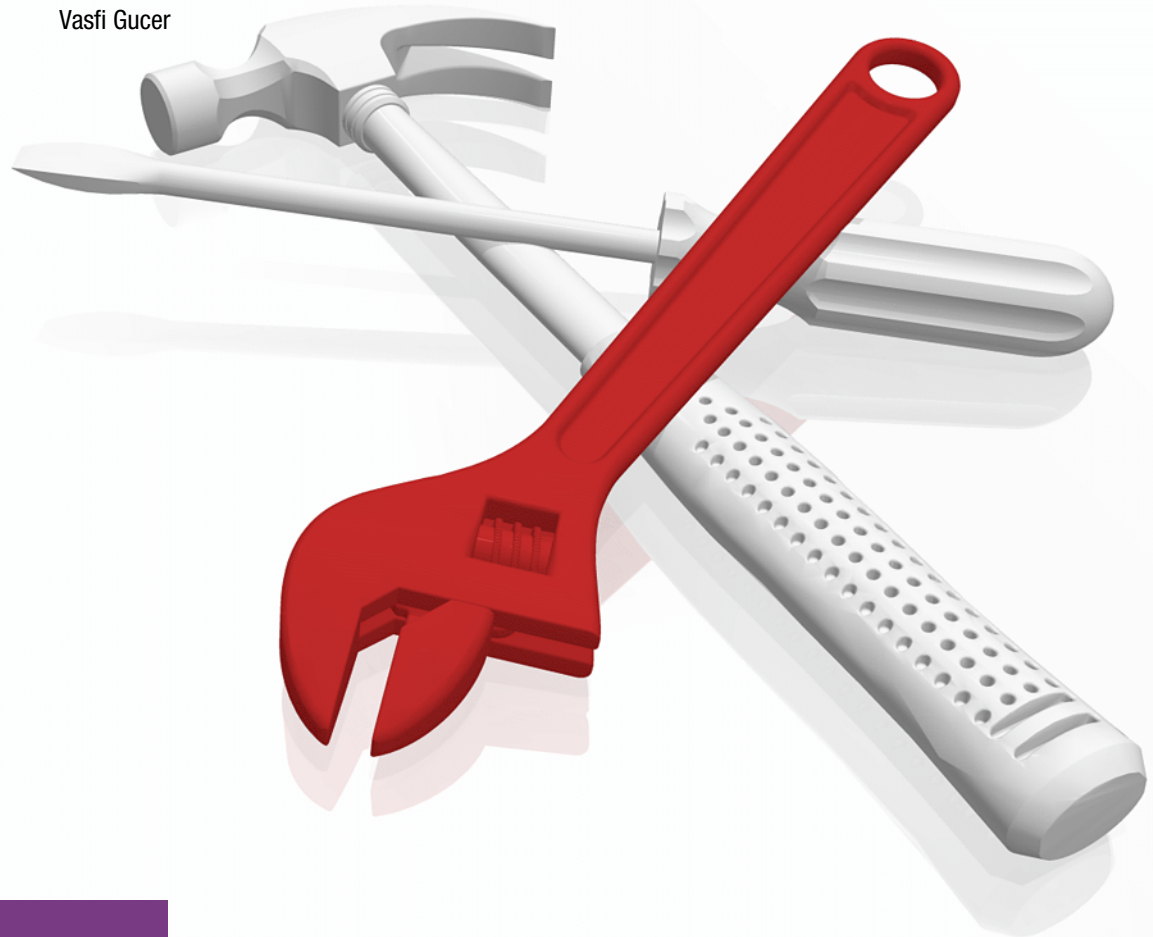


IBM FlashSystem Best Practices and Performance Guidelines

Anil K Nayak
Antonio Rainero
Barry Whyte
Chris Hoffmann
Danilo Morelli Miyasiro
David Green
Duane Bolland
Jackson Shea
Jon Herd
Jordan Fincher
Marcelo Avalos Del Carpio
Sergey Kubin
Sidney Varoni Junior
Thales Noivo Ferreira

Vasfi Gucer



Storage



IBM Redbooks

**IBM FlashSystem Best Practices and Performance
Guidelines**

May 2021

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

First Edition (May 2021)

This edition applies to IBM Spectrum Virtualize Version 8.4.

© Copyright International Business Machines Corporation 2021. 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	xi
Tables	xvii
Examples	xix
Notices	xxi
Trademarks	xxii
Preface	xxiii
Authors	xxiii
Now you can become a published author, too!	xxvii
Comments welcome	xxvii
Stay connected to IBM Redbooks	xxviii
Chapter 1. IBM FlashSystem introduction	1
1.1 IBM FlashSystem supported product range	2
1.1.1 What's new in V8.4	2
1.1.2 Products supported	3
1.2 IBM FlashSystem high-level features	4
1.3 IBM Storwize and IBM FlashSystem product range	6
1.3.1 Clustering rules and upgrades	25
1.3.2 Mixed clustering rules and licensing	26
1.3.3 IBM FlashSystem 9200R Rack Solution overview	26
1.4 Advanced functions for data reduction	28
1.4.1 FlashCore Modules (FCM)	28
1.4.2 Data reduction pools (DRP)	28
1.4.3 Deduplication	29
1.4.4 Thin provisioning	29
1.4.5 Thin-provisioned FlashCopy snapshots	29
1.5 Advanced software features	29
1.5.1 Data migration	29
1.5.2 Copy services	30
1.5.3 Easy Tier	30
1.5.4 External virtualization	31
1.5.5 IBM HyperSwap	31
1.5.6 Licensing	31
Chapter 2. Storage area network	33
2.1 SAN topology general guidelines	34
2.1.1 SAN performance and scalability	34
2.1.2 ISL considerations	35
2.2 SAN topology-specific guidelines	36
2.2.1 Single switch SANs	36
2.2.2 Basic core-edge topology	37
2.2.3 Edge-core-edge topology	38
2.2.4 Full MeSH topology	39
2.2.5 IBM FlashSystem as a SAN bridge	40
2.2.6 Device placement	41

2.2.7 SAN partitioning	43
2.3 IBM FlashSystem 9200 controller ports	44
2.3.1 Slots and ports identification	44
2.3.2 Port naming and distribution	45
2.4 Zoning	47
2.4.1 Types of zoning	47
2.4.2 Pre-zoning tips and shortcuts	49
2.4.3 IBM FlashSystem 9200 internode communications zones	50
2.4.4 IBM FlashSystem 9200 host zones	51
2.5 Distance extension for Remote Copy services	55
2.5.1 Optical multiplexors	56
2.5.2 Long-distance SFPs or XFPs	56
2.5.3 Fibre Channel over IP	56
2.5.4 SAN extension with Business Continuity configurations	57
2.5.5 Native IP replication	59
2.6 Tape and disk traffic that share the SAN	61
2.7 Switch interoperability	61
Chapter 3. Storage backend	63
3.1 Internal storage types	64
3.1.1 NVMe storage	64
3.1.2 SAS drives	67
3.1.3 Internal storage considerations	68
3.2 Arrays	72
3.2.1 Supported RAID types	72
3.2.2 Array considerations	73
3.2.3 Compressed array monitoring	76
3.3 General external storage considerations	78
3.3.1 Storage controller path selection	78
3.3.2 Guidelines for creating optimal backend configuration	80
3.3.3 Considerations for compressing and deduplicating back-end	82
3.4 Controller-specific considerations	83
3.4.1 Considerations for DS8000 series	83
3.4.2 Considerations for IBM XIV Storage System	91
3.4.3 Considerations for IBM FlashSystem A9000/A9000R	93
3.4.4 Considerations for FlashSystem 5000, 5100, 7200, 9100, and 9200	95
3.4.5 Considerations for IBM FlashSystem 900	97
3.4.6 Path considerations for third-party storage with EMC VMAX and Hitachi Data Systems	98
3.5 Quorum disks	99
Chapter 4. Storage pools	101
4.1 Introduction to pools	102
4.1.1 Standard pool	102
4.1.2 Data reduction pools	107
4.1.3 Standard pools versus data reduction pools	113
4.1.4 Data reduction estimation tools	115
4.1.5 Understanding capacity use in a data reduction pool	120
4.2 Storage pool planning considerations	123
4.2.1 Planning for availability	123
4.2.2 Planning for performance	124
4.2.3 Planning for capacity	126
4.2.4 Extent size considerations	127

4.2.5 External pools	128
4.3 Data reduction pools best practices	130
4.3.1 Data reduction pools with IBM FlashSystem NVMe attached drives	130
4.3.2 DRP and external storage considerations.	131
4.3.3 Data reduction pools and VMware vVols	132
4.3.4 Data reduction pool configuration limits	132
4.3.5 DRP provisioning considerations	133
4.3.6 Standard and DRP pools coexistence	135
4.3.7 Data migration with DRP.	135
4.4 Operations with storage pools.	137
4.4.1 Creating data reduction pools	137
4.4.2 Adding external MDisks to existing storage pools.	140
4.4.3 Renaming MDisks.	141
4.4.4 Removing MDisks from existing storage pools	142
4.4.5 Remapping managed MDisks.	146
4.4.6 Controlling extent allocation order for volume creation.	147
4.5 Considerations when using encryption	148
4.5.1 General considerations.	148
4.5.2 Hardware and software encryption	149
4.5.3 Encryption at rest with USB keys	151
4.5.4 Encryption at rest with key servers	152
4.6 Easy Tier, tiered and balanced storage pools.	159
4.6.1 Easy Tier concepts	159
4.6.2 Easy Tier definitions	161
4.6.3 Easy Tier operating modes	163
4.6.4 MDisk tier types	166
4.6.5 Changing the tier type of an MDisk.	170
4.6.6 Easy Tier overload protection	172
4.6.7 Removing an MDisk from an Easy Tier pool.	172
4.6.8 Easy Tier implementation considerations	173
4.6.9 Easy Tier settings	174
4.6.10 Monitoring Easy Tier using the GUI	180
Chapter 5. Volumes	187
5.1 Overview of volumes	188
5.2 Guidance for creating volumes	188
5.3 Thin-provisioned volumes	192
5.3.1 Compressed volumes	195
5.3.2 Deduplicated volumes.	196
5.3.3 Thin-provisioning considerations.	198
5.3.4 Limits on virtual capacity of thin-provisioned volumes	200
5.4 Mirrored volumes	200
5.4.1 Write fast failovers	203
5.4.2 Read fast failovers	204
5.4.3 Maintaining data integrity of mirrored volumes	204
5.5 HyperSwap volumes	205
5.6 VMware virtual volumes	206
5.7 Cloud volumes	209
5.7.1 Transparent cloud tiering configuration limitations and rules	210
5.7.2 Restore to the production volume	211
5.7.3 Restore to a new volume	211
5.8 Volume migration	212
5.8.1 Image-type to striped-type volume migration	212

5.8.2	Migrating to image-type volume	213
5.8.3	Migrating with volume mirroring	214
5.8.4	Migration from Standard Pool to Data Reduction Pool	215
5.9	Preferred paths to a volume	216
5.10	Moving a volume between I/O groups and nodes	217
5.10.1	Changing the preferred node of a volume within an I/O group	217
5.10.2	Moving a volume between I/O groups	218
5.11	Volume throttling	218
5.12	Volume cache mode	221
5.13	Additional considerations	224
5.13.1	Volume protection	224
5.13.2	Volume resizing	225
5.13.3	Migrating from Fibre Channel connections to RDMA over Ethernet connections between nodes	227
Chapter 6.	Copy services	229
6.1	Introduction to copy services	230
6.1.1	FlashCopy	230
6.1.2	Metro Mirror and Global Mirror	230
6.1.3	Volume Mirroring	230
6.2	FlashCopy	231
6.2.1	FlashCopy use cases	231
6.2.2	FlashCopy capabilities overview	233
6.2.3	FlashCopy functional overview	239
6.2.4	FlashCopy planning considerations	247
6.3	Remote Copy services	257
6.3.1	Remote Copy use cases	258
6.3.2	Remote Copy functional overview	258
6.3.3	Remote Copy network planning	274
6.3.4	Remote Copy services planning	288
6.3.5	Multiple site remote copy	299
6.3.6	1920 error	302
6.4	Native IP replication	314
6.4.1	Native IP replication technology	314
6.4.2	IP partnership limitations	316
6.4.3	VLAN support	317
6.4.4	IP Compression	318
6.4.5	Remote Copy groups	319
6.4.6	Supported configurations examples	320
6.4.7	Native IP replication performance consideration	329
6.5	Volume Mirroring	330
6.5.1	Read and write operations	331
6.5.2	Volume mirroring use cases	332
6.5.3	Mirrored volume components	334
6.5.4	Volume Mirroring synchronization options	334
6.5.5	Volume Mirroring performance considerations	335
6.5.6	Bitmap space for out-of-sync volume copies	337
Chapter 7.	Business continuity	339
7.1	Business continuity with HyperSwap	340
7.2	Third site and IP quorum	344
7.2.1	Quorum modes	345
7.3	HyperSwap Volumes	346

7.4 Other considerations and general recommendations	348
Chapter 8. Hosts	349
8.1 General configuration guidelines	350
8.1.1 Number of paths	350
8.1.2 Host ports	350
8.1.3 Port masking	350
8.1.4 N-port ID virtualization (NPIV)	350
8.1.5 Host to I/O group mapping	350
8.1.6 Volume size as opposed to quantity	351
8.1.7 Host volume mapping	351
8.1.8 Server adapter layout	352
8.1.9 Host status improvements	352
8.1.10 Considerations for NVMe over Fibre Channel host attachments	352
8.1.11 Considerations for iSER host attachments	352
8.2 Host pathing	353
8.2.1 Path selection	353
8.3 I/O queues	353
8.3.1 Queue depths	354
8.4 Host clusters	354
8.4.1 Persistent reservations	356
8.4.2 Clearing reserves	357
8.5 AIX hosts	357
8.5.1 Multipathing support	357
8.5.2 Configuration recommendations for AIX	357
8.6 Virtual I/O server hosts	358
8.6.1 Multipathing support	358
8.6.2 Physical and logical volumes	358
8.6.3 Methods to identify a disk for use as a virtual SCSI disk	358
8.7 Windows hosts	358
8.7.1 Multipathing support	358
8.7.2 Windows configuration	359
8.8 Linux hosts	359
8.9 Oracle Solaris hosts	359
8.9.1 Solaris MPxIO	359
8.9.2 Symantec Veritas Volume Manager	360
8.9.3 DMP multipathing	360
8.10 VMware ESXi server hosts	361
8.10.1 Configuring VMware	361
8.10.2 Multipathing configuration maximums	361
Chapter 9. Monitoring	363
9.1 Generic monitoring	364
9.1.1 Monitoring with the GUI	364
9.1.2 Monitoring using quotas and alert	366
9.2 Performance monitoring	367
9.2.1 Performance monitoring with the GUI	367
9.2.2 Performance monitoring with IBM Spectrum Control	369
9.2.3 Performance monitoring with IBM Storage Insights	373
9.3 Capacity metrics for block storage systems	387
9.3.1 Storage system capacity metrics	388
9.3.2 Pool capacity metrics	394
9.3.3 Volume capacity metrics	399

9.4	Creating alerts for IBM Spectrum Control and IBM Storage Insights	400
9.4.1	Alert examples	401
9.4.2	Alert to monitor back-end capacity: Available Physical Space (%)	401
9.5	Error condition example with IBM Spectrum Control: FC port	409
9.6	Important metrics	411
9.7	Performance support package	412
9.8	Metro and Global Mirror monitoring with IBM Copy Services Manager and scripts	414
9.8.1	Monitoring MM and GM with scripts	415
9.9	Monitoring Tier1 SSD	415
Chapter 10.	Maintenance	417
10.1	User interfaces	419
10.1.1	Management GUI	419
10.1.2	Service Assistant Tool GUI	419
10.1.3	Command line interface	420
10.2	Users and groups	421
10.3	Volumes	423
10.4	Hosts	424
10.5	Software updates	425
10.5.1	Deciding the target software level	426
10.5.2	Hardware considerations	427
10.5.3	Update sequence	428
10.5.4	SAN fabrics preparation	428
10.5.5	Storage controllers preparation	429
10.5.6	Hosts preparation	429
10.5.7	Copy services considerations	429
10.5.8	Running the Upgrade Test Utility	430
10.5.9	Updating the software	432
10.6	Drive firmware updates	435
10.7	SAN modifications	437
10.7.1	Cross-referencing WWPN	437
10.7.2	Cross-referencing LUN ID	439
10.8	Server HBA replacement	439
10.9	Hardware upgrades	440
10.9.1	Adding control enclosures	441
10.9.2	Upgrading nodes in an existing cluster	444
10.9.3	Upgrading NVMe drives	450
10.9.4	Moving to a new IBM FlashSystem cluster	450
10.9.5	Splitting an IBM FlashSystem cluster	451
10.9.6	Adding expansion enclosures	452
10.9.7	Removing expansion enclosures	456
10.9.8	IBM FlashWatch	457
10.10	I/O Throttling	459
10.10.1	General information on I/O throttling	459
10.10.2	I/O throttling on front-end I/O control	459
10.10.3	I/O Throttling on back-end I/O control	460
10.10.4	Overall benefits of using I/O throttling	460
10.10.5	Considerations for I/O throttling	461
10.10.6	Configuring I/O throttling using the CLI	461
10.10.7	Configuring I/O throttling using the GUI	462
10.10.8	Creating a volume throttle	462
10.10.9	Creating a host throttle	463
10.10.10	Creating a host cluster throttle	463

10.10.11	Creating a storage pool throttle	464
10.10.12	Creating an offload throttle	464
10.11	Automation	465
10.11.1	Red Hat Ansible	465
10.11.2	RESTful API	466
10.12	Documenting IBM FlashSystem and SAN environment	467
10.12.1	Naming conventions	468
10.12.2	SAN fabric documentation	471
10.12.3	IBM FlashSystem documentation	473
10.12.4	Storage documentation	475
10.12.5	Technical support information	476
10.12.6	Tracking incident and change tickets	477
10.12.7	Automated support data collection	478
10.12.8	Subscribing to IBM FlashSystem support	478
Chapter 11.	Troubleshooting and diagnostics	479
11.1	Starting troubleshooting	480
11.1.1	Recommended actions and fix procedure	482
11.2	Diagnostic data collection	485
11.2.1	IBM FlashSystem data collection	485
11.2.2	Host multipath software data collection	487
11.2.3	Additional data collection	488
11.3	Common problems and isolation techniques	489
11.3.1	Host problems	490
11.3.2	SAN problems	494
11.3.3	Storage subsystem problems	495
11.3.4	Native IP replication problems	500
11.3.5	Remote Direct Memory Access based clustering	501
11.3.6	Advanced Copy services problems	501
11.3.7	Health status during upgrade	503
11.3.8	Managing physical capacity of over provisioned storage controllers	503
11.3.9	Replacing a failed flash drive	505
11.3.10	Recovering from common events	505
11.4	Remote Support Assistance	506
11.5	Call Home Connect Cloud and Health Checker feature	507
11.5.1	Health Checker	509
11.6	IBM Storage Insights	509
11.6.1	Storage Insights Customer Dashboard	512
11.6.2	Customized dashboards to monitor your storage	512
11.6.3	Creating support tickets	512
11.6.4	Updating support tickets	520
11.6.5	SI Advisor	523
Appendix A.	IBM i considerations	525
	IBM i Storage management	526
	Single-level storage	527
	IBM i response time	529
	Planning for IBM i storage capacity	532
	Storage connection to IBM i	533
	Native attachment	534
	VIOS attachment	535
	Setting of attributes in VIOS	539
	FC adapter attributes	539

Disk device attributes	539
Guidelines for Virtual I/O Server resources.	540
Disk drives for IBM i	540
Defining LUNs for IBM i	543
Data layout	544
Fibre Channel adapters in IBM i and VIOS	545
Zoning SAN switches	546
IBM i Multipath	546
Boot from SAN	547
IBM i mirroring	547
Copy services considerations	548
Remote replication	548
FlashCopy	549
HyperSwap	550
Db2 mirroring for IBM i	552
Related publications	559
IBM Redbooks	559
Online resources	559
Help from IBM	560

Figures

1-1 IBM FlashSystem and IBM Storwize products that support Spectrum Virtualize software v8.4.	2
1-2 LFF expansion enclosure	5
1-3 SFF expansion enclosure	6
1-4 LFF HD expansion enclosure	6
1-5 IBM Storwize V5100 front view	7
1-6 IBM Storwize V7000 Generation 2 (2076-524) and Generation2+ (2076-624) SFF.	9
1-7 IBM Storwize V7000 Generation 3 (2076-724)	9
1-8 IBM FlashSystem 5015 and 5035 LFF control enclosure front view.	12
1-9 IBM FlashSystem 5015 and 5035 SFF control enclosure front view	13
1-10 IBM FlashSystem 5100 front view showing the 24 NVMe drives installed	16
1-11 IBM FlashSystem 5200 control enclosure front and 3/4 ISO view	18
1-12 IBM FlashSystem 7200 control enclosure front view	20
1-13 IBM FlashSystem 9100 control enclosure with one NVMe drives partially removed	21
1-14 IBM FlashSystem 9200 control enclosure	23
2-1 ISL data flow	36
2-2 Single Switch Topology	37
2-3 Core/Edge Topology	38
2-4 Edge-Core-Edge Topology	39
2-5 Full MeSH topology.	40
2-6 FlashSystem as a SAN Bridge	41
2-7 Storage and hosts attached to the same SAN switch.	42
2-8 Segregation using edge-core-edge.	43
2-9 Port location in FlashSystem 9200 rear view	44
2-10 FlashSystem 9200 port distribution.	45
2-11 Port masking configuration on FlashSystem 9200	46
2-12 Typical host to FlashSystem 9200 zoning	51
2-13 Four port host zoning	52
2-14 ESX Cluster zoning.	53
2-15 LPARs SAN connections	54
2-16 Live partition migration	55
2-17 Configuration 1: Physical Paths Shared Among Fabrics	58
2-18 Configuration 2: physical paths not shared among the fabrics	59
2-19 Effect of distance on packet loss	60
3-1 FCM capacity monitoring with GUI	70
3-2 Array capacity monitoring with GUI.	77
3-3 SSIC example.	79
3-4 DS8900 virtualization concepts focus to IBM FlashSystem	85
3-5 DA pair reduced bandwidth configuration	86
3-6 DA pair correct configuration	87
3-7 The lsarray and lsrank command output.	88
3-8 Four DS8900F extent pools as one IBM FlashSystem storage pool	91
3-9 XIV rack configuration: 281x-214	92
4-1 Standard and data reduction pool - volumes	111
4-2 Garbage Collection principle.	112
4-3 Capacity savings analysis.	116
4-4 Customized view.	117
4-5 Data reduction pool capacity use example	121

4-6	Example dashboard capacity view	126
4-7	Compression Savings dashboard report.	127
4-8	Create pool page	137
4-9	Right-click parent pool actions menu	138
4-10	Create child pool page	138
4-11	Create Volume page	139
4-12	The Create Mapping page	140
4-13	FlashSystem volume details	145
4-14	FlashSystem volume details for host maps	145
4-15	IBM SAN Volume Controller MDisk details for IBM FlashSystem volumes	146
4-16	Encryption placement in lower layers of the IBM FlashSystem software stack	149
4-17	Mixed encryption in a storage pool	150
4-18	Update certificate on IBM FlashSystem	154
4-19	Create self-signed certificate on IBM Security Key Lifecycle Manager server	155
4-20	Create device group for IBM FlashSystem	155
4-21	SKLM Replication Schedule	156
4-22	Keys associated to a device group	157
4-23	Easy Tier single volume, multiple tiers	160
4-24	Easy Tier extent migration types.	163
4-25	Single tier storage pool with striped volume	167
4-26	Multitier storage pool with striped volume.	168
4-27	Change the MDisk tier	171
4-28	Select wanted MDisk tier	171
4-29	Customizing the title row to show the tier column.	172
4-30	Easy Tier Data Movement page	181
4-31	Easy Tier Movement description page	182
4-32	Easy Tier - single tier pool - composition report page.	183
4-33	Easy tier - multi-tier pool - composition page	184
4-34	Workload skew - single tier pool	185
4-35	Workload skew - multi-tier configuration	185
5-1	Volumes format option	189
5-2	Thin-provisioned volume.	192
5-3	Different kinds of volumes in DRP	192
5-4	Conceptual diagram of thin-provisioned volume.	193
5-5	Modifying capacity savings of a volume nondisruptively.	195
5-6	Customized view.	196
5-7	Creating deduplicated volumes.	197
5-8	Mirrored volume creation	202
5-9	Adding a volume copy.	203
5-10	Overall HyperSwap diagram.	205
5-11	Master and Auxiliary volumes.	206
5-12	Overview of the key components of VMware environment.	207
5-13	VMware vSphere Storage APIs Array Integration (VAAI)	207
5-14	Enable VVOL window	208
5-15	Cloud volumes - Transparent Cloud Tiering	210
5-16	Migration with Volume Mirroring	215
5-17	Converting volumes with Volume Mirroring.	216
5-18	Write operations from a host.	217
5-19	Volume throttling for each LUN.	218
5-20	Volume Throttling	219
5-21	Edit bandwidth and IOPS limit	219
5-22	Cache activated	221
5-23	Cache deactivated	222

5-24	Edit cache mode	223
5-25	Volume Protection.	225
5-26	Expanding a volume	226
5-27	Shrinking a volume	226
6-1	FlashCopy mapping	233
6-2	Multiple volumes mapping in a Consistency Group	235
6-3	Incremental FlashCopy	236
6-4	Multiple Target FlashCopy	237
6-5	Cascaded FlashCopy	237
6-6	FlashCopy mapping states diagram	241
6-7	New cache architecture	244
6-8	Logical placement of the FlashCopy indirection layer.	245
6-9	Interaction between Multiple Target FlashCopy mappings	246
6-10	GUI Flashcopy Presets.	250
6-11	Remote Copy components and applications	260
6-12	Remote Copy partnership.	261
6-13	Role and direction changes	262
6-14	Conceptualization of layers.	264
6-15	Supported topologies for Remote Copy partnerships	265
6-16	Metro Mirror write sequence	267
6-17	Global Mirror relationship write operation	269
6-18	Colliding writes	271
6-19	Global Mirror with Change Volumes	272
6-20	Global Mirror with Change Volumes uses FlashCopy point-in-time copy technology	273
6-21	Standard SCSI read operation	275
6-22	Standard SCSI write operation	276
6-23	Spectrum Virtualize remote copy write	277
6-24	Zoning scheme for >80 ms Remote Copy partnerships	285
6-25	Typical Remote Copy network configuration	286
6-26	Configuration 1: physical paths shared among the fabrics	287
6-27	Configuration 2: physical paths not shared among the fabrics	288
6-28	Remote Copy resources that are not optimized	293
6-29	Optimized Global Mirror resources	293
6-30	Using three-way copy services	300
6-31	Cascading-like infrastructure	301
6-32	Effect of packet size (in bytes) versus the link size.	312
6-33	Typical Ethernet network data flow	315
6-34	Optimized network data flow by using Bridgeworks SANSlide technology.	315
6-35	Only one Remote Copy group on each system and nodes with failover ports configured	321
6-36	Multinode systems single inter-site link with only one Remote Copy port group	322
6-37	Multinode systems single inter-site link with only one Remote Copy port group	323
6-38	Dual links with two Remote Copy groups on each system configured	324
6-39	Multinode systems with dual inter-site links between the two systems.	326
6-40	Multinode systems with dual inter-site links between the two systems.	328
6-41	1 Gbps port throughput trend	330
6-42	Volume Mirroring overview	331
7-1	Typical HyperSwap configuration with IBM FlashSystem	341
7-2	IBM FlashSystem HyperSwap in a storage failure scenario	342
7-3	IBM FlashSystem HyperSwap in a site failure scenario	343
7-4	IP Quorum network layout	344
7-5	HyperSwap Volume	346
8-1	SCSI ID assignment on volume mappings	355

9-1	Email notification options	364
9-2	Syslog Configuration	366
9-3	Monitoring GUI example	368
9-4	Selecting metrics	368
9-5	System -> Dashboard	369
9-6	Change day of reference	370
9-7	Metric exceeding best practice	371
9-8	Changed chart due to iogrp selection	371
9-9	DashBoard- Key Performance Indicators	372
9-10	IBM Storage Insight	374
9-11	IBM Storage Insight registration screen	382
9-12	Choose IBM Storage Insights or IBM Storage Insights for Spectrum Control	382
9-13	Registration login screen	383
9-14	Create an IBM account	384
9-15	Registration - ID and password	384
9-16	IBM Storage Insights registration form	385
9-17	Understanding capacity information	388
9-18	Allocated Space	389
9-19	Assigned Volume Space (GiB)	389
9-20	Assigned Volume Space - graph	390
9-21	Physical Allocation (%)	391
9-22	Shortfall	392
9-23	Zero Capacity trend	394
9-24	Physical Allocation	396
9-25	Create new Alert Policy	402
9-26	Default Alert Policy	403
9-27	Copy existing Policy and create a new one	403
9-28	Store copied policy	404
9-29	New Policy with inherited Alert Definitions	404
9-30	Alert Definition RAID Array > Capacity	405
9-31	Alert Definition 15% or less Available Physical Space - Critical	406
9-32	Alert Definition - Critical - change frequency	406
9-33	Alert Definition 20% or less Available Physical Space - Warning	407
9-34	Alert Definition 30% or less Available Physical Space - Notification	407
9-35	Change Notification Settings	408
9-36	Notification Settings: Details	408
9-37	Error condition spotted on a storage subsystem with IBM Spectrum Control	409
9-38	FlashSystem error condition - internal resources - ports	410
9-39	FC ports stopped: detail view	410
9-40	Performance support package creation	413
9-41	Package files example	413
10-1	Restricted Dashboard view for a user in an ownership group	423
10-2	Update System output panel	426
10-3	Fix Central software packages	427
10-4	IBM FlashSystem Upgrade Test Utility using the GUI	430
10-5	IBM FlashSystem Upgrade Test Utility completion panel	431
10-6	Drive firmware upgrade	436
10-7	IBM FlashSystem performance statistics (IOPS)	445
10-8	Cabling for adding four expansion enclosures in two SAS chains	453
10-9	Distribution of controller resources before and after I/O throttling	461
10-10	Creating a volume throttle in the GUI	462
10-11	Creating a host throttle in the GUI	463
10-12	Creating a host cluster throttle in the GUI	463

