimable_capacity 0.00MB
```

To get the Available Capacity value through the command-line interface, the `physical_free_capacity` have to be subtracted by `physical_capacity`. Example 9-10 shows how to get these value filtered from **lssystem**.

Example 9-10 Physical_capacity and physical_free_capacity from lssystem command

```
IBM_FlashSystem:FS9xxx:superuser>lssystem |grep physical
physical_capacity 21.39TB
physical_free_capacity 21.39TB
```

To get the Total Capacity through the command-line interface, the `physical_capacity` value needs to be filtered from **lssystem**, shown in Example 9-11.

Example 9-11 Physical_capacity from lssystem command

```
IBM_FlashSystem:FS9xxx:superuser>lssystem |grep physical_capacity
physical_capacity 21.39TB
```

Capacity Savings (Figure 9-44 on page 591) view shows the amount of capacity that is saved on the system by using compression, deduplication, and thin-provisioning.

Total Provisioned shows the total capacity of all volumes provisioned in the system.

Data Reduction shows the ratio of written capacity to stored capacity, where the written data has been compressed and/or depduplicated.

Total Savings shows the ratio of provisioned capacity to stored capacity, which includes the capacity savings achieved through compression, deduplication and thin-provisioning.

Deduplication indicates the total capacity savings that the system is saved from all deduplicated volumes. Thin-Provisioning displays the total capacity savings for all thin-provisioned volumes on the system. You can view all the volumes that use each of these technologies. Different system models can have more requirements to use compression or deduplication. Verify that all system requirements before these functions are used.

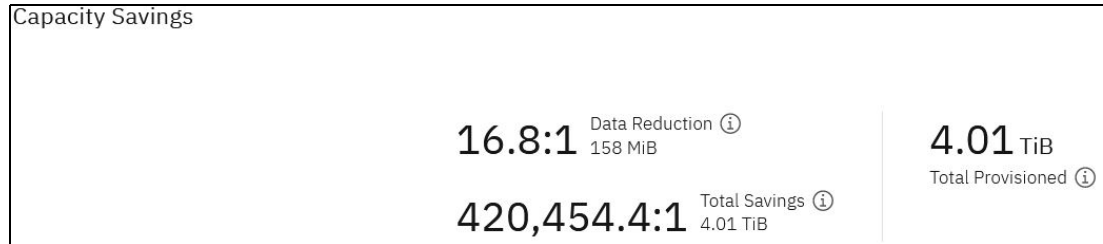


Figure 9-44 Capacity Savings window

Example 9-12 shows deduplication and compression savings and used capacity before and after reduction on CLI.

Example 9-12 Deduplication and compression savings and used capacity

```
IBM_FlashSystem:FS9xxx:superuser>lssystem |grep deduplication
deduplication_capacity_saving 0.00MB
IBM_FlashSystem:FS9xxx:superuser>lssystem |grep compression_virtual_ && lssystem
|grep compression_compressed_ && lssystem |grep compression_uncompressed_
compression_virtual_capacity 0.00MB
compression_compressed_capacity 0.00MB
compression_uncompressed_capacity 0.00MB
IBM_FlashSystem:FS9xxx:superuser>lssystem |grep reduction
used_capacity_before_reduction 0.00MB
used_capacity_after_reduction 0.00MB
```

Capacity view at the pool level

The capacity dashboard aggregates capacity information for the entire system. The pool capacity views can be used to display capacity information for a specific pool, as shown in Figure 9-45 on page 592.

Similar to the system view, this view shows the relevant capacity metrics for a specific pool.

For example, this pool is using multiple data reduction features and shows the capacity savings for each of them.

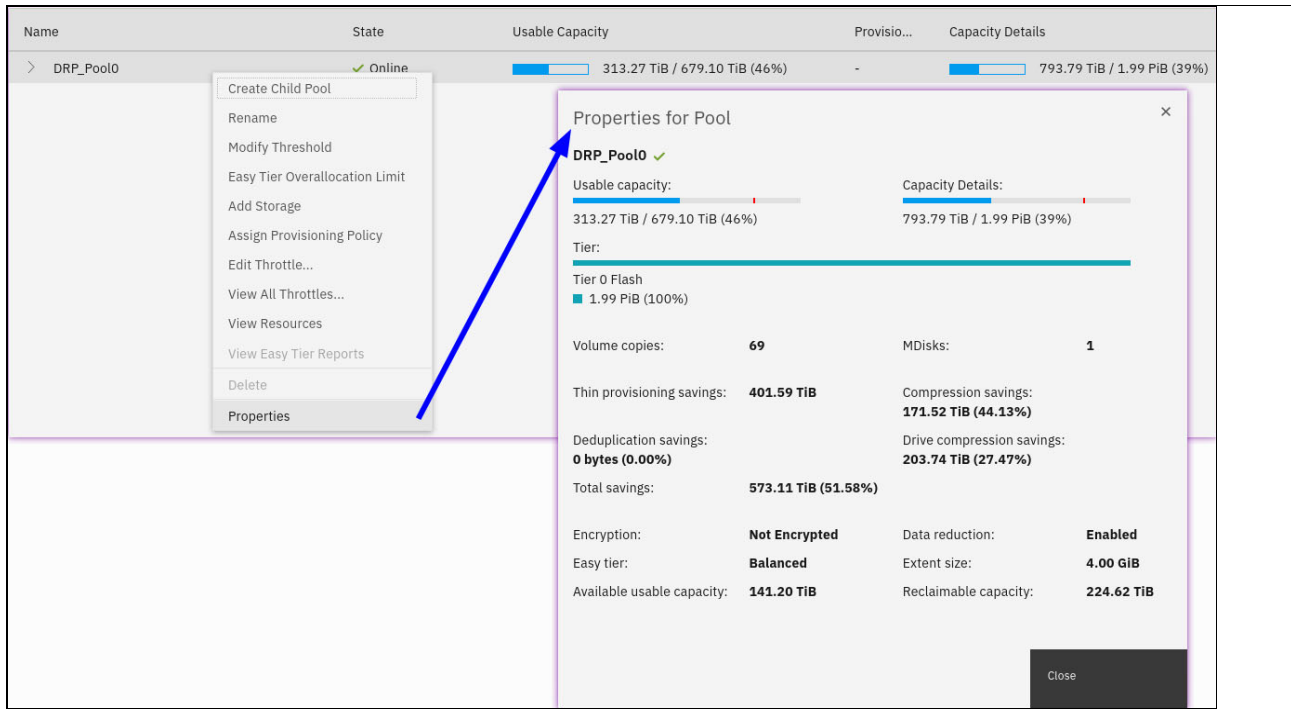


Figure 9-45 Sidebar > Pools > Properties > Properties for Pool

Capacity view at the MDisk level for compressing arrays

The capacity metrics of non-overallocated MDisks can be taken at face value. However, compressed arrays have more details to account for their physical and logical capacities.

For example, Figure 9-46 on page 593 shows that with a usable capacity of 679.10 TiB, this compressed array could store up to 1.99 PiB of addressable logical data (Written Capacity Limit).

Name	State	Usable Capacity	Written Capacity Limit
Unassigned MDisks (0)			
DRP_Pool0	Online	313.27 TiB / 679.10 TiB (46%)	793.75 TiB / 1.99 PiB (39%)
MDisk1	Online	679.10 TiB	1.99 PiB

- Rename
- Swap Drive
- Set Rebuild Areas Goal
- Expand...
- Delete
- Dependent Volumes
- Drives
- View Provisioning Group
- Properties**

Properties for MDisk MDisk1

Name: MDisk1
 State: Online
 ID: 0
 Written capacity limit: 741.63 TiB Used / 1.99 PiB Total (36.3%)
 Pool: DRP_Pool0
 Mode: Array
 Write protected: No
 Tier: Tier 0 Flash
 Encryption: Not Encrypted
 Deduplication: Not Active
 Fast-Write state: Not Empty
 Thin-Provisioned: Yes
 Supports unmap: Yes
 Usable capacity: 679.10 TiB
 Available usable capacity: 141.20 TiB
 Drive compression savings: 203.74 TiB
 Provisioning group: 0
 RAID state: Online

Close

Properties for MDisk MDisk1

Name: MDisk1
 State: Online
 ID: 0
 Written capacity limit: 741.63 TiB Used / 1.99 PiB Total (36.3%)
 Pool: DRP_Pool0
[View more details](#)

Close

Figure 9-46 Sidebar > Pools > MDisks by Pools > Properties > More details

Easy Tier Overallocation

Code level 8.2.1 introduced support for a parameter that is called *Easy Tier Overallocation Limit*. This parameter was initially available only on the CLI. Code level 8.4.2.0 and later introduced GUI support for this feature, as shown in Figure 9-47.

The default value is 100%, which helps to reduce Easy Tier swapping of less compressible extents from non-compressing arrays with highly compressible extents on compressing arrays. This approach reduces the likelihood of Easy Tier causing the compressing array to run out of physical space.

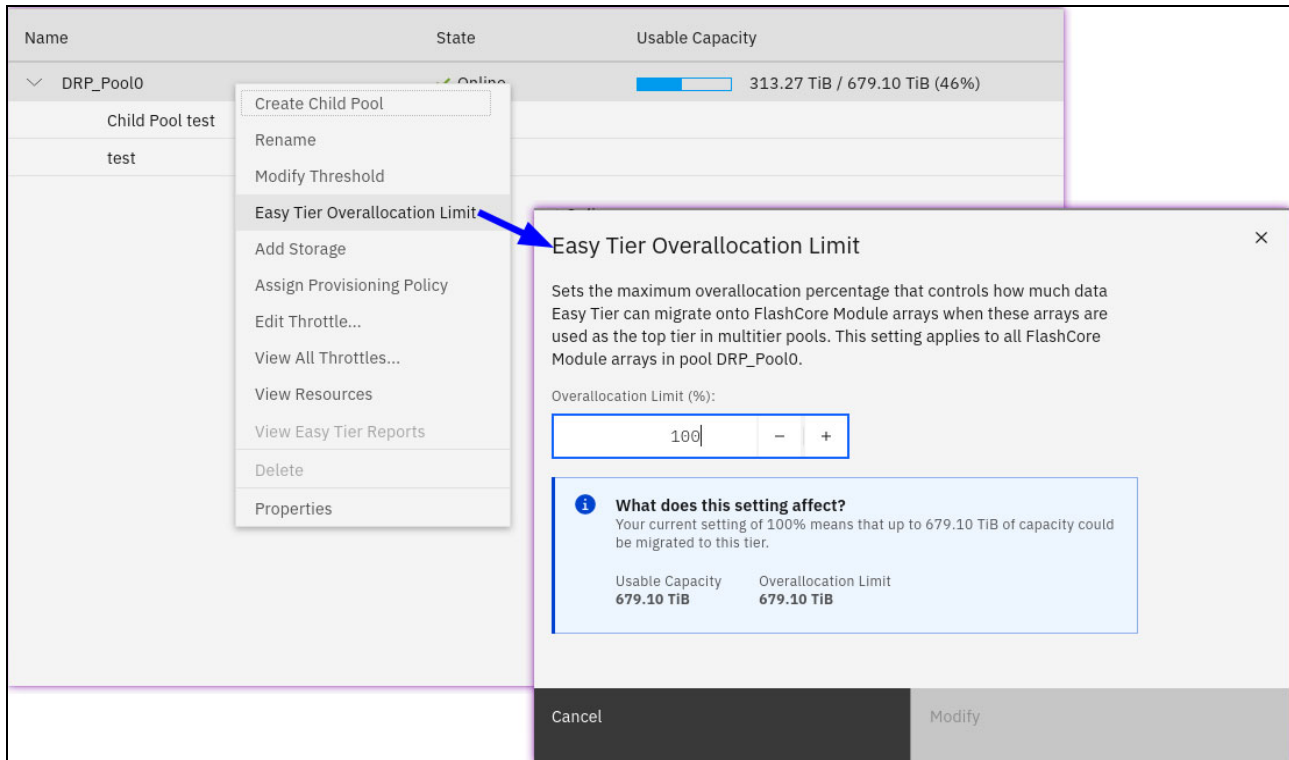


Figure 9-47 Easy Tier Overallocation Limit GUI support

Note: As of code level 8.5.0, a user is prevented from creating two compressing arrays in the same pool. **1sarrayrecommndation** does not recommend a second compressing array in a pool that already contains one. Supported systems with two or more compressing arrays in the same pool that were created pre-8.5.0 are allowed to upgrade to 8.5.0.

Compressed array capacity events

There are cases when compressed arrays (FCM arrays) might be low on usable capacity. There are specific event IDs and error codes for each of the scenarios, as shown in Table 9-1. It is important to be aware of them and to be familiar with the concept of out of space handling and recovery.

Table 9-1 Low space warning percentages for compressed arrays

Condition	Event ID	Error code	Percentage of usable capacity used
Compressed array is running low on usable capacity.	020009	1246 ^a	90% used

Condition	Event ID	Error code	Percentage of usable capacity used
Critical level of compressed array is running low on usable capacity.	020010	1246 ^a	96%
1% usable capacity that is left for a compressed array.	020011	1242	99%
A compressed array out of usable capacity.	020012	1241	100%

a. Error 1246 is not raised on Easy Tier pools or DRPs.

9.3.2 Capacity monitoring by using IBM Spectrum Control or IBM Storage Insights

The Capacity section of IBM Spectrum Control and IBM Storage Insights provides an overall view of system capacity. This section displays usable capacity, provisioned capacity, and capacity savings.

- ▶ The Capacity chart at the top of the Overview page shows how much capacity is used and how much capacity is available for storing data.
- ▶ The Provisioned Capacity chart shows the written capacity values in relation to the total provisioned capacity values before data reduction techniques are applied.
- ▶ A breakdown of the total capacity savings that are achieved when the written capacity is stored on the thin-provisioned volumes is also provided.

The Capacity chart (see Figure 9-48) of IBM Spectrum Control at the top of the Overview page (select **IBM Spectrum Control GUI** → **Storage** → **Block Storage Systems**, and then double-click the device) shows how much capacity is used and how much capacity is available for storing data.



Figure 9-48 IBM Spectrum Control overview page

IBM Storage Insights

In IBM Storage Insights, the Capacity chart shows the capacity usage (see Figure 9-49) on the Dashboards page (select **IBM Storage Insights GUI** → **Dashboards**, and then click the device).

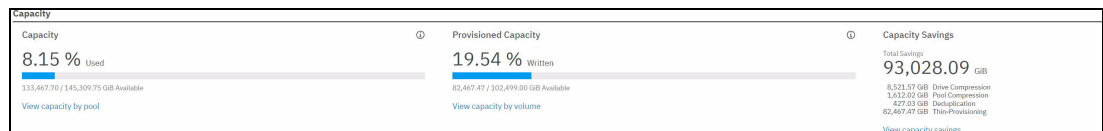


Figure 9-49 IBM Storage Insights overview page

The Provisioned Capacity chart shows the written capacity values in relation to the total provisioned capacity values before data reduction techniques are applied. The following values are shown:

- ▶ The capacity of the data that is written to the volumes as a percentage of the total provisioned capacity of the volumes.
- ▶ The amount of capacity that is still available for writing data to the thin-provisioned volumes in relation to the total provisioned capacity of the volumes. Available capacity is the difference between the provisioned capacity and the written capacity, which is the thin-provisioning savings.
- ▶ A breakdown of the total capacity savings that are achieved when the written capacity is stored on the thin-provisioned volumes is also provided.

In the Capacity Overview chart, a horizontal bar is shown when a capacity limit is set for the storage system. Hover your cursor over the chart to see what the capacity limit is and how much capacity is left before the capacity limit is reached.

For a breakdown of the capacity usage by pool or volume, click the links (see Figure 9-48 on page 595 and Figure 9-49 on page 595).

Capacity views and their metrics

In this section, we describe the metrics of the Capacity View of IBM Spectrum Control and IBM Storage Insights for block storage systems. To see the capacity views, complete the following steps:

1. To open the Capacity View in IBM Spectrum Control, click **Storage**, and then click **Block Storage Systems**. Right-click one or more storage systems and click **View Capacity** (see Figure 9-50).

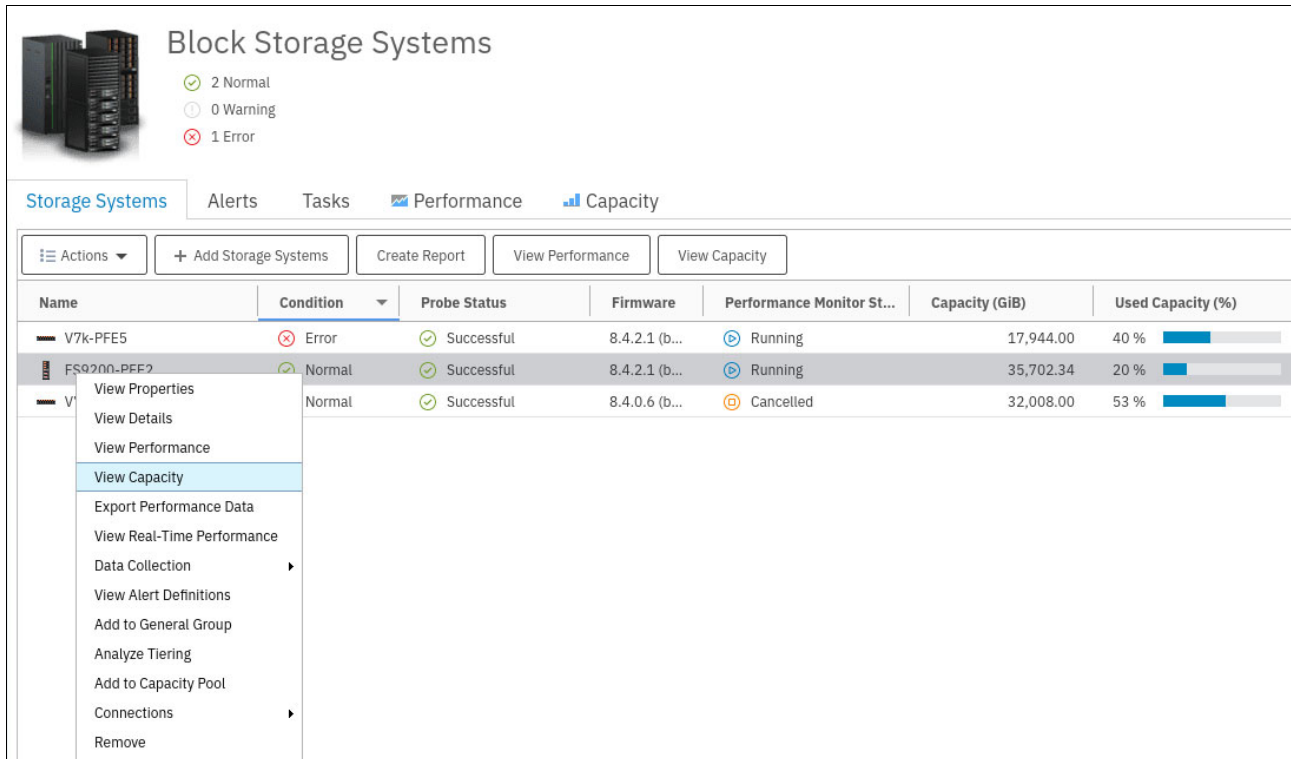


Figure 9-50 Block Storage Systems overview

2. You can also click **View Capacity** on the **Actions** menu (see Figure 9-51 on page 597) of each device.

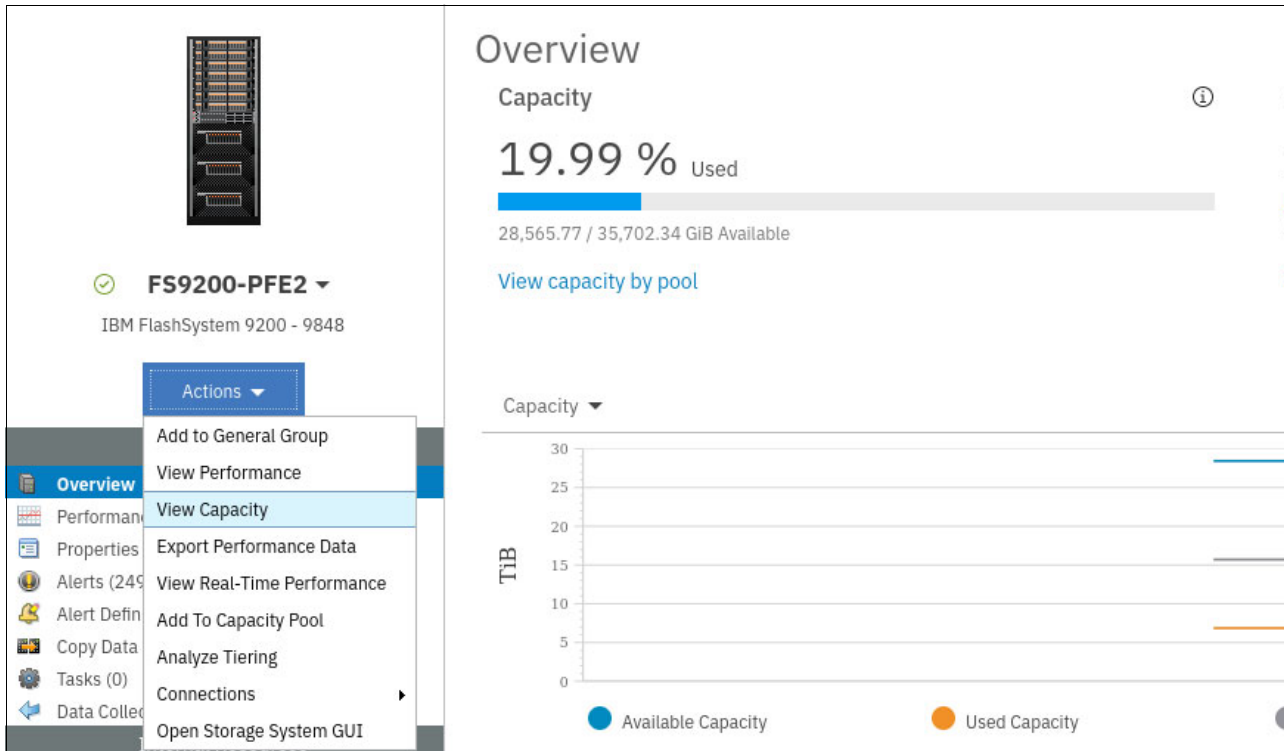


Figure 9-51 Capacity overview of Storage System

- To open the Capacity view in IBM Storage Insights, use the **Resources** menu. The rest of the sequence is similar to IBM Spectrum Control with some minor cosmetic differences (see Figure 9-50 on page 596 and Figure 9-51).

Storage system, pool capacity, and volume capacity metrics

Used Capacity (%) shows the percentage of physical capacity in the pools that is used by the standard-provisioned volumes, thin-provisioned volumes, and volumes that are in child pools. Check the value for used capacity percentage to see the following information:

- ▶ Whether the physical capacity of the pools is fully allocated, that is, the value for used capacity is 100%.
- ▶ Whether sufficient capacity is available to perform the following actions:
 - Provision new volumes with storage.
 - Allocate to the compressed and thin-provisioned volumes in the pools.

The following formula is used to calculate Used Capacity (%), as shown in Figure 9-52:

$$[(\text{Used Capacity} \div \text{Capacity}) * 100]$$

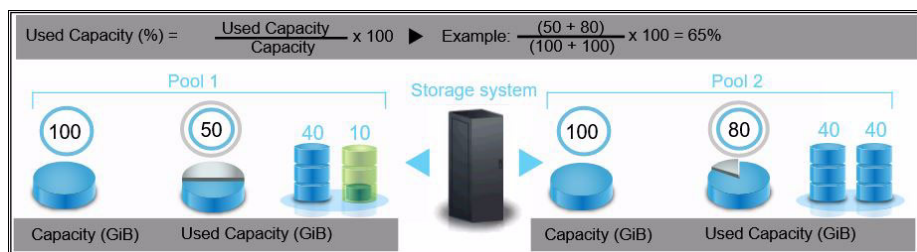


Figure 9-52 Used Capacity

Note: Used Capacity (%) was previously known as Physical Allocation.

Used Capacity (GiB) shows the amount of space that is used by the standard and thin-provisioned volumes in the pools. If the pool is a parent pool, the amount of space that is used by the volumes in the child pools also is calculated.

The capacity that is used by for thin-provisioned volumes is less than their provisioned capacity, which is shown in the Provisioned Capacity (GiB) column. If a pool does not have thin-provisioned volumes, the value for used capacity is the same as the value for provisioned capacity.

Note: Used Capacity (GiB) was previously known as Allocated Space.

Adjusted Used Capacity (%) shows the amount of capacity that can be used without exceeding the capacity limit.

The following formula is used to calculate Adjusted Used Capacity (%):

$$[(\text{Used Capacity in GiB} \div \text{Capacity Limit in GiB}) * 100]$$

For example, if the capacity is 100 GiB, the used capacity is 40 GiB, and the capacity limit is 80% or 80 GiB, the value for Adjusted Used Capacity (%) is $(40 \text{ GiB} / 80 \text{ GiB}) * 100$ or 50%.

Therefore, in this example, you can use 30% or 40 GiB of the usable capacity of the resource before you reach the capacity limit (see Figure 9-53).

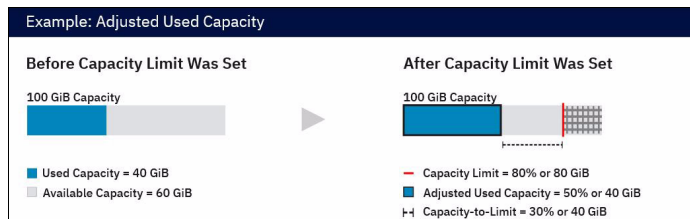


Figure 9-53 Example of Adjusted Used Capacity

If the used capacity exceeds the capacity limit, the value for Adjusted Used Capacity (%) is over 100%.

To add the Adjusted Used Capacity (%) column, right-click any column heading on the Block Storage Systems window.

Available Capacity (GiB) shows the total amount of the space in the pools that is not used by the volumes in the pools. To calculate available capacity, the following formula is used:

$$[\text{pool capacity} - \text{used capacity}]$$

Note: Available Capacity was previously known as Available Pool Space.

Available Volume Capacity (GiB) shows the total amount of remaining space that can be used by the volumes in the pools. The following formula is used to calculate this value:

$$[\text{Provisioned Capacity} - \text{Used Capacity}]$$

The capacity that is used by thin-provisioned volumes is typically less than their provisioned capacity. Therefore, the available capacity represents the difference between the provisioned capacity and the used capacity for all the volumes in the pools. For Hitachi VSP non-thin-provisioned pool capacity, the available capacity is always zero.

Note: Available Volume Capacity (GiB) was previously known as Effective Unallocated Volume Space.

Capacity (GiB) shows the total amount of storage space in the pools. For XIV systems and IBM Spectrum Accelerate, capacity represents the physical (“hard”) capacity of the pool, not the provisioned (“soft”) capacity. Pools that are allocated from other pools are not included in the total pool space.

Note: Capacity was previously known as Pool Capacity.

Capacity Limit (%) and *Capacity Limit (GiB)* can be set for the capacity that is used by your storage systems. For example, the policy of your company is to keep 20% of the usable capacity of your storage systems in reserve. Therefore, you log in to the GUI as Administrator and set the capacity limit to 80% (see Figure 9-54).



Figure 9-54 Capacity limit example

Capacity-to-Limit (GiB) shows the amount of capacity that is available before the capacity limit is reached.

The formula for calculating Capacity-to-Limit (GiB) is as follows:

$$(\text{Capacity Limit in GiB} - \text{Used Capacity in GiB})$$

For example, if the capacity limit is 80% or 80 GiB and the used capacity is 40 GiB, the value for Capacity-to-Limit (GiB) is (80 GiB - 40 GiB or 80% - 50%), which is 30% or 40 GiB (see Figure 9-55).



Figure 9-55 Capacity-to-Limit

Compression Savings (%) are the estimated amount and percentage of capacity that is saved by using data compression across all pools on the storage system. The percentage is calculated across all compressed volumes in the pools and does not include the capacity of non-compressed volumes.

For storage systems with drives that use inline data compression technology, the Compression Savings does not include the capacity savings that are achieved at the drive level.

The following formula is used to calculate the amount of storage space that is saved:

[written capacity - compressed size]

The following formula is used to calculate the percentage of capacity that is saved:

$((\text{written capacity} - \text{compressed size}) \div \text{written capacity}) \times 100$

For example, the written capacity, which is the amount of data that is written to the volumes before compression, is 40 GiB. The compressed size, which reflects the size of compressed data that is written to disk, is 10 GiB. Therefore, the compression savings percentage across all compressed volumes is 75%.

Note: The Compression Savings (%) metric is available for resources that run IBM Storage Virtualize.

Exception: For compressed volumes that are also deduplicated, this column is blank on storage systems that run IBM Storage Virtualize.

Deduplication Savings (%) shows the estimated amount and percentage of capacity that is saved by using data deduplication across all DRPs on the storage system. The percentage is calculated across all deduplicated volumes in the pools, and it does not include the capacity of volumes that are not deduplicated.

The following formula is used to calculate the amount of storage space that is saved:

written capacity - deduplicated size

The following formula is used to calculate the percentage of capacity that is saved:

$((\text{written capacity} - \text{deduplicated size}) \div \text{written capacity}) \times 100$

For example, the written capacity, which is the amount of data that is written to the volumes before deduplication, is 40 GiB. The deduplicated size, which reflects the size of deduplicated data that is written to disk, is just 10 GB. Therefore, data deduplication reduced the size of the data that is written by 75%.

Note: The Deduplication Savings (%) metric is available for IBM FlashSystem A9000, IBM FlashSystem A9000R, and resources that run IBM Storage Virtualize 8.1.3 or later.

Drive Compression Savings (%) shows the amount and percentage of capacity that is saved with drives that use inline data compression technology. The percentage is calculated across all compressed drives in the pools.

The amount of storage space that is saved is the sum of drive compression savings.

The following formula is used to calculate the percentage of capacity that is saved:

$((\text{used written capacity} - \text{compressed size}) \div \text{used written capacity}) \times 100$

Note: The Drive Compression Savings (%) metric is available for storage systems that contain FCMs with hardware compression.

Mapped Capacity (GiB) shows the total volume space in the storage system that is mapped or assigned to host systems, including child pool capacity.

Note: Mapped Capacity (GiB) was previously known as Assigned Volume Space.

Overprovisioned Capacity (GiB) shows the capacity that cannot be used by volumes because the physical capacity of the pools cannot meet the demands for provisioned capacity. The following formula is used to calculate this value:

[Provisioned Capacity - Capacity]

Note: Overprovisioned Capacity (GiB) was previously known as Unallocatable Volume Space.

Shortfall (%) shows the difference between the remaining unused volume capacity and the available capacity of the associated pool, which is expressed as a percentage of the remaining unused volume capacity. The shortfall represents the relative risk of running out of space for overallocated thin-provisioned volumes. If the pool has sufficient available capacity to satisfy the remaining unused volume capacity, no shortfall exists. As the remaining unused volume capacity grows or as the available pool capacity decreases, the shortfall increases, and the risk of running out of space becomes higher. If the available capacity of the pool is exhausted, the shortfall is 100%, and any volumes that are not yet fully allocated have run out of space.

If the pool is not thin-provisioned, the shortfall percentage equals zero. If the shortfall percentage is not calculated for the storage system, the field is left blank.

The following formula is used to calculate this value:

[Overprovisioned Capacity ÷ Committed but Unused Capacity]

You can use this percentage to determine when the amount of over-committed space in a pool is at a critically high level. Specifically, if the physical space in a pool is less than the committed provisioned capacity, then the pool does not have enough space to fulfill the commitment to provisioned capacity. This value represents the percentage of the committed provisioned capacity that is not available in a pool. As more space is used over time by volumes while the pool capacity remains the same, this percentage increases.

For example, the remaining physical capacity of a pool is 70 GiB, but 150 GiB of provisioned capacity was committed to thin-provisioned volumes. If the volumes are using 50 GiB, then 100 GiB is still committed to the volumes (150 GiB - 50 GiB) with a shortfall of 30 GiB (70 GiB remaining pool space - 100 GiB remaining commitment of volume space to the volumes). Because the volumes are overcommitted by 30 GiB based on the available capacity in the pool, the shortfall is 30% when the following calculation is used:

$$\frac{[(100 \text{ GiB unused volume capacity} - 70 \text{ GiB remaining pool capacity}) \div 100 \text{ GiB unused volume capacity}] \times 100}{}$$

Note: Shortfall (%) is available for DS8000, Hitachi Virtual Storage Platform, and storage systems that run IBM Storage Virtualize.

For IBM FlashSystem A9000 and IBM FlashSystem A9000R, this value is not available.

Provisioned Capacity (%) shows the percentage of the physical capacity that is committed to the provisioned capacity of the volumes in the pools. If the value exceeds 100%, the physical capacity does not meet the demands for provisioned capacity. To calculate the provisioned capacity percentage, the following formula is used:

$$[(\text{provisioned capacity} \div \text{pool capacity}) \times 100]$$

For example, if the provisioned capacity percentage is 200% for a storage pool with a physical capacity of 15 GiB, then the provisioned capacity that is committed to the volumes in the pools is 30 GiB.

Twice as much space is committed to the pools than is physically available to the pools. If the provisioned capacity percentage is 100% and the physical capacity is 15 GiB, then the provisioned capacity that is committed to the pools is 15 GiB. The total physical capacity that is available to the pools is used by the volumes in the pools.

A provisioned capacity percentage that is higher than 100% is considered to be aggressive because insufficient physical capacity is available to the pools to satisfy the allocation of the committed space to the compressed and thin-provisioned volumes in the pools. In such cases, you can check the Shortfall (%) value to determine how critical the shortage of space is for the storage system pools.

Note: Provisioned Capacity (%) was previously known as Virtual Allocation.

Provisioned Capacity (GiB) shows the total amount of provisioned capacity of volumes within the pool. If the pool is a parent pool, it also includes the storage space that can be made available to the volumes in the child pools.

Note: Provisioned Capacity (GiB) was previously known as Total Volume Capacity.

Safeguarded Capacity (GiB) shows the total amount of capacity that is used to store volume backups that are created by the Safeguarded Copy feature in DS8000.

Total Capacity Savings (%) shows the estimated amount and percentage of capacity that is saved by using data deduplication, pool compression, thin provisioning, and drive compression, across all volumes in the pool.

The following formula is used to calculate the amount of storage space that is saved:

Provisioned Capacity ? Used Capacity

The following formula is used to calculate the percentage of capacity that is saved:

$((\text{Provisioned Capacity} - \text{Used Capacity}) \div \text{Provisioned Capacity}) \times 100$

Note: Total Capacity Savings (%) was previously known as Total Data Reduction Savings, and it is available for IBM FlashSystem A9000 and IBM FlashSystem A9000R, IBM Spectrum Accelerate, XIV storage systems with firmware 11.6 or later, and resources that run IBM Storage Virtualize.

Unmapped Capacity (GiB) shows the total amount of space in the volumes that are not assigned to hosts.

Note: Unmapped Capacity (GiB) was previously known as Unassigned Volume Space.

In the *Zero Capacity* column (see Figure 9-56 on page 603) on the Pools page, you can see the date, based on the storage usage trends for the pool, when the pool will run out of available capacity.

Zero Capacity: The capacity information that is collected over 180 days is analyzed to determine, based on historical storage consumption, when the pools will run out of capacity. The pools that ran out of capacity are marked as depleted. For the other pools, a date is provided so that you know when the pools are projected to run out of capacity.

If sufficient information is not collected to analyze the storage usage of the pool, None is shown as the value for zero capacity. If a capacity limit is set for the pool, the date that is shown in the Zero Capacity column is the date when the available capacity based on the capacity limit will be depleted.

For example, if the capacity limit for a 100 GiB pool is 80%, it is the date when the available capacity of the pool is less than 20 GiB. Depleted is shown in the column when the capacity limit is reached.

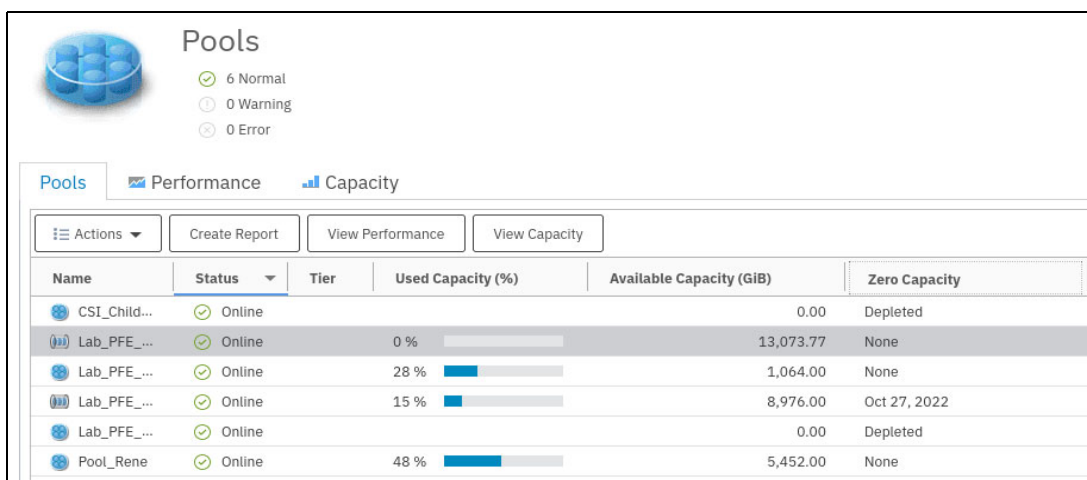


Figure 9-56 Zero Capacity

The following values can be shown in the Zero Capacity column:

- ▶ A date
 - The data that is based on space usage trends for the pool when the capacity runs out (projected).
- ▶ None
 - Based on the current trend, no date can be calculated for when the pool is to be filled (for example, if the trend is negative) as data is moved out of the pool.
- ▶ Depleted
 - The pool is full.

The following metrics can be added to capacity charts for storage systems within capacity planning. Use the charts to detect capacity shortages and space usage trends.

- ▶ *Available Repository Capacity (GiB)* shows the available, unallocated storage space in the repository for Track Space-Efficient (TSE) thin-provisioning.

Note: Available for DS8000 thin-provisioned pools.

- ▶ *Soft Capacity (GiB)* shows the amount of virtual storage space that is configured for the pool.

Note: Soft Capacity (GiB) is available for XIV systems and IBM Spectrum Accelerate storage systems.

- ▶ *Available Soft Capacity (GiB)* shows the amount of virtual storage space that is available to allocate to volumes in a storage pool.

Note: Available for XIV systems and IBM Spectrum Accelerate storage systems.

- ▶ *Written Capacity (GiB)* shows the amount of data that is written from the assigned hosts to the volume before compression or data deduplication are used to reduce the size of the data. For example, the written capacity for a volume is 40 GiB. After compression, the volume used space, which reflects the size of compressed data that is written to disk, is only 10 GiB.

Note: Written Capacity (GiB) was previously known as Written Space.

- ▶ *Available Written Capacity (GiB)* shows the amount of capacity that can be written to the pools before inline compression is applied. If the pools are not compressed, this value is the same as Available Capacity.

Note: Available Written Capacity (GiB) was previously known as Effective Used Capacity.

Because data compression is efficient, a pool can run out of Available Written Capacity while physical capacity is still available. To stay aware of your capacity needs, monitor this value and Available Capacity.

- ▶ *Enterprise hard disk drive (HDD) Available Capacity (GiB)* shows the amount of storage space that is available on the Enterprise HDDs that can be used by Easy Tier for retiring the volume extents in the pool.

Note: Enterprise HDD Available Capacity (GiB) is available for DS8000 and storage systems that run IBM Storage Virtualize.

- ▶ *Enterprise HDD Capacity (GiB)* shows the total amount of storage space on the Enterprise HDDs that can be used by Easy Tier for retiring the volume extents in the pool.

Note: Enterprise HDD Capacity (GiB) is available for DS8000 and storage systems that run IBM Storage Virtualize.

- ▶ *Nearline HDD Available Capacity (GiB)* shows the amount of storage space that is available on the Nearline HDDs that can be used by Easy Tier for retiring the volume extents in the pool.

Note: Nearline HDD Available Capacity (GiB) is available for DS8000 and storage systems that run IBM Storage Virtualize.

- ▶ *Nearline HDD Capacity (GiB)* shows the total amount of storage space on the Nearline HDDs that can be used by Easy Tier for retiring the volume extents in the pool.

Note: Nearline HDD Available Capacity (GiB) is available for DS8000 and storage systems that run IBM Storage Virtualize.

- ▶ *Repository Capacity (GiB)* shows the total storage capacity of the repository for Track Space-Efficient (TSE) thin-provisioning.

Note: Repository Capacity (GiB) is available for DS8000 thin-provisioned pools.

- ▶ *Reserved Volume Capacity* shows the amount of pool capacity that is reserved but is not yet used to store data on the thin-provisioned volume.

Note: Reserved Volume Capacity was known as Unused Space, and it is available for resources that run IBM Storage Virtualize.

- ▶ *SCM Available Capacity (GiB)* shows the available capacity on storage-class memory (SCM) drives in the pool. Easy Tier can use these drives to retier the volume extents in the pool.

Note: SCM Available Capacity (GiB) is available for IBM Storage Virtualize systems, such as IBM FlashSystem 9100, IBM FlashSystem 7200, and IBM Storwize family storage systems that are configured with block storage.

- ▶ *SCM Capacity (GiB)* shows the total capacity on SCM drives in the pool. Easy Tier can use these drives to retier the volume extents in the pool.

Note: SCM Capacity (GiB) is available for IBM Storage Virtualize systems, such as IBM FlashSystem 9100, IBM FlashSystem 7200, and IBM Storwize family storage systems that are configured with block storage.

- ▶ *Tier 0 Flash Available Capacity (GiB)* shows the amount of storage space that is available on the Tier 0 flash solid-state drives (SSDs) that can be used by Easy Tier for retiering the volume extents in the pool.

Note: Tier 0 Flash Available Capacity (GiB) is available for DS8000 and storage systems that run IBM Storage Virtualize.

- ▶ *Tier 0 Flash Capacity (GiB)* shows the total amount of storage space on the Tier 0 flash SSDs that can be used by Easy Tier for retiering the volume extents in the pool.

Note: Tier 0 Flash Capacity (GiB) is available for DS8000 and storage systems that run IBM Storage Virtualize.

- ▶ *Tier 1 Flash Available Capacity (GiB)* shows the amount of storage space that is available on the Tier 1 flash, which is read-intensive (RI) SSDs that can be used by Easy Tier for retiering the volume extents in the pool.

Note: Tier 1 Flash Available Capacity (GiB) is available for DS8000 and storage systems that run IBM Storage Virtualize.

- ▶ *Tier 1 Flash Capacity (GiB)* shows the total amount of storage space on the Tier 1 flash, which is RI SSDs that can be used by Easy Tier for retiring the volume extents in the pool.

Note: Tier 1 Flash Capacity (GiB) is available for DS8000 and storage systems that run IBM Storage Virtualize.

- ▶ *Tier 2 Flash Available Capacity (GiB)* shows the available capacity on Tier 2 flash, which are high-capacity drives in the pool. Easy Tier can use these drives to retire the volume extents in the pool.

Note: Tier 2 Flash Available Capacity (GiB) is available for DS8000 storage systems.

- ▶ *Tier 2 Flash Capacity (GiB)* shows the total capacity on Tier 2 flash, which is high-capacity drives in the pool. Easy Tier can use these drives to retire the volume extents in the pool.

Note: Tier 2 Flash Capacity (GiB) is available for DS8000 storage systems.

9.4 Creating alerts for IBM Storage Control and IBM Storage Insights

In this section, we provide information about alerts with IBM Spectrum Control and IBM Storage Insights. The no-charge version of IBM Storage Insights does not support alerts.

New data reduction technologies add intelligence and capacity savings to your environment. If you use data reduction on different layers, such as hardware compression in the IBM FlashSystem 9500 FCMs (if a IBM FlashSystem 9500 is virtualized by the SVC) and in the DRPs, ensure that you do not have insufficient space in the back-end storage device.

First, distinguish between thin-provisioning and over-allocation (over-provisioning). Thin-provisioning is a method for optimizing the usage of available storage. It relies on allocation of blocks of data on-demand versus the traditional method of allocating all the blocks up front. This method eliminates almost all white space, which helps avoid the poor usage rates (often as low as 10%) that occur in the traditional storage allocation method. Traditionally, large pools of storage capacity are allocated to individual servers, but remain unused (not written to).

Over-provisioning means that in total that more space is being assigned and promised to the hosts. They can possibly try to store more data on the storage subsystem as physical capacity is available, which results in an out-of-space condition.

Remember: You must constantly monitor your environment to avoid over-provisioning situations that can be harmful to the environment and can cause access loss.

Keep at least 15% free space for garbage collection in the background. For more information, see 4.1.2, “Data reduction pools” on page 249.

Data reduction technologies conserve some physical space. If the space that is used for the data can be reduced, the conserved space can be used for other data. Depending on the type of data, deletion might not free up much space.

Imagine that you have three identical or almost identical files on a file system that were deduplicated. This issue resulted in getting a good compression ratio (CR) (three files, but stored only once). If you now delete one file, you do not gain more space because the deduplicated data must stay on the storage (because two other versions refer to the data). Similar results can be seen when several FlashCopies of one source are used.

9.4.1 Alert examples

Table 9-2 shows an alert for SVC based on the pool level.

Table 9-2 Event examples for SAN Volume Controller

System	Entity	Resource type	Event
SVC	Pool	Used Pool Capacity	Used Capacity >= nn%

Other alerts are possible as well, but generally percentage alerts are best suited because the alert definition applies to all pools in a storage system.

9.4.2 Alert example to monitor pool capacity: Used Capacity

The following example shows how to create an alert to get status information about the remaining physical space in an IBM Storage Virtualize system.

Assign a severity to an alert. Assigning a severity can help you more quickly identify and address the critical conditions that are detected on resources. The severity that you assign depends on the guidelines and procedures within your organization. Default assignments are provided for each alert.

Table 9-3 lists the possible alert severities.

Table 9-3 Alert severities

Option	Description
Critical	The alert is critical and must be resolved. For example, alerts that notify you when the amount of available space on a file system falls below a specified threshold.
Warning	Alerts that are not critical, but represent potential problems. For example, alerts that notify you when the status of a data collection job is not normal.
Informational	Alerts that might not require any action to resolve and are primarily for informational purposes. For example, alerts that are generated when a new pool is added to a storage system.

In this example, we created the following thresholds:

- ▶ Critical (95% space usage in the pool)
- ▶ Warning (90% space usage in the pool)
- ▶ Information (85% space usage in the pool)

Adjust the percentage levels to the required levels as needed. The process to extend storage might take some time (ordering, installation, provisioning, and so on).

The advantage of this way of setting up an Alert Policy is that you can add various IBM Storage Virtualize systems to this customized alert.

Figure 9-57 shows how to start creating an Alert Policy in IBM Spectrum Control.

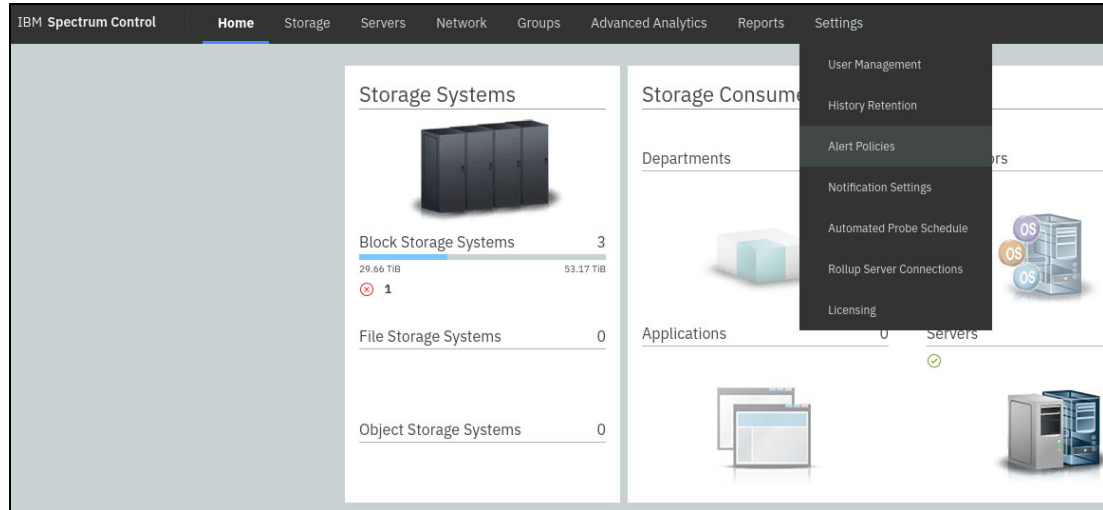


Figure 9-57 IBM Spectrum Control Alert policies

For IBM Storage Insights, Figure 9-58 shows how to start creating an Alert Policy.

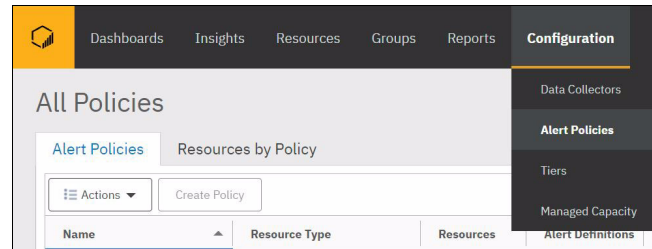


Figure 9-58 IBM Storage Insights Alert policies

The following example shows how to create an Alert Policy by copying the existing policy. You also might need to change an existing Alert Policy (in our example, the Default Policy). Consider that a storage subsystem can be active in only one Alert Policy.

Figure 9-59 on page 609 shows the Default IBM FlashSystem Family policy in IBM Spectrum Control 5.4.6.

Name	Resource Type	Resources	Alert Definitions
Default FlashSystem Family policy	FlashSystem Family	0	24
Default FlashSystem Family policy Custom	FlashSystem Family	1	24
Default Linux policy	Linux	1	13
Default Storwize policy	Storwize	2	24

Figure 9-59 All alert policies in IBM Spectrum Control

Note: Unless otherwise noted, IBM Storage Insights and IBM Spectrum Control do not differ for the steps that are described next.

Figure 9-60 describes how to copy a policy to create one. Hover your cursor over the policy that you want to copy, click the left mouse button, and select **Copy Policy**.

Name	Resource Type	Resources	Alert Definitions
Default FlashSystem Family policy	FlashSystem Family	0	24
Default FlashSystem Family policy Custom	FlashSystem Family	1	24
Default Linux policy	Linux	1	13
Default Storwize policy	Storwize	2	24

Figure 9-60 Copying a policy in IBM Spectrum Control

Figure 9-61 on page 610 shows how to rename the previously copied policy. The new policy is stored as another policy. One IBM Storage Virtualize system can be added to a single policy only. You can add the system later if you are unsure now (optionally, select **Resource**, and then select the option).

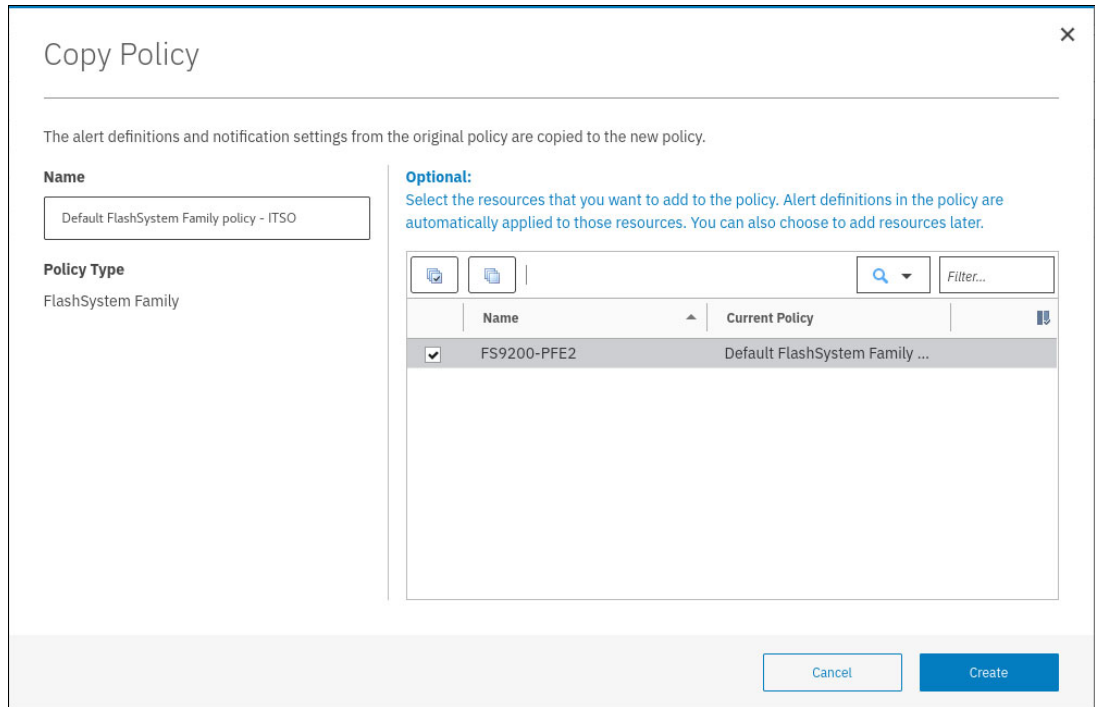


Figure 9-61 Copy Policy window

Figure 9-62 shows the newly created Alert Policy Default IBM FlashSystem Family policy - ITS0 with all alerts that were inherited from the default policy.

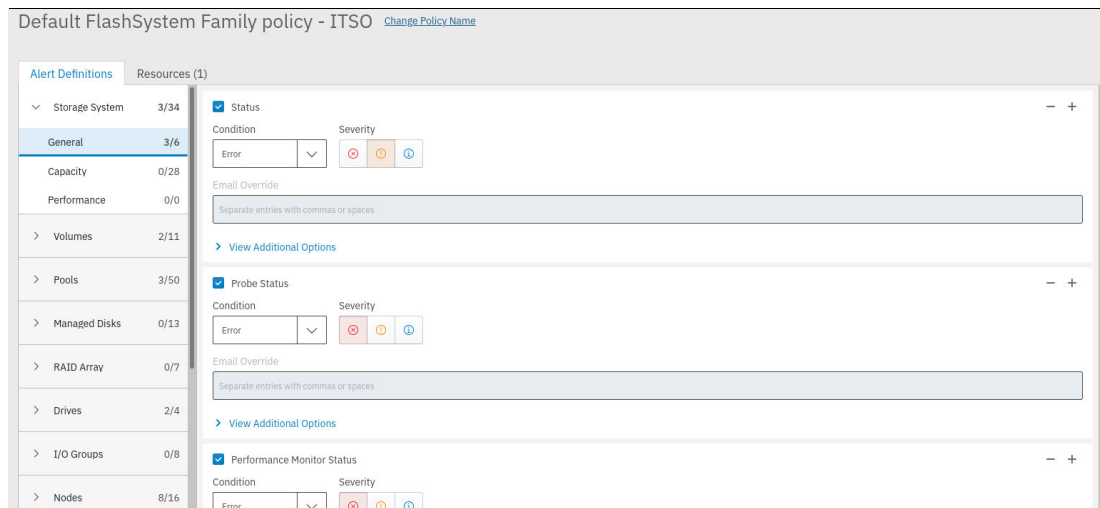


Figure 9-62 New policy with inherited alert definitions

Figure 9-63 shows how to choose the required alert definitions by selecting **Pool** → **Capacity**.

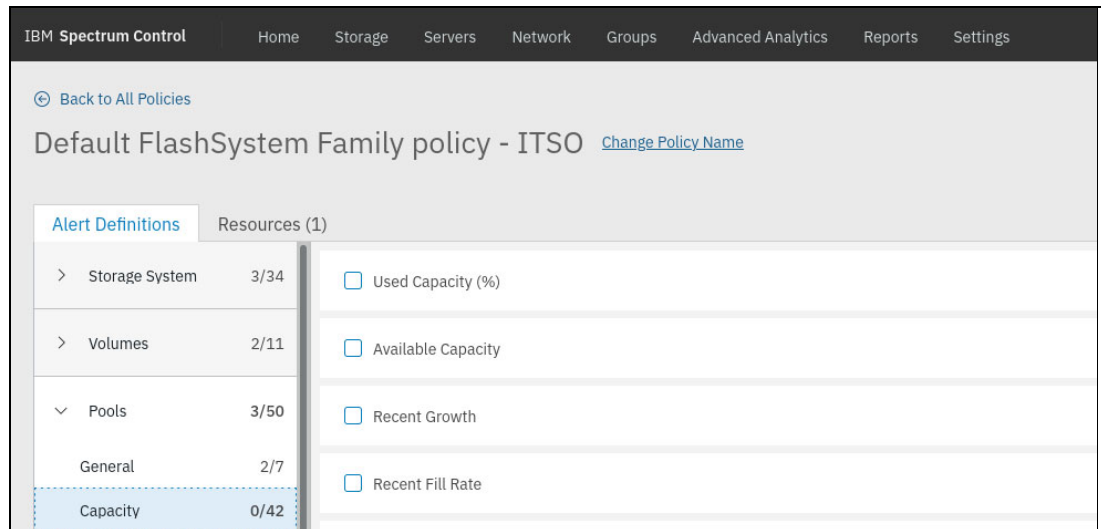


Figure 9-63 Choosing the required alert definitions

Figure 9-64 denotes the tasks for setting up the Critical definition by monitoring the Used Capacity (%) and releasing Policy Notifications at 95%.

Pre-defined notification methods can be one of the following options:

- ▶ Email addresses
- ▶ SNMP
- ▶ IBM Netcool/OMNibus
- ▶ Windows Event Log or UNIX syslog
- ▶ Run script

These methods must be defined before you can choose them if your environment does not include pre-defined methods.

Figure 9-64 shows how to set the operator, value, and severity for the alert. It also shows how to modify the notification frequency and select notification methods.

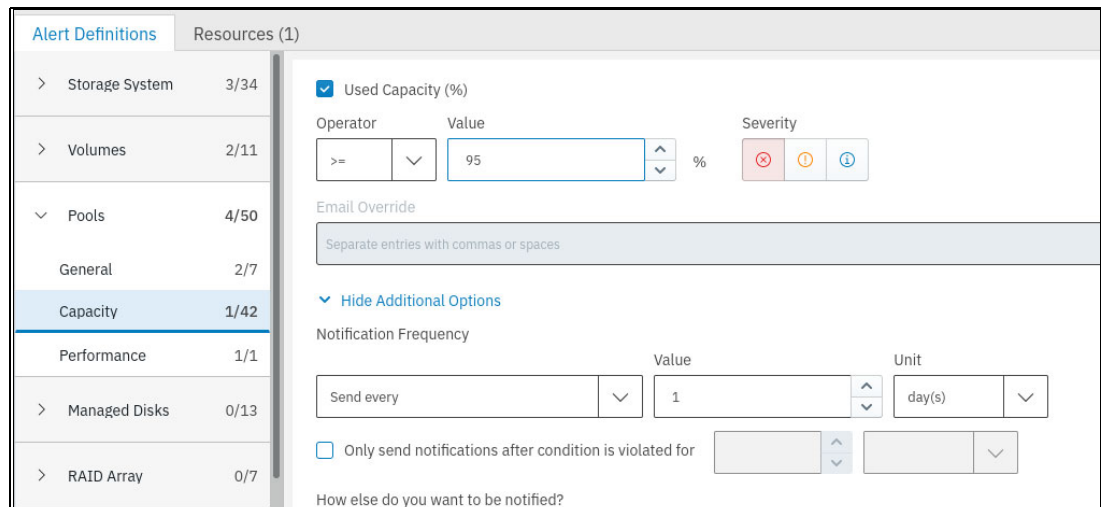


Figure 9-64 Alert parameters

Figure 9-65 shows how to set up the Warning level at 90% for Used Capacity (%). To proceed, choose the plus sign at the previously defined Definition (Critical) and complete the information, as shown in Figure 9-65 (Operator: ">=", Value: "90%", and Severity "Warning").

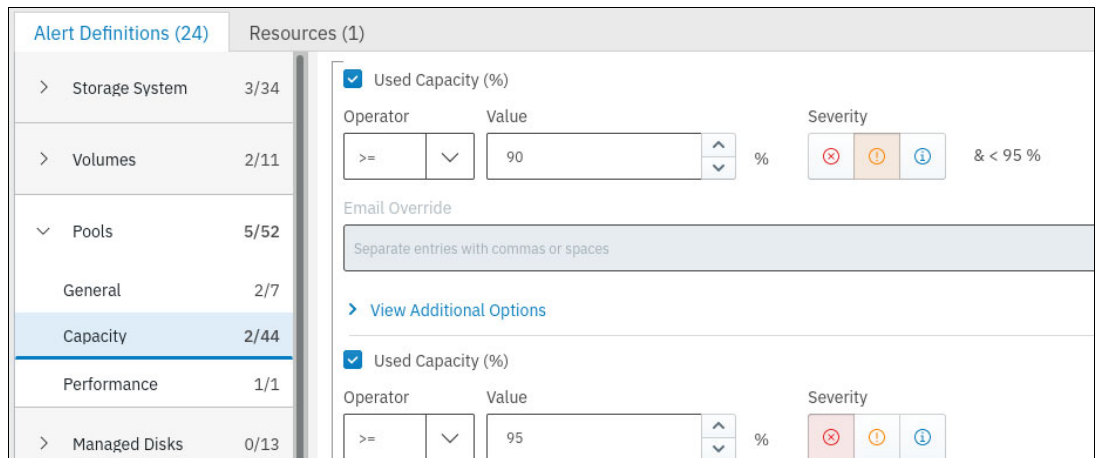


Figure 9-65 Setting up the Warning level

Figure 9-66 shows how to set up the Informational Threshold at 85%.

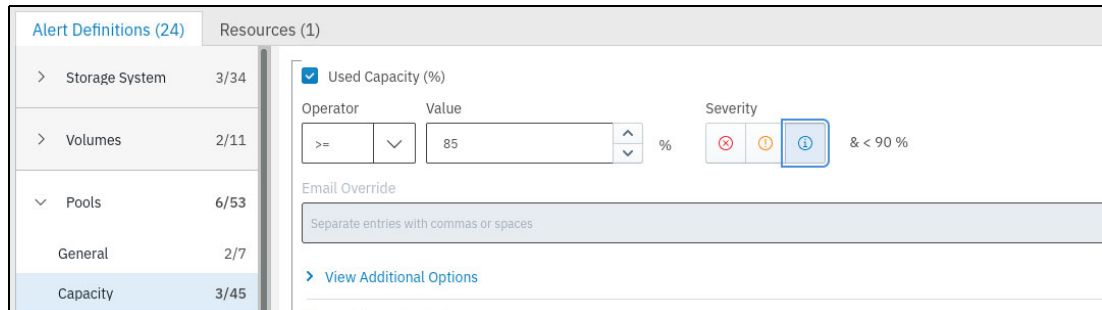


Figure 9-66 Setting up the informational threshold

Figure 9-67 shows how to open the Notification Settings in IBM Spectrum Control.

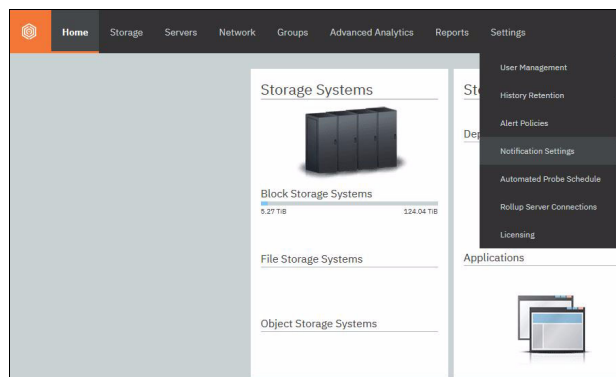


Figure 9-67 IBM Spectrum Control notification settings

Note: With IBM Storage Insights, you can send notifications with the email method only. Call Home information is also leveraged to provide more accurate predictions of when a DRP or a compressing array will run out of space. These notifications are pushed to the **Insights** → **Advisor** view in IBM Storage Insights, as described in Figure 9-40 on page 587.

9.5 Health monitoring

In this section, we review methods to monitor the health status of the system.

9.5.1 Health monitoring in the IBM Storage Virtualize GUI

This section covers the health monitoring in the IBM Storage Virtualize GUI.

System health

By using the management GUI dashboard, you can detect errors in the System Health page.

System components are separated into the following categories:

- ▶ *Hardware components* display the health of all components that are specific to the physical hardware.
- ▶ *Logical components* display the health of all logical and virtual components in the management GUI.
- ▶ *Connectivity components* display the health of all components that are related to the system's connectivity and the relationship between other components or systems.

For more information, see [System Health Tiles for SAN Volume Controller](#) and [System Health Tiles for FlashSystem 9x00](#).

There are tiles for each subset of component within each category that shows the health state of the category.

Tiles with errors and warnings are displayed first so that components that require attention have higher visibility. Healthy pages are sorted in order of importance in day-to-day use.

The System Health page in Figure 9-68 shows the three categories of system components.



Figure 9-68 System Health state of management GUI Dashboard

By expanding the Hardware Components page, you can see the type of hardware components and the respective health states, as shown in Figure 9-69 on page 614.

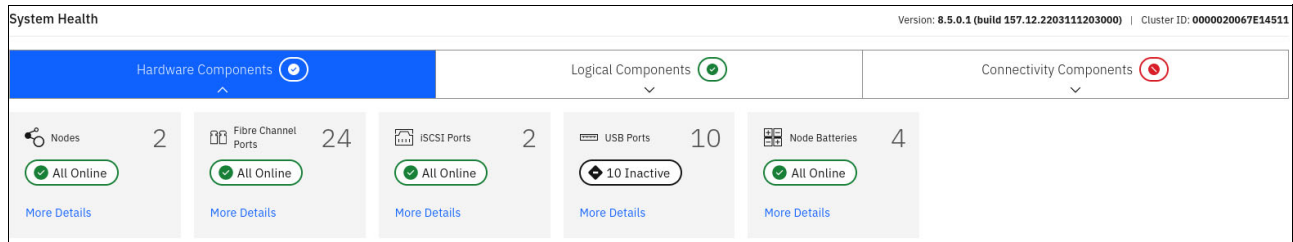


Figure 9-69 Expanded Hardware Components view for a SAN Volume Controller Cluster

Note: The More Details view shows a tabular view of individual components with more detail. The number of tiles also might vary between systems. For example, Figure 9-70 shows that an enclosure-based system typically has more types of hardware components compared to SVC systems.

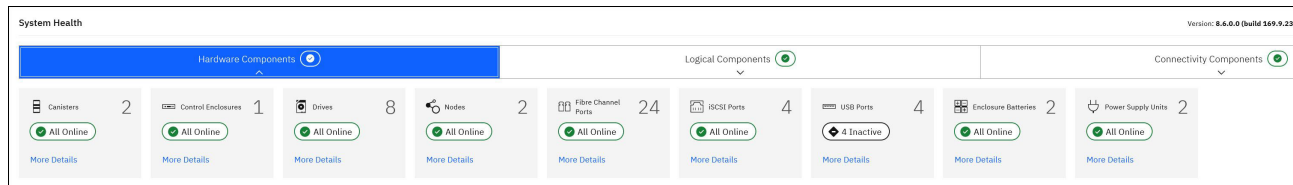


Figure 9-70 Expanded Hardware Components view for IBM FlashSystem 9100

The pages that have errors or warnings sort the tiles in an order that draws the most attention to the tiles that are not optimal. For example, in Figure 9-71, Call Home and Support Assistance are in the error status and appear at the left.

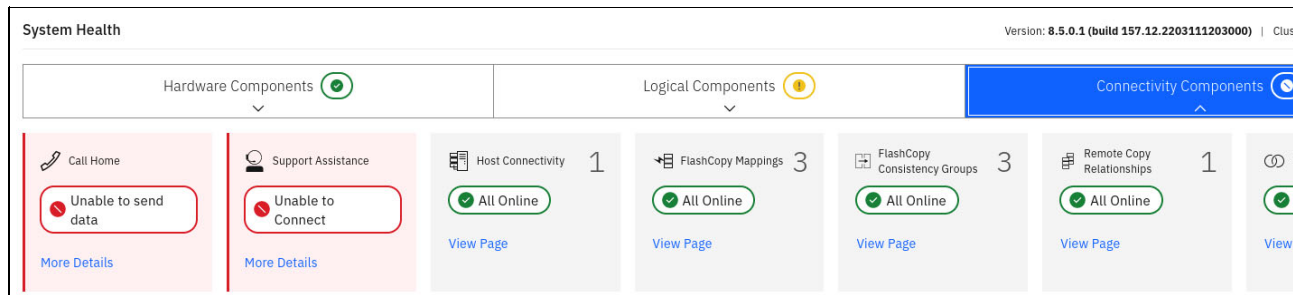


Figure 9-71 Prioritizing tiles that need attention

Event log

A user might become aware of a problem on a system through active monitoring of the System Health dashboard or by receiving an alert through one of the configured notification methods.

The dashboard is intended to get the user's attention, and it is an entry point that directs the user to the relevant event in **Monitoring** → **Events** and the associated fix procedure.

For example, Figure 9-71 displays a status of *Call Home - Unable to send data*.

Clicking **More Details** leads the customer to the specific event log entry, as shown in Figure 9-72 on page 615. The Run Fix Procedure option provides instructions that the user can follow to resolve the issue.

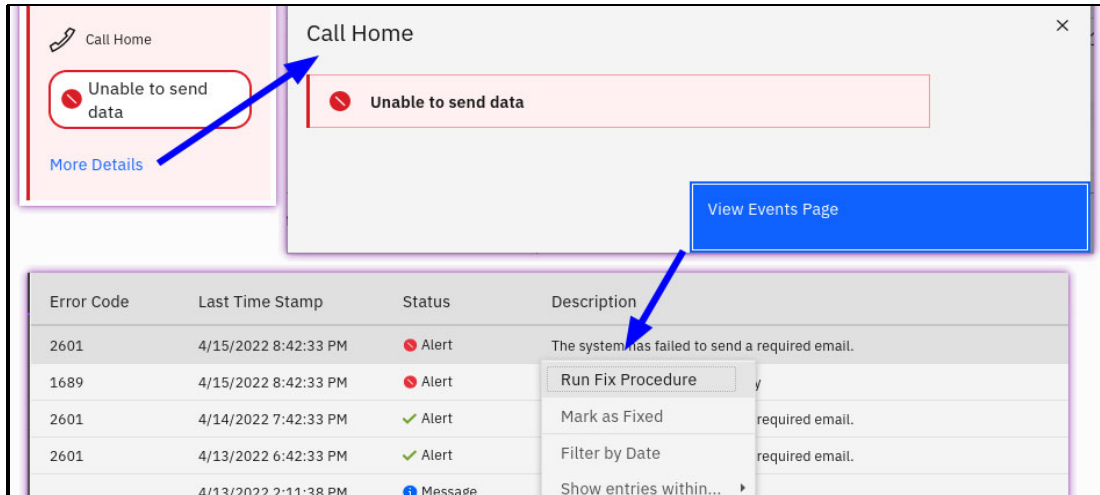


Figure 9-72 Dashboard entry point drills down to the event log

The Events by Priority icon in the upper right of the GUI navigation area also provides a similar entry into events that need attention, as shown in Figure 9-73.

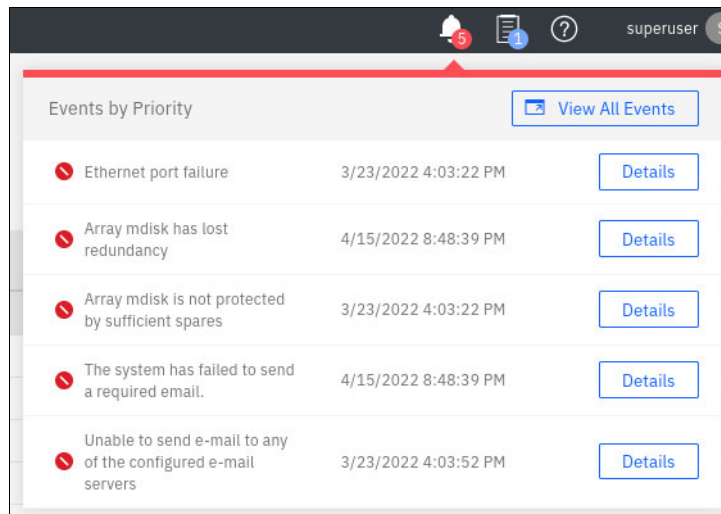


Figure 9-73 Events by Priority

9.5.2 Health monitoring through the IBM Storage Virtualize command-line interface (CLI)

This section covers the health monitoring through the IBM Storage Virtualize command-line interface (CLI).

Port monitoring

With introducing IBM Storage Virtualize 8.4.0, a new command was added to the CLI to monitor the node port error counters.

Use the `lspportstats` command to view the port transfer and failure counts and Small Form-factor Pluggable (SFP) diagnostics data that is recorded in the statistics file for a node.

Example 9-13 shows the TX and RX values, as well as the zero buffer-to-buffer credits and the amount of CRC errors for port 3 of node 1.

Example 9-13 CLI output example for lspportstats command to show the TX & RX power

```

IBM_FlashSystem:FS9xx0:superuser>lspportstats -node node1 3|grep
fc_wwpn&&lspportstats -node node1 3|grep power&&lspportstats -node node
1 3|grep buffer-buffer&&lspportstats -node node1 3|grep CRC
fc_wwpn                    ="0x500507681013xxxx"
TX power (uW)              ="588"
TX power low alarm threshold ="126"
RX power (uW)              ="544"
RX power low alarm threshold ="31"
zero buffer-buffer credit timer (uS) ="0"
invalid CRC error count    ="0"

```

9.5.3 Health monitoring in IBM Spectrum Control

The IBM Spectrum Control Dashboard provides a summary for each type of device that it monitors. For example, in the upper left of Figure 9-74, we can see that IBM Spectrum Control is monitoring three block storage systems and that one of them is in the *Error* state.

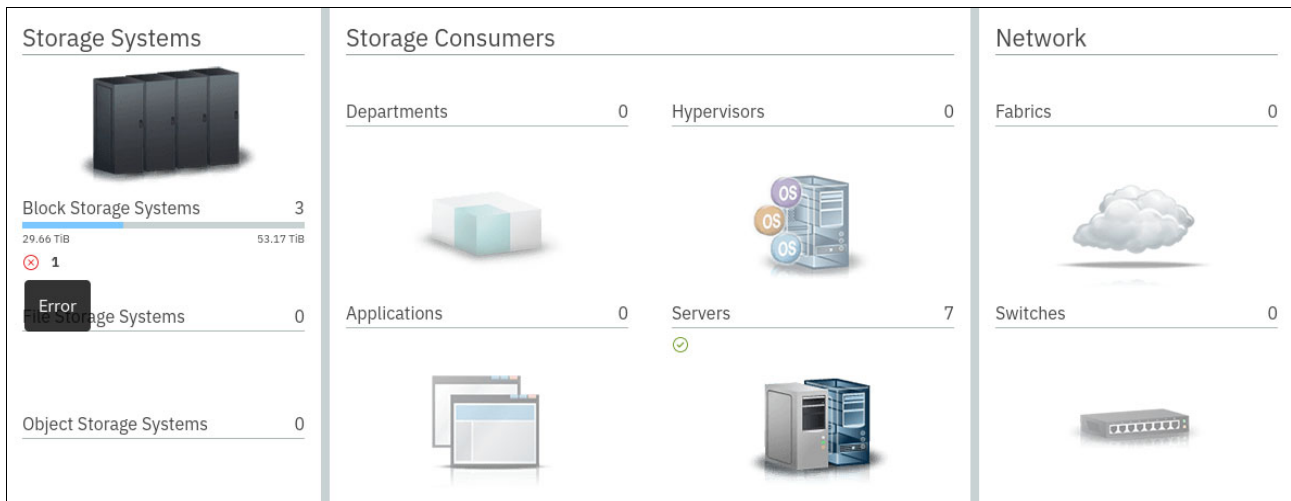


Figure 9-74 IBM Spectrum Control Dashboard summary

Clicking **Block Storage Systems** takes you to the list of monitored systems to identify the specific system that is in the *Error* state. For example, in Figure 9-75 on page 617, you can see that the first system in the list is in the *Error* state.

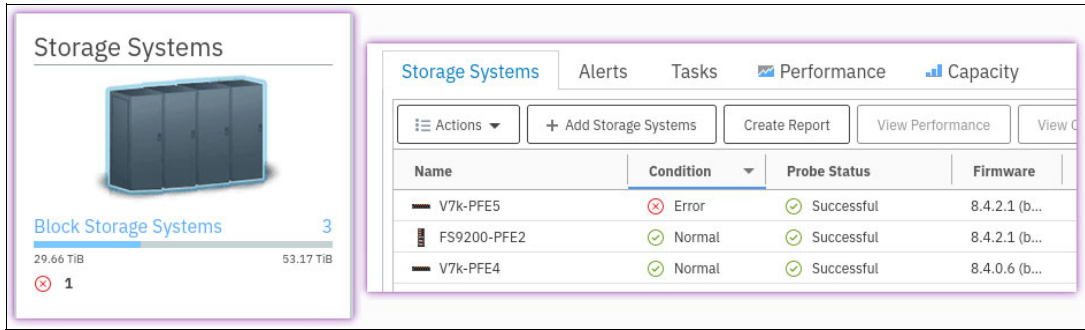


Figure 9-75 IBM Spectrum Control Block Storage Systems

For more information about the error condition, right-click the system and select **View Details**, which displays the list of internal components and their status. Figure 9-76 shows that the error status is due to a problem with one or more volumes.

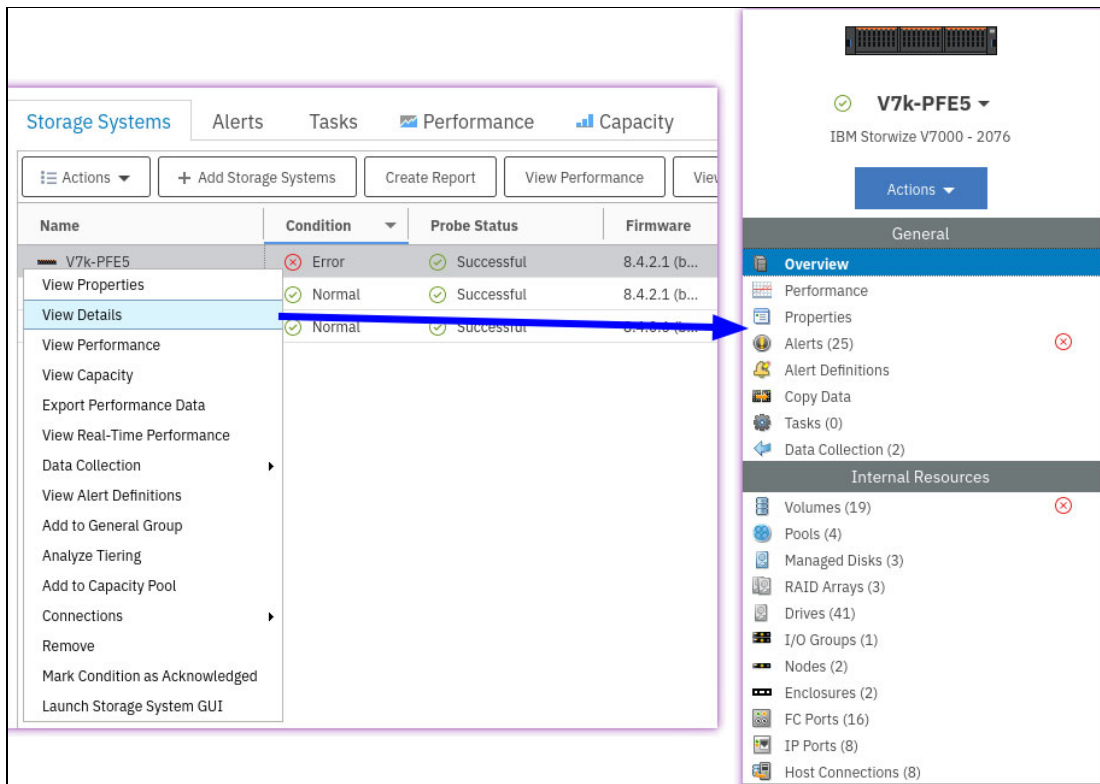


Figure 9-76 Detailed Block Storage System view

Clicking **Volumes** reveals the offending volumes, as shown in Figure 9-77 on page 618.

Note: Use the column sorting function or filter to display only the objects of interest.

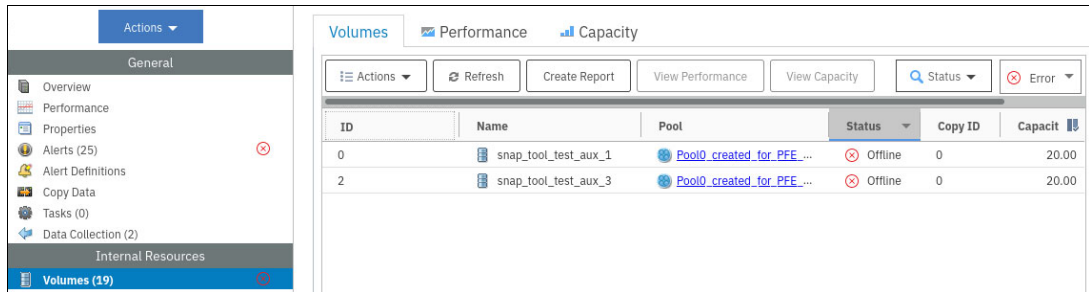


Figure 9-77 Offline volumes

In this specific case, the offline state is expected because the volumes are auxiliary volumes of *inconsistent stopped* Global Mirror (GM) relationships. Therefore, the status can be marked as *acknowledged*, as shown in Figure 9-78.

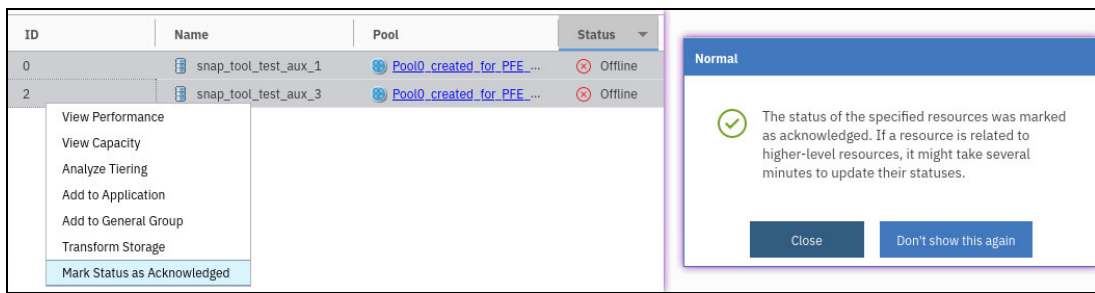


Figure 9-78 Marking the status as acknowledged

The system and volume status no longer reports the *Error* state, as shown in Figure 9-79.

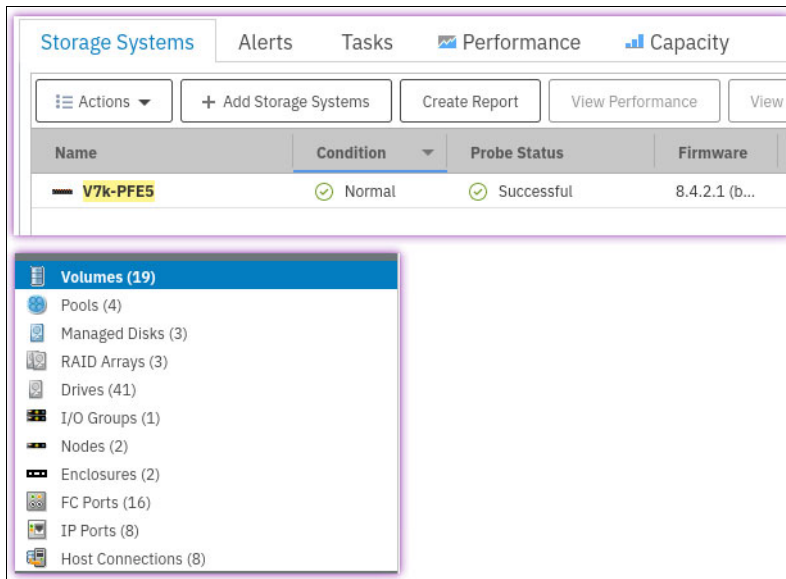


Figure 9-79 Error status cleared

Note: If IBM Spectrum Control and IBM Storage Insights are monitoring the environment, the acknowledgment must be set in both instances.

Other use cases might exist in which you must replace hardware after you open a ticket in your internal system with the vendor. In these instances, you still acknowledge the status so that any other errors change the storage system from green to red again and you see that a second event occurred.

9.5.4 Health monitoring in IBM Storage Insights

Monitoring in IBM Storage Insights is essentially the same as IBM Spectrum Control with some minor cosmetic differences.

One significant difference is that IBM Storage Insights has an Operations dashboard, a NOC dashboard, and custom dashboards that can show the health status in a grid or list of tiles. When the system of interest is selected, the detailed view shows a Component Health page, which is similar in appearance to the corresponding view in the IBM Storage Virtualize management GUI.

For example, Figure 9-80 shows that the system is in the *Error* state because of a volume error.

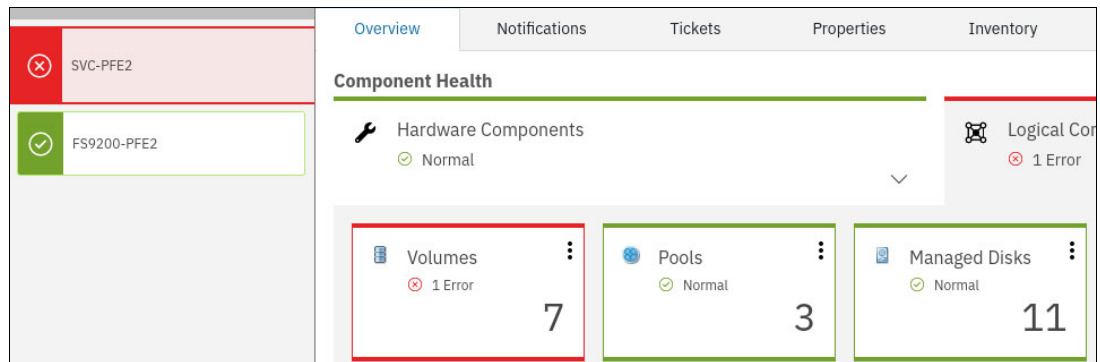


Figure 9-80 IBM Storage Insights dashboard showing a volume error

The Volumes tile in the *Error* state allows two possible actions, as shown in Figure 9-81:

- ▶ **View List:** Similar to IBM Spectrum Control, as shown in Figure 9-77 on page 618.
- ▶ **Launch Storage System GUI:** Launches the system’s native GUI to identify the source of the error, as shown in 9.5.1, “Health monitoring in the IBM Storage Virtualize GUI” on page 613.

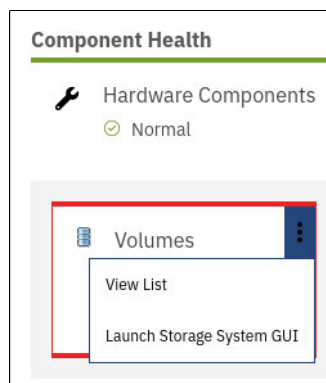


Figure 9-81 Actions available from the Volume tile

9.6 Important performance metrics

IBM Storage Virtualize systems track and record a number of raw metrics. These raw metrics are also transmitted to monitoring systems like IBM Spectrum Control and IBM Storage Insights. These monitoring systems derive metrics that are more meaningful.

Table 9-4 lists some useful metrics for assessing performance problems at the volume level. A few typical use cases are as follows:

- ▶ Diagnose the elevated volume write response time that is caused by GM.
- ▶ Diagnose the elevated volume write response time that is caused by slowness in IBM Real-time Compression (RtC).

Table 9-4 VDisk and volume metrics (front end)

Metric name	Description	Unit
Volume workload metrics		
▶ Read I/O rate	▶ The average number of read operations per second. This value includes both sequential and nonsequential read operations.	▶ IOPS
▶ Write I/O rate	▶ The average number of write operations per second. This value includes both sequential and nonsequential write operations. Also includes write I/Os on remote copy secondary volumes (except HyperSwap).	▶ IOPS
▶ Total I/O rate	▶ Read I/O rate, write I/O rate, and unmap I/O rate combined.	▶ IOPS
▶ Unmap I/O rate	▶ The average number of unmap operations per second. This metric corresponds to the collected uo statistic.	▶ IOPS
▶ Read data rate	▶ The average number of MiBs per second that are transferred for read operations. Does not include data that is moved by using XCOPY.	▶ MBps
▶ Write data rate	▶ The average number of MiBs per second that are transferred for write operations. Also includes write I/Os on remote copy secondary volumes (except HyperSwap). Does not include data that is written by using XCOPY.	▶ MBps
▶ Total data rate	▶ Read data rate, write data rate, and unmap data rate that is combined.	▶ MBps
▶ Unmap data rate	▶ The average number of MiBs per second that were unmapped. This metric corresponds to the collected ub statistic.	▶ MBps
▶ Read transfer size	▶ The average number of KiB that are transferred per read operation.	▶ KB
▶ Write transfer size	▶ The average number of KiB that are transferred per write operation.	▶ KB
▶ Overall transfer size	▶ The average number of KiB that are transferred per I/O operation. This value includes both read/write operations, but excludes unmap operations.	▶ KB
▶ Unmap transfer size	▶ Average size of unmap requests that are received.	▶ KB

Metric name	Description	Unit
Volume latency metrics		
▶ Read response time	▶ The average number of milliseconds to complete a read operation.	▶ ms
▶ Write response time	▶ The average number of milliseconds to complete a write operation.	▶ ms
▶ Overall response time	▶ The average number of milliseconds to complete an I/O operation. This value includes read/write operations.	▶ ms
▶ Unmap response time	▶ The average number of milliseconds that is required to complete an unmap operation. This metric corresponds to the collected ul statistic.	▶ ms
▶ Peak read response time	▶ The worst response time that is measured for a read operation in the sample interval. ^a	▶ ms
▶ Peak write response time	▶ The worst response time measured for a write operation in the sample interval. ^b	▶ ms
▶ Peak unmap response time	▶ The worst response time that is measured for an unmap operation in the sample interval. This metric corresponds to the collected ulw statistic. ^b	▶ ms
▶ Overall Host Attributed Response Time Percentage	▶ The percentage of the average response time that can be attributed to delays from host systems. This value includes both read response times and write response times, and can help you diagnose slow hosts and fabrics that are not working efficiently. For read response time, the value is based on the time that it takes for hosts to respond to transfer-ready notifications from the nodes. For write response time, the value is based on the time that it takes for hosts to send the write data after the node responds to a transfer-ready notification.	▶ %
Volume cache metrics		
▶ Read cache hits	▶ The percentage of all read operations that find data in the cache. This value includes both sequential and random read operations, and read operations in UCA and LCA where applicable.	▶ %
▶ Write cache hits	▶ The percentage of all write operations that are handled in the cache. This value includes both sequential and random write operations, and writes operations in UCA and LCA where applicable.	▶ %
▶ Upper cache (UCA) destage latency average ^b	▶ The average number of milliseconds that it took to complete each destage operation in the volume cache, that is, the time that it took to do write operations from the volume cache to the disk.	▶ μs
▶ Lower cache (LCA) destage latency average ^c	▶ The average number of milliseconds that it took to complete each destage operation in the volume copy cache, that is, the time that it took to do write operations from the volume copy cache to the disk.	▶ μs

Metric name	Description	Unit
Volume GM metrics		
<ul style="list-style-type: none"> ▶ Global Mirror Secondary Writes ▶ Global Mirror Secondary Write Lag 	<ul style="list-style-type: none"> ▶ The number of writes on the GM secondary per second. ▶ This stat has a different meaning depending on whether the VDisk is the primary or the secondary in a GM relationship: <ul style="list-style-type: none"> – Primary: The time between completing a write I/O on the primary (that is, getting the acknowledgment back from cache on the primary) and completing the write on the secondary (that is, getting the acknowledgment from the secondary that the write completed on the secondary). If the secondary write completes before the primary write, then that particular I/O does not contribute to this stat. – Secondary: The time between receiving a write from the primary and submitting the I/O down the stack (to cache). This time includes only the time spent processing the I/O in the remote copy component. It does not include the time to write the data to cache on the secondary. The latter is recorded in the Write Response Time stat. 	<ul style="list-style-type: none"> ▶ IOPS ▶ ms
Global Mirror Overlapping Writes	<ul style="list-style-type: none"> ▶ The number of overlapping writes on a GM primary volume per interval (not per second). An overlapping write is a write where the logical block address (LBA) range of the request collides with another outstanding write I/O to the same LBA range, the write is still outstanding to the secondary site, and the write could not be handled by the journal. 	<ul style="list-style-type: none"> ▶ I/Os per stats interval ^d

- a. The average of peaks in IBM Storage Insights when IBM Spectrum Control is also monitoring the system with a lower statistics interval, for example, 1 minute.
- b. UCA is equivalent to VDisk cache (Vc) in IBM Spectrum Control and IBM Storage Insights.
- c. LCA is equivalent to VDisk copy cache (Vcc) in IBM Spectrum Control and IBM Storage Insights.
- d. IBM Spectrum Control and IBM Storage Insights also report this value as a percentage.

Note: It is important to characterize the host workload before assessing performance. I/O-rate-intensive workloads usually require low latency, and data-rate-intensive workloads aim to achieve maximum throughput. The latter typically also uses higher transfer sizes, so it is not uncommon to observe higher latency for this type of workload.

Table 9-5 lists some useful metrics for assessing performance problems at the MDisk and drive level.

This stage of performance analysis is appropriate when a preliminary analysis at the volume level showed that delays were below lower cache (LCA). A few typical use cases are as follows:

- ▶ Diagnose a back-end overload that is causing elevated volume read response times.
- ▶ Diagnose potential SAN communication problems for external MDisks.

Table 9-5 MDisk and drive metrics

Metric name	Description	Unit
MDisk workload metrics		
<ul style="list-style-type: none"> ▶ Back-end read I/O rate ▶ Back-end write I/O rate ▶ Overall back-end I/O rate 	<ul style="list-style-type: none"> ▶ The number of read I/O commands that are submitted per second to back-end storage. ▶ The number of write I/O commands that are submitted per second to back-end storage. ▶ Back-end read I/O rate and back-end write I/O rate that is combined. 	<ul style="list-style-type: none"> ▶ IOPS ▶ IOPS ▶ IOPS
<ul style="list-style-type: none"> ▶ Back-end read data rate ▶ Back-end write data rate ▶ Overall back-end data rate 	<ul style="list-style-type: none"> ▶ The amount of data that is read per second from back-end storage. ▶ The amount of data that is written per second to back-end storage. ▶ Back-end read data rate and back-end write data rate combined. 	<ul style="list-style-type: none"> ▶ MBps ▶ MBps ▶ MBps
<ul style="list-style-type: none"> ▶ Back-end read transfer size ▶ Back-end write transfer size ▶ Overall back-end transfer size 	<ul style="list-style-type: none"> ▶ The average I/O size of all back-end reads that are submitted within a stats interval. ▶ The average I/O size of all back-end writes that are submitted within a stats interval. ▶ Average of back-end read transfer size and back-end read transfer size. 	<ul style="list-style-type: none"> ▶ KB ▶ KB ▶ KB
MDisk latency metrics		
<ul style="list-style-type: none"> ▶ Back-end read response time ▶ Back-end write response time ▶ Overall back-end response time 	<ul style="list-style-type: none"> ▶ The average number of milliseconds for the back-end storage resources to respond to a read operation.^a ▶ The average number of milliseconds for the back-end storage resources to respond to a write operation.^a ▶ Average of MDisk read response time and MDisk write response time.^a 	<ul style="list-style-type: none"> ▶ ms ▶ ms ▶ ms
Drive workload metrics		
<ul style="list-style-type: none"> ▶ Back-end read I/O rate ▶ Back-end write I/O rate ▶ Overall back-end I/O rate 	<ul style="list-style-type: none"> ▶ The number of read I/O commands that are submitted per second per drive. ▶ The number of write I/O commands that are submitted per second per drive. ▶ Back-end read I/O rate and back-end write I/O rate combined. 	<ul style="list-style-type: none"> ▶ IOPS ▶ IOPS ▶ IOPS

Metric name	Description	Unit
<ul style="list-style-type: none"> ▶ Back-end read data rate ▶ Back-end write data rate ▶ Overall back-end data rate 	<ul style="list-style-type: none"> ▶ The amount of data that is read per second per drive. ▶ The amount of data that is written per second per drive. ▶ Back-end read data rate and back-end write data rate combined. 	<ul style="list-style-type: none"> ▶ MBps ▶ MBps ▶ MBps
<ul style="list-style-type: none"> ▶ Back-end read transfer size ▶ Back-end write transfer size ▶ Overall back-end transfer size 	<ul style="list-style-type: none"> ▶ The average I/O size per drive of reads that are submitted within a stats interval. ▶ The average I/O size per drive of writes that are submitted within a stats interval. ▶ Average of back-end read transfer size and back-end read transfer size. 	<ul style="list-style-type: none"> ▶ KB ▶ KB ▶ KB
Drive latency metrics		
<ul style="list-style-type: none"> ▶ Drive read response time ▶ Drive write response time ▶ Overall drive response time ▶ Drive read queue time ▶ Drive write queue time ▶ Overall drive queue time ▶ Peak drive read response time ▶ Peak drive write response time 	<ul style="list-style-type: none"> ▶ The average number of milliseconds for the drive resources to respond to a read operation. ▶ The average number of milliseconds for the drive resources to respond to a write operation. ▶ Average of drive read response time and drive write response time. ▶ The average number of milliseconds that a read operation spends in the queue before the operation is sent to the drive.^b ▶ The average number of milliseconds that a write operation spends in the queue before the operation is sent to the back-end storage resources.^b ▶ Average of drive read queue time and drive write queue time.^b ▶ The response time of the slowest read per drive in a specific interval.^c ▶ The response time of the slowest write per drive in a specific interval.^c 	<ul style="list-style-type: none"> ▶ ms ▶ ms ▶ ms ▶ ms ▶ ms ▶ ms ▶ ms ▶ ms

a. Includes the latency in a redundant array of independent disks (RAID) for array MDisks.

b. High values here are indicative of an overloaded drive.

c. Peak response times are calculated as an average of the peaks in IBM Storage Insights when the system is also monitored by IBM Spectrum Control with a lower interval, for example, 1 minute.

Note: The concept of abstraction in the IBM Storage Virtualize I/O stack requires careful consideration when evaluating performance problems. For example, the back-end could be overloaded even though the host workload is moderate. Other components within the I/O stack could be generating back-end workload, for example, FlashCopy background copy, Easy Tier extent migration, or DRP garbage collection. It might be necessary to review other metrics that record these workloads at their respective points in the I/O stack. For example, in IBM Storage Insights, Fast-Write Writes Data Rate (Vc) records the workload entering upper cache, Fast-Write Writes Data Rate (Vcc) records the write workload entering lower cache, and Data Movement Rate records the read/write workload of garbage collection. By evaluating the workload at various points, you can determine the cause of back-end overloading.

Table 9-6 lists some useful metrics for assessing performance problems at a node level. A few typical use cases are as follows:

- ▶ Diagnose local internode delays that are causing elevated volume write response time.
- ▶ Diagnose remote internode delays that are causing GM interruptions.
- ▶ Diagnose high CPU core usage that is affecting multiple components in the I/O stack adversely.

Table 9-6 Node metrics

Metric name	Description	Unit
Node workload metrics^a		
<ul style="list-style-type: none"> ▶ Port to Local Node Send Data Rate ▶ Port to Local Node Receive Data Rate ▶ Total Port to Local Node Data Rate 	<ul style="list-style-type: none"> ▶ The actual amount of data tht is sent from the node to the other nodes in the local cluster. ▶ The actual amount of data that is received by the node from the other nodes in the local cluster. ▶ Port to node send data rate and port to node receive data rate combined. 	<ul style="list-style-type: none"> ▶ MBps ▶ MBps ▶ MBps
<ul style="list-style-type: none"> ▶ Port to Remote Node Send Data Rate ▶ Port to Remote Node Receive Data Rate ▶ Total Port to Remote Node Data Rate 	<ul style="list-style-type: none"> ▶ The actual amount of data that is sent from the node to the other nodes in the partner cluster. ▶ The actual amount of data received by the node from the other nodes in the partner cluster. ▶ Port to node send data rate and port to node receive data rate combined. 	<ul style="list-style-type: none"> ▶ MBps ▶ MBps ▶ MBps
Node latency metrics^b		
<ul style="list-style-type: none"> ▶ Port to local node send response time ▶ Port to local node send queue time ▶ Port to local node receive queue time 	<ul style="list-style-type: none"> ▶ The average number of milliseconds to complete a send operation to another node that is in the local cluster. This value represents the external response time of the transfers. ▶ The average time in milliseconds that a send operation spends in the queue before the operation is processed. This value represents the queue time for send operations that are issued to other nodes that are in the local cluster. ▶ The average time in milliseconds that a receive operation spends in the queue before the operation is processed. This value represents the queue time for receive operations that are issued from other nodes that are in the local cluster. 	<ul style="list-style-type: none"> ▶ ms ▶ ms ▶ ms
<ul style="list-style-type: none"> ▶ Port to remote node send response time ▶ Port to remote node send queue time ▶ Port to remote node receive queue time 	<ul style="list-style-type: none"> ▶ The average number of milliseconds to complete a send operation to a node that is in the remote cluster. This value represents the external response time of the transfers. ▶ The average time in milliseconds that a send operation spends in the queue before the operation is processed. This value represents the queue time for send operations that are issued to a node that is in the remote cluster. ▶ The average time in milliseconds that a receive operation spends in the queue before the operation is processed. This value represents the queue time for receive operations that are issued from a node that is in the remote cluster. 	<ul style="list-style-type: none"> ▶ ms ▶ ms ▶ ms

Metric name	Description	Unit
Node CPU utilization metrics		
▶ System CPU utilization	▶ The average percentage of time that the processors on nodes are busy doing system I/O tasks.	▶ %
▶ Compression CPU utilization	▶ The average percentage of time that the processors that are used for data compression I/O tasks are busy. ^c	▶ %
▶ System CPU utilization per core	▶ The approximate percentage of time that a processor core was busy with system I/O tasks.	▶ %
▶ Compression CPU utilization per core	▶ The approximate percentage of time that a processor core was busy with data compression tasks. ^c	▶ %

- a. The logical data rates have separate metrics that include zero data, unmap data, and pointers to other data.
- b. IBM Spectrum Control and IBM Storage Insights categorize these metrics as port latency metrics, but they are available only at the node level.
- c. This metric is relevant only when RtC is in use.

Table 9-7 lists some useful metrics for assessing performance problems at an FC port level. A few typical use cases are as follows:

- ▶ Diagnose ports at their practical bandwidth limit, which causes delayed transfers.
- ▶ Diagnose ports being excluded because of nonzero cyclic redundancy check (CRC) rates.
- ▶ Diagnose MDisk path exclusions due to nonzero CRC rates on ports that are used for back-end connectivity.
- ▶ Diagnose an impending small form-factor pluggable (SFP) failure due to excessive temperature.
- ▶ Diagnose low SFP Rx power, which indicates a potentially defective FC cable.

Table 9-7 Port metrics

Metric name	Description	Unit
FC port workload metrics		
▶ Receive data rate	▶ The average rate at which data is transferred to the port (ingress).	▶ MBps
▶ Send data rate	▶ The average rate at which data is transferred through from the port (egress).	▶ MBps
▶ Total data rate	▶ The sum of receive data rate and send data rate. ^a	▶ MBps
Overall port bandwidth percentage	Approximate bandwidth utilization percentage for send and receive operations per port for a given interval. ^b	%
FC port SFP metrics		
▶ SFP temperature	▶ The temperature of the SFP transceiver plugged into a physical port in degrees Celsius (^o C). Use this metric to watch for fluctuating and high temperatures of an SFP to monitor its environmental health.	▶ ^o C
▶ Tx power	▶ The power in micro watts (μ W) at which the SFP transmits its signal.	▶ μ W
▶ Rx power	▶ The power in micro watts (μ W) at which the SFP receives a signal.	▶ μ W

Metric name	Description	Unit
FC port error rate metrics		
▶ Zero buffer credit percentage.	▶ The amount of time, as a percentage, that the port was not able to send frames between ports because of insufficient buffer-to-buffer credit. In FC technology, buffer-to-buffer credit is used to control the flow of frames between ports.	▶ %
▶ Port send delay time	▶ The average number of milliseconds of delay that occur on the port for each send operation. The reason for these delays might be a lack of buffer credits.	▶ ms
▶ Port send delay I/O percentage	▶ The percentage of send operations where a delay occurred, relative to the total number of send operations that were measured for the port. Use this metric with the Port Send Delay Time metric to distinguish a few long delays from many short delays.	▶ %
▶ CRC error rate	▶ The average number of frames per second that are received in which a cyclic redundancy check (CRC) error is detected. A CRC error is detected when the CRC in the transmitted frame does not match the CRC computed by the receiver.	▶ rate
▶ Invalid transmission word rate	▶ The average number of bit errors per second that are detected.	▶ rate

- a. 8 Gbps and higher fiber FC adapters are full-duplex, so they can send and receive simultaneously at their practical limit. It is more appropriate to evaluate the send or receive metrics separately.
- b. Based on the assumption of full duplex behavior, this metric is approximated by using the maximum of the send and receive bandwidth percentages.

Table 9-8 lists some useful miscellaneous metrics. A few typical use cases are as follows:

- ▶ Diagnose an elevated volume write response time due to a full cache partition.
- ▶ Diagnose elevated CPU core utilization due to aggressive garbage collection.

Table 9-8 Miscellaneous metrics

Metric name	Description	Unit
Node cache fullness metrics		
▶ Max read cache fullness ^a	▶ The maximum amount of fullness for the amount of node memory that is designated for the read cache.	▶ %
▶ Max write cache fullness ^a	▶ The maximum amount of fullness for the amount of node memory that is designated for the write cache.	▶ %
Cache partition fullness metrics		
Max write cache fullness ^b	The maximum amount of the lower cache that the write cache partitions on the nodes that manage the pool are using for write operations. If the value is 100%, one or more cache partitions on one or more pools is full. The operations that pass through the pools with full cache partitions are queued, and I/O response times increase for the volumes in the affected pools.	%

Metric name	Description	Unit
Garbage-collection metrics		
▶ Data movement rate	▶ The capacity, in MiBs per second, of the valid data in a reclaimed volume extent that garbage collection moved to a new extent in the DRP on the node. The valid data must be moved so that the whole extent can be freed up or reused to write new data. This metric corresponds to the collected mm statistic. ^a	▶ MBps
▶ Recovered capacity rate	▶ The capacity in number of MiBs per second that was recovered by garbage collection for reuse in the DRPs on the node. This metric corresponds to the collected rm statistic. ^c	▶ MBps

a. Measured at a node level.

b. Measured at a pool or partition level.

c. Measures the rate at which reclaimable capacity is recovered.

The complete list of raw metrics that are collected by IBM Storage Virtualize systems can be found at [Starting statistics collection](#).

The complete list of metrics that are derived by IBM Spectrum Control and IBM Storage Insights can be found at [IBM Spectrum Control Statistics](#) and [IBM Storage Insights Statistics](#).

9.7 Performance diagnostic information

If you experience performance issues on your system at any level (host, volume, nodes, pools, etc.), consult IBM Support. IBM Support requires detailed performance data about the IBM Storage Virtualize system to diagnose the problem. Generate a support package by using the IBM Storage Virtualize GUI or export performance data by using IBM Spectrum Control.

For more information about which type of support package to collect, see [What Data Should You Collect for a Problem on IBM Storage Virtualize Systems?](#)

9.7.1 Performance diagnostic information included in a support package

During the process of generating a support package, which you can generate by selecting **Settings** → **Support Package** → **Download Support Package**, all performance diagnostic statistics of each node also are captured.

A maximum of 16 files are stored on each node at any one time for each statistics file type.

The total statistics coverage depends on the statistics interval. For example, the default setting of 15 minutes has a coverage of 4 hours; however, a 15-minute sample time is too coarse to perform detailed performance analysis. If the system is not monitored by IBM Spectrum Control or IBM Storage Insights, then setting the statistics interval to 5 minutes strikes a good balance between statistics coverage and statistics granularity.

Use the **startstats** command to modify the interval at which statistics are collected.

If an interval of 5 minutes is configured, a coverage of 80 minutes (5 minutes x 16 = 80 minutes) is achieved (see Example 9-14).

Note: If the system is monitored by IBM Spectrum Control and you change the statistics interval on the IBM Storage Virtualize system, IBM Spectrum Control reverts the change automatically.

Example 9-14 CLI example to change the interval

```
IBM_FlashSystem:FS9xxx:superuser>lssystem |grep frequency
statistics_frequency 5
IBM_FlashSystem:FS9xxx:superuser>startstats -interval 3
IBM_FlashSystem:FS9xxx:superuser>lssystem |grep frequency
statistics_frequency 3
```

9.7.2 Performance diagnostic information exported from IBM Spectrum Control

You can export performance diagnostic data for a managed resource. If you contact IBM Support to help you analyze a performance problem with storage systems or fabrics, you might be asked to provide this data.

The performance data files might be large, especially if the data is for storage systems that include many volumes, or the performance monitors are running with a 1-minute sampling frequency. If the time range for the data is greater than 12 hours, volume data and 1-minute sample data are automatically excluded from the performance data, even if it is available.

To include volume data and 1-minute sample data, select the **Advanced export** option (see Figure 9-83 on page 630) when you export performance data.

When you export performance data, you can specify a time range. The time range cannot exceed the history retention limit for sample performance data. By default, this history retention limit is two weeks.

To export hourly or daily performance data, use the **exportPerformanceData** script. However, the time range still cannot exceed the history retention limits for the type of performance data.

Complete the following steps:

1. In the menu bar, select the type of storage system.
For example, to create a compressed file for a block storage system, select **Storage** → **Block** → **Storage Systems**. (To create a compressed file for a fabric, select **Network** → **Fabrics**).
2. Right-click the storage resource, and then click **Export Performance Data** (see Figure 9-82 on page 630).

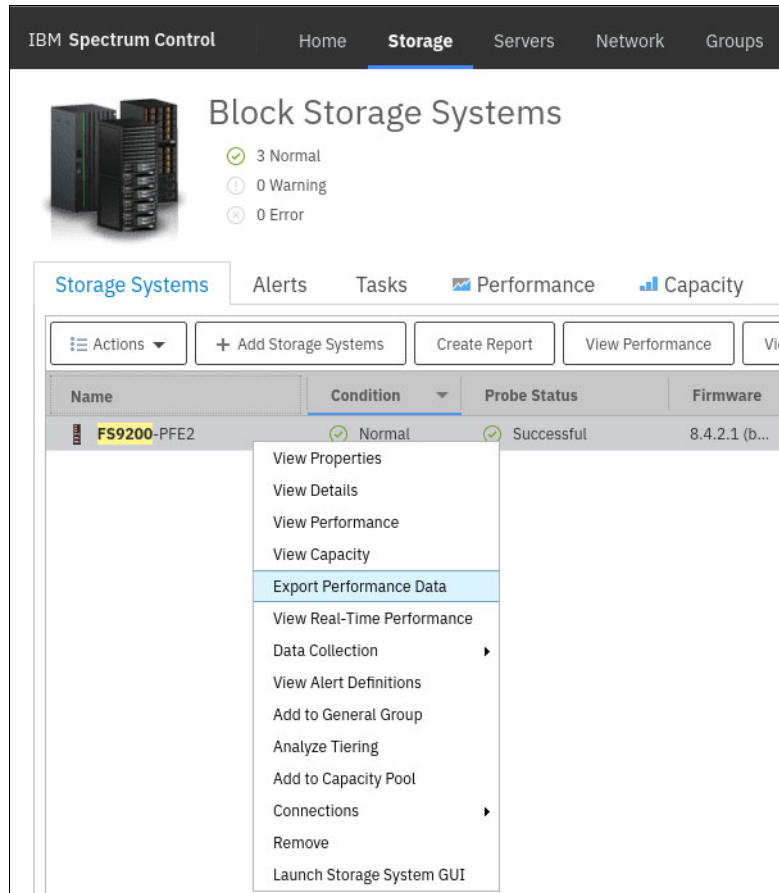


Figure 9-82 IBM Spectrum Control: Export Performance Data

3. Select the time range of the performance data that you want to export. You can use the quick select options for the previous 4, 8, or 12 hours, or specify a custom time range by clicking the time and date. Click **Create** (see Figure 9-83).

Note: To include volume data if the time range that you selected is greater than 12 hours, click **Advanced export**.

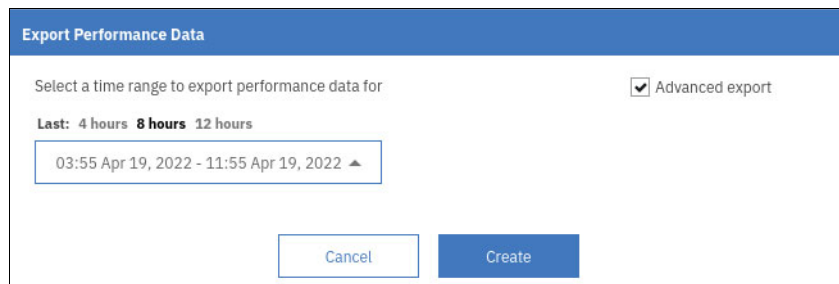


Figure 9-83 IBM Spectrum Control: Export Performance Data - Advanced Export

After the package is created, the compressed file can be downloaded by using the browser. The package includes different reports in csv format, as shown in Figure 9-84.

Name	Size
PerfReport_0000020421A0B714_Volumes_20220419-035514_8hrs0mins.csv	65.0 MB
PerfReport_0000020421A0B714_StorageSystem_20220419-035514_8hrs0mins.csv	611.9 kB
PerfReport_0000020421A0B714_StoragePorts_20220419-035514_8hrs0mins.csv	1.9 MB
PerfReport_0000020421A0B714_Pools_20220419-035514_8hrs0mins.csv	3.7 kB
PerfReport_0000020421A0B714_Nodes_20220419-035514_8hrs0mins.csv	1.2 MB
PerfReport_0000020421A0B714_ManagedDisks_20220419-035514_8hrs0mins.csv	236.2 kB
PerfReport_0000020421A0B714_IOGroups_20220419-035514_8hrs0mins.csv	592.2 kB
PerfReport_0000020421A0B714_HostConnections_20220419-035514_8hrs0mins...	411.2 kB
PerfReport_0000020421A0B714_Disks_20220419-035514_8hrs0mins.csv	2.7 MB

Figure 9-84 IBM Spectrum Control: Package files example

For more information about how to create a performance support package, see [Exporting performance data for storage systems and fabrics](#).

For more information about older versions of IBM Spectrum Control, see [Performance data collection with TPC, IBM VSC and IBM Spectrum Control](#).

9.7.3 Performance diagnostic information exported from IBM Storage Insights

To help resolve performance issues with storage systems, complete the following steps to export performance data for the resource to a compressed file from IBM Storage Insights:

1. To export the performance data, select the type of storage system in the menu bar.

For example, to create a compressed file for a block storage system, select **Resources** → **Block Storage Systems** (see Figure 9-85).

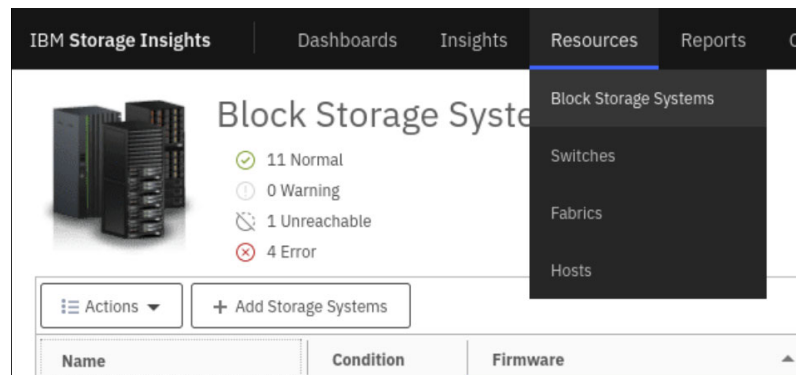


Figure 9-85 Selecting Block Storage Systems

2. Right-click the storage system and select **Export Performance Data** (see Figure 9-86 on page 632).

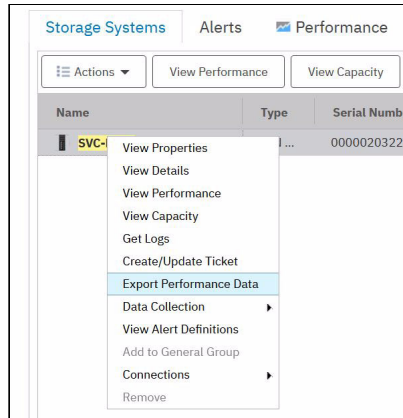


Figure 9-86 Selecting Export Performance Data

3. Select the time range of the performance data that you want to export.

You can select a time range of the previous 4, 8, or 12 hours through the quick select options, or specify a custom time range by clicking the time and date. Click **Create**. A task is started and shown in the *running tasks* icon in the menu bar.

Note: To include volume data if the time range that you selected is greater than 12 hours, click **Advanced export**.

4. When the task is complete, click the **Download** icon in the running tasks list in the task to save the file locally.

For more information about how to create a performance support package, see [Exporting performance data for storage systems](#).

Note: This customer option is available only in IBM Storage Insights Pro and IBM Storage Insights for IBM Spectrum Control. IBM Support can also perform this task for systems that are registered for the no-charge edition of IBM Storage Insights.

9.8 Metro Mirror and Global Mirror monitoring

In this section, we cover Metro Mirror (MM) and GM monitoring.

9.8.1 Monitoring with IBM Copy Services Manager

IBM Copy Services Manager (IBM CSM) is a separate licensed product that is used to administer copy services in IBM storage environments. Copy services are features that are used by storage systems such as IBM Storage Virtualize systems to configure, manage, and monitor data-copy functions. Copy services include IBM FlashCopy, MM, GM, and Global Mirror with Change Volumes (GMCV).

You can use IBM CSM to complete the following data replication tasks and help reduce the downtime of critical applications:

- ▶ Plan for replication when you are provisioning storage.
- ▶ Keep data on multiple related volumes consistent across storage systems if there is a planned or unplanned outage.
- ▶ Monitor and track replication operations.
- ▶ Automate the mapping of source volumes to target volumes.

Figure 9-87 is an example of the initial sync progress after CSM MM and GM sessions are created and started.

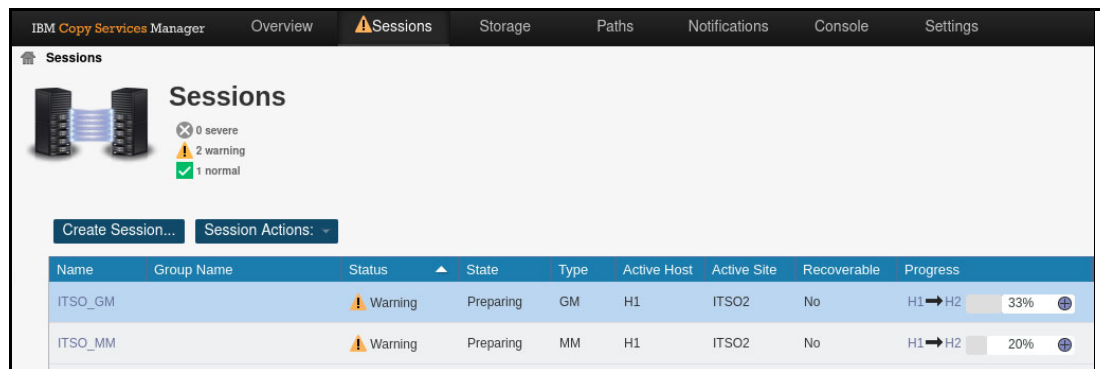


Figure 9-87 CSM sessions preparing

Figure 9-88 shows the CSM sessions after they complete their initial sync.

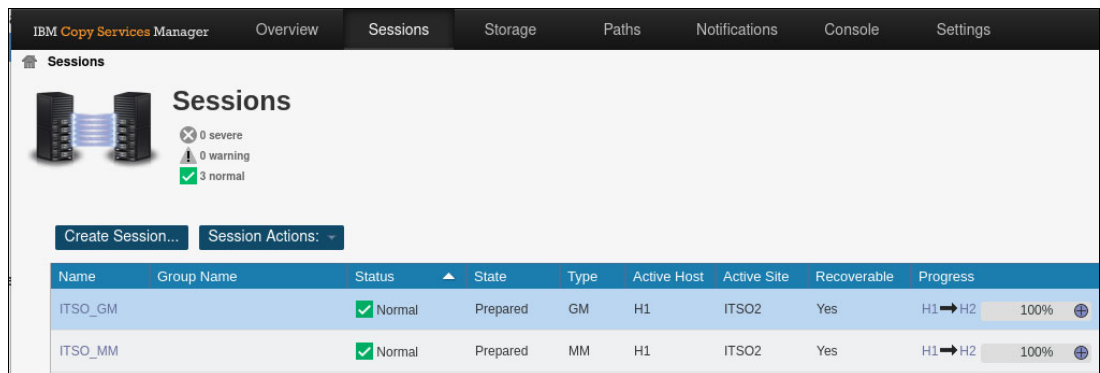


Figure 9-88 CSM sessions that are prepared and 100% synced

Note: Recoverable is now Yes, which indicates that there is a consistent recovery point.

One of the most important events that must be monitored when IBM Storage Virtualize systems are implemented in a disaster recovery (DR) solution with GM functions is checking whether GM was suspended because of a 1920 or 1720 error.

IBM Storage Virtualize can suspend the GM relationship to protect the performance on the primary site when GM starts to affect write response time. That suspension can be caused by several factors.

IBM Storage Virtualize systems do not restart the GM automatically. They must be restarted manually.

IBM Storage Virtualize systems alert monitoring is explained in 9.1.1, “Monitoring by using the management GUI” on page 552. When MM or GM is managed by CSM and a 1920 error occurs, CSM can automatically restart GM sessions. The delay time on the automatic restart option is configurable. This delay allows some time for the underlying cause to dissipate. Automatic restart is disabled in CSM by default.

Figure 9-89 shows the path to enable automatic restart of GM sessions. You select **Sessions** → **Select Session** → **Session Actions** → **View/Modify** → **Properties** → **H1-H2 options**.

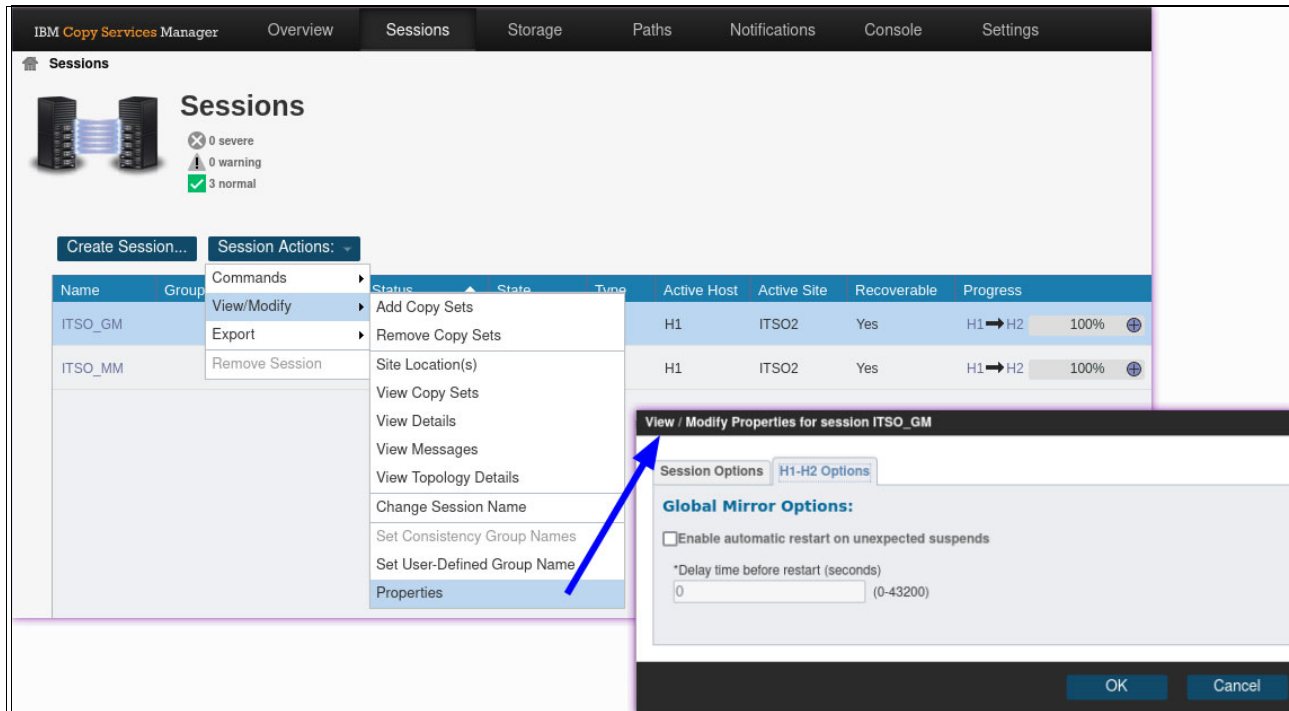


Figure 9-89 CSM automatic restart is disabled by default

If you have several sessions, you can stagger the delay time so that they do not all restart at the same time, which can affect system performance. Choose the set delay time feature to define a time in seconds for the delay between when IBM CSM processes the 1720/1920 event and when the automatic restart is issued.

An automatic restart is attempted for every suspend with reason code 1720 or 1920 up to a predefined number of times within a 30-minute period.

The number of times that a restart is attempted is determined by the storage system’s **gmlinktolerance** value. If the number of allowable automatic restarts is exceeded within the period, the session does not restart automatically on the next unexpected suspend. Issue a **Start** command to restart the session, clear the automatic restart counters, and enable automatic restarts.

Warning: When you enable this option, the session is automatically restarted by the CSM server. When this situation occurs, the secondary site is not consistent until the relationships are fully resynched.

You can specify the amount of time (in seconds) that the CSM server waits after an unexpected suspend before automatically restarting the session. The range of possible values is 0 - 43,200. The default is 0, which specifies that the session is restarted immediately after an unexpected suspend.

Figure 9-90 displays the secondary consistency warning when automatic GM restart is enabled.

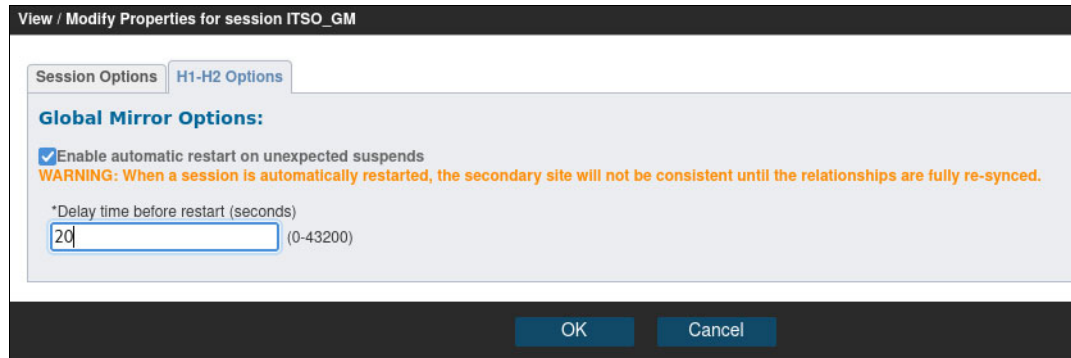


Figure 9-90 Secondary consistency warning when automatic restart is enabled

9.8.2 Monitoring MM and GM with scripts

An IBM Storage Virtualize system provides a CLI that you can use to interact with your systems by using scripts. Those scripts can run in the IBM Storage Virtualize shell, but with a restricted set of bash commands that are available, or they can run out of the shell by using any scripting language that you prefer.

An example of script usage is one to check at a specific interval time whether MM or GM are still active, if any 1920 errors occurred, or to react to an SNMP or email alert that is received. Then, the script can start some specific recovery action based on your recovery plan and environment.

Customers who do not use IBM Copy Service Manager often create their own scripts. These scripts are sometimes supported by IBM as part of ITS professional services or IBM System Lab Services. Tell your IBM representative what kind of monitoring that you want to implement with scripts, and together you can try to find whether a solution is in the IBM Intellectual Capital Management repository that can be reused.

An example of such a script can be found at [Example 2: Restarting any stopped Remote Copy relationships every 10 minutes](#).

9.9 Monitoring Tier 1 SSDs

Tier 1 SSDs require that you pay special attention to the endurance events that can be triggered. For monitoring purposes, stay alert for the new fields that are listed in Table 9-9 on page 636.

Table 9-9 Field changes to drive and array devices

Field	Description
write_endurance_used	<p>Indicates the drive writes per day (DWPD). This value is blank for drives that are not SSD drives. The value must be 0 - 255.</p> <p>This value indicates the percentage of life that is used by the drive. The value 0 indicates that full life remains, and 100 indicates that the drive is at or past its end of life.</p> <p>The drive must be replaced when the value exceeds 100.</p> <p>This value is blank for drives that are either one of the following:</p> <ol style="list-style-type: none"> 1. Not SSDs. 2. SSDs that predate support of the endurance indicator. <p>This value also applies to drives that are yet to be polled, which can take up to 24 hours.</p>
write_endurance_usage_rate	<p>Indicates the DWPD usage rate. The values are the following ones:</p> <ul style="list-style-type: none"> ▶ Measuring: No rate information available. ▶ High: The drives will not last as expected (~4.5 years). ▶ Marginal: The drives will last as expected (~4.5 - 5.5 years). ▶ Low: The drives will last as expected (~5.5 years or more). <p>This value is blank for non-SSD drives.</p> <p>This field displays a value only when the write_endurance_used value changes.</p>
replacement_date	<p>Indicates the date of a potential drive failure. The format must be YYMMDD. This value is blank for non-SSD drives.</p>

When write_endurance_usage_rate is high, an event is reported with error code 2560 and event code 010126. The description of this event is as follows:

The usage rate for a flash drive is high, which can affect the expected lifespan of the drive.

When write_endurance_used is greater than 95%, an event is reported with error code 2560 and event ID 010125. The description of this event is as follows:

A flash drive is expected to fail due to the write endurance value exceeding 95.



Maintaining an IBM Storage Virtualize infrastructure

As an IT environment grows and is renewed, so must the storage infrastructure. One of the many benefits that the IBM Storage Virtualize family provides is to simplify the device management tasks that system administrators must perform.

This chapter provides guidance about the maintenance activities of IBM Storage Virtualize software. This guidance can help you maintain an infrastructure with the levels of availability, reliability, and resiliency that are required by complex applications, and to keep up with environmental growth needs.

You can also find tips and guidance to simplify the storage area network (SAN) administration tasks that are used daily, such as adding users, storage allocation and removal, adding or removing a host from the SAN, and create procedures to manage your environment.

The discussion in this chapter focuses on the IBM FlashSystem 9500, and uses screen captures and command outputs from this model. The recommendations and practices that are described in this chapter are applicable to the following device models:

- ▶ IBM FlashSystem 5015
- ▶ IBM FlashSystem 5035
- ▶ IBM FlashSystem 5045
- ▶ IBM FlashSystem 5100
- ▶ IBM FlashSystem 5200
- ▶ IBM FlashSystem 7200
- ▶ IBM FlashSystem 7300
- ▶ IBM FlashSystem 9100
- ▶ IBM FlashSystem 9200
- ▶ IBM FlashSystem 9500
- ▶ IBM SAN Volume Controller (SVC) SV2 - SV3 - SA2

Note: The practices that are described in this chapter were effective in many deployments of different models of the IBM Storage Virtualize family. These deployments were performed in various business sectors for various international organizations. They all had one common need: to manage their storage environment easily, effectively, and reliably.

This chapter includes the following topics:

- ▶ “User interfaces” on page 638
- ▶ “Users and groups” on page 641
- ▶ “Volumes” on page 643
- ▶ “Hosts” on page 644
- ▶ “Software updates” on page 644
- ▶ “Non-disruptive security and software patching” on page 658
- ▶ “Drive firmware updates for IBM FlashSystem” on page 661
- ▶ “Remote Code Load” on page 663
- ▶ “Replacing IBM FlashCore Module in IBM FlashSystem” on page 666
- ▶ “SAN modifications” on page 667
- ▶ “Server HBA replacement” on page 669
- ▶ “Hardware upgrades” on page 671
- ▶ “I/O throttling” on page 682
- ▶ “Documenting an IBM Storage Virtualize and SAN environment” on page 688

10.1 User interfaces

The IBM Storage Virtualize family allows users to easily manage and maintain the infrastructure with a robust and powerful GUI, which provides different sets of facilities to help resolve situations that you might encounter. The interfaces for servicing your system connect through the Ethernet ports that are accessible from port 1 of each canister.

- ▶ Use the management GUI to monitor and maintain the configuration of your hardware.
- ▶ Use the Service Assistant Tool GUI to complete service procedures.
- ▶ Use the command-line interface (CLI) to manage your system.

A best practice is to use the interface most appropriate to the task that you are attempting to complete. For example, a manual software update is best performed by using the service assistant GUI rather than the CLI. Running fix procedures to resolve problems or configuring expansion enclosures can be performed only by using the management GUI. Creating many volumes with customized names is best performed by using a script on the CLI. To ensure efficient storage administration, it is a best practice to become familiar with all available user interfaces.

10.1.1 Management GUI

The management GUI is the primary tool that is used to service your system and check its status. Problem investigation, user creation and deletion, and minor configurations can be easily done by using this interface. Use the views that are available in the management GUI to verify the performance of the system, hardware-specific information, physical storage, and the available volumes and hosts.

Note: Taking advantage of your system GUI, you can easily verify performance information such as input/output operations per second (IOPS), latency, port utilization, host status, and several other sensitive information from your system. Graphics are also used to compare past statuses of your system.

To access the management GUI, start a supported web browser and go to https://<system_ip_address>, where <system_ip_address> is the management IP address that was set when the system is created.

For more information about the task menus and functions of the management GUI, see Chapter 4, “IBM Spectrum Virtualize GUI”, of *Implementation Guide for IBM Storage FlashSystem and IBM SAN Volume Controller: Updated for IBM Storage Virtualize Version 8.6*, SG24-8542.

10.1.2 Service Assistant Tool GUI

The service assistant interface is a browser-based GUI that can be used to service individual node canisters in the control enclosures.

Important: If used incorrectly, the service actions that are available through the service assistant can cause loss of access to data or even data loss.

You can connect to the service assistant on one node canister by entering the service IP address. If there is a working communications path between the node canisters, you can view status information and perform service tasks on the other node canister by making the other node canister the current node. You do not have to reconnect to the other node. On the system itself, you also can access the service assistant interface by using the technician port.

The service assistant provides facilities to help you service only control enclosures. Always service the expansion enclosures by using the management GUI.

You can also complete the following actions by using the service assistant:

- ▶ Collect logs to create and download a package of files to send to support personnel.
- ▶ Provide detailed status and error summaries.
- ▶ Remove the data for the system from a node.
- ▶ Recover a system if it fails.
- ▶ Install a code package from the support site or rescue the code from another node.
- ▶ Update code on node canisters manually.
- ▶ Configure a control enclosure chassis after replacement.
- ▶ Change the service IP address that is assigned to Ethernet port 1 for the current node canister.
- ▶ Install a temporary Secure Shell (SSH) key if a key is not installed and CLI access is required.
- ▶ Restart the services that are used by the system.
- ▶ Halt the system for maintenance (parts replacement).

To access the Service Assistant Tool GUI, start a supported web browser and go to `https://<system_ip_address>/service`, where `<system_ip_address>` is the service IP address for the node canister or the management IP address for the system on which you want to work.

10.1.3 Command-line interface

The system CLI is intended for use by advanced users who are confident about using commands. Up to 32 simultaneous interactive Secure Shell (SSH) sessions to the management IP address are supported.

Nearly all the functions that are offered by the CLI are also available through the management GUI. However, the CLI does not provide the fix procedures or the performance graphics that are available in the management GUI. Alternatively, use the CLI when you require a configuration setting that is unavailable in the management GUI.

Running **he1p** in a CLI displays a list of all available commands. You have access to a few other UNIX commands in the restricted shell, such as **grep** and **more**, which are useful in formatting output from the CLI commands. Reverse-i-search (Ctrl+R) is also available.

Table 10-1 shows a list of UNIX commands:

Table 10-1 UNIX commands available in the CLI

UNIX command	Description
grep	Filter output by keywords.
more	Moves through output one page at a time.
sed	Filters output.
sort	Sorts output.
cut	Removes individual columns from output.
head	Display only first lines.
less	Moves through the output one page at a time.
tail	Display only last lines.
uniq	Hides any duplicates in the output.
tr	Translates characters.
wc	Counts lines, words, and characters in the output.
history	Display command history.
scp	Secure copy protocol.

For more information about command reference and syntax, see the following resources:

- ▶ [Command-line interface IBM FlashSystem 9500, 9200 and 9100](#)
- ▶ [IBM Storage Virtualize for SAN VolumeController and FlashSystem Family - Command-Line Interface User's Guide](#)

Service command-line interface

Service CLI commands also can run on a specific node. To run such a command in this way, log in to the service IP address of the node that requires servicing.

For more information about the use of the service CLI, see [Service command-line interface](#).

USB command interface

When a USB flash drive is inserted into one of the USB ports on a node, the software searches for a control file (`satask.txt`) on the USB flash drive and runs the command that is specified in the file. Using the USB flash drive is required in the following situations:

- ▶ When you cannot connect to a node canister in a control enclosure by using the service assistant and you want to see the status of the node.

- ▶ When you do not know, or cannot use, the service IP address for the node canister in the control enclosure and must set the address.
- ▶ When you have forgotten the superuser password and must reset the password.

For more information about the usage of the USB port, see [Procedure: Getting node canister and system information by using a USB flash drive](#).

Technician port

The technician port is an Ethernet port on the back window of the controller canister. You can use it to configure the node. The technician port can be used to do most of the system configuration operations, which include the following tasks:

- ▶ Defining a management IP address
- ▶ Initializing a new system
- ▶ Servicing the system

For more information about the usage of the technician port, see [The technician port](#).

10.2 Users and groups

Almost all organizations have IT security policies that enforce the usage of password-protected user IDs when their IT assets and tools are used. However, some storage administrators still use generic shared IDs, such as superuser, admin, or root in their management consoles to perform their tasks. They might even use a factory-set default password. Their justification might be a lack of time, forgetfulness, or the fact that their SAN equipment does not support the organization's authentication tool.

SAN storage equipment management consoles often do not provide direct access to stored data, but you can easily shut down (accidentally or deliberately) a shared storage controller and any number of critical applications along with it. Moreover, having individual user IDs set for your storage administrators allows much better auditing of changes if you must analyze your logs.

The IBM Storage Virtualize 8.6 family supports the following authentication methods:

- ▶ Local authentication by using a password
- ▶ Local authentication by using SSH keys
- ▶ Remote authentication by using Lightweight Directory Access Protocol (LDAP) (Microsoft Active Directory or IBM Security Directory Server)
- ▶ Multifactor authentication support
- ▶ Single sign-on (SSO) support

Local authentication is appropriate for small, single-enclosure environments. Larger environments with multiple clusters and enclosures benefit from the ease of maintenance that is achieved by using SSO that uses remote authentication by using LDAP, for example.

By default, the following user groups are defined:

- ▶ **Security Admin:** Users with this role can access all functions on the system, including managing users, user groups, all aspects of security and user authentication.
- ▶ **Administrator:** Users with this role can access all functions on the system except those that deal with managing users, user groups, and authentication. This is the standard role

to assign users who administer the system and perform tasks such as provisioning storage.

- ▶ **Copy Operator:** Users with this role have monitor role privileges and can create, change, and manage all Copy Services functions but cannot create consistency groups or modify host mappings.
- ▶ **Service:** Users can delete dump files, add and delete nodes, apply service, and shut down the system. Users can also perform the same tasks as users in the monitor role.
- ▶ **Monitor:** Users with this role can view objects, but cannot manage the system or its resources. Support personnel can be assigned this role to monitor the system and to determine the cause of problems. This role is suitable for use by automation tools such as the IBM Storage Insights data collector for collecting status about the system. For more information about IBM Storage Insights, see Chapter 9, “Implementing a storage monitoring system” on page 551.
- ▶ **Restricted Administrator:** Users with this role can perform the same tasks as the Security Administrator role, but are restricted from deleting certain objects. Support personnel can be assigned this role to solve problems
- ▶ **3-Site Administrator:** Users with this role can configure, manage, and monitor 3-site replication configurations through certain command operations only available on the 3-Site Orchestrator. This is the only role intended to be used with the 3-site Orchestrator.
- ▶ **VASA Provider:** Users with this role can manage virtual volumes or vVols that are used by VMware vSphere and managed through a VASA Provider.
- ▶ **FlashCopy Administrator:** Users can create, change, and delete all the existing FlashCopy mappings and consistency groups as well as create and delete host mappings. For more information, see [FlashCopy commands](#).

In addition to standard groups, you also can configure ownership groups to manage access to resources on the system. An ownership group defines a subset of users and objects within the system. You can create ownership groups to further restrict access to specific resources that are defined in the ownership group. For more details, see [Ownership groups](#).

Users within an ownership group can view or change only resources within the ownership group in which they belong. For example, you can create an ownership group for database administrators to provide monitor-role access to a single pool that is used by their databases. Their views and privileges in the management GUI are automatically restricted, as shown in Figure 10-1 only the child pool with the associated volume is listed.

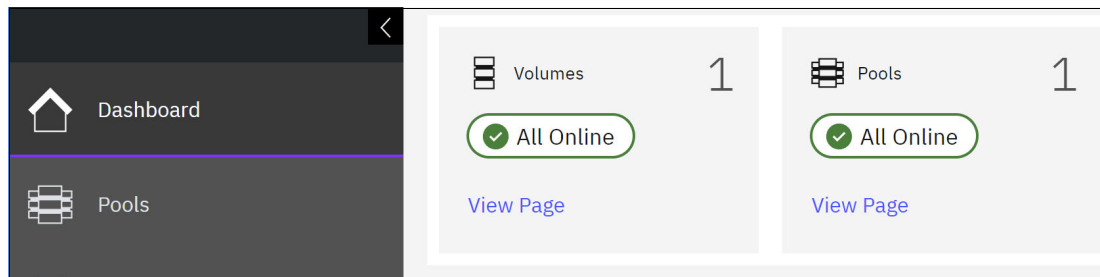


Figure 10-1 Example for restricted view for ownership groups

Regardless of the authentication method that you choose, complete the following tasks:

- ▶ Create individual user IDs for your Storage Administration staff. Choose user IDs that easily identify the user and meet your organization’s security standards.
- ▶ Include each individual user ID into the UserGroup with only enough privileges to perform the required tasks. For example, your first-level support staff probably requires only

Monitor group access to perform their daily tasks, but second-level support might require Restricted Administrator access. Consider using Ownership groups to further restrict privileges.

- ▶ If required, create generic user IDs for your batch tasks, such as Copy Services or Monitoring. Include them in a Copy Operator or Monitor UserGroup. Never use generic user IDs with the SecurityAdmin privilege in batch tasks.
- ▶ Create unique SSH public and private keys for each administrator requiring local access.
- ▶ Store your superuser password in a safe location in accordance to your organization's security guidelines and use it only in emergencies.
- ▶ For users with local authentication, it is a best practice to enable a password policy (length/expiry) that respects security standards.
- ▶ Enable multi-factor authentication (MFA).
- ▶ Use single sign-on (SSO) access if it is supported by your organization.

10.3 Volumes

A *volume* is a logical disk that is presented to a host by an I/O group (pair of nodes), and within that group a preferred node serves I/O requests to the volume.

When you allocate and deallocate volumes to hosts, consider the following guidelines:

- ▶ Before you allocate new volumes to a server with redundant disk paths, verify that these paths are working well, and that the multipath software is free of errors. Fix disk path errors that you find in your server before you proceed.
- ▶ When you plan for future growth of space-efficient volumes, determine whether your server's operating system supports the particular volume to be extended online. AIX 6.1 TL2 and earlier, for example, do not support online expansion of rootvg logical unit numbers (LUNs). Test the procedure in a non-production server first.
- ▶ Always cross-check the host LUN ID information with the `vdisk_uid` of IBM Storage Virtualize. Do not assume that the operating system recognizes, creates, and numbers the disk devices in the same sequence or with the same numbers as you created them in IBM Storage Virtualize.
- ▶ Ensure that you delete any volume or LUN definition in the server *before* you unmap it in IBM Storage Virtualize. For example, in AIX, remove the `hdisk` from the volume group (`reducevg`) and delete the associated `hdisk` device (`rmdev`).
- ▶ Consider keeping volume protection enabled. If this option is not enabled on your system, use the command `chsystem vdiskprotectionenabled yes -vdiskprotectiontime <value_in_minutes>`. Volume protection ensures that some CLI actions (mostly the ones that either explicitly or implicitly remove host-volume mappings or delete volumes) are policed to prevent the removal of mappings to volumes or deletion of volumes that are considered *active*, that is, the system detected I/O activity to the volume from any host within a specified period (15 - 1440 minutes).

Note: Volume protection cannot be overridden by using the `-force` flag in the affected CLI commands. Volume protection must be disabled to perform an activity that is blocked.

- ▶ Ensure that you explicitly remove a volume from any volume-to-host mappings and any copy services relationship to which it belongs *before* you delete it.

Attention: Avoid using the `-force` parameter in `rmvdisk`.

- ▶ If you issue the `svctask rmvdisk` command and IBM Storage Virtualize still has pending mappings, the system prompts you to confirm the action, which is a hint that you might have done something incorrectly.
- ▶ When you are deallocating volumes, plan for an interval between unmapping them to hosts (`rmvdiskhostmap`) and deleting them (`rmvdisk`). One conservative recommendation is a minimum of a 48-hour period, and having at least one business day interval between unmapping and deleting so that you can perform a quick backout if you later realize you still need some data on that volume.

For more information about volumes, see Chapter 5, “Volumes” on page 319.

10.4 Hosts

A host is a physical or virtual computer that is mapped inside your system that is directly attached or added by using your SAN (switch) through Fibre Channel (FC), internet Small Computer Systems Interface (iSCSI), and other protocols.

When you add and remove hosts in IBM Storage Virtualize, consider the following guidelines:

- ▶ Before you map new servers to IBM Storage Virtualize, verify that they are all error-free. Fix errors that you find in your server and IBM Storage Virtualize before you proceed. In IBM Storage Virtualize, pay special attention to anything inactive in the `lsfabric` command.
- ▶ Plan for an interval between updating the zoning in each of your redundant SAN fabrics, such as at least 30 minutes. This interval allows for failover to occur and stabilize, and for you to be notified if unexpected errors occur.
- ▶ After you perform the SAN zoning from one server’s Host Bus Adapter (HBA) to IBM Storage Virtualize, you should list the host’s worldwide port name (WWPN) by using the `lshbaportcandidate` command. Use the `lsfabric` command to certify that it was detected by the IBM Storage Virtualize nodes and ports that you expected. When you create the host definition in the IBM Storage Virtualize (`mkhost`), try to avoid the `-force` parameter. If you do not see the host’s WWPNs, it might be necessary to scan the fabric from the host. For example, use the `cfgmgr` command in AIX.

For more information about hosts, see Chapter 8, “Hosts” on page 519.

10.5 Software updates

Because the IBM Storage Virtualize hardware might be at the core of your disk and SAN storage environment, the software update procedure requires planning, preparation, and verification. However, with the appropriate precautions, an update to your servers and applications can be conducted easily and transparently. This section highlights the applicable guidelines for the IBM Storage Virtualize update.

Most of the following sections explain how to prepare for the software update. These sections also present version-independent guidelines about how to update the IBM Storage Virtualize family systems and flash drives.

Before you update the system, ensure that the following requirements are met:

- ▶ Download the latest system update package and update test utility. The latest package can be obtained directly using the new download function at the IBM Storage Virtualize system or by using the FixCentral download option. For more details, see [Obtaining the software packages](#).
- ▶ All node canisters are online.
- ▶ All errors in the system event log are addressed and marked as fixed.
- ▶ There are no volumes, managed disks (MDisks), or storage systems with Degraded or Offline status.
- ▶ The service assistant IP address is configured on every node in the system.
- ▶ The system superuser password is known.
- ▶ The system configuration is backed up and saved (preferably off-site), as shown in Example 10-16 on page 694.
- ▶ You can physically access the hardware.

The following actions are not required, but are recommended to reduce unnecessary load on the system during the update:

- ▶ Stop all Metro Mirror (MM), Global Mirror (GM), or HyperSwap operations.
- ▶ Avoid running FlashCopy operations.
- ▶ Avoid migrating or formatting volumes.
- ▶ Stop collecting IBM Spectrum Control performance data for the system.
- ▶ Stop automated jobs that access the system.
- ▶ Ensure that no other processes are running on the system.
- ▶ If you want to update without host I/O, then shut down all hosts.

For additional information see: [IBM Storage Virtualize Family of Products Upgrade Planning](#).

Note: For customers who purchased the IBM Storage Virtualize they may select from IBM Storage Expert Care Basic, IBM Storage Expert Care Advanced or IBM Storage Expert Care Premium. Storage Expert Care is designed to simplify and standardize the support approach on the IBM Storage FlashSystem portfolio. Customers can select their preferred level of service and support at the time of the system purchase. For more details see: and [IBM Storage Expert Care](#).

If a customer has purchased the IBM Storage Expert Care Premium, two code upgrades per year, which are performed by IBM are included. These upgrades are done by the IBM dedicated Remote Code Load (RCL) team or, where remote support is not allowed or enabled, by an onsite IBM Systems Service Representative (IBM SSR).

For more information about Remote Code Load, see 10.8, “Remote Code Load” on page 663.

10.5.1 Determining the current and target software level

The first step is to determine your current and your target IBM Storage Virtualize code level.

Using the example of an IBM Storage Virtualize 9500, log in to the web-based GUI and find the current version.

You can find the current version by doing either of the following actions:

- ▶ At the upper right, click the question mark symbol (?) and select **About IBM Storage Virtualize 9500** to display the current version.
- ▶ Select **Settings** → **System** → **Update System** to display both the current and target versions.

Figure 10-2 shows the Update System output window and displays the current and latest code levels. In this example, the code level is 8.5.0.7.

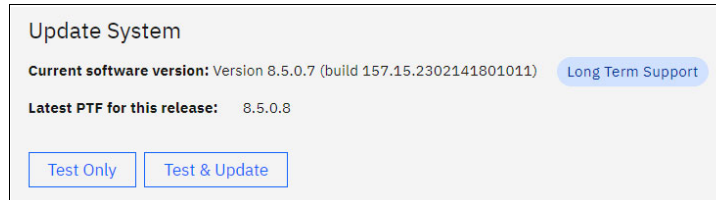


Figure 10-2 Current software version

Figure 10-3 shows the Update System output where code level is up to date. In this example, the code level is 8.6.0.0.

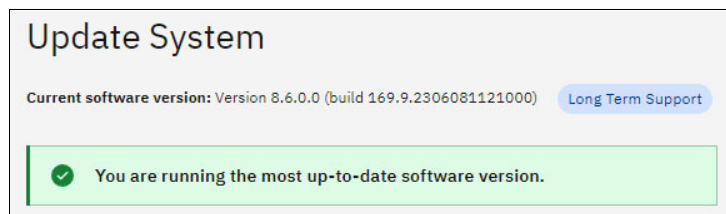


Figure 10-3 Up-to-date software version

Alternatively, if you use the CLI, run the `lssystem` command. Example 10-1 shows the output of the `lssystem` CLI command and where the code level output can be found.

Example 10-1 The lssystem command