10-13	Creating a storage pool throttle in the GUI	464
10-14	Creating system offload throttle in the GUI	464
10-15	A poorly formatted SAN diagram	472
10-16	Brocade SAN Health Options window	472
10-17	Creating a subscription to IBM FlashSystem 9200 notifications	478
11-1	Events icon in GUI	481
11-2	Dashboard showing system health	481
11-3	System Health expanded section in dashboard	482
11-4	Monitoring > Events panel	483
11-5	Properties and sense data for event window	484
11-6	<i>Upload Support Package</i> panel	486
11-7	Remote Support options	506
11-8	Call Home Web	508
11-9	Call Home Web details panel	509
11-10	Storage Insights versus Storage Insights Pro comparison	511
11-11	Storage Insights Main Dashboard	512
11-12	SI Create / Update a support Ticket	513
11-13	Create ticket	514
11-14	Collecting ticket information	515
11-15	Adding problem description and any additional information	516
11-16	Set severity level	517
11-17	Review the ticket information	518
11-18	Final summary before ticket creation	519
11-19	Ticket summary	520
11-20	SI Update Ticket	521
11-21	Entering the PMR ticket number	522
11-22	Log type selection	523
11-23	SI Advisor menu	524
11-24	Advisor detailed summary of recommendations	524
A-1	IBM i storage management spreads objects across LUNs	526
A-2	Virtual address space	527
A-3	IBM i auxiliary storage pools architecture	528
A-4	TIMI atomicity	529
A-5	Disk subsystem	531
A-6	Disk I/O on IBM i	531
A-7	IBM i with different sector sizes	533
A-8	IBM i SAN access using NPIV	536
A-9	Diagram of sizing and modeling for IBM i using Disk Magic	542
A-10	SAN switch zoning for IBM i with IBM Spectrum Virtualize storage	546
A-11	IBM i full system replication with IBM Spectrum Virtualize	548
A-12	IBM i IASP replication with IBM Spectrum Virtualize	549
A-13	IBM i HyperSwap SAN fabric connection example	551
A-14	Db2 Mirror environment with one IBM Spectrum Virtualize storage	556
A-15	Db2 Mirror environment with two IBM Spectrum Virtualize storages	557
A-16	DB2 Mirror and full system replication	558

Tables

1-1 IBM Storwize V5100 host, drive capacity, and functions summary	8
1-2 IBM Storwize V7000 host, drive capacity and functions Summary	10
1-3 Machine type and model comparison for the IBM FlashSystem 5000	13
1-4 IBM FlashSystem 5015 host, drive capacity and functions summary	13
1-5 2.5 inch supported drives for the IBM FlashSystem 5000	14
1-6 3.5 inch supported drives for the IBM FlashSystem 5000	14
1-7 IBM FlashSystem 5035 host, drive capacity and functions summary	15
1-8 IBM FlashSystem 5100 host, drive capacity, and functions summary	16
1-9 IBM FlashSystem 5200 host, drive capacity, and functions summary	18
1-10 IBM FlashSystem 7200 host, drive capacity and functions summary	20
1-11 IBM FlashSystem 9100 host, drive capacity, and functions summary	22
1-12 IBM FlashSystem 9200 host, drive capacity, and functions summary	24
1-13 Clustering control nodes matrix	25
1-14 IBM FlashSystem 9200R Rack Solution combinations	27
2-1 FlashSystem 9200	44
2-2 Alias names examples	49
2-3 Template examples	50
3-1 FlashCore module capacities	66
3-2 Supported SCM drive capacities	67
3-3 Maximum number of drive slots per SAS expansion chain	68
3-4 Supported RAID levels	72
3-5 XIV minimum volume size and quantity recommendations	93
3-6 Host connections for A9000	94
3-7 Host connections for A9000R	94
4-1 Compression ratios of common data types	114
4-2 DRP Capacity Uses	121
4-3 Capacity Terminology in 8.4.0	126
4-4 Data reduction pool properties	132
4-5 Pool size by extent size and IO group number	134
4-6 Minimum recommended pool size by extent size and IO group number	134
4-7 Easy Tier settings	165
4-8 Recommended 3-tier Easy Tier mapping policy	168
4-9 4 and 5 Tier mapping policy4 and 5 Tier mapping policy	169
4-10 Unsupported temporary 4 and 5 Tier mapping policy	169
4-11 Migration target tier priorities	173
5-1 Maximum number of volumes in IBM FlashSystem	190
5-2 Migration types and associated commands	212
5-3 Sample synccrate values	214
6-1 Relationship between the rate and data rate per second	234
6-2 Summary table of the FlashCopy indirection layer algorithm	243
6-3 FlashCopy properties and maximum configurations	247
6-4 Relationship of bitmap space to FlashCopy address space for the specified I/O Group	248
6-5 Workload distribution for back-end I/O operations	254
6-6 Maximum round trip	277
6-7 IBM Spectrum Virtualize intersystem heartbeat traffic (megabits per second)	278
6-8 Remote Copy maximum limits	289
6-9 IP replication limits	317

6-10	Relationship between the rate value and the data copied per second	335
6-11	Relationship of bitmap space to Volume Mirroring address space	337
9-1	Features in IBM Storage Insights and IBM Storage Insights Pro	376
9-2	Feature comparison	379
9-3	Event examples for IBM Flash System	401
9-4	Alert severities	402
9-5	Field changes to drive and array devices	415
10-1	UNIX commands available in the CLI	420
10-2	Available memory configuration for one node in a control enclosure	445
10-3	Base memory features	446
10-4	Additional memory features	447
10-5	IBM FlashSystem 7200 memory options	447
10-6	Memory options	447
10-7	IBM FlashSystem 9200 control enclosure adapter card options.	448
10-8	IBM FlashSystem 9100 control enclosure adapter card options.	448
10-9	IBM FlashSystem 5000 family configurations.	449
10-10	IBM FlashSystem 5000 family adapter cards	449
10-11	IBM FlashWatch product matrix for IBM FlashSystem products.	458
10-12	Files created by the backup process	474
A-1	Comparing IBM i native and Virtual I/O Server attachment	534
A-2	Limits increased for Max Disk Arms and LUN size.	544
A-3	Throughput of Fibre Channel adapters.	545

Examples

3-1 Manual FCM format	69
3-2 FCM capacity monitoring with CLI	70
3-3 Array capacity monitoring with CLI	77
3-4 Round robin enabled storage controller	79
3-5 MDisk group balanced path selection (no round robin enabled) storage controller	80
3-6 Command output for the lsarray and lsrank commands	86
3-7 Output of the showvolgrp command	89
3-8 Output for the lshostconnect command	90
3-9 Output of the lscontroller command	90
3-10 The lsquorum command	100
4-1 Results of capacity savings analysis	116
4-2 DRET command line	118
4-3 Listing of volumes that have extents on an MDisk to be deleted	142
4-4 DS8000 UID example	143
4-5 The lscontroller command	144
4-6 Command to declare or identify a self-encrypted MDisk from a virtualized external storage	151
4-7 Example of a SKLMConfig.properties configuration file	153
4-8 Manually triggered replication	156
4-9 Verify key state	157
4-10 Commands to enable key server encryption option on a system upgraded from pre-7.8.0	158
4-11 Changing MDisk tier	170
4-12 Changing Easy Tier setting	175
5-1 Volume creation without auto formatting option	189
5-2 Thin-provisioned volume creation	194
5-3 Creating thin-provisioned volume with deduplication option	197
5-4 Creating compressed volume with deduplication option	197
5-5 The migratevdisk command	212
5-6 Monitoring the migration process	213
5-7 Throttle commands example	220
5-8 Changing the cache mode of a volume	223
6-1 Changing gmlinktolerance to 30 and gmmaxhostdelay to 100	298
6-2 Changing rbuffersize to 64 MB	298
6-3 maxreplicationdelay and partnershipexclusionthreshold setting	304
6-4 A lsiogrp and chiogrp command example	337
8-1 The lshostvdiskmap command	351
8-2 Output of using the lshostvdiskmap command	351
8-3 Determining the Symantec Veritas server configuration	360
8-4 Symantec Volume Manager using SDD pass-through mode ASL	360
8-5 IBM Spectrum Virtualize that is configured by using native ASL	360
10-1 lssystem command	426
10-2 Copying the upgrade test utility to IBM FlashSystem 9200	431
10-3 Upgrade test using the CLI	432
10-4 List firmware level for drives 0,1, 2 and 3	436
10-5 Output of the pcmpath query WWPN command	437
10-6 Output of the lshost <hostname> command	438
10-7 Cross-referencing information with SAN switches	438

10-8	Results of running the lshostvdiskmap command	439
10-9	Mapping the host to I/O groups	442
10-10	Creating a throttle using the mkthrottle command in the CLI	461
10-11	Sample svcconfig backup command output	474
10-12	Saving the config backup files to your workstation	474
11-1	The svc_livedump command	486
11-2	The sddgetdata.bat tool	488
11-3	lshost command	490
11-4	lshost <host_id_or_name> command	490
11-5	lsfabric -host <host_id_or_name> command	491
11-6	Incorrect WWPN zoning	495
11-7	Correct WWPN zoning	495
11-8	Issuing an lsmdisk command	496
11-9	Output of the svcinfo lscontroller command	498
11-10	Determining the ID for the MDisk	500

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 <http://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 Garage™	PowerHA®
Db2®	IBM Research®	PowerVM®
DB2®	IBM Security™	Redbooks®
DS8000®	IBM Spectrum®	Redbooks (logo)  ®
Easy Tier®	Insight®	Service Request Manager®
FlashCopy®	Interconnect®	Storwize®
Global Technology Services®	MicroLatency®	SystemMirror®
HyperSwap®	Netcool®	Tivoli®
IBM®	Orchestrate®	XIV®
IBM Cloud®	POWER7®	z/OS®
IBM FlashCore®	POWER8®	
IBM FlashSystem®	POWER9™	

The following terms are trademarks of other companies:

Intel, 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, 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 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 of the preferred practices and describes the performance gains that can be achieved by implementing the IBM FlashSystem® products. These practices are based on field experience.

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

It explains how you can optimize disk performance with the IBM System Storage Easy Tier® function. It also provides preferred practices for monitoring, maintaining, and troubleshooting.

This book is intended for experienced storage, SAN, IBM FlashSystem, SAN Volume Controller (SVC), and IBM Storwize® administrators and technicians. Understanding this book requires advanced knowledge of these environments.

Authors

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



Anil K Nayak is a Level 3 Development Support Engineer at IBM India Systems Development Lab providing support to IBM FlashSystem/SVC worldwide installed base. In the current role as a technical team leader, his responsibilities include troubleshooting of complex field issues, coaching ecosystem for quality problem resolutions, being a technical advocate to many large spectrum virtualize deployments.

Anil joined IBM in 2012. He has 16 years of industry experience in Systems and Storage domain. Prior to joining IBM, Anil was a device driver developer at Hewlett Packard(HP). His core skill sets include I/O stack, Mass storage protocols, Server/Storage virtualization, and backup/recovery software. He has Bachelor of Engineering [B.E.] Computer Science and Engineering (CSE) from BPUT, Odisha and a corporate MBA in Marketing from Symbiosis, Maharashtra, India.



Antonio Rainero is an Executive Technical Specialist working for the IBM Global Technology Services® organization in IBM® Italy. He joined IBM in 1998, and has more than 20 years of experience in the delivery of storage services for Open Systems and IBM z/OS® clients. His areas of expertise include storage systems implementation, SANs, storage virtualization, performance analysis, disaster recovery, and high availability solutions. He has co-authored several IBM Redbooks publications. Antonio holds a degree in Computer Science from University of Udine, Italy.



Barry Whyte is an “IBM Master Inventor” working in the IBM Systems Group. Based in Auckland, New Zealand, Barry is an IBM Advanced Technical Specialist team covering Storage across the Asia Pacific region. Barry primarily works with the Spectrum Virtualize (IBM SAN Volume Controller, Storwize, and FlashSystem) family of virtual disk systems. Barry graduated from The University of Glasgow in 1996 with a B.Sc (Hons) in Computing Science. In his 23 years at IBM he has also worked on the successful Serial Storage Architecture (SSA) and the IBM DS8000® products. Barry joined the SVC development team soon after its inception and has held many positions before he took on the role as performance architect. In 2015 Barry moved to New Zealand, but maintains a part-time working role for the Hursley team.



Chris Hoffmann has 30 years experience in IT in a variety of areas, concentrating on backup and storage for the last 15 years. He is currently employed by Advent One, an IBM Platinum Business Partner in Australia. In his current role over the last 2 years he has been involved in the planning, installation and configuration of FlashSystem 9150, V7000, and V5000 storage systems.



Danilo Morelli Miyasiro has more than 13 years working with distributed disk storage systems and SAN networking. He is currently working as SAN administrator and also as system administrator (UNIX, Linux, Windows, and VMware) in IBM Silicon Valley and Tucson labs.



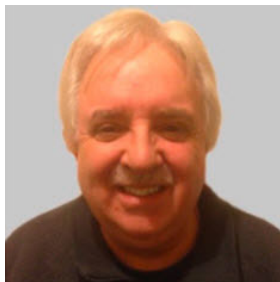
David Green works with the IBM SAN Central team troubleshooting performance and other problems on storage networks. He has authored, or contributed to, a number of IBM Redbooks publications. He is a regular speaker at IBM Technical University. You can find his blog at Inside IBM Storage Networking where he writes about all things related to Storage Networking and IBM Storage Insights.



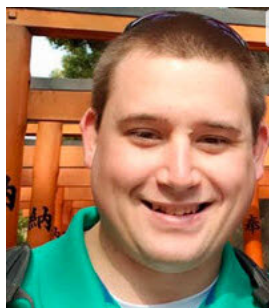
Duane Bolland is an IBM Storage Technical Advisor working from Portland, Oregon. His first run-in with IBM was as an AIX® administrator in the 1990s. He has been with IBM for 20 years. His past assignments include roles in the Open Systems Lab, SVC PFE, and xSeries Lab Services. He is a technical focal for Spectrum Virtualize having worked with the products since 2006. This is his second IBM Redbooks Publication. Duane likes to photograph off-trail waterfalls.



Jackson Shea is a Level 2 certified IBM Information Technology Specialist/Architect performing design and implementation engagements through Lab Services. He has been with IBM since April 2010. He was a Lead Storage Administrator with a large health insurance consortium in the Pacific Northwest, and has been working with IBM equipment since 2002. He has had over 12 years of experience with IBM Spectrum® Virtualize, formerly known as the IBM SAN Volume Controller and related technologies. Jackson is based out of Portland, Oregon. He received his Bachelor of Science degree in Philosophy with minors in Communications and Chemistry from Lewis and Clark College. Jackson's professional focus is IBM Spectrum Virtualize, but he is conversant with storage area network design, implementation, extension, and storage encryption.



Jon Herd is an IBM Executive Technical Advisor working for the ESCC, Germany. He covers the United Kingdom, Ireland, and Sweden, advising customers on a portfolio of IBM storage products, including IBM FlashSystem products. Jon has been with IBM for more than 45 years, and has held various technical roles, including Europe, Middle East, and Africa (EMEA) level support on mainframe servers and technical education development. He has written many IBM Redbooks® publications about IBM FlashSystem products and is an IBM Redbooks Platinum level author. He holds IBM certifications in Product Services at a thought leader L3 level, and Technical Specialist at an experienced level. He is also 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.



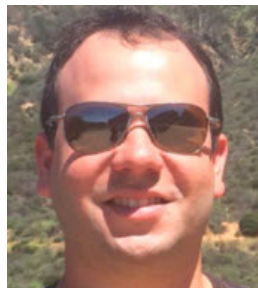
Jordan Fincher is a SVC and FlashSystems Level 3 Support Engineer. He has contributed to several IBM Redbooks publications and periodically speaks at IBM Technical University events. You can find his blog “Supporting Spectrum Virtualize” (<https://www.specvsupport.com/>), where he writes about various support cases.



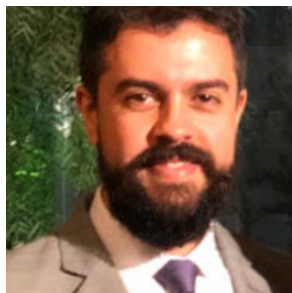
Marcelo Avalos Del Carpio is an Infrastructure Architect for IBM i, he is based in IBM Uruguay. During his 7 plus years experienced IT background, he has held leadership roles, handled demanding relevant IBM customers accounts, critical situations, decision capability, and hands-on experience delivering deployment of IBM semiconductor, Power Systems, storage systems (High-End & Mid-range), and system software, in Banking Industry in a couple of countries in South America. He received his degree in Electronic Systems Engineering from Escuela Militar de Ingeniería in Bolivia and is currently pursuing a master's degree in Project Management at GSPM UCI - Costa Rica.



Sergey Kubin is a subject matter expert (SME) for IBM Storage and SAN technical support. He holds an Electronics Engineer degree from Ural Federal University in Russia and has more than 15 years of experience in IT. At IBM, he works for IBM Technology Support Services, where he provides support and guidance about IBM Spectrum Virtualize family systems for customers in Europe, the Middle East, and Russia. His expertise includes SAN, block-level, and file-level storage systems and technologies. He is an IBM Certified Specialist for IBM FlashSystem Family Technical Solutions.



Sidney Varoni Junior is a Storage Technical Advisor for IBM Systems in IBM Brazil. He has over 14 years of experience working with complex IT environments, having worked with both Mainframe and Open Systems platforms. He currently works with clients from Brazil and other countries in Latin America, advising them on how to take the best out of their IBM Storage products. He holds a bachelor degree in Computer Science from Faculdade Politecnica de Jundiai. His areas of expertise include High Availability, Disaster Recovery, Business Continuity solutions, and performance analysis.



Thales Noivo Ferreira has been working as SAN Admin for the past 8 years at IBM, working on various accounts and technologies such as Netapp, IBM Storwize Family, DS8K Family, XiV, Cisco Switches, and Brocade Switches. Currently he is working on environments with SVC and FlashSystem family.



Vasfi Gucer is an IBM Technical Content Services Project Leader with IBM Garage™ for Systems. 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 cloud computing, including cloud storage technologies for the last 6 years. Vasfi is also an IBM Certified Senior IT Specialist, Project Management Professional (PMP), IT Infrastructure Library (ITIL) V2 Manager, and ITIL V3 Expert.

Thanks to the following for their contributions that made this book possible:

Lucy Harris, Matthew Smith, Suri Poliseti
IBM Hursley, UK

Dharmesh Kamdar, Gary Domrow, Ian MacQuarrie
IBM US

Petar Kalachev
IBM Bulgaria

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 Facebook:
<http://www.facebook.com/IBMRedbooks>
- ▶ Follow us on Twitter:
<http://twitter.com/ibmredbooks>
- ▶ Look for us on LinkedIn:
<http://www.linkedin.com/groups?home=&gid=2130806>
- ▶ Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:
<https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm>
- ▶ Stay current on recent Redbooks publications with RSS Feeds:
<http://www.redbooks.ibm.com/rss.html>



IBM FlashSystem introduction

This chapter introduces the IBM FlashSystem storage subsystem range that is supported by the new Spectrum Virtualize software v8.4. It details all the relevant models, their key features, benefits, and technology.

This chapter includes the following topics:

- ▶ 1.1, “IBM FlashSystem supported product range” on page 2
- ▶ 1.2, “IBM FlashSystem high-level features” on page 4
- ▶ 1.3, “IBM Storwize and IBM FlashSystem product range” on page 6
- ▶ 1.4, “Advanced functions for data reduction” on page 28
- ▶ 1.5, “Advanced software features” on page 29

1.1 IBM FlashSystem supported product range

This section details the IBM FlashSystem products that are currently supported by the new Spectrum Virtualize software v8.4.

Figure 1-1 shows the range of IBM FlashSystem and IBM Storwize products that support Spectrum Virtualize software v8.4.

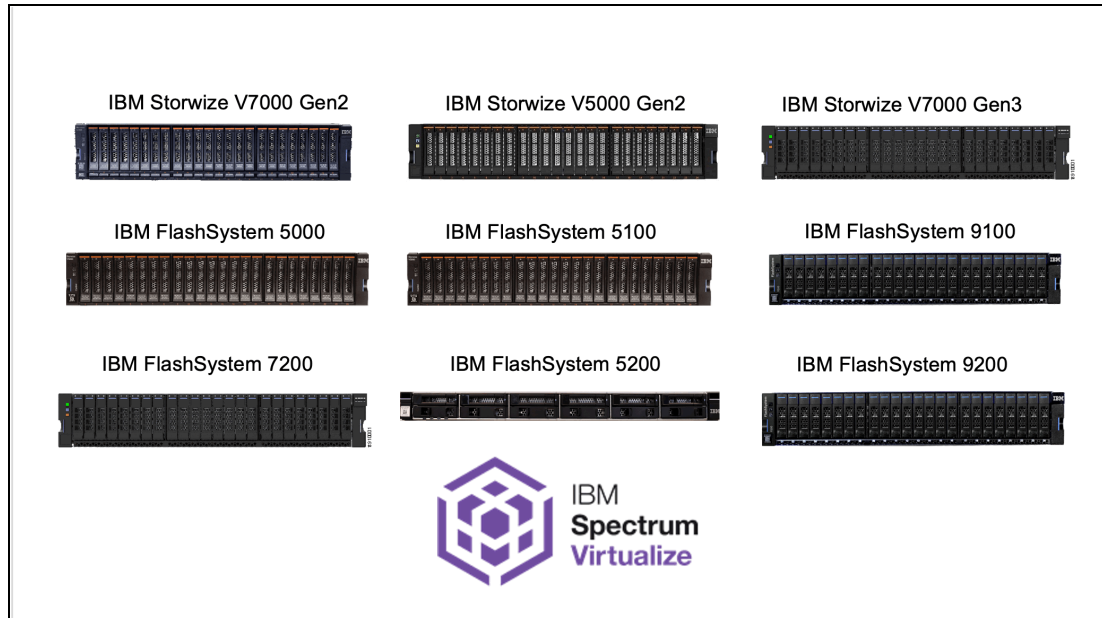


Figure 1-1 IBM FlashSystem and IBM Storwize products that support Spectrum Virtualize software v8.4

1.1.1 What's new in V8.4

IBM Spectrum Virtualize 8.4 provides additional features and updates to the IBM Spectrum Virtualize family of products which IBM FlashSystem is part of.

The major software changes in version 8.4 are:

- ▶ Data Reduction Pool (DRP) improvements;
 - Data Reduction Child Pool: This improvement allows for more flexibility such as multi-tenancy.
 - FlashCopy® with redirect-on-write support: This improvement uses DRP's internal deduplication referencing capabilities to reduce overheads by creating references instead of copying the data. Redirect-on-write (RoW) is an alternative to the existing copy-on-write (CoW) capabilities.

Note: At the time of writing, this capability might be used only for volumes with supported deduplication without mirroring relationships and within the same pool and I/O group. The mode selection (RoW/CoW) is automatic based on these conditions.

- Compressimator always on; This improvement allows the systems to sample each volume at regular intervals, providing the ability to display the compressibility of the data in the GUI and IBM Storage Insights at any time.
- RAID Reconstruct Read; This improvement increases reliability and availability by reducing chances of DRP going offline because of fixable array issues, leveraging RAID capabilities, DRP asks for a specific data block reconstruction when detecting a potential corruption.
- ▶ Distributed RAID 1 (DRAID 1) support provides the ability to extend distributed RAID advantages to smaller pools of drives. This improves performance over traditional RAID 1 implementations, allowing a better use of flash technology. These distributed arrays can support as few as two drives, with no rebuild area, and 3 - 16 drives, with a single rebuild area.

Note: At the time of writing, DRAID 1 is only supported on IBM FlashSystems 5015, 5035, 5200, 7200, and 9200 and is not available for FlashCore Modules (FCM-XL) of 38.4 TB

- ▶ With 8.4 FlashSystem 5100, 5200, 7200, and 9200, systems can support up to 12 Storage Class Memory (SCM) devices per enclosure with no slot restrictions. Previously, the limit for all SCM drives was four per enclosure in drive slots 21 - 24.

Note: With 8.4 code FlashSystem 5100, 5200, 7200, and 9200, systems can now support up to 12 zSSD SCM drives or Optane SCM drives.

- ▶ Expansion of mirrored volumes (also known as vDisks) allows the volumes capacity to be expanded or reduced online, without requiring an offline format and sync. This improves the availability of the volume because the new capacity is available immediately.
- ▶ Three-site replication with HyperSwap® support provides improved availability for data in three-site implementations. This expands on the disaster recovery capabilities inherent in this topology.

Important: Three-site replication using Metro Mirror was previously supported on version 8.3.1 only in limited installations via RPQ process. With 8.4.0, this implementation is generally available.

- ▶ Host attachment support with FC-NVMe in HyperSwap systems.
- ▶ DNS support for LDAP and NTP with full DNS length (that is, 256 characters).
- ▶ Updates to maximum configuration limits. This doubles FlashCopy mapping from 5,000 to 10,000 and increases HyperSwap volumes limit from 1,250 to 2,000.
- ▶ Password and login changes on the IBM FlashSystem v8.4 GUI to meet today's extra regulatory compliance with expiry and security enhancements.
- ▶ Support for internal proxy servers (also known as customer web proxy) uses IBM Call Home with cloud services and log upload features.

1.1.2 Products supported

The following IBM FlashSystem and IBM Storwize products are supported to run the Spectrum Virtualize software v8.4 software. The following section lists the IBM Storwize and

IBM FlashSystem series name and then the hardware machine type and model for extra clarity:

Storwize products

- ▶ IBM Storwize V5100
 - 2077-AF4, 2077-424, 2077-U5B, 2078-AF4, 2078-424, or 2078-U5B
- ▶ IBM Storwize V7000
 - 2076-724, 2076-U7B, 2076-U7A, 2076-AF6, 2076-624, or 2076-524

FlashSystem products

- ▶ IBM FlashSystem 5000
 - IBM FlashSystem 5010 and IBM FlashSystem 5030
 - (formerly known as IBM Storwize V5010E and Storwize V5030E)
 - 2072-2H2, 2072-U12, 2072-2H4, 2072-U24, 2072-3H2, 2072-V12, 2072-3H4, or 2072-V24
 - IBM FlashSystem 5015 and IBM FlashSystem 5035
 - 2072-2N2, 2072-U12, 2072-2N4, 2072-U24, 2072-3N2, 2072-V12, 2072-3N4, or 2072-V24
- ▶ IBM FlashSystem 5100
 - 2077-4H4, 2078-4H4, or 2078-UHB
- ▶ IBM FlashSystem 5200
 - 4662-6H2 or 4662-UH6
- ▶ IBM FlashSystem 7200
 - 2076-824 or 2076-U7C
- ▶ IBM FlashSystem 9100
 - 9846-AF7, 9848-AF7, 9848-UF7, 9846-AF8, 9848-AF8, or 9848-UF8
- ▶ IBM FlashSystem 9200
 - 9846-AG8, 9848-AG8, or 9848-UG8

1.2 IBM FlashSystem high-level features

This IBM Redbooks publication describes and focuses on the best practice and options to gain the optimum performance from the product, including the set of software-defined storage features.

It also describes data-reduction techniques, including deduplication, compression, dynamic tiering, thin provisioning, snapshots, cloning, replication, data copy services, and IBM HyperSwap for high availability.