```
IBM_FlashSystem:FS9500:superuser>lssystem |grep code
code_level 8.5.0.7 (build 157.15.2302141801011)
```

IBM Storage Virtualize software levels are specified by 4 digits in the following format (in our example V.R.M.F = 8.5.0.7):

- ▶ V is the major version number.
- ▶ R is the release level.
- ▶ M is the modification level.
- ▶ F is the fix level.

To update your system by using the most suitable code version, check the following examples and define which version to use:

- ▶ The specific version of an application or other component of your SAN Storage environment has a known problem or limitation.
- ▶ The latest IBM Storage Virtualize software release is not yet cross-certified as compatible with another key component of your SAN storage environment.

- ▶ Your organization has mitigating internal policies, such as the usage of the “latest release minus 1” or requiring “seasoning” in the field before implementation in a production environment.

For more information, see [IBM Storage Virtualize Family of Products Upgrade Planning](#).

10.5.2 Obtaining the software packages

Before you can update the system software, you need to obtain the current version of the software. This can be done directly or manually. For more information see: [Obtaining the software packages](#).

Obtaining the package manually

To obtain a new release of software for a system update, see [IBM Fix Central](#) and either find your product by typing its name or choose it from a list that is provided by IBM. (in this case we are looking for IBM FlashSystem 9500):

1. From the **Find Product** tab, type IBM FlashSystem 9500 (or whatever model is appropriate in your environment).
2. From the Installed Version list, select the appropriate software version level that was determined in 10.5.1, “Determining the current and target software level” on page 645.
3. Select **Continue**.

To download the software package, complete the following steps:

4. In the Product Software section, select **IBM Storage Virtualize Software Upgrade Test Utility**, and also select the **8.6.0.0 Fix Pack** that is listed for your system.
5. Select **Continue**.
6. Click the option button for your preferred download options and click **Continue**.
7. Enter your machine type and serial number.
8. Click **Continue**.
9. Read the terms and conditions, and then select **I Agree**.
10. Select **Download Now** and save the three files onto your management computer. Figure 10-4 on page 648 shows the window for downloading the package.

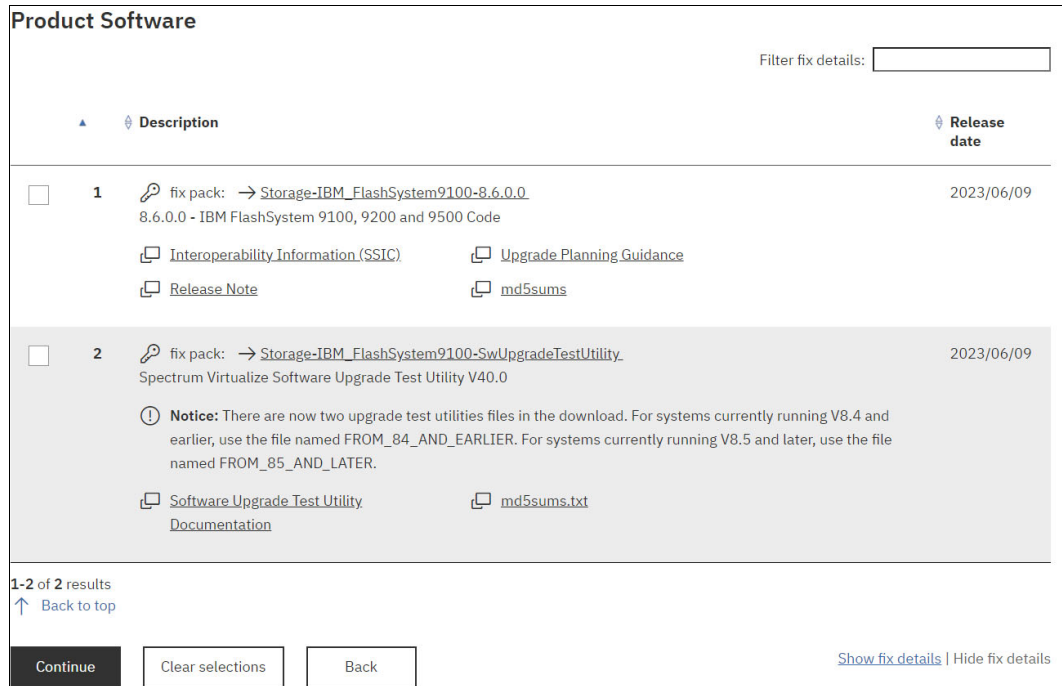


Figure 10-4 Fix Central download

Provide the package manually

Use this option to upload the update package, test utility or patch file to the storage system manually. This is required to unhide the corresponding button **Test only**, **Test & Upgrade** or **Install patches**.

Note: The option to upload packages manually only applies to IBM Storage Virtualize Software 8.5.4 and above. On IBM Storage Virtualize Systems running below 8.5.4 the step “Provide the package manually” can be skipped.

To provide the package manually, complete these steps:

Select **Settings > System > Update System**.

Click **Upload** as shown in Figure 10-5.

On the Upload page, drag and drop the files or click **Select files**. Select the **test utility**, **update package**, or **patch files** that you want to update.

Click **Upload**.

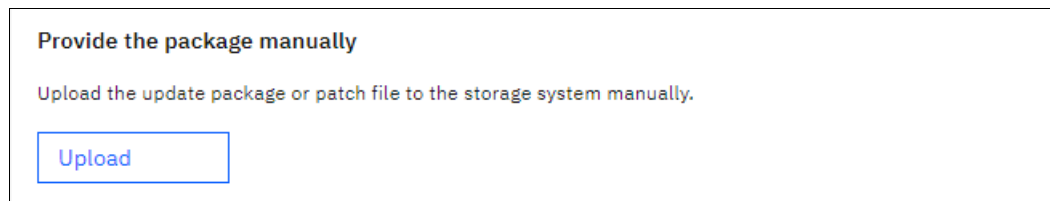


Figure 10-5 Upload package manually

After uploading a valid **upgrade package, test utility or patch**, the management interface will unhide the corresponding **Test only**, **Test & Upgrade** or **Install patches** buttons as shown in Figure 10-6.

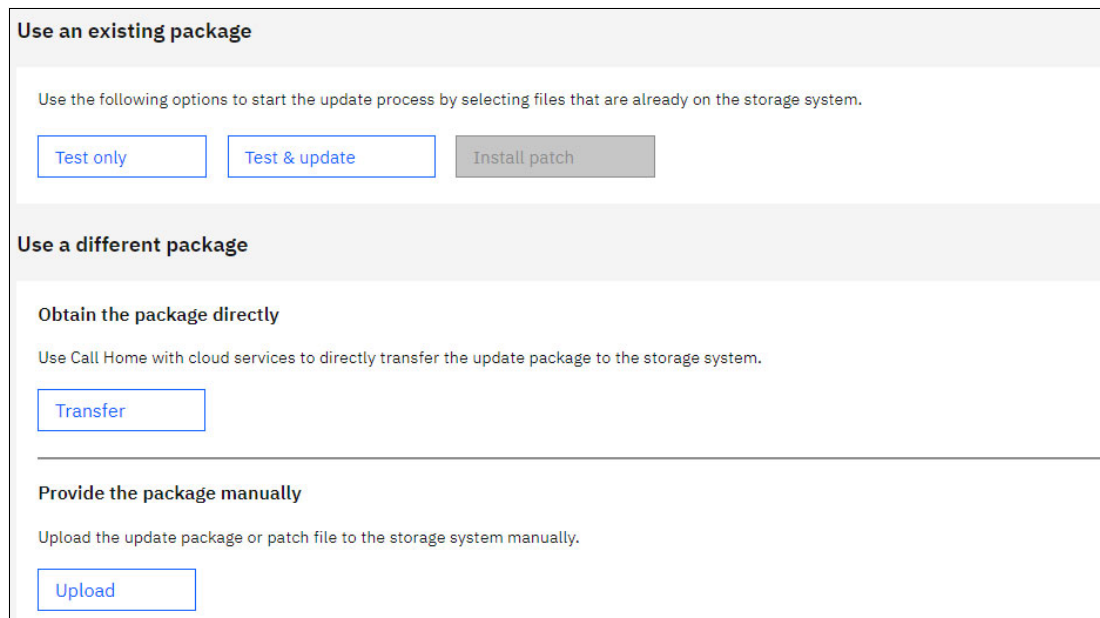


Figure 10-6 Unhide GUI buttons

Obtain the package directly

To use this option, enable Call Home with cloud services to connect to support, and directly transfer the latest update available to the storage system.

Note: The option to upload packages directly only applies to IBM Storage Virtualize Software 8.5.4 and above. For IBM Storage Virtualize Systems running below 8.5.4 the option “Obtain the package directly” is not available.

To obtain the package directly, complete these steps:

1. Select **Settings** → **System** → **Update System**.
2. If Transfer is disabled, click **Call Home** to navigate to the access panel and enable Call Home with Cloud Services. For more information, see [Setting up Call Home with cloud services](#).
3. On the Update System page, click **Transfer**.
4. On the Transfer Update Package page. Enter your IBMid and password along with either Version ID or Fix ID. Example is shown in Figure 10-7 on page 650 and Figure 10-8 on page 650
5. Click **Transfer**, a progress bar is displayed to provide the status of the transferring package. Click **Cancel** to cancel the transfer of package.
 - After the transfer is complete, click **Check** for files.
If the transfer is successful, then the transferred package appears in Use an existing package.
 - If the transfer is unsuccessful, then No files found is displayed.
6. To test and update, go to Use an existing package.

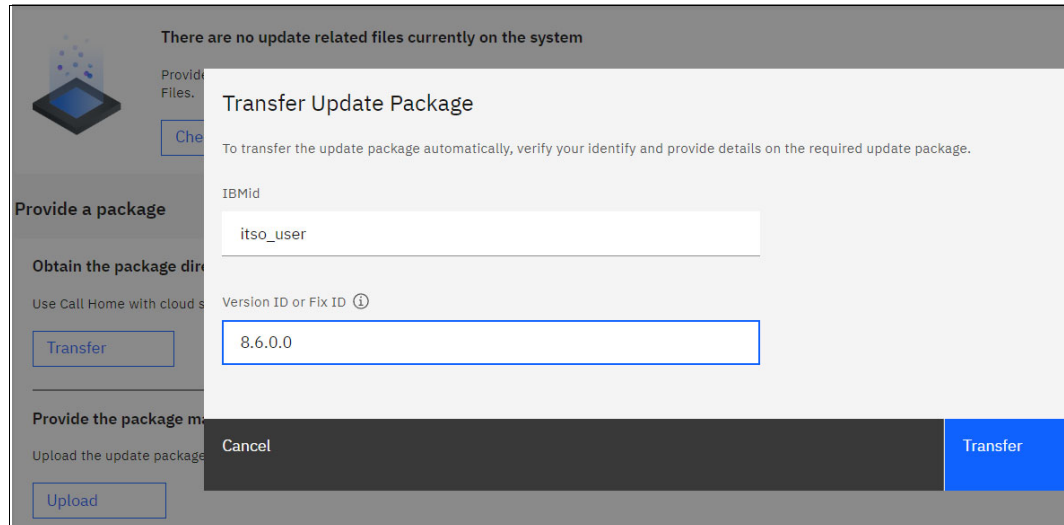


Figure 10-7 Transfer update package

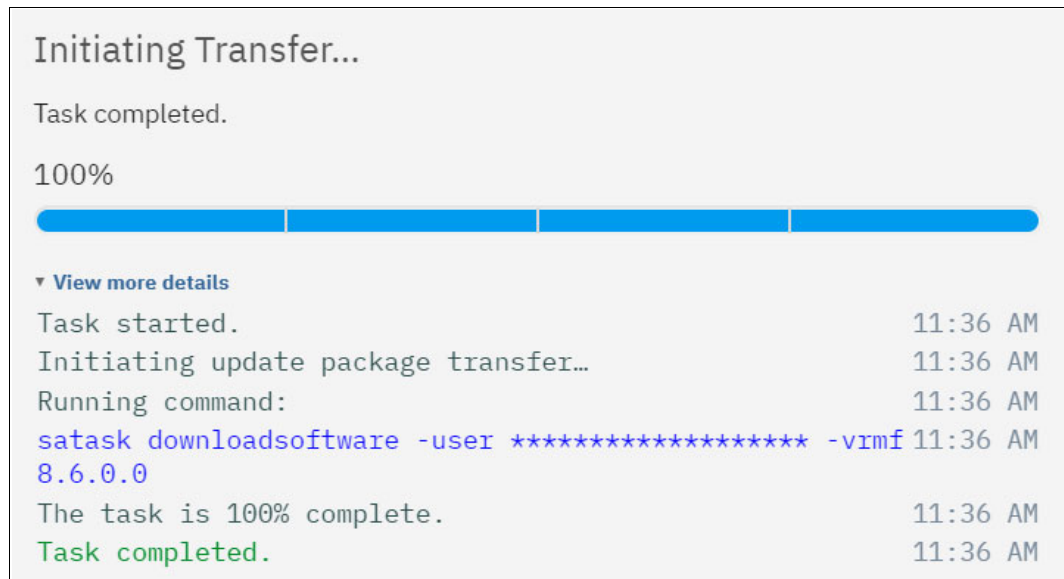


Figure 10-8 Initiating transfer of the update package

10.5.3 Hardware considerations

Before you start the update process, always check whether your IBM Storage Virtualize hardware and target code level are compatible.

If part or all your current hardware is not supported at the target code level that you want to update to, replace the unsupported hardware with newer models before you update to the target code level.

Conversely, if you plan to add or replace hardware with new models to an existing cluster, you might have to update your IBM Storage Virtualize code level first.

For more information about Code and Hardware Interoperability for IBM Storage Virtualize see: [Concurrent Compatibility and Code Cross Reference for IBM Storage Virtualize](#).

10.5.4 Update sequence

Check the compatibility of your target IBM Storage Virtualize code level with all components of your SAN storage environment (SAN switches, storage controllers, and server HBAs) and its attached servers and storage (operating systems and applications).

Applications often certify only the operating system that they run under and leave the task of certifying its compatibility with attached components (such as SAN storage) to the operating system provider. However, various applications might use special hardware features or raw devices and certify the attached SAN storage. If you have this situation, consult the compatibility matrix for your application to certify that your IBM Storage Virtualize target code level is compatible.

By cross-checking that the version of IBM Storage Virtualize is compatible with the versions of your SAN environment components, you can determine which one to update first. By checking a component's update path, you can determine whether that component requires a multistep update.

For IBM Storage Virtualize systems, if you are not making major version or multi-step updates in any components, the following update order is recommended to avoid problems:

1. Back-end storage controllers (if present)
2. Host HBA microcode, driver, and multipath software
3. IBM Storage Virtualize system
4. IBM Storage Virtualize internal SAS and NVMe drives (if present)

Attention: Do *not* update two components of your IBM Storage Virtualize system simultaneously, such as an IBM SAN Volume Controller model SV3 and one backend storage controller. This caution is true even if you intend to perform this update with your system offline. An update of this type can lead to unpredictable results, and an unexpected problem is much more difficult to debug.

10.5.5 SAN fabrics preparation

If you are using symmetrical, redundant, or independent SAN fabrics, preparing these fabrics for an IBM Storage Virtualize update can be safer than hosts or storage controllers. This statement is true assuming that you follow the guideline of a 30-minute minimum interval between the modifications that you perform in one fabric to the next. Even if an unexpected error brings down your entire SAN fabric, the IBM Storage Virtualize environment continues working through the other fabric, and your applications remain unaffected.

Because you are updating IBM Storage Virtualize, also update your SAN switches code to the latest supported level. Start with your principal core switch or director, continue by updating the other core switches, and update the edge switches last. Update one entire fabric (all switches) before you move to the next one so that a problem you might encounter affects only the first fabric. Begin your other fabric update only after you verify that the first fabric update has no problems.

If you are not running symmetrical, redundant, or independent SAN fabrics, you should be because a lack of them represents a single point of failure.

10.5.6 Storage controllers preparation

Much like with the attached hosts, the attached storage controllers must correctly handle the failover of MDisk paths. Therefore, they must be running supported microcode levels, and their own SAN paths to IBM Storage Virtualize must be free of errors.

10.5.7 Host preparation

If the appropriate precautions are taken, the IBM Storage Virtualize update is not apparent to the attached host and their applications. The automated update procedure updates one IBM Storage Virtualize node at a time while the other node in the I/O group covers for its designated volumes.

However, to ensure that this feature works, the failover capability of your multipath software must be working correctly. This capability can be mitigated by enabling N_Port ID Virtualization (NPIV) if your current code level supports this function. For more information about NPIV, see Chapter 2, “Storage area network guidelines” on page 123.

Before you start the IBM Storage Virtualize update preparation, check the following items for every host that is attached to IBM Storage Virtualize that you update:

- ▶ The operating system type, version, and maintenance or fix level
- ▶ The make, model, and microcode level of the HBAs
- ▶ The multipath software type, version, and error log

Fix every problem or “suspect” that you find with the disk path failover capability. Because a typical IBM Storage Virtualize environment can have hundreds of hosts that are attached to it, a spreadsheet might help you with the Attached Hosts Preparation tracking process. If you have some host virtualization, such as VMware ESX, AIX logical partitions (LPARs), IBM Virtual I/O Server (VIOS), or Solaris containers in your environment, verify the redundancy and failover capability in these virtualization layers.

10.5.8 Copy services considerations

When you update an IBM Storage Virtualize family product that participates in an intercluster Copy Services relationship, do *not* update both clusters in the relationship simultaneously. This situation is not verified or monitored by the automatic update process, and it might lead to a loss of synchronization and unavailability.

You must successfully finish the update in one cluster before you start the next one. Try to update the next cluster as soon as possible to the same code level as the first one. Avoid running them with different code levels for extended periods. The recommendation is to start with the AUX cluster.

10.5.9 Running the Upgrade Test Utility

It is a requirement that you install and run the latest IBM Storage Virtualize Upgrade Test Utility before you update the IBM Storage Virtualize software. For more information, see [Software Upgrade Test Utility](#). This tool verifies the health of your IBM Storage Virtualize system for the update process. It checks for unfixed errors, degraded MDisk, inactive fabric connections, configuration conflicts, hardware compatibility, drive firmware, and many other issues that might otherwise require cross-checking a series of command outputs.

You can use the management GUI or the CLI to install and run the Upgrade Test Utility.

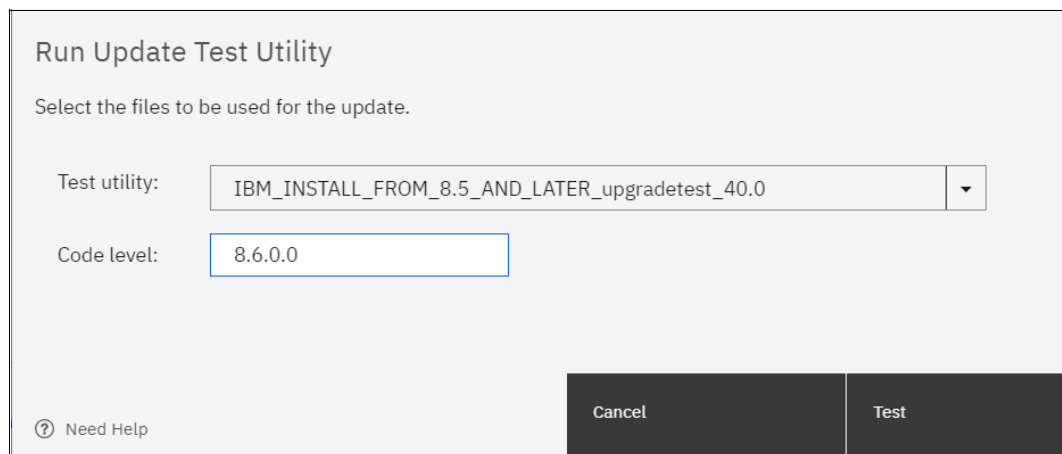
Using the management GUI

To test the software on the system, complete the following steps:

1. In the management GUI, select **Settings** → **System** → **Update System**.
2. Click **Test Only**.

Note: In case the IBM Storage Virtualize Systems is running 8.5.4 or above, you have to upload the package first to unhide the **Test only** button. For details see, “**Provide the package manually**” on page 648.

3. Select the test utility that you downloaded from the Fix Central support site. Upload the Test utility file and enter the code level to which you are planning to update. Figure 10-9 shows the IBM Storage Virtualize management GUI window that is used to install and run the Upgrade Test Utility.



Run Update Test Utility

Select the files to be used for the update.

Test utility: IBM_INSTALL_FROM_8.5_AND_LATER_upgradetest_40.0

Code level: 8.6.0.0

? Need Help

Cancel Test

Figure 10-9 IBM Storage Virtualize Upgrade Test Utility by using the GUI

4. Click **Test**. The test utility verifies that the system is ready to be updated. After the Update Test Utility completes, you are presented with the results. The results state that no warnings or problems were found, or directs you to more information about known issues that were discovered on the system.

Figure 10-10 on page 654 shows a successful completion of the Upgrade Test Utility.

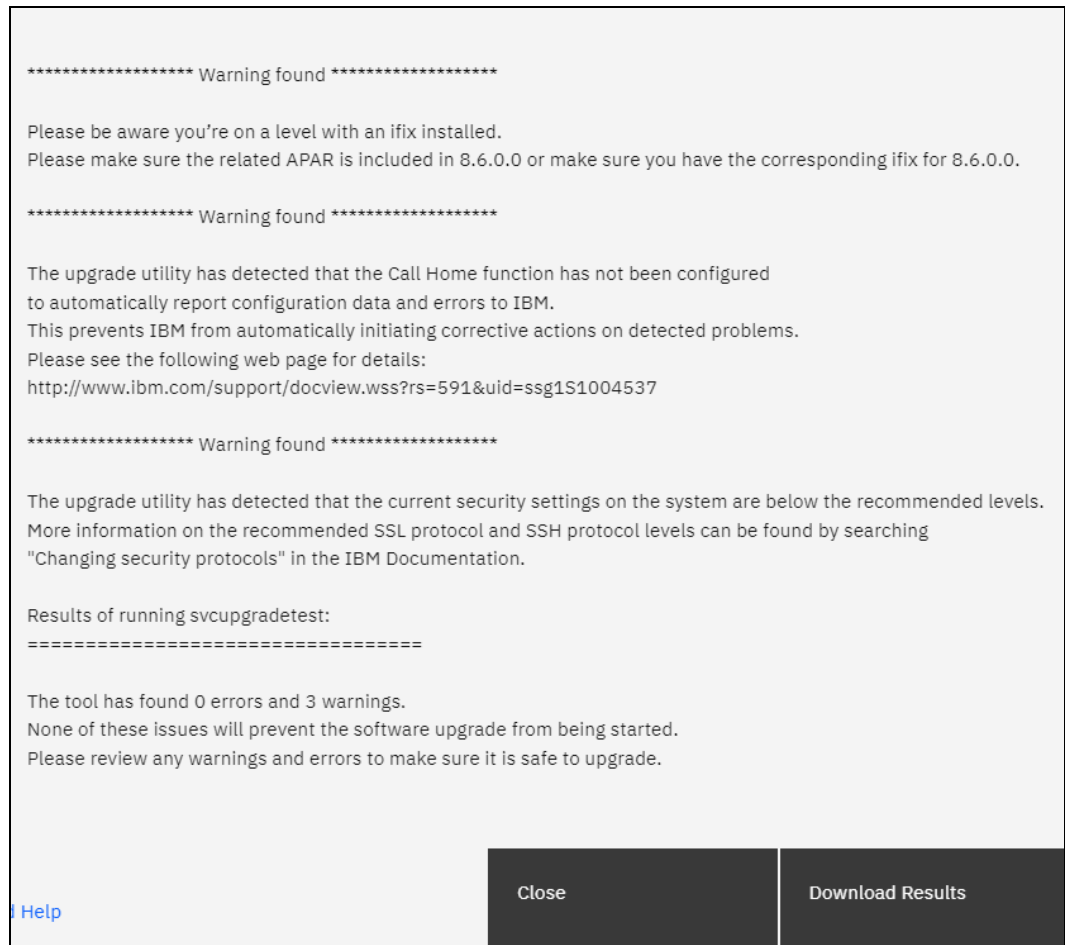


Figure 10-10 Example result of the IBM Storage Virtualize Upgrade Test Utility

5. Click **Download Results** to save the results to a file.
6. Click **Close**.

Using the command-line interface

To test the software on the system, complete the following steps:

1. Use the OpenSSH **scp** or PuTTY **pscp**, and copy the Test Utility file and the Software Update package to the `/home/admin/upgrade` directory by using the management IP address of IBM Storage Virtualize. Documentation and online help might refer to the `/home/admin/update` directory, which points to the same location on the system.

An example for IBM Storage Virtualize is shown in Example 10-2.

Example 10-2 Copying the upgrade test utility to IBM Storage Virtualize

```
C:\Program Files\PuTTY>pscp -unsafe
C:\Users\120688724\Downloads\IBM_INSTALL_FROM_8.5_AND_LATER_upgradetest_40.0
superuser@9.42.162.153:/home/admin/upgrade
Keyboard-interactive authentication prompts from server:
| Password:
End of keyboard-interactive prompts from server
IBM_INSTALL_FROM_8.5_AND_ | 411 kB | 411.7 kB/s | ETA: 00:00:00 | 100%
```

2. Ensure that the file was successfully copied, by checking the output of the `lsdumps -prefix /home/admin/upgrade` command at the Storage Virtualize CLI. An example is shown in Example 10-3

Example 10-3 Ensure that files are uploaded

```
IBM_FlashSystem:FS9500:superuser> lsdumps -prefix /home/admin/upgrade
id filename
0 IBM_INSTALL_FROM_8.5_AND_LATER_upgradetest_40.0
IBM_FlashSystem:FS9110:superuser>
```

3. Install and run Upgrade Test Utility in the CLI, as shown in Example 10-4. In this case, the Upgrade Test Utility found no errors but one warning and completed successfully.

Example 10-4 Upgrade test by using the CLI

```
IBM_FlashSystem:FS9500:superuser>svctask applysoftware -file
IBM_INSTALL_FROM_8.5_AND_LATER_upgradetest_40.0
CMMVC9001I The package installed successfully.
```

```
IBM_FlashSystem:FS9500:superuser>svcupgradetest -v 8.6.0.0
svcupgradetest version 40.0
```

Please wait, the test may take several minutes to complete.

This version of svcupgradetest believes that this system is already running the latest available level of code.

No upgrade is required at this time.

If you believe this is not correct, please check the support website to see if a newer version of this tool is available.

If you are installing an ifix, and no further issues are reported below, please continue with the upgrade.

***** Warning found *****

The upgrade utility has detected that the current security settings on the system are below the recommended levels.

More information on the recommended SSL protocol and SSH protocol levels can be found by searching

"Changing security protocols" in the IBM Documentation.

Results of running svcupgradetest:

=====

The tool has found 0 errors and 1 warnings.

None of these issues will prevent the software upgrade from being started.

Please review any warnings and errors to make sure it is safe to upgrade.

Review the output to check whether there were any problems that were found by the utility. The output from the command either states that no problems were found, or it directs you to details about known issues that were discovered on the system.

Note: Always use the latest available Upgrade Test Utility version. This tool runs tests to certify that your hardware can receive the code level that you are planning to install.

10.5.10 Updating the software

IBM Storage Virtualize software can be updated by using one of the following methods:

- ▶ **GUI:** During a standard update procedure in the management GUI, the system updates each of the nodes systematically. This method is the recommended one for updating software on nodes.
- ▶ **CLI:** The CLI provides more control over the automatic upgrade process. You can resolve multipathing issues when nodes go offline for updates. You can also override the default 30-minute mid-point delay, pause an update, and resume a stalled update.
- ▶ **Manual:** To provide even more flexibility in the update process, you can manually update each node individually by using the Service Assistant Tool GUI. When upgrading the software manually, you remove a node from the system, update the software on the node, and return the node to the system. You repeat this process for the remaining nodes until the last node is removed from the system. Now, the remaining nodes switch to running the new software. When the last node is returned to the system, it updates and runs the new level of software. This action cannot be performed on an active node. To update software manually, the nodes must either be candidate nodes (a candidate node is a node that is not in use by the system and cannot process I/O) or in a service state. During this procedure, every node must be updated to the same software level, and the node becomes unavailable during the update.

Whichever method (GUI, CLI, or manual) that you choose to perform the update, make sure that you adhere to the following guidelines for your IBM Storage Virtualize software update:

- ▶ Schedule the IBM Storage Virtualize software update for a low I/O activity time. The update process puts one node at a time offline. It also disables the write cache in the I/O group that node belongs to until both nodes are updated. Therefore, with lower I/O, you are less likely to notice performance degradation during the update.
- ▶ Never power off, restart, or reset an IBM Storage Virtualize node during a software update unless you are instructed to do so by IBM Support. Typically, if the update process encounters a problem and fails, it backs out. The update process can take 1 hour per node with a further, optional, 10 minute mid-point delay.
- ▶ If you are planning for a major IBM Storage Virtualize version update, see the [Code Cross Reference](#).
- ▶ Check whether you are running a web browser type and version that is supported by the IBM Storage Virtualize target software level on every computer that you intend to use to manage your IBM Storage Virtualize.

This section describes the steps that are required to update the software.

Using the management GUI

To update the software on the system automatically, complete the following steps:

1. In the management GUI, select **Settings** → **System** → **Update System**.
2. Click **Test & Update**.
3. Select the test utility and the software package that you downloaded from the Fix Central support site or via direct download. The test utility verifies (again) that the system is ready to be updated.
4. Click **Next**. Select **Automatic update**.

5. Select whether you want to create intermittent pauses in the update to verify the process. Select one of the following options.
 - Full automatic update without pauses (recommended).
 - Pausing the update after half of the nodes are updated.
 - Pausing the update before each node updates.
6. Click **Finish**. As the canisters on the system are updated, the management GUI displays the progress for each canister.
7. Monitor the update information in the management GUI to determine when the process is complete.

Using the command-line interface

To update the software on the system automatically, complete the following steps:

1. You must run the latest version of the test utility to verify that no issues exist with the current system, as shown in Example 10-4 on page 655.
2. Copy the software package to IBM Storage Virtualize by using the same method that is described in Example 10-2 on page 654.
3. Before you begin the update, you must be aware of the following situations, which are *valid for both the CLI and GUI*:
 - The installation process fails under the following conditions:
 - If the software that is installed on the system is not compatible with the new software or if an inter-system communication error does not allow the system to check that the code is compatible.
 - If any node in the system has a hardware type that is not supported by the new software.
 - If the system determines that one or more volumes in the system would be taken offline by restarting the nodes as part of the update process. You can find details about which volumes would be affected by using the `1sdependentvdisks` command. If you are prepared to lose access to data during the update, you can use the `-force` flag to override this restriction.
 - The update is distributed to all the nodes in the system by using internal connections between the nodes.
 - Nodes are updated one at a time.
 - Nodes run the new software concurrently with normal system activity.
 - While the node is updated, it does not participate in I/O activity in the I/O group. As a result, all I/O activity for the volumes in the I/O group is directed to the other node in the I/O group by the host multipathing software.
 - There is a 10-minute delay between node updates. The delay allows time for the host multipathing software to rediscover paths to the nodes that are updated. There is no loss of access when another node in the I/O group is updated.
 - The update is not committed until all nodes in the system are successfully updated to the new software level. If all nodes are successfully restarted with the new software level, the new level is committed. When the new level is committed, the system vital product data (VPD) is updated to reflect the new software level.
 - Wait until all member nodes are updated and the update is committed before you invoke the new functions of the updated software.

- Because the update process takes some time, the installation command completes when the software level is verified by the system. To determine when the update is completed, you must either display the software level in the system VPD or look for the Software update complete event in the error/event log. If any node fails to restart with the new software level or fails at any other time during the process, the software level is backed out.
 - During an update, the version number of each node is updated when the software is installed and the node is restarted. The system software version number is updated when the new software level is committed.
 - When the update starts, an entry is made in the error or event log and another entry is made when the update completes or fails.
4. Issue the following CLI command to start the update process:

```
applysoftware -file <software_update_file>
```

where <software_update_file> is the file name of the software update file. If the system identifies any volumes that would go offline as a result of restarting the nodes as part of the system update, the software update does not start. An optional **-force** parameter can be used to indicate that the update continues regardless of the problem identified. If you use the **-force** parameter, you are prompted to confirm that you want to continue.

5. Issue the following CLI command to check the status of the update process:

```
lupdate
```

This command displays success when the update is complete.

6. To verify the node software version issue the **lnodecanistervpd** on each node. To confirm the cluster version issue the **lssystem** command. See Example 10-5.

Example 10-5 Verify software version

```
IBM_FlashSystem:FS9500:superuser>lnodecanistervpd 1 |grep code_level
code_level 8.6.0.0 (build 169.9.2306081121000)
```

```
IBM_FlashSystem:FS9500:superuser>lssystem |grep code_level
code_level 8.6.0.0 (build 169.9.2306081121000)
```

For more details about software upgrade, see [Updating the system software](#).

10.6 Non-disruptive security and software patching

The security and software patching allows Storage Virtualize products to change non-Storage Virtualize code without disruption to the IO path of the nodes. Code not directly used by the IO stack may be patched to fix published security issues.

Initially, patching will only be used to update software outside of the I/O path and the installation of a patch will have no impact to I/O.

Examples where a patch might be required are:

- ▶ Linux utilities
- ▶ Open-source software (open SSL, Apache, Tomcat)
- ▶ Device firmware

A Patch will modify/create/delete one or more files on the boot disk. Several Patches can be installed, but each Patch is on top of the already existing patch and can only be reverted in order.

For more information, see [Managing patches](#).

10.6.1 Installing a software patch

A patch can be installed on a single node or on the cluster (all nodes) either by GUI or CLI. Upload the patch first as described in “**Provide the package manually**” on page 648.

Installing a patch using CLI

To install a patch on a single node use `satask installsoftware` to install a patch on all nodes use the `svctask applysoftware`. An example is shown in Example 10-6.

Example 10-6 Installing a patch on all nodes

```
IBM_FlashSystem:FS9500:superuser>svctask applysoftware -file  
IBM_PATCH_SVT00001_1.0
```

To check and verify if and what patch is installed see 10.6.2, “Verifying a software patch” on page 659.

Installing a patch using GUI

To install the patch using the GUI, complete the following steps:

1. Upload the Patch as described in “**Provide the package manually**” on page 648.
2. select **Settings** → **System** → **Update System** → **Install patch**.
3. Select the patch file that you want to install.
4. Click install to start the installation.

See Figure 10-11 for details.

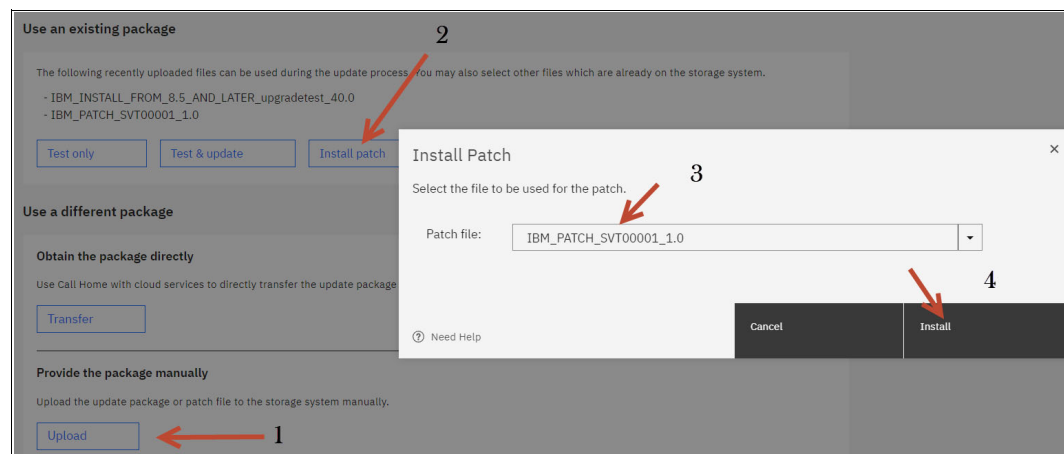


Figure 10-11 Installing the patch using the GUI

10.6.2 Verifying a software patch

To check whether a patch is installed, you can use either the GUI or CLI.

Patch verification using CLI

To check whether a patch is installed, you can work with the commands `sainfo` `lsservicestatus` and `lspatch` as shown in Example 10-7.

Example 10-7 Verify patch installation using lsservicestatus

```
IBM_FlashSystem:FS9500:superuser>sainfo lsservicestatus |grep -i patch
patch_count 1
patch_name IBM_PATCH_SVT00001_1.0
patch_description SVT patch 01
patch_status installed
IBM_FlashSystem:FS9500:superuser>lspatch 1
id node_id node_name panel_name patch_name patch_status
1 1 node1 78E003K-1 IBM_PATCH_SVT00001_1.0 installed
IBM_FlashSystem:FS9500:superuser>lspatch 2
id node_id node_name panel_name patch_name patch_status
1 2 node2 78E003K-2 IBM_PATCH_SVT00001_1.0
installed
```

Patch verification using GUI

To list all installed patches using the GUI, complete the following steps:

Select **Settings** → **Update System** → **View installed patches**.

An example of the result is shown in Figure 10-12.

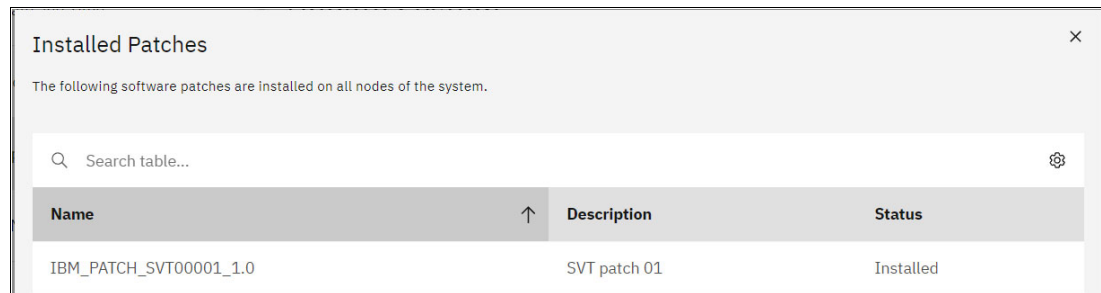


Figure 10-12 list installed patches in GUI

10.6.3 Removing a software patch

When a code upgrade happens the code gets installed as usual. The new code has the information about all patches included. Patches can have the status of "allowed" or "included". Allow means that the patch can be reinstalled. Included means the current code has the fix already included and there is no need to reapply the patch.

Note: patches (that are not obsolete) are retained across code upgrades.

Issue the CLI command `satask removepatch -patch <patch_file> <panel_name>` to remove patches on one or more nodes. `lsservicenodes` will list error 843 (patch mismatch) in case patches are not consistent within the cluster. See Example 10-8 on page 661 for more details.

Example 10-8 Removing patch on single nodes

```
IBM_FlashSystem:FS9500:superuser>satask removepatch -patch IBM_PATCH_SVT00001_1.0 01-2  
CMMVC8222I Patch operation succeeded.
```

```
IBM_FlashSystem:FS9500:superuser>sainfo lsservicenodes  
panel_name cluster_id      cluster_name node_id node_name relation node_status  
error_data  
01-2      0000020421600428 FS9110      2      node2      local      Active  
843 1 IBM_PATCH_SVT00001_1.0  
01-1      0000020421600428 FS9110      1      node1      partner   Active
```

To remove all patches from a node in the system, enter the following command as shown in Example 10-9:

```
satask purgepatches -force panel_name
```

Example 10-9 Removing all patches from a node in the system

```
IBM_FlashSystem:FS9500:superuser>satask purgepatches -force 01-1  
CMMVC8237I All patches removed from node.
```

10.7 Drive firmware updates for IBM FlashSystem

Updating drive firmware is a concurrent process that can be performed online while the drive is in use, whether it is NVMe or storage-class memory (SCM) drives in the control enclosure or SSD drives in any SAS-attached expansion enclosures.

When used on an array member drive, the update checks for volumes that depend on the drive and refuses to run if any are found. Drive-dependent volumes are usually caused by non-redundant or degraded redundant array of independent disks (RAID) arrays. Where possible, you should restore redundancy to the system by replacing any failed drives before upgrading the drive firmware. When this task is not possible, you can either add redundancy to the volume by adding a second copy in another pool, or use the **-force** parameter to bypass the dependent volume check. Use **-force** only if you are willing to accept the risk of data loss on dependent volumes (if the drive fails during the firmware update).

Note: Due to some system constraints, it is not possible to produce a single NVMe firmware package that works on all NVMe drives on all IBM Storage Virtualize code levels. Therefore, you find three different NVMe firmware files that are available for download depending on the size of the drives that you installed.

Using the management GUI

To update the drive firmware automatically, complete the following steps:

1. Select **Pools** → **Internal Storage** → **Select the drives you want to upgrade** → **Actions** → **Upgrade**.
2. As shown in Figure 10-13, browse to the test utility and drive firmware package you downloaded, as described in “Obtaining the software packages” on page 647. To upload the required files use the click **Add File** or use the option **File Upload** as described in “Provide the package manually” on page 648.

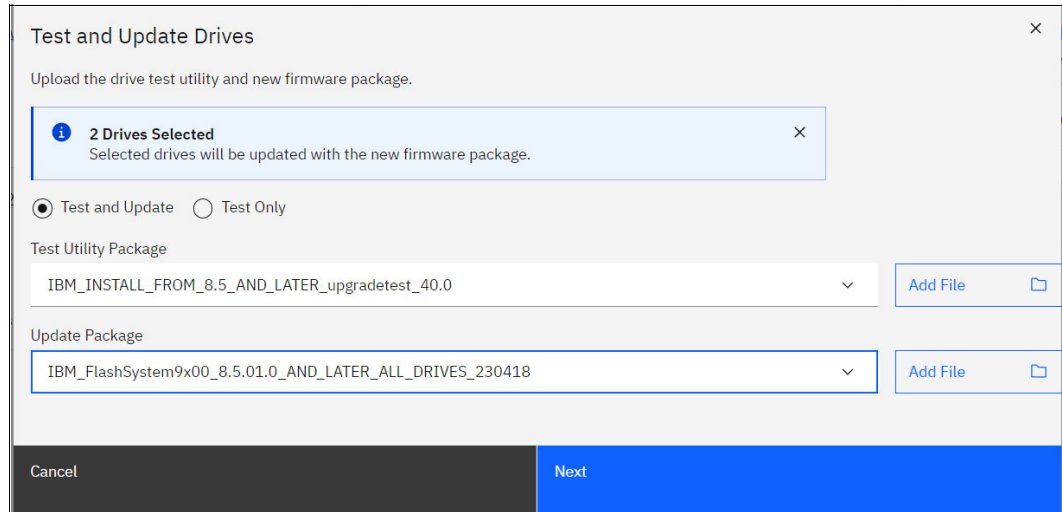


Figure 10-13 Drive firmware upgrade

3. Click **Next** to start the upgrade test utility. An example of the result is shown in Figure 10-14.

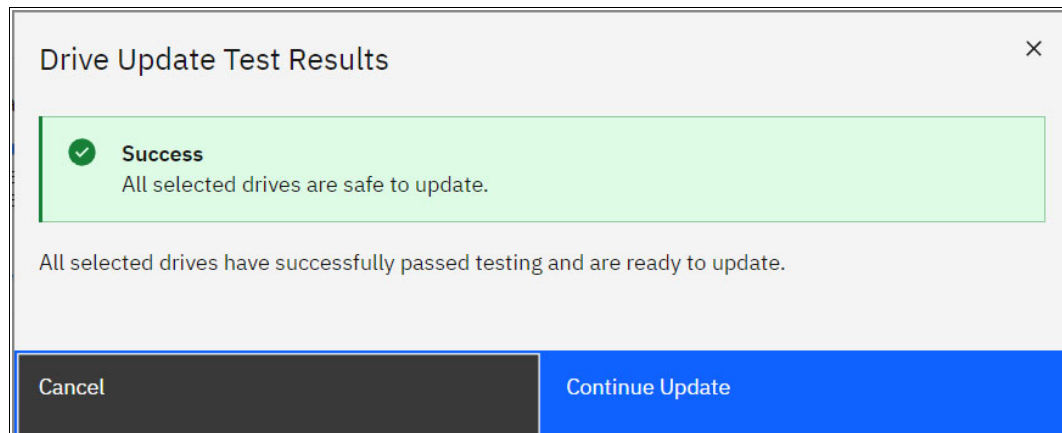


Figure 10-14 Drive update test result

4. Click **Continue Upgrade** to start the upgrade process of the selected drives.
5. To monitor the progress of the upgrade, select **Monitoring** → **Background Tasks**.

For more information, see [IBM Documentation - Updating Drive Firmware](#).

Using the command-line interface

To update the software on the system manually, complete the following steps:

1. Copy the drive firmware package to IBM Storage Virtualize by using the same method that is described in Example 10-2 on page 654.
2. Issue the following CLI command to start the update process for all drives:

```
applydrivesoftware -file <software_update_file> -type firmware -all
```

where `<software_update_file>` is the file name of the software update file. The usage of the `-all` option updates firmware on all eligible drives, including quorum drives, which is a slight risk. To avoid this risk, use the `-drive` option instead and make sure that the quorum is moved by using the `1squorum` and `chquorum` commands in between `applydrivesoftware` invocations.

Note: The maximum number of drive IDs that can be specified on a CLI by using the `-drive` option is 128. If you have more than 128 drives, use the `-all` option or run multiple invocations of `applydrivesoftware` to complete the update.

3. Issue the following CLI command to check the status of the update process:

```
1sdriveupgradeprogess
```

This command displays success when the update is complete.

4. To verify that the update successfully completed, issue the `1sdrive` command for each drive in the system. The `firmware_level` field displays the new code level for each drive. Example 10-10 demonstrates how to list the firmware level for four specific drives.

Example 10-10 Listing the firmware level for drives 0, 1, 2, and 3

```
IBM_FlashSystem:FS9500:superuser>for i in 0 1 2 3; do echo "Drive $i = `1sdrive $i|grep firmware`"; done
Drive 0 = firmware_level 3_1_8
Drive 1 = firmware_level 3_1_8
Drive 2 = firmware_level 3_1_8
Drive 3 = firmware_level 3_1_8
```

10.8 Remote Code Load

Remote Code Load (RCL) is a service offering that is provided by IBM, which allows code updates to be performed by remote support engineers instead of an onsite IBM SSR.

IBM Assist on-Site (AOS) or remote support center, or Secure Remote Access (SRA), including Call Home enablement, are required to enable RCL. With AOS enabled, a member of the IBM Support team can view your desktop and share control of your mouse and keyboard to get you on your way to a solution. The tool can also speed up problem determination, collection of data, and your problem solution.

For more information about configuring support assistance, see [IBM Documentation - Remote Code Load](#).

To request RCL for your system, go to [IBM Support - Remote Code Load](#) and select your product type. Then, complete the following steps:

1. At the IBM Remote Code Load web page, select **Product type** → **Book Now - IBM FlashSystem or SVC Remote Code Load**.

2. Click **Schedule Service** to start scheduling the service, as shown in Figure 10-15.



Figure 10-15 IBM FlashSystem RCL Schedule Service page

3. Select the Product type for RCL. Choose your device Model and Type, and click **Select**. The example in Figure 10-16 is an SVC - 2147.

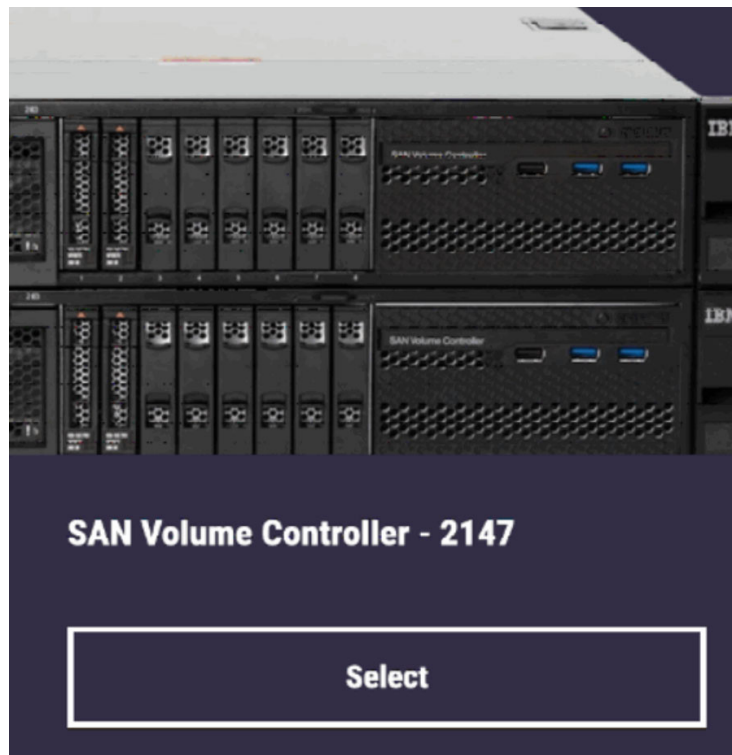


Figure 10-16 RCL Product type page

4. In the RCL timeframe option, select the date (Figure 10-17) and timeframe (Figure 10-18).

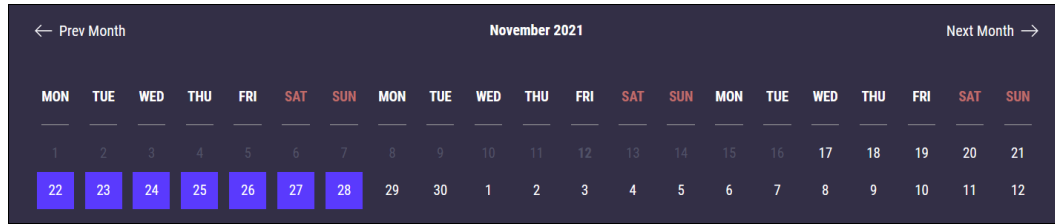


Figure 10-17 Timeframe selection page



Figure 10-18 RCL Time selection page

5. Enter your booking details into the RCL booking information form (Figure 10-19).

You can use social media to continue

Existing clients; please sign up here

Email:

Email

Password:

Password

[Remind password](#)

Sign In And Continue

New clients: please sign up here

Name: *

Name

Email: *

Email

Phone: *

Phone

Password: *

Password

Be one of the first to receive messages about our promotions and other cool stuff.

Sign Up And Continue

Figure 10-19 RCL booking information page

6. RCL will confirm booking via email and then nearer the date they will contact the client to outline the RCL process and what to expect.

10.9 Replacing IBM FlashCore Module in IBM FlashSystem

Replacing an IBM FlashCore Module (FCM) in your IBM FlashSystem requires special attention to avoid out-of-scope procedures that can damage your system. Before you start the FCM replacement procedure, review the following items to prevent any damage to your system and FCM or to your system data:

- ▶ Do not replace, reseal, or run any task on IBM Storage Virtualize if you are not sure or not comfortable with the procedure. Always engage IBM Support in case of issues or any problems during any procedure that you run.
- ▶ Call IBM Support to ensure that the logs were verified and a correct action plan was provided to replace the failed FCM.
- ▶ Confirm with the IBM SSR whether the correct FRU number was shipped to be replaced and that the action plan that was provided by IBM Support was revised.
- ▶ Confirm that your system does not have any other FCM failure or error messages in the error log tab *before* conducting the replacement procedure.

If you have the IBM Expert Care coverage feature for your IBM FlashSystem, make sure that your Technical Account Manager (TAM) is aware of the procedure and engaged with the service team to proceed with the replacement.

For more information, see [IBM FlashSystem documentation - Removing and Replacing a Drive](#).

Note: Reseating an FCM can reformat the module in specific instances. All FCM drive failure alerts must be addressed before any re-seat or replacement procedure is done. On receiving any error message for the FCM drives, escalate the problem to IBM Support.

10.10 SAN modifications

When you administer shared storage environments, human error can occur when a failure is fixed or a change is made that affects one or more servers or applications. Then, that error can affect other servers or applications because suitable precautions were not taken.

Human error can include some of the following examples:

- ▶ Disrupting or disabling the working disk paths of a server while trying to fix failed ones.
- ▶ Disrupting a neighbor SAN switch port while inserting or removing an FC cable or small form-factor pluggable (SFP).
- ▶ Disabling or removing the working part in a redundant set instead of the failed one.
- ▶ Making modifications that affect both parts of a redundant set without an interval that allows for automatic failover during unexpected problems.

Adhere to the following guidelines to perform these actions with assurance:

- ▶ Uniquely, and correctly identify the components of your SAN.
- ▶ Use the proper failover commands to disable only the failed parts.
- ▶ Understand which modifications are necessarily disruptive, and which can be performed online with little or no performance degradation.

10.10.1 Cross-referencing WWPNS

With the WWPNS of an HBA, you can uniquely identify one server in the SAN. If the name of the server is changed at the operating system level and not at the IBM Storage Virtualize host definitions, the server continues to access its mapped volumes exactly because the WWPNS of the HBA did not change.

Alternatively, if the HBA of a server is removed and installed in a second server and the SAN zones for the first server and the IBM Storage Virtualize host definitions are not updated, the second server can access volumes that it probably should not access.

To cross-reference HBA WWPNS, complete the following steps:

1. In your server, verify the WWPNS of the HBAs that are used for disk access. Typically, you can complete this task by using the SAN disk multipath software of your server.

If you are using AIX Path Control Module (AIXPCM), run the **pcmpath query wwpn** command to see output like what is shown in Example 10-11.

Example 10-11 Output of the pcmpath query wwpn command

```
[root@Server127]> pcmpath query wwpn
Adapter Name PortWWN
fscsi0       10000090FA021A13
fscsi1       10000090FA021A12
```

If your organization uses a change management tracking tool, perform all your SAN storage allocations under approved change requests with the servers' WWPNs that are listed in the Description and Implementation sections.

10.10.2 Cross-referencing LUN IDs

Always cross-reference the IBM Storage Virtualize `vdisk_uid` with the server LUN ID before you perform any modifications that involve IBM Storage Virtualize volumes.

If your organization uses a change management tracking tool, include the `vdisk_uid` and LUN ID information in every change request that performs SAN storage allocation or reclamation.

Note: Because a host can have many volumes with the same `scsi_id`, always cross-reference the IBM Storage Virtualize volume unique identifier (UID) with the host volume UID and record the `scsi_id` and LUN ID of that volume.

10.11 Server HBA replacement

Replacing a failed HBA in a server is a fairly trivial and safe operation if it is performed correctly. However, more precautions are required if your server has multiple, redundant HBAs on different SAN fabrics and the server hardware permits you to “hot” replace it (with the server still running).

To replace a failed HBA and retain the working HBA, complete the following steps:

1. In your server, identify the failed HBA and record its WWPNs. (For more information, see 10.10.1, “Cross-referencing WWPNs” on page 667.) Then, place this HBA and its associated paths offline (gracefully if possible). This approach is important so that the multipath software stops attempting to recover the HBA. Your server might even show a degraded performance while you perform this task.
2. Some HBAs have an external label that shows the WWPNs. If you have this type of label, record the WWPNs before you install the new HBA in the server.
3. If your server does not support HBA hot-swap, power off your system, replace the HBA, connect the used FC cable to the new HBA, and power on the system.

If your server does support hot-swap, follow the appropriate procedures to perform a “hot” replace of the HBA. Do *not* disable or disrupt the working HBA in the process.
4. Verify that the new HBA successfully logged in to the SAN switch. If it logged in successfully, you can see its WWPNs logged in to the SAN switch port. Otherwise, fix this issue before you continue to the next step.

Cross-check the WWPNs that you see in the SAN switch with the one that you noted in step 1, and make sure that you did not record the wrong worldwide node name (WWNN).
5. In your SAN zoning configuration tool, replace the old HBA WWPNs for the new ones in every alias and zone to which they belong. Do *not* touch the other SAN fabric (the one with the working HBA) while you perform this task.

Only one alias should use each WWPN, and zones must reference this alias.

If you are using SAN port zoning (though you should not be) and you did not move the new HBA FC cable to another SAN switch port, you do not need to reconfigure zoning.
6. Verify that the new HBA’s WWPNs appear in IBM Storage Virtualize by using the **lshostcandidate** command.

If the WWPNs of the new HBA do not appear, troubleshoot your SAN connections and zoning.
7. Add the WWPNs of this new HBA in the IBM Storage Virtualize host definition by using the **addhostport** command. Do *not* remove the old one yet. Run the **lshost <servername>** command. Then, verify that the working HBA shows as *active*, and that the failed HBA shows as *inactive* or *offline*.
8. Use software to recognize the new HBA and its associated SAN disk paths. Certify that all SAN LUNs have redundant disk paths through the working HBA and the new HBA.
9. Return to IBM Storage Virtualize and verify again (by using the **lshost <servername>** command) that both the working and the new HBA’s WWPNs are active. In this case, you can remove the old HBA WWPNs from the host definition by using the **rmhostport** command.
10. Do not remove any HBA WWPNs from the host definition until you ensure that you have at least two active ones that are working correctly.

By following these steps, you avoid removing your only working HBA by mistake.

10.12 Hardware upgrades

The IBM Storage Virtualize family scalability features allow significant flexibility in its configuration. The IBM Storage Virtualize family is based on control enclosures and expansion enclosures.

- ▶ Control enclosures (IBM FlashSystem 9500) manage your storage systems, communicate with the host, and manage interfaces. In addition, they can also house up to 48 NVMe-capable flash drives.
- ▶ Control enclosures (IBM SVC) manage your storage systems, communicate with the host, and manage interfaces. Each IBM SVC control enclosure contains a single node. To build a redundant SVC cluster, two SVC control enclosures are required.
- ▶ Expansion enclosures increase the available capacity of an IBM Storage Virtualize cluster. They communicate with the control enclosure through a dual pair of 12 Gbps serial-attached SCSI (SAS) connections. These expansion enclosures can house many flash (SSD) SAS type drives.

A basic configuration of an IBM Storage Virtualize storage platform consists of one IBM Storage Virtualize control enclosure. For a balanced increase of performance and scale, up to four (depending on model) IBM Storage Virtualize control enclosures can be clustered into a single storage system. Similarly, to increase capacity, up to two chains (depending on model) of expansion enclosures can be added per control enclosure. Therefore, several scenarios are possible for its growth.

10.12.1 Adding control enclosures

If your IBM Storage Virtualize cluster is below the maximum I/O groups limit for your specific product and you intend to upgrade it, you can install another control enclosure. It is also feasible that you might have traditional cluster nodes that you want to add the IBM Storage Virtualize enclosures to because the latter are more powerful than your existing ones. Therefore, your cluster can include different node models in different I/O groups. For more information, see [Adding a control enclosure to an existing system](#).

To install these control enclosures, determine whether you need to upgrade your IBM Storage Virtualize first.

For more information, see 10.5.3, “Hardware considerations” on page 650.

Note: If exactly two control enclosures are in a system, you must set up a quorum disk or application outside of the system. If the two control enclosures lose communication with each other, the quorum disk prevents both I/O groups from going offline.

After you install the new nodes, you might need to redistribute your servers across the I/O groups. Consider the following points:

- ▶ Moving a server’s volume to different I/O groups can be done online because of a feature called Non-Disruptive Volume Movement (NDVM). Although this process can be done without stopping the host, careful planning and preparation are advised.

Note: You cannot move a volume that is in a type of remote copy relationship.


```

IBM_FlashSystem:FS9500:superuser>rmhostiogrp -iogrp 3 Win2012srv1

IBM_FlashSystem:FS9500:superuser>lshostiogrp Win2012srv1
id name
0 io_grp0
1 io_grp1
2 io_grp2

IBM_FlashSystem:FS9500:superuser>addhostiogrp -iogrp io_grp3 Win2012srv1

IBM_FlashSystem:FS9500:superuser>lshostiogrp Win2012srv1
id name
0 io_grp0
1 io_grp1
2 io_grp2
3 io_grp3

IBM_FlashSystem:FS9500:superuser>lsiogrp
id name node_count vdisk_count host_count site_id site_name
0 io_grp0 2 11 2
1 io_grp1 0 0 2
2 io_grp2 0 0 2
3 io_grp3 0 0 2
4 recovery_io_grp 0 0 0

```

- If possible, avoid setting a server to use volumes from different I/O groups that have different node types for extended periods. Otherwise, as this server's storage capacity grows, you might experience a performance difference between volumes from different I/O groups. This mismatch makes it difficult to identify and resolve eventual performance problems.

10.12.2 Adding IBM SVC nodes

You can add nodes to replace the existing nodes of your SVC cluster with newer ones, and the replacement procedure can be performed non-disruptively. The new node can assume the WWNN of the node you are replacing, which requires no changes in host configuration, SAN zoning, or multipath software. For more information about this procedure, see [IBM Documentation - Installing Node](#).

Alternatively, you can add nodes to expand your system. If your SVC cluster is below the maximum I/O groups limit for your specific product and you intend to upgrade it, you can install another I/O group.

For more information, see 10.5.3, "Hardware considerations" on page 650.

Note: If I/O groups are present in a system, you must set up a quorum disk or application outside of the system. If the I/O groups lose communication with each other, the quorum disk prevents split brain scenarios.

For more information about adding a node to an SVC cluster, see Chapter 3, "Initial configuration", of *Implementation Guide for IBM Storage FlashSystem and IBM SAN Volume Controller: Updated for IBM Storage Virtualize Version 8.6*, SG24-8542.

Note: Use a consistent method (only the management GUI or only the CLI) when you add, remove, and re-add nodes. If a node is added by using the CLI and later re-added by using the GUI, it might get a different node name than it originally had.

After you install the newer nodes, you might need to redistribute your servers across the I/O groups. Consider the following points:

- ▶ Moving a server's volume to different I/O groups can be done online because of a feature called Non-Disruptive Volume Movement (NDVM). Although this process can be done without stopping the host, careful planning and preparation are advised.

Note: You cannot move a volume that is in any type of remote copy relationship.

- ▶ If each of your servers is zoned to only one I/O group, modify your SAN zoning configuration as you move its volumes to another I/O group. As best you can, balance the distribution of your servers across I/O groups according to I/O workload.
- ▶ Use the `-iogrp` parameter with the `mkhost` command to define which I/O groups of the SVC that the new servers will use. Otherwise, SVC by default maps the host to all I/O groups, even if they do not exist and regardless of your zoning configuration. Example 10-14 on page 672 shows this scenario and how to resolve it by using the `rmhostiogrp` and `addhostiogrp` commands.
- ▶ If possible, avoid setting a server to use volumes from different I/O groups that have different node types for extended periods. Otherwise, as this server's storage capacity grows, you might experience a performance difference between volumes from different I/O groups. This mismatch makes it difficult to identify and resolve eventual performance problems.

Adding hot-spare nodes

To reduce the risk of a loss of redundancy or degraded system performance, hot-spare nodes can be added to the system. A hot-spare node has active system ports, but no host I/O ports, and it is not part of any I/O group. If any node fails or is upgraded, this spare node automatically joins the system and assumes the place of the failed node, and restores redundancy.

The hot-spare node uses the same NPIV WWPNs for its FC ports as the failed node, so host operations are not disrupted. After the failed node returns to the system, the hot-spare node returns to the Spare state, which indicates it can be automatically swapped for other failed nodes on the system.

The following restrictions apply to the usage of hot-spare nodes on the system:

- ▶ Hot-spare nodes can be used with FC-attached external storage only.
- ▶ Hot-spare nodes cannot be used in the following situations:
 - In systems that use Remote Direct Memory Access (RDMA)-capable Ethernet ports for node-to-node communications.
 - On enclosure-based FlashSystem storage systems.
 - With SAS-attached storage.
 - With iSCSI-attached storage.
 - With storage that is directly attached to the system.
- ▶ A maximum of four hot-spare nodes can be added to the system.

Using the management GUI

If your nodes are configured on your systems and you want to add hot-spare nodes, you must connect the extra nodes to the system. After hot-spare nodes are configured correctly on the system, you can add the spare node to the system configuration by completing the following steps:

1. In the management GUI, select **Monitoring** → **System Hardware**.
2. On the System Hardware - Overview window, click **Add Nodes**.
3. On the Add Node page, select the hot-spare node to add to the system.

If your system uses stretched or HyperSwap system topology, hot-spare nodes must be designated per site.

Using the command-line interface

To add a spare node to the system, run the following command:

```
addnode -panelname <panel_name> -spare
```

Where *<panel_name>* is the name of the node that is displayed in the Service Assistant or in the output of the `lsnodecandidate` command.

For more information, see *IBM Spectrum Virtualize: Hot-Spare Node and NPIV Target Ports*, REDP-5477.

10.12.3 Upgrading nodes in an existing cluster

If you want to upgrade the nodes or canisters of your existing IBM Storage Virtualize system, you can increase the cache memory size or the adapters in each node. This task can be done one node at a time to be nondisruptive to systems operations.

When evaluating cache memory upgrades, consider the following points:

- ▶ As your working set and total capacity increases, you should consider increasing your cache memory size. A *working set* is the most accessed workloads, excluding snapshots and backups. *Total capacity* implies more or larger workloads and a larger working set.
- ▶ If you are consolidating from multiple controllers, consider at least matching the amount of cache memory across those controllers.
- ▶ When externally virtualizing controllers (such as a switched virtual circuit), a large cache can accelerate older controllers with smaller caches.
- ▶ If you are using a data reduction pool (DRP), then maximize the cache size and consider adding SCM drives with Easy Tier for the best performance.
- ▶ If you are making heavy use of copy services, consider increasing the cache beyond just your working set requirements.
- ▶ A truly random working set might not benefit greatly from the cache.

Important: Do not power on a node that is shown as offline in the management GUI if you powered off the node to add memory to increase total memory. Before you increase memory, you must remove the node from the system so that it is not showing in the management GUI or in the output from the `svcinfo lsnode` command.

Do not power on a node that is still in the system and showing as offline with more memory than the node had when it powered off. Such a node can cause an immediate outage or an outage when you update the system software.

When evaluating adapter upgrades, consider the following points:

- ▶ A single 32-Gb FC port can deliver over 3 GBps (allowing for overhead).
- ▶ A 32-Gb FC card in each canister with eight ports can deliver more than 24 GBps.
- ▶ An FCM NVMe device can perform at over 1 GBps.
- ▶ A single 32-Gb FC port can deliver 80,000 - 125,000 IOPS with a 4k block size.
- ▶ A 32-Gb FC card in each canister with eight ports can deliver up to 1,000,000 IOPS.
- ▶ A IBM FlashSystem 9500 can deliver up to 8,000,000 4k read hit IOPS.
- ▶ If you have more than 12 NVMe devices, consider using two FC cards per canister. A third FC card allows you to achieve up to 45 GBps.
- ▶ If you want to achieve more than 800,000 IOPS, use at least two FC cards per canister.
- ▶ If IBM Storage Virtualize is performing remote copy or clustering, consider using separate ports to ensure that there is no conflict with host traffic.
- ▶ iSCSI Extensions for RDMA (iSER) through 25-gigabit Ethernet (GbE) ports has similar capabilities as 16-Gb FC ports, but with less overall ports available. If you are planning on using 10-Gb iSCSI, ensure that it can service your expected workloads.

Real-time performance statistics are available in the management GUI by selecting **Monitoring** → **Performance**, as shown in Figure 10-20.

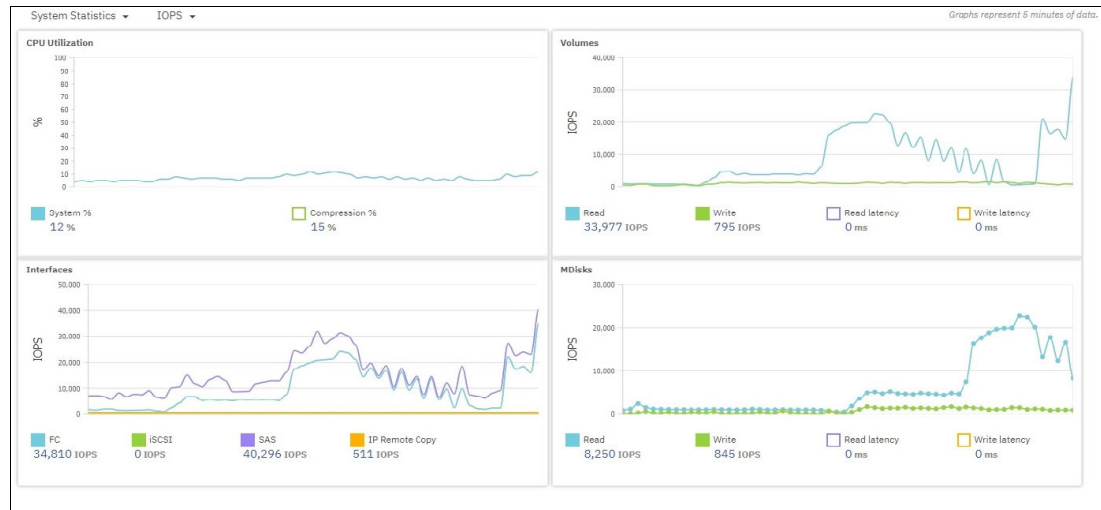


Figure 10-20 IBM Storage Virtualize performance statistics (IOPS)

Memory options for an IBM FlashSystem 9500 control enclosure

On the system board, the DIMM slots are labeled according to their memory channel and slot. They are associated with the CPU nearest to their DIMM slots. You can install three distinct memory configurations in those 24 DIMM slots in each node canister.

Important: The memory in both node canisters must be configured identically to create the total enclosure memory size.

Table 10-2 on page 677 shows the available memory configuration for each IBM FlashSystem and SVC control enclosure. Each column gives the valid configuration for each total enclosure memory size. DIMM slots are listed in the same order that they appear in the node canister.

To ensure proper cooling and a steady flow of air from the fan modules in each node canister, blank DIMMs must be inserted in any slot that does not contain a memory module.

Table 10-2 Available memory configurations for one node in a control enclosure

DIMM slot	Total enclosure memory 1024 GB	Total enclosure memory 2048 GB	Total enclosure memory 3072 GB
1 (CPU1)	Blank	64 GB	64 GB
2 (CPU1)	64 GB	64 GB	64 GB
3 (CPU1)	Blank	Blank	64 GB
4 (CPU1)	Blank	64 GB	64 GB
5 (CPU1)	64 GB	64 GB	64 GB
6 (CPU1)	Blank	Blank	64 GB
7 (CPU0)	Blank	Blank	64 GB
8 (CPU0)	64 GB	64 GB	64 GB
9 (CPU0)	Blank	Blank	64 GB
10 (CPU0)	Blank	64 GB	64 GB
11 (CPU0)	64 GB	64 GB	64 GB
12 (CPU0)	Blank	Blank	64 GB
13 (CPU0)	Blank	64 GB	64 GB
14 (CPU0)	64 GB	64 GB	64 GB
15 (CPU0)	Blank	Blank	64 GB
16 (CPU0)	Blank	64 GB	64 GB
17 (CPU1)	64 GB	64 GB	64 GB
18 (CPU1)	Blank	Blank	64 GB
19 (CPU1)	Blank	Blank	64 GB
20 (CPU1)	64 GB	64 GB	64 GB
21 (CPU1)	Blank	64 GB	64 GB
22 (CPU1)	Blank	Blank	64 GB
23 (CPU1)	64 GB	64 GB	64 GB
24 (CPU1)	Blank	64 GB	64 GB

Adapter options for an IBM FlashSystem 9500/SV3 control enclosure

For the IBM FlashSystem 9500, these adapters are added as a pair (one card in each node). Six Peripheral Component Interconnect Express (PCIe) slots are available for port expansions in the IBM FlashSystem 9500 control enclosure. Each canister has three PCIe adapter cages with two slots each, and both canisters must have the same configuration. The PCIe adapter feature codes offer a pair of adapters to ensure that they are supplied symmetrically in each canister.

The control enclosure can be configured with three I/O adapter features to provide up to forty-eight 32-Gb FC ports or up to ten 25-GbE (iSCSI or iSER capable) ports. The control enclosure also includes eight 10-GbE ports as standard for iSCSI connectivity and two 1-GbE ports for system management. A feature code also is available to include the SAS Expansion card if the user wants to use optional expansion enclosures (only valid for FS9500). The options for the features that are available are shown in Table 10-3.

Table 10-3 IBM FlashSystem 9500 control enclosure adapter options

Number of control enclosures	32 Gbps FC	100 GbE	25 GbE
1	12	6	10
2	24	12	20

For more information about the feature codes, memory options, and functions of each adapter, see *IBM FlashSystem 9500 Product Guide*, REDP-5669.

Adapter options for an IBM FlashSystem 7300 control enclosure

You can also add new adapters to the IBM FlashSystem 7300 nodes. These adapters are added as a pair (one card in each node). Six PCIe slots are available for port expansions in the IBM FlashSystem 7300 control enclosure. Each canister has three PCIe adapter slots each, and both canisters must have the same configuration. The PCIe adapter feature codes offer a pair of adapters to ensure that they are supplied symmetrically in each canister.

The control enclosure can be configured with three I/O adapter features to provide up to twenty-four 32-Gb FC ports or up to twelve 25-GbE (iSCSI or iSER capable) ports. The control enclosure also includes eight 10-GbE ports as standard for iSCSI connectivity and two 1-GbE ports for system management. A feature code is also available to include the SAS Expansion card if the user wants to use optional expansion enclosures. The options for the features available are shown in Table 10-4.

Table 10-4 IBM FlashSystem 7300 control enclosure adapter options

Number of control enclosures	16 Gbps/ 32 Gbps FC	100 GbE	25 GbE
1	12	6	6
2	24	12	12
3	36	18	18
4	48	24	24

For more information about the feature codes, memory options, and functions of each adapter, see *IBM FlashSystem 7300 Product Guide*, REDP-5668.

Adapter options for an IBM FlashSystem 5x00 control enclosure

All the IBM FlashSystem 5000 control enclosures include 1-GbE or 10-GbE ports as standard for iSCSI connectivity. The standard connectivity can be extended with extra ports or enhanced with more connectivity through an optional I/O adapter feature. Table 10-5 lists which configurations are standard for the IBM FlashSystem 5015, 5035, 5045 and 5200.

Table 10-5 IBM FlashSystem 5000 family standard configurations

Platform	IBM FlashSystem 5015	IBM FlashSystem 5035/5045	IBM FlashSystem 5200
iSCSI	One 1-GbE tech port + iSCSI One 1-GbE iSCSI only	One 1-GbE dedicated tech port	One 1-GbE dedicated tech port
iSCSI	N/A	Two 10-GbE (iSCSI only)	Four 10-GbE (iSCSI only)
SAS	One 12-Gb SAS expansion	Two 12-Gb SAS expansion	N/A

Table 10-6 lists the possible adapter installation for IBM FlashSystem 5015, 5035, 5045 and 5200. Only one interface card can be installed per canister, and the interface card must be the same in both canisters.

Table 10-6 IBM FlashSystem 5000 family adapters

Platform	IBM FlashSystem 5015	IBM FlashSystem 5035/5045	IBM FlashSystem 5200
FC	4-port 16-Gb FC or	4-port 16-Gb FC or	4-port 16-Gb FC or FC NVMeoF
iSCSI	4-port 10-GbE iSCSI or	4-port 10-GbE iSCSI or	2-port 25-GbE ROCE ISER or iSCSI
iSCSI	2-port 25-GbE iSCSI or	2-port 25-GbE iSCSI or	2-port 25-GbE internet Wide-area RDMA Protocol (iWARP) ISER or iSCSI
SAS	4-port 12-Gb SAS host attach	4-port 12-Gb SAS host attach	2-port 12-Gb SAS to allow SAS expansions

IBM FlashSystem 5015 and 5035/5045 control enclosures include 1-GbE or 10-GbE ports as standard for iSCSI connectivity. The standard connectivity can be extended by using more ports or enhanced with more connectivity through an optional I/O adapter feature. For more information, see [Family 2072+06 IBM FlashSystem 5015 and 5035](#).

IBM FlashSystem 5015 control enclosure models 2N2 and 2N4

IBM FlashSystem 5015 control enclosure models 2N2 and 2N4 include the following features:

- ▶ Two node canisters, each with a 2-core processor.
- ▶ 32-GB cache (16 GB per canister) with an optional 64-GB cache (32 GB per canister).
- ▶ 1-Gb iSCSI connectivity standard with optional 16-Gb FC, 12-Gb SAS, 10-Gb iSCSI (optical), 25-Gb iSCSI (optical), and 1-Gb iSCSI connectivity.

- ▶ 12-Gb SAS ports for expansion enclosure attachment.
- ▶ Twelve slots for 3.5-inch large form factor (LFF) SAS drives (Model 2N2) and 24 slots for 2.5-inch small form factor (SFF) SAS drives (Model 2N4).
- ▶ 2U 19-inch rack mount enclosure with 100 - 240 V AC or -48 V DC power supplies.

IBM FlashSystem 5035 control enclosure models 3N2 and 3N4

IBM FlashSystem 5035 control enclosure models 3N2 and 3N4 include the following features:

- ▶ Two node canisters, each with a 6-core processor.
- ▶ 32-GB cache (16 GB per canister) with an optional 64-GB cache (32 GB per canister).
- ▶ 10-Gb iSCSI (copper) connectivity standard with optional 16-Gb FC, 12-Gb SAS, 10-Gb iSCSI (optical), and 25-Gb iSCSI (optical) connectivity.
- ▶ 12-Gb SAS ports for expansion enclosure attachment.
- ▶ Twelve slots for 3.5-inch LFF SAS drives (Model 3N2) and 24 slots for 2.5-inch SFF SAS drives (Model 3N4).
- ▶ 2U 19-inch rack mount enclosure with 100 - 240 V AC or -48 V DC power supplies.

IBM FlashSystem 5045 control enclosure models 3P2 and 3P4

IBM FlashSystem 5045 control enclosure models 3P2 and 3P4 include the following features:

- ▶ Two node canisters, each with a 6-core processor.
- ▶ 32 GB and 64 GB cache options (16 GB or 32 GB per canister)
- ▶ 10 Gb iSCSI (copper) connectivity standard with optional 16 Gb FC, 12 Gb SAS, and
- ▶ 25/10 Gb iSCSI (optical) connectivity 12-Gb SAS ports for expansion enclosure attachment.
- ▶ 12 Gb SAS ports for expansion enclosure attachment
- ▶ Twelve 3.5-inch LFF or twenty-four 2.5-inch SFF drive slots within the enclosure
- ▶ A 2U, 19-inch rackmount enclosure with either AC or DC power

IBM FlashSystem 5200 control enclosure models 6H2

- ▶ Two node canisters, each with an eight-core processor
- ▶ 64 GB, 256 GB, and 512 GB cache options (32 GB, 128 GB, or 256 GB per canister)
- ▶ 10 Gb iSCSI (copper) connectivity standard with optional 32/16 Gb FC, 12 Gb SAS, and 25/10 Gb iSCSI (optical) connectivity
- ▶ 12 Gb SAS ports for expansion enclosure attachment
- ▶ Twelve slots for 2.5-inch SFF NVMe drives
- ▶ 1U, 19-inch rack mount enclosure with 200 - 240 V AC power supplies

10.12.4 Upgrading NVMe drives

To provide ultra-low latency for performance-sensitive but less cache-friendly workloads, SCM drives from Intel and Samsung are available as a persistent storage tier for the IBM Storage Virtualize family. SCM is a substantial step forward in memory technology, offering nonvolatile, ultra-low latency memory for a fraction of the cost of traditional memory chips.

IBM FlashSystem products support SCM drives over NVMe to improve overall storage performance, or offer a higher performance storage pool. SCM drives can be used for small workloads that need exceptional levels of performance at the lowest latencies, or they can be combined with other NVMe drives by using Easy Tier to accelerate much larger workloads. Like FCM, SCM drives are also available as upgrades for the previous generation of all flash arrays.

IBM Storage Virtualize 8.6 supports up to 12 SCM drives in a control enclosure for the IBM FlashSystem 9000, 7000, and 5200 families.

For more information about SCM, see Chapter 3, “Storage back-end” on page 191.

10.12.5 Splitting an IBM Storage Virtualize cluster

Splitting an IBM Storage Virtualize cluster might become a necessity if you have one or more of the following requirements:

- ▶ To grow the environment beyond the maximum number of I/O groups that a clustered system can support.
- ▶ To grow the environment beyond the maximum number of attachable subsystem storage controllers.
- ▶ To grow the environment beyond any other maximum system limit.
- ▶ To achieve new levels of data redundancy and availability.

By splitting the clustered system, you no longer have one IBM Storage Virtualize cluster that handles all I/O operations, hosts, and subsystem storage attachments. The goal is to create a second IBM Storage Virtualize cluster so that you can equally distribute the workload over the two systems.

After safely removing enclosures from the existing cluster and creating a second IBM Storage Virtualize cluster, choose from the following approaches to balance the two systems:

- ▶ Attach new storage subsystems and hosts to the new system and start adding only new workloads on the new system.
- ▶ Migrate the workload onto the new system by using the approach that is described in Chapter 3, “Storage back-end” on page 191.

10.12.6 IBM FlashWatch

Driven by the concept of “Storage Made Simple,” *IBM FlashWatch* is a suite of programs that enhances your experience of owning an IBM Storage Virtualize storage. Bringing together programs, which span the acquisition, operation, and migration phases, this suite aims to reduce deployment and operational risks to improve your support experience, and to offer a fully flexible, commitment-free hardware refresh. IBM FlashWatch is an offering from IBM to complement the purchase of the IBM Storage Virtualize product. IBM FlashWatch provides the following features that are included in the purchase of the product:

- ▶ IBM Flash Momentum

IBM Flash Momentum is a storage upgrade program, which you can use to replace your controller and storage every 3 years with full flexibility. Before the expiration of the agreement period, you decide whether to keep IBM Storage Virtualize, refresh it, or walk away. You can refresh IBM Storage Virtualize for the same monthly price or less, or upsize or downsize your system to meet your needs.

- ▶ High availability (HA) guarantee

Robust IBM Storage Virtualize software has a measured availability of 99.9999%, and IBM offers an optional 100% availability commitment when HyperSwap is also used.
- ▶ Data reduction guarantee

A 3:1 data reduction is ensured, and you must self-certify that the data you are writing can be reduced (for example, not encrypted or compressed). Up to 5:1 data reduction can be ensured with more detailed profiling of your workload.
- ▶ All-inclusive licensing

All storage functions available are included in the licensing cost for internal storage.
- ▶ Comprehensive care

Up to 7 years of 24x7 support, with 3 years of IBM Technical Account Managers (TAMs), enhanced response times of 30 minutes for Severity 1 incidents, and six managed code upgrades over 3 years. However, this feature is not available for all IBM Storage Virtualize models. For more information, see [IBM FlashWatch FAQ](#).
- ▶ IBM Storage Insights

IBM Storage Insights is included at no extra cost to proactively manage your environment.
- ▶ Flash endurance guarantee

Flash media is covered for all workloads while under warranty or maintenance.
- ▶ IBM Storage Utility pricing

The IBM Storage Utility pricing solution delivers 3 years of your planned capacity needs on day one. To predict and control your future needs, IBM uses IBM Storage Insights to help you easily meet your capacity needs without interrupting your data center. The IBM Storage Virtualize 9200 (9848-UG8) is leased through IBM Global Finance on a 3-year lease, which entitles the customer to use approximately 30 - 40% of the total system capacity at no extra cost. If the storage must increase beyond that initial capacity, usage is billed on a quarterly basis based on the average daily provisioned capacity per terabyte per month.
- ▶ No cost migration

For a 90-day period, from the date of installation, you can migrate data from over 500 older storage systems (IBM and other vendors) to your IBM Storage Virtualize product by using an approach of your choice, without having to pay any extra external licensing.

For more information, see [IBM Flashwatch](#).

10.13 I/O throttling

I/O throttling is a mechanism that you can use to limit the volume of I/O that is processed by the storage controller at various levels to achieve quality of service (QoS). If a throttle is defined, the system either processes the I/O, or delays the processing of the I/O to free resources for more critical I/O. Throttling is a way to achieve a better distribution of storage controller resources.

IBM Storage Virtualize 8.3 and later brings the possibility for you to set the throttling at volume, host, host cluster, or storage pool levels, and offload throttling by using the GUI. This section describes some details of I/O throttling and shows how to configure the feature in your system.

10.13.1 General information about I/O throttling

I/O throttling has the following characteristics:

- ▶ IOPS and bandwidth throttle limits can be set.
- ▶ It is an upper bound QoS mechanism.
- ▶ No minimum performance is guaranteed.
- ▶ Volumes, hosts, host clusters, and MDisk groups can be throttled.
- ▶ Queuing occurs at microsecond granularity.
- ▶ Internal I/O operations (FlashCopy, cluster traffic, and so on) are not throttled.
- ▶ Reduces I/O bursts and smooths the I/O flow with variable delay in throttled I/Os.
- ▶ The throttle limit is a per-node value.

10.13.2 I/O throttling on front-end I/O control

You can use throttling for better front-end I/O control at the volume, host, host cluster, and offload levels:

- ▶ In a multi-tenant environment, hosts can have their own defined limits.
You can use this feature to allow restricted I/Os from a data-mining server and a higher limit for an application server.
- ▶ An aggressive host consuming bandwidth of the controller can be limited by a throttle.
For example, a video streaming application can have a limit that is set to avoid consuming too much of the bandwidth.
- ▶ Restrict a group of hosts by their throttles.
For example, Department A gets more bandwidth than Department B.
- ▶ Each volume can have a defined throttle.
For example, a volume that is used for backups can be configured to use less bandwidth than a volume that is used for a production database.
- ▶ When performing migrations in a production environment, consider using host or volume level throttles.
- ▶ Offloaded I/Os.

Offload commands, such as **UNMAP** and **XCOPY**, free hosts and speed the copy process by offloading the operations of certain types of hosts to a storage system. These commands are used by hosts to format new file systems or copy volumes without the host needing to read and then write data. Throttles can be used to delay processing for offloads to free bandwidth for other more critical operations, which can improve performance but limits the rate at which host features, such as VMware VMotion, can copy data.

10.13.3 I/O throttling on back-end I/O control

You can also use throttling to control back-end I/O by throttling the storage pools, which can be useful in the following scenarios:

- ▶ Each storage pool can have a defined throttle.
- ▶ Allows control of back-end I/Os from IBM Storage Virtualize.
- ▶ Useful to avoid overwhelming any external back-end storage.
- ▶ Useful in VVOLs because a VVOL is created in a child pool. A child pool (**mdiskgrp**) throttle can control I/Os coming from that VVOL.

- ▶ Only parent pools support throttles because only parent pools contain MDisks from internal or external back-end storage. For volumes in child pools, the throttle of the parent pool is applied.
- ▶ If more than one throttle applies to an I/O operation, the lowest and most stringent throttle is used. For example, if a throttle of 100 MBps is defined on a pool and a throttle of 200 MBps is defined on a volume of that pool, the I/O operations are limited to 100 MBps.

10.13.4 Overall benefits of using I/O throttling

The overall benefits of using I/O throttling for a better distribution of all system resources include the following ones:

- ▶ Avoids overwhelming the controller objects.
- ▶ Avoids starving the external entities, like *hosts*, from their share of the controller.
- ▶ Creates scheme of distribution of controller resources that results in better utilization of external resources, such as host capacities.

With throttling not enabled, a scenario exists where Host1 dominates the bandwidth, and after enabling the throttle, a much better distribution of the bandwidth among the hosts results, as shown in Figure 10-21.

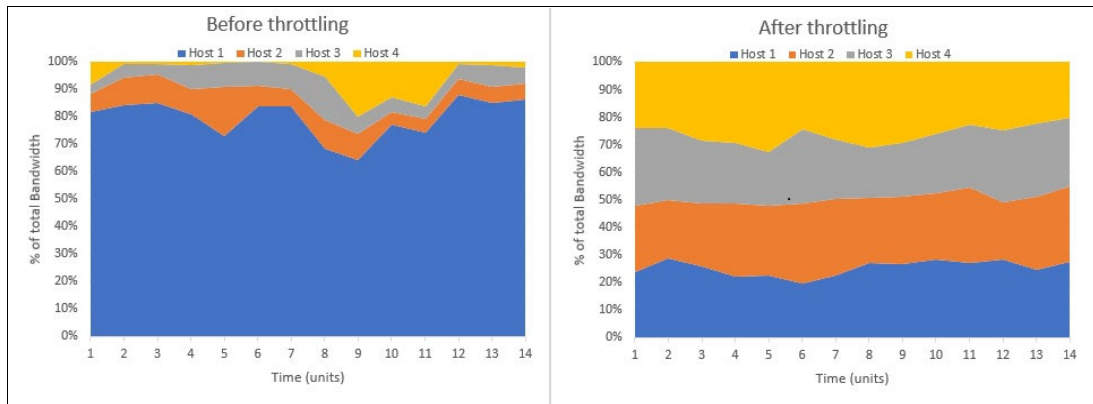


Figure 10-21 Distribution of controller resources before and after I/O throttling

10.13.5 Considerations for I/O throttling

Consider the following points when you are planning to use I/O throttling:

- ▶ The throttle cannot be defined for the host if it is part of a host cluster that has a host cluster throttle.
- ▶ If the host cluster does not have a throttle that is defined, its member hosts can have their individual host throttles defined.
- ▶ If a volume has multiple copies, then throttling is done for the storage pool serving the primary copy. The throttling is not applicable on the secondary pool for mirrored volumes and stretched cluster implementations.
- ▶ A host cannot be added to a host cluster if both have their individual throttles defined. If only one of the host or host cluster throttles is present, the command succeeds.
- ▶ A seeding host that is used for creating a host cluster cannot have a host throttle that is defined for it.

Note: Throttling is applicable only for the I/Os that IBM Storage Virtualize receives from hosts and host clusters. The I/Os that are generated internally, such as mirrored volume I/Os, cannot be throttled.

10.13.6 Configuring I/O throttling by using the CLI

To create a throttle by using the CLI, use the `mkthrottle` command, as shown in Example 10-15. The bandwidth limit is the maximum amount of bandwidth that the system can process before the system delays I/O processing. Similarly, the `iops_limit` is the maximum amount of IOPS that the system can process before the system delays I/O processing.

Example 10-15 Creating a throttle by using the `mkthrottle` command in the CLI

Syntax:

```
mkthrottle -type [offload | vdisk | host | hostcluster | mdiskgrp]
            [-bandwidth bandwidth_limit_in_mb]
            [-iops iops_limit]
            [-name throttle_name]
            [-vdisk vdisk_id_or_name]
            [-host host_id_or_name]
            [-hostcluster hostcluster_id_or_name]
            [-mdiskgrp mdiskgrp_id_or_name]
```

Usage examples:

```
IBM_FlashSystem:FS9500:superuser>mkthrottle -type host -bandwidth 100 -host
ITS0_HOST3
```

```
IBM_FlashSystem:FS9500:superuser>mkthrottle -type host cluster -iops 30000
-hostcluster ITS0_HOSTCLUSTER1
```

```
IBM_FlashSystem:FS9500:superuser>mkthrottle -type mdiskgrp -iops 40000 -mdiskgrp 0
```

```
IBM_FlashSystem:FS9500:superuser>mkthrottle -type offload -bandwidth 50
```

```
IBM_FlashSystem:FS9500:superuser>mkthrottle -type vdisk -bandwidth 25 -vdisk
volume1
```

```
IBM_FlashSystem:FS9500:superuser>lsthrottle
```

```
throttle_id throttle_name object_id object_name      throttle_type IOPs_limit
bandwidth_limit_MB
```

```
0          throttle0    2          ITS0_HOST3      host          100
1          throttle1    0          ITS0_HOSTCLUSTER1 host cluster
30000
2          throttle2    0          Pool0          mdiskgrp
40000
3          throttle3                    offload          50
4          throttle4    10         volume1        vdisk          25
```

Note: You can change a throttle parameter by using the `chthrottle` command.

10.13.7 Creating a throttle by using the GUI

The system supports throttles on volumes, hosts, host clusters, storage pools and copy offload operations.

Creating a volume throttle

To create a volume throttle, select **Volumes** → **Volumes** → **Actions**, select **Edit Throttle**, as shown in Figure 10-22. The bandwidth can be set 1 MBps - 256 TBps, and IOPS can be set 1 - 33,254,432.

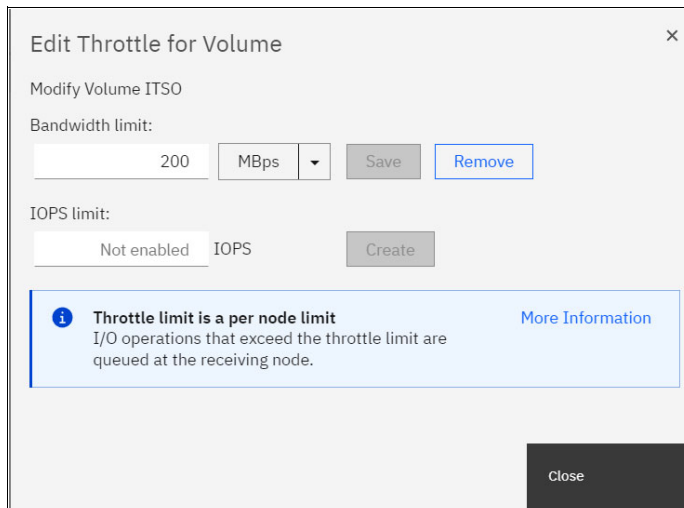


Figure 10-22 Creating a volume throttle in the GUI

If a throttle already exists, the dialog box that is shown in Figure 10-22 also shows a **Remove** button that you can use to delete the throttle.

Creating a host throttle

To create a host throttle, select **Hosts** → **Hosts** → **Actions**, select **Edit Throttle**, as shown in Figure 10-23.

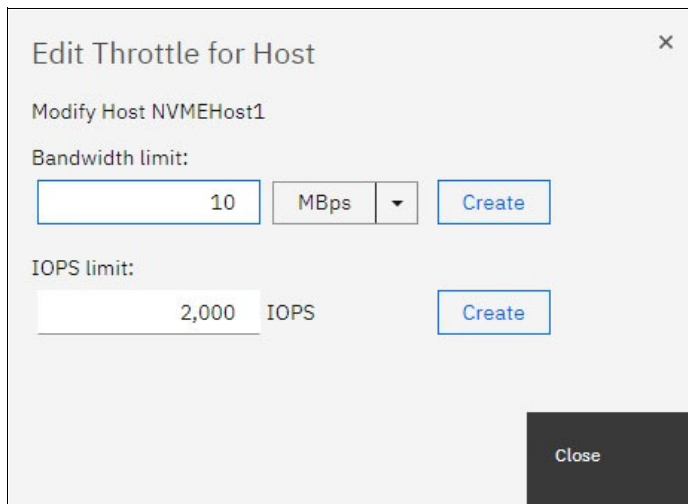


Figure 10-23 Creating a host throttle in the GUI

Creating a host cluster throttle by using the GUI

To create a host cluster throttle, select **Hosts** → **Host Clusters** → **Actions**, select **Edit Throttle**, as shown in Figure 10-24.

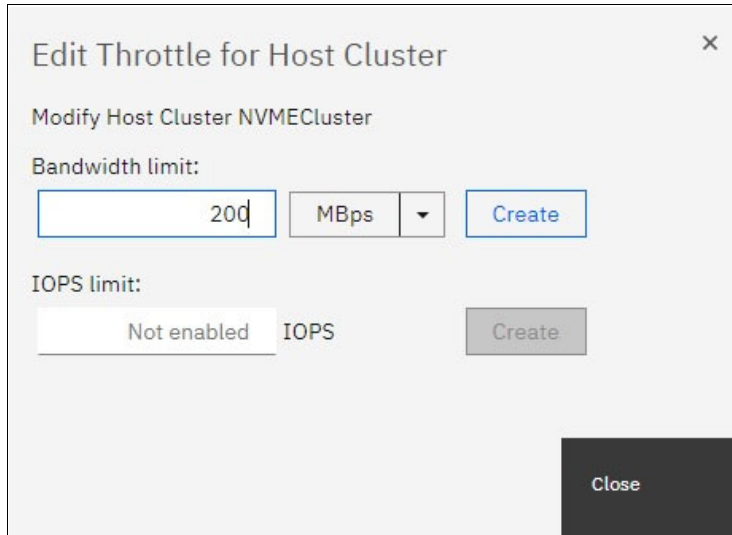


Figure 10-24 Creating a host cluster throttle in the GUI

Creating a storage pool throttle by using the GUI

To create a storage pool throttle, select **Pools** → **Pools** → **Actions**, select **Edit Throttle**, as shown in Figure 10-25.

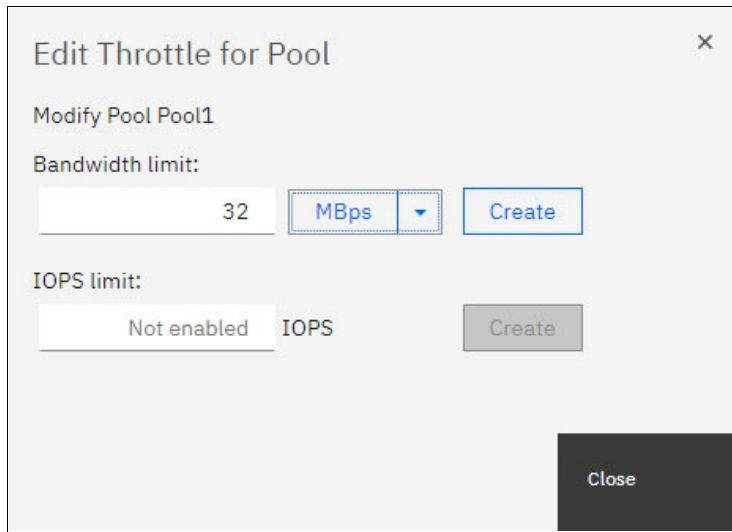


Figure 10-25 Creating a storage pool throttle in the GUI

Creating an offload throttle by using the GUI

To create an offload throttle, select **Monitoring** → **System Hardware** → **Actions**, and then select **Edit System Offload Throttle**, as shown in Figure 10-26.

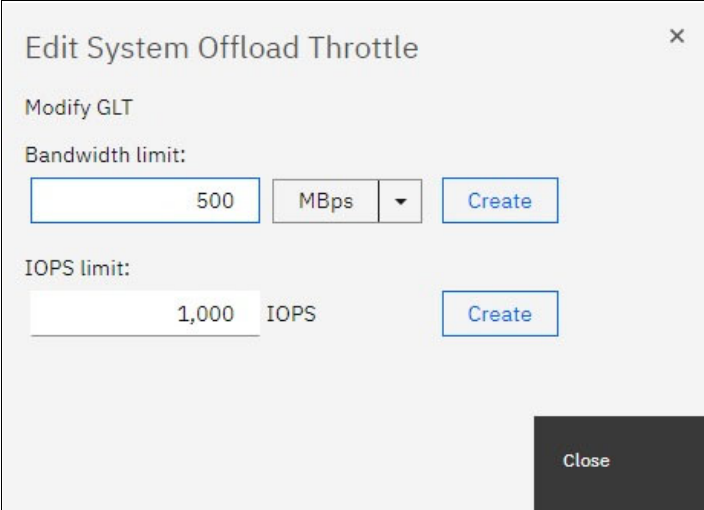


Figure 10-26 Creating a system offload throttle in the GUI

10.14 Documenting an IBM Storage Virtualize and SAN environment

This section focuses on the challenge of automating the documentation that is needed for an IBM Storage Virtualize solution. Consider the following points:

- ▶ Several methods and tools are available to automate the task of creating and updating the documentation. Therefore, the IT infrastructure might handle this task.
- ▶ Planning is key to maintaining sustained and organized growth. Accurate documentation of your storage environment is the blueprint with which you plan your approach to short-term and long-term storage growth.
- ▶ Your storage documentation must be conveniently available and easy to consult when needed. For example, you might need to determine how to replace your core SAN directors with newer ones, or how to fix the disk path problems of a single server. The relevant documentation might consist of a few spreadsheets and a diagram.
- ▶ Include photographs in the documentation where appropriate.

Storing documentation: Avoid storing IBM Storage Virtualize and SAN environment documentation only in the SAN. If your organization has a disaster recovery (DR) plan, include this storage documentation in it. Follow its guidelines about how to update and store this data. If no DR plan exists and you have the proper security authorization, it might be helpful to store an updated copy offsite.

In theory, this IBM Storage Virtualize and SAN environment documentation should be written at a level sufficient for any system administrator who has average skills in the products to understand. Make a copy that includes all your configuration information.

Use the copy to create a functionally equivalent copy of the environment by using similar hardware without any configuration, off-the-shelf media, and configuration backup files. You might need the copy if you ever face a DR scenario, which is also why it is so important to run periodic DR tests.

Create the first version of this documentation (“as-built documentation”) as you install your solution. If you completed forms to help plan the installation of your IBM Storage Virtualize solution, use these forms to help you document how your IBM Storage Virtualize solution was first configured. Minimum documentation is needed for an IBM Storage Virtualize solution. Because you might have more business requirements that require other data to be tracked, the following sections do not address every situation.

10.14.1 Naming conventions

Whether you are creating your IBM Storage Virtualize and SAN environment documentation or you are updating what is already in place, first evaluate whether you have a good naming convention in place. With a good naming convention, you can quickly and uniquely identify the components of your IBM Storage Virtualize and SAN environment. Then, system administrators can determine whether a name belongs to a volume, storage pool, MDisk, host, or HBA by looking at it.

Because error messages often point to the device that generated an error, a good naming convention quickly highlights where to start investigating when an error occurs. Typical IBM Storage Virtualize and SAN component names limit the number and type of characters that you can use. For example, IBM Storage Virtualize names are limited to 63 characters, which make creating a naming convention easier

Many names in an IBM Storage Virtualize and SAN environment can be modified online. Therefore, you do not need to worry about planning outages to implement your new naming convention. The naming examples that are used in the following sections are effective in most cases, but might not be fully adequate for your environment or needs. The naming convention to use is your choice, but you must implement it in the whole environment.

Enclosures, node canisters, and external storage controllers

IBM Storage Virtualize names its internal canisters or nodes as `nodeX`, with `X` being a sequential decimal number. The range can be 2 - 8, dependent on the number of installed I/O groups. In a FS9500 only 4 nodes (2 I/O groups) are supported.

If multiple external controllers are attached to your IBM Storage Virtualize solution, these controllers are detected as `controllerX`, so you might need to change the name so that it includes, for example, the vendor name, the model, or its serial number. Therefore, if you receive an error message that points to `controllerX`, you do not need to log in to IBM Storage Virtualize to know which storage controller to check.

Note: IBM Storage Virtualize detects external controllers based on their worldwide node name (WWNN). If you have an external storage controller that has one WWNN for each WWPN, this configuration might lead to many `controllerX` names pointing to the same physical box. In this case, prepare a naming convention to cover this situation.

MDisks and storage pools

When IBM Storage Virtualize detects new MDisks, it names them by default as `mdiskXX`, where `XX` is a sequential number. You should change the `XX` value to something more meaningful. MDisks are either arrays (distributed RAID (DRAID)) from internal storage or volumes from an external storage system.

Ultimately, it comes down to personal preference and what works in your environment. The main “convention” you should follow is to avoid the usage of special characters in names, apart from the underscore, the hyphen, and the period, which are permitted, and spaces (which can make scripting difficult).

For example, you can change a name to include the following information:

- ▶ For internal MDisks, refer to the IBM Storage Virtualize system or cluster name.
- ▶ A reference to the external storage controller it belongs to (such as its serial number or last digits).
- ▶ The extpool, array, or RAID group that it belongs to in the storage controller.
- ▶ The LUN number or name that it has in the storage controller.

Consider the following examples of MDisk names with this convention:

- ▶ IBM FlashSystem 9500CL01-MD03, where IBM FlashSystem 9500CL01 is the system or cluster name, and MD03 is the MDisk name.
- ▶ 23K45_A7V10, where 23K45 is the serial number, 7 is the array, and 10 is the volume.
- ▶ 75VXYZ1_02_0206, where 75VXYZ1 is the serial number, 02 is the extpool, and 0206 is the LUN.

Storage pools have several different possibilities. One possibility is to include the storage controller, the type of back-end disks if they are external, the RAID type, and sequential digits. If you have dedicated pools for specific applications or servers, another possibility is to use them instead.

Consider the following examples:

- ▶ IBM FlashSystem 9500-P00L01: IBM FlashSystem 9500 is the system or cluster name, and P00L01 is the pool.
- ▶ P05XYZ1_3GDR6: Pool 05 from serial 75VXYZ1, LUNs with 300 GB DDMs, and DRAID 6.
- ▶ P16XYZ1_EX01: Pool 16 from serial 75VXYZ1, and pool 01 is dedicated to Exchange Mail servers.
- ▶ XIV01_F7302_ET: A pool with disks from XIV that are named XIV01 and IBM Storage Virtualize 7300 named F7302, both of which are managed by Easy Tier.

Volumes

Volume names should include the following information:

- ▶ The host or cluster to which the volume is mapped.
- ▶ A single letter that indicates its usage by the host, as shown in the following examples:
 - B: For a boot disk, or R for a rootvg disk (if the server boots from SAN)
 - D: For a regular data disk
 - Q: For a cluster quorum disk (do not confuse with IBM Storage Virtualize quorum disks)
 - L: For a database log disk
 - T: For a database table disk
- ▶ A few sequential digits, for uniqueness.
- ▶ Sessions standard for VMware datastores:
 - esx01-sessions-001: For a data stores composed of a single volume

- esx01-sessions-001a and esx01-sessions-001b: For a data stores composed of two volumes

For example, ERPNY01-T03 indicates a volume that is mapped to server ERPNY01 and database table disk 03.

Hosts

In today's environment, administrators deal with large networks, the internet, and cloud computing. Use good server naming conventions so that they can quickly identify a server and determine the following information:

- ▶ Where it is (to know how to access it).
- ▶ What kind it is (to determine the vendor and support group in charge).
- ▶ What it does (to engage the proper application support and notify its owner).
- ▶ Its importance (to determine the severity if problems occur).

Changing a server's name in IBM Storage Virtualize is as simple as changing any other IBM Storage Virtualize object name. However, changing the name on the operating system of a server might have implications for application configuration and DNS, and such a change might require a server restart. Therefore, you might want to prepare a detailed plan if you decide to rename several servers in your network. The following example is for a server naming convention of LLAATRFNN, where:

- ▶ LL is the location, which might designate a city, data center, building floor, or room.
- ▶ AA is a major application, for example, billing, error recovery procedure (ERP), and Data Warehouse.
- ▶ T is the type, for example, UNIX, Windows, and VMware.
- ▶ R is the role, for example, Production, Test, Q&A, and Development.
- ▶ FF is the function, for example, DB server, application server, web server, and file server.
- ▶ NN is numeric.

SAN aliases and zones

SAN aliases often need to reflect only the device and port that is associated to them. Including information about where one particular device port is physically attached on the SAN might lead to inconsistencies if you make a change or perform maintenance and then forget to update the alias. Create one alias for each device port WWPN in your SAN and use these aliases in your zoning configuration. Consider the following examples:

- ▶ AIX_NYBIXTDB02_FC2: Interface fcs2 of AIX server NYBIXTDB02
- ▶ LIN-POKBIXAP01-FC1: Interface fcs1 of Linux server POKBIXAP01
- ▶ WIN_EXCHSRV01_HBA1: Interface HBA1 of physical Windows server EXCHSRV01
- ▶ ESX_NYVMCLUSTER01_VMHBA2: Interface vmhba2 of VMware ESX server NYVMCLUSTER01
- ▶ IBM-NYIBM FlashSystem 9500-N1P1_HOST: Port 1 of Node 1 from IBM FlashSystem 9500 Cluster NYIBM FlashSystem 9500 dedicated for hosts
- ▶ IBM-NYIBM FlashSystem 9500-N1P5_INTRA: Port 5 of Node 1 from IBM FlashSystem 9500 Cluster NYIBM FlashSystem 9200 dedicated to intracluster traffic
- ▶ IBM-NYIBM FlashSystem 9500-N1P7_REPL: Port 7 of Node 1 from IBM FlashSystem 9500 Cluster NYIBM FlashSystem 9500 dedicated to replication

Be mindful of the IBM Storage Virtualize 9500 port aliases. There are mappings between the last digits of the port WWPN and the node FC port.

- ▶ IBM_D88870_75XY131_I0301: DS8870 serial number75XY131, port I0301

- ▶ TS4500-TD06: TS4500 Tape Library, tape drive 06
- ▶ EMC_VNX7500_01_SPA2: EMC VNX7500 hostname VNX7500_01, SP A, port 2

If your SAN does not support aliases (for example, in heterogeneous fabrics with switches in some interoperation modes), use WWPNs in your zones. However, update every zone that uses a WWPN if you change it.

Your SAN zone name should reflect the devices in the SAN that it includes (normally in a one-to-one relationship), as shown in the following examples:

- ▶ SERVERALIAS_T1_IBM FlashSystem 9500CLUSTERNAME (from a server to the IBM Storage Virtualize 9500, where you use T1 as an ID to zones that uses, for example, node ports P1 on Fabric A, and P2 on Fabric B)
- ▶ SERVERALIAS_T2_IBM FlashSystem 9500CLUSTERNAME (from a server to the IBM Storage Virtualize 9500, where you use T2 as an ID to zones that uses, for example, node ports P3 on Fabric A, and P4 on Fabric B)
- ▶ IBM_DS8870_75XY131_IBM FlashSystem 9500CLUSTERNAME (zone between an external back-end storage and the IBM Storage Virtualize 9500)
- ▶ NYC_IBM FlashSystem 9500_POK_IBM FlashSystem 9500_REPLICATION (for remote copy services)

10.14.2 SAN fabric documentation

The most basic piece of SAN documentation is a SAN diagram. It is likely to be one of the first pieces of information that you need if you ever seek support from your SAN switches vendor. Also, a good spreadsheet with ports and zoning information eases the task of searching for detailed information, which if it is included in the diagram makes the diagram easier to use.

Brocade SAN Health

The *Brocade SAN Health Diagnostics Capture tool* is a no-cost, automated tool that can help you retain this documentation. SAN Health consists of a data collection tool that logs in to the SAN switches that you indicate and collects data by using standard SAN switch commands. Then, the tool creates a compressed file with the data collection. This file is sent to a Brocade automated machine for processing by secure web or email.

After some time (typically a few hours), you receive an email with instructions about how to download the report. The report includes a Visio diagram of your SAN and an organized Microsoft Excel spreadsheet that contains all your SAN information. For more information and to download the tool, see [Brocade SAN Health](#).

The first time that you use the SAN Health Diagnostics Capture tool, explore the options that are provided to learn how to create a well-organized and useful diagram.

Figure 10-27 on page 693 shows an example of a poorly formatted diagram.

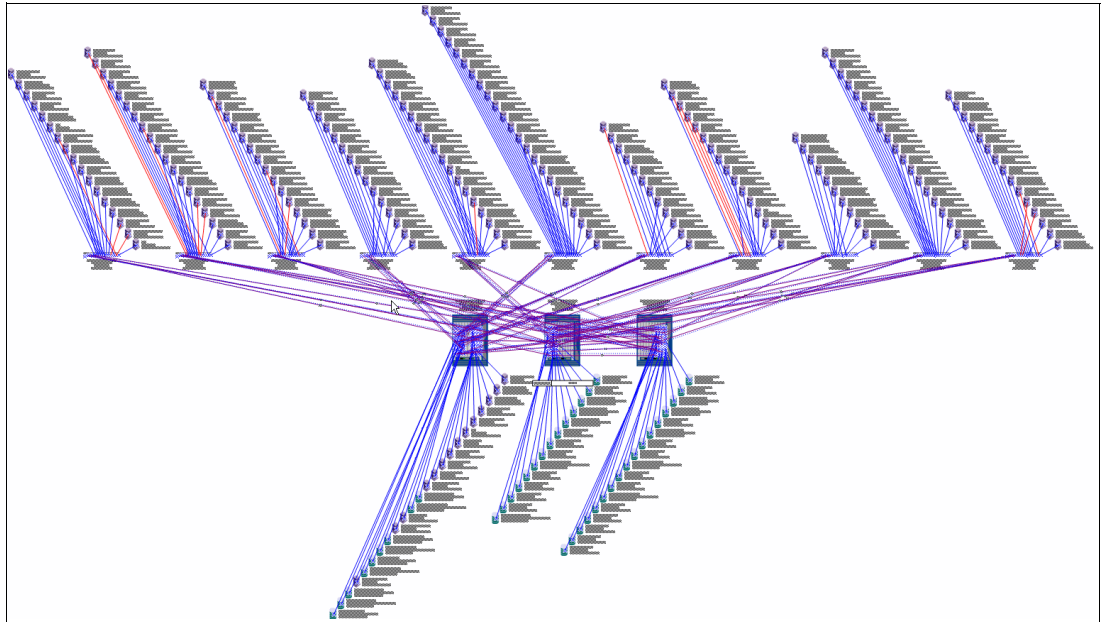


Figure 10-27 Poorly formatted SAN diagram

Figure 10-28 shows a tab of the SAN Health Options window in which you can choose the format of SAN diagram that best suits your needs. Depending on the topology and size of your SAN fabrics, you might want to manipulate the options in the **Diagram Format** or **Report Format** tabs.

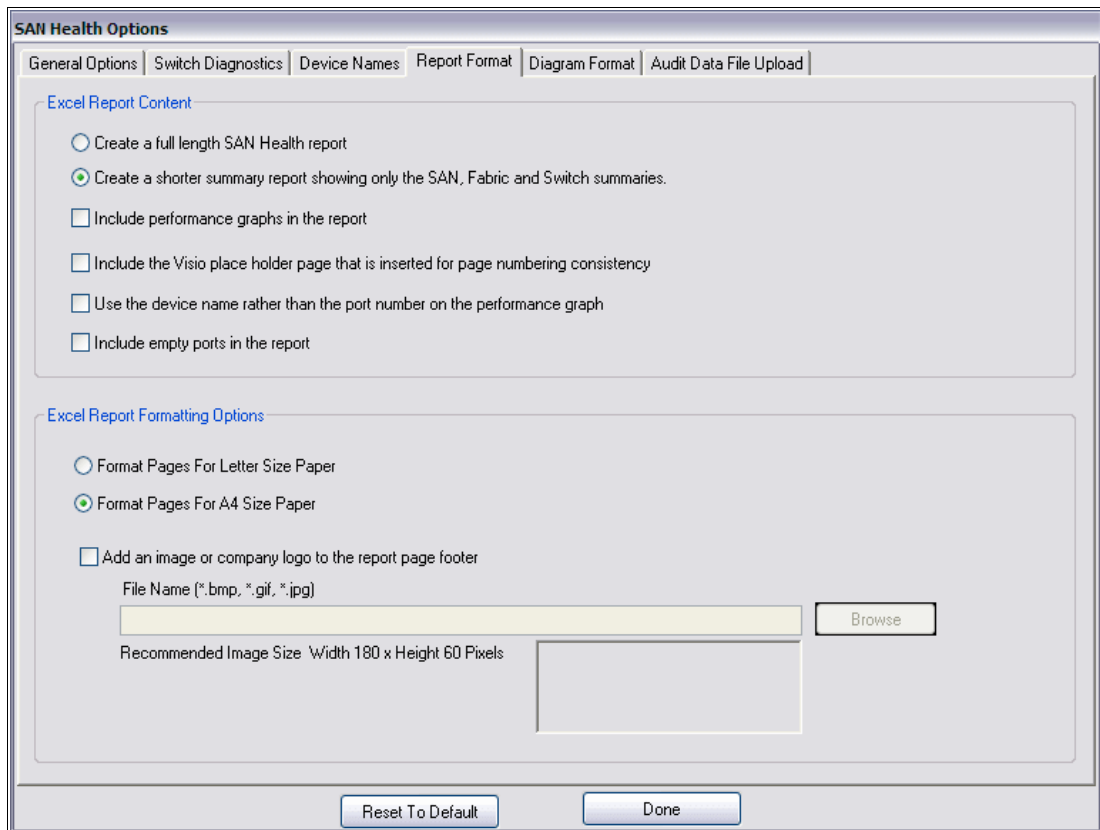


Figure 10-28 Brocade SAN Health Options window

SAN Health supports switches from manufacturers other than Brocade, such as Cisco. Both the data collection tool download and the processing of files are available at no cost. You can download Microsoft Visio and Excel viewers at no cost from the Microsoft website.

Another tool, which is known as *SAN Health Professional*, is also available for download at no cost. With this tool, you can audit the reports in detail by using advanced search functions and inventory tracking. You can configure the SAN Health Diagnostics Capture tool as a Windows scheduled task. To download of the SAN Health Diagnostics Capture tool, see this [Broadcom web page](#).

Tip: Regardless of the method that is used, generate a fresh report at least once a month or after any major changes. Keep previous versions so that you can track the evolution of your SAN.

IBM Spectrum Control reporting

If you have IBM Spectrum Control running in your environment, you can use it to generate reports on your SAN. For more information about how to configure and schedule IBM Spectrum Control reports, see [IBM Spectrum Control 5.4.10](#).

For more information about how to configure and set up IBM Spectrum Control, see Chapter 9, “Implementing a storage monitoring system” on page 551.

Ensure that the reports that you generate include all the information that you need. Schedule the reports with a period that you can use to backtrack any changes that you make.

10.14.3 IBM Storage Virtualize documentation

You can back up the configuration data for an IBM Storage Virtualize system after preliminary tasks complete. Configuration data for the system provides information about your system and the objects that are defined in it. This data contains the configuration data of arrays, pools, volumes, and so on. The backup does not contain any data from the volumes themselves.

Before you back up your configuration data, the following prerequisites must be met:

- ▶ Independent operations that change the configuration for the system cannot be running while the **backup** command is running.
- ▶ Object names cannot begin with an underscore character (_).

Note: The system automatically creates a backup of the configuration data each day at 1 AM. This backup is known as a *cron backup*, and on the configuration node it is copied to `/dumps/svc.config.cron.xml_<serial#>`.

To generate a manual backup at any time, complete the following steps:

1. Issue the **svcconfig backup** command to back up your configuration. The command displays messages similar to the ones in Example 10-16.

Example 10-16 Sample svcconfig backup command output

```
IBM_FlashSystem:FS9500:superuser>svcconfig backup
.....
CMMVC6155I SVCCONFIG processing completed successfully
```

The **svconfig backup** command creates three files that provide information about the backup process and the configuration. These files are created in the `/tmp` directory and copied to the `/dumps` directory of the configuration node. You can use the **lsdumps** command to list them. Table 10-7 describes the three files that are created by the backup process.

Table 10-7 Files that are created by the backup process

File name	Description
<code>svc.config.backup.xml_<serial#></code>	Contains your configuration data.
<code>svc.config.backup.sh_<serial#></code>	Contains the names of the commands that were issued to create the backup of the system.
<code>svc.config.backup.log_<serial#></code>	Contains details about the backup, including any reported errors or warnings.

2. Check that the **svconfig backup** command completes successfully, and examine the command output for any warnings or errors. The following output is an example of the message that is displayed when the backup process is successful:


```
CMMVC6155I SVCCONFIG processing completed successfully
```
3. If the process fails, resolve the errors and run the command again.
4. Keep backup copies of the files outside the system to protect them against a system hardware failure. With Microsoft Windows, use the PuTTY **pscp** utility. With UNIX or Linux, you can use the standard **scp** utility. By using the **-unsafe** option, you can use a wildcard to download all the `svc.config.backup` files with a single command. Example 10-17 shows the output of the **pscp** command.

Example 10-17 Saving the configuration backup files to your workstation