Note: The detailed technical explanations, and theory of operations, of these features are not covered in this publication. If you need extra information in this area, see the following Redbook publications:

- ▶ *Implementing the IBM SAN Volume Controller with IBM Spectrum Virtualize V8.4*, SG24-8467
- ▶ *Implementing the IBM FlashSystem with IBM Spectrum Virtualize V8.4*, SG24-8465
- ▶ *IBM FlashSystem 9200 Product Guide*, REDP-5586
- ▶ *IBM FlashSystem 9100 Product Guide*, REDP-5524
- ▶ *IBM FlashSystem 7200 Product Guide*, REDP-5587
- ▶ *IBM FlashSystem 5200 Product Guide*, REDP-5617
- ▶ *IBM FlashSystem 5000 and 5100 for Mid-Market*, REDP-5594
- ▶ *IBM Flashsystem 5000 and 5200 for Mid-Market*, REDP-5630
- ▶ *IBM DS8870 Easy Tier Heat Map Transfer*, REDP-5015

The following two types of enclosures are part of the IBM FlashSystem products that run Spectrum Virtualize:

- ▶ A *control enclosure* manages your storage systems, communicates with the host, and manages interfaces. In addition, it can also house up to 24 drives. These drives can be either industry-standard NVMe type, the exclusive IBM NVMe FlashCore Modules (FCM), standard flash (SSD) Serial Attached SCSI (SAS) type drives, or hard disk drives (HDD), depending on which model of control enclosure is ordered.

Each control enclosure is either a standard 2U high, or 1U high for the IBM FlashSystem 5200, 19" rack-mounted unit.

- ▶ An *expansion enclosure* enables you to increase the available capacity of the IBM FlashSystem cluster communicates with the control enclosure via a pair of 12 Gbps SAS connections. These expansion enclosures can house many flash (SSD) SAS type drives or hard disk drives (HDD), depending on which model of expansion enclosure is ordered.

Expansion enclosures are generally of the three following types:

- Large form factor (LFF). Figure 1-2 shows the LFF expansion enclosure that can hold twelve 3.5" drives and is 2U high.



Figure 1-2 LFF expansion enclosure

- Small form factor (SFF). Figure 1-3 on page 6 shows the SFF expansion enclosure that can hold 24 2.5" drives and is 2U high.

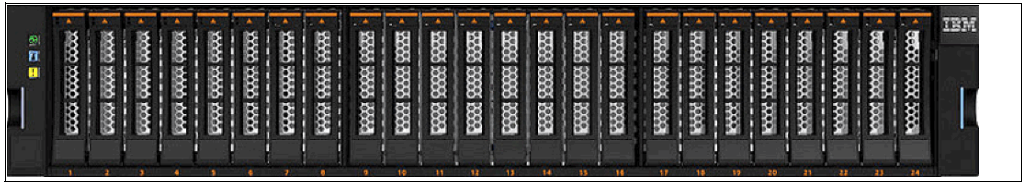


Figure 1-3 SFF expansion enclosure

- Large form factor high density (LFF HD). Figure 1-4 shows the large form factor high density (LFF HD) expansion enclosure that can hold ninety-two 3.5” drives (or ninety-two 2.5” drives in carriers) and is 5U high.

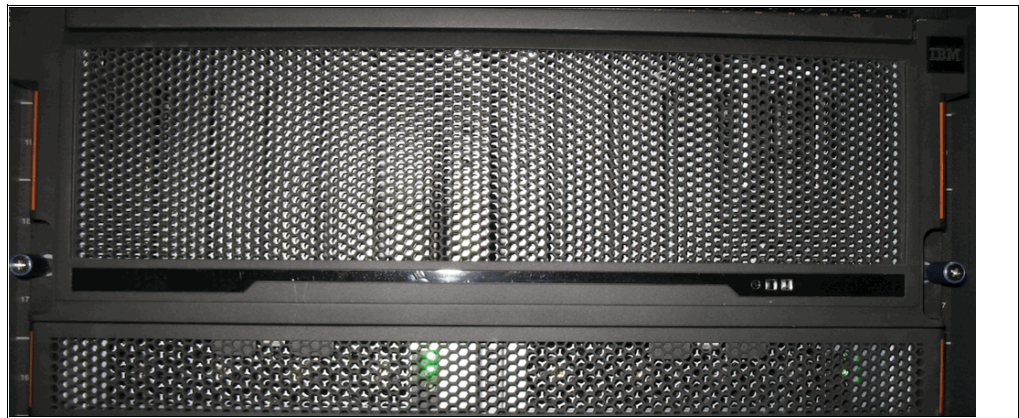


Figure 1-4 LFF HD expansion enclosure

The type and models of expansion enclosures that can attach to the relevant control enclosure is model-dependent and is shown in the tables in 1.3, “IBM Storwize and IBM FlashSystem product range” on page 6.

1.3 IBM Storwize and IBM FlashSystem product range

The following section describes the various IBM Storwize and IBM FlashSystem products that are supported in IBM Spectrum Virtualize software v8.4. It includes in-depth information about each product, its capabilities, features, and functions. Also supplied for each product range are links to information on the configuration limits and restrictions, so the customer can research information or values needed for optimum performance and adhere to the best practices.

IBM Storwize V5100

IBM Storwize V5100 is a virtualized, software-defined storage system comprised of hardware components and a requisite licensed software product, IBM Spectrum Virtualize Software. All functional capabilities for the Storwize V5100 system are provided through IBM Spectrum Virtualize software v8.4.

Note: v8.4 or later does not support Storwize V5015, V5020, or V5030 systems.

Figure 1-5 on page 7 shows the front view of the IBM Storwize V5100 control enclosure.



Figure 1-5 IBM Storwize V5100 front view

IBM Storwize V5100 models 424 and AF4 are designed to meet modern high-performance storage requirements, including ultra-low latency, cost-effectiveness, operational efficiency, and mission-critical reliability. It is built on a flash-optimized design, with an end-to-end NVMe strategy to bring extremely low latencies to organizations of all sizes.

- ▶ Storwize V5100 model AF4
 - All-flash storage system that supports NVMe FlashCore Modules and industry-standard NVMe flash drives in the control enclosure.
 - Model AF4 attaches to expansion enclosure models AFF and A9F that support SAS Flash drives.
- ▶ Storwize V5100 model 424
 - Hybrid storage system that supports NVMe FlashCore Modules and industry-standard flash NVMe drives in the control enclosure.
 - Model 424 attaches to expansion enclosure models 12F, 24F, and 92F that support SAS Flash drives and SAS HDD Drives.
- ▶ Storwize V5100 model U5B
 - Storwize V5100 hardware component to be used in the Storage Utility Offering space. It is physically and functionally identical to V5100 models 424 and AF4 except for target configurations and variable capacity billing. The variable capacity billing uses IBM Storage Insights to monitor the system usage, enabling allocated storage usage that exceeds a base subscription rate to be billed per terabyte, per month.

Existing machine type 2077 and 2078 expansion enclosure models 12F, 24F, and 92F can be attached to a Storwize V5100 model 424. Models AFF and A9F can be attached to a Storwize V5100 model AF4.

Storwize V5100 systems can be clustered with another V5100 system. All systems within a cluster must be using the same version of Storwize V5000 software.

For a comprehensive list of supported environments, devices, and configurations, see: [IBM System Storage Interoperation Center \(SSIC\)](#).

Table 1-1 on page 8 shows the IBM Storwize V5100 host, drive capacity, and functions

Note: The V5100 systems are now End of Marketing (EOM) since April 2020, and as such not available to purchase from IBM anymore. They are included in this Redbook for completeness as they support running the Spectrum Virtualize software v8.4.

summary.

Table 1-1 IBM Storwize V5100 host, drive capacity, and functions summary

Feature / Function	Description
Host interface	<ul style="list-style-type: none"> ▶ 10 Gbps Ethernet (iSCSI) ▶ 25 Gbps Ethernet (iSCSI, iWARP, RoCE) ▶ 16 Gbps Fibre Channel (FC, FC-NVMe) ▶ 32 Gbps Fibre Channel (FC-NVMe)
Control Enclosure Supported drives	<ul style="list-style-type: none"> ▶ 2.5-inch NVMe FCMs <ul style="list-style-type: none"> – 4.8 TB, 9.6 TB, and 19.2 TB compressing FCMs ▶ 2.5-inch NVMe flash drives <ul style="list-style-type: none"> – 800 GB, 1.92 TB, 3.84 TB, 7.68 TB, and 15.36 TB
SAS expansion enclosures 760 per control enclosure; 1,520 per clustered system	<ul style="list-style-type: none"> ▶ Control enclosure model 424 supports the following expansions: <ul style="list-style-type: none"> – Model 12F/24F 2U 12 or 24 drive – Model 92F 5U 92 drive – 2.5-inch flash drives supported: <ul style="list-style-type: none"> • 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 and 900 GB 15k SAS drive • 600 GB, 900 GB, 1.2 TB, 1.8 TB, and 2.4 TB 10k SAS disk • 2 TB 7.2k nearline SAS disk – 3.5-inch disk drives supported: <ul style="list-style-type: none"> • 4 TB, 6 TB, 8 TB, 10 TB, 12 TB, and 14 TB 7.2k nearline SAS disk ▶ Control enclosure model AF4 supports the following expansions: <ul style="list-style-type: none"> – Model AFF 2U 24 drive – Model A9F 5U 92 drive – 2.5-inch flash drives supported: <ul style="list-style-type: none"> • 800 GB, 1.6 TB, 1.92 TB, 3.84 TB, 7.68 TB, 15.36 TB, and 30.72 TB
RAID levels	<ul style="list-style-type: none"> ▶ DRAID 5 (CLI-only) and 6, TRAITD 10
Advanced features included with each system	<ul style="list-style-type: none"> ▶ Virtualization of internal storage ▶ Data migration ▶ Data Reduction Pools with thin provisioning ▶ UNMAP ▶ Compression and deduplication ▶ Metro Mirror (synchronous) and Global Mirror (asynchronous)
Additional available advanced features	<ul style="list-style-type: none"> ▶ Remote mirroring ▶ Easy Tier compression ▶ External virtualization ▶ Encryption ▶ FlashCopy ▶ IBM Spectrum Control ▶ IBM Spectrum Protect Snapshot

For more information, see [V8.4.0.x Configuration Limits and Restrictions for IBM Storage V5000E and V5100](#).

IBM Storwize V7000

IBM Storwize V7000 is a virtualized storage system to complement virtualized server environments that provides unmatched performance, availability, advanced functions, and highly scalable capacity never seen before in midrange disk systems.

IBM Storwize V7000 is a powerful midrange disk system that has been designed to be easy to use and enable rapid deployment without additional resources. IBM Storwize V7000 is virtual storage that offers greater efficiency and flexibility through built-in SSD optimization and “thin provisioning” technologies.

The following three generations of systems in the IBM Storwize V7000 are supported by Spectrum Virtualize software v8.4:

- ▶ IBM Storwize V7000 Generation 2 (2076-524)
- ▶ IBM Storwize V7000 Generation 2+ (2076-624 and 2076-U7A)
- ▶ IBM Storwize V7000 Generation 3 (2075-724 and 2076 -U7B)

Note: The V7000 systems are End of Marketing (EOM) since August 2020, and no longer available to purchase from IBM. They are included in this Redbook for completeness as they support running the Spectrum Virtualize software v8.4.

Figure 1-6 shows the IBM Storwize V7000 Generation 2 (2076-524) and IBM V7000 Generation 2+ (2076-624) front view.



Figure 1-6 IBM Storwize V7000 Generation 2 (2076-524) and Generation2+ (2076-624) SFF

Figure 1-7 shows the IBM Storwize V7000 Generation 3 (2076-724) front view.



Figure 1-7 IBM Storwize V7000 Generation 3 (2076-724)

- ▶ The IBM 2076 SFF Control Enclosure Model 524 and 624 features are as follows:

- Two node canisters and up to 256 GB cache (system total) in a 2U, 19-inch rack mount enclosure.
 - 1 Gb iSCSI connectivity is standard, with options for 16 Gb FC, 10 Gb iSCSI/FCoE, and 25 Gb iSCSI connectivity.
 - It holds up to twenty-four 2.5-inch SAS flash drives and supports the attachment of up to 20 Storwize V7000 expansion enclosures.
- ▶ The IBM 2076 Model 724 SFF NVMe Control Enclosure features are as follows:
- Two node canisters and up to 1 TB cache (system total) in a 2U, 19-inch rack mount enclosure.
 - 10 Gb iSCSI connectivity is standard, with options for 32 Gb FC, 16 Gb FC, and 25 Gb iSCSI connectivity.
 - It holds up to twenty-four 2.5-inch NVMe FlashCore Modules or industry-standard flash drives and supports the attachment of up to 20 Storwize V7000 expansion enclosures.

The IBM 2076 Models U7A and U7B are the Storwize V7000 hardware components to be utilized in the Storage Utility Offering space. They are physically and functionally identical to the V7000 model 624 and 724 respectively, except for target configurations and variable capacity billing. The variable capacity billing uses IBM Spectrum Control Storage Insights to monitor the system usage, allowing allocated storage usage that exceeds a base subscription rate to be billed per TB, per month.

Table 1-2 gives a summary of the host connections, drive capacities, features, and standard with Spectrum Virtualize that are available on the IBM Storwize V7000.

Table 1-2 IBM Storwize V7000 host, drive capacity and functions Summary

Feature / Function	Description
Host interface	<ul style="list-style-type: none"> ▶ Gen 2 and Gen 2+ <ul style="list-style-type: none"> - 1 Gb iSCSI connectivity is standard, with options for 16 Gb FC, 10 Gb iSCSI/FCoE, and 25 Gb iSCSI connectivity ▶ Gen 3 <ul style="list-style-type: none"> • 10 Gb iSCSI connectivity is standard, with options for 16 Gb FC and 25 Gb iSCSI connectivity
Control Enclosure Supported drives	<ul style="list-style-type: none"> ▶ Gen 2 / Gen2+ <ul style="list-style-type: none"> - NL SAS 7.2K RPM <ul style="list-style-type: none"> • 2 TB, 4 TB, 6 TB, 8 TB, and 10 TB - Flash (SSD) drives <ul style="list-style-type: none"> • 400 GB, 800 GB, 1.6 TB, 1.92 TB, 3.2 TB, 3.84 TB, 7.68 TB, and 15.36 TB - Enterprise class disk drives <ul style="list-style-type: none"> • 15K RPM - 300 GB, 600 GB, and 900 GB 15,000 rpm • 10K RPM _ 900 GB, 1.2 TB, 1.8 TB, and 2.4 TB ▶ Gen 3 <ul style="list-style-type: none"> - 2.5-inch SFF NVMe FlashCore Modules (FCM) <ul style="list-style-type: none"> • 4.8 TB, 9.6 TB, and 19.2 TB • 2.5-inch SFF NVMe industry-standard drives • 800 GB 2.5-inch 3DWPD - NVMe flash drive <ul style="list-style-type: none"> • 1.92 TB, 3.84 TB, 7.68 TB, and 15.36 TB

Feature / Function	Description
SAS expansion enclosures <ul style="list-style-type: none"> ▶ Storwize V7000 LFF Expansion Enclosure Model 12F <ul style="list-style-type: none"> – Twelve slots for 3.5-inch SAS drives ▶ Storwize V7000 SFF Expansion Enclosure Model 24F <ul style="list-style-type: none"> – Twenty-four slots for 3.5-inch SAS drives ▶ Storwize V7000 HD LFF Expansion Enclosure Model 92F <ul style="list-style-type: none"> – Ninety-two slots for SAS drives in a 3.5-inch carrier 	<ul style="list-style-type: none"> ▶ NL SAS 7.2K RPM <ul style="list-style-type: none"> – 2 TB, 4 TB, 6 TB, 8 TB, and 10 TB ▶ Flash (SSD) drives <ul style="list-style-type: none"> – 400 GB, 800 GB, 1.6 TB, 1.92 TB, 3.2 TB, 3.84 TB, 7.68 TB, and 15.36 TB ▶ Enterprise class disk drives <ul style="list-style-type: none"> – 15K RPM - 300 GB, 600 GB, and 900 GB 15,000 rpm – 10K RPM _ 900 GB, 1.2 TB, 1.8 TB and 2.4 TB
RAID levels	<ul style="list-style-type: none"> ▶ Gen 2 and 2+ <ul style="list-style-type: none"> – RAID 0, 1, 5, 6, and 10 ▶ Gen 3 <ul style="list-style-type: none"> – DRAID5, DRAID6, TRAI0, TRAI1 and TRAI10. – For compressed drives only DRAID5 and DRAID6 are supported.
Advanced features included with each system	<ul style="list-style-type: none"> – IBM System Storage Easy Tier – IBM FlashCopy – Thin provisioning
Additional available advanced features	<ul style="list-style-type: none"> – Remote Mirroring – External Virtualization – IBM FlashCopy Manager – IBM Systems Director

For more information on the V8.4.0.x Configuration Limits and Restrictions for IBM Storwize V7000 Generation 2 and 2+ and IBM Storwize V7000 Generation 3, see [V8.4.0.x Configuration Limits and Restrictions for IBM Storwize V7000 and V7000F](#).

IBM FlashSystem 5000

The IBM FlashSystem 5000 is a member of the IBM FlashSystem family of storage solutions. The IBM FlashSystem 5000 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 technologies to complement and enhance virtual environments, which deliver a simpler, more scalable, and cost-efficient IT infrastructure. The IBM FlashSystem 5000 features two node canisters in a compact, 2U 19-inch rack mount enclosure.

Note: At the time of writing, the IBM FlashSystem 5010 and IBM FlashSystem 5030 are End of Marketing (EOM) and have been replaced by the IBM FlashSystem 5015 and IBM FlashSystem 5035 respectively. The IBM FlashSystem 5015/5035 offer superior CPU power and memory options, but the features and functions remain the same. The IBM FlashSystem 5015/5035 charts are included only as a reference.

The new IBM FlashSystem 5015 and IBM FlashSystem 5035 are similar to the older 5010 and 5030 models, but with higher-specification CPU and memory options. The new models also include all-flash and hybrid-flash solutions designed to provide enterprise-grade functionality without compromising affordability or performance, built with the rich features of IBM Spectrum Virtualize 8.4. The IBM FlashSystem 5000 helps make modern technologies, such as artificial intelligence, accessible to enterprises of all sizes.

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 Spectrum Virtualize features and offers multiple flash and disk drive storage media and expansion options.

Figure 1-8 shows IBM FlashSystem 5015 and 5035 LFF control enclosure front view.



Figure 1-8 IBM FlashSystem 5015 and 5035 LFF control enclosure front view

Figure 1-9 on page 13 shows IBM FlashSystem 5015 and 5035 SFF control enclosure front view.



Figure 1-9 IBM FlashSystem 5015 and 5035 SFF control enclosure front view

This next section provides hardware information about the IBM FlashSystem 5000 models and the feature set of each one.

Table 1-3 shows the model comparison chart for the IBM FlashSystem 5000 range.

Table 1-3 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

Table 1-4 shows a summary of the host connections, drive capacities, features, and standard options with Spectrum Virtualize that are available on the IBM FlashSystem 5015.

Table 1-4 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 motherboard) ▶ 16 Gb/s Fibre Channel ▶ 12 Gb/s SAS ▶ 25 Gb/s iSCSI (iWARP or RoCE) ▶ 10 Gb/s iSCSI
Control Enclosure and SAS expansion enclosures supported drives	<ul style="list-style-type: none"> ▶ For SFF enclosures see Table 1-5 ▶ For LFF enclosures see Table 1-6
Cache per control enclosure / clustered system	32 GB or 64 GB
RAID levels	distributed DRAID 1, 5 and 6

Feature / Function	Description
Maximum expansion enclosure capacity	Up to 10 standard expansion enclosures per controller Up to 4 high-density expansion enclosures per controller
Advanced functions included with each system	<ul style="list-style-type: none"> ▶ Virtualization of internal storage ▶ Data reduction pools with thin provisioning and UNMAP ▶ One-way data migration
Additional available advanced features	<ul style="list-style-type: none"> ▶ Easy Tier ▶ FlashCopy ▶ Remote mirroring

Table 1-5 shows the capacity for the 2.5-inch supported drives for the IBM FlashSystem 5000.

Table 1-5 2.5 inch supported drives for the IBM FlashSystem 5000

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 (7.2k rpm)	2 TB					

Table 1-6 shows the speed and capacity for the 3.5-inch supported drives for the IBM FlashSystem 5000.

Table 1-6 3.5 inch supported drives for the IBM FlashSystem 5000

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

IBM FlashSystem 5035

IBM FlashSystem 5035 provides greater functionality, including powerful encryption capabilities and data reduction pools with compression, deduplication, thin provisioning, and the ability to cluster for scale-up and scale-out.

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

- ▶ Block deduplication that works across all the storage in a data reduction pool to minimize the number of identical blocks
- ▶ New compression technology that provides guaranteed 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 1-7 shows a summary of the host connections, drive capacities, features, and standard options with Spectrum Virtualize that are available on the IBM FlashSystem 5035.

Table 1-7 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 motherboard) ▶ 16 Gb/s Fibre Channel ▶ 12 Gb/s SAS ▶ 25 Gb/s iSCSI (iWARP or RoCE) ▶ 10 Gb/s iSCSI
Control Enclosure and SAS expansion enclosures Supported drives	<ul style="list-style-type: none"> ▶ For SFF enclosures see Table 1-5 ▶ For LFF enclosures see Table 1-6
Cache per control enclosure / clustered system	32 GB or 64 GB / 64 GB or 128 GB
RAID levels	Distributed 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 included with each system	<ul style="list-style-type: none"> ▶ Virtualization of internal storage ▶ Data reduction pools with thin provisioning ▶ UNMAP, compression, and deduplication ▶ One-way data migration ▶ Dual-system clustering
Additional available advanced features	<ul style="list-style-type: none"> ▶ Easy Tier ▶ FlashCopy ▶ Remote mirroring ▶ Encryption

For more information on the V8.4.0.x Configuration Limits and Restrictions for IBM FlashSystem 5015 and 5035, see [V8.4.0.x Configuration Limits and Restrictions for IBM FlashSystem 5x00](#).

IBM FlashSystem 5100

The IBM FlashSystem 5100 supports the following enhanced features, capacities, and enclosures:

- ▶ NVMe-accelerated flash arrays with control enclosures that are 100 percent, end-to-end NVMe-enabled, and Storage Class Memory (SCM)-capable.
- ▶ The systems offer industry-leading performance and scalability and support physical, virtual, and Docker environments.
- ▶ Hybrid-flash array enabled with multiple expansion enclosure options based on 12 Gbs SAS that support both SSDs and hard disk drives.
- ▶ AI-enhanced with the IBM Storage Insights analytics, resource management, and support platform. Also, IBM Spectrum Virtualize functions include AI-based data placement for optimal data center performance and zero-downtime data migration.

- ▶ Hybrid-cloud ready and can support private, hybrid, or public cloud deployments. The solutions come with ready-to-use, proven, validated “cloud blueprints” with support for cloud API automation, replication, and secondary data orchestration software.

The IBM FlashSystem 5100 control enclosure supports up to 24 2.5” NVMe capable flash drives in a 2U high form factor.

The IBM FlashSystem 5100 has two standard models (2077-4H4 and 2078-4H4) and one utility model (2078-UHB).

Figure 1-10 shows the IBM FlashSystem 5100 control enclosure front view.



Figure 1-10 IBM FlashSystem 5100 front view showing the 24 NVMe drives installed

Table 1-8 shows a summary of the host connections, drive capacities, features, and standard with Spectrum Virtualize that are available on the IBM FlashSystem 5100.

Table 1-8 IBM FlashSystem 5100 host, drive capacity, and functions summary

Feature / Function	Description
Host interface	<ul style="list-style-type: none"> ▶ 10 Gbps Ethernet (iSCSI) ▶ 25 Gbps Ethernet (iSCSI, iSER - iWARP, RoCE) ▶ 16 Gbps Fibre Channel (FC, FC-NVMe) ▶ 32 Gbps Fibre Channel (FC, FC-NVMe)
Control Enclosure Supported drives	<ul style="list-style-type: none"> ▶ 2.5-inch NVMe self-compressing FCM 2.5-inch <ul style="list-style-type: none"> – 4.8 TB, 9.6 TB, 19.2 TB, and 38.4 TB ▶ NVMe flash drives <ul style="list-style-type: none"> – 800 GB, 1.92 TB, 3.84 TB, 7.68 TB, and 15.36 TB
SAS expansion enclosures 760 per control enclosure; 1520 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 <ul style="list-style-type: none"> – 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 7.2k nearline SAS disk ▶ 3.5-inch disk drives supported <ul style="list-style-type: none"> – 4 TB, 6 TB, 8 TB, 10 TB, 12 TB, 14 TB, 16 TB, and 18 TB 7.2k nearline SAS disk
RAID levels	Distributed RAID 5 and 6, TRAIID 1 and 10

Feature / Function	Description
Advanced features included with each system	<ul style="list-style-type: none"> – Virtualization of internal storage – Data migration – Data Reduction Pools with thin provisioning – UNMAP – Compression and deduplication – Metro Mirror (synchronous) and Global Mirror (asynchronous)
Additional available advanced features	<ul style="list-style-type: none"> – Remote mirroring – Easy Tier compression – External virtualization – Encryption – FlashCopy – IBM Spectrum Control – IBM Spectrum Protect Snapshot

For more information on the V8.4.0.x Configuration Limits and Restrictions for IBM FlashSystem 5100, see [V8.4.0.x Configuration Limits and Restrictions for IBM FlashSystem 5x00](#).

IBM FlashSystem 5200

IBM FlashSystem 5200 allows you to be ready for the 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 Spectrum 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 FlashSystem family is based on a common storage software platform, IBM Spectrum Virtualize, that provides powerful all-flash and hybrid-flash solutions, offering feature-rich, cost effective, enterprise-grade storage solutions.

The industry-leading capabilities of IBM Spectrum Virtualize include a wide range of data services that can be extended to more than 500 heterogeneous storage systems. For example:

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

Available on IBM Cloud® and AWS, IBM Spectrum Virtualize for Public Cloud works together with IBM FlashSystem 5200 to deliver consistent data management between on-premises storage and public cloud. For example:

- ▶ Moving data and applications between on-premises and public cloud
- ▶ Implementing new DevOps strategies
- ▶ Using public cloud for disaster recovery without the cost of a second data center
- ▶ Improving cyber resiliency with “air gap” cloud snapshots

IBM FlashSystem 5200 offers world-class customer support, product upgrade, and guarantee programs:

- ▶ The IBM Storage Expert Care service and support is 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.
- ▶ The IBM Data Reduction Guarantee helps reduce planning risks and lower storage costs with baseline levels of data compression effectiveness in IBM Spectrum 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” NVMe capable flash drives in a 1U high form factor.

The IBM FlashSystem 5200 has one standard model (4662-6H2) and one utility model (4662-UH6).

Figure 1-11 shows the IBM FlashSystem 5200 control enclosure front view with 12 NVMe drives and a 3/4 ISO view.

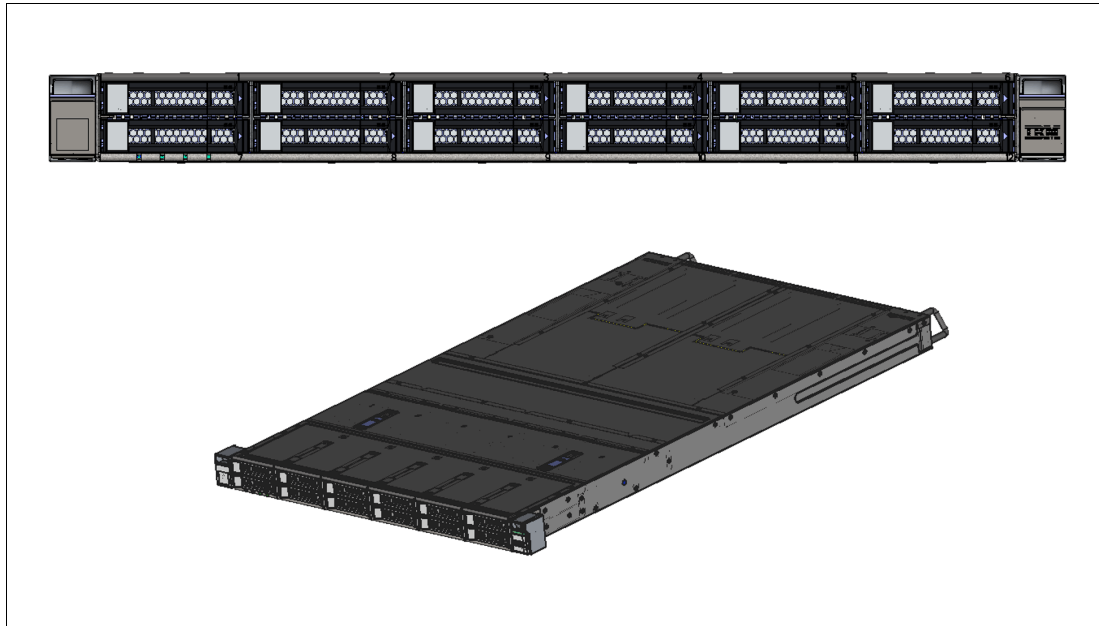


Figure 1-11 IBM FlashSystem 5200 control enclosure front and 3/4 ISO view

Table 1-9 gives a summary of the host connections, drive capacities, features, and standard options with Spectrum Virtualize that are available on the IBM FlashSystem 5200.

Table 1-9 IBM FlashSystem 5200 host, drive capacity, and functions summary

Feature / Function	Description
Host interface	<ul style="list-style-type: none"> ▶ 10 Gbps Ethernet (iSCSI) ▶ 25 Gbps Ethernet (iSCSI, iSER - iWARP, RoCE) ▶ 16 Gbps Fibre Channel (FC, FC-NVMe) ▶ 32 Gbps Fibre Channel (FC, FC-NVMe)
Control Enclosure Supported drives (12 maximum)	<ul style="list-style-type: none"> ▶ 2.5-inch NVMe self-compressing FCMs <ul style="list-style-type: none"> – 4.8 TB, 9.6 TB, 19.2 TB, and 38.4 TB ▶ NVMe flash drives <ul style="list-style-type: none"> – 800 GB, 1.92 TB, 3.84 TB, 7.68 TB, and 15.36 TB

Feature / Function	Description
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 <ul style="list-style-type: none"> – 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 7.2k nearline SAS disk ▶ 3.5-inch disk drives supported <ul style="list-style-type: none"> – 4 TB, 6 TB, 8 TB, 10 TB, 12 TB, 14 TB, 16 TB, and 18 TB 7.2k nearline SAS disk
RAID levels	Distributed RAID 5 and 6, TRAIID 1 and 10
Advanced features included with each system	<ul style="list-style-type: none"> – Virtualization of internal storage – Data migration – Data Reduction Pools with thin provisioning – UNMAP – Compression and deduplication – Metro Mirror (synchronous) and Global Mirror (asynchronous)
Additional available advanced features	<ul style="list-style-type: none"> – Remote mirroring – Easy Tier compression – External virtualization – Encryption – FlashCopy – IBM Spectrum Control – IBM Spectrum Protect Snapshot

For more information on the V8.4.0.x Configuration Limits and Restrictions for IBM FlashSystem 5200, see [V8.4.0.x Configuration Limits and Restrictions for IBM FlashSystem 5x00](#).

IBM FlashSystem 7200

To take advantage of artificial intelligence (AI)-enhanced applications, real-time Big Data analytics, and cloud architectures that require higher levels of system performance and storage capacity, enterprises around the globe are rapidly moving to modernize legacy IT infrastructures.

For many organizations, staff resources and expertise are not abundant, and cost-efficiency is a top priority. These organizations have important investments in existing infrastructure that they want to maximize. They need enterprise-grade solutions that optimize cost-efficiency while simplifying the pathway to modernization. The new IBM FlashSystem 7200 is designed specifically for these requirements and use cases.

The highlights of the IBM FlashSystem 7200 are:

- ▶ Deploy enterprise-grade functionality
- ▶ Leverage NVMe performance in one cost efficient system
- ▶ Build easy-to-manage, high-performance hybrid cloud environments
- ▶ Extend data services across more than 500 heterogeneous systems
- ▶ Transform data economics using sophisticated data reduction
- ▶ Leverage AI to optimize storage management and streamline issue resolution
- ▶ Deploy leading-edge storage solutions with confidence using IBM FlashWatch
- ▶ Increase cost-efficiency with IBM Storage Utility programs

Figure 1-12 on page 20 shows the IBM FlashSystem 7200 control enclosure front view.

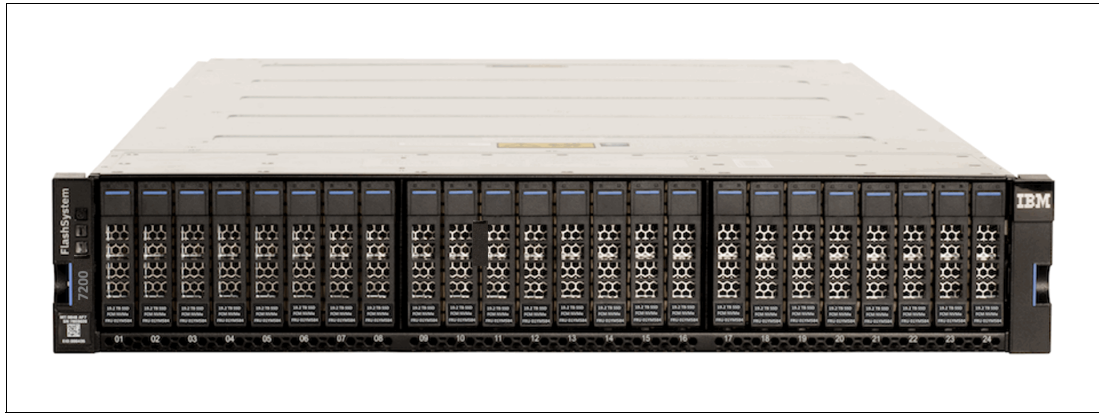


Figure 1-12 IBM FlashSystem 7200 control enclosure front view

Table 1-10 shows a summary of the host connections, drive capacities, features, and standards with Spectrum Virtualize that are available on the IBM FlashSystem 7200.

Table 1-10 IBM FlashSystem 7200 host, drive capacity and functions summary

Feature / Function	Description
Host interface	Per control enclosure <ul style="list-style-type: none"> ▶ Up to 24 x 16 Gbps Fibre Channel (FC, NVMeoF) ▶ Up to 24 x 32 Gbps Fibre Channel (FC, NVMeoF) ▶ 8 x 10 Gbps Ethernet (iSCSI) ▶ Up to 12 x 25 Gbps Ethernet (iSCSI, iSER - iWARP, RoCE)
Maximum drives supported	<ul style="list-style-type: none"> ▶ 24 x 2.5" NVMe drives per control enclosure ▶ 12 x 3.5" SAS drives per 12G expansion enclosure ▶ 24 x 2.5" SAS drives per 24G expansion enclosure ▶ 92 x 2.5" or 3.5" SAS drives per 92G expansion enclosure ▶ Up to a maximum of 760 SAS drives in expansion enclosures per control enclosure
Supported NVMe drives	<ul style="list-style-type: none"> ▶ FlashCore Modules (FCM): <ul style="list-style-type: none"> – 4.8 TB, 9.6 TB, 19.2 TB, and 38.4 TB with hardware compression ▶ Storage Class Memory (SCM) <ul style="list-style-type: none"> – 375 GB, 750 GB, 800 GB, 1.6 TB ▶ Industry Standard NVMe <ul style="list-style-type: none"> – 800 GB, 1.92 TB, 3.84 TB, 7.68 TB, and 15.36 TB
Supported SAS drives	<ul style="list-style-type: none"> ▶ Supported SAS drives 2.5-inch SAS SSD <ul style="list-style-type: none"> – 800 GB, 1.6 TB, 1.92 TB, 3.84 TB, 7.68 TB, 15.36 TB, and 30.72 TB ▶ 2.5-inch SAS HDD <ul style="list-style-type: none"> – 1.2 TB, 1.8 TB, and 2.4 TB 10k SAS – 2 TB 7.2k nearline SAS ▶ 3.5-inch disk drives supported: <ul style="list-style-type: none"> – 4 TB, 6 TB, 8 TB, 10 TB, 12 TB, 14 TB, 16 TB, and 18 TB 7.2k nearline SAS
RAID levels	DRAID 1, 5, and 6 with dynamic DRAID expansion; and TRAIID 1 and 10
Maximum IOPS (4K read hit)	2.3 million
Minimum latency (4K read hit)	<70 μ s

Feature / Function	Description
Maximum IOPS (4K read miss)	700 k
Maximum bandwidth (256Kb readmiss)	35 GB/s
Advanced features included with each system	<ul style="list-style-type: none"> – Virtualization of internal storage – Data migration – Data Reduction Pools with thin provisioning – UNMAP – Compression and deduplication – Metro Mirror (synchronous) and Global Mirror (asynchronous)
Additional available advanced features	<ul style="list-style-type: none"> – Remote mirroring – Easy Tier compression – External virtualization – Encryption – FlashCopy – IBM Spectrum Control – IBM Spectrum Protect Snapshot

For more information, see [V8.4.0.x Configuration Limits and Restrictions for IBM FlashSystem 7200](#).

IBM FlashSystem 9100

Figure 1-13 shows the IBM FlashSystem 9100 Control Enclosure with one of the IBM NVMe drives partially removed.



Figure 1-13 IBM FlashSystem 9100 control enclosure with one NVMe drives partially removed

The IBM FlashSystem 9100 control enclosure supports up to 24 NVMe capable flash drives in a 2U high form factor.

There are two standard models of The IBM FlashSystem 9100 has two standard models (9110-AF7 and 9150-AF8).

These numbers are the sales models, and each one is available as either a one-year (hardware machine type 9846), or a three-year (hardware machine type 9848) warranty product.

The IBM FlashSystem 9100 also has two utility models (9110-UF7 and 9150-UF8).

Note: The IBM 9110-UF7 and 9150-UF8 are the IBM FlashSystem 9100 with a three-year warranty only. These models are physically and functionally identical to the IBM FlashSystem 9848-AF7 and AF8 respectively, except for target configurations and variable capacity billing.

The variable capacity billing uses IBM Spectrum Control Storage Insights to monitor the system usage, allowing allocated storage usage that exceeds a base subscription rate to be billed per TB, actually written is considered used. For thick provisioning, total allocated volume space is considered used.

Table 1-11 shows a summary of the host connections, drive capacities, features, and standard options with Spectrum Virtualize that are available on the IBM FlashSystem 9100.

Table 1-11 IBM FlashSystem 9100 host, drive capacity, and functions summary

Feature / Function	Description
Host interface	<ul style="list-style-type: none"> ▶ 24 ports 16 Gb or 32 Gb Fibre Channel (FC, FC-NVMe) ▶ 8 ports 10 GbE iSCSI ▶ 12 ports 25 GbE iWARP or RoCE
Maximum drives supported	2.5-inch NVMe FCMs <ul style="list-style-type: none"> – 4.8 TB, 9.6 TB, and 19.2 TB compressing FCMs 2.5-inch NVMe flash drives <ul style="list-style-type: none"> – 1.92 TB, 3.84 TB, 7.68 TB, and 15.36 TB
Control enclosure supported NVMe drives	2.5-inch NVMe FCMs <ul style="list-style-type: none"> – 4.8 TB, 9.6 TB, and 19.2 TB compressing FCMs 2.5-inch NVMe flash drives <ul style="list-style-type: none"> – 1.92 TB, 3.84 TB, 7.68 TB, and 15.36 TB
Expansion enclosure supported SAS drives	<ul style="list-style-type: none"> ▶ Model AFF 2U 24 drive ▶ Model A9F 5U 92 drive ▶ 2.5-inch flash drives supported <ul style="list-style-type: none"> – 1.92 TB, 3.2 TB, 3.84 TB, 7.68 TB, 15.36 TB, and 30.72 TB
RAID levels	<ul style="list-style-type: none"> ▶ FCM drives <ul style="list-style-type: none"> – Distributed RAID 6 (recommended),: Distributed RAID5 (supported) ▶ NVMe flash drives <ul style="list-style-type: none"> – Traditional RAID 10 and Distributed RAID 6 (recommended), Distributed RAID 5 (supported)
Maximum IOPS (4K read hit)	3,800,000
Minimum write latency	120us
Maximum IOPS (4K read miss with hardware compression)	1,200,000
Maximum bandwidth (256Kb read miss)	34 GB/s
Advanced features included with each system	<ul style="list-style-type: none"> ▶ Virtualization of internal storage ▶ Data migration ▶ Data Reduction Pools with thin provisioning ▶ UNMAP ▶ Compression and deduplication ▶ Metro Mirror (synchronous) and Global Mirror (asynchronous)

Feature / Function	Description
Additional available advanced features	<ul style="list-style-type: none"> ▶ Remote mirroring ▶ Easy Tier compression ▶ External virtualization ▶ Encryption ▶ FlashCopy ▶ IBM Spectrum Control ▶ IBM Spectrum Protect Snapshot

For more information, see [V8.4.0.x Configuration Limits and Restrictions for IBM FlashSystem 9100](#).

IBM FlashSystem 9200

Some applications exist that are foundational to the operations and success of an enterprise. These applications might function as prime revenue generators, might guide or control important tasks, or might provide crucial business intelligence, among many other jobs. Whatever their purpose, they are mission-critical to the organization. They demand the highest levels of performance, functionality, security, and availability. To support mission-critical applications, enterprises of all types and sizes turn to IBM FlashSystem 9200.

IBM FlashSystem 9200 combines the performance of flash and a Non-Volatile Memory Express (NVMe)-optimized architecture with the reliability and innovation of IBM FlashCore technology and the rich feature set and high availability of IBM Spectrum Virtualize. This powerful new storage platform provides:

- ▶ The option to use large capacity IBM FlashCore modules (FCM) with inline-hardware compression, data protection, and innovative flash-management features; industry standard NVMe drives; or Storage Class Memory (SCM) drives.
- ▶ The software-defined storage functionality of IBM Spectrum Virtualize with a full range of industry-leading data services such as dynamic tiering,
- ▶ IBM FlashCopy management, data mobility, and high-performance data encryption.
- ▶ Innovative data reduction pool (DRP) technology that includes deduplication and hardware-accelerated compression technology, with SCSI UNMAP support and all the thin provisioning, copy management, and efficiency you'd expect from IBM Spectrum Virtualize-based storage.

Figure 1-14 shows the IBM FlashSystem 9200 Control Enclosure with 24 NVMe FCM type drives installed.

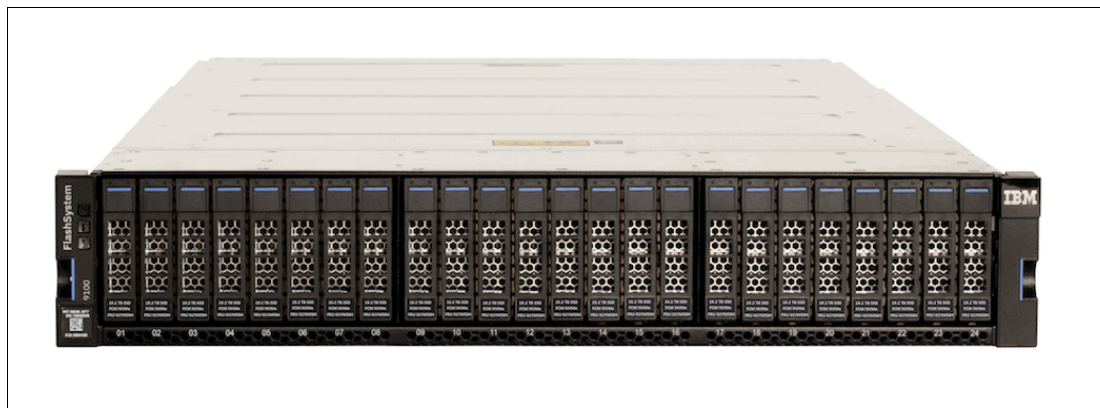


Figure 1-14 IBM FlashSystem 9200 control enclosure

The IBM FlashSystem 9200 solutions provide a single enterprise class platform to address the full spectrum of 21st-century data storage requirements. IBM FlashSystem 9200 is designed to simplify storage and accelerate business productivity, with the following benefits:

- ▶ NVMe-powered all-flash performance and IBM FlashCore reliability
- ▶ Easy integration and almost unlimited scalability
- ▶ Data services that can transform and modernize existing systems

Table 1-12 gives a summary of the host connections, drive capacities, features, and standard options with Spectrum Virtualize that are available on the IBM FlashSystem 9200.

Table 1-12 IBM FlashSystem 9200 host, drive capacity, and functions summary

Feature / Function	Description
Host interface	<ul style="list-style-type: none"> ▶ Up to 24 x 16 Gbps Fibre Channel (FC, NVMeoF) ▶ Up to 24 x 32 Gbps Fibre Channel (FC, NVMeoF) ▶ 8 x 10 Gbps Ethernet (iSCSI) ▶ Up to 12 x 25 Gbps (iSCSI, iSER - iWARP, RoCE)
Maximum drives supported	<ul style="list-style-type: none"> ▶ 24 NVMe drives per control enclosure ▶ 24 2.5" SAS drives per AFF expansion enclosure ▶ 92 2.5" SAS drives per A9F expansion enclosure ▶ Up to a maximum of 760 SAS drives in expansion enclosures per control enclosure
Supported NVMe drives	<ul style="list-style-type: none"> ▶ FlashCore Modules <ul style="list-style-type: none"> – 4.8 TB, 9.6 TB, 19.2 TB, and 38.4 TB with hardware compression ▶ Storage Class Memory (SCM) <ul style="list-style-type: none"> – 375 GB, 750 GB, 800 GB, 1.6 TB ▶ Industry-standard NVMe: <ul style="list-style-type: none"> – 800 GB, 1.92 TB, 3.84 TB, 7.68 TB, and 15.36 TB
Supported SAS drives	<ul style="list-style-type: none"> ▶ 2.5-Inch SAS SSD <ul style="list-style-type: none"> – 1.6 TB, 1.92 TB, 3.84 TB, 7.68 TB, 15.36 TB, and 30.72 TB
RAID levels	DRAID 1, 5, and 6 with dynamic DRAID expansion and TR RAID 1 and 10
Maximum IOPS (4K read hit)	4.5 million
Minimum latency (4K read hit)	<70 µs
Maximum IOPS (4K read miss)	1.2 million
Maximum bandwidth (256Kb readmiss)	45 GB/s
Advanced features	<ul style="list-style-type: none"> ▶ Data reduction via thin provisioning ▶ UNMAP ▶ Compression and deduplication ▶ Data-at-rest AES-XTS 256 encryption ▶ Easy Tier ▶ Data migration ▶ External virtualization
Replication features	<ul style="list-style-type: none"> ▶ FlashCopy ▶ Metro Mirror (synchronous) ▶ Global Mirror (asynchronous) ▶ Global Mirror with change volumes ▶ Three sites replication ▶ IBM HyperSwap (high availability)

Feature / Function	Description
Additional available advanced features	<ul style="list-style-type: none"> ▶ IBM Storage Insights Pro ▶ IBM Spectrum Virtualize for Public Cloud ▶ IBM Spectrum Control ▶ IBM Spectrum Protect Snapshot

For more information, see [V8.4.0.x Configuration Limits and Restrictions for IBM FlashSystem 9200](#).

1.3.1 Clustering rules and upgrades

The IBM Storwize and IBM FlashSystem products can be clustered with up to four control enclosures, using four I/O groups. There are some restrictions in the lower-level products, due to hardware and memory constraints, but in the main most, systems can cluster.

All the options and rules for clustering control nodes are show in Table 1-13

Table 1-13 Clustering control nodes matrix

Product Name	Machine Types / Models	Clustering Rules
IBM Storwize V5100	2077-AF4, 2077-424, 2077-U5B, 2078-AF4, 2078-424, 2078-U5B	<ul style="list-style-type: none"> ▶ A maximum of 2 I/O groups can be clustered. ▶ Storwize V5100 systems can be clustered only with another Storwize V5100 system.
IBM Storwize V7000	2076-724, 2076-U7B, 2076-U7A, 2076-AF6, 2076-624, 2076-524	<ul style="list-style-type: none"> ▶ A maximum of 4 I/O groups can be clustered. ▶ Any combination of V7000 Gen 2 or Gen2+, FS7200, FS9100, or FS9200 enclosures are permitted.
IBM FlashSystem 5015	2072-2N2, 2072-U12, 2072-2N4, 2072-U24,	<ul style="list-style-type: none"> ▶ Not supported - cannot cluster
IBM FlashSystem 5035	2072-3N2, 2072-V12, 2072-3N4, 2072-V24	<ul style="list-style-type: none"> ▶ A maximum of 2 I/O groups can be clustered. ▶ FlashSystem 5035 systems can be clustered with another FlashSystem 5035 system only.
IBM FlashSystem 5100 (FS5100)	2077-4H4, 2078-4H4, 2078-UHB	<ul style="list-style-type: none"> ▶ A maximum of 4 I/O groups can be clustered. ▶ FlashSystem 5100 systems can be clustered only with another FlashSystem 5100 system.
IBM FlashSystem 7200 (FS7200)	2076-824 2076-U7C	<ul style="list-style-type: none"> ▶ A maximum of 4 I/O groups can be clustered. ▶ Any combination of V7000 Gen 2 or Gen2+, FS7200, FS9100, or FS9200 enclosures are permitted. ▶ The FS7200 will NOT cluster with the Flash System V9000, Storwize V5035, or FlashSystem 5000.

Product Name	Machine Types / Models	Clustering Rules
IBM FlashSystem 9100 (FS9100)	9846-AF7, 9848-AF7, 9848-UF7, 9846-AF8, 9848-AF8, 9848-UF8	<ul style="list-style-type: none"> ▶ A maximum of 4 I/O groups can be clustered. ▶ Any combination of V7000 Gen 2 or Gen2+, FS7200, FS9100, or FS9200 enclosures are permitted. ▶ The FS9100 will NOT cluster with the Flash System V9000, Storwize V5035, or FlashSystem 5000.
IBM FlashSystem 9200 (FS9200)	9846-AG8, 9848-AG8, 9848-UG8	<ul style="list-style-type: none"> ▶ A maximum of 4 I/O groups can be clustered. ▶ Any combination of V7000 Gen 2 or Gen2+, FS7200, FS9100, or FS9200 enclosures are permitted. ▶ The FS9200 will NOT cluster with the Flash System V9000, Storwize V5035, or FlashSystem 5000.

1.3.2 Mixed clustering rules and licensing

From Spectrum Virtualize software version 8.2.0 onwards, when you cluster dissimilar models in a system, the resulting licensing scheme used for the system is overwritten by the licensing scheme of the most capable. For example, V7000 is over-ridden by FS7200, which is over-ridden by FS9200.

The extended rule is that the new or highest system overrules anything else in the cluster.

The explicit order of priority is as follows:

FS9200 > FS9100 > FS7200 > V7000

Example: When you add an FS7200 I/O group to an FS7200. If you then add an FS9100, the resulting cluster is an FS9100. If you then add an FS9200, the cluster reports as an FS9200.

The following points must be observed:

- ▶ All systems must have the same level of IBM Spectrum Virtualize software installed to be able to cluster.
- ▶ To cluster the Storwize V7000 systems, it must have an all-inclusive license.
- ▶ Migration must be done through additional I/O groups.
- ▶ The default layer is storage, but a replication layer is also supported for clustering.
- ▶ The systems that are listed in Table 1-13 on page 25 cannot be clustered with the IBM FlashSystem V9000 or the IBM Storage Virtualization Controller (SVC).

1.3.3 IBM FlashSystem 9200R Rack Solution overview

IBM FlashSystem 9200R is a pre-cabled, pre-configured rack solution that contains multiple IBM FlashSystem 9200 Control Enclosures and uses IBM Spectrum Virtualize to linearly scale the performance and capacity through clustering. For more information about this product, see *IBM FlashSystem 9200R Rack Solution Product Guide*, REDP-5593.

The IBM FlashSystem 9200R Rack Solution system has a dedicated FC network for clustering and optional expansion enclosures, which are delivered ready-assembled in a rack. Available with two, three, or four clustered IBM FlashSystem 9200 systems and up to four

expansion enclosures, it can be ordered as a IBM FlashSystem 9202R, IBM FlashSystem 9203R, or IBM FlashSystem 9204R system with the last number denoting the number of 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 9200 control enclosure and its expansion enclosures, the additional components of this solution are not covered under Enterprise Class Support (ECS). Instead, the components have their own warranty, maintenance terms, and conditions.

Rack rules

The IBM FlashSystem 9200R Rack Solution product represents a limited set of possible configurations. Each IBM FlashSystem 9200R Rack Solution order must contain the following components:

- ▶ Two, three, or four 9848 Model AG8 Control Enclosures.
- ▶ Two IBM SAN24B-6 or two IBM SAN32C-6 FC switches.
- ▶ Optionally, 0 - 4 9848 Model AFF Expansion Enclosures, with no more than one expansion enclosure per Model AG8 Control Enclosure and no mixing with the 9848 Model A9F Expansion Enclosure.
- ▶ Optionally, 0 - 2 9848 Model A9F Expansion Enclosures, with no more than one expansion enclosure per Model AG8 Control Enclosure and no mixing with 9848 Model A9F Expansion Enclosure.
- ▶ One 7965-S42 rack with the appropriate 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 AG8, AFF, and A9F Control Enclosures, the first and largest capacity enclosure includes feature code #AL01, with subsequent enclosures that use #AL02, #AL03, and #AL04 in capacity order. The 9848 Model AG8 Control Enclosure with #AL01 must also have #AL0R included.

Following the initial order, each 9848 Model AG8 Control Enclosures can be upgraded through MES.

More components can be ordered separately and added to the rack within the configuration limitations of the IBM FlashSystem 9200 system. Clients 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.

Table 1-14 shows the IBM FlashSystem 9200R Rack Solution combinations, the MTMs, and their associated feature codes.

Table 1-14 IBM FlashSystem 9200R Rack Solution combinations

Machine type and model	Description	Quantity
7965-S42	IBM Enterprise Slim Rack	1
8960-F24	IBM SAN24B-6 Fibre Channel switch (Brocade)	2 ^a

Machine type and model	Description	Quantity
8977-T32	IBM SAN32C-6 Fibre Channel switch (Cisco)	2 ^a
9848-AFF	IBM FlashSystem 9000 2U SFF Expansion Enclosure with 3-year Warranty and ECS	0 - 4 ^b
9848-AG8	IBM FlashSystem 9200 Control Enclosure with 3-year Warranty and ECS	2, 3, or 4
9848-A9F	IBM FlashSystem 9000 5U LFF high-density Expansion Enclosure with 3-year Warranty and ECS	0 - 2 ^b

a. For the FC switch, choose either two of machine type (MT) 8977 or two of MT 8960.

b. For extra expansion enclosures, choose either model AFF, model A9F, or none. You cannot use both.

For more information on the FlashSystem 9200R solution, see the following IBM Redbooks:

- ▶ *IBM FlashSystem 9200R Rack Solution Product Guide*, REDP-5593
- ▶ *Implementing the IBM FlashSystem with IBM Spectrum Virtualize V8.4*, SG24-8467

1.4 Advanced functions for data reduction

The IBM FlashSystem range can function as a feature-rich, software-defined storage layer that virtualizes and extends the functionality of all managed storage. These include data reduction, dynamic tiering, copy services, and high-availability configurations. In this capacity, the IBM FlashSystem acts as the virtualization layer between the host and other external storage systems, providing flexibility, and extending functionality to the virtualized external storage capacity.

The IBM FlashSystem 5100, 7200, 9100 and 9200 all employ several features to assist with data reduction and the ability to increase their effective capacity.

1.4.1 FlashCore Modules (FCM)

The IBM FlashSystem 5100, 7200, 9100, and 9200 all have the option to be supplied with either FCMs or industry-standard NVMe drives. If the FCM option is chosen, then the user can take advantage of the built-in hardware compression, which will automatically try to compress the stored data when written to the drives. These FCMs can be used either with standard pools or DRP pools.

1.4.2 Data reduction pools (DRP)

Data reduction pools (DRP) represent a significant enhancement to the storage pool concept. This is because the virtualization layer is primarily a simple layer that executes the task of lookups between virtual and physical extents. Now with the introduction of data reduction technology, compression, and deduplication, it is more of a requirement to have an uncomplicated way to stay “thin”. The pools enable you to automatically de-allocate (not to be

confused with deduplicate) and reclaim capacity of thin-provisioned volumes containing deleted data.

1.4.3 Deduplication

Deduplication can be configured with thin-provisioned and compressed volumes in data reduction pools 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 signature of the new chunk 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, resulting in a sizeable reduction of the amount of data that must be stored.

1.4.4 Thin provisioning

In a shared storage environment, thin provisioning is a method for optimizing the use of available storage. It relies on allocation of blocks of data on demand versus the traditional method of allocating all of the blocks up front.

This methodology 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).

1.4.5 Thin-provisioned FlashCopy snapshots

Thin-provisioned IBM FlashCopy (or snapshot function in the GUI) uses disk space only when updates are made to the source or target data, and not for the entire capacity of a volume copy.

1.5 Advanced software features

Some of the advanced software features of the IBM FlashSystem 5100, 5200, 7200, 9100, and 9200 are as follows:

- ▶ Data Migration
- ▶ Copy Services
 - Metro Mirror
 - Global Mirror
 - Global Mirror with Change Volumes
 - FlashCopy
 - Remote Mirroring
- ▶ External Virtualization
- ▶ Easy Tier

1.5.1 Data migration

The IBM FlashSystem range provides online volume migration while applications are running, which is possibly the greatest single benefit for storage virtualization. This capability enables data to be migrated on and between the underlying storage subsystems without any effect on

the servers and applications. In fact, this migration is performed without the knowledge of the servers and applications that it even occurred. The IBM FlashSystem deliver these functions in a homogeneous way on a scalable and highly available platform over any attached storage and to any attached server.

1.5.2 Copy services

Advanced copy services are a class of functionality within storage arrays and storage devices that enable various forms of block-level data duplication locally or remotely. By using advanced copy services, you can make mirror images of part or all of your data eventually between distant sites. Copy services functions are implemented within an IBM FlashSystem (FlashCopy and Image Mode Migration), or between one IBM FlashSystem and another IBM FlashSystem, or any other member of the IBM Spectrum Virtualize family, in three different modes:

- ▶ *Metro Mirror* is the IBM branded term for synchronous Remote Copy function.
- ▶ *Global Mirror* is the IBM branded term for the asynchronous Remote Copy function.
- ▶ *Global Mirror with Change Volumes* is the IBM branded term for the asynchronous Remote Copy of a locally and remotely created FlashCopy.