```
C:\>
pscp -unsafe superuser@9.10.11.12:/dumps/svc.config.backup.* C:\
Using keyboard-interactive authentication.
Password:
svc.config.backup.log_78E | 33 KB | 33.6 KB/s | ETA: 00:00:00 | 100%
svc.config.backup.sh_78E0 | 13 KB | 13.9 KB/s | ETA: 00:00:00 | 100%
svc.config.backup.xml_78E | 312 KB | 62.5 KB/s | ETA: 00:00:00 | 100%
C:\>
```

The configuration backup file is in Extensible Markup Language (XML) format and can be inserted as an object into your IBM Storage Virtualize documentation spreadsheet. The configuration backup file might be large. For example, it contains information about each internal storage drive that is installed in the system.

Note: Directly importing the file into your IBM Storage Virtualize documentation spreadsheet might make the file unreadable.

Also, consider collecting the output of specific commands. At a minimum, you should collect the output of the following commands:

- ▶ **svcinfo lsfabric**
- ▶ **svcinfo lssystem**
- ▶ **svcinfo lsdisk**
- ▶ **svcinfo lsdiskgrp**
- ▶ **svcinfo lsvdisk**

- ▶ **svcinfo lshost**
- ▶ **svcinfo lshostvdiskmap**

Note: Most CLI commands that are shown here work without the **svcinfo** prefix; however, some commands might not work with only the short name, and therefore require the **svcinfo** prefix to be added.

Import the commands into the master spreadsheet, preferably with the output from each command on a separate sheet.

One way to automate either task is to first create a batch file (Windows), shell script (UNIX or Linux), or playbook (Ansible) that collects and stores this information. Then, use spreadsheet macros to import the collected data into your IBM Storage Virtualize documentation spreadsheet.

When you are gathering IBM Storage Virtualize information, consider the following best practices:

- ▶ If you are collecting the output of specific commands, use the **-delim** option of these commands to make their output delimited by a character other than tab, such as comma, colon, or exclamation mark. You can import the temporary files into your spreadsheet in comma-separated value (CSV) format, specifying the same delimiter.

Note: Use a delimiter that is not already part of the output of the command. Commas can be used if the output is a particular type of list. Colons might be used for special fields, such as IPv6 addresses, WWPNs, or iSCSI names.

- ▶ If you are collecting the output of specific commands, save the output to temporary files. To make your spreadsheet macros simpler, you might want to preprocess the temporary files and remove any “garbage” or unwanted lines or columns. With UNIX or Linux, you can use commands such as **grep**, **sed**, and **awk**. Freeware software is available for Windows with the same commands, or you can use any batch text editor tool.

The objective is to fully automate this procedure so you can schedule it to regularly run automatically. Make the resulting spreadsheet easy to consult and have it contain only the information that you use frequently. The automated collection and storage of configuration and support data (which is typically more extensive and difficult to use) are described in 10.14.7, “Automated support data collection” on page 698.

10.14.4 Storage documentation

You must generate documentation of your back-end storage controllers after configuration. Then, you can update the documentation when these controllers receive hardware or code updates. As such, there is little point to automating this back-end storage controller documentation. The same applies to the IBM Storage Virtualize internal drives and enclosures.

Any portion of your external storage controllers that is used outside the IBM Storage Virtualize solution might have its configuration changed frequently. In this case, for more information about how to gather and store the information that you need, see your back-end storage controller documentation.

Fully allocate all the available space in any of the optional external storage controllers that you might use as more back-ends to the IBM Storage Virtualize solution. This way, you can perform all your disk storage management tasks by using the IBM Storage Virtualize user interface.

10.14.5 Technical support information

If you must open a technical support incident for your storage and SAN components, create and keep available a spreadsheet with all relevant information for all storage administrators. This spreadsheet should include the following information:

- ▶ Hardware information:
 - Vendor, machine and model number, and serial number (for example, IBM 9848-AF8 S/N 7812345).
 - Configuration, if applicable.
 - Current code level.
- ▶ Physical location:
 - Data center, including the complete street address and phone number.
 - Equipment physical location (room number, floor, tile location, and rack number).
 - Vendor’s security access information or procedure, if applicable.
 - Onsite person’s contact name and phone or page number.
- ▶ Support contract information:
 - Vendor contact phone numbers and website.
 - Customer’s contact name and phone or page number.
 - User ID to the support website, if applicable.
 - Do not store the password in the spreadsheet under any circumstances.
 - Support contract number and expiration date.

By keeping this data on a spreadsheet, storage administrators have all the information that they need to complete a web support request form or to provide to a vendor’s call support representative. Typically, you are asked first for a brief description of the problem and then asked later for a detailed description and support data collection.

10.14.6 Tracking incident and change tickets

If your organization uses an incident and change management and tracking tool (such as IBM Tivoli® Service Request Manager®), you or the storage administration team might need to develop proficiency in its use for several reasons:

- ▶ If your storage and SAN equipment are not configured to send Simple Network Management Protocol (SNMP) traps to this incident management tool, you should manually open incidents whenever an error is detected.
- ▶ IBM Storage Virtualize can be managed by the IBM Storage Insights tool that is available free of charge to owners of IBM storage systems. With the IBM Storage Insights tool, you can monitor all the IBM storage device information on IBM Storage Insights. For more information, see Chapter 9, “Implementing a storage monitoring system” on page 551.
- ▶ Call Home Connect Cloud (CHCC) provides an enhanced live view of your assets, including the status of cases, warranties, maintenance contracts, service levels, and end of service information.

Additionally, Call Home Connect Cloud offers links to other online tools and security documents. For more details, see [Introducing Call Home Connect Cloud](#).

- ▶ Disk storage allocation and deallocation and SAN zoning configuration modifications should be handled under properly submitted and approved change requests.
- ▶ If you are handling a problem yourself or calling your vendor's technical support desk, you might need to produce a list of the changes that you recently implemented in your SAN or that occurred since the documentation reports were last produced or updated.

When you use incident and change management tracking tools, adhere to the following guidelines for IBM Storage Virtualize and SAN Storage Administration:

- ▶ Whenever possible, configure your storage and SAN equipment to send SNMP traps to the incident monitoring tool so that an incident ticket is automatically opened and the proper alert notifications are sent. If you do not use a monitoring tool in your environment, you might want to configure email alerts that are automatically sent to the mobile phones or pagers of the storage administrators on duty or on call.
- ▶ Discuss within your organization the risk classification that a storage allocation or deallocation change request should have. These activities are typically safe and nondisruptive to other services and applications when properly handled.

However, activities might cause collateral damage if human error or an unexpected failure occurs during implementation. Your organization might decide to assume more costs with overtime and limit such activities to off-business hours, weekends, or maintenance windows if they assess that the risks to other critical applications are too high.

- ▶ Use templates for your most common change requests, such as storage allocation or SAN zoning modification to facilitate and speed up their submission.
- ▶ Do not open change requests in advance to replace failed, redundant, or hot-pluggable parts, such as disk drive modules (DDMs) in storage controllers with hot spares, or SFPs in SAN switches or servers with path redundancy.

Typically, these fixes do not change anything in your SAN storage topology or configuration, and they do not cause any more service disruption or degradation than you already had when the part failed. Handle these fixes within the associated incident ticket because it might take longer to replace the part if you must submit, schedule, and approve a non-emergency change request.

An exception is if you must interrupt more servers or applications to replace the part. In this case, you must schedule the activity and coordinate support groups. Use good judgment and avoid unnecessary exposure and delays.

- ▶ Keep handy the procedures to generate reports of the latest incidents and implemented changes in your SAN storage environment. Typically, you do not need to periodically generate these reports because your organization probably already has a Problem and Change Management group that runs such reports for trend analysis purposes.

10.14.7 Automated support data collection

In addition to the easier-to-use documentation of your IBM Storage Virtualize and SAN Storage environment, collect, and store for some time the configuration files and technical support data collection for all your SAN equipment.

For IBM Storage Virtualize, this information includes **snap** data. For other equipment, for more information about how to gather and store the support data that you might need, see the related documentation.

You can create procedures that automatically create and store this data on scheduled dates, delete old data, or transfer the data to tape.

You can use IBM Storage Insights to create support tickets and then attach the **snap** data to this record from within the IBM Storage Insights GUI. For more information, see Chapter 11, “Storage Virtualize Troubleshooting and diagnostics” on page 701.

10.14.8 Subscribing to IBM Storage Virtualize support

Subscribing to IBM Storage Virtualize support is probably the most overlooked practice in IT administration, and yet it is the most efficient way to stay ahead of problems. With this subscription, you can receive notifications about potential threats before they can reach you and cause severe service outages.

To subscribe to this support and receive support alerts and notifications for your products, subscribe for notifications at [Sign up for Notifications](#).

If you do not have an IBMid, create one.

You can subscribe to receive information from each vendor of storage and SAN equipment from the IBM website. You can often quickly determine whether an alert or notification is applicable to your SAN storage. Therefore, open them when you receive them and keep them in a folder of your mailbox.

Sign up and tailor the requests and alerts that you want to receive. For example, type IBM FlashSystem in the Product lookup text box and then click **Subscribe** to subscribe to IBM FlashSystem 9x00 notifications, as shown in Figure 10-29.

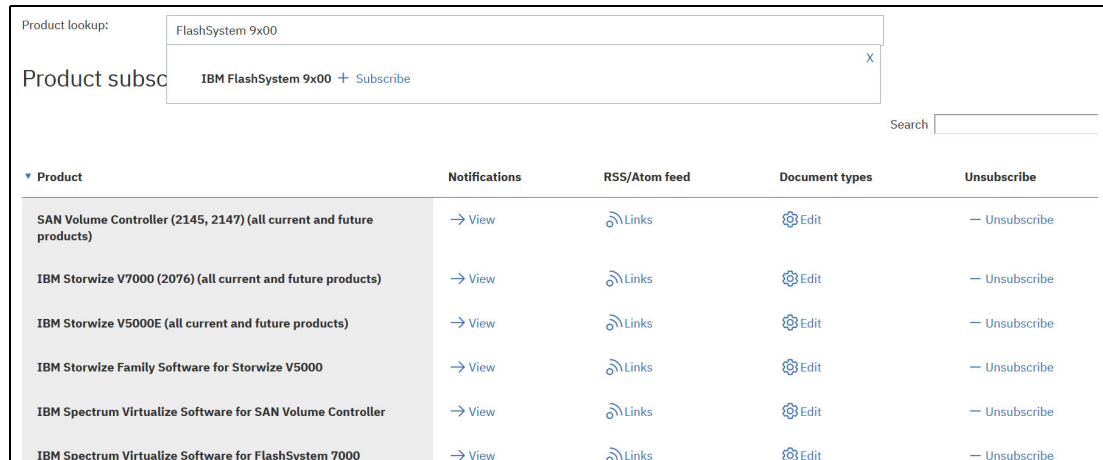



Figure 10-29 Creating a subscription to IBM Storage Virtualize notification



Storage Virtualize Troubleshooting and diagnostics

This chapter provides troubleshooting information for common problems that can occur in an IBM Storage Virtualize 8.6.0 environment. It describes situations that are related to IBM SAN Volume Controller (SVC), IBM FlashSystems, the storage area network (SAN) environment, optional external storage subsystems, and hosts. It also explains how to collect necessary problem determination data.

This chapter includes the following topics:

- ▶ “Troubleshooting” on page 702
- ▶ “Collecting diagnostic data” on page 710
- ▶ “Common problems and isolation techniques” on page 722
- ▶ “Remote Support Assistance” on page 746
- ▶ “Call Home Connect Cloud and Health Checker feature” on page 747
- ▶ “IBM Storage Insights” on page 748

11.1 Troubleshooting

Troubleshooting should follow a systematic approach to solve a problem. The goal of troubleshooting or problem determination is to understand, why something does not work as expected and create a resolution to resolve this. An important step therefore is to make a proper problem description, which should be as accurate as possible. Then you need to collect the support data from all involved components of the environment for analysis. This might include a *snap* from the IBM Storage Virtualize system, logs from SAN or network switches and host OS logs.

An effective problem report ideally describes these items:

- ▶ the expected behavior
- ▶ the actual behavior
- ▶ if possible, how to reproduce the behavior
- ▶ a precise timeline

The following questions help define the problem for effective troubleshooting:

- ▶ What are the symptoms of the problem?
 - What is reporting the problem?
 - Which error codes and messages were observed?
 - What is the business impact of the problem?
 - Where does the problem occur?
 - Which exact component is affected, the whole system or for instance certain hosts, IBM Storage Virtualize nodes
 - Is the environment and configuration supported?
- ▶ When does the problem occur?
 - How often does the problem happen?
 - Does the problem happen only at a certain time of day or night?
 - What kind of activities was ongoing at the time the problem was reported?
 - Did the problem happen after a change in the environment, such as a code upgrade or installing software or hardware?
- ▶ Under which conditions does the problem occur?
 - Does the problem always occur when the same task is being performed?
 - Does a certain sequence of events need to occur for the problem to surface?
 - Do any other applications fail at the same time?
- ▶ Can the problem be reproduced?
 - Can the problem be recreated, for example by running a single command, a set of commands, or a particular application?
 - Are multiple users or applications encountering the same type of problem?
 - Can the problem be reproduced on any other system?

Note: Collecting log files as close as possible to the time of the incident, and providing an accurate problem description and timeline are essential for effective troubleshooting.

11.1.1 Using the GUI

The IBM Storage Virtualize GUI is a good entry point to start troubleshooting with. There are two essential icons at the top that are accessible from any GUI panel.

In Figure 11-1, the first icon shows IBM Storage Virtualize events, such as an error or a warning, and the second icon shows suggested, running or recently completed background tasks.

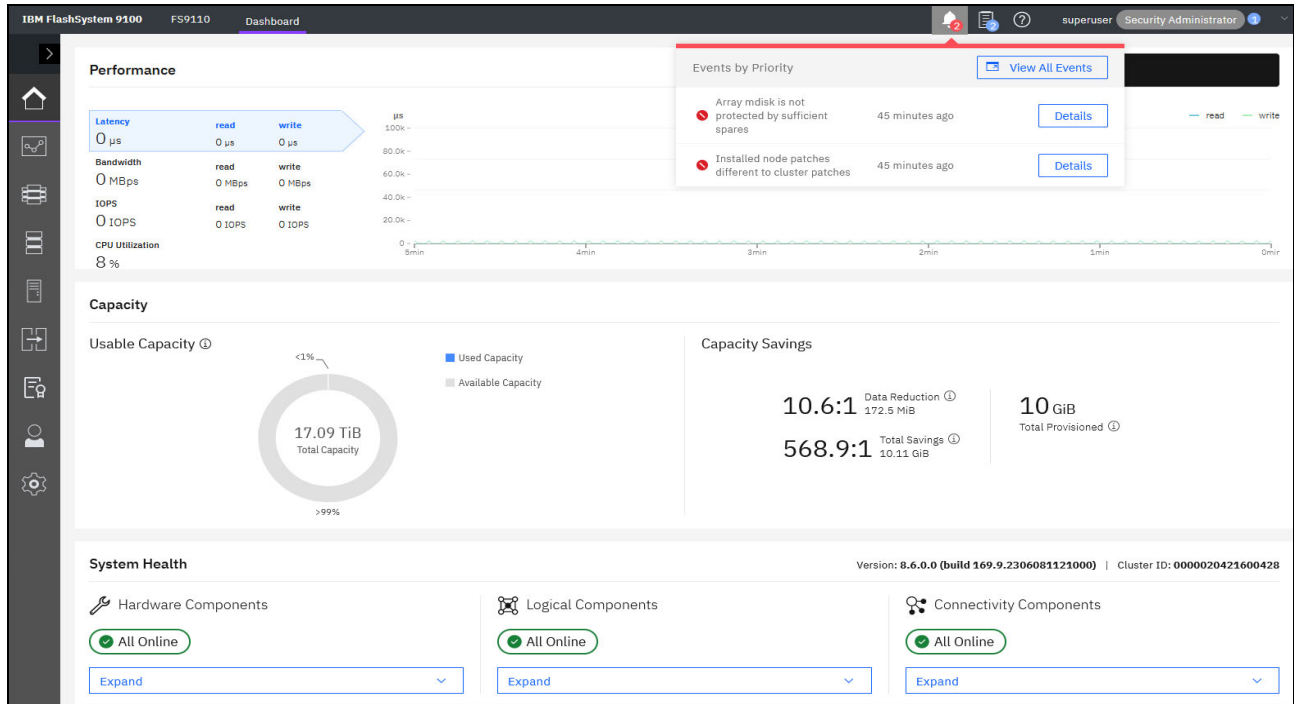


Figure 11-1 Events icon in the GUI

The GUI dashboard provides an at-a-glance view of the system's condition and notifies you of any circumstances that require immediate action. It contains sections for performance, capacity, and system health that provide an overall understanding of what is going on in the system.

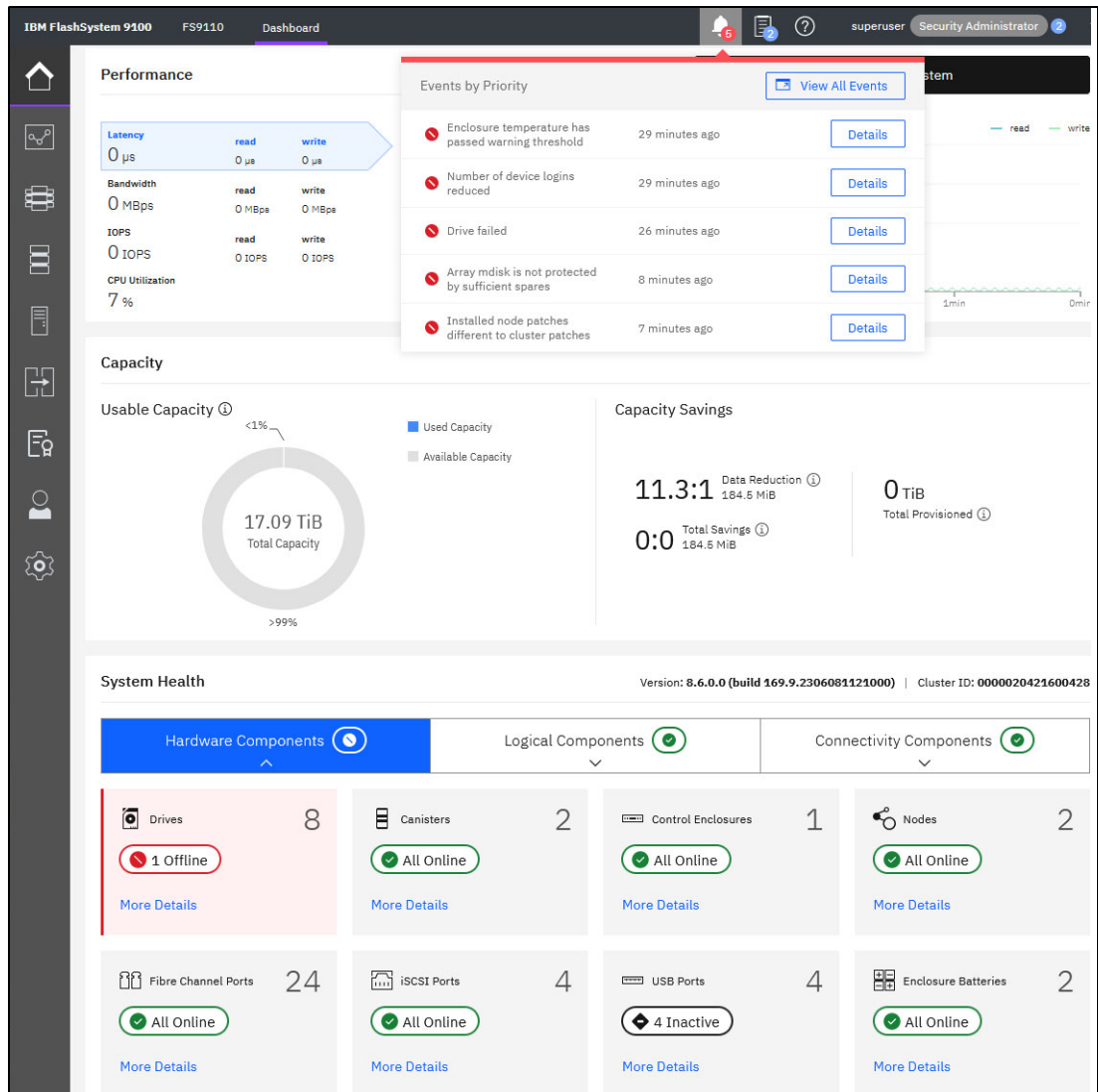


Figure 11-2 GUI Dashboard displaying system health events and hardware components

The System Health section in the bottom part of the dashboard provides information about the health status of hardware, logical, and connectivity components. If you click **Expand** in each of these categories, the status of the individual components is shown (see Figure 11-3 on page 705). Clicking on **More Details** takes you to the GUI panel related to that specific component, or shows more information about it.

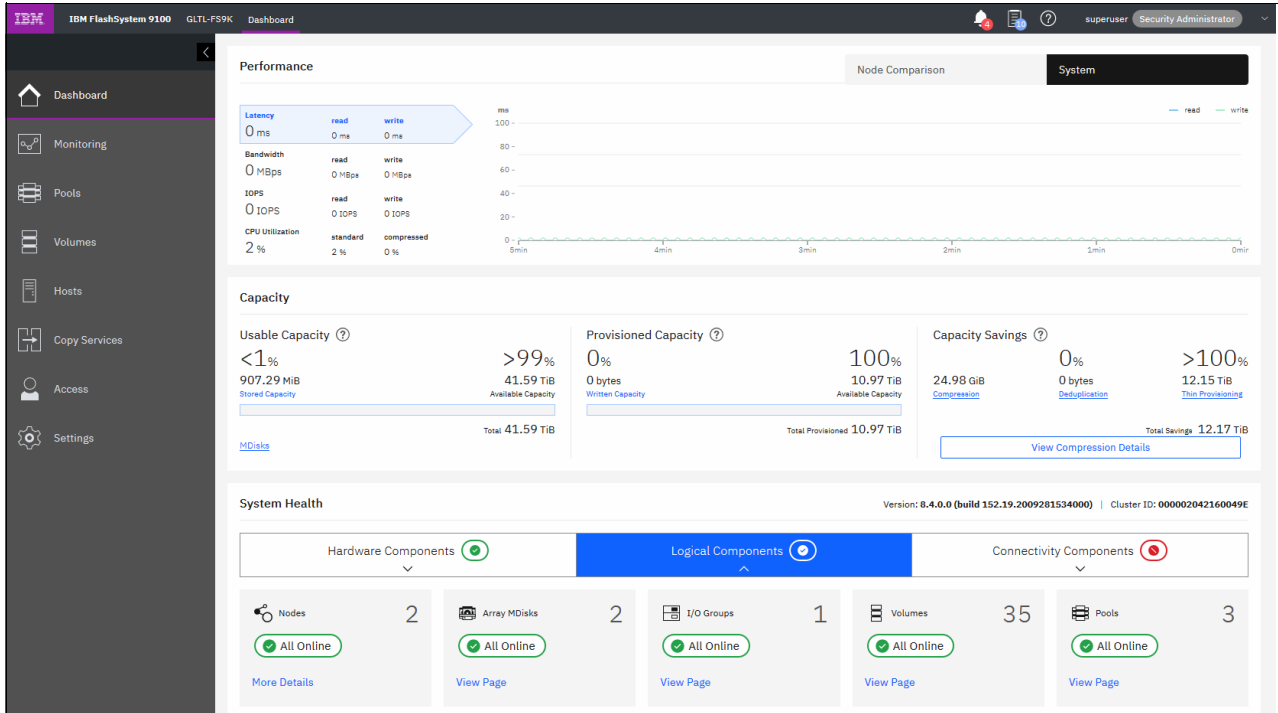


Figure 11-3 System Health expanded section in the dashboard

For more information about the components in each category and for troubleshooting, see [Troubleshooting](#).

11.1.2 Recommended actions and fix procedure

A *Fix procedure*, sometimes also referred to as *Directed Maintenance Procedure*, assists you in fixing a problem without doing any harm. Whenever one or multiple unfixed errors need to be addressed, the management GUI provides the means to run the recommended fix procedure. Therefore, the first step in troubleshooting is to check the event log for Recommended Actions in **Monitoring** → **Events**. The highest priority event, that is the event log entry with the lowest four-digit error code, will be highlighted to be addressed first as shown in Figure 11-4.

The screenshot shows the Events section of the dashboard. A red alert box highlights the highest priority event: "Error 1098 : Enclosure temperature has passed warning threshold". Below this, a table lists several events with their error codes, timestamps, statuses, descriptions, object types, and object IDs.

Error Code	Last Time Stamp	Status	Description	Object Type	Object ID
1098	20/6/2023 14:50:53	Alert	Enclosure temperature has passed warning threshold	enclosure	1
1630	20/6/2023 14:50:28	Alert	Number of device logins reduced	controller	0
1690	20/6/2023 12:03:08	Alert	Array mdisk is not protected by sufficient spares	mdisk	16
2015	20/6/2023 12:03:23	Alert	Installed node patches different to cluster patches	node	1

Figure 11-4 Recommended actions

Click on **Run Fix** to launch the Fix Procedure for this particular event. Fix Procedures help resolving a problem. In the background, a Fix Procedure analyzes the status of the system and its components and provides further information about the nature of the problem. This is to ensure, that the actions taken do not lead to undesirable results, as for instance volumes becoming inaccessible to the hosts. The Fix Procedure then automatically performs the actions required to return the system to its optimal state. This may include checking for dependencies, resetting internal error counters and apply updated to the system configuration. Whenever user interaction is required, you will be shown suggested actions to take and guided through the same. If the problem can be fixed, the related error in the event log eventually will be marked as fixed. Also, an associated alert in the GUI will be cleared.

Error codes along with their detailed properties in the event log provide reference information when a service action is required. The four-digit *Error Code* is visible in the event log. They are accompanied by a six-digit *Event ID* which provides additional details about this event.

Three-digit *Node Error Codes* are visible in the node status in the Service Assistant GUI. For more information about messages and codes, see [Messages and Codes](#).

An IBM Storage Virtualize system might encounter various kinds of failure recovery in certain conditions. These are known as Tier 1 (T1) through Tier 4 (T4) recovery.

- ▶ A T1 or Tier 1 Recovery - Node warmstart (node assert) will be logged with error code 2030 in the event log.

A single node assert is a recovery condition that is deployed by the IBM Storage Virtualize software, when a single node attempts to run an invalid code path or detects a transient hardware problem.

A T1 recovery alias single-node warmstart, is performed without suspending I/O. This task can be accomplished because the cluster is configured into redundant pairs of nodes or node canisters, and the clustering software ensures the deployment of a “replicated hardened state” across nodes. A single node can encounter an *assert condition*, perform a software restart recovery action (capturing first-time debug data), and return to the clustered system without the suspension of I/O.

On warm restart, the assert condition is cleared and the node rejoins the cluster automatically. Typically, a single node assert restart takes 1 - 5 minutes. Host data I/O continues as the host OS multipath software redirects the I/O to the partner node of the same I/O group.

The event with error id 2030 will be logged upon return of the cluster node to the system.

Right-click the **2030 event** and mark it as *Fixed* to prevent repeated alerts and notifications for the same event.

- ▶ A T2 or Tier 2 recovery is reported in the event log with error code 1001.

The cluster has asserted. In this case, all cluster nodes of the system have undergone a warmstart at the same time to recover from a condition which could not have been resolved otherwise. Error code 1001 means that system recovery was successful, and then the cluster resumes I/O, no data is lost. There was a temporary loss of access until the recovery completed, so host applications probably must be restarted. It is advisable to conduct a sanity check of the hosts' file systems afterwards. FlashCopy mappings and remote copy (Metro Mirror (MM) and Global Mirror (GM)) relationships are restored, along with the other essential cluster state information.

After a T2 recovery all configuration commands are blocked until you re-enable them, so that the unfixed event log entry with error code 1001 is being marked as fixed. It is recommended that they are not re-enabled until the recovery dumps and trace files from all nodes have been collected and were reviewed by IBM Support to confirm that it is safe to do so.

The Service GUI is the preferred method for collecting logs of each node. Open a browser session to the Service GUI at https://<cluster_ip>/service. Select the **Collect Logs** pane from the left navigation bar, and then select the option to create a support package with the latest statesave.

- ▶ A Tier 3 or T3 Recovery is required when there is no more active cluster node and all nodes of the clustered system report node error 550 and/or 578. The *Recover System Procedure* recovers the system if the system state is lost from all cluster nodes.

The T3 Recovery procedure re-creates the system configuration to the state from before the incident that led to this situation. Depending on the type of the IBM Storage Virtualize system and the configuration, this is achieved by ingesting the configuration and hardened system data. This data is stored on either a quorum MDisk, quorum drive, or an IP quorum set up to store metadata. In combination with the information stored in the configuration backup `svc.config.backup.xml`, the system's configuration and state will be restored.

Note: Attempt to run the *Recover System Procedure* only after a complete and thorough investigation of the cause of the system failure. Attempt to resolve those issues by using other service procedures.

Selecting **Monitoring** → **Events** shows information messages, warnings, and issues about the IBM Storage Virtualize system. Therefore, this area is a good place to check for problems in the system.

To display the most important events that must be fixed, use the **Recommended Actions** filter.

If an important issue must be fixed, look for the **Run Fix** button in the upper left with an error message that indicates which event must be fixed as soon as possible. This fix procedure helps resolve problems. It analyzes the system, provides more information about the problem, suggests actions to take with the steps to follow, and finally checks to see whether the problem is resolved.

Always use the fix procedures to resolve errors that are reported by the system, such as system configuration problems or hardware failures.

Note: IBM Storage Virtualize systems detect and report error messages; however, events may have been triggered by factors external to the system, for example back-end storage devices or the storage area network (SAN).

It is safe to mark events as fixed; if error condition still exists or errors reoccur, a new event will be logged. You can select multiple events in the table by pressing and holding the CTRL-key and clicking the events to be fixed with mouse.

Figure 11-5 on page 708 shows **Monitoring** → **Events** window with Recommended Run Fix.

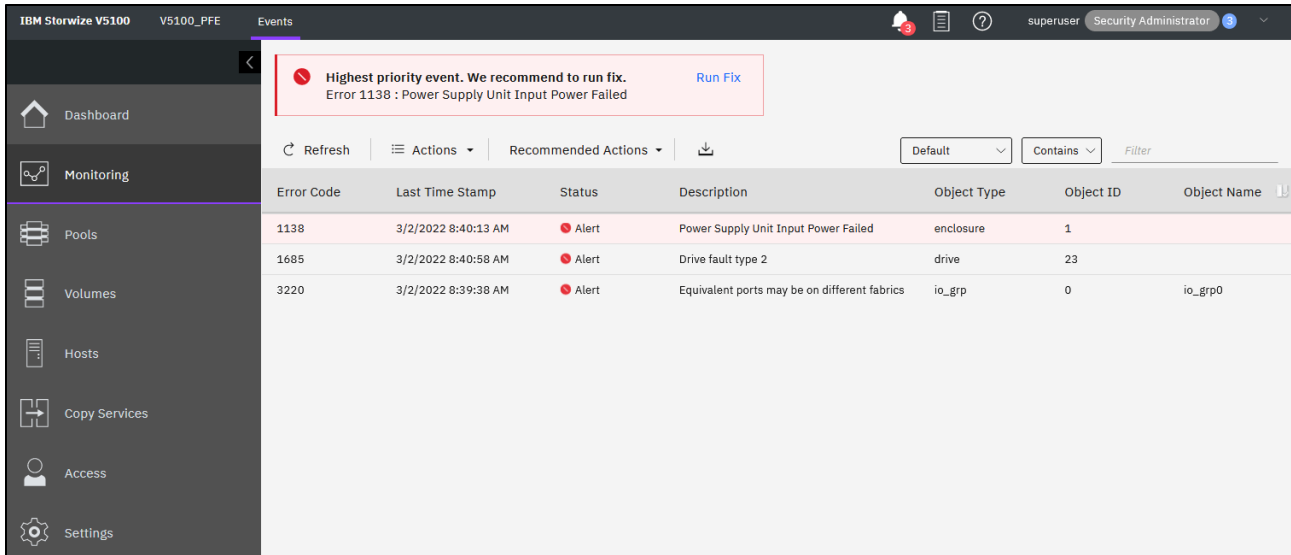


Figure 11-5 Monitoring → Events window

Resolve alerts in a timely manner: When an issue or a potential issue is reported, resolve it as quickly as possible to minimize its impact, and potentially avoid more serious problems with your system.

To obtain more information about any event, double-click or select an event in the table, and select **Actions** → **Properties**. You can also select **Run Fix Procedure** and properties by right-clicking an event.

The properties and details are displayed in a pane, as shown in Figure 11-6. Sense Data is available in an embedded tab. You can review and click **Run Fix** to run the fix procedure.

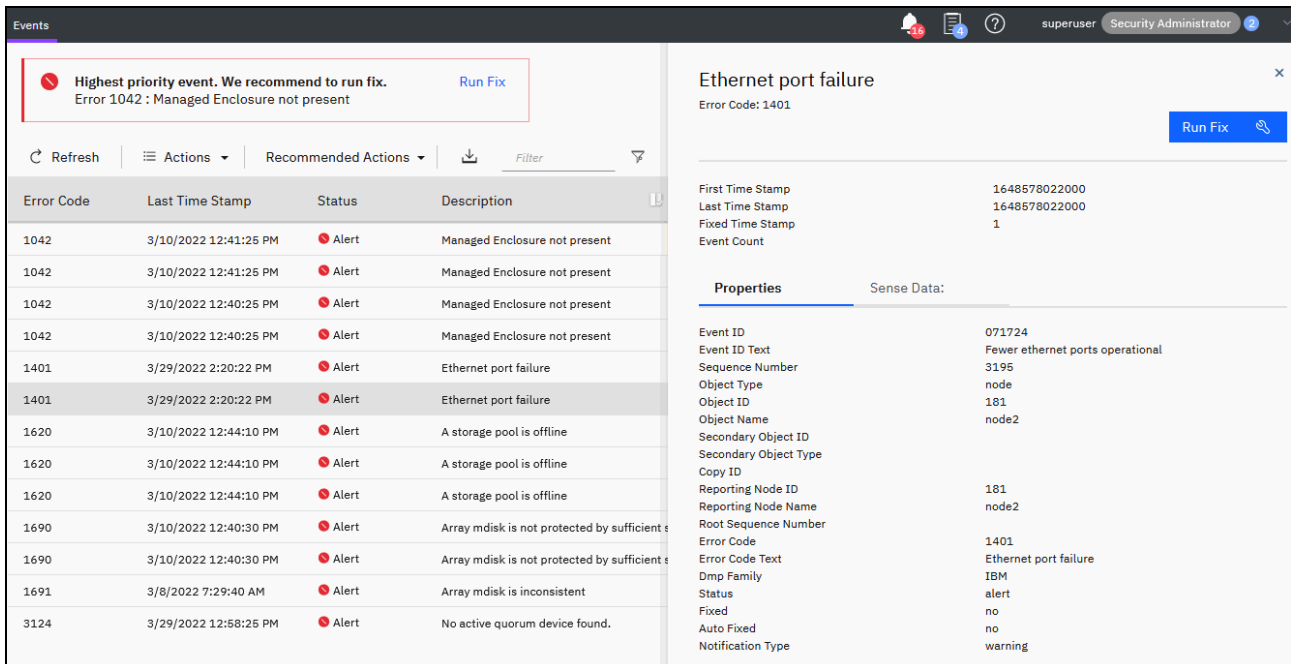


Figure 11-6 Properties and Sense Data for an event

11.1.3 Using the command-line interface

Another option to investigate and resolve issues is to use the IBM Storage Virtualize command-line interface (CLI). Although the *fix procedures* automatically perform the necessary steps, it may be sometimes faster and more convenient to run these commands directly through the CLI. This is particularly the case, when numerous events of the same kind need to be fixed, as this can be a strenuous task to click each individual event in the GUI.

Run the commands when you encounter the following issues:

- ▶ You experience a back-end storage or internode issue. For example:
 - Error code 1370: A managed disk (MDisk) error recovery procedure (ERP) has occurred.
 - Error code 1630: The number of device logins was reduced.
 - Error code: 1230 or 1231 Login Excluded.
- ▶ You performed maintenance on the following items:
 - Back-end storage subsystems.
 - SAN devices like switches, cables, optical transceivers as SFPs.

Important: Execute these commands when any type of change that is related to the communication between IBM Storage Virtualize systems and back-end storage subsystem occurs such as back-end storage is configured or a SAN zoning change occurred). This process ensures that IBM Storage Virtualize recognizes the changes.

Common error recovery involves the following IBM Storage Virtualize CLI commands:

- ▶ **detectmdisk**
Discovers changes in the SAN and back-end storage.
- ▶ **lscontroller** and **lsmdisk**
Provides the status of all controllers and MDisks. Pay attention to status values other than online, for instance *offline* or *degraded*.
- ▶ **lscontroller <controller_id_or_name>**
Checks the controller that was causing the issue and verifies that all the worldwide port names (WWPNs) are listed as you expect. Also check if the `path_counts` are distributed evenly across the WWPNs.
- ▶ **lsmdisk**
Determines whether all MDisks are online.

Note: When an issue is resolved by using the CLI, verify that the error disappears by selecting **Monitoring** → **Events**. If not, make sure to mark the error as fixed.

11.2 Collecting diagnostic data

Problem source identification and problem determination (PSI/PD) may be a challenging task in complex and heterogeneous IT environments. It is crucial hence to collect the right diagnostic data at the right time to enable the IBM Remote Support teams to assist you in resolving a problem. The following section outlines the steps to enable you to collect diagnostic data to find and isolate problems in an IBM Storage Virtualize environment.

11.2.1 IBM Storage Virtualize systems data collection

When you encounter a problem with an IBM Storage Virtualize system and you need to open a case with IBM support, you most likely will be asked to provide a *support package* from the system. The support package may interchangeably be referred to as *Snap* or *data collection*.

Checking for an automatically opened Call Home case

IBM Storage Virtualize system configured for Call Home automatically report events to IBM. A support case is automatically opened depending on the type of event. It is a good practice to check first, whether a case already exists. To do so, log in with your IBMid to [IBM Call Home Connect Cloud](#). Alternatively, you may check the My Cases section in the [IBM My Support](#) portal. Storage Virtualize systems configured to be monitored in [IBM Storage Insights](#), will show associated support cases as well.

What data to collect on IBM Storage Virtualize systems

The data that is needed for analysis depends on the kind of problem to be analyzed. An IBM Storage Virtualize system stores different kind of log files, message files, statistics, and traces. The following terms are commonly used in related publications:

- ▶ **Dump:** A node dump or full dump s collected whenever the software restarts for some reason. It is similar in nature to a core dump file, and it can be used by IBM Remote Technical Support and development teams to investigate a problem
- ▶ **Livedump:** A livedump is a binary data capture of the current state of the software. It causes only minimal impact to I/O operations. The contents of a livedump are similar to the contents of a dump with slightly less detailed information.
- ▶ **Statesave:** The term statesave is interchangeably used for either a dump or a livedump.

Four different types of *Snap* can be collected, *Snap Type 1* through *Snap Type 4*, colloquially often referred to as *Snap/1*, *Snap/2*, *Snap/3* or *Snap/4*. The Snap types vary in the amount of diagnostic information that is contained in the package:

- ▶ **Snap/1:** Standard logs including performance stats
 - fastest, smallest, no node dumps
- ▶ **Snap/2:** Same as Snap/1 plus one existing statesave, the most recently created dump or livedump from the current config node
 - slightly slower than snap/1, big
- ▶ **Snap/3:** Same as Snap/1 plus the most recent dump or livedump from each active member node in the clustered system
- ▶ **Snap option4:** Same as Snap/1 plus fresh livedump from each active member node in the clustered, which are created upon triggering the data collection

Statesaves, dumps and livedumps

Statesaves or dumps: A dump is created and written to a cluster node's boot driven, when the Storage Virtualize software stack restarts for some reason. It is similar in nature to a core

dump file. It can be used by IBM Remote Technical Support and development teams to understand, why the software restarted.

There are different kinds of a dump:

- ▶ Livedump: A livedump is a binary data capture of the current state of the software. It causes only minimal impact to I/O operations. The contents of a livedump are similar to the contents of a dump with slightly less detailed information.

Which Snap to collect

Two major factors play a part in deciding which snap to collect for a specific support case:

- ▶ Speed of collection
 - A Snap/1 is generated more rapidly, and it is much smaller.
- ▶ Amount of data
 - Collecting a Snap/4 as soon as possible after a problem has occurred increases the likelihood that the livedump contains the data required to diagnose the problem.

Consider the following points:

- ▶ For issues that are related to interoperability with hosts or storage, collect Snap Type 4.
- ▶ For critical performance issues, collect Snap/1 and then collect Snap/4.
- ▶ For general performance issues, collect Snap/4.
- ▶ For issues that are related to replication (including 1920 errors), collect Snap/4 from both systems in the replication partnership.
- ▶ For issues that are related to Data Reduction Pools, collect Snap/4.
- ▶ For 2030, 1196, or 1195 errors, collect Snap/3.
- ▶ For all other issues, collect Snap/4.

Tip: For urgent cases, start with collecting and uploading a Snap/1 followed by a Snap/4. This enables IBM Remote Support to quicker commence the analysis, while the more detailed Snap/4 is being collected and uploaded.

For more information about the required support package that is most suitable to diagnose different type of issues and their content, see [What data should you collect for a problem on IBM Storage Virtualize systems?](#)

Note: After an issue is solved, it is a best practice to do some housekeeping and delete old dumps on each node by running the following command:

```
cleardumps -prefix /dumps node_id | node_name
```

Support package collection and upload

The most commonly used method to collect and upload a support package is using the IBM Storage Virtualize GUI. However, using IBM Storage Insights is even easier, as it reduces the efforts to upload the collected package to IBM. This method is described in 11.6.4, “Updating a support ticket” on page 758.

By default, use IBM Storage Virtualize to automatically upload the support packages from IBM Storage Virtualize by using the GUI or CLI. The support packages are collected and uploaded to the IBM Support center automatically by using IBM Storage Virtualize or downloading the package from the device and manually uploading to IBM.

Collecting data by using the GUI

To collect data by using the GUI, complete the following steps:

1. Select **Settings** → **Support** → **Support Package**. You can choose to download the support package to your workstation or upload it to the IBM support center. The latter option requires internet connectivity for the IBM Storage Virtualize system, either directly or through a configured *Proxy Server*.
2. To automatically upload the support packages, click **Upload Support Package**.
3. Select **Create New Package and Upload**.
4. In the pop-up window, enter the IBM Support case number (TSxxxxxxx) and the type of support package to upload to the IBM Support Center.

The Upload Support Package window is shown in Figure 11-7.

Upload Support Package

Your system will generate and upload a new package to the IBM support center.

Case number: [Don't have a case number?](#)

TS012345678

Select the type of new support package to generate and upload to the IBM support center:

- Snap Type 1: Standard logs
Contains the most recent logs for the system, including the event and audit logs.
- Snap Type 2: Standard logs plus one existing statesave
Contains all the standard logs plus one existing statesave from any of the nodes in the system.
- Snap Type 3: Standard logs plus most recent statesave from each node
Contains all the standard logs plus each node's most recent statesave.
- Snap Type 4: Standard logs plus new statesaves
Contains all the standard logs and generate a new statesave on each node in the system.

[Need Help](#) **Cancel** **Upload**

Figure 11-7 Upload Support Package window

You can monitor the progress of the individual sub-tasks by clicking on View more details as shown in Figure 11-8 on page 713.

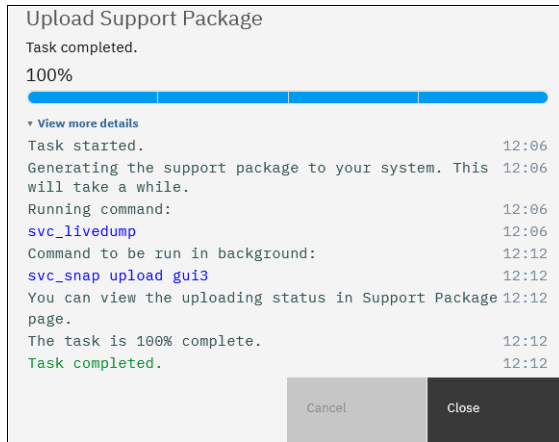


Figure 11-8 Upload Support Package details

Collecting data by using the CLI

Log in to the CLI and run the command that matches the type of snap that is requested:

- ▶ Standard logs (Snap type 1):
`svc_snap gui1 upload pmr=TSxxxxxxxx`
- ▶ Standard logs plus one existing statesave from current config node (Snap type 2):
`svc_snap gui2 upload pmr=TSxxxxxxxx`
- ▶ Standard logs plus most recent statesave from each active cluster node (Snap type 3):
`svc_snap gui3 upload pmr=TSxxxxxxxx`
- ▶ Standard logs plus new statesaves (Snap type 4):
`svc_livedump -nodes all -yes`
`svc_snap gui3 upload pmr=TSxxxxxxxx`

To collect a Snap type 4 using the CLI, a livedump of each active node must be generated by using the `svc_livedump` command. Then, the log files and newly generated dumps are uploaded by using the `svc_snap gui3` command, as shown in Example 11-1 on page 714. To verify whether the support package was successfully uploaded, use the `sainfo lscmdstatus` command (TSXXXXXX is the case number).

Note: The use of Service Assistant commands as `sainfo` or `satask` requires superuser privileges.

Example 11-1 The svc_livedump command

```
IBM_FlashSystem:FS9110:superuser>svc_livedump -nodes all -yes
Livedump - Fetching Node Configuration
Livedump - Checking for dependent vdisks
Livedump - Check Node status
Livedump - Preparing specified nodes - this may take some time...
Livedump - Prepare node 1
Livedump - Prepare node 2
Livedump - Trigger specified nodes
Livedump - Triggering livedump on node 1
Livedump - Triggering livedump on node 2
Livedump - Waiting for livedumps to complete dumping on nodes 1,2
Livedump - Waiting for livedumps to complete dumping on nodes 1
Livedump - Successfully captured livedumps on nodes 1,2

IBM_IBM FlashSystem:FLASHPF95:superuser>svc_snap gui3 upload pmr=TSxxxxxxxxx
Collecting data
Packaging files
Snap data collected in /dumps/snap.serial.YYMMDD.HHMMSS.tgz

IBM_FlashSystem:FS9110:superuser>sainfo lscmdstatus
last_command satask supportupload -pmr TSxxxxxxxxx -filename
/dumps/snap.serial.YYMMDD.HHMMSS.tgz
last_command_status CMMVC8044E Command completed successfully.
T3_status
T3_status_data
cpfiles_status Complete
cpfiles_status_data Copied 160 of 160
snap_status Complete
snap_filename /dumps/snap.serial.YYMMDD.HHMMSS.tgz
installcanistersoftware_status
supportupload_status Active
supportupload_status_data Uploaded 267.5 MiB of 550.2 MiB
supportupload_progress_percent 48
supportupload_throughput_KBps 639
supportupload_filename /dumps/snap.serial.YYMMDD.HHMMSS.tgz
```

If you do not want to automatically upload the snap to IBM, omit the **upload pmr=TSxxxxxxxxx** command option. When the snap creation completes, all collected files are packaged into a gzip-compressed tarball that uses the following format:

```
/dumps/snap.<panel_id>.YYMMDD.hhmmss.tgz
```

The creation of the Snap archive takes a few minutes to complete. Depending on the size of the system and the configuration, it can take considerably longer, particularly if fresh livedumps are being created.

The generated file can be retrieved from the GUI by selecting **Settings** → **Support** → **Manual Upload Instructions** → **Download Support Package**, and then clicking **Download Existing Package**. Find the exact name of the snap that was generated by running the **svc_snap** command that was run earlier. Select that file, and click **Download**.

In certain circumstances it may be necessary to collect a livedump from an individual node of a clustered system at a certain point in time. This can be achieved through CLI commands **preplivedump** and **triggerlivedump**, followed by the targeted node's numeric id or name. The

livedump status of a node can be checked with command `lslivedump`. Once the status has changed from *dumping* to *inactive*, the livedump file is ready to be copied off the system using either the GUI or `scp` command.

Example 11-2 preplivedump and lslivedump commands

```
IBM_FlashSystem:FS9110:superuser>preplivedump 2
IBM_FlashSystem:FS9110:superuser>lslivedump 2
status
prepared
IBM_FlashSystem:FS9110:superuser>triggerlivedump 2
IBM_FlashSystem:FS9110:superuser>lslivedump 2
status
dumping
IBM_FlashSystem:FS9110:superuser>lslivedump 2
status
inactive
IBM_FlashSystem:FS9110:superuser>lsdumps 2
[...]
livedump.panel_id.YYMMDD.HHMMSS
[...]
```

11.2.2 Host multipath software

If a problem occurs that is related to host communication with an IBM Storage Virtualize system, collecting data from hosts and their multipath software is useful.

Linux using device-mapper-multipath (dmmp)

To troubleshoot by using the multipathd CLI, issue the `multipath -ll` command, which shows detailed information about the multipath devices.

Example 11-3 shows the output for the command `multipath -ll`, including the following information:

- ▶ Name of the mpath device (mpatha / mpathb).
- ▶ UUID of the mpath device.
- ▶ Discovered paths for each mpath device, including the name of the sd-device, the priority, and state information.

Example 11-3 Output for the multipath -ll command

```
root@myServer ~]# multipath -ll
mpatha (3600507680185801aa00000000000b79) dm-3 IBM ,2145
size=100G features='1 queue_if_no_path' hwhandler='0' wp=rw
| -+- policy='service-time 0' prio=50 status=active
| | - 16:0:0:2 sdl 8:176 active ready running
| | ~- 18:0:0:2 sdm 8:192 active ready running
| ~-+- policy='service-time 0' prio=10 status=enabled
| | - 16:0:1:2 sdg 8:96 active ready running
| | ~- 18:0:1:2 sdt 65:48 active ready running
mpathb (3600507680185801aa00000000000b78) dm-4 IBM ,2145
size=100G features='1 queue_if_no_path' hwhandler='0' wp=rw
| -+- policy='service-time 0' prio=50 status=active
| | - 16:0:1:1 sde 8:64 active ready running
| | ~- 18:0:1:1 sds 65:32 active ready running
```

```

^-+- policy='service-time 0' prio=10 status=enabled
  |- 16:0:0:1 sdj 8:144 active ready running
  ^- 18:0:0:1 sdk 8:160 active ready running

```

Expand the command to **multipath -ll -v3** to print debug information.

You can also use the multipathd interactive console for troubleshooting. The **multipath -k** command opens an interactive interface to the multipathd daemon.

Entering this command opens an interactive multipath console. After running this command, it is possible to enter help to get a list of available commands, which can be used within the interactive console. To exit the console, press Ctrl-d.

To display the current configuration, including the defaults, issue **show config** within the interactive console.

AIX using multipath I/O

Table 11-1 shows some of the useful AIX **lspath** commands.

Table 11-1 Useful AIX *lspath* commands

Command	Result
lspath	Lists all paths for all hdisks with their status and parent fscsi device information.
lspath -H -l hdisk1	List all paths for the specified hdisk with its status and corresponding fscsi device information. The output includes a column header.
lspath -l hdisk1 -HF "name path_id parent connection path_status status"	Lists more detailed information about the specified hdisk the parent fscsi device and its path status.
lspath -s disabled	Lists all paths whose operational status is disabled.
lspath -s failed	Lists all paths whose operational status is failed.
lspath -AHE -l hdisk0 -p vscsi0 -w "810000000000"	Display attributes for a path and connection (-w) (-A is like lsattr for devices. If only one path exists to the parent device, the connection can be omitted by running: lspath -AHE -l hdisk0 -p vscsi0)
lspath -l hdisk1 -a priority -F value -p fscsi0 -w 500507680d7e1264,0	Lists the priority for a specific path.

Table 11-2 shows some of the useful AIX **lsmPIO** commands.

Table 11-2 Useful AIX *lsmPIO* commands

Command	Result
lsmPIO	Shows all disks and corresponding paths with state, parent, and connection information.
lsmPIO -q	Shows all disks with vendor ID, product ID, size, and volume name.

Command	Result
<code>lsmPIO -q1 hdisk0</code>	Shows detailed disk information like: <ul style="list-style-type: none"> ▶ Vendor ID ▶ Product ID ▶ Capacity ▶ Machine Type ▶ Model Number ▶ Host Group ▶ Volume Name ▶ Volume Serial Number
<code>lsmPO -S1 hdisk0 grep Path</code>	Shows path statistics.
<code>lsmPIO -ar</code>	Lists the parent adapter and remote port information (-a: adapter (local), and -r: remote port).
<code>lsmPIO -are</code>	Lists the parent adapter and remote port error statistics (-e: error).
<code>lsmPIO -z</code>	Lists all multipath I/O (MPIO) statistics.

Windows using MPIO

Because IBM Storage Virtualize 8.3.0 is the last version that supports *Subsystem Device Driver Device Specific Module* (SDDDSM), you must use native Windows multipathing, which is provided by the installable feature MPIO.

Besides managing the multipathing configuration by using the Windows GUI, it is possible to use the CLI by using the tool `mpclaim.exe`, which is installed by default.

Table 11-3 shows some of the useful Windows `mpclaim.exe` commands.

Table 11-3 Useful Windows `mpclaim.exe` commands

Command	Result
<code>mpclaim.exe -e</code>	View the storage devices that are discovered by the system.
<code>mpclaim.exe -r -i -d _IBM_2145</code>	Manages FC devices with MPIO.
<code>mpclaim.exe -r -u -d _IBM_2145</code>	Removes MPIO management of FC devices.
<code>mpclaim.exe -r -i -d"MSFT2005iSCSIBusType_0x9"</code>	Manages internet Small Computer Systems Interface (iSCSI) devices with MPIO.
<code>mpclaim.exe -r -u -d"MSFT2005iSCSIBusType_0x9"</code>	Removes MPIO management of iSCSI devices.
<code>mpclaim.exe -r -i -a""</code>	Manages all storage devices with MPIO.
<code>mpclaim.exe -r -u -a""</code>	Removes MPIO management for all devices.
<code>mpclaim.exe -r</code>	View storage devices that are managed by Microsoft DSM.
<code>mpclaim.exe -L -M<num></code>	Modifies the load-balancing policy.
<code>mpclaim.exe -s -d</code>	Checks the policy that your volumes are currently using.
<code>mpclaim.exe -s -d <number></code>	Checks the policy for a specific disk.

Generic MPIO settings can be listed and modified by using Windows PowerShell cmdlets. Table 11-4 shows the PowerShell cmdlets, which may be used to list or modify generic Windows MPIO settings.

Table 11-4 Useful Windows PowerShell cmdlets

Command	Result
Get-MSDSMSupportedHW	The cmdlet lists hardware IDs in the Microsoft Device Specific Module (MSDSM) supported hardware list.
Get-MPIOSetting	The cmdlet gets Microsoft MPIO settings. The settings are as follows: <ul style="list-style-type: none"> ▶ PathVerificationState ▶ PathVerificationPeriod ▶ PDORemovePeriod ▶ RetryCount ▶ RetryInterval ▶ UseCustomPathRecoveryTime ▶ CustomPathRecoveryTime ▶ DiskTimeoutValue
Set-MPIOSetting	The cmdlet changes Microsoft MPIO settings. The settings are as follows: <ul style="list-style-type: none"> ▶ PathVerificationState ▶ PathVerificationPeriod ▶ PDORemovePeriod ▶ RetryCount ▶ RetryInterval ▶ UseCustomPathRecoveryTime ▶ CustomPathRecoveryTime ▶ DiskTimeoutValue

VMware using VMware native multipathing

There are two methods that are used to obtain the multipath information from the VMware ESX host:

- ▶ ESXi CLI: Use the CLI to obtain the multipath information when performing troubleshooting procedures.
- ▶ vSphere Client and vSphere Web Client: Use this option when you are performing system maintenance.

Command-line interface

To obtain logical unit number (LUN) multipathing information from the ESXi host CLI, complete the following steps:

1. Log in to the ESXi host console.
2. To get detailed information about the paths, run **esxcli storage core path list**.

Example 11-4 shows an example for the output of the **esxcli storage core path list** command.

Example 11-4 Output of “esxcli storage core path” list command

```
fc.5001438028d02923:5001438028d02922-fc.500507680100000a:500507680120000a-naa.6
00507680185801aa000000000000a68
UID:
fc.5001438028d02923:5001438028d02922-fc.500507680100000a:500507680120000a-naa.6
00507680185801aa000000000 000a68
Runtime Name: vmhba2:C0:T1:L54
```

```

Device: naa.600507680185801aa000000000000a68
Device Display Name: IBM Fibre Channel Disk
(naa.600507680185801aa000000000000a68)
Adapter: vmhba2
Channel: 0
Target: 1
LUN: 54
Plugin: NMP
State: active
Transport: fc
Adapter Identifier: fc.5001438028d02923:5001438028d02922
Target Identifier: fc.500507680100000a:500507680120000a
Adapter Transport Details: WWNN: 50:01:43:80:28:d0:29:23 WWPNN:
50:01:43:80:28:d0:29:22
Target Transport Details: WWNN: 50:05:07:68:01:00:00:0a WWPNN:
50:05:07:68:01:20:00:0a
Maximum I/O Size: 33553920

```

- To list detailed information for all the corresponding paths for a specific device, run **esxcli storage core path list -d <naaID>**.

Example 11-5 shows the output for the specified device with the ID naa.600507680185801aa000000000000972, which is attached with eight paths to the ESXi server. The output was omitted for brevity.

Example 11-5 Output of esxcli storage core path list -d <naaID>

```

fc.5001438028d02923:5001438028d02922-fc.500507680100037e:500507680120037e-naa.600507680185801aa0
00000000000972
UID:
fc.5001438028d02923:5001438028d02922-fc.500507680100037e:500507680120037e-naa.600507680185801aa0
00000000000972
Runtime Name: vmhba2:C0:T3:L9
Device: naa.600507680185801aa000000000000972
Device Display Name: IBM Fibre Channel Disk (naa.600507680185801aa000000000000972)
Adapter: vmhba2
Channel: 0
Target: 3
LUN: 9
Plugin: NMP
State: active
Transport: fc
Adapter Identifier: fc.5001438028d02923:5001438028d02922
Target Identifier: fc.500507680100037e:500507680120037e
Adapter Transport Details: WWNN: 50:01:43:80:28:d0:29:23 WWPNN: 50:01:43:80:28:d0:29:22
Target Transport Details: WWNN: 50:05:07:68:01:00:03:7e WWPNN: 50:05:07:68:01:20:03:7e
Maximum I/O Size:
33553920fc.5001438028d02923:5001438028d02922-fc.500507680100037e:500507680130037e-naa.6005076801
85801aa000000000000972
UID:

fc.5001438028d02923:5001438028d02922-fc.500507680100037e:500507680130037e-naa.600507680185801aa0
00000000000972
Runtime Name: vmhba2:C0:T2:L9
Device: naa.600507680185801aa000000000000972
Device Display Name: IBM Fibre Channel Disk (naa.600507680185801aa000000000000972)
Adapter: vmhba2

```

Channel: 0
Target: 2
LUN: 9
Plugin: NMP
State: active
Transport: fc
Adapter Identifier: fc.5001438028d02923:5001438028d02922
Target Identifier: fc.500507680100037e:500507680130037e
Adapter Transport Details: WWNN: 50:01:43:80:28:d0:29:23 WWPN: 50:01:43:80:28:d0:29:22
Target Transport Details: WWNN: 50:05:07:68:01:00:03:7e WWPN: 50:05:07:68:01:30:03:7e
Maximum I/O Size:
33553920fc.5001438028d02921:5001438028d02920-fc.500507680100037e:500507680110037e-naa.600507680185801aa000000000000972
UID:

4. The command `esxcli storage nmp device list` lists the LUN multipathing information for all attached disks.

Example 11-6 shows the output for one of the attached disks. All other output was omitted for brevity.

Example 11-6 Output for esxcli storage nmp device list

```
naa.600507680185801aa00000000000a68
  Device Display Name: IBM Fibre Channel Disk
(naa.600507680185801aa00000000000a68)
  Storage Array Type: VMW_SATP_ALUA
  Storage Array Type Device Config: {implicit_support=on;
explicit_support=off; explicit_allow=on; alua_followover=on;
action_OnRetryErrors=off; {TPG_id=1,TPG_state=ANO}{TPG_id=0,TPG_state=A0}}
  Path Selection Policy: VMW_PSP_RR
  Path Selection Policy Device Config:
{policy=rr,iops=1000,bytes=10485760,useANO=0; lastPathIndex=1;
NumIOsPending=0,numBytesPending=0}
  Path Selection Policy Device Custom Config:
  Working Paths: vmhba2:C0:T3:L54, vmhba1:C0:T2:L54, vmhba1:C0:T3:L54,
vmhba2:C0:T2:L54
  Is USB: false
```

vSphere Client HTML5 and Web Client

To obtain multipath settings for your storage in the HTML5 client, complete the following steps:

1. Select an ESXi host, and click the **Configure** tab.
2. Click **Storage Devices**.
3. Select the storage device that you want to verify.
4. Scroll down in the Properties tab and click **Edit multipathing...**

vSphere Client (Thick Client for 6.x)

To obtain multipath settings for your storage in vSphere Client, complete the following steps:

1. Select an ESXi host, and click the **Configuration** tab.
2. Click **Storage**.
3. Select a data store or mapped LUN.

4. Click **Properties**.
5. In the Properties dialog, select the extent, if necessary.
6. Select **Extent Device** → **Manage Paths** and obtain the paths from the Manage Path dialog.

11.2.3 Drive data collection: drivedumps

IBM FlashCore Module modules (FCMs) are a family of high-performance flash drives. The FCM design uses the NVMe protocol, a Peripheral Component Interconnect Express (PCIe) interface, and high-speed NAND memory that is driven by FCM field programmable gate array (FPGA) to provide high throughput, inline compression, and input/output operations per second (IOPS) with consistent and predictable latency.

For deeper analysis in cases where drives or FCM are involved, drivedumps are often useful. Their data can help you understand problems with the drive, and they do *not* contain any data that applications write to the drive. In some situations, drivedumps are automatically triggered by the system. To collect support data from a disk drive, run the **triggerdrivedump drive_id** command. The output is stored in a file in the `/dumps/drive` directory. This directory is on one of the nodes that are connected to the drive.

Example 11-7 shows the usage of the **triggerdrivedump** command.

Example 11-7 The triggerdrivedump command

```
IBM_IBM FlashSystem:FS9110:superuser>triggerdrivedump 1
Drive dump on node id [5] successfully created
IBM_IBM FlashSystem:FS9110:superuser>

IBM_IBM FlashSystem:FS9110:superuser>lsdumps -prefix /dumps/drive
id filename
0 drivedump_7812345-1_1_220411_055205
IBM_IBM FlashSystem:FS9110:superuser>
```

Any snap that is taken after the trigger command contains the stored drivedumps. It is sufficient to provide Snap Type 1: Standard logs for drivedumps.

11.2.4 More data collection

Data collection methods vary by storage platform, SAN switch, and operating system.

For an issue in a SAN environment when it is not clear where the problem is occurring, you might need to collect data from several devices in the SAN.

The following basic information must be collected for each type of device:

- ▶ Hosts:
 - Operating system: Version and level
 - Host Bus Adapter (HBA): Driver and firmware level
 - Multipathing driver level
- ▶ SAN switches:
 - Hardware model
 - Software version
- ▶ Storage subsystems:
 - Hardware model
 - Software version

For performance-related issues, it is helpful to have corresponding monitoring. Section 9.2, “Performance monitoring” on page 557 describes IBM Storage Insights and IBM Spectrum Control. If required, you can export performance data from there for the related period, as described in 9.7, “Performance diagnostic information” on page 628.

11.3 Common problems and isolation techniques

Environment comprised of SANs, storage subsystems, host systems and other components can be complex. They often consist of hundreds or thousands of disks, multiple redundant subsystem controllers, virtualization engines, and different types of SAN switches. All these components must be configured, monitored, and managed correctly. If issues occur, administrators must know what to look for and where to look.

IBM Storage Virtualize storage systems feature useful error logging and notification mechanisms. The system tracks its internal events and informs the user about issues in the SAN or storage subsystem. It also helps to isolate problems with the attached host systems. Therefore, by using these functions, administrators can easily locate any issue areas and take appropriate action to resolve any issue.

In many cases, IBM Storage Virtualize system and its service and maintenance features guide administrators directly, provide help, and suggest remedial actions. Furthermore, IBM Storage Virtualize determines whether or not a problem does still persist.

Another feature that helps administrators to isolate and identify issues that might be related to IBM Storage Virtualize systems is the ability of their nodes to maintain a database of other devices that communicate with the IBM Storage Virtualize system’s devices. Devices, such as hosts and optional back-end storages, are added or removed from the database as they start or stop communicating to IBM Storage Virtualize systems.

Although an IBM Storage Virtualize system’s node hardware and software events can be verified in the GUI or CLI, external events, such as failures in the SAN zoning configuration, hosts, and back-end storages, are common. You must troubleshoot these failures outside of the IBM Storage Virtualize systems.

For example, a misconfiguration in the SAN zoning might lead to the IBM Storage Virtualize cluster not working correctly. This problem occurs because the IBM Storage Virtualize cluster nodes communicate with each other by using the FC SAN fabrics.

In this case, check the following areas from an IBM Storage Virtualize system’s perspective:

- ▶ The attached hosts. For more information, see 11.3.2, “Host problems” on page 723.
- ▶ The SAN. For more information, see 11.3.3, “Fibre Channel SAN and IP SAN problems” on page 728.
- ▶ The attached storage subsystem. For more information, see 11.3.5, “Storage subsystem problems” on page 731.
- ▶ The local FC port masking and portsets. For more information, see 8.8, “Portsets” on page 530.

11.3.1 Interoperability

When you experience events in an IBM Storage Virtualize environment, as an initial step, ensure that all components that comprise the storage infrastructure are interoperable, which applies to hosts, host OS, Host Bus Adapter (HBA), driver, firmware, SAN devices, and back-end devices. In an IBM Storage Virtualize environment, the product support matrix is the main source for this information. For the latest IBM Storage Virtualize systems support matrix, see [IBM System Storage Interoperation Center \(SSIC\)](#).

It is crucial, to maintain up to date HBA firmware and device driver levels. This equally applies to multipath software and host OS patch levels. Failing to do so may lead to connectivity or interoperability issues, for instance host logins fail to reestablish after SAN maintenance activities. In worst case, this may lead to access loss.

11.3.2 Host problems

From host perspective, you can experience various situations that range from performance degradation to inaccessible disks. The first step in troubleshooting such issues is to check whether any potential interoperability issues exist.

After interoperability is verified, check the configuration of the host on the IBM Storage Virtualize system's side. The Hosts window in the GUI or the following CLI commands can be used to start a verification of potential host-related issues:

► **lshost**

Note: Depending on the connection type of the host (FC, FC direct attach, iSCSI, or NVMe, SAS) the output slightly differs in detail from each other.

This command displays the status of all configured host objects.

- Status *online*: the host ports are online in both nodes of an I/O group.
- Status *offline*: the host ports are offline in both nodes of an I/O group.
- Status *inactive*: the host has volumes that are mapped to it, but all its ports did not receive Small Computer System Interface (SCSI) commands in the last 5 minutes.
- Status *degraded*: there are no redundant logins to all nodes of an I/O group.

Example 11-8 shows the **lshost** command output.

Example 11-8 The lshost command

```
IBM_IBM FlashSystem:FLASHPFE95:superuser>lshost
0 Win2K8 2 4 degraded
1 ESX_67_B 2 4 online
2 ESX_67_A 2 1 offline
3 Server127 2 1 degraded
```

► **lshost <host_id_or_name>**

This command shows more information about a specific host. It is often used when you must identify which host port is not online in an IBM Storage Virtualize system node.

Example 11-9 shows the **lshost <host_id_or_name>** command output.

Example 11-9 The lshost <host_id_or_name> command

```
IBM_IBM FlashSystem:FLASHPFE95:superuser>lshost Win2K8
id 0
name Win2K8
```

For more information about managing hosts on IBM Storage Virtualize systems, see 8.10, “I/O queues” on page 534.

Apart from hardware-related situations, problems can exist in such areas as the operating system or the software that is used on the host. These problems normally are handled by the host administrator or the service provider of the host system. However, the multipathing driver that is installed on the host and its features can help to determine possible issues.

For example, a volume path issue is reported, which means that a specific HBA on the server side cannot reach all the nodes in the I/O group to which the volumes are associated.

Note: Subsystem Device Driver Device Specific Module (SDDDSM) and Subsystem Device Driver Path Control Module (SDDPCM) reached end of service (EOS). Therefore, migrate SDDDSM to MSDSM on Windows platform and SDDPCM to AIX Path Control Module (AIXPCM) on AIX and Virtual I/O Server (VIOS) platforms.

For more information, see [IBM Spectrum Virtualize Multipathing Support for AIX and Windows Hosts](#).

Faulty paths can be caused by hardware and software problems, such as the following examples:

- ▶ Hardware:
 - A faulty small form-factor pluggable transceiver (SFP) on the host or SAN switch.
 - Faulty fiber optic cables, for example damaged cables by exceeding the minimum permissible bend radius.
 - A faulty HBA.
 - Faulty SAN switch.
 - Contaminated SFP or cable connectors.
 - Patch panels
- ▶ Software or configuration:
 - Incorrect zoning, portset, or portmask.
 - Incorrect host-to-VDisk mapping.
 - Outdated HBA firmware or driver.
 - A back-level multipathing configuration or driver.

Based on field experience, it is a best practice that you complete the following hardware checks first:

- ▶ Whether connection error indicators are lit on the host, SAN switch or the IBM Storage Virtualize system.
- ▶ Whether all the parts are seated correctly. For example, cables are securely plugged in to the SFPs and the SFPs are plugged all the way into the switch port sockets.
- ▶ Ensure that fiber optic cables are not damaged. If possible, swap a suspicious cable with a known good cable.

Note: When replacing or relocating fibre channel cables, always clean their pluggable connectors using proper cleaning tools. This even applies to brand new cables taken from sealed bags.

After the hardware check, continue to check the following aspects of the software setup:

- ▶ Whether the HBA driver level and firmware level are at the preferred and supported levels.
- ▶ Verify your SAN zoning configuration.
- ▶ The general SAN switch status and health for all switches in the fabric.
- ▶ The multipathing driver, and make sure that it is at the preferred configuration and supported level.
- ▶ Link layer errors reported by the host or the SAN switch may indicate so far undiscovered cable or SFP issues.

iSCSI or iSCSI Extensions for Remote Direct Memory Access configuration and performance issues

This section describes the internet Small Computer Systems Interface (iSCSI) and iSCSI Extensions for Remote Direct Memory Access (RDMA) (iSER) configuration and performance issues.

Link issues

If the Ethernet port link does not come online, check whether the SFP or cables and the port support auto-negotiation with the switch. This issue is especially true for SFPs, which support 25 Gb and higher port speeds because a mismatch might exist in Forward Error Correction (FEC) that might prevent a port to auto-negotiate.

Another potential source of problems is 4X-splitter-cables and Direct Attach Copper (DAC) cables.

Longer cables are not only exposed to more noise or interference (high Bit Error Ratio (BER)); therefore, they require more powerful error correction codes.

Two IEEE 802.3 FEC specifications are important. For an auto-negotiation issue, verify whether a compatibility issue exists with SFPs at both end points:

- ▶ Clause 74: Fire Code (FC-FEC) or BASE-R (BR-FEC) (16.4 dB loss specification)
- ▶ Clause 91: Reed-Solomon (RS-FEC) (22.4 dB loss specification)

Use the `lshostiplogin` command to list the login session type, such as associated host object, login counts login protocol, and other details, for hosts that are identified by their iSCSI Qualified Name (IQN). The output is provided for ports, which logged in to Ethernet ports that are configured with IP addresses. The output shows, among other things, the protocol that is used.

The output in the protocol field indicates the connection protocol that is used by the configured IP host IQN to establish a login session that is referred by the login field. This value can be one of the following values:

- ▶ iSCSI
- ▶ iSER

Priority flow control

Priority flow control (PFC) is an Ethernet protocol that supports the ability to assign priorities to different types of traffic within the network. On most Data Center Bridging Capability Exchange (DCBX) protocol supported switches, verify whether Link Layer Discovery Protocol (LLDP) is enabled. The presence of a virtual local area network (VLAN) is a prerequisite for the configuration of PFC. It is recommended to set the priority tag 0 - 7.

A DCBX-enabled switch and a storage adapter exchange parameters that describe traffic classes and PFC capabilities.

In IBM Storage Virtualize systems, Ethernet traffic is divided into the following classes of service based on the feature use case:

- ▶ Host attachment (iSCSI or iSER)
- ▶ Back-end storage (iSCSI)
- ▶ Node-to-node communication (Remote Direct Memory Access (RDMA) clustering)

If challenges occur as the PFC is configured, verify the following attributes to determine the issue:

- ▶ Configure the IP address or VLAN by using `mkip`.
- ▶ Configure the class of service (COS) by using `chsystemethernet`.
- ▶ Ensure that the priority tag is enabled on the switch.
- ▶ Ensure that the `lspport ip` output is as follows:
`dcbx_state, pfc_enabled_tags`
- ▶ The Enhanced Transmission Selection (ETS) setting is recommended if a port is shared.

For more information about problem solving, see [Resolving a problem with PFC settings](#).

Standard network connectivity check

Verify that the required TCP/UDP ports are allowed in the network firewall. The following ports can be used for various host attachments:

- ▶ Software iSCSI requires TCP port 3260.
- ▶ iSER or RDMA over Converged Ethernet (RoCE) host requires TCP port 3260.
- ▶ iSER or iWRAP host requires TCP port 860.

A comprehensive list of TCP/IP address and port requirements is available at IBM Documentation [IP address allocation and usage](#).

Verify that the IP addresses are reachable and the TCP ports are open.

Note: iSER host attachment is not supported on IBM FlashSystem 9500, IBM FlashSystem 7300, and IBM SAN Volume Controller SV3. It is, however, supported on other IBM Storage Virtualize products.

iSCSI performance issues

Here are some of the attributes and host parameters that might affect iSCSI performance:

- ▶ TCP delayed acknowledgment (ACK).
- ▶ Ethernet jumbo frames.
- ▶ Network bottleneck or oversubscription.
- ▶ iSCSI session login balance.
- ▶ PFC setting and bandwidth allocation for iSCSI in the network.
- ▶ iSCSI bandwidth performance is reduced if an iSCSI or iSER host is mapped to a portset that contains IP addresses that are configured on 100 Gb ports.
- ▶ iSCSI I/O should use dedicated Ethernet ports or portsets on the IBM Storage Virtualize system, separate it from other kinds of usage, for instance IP replication.

In specific situations, the TCP/IP layer might attempt to combine several ACK responses into a single response to improve performance. However, that combination can negatively affect iSCSI read performance as the storage target waits for the response to arrive. This issue is observed when the application is single-threaded and has a low queue depth.

It is a best practice to disable the **TCPDelayedAck** parameter on the host platforms to improve overall storage I/O performance. If the host platform does not provide a mechanism to disable **TCPDelayedAck**, verify whether a smaller “Max I/O Transfer Size” with more concurrency (queue depth > 16) improves overall latency and bandwidth usage for the specific host workload. In most Linux distributions, this Max I/O Transfer Size is controlled by the **max_sectors_kb** parameter with a suggested transfer size of 32 kiB.

In addition, review network switch diagnostic data to evaluate potential issues as packet drop or packet retransmission. It is advisable to enable flow control or PFC to enhance the reliability of the network delivery system to avoid packet loss, aiming to enhance the overall performance.

For more information about iSCSI performance analysis and tuning, see [iSCSI performance analysis and tuning](#).

11.3.3 Fibre Channel SAN and IP SAN problems

It is not a difficult task to introduce IBM Storage Virtualize systems into your SAN environment and use its virtualization functions. However, before you can use IBM Storage Virtualize systems in your environment, you must follow some basic rules. These rules are not complicated, but you can make mistakes that lead to accessibility issues or a reduction in the performance experienced.

Various types of SAN zones are needed to run IBM Storage Virtualize systems in your environment: a *host zone*, and a *storage zone* for virtualized back-end storage systems. Additionally, you must have an IBM Storage Virtualize systems zone that contains all the IBM Storage Virtualize node ports used for communication between the clustered system's nodes or node canisters. Dedicated SAN zoning is also needed, if the IBM Storage Virtualize system does utilize Remote Copy Services for volume replication with another Storage Virtualize system.

For more information and important points about setting up IBM Storage Virtualize systems in a SAN fabric environment, see Chapter 2, “Storage area network guidelines” on page 123.

Because IBM Storage Virtualize systems are a major component of the SAN and connect the host to the storage subsystem, check and monitor the SAN fabrics.

FC ports intended for *intra-cluster* or *node-to-node* communication usage, should not be used for replication, host or back-end storage traffic. Equally, ports used for inter-system traffic for replication with other Storage Virtualize systems, should be dedicated for this purpose only.

Some situations of performance degradation and buffer-to-buffer credit exhaustion in FC-based configurations can be caused by suboptimally configured *local_fc_port_mask* and *partner_fc_port_mask*. To ensure optimal operation of your IBM Storage Virtualize systems, make sure to allow inter-node respectively inter-system traffic by properly SAN zoning and FC port mask configuration. Inter-node zoning within the same clustered system does police the logins between the FC ports, in addition the *local_fc_port_mask* policies the use of the logins. Further details on this can be found in [Fibre Channel port masking](#).

Some situations can cause issues in the SAN fabric and SAN switches. Problems can be related to a hardware fault or to a software problem on the switch. The following hardware defects are normally the easiest problems to find:

- ▶ Switch power, fan, or cooling units
- ▶ Installed SFP modules
- ▶ Fiber optic cables

Software failures are more difficult to analyze. In most cases, you must collect data and involve IBM Support. However, before you take further action, check the installed code level for any known issues in the switch vendor's release notes for the switch firmware and software. Also, check whether a new code level is available that resolves the problem that you are experiencing.

SAN connectivity issues are commonly related to zoning. For example, a wrong WWPN for a host zone may have been chosen, such as when two IBM Storage Virtualize System ports must be zoned to one HBA with one port from each IBM Storage Virtualize system node. Another example could be a host port WWPN that was omitted in the host zoning, causing the host object to be displayed as *degraded*.

SAN zoning therefore should be done after thorough planning, using a unified naming schema for aliases, zones et cetera. It is equally beneficial for both the user and any supporting function, if the SAN switches follow a clear naming structure as well as a coordinated domain id assignment. While it is not a problem from a technical point of view to reuse switch domain ids across SAN fabrics, it unnecessarily complicates troubleshooting, as the switch nPort IDs (nPID) will show up multiple times in diagnostic data and the output of CLI commands as for instance **lspportfc**, **lstargetportfc** and **lsfabric**.

Existing SAN environment often have developed organically over time with no or little documentation. It definitely pays to create a documentation and graphical SAN layouts, to enable faster problem analysis and resolution.

On IBM Storage Virtualize systems, a part of the worldwide port names (WWPN) is derived from the worldwide node name (WWNN) of the node canister in which the adapter is installed. The WWNN is part of each node's Vital Product Data (VPD), it impacts the WWPN's last four digits. The ports' WWPN also are derived from the PCIe slot the adapter is installed in and its port id. For more information, see [Worldwide node and port names](#).

So, the WWPNs for the different ports of the same node differ in the 6th and 5th last digit. For example:

```
50:05:07:68:10:13:37:dc
50:05:07:68:10:14:37:dc
50:05:07:68:10:24:37:dc
```

The WWPNs for ports on different nodes differ in the last 4 digits. For example, here are the WWPNs for port 3 and 4 on each node of a IBM FlashSystem:

```
50:05:07:68:10:13:37:dc
50:05:07:68:10:14:37:dc

50:05:07:68:10:13:37:e5
50:05:07:68:10:14:37:e5
```

As shown in Example 11-11 on page 730, two ports that belong to the same Storage Virtualize node are zoned to a host FC port. Therefore, the result is that the host port will not log in to both nodes of that I/O group and the multipathing driver will not see redundant paths:

Example 11-11 Incorrect WWPN zoning

```
zone: z_Win2k19_FS9110_iogrp0
      50:05:07:68:10:13:37:dc
      50:05:07:68:10:14:37:dc
      20:00:00:e0:8b:89:cc:c2
```

The correct zoning must look like the zoning that is shown in Example 11-12.

Example 11-12 Correct WWPN zoning

```
zone: z_Win2k19_FS9110_iogrp0
      50:05:07:68:10:14:37:e5
      50:05:07:68:10:14:37:dc
      20:00:00:e0:8b:89:cc:c2
```

The following IBM FlashSystem error codes are related to the SAN environment:

- ▶ Error 1060: Fibre Channel ports are not operational.
- ▶ Error 1220: A remote port is excluded.

A bottleneck is another common issue that is related to SAN switches. The bottleneck can be present in a port where a host, storage subsystem, or IBM Storage Virtualize device is connected, or in Inter-Switch Link (ISL) ports. The bottleneck can occur in some cases, such as when a device that is connected to the fabric is slow to process received frames, or if a SAN switch port cannot transmit frames at a rate that is required by a device that is connected to the fabric.

These cases can slow down communication between devices in your SAN. To resolve this type of issue, see the SAN switch documentation to investigate and identify what is causing the bottleneck and how to fix it.

If you cannot fix the issue with these actions, use the method that is described in 11.2, “Collecting diagnostic data” on page 710, collect the SAN switch debugging data, and then contact the vendor for assistance or open a case with the vendor.

11.3.4 Port issues and transceiver statistics

Ports and their connections are involved in many support cases. Starting with IBM Storage Virtualize 8.4, a new support command is available. The **lspportstats** command supports the administrator in troubleshooting ports of any kind on the IBM Storage Virtualize systems side. The output of the command contains many different details like the port type, WWPN, IQN, send and receive statistics, and SFP details. For example, you can check the physical error counter for a port and other interesting values of an SFP.

Example 11-13 shows the output for an FC port.

Example 11-13 lspportstats command output

```
IBM_FlashSystem:FS9110:superuser>lspportstats -node node1
Nn_stats_78E003K-1_230623_151121
<port id="1"
type="FC"
type_id="1"
wwpn="0x5005076810110214"
fc_wwpn="0x5005076810110214"
fcoe_wwpn=""
```

```

sas_wnn=""
iqn=""
hbt="80487" hbr="0" het="0" her="1465"
cbr="0" cbr="612" cet="260" cer="0"
lnbt="0" lnbr="0" lnet="955316" lner="955332"
rmbt="0" rnbr="0" rmet="0" rmer="0"
dtdt="242" dtdc="6" dtdm="956797"
dtdt2="242" dtdc2="6"
lf="14" lsy="21" lsi="0" pspe="0"
itw="54" icrc="0" bbcz="0"
tmp="46" tmpht="85"
txpwr="596" txpwr1t="126"
rxpwr="570" rxpwr1t="31"
hsr="0" hsw="0" har="0" haw="0"
/>
<port id="2"
type="FC"
type_id="2"

z
type_id="2"
[...]
```

Table 11-5 shows some of the most interesting attributes and their meanings.

Table 11-5 Selected attributes of the `lsportstats` output

Attribute	Information
Nn_stats_78F13MY-1_220413_152655	Data source stats file of the output.
lsy	Indicates the loss of sync error count.
itw	Invalid transmissionword error count.
icrc	Indicates the invalid cyclic redundancy check (CRC) error count.
txpwr	SFP TX power in microwatts.
rxpwr	SFP RX power in microwatts.

It is not possible to reset or clear the shown counter with a command at the moment. To examine the current trend of the values or whether they are increasing, a best practice is to compare two outputs of the command for differences. Allow some run time between the two iterations of the command.

For more information about the `lsportstats` command, see [lsportstats](#).

11.3.5 Storage subsystem problems

Today, various heterogeneous storage subsystems are available. All these subsystems have different management tools, different setup strategies, and possible problem areas depending on the manufacturer. To support a stable environment, all subsystems must be correctly configured by following best practices and have no existing issues.

If you experience a storage-subsystem-related issue, check the following areas:

- ▶ Always check the [SSIC](#) to see whether the subsystem is supported.
- ▶ Storage subsystem configuration: Ensure that a valid configuration and best practices are applied to the subsystem.
- ▶ Storage subsystem controllers: Check the health and configurable settings on the controllers.
- ▶ Storage subsystem array: Check the state of the hardware, such as an FCM, solid-state drive (SSD), or disk drive module (DDM) failure or enclosure alerts.
- ▶ Storage volumes: Ensure that the LUN masking is correct.
- ▶ Host attachment ports: Check the status, configuration, and connectivity to storage SAN switches.
- ▶ Layout and size of redundant array of independent disks (RAID) arrays and LUNs: Performance and redundancy are contributing factors.

IBM Storage Virtualize has several CLI commands that you can use to check the status of the system and attached storage subsystems. Before you start a complete data collection or problem isolation on the SAN or subsystem level, first use the following commands and check the status from the IBM Storage Virtualize perspective:

▶ **lscontroller <controller_id_or_name>**

Checks that multiple WWPNs that match the back-end storage subsystem controller ports are available.

Checks that the path_counts are evenly distributed across each storage subsystem controller, or that they are distributed correctly based on the preferred controller. The total of all path_counts must add up to the number of MDisks multiplied by the number of IBM Storage Virtualize nodes.

▶ **lsmdisk**

Checks that all MDisks are online (not degraded or offline).

▶ **lsmdisk <MDisk_id_or_name>**

Checks several of the MDisks from each storage subsystem controller. Are they online? Do they all have path_count = number of back-end ports in the zone to IBM Storage Virtualize x number of nodes? An example of the output from this command is shown in Example 11-14. MDisk 0 is a local MDisk in an IBM FlashSystem, and MDisk 1 is provided by an external, virtualized storage subsystem.

Example 11-14 Issuing a lsmdisk command

```
IBM_IBM FlashSystem:FLASHPFE95:superuser>lsmdisk 0
id 0
name MDisk0
status online
mode array
MDisk_grp_id 0
MDisk_grp_name Pool0
capacity 198.2TB
quorum_index
block_size
controller_name
ctrl_type
ctrl_wwnn
controller_id
path_count
max_path_count
```



```
ctrl_LUN_#
UID
preferred_WWPN
active_WWPN
fast_write_state empty
raid_status online
raid_level raid6
redundancy 2
strip_size 256
spare_goal
spare_protection_min
balanced exact
tier tier0_flash
slow_write_priority latency
fabric_type
site_id
site_name
easy_tier_load
encrypt no
distributed yes
drive_class_id 0
drive_count 8
stripe_width 7
rebuild_areas_total 1
rebuild_areas_available 1
rebuild_areas_goal 1
dedupe no
preferred_iscsi_port_id
active_iscsi_port_id
replacement_date
over_provisioned yes
supports_unmap yes
provisioning_group_id 0
physical_capacity 85.87TB
physical_free_capacity 78.72TB
write_protected no
allocated_capacity 155.06TB
effective_used_capacity 16.58TB.
```

```
IBM_IBM FlashSystem:FLASHPFE95:superuser>lsmdisk 1
id 1
name flash9h01_itsosvcc11_0
status online
mode managed
MDisk_grp_id 1
MDisk_grp_name Pool1
capacity 51.6TB
quorum_index
block_size 512
controller_name itsoflash9h01
ctrl_type 6
ctrl_WWNN 500507605E852080
controller_id 1
path_count 16
max_path_count 16
```

```
ctrl_LUN_# 0000000000000000
UID 6005076441b53004400000000000000100000000000000000000000000000000
preferred_WWPN
active_WWPN many
```

NOTE: lines removed for brevity

Example 11-14 on page 732 shows that for MDisk 1 that the external storage controller has eight ports that are zoned to IBM Storage Virtualize systems, which have two nodes (8 x 2 = 16).

► **lsvdisk**

Checks that all volumes are online (not degraded or offline). If the volumes are degraded, are there stopped FlashCopy jobs present? Restart stopped FlashCopy jobs or seek IBM Storage Virtualize systems support guidance.

► **lsfabric**

Use this command with the various options, such as **-controller controllerid**. Also, check different parts of the IBM Storage Virtualize systems configuration to ensure that multiple paths are available from each IBM Storage Virtualize node port to an attached host or controller. Confirm that IBM Storage Virtualize systems node port WWPNs are also consistently connected to an external back-end storage.

Determining the number of paths to an external storage subsystem

By using CLI commands, the total number of paths to an external storage subsystem can be determined. To determine the value of the available paths, use the following formulas:

Number of MDisks x Number of nodes per Cluster = Number of paths
MDisk_link_count x Number of nodes per Cluster = Sum of path_count

Example 11-15 shows how to obtain this information by using the **lscontroller <controllerid>** and **svcinfo lsnode** commands.

Example 11-15 Output of the svcinfo lscontroller command

```
IBM_IBM FlashSystem:FLASHPFE95:superuser>lscontroller 1
id 1
controller_name itsof9h01
WWNN 500507605E852080
MDisk_link_count 16
max_MDisk_link_count 16
degraded no
vendor_id IBM
product_id_low FlashSys
product_id_high tem-9840
product_revision 1430
ctrl_s/n 01106d4c0110-0000-0
allow_quorum yes
fabric_type fc
site_id
site_name
WWPN 500507605E8520B1
path_count 32
max_path_count 32
WWPN 500507605E8520A1
path_count 32
```

```

max_path_count 64
WWPN 500507605E852081
path_count 32
max_path_count 64
WWPN 500507605E852091
path_count 32
max_path_count 64
WWPN 500507605E8520B2
path_count 32
max_path_count 64
WWPN 500507605E8520A2
path_count 32
max_path_count 64
WWPN 500507605E852082
path_count 32
max_path_count 64
WWPN 500507605E852092
path_count 32
max_path_count 64

```

```
IBM_IBM FlashSystem:FLASHPFE95:superuser>svcinflsnode
```

```

id name UPS_serial_number WWNN status IO_group_id IO_group_name
config_node UPS_unique_id hardware iscsi_name
iscsi_alias panel_name enclosure_id canister_id enclosure_serial_number site_id
site_name

```

```

1 node1 500507681000000A online 0 io_grp0 no
AF8 iqn.1986-03.com.ibm:2145.flashpfe95.node1 01-2 1
2 F313150
2 node2 5005076810000009 online 0 io_grp0 yes
AF8 iqn.1986-03.com.ibm:2145.flashpfe95.node2 01-1 1
1 F313150

```

```
IBM_IBM
FlashSystem:FLASHPFE95:superuser>
```

Example 11-15 on page 734 shows that 16 MDisks are present for the external storage subsystem controller with ID 1, and two IBM Storage Virtualize nodes are in the cluster. In this example, the path_count is $16 \times 2 = 32$.

IBM Storage Virtualize has useful tools for finding and analyzing back-end storage subsystem issues because it includes a monitoring and logging mechanism.

Typical events for storage subsystem controllers include incorrect configuration, which results in a 1625 - A controller configuration is not supported error code. Other issues that are related to the storage subsystem include failures that point to the MDisk I/O (error code 1310), disk media (error code 1320), and error recovery procedure (error code 1370).

However, all messages do not have only one specific reason for being issued. Therefore, you must check several areas for issues, not only the storage subsystem.

To determine the root cause of a problem, complete the following steps:

A system can be part of only two IP partnerships. IBM Storage Virtualize systems with pre-8.4.2.0 firmware are still limited to one IP partnership. Partnerships on low memory platform nodes share memory resources, which can lead to degraded performance.

Portsets replace the requirement for creating remote-copy groups for IP partnerships. Dedicated portsets can be created for remote copy traffic. The dedicated portsets provide a group of IP addresses for IP partnerships.

During updates of the software, any IP addresses that are assigned to remote-copy groups with an IP partnership are automatically moved to a corresponding portset. For example, if remote-copy group 1 is defined on the system before the update, IP addresses from that remote-copy group are mapped to portset 1 after the update. Similarly, IP addresses in remote-copy group 2 are mapped to portset 2.

The native IP replication feature uses the following TCP/IP ports for remote cluster path discovery and data transfer, therefore, these ports need to be open:

- ▶ IP partnership management IP communication: TCP port 3260
- ▶ IP partnership data path connections: TCP port 3265

If a connectivity issue exists between the cluster in the management communication path, the cluster reports error code 2021: Partner cluster IP address unreachable. However, when a connectivity issue exists in the data path, the cluster reports error code 2020: IP Remote Copy link unavailable.

For more information, see [Resolving a not_present state for an IP partnership](#).

If the IP addresses are reachable and TCP ports are open, verify whether the end-to-end network supports a maximum transmission unit (MTU) of 1500 bytes without packet fragmentation. When an external host-based ping utility is used to validate end-to-end MTU support, use the “do not fragment” qualifier.

Fix the network path so that traffic can flow correctly. After the connection is made, the error auto-corrects.

The network quality of service largely influences the effective bandwidth usage of the dedicated link between the cluster. Bandwidth usage is inversely proportional to round-trip time (RTT) and the rate of packet drop or retransmission in the network.

Note: For standard block traffic, a packet drop or retransmission of 0.5% or more can lead to unacceptable usage of the available bandwidth.

Work with the network team to investigate over-subscription or other quality of service (QoS) issues of the link, with an objective of having the lowest possible (less than 0.1%) packet-drop percentage.

For more information about the configuration, see 6.5, “Native IP replication” on page 458. For more information about performance contributors, see 6.5.8, “Native IP replication performance considerations” on page 475.

11.3.7 Remote Direct Memory Access-based clustering

RDMA technology supports zero-copy networking, which makes it possible to read data directly from the main memory of one computer and write that data directly to the main memory of another computer. This technology bypasses CPU intervention during I/O, leading to lower latency and a faster rate of data transfer.

An IBM Storage Virtualize cluster can be formed by using RDMA-capable NICs that use RoCE or internet Wide-area RDMA Protocol (iWARP) technology. Consider the following points:

- ▶ Inter-node Ethernet connectivity can be done over identical ports only, and such ports must be connected within the same switching fabric.
- ▶ To ensure best performance and reliability, a minimum of two dedicated RDMA-capable Ethernet ports are required for node-to-node communications. These ports must be configured for inter-node traffic only and must not be used for host attachment, virtualization of Ethernet-attached external storage, or IP replication traffic.
- ▶ If the cluster will be created without an ISL (up to 300 meters (984 feet)), deploy independent (isolated) switches.
- ▶ If the cluster will be created on a short-distance ISL (up to 10 km (6.2 miles)), provision as many ISL between switches as there are RDMA-capable cluster ports.
- ▶ For a long-distance ISL (up to 100 km (62 miles)), the Dense Wavelength Division Multiplexing (DWDM) and Coarse Wavelength Division Multiplexing (CWDM) methods are applicable for L2 networks. Packet-switched or VXLAN methods are deployed for an L3 network because this equipment comes with deeper buffer “pockets”.

The following ports must be opened in the firewall for IP-based RDMA clustering:

- ▶ TCP 4791, 21451, 21452, and 21455
- ▶ UDP 4791, 21451, 21452, and 21455

For more information, see [Configuration details for using RDMA-capable Ethernet port for node-to-node communications](#).

Best practices to manage RDMA-capable Ethernet ports

The basic node tasks, such as adding a node or removing a node, are the same for both FC-based and RDMA-based connections between nodes. But, you might need to complete management actions on the RDMA-capable Ethernet ports before completing node-level management tasks.

Before completing managing tasks that are related to RDMA-capable Ethernet ports on a node, use the following best practices to manage these ports:

- ▶ If you already have a system that is configured to use RDMA-capable Ethernet ports, you must ensure that one redundant path is available before adding, removing, or updating settings for RDMA-capable Ethernet ports.
- ▶ Add, remove, or update settings on only one RDMA-capable Ethernet port at a time. Wait 15 seconds between these changes before updating other RDMA-capable Ethernet ports.
- ▶ If you are using a VLAN to create physical separation of networks, ensure that you follow these extra guidelines when completing management-related tasks:
 - VLAN IDs cannot be updated or added independently of other settings on a RDMA-capable Ethernet port, such as an IP address.
 - Before adding or updating VLAN ID information to RDMA-capable Ethernet ports, you must configure VLAN support on all the Ethernet switches in your network. For example, on each switch, set VLAN to “Trunk” mode, and specify the VLAN ID for the RDMA-capable Ethernet ports that will be in the same VLAN.

Problem determination

The first step is to review whether the node IP address is reachable and verify that the required TCP/UDP ports are accessible in both directions.

The following CLI command lists the port level connectivity information for node to node or clustering connectivity, and can be helpful to find the reason for connectivity error:

```
sainfo lsnodeipconnectivity
```

This command lists the port level connectivity information for node to node or clustering connectivity.

The [IBM Documentation for the lsnodeipconnectivity command](#) lists the different error_data values with a description, and provides possible corrective actions.

11.3.8 Advanced copy services or data reduction related problems

Performance of a specific storage feature or overall storage subsystem is generally interlinked, meaning that a bottleneck in one software or hardware layer can propagate to other layers. Therefore, problem isolation is a critical part of performance analysis.

The first thing to check is whether any unfixed events exist that require attention. After the fix procedure is followed to correct the alerts, the next step is to check the audit log to determine whether any activity exists that can trigger the performance issue. If that information correlates, more analysis can be done to check whether that specific feature is used.

The most common root causes for performance issues are SAN congestion, configuration changes, incorrect sizing or estimation of advanced copy services (replication, FlashCopy, and volume mirroring), or I/O load change.

The following sections are a quick reference to common misconfigurations.

Volume mirroring

The write-performance of the mirrored volumes is dictated by the slowest copy. Reads are served from the primary copy of the volume (in a stretched cluster topology, both copies can serve reads, which are dictated by the host site attribute). Therefore, size the solution as needed.

The mirroring layer maintains a bitmap copy on the quorum device. If a quorum disk is not accessible and volume mirroring cannot update the state information, a mirrored volume might need to be taken offline to maintain data integrity. Similarly, slow access to the quorum can affect the performance of mirroring volumes.

Problems sometimes occur during the creation of a mirrored volume or in relation to the duration of the synchronization. Helpful details and best practices are described in 6.6.6, “Bitmap space for out-of-sync volume copies” on page 484 and 6.6.5, “Volume mirroring performance considerations” on page 482.

FlashCopy

FlashCopy is a function that you can use to create a point-in-time copy of one of your volumes. Section 6.2.4, “FlashCopy planning considerations” on page 381 provides technical background and details for FlashCopy configurations. Review the provided recommendations and best practices for performance-related issues in 6.2.4, “FlashCopy planning considerations” on page 381.

Consider the following points for FlashCopy troubleshooting:

- ▶ Verify that the preferred node of the FlashCopy source and target volumes is the same to avoid excessive internode communication, except for:

- A clustered IBM Storage Virtualize system with multiple I/O groups in HyperSwap, where the source volumes are evenly spread across all the nodes. In this case, the preferred node placement is to follow the location of the source and target volumes on the back-end storage to avoid the re-direction of the FlashCopy write operation through the node-to-node network.
- A clustered IBM FlashSystem system with multiple control enclosures, where the source volumes are evenly spread across all the canisters. In this case, the preferred node placement is to follow the location of source and target volumes on the internal storage.
- ▶ High background copy rate and clean rate of FlashCopy relationships can cause back-end overload.
- ▶ Port saturation or node saturation. Review if the values are correctly sized.
- ▶ Check the number of FC relationships in any FlashCopy consistency group (CG). The larger the number of relationships, the higher the I/O pause time (Peak I/O Latency) when the CG starts.
- ▶ Consider using multiple CGs to spread and better use the available resources.
- ▶ If the host I/O pattern is small and random, evaluate whether reducing the FlashCopy grain size to 64 KB provides any improvement in latency compared to the default grain size of 256 KB.

For more information, see [FlashCopy function](#).

Policy-based replication

Policy-based replication helps you to replicate data between systems with minimal management, significantly higher throughput, and reduced latency compared to the asynchronous remote copy function.

Policy-based replication simplifies asynchronous replication with the following key advantages:

- ▶ Uses volume groups instead of consistency groups. With volume groups, all volumes are replicated based on the assigned policy.
- ▶ Simplifies administration by removing the need to manage relationships and change volumes.
- ▶ Automatically manages provisioning on the remote system.
- ▶ Supports easier visualization of replication during a site failover.
- ▶ Automatically notifies you when the recovery point objective (RPO) is exceeded.
- ▶ Easy-to-understand status and alerts on the overall health of replication.

Checking the status and RPO using the GUI

You can view the status of replication on the Policies tab of a volume group. To check the status in management GUI, complete these steps:

- ▶ In the management GUI, select **Volumes** → **Volume Groups**.
- ▶ Select the volume group to view.
- ▶ On the Volume Groups page, select **Policies**.
- ▶ On the Policies page, the following statuses as shown in Figure 11-9 on page 741 can be displayed:

Replication state	Description	Action required
Replication running	Indicates that data is currently being replicated between systems.	No action required
◇ Independent access	Indicates that replication is stopped and each copy of the volume group is accessible for I/O.	To resume replication, choose the system you would like to use the data and configuration from. Make this system the production copy.
⌚ Replication suspended	Indicates that replication is suspended due to an error on one of the systems. Replication will automatically resume when all errors are resolved.	Review the event log and address errors.
🔴 System disconnected	Indicates that the connection between systems is unavailable.	Restore connectivity between the systems.

Figure 11-9 PBR replication status

If a volume group exceeds the RPO of the associated replication policy, then an alert is generated in the event log. Depending on notification settings, a notification is sent by either an email, syslog, or SNMP. The management GUI displays following RPO statuses as shown in Figure 11-10:

RPO status	Description
✓ Within policy's RPO	Indicates that replication is within the RPO value set in the policy.
🔴 Outside policy's RPO	Indicates that the data on the recovery copy is outside the RPO value set in the policy.
🕒 Initial copy in progress	Indicates that the replication is in progress for the first time.
🕒 Initial copy incomplete	Indicates that the replication is incomplete and replication is suspended.

Figure 11-10 RPO statuses

Checking the status and RPO using the CLI

Run the following command to check the status of replication for a volume group:
lsvolume group replication

More details see: [Checking the status and RPO for policy-based replication.](#)

Safeguarded Copy

Safeguarded Copy on IBM Storage Virtualize supports the ability to create cyber-resilient point-in-time copies of volumes that cannot be changed or deleted through user errors, malicious actions, or ransomware attacks. The system integrates with IBM Copy Services Manager (IBM CSM) to provide automated backup copies and data recovery.

The online documentation of IBM CSM provides a dedicated chapter for troubleshooting and support.

For more information, see *IBM FlashSystem Safeguarded Copy Implementation Guide*, REDP-5654 and [IBM Copy Services Manager -> Troubleshooting and support.](#)

HyperSwap

With HyperSwap, a fully independent copy of the data is maintained at each site. When data is written by hosts at either site, both copies are synchronously updated before the write operation is completed. The HyperSwap function automatically optimizes itself to minimize data that is transmitted between two sites, and to minimize host read/write latency.

Verify that the link between the sites is stable and has enough bandwidth to replicate the peak workload. Also, check whether a volume must frequently change the replication direction from one site to another one. This issue occurs when a specific volume is being written by hosts from both the sites. Evaluate whether this issue can be avoided to reduce frequent direction changes. Ignore this issue if the solution is designed for active/active access.

If a single volume resynchronization between the sites takes a long time, review the partnership `link_bandwidth_mbits` and per `relationship_bandwidth_limit` parameters.

Data reduction pools

Data reduction pools (DRPs) internally implement a Log Structured Array (LSA), which means that writes (including over-writes or updates) always allocate newer storage blocks. The older blocks (with invalid data) are marked for garbage collection later.

The garbage-collection process is designed to defer the work as much as possible because the more it is deferred, the higher the chance of having to move only a small amount of valid data from the block to make that block available to the free pool. However, when the pool reaches more than 85% of its allocated capacity, garbage collection must speed up to move valid data more aggressively to make space available sooner. This issue might lead to increased latency because of increased CPU usage and load on the back-end. Therefore, it is a best practice to manage storage provisioning to avoid such scenarios.

Note: If the usable capacity of a DRP exceeds more than 85%, I/O performance can be affected. The system needs 15% of usable capacity that is available in DRPs to ensure that capacity reclamation can be performed efficiently.

Users are encouraged to pay close attention to any GUI notifications and use best practices for managing physical space. Use data reduction only at one layer (at the virtualization layer or the back-end storage or drives) because no benefit is realized by compressing and deduplicating the same data twice.

Because encrypted data cannot be compressed, data reduction must be done before the data is encrypted. Correct sizing is important to get the best performance from data reduction; therefore, use data reduction tools to evaluate system performance and space saving.

IBM Storage Virtualize systems use the following types of data reduction techniques:

- ▶ IBM FlashSystem that use FCM NVMe drives have built-in hardware compression.
- ▶ IBM FlashSystem that use industry-standard NVMe drives and SVC rely on the IBM Storage Virtualize software and DRP pools to deliver data reduction.

For more information about DRPs, see *Introduction and Implementation of Data Reduction Pools and Deduplication*, SG24-8430.

Compression

Starting with IBM Storage Virtualize 8.4, the integrated Comprestimator is always enabled and running continuously, thus providing up-to-date compression estimation over the entire cluster, both in the GUI and IBM Storage Insights. To display information for the thin-provisioning and compression estimation analysis report for all volumes, run the `lsdiskanalysis` command.

11.3.9 Health status during an upgrade

During the software upgrade process, alerts that indicate that the system is not healthy are reported. These alerts are normal behavior because the IBM Storage Virtualize systems node canisters go offline during this process; therefore, the system triggers these messages.

Normal behavior alerts for hardware, logical, and connectivity components during an upgrade of IBM Storage Virtualize storage systems are as follows:

- ▶ Degraded Host Connectivity
- ▶ Degraded Array MDisks
- ▶ Degraded Volumes
- ▶ Degraded connectivity to the internal disks
- ▶ Degraded Control Enclosures
- ▶ Degraded Expansion Enclosures
- ▶ Degraded Drives
- ▶ Node offline
- ▶ FC Ports offline
- ▶ Serial-attached SCSI (SAS) Ports offline
- ▶ Enclosure Batteries offline
- ▶ Node added
- ▶ Node restarted
- ▶ Number of device logins reduced on IBM Storage Virtualize System (for example, when an IBM Storage Virtualize system is updated, it is used as back-end storage for the SVC)

When you attempt to upgrade an IBM Storage Virtualize system, you also might receive a message, such as an error occurred in verifying the signature of the update package. This message does not mean that an issue exists in your system. Sometimes, this issue occurs because not enough space is available on the system to copy the file, or the package is incomplete or contains errors. In this case, open a Support Ticket with IBM Support and follow their instructions.

11.3.10 Managing the physical capacity of overprovisioned storage controllers

Drives and back-end controllers exist that include built-in hardware compression and other data reduction technologies that allow capacity to be provisioned over the available real physical capacity. Different data sets lead to different capacity savings, and some data, such as encrypted data or compressed data, does not compress. When the physical capacity savings do not match the expected or provisioned capacity, the storage can run out of physical space, which leads to a write-protected drive or array.

To avoid running out of space on the system, the usable capacity must be monitored carefully by using the GUI of the IBM Storage Virtualize system. The IBM Storage Virtualize GUI is the only capacity dashboard that shows the physical capacity.

Monitoring is especially important when migrating substantial amounts of data onto IBM Storage Virtualize systems, which typically occur during the first part of the workload lifecycle as data is on-boarded or initially populated into the storage system.

IBM encourages users to configure Call Home on the IBM Storage Virtualize system. Call Home monitors the physical free space on the system and automatically opens a service call for systems that reach 99% of their usable capacity. IBM Storage Insights also can monitor and report on any potential out-of-space conditions, and the new Advisor function warns when the IBM Storage Virtualize system almost at full capacity. For more information, see 11.6.5, “IBM Storage Insights Advisor” on page 763.

When the IBM Storage Virtualize system pool reaches an out-of-space condition, the device drops into a read-only state. An assessment of the data compression ratio (CR) and the re-planned capacity estimation should be done to determine how much outstanding storage demand might exist. This extra capacity must be prepared and presented to the host so that recovery can begin.

The approaches that can be taken to reclaim space on the IBM Storage Virtualize system in this scenario vary by the capabilities of the system, optional external back-end controllers, the system configuration, and planned capacity overhead needs.

In general, the following options are available:

- ▶ Add capacity to the IBM Storage Virtualize system. Customers are encouraged to plan to add capacity to the system when needed.
- ▶ Reserve space in the IBM Storage Virtualize system that makes it “seem” fuller than it really is, and that you can free up in an emergency situation. IBM Storage Virtualize can create a volume that is not compressed, deduplicated, or thin-provisioned (a fully allocated volume). Create some of these volumes to reserve an amount of physical space, and give them a descriptive name (for example, “emergency buffer space”). If you are reaching the limits for physical capacity, you can delete one or more of these volumes to give yourself a temporary reprieve.

Important: Running out of space can be a serious situation. Recovery can be time-consuming. For this reason, it is imperative that suitable planning and monitoring be done to avoid reaching this condition.

Next, we describe the process for recovering from an out-of-space condition.

Analyzing the situation

This stage of the recovery gathers the pertinent details of the state of the system to form the recovery strategy.

Reclaiming and unlocking

After you assess and account for storage capacity, contact IBM Support, who can help unlock the read-only mode and restore operations. The reclamation task can take a long time to run, and larger flash arrays take longer to recover than smaller ones.

Freeing up space

You can reduce the amount of used space by using several methods, which are described in the following sections.

Reclaiming space in a standard pool

To recover from out of space conditions on standard pools, complete the following steps:

1. Add storage to the system, if possible.
2. Migrate extents from the write-protected array to other non write-protected MDisk with enough extents, such as an external back-end storage array.

3. Migrate volumes with extents on the write-protected array to another pool. If possible, moving volumes to another pool can free up space in the affected pool to allow for space reclamation.
4. As this volume moves into the new pool, its previously occupied flash extents are freed (by using SCSI **unmap**), which then provides more free space to the IBM FlashSystem enclosure to be configured to a proper provisioning to support the CR.
5. Delete dispensable volumes to free up space. If possible, within the pool (MDisk group) on the IBM Storage Virtualize system, delete unnecessary volumes. IBM Storage Virtualize systems support SCSI **unmap**, so deleting volumes results in space reclamation benefits by using this method.
6. Bring the volumes in the pool back online by using a Directed Maintenance Procedure.

Reclaiming space in a data reduction pool

If the out of space storage is in a DRP, contact IBM Support. Due to the basic organization of a DRP, extent-level migrations are not possible. However, DRPs make full usage of the SCSI **unmap** function, including a garbage-collection process that can be used to reclaim space. Because of this capability, there are a few different methods that can be used to reclaim space.

Note: Power off all hosts accessing the pool to avoid host writes from impacting the success of the recovery plan.

For more information about the types of recovery for out of space situations, including standard pools and DRPs, see [Handling out of physical space conditions](#).

11.3.11 Replacing a failed flash drive

When IBM FlashSystem detects a failed FCM, NVMe drive, or storage-class memory (SCM) drive, it automatically generates an error in the Events window. To replace the failed drive, select **Monitoring** → **Events**, and then run the Fix Procedure for this event.

The Fix Procedure helps you to identify the enclosure and slot where the bad drive is located, and guides you to the correct steps to follow to replace it.

When a flash drive fails, it is removed from the array, and the rebuild process to the available rebuild areas starts. After the failed flash drive is replaced and the system detects the replacement, it reconfigures the new drive, a copy-back starts, and the new drive is used to fulfill the array membership goals of the system.

11.3.12 Recovering from common events

You can recover from several of the more common events that you might encounter by using the Recommended Action feature. In all cases, you must read and understand the current product limitations to verify the configuration and determine whether you must upgrade any components or install the latest fixes.

To obtain IBM product support, see [IBM Support](#).

If the problem is caused by IBM Storage Virtualize and you cannot fix it by using the Recommended Action feature or by examining the event log, collect the IBM Storage Virtualize support package, as described in 11.2.1, “IBM Storage Virtualize systems data collection” on page 710.

11.4 Remote Support Assistance

Remote Support Assistance (RSA) enables IBM Support to access an IBM Storage Virtualize systems device to perform troubleshooting and maintenance tasks. Support assistance can be configured to support personnel work onsite only, or to access the system both onsite and remotely. Both methods use secure connections to protect data in the communication between support center and system. Also, you can audit all actions that support personnel conduct on the system.

To set up the remote support options by using the GUI, select **Settings** → **Support** → **Support Assistance** → **Reconfigure Settings**, as shown in Figure 11-11.

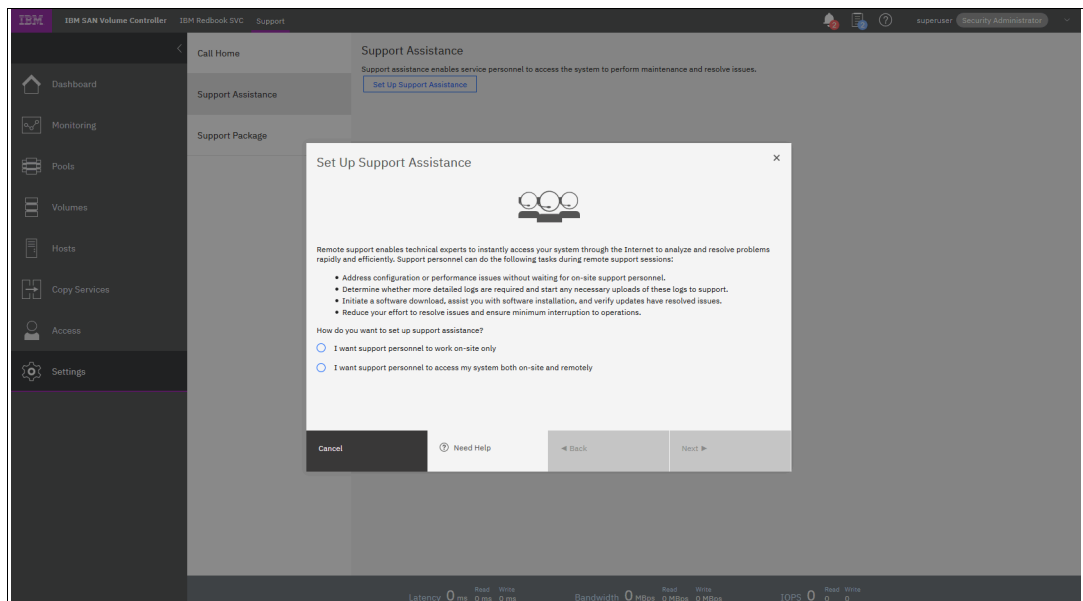


Figure 11-11 Remote Support options

You can use local support assistance if you have security restrictions that do not allow support to connect remotely to your systems. With RSA, support personnel can work onsite and remotely by using a secure connection from the support center.

They can perform troubleshooting, upload support packages, and download software to the system with your permission. When you configure RSA in the GUI, local support assistance also is enabled.

The following access types are in the RSA method:

- ▶ At any time
Support center can start remote support sessions at any time.
- ▶ By permission only
The support center can start a remote support session only if permitted by an administrator. A time limit can be configured for the session.

Note: Systems that are purchased with a 3-year warranty include enterprise-class support (ECS), and they are entitled to IBM Support by using RSA to quickly connect and diagnose problems. However, IBM Support might choose to use this feature on non-ECS systems at their discretion; therefore, we recommend configuring and testing the connection on all systems.

To configure RSA, the following prerequisites must be met:

- ▶ Cloud Call Home or a valid email server is configured. Cloud Call Home is used as the primary method to transfer the token when you initiate a session, with email as a backup.
- ▶ A valid service IP address is configured on each node on the system.
- ▶ A DNS server must be configured on your system.
- ▶ You can configure your firewall to allow traffic to pass directly from the system or you can route traffic through an HTTP proxy server within your environment.
- ▶ Uploading support packages and downloading software require direct connections to the internet. A DNS server must be defined on your system for both of these functions to work.
- ▶ If the storage nodes are directly connected to the internet, request your network administrator to allow connections to 170.225.126.11, 170.225.126.12, 170.225.127.11, and 170.225.127.12 on port 22 or 443.

For more information, see [Configuring support assistance](#).

11.5 Call Home Connect Cloud and Health Checker feature

Formerly known as *Call Home Web*, the new Call Home Connect Cloud is a cloud-based version with improved features to view Call Home information on the web.

Call Home is available in several IBM systems, including IBM Storage Virtualize systems, which allows them to automatically report problems and statuses to IBM.

In addition, Call Home Connect Cloud provides an app that is called *Call Home Connect Anywhere*, a mobile version to monitor your systems from anywhere.

The IBM Call Home Connect Anywhere mobile app is available on iOS and Android, and it provides a live view of your IBM assets, including cases, alerts, and support statuses. The mobile app, which is available within the Apple App Store and the Google Play Store, is a companion application to IBM Call Home Connect Cloud. If you do not already have assets that are registered, you are directed to IBM Call Home Connect Cloud to register assets to be viewed in the mobile app.

Call Home Connect Cloud provides the following information about IBM systems:

- ▶ Automated tickets
- ▶ Combined ticket view
- ▶ Warranty and contract status
- ▶ Health check alerts and recommendations
- ▶ System connectivity heartbeat
- ▶ Recommended software levels
- ▶ Inventory
- ▶ Security bulletins

IBM Call Home Connect Anywhere provides the following functions:

- ▶ An inventory-based user interface that provides a quick search of your assets.
- ▶ Live updates for your assets, ensuring that you always see the latest data.
- ▶ Case summaries for cases with IBM Support.
- ▶ Proactive alerts when important conditions are detected for your assets.
- ▶ IBM Call Home status and the last contact for your assets.

- ▶ Detailed information on warranties, maintenance contracts, service levels, and end of service information for each of your assets.
- ▶ Recommended software levels for each of your assets.

For more information about Call Home Connect Cloud (Call Home Web), see [IBM Support](#).

For more information about how to set up and use Call Home Connect Cloud, see [Introducing Call Home Connect Cloud](#).

11.5.1 Health Checker

A new feature of Call Home Connect Cloud is the Health Checker, which is a tool that runs in the IBM Cloud.

It analyzes Call Home and inventory data of systems that are registered in Call Home Connect Cloud and validates their configuration. Then, it displays alerts and provides recommendations in the Call Home Connect Cloud tool.

Note: Use Call Home Connect Cloud because it provides useful information about your systems. The Health Checker feature helps you to monitor the system, and operatively provides alerts and creates recommendations that are related to them.

Some of the functions of the IBM Call Home Connect Cloud and Health Checker were ported to IBM Storage Insights, as described in 11.6, “IBM Storage Insights” on page 748.

11.5.2 Configuring Call Home by using Ansible

Ansible Modules and Playbooks provide automation capabilities for configuring Call Home and RSA on IBM Storage Virtualize systems. Automation software allows for the creation of repeatable sets of instructions. It also reduces the need for human interaction with computer systems. For more details see *Do More with Less: Automating IBM Storage FlashSystem Tasks with REST APIs, Scripting, and Ansible*, REDP-5736.

11.6 IBM Storage Insights

IBM Storage Insights is an important part of monitoring and ensuring continued availability of IBM Storage Virtualize systems.

Available at no addition charge, cloud-based IBM Storage Insights provides a single dashboard that gives you a clear view of all your IBM block storage. You can make better decisions by seeing trends in performance and capacity. Storage health information enables you to focus on areas that need attention.

In addition, when IBM Support is needed, IBM Storage Insights simplifies uploading logs, speeds resolution with online configuration data, and provides an overview of open tickets all in one place.

The following features are included:

- ▶ A unified view of IBM systems, which provides:
 - A single window to see all your system’s characteristics.
 - A list of all your IBM storage inventory.

- A live event feed so that you know, up to the second, what is going on with your storage, which enables you to act fast.
- ▶ IBM Storage Insights collects telemetry data and Call Home data, and provides up-to-the-second system reporting of capacity and performance.
- ▶ Overall storage monitoring:
 - The overall health of the system.
 - A configuration to see whether it meets the best practices.
 - System resource management: Determine whether the system is being overly taxed, and provide proactive recommendations to fix it.
- ▶ IBM Storage Insights provides advanced customer service with an event filter that enables the following functions:
 - The ability for you and IBM Support to view, open, and close support tickets, and track trends.
 - An auto log collection capability enables you to collect logs and send them to IBM before support starts looking into the problem. This feature can save as much as 50% of the time to resolve the case.
- ▶ Virtual infrastructure monitoring:
 - Monitoring of VMware hosts and virtual machines (VMs).
 - Collect and view performance, capacity, configuration, and status metadata about the VMware ESXi hosts and VMs in your environment.
- ▶ Monitoring IBM Safeguarded Copy helps to:
 - Understand how much of the capacity of the storage systems and pools is being consumed by Safeguarded copies.
 - Verify that the correct volumes are being protected.
 - Generate reports to see how much of your capacity is protected.

In addition to the no additional charge IBM Storage Insights, you also can use IBM Storage Insights Pro. This service is a subscription service that provides longer historical views of data, offers more reporting and optimization options, and supports IBM file and block storage with EMC VNX and VMAX.

Figure 11-12 shows a comparison of IBM Storage Insights and IBM Storage Insights Pro.

Product Comparison			
	Capability	IBM Storage Insights (Free)	IBM Storage Insights Pro (Subscription)
Monitoring	Health, Performance and Capacity	✓	✓
	Filter events to quickly isolate trouble spots	✓	✓
	Drill down performance workflows to enable deep troubleshooting		✓
	Application / server storage performance troubleshooting		✓
	Customizable multi-conditional alerting		✓
Support Services	Simplified ticketing / log workflows and ticket history	✓	✓
	Proactive notification of risks (select systems)	✓	✓
Device Analytics	Part failure prediction	✓	✓
	Configuration best practice	✓	✓
	Customized upgrade recommendation	✓	✓
TCO Analytics	Capacity planning		✓
	Performance planning		✓
	Application / server storage consumption		✓
	Capacity optimization with reclamation planning		✓
	Data optimization with tier planning		✓

Figure 11-12 IBM Storage Insights versus IBM Storage Insights Pro

For more information regarding the features that are included in the available editions of IBM Storage Insights, see [IBM Storage Insights](#).

IBM Storage Insights provides a lightweight data collector that is deployed on a customer-supplied server. This server can be a Linux, Windows, or AIX server, or a guest in a VM (for example, a VMware guest).

For more information about the supported operating systems for a data collector, see the IBM Storage Insights documentation at [Managing data collectors](#).

The data collector streams performance, capacity, asset, and configuration metadata to your IBM Cloud instance.

The metadata flows in one direction: from your data center to IBM Cloud over HTTPS. In the IBM Cloud, your metadata is protected by physical, organizational, access, and security controls. IBM Storage Insights is ISO/IEC 27001 Information Security Management certified.

To make your data collection services more robust, install two or more data collectors on separate servers or VMs in each of your data centers.

When you add storage devices, the data collectors that you deploy are tested to see whether they can communicate with those devices. If multiple data collectors can communicate with a device, then the data collector with the best response time collects the metadata. If the collection of metadata is interrupted, the data collectors are tested again, and the data collectors with the best response times take over.

Figure 11-13 on page 751 shows an installation of multiple data collectors.

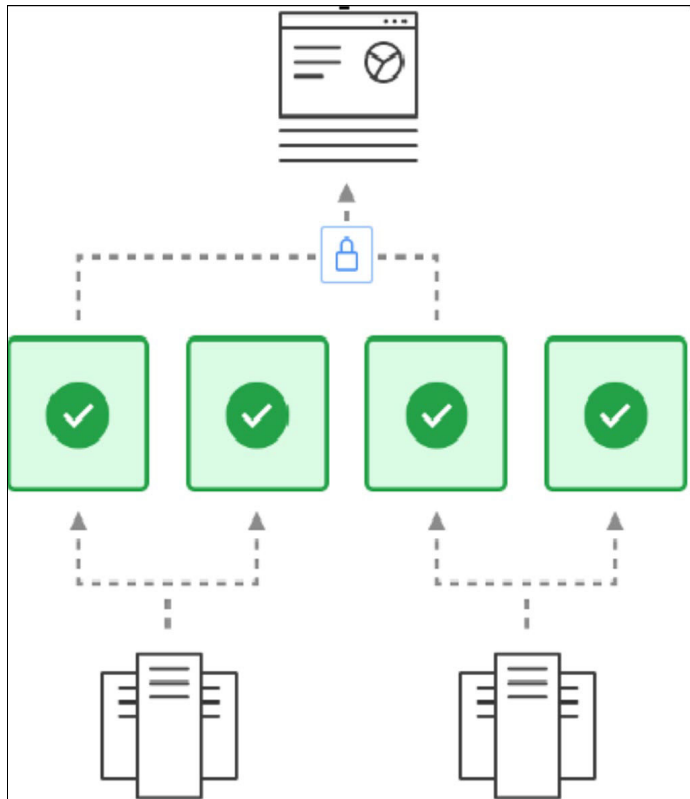


Figure 11-13 Illustration of multiple data collectors

Collected metadata

The following metadata about the configuration and operations of storage resources is collected:

- ▶ Name, model, firmware, and type of storage system.
- ▶ Inventory and configuration metadata for the storage system's resources, such as volumes, pools, disks, and ports.
- ▶ Capacity values, such as capacity, unassigned space, used space, and the CR.
- ▶ Performance metrics, such as read/write data rates, I/O rates, and response times.

The application data that is stored on the storage systems cannot be accessed by the data collector.

Accessing the metadata

Access to the metadata that is collected is restricted to the following users:

- ▶ The customer who owns the dashboard.
- ▶ The administrators who are authorized to access the dashboard, such as the customer's operations team.
- ▶ The IBM Cloud team that is responsible for the day-to-day operation and maintenance of IBM Cloud instances.
- ▶ IBM Support for investigating and closing service tickets.

11.6.1 IBM Storage Insights customer main dashboard

Figure 11-14 shows a view of the IBM Storage Insights main dashboard and the systems that it is monitoring.

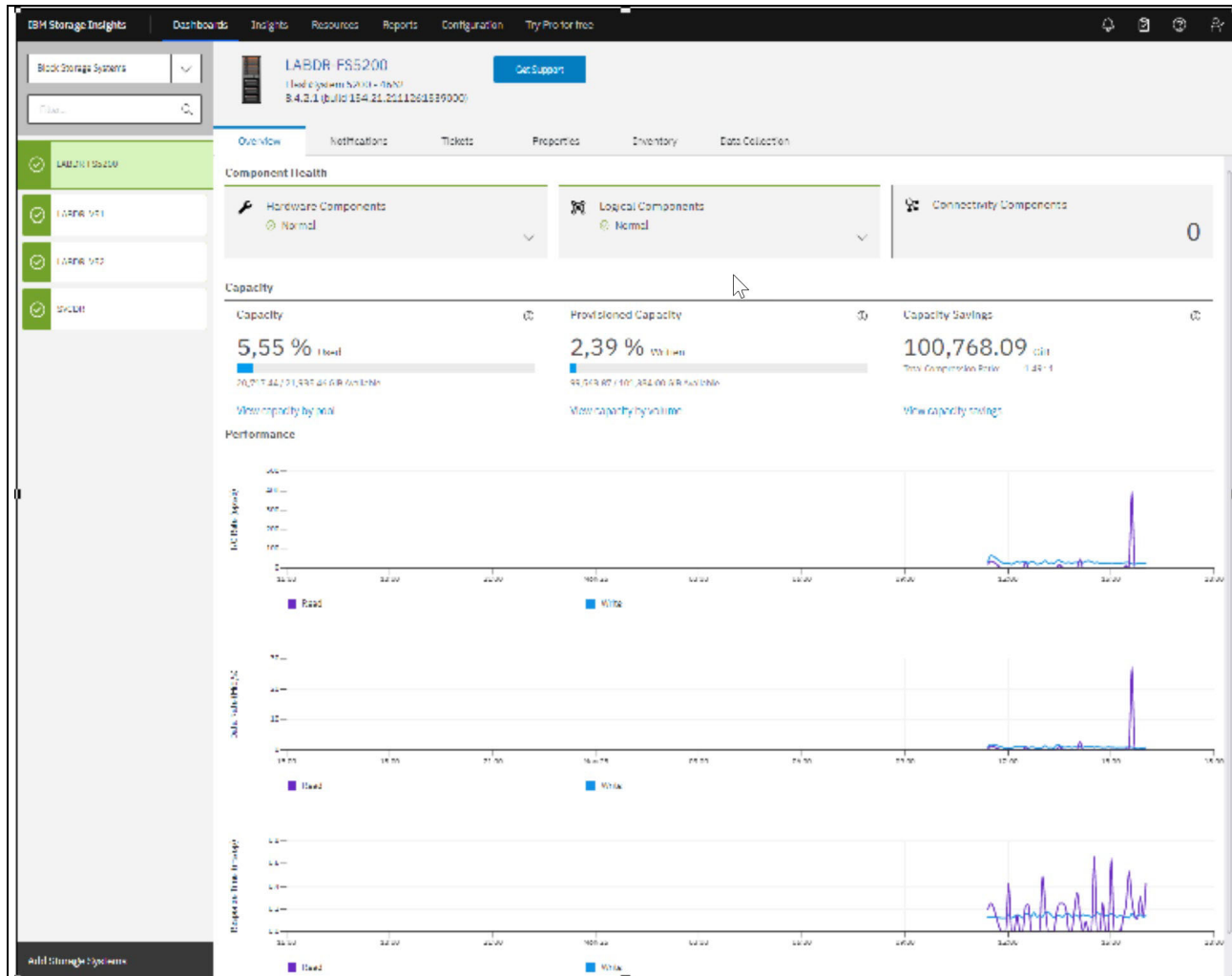


Figure 11-14 IBM Storage Insights main dashboard

11.6.2 Customized dashboards to monitor your storage

With the latest release of IBM Storage Insights, you can customize the dashboard to show only a subset of the systems that are monitored. This feature is useful for customers that might be cloud service providers (CSPs) and want only a specific user to see those machines for which they are paying.

For more information about setting up the customized dashboard, see [Creating customized dashboards to monitor your storage](#).

11.6.3 Creating a support ticket

The IBM Storage Insights dashboard GUI can be used to create a support ticket for any of the systems that IBM Storage Insights generates reports.

Complete the following steps:

1. In the IBM Storage Insights dashboard, choose the system for which you want to create the ticket. Then, select **Get Support** (see Figure 11-15).

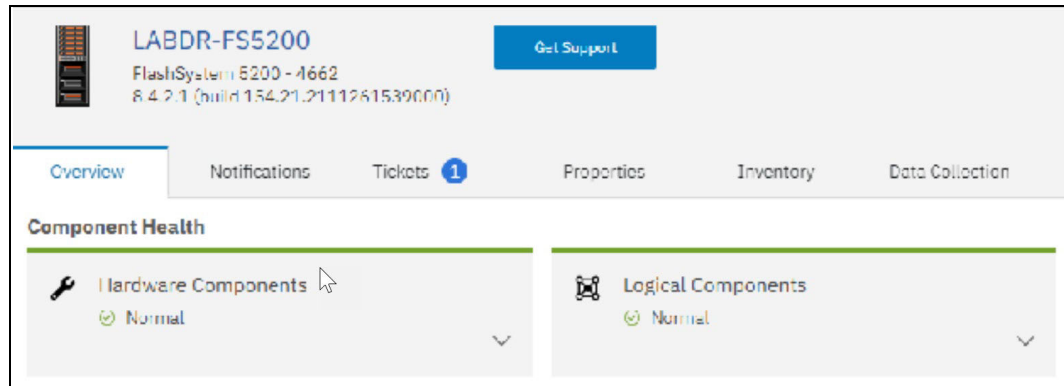


Figure 11-15 IBM Storage Insights Get Support option

2. Select **Create Ticket** (see Figure 11-16). Several windows open in which you enter information about the machine, a problem description, and the option to upload logs.

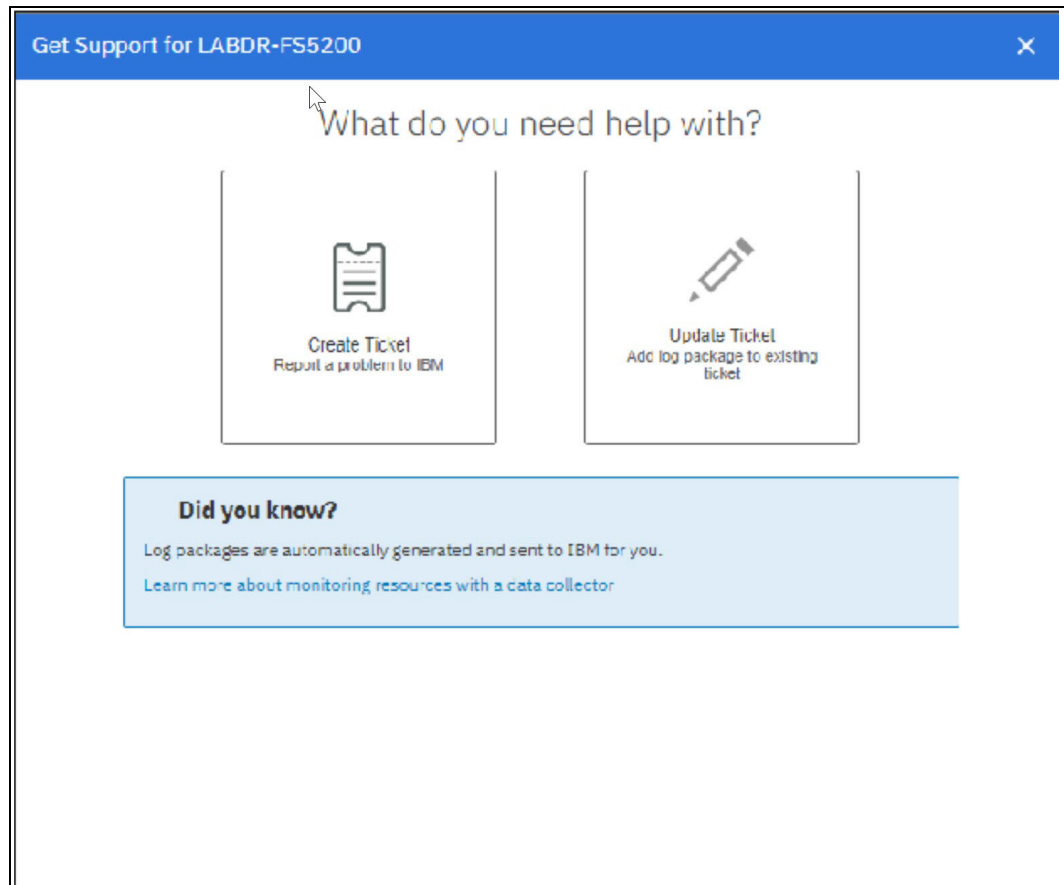


Figure 11-16 Create Ticket option

Note: The “Permission given” information box (see Figure 11-17 on page 754) is an option that the customer must enable in the IBM Storage Virtualize systems GUI. For more information, see 11.4, “Remote Support Assistance” on page 746.

Figure 11-17 shows the ticket data collection that is done by the IBM Storage Insights application.

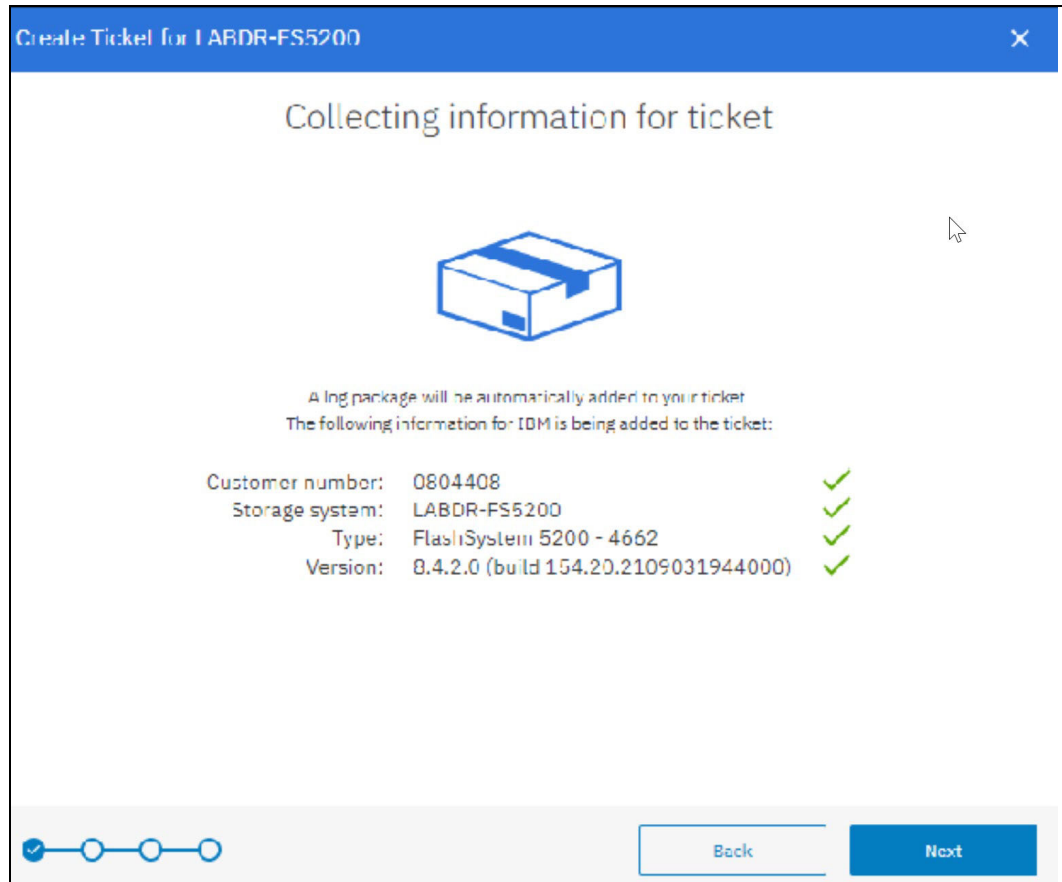


Figure 11-17 Collecting ticket information

As shown in Figure 11-18, you can add a problem description and attach other files to support the ticket, such as error logs or window captures of error messages.

The screenshot shows a web interface for creating a ticket. The title bar reads "Create Ticket for LABDR-FS5200". The main heading is "Add a note or attachment". There are two text input fields. The first field contains the text "Device is showing a failure alert" and has a character count of "39". Below it is a hint: "Hint: Include what happened and the error code, if any." The second field is empty and has a character count of "3000". Below it is a hint: "Hint: Include the time the problem or error occurred, the affected resources, and details of any maintenance or other activities that occurred before the problem." There are two options for attaching files: a "Browse" button under the heading "Attach Image or File:" and a dashed box with an upward arrow and the text "Drag file here". At the bottom left, there is a progress indicator with four circles, the second of which is checked. At the bottom right, there are "Back" and "Next" buttons.

Figure 11-18 Adding a problem description and any other information

3. You are prompted to set a severity level for the ticket, as shown in Figure 11-19. Severity levels range from Severity 1 (for a system that is down or extreme business impact) to Severity 4 (noncritical issue).

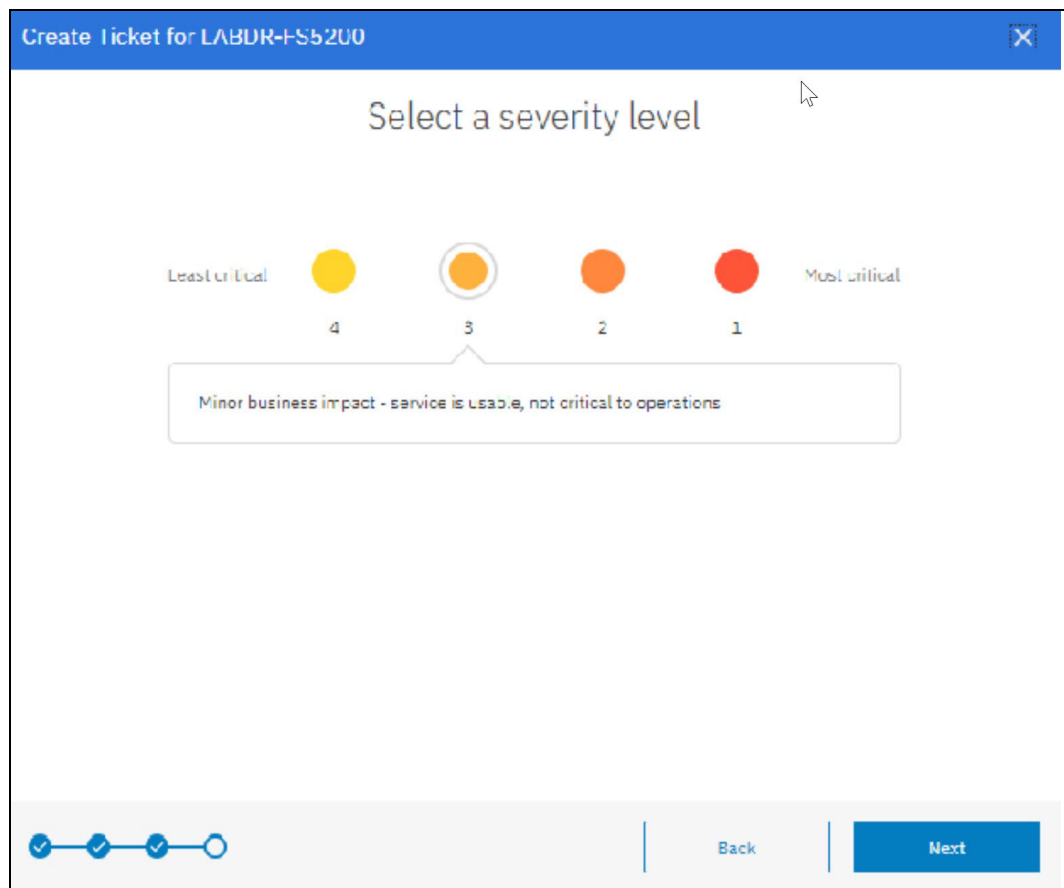


Figure 11-19 Setting the severity level

- The final summary window (see Figure 11-20) includes the option to add logs to the ticket. When completed, click **Create Ticket** to create the support ticket and send it to IBM. The ticket number is created by the IBM Support system and returned to your IBM Storage Insights instance.

Figure 11-20 Final summary before ticket creation

- Figure 11-21 shows how to view the summary of the open and closed ticket numbers for the system that is selected by using the **Action** menu option.

Figure 11-21 Support ticket summary

11.6.4 Updating a support ticket

After a support ticket is created, the IBM Storage Insights dashboard GUI can be used update tickets.

Complete the following steps:

1. In IBM Storage Insights dashboard, select **Update Ticket** (see Figure 11-22).

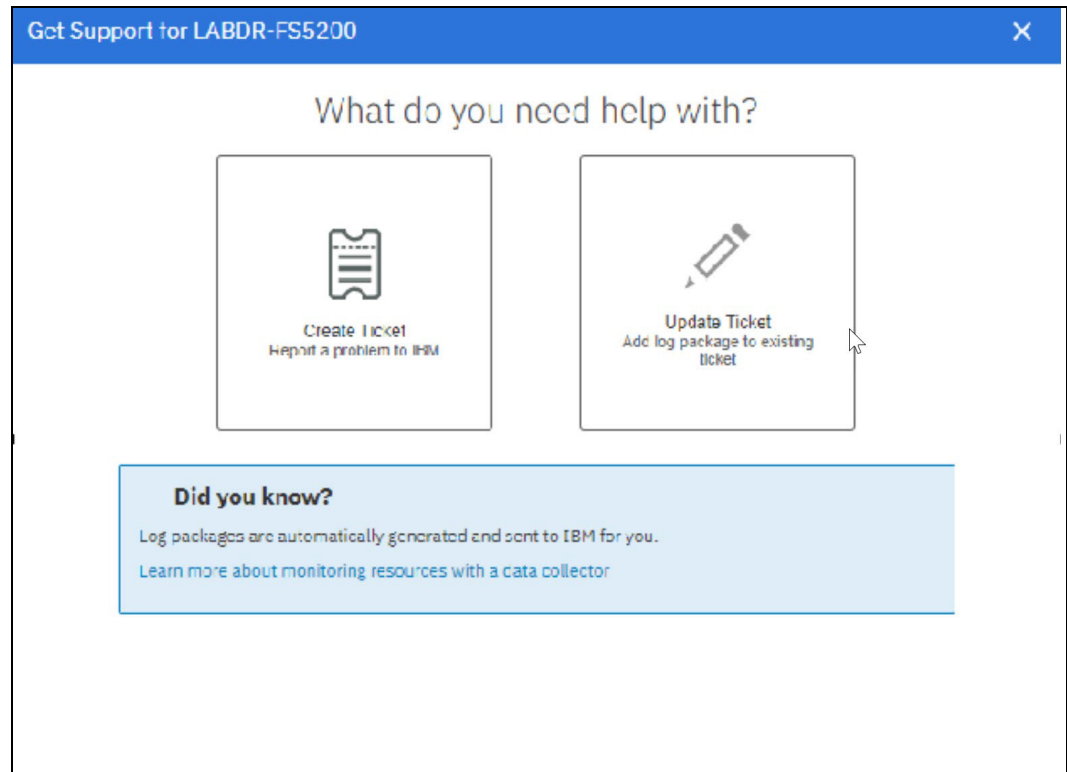


Figure 11-22 IBM Storage Insights Update Ticket

2. Enter the IBM Support case number, and then click **Next** (see Figure 11-23).
The IBM Support case number uses the following format:

TS000XXXXX

These details were supplied when you created the ticket or by IBM Support if the Problem Management Record (PMR) was created by a problem Call Home event (assuming that Call Home is enabled).

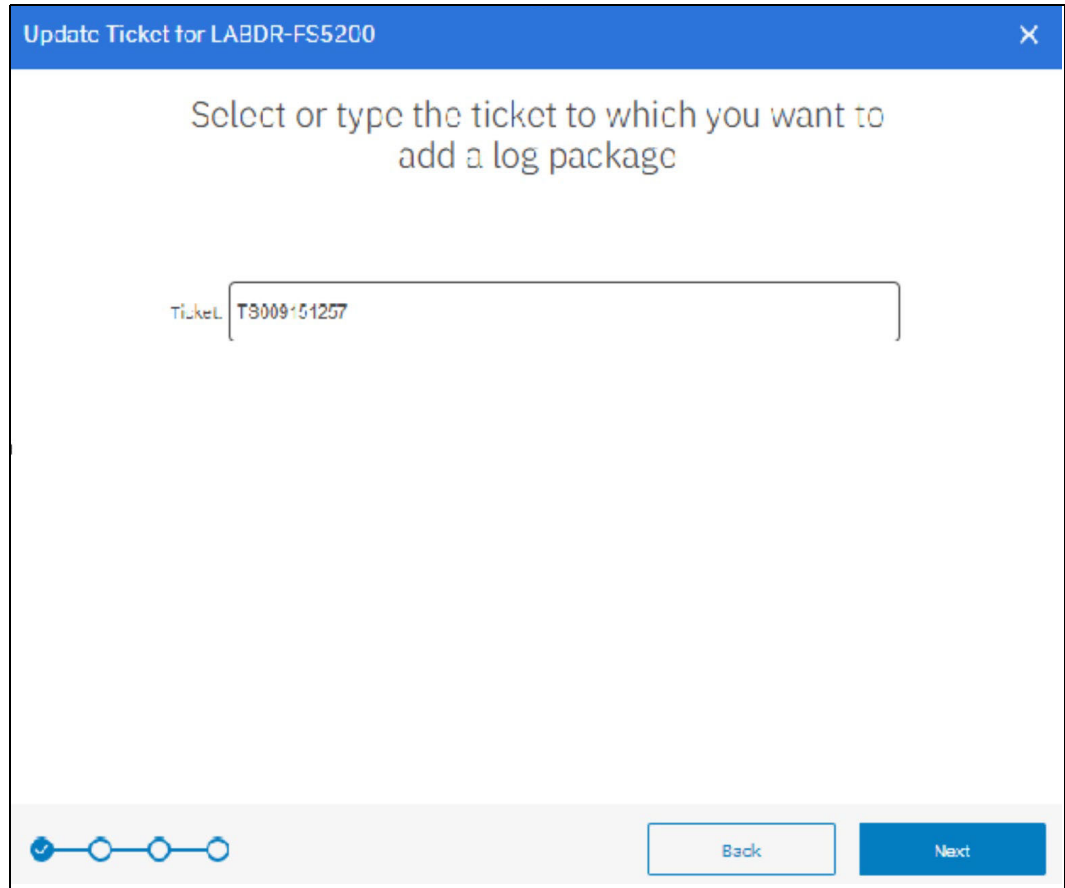


Figure 11-23 Entering the IBM Support or PMR case number

A window opens in which you can choose the log type to upload. The window and the available options are shown in Figure 11-24 on page 760.

The following options are available:

- Type 1 - Standard logs.
For general problems, including simple hardware and simple performance problems.
- Type 2 - Standard logs and the most recent statesave log.
- Type 3 - Standard logs and the most recent statesave log from each node.
For 1195 and 1196 node errors and 2030 software restart errors.
- Type 4 - Standard logs and new statesave logs.
For complex performance problems, and problems with interoperability of hosts or storage systems, compressed volumes, and remote copy operations, including 1920 errors.

You can allow IBM Support to collect and upload packages from your storage systems without requiring permission from your organization. If you grant this permission, it can help IBM Support resolve your support tickets faster.

If IBM Support does not have that permission, a notification appears within the window.

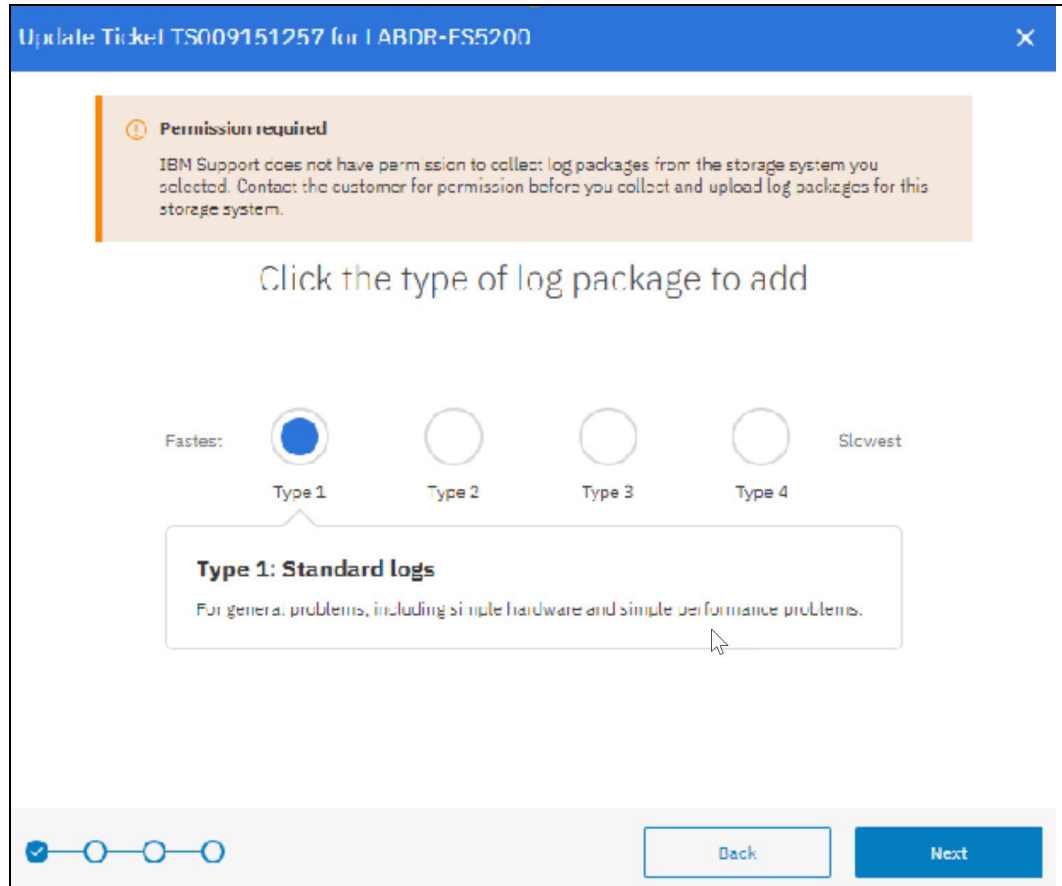


Figure 11-24 Log type selection

If you are unsure about which log type to upload, contact IBM Support for guidance. The most common type to use is type 1, which is the default type. The other types are more detailed logs and for issues in order of complexity.

3. After the type of log is selected, click **Next**. The log collection starts. When completed, the log completion window is displayed, as shown in Figure 11-25.

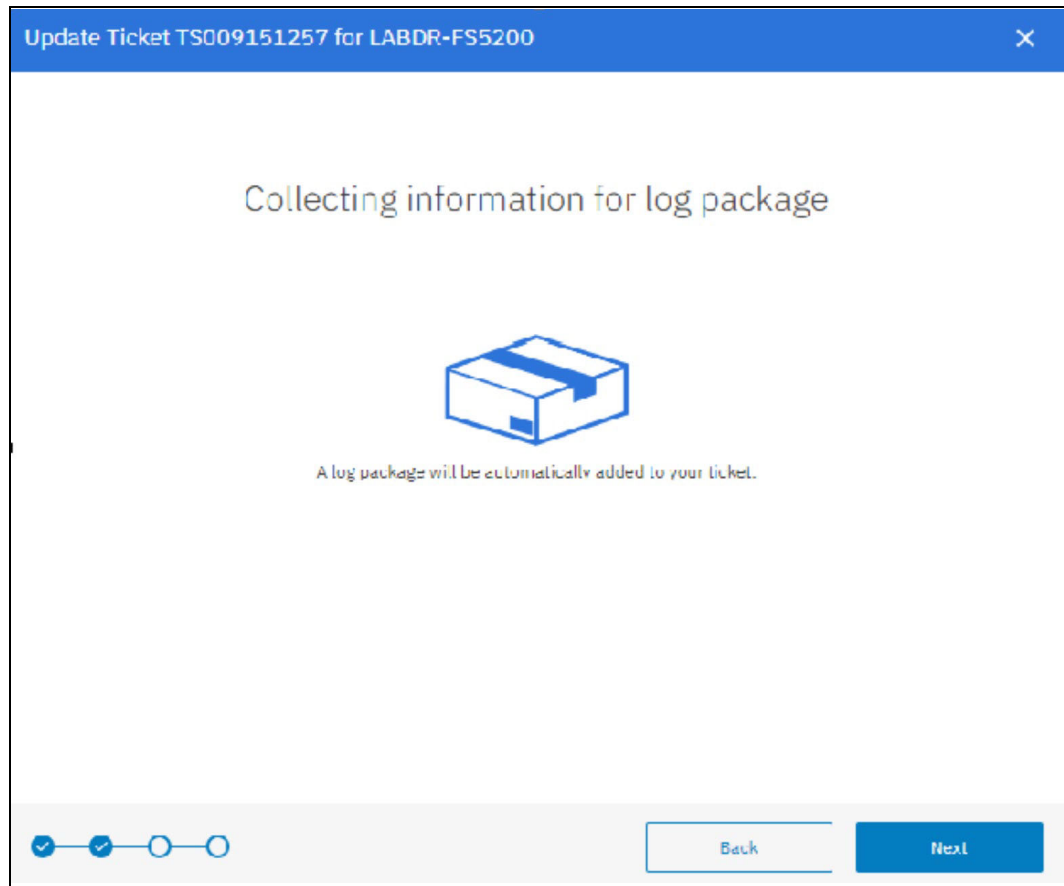


Figure 11-25 Collecting new logs

4. After clicking **Next**, it is possible to provide more information in a text field, or upload more files, as shown in Figure 11-26.

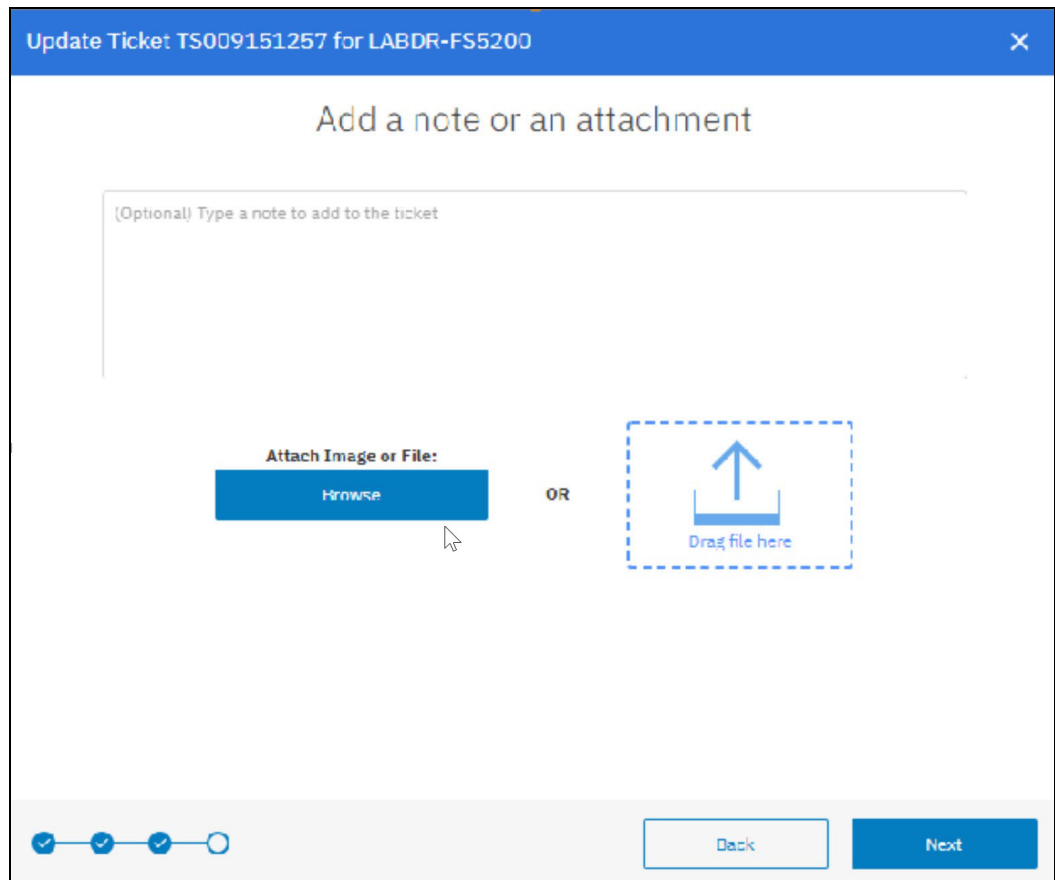


Figure 11-26 Add a note or attachment

5. By clicking **Next**, the update of the support ticket is completed, as shown in Figure 11-27.

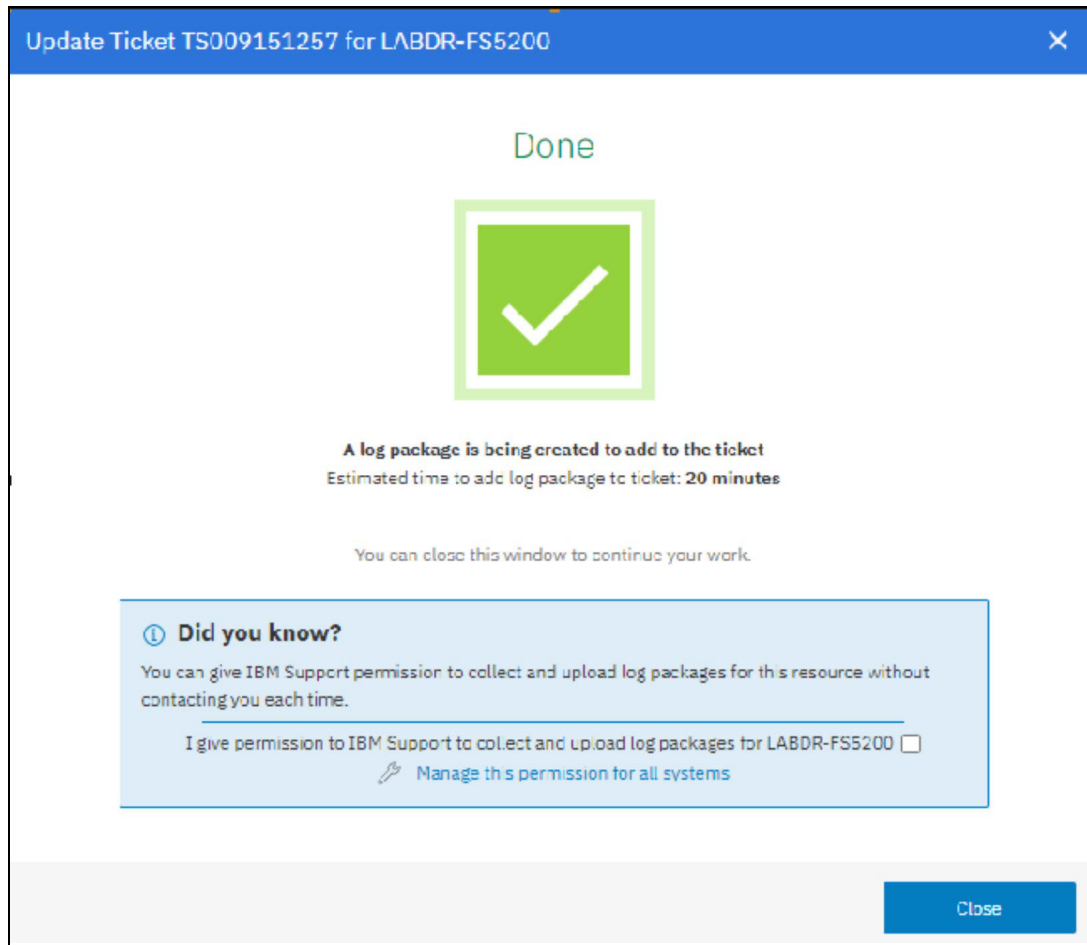


Figure 11-27 Update of the support ticket is complete

11.6.5 IBM Storage Insights Advisor

IBM Storage Insights continually evolves, and its latest addition is a new option from the action menu that is called *Advisor*.

IBM Storage Insights analyzes your device data to identify violations of best practice guidelines and other risks, and to provide recommendations about how to address these potential problems.

To view these recommendations, select **Insights** → **Advisor**. To see more information about a recommendation or to acknowledge it, double-click the recommendation.

All advice that is categorized as Error, Warning, Informational, or Acknowledged is shown for all attached storage within a table. Using a filter, it is possible, for example, to display only advisories for a specific IBM Storage Virtualize system.

Figure 11-28 on page 764 shows the initial IBM Storage Insights Advisor menu.

Advisor

Keep your infrastructure healthy with these recommendations.

Unacknowledged Recommendations: 24 24 Informational | 25 Acknowledged

Actions Acknowledge

Event	Severity	Time	Device Name	More Information
Hosts have more than the recommended number of paths to volumes	Informational	Dec 13, 2021, 21:58:54	v7000-ctr-60	Based on our general experience, it is best to limit the total number of paths from...
Consistency Protection can be configured for Global Mirror and Metro Mirror relationships	Informational	Dec 13, 2021, 21:58:54	v7000-ctr-60	IBM Spectrum Virtualize V7.8.1 introduced the Remote Copy Consistency Prote...
The system has not been configured to use Network Time Protocol (NTP)	Informational	Dec 17, 2021, 13:54:40	V7k-PFE5	The system is not currently configured to use NTP to keep the system clock sho...
Performance statistics configuration outside of recommended parameters	Informational	Dec 17, 2021, 13:54:40	V7k-PFE5	Collection of performance statistics is either disabled or the frequency is set to ...
Defined hosts show as offline status.	Informational	Dec 17, 2021, 13:58:24	V7k-PFE5	There are one or most hosts which are configured in the system that show an of...
Unused volumes identified in the system	Informational	Dec 17, 2021, 13:58:24	V7k-PFE5	Volumes were found in storage pools that appear to be unused. These may hav...
Consider enabling volume protection	Informational	Dec 17, 2021, 13:58:24	V7k-PFE5	To prevent an active volume from being deleted unintentionally, administrators ...
Preferred node optimization for FlashCopies	Informational	Jan 6, 2022, 10:10:29	V7k-PFE4	FlashCopy mappings were identified that have source and target volumes in the...
Unused volumes identified in the system	Informational	Jan 6, 2022, 10:10:29	V7k-PFE4	Volumes were found in storage pools that appear to be unused. These may hav...
The system has not been configured to use Network Time Protocol (NTP)	Informational	Jan 12, 2022, 10:16:43	SVC-PFE1	The system is not currently configured to use NTP to keep the system clock sho...
Performance statistics configuration outside of recommended parameters	Informational	Jan 12, 2022, 10:16:43	SVC-PFE1	Collection of performance statistics is either disabled or the frequency is set to ...
Preferred node optimization for FlashCopies	Informational	Jan 27, 2022, 21:36:42	F59200-PF...	FlashCopy mappings were identified that have source and target volumes in the...
Software Update Recommendation	Informational	Feb 1, 2022, 06:08:56	V3700_PFE1	
Software Update Recommendation	Informational	Feb 1, 2022, 06:09:35	V3700_PFE1	
Software Update Recommendation	Informational	Feb 3, 2022, 06:07:26	V5030F_PF...	
Software Update Recommendation	Informational	Feb 3, 2022, 06:07:51	V5030F_PF...	
Software Update Recommendation	Informational	Feb 4, 2022, 06:10:07	V3700_PFE1	
Enable Configuration Reporting for automated system analysis	Informational	Mar 15, 2022, 21:00:35	F59200-PF...	IBM systems automatically check your configuration for known issues and best ...

Figure 11-28 IBM Storage Insights Advisor menu

Figure 11-29 shows an example of the detailed IBM Storage Insights Advisor recommendations.

Advisor

Keep your infrastructure healthy with these recommendations.

Unacknowledged Recommendations: 7 7 Informational | 3 Acknowledged

Actions Acknowledge

Filter...

Event	Severity	Time
Software Update Recommendation	Informational	Mar 9,
Software Update Recommendation	Informational	Mar 5,
Software Update Recommendation	Informational	Mar 5,
Software Update Recommendation	Informational	Mar 4,
Software Update Recommendation	Informational	Mar 3,
Software Update Recommendation	Informational	Mar 3,
Software Update Recommendation	Informational	Mar 3,
Running out of space (10% or less remaining)	Warning	Mar 2,
Running out of space (5% or less remaining)	Warning	Jan 14
Running out of space (5% or less remaining)	Warning	Jan 9,

Running out of space (5% or less remaining) X

Jan 14, 2019, 05:24:02

[Acknowledge](#) [Unacknowledge](#)

The total system capacity has less than 5% available space.

Percent Free Space	Total Free Space
0.05280	7256776743321

Consider adding more capacity to the system, or cleaning up unneeded volumes.

Figure 11-29 Advisor detailed summary of recommendations

As shown in Figure 11-29, the details of a Running out of space recommendation is shown the Advisor page. In this scenario, the user clicked the **Warning** tag to focus only on the recommendations that feature a severity of Warning.

For more information about setting and configuring the Advisor options, see [Monitoring recommended actions in the Advisor](#).



IBM i considerations

The IBM Storage Virtualize family of block storage systems, including IBM SAN Volume Controller (SVC), IBM FlashSystem 5000 series, IBM FlashSystem 7200 and 7300, and IBM FlashSystem 9200, 9200R, 9500, and 9500R, provides a broad range of flexible and scalable storage area network (SAN) storage solutions.

These solutions can meet demands of IBM i customers for entry to high-end storage infrastructure solutions.

All family members that are based on IBM Storage Virtualize software use a common management interface. They also provide a comprehensive set of advanced functions and technologies, such as advanced Copy Services functions, encryption, compression, storage tiering, Non-Volatile Memory Express (NVMe) flash, storage-class memory (SCM) devices, and external storage virtualization. Many of these advanced functions and technologies also are of interest to IBM i customers who are looking for a flexible, high-performing, and highly available (HA) SAN storage solution.

This appendix provides important considerations and guidelines for successfully implementing the IBM Storage Virtualize family and its advanced functions with IBM i.

Unless otherwise stated, the considerations also apply to previous generations of products, such as the IBM Storwize family, the IBM FlashSystem 9100 series, and IBM FlashSystem V9000.

Note: For the most recent IBM i functional enhancements, see [IBM i Functional Enhancements Summary](#).

This appendix includes the following topics:

- ▶ “IBM i Storage management” on page 767
- ▶ “Single-level storage” on page 768
- ▶ “IBM i response time” on page 770
- ▶ “Planning for IBM i storage capacity” on page 774
- ▶ “Storage connection to IBM i” on page 775
- ▶ “Setting attributes in VIOS” on page 779
- ▶ “Disk drives for IBM i” on page 781

- ▶ “Defining LUNs for IBM i” on page 783
- ▶ “Data layout” on page 785
- ▶ “Fibre Channel adapters in IBM i and VIOS” on page 786
- ▶ “Zoning SAN switches” on page 786
- ▶ “IBM i multipath” on page 787
- ▶ “Booting from SAN” on page 788
- ▶ “IBM i mirroring” on page 788
- ▶ “Copy Services considerations” on page 788
- ▶ “SAN Volume Controller stretched cluster” on page 797
- ▶ “Db2 mirroring for IBM i” on page 801

IBM i Storage management

Because of the unique IBM i storage architecture, special considerations for planning and implementing a SAN storage solution are required (also with IBM Storage Virtualize-based storage). This section describes how IBM i storage management manages its available disk storage.

Many host systems require the user to take responsibility for how information is stored and retrieved from the disk units. An administrator also must manage the environment to balance disk usage, enable disk protection, and maintain balanced data to be spread for optimum performance.

The IBM i architecture is different in the way that the system takes over many of the storage management functions, which are the responsibility of a system administrator on other platforms.

IBM i, with its Technology Independent Machine Interface (TIMI), largely abstracts the underlying hardware layer from the IBM i operating system and its users and manages its system and user data in IBM i disk pools, which are also called *auxiliary storage pools* (ASPs).

When you create a file, you do not assign it to a storage location. Instead, the IBM i system places the file in the location that ensures the best performance from an IBM i perspective (see Figure A-1).

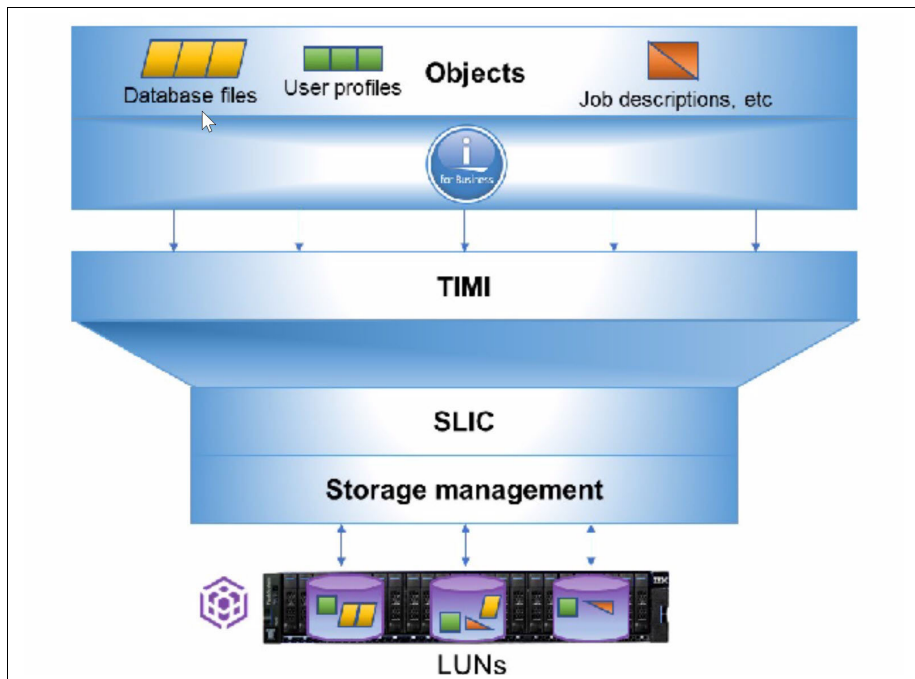


Figure A-1 IBM i storage management spreads objects across logical unit numbers

Note: When a program presents instructions to the machine interface for execution, the interface appears to the program as the system hardware, but it is not. The instructions that are presented to TIMI pass through a layer of microcode before they are understood by the hardware. Therefore, TIMI and System Licensed Internal Code (SLIC) allow IBM Power with IBM i to take technology in stride.

As a component of the IBM i SLIC, IBM i storage management normally spreads the data in the file across multiple disk units (logical unit numbers (LUNs) when external storage is used). When you add records to the file, the system automatically assigns more space on one or more disk units or LUNs.

Single-level storage

IBM i uses a single-level storage, object-orientated architecture. It sees all disk space and the main memory or main storage as one address space. It also uses the same set of virtual addresses to cover main memory and disk space. Paging the objects in this virtual address space is performed in 4 KB pages, as shown in Figure A-2. After a page is written to disk, it is stored with metadata, including its unique virtual address. For this purpose, IBM i originally used a proprietary 520 bytes per sector disk format.

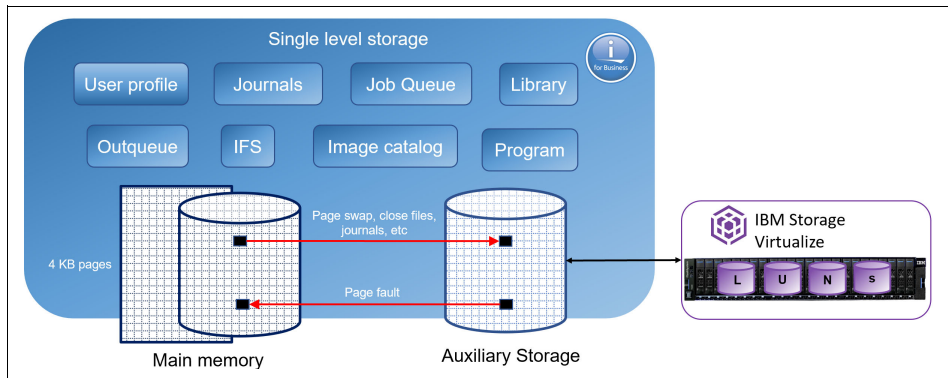


Figure A-2 Virtual address space

Note: The system storage that is conformed with main storage or main memory and auxiliary storage is addressed in the same way. This single, device-independent addressing mechanism means that objects are referred to by name or name and library, and *never* by disk location. The virtual addressing of IBM i is independent of the physical location of the object, type, capacity, and the number of disks units or LUNs on the system.

The IBM i disk storage space is managed by using ASPs. Each IBM i system has a system ASP (ASP 1), which includes the load source (also known as *boot volume* on other systems) as disk unit 1, and optional user ASPs (ASPs 2 - 33). The system ASP and the user ASPs are designated as SYSBAS, and they constitute the system database.

The single-level storage with its unique virtual addresses also implies that the disk storage that is configured in SYSBAS of an IBM i system must be available in its entirety for the system to remain operational. It cannot be shared for simultaneous access by other IBM i systems.

To allow for sharing of IBM i disk storage space between multiple IBM i systems in a cluster, switchable independent auxiliary storage pools (IASPs) can be configured. The IBM i ASPs' architecture is shown in Figure A-3 on page 769.

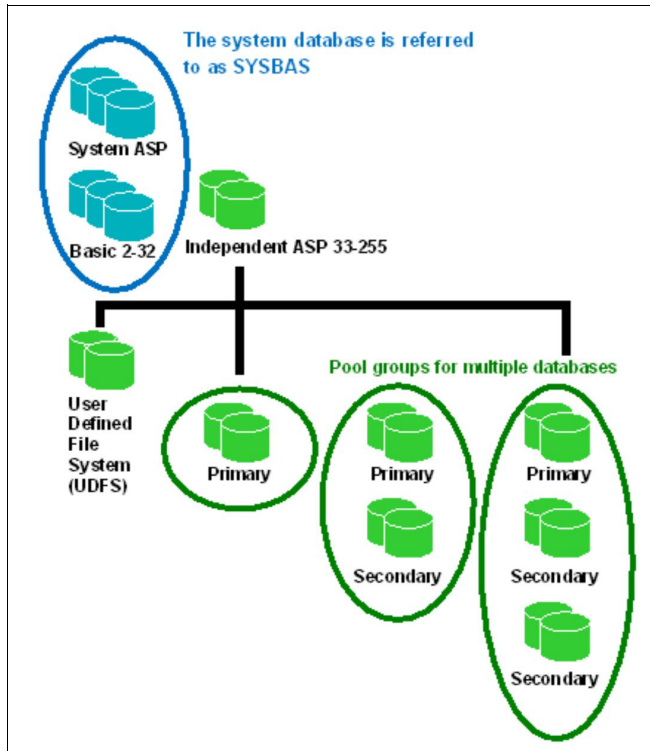


Figure A-3 IBM i auxiliary storage pools architecture

Single-level storage makes main memory work as a large cache. Reads are done from pages in main memory, and requests to disk are done only when the needed page is not there yet.

Writes are done to main memory or main storage, and write operations to disk are performed as a result of swap, file close, or forced write. Application response time depends on disk response time and many other factors.

Other storage-related factors include the IBM i storage pool configuration for the application, how frequently the application closes files, and whether it uses journaling. An example is shown in Figure A-4.

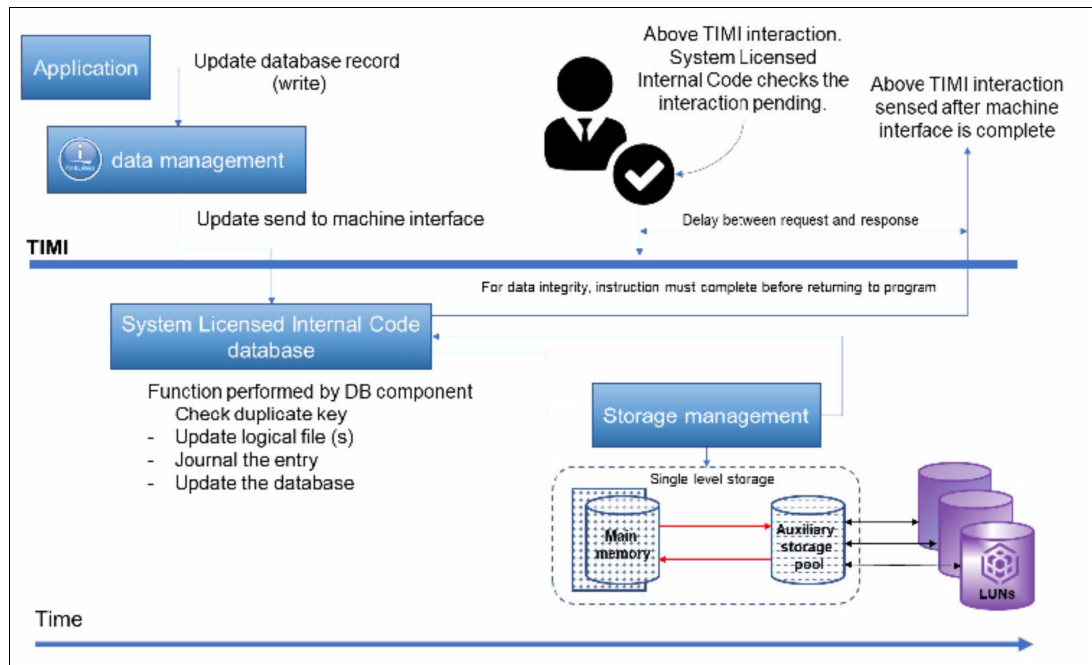


Figure A-4 TIMI atomicity

Note: Figure A-4 illustrates the ASP, comprised of assigned LUNs from IBM Storage Virtualize to the IBM i system. It depicts an application request and the subsequent update to a database record. While the TIMI task is in progress, interactions above TIMI can occur, but they will not proceed until the TIMI task completes.

IBM i response time

IBM i customers often are concerned about the following types of performance:

- ▶ **Application response time:** The response time of an application transaction. This time often is critical for the customer.
- ▶ **Duration of batch jobs:** Batch jobs often are run during the night or other off-peak periods. The duration of a batch job is critical for the customer because it must be finished before regular daily transactions start.
- ▶ **Disk response time:** Disk response time is the time that is needed for a disk I/O operation to complete. It includes the service time for I/O processing and the wait time for potential I/O queuing on the IBM i host.

Disk response time can influence application response time and the duration of a batch job. Because the performance of the disk subsystem affects overall system performance, this issue is described in “Disk performance considerations” on page 771.

Disk performance considerations

Disk subsystem performance affects overall IBM i system performance, especially in a commercial data processing environment where a large volume of data often must be processed. Disk drives or the LUNs' response times contribute to a major portion of the overall response time (online transaction processing (OLTP)) or run time (batch).

Also, disk subsystem performance is affected by the type of protection (redundant array of independent disks (RAID), distributed RAID (DRAID), or mirroring).

The amount of free space (GB) on the drives and the extent of fragmentation also has an effect. The reason is the need to find suitable contiguous space on the disks to create objects or extend objects. Disk space often is allocated in extents of 32 KB. If a 32 KB contiguous extent is not available, two extents of 16 KB are used.

The following disk performance considerations are described in the following sections:

- ▶ Disk I/O requests
- ▶ Disk subsystems
- ▶ Disk operation
- ▶ Asynchronous I/O wait
- ▶ Disk protection
- ▶ Logical database I/O versus physical disk I/O

Disk I/O requests

Many disk requests often occur if a request for information cannot be satisfied by what is in memory. Requests to bring information into memory also result in disk I/O. Memory pages also can be purged periodically, which results in disk I/O activity.

Note: The Set Object Access (**SETOBJACC**) command on IBM i temporarily changes the speed of access to an object by bringing the object into a main storage pool or purging it from all main storage pools. An object can be kept in main storage by selecting a pool for the object that has available space and does not have jobs that are associated with it.

For more information, see [Set Object Access \(SETOBJACC\)](#).

Disk subsystems

Typically, an external disk subsystem (storage system) connects a server through a SAN, as shown in Figure A-5.

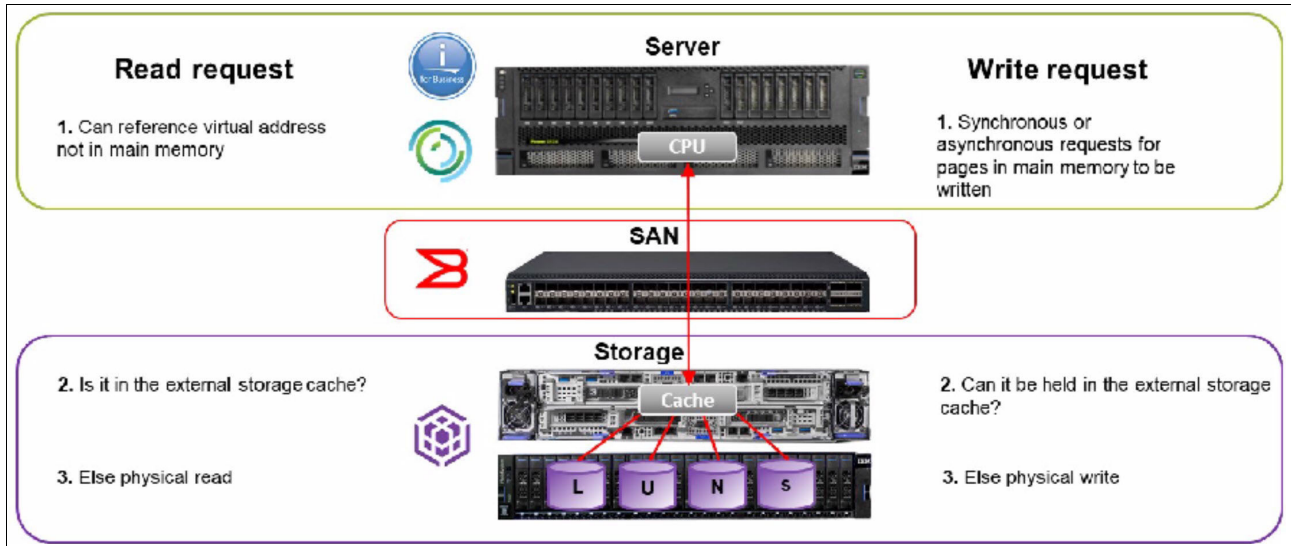


Figure A-5 Disk subsystem

An information request (data or instructions) from the CPU that is based on user interactions is submitted to the disk subsystem if it cannot be satisfied from the contents of main memory. If the request can be satisfied from the disk subsystem cache, it responds or forwards the request to the disk drives or LUNs.

Similarly, a write request is retained in memory unless the operating system determines that it must be written to the disk subsystem. Then, the operating system attempts to satisfy the request by writing to the controller cache.

Note: The QAPMDISKRB from the collections services data files in IBM i includes disk file response bucket entries. It also contains one record for each device resource name. It is intended to be used with the QAPMDISK file.

For more information, see [Collection Services data files: QAPMDISKRB](#).

Disk operation

On IBM i, physical disk I/O requests are categorized as database (physical or logical files) or non-database I/Os, as shown in Figure A-6.

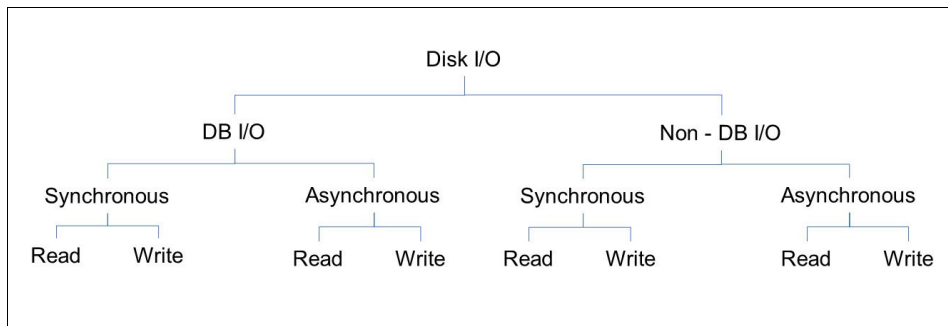


Figure A-6 Disk I/O on IBM i

The time that is taken to respond to synchronous disk I/Os contributes to the OLTP response time or batch run time. With asynchronous I/O, the progress of a request does not wait for the completion of I/O.

Often, write requests are asynchronous, including journal deposits with commitment control. However, the writes become synchronous if journaling is active without commitment control.

Asynchronous I/O wait

On IBM i, jobs might have to wait at times for asynchronous I/O requests to complete. The job issues a request but requires the data sooner than it can be made available by the disk subsystem. When a job waits for asynchronous I/O requests to complete, the I/O portion of the operation becomes synchronous. The time is recorded as asynchronous disk I/O wait in the QAPMJOBL file.

JBWIO is the number of times that the process waited for outstanding asynchronous I/O operations to complete. For more information, see [Collection Services data files: QAPMJOBS and QAPMJOBL](#).

This issue might be caused by faster processors that are running with relatively poor disk subsystems performance. Disk subsystem performance can be affected by busy or slow disks or small I/O cache.

Disk protection

For more information about external storage considerations for setting up your RAID protection, see Chapter 4, “Storage pools” on page 243.

Note: If you need high I/O performance on your IBM i workload, an option is to create a DRAID 1 on your supported storage system, such as IBM FlashSystem 7200 or 9200 with IBM Storage Virtualize 8.4 or later. In this configuration, the rebuild area is distributed over all member drives. The minimum extent size for this type of DRAID is 1024 MB.

Logical database I/O versus physical disk I/O

Information in partition buffer memory is available for use by any job or thread. Commonly, information is available in the partition buffer as a block of data rather than individual records. Data in a job buffer is available for use by the job only.

When an application program requests data, storage management checks whether the data is available in memory. If so, the data is moved to the open data path in the job buffer. If the data is not in memory, the request is submitted to the disk subsystem as a read command.

In that context, logical database I/O information is moved between the open data path of the user program and the partition buffer. This information is a count of the number of buffer movements, and not a reflection of the records that are processed.

For more information, see:

- ▶ [Sharing an Open Data Path](#)
- ▶ [Monitoring the metrics](#)

Physical disk I/O occurs when information is read or written as a block of data to or from the disk. It involves the movement of data between the disk and the partition buffer in memory. For more information, see [IBM i 7.5: Performance](#).

Planning for IBM i storage capacity

To correctly plan the storage capacity that is provided by IBM Storage Virtualize family systems for IBM i, you must be aware of IBM i block translation for external storage that is formatted in 512-byte blocks. IBM i internal disks use a block size of 520 or 4160 bytes.

IBM Storage Virtualize storage for hosts is formatted with a block size of 512 bytes; therefore, a translation or mapping is required to attach it to IBM i. IBM i changes the data layout to support 512-byte blocks (sectors) in external storage by using an extra ninth sector to store the headers for every page.

The eight 8-byte headers from each 520-byte sectors of a page are stored in the ninth sector, which is different than 520-byte sector storage where the 8 bytes are stored continuous with the 512 bytes of data to form the 520-byte sector.

The data that was stored in eight sectors is now stored by using nine sectors, so the required disk capacity on IBM Storage Virtualize based systems is 8/9ths of the IBM i usable capacity. Similarly, the usable capacity in IBM i is 8/9ths of the allocated capacity in these storage systems.

When attaching IBM Storage Virtualize family storage to IBM i, plan for extra capacity on the storage system so that the 8/9ths of the effective storage capacity that is available to IBM i covers the capacity requirements for the IBM i workload.

The performance impact of block translation in IBM i is small or negligible.

Figure A-7 shows the byte sectors for IBM i.

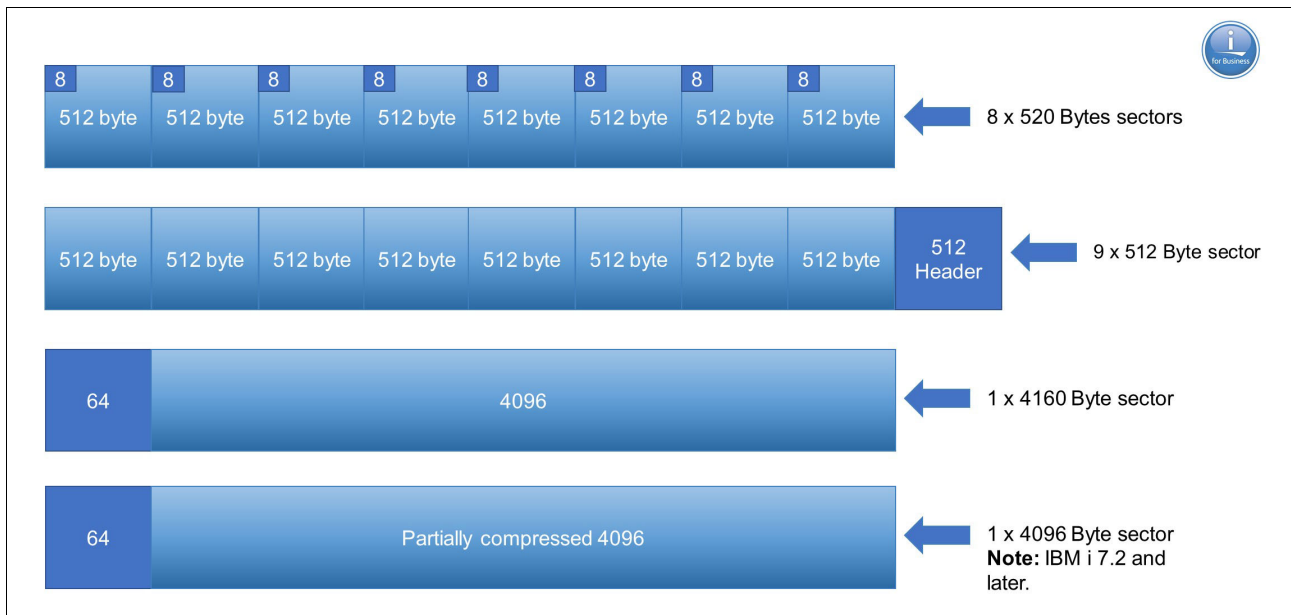


Figure A-7 IBM i with different sector sizes

Storage connection to IBM i

IBM Storage Virtualize storage can be attached to IBM i in the following ways:

- ▶ Native connection without using the IBM PowerVM® Virtual I/O Server (VIOS)
- ▶ Connection with VIOS in N_Port ID Virtualization (NPIV) mode
- ▶ Connection with VIOS in virtual Small Computer System Interface (SCSI) mode

The decision for IBM i native storage attachment or VIOS attachment is based on the customer's requirements. Native attachment has its strength in terms of simplicity, and it can be a preferred option for static and smaller IBM i environments with only a few partitions. It does not require extra administration and configuration of a VIOS environment. However, it also provides the least flexibility and cannot be used with PowerVM advanced functions, such as Live Partition Mobility (LPM) or remote restart.

Table A-1 lists the key criteria to help you with the decision of selecting an IBM i storage attachment method.

Table A-1 Comparing IBM i native and Virtual I/O Server attachment

Criteria	Native attachment	VIOS attachment
Simplicity (configuration, maintenance, and failure analysis)	✓	More complex
Performance	✓	✓ (with NPIV)
Consolidation (storage and network adapters)	More limited	✓
PowerVM advanced functions (partition mobility, suspend and resume, remote restart, and private cloud deployment)	Not available	✓
Hardware support (storage and network adapters, and entry level servers)	More limited	✓

The next sections describe the guidelines and best practices for each type of connection.

Note: For more information about the current requirements, see the following web pages:

- ▶ [IBM System Storage Interoperation Center \(SSIC\)](#)
- ▶ [IBM i POWER External Storage Support Matrix Summary](#)

Native attachment

Refer to [SSIC](#) for the native connection support for IBM i with IBM Storage Virtualize storage. This website provides the most current information for the supported versions.

Native connection *with* SAN switches can be done by using the following adapters:

- ▶ 32 Gb PCIe3 2-port Fibre Channel (FC) adapters (Feature Code #EN1A or #EN1B (IBM POWER9™ processor-based servers only))
- ▶ 16 Gb PCIe3 4-port FC adapters (Feature Code #EN1C or #EN1D (POWER9 processor-based servers only))
- ▶ 16 Gb PCIe3 2-port FC adapters (Feature Code #EN0A or #EN0B)

- ▶ 8-Gb Peripheral Component Interconnect Express (PCIe) 2-port FC adapters (Feature Code #5735 or #5273)
- ▶ 4-Gb PCIe 2-port FC adapters (Feature Code #5774 or #5276)

Direct native connection *without* SAN switches can be done by using the following adapters:

- ▶ 16-Gb adapters in IBM i connected to 16-Gb adapters in IBM Storage Virtualize 7.5 or later based storage with non-NPIV target ports
- ▶ 4-Gb FC adapters in IBM i connected to 8-Gb adapters in IBM Storage Virtualize based storage with non-NPIV target ports

For resiliency and performance reasons, connect IBM Storage Virtualize storage to IBM i with multipathing that uses two or more FC adapters. Consider the following points:

- ▶ You can define a maximum of 127 LUNs (up to 127 active + 127 passive paths) to a 16- or 32-Gb port in IBM i with IBM i 7.2 TR7 or later, and with IBM i 7.3 TR3 or later.
- ▶ You can define a maximum of 64 LUNs (up to 64 active + 64 passive paths) to a 16- or 32-Gb port with IBM i release and TR lower than IBM i 7.2 TR7 and IBM i 7.3 TR3.
- ▶ You can define a maximum of 64 LUNs (up to 64 active + 64 passive paths) to a 4- or 8-Gb port, regardless of the IBM i level.

The LUNs report in IBM i as disk units with type 2145.

IBM i enables SCSI command tag queuing in the LUNs from natively connected IBM Storage Virtualize storage. The IBM i queue depth per LUN and path with this type of connection is 16.

VIOS attachment

The following FC adapters are supported for VIOS attachment of IBM i to IBM Storage Virtualize storage:

- ▶ 32 Gb PCIe3 2-port FC adapter (Feature Code #EN1A or #EN1B (POWER9 processor-based servers only))
- ▶ 16 Gb PCIe3 4-port FC adapter (Feature Code #EN1C or #EN1D (POWER9 processor-based servers only))
- ▶ 16 Gb PCIe3 2-port FC adapter (Feature Code #EN0A or #EN0B)
- ▶ 8-Gb PCIe 2-port FC adapter (Feature Code #5735 or #5273)
- ▶ 8 Gb PCIe2 2-port FC adapter (Feature Code #EN0G or #EN0F)
- ▶ 8 Gb PCIe2 4-port FC adapter (Feature Code #5729)
- ▶ 8 Gb PCIe2 4-port FC adapter (Feature Code #EN12 or #EN0Y)

Important: For more information about the current requirements, see:

- ▶ [IBM System Storage Interoperation Center \(SSIC\)](#)
- ▶ [IBM i POWER External Storage Support Matrix Summary](#)

Connecting with VIOS NPIV

IBM i storage attachment support that uses PowerVM VIOS NPIV was introduced with POWER6 technology. With NPIV, volumes (LUNs) from the IBM Storage Virtualize storage system are directly mapped to the IBM i server. VIOS does not see NPIV-connected LUNs; instead, it is an FC pass-through.

The storage LUNs are presented to IBM i with their native device type of 2145 for IBM Storage Virtualize based storage. NPIV attachment requires 8 Gb or later generation FC adapter technology, and SAN switches that must be NPIV-enabled (see Figure A-8).

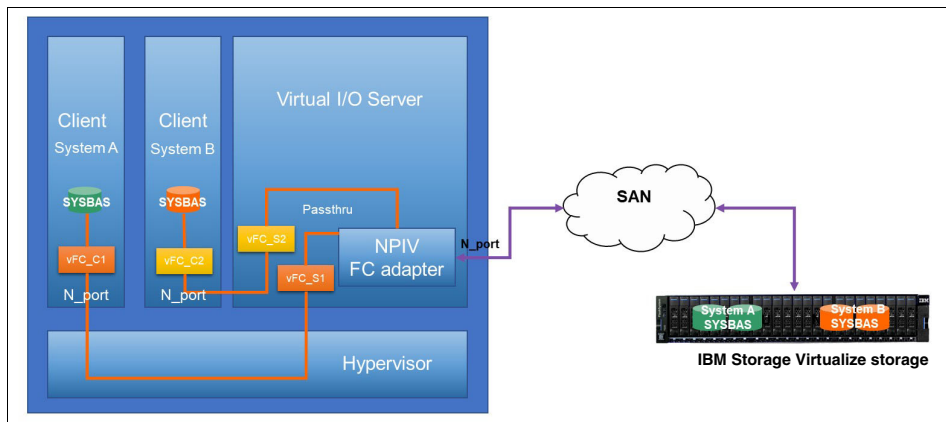


Figure A-8 IBM i SAN access by using NPIV

Redundant VIOS with NPIV

For resiliency and performance reasons, connect IBM Storage Virtualize storage to IBM i by using multipathing across two or more VIOS servers.

Observe the following rules for mapping IBM i server virtual FC (vFC) client adapters to the physical FC ports in VIOS when implementing an NPIV connection:

- ▶ Up to 64 vFC adapters can be mapped to the same physical FC adapter port in VIOS. With VIOS 3.1 and later, this limit was increased to support mapping of up to 255 vFC adapters to a 32-Gb physical FC adapter port.
- ▶ Mapping of more than one NPIV client vFC adapter from the *same* IBM i system to a VIOS physical FC adapter port is supported since IBM i 7.2 TR7 and i 7.3 TR3. However, when PowerVM partition mobility is used, only a single vFC adapter can be mapped from the *same* IBM i system to a VIOS physical FC adapter port.
- ▶ The same port can be used in VIOS for NPIV mapping and connecting with VIOS virtual Small Computer System Interface (VSCSI).
- ▶ If PowerHA solutions with IBM i IASPs are implemented, different vFC adapters must be used for attaching the IASP LUNs, and an adapter is not shared between SYSBAS and IASP LUNs.

A maximum of 127 LUNs (up to 127 active + 127 passive paths) can be configured to a vFC adapter with IBM i 7.2 TR7 or later, and with IBM i 7.3 TR3 or later.

A maximum of 64 LUNs (up to 64 active + 64 passive paths) can be configured to a vFC adapter with IBM i release and TR earlier than i 7.2 TR7 and i 7.3 TR3.

IBM i enables SCSI command tag queuing for LUNs from a VIOS NPIV that is connected to IBM Storage Virtualize storage. The IBM i queue depth per LUN and path with this type of connection is 16.

Note: If you encounter issues with NPIV or vFC of IBM i that is attached to IBM Storage Virtualize, such as missing paths and missing disk units, consider the following best practices:

- ▶ For System Snapshot (SYSSNAP), be sure to include Licensed Internal Code (LIC) logs, history log message queue (QHST), and product activity logs (PALs). Change the date range to include the date range of the problem. For more information, see [QMGTOOLS: System Snapshot \(SYSSNAP\)](#).
- ▶ VIOS snapshots can be collected from the VIOS partitions as part of the SYSSNAP or separately. For more information, see [How to Collect Snap From a PowerVM Virtual I/O Server \(VIOS\)](#).
- ▶ Collect switch logs as close as possible to the time of the problem.
- ▶ Collect the applicable state snapshot from IBM Storage Virtualize at the time that the problem is occurring. This information is needed by the storage support team.

NPIV acceleration

VIOS 3.1.2 or later strengthened FC NPIV to provide multiqueue support. This enhanced performance, including more throughput, reduced latency, and higher input/output operations per second (IOPS), spreads the I/O workload across multiple work queues.

The following FC adapter feature codes are supported:

- ▶ 32 Gb PCIe3 2-port FC adapters (Feature Code #EN1A or #EN1B (POWER9 processor-based servers only))
- ▶ 16 Gb PCIe3 4-port FC adapters (Feature Code #EN1C or #EN1D (POWER9 processor-based servers only))