Remote replication can be implemented using both Fibre Channel and Internet Protocol (IP) network methodologies.

For more, see Chapter 6, “Copy services” on page 229.

FlashCopy

FlashCopy is the IBM branded name for point-in-time copy, which is sometimes called time-zero (T0) copy. This function makes a copy of the blocks on a source volume and can duplicate them on 1 - 256 target volumes.

Remote mirroring

The three remote mirroring modes are implemented at the volume layer within the IBM FlashSystem family. They are collectively referred to as Remote Copy capabilities. In general, the purpose of these functions is to maintain two copies of data.

Often, but not necessarily, the two copies are separated by distance. The Remote Copy can be maintained in one of two modes: synchronous or asynchronous, with a third asynchronous variant:

- ▶ Metro Mirror
- ▶ Global Mirror
- ▶ Global Mirror with Change Volumes

1.5.3 Easy Tier

Easy Tier is a performance function that automatically migrates or moves extents of a volume from one storage tier to another storage tier. With IBM FlashSystem, Easy Tier supports four kinds of storage tiers.

Consider the following information about Easy Tier:

- ▶ Easy Tier monitors the host volume I/O activity as extents are read, and migrates the most active extents to higher performing tiers.

- ▶ The monitoring function of Easy Tier is continual but, in general, extents are migrated over a 24-hour period. As extent activity cools, Easy Tier moves extents to slower performing tiers.
- ▶ Easy Tier creates a migration plan that organizes its activity to decide how to move extents. This plan can also be used to predict how extents will be migrated.

1.5.4 External virtualization

The IBM FlashSystem range includes data virtualization technology to help insulate hosts, hypervisors, and applications from physical storage. This enables them to run without disruption, even when changes are made to the underlying storage infrastructure. The IBM FlashSystem functions benefit all virtualized storage.

For example, Easy Tier and Data Reduction Pools with compression help improve performance and increase effective capacity, where high-performance thin provisioning helps automate provisioning. These benefits can help extend the useful life of existing storage assets, reducing costs. Additionally, because these functions are integrated into the IBM FlashSystem 5100, 7200, 9100, and 9200, they can operate smoothly together, reducing management effort.

1.5.5 IBM HyperSwap

HyperSwap capability enables each volume to be presented by two IBM FlashSystem family I/O groups. The configuration tolerates combinations of node and site failures, using host multipathing driver based on the one that is available for the IBM FlashSystem family. The IBM FlashSystem provides GUI and CLI management of the HyperSwap function.

A more detailed overview and explanation of the HyperSwap function can be found in Appendix A, “IBM i considerations” on page 525.

For more information, see [IBM FlashSystem 9200 8.4.0 Documentation - HyperSwap function](#).

1.5.6 Licensing

The base license that is provided with the system includes the use of 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 on the system.

For a more detailed overview and explanation of the licensing on the IBM FlashSystem 5100, 7200, 9100 and 9200, see the “Licensing and Features” chapter in *Implementing the IBM FlashSystem with IBM Spectrum Virtualize V8.4*, SG24-8467.



Storage area network

The storage area network (SAN) is one of the most important aspects when implementing and configuring IBM Spectrum Virtualize and IBM FlashSystem.

This chapter does not describe how to design and build a flawless SAN from the beginning. Rather, it provides guidance to connect IBM Spectrum Virtualize and Storwize in an existing SAN to achieve a stable, redundant, resilient, scalable, and performance-likely environment. However, you can take the principles here into account when building your SAN.

Important: This chapter was written specifically for IBM FlashSystem 9200, however most of the general principles apply to the IBM FlashSystem 9100.

If you are in doubt as to whether the principles are applicable to the FlashSystem 9100, contact your local IBM representative.

For more information, see the FlashSystem 9200 Product Guide at:

<https://www.redbooks.ibm.com/abstracts/redp5586.html?Open>

This chapter includes the following sections:

- ▶ 2.1, “SAN topology general guidelines” on page 34
- ▶ 2.2, “SAN topology-specific guidelines” on page 36
- ▶ 2.3, “IBM FlashSystem 9200 controller ports” on page 44
- ▶ 2.4, “Zoning” on page 47
- ▶ 2.5, “Distance extension for Remote Copy services” on page 55
- ▶ 2.6, “Tape and disk traffic that share the SAN” on page 61
- ▶ 2.7, “Switch interoperability” on page 61

2.1 SAN topology general guidelines

The SAN topology requirements for IBM FlashSystem do not differ too much from any other SAN. Remember that a well-sized and designed SAN allows you to build a redundant and failure-proof environment, as well as minimizing performance issues and bottlenecks. Therefore, before installing any of the products 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 SAN design and preferred practices, see the [SAN Fabric Administration Best Practices Guide Support Perspective](#).

A topology is described in terms of how the switches are interconnected. Several different SAN topologies exist, such as core-edge, edge-core-edge, and full MeSH. Each topology has its uses, scalability, and also its cost, so one topology will be 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 well, redundant, and resilient.

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 when troubleshooting problems.

Because most SAN installations continue to grow over the years, the main SAN industry-lead companies design their products in a way to support a certain growth. Keep in mind that 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
- ▶ Monitor for slow drain device issues

From the scalability standpoint, ensure that your SAN will support the new storage and host traffic. Make sure that the chosen topology will also support a growth not only in performance, but also in port density.

If new ports need to 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 additional information, see 2.1.2, “ISL considerations” on page 35.

Consider the use of SAN director-class switches. They reduce the number of switches in a SAN and provide the best scalability available. Most of the SAN equipment vendors provide high port density switching devices.

Therefore, if possible, plan for the maximum size configuration that you expect your IBM FlashSystem installation to reach. Planning for the maximum size does not mean that you must purchase all of the SAN hardware initially. It only requires you to design the SAN to be able 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 as the core of a SAN topology. Consequently, 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 FlashSystem installation, consider the following practices to your SAN Inter-switch link design:

- ▶ Beware of the ISL oversubscription ratio
 - The standard recommendation is up to 7:1 (seven hosts using 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%, schedule fabric changes to distribute the load further.
- ▶ Avoid unnecessary ISL traffic
 - If you plan to use external virtualized storages, connect all FlashSystem canister ports in a clustered system to the same SAN switches/Directors as all of the storage devices with which the clustered system of FlashSystem is expected to communicate. Conversely, storage traffic and internode traffic must never cross an ISL, except during migration scenarios.
 - Keep high-bandwidth utilization servers and I/O Intensive application on the same SAN switches as the FlashSystem 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 inter-switch link 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 the FlashSystem 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 going to the switch where the Flash System 9200 and 9200 is connected will also handle the traffic coming from the switches on the edges, as shown in Figure 2-1.

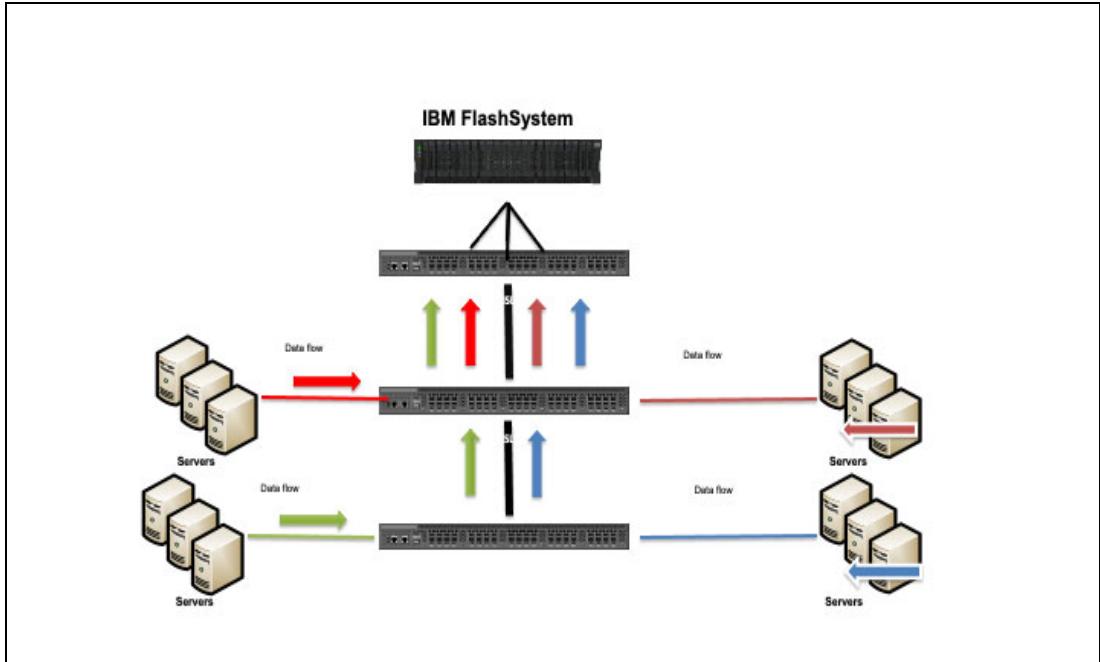


Figure 2-1 ISL data flow

- ▶ 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 you interconnect SAN directors through ISL, spread the Inter-Switch Link (ISL) cables across different directors' blades. In a situation where an entire blade fails, the ISL will still be redundant through the links connected to other blades.
- ▶ Plan for the peak load, not for the average load.

2.2 SAN topology-specific guidelines

Some preferred practices apply to all SANs, as described in 2.1, “SAN topology general guidelines” on page 34. However each SAN topology has its own specific preferred practices requirements. The following topic shows the difference between the different kinds of topology and highlights the specific considerations for each of them.

This section covers the following topologies:

- ▶ Single switch fabric
- ▶ Core-edge fabric
- ▶ Edge-core-edge
- ▶ Full MeSH

2.2.1 Single switch SANs

The most basic IBM FlashSystem 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 is a low-cost design solution that has the advantage of simplicity and is a sufficient architecture for small-to-medium FlashSystem installations.

One of the advantages of a single switch SAN is that when all servers and storages are connected to the same switches, there is no hop.

Note: To meet redundancy and resiliency requirements, a single switch solution needs at least two SAN switches or directors, with one per different fabric.

The preferred practice is to use a multislot director-class single switch rather than 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 34.

The single switch topology, as shown in Figure 2-2, has only two switches, so the FlashSystem ports must be equally distributed on both fabrics.

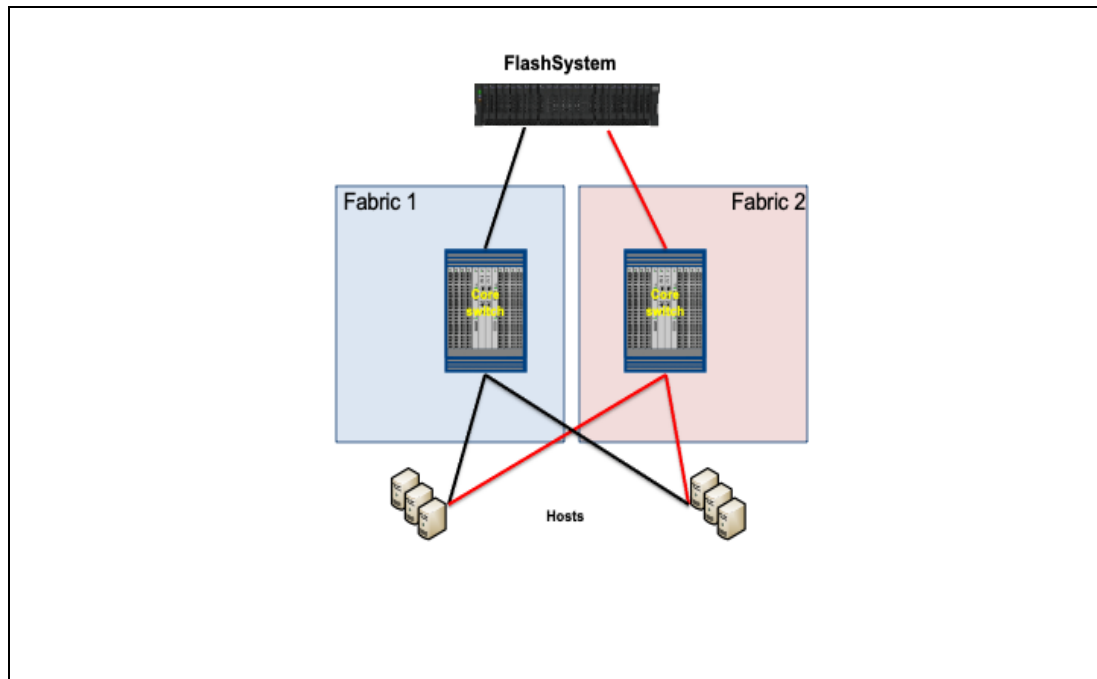


Figure 2-2 Single Switch Topology

2.2.2 Basic core-edge topology

The core-edge topology (as shown in Figure 2-3 on page 38) 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 FlashSystem 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.

When the FlashSystem and servers are connected to different switches, the hop count for this topology is one.

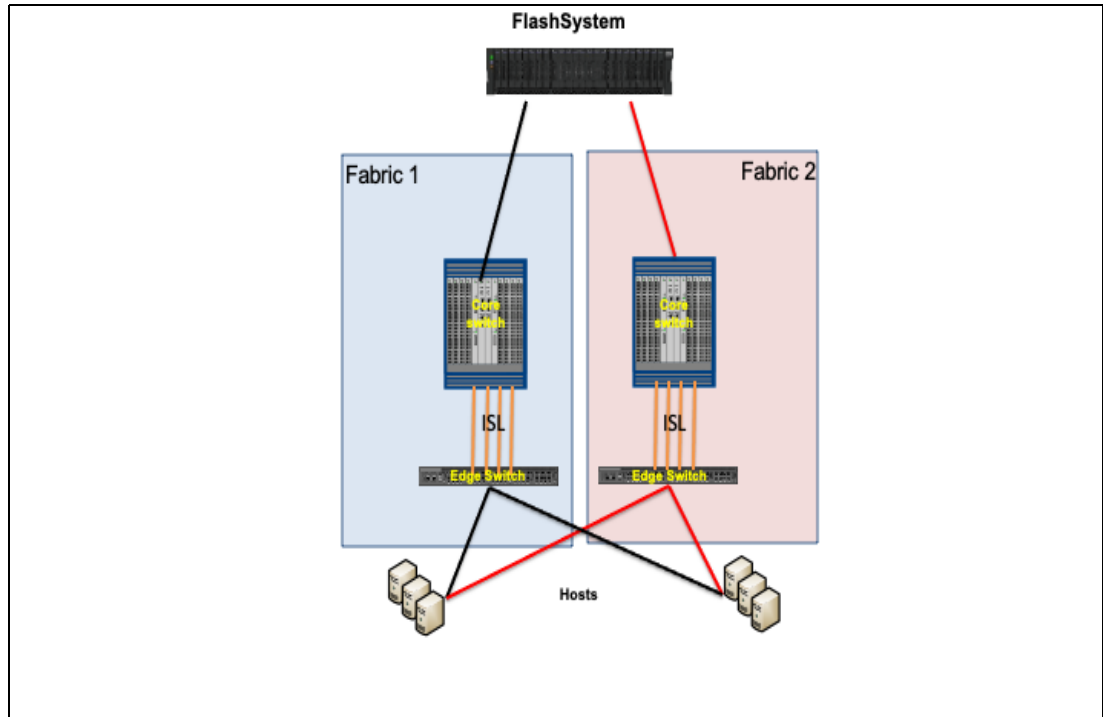


Figure 2-3 Core/Edge Topology

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 made up of multislot director-class SAN switches is insufficient. This design is useful for large, multiclustered system installations. Similar to a regular core-edge, the edge switches can be of any size, and multiple ISLs must be installed per switch.

Figure 2-4 on page 39 shows an edge-core-edge topology with two different edges, one of which is exclusive for the FlashSystem and high-bandwidth servers. The other pair is exclusively for servers.

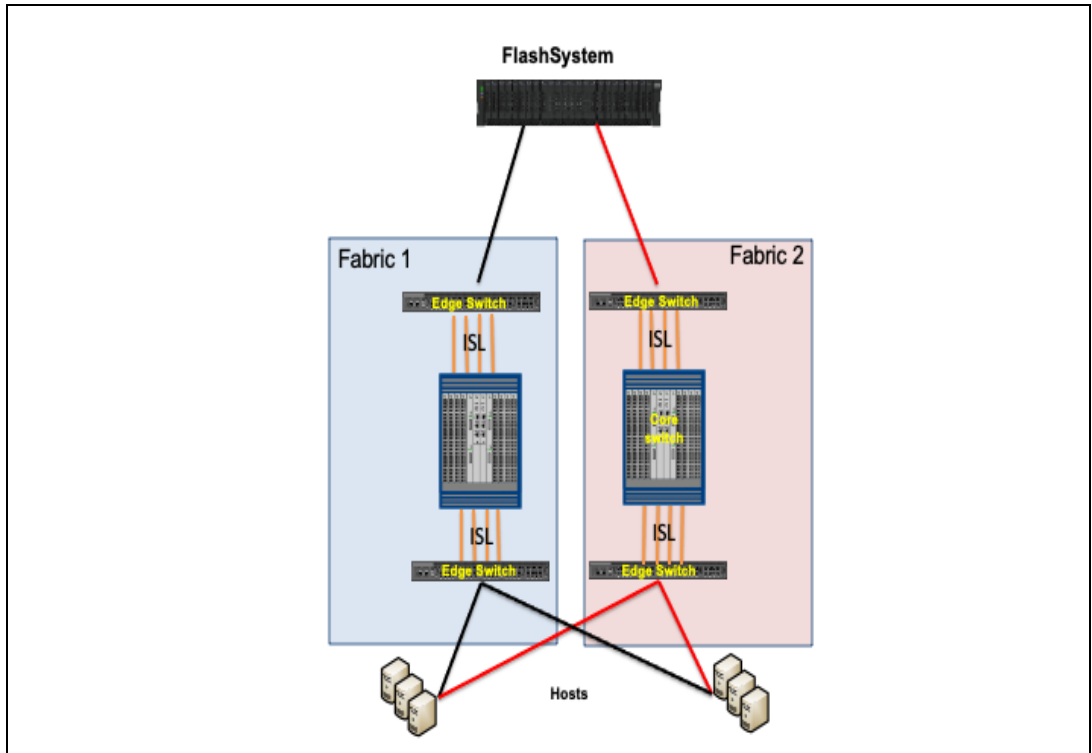


Figure 2-4 Edge-Core-Edge Topology

Edge-core-edge fabrics allow better isolation between tiers. For additional information, see 2.2.6, “Device placement” on page 41.

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 if the number of hops is not more than one. Figure 2-5 on page 40 shows a full MeSH topology.

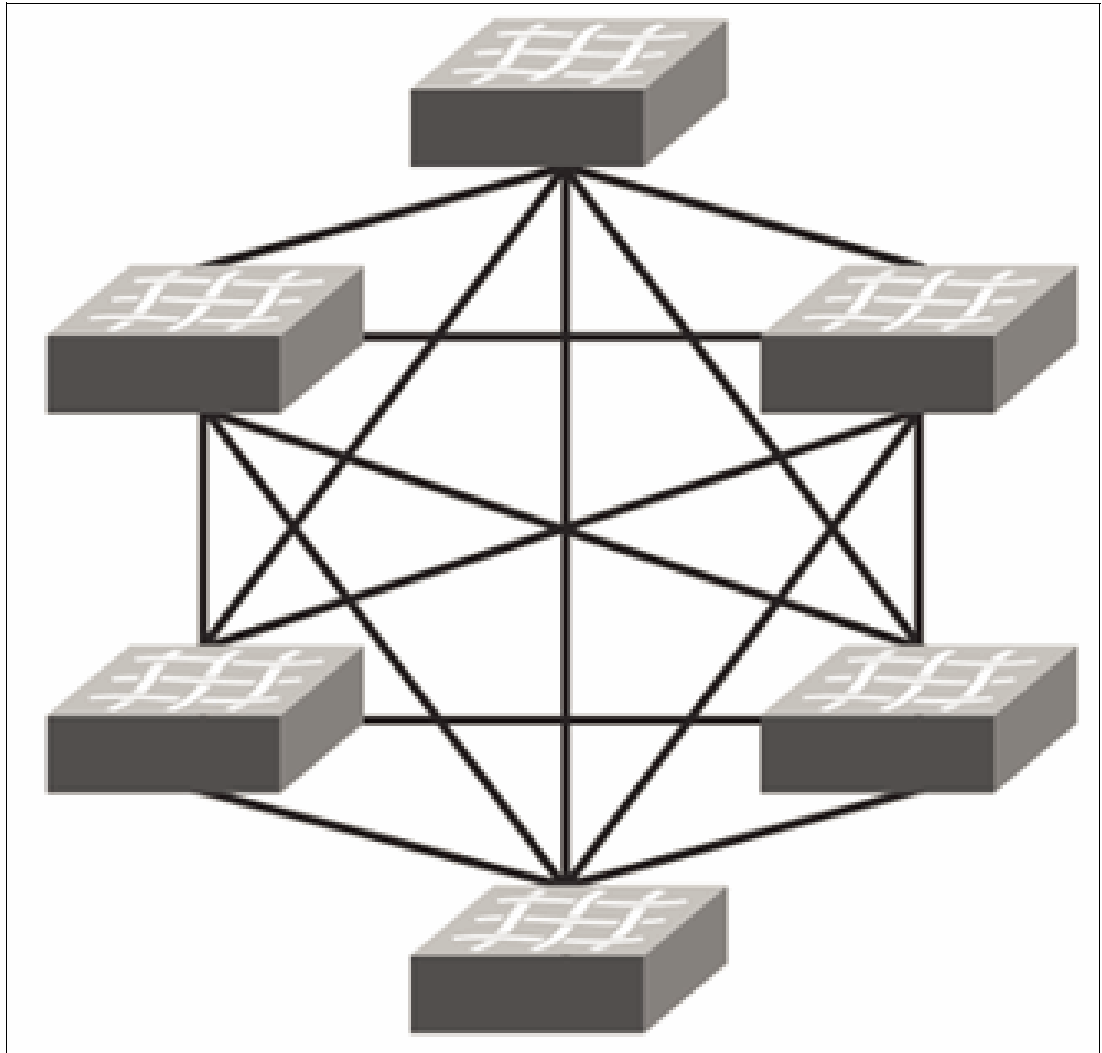


Figure 2-5 Full MeSH topology

2.2.5 IBM FlashSystem as a SAN bridge

IBM FlashSystem now has a maximum of 24 ports. In addition to the increased throughput capacity, this number of ports enables new possibilities and allows different kinds of topologies and migration scenarios.

One of these topologies is the use of a FlashSystem as a bridge between two isolated SANs. This configuration is useful for storage migration or sharing resources between SAN environments without merging them. Another use is if you have devices with different SAN requirements in your installation.

Figure 2-6 shows an example of an FlashSystem as a SAN bridge.

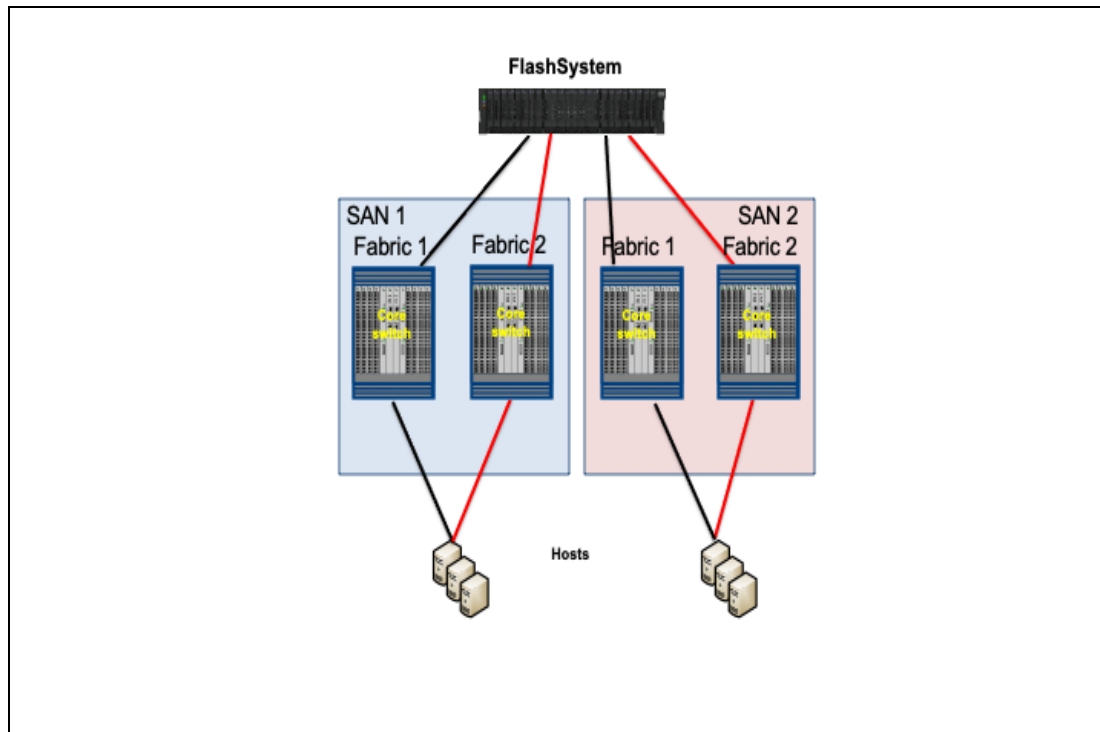


Figure 2-6 FlashSystem as a SAN Bridge

Notice in Figure 2-6 that both SANs (Blue and Pink) are isolated and there is no communication through ISLs. When connected to both fabrics, FlashSystem is able to serve hosts and virtualize storages from either fabrics. They can provide disks to hosts from both SAN and when it has external virtualized storage, it can provide disks from storage on the Pink SAN (right), for example, to hosts on blue SAN (left).

2.2.6 Device placement

In a well-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 certain 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.

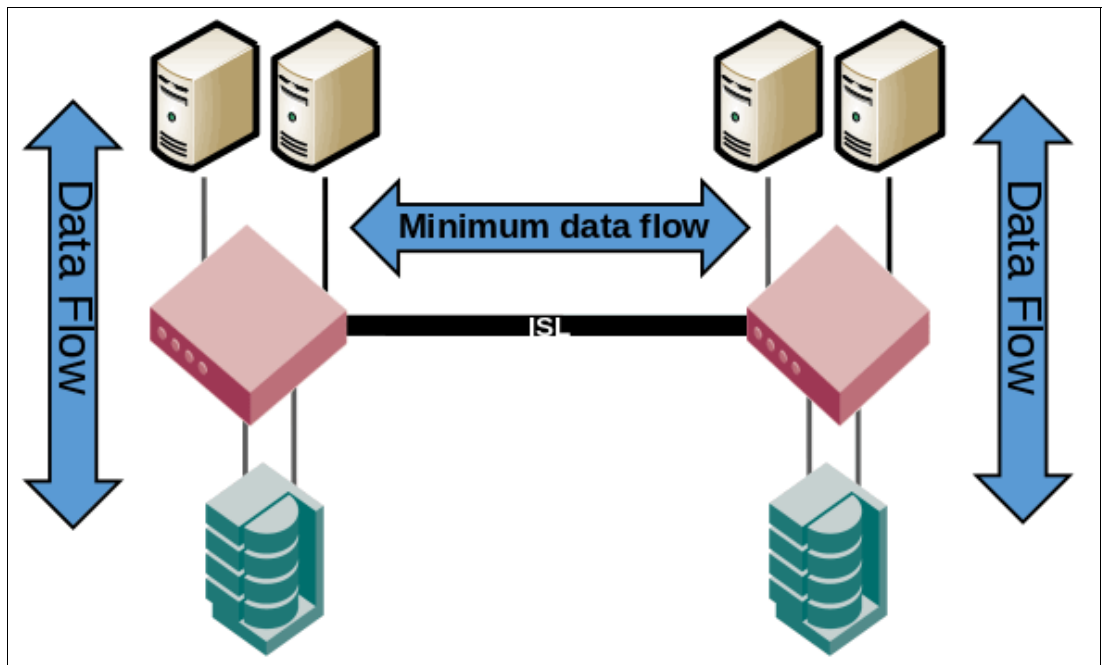


Figure 2-7 Storage and hosts attached to the same SAN switch

This solution can fit perfectly in small and medium SANs. However, it is not as scalable as other available topologies. As stated in 2.2.3, “Edge-core-edge topology” on page 38, the most scalable SAN topology is the edge-core-edge. Besides scalability, this topology provides various resources to isolate the traffic and reduce possible SAN bottlenecks.

Figure 2-8 shows an example of traffic segregation on the SAN using edge-core-edge topology.

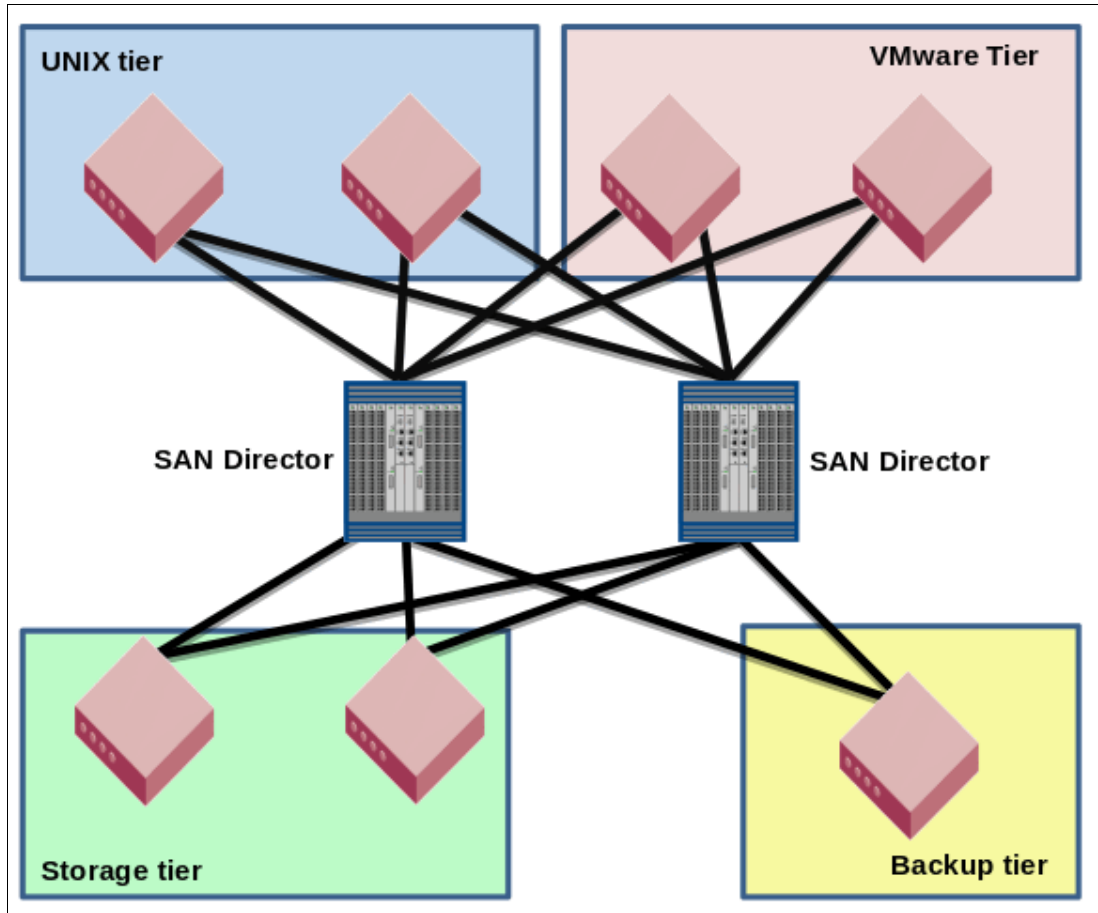


Figure 2-8 Segregation using edge-core-edge

Even when sharing the same core switches, it is possible to use virtual switches (see 2.2.7, “SAN partitioning” on page 43 for details) to isolate one tier from the other. This configuration helps avoid traffic congestion 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, respectively, *Virtual Fabric* and *Virtual SAN (VSAN)*.

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.” Virtual switches are then 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 completely transparent and so the same guidelines and practices that apply to physical switches apply also to the virtual ones.

While 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 based on FlashSystem.

For a description of the IBM FlashSystem business continuity solutions, see Chapter 7, “Business continuity” on page 339.

2.3 IBM FlashSystem 9200 controller ports

Port connectivity options are significantly increased with IBM FlashSystem 9200 hardware. Models 9846-AG8 and 9848-AG8 deliver up to 12x16 Gb or 12x32GbFC ports per node canister as shown in Table 2-1.

Table 2-1 FlashSystem 9200

Feature	FlashSystem 900
Fibre Channel HBA	3x Quad 16 Gb or 3x Quad 32Gb
Ethernet I/O	2x Dual 25Gb iWARP/RoCE for iSCSI or iSER
Built in ports	4x 10 Gb for internet small computer systems interface (iSCSI)
Serial attached SCSI (SAS) expansion ports	1x Quad 12 Gb SAS (2 ports active)

Note: FlashSystem 9200 node canisters have three peripheral component interconnect express (PCIe) slots which you can combine the cards as needed. If expansions is used, one of the slots must have the SAS expansion card. Then 2 ports are left for fiber channel Host Bus Adapter (HBA) cards, Internet Wide-area RDMA Protocol (iWARP) or RDMA over Converged Ethernet (RoCE) Ethernet cards. For more information see [IBM FlashSystem 9200 8.4.0 Documentation - Technical Overview](#).

This section describes some preferred practices and use cases that show how to connect a FlashSystem on the SAN to use this increased capacity.

2.3.1 Slots and ports identification

The IBM FlashSystem 9200 can have up to three quad Fibre Channel (FC) HBA cards (12 FC ports) per node canister. Figure 2-9 shows the port location in the rear view of the FlashSystem 9200 node canister.

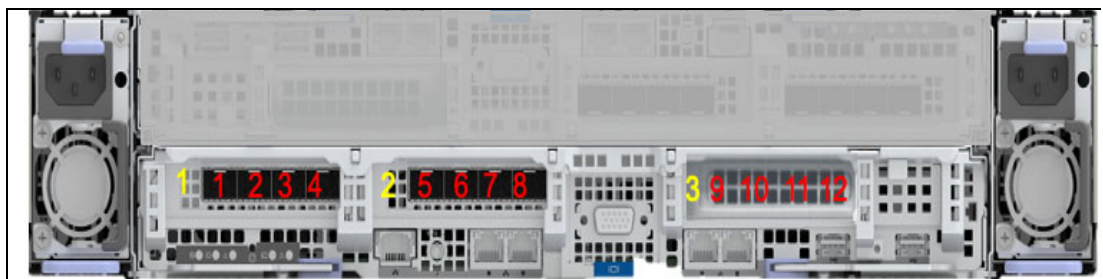


Figure 2-9 Port location in FlashSystem 9200 rear view

For maximum redundancy and resiliency, spread the ports across different fabrics. Because the port count varies according to the number of cards included in the solution, try to keep the port count equal on each fabric.

2.3.2 Port naming and distribution

In the field, fabric naming conventions vary. However, it is common to find fabrics with names such as 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 management and troubleshooting, after their denomination followed by 1 and 2 or A and B, which specifies that the devices 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 2-10.

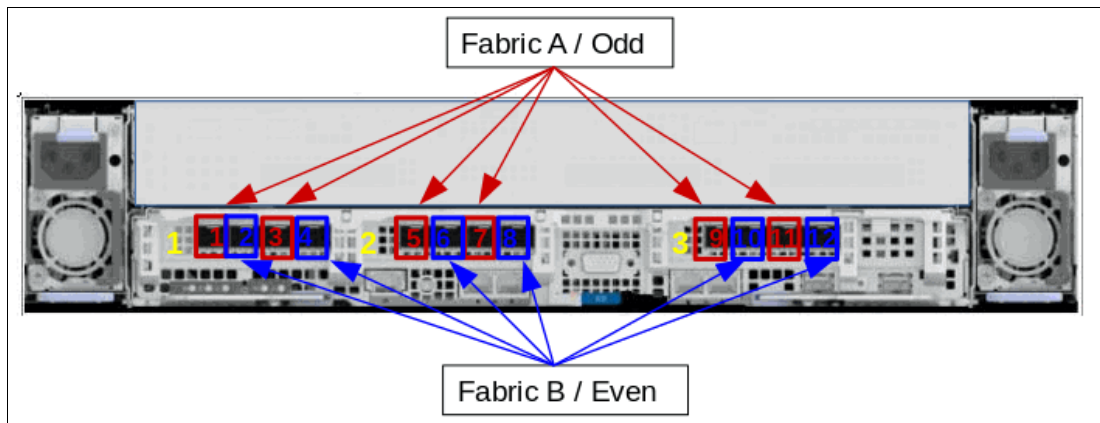


Figure 2-10 FlashSystem 9200 port distribution

As a preferred practice, assign specific uses to specific FlashSystem 9200 ports. This technique helps to optimize the port utilization by aligning the internal allocation of hardware CPU cores and software I/O threads to those ports.

Figure 2-11 on page 46 shows the specific port use guidelines for the FlashSystem 9200.

Card / Port	4 ports	8 ports	12 ports
Card 1 Port 1	Host/Storage/Inter-node	Host/Storage	Host/Storage
Card 1 Port 2	Host/Storage/Inter-node	Host/Storage	Host/Storage
Card 1 Port 3	Host/Storage/Replication*	Inter-node	Inter-node
Card 1 Port 4	Host/Storage/Replication*	Inter-node	Inter-node
Card 2 Port 1		Host/Storage	Host/Storage
Card 2 Port 2		Host/Storage	Host/Storage
Card 2 Port 3		Host/Storage/Replication*	Host/Storage/Replication*
Card 2 Port 4		Host/Storage/Replication*	Host/Storage/Replication*
Card 3 Port 1			Host/Storage
Card 3 Port 2			Host/Storage
Card 3 Port 3			Host/Storage
Card 3 Port 4			Host/Storage
localfcportmask	0011	00001100	000000001100
partnerfcportmask	1100	11000000	000011000000
* Use for host/storage in case no replication is in place. ** Do not use the same port for replication and inter-node traffic. *** For HyperSwap, dedicate ports for inter-node traffic			

Figure 2-11 Port masking configuration on FlashSystem 9200

Note:

- ▶ If you are using an IBM FlashSystem 9200 system with a single I/O group, the system keeps `localfcportmask` set to `111111111111` and does not allow you to change this. However, this is not a problem because the inter-node traffic happens on the internal PCI midplane link. The port-masking recommendations described in Figure 2-11 apply to systems with more than one I/O group.
- ▶ Depending on the workload or number of I/O groups, you can reserve ports 1 and 2 from card 3 for inter-node traffic. In this case, you will have four ports for inter-node traffic, and the `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 generally the preferred practice. However, traffic segregation can also provide some benefits in terms of troubleshooting and host zoning management. Consider, for instance, SAN congestion conditions due to a slow draining device.

In this case, segregating the ports simplifies the identification of the device causing the problem. At the same time, it limits 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 advised to implement the port traffic segregation with configurations with 12 ports only.

Buffer credits

FlashSystem 9200 has a predefined number of buffer credits. The number of buffer credits determines the available throughput over distances as follows: 4-port 16 Gbps adapters have 40 credits available per port, saturating links at up to 5 km at 16 Gbps.

Switch port buffer credit: For stretched cluster and IBM HyperSwap configurations not using ISLs for the internode communication, you should set the switch port buffer credits to match the IBM FlashSystem 9200 port.

Performance recommendation: Balance your bandwidth and make sure you have enough incoming bandwidth to saturate your backend bandwidth. Follow the simple guidance on where to plug them in, and try to balance your io across the ports.

2.4 Zoning

This section describes the zoning recommendations for FlashSystem 9200. For external storage virtualization zoning information, see *IBM System Storage SAN Volume Controller, IBM Storwize V7000, and IBM FlashSystem 7200 Best Practices and Performance Guidelines*, SG24-7521 as the recommendations are the same.

Important: Errors that are caused by improper FlashSystem 9200 zoning are often difficult to isolate and the steps to fix them can impact the SAN environment. Therefore, create your zoning configuration carefully.

The initial configuration for FlashSystem 9200 requires the following three zone types:

- ▶ Internode and intra-cluster zones
- ▶ Replication zones (if using replication)
- ▶ Host to FlashSystem 9200 zoning

Each zone type has its own guidelines, which are detailed in 2.4.1, “Types of zoning” on page 47.

Note: Although an internode or intra-cluster zone is not necessary for non clustered FlashSystem 9200 family, it is generally preferred to have **one**.

2.4.1 Types of zoning

Modern SAN switches have two types of zoning available: Port zoning, and worldwide port name (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 particular SAN switch port.

Zoning types and NPIV: Avoid the use of a zoning configuration that has a mix of port and WWPN zoning. For NPIV configurations, host zoning must use the WWPN zoning type.

Traditional zone design preferred practice calls for *single initiator* zoning. This means that a zone can consist of many target devices but only one initiator. This is because target devices will usually wait for an initiator device to connect to them, while initiators will actively attempt to connect to each device to which they are zoned. The single-initiator approach removes the possibility that a misbehaving initiator will affect other initiators.

The drawback to single initiator zoning is that on a large SAN having many zones can make the SAN administrators job more difficult, and the number of zones on a large SAN can exceed the zone database size limits.

Cisco and Brocade have both developed features that can reduce the number of zones by allowing the SAN administrator to control which devices in a zone can talk to other devices in the zone. The features are called Cisco Smart Zoning and Brocade Peer Zoning. Both Cisco Smart Zoning and Brocade Peer Zoning are supported with IBM Spectrum Virtualize and Storwize systems. A brief overview of both is provided below.

Cisco Smart Zoning

Cisco Smart Zoning is a feature that, when enabled, restricts the initiators in a zone to communicating only with target devices in the same zone. For our cluster example, this would allow a SAN administrator to zone all of the host ports for a VMware cluster in the same zone with the storage ports that all the hosts need access to. Smart Zoning configures the access control lists in the fabric routing table to only allow the hosts to communicate with target ports.

For more information about Smart Zoning, see [Cisco - 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 similar functionality of restricting 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 only communicate with the principal device, not with each other.

As with Cisco, the communication is enforced in the fabric routing table. You can see more information about Peer Zoning on chapter 4.2.3 of *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. For intracluster, back-end, and replication zoning, use traditional zoning instead.

Simple zone for small environments

As an option for small environments, the IBM FlashSystem-based system supports 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, like the IBM Spectrum Virtualize based system ports which are target + initiator.

So, a valid zone can be 32 host ports plus 8 IBM FlashSystem based system ports. Include exactly 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.

For more information, see: [IBM FlashSystem 9200 8.4.0 Documentation - Zoning Details](#).

2.4.2 Pre-zoning tips and shortcuts

Several tips and shortcuts are available for FlashSystem 9200 zoning.

Naming convention and zoning scheme

When you create and maintain a FlashSystem 9200 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.

Remember that environments have different requirements, which means that the level of detailing 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 FlashSystem 9200 naming convention, see 10.12.1, “Naming conventions” on page 468.

Aliases

Use zoning aliases when you create your FlashSystem 9200 zones if they are available on your particular type of SAN switch. Zoning aliases makes your zoning easier to configure and understand, and causes fewer possibilities for errors, see Table 2-2.

Table 2-2 Alias names examples

Port/WWPN	Use	Alias
Card 1 Port 1 physical WWPN	External Storage back-end	FS9200_N1P1_STORAGE
Card 1 Port 1 NPIV WWPN	Host attachment	FS9200_N1P1_HOST_NPIV
Card 1 Port 2 physical WWPN	External Storage back-end	FS9200_N1P2_STORAGE
Card 1 Port 2 NPIV WWPN	Host attachment	FS9200_N1P2_HOST_NPIV
Card 1 Port 3 physical WWPN	Inter-node traffic	FS9200_N1P3_CLUSTER
Card 1 Port 3 NPIV WWPN	No use	No alias
Card 1 Port 4 physical WWPN	Inter-node traffic	FS9200_N1P4_CLUSTER
Card 1 Port 4 NPIV WWPN	No use	No alias
Card 2 Port 3 physical WWPN	Replication traffic	FS9200_N1P7_REPLICATION
Card 2 Port 3 NPIV WWPN	No use	No alias
Card 2 Port 4 physical WWPN	Replication traffic	FS9200_N1P8_REPLICATION
Card 2 Port 4 NPIV WWPN	No use	No alias

Note: In Table 2-2 not all ports are used as example for aliases. Remember that NPIV ports can be used for host attachment only. If you are using external virtualized back-ends, use the physical port WWPN. For replication and inter-node, also use the physical WWPN. On the alias examples in Table 2-2, the N stands for node, and all examples are from node 1. An N2 example is *FS9200_N2P4_CLUSTER*.

One approach is to create template zones from the host to the FlashSystem 9200. The zoning should contain one alias from the host, and this alias must contain one initiator, and one alias from each node canister from the FlashSystem 9200, preferably the same port.

As an example, create the following zone aliases:

- ▶ One zone alias for each FlashSystem 9200 port
- ▶ One alias for each host initiator
- ▶ One host initiator alias to FlashSystem 9200 port 1 from node 1, and to port 1 from node 2. Then, name this zone *HOST1_HBA1_T1_FS92009200*
- ▶ (Optional) A second host initiator alias to FlashSystem 9200 port 3 from node 1 and to port 3 from node 2. Then, name this zone *HOST2_HBA1_T2_FS9200*

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 FlashSystem 9200 ports. Table 2-3 shows how the aliases are distributed if you create template zones described in the example.

Table 2-3 Template examples

Template	FS9200 ports on Fabric A	FS9200 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

Note: The number of templates varies depending on how many fiber ports are in your system, and how many are dedicated to host access. The port numbers are examples; you can use different ports depending on the number of HBA cards. Plan accordingly.

2.4.3 IBM FlashSystem 9200 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 node canisters. In the FlashSystem family, internode communication takes place primarily through the internal PCI connectivity between the two canisters of a control enclosure. However, for the clustered IBM FlashSystem, the internode communication requirements are similar to those of the SAN Volume Controller.

To establish efficient, redundant, and resilient intracluster communication, the intracluster zone must contain at least two ports from each node/canister.

For FlashSystem 9200 clusters with two I/O groups or more with eight ports, you should isolate the intracluster traffic by dedicating node ports specifically to inter-node communication. The ports to be used for intracluster communication varies according to the port count. See Figure 2-12 on page 51 for port assignment recommendations.

NPIV configurations: On NPIV-enabled configurations, use the physical WWPN for the intracluster zoning.

Only 16 port logins are allowed from one node to another node in a SAN fabric. Ensure that you apply the proper port masking to restrict the number of port logins. For FlashSystem 9200 clusters with two or more I/O groups, without port masking, any FlashSystem 9200 port and any member of the same zone can be used for intracluster communication. This includes the port members of FlashSystem 9200 that connect to host and external virtualized back-ends.

2.4.4 IBM FlashSystem 9200 host zones

The preferred practice to connect a host into a FlashSystem 9200 is to create a single zone to each host port. This zone must contain the host port and *one* port from each FlashSystem 9200 node canister that the host must access, as shown in Figure 2-12.

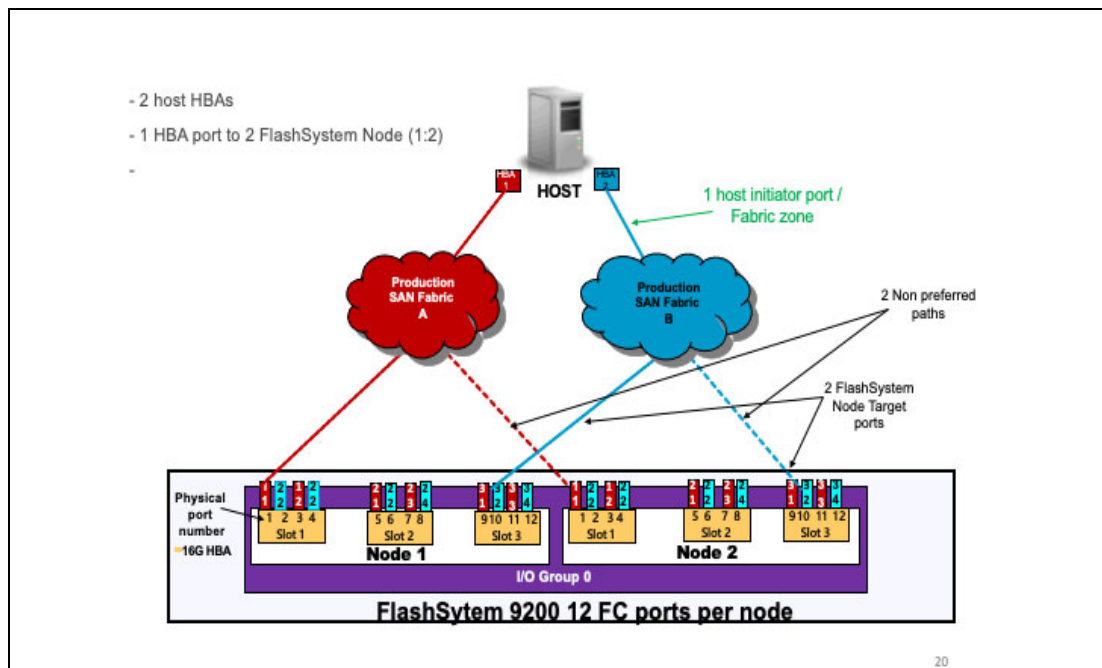


Figure 2-12 Typical host to FlashSystem 9200 zoning

This configuration provides four paths to each volume: two preferred paths (one per fabric) and two non-preferred paths. Multipathing software (such as AIXPCM, SDDDSM, VMWare NMP, and the FlashSystem 9200) is optimized to work such with four paths per volume.

NPIV consideration: The recommendations in this section also apply to NPIV-enabled configurations. For a list of the systems supported by the NPIV, see [V8.4.0.x Configuration Limits and Restrictions for IBM FlashSystem 9200](#).

When the recommended number of paths to a volume are exceeded, path failures sometimes are not recovered in the required amount of time. In some cases, too many paths to a volume can cause excessive I/O waits, resulting in application failures and, under certain circumstances, it can reduce performance.

Note: The option of having eight paths by volume is also supported. However, this design provides no performance benefit and, in some circumstances, can reduce performance. Also, it does not significantly improve reliability nor 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 to each FlashSystem 9200 port must be equal. Having the same amount of workload typically involves zoning approximately the same number of host FC ports to each FlashSystem 9200 FC port.

Hosts with four or more host bus adapters

If you have four HBAs in your host instead of two HBAs, more planning is required. Since eight paths is not an optimum number, configure your FlashSystem 9200 host definitions (and zoning) as though the single host is two separate hosts. During volume assignment, you alternate which volume was assigned to one of the “pseudo hosts.”

The reason for not assigning one HBA to each path is because the FlashSystem 9200 I/O group works as a cluster. When a volume is created, one node is assigned as preferred and the other node solely serves as a backup node for that specific volume. It 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 2-13 shows an example of a four port host zoning.

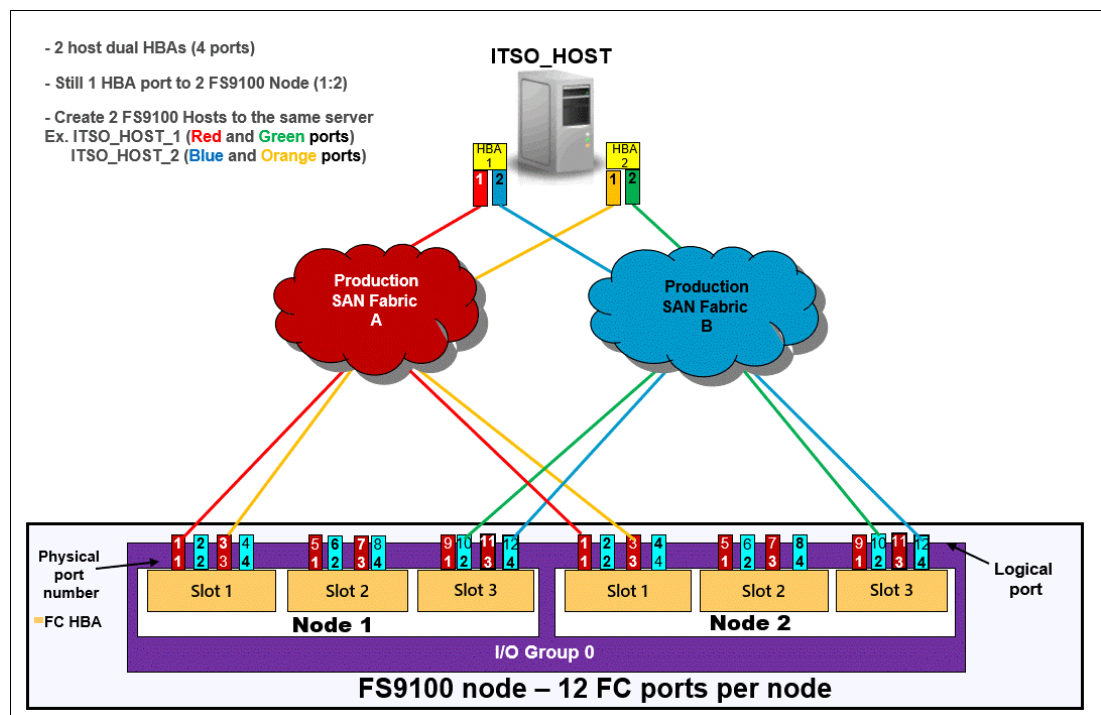


Figure 2-13 Four port host zoning

Because the optimal number of volume paths is four, you must create two or more hosts on FlashSystem 9200. During volume assignment, alternate which volume is assigned to each of the “pseudo-hosts,” in a round-robin fashion.

Note: Pseudo-hosts is not a defined function or feature of SAN Volume Controller/ Storwize. To create a pseudo-host, you simply need to add another host ID to the SAN Volume Controller and Storwize host configuration. Instead of creating one host ID with four WWPNs, you define two hosts with two WWPNs, therefore you need to pay extra attention to the scsi ids assigned to each of the pseudo-hosts to avoid having 2 different volumes from the same storage subsystem with the same scsi id.

ESX Cluster zoning

For ESX Clusters, you must create separate zones for each host node in the ESX Cluster as shown in Figure 2-14.

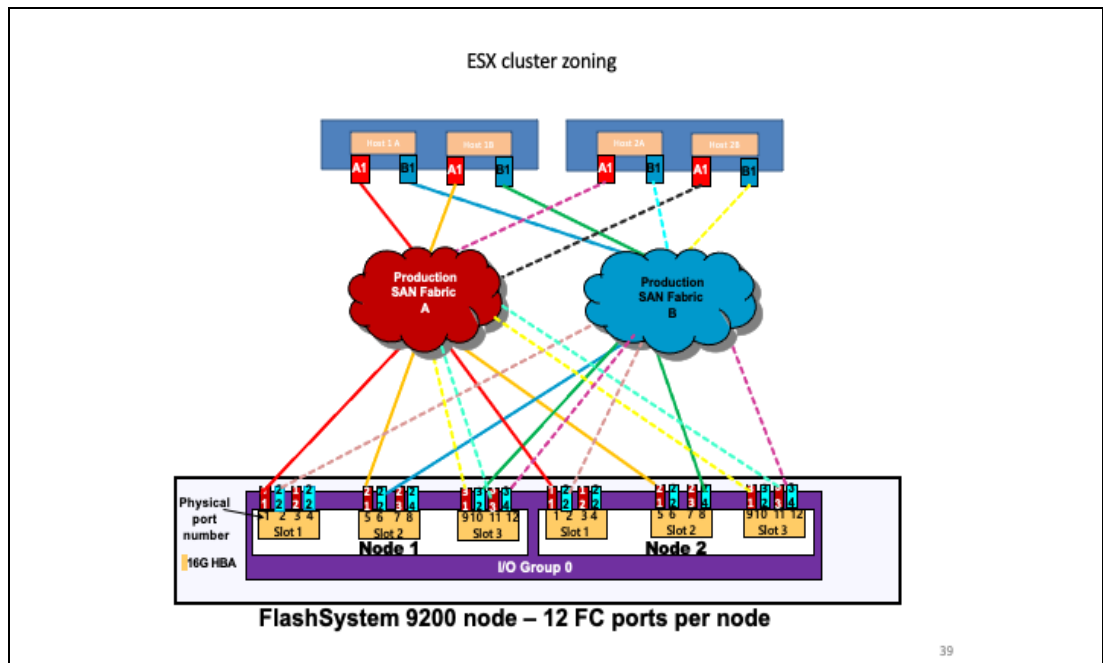


Figure 2-14 ESX Cluster zoning

Ensure that you apply the following preferred practices to your ESX VMware clustered hosts configuration:

- ▶ Zone a single ESX cluster in a manner that avoids ISL I/O traversing.
- ▶ Spread multiple host clusters evenly across the FlashSystem 9200 node ports and I/O Groups.
- ▶ Create one host entity for each host node in FlashSystem 9200 and group them in a *hostcluster* entity.
- ▶ Create separate zones for each host node in FlashSystem 9200 and on the ESX cluster.

When you allocate a LUN/volume to a clustered system, you should use a host cluster on FlashSystem 9200, this way you will have your hosts with the same SCSI ID for every volume, which will avoid outages due to SCSI mismatch.

AIX VIOs: LPM zoning

When zoning IBM AIX VIOs to IBM FlashSystem 9200, you must plan carefully. Because of its complexity, it is common to create more than four paths to each Volume or not provide for

proper redundancy. The following preferred practices can help you to have a non-degraded path error on IBM Spectrum Virtualize/Storage with four paths per volume:

- ▶ Create two separate and isolated zones on each fabric for each LPAR.
- ▶ Do not put both the active and inactive LPAR WWPNs in either the same zone or same IBM FlashSystem 9200 host definition.
- ▶ Map LUNs to the virtual host FC HBA port WWPNs, not the physical host FCA adapter WWPN.
- ▶ When using NPIV, generally make no more than a ratio of one physical adapter to eight Virtual ports. This configuration avoids I/O bandwidth oversubscription to the physical adapters.
- ▶ Create a pseudo host in IBM Spectrum Virtualize/Storage host definitions that contain only two virtual WWPNs, one from each fabric as shown in Figure 2-15.