- ▶ 16 Gb PCIe3 2-port FC adapters (Feature Code #EN2A or #EN2B)

Note: NPIV acceleration is supported by IBM i 7.2 or later, and by IBM POWER9 firmware 940 or later.

Connecting with VIOS VSCSI

IBM i storage attachment by using the PowerVM VIOS connection that uses VSCSI was introduced with IBM POWER6 technology.

When deciding on an PowerVM VIOS storage attachment for IBM i, NPIV attachment is often preferred over VSCSI attachment for the following reasons:

- ▶ With VSCSI, an emulation of generic SCSI devices is performed by VIOS for its client partitions, such as IBM i, which requires extra processing and adds a small delay to I/O response times.
- ▶ VSCSI provides much lower scalability in terms of the maximum supported LUNs per virtual adapter than NPIV. It also requires more storage management, such as multipath configuration and customization at the VIOS layer, which adds complexity.
- ▶ Because of the VSCSI emulation unique device characteristics of the storage device, such as device type (or in the case of tape devices, media type) and other device attributes are no longer presented to the IBM i client.
- ▶ VSCSI attachment is not supported for PowerHA LUN-level switching technology, which is required for IASP HyperSwap solutions with IBM Storage Virtualize.

Similar considerations for NPIV apply regarding the usage of IBM i multipathing across two or more VIOS to improve resiliency and performance. However, because with VSCSI multipathing also is implemented at the VIOS layer, the following considerations apply:

- ▶ IBM i multipathing is performed with two or more VSCSI client adapters, each of them assigned to a VSCSI server adapter in a different VIOS. With VSCSI, volumes (LUNs) from the IBM Storage Virtualize system are not mapped directly to an IBM i host, but to the two or more VIOS servers. These LUNs, which are detected as hard disk drives (HDDs) on each VIOS, must be mapped as a virtual target device to the relevant VSCSI server adapters to be used by the IBM i client.
- ▶ In addition to IBM i multipathing across multiple VIOS servers, with VSCSI, multipathing is implemented at the VIOS server layer to provide further I/O parallelism and resiliency by using multiple physical FC adapters and SAN fabric paths from each VIOS server to its storage.
- ▶ The IBM recommended multipath driver for IBM Storage Virtualize based storage running microcode 7.6.1 or later is the VIOS built-in AIX Path Control Module (AIXPCM) multipath driver, which replaces the previously recommended Subsystem Device Driver Path Control Module (SDDPCM) multipath driver.

For more information, see [The Recommended Multi-path Driver to use on IBM AIX and VIOS When Attached to SVC and Storwize storage](#).

Up to 4095 LUNs can be connected per target, and up to 510 targets per port in a physical adapter in VIOS. With IBM i 7.2 and later, a maximum of 32 disk LUNs can be attached to a VSCSI adapter in IBM i.

With IBM i releases before 7.2, a maximum of 16 disk LUNs can be attached to a VSCSI adapter in IBM i. The LUNs are reported in IBM i as generic SCSI disk units of type 6B22.

IBM i enables SCSI command tag queuing in the LUNs from a VIOS VSCSI adapter that is connected to IBM Storage Virtualize storage. A LUN with this type of connection features a queue depth of 32.

Setting attributes in VIOS

This section describes the values of specific device attributes in VIOS, which must be configured for resiliency and performance.

FC adapter attributes

With a VIOS VSCSI connection or NPIV connection, use the VIOS **chdev** command to specify the following attributes for each SCSI I/O Controller Protocol Device (fscsi) device that connects an IBM Storage Virtualize storage LUN to IBM i:

- ▶ The attribute **fc_err_recov** should be set to **fast_fail**.
- ▶ The attribute **dyntrk** should be set to **yes**.

The specified values for the two attributes specify how the VIOS FC adapter driver or VIOS disk driver handle specific types of fabric-related failures and dynamic configuration changes. Without setting these values for the two attributes, the way these events are handled is different, which causes unnecessary retries or manual actions.

Note: These attributes also are set to the recommended values when applying the default rules set that is available with VIOS 2.2.4.x or later.

Disk device attributes

With a VIOS VSCSI connection, use the VIOS **chdev** command to specify the following attributes for each hdisk device that represents an IBM Storage Virtualize storage LUN that is connected to IBM i:

- ▶ If IBM i multipathing across two or more VIOS servers is used, the attribute **reserve_policy** is set to `no_reserve`.
- ▶ The attribute **queue_depth** is set to 32.
- ▶ The attribute **algorithm** is set to `shortest_queue`.

Consider the following points:

- ▶ To prevent SCSI reservations on the hdisk device, **reserve_policy** must be set to `no_reserve` in each VIOS if multipathing with two or more VIOSs is implemented.
- ▶ Set **queue_depth** to 32 for performance reasons. Setting this value ensures that the maximum number of I/O requests that can be outstanding on an HDD in the VIOS at a time matches the maximum number of 32 I/O operations that IBM i operating system allows at a time to one VIOS VSCSI-connected LUN.
- ▶ Set **algorithm** to `shortest_queue` for performance reasons. Setting this value allows the AIXPCM driver in VIOS to use dynamic load-balancing instead of the default path failover algorithm for distributing the I/O across the available paths to IBM Storage Virtualize storage.
- ▶ Setting a physical volume identifier (PVID) for HDD devices that are used for VSCSI attachment of IBM i client partitions is not recommended because it makes those devices ineligible for a possible later migration to NPIV or native attachment.

Important: While working with SCSI and NPIV, you cannot use both for the paths to the same LUN. However, VIOS supports NPIV and SCSI concomitantly, that is, some LUNs can be attached to the virtual worldwide port names (WWPNs) of the NPIV FC adapter. At the same time, the VIOS also can provide access to LUNs that are mapped to virtual target devices and exported as VSCSI devices.

One or more VIOSs can provide the pass-through function for NPIV. Also, one or more VIOSs can host VSCSI storage. Therefore, the physical Host Bus Adapter (HBA) in the VIOS supports NPIV and VSCSI traffic.

Guidelines for Virtual I/O Server resources

Be aware of the memory requirements of the hypervisor when determining the overall memory of the system. Above and beyond the wanted memory for each partition, you must add memory for virtual resources (VSCSI and vFC) and hardware page tables to support the maximum memory value for each partition.

A best practice is to use the IBM Workload Estimator tool to estimate the needed VIOS resources. However, as a starting point in context of CPU and memory for VIOS, see [Sizing the Virtual I/O Server \(VIOS\) - My Rules of Thumb](#).

Disk drives for IBM i

This section describes how to implement internal disk drives in IBM Storage Virtualize storage or externally virtualized back-end storage for an IBM i host. These suggestions are based on the characteristics of a typical IBM i workload, such as a relatively high write ratio, a relatively high-access density, and a small degree of I/O skew because of the spreading of data by IBM i storage management.

Considering these characteristics and typical IBM i customer expectations for low I/O response times, we expect that many SAN storage configurations for IBM i will be based on an all-flash storage configuration.

If for less demanding workloads or for commercial reasons a multitier storage configuration that uses enterprise class (`tier0_flash`) and high-capacity (`tier1_flash`) flash drives or even enterprise HDDs (`tier2_HDD`) is preferred, ensure that a sufficiently large part of disk capacity is on flash drives. As a rule, for a multitier configuration with the typically low IBM i I/O skew, at least 20% of IBM i capacity should be based on the higher tier flash storage technology.

Even if specific parts of IBM i capacity are on flash drives, it is important that you provide enough HDDs with high rotation speed for a hybrid configuration with flash drives and HDDs. Preferably, use 15 K RPM HDDs of 300 GB or 600 GB capacity, along with flash technology.

IBM i transaction workload often achieves the best performance when disk capacity is used entirely from enterprise class flash (`tier0_flash`) storage.

The usage of a multitier storage configuration by IBM Storage Virtualize storage is achieved by using Easy Tier.

For more information, see *Implementing the IBM FlashSystem with IBM Spectrum Virtualize Version 8.4.2*, SG24-8506.

Even if you do not plan to use a multitier storage configuration or currently have no multitier storage configuration that is installed, you can still use Easy Tier for intra-tier rebalancing. You also can evaluate your workload with its I/O skew, which provides information about the benefit that you might gain by adding flash technology in the future.

Compression considerations

If compression is wanted, the preferred choice for using compression at the IBM Storage Virtualize storage system layer for a performance-critical IBM i workload is by using IBM FlashCore Module (FCM) hardware compression technology at the disk drive level within IBM Storage Virtualize standard pools or data reduction pools (DRPs) with fully allocated volumes. These configuration options do not affect performance compared to other compression technologies, such as DRP compressed volumes or IBM Real-time Compression (RtC) at the storage subsystem level.

Important: Data reduction or deduplication can be used with IBM i, which affects performance positively.

Nevertheless, the performance is tremendously affected and different whenever something is touched, such as 30 minutes taking 3 - 18 hours. The data is affected whenever something is created, changed, or used. The integrity of the objects is maintained.

However, if a physical page on disk is corrupted, potentially hundreds or thousands of objects become corrupted instead of only one. Another consideration is the amount of wear that occurs on the drives from so much read/write activity.

If you plan to use deduplication for archival or test purposes, deduplication might be a viable solution for saving huge amounts of storage. If the deduplication solution is planned for a production or development environment, we recommend that you test it thoroughly before committing.

Storage sizing and performance modeling

IBM provides tools, such as IBM Storage Modeller (StorM) and IntelliMagic Disk Magic for IBM representatives and IBM Business Partners, which are recommended to be used for performance modelling and sizing before implementing an IBM Storage Virtualize storage configuration for IBM i. These tools allow the user to enter the performance data of the current IBM i workload manually or by using file import from IBM i (5770-PT1 reports or PDI data) or from IBM Spectrum Control performance data. Enter the current storage configuration and model the wanted configuration.

When modeling Easy Tier, specify the lowest skew level for IBM i workload or import an I/O skew curve from available Easy Tier reports. The steps that are taken for sizing and modeling IBM i are shown in Figure A-9.

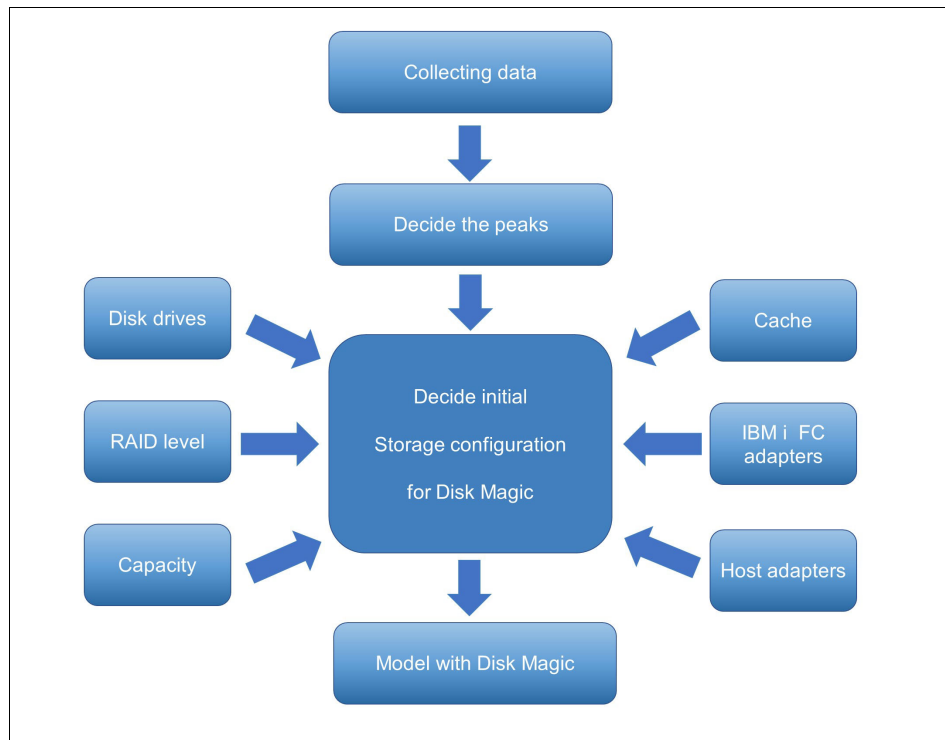


Figure A-9 Sizing and modeling for IBM i by using Disk Magic

The modeling helps to ensure an adequate solution sizing by providing predictions for the modeled IBM Storage Virtualize storage resource usage, the predicted disk response time for IBM i, and the usage and response times of workload growth.

Note: Contact your IBM representative or IBM Business Partner to discuss a performance modeling and sizing for a planned IBM Storage Virtualize storage solution for IBM i.

IBM i unmap support

To better use IBM Storage Virtualize storage flash technology with efficient storage space allocation and deallocation, IBM i supports storage system unmap capabilities by using corresponding host unmap functions.

Initially, IBM i unmap support that was implemented by using the SCSI Write Same command was introduced with IBM i 7.2 TR8 and IBM i 7.3 TR4 for LUN initialization only, that is, for the Add Disk Units to ASP function.

With IBM i 7.3 TR9 and IBM i 7.3 TR5, runtime support was added, which also supports synchronous unmap for scenarios, such as object deletion and journal clearance. The runtime unmap algorithm was further enhanced supported by IBM i 7.3 TR7 and IBM i 7.4 TR1, which implements an asynchronous periodic free-space cleaning.

IBM Storage Virtualize 8.1.1 and later storage systems can use the unmap function to efficiently deallocate space, such as for volume deletion, on their back-end storage by sending SCSI **unmap** commands to specific supported internal solid-state drives (SSDs) and FCMs, and selected virtualized external flash storage.

Space reclamation that is triggered by host **unmap** commands is supported by IBM Storage Virtualize 8.1.2 and later for DRP thin-provisioned volumes, which can increase the free capacity in the storage pool so that it becomes available for use by other volumes in the pool.

For more information about IBM Storage Virtualize storage SCSI unmap support, see 4.4, “Data reduction pools best practices” on page 270 and [SCSI Unmap support in Spectrum Virtualize systems](#).

Defining LUNs for IBM i

LUNs for an IBM i host are defined from IBM Storage Virtualize block-based storage. The LUNs are created from available extents within a storage pool, much like for open system hosts.

Although IBM i supports a usable, large-size LUN of up to 2 TB (1 byte for IBM Storage Virtualize storage), the usage of only a few large-size LUNs for IBM i is not recommended for performance reasons.

In general, the more LUNs that are available to IBM i, the better the performance for the following reasons:

- ▶ If more LUNs are attached to IBM i, storage management uses more threads and enables better performance.
- ▶ More LUNs provide a higher I/O concurrency, which reduces the likelihood of I/O queuing and the wait time component of the disk response time, which results in lower latency of disk I/O operations.

For planning purposes, consider that many LUNs might also require more physical or vFC adapters on IBM i based on the maximum number of LUNs that is supported by IBM i per FC adapter port.

The sizing process helps to determine a reasonable number of LUNs that are required to access the needed capacity while meeting performance objectives. Regarding both these aspects and best practices, we suggest the following guidelines:

- ▶ For any IBM i disk pool (ASP), define all the LUNs as the same size.
- ▶ 40 GB is the preferred minimum LUN size.
- ▶ You should not define LUNs larger than about 200 GB.

Note: This rule is not fixed because it is important that enough LUNs are configured, with which this guideline helps. Selecting a larger LUN size should not lead to configurations, such as storage migrations, with fewer LUNs being configured, with possibly detrimental effects on performance.

- ▶ A minimum of eight LUNs for each ASP is preferred for small IBM i partitions, a couple of dozen LUNs for medium partitions, and up to a few hundreds for large partitions.

When defining LUNs for IBM i, consider the following required minimum capacities for the load source (boot disk) LUN:

- ▶ With IBM i 7.1, the minimum capacity is 20 GB.
- ▶ With IBM i 7.2 before TR1, the minimum capacity is 80 GB in IBM i.
- ▶ With IBM i 7.2 TR1 and later, the minimum capacity is 40 GB in IBM i.

IBM Storage Virtualize dynamic volume expansion is supported for IBM i with IBM i 7.3 TR4 and later. An IBM i initial program load (IPL) is required to use the extra volume capacity.

Tip: For more information about cross-referencing IBM i disks units with IBM Storage Virtualize LUNs by using NPIV, see [Cross-Referencing \(Device Mapping\) IBM i Disks with SAN Disks Using N-Port ID Virtualization \(NPIV\)](#).

Disk arms and maximum LUN size

Selected limits that are related to disk arms and LUN sizes were increased in IBM i 7.4, as listed in Table A-2.

Table A-2 Limits increased for maximum disk arms and LUN sizes

System limits	IBM i 7.2	IBM i 7.3	IBM i 7.4	IBM i 7.5
Disk arms in all basic ASPs (ASPs 1 - 32) per logical partition (LPAR)	2047	2047	3999	3999
Disk arms in all IASPs (IASPs 33 - 255) in all nodes in a cluster	2047	2047	5999	5999
Maximum combined number of disk arms and redundant connections to disk units	35.600	35.600	35.600	35.600
512-byte block size LUNs ^a	2 TB	2 TB	2 TB	2 TB
4096-byte block size LUNs ^b	2 TB	2 TB	16 TB	16 TB

a. The limit is one block short of the maximum that is listed in Table A-2. For all 512 block LUNs, the maximum is still up to 2 TB, including IBM Storwize LUNs and SVC LUNs.

- b. This size includes IBM FlashSystem LUNs, and 4 K block serial-attached SCSI (SAS) disks (VSCSI-attached).

Note: For more information about these limits and other limits, see [Miscellaneous limits](#).

Data layout

Spreading workloads across all IBM Storage Virtualize storage components maximizes the usage of the hardware resources in the storage subsystem. I/O activity must be balanced between the two nodes or controllers of the IBM Storage Virtualize storage system I/O group, which often is addressed by the alternating preferred node volume assignments at LUN creation.

However, performance problems might arise when sharing resources because of resource contention, especially with incorrect sizing or unanticipated workload increases.

Some isolation of workloads, at least regarding a shared back-end storage, can be accomplished by using a configuration in which each IBM i ASP or LPAR has its own managed storage pool. Such a configuration with dedicated storage pools results in a trade-off between accomplishing savings from storage consolidation and isolating workloads for performance protection. This result occurs because a dedicated storage pool configuration likely requires more back-end storage hardware resources because it cannot use the averaging effect of multiple workloads, typically showing their peaks at different time intervals.

Consider the following data layouts:

- ▶ For all-flash storage configurations, assuming a correctly sized storage back end, often no reason exists for not sharing the disk pool among multiple IBM i workloads.
- ▶ For hybrid configurations with Easy Tier on mixed HDD and flash disks, the storage pool also might be shared among IBM i workloads. Only large performance-critical workloads are configured in isolated disk pools.
- ▶ For HDD only pools, make sure that you isolate performance-critical IBM i workloads in separate storage pools.
- ▶ Avoid mixing IBM i LUNs and non-IBM i LUNs in the same disk pool.

Apart from the usage of Easy Tier on IBM Storage Virtualize for managing a multitier storage pool, an option is available to create a separate storage pool for different storage tiers on IBM Storage Virtualize storage and create different IBM i ASPs for each tier. IBM i applications that have their data in an ASP of a higher storage tier experience a performance boost compared to the ones that use an ASP with a lower storage tier.

IBM i internal data relocation methods, such as the ASP balancer hierarchical storage management function and IBM Db2 media preference, are not available to use with IBM Storage Virtualize flash storage.

Fibre Channel adapters in IBM i and VIOS

When you size the number of FC adapters for an IBM i workload for native or VIOS attachment, consider the maximum I/O rate (IOPS) and data rate (MBps) that a port in a specific adapter can sustain at 70% utilization. Also, consider the I/O rate and data rate of the IBM i workload.

If multiple IBM i partitions connect through the same FC port in VIOS, consider the maximum rate of the port at 70% utilization and the sum of I/O rates and data rates of all connected LPARs.

For sizing, you might consider the throughput that is listed in Table A-3, which shows the throughput of a port in a specific adapter at 70% utilization.

Table A-3 Throughput of FC adapters

Maximal I/O rate per port	16 Gb 2-port adapter	8 Gb 2-port adapter
IOPS per port	52,500 IOPS	23,100 IOPS
Sequential throughput per port	1,330 MBps	770 MBps
Transaction throughput per port	840 MBps	371 MBps

Make sure to plan for the usage of separate FC adapters for IBM i disk and tape attachment. This separation is recommended because of the required IBM i virtual input/output processor (IOP) reset for tape configuration changes and for workload performance isolation.

Zoning SAN switches

With IBM i native attachment or VIOS NPIV attachment, zone the SAN switches so that one IBM i FC initiator port is in a zone with two FC ports from the IBM Storage Virtualize storage target, with each port from one node canister of the I/O group, as shown in Figure A-10. This configuration provides resiliency for the I/O to and from a LUN that is assigned to the IBM i FC initiator port. If the preferred node for that LUN fails, the I/O continues to use the non-preferred node.

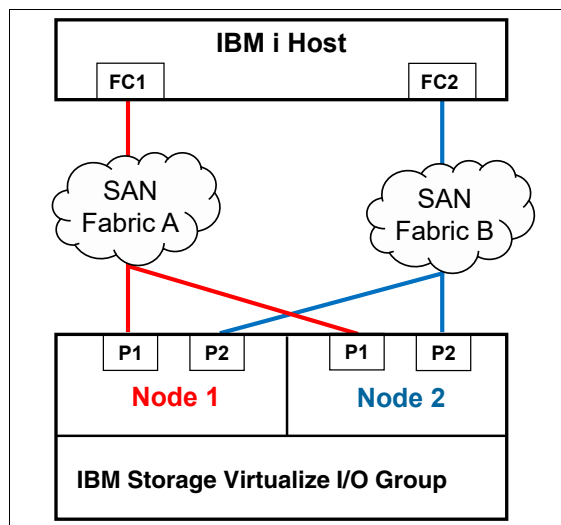


Figure A-10 SAN switch zoning for IBM i with IBM Storage Virtualize storage

For VIOS VSCSI attachment, zone one physical port in VIOS with one or more available FC ports from each of both node canisters of the IBM Storage Virtualize storage I/O group. SVC or Storwize ports that are zoned with one VIOS port should be evenly spread between both node canisters. A maximum of eight host paths is supported from VIOS to IBM Storage Virtualize storage.

IBM i multipath

Multipath provides greater resiliency for SAN-attached storage, and it can improve performance as well. IBM i supports up to eight active paths and up to eight passive paths to each LUN. In addition to availability considerations, lab performance testing shows that two or more paths provide performance improvements when compared to a single path.

Typically, two active paths to a LUN are a good balance of price and performance. The scenario that is shown in Figure A-10 on page 786 results in two active and two passive paths to each LUN for IBM i. However, you can implement more than two active paths for workloads where high I/O rates are expected to the LUNs (a high I/O access density is expected).

It is important to understand that IBM i multipathing for a LUN is achieved by connecting the LUN to two or more FC ports that belong to different adapters in an IBM i partition. Adding more than one FC port from the same IBM Storage Virtualize storage node canister to a SAN switch zone with an IBM i FC initiator port does not provide more active paths because an IBM i FC initiator port, by design, logs in to only one target port of a node.

With IBM i native attachment, the ports for multipath must be from different physical FC adapters in IBM i. With VIOS NPIV, the vFC adapters for multipath must be assigned to different VIOSs for redundancy. However, if more than two active paths are used, you can use two VIOSs and split the paths among them. With VIOS VSCSI attachment, the VSCSI adapters for IBM i multipath must be assigned to different VIOSs.

IBM Storage Virtualize storage uses a redundant dual active controller design that implements SCSI Asymmetric Logical Unit Access (ALUA). Some of the paths to a LUN are presented to the host as optimized and others as non-optimized.

With an ALUA-aware host such as IBM i, the I/O traffic to and from a specific LUN normally goes through only the optimized paths, which often are associated with a specific LUN of a preferred node. The non-optimized paths, which often are associated with the non-preferred node, are not actively used.

In an IBM Storage Virtualize storage topology, such as HyperSwap or IBM SAN Volume Controller Enhanced Stretched Cluster (ESC) that implements host site awareness, the optimized paths are not necessarily associated with a preferred node of a LUN but with the node of the I/O group that includes the same site attributes as the host.

If the node with the optimized paths fails, the other node of the I/O group takes over the I/O processing. With IBM i multipath, all the optimized paths to a LUN are reported as *active* on IBM i, while the non-optimized paths are reported as *passive*. IBM i multipath employs its load-balancing among the active paths to a LUN and starts to use the passive paths if all the active paths failed.

Booting from SAN

All IBM i storage attachment options that are native (VIOS NPIV and VIOS VSCSI) support IBM i boot from SAN. The IBM i load source is on an IBM Storage Virtualize storage LUN that is connected in the same manner as the other LUNs.

Apart from a required minimum size, the load source LUN does not include any special requirements. The FC or SCSI I/O adapter for the load source must be *tagged* (that is, specified) by the user in the IBM i partition profile on the IBM Power Hardware Management Console (HMC). When installing the IBM SLIC with disk capacity on IBM Storage Virtualize storage, the installation prompts you to select one of the available LUNs for the load source.

IBM i mirroring

Some customers prefer to use IBM i mirroring functions for resiliency. For example, they use IBM i mirroring between two IBM Storage Virtualize storage systems, each connected with one VIOS.

When setting up IBM i mirroring with VIOS-connected IBM Storage Virtualize storage, complete the following steps to add the LUNs to the mirrored ASP:

1. Add the LUNs from two virtual adapters, with each adapter connecting one to-be mirrored half of the LUNs.
2. After mirroring is started for those LUNs, add the LUNs from another two new virtual adapters, each adapter connecting one to-be mirrored half, and so on. This way, you ensure that IBM i mirroring is started between the two IBM Storage Virtualize storage systems and not among the LUNs from the same storage system.

Copy Services considerations

This section covers IBM Storage Copy Services considerations for usage with IBM i.

Remote replication

The IBM Storage Virtualize family supports Metro Mirror (MM) synchronous remote replication and Global Mirror (GM) asynchronous remote replication.

Two options are available for GM: *Standard* GM, and Global Mirror with *Change Volumes* (GMCV), which allows for a flexible and configurable recovery point objective (RPO) that allows data replication to be maintained during peak periods of bandwidth constraints, and data consistency at the remote site to be maintained and during resynchronization.

Regarding the usage of IBM Storage Virtualize Copy Services functions, the IBM i single-level storage architecture requires that the disk storage of an IBM i system is treated as a single entity, that is, the scope of copying or replicating an IBM i disk space must include SYSBAS (referred to as *full system replication*) or an IASP (referred to as *IASP replication*).

Full system replication is used for disaster recovery (DR) purposes where an IBM i standby server is used at the DR site, as shown in Figure A-11.

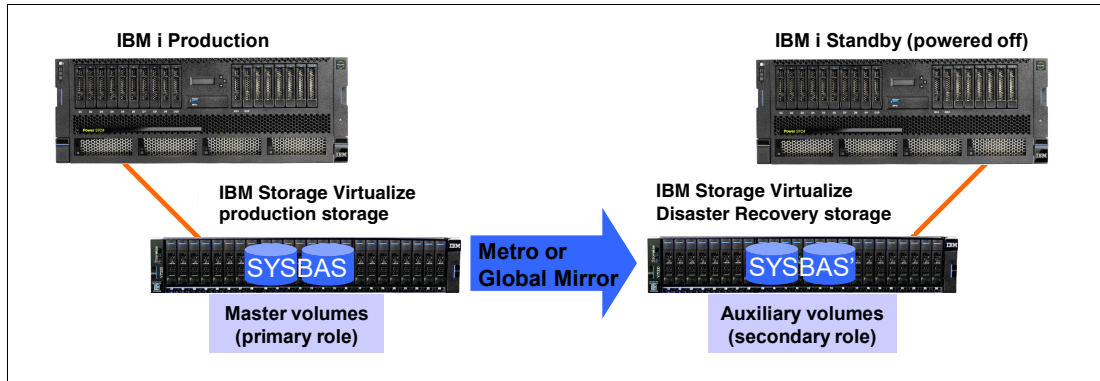


Figure A-11 IBM i full system replication with IBM Storage Virtualize

When a planned or unplanned outage occurs for the IBM i production server, the IBM i standby server can be started (undergo an IPL) from the replicated SYSBAS volumes, and then on IBM Storage Virtualize, they take on the primary role to become accessible to the IBM i standby host.

IASP-based replication for IBM i is used for a high availability (HA) solution where an IBM i production and an IBM i backup node are configured in an IBM i cluster and the IASP that is replicated by IBM Storage Virtualize remote replication is switchable between the two cluster nodes, as shown in Figure A-12.

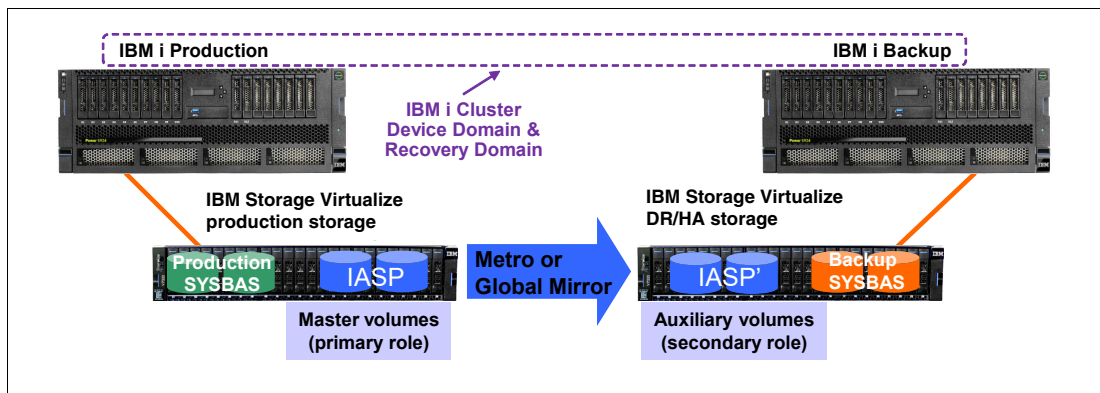


Figure A-12 IBM i IASP replication with IBM Storage Virtualize

In this scenario, the IBM i production system and the IBM i backup system each have their own non-replicated SYSBAS volumes and only the IASP volumes are replicated. This solution requires IBM PowerHA SystemMirror® for i Enterprise Edition (5770-HAS *BASE and option 1) to manage the IBM i cluster node switch and failovers and the IBM Storage Virtualize storage remote replication switching.

For more information about IBM i HA solutions with IBM Storage Virtualize Copy Services, see *PowerHA SystemMirror for IBM i Cookbook*, SG24-7994.

The sizing of the required replication link bandwidth for MM or GM must be based on the peak write data rate of the IBM i workload to avoid affecting production performance. For more information, see 6.4.3, “Remote copy network planning” on page 415.

For more information about current IBM Storage Virtualize storage zoning guidelines, see 2.3, “IBM Storage Virtualize system ports” on page 134.

For environments that use remote replication, a minimum of two FC ports is suggested on each IBM Storage Virtualize storage node that is used for remote mirroring. The remaining ports on the node should not have any visibility to any other IBM Storage Virtualize cluster. Following these zoning guidelines helps you avoid configuration-related performance issues.

FlashCopy

When planning for FlashCopy with IBM i, make sure that enough disk drives are available to the FlashCopy target LUNs to maintain a good performance of the IBM i production workload while FlashCopy relationships are active. This guideline is valid for FlashCopy with background copy and without background copy.

When FlashCopy is used with thin-provisioned target LUNs, make sure that sufficient capacity is available in the storage pool to be dynamically allocated when needed for the copy-on-write (CoW) operations. The required thin target LUN capacity depends on the amount of write operations to the source and target LUNs, the locality of the writes, and the duration of the FlashCopy relationship.

FlashCopy temperature and considerations for IBM i

FlashCopy temperature indicates the amount of disruption to the source system and the quality of the FlashCopy target. FlashCopy copies what was sent to disk. Updates that are sitting in memory on the IBM i are not known to the storage system.

FlashCopy cold

The following considerations apply to FlashCopy cold:

- ▶ All memory is flushed to disk.
- ▶ The source IASP must be varied off before performing a FlashCopy.
- ▶ This method is the only method to ensure that all writes are sent out to disk and included.

FlashCopy warm

The following considerations apply to FlashCopy warm:

- ▶ No memory is flushed to disk.
- ▶ Writes in memory are excluded from the FlashCopy target.
- ▶ Zero disruption to IBM i source system.

FlashCopy quiesced

IBM i provides a quiesce function that can suspend database transactions and database and Integrated File System (IFS) file change operations for the system and configured basic ASPs or IASPs.

The following considerations apply to FlashCopy quiesced:

- ▶ Some memory flushed to disk.
- ▶ Attempts to flush writes to disk and suspend DB I/O, and reach commitment control boundaries.
- ▶ Minimal disruption to source is the best practice, and a better quality than warm.

IBM Lab Services PowerHA Tools for IBM i

PowerHA Tools for IBM i are a set of tools that are developed by IBM Lab Services that focus on high availability and disaster recovery (HADR) and offline backup solutions by using external storage and PowerHA for IBM i.

These tools extend the functions that are included in the base PowerHA code, which contributes to the automation and usability of this kind of solution.

They are an IBM Lab Services asset, and they are delivered with the corresponding implementation, support, and training services.

Some of these tools are enabled to work with IBM Storage Virtualize products.

Full System Replication Manager

Full System Replication Manager is a set of tools that are designed for the management and automation of replication solutions for IBM i systems based on external storage, where the entire IBM i system is replicated in its entirety. It implements a storage-based replication solution for an IBM i system so that you do not need to migrate the customer environment to an IASP environment.

These tools are useful both for the implementation and administration of the solution and for the automation of switchover tasks.

The solution has a control LPAR (IBM i) that communicates with the storage units, HMCs, and IBM i LPARs (primary and secondary nodes).

The control LPAR is used to monitor, manage, and switch replication by using the commands and menus that are provided by this toolkit.

For redundancy reasons, a best practice is to have one dedicated control LPAR per site. The synchronization between these LPARs is done through the PowerHA clustering functions.

To ensure that unnecessary switchover is never performed, for example, in situations when other corrective actions are more appropriate, automatic switchovers are not allowed, and the switchover task must be initiated manually by running a toolkit command.

Figure A-13 shows an overview of the Full System Replication Manager.

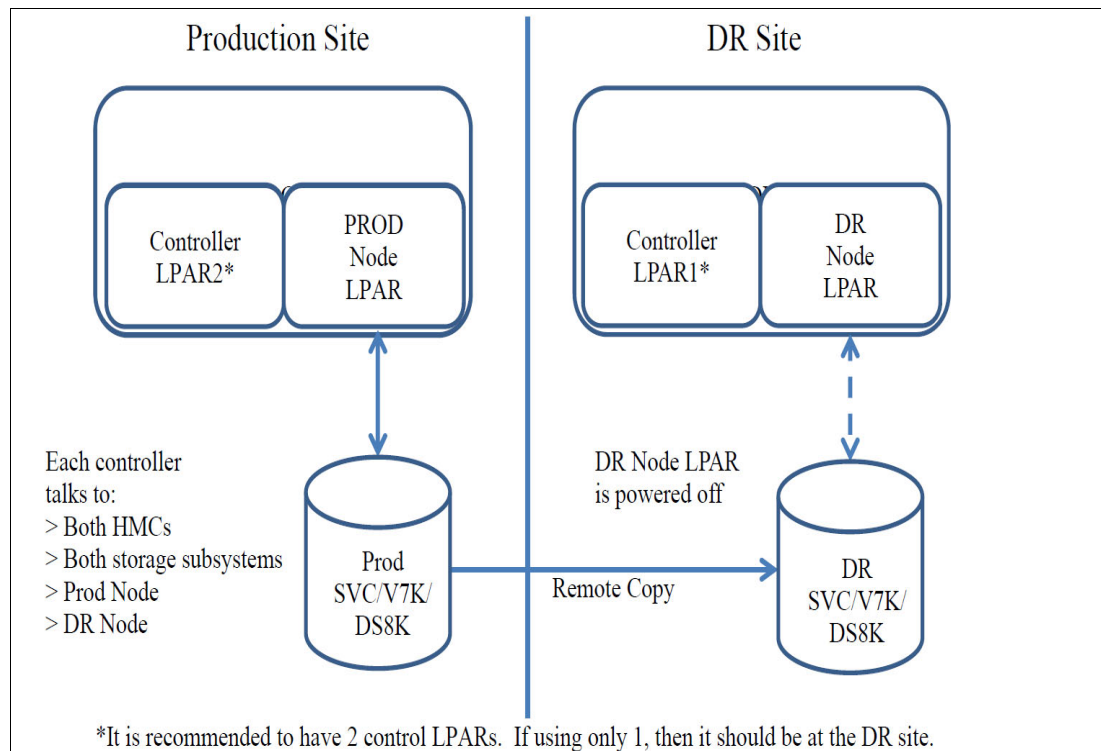


Figure A-13 IBM Lab Services PowerHA Tools for IBM i: Full System Replication Manager

Note: For more information about the Full System Replication toolkit, see [PowerHA Tools for IBM i - Full System Replication](#).

Full System FlashCopy

PowerHA Full System FlashCopy (FSFC) for IBM i is used to automate taking backups offline while minimizing the disruption to the production system.

This tool provides access to the HMCs to manage IBM i partitions, the IBM i function to pause database activity on a commit boundary, and the IBM Storage Virtualize FlashCopy feature to create a copy of the production LPAR from which backups such as full system saves can be taken in a restricted state, all while minimizing the impact on users. The users experience a pause in database activity that is generally less than 30 seconds.

FSFC Manager for IBM i copies the whole system ASP so that you can implement an FlashCopy based offline backup solution for an IBM i system, which avoids the need to migrate the customer environment to an IASP environment.

The PowerHA FSFC for IBM i Manager operation has the following steps:

1. On the production LPAR, the database operations are paused on a transaction boundary, and the information in main memory is flushed to disks.
2. The FlashCopy relationships are started in the external storage unit.
3. Database activity is resumed on the production LPAR.
4. An IPL is performed in the backup LPAR, and then the configured backups are started.
5. If necessary, the Backup Recovery and Media Services (BRMS) information is updated in the production partition by transferring it from the backup partition.

Figure A-14 shows the Full System FlashCopy Manager.

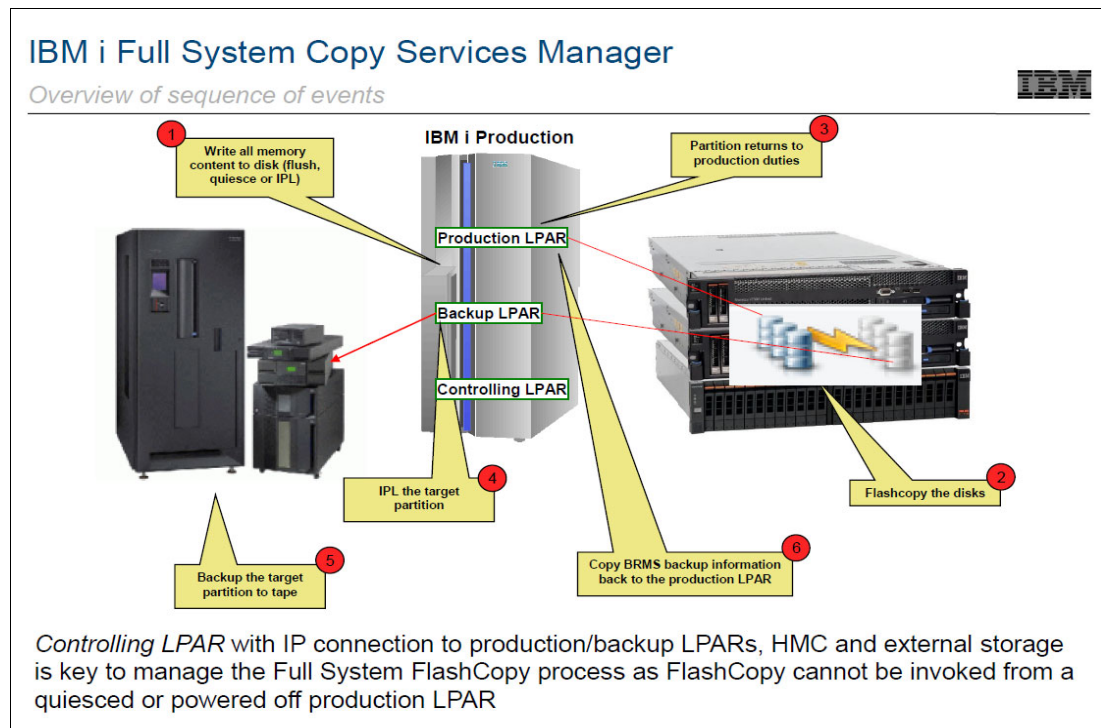


Figure A-14 IBM Lab Services PowerHA Tools for IBM i: Full System FlashCopy

Note: For more information about the FSFC toolkit, see [PowerHA Tools for IBM i - Full System FlashCopy](#).

IASP Manager

PowerHA Tools IASP Manager is a product that is designed for IBM i customers that use PowerHA, IASP, and external storage solutions.

Its main objective is to enhance the automation, monitoring, management, testing, and customization capabilities of these kinds of environments, which complements the PowerHA functions with command-line interface (CLI) commands and automated scripts.

With IBM Storage Virtualize, the PowerHA Tools IASP Manager-FlashCopy is available, and it provides functions to assist with the automation and management of FlashCopy through a set of commands you can use to create a point in time copy of an IASP.

In IBM Storage Virtualize, IASP Manager is available the PowerHA Tools IASP Manager-FlashCopy toolkit, which is responsible for completely automating the entire FlashCopy process by running the following tasks:

- ▶ Vary off the production IASP or quiesce its database activity.
- ▶ Start the FlashCopy relationships on the external storage box.
- ▶ Vary on the production IASP or resume its database activity.
- ▶ Connect or disconnect the host connections for the FlashCopy target LUNs.
- ▶ Vary on the IASP on FlashCopy node.
- ▶ Integrate a customized backup program on the FlashCopy target.

- ▶ If required, vary off the IASP on the FlashCopy node and remove the FlashCopy relationships on the external storage box after backups are complete.
- ▶ Integrate with remote copy to ensure that the FlashCopy copy has valid data.

Note: For more information about the IASP Manager-FlashCopy toolkit, see [PowerHA Tools for IBM i - IASP Manager](#).

Smart Assist for PowerHA on IBM i

Smart Assist for PowerHA on IBM i is a set of tools that were created by IBM Lab Services based on feedback and the needs of numerous customer engagements over many years.

Many of these tools were created in response to customers' automation requirements. The IBM Lab Services team continues enhancing existing tools and adding new ones regularly.

These tools are useful for many implementations of solutions that are based on PowerHA. When specifically referring to implementations of solutions for IBM i with IBM Storage Virtualize, we can highlight the following:

- ▶ IBM i Independent ASP (IASP) migration and management
- ▶ PowerHA-managed Geographic Mirroring
- ▶ PowerHA-managed IBM Storage Virtualize based FlashCopy
- ▶ PowerHA Tools for IBM i IASP Manager managed implementations

Some of the benefits that Smart Assist can provide through its various commands are as follows:

- ▶ Extra functions to make the setup and installation phases of the environment easier.
- ▶ Programming interfaces to help monitor the environment.
- ▶ Command utilities to help with the daily administration of the environment:
 - IASP Management
 - PowerHA Cluster Management
 - Admin Domain Management
 - PowerHA Environments
 - SVC Based Management
 - IASP Manager Environments

Note: For more information about Smart Assist for PowerHA on IBM i, see [Smart Assist for PowerHA on IBM i](#).

HyperSwap

IBM Storage Virtualize HyperSwap as an active-active remote replication solution is supported for IBM i full system replication with IBM i 7.2 TR3 or later. It is supported for native and VIOS NPIV attachment.

HyperSwap for IBM i IASP replication is supported by IBM i 7.2 TR5 or later and IBM i 7.3 TR1 or later. With this solution, you must install IBM PowerHA SystemMirror for i Standard Edition (5770-HAS *BASE and option 2), which enables LUN level switching to site 2. It is supported for native and VIOS NPIV attachment.

HyperSwap relies on the SCSI ALUA-aware IBM i host multipath driver to manage the paths to the local and remote IBM Storage Virtualize storage systems, which are logically configured as a single clustered system.

From a SAN switch zoning perspective, HyperSwap requires that the IBM i host is zoned with both IBM Storage Virtualize nodes of the I/O group on each site. For a balanced configuration, the SAN switches from a dual-fabric configuration must be used evenly.

Figure A-15 shows an example of the SAN fabric connections for IBM i HyperSwap with VIOS NPIV attachment. This configuration example results in four active paths and 12 passive paths that are presented on IBM i for each HyperSwap LUN.

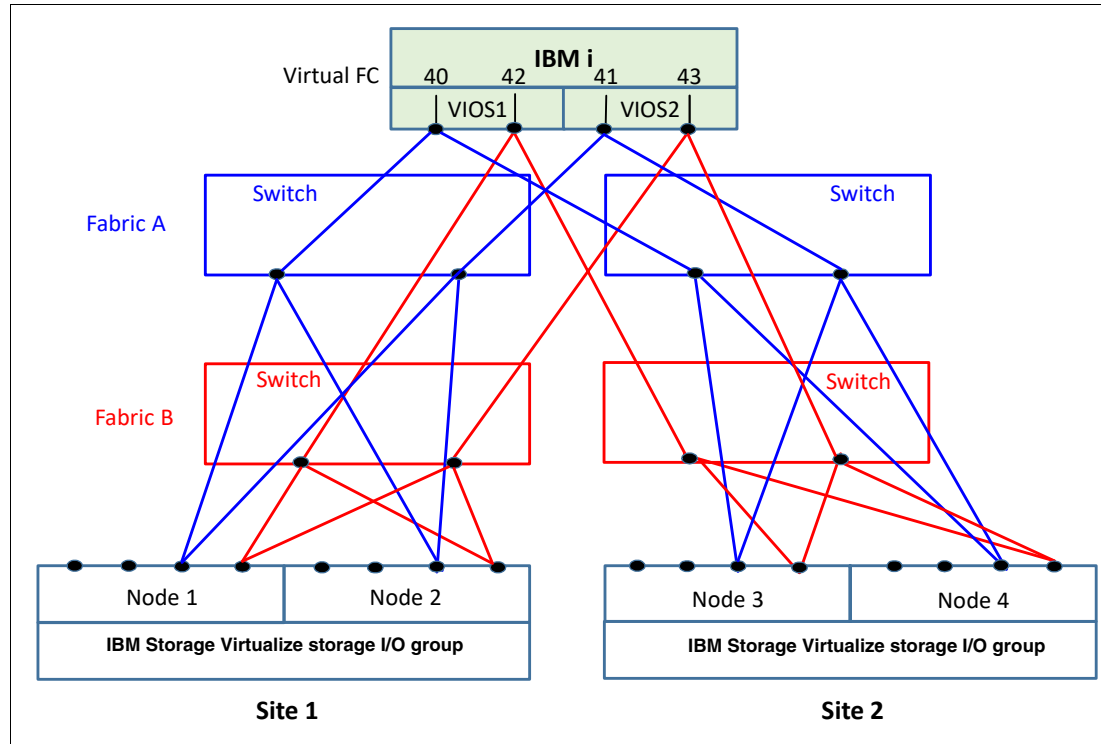


Figure A-15 IBM i HyperSwap SAN fabric connection example

Next, we briefly describe some HA scenarios that use HyperSwap for IBM i.

Outage of an IBM Storage Virtualize I/O group at site 1

In this scenario, the entire IBM i storage capacity is on HyperSwap LUNs.

After the outage of an I/O group at site 1 occurs, the I/O rate automatically transfers to the IBM Storage Virtualize nodes at site 2. The IBM i workload keeps running, and no relevant messages exist in the IBM i message queues.

When the outage completes, the IBM i I/O rate automatically transfers to nodes on site 1. The IBM i workload keeps running without interruption.

Disaster at site 1 with a full system HyperSwap

In this scenario, we use a prepared IBM i standby system at site 2. The entire IBM i storage capacity is on HyperSwap LUNs. Two hosts are defined in the IBM Storage Virtualize storage cluster: one host with the WWPNs of IBM i at site 1, and one with the WWPNs of site 2.

After a failure of site 1, including a failure of the IBM i production system and the storage at site 1, the IBM i LUNs are still available from the IBM Storage Virtualize nodes at site 2. In the HyperSwap cluster, we manually unmap the HyperSwap LUNs from the IBM i production host at site 1, map the LUNs to the IBM i standby host at site 2, and perform an IPL of the IBM i standby host at site 2. After the IPL finishes, we can resume the workload on site 2.

After the outage of site 1 completes, we power down IBM i at site 2, unmap the IBM i LUNs from the host at site 2, and then map the LUNs to the host at site 1. We perform an IPL of IBM i at site 1 and resume the workload. The I/O rate is transferred to the IBM Storage Virtualize storage nodes at site 1.

Disaster at site 1 with IASP HyperSwap

This scenario requires IBM PowerHA SystemMirror for IBM i software to be installed, and the corresponding IBM i setup, which consists of two IBM i partitions in a cluster and a switchable IASP on IBM i at site 1, a PowerHA cluster resource group, and PowerHA copy description. The workload is running in the IASP.

For more information about the PowerHA for IBM i setup, see *IBM PowerHA SystemMirror for i: Preparation (Volume 1 of 4)*, SG24-8400.

In this scenario, ensure that all IBM i LUNs (not only the IASP LUNs) are HyperSwap volumes.

If a disaster occurs at site 1, PowerHA automatically switches the IASP to the system at site 2, and the workload can be resumed at site 2.

After the failure at site 1 is fixed, use PowerHA to switch the IASP back to site 1 and resume the workload at this site.

Planned outage with Live Partition Mobility

PowerVM LPM allows you to move a running LPAR, including its operating system and running applications, from one system to another one without a shutdown or disrupting the operation of that LPAR.

In this scenario, we combine LPM with HyperSwap to transfer the workload onto site 2 during a planned outage of site 1. This combination requires VIOS NPIV attachment and all IBM i LUNs configured as HyperSwap LUNs.

For more information about LPM and its requirements, see *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940.

To use LPM, you must define the IBM i host in IBM Storage Virtualize with the WWPNs of the second port of the vFC adapters. As a best practice, create a separate host object definition for the secondary ports to specify site 2 for this host object. Then, enable the I/O rate to be transferred to the nodes at site 2 after migrating the IBM i partition with LPM.

After the outage is complete, you can use LPM again to transfer the IBM i partition back to site 1. After the migration, the I/O rate automatically moves to the nodes at site 1.

Important: LPM now supports multiple client vFC adapter ports that are mapped to a single physical FC port. Each client vFC must be mapped to a separate physical port in advance, whether LPM with FC NPIV is used. That restriction was removed for the use of VIOS 3.1.2.10 or later and IBM i 7.2 or later. Therefore, the same physical port can be double-mapped to the same IBM i client partition. This configuration allows for better adapter usage.

SAN Volume Controller stretched cluster

SVC is a hardware and software storage solution that implements IBM Storage Virtualize. SVC appliances map physical volumes in the storage device to virtualized volumes, which makes them visible to host systems (for example, IBM i). SVC also provides Copy Services functions that can be used to improve availability and support DR, including MM, GM, and FlashCopy.

Therefore, the IBM PowerHA SystemMirror for IBM i interface is compatible with SVC. After the basic SVC environment is configured, PowerHA can create a copy session with the volumes.

The usage of PowerHA with SVC management creates an automated HADR solution with minimal extra configurations. PowerHA and SVC interfaces are compatible with hardware that is running IBM Storage Virtualize and IBM Storwize series.

Full system replication in a stretched cluster

This HA storage solution with SVC uses a stretched cluster topology and volume mirroring. For more information about stretched clusters, see Chapter 7, “Ensuring business continuity” on page 501.

A scenario that uses full system replication with IBM Storage Virtualize in SVC presents full system replication by using volume mirroring is shown in Figure A-16.

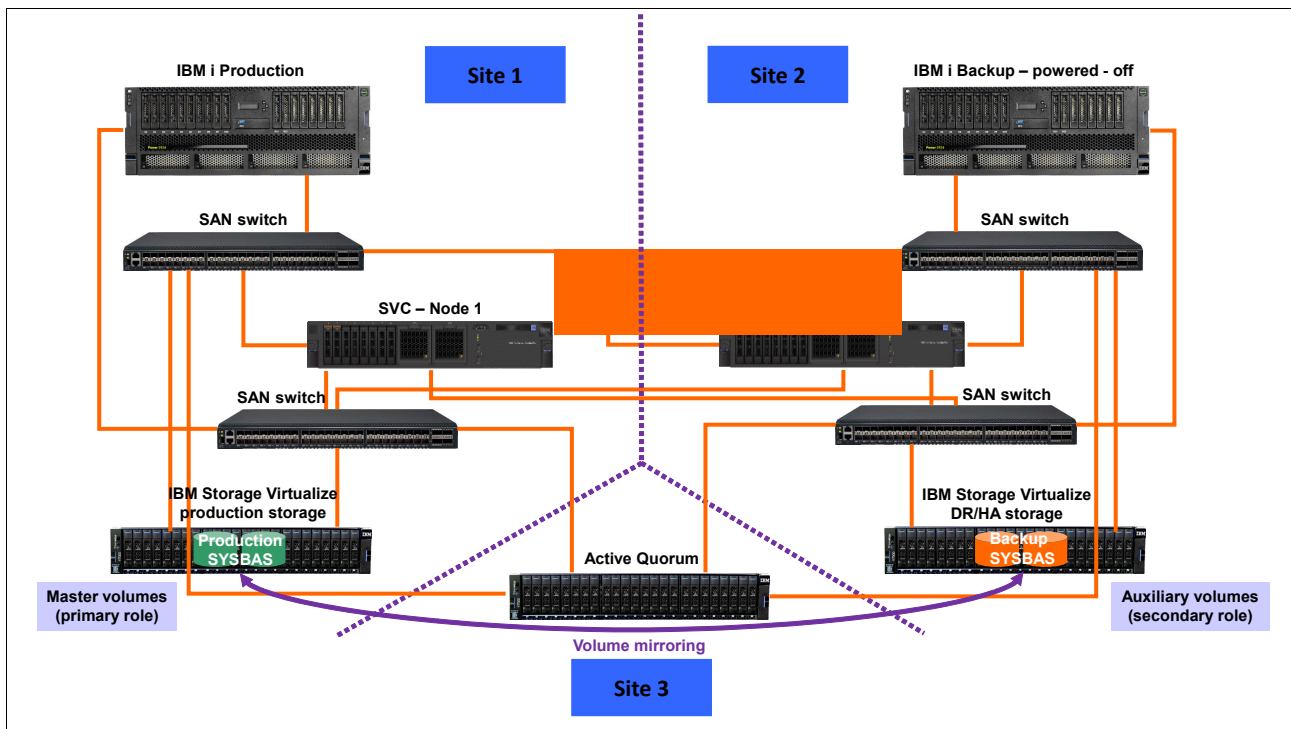


Figure A-16 Full system replication that uses SAN Volume Controller volume mirroring

The scenario that is shown in Figure A-16 shows an IBM i production system at site 1, a prepared IBM i backup system at site 2 that is powered off, and a third site that is the active quorum.

Two nodes of SVC are in a stretched cluster topology that is called a split-cluster.

The LUNs of production SYSBAS are in an IBM Storage Virtualize production storage system at site 1. Those LUNs include a copy at site 2 in a second IBM Storage Virtualize DR/HA storage system, and this copy is done by using volume mirroring. Therefore, the SVC stretch cluster configuration provides a continuous availability platform in which IBM i access is maintained, whether any single failure domain is lost (in this example, three domains exist).

Simultaneous IBM i access to both copies must be prevented, that is, the IBM i backup system must be powered off when the IBM i production system is active, and vice versa.

In our example, after a failure of site 1 (including a failure of the IBM i production system and the storage at site 1), the IBM i LUNs are still available because of the two data copies (the second at site 2).

An abnormal IPL is done at the IBM i backup systems. Later, the IPL ends, and we can resume the workload at site 2.

After the outage of site 1 is finished, we power down the IBM i backup system at site 2, and the resynchronization between both copies is incremental and started by the SVC automatically. Volume mirroring is below the cache and copy services. Then, we restart the workload of IBM i production at site 1.

Note: For this example, HA testing and configuration changes are more challenging than with remote copy. For example, manual assignment is needed for the preferred node to enable local reads. Therefore, ESC, which was introduced with SVC 7.2, adds a site awareness feature (reads always locally) and DR capability if simultaneous site and active quorum failures occur.

LUN-level switching

This solution uses a single copy of an IASP group that can be switched between two IBM i systems. Likewise, LUN-level switching is supported for NPIV attachment and native attachment for storage that is based on IBM Storage Virtualize or IBM Storwize series.

LUN-level switching also is supported for an SVC. This solution engages in heterogeneous environments where an SVC stretched cluster is used as the basis for a cross-platform, two-site HA solution.

Consider the following points regarding LUN-level switching features:

- ▶ LUN-level switching uses a single copy of an IASP on a single storage system.
- ▶ The IASP LUNs are assigned to host object definitions for the primary IBM i system.
- ▶ A second set of host object is defined for the secondary IBM i partition, but no LUNs are presented to the secondary system.
- ▶ The primary and secondary partitions are nodes in a cluster device domain.
- ▶ A device description for the IASP must be created on the secondary node before IBM PowerHA is configured.

- ▶ LUN-level switching can be used alone or together with MM or GM in a 3-node cluster. In a 3-node cluster, LUN-level switching provides local HA. Whether the whole local site goes down, MM or GM provides a DR option to a remote site.

Note: LUN-level switching plus MM or GM (IASP) in the 3-site solutions is not available for IBM Storage Virtualize at the time of writing.

- ▶ If you want to add LUN-level switching to MM or GM, you do not need to create the cluster or IASP, or change the cluster administrative domain. You must create the IASP device description on the backup system from the LUN switching perspective.

The following IBM i licensed programs are required:

- ▶ IBM PowerHA SystemMirror for i (5770-HAS).
- ▶ Option 41 (HA Switchable Resources), which is installed in all relevant IBM i systems.
- ▶ Option 33 (5770-SS1 - Portable Application Solutions Environment), which is installed in all relevant IBM i systems.
- ▶ IBM Portable Utilities for i and OpenSSH, Open SSI, zlib (5733-SC1 base and option 1), which is installed in all relevant IBM i systems.

The LUN-level switching IBM PowerHA SystemMirror for IBM i editions are listed in Table A-4.

Table A-4 LUN-level switching IBM PowerHA SystemMirror for i editions

IBM i HADR clustering	Express Edition	Standard Edition	Enterprise Edition
Cluster admin domain	No	Yes	Yes
Cluster device domain	No	Yes	Yes
Integrated heartbeat	No	Yes	Yes
Application monitoring	No	Yes	Yes
IBM i event and error management	No	Yes	Yes
Automated planned failover	No	Yes	Yes
Managed unplanned failover	No	Yes	Yes
Centralized FlashCopy	No	Yes	Yes
LUN-level switching	No	Yes	Yes
Geomirror Sync mode	No	Yes	Yes
Geomirror Async mode	No	No	Yes
Multisite HADR management	No	No	Yes
MM	No	No	Yes
GM	No	No	Yes
IBM HyperSwap	Yes	Yes	Yes

IBM PowerHA LUN-level switching for SAN Volume Controller in a stretched cluster

This HA solution (see Figure A-17) supports two sites by using a combination of IBM i PowerHA LUN-level switching for server redundancy and SVC in a stretched cluster that uses volume mirroring for storage redundancy.

The following requirements must be met:

- ▶ IBM i 7.1 TR6 or later
- ▶ NPIV or native attachment of SVC

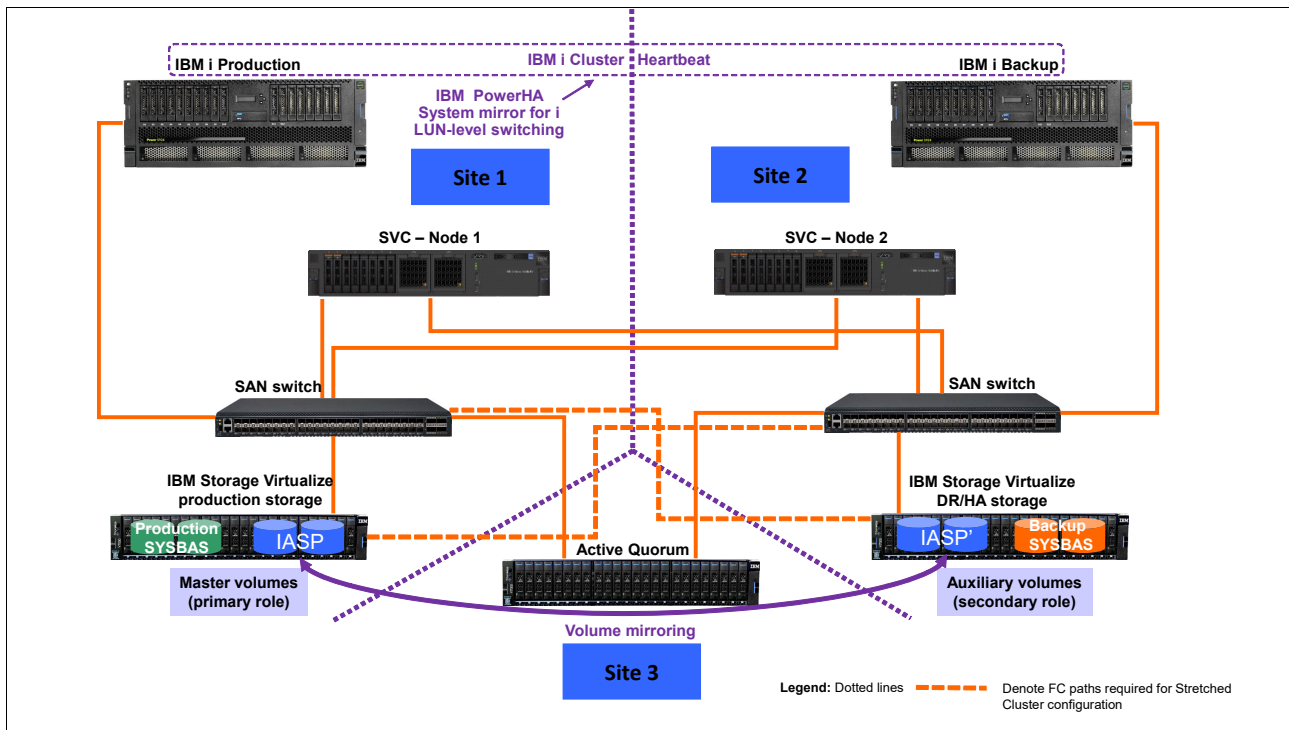


Figure A-17 IBM PowerHA System Mirror for i LUN-level switching with SAN Volume Controller stretched cluster

The scenario that uses IBM PowerHA System Mirror for i LUN-level switching with an SVC stretched cluster provides the following benefits over IBM i full system replication with an SVC stretched cluster:

- ▶ High degree of automation for planned and unplanned site switches and failovers
- ▶ Shorter recovery times by using IASP
- ▶ Reduced mirroring bandwidth requirements by using IASP (Temporary writes in SYSBAS, such as for index builds, are not mirrored.)

As shown in Figure A-17, availability is achieved through the inherent active architecture of SVC with volume mirroring.

During a failure, the SVC nodes and associated mirror copy of the data remain online and available to service all host I/O. The two data copies are placed in different managed disk (MDisk) groups or IBM Storage Virtualize storage systems. The resynchronization between both copies of IASP is incremental. Mirrored volumes feature the same functions and behavior as a standard volume.

Note: Because HyperSwap was introduced with IBM Storage Virtualize 7.5 on Storwize and SVC, the scenario that uses the topology of HyperSwap with SVC also is valid for IBM i.

Even with HyperSwap, we can use consistency groups (CGs) that are enabled by using the IBM i multipath driver, but not in stretched cluster scenarios. A remote mirroring license is required for the usage of HyperSwap with IBM i.

For more information, see *Storwize HyperSwap with IBM i*, REDP-5490.

For more information about limits and restrictions for SVC, see [V8.6.0.x Configuration Limits and Restrictions for IBM System Storage SAN Volume Controller](#).

Db2 mirroring for IBM i

The Db2 Mirror base configuration consists of two systems that are in the same data center. This configuration does not span locations because it is active-active read/write, which means that by definition all write operations are synchronous (by using a Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE) network) to the application state. All write operations between two systems necessitates that the distance between the systems is limited to not affect performance.

The following broad approaches can be used to deploy active-active solutions:

- ▶ Use distributed lock management where multiple application servers can access the common or shared database. The servers are prevented from stepping on each other by the distributed lock management, which locks out the other users while you perform an update.
- ▶ The replication approach is used when each update of any type is synchronous to the application state. Therefore, when an application performs an update, it does not proceed to the next application step until the current write operations complete on the primary and secondary objects, that is, a two-phase commit exists between the two systems.

Note: Applications can be deployed in an active-active manner, in which each application server has simultaneous access to the database on both systems in the two-node active-active complex. If one of the database servers fails, the application servers continue performing I/O operations to the other system in the mirrored pair. This configuration includes the added benefit of enabling workload balancing.

However, applications also can be deployed in an active-passive manner where application servers conduct write operations to one of the two systems in the two-system complex. If the primary system is removed, the application groups are switched to the secondary system. The active-active case necessitates that the application servers are hosted separately from the database servers and connected through a client/server construct, such as Java Database Connectivity (JDBC).

Note: IBM i JDBC drivers now contain alternative server failover support to automatically send the JDBC request between systems when one connection is no longer available. For many IBM i application workloads, deployment is completed through the traditional 5250 emulation window and contained in the same LPAR as the operating system and database. In this case, if the primary fails, the database is continuously replicated to the secondary system synchronously and immediately available. The application must be restarted on the secondary system before the workload processing resumes.

When one of the systems in the IBM DB2® Mirror configuration is not available, Db2 Mirror tracks all update, change, and delete operations to the database table and all other mirror-eligible objects. When the pair is reconnected, all changes are synchronized between the systems. This process includes databases that are in an IASP or as part of the base system storage.

Db2 Mirror is compatible with IASPs and uses IASPs for IFS support within the Db2 Mirror configuration. For non-IFS objects, IASPs can be used, but are not required.

Also, Db2 Mirror supports applications that use traditional record-level access or SQL-based database access. Support for IFS and IFS journals is accomplished through deployment into an IASP, which can be configured as a switchable LUN, or in a mirrored pair of IASPs through storage replication.

This solution requires a POWER8 processor-based server or later and IBM i 7.4 or higher. For more information about software requirements for Db2 Mirror, see this [IBM i operating system Software requirements](#).

DR can be achieved by using various options, such as the IBM PowerHA SystemMirror for i Enterprise Edition, full system replication, or logical replication.

Important: Db2 Mirror local continuous availability can be combined with HADR replication technologies. Consider the following points:

- ▶ Remote replication for DR can be implemented by storage-based replication, that is, by using IBM Storage Virtualize Copy Services software.
- ▶ Any IFS IASP must remain switchable between both local Db2 Mirror nodes by choosing a DR topology that is supported by IBM PowerHA SystemMirror for IBM i.
- ▶ Any DB IASP is available on both local nodes (no switch between local nodes).
A DB IASP is not required for local Db2 Mirror database replication, but might be preferred for implementing a remote replication solution with shorter recovery times compared to SYSBAS replication.
- ▶ For a complete business continuity solution at the DR site, a remote Db2 Mirror node pair can be configured for a 4-node Db2 Mirror PowerHA cluster configuration. IFS IASPs and DB IASPs must be registered with the remote Db2 Mirror pair (by using the SHADOW option for the DB IASP to maintain its Db2 Mirror configuration data, such as default inclusion state and Remote Code Load (RCL)).

For more information, see *IBM Db2 Mirror for i Getting Started*, REDP-5575.

Setup process overview

During the setup and configuration process for Db2 Mirror, the following nodes are referred to:

- ▶ Managing node
- ▶ Setup source node

► Setup copy mode

For more information about these nodes, the setup process, and configuration, see [Overview of the setup process](#).

Db2 Mirror is initially configured on a single partition that is called the *setup source node*. During the setup and configuration process, the setup source node is cloned to create the second node of the Db2 Mirror pair, which is called the *setup copy node*. The setup copy node is configured and initialized automatically by Db2 Mirror during its first IPL.

The Db2 Mirror configuration and setup process supports external and internal storage. External storage systems are used during the cloning process, and IBM storage systems are recommended rather than non-IBM external storage because the cloning process is automated, that is, Db2 Mirror automates the cloning for the IBM Storage Virtualize family.

The cloning technologies that are used for IBM storage systems are FlashCopy (cold and warm) and remote copy. FlashCopy is used when both Db2 Mirror nodes connect to the same IBM Storage Virtualize storage system. Cold cloning requires that the setup source node is shut down during the cloning portion of the setup process. A warm clone allows the setup source node to remain active during the entire Db2 Mirror setup and configuration process.

Remote copy is used when the Db2 Mirror nodes are connected to different IBM Storage Virtualize storage systems. However, a manual copy also is available. For more information, see [Storage requirements](#).

Note: Volume mirroring that is supported in IBM FlashSystem 9200 and SVC is a valid cloning method for Db2 Mirror for the manual copy category. It is *not* automated like when you use FlashCopy, MM, or GM.

IBM Storage Virtualize and Db2 Mirror

IBM Storage Virtualize storage systems establish communication with Db2 Mirror by using Secure Shell (SSH) to manage Copy Services functions. IBM Storage Virtualize user IDs must have the user role of administrator.

The following products are required for a managing node:

- 5733SC1 *BASE IBM Portable Utilities for i
- 5733SC1 Option 1 OpenSSH, OpenSSL, zlib
- 5770SS1 Option 33 Portable Application Solutions Environment

Note: For more information about creating an SSH key pair, see [Creating an SSH key file for accessing IBM Storage Virtualize storage](#). After an SSH key pair is created, attach the SSH public key to the IBM Storage Virtualize storage system. The corresponding private key file must be uploaded to the managing node so that it can be used during the Db2 Mirror setup process.

Virtual I/O Server and native attachment

The Db2 Mirror storage cloning process for IBM Storage Virtualize requires FC adapters with native attachment or attachment with VIOS NPIV.

Host object definition and volume planning

Before you set up Db2 Mirror, you must define the host object, and assign volumes to the hosts to be used by the setup copy node. The same number of host objects and volumes, and the same-sized volumes must be defined for the setup source node and setup copy node.

Later, the Db2 Mirror cloning process pairs storage volumes between the setup source node and setup copy node, which applies for SYSBAS and IASPs.

Consider the following points:

- ▶ The setup source node and set-up copy node must have the same number and sizes of LUNs or disks in SYSBAS.
- ▶ The host object and volumes for any database IASPs must be predefined for the setup copy node before a database IASP is added to Db2 Mirror.

Remote copy cloning

In this case, Db2 Mirror remote copy cloning uses the following IBM Storage Virtualize Copy Services operations to copy the setup source node volumes to the setup copy nodes volumes:

- ▶ GM for a cold clone
- ▶ GMCV for a warm clone

Whether you plan to perform the remote copy during a planned outage window, you must ensure that your bandwidth between storage systems is sufficient to complete the remote copy during that period. The Db2 Mirror cloning process cannot pause the cloning and then resume it later. Therefore, you must plan for enough time for the remote copy to complete.

Important: For IBM Storage Virtualize, the Copy Services partnership between storage systems must be manually created before Db2 Mirror is configured.

Architectures and considerations for Db2 Mirror

Because of the synchronous design of Db2 Mirror, the distance between the nodes is limited to be within a data center for most cases. Multiple configurations are supported for a data center Db2 Mirror implementation and the addition of a DR solution.

Several options are described in this section as examples with IBM Storage Virtualize storage systems. A specific implementation depends on your business resilience requirement.

Note: The Db2 Mirror configuration and setup process supports SVC topologies, such as ESC and HyperSwap.

Db2 Mirror environment with one IBM Storage Virtualize storage

In this example, one IBM Storage Virtualize storage system is used as a basic configuration for using Db2 Mirror. This configuration features some key advantages.

By using one storage system, you can take advantage of FlashCopy to set up your configuration rapidly. This solution might be considered for a DR strategy to provide storage resiliency.

As shown in Figure A-18, two IBM Power servers are used (at least one RoCE adapter per server). However, you can reduce the cost decreased resiliency of this scenario by implementing Db2 Mirror across two IBM i LPARs on the same IBM Power server. For this example, a SYSBAS is cloned, and IASP also can be added by using another set of volumes.

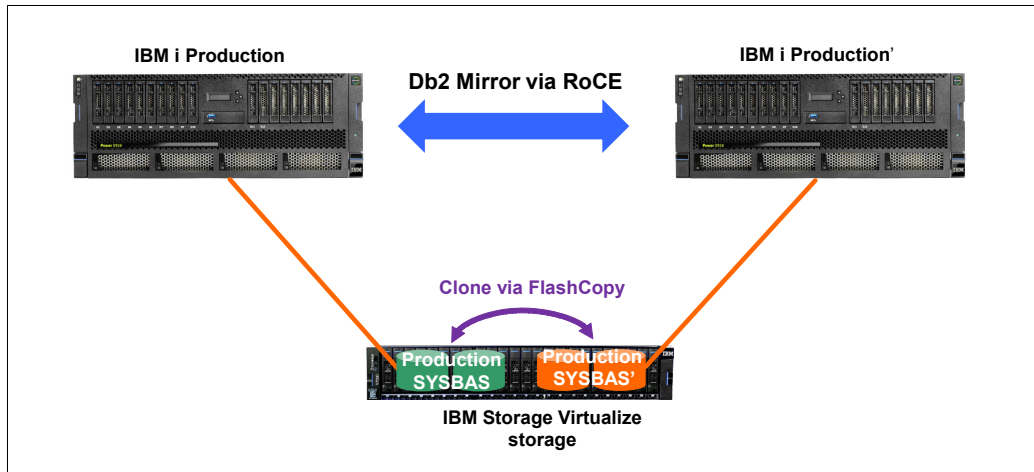


Figure A-18 Db2 Mirror environment with one IBM Storage Virtualize storage system

Db2 Mirror environment with two IBM Storage Virtualize storage systems

The usage of two IBM Storage Virtualize storage systems provides further redundancy by helping to ensure that the active node remains running and available during a storage outage. In this example, two IBM Power servers and IBM Storage Virtualize storage systems are used. Also, remote copy is used to set up Db2 Mirror.

As shown in Figure A-19, the set of volumes for SYSBAS and the set of volumes for IASP are replicated. GM also can be used.

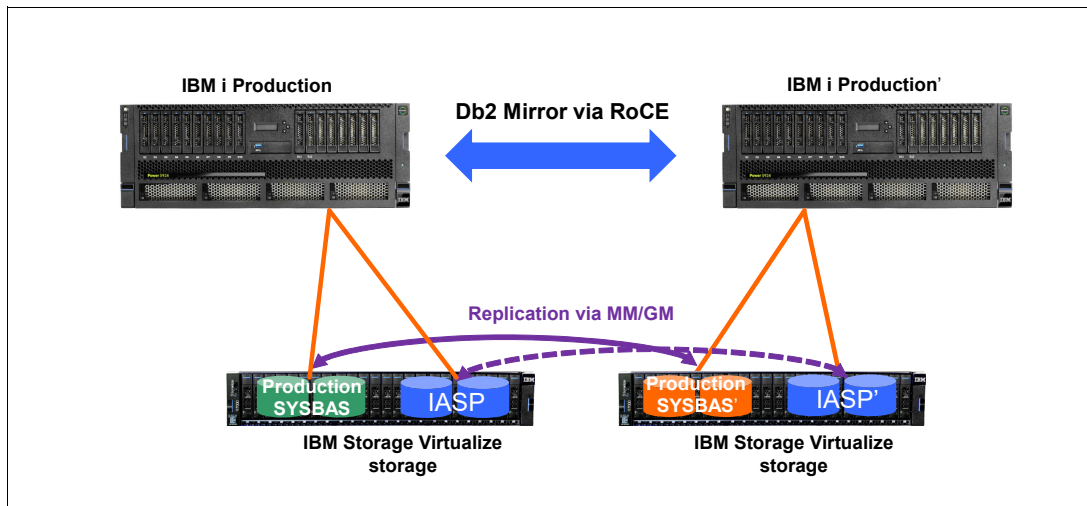


Figure A-19 Db2 Mirror environment with two IBM Storage Virtualize storage systems

Db2 Mirror and DR considerations

Db2 Mirror is a continuous availability solution, so it is *not* considered a DR solution. Db2 Mirror can be used within your DR strategy to improve your availability, even in a disaster situation.

The Db2 Mirror nodes must be close to each other because the maximum distance between the IBM Power servers is 200 meters (656 feet). At site 1, Db2 Mirror nodes are used, and at site 2 (the DR location), we can have a single server or multiple servers with Db2 Mirror nodes, and a unique or multiple IBM Storage Virtualize storage systems.

The communication between the continuous availability at site 1 and the DR at site 2 can be achieved by using technologies such as IBM PowerHA SystemMirror for use with MM or GM with IASPs, full system replication, and logical replication from a third-party vendor.

Db2 Mirror and full system replication

The usage of a mirrored pair within the disaster site provides extra protection if you are required to role swap to the DR location. With this scenario, a continuously available environment exists, even in DR.

A topology with multiple IBM Storage Virtualize storage systems and multiple IBM Power servers is shown in Figure A-20.

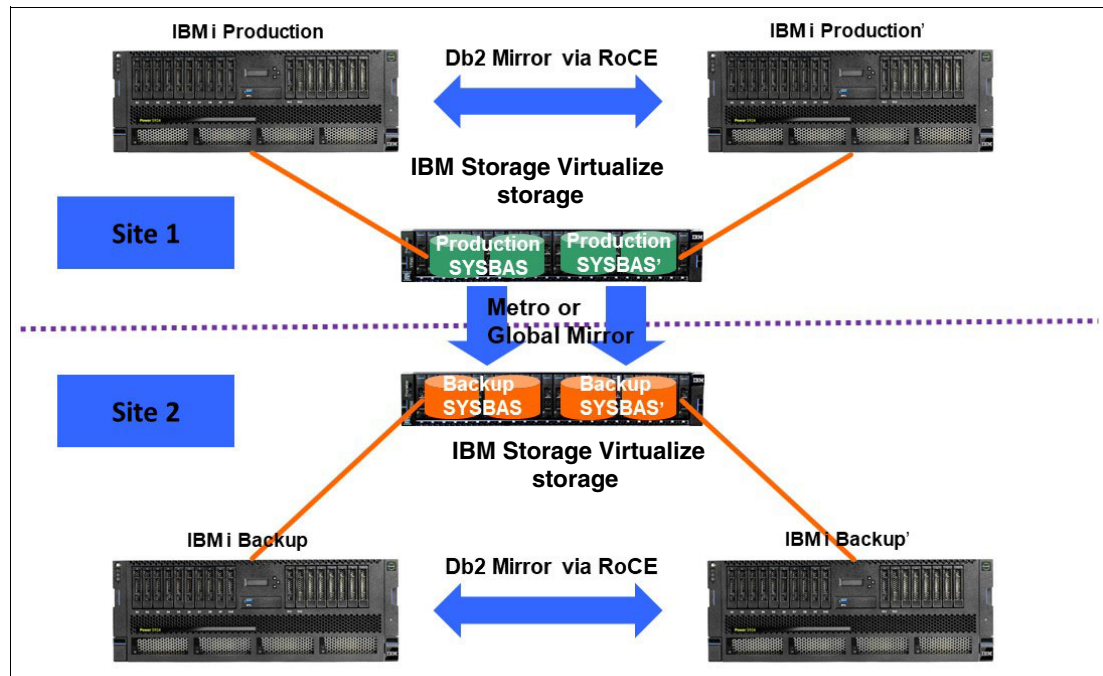


Figure A-20 Db2 Mirror and full system replication

Full system replication is fully supported. If you are not using IASP, this type of replication can be done for IBM i at the IBM Storage Virtualize storage level.

At site 1, an active side exists because of full system replication. However, at site 2, the IBM i systems are powered off, and the replication is active across sites.

Two copies are at a DR location because if one side fails, the other side must continue replicating. If only three nodes are replicating, you cannot predict which side fails and does not have a valid copy of the storage data to switch.

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this book.

IBM Redbooks

The following IBM Redbooks publications provide additional information about the topic in this document. Note that some publications referenced in this list might be available in softcopy only.

- ▶ *Implementation Guide for IBM Storage FlashSystem and IBM SAN Volume Controller: Updated for IBM Storage Virtualize Version 8.6*, SG24-8542
- ▶ *Do More with Less: Automating IBM Storage FlashSystem Tasks with REST APIs, Scripting, and Ansible*, REDP-5736.
- ▶ *IBM Storage Virtualize and VMware: Integrations, Implementation and Best Practices*, SG24-8549
- ▶ *Policy-Based Replication with IBM Storage FlashSystem, IBM SAN Volume Controller and IBM Storage Virtualize*, REDP-5704
- ▶ *IBM Spectrum Virtualize 3-Site Replication*, SG24-8504
- ▶ *Introduction and Implementation of Data Reduction Pools and Deduplication*, SG24-8430
- ▶ *IBM Storage as a Service (STaaS) Offering Guide*, REDP-5644
- ▶ *IBM FlashSystem Safeguarded Copy Implementation Guide*, REDP-5654
- ▶ *Automate and Orchestrate Your IBM FlashSystem Hybrid Cloud with Red Hat Ansible*, REDP-5598
- ▶ *IBM Spectrum Virtualize and SAN Volume Controller Enhanced Stretched Cluster with VMware*, SG24-8211
- ▶ *IBM Storwize V7000, Spectrum Virtualize, HyperSwap, and VMware Implementation*, SG24-8317
- ▶ *IBM Storage Virtualize, IBM Storage FlashSystem, and IBM SAN Volume Controller Security Feature Checklist*, REDP-5717
- ▶ *IBM Spectrum Virtualize HyperSwap SAN Implementation and Design Best Practices*, REDP-5597
- ▶ *IBM SAN Volume Controller Stretched Cluster with PowerVM and PowerHA*, SG24-8142
- ▶ *Implementing IBM FlashSystem 900*, SG24-8271

You can search for, view, download or order these documents and other Redbooks, Redpapers, Web Docs, draft and additional materials, at the following website:

ibm.com/redbooks

Online resources

These websites are also relevant as further information sources:

- ▶ IBM Redbooks Storage videos
<https://www.redbooks.ibm.com/feature/storagevideos>

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services



Redbooks

Performance and Best Practices Guide for IBM Storage

SG24-8543-00

ISBN 0738461644

(1.5" spine)
1.5" <-> 1.998"
789 <-> 1051 pages



Redbooks

Performance and Best Practices Guide for IBM Storage

SG24-8543-00

ISBN 0738461644

(1.0" spine)
0.875" <-> 1.498"
460 <-> 788 pages



Redbooks

Performance and Best Practices Guide for IBM Storage

SG24-8543-00

ISBN 0738461644

(0.5" spine)
0.475" <-> 0.873"
250 <-> 459 pages



Redbooks

Performance and Best Practices Guide for IBM Storage FlashSystem and

(0.2" spine)

0.17" <-> 0.473"

90 <-> 249 pages

(0.1" spine)

0.1" <-> 0.169"

53 <-> 89 pages



Performance and Best Practices Guide for IBM

SG24-8543-00

ISBN 0738461644

(2.5" spine)
2.5" <-> mmm.n"
1315 <-> mmm pages



Performance and Best Practices Guide for IBM Storage FlashSystem and IBM SAN

SG24-8543-00

ISBN 0738461644

(2.0" spine)
2.0" <-> 2.498"
1052 <-> 1314 pages





SG24-8543-00

ISBN 0738461644

Printed in U.S.A.

Get connected