Figure 2-15 shows a correct SAN connection and zoning for LPARs.

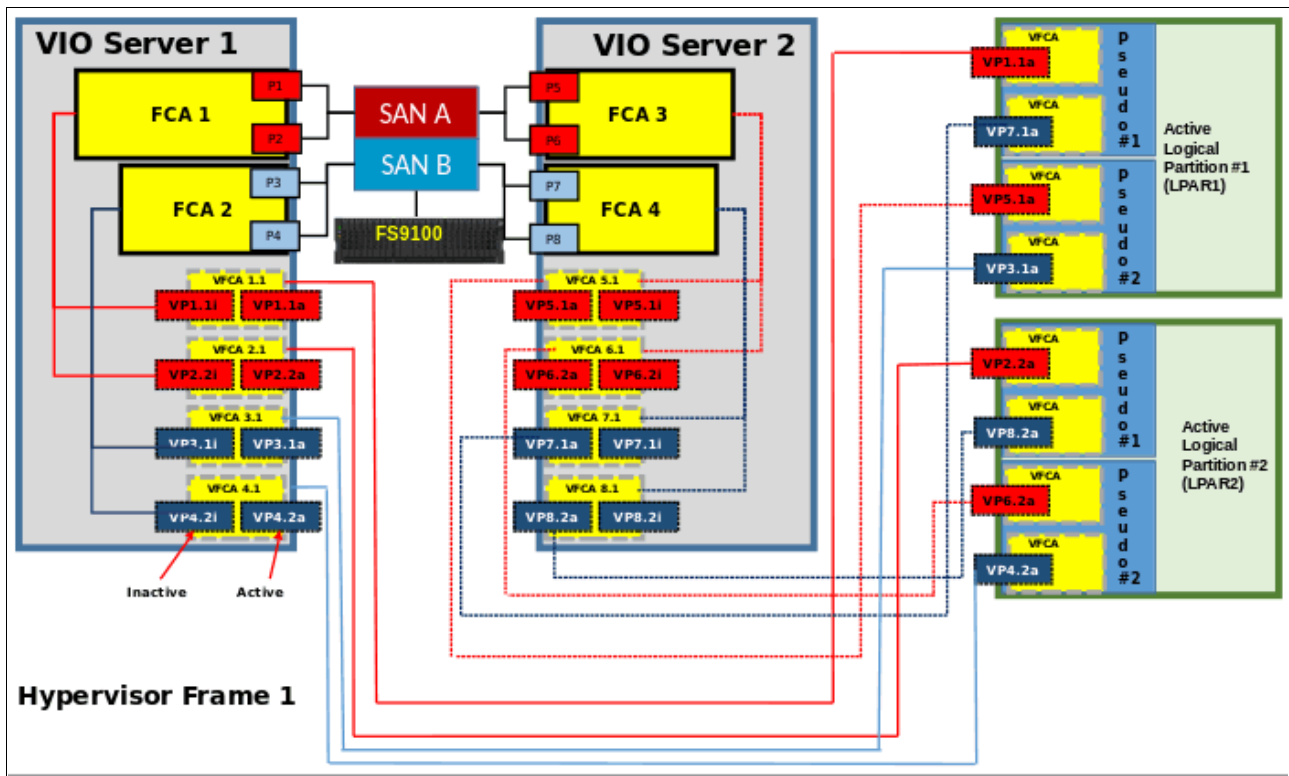


Figure 2-15 LPARs SAN connections

During Live Partition Migration (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 2-16 shows a Live partition migration from the hypervisor frame to another frame.

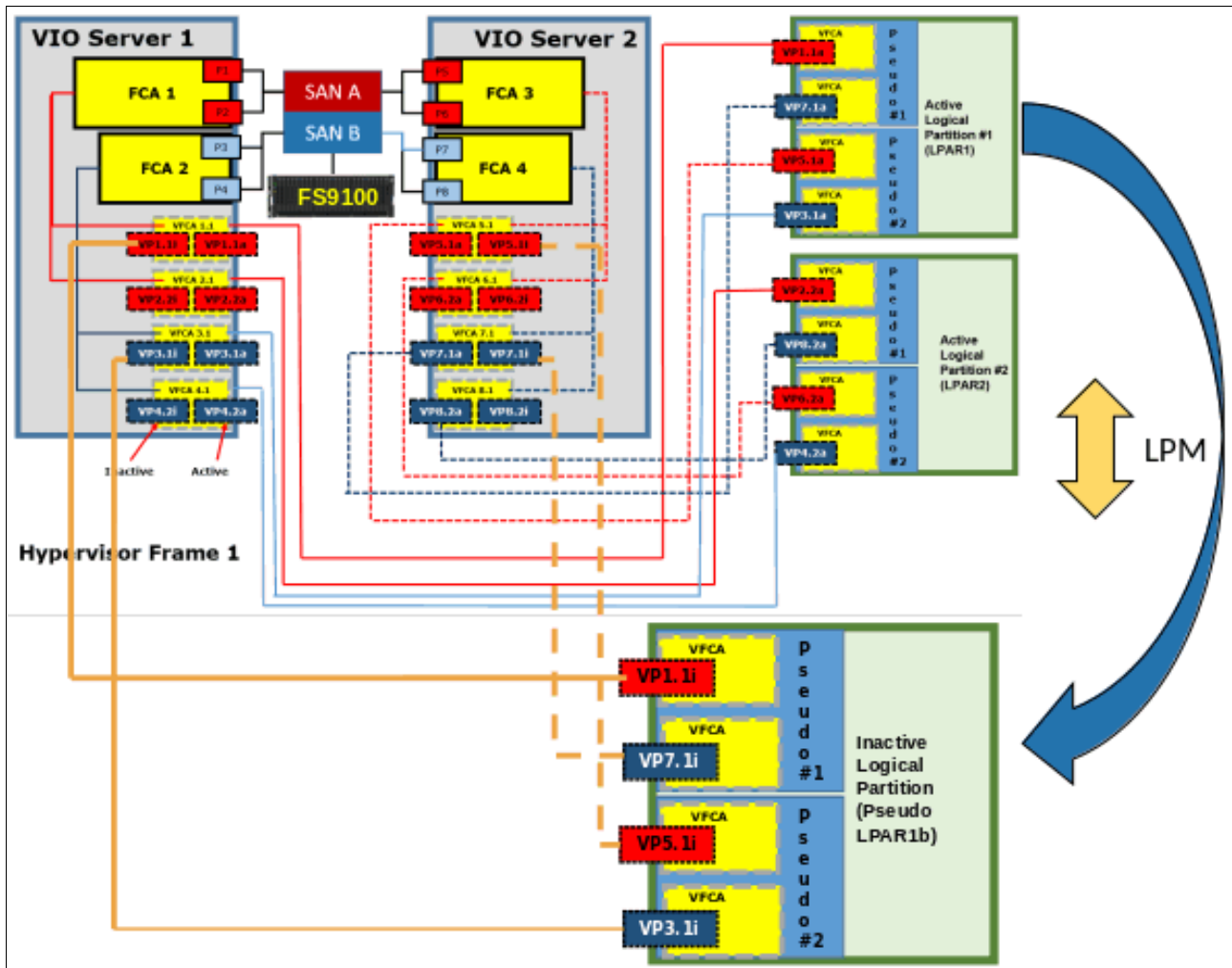


Figure 2-16 Live partition migration

Note: During LPM, the number of paths doubles from 4 to 8. Starting with eight paths per LUN/volume results in an unsupported 16 paths during LPM, which can lead to I/O interruption.

2.5 Distance extension for Remote Copy services

To implement Remote Copy services over distance, the following options are available:

- ▶ Optical multiplexors, such as Dense Wavelength Division Multiplexing (DWDM) or Coarse Wavelength Division Multiplexing (CWDM) devices
- ▶ Long-distance SFPs and XFPs
- ▶ FC-to-IP conversion boxes
- ▶ Native IP-based replication with Spectrum Virtualize code

Of these options, the optical varieties of distance extension are preferred. IP distance extension introduces more complexity, is less reliable, and has performance limitations. However, optical distance extension is impractical in many cases 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. The FlashSystem 9200 has no allegiance to a particular model of optical multiplexor.

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.

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*. Each switch vendor supports only a specific set of small form-factor pluggable (SFP) or Ten Gigabit Small Form Factor Pluggable (XFP) transceivers, so it is unlikely that Cisco SFPs will 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 scenarios with some kind of data replication between the primary and secondary site.

FCIP is a tunneling technology, which means FC frames are encapsulated in the TCP/IP packets. As such, it is not apparent to 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 either through stand-alone devices (Multiprotocol routers) or using blades integrated in the director class product. FlashSystem 9200 supports 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 e-mail, do not assume that this is the only type of traffic. The design of FC is sensitive to congestion. You do not want a spyware problem or a DDOS attack on an IP network to disrupt your FlashSystem 9200.

Also, when you are communicating 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 approximately 15 MBps to your FlashSystem 9200. 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 steps 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 that 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/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 Global Mirror with Change Volumes (GMCV), make sure that you have sized your TCP/IP link to sustain peak workloads.
- ▶ The use of FlashSystem 9200 internal Global Mirror (GM) simulation options can help you test your applications before production implementation. You can simulate the GM environment within one FlashSystem 9200 system without partnership with another. Use the `chsystem` command with the following parameters to perform GM testing:
 - `gminterdelaysimulation`
 - `gmintradelaysimulation`

For more information on GM planning, see Chapter 6, “Copy services” on page 229.

- ▶ 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 for your TCP/IP link, consider adhering to the following preferred practices:

- ▶ Set the link bandwidth and background copy rate of partnership between your replicating FlashSystem 9200 to a value *lower* than your TCP/IP link capacity. Failing to do this 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.
- ▶ 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 [IBM/Cisco Multiprotocol Routing: An Introduction and Implementation](#).

2.5.4 SAN extension with Business Continuity configurations

FlashSystem 9200 HyperSwap technology provides Business Continuity solutions over metropolitan areas with distances up to 300 km. Usually these solutions are achieved using SAN extension over WDM technology. Furthermore, in order 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 set up.

In this configuration, the intercluster communication is isolated in a Private SAN that interconnects Site A and Site B through a SAN extension infrastructure consisting of two DWDMs. Let's assume that, for redundancy reasons, two ISLs are used for each fabric for the Private SAN extension.

Two possible configurations to interconnect the Private SANs are as follows:

- ▶ In Configuration 1, shown in Figure 2-17, 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.

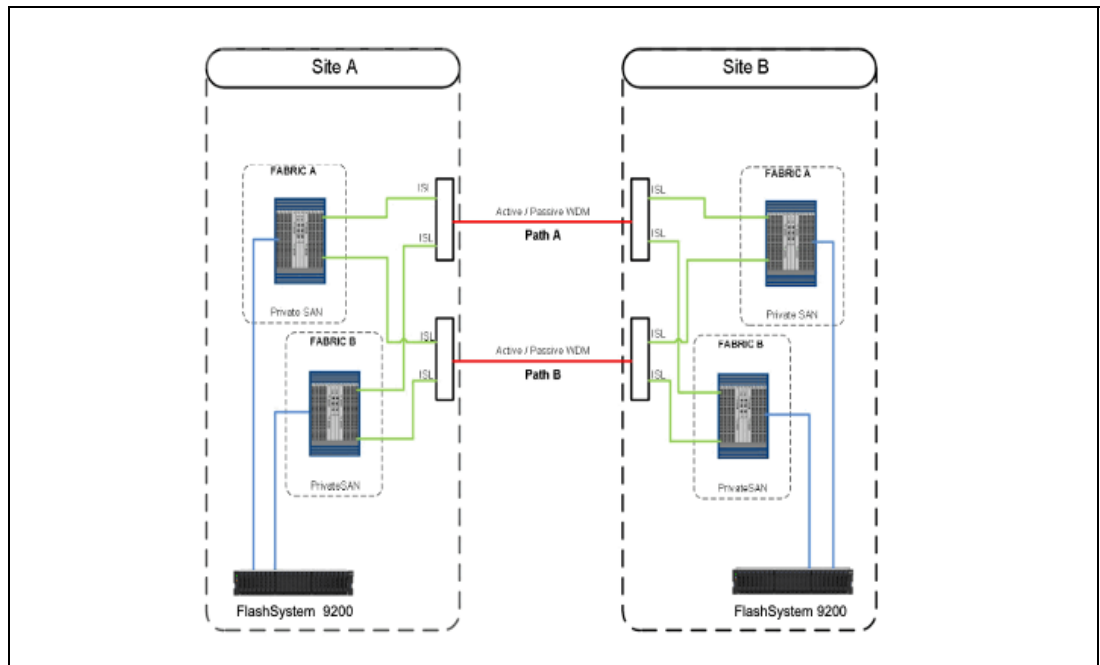


Figure 2-17 Configuration 1: Physical Paths Shared Among Fabrics

- ▶ In Configuration 2, shown in Figure 2-18 on page 59, 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.

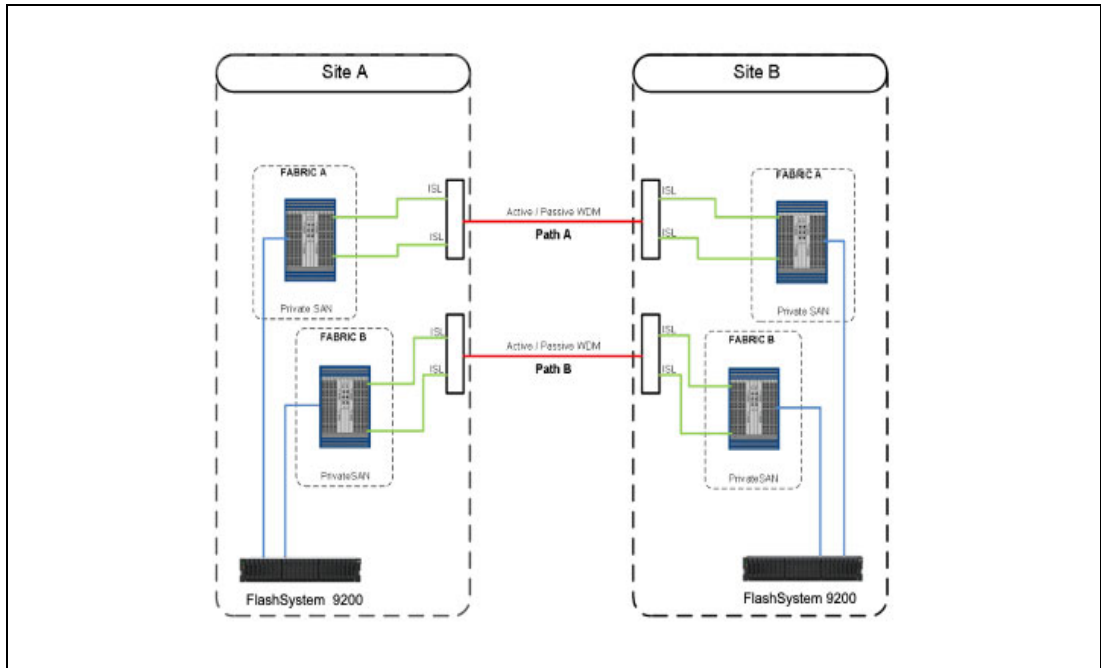


Figure 2-18 Configuration 2: physical paths not shared among the fabrics

With Configuration 1, in case of failure of one of the physical paths, both fabrics are simultaneously affected and a fabric reconfiguration occurs because of an ISL loss. This situation could lead to a temporary disruption of the intracluster communication and, in the worst case, to a split brain 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 intracluster communication would be guaranteed through the unaffected fabric.

To summarize, the recommendation is to fully understand the implication of a physical path or DWDM loss in the SAN extension infrastructure and implement the appropriate architecture in order to avoid a simultaneous impact.

2.5.5 Native IP replication

To enable native IP replication, FlashSystem 9200 implements the Bridgeworks SANSlide network optimization technology. For more information about this solution, see [IBM SAN Volume Controller and Storwize Family Native IP Replication](#).

It is possible to implement native IP-based replication on the FlashSystem 9200. *Native* means the FlashSystem 9200 does not need any FCIP routers to create a replication partnership. This partnership is based on the Internet Protocol network (IP) as opposed to the Fibre Channel (FC) network.

The main design point for the initial SANSlide implementation as well as subsequent enhancements including the addition of replication compression is to reduce link utilization in order to allow the links to run closer to their respective line speed at distance and over poor quality links. IP replication compression will 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 that 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.

It is important to understand that the effective bandwidth of an IP link is highly dependent on latency and the quality of the link in terms of the rate of packet loss. Even a small amount of packet loss and resulting retransmits will significantly degrade the bandwidth of the link.

Figure 2-19 shows the effects distance and packet loss have on the effective bandwidth of the links in MBps. Numbers reflect 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 2-19 Effect of distance on packet loss

Note 1: The maximum bandwidth for a typical IP replication configuration consisting of two 1 Gb links is approximately 244 MBps at zero latency and zero packet loss.

Note 2: When using two links replication will perform 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 will be 160 MBps.

Note 3: 10 Gb links should not be used with latencies beyond 10 ms. Beyond 10 ms a 1 Gb link begins to outperform a 10 Gb link.

Note 4: The FlashSystem 9200 supports volume compression. However, replication runs above volume compression in the IBM Spectrum Virtualize software stack, which means volumes are replicated at their full uncompressed capacity. This differs from some storage systems such as the IBM XIV® where replication runs below volume compression and therefore replicate the compressed capacity of the volumes. This difference needs to be taken into account when sizing workloads that are moving from one storage system technology to another.

2.6 Tape and disk traffic that share the SAN

If you have free ports on your core switch, you can place tape devices (and their associated backup servers) on the FlashSystem 9200 SAN. However, do not put tape and disk traffic on the same FC HBA.

Do not put tape ports and backup servers on different switches. Modern tape devices have high-bandwidth requirements. Placing tape ports and backup servers on different switches can quickly lead to SAN congestion over the ISL between the switches.

2.7 Switch interoperability

Note: For complete interoperability information see [IBM System Storage Interoperation Center \(SSIC\)](#).

FlashSystem 9200 is flexible as far as switch vendors are concerned. All of the node canister connections on a particular FlashSystem 9200 single or 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.

FlashSystem 9200 supports 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.

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 N-Port ID Virtualization (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 backend

This chapter describes the aspects and 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 Spectrum Virtualize by virtualizing a separate IBM or third-party storage system, attached with FC or iSCSI.

Chapter also provides information on traditional quorum disks. For information on IP Quorum, see Chapter 7, “Business continuity” on page 339.

This chapter contains the following sections:

- ▶ 3.1, “Internal storage types” on page 64
- ▶ 3.2, “Arrays” on page 72
- ▶ 3.3, “General external storage considerations” on page 78
- ▶ 3.4, “Controller-specific considerations” on page 83
- ▶ 3.5, “Quorum disks” on page 99

3.1 Internal storage types

A system supports the following three types of devices attached with Non-Volatile Memory express (NVMe) protocol:

- ▶ Storage-class memory drives
- ▶ Industry-standard NVMe flash drives
- ▶ IBM FlashCore Modules

With a serial attached SCSI (SAS) attachment, flash (solid state) and spinning disk drives are supported. The set of supported drives depends on the platform.

3.1.1 NVMe storage

FlashSystem 5100, FlashSystem 7200, FlashSystem 9100, and FlashSystem 9200 control-enclosures have 24 x 2.5" slots to populate with NVMe storage.

NVMe protocol

NVMe is an optimized, high-performance scalable host controller interface designed to address the needs of systems that utilize 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 extremely fast storage media, such as flash-based solid-state drives (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 additional I/O controller between the CPU and the storage device. This architecture results in lower latency, throughput scalability, and simpler system designs.

NVMe protocol supports multiple I/O queues, versus legacy SAS and Serial Advanced Technology Attachment (SATA) protocols, which use only a single queue.

NVMe as a protocol, is similar to SCSI. It allows for discovery, error recovery, and read and write operations. However, NVMe uses RDMA over new or existing physical transport layers such as PCIe, Fibre Channel, or Ethernet. The major advantage of an NVMe-drive attachment is that this is usually via PCIe connectivity, thus the drives are physically connected to the CPU itself via a high-bandwidth PCIe connection, rather than using a “middle man”, such as a SAS controller chip which will limit total bandwidth to that available to the PCIe connection into the SAS controller. Where a SAS controller might have used 8 or 16 PCIe lanes in total, each NVMe drive has its own dedicated pair of PCIe lanes. This means a single drive can achieve data rates in excess of multiple GiB/s rather than hundreds of MiB/s when compared with SAS.

Overall latency can be improved by the adoption of larger parallelism and the modern device drivers used to control NVMe interfaces. For example, NVMe over Fibre Channel versus SCSI over Fibre Channel are both bound by the same Fibre Channel network speeds and bandwidths. However, the overhead on legacy SCSI device drivers (for example, reliance on kernel-based interrupt drivers) means that the software functionality in the device driver might limit its capability when compared with an NVMe driver. This is 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 it 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. Interrupt-driven kernel drivers therefore waste time in switching between kernel and user space. As a result, all useful work is prevented from occurring on the CPU while the interrupt is handled. This adds latency and reduces total throughput, and therefore the amount of work 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 Spectrum Virtualize Fibre Channel and SAS drivers have always been 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 above bandwidth increases are seen when a SAS controller is switched to a to PCIe-attached drive.

The majority of the advantages of using an end-to-end NVMe solution, when attaching to a Spectrum Virtualize based system, are seen as a reduction in the CPU cycles that are needed to handle the interrupts in the host server where the Fibre Channel HBA resides. Most SCSI device drivers remain interrupt driven, therefore switching to NVMe over Fibre Channel will result in the same latency reduction. CPU cycle reduction and general parallelism improvements have been enjoyed inside Spectrum Virtualize products since 2003.

Industry-standard NVMe drives

FlashSystem 5100, FlashSystem 7200, FlashSystem 9100, and FlashSystem 9200 control enclosures provide an option to use self-encrypting industry-standard (IS) NVMe flash drives, which are available with capacity from 800 GB to 15.36 TB.

Supported IS NVMe SSD drives are built in 2.5-inch form factor (SFF) and use a dual-port PCIe Gen3 interface to connect to the midplane.

NVMe FlashCore modules

At the heart of the IBM FlashSystem system is IBM FlashCore technology. IBM FlashCore Modules (FCMs) is a family of high-performance flash drives, that provide performance-neutral, hardware-based data compression and self-encryption.

FlashCore modules introduce the following features:

- ▶ Hardware-accelerated architecture that is engineered for flash, with a hardware-only data path
- ▶ Modified dynamic GZIP algorithm for data compression and decompression, implemented completely in drive hardware
- ▶ Dynamic SLC cache for reduced latency
- ▶ Cognitive algorithms for wear leveling and heat segregation

Variable stripe redundant array of independent disks (RAID) (VSR) stripes data across more granular, sub-chip levels. This allows for failing areas of a chip to be identified and isolated without failing the entire chip. Asymmetric wear-leveling understands 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.

Bit errors caused by electrical interference are continually scanned for, and if any are found will be corrected by an enhanced Error Correcting Code (ECC) algorithm. If an error cannot be corrected, then the FlashSystem DRAID layer will be used to rebuild the data.

NVMe FlashCore Modules use inline hardware compression to reduce the amount of physical space required. Compression cannot be disabled (and there is no reason to do that). If the written data cannot be compressed further, or compressing the data causes it to grow in size, the uncompressed data will be written. In either case, because the FCM compression is done in the hardware there will be no performance impact.

FlashSystem FlashCore Modules are not interchangeable with the flash modules that are used in FlashSystem 900 storage enclosures, as they have a different form factor and interface.

Modules that are used in FlashSystem 5100, 7200, 9100, and 9200 are built in 2.5-inch U.2 dual-port form factor.

FCMs are available in their physical capacity (4.8, 9.6, 19.2 and 38.4 TB sizes) or their usable capacity. The *usable capacity* is a factor of how many bytes the flash chips can hold.

They also have a maximum effective capacity (or virtual capacity), beyond which they cannot be filled. *Effective capacity* is the total amount of user data that could be stored on a module, assuming the compression ratio 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.

Module capacities are shown in Table 3-1.

Table 3-1 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	43.98 TB
38.4 TB	2.3: 1	87.96 TB

4.8 TB FCM has a higher compression ratio as it has the same amount of metadata space as the 9.6 TB.

Usable and effective capacities as discussed later in this chapter, see 3.1.3, “Internal storage considerations” on page 68.

At the moment of writing, IBM offers the second generation of IBM FlashCore modules, FCM2.0, which provides better performance and lower latency than FlashCore Module gen1. FCMs can be intermixed between generations within one system and within a single array. If needed, FCM1.0 can be replaced with an FCM2.0 of the same capacity.

Note: An array with intermixed FCM1.0 and FCM2.0 drives will perform like an FCM1.0 array.

Storage-class memory drives

Storage Class Memory (SCM) is a term that is used to describe non-volatile memory devices that perform faster (~10µs) than traditional NAND SSDs (100µs), but slower than DRAM (100ns).

IBM FlashSystem supports SCM drives that are built on two different technologies:

- ▶ 3D XPoint technology from Intel, developed by Intel and Micron (Intel Optane drives)
- ▶ zNAND technology from Samsung (Samsung zSSD)

Available SCM drive capacities are shown in Table 3-2 on page 67.

Table 3-2 Supported SCM drive capacities

Technology	Small capacity	Large capacity
3D XPoint	350 GB	750 GB
zNAND	800 GB	1800 GM

SCM drives have their own technology type and drive class in Spectrum 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 Storage Class Memory drives can be installed per control enclosure.

Note: If you want a limit of 12 SCMs, you must have Spectrum Virtualize code version 8.4.0 or above. On previous code versions, only four SCM drives were allowed and only in the last four slots of IBM FlashSystem enclosure.

3.1.2 SAS drives

IBM FlashSystem 5100, 7200, 9100, and 9200 control enclosures have NVMe drive slots. They can be scaled up by attaching SAS expansion enclosures with SAS drives.

IBM FlashSystem 5010 and 5030 control enclosures have twelve 3.5-inch LFF or twenty-four 2.5-inch SFF SAS drive slots, and also can be scaled up by connecting SAS expansion enclosures.

A single FlashSystem 5100, 7200, 9100, and 9200 control enclosure supports attachment of up to 20 expansion enclosures with a maximum of 760 drives (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 FlashSystem 5100 and a maximum of 3040 drives for FlashSystem 7200 and 9x00.

FlashSystem 5030 control enclosure supports up to 20 expansion enclosures with a maximum of 504 drives (including drives in the control enclosure). With two-way clustering, available for FlashSystem 5030, it allows up to 1008 drives per system.

FlashSystem 5010 control enclosure supports up to 10 expansions and 392 drives maximum.

Expansion enclosures are designed to be dynamically added without downtime, helping to quickly and seamlessly respond to growing capacity demands.

Three types of SAS-attached expansion enclosures are available for 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 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 can be intermixed with each other.

IBM FlashSystem 5030, 5100, 7200, and 9x00 control enclosures have two SAS chains for attaching expansion enclosures. Aim to 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 5010 has only a single SAS chain.

The number of drive slots per SAS chain is limited to 368. To achieve this, you need four 5U high-density enclosures. Table 3-3 shows maximum number of drives that are allowed when different enclosures are attached and intermixed. For example, if there are three 5U enclosures attached to a chain, you cannot connect more than two 2U enclosures to the same chain, and get 324 drive slots as the result.

Table 3-3 Maximum number of drive slots per SAS expansion chain

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	--	--	--	--	--	--	--	--	--	--

IBM FlashSystem 5010 and 5030 node canisters have onboard SAS ports for expansions. IBM FlashSystem 5100, 7200, and 9x00 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 spinning drives (high-performance enterprise-class disk drives or high-capacity nearline disk drives) or with solid state (flash) drives.

A set of allowed drive types depends on the system:

- ▶ 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.

Drive capacities vary from less than 1 TB to more than 30 TB.

3.1.3 Internal storage considerations

The following practices should be considered when planning and managing IBM FlashSystem internal storage.

- ▶ SAS enclosures are used to scale capacity within the performance envelope of a single controller enclosure. Clustering (up to four control enclosures) scales performance with the additional NVMe storage.
- ▶ Drives of the same form factor and connector type can be intermixed within an expansion enclosure.
- ▶ NVMe and SAS drives in the same system can be mixed, but NVMe drives can only exist in the control enclosure, and SAS drives can only exist in SAS expansion enclosures.
- ▶ Industry-standard NVMe drives start at a smaller capacity point, allowing for a smaller system.

- ▶ Within a control enclosure, NVMe drives of different capacities can be intermixed, also industry-standard NVMe drives and SCMs can be intermixed with FlashCore modules. However, within a DRAID array NVMe drives must all be the same size and should use the same type of drive. It is not allowed to mix IS NVMe drives and SCMs or FCMs in a single array.
- ▶ FlashCore modules need to be formatted before they can be used. Format is important because when array is created, its members must have zero used capacity. Drives will automatically format when being changed to candidate. While they are formatting, they will appear as offline candidates. If you attempt to create an array before format is complete, the create command is delayed until all formatting is done. Once this happens, the command will complete.

If a drive fails to format, it will go offline. In this case, 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_FlashSystem:FS9100-ITS0:superuser>lsdrive
id status error_sequence_number use tech_type ....
....
13 offline 118 candidate tier0_flash ....
IBM_FlashSystem:FS9100-ITS0:superuser>chdrive -task format 13
IBM_FlashSystem:FS9100-ITS0:superuser>

```

An FCM is expected to format in under 70 seconds.

- ▶ When a drive is replaced in the same slot, the system tries to take the drive as a replacement for the array member. The drive will have to format first, so this might take some time for an FCM. If the drive was previously encrypted in another array, it will come up as failed since this system won't have the required keys. The drive must be manually formatted to make it a candidate.
- ▶ Formatting an SCM drive takes much longer than an FCM or IS NVMe drive. On Intel Optane, drive formatting can take 15 minutes.
- ▶ All SCM, FCM, and IS NVMe drives that are used in the system are self-encrypting. For SAS drives, encryption is performed, by SAS chip in control enclosure.
- ▶ For IS 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.

- ▶ The IBM FlashSystem GUI (as shown in Figure 3-1 on page 70) and CLI (as shown in Example 3-2 on page 70) allows you to monitor effective and physical capacity for each FCM.

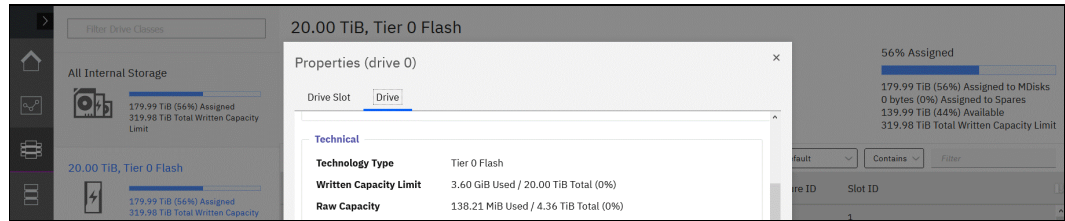


Figure 3-1 FCM capacity monitoring with GUI

Example 3-2 FCM capacity monitoring with CLI

```
IBM_FlashSystem:FS9100-ITS0:superuser>lsdrive 0
id 0
...
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 same 4.8 TB FCM with maximum effective capacity of 20 TiB (or 21.99 TB).

To calculate actual compression ratio, divide effective used capacity by 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.
- ▶ If FCM drive is an active member of a RAID array (drive *use* property shows *member*), it must not be reseated unless directly advised by IBM Support. Reseating drives that are still in use by an array can cause undesired consequences.
- ▶ Plan and 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 be able 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 causing 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).

Note: For more information on physical flash provisioning, see [Do not provision 100% of the physical flash](#).

- ▶ When using FCMs, data compression ratios should be thoroughly planned and monitored. If highly compressible data is written to an FCM, it will still become full when it reaches the maximum effective capacity. Any spare data space remaining at this point will be used to improve the performance of the module and extend the wear.

Example #1: 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 will be 100% full for the system. At the same time, data will use only 2 TiB of the physical space. 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 will fill up until the maximum physical capacity is reached.

Example #2: 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 will be used. After 1.2 compression, it will occupy 16.7 TiB of physical capacity, making the module 95% full, and potentially impacting the module's performance.

Pool-level and array-level warnings can be set to alert and prevent compressed drive overfill.

- ▶ Drive Writes Per Day (DWPD) is a term that is used to express the number of times that the total capacity of a drive may 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 will alert that the drive is wearing faster than expected. As DWPD is taken into account during system sizing, it usually means that workload differs from what was expected on the given array and it needs to be revised.

DWPD numbers are important with SSD drives of smaller sizes. With drive capacities below 1 terabyte (TB), it is possible to write the total capacity of a drive several times a day. When a single SSD provides TBs, it is unlikely that you will be capable of overrunning the DWPD measurement. The DWPD measurement is therefore less relevant for FCMs and large SSDs.

- ▶ 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 will last the five-year warranty.

Example: 3.84 TB RI SAS drive is rated for 1 DWPD, which means 3840000 MB of data may be written on it every 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 $3.84 \times 1 \text{ DWPD} \times 365 \times 5 = 6.8$ PB.

- ▶ FCM2.0 drives are rated with two DWPD over five years, measured in usable capacity. This means that if data is compressible, for example, 2:1, then the DWPD doubles.

Example: 19.2 TB FCM is rated for 2 DWPD. Its effective capacity is nearly 44 TB = 40 TiB, so considering 2.3:1 compression, to reach DWPD limit average daily workload over 5 years must be around 1 GB/s. Total cumulative writes over a 5-year period are more than 140 PB.

- ▶ SCM drives are rated with 30 DWPD over five years.
- ▶ It is acceptable to see "write endurance usage rate is high" warnings, which indicate that write data rate exceeds expected for the given drive type, during the initial phase of system implementation or during continuous testing. Afterwards, the system's workload will reduce to what is sized for it, and the system recalculates the usage rate and removes the warnings. Because the calculation is based on a long-run average, it can take time (up to one month) for them to be automatically cleared.

3.2 Arrays

In order to use internal IBM FlashSystem drives in storage pools and provision their capacity to hosts, they need to be joined into RAID arrays to form array-type MDisks.

3.2.1 Supported RAID types

RAID provides two key design goals:

- ▶ Increased data reliability
- ▶ Increased input/output (I/O) performance

The IBM FlashSystem system supports two RAID types:

- ▶ **Traditional RAID (TRAIID):** In a traditional RAID approach, data is spread amongst 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 written to a single spare drive.
- ▶ **Distributed RAID (DRAID):** With distributed RAID (DRAID), spare capacity is used instead of the idle spare drives from a traditional RAID. The spare capacity is spread across the disk drives. Because no drives are idling, all drives contribute to array performance. In the case of a drive failure, the rebuild load is distributed across multiple drives. By doing this, DRAID addresses two main disadvantages of a traditional RAID approach: it reduces rebuild times by eliminating the bottleneck of one drive and increases array performance by increasing the number of drives sharing the workload.

IBM FlashSystem implementation of DRAID allows effectively spread workload across multiple node canister CPU cores, which provides significant performance improvement over single-threaded traditional RAID arrays.

NVMe FlashCore Modules that are installed in IBM FlashSystem can be aggregated into DRAID 6, DRAID 5 or DRAID 1. Traditional RAID levels 5 and 6, as well as RAID 0, 1, and 10, are not supported on FCMs.

SCM drives support DRAID levels 6 and 5, as well as DRAID 1, and TRAIID 0.

IS NVMe drives and SAS drives in expansion enclosures can be aggregated into DRAID 6 and DRAID 5 arrays, and also can form RAID 1 and RAID 10 arrays. TRAIID 5 and 6 are not supported.

Note: Distributed RAID 1 is supported on only IBM FlashSystem 7200 and IBM FlashSystem 9200.

Table 3-4 summarizes the supported drives, array types, and RAID levels.

Table 3-4 Supported RAID levels

Supported drives	Non-distributed arrays (traditional RAID)					Distributed arrays (DRAID)		
	RAID 0	RAID 1	RAID 5	RAID 6	RAID 10	RAID 1 *	RAID 5	RAID 6
SAS flash drives	Yes	Yes	No	No	Yes	Yes	Yes	Yes

Supported drives	Non-distributed arrays (traditional RAID)					Distributed arrays (DRAID)		
NVMe drives	Yes	Yes	No	No	Yes	Yes	Yes	Yes
FlashCore Modules	No	No	No	No	No	Yes **	Yes	Yes
SCM drives	No	Yes	No	No	Yes	Yes	Yes	Yes

* Where supported by IBM FlashSystem platform

** XL (38.4 TB) FCMs are not supported by DRAID 1

3.2.2 Array considerations

The following practices should be considered when planning and managing drive arrays in IBM FlashSystem environment.

RAID level

Consider the following points when determining which RAID level to use:

- ▶ DRAID 6 is strongly recommended for all arrays with more than 6 drives.
Traditional RAID levels 5 and 6 are not supported on the current generation of IBM FlashSystem, as DRAID is superior to them in all aspects.
For most use cases, DRAID5 has no performance advantage compared to DRAID6. At the same time, DRAID6 offers protection from the second drive failure, which is vital as rebuild times are increasing together with the drive size. As DRAID6 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 recommended to use DRAID 1 or DRAID 5, depending on capacity and performance requirements.
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 that of 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 traditional RAID 1 or RAID 10, as they will not perform as well as distributed RAID type.
- ▶ On platforms that do not support DRAID 1, for NVMe SCM drives recommended RAID level is TRAIID10 for arrays of 2 drives, and DRAID 5 for arrays of four or five drives.
- ▶ RAID configurations that differ from the recommendations that are listed above are not available with the system GUI. If they are still required, arrays with desired supported RAID levels may be created with the system CLI.

RAID geometry

Consider the following points when determining your RAID geometry:

- ▶ Data, parity, and spare space need to be striped across the number of devices available. The higher the number of devices, the lower the percentage of overall capacity the spare

and parity devices will consume, and the more bandwidth that will be available during rebuild operations.

Fewer devices are acceptable for smaller capacity systems that don't have a high-performance requirement, but solutions with a small number of 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 comparing 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 traditional RAID, for example, to create four or eight arrays to spread load across multiple CPU threads, do not apply to DRAID.
- ▶ For IS NVMe drives and FCMs, optimal number of drives in an array is 16 to 24. This ensures a balance between performance, rebuild times and usable capacity. Array of NVMe drives cannot have more than 24 members.
 - For SCM, maximum number of drives in an array is 12.
 - For SAS SSD drives, optimal array size is 24-36 drives per DRAID 6 array.
 - For SAS HDDs, the typical best benefit for rebuild times is around 48 to 64 HDD drives in a single DRAID6.
- ▶ Distributed spare capacity, or rebuild areas, are configured with the following guidelines:
 - DRAID 1 with two members: it is the only DRAID type, which is allowed without spare capacity (zero rebuild areas);
 - DRAID 1 with 3-16 members: array must have one rebuild area, it cannot have zero and cannot have more;
 - DRAID 5 or 6: minimum recommendation is one rebuild area per every 36 drives, optimal is one rebuild area per 24 drives.
 - Arrays with FCM drives cannot have more than one rebuild area per array.
- ▶ DRAID stripe width is set during array creation and indicates the width of a single unit of redundancy within a distributed set of drives. Note that reducing the stripe width will not enable the array to tolerate more failed drives. DRAID 6 will not get more redundancy than determined for level 6, independently of the width of a single redundancy unit.

Reduced width increases capacity overhead, but also increases rebuild speed, as there is a smaller amount of data that RAID needs to read in order to reconstruct the missing data. For example, rebuild on DRAID with 14+P+Q geometry (width = 16) would be slower, or have a higher write penalty, than rebuild on DRAID with the same number of drives but 3+P+Q geometry (width = 5). In return, usable capacity for an array with width = 5 will be smaller than for an array with width = 16.

Default stripe width settings (12 for DRAID6) provide an optimal balance between those parameters.

- ▶ Array strip size must be 256 KiB. With Spectrum Virtualize code releases before 8.4.x, it was possible to choose between 128 KiB and 256 KiB if DRAID member drive size was below 4 TB. From 8.4.x and above, it is recommended that you use only strip size 258 KiB for all arrays.

Arrays that were created on previous code levels, with strip size 128 KiB, are still fully supported.

- ▶ Stripe width together with strip size determine Full Stride Write (FSW) size. With fFSW, there is no need to read existing data in a stride, so the RAID I/O penalty is massively reduced. For better performance, it is a good idea to set the host file system block size to

the same value as either the FSW size or a multiple of the FSW stripe size. However, IBM FlashSystem cache is designed to perform FSW whenever possible, so in most scenarios there isn't a noticeable difference in the performance of the host.

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 8+P+Q DRAID6 array (8 data disks x 256 KiB stripe size, 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 a “superior” NVMe drive (that is, one with a greater capacity),
 - For SAS drives: Members of an array do not require the same physical and logical capacity. You can replace a SAS drive with “superior” SAS drive.
- ▶ 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 a SAS drive from different control enclosures in a system into one array. One DRAID6 can span across multiple expansion enclosures.

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 re-striped to maintain the algorithm placement of strips 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 it was set during array creation.
- ▶ RAID-member count cannot be decreased: it is not possible to shrink an array.
- ▶ DRAID 5, DRAID 6 and DRAID 1 can be expanded. Traditional RAID arrays do not support expansion.
- ▶ Only one expansion process can run on array at one 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, as 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 designed to run in background and can take significant time. It can affect host performance and latency, especially when expanding an array of spinning drives. Do not expand an array when the array has over 50% load. If you do not reduce host I/O load, the amount of time that is needed to complete the expansion increases greatly.

RAID capacity

Consider the following points when determining RAID capacity:

- ▶ To find out capacity of a potential array for a specified drive count, drive class, and RAID level in the specified pool, use the **lspotentialarraysize** CLI command.
- ▶ To get an approximate amount of available space in DRAID6 array, use the following formula:

$$\text{Array Capacity} = D / ((W * 256) + 16) * ((N - S) * (W - 2) * 256)$$

Where:

D - Drive capacity

N - Drive count

S - Rebuild areas (spare count)

W - Stripe width

Example #1:

Capacity of DRAID6 array out of 16 x 9.6 TB FlashCore modules.

D = 9.6 TB = 8.7 TiB

N = 16

S = 1

W = 12

$$\begin{aligned} \text{Array capacity} &= 8.7 \text{ TiB} / ((12*256)+16) * ((16-1) * (12-2) * 256) = \\ &= 8.7 \text{ TiB} / 3088 * 38400 = 108.2 \text{ TiB} \end{aligned}$$

3.2.3 Compressed array monitoring

DRAID arrays on FlashCore modules need to be carefully monitored and well-planned, as they are over-provisioned, which means they are susceptible to an out-of-space condition.

To minimize the risk of an out of space condition, ensure the following:

- ▶ The data compression ratio is known and taken into account when planning for array physical and effective capacity.
- ▶ Monitor array free space and avoid filling it up 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, “Monitoring” on page 363.

IBM FlashSystem GUI and CLI will also display used and available effective and physical capacities. For examples, see Figure 3-2 on page 77 and Example 3-3 on page 77.

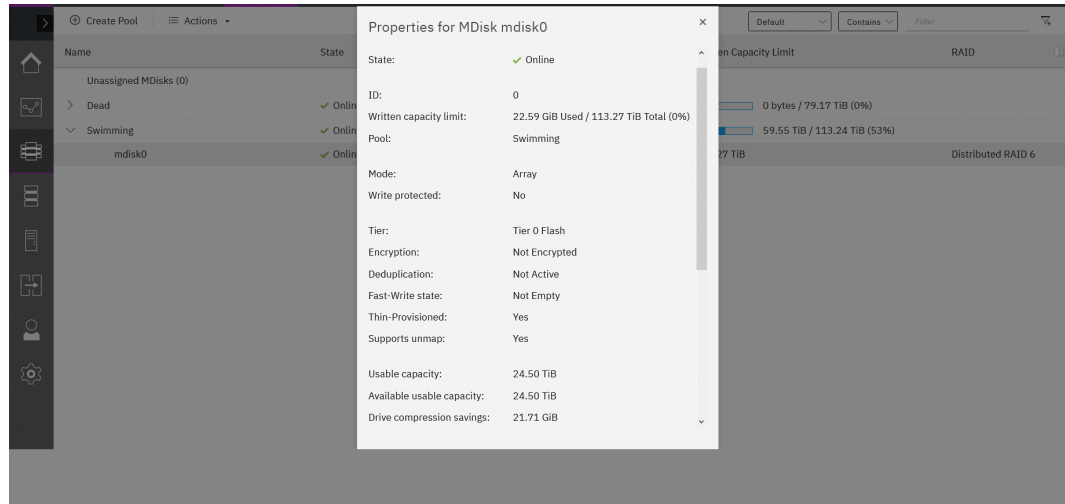


Figure 3-2 Array capacity monitoring with GUI

Example 3-3 Array capacity monitoring with CLI

```

IBM_FlashSystem:FS9100-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 FlashSystem raises event ID 1241 “1% physical space left for compressed array”. This is a call for immediate action.

To prevent running out of space, one or a combination of corrective actions should be taken:

- Add more storage to the pool and wait while data is balanced between arrays by Easy Tier.
- Migrate volumes with extents on the managed disk 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 managed disks that have sufficient extents.
- Delete or migrate data from the volumes using a host that supports UNMAP commands. IBM FlashSystem will issue UNMAP to the array and space will be released.

For more information on 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. This means that if array capacity is 85% full, each array FCM is used for less than that due to spare space reserve. When DRAID is rebuilding this space becomes used.

After the rebuild is complete, the extra space is filled up and the drives can 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 FlashSystem can virtualize external storage and make it available to the system. External back-end storage systems (or controllers in Spectrum Virtualize terminology) provide their logical volumes (LUs), 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 virtualized by IBM FlashSystem.

External back-end storage can be connected to IBM FlashSystem with FC (SCSI) or iSCSI. NVMe-FC back-end attachment is not supported, as it provides no performance benefits for IBM FlashSystem. For more information, see “NVMe protocol” on page 64.

3.3.1 Storage controller path selection

When a managed disk (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 Fibre Channel 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 FlashSystem node canisters select a set of ports to access the storage system, the two types of path selection 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 find out which algorithm is used for a particular back-end system, see [System Storage Interoperaton Center \(SSIC\)](#), as shown in Figure 3-3 on page 79.



Figure 3-3 SSIC example

Round-robin path algorithm

With the round-robin path algorithm, each MDisk uses one path per target port per IBM FlashSystem node. This means that in cases of storage systems without a preferred controller such as XIV or DS8000, each MDisk uses all of 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 therefore MDisk, uses all available IBM FlashSystem 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.

Preferred 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 FlashSystem supports up to 16 FC ports per storage controller. See your storage system documentation for FC port connection and zoning guidelines.

Example 3-4 shows a storage controller that supports round-robin path selection.

Example 3-4 Round robin enabled storage controller

```
IBM_FlashSystem:FS9100-ITS0:superuser>lsmdisk 4
id 4
name mdisk4
...
preferred_WPN
active_WPN 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 such as IBM DS5000 and DS3000 (or other Active-Passive type systems), IBM FlashSystem accesses an MDisk LU through one of the ports on the preferred controller. In order to best utilize the back-end storage, it is important to make sure that the number of LUs 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 MDisk group balanced path selection.

Example 3-5 MDisk group balanced path selection (no round robin enabled) storage controller

```
IBM_FlashSystem:FS9100-ITS0:superuser>lsmdisk 5
id 5
name mdisk5
...
preferred_WWPN
active_WWPN 20110002AC00C202 · <<< indicates Mdisk group balancing
```

3.3.2 Guidelines for creating optimal backend configuration

Most of the backend controllers aggregate spinning or solid state drives into RAID arrays, then join arrays into pools. Logical volumes are created on those pools and provided to hosts. When connected to external backend storage, IBM FlashSystem acts as a host. It is important to create backend controller configuration that provides performance and resiliency, as IBM FlashSystem will rely on back-end storage when serving I/O to attached host systems.

If your backend system has homogeneous storage, create the required number of RAID arrays (usually RAID 6 or RAID 10 are recommended) with equal number of drives. The type and geometry of an array depends on the backend controller vendor's recommendations. If your backend controller can spread the load stripe across multiple arrays in a resource pool (for example, by striping), create a single pool and add all arrays there.

On backend systems with mixed drives, create a separate resource pool for each drive technology. Keep the drive-technology type in mind, as you will need to assign the correct tier for an MDisk when it is used by IBM FlashSystem.

Create a set of fully allocated logical volumes from the back-end system storage pool (or pools). Each volume is detected as MDisk on IBM FlashSystem. The number of logical volumes to create depends the type of drives, used by your back-end controller.

Back-end controller with spinning drives

If your back-end is using spinning drives, 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, spinning drives need 8-10 concurrent I/O at the device, and this doesn't change with drive rotation speed. Make sure in a highly loaded system, that any given IBM FlashSystem MDisk can queue up approximately 8 I/O per back-end system drive.

IBM FlashSystem queue depth per MDisk is approximately 60. The exact maximum seen on a real system might vary depending on the circumstances. However, for the purpose of 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 8 I/O per drive and with queue depth 60 per MDisk from IBM FlashSystem, a back-end array with $60/8 = 7.5$ that is approximately equal to 8 physical drives is optimal, or we need one logical volume per every 8 drives in an array.

Example #1:

Backend controller to be virtualized is IBM Storwize V5030 with 64 NL-SAS 8 TB drives.

System is homogeneous. According to recommendations given in 3.2.2, “Array considerations” on page 73, create a single DRAID6 array at Storwize and put in a storage pool. Use the HDD rule of 8, tells us we want $64/8 = 8$ MDisks, so create 8 volumes from a pool to present to IBM FlashSystem and assign then to nearline tier.

All-flash backend controllers

For All-flash controllers, the considerations are more of I/O distribution across IBM FlashSystem ports and processing threads, than of queue depth per drive. Since most All-flash arrays that are put behind virtualizer have very high I/O capabilities, make sure that IBM FlashSystem is given the optimal chance to spread the load and evenly make use of its internal resources, so queue depths are of less a concern here (because of the lower latency per I/O).

For all-flash backend arrays, IBM recommends creating 32 logical volumes from the array capacity, as it allows to keep the queue depths high enough and spreads the work across the virtualizer resources. For smaller setups with a low number of solid state drives this number can be reduced to 16 logical volumes (which results in 16 MDisks) or even 8 volumes.

Example #2:

Backend controllers to be virtualized are IBM FlashSystem 5030 with 24 Tier1 7.6 TB drives IBM FlashSystem 900. Virtualizer needs a pool with two storage tiers. On IBM FlashSystem 5030, create a single DRAID6 array and add it to a storage pool. Using all-flash rule, we need to create 32 volumes to present as MDisks. However, as it is rather small setup, we can reduce a number of volumes to 16. On IBM FlashSystem 900, join all micro-latency modules into a RAID5 array and add it to a storage pool. FlashSystem 900 is Tier0 solution, so use all-flash rule, and create 32 volumes to present as MDisks. On virtualizer, add 16 MDisks from IBM FlashSystem 5030 as Tier1 flash, and 32 MDisks as Tier0 flash, to a single multi-tier pool.

Large setup considerations

For controllers like IBM DS8000 and XIV, you can use 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 with the following formula:

$$Q = ((P \times C) / N) / M$$

Where:

- Q** Calculated queue depth for each MDisk
- P** Number of backend controller host ports (unique 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 like XIV or DS8000
- N** Number of nodes in the IBM FlashSystem cluster (2, 4, 6, or 8)
- M** Number of volumes that are presented by back-end controller and detected as MDisks

For a result of $Q = 60$, calculate the number of volumes needed to create as $M = (P \times C) / (N \times Q)$ $?(16 \times P) / N$.

Example #3:

4-node IBM FlashSystem 9200 is used with 12 host ports on the IBM XIV System. By using the formula above, we need to create $M = (16 \times 12) / 4 = 48$ volumes on IBM XIV to obtain balanced high-performing configuration.

3.3.3 Considerations for compressing and deduplicating back-end

IBM FlashSystem supports over-provisioning on selected back-end controllers. This means that if back-end storage performs data deduplication or data compression on LUs provisioned from it, the LUs still can be used as external MDisks on IBM FlashSystem.

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.

The IBM FlashSystem will detect:

- ▶ 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 and 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 and it varies based on the nature of the data. Always use a conservative data reduction ratio for the initial configuration.

Using the inappropriate ratio for capacity assignment could cause an out of space situation. If the MDisks do not provide enough capacity, IBM FlashSystem disables access to all the volumes in the storage pool.

Example:

Assumption 1: Sizing is performed with an optimistic 5:1 rate

Assumption 2: Real rate is 3:1

Physical Capacity: 20 TB

Calculated capacity: $20 \text{ TB} \times 5 = 100 \text{ TB}$

Volume assigned from compression or deduplication enabled storage subsystem to SAN

Volume Controller or Storwize is 100 TB

Real usable capacity: $20 \text{ TB} \times 3 = 60 \text{ TB}$

If the hosts try to write more than 60 TB data to the storage pool, the storage subsystem cannot provide any more capacity, and all volumes that are used as IBM Spectrum Virtualize or Storwize Managed Disks 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.

Also, the 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.

3.4 Controller-specific considerations

This section discusses implementation specifics related to different supported back-end systems. For general requirements, see [IBM FlashSystem 9200 8.4.0 Documentation - Configuring and servicing storage systems](#).

3.4.1 Considerations for DS8000 series

Interaction between DS8000 and IBM FlashSystem

It is important to know 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 the IBM FlashSystem.

In this regard, the basis for virtualization begins with the physical drives of DS8000, which are mounted in storage enclosures. Virtualization builds upon 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 the RAID 6 is highly preferred, and is 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 [DS8900 9.1.1 Documentation - RAID implementation](#).

A rank, which is a logical representation for the physical array, is relevant for an IBM FlashSystem 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.

A fixed-block rank features an extent size of one of the following:

- ▶ 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). A fixed-block 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 Managed Disks in an IBM FlashSystem. Use IBM Spectrum Virtualize Copy Services instead. Based on the considerations, the maximum LUN sizes to create at DS8900F and present to IBM FlashSystem are as follows:

- ▶ 16 TB LUN with large extents (1 GiB)
- ▶ 16 TB LUN with small extent (16 MiB) for DS8880F with version or edition R8.5 or later, and for DS8900F R9.0 or later.

Logical subsystems (or LSS) are another logical construct, and mostly used in conjunction with fixed-block volumes. Thus, a maximum of 255 LSSs can exist on DS8900F. For more information, see [DS8900 9.1.1 Documentation - Actions on the Volumes by LSS page](#).

The concepts of virtualization of DS8900F for IBM FlashSystem or IBM SVC are schematically shown in Figure 3-4 on page 85.

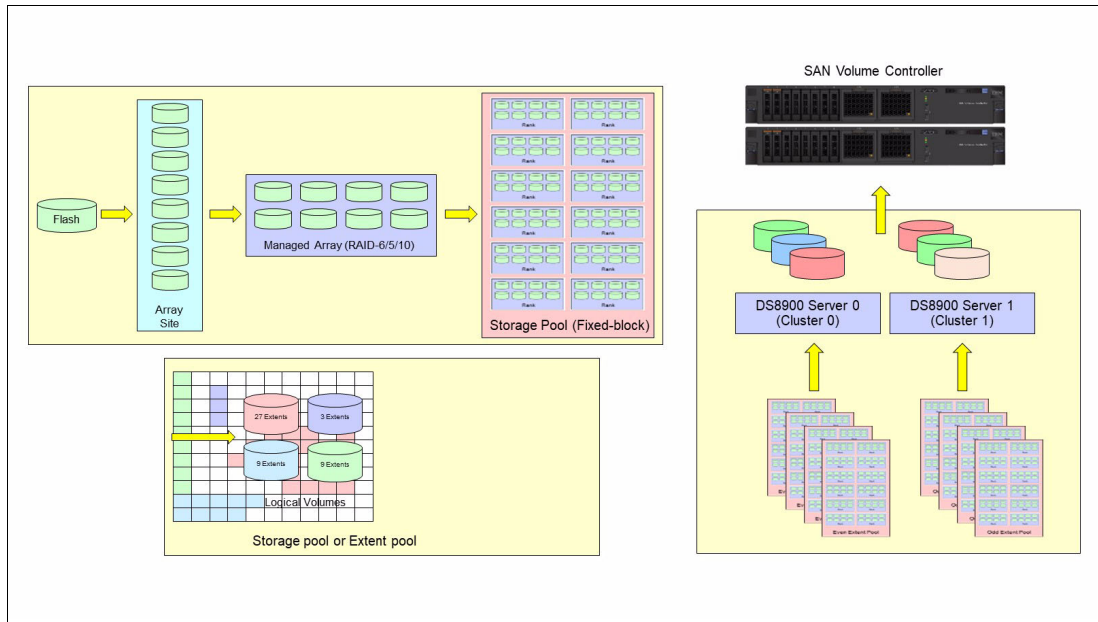


Figure 3-4 DS8900 virtualization concepts focus to IBM FlashSystem

Connectivity considerations

The number of DS8000 ports to be used is at least eight. With large and workload intensive configurations, consider using more ports, up to 16, which is the maximum 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 LIC update, a host adapter port might need to be taken offline. This configuration allows the IBM FlashSystem I/O to survive a hardware failure on any component on the SAN path.

For more information about SAN preferred practices and connectivity, see Chapter 2, “Storage area network” on page 33.

Defining storage

To optimize the DS8000 resource utilization, use the following guidelines:

- ▶ Distribute capacity and workload across device adapter 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 ranks on a device adapter (DA) pair are evenly balanced between odd and even extent pools. If you

do not ensure that the ranks are balanced, uneven device adapter 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. Notice how 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.

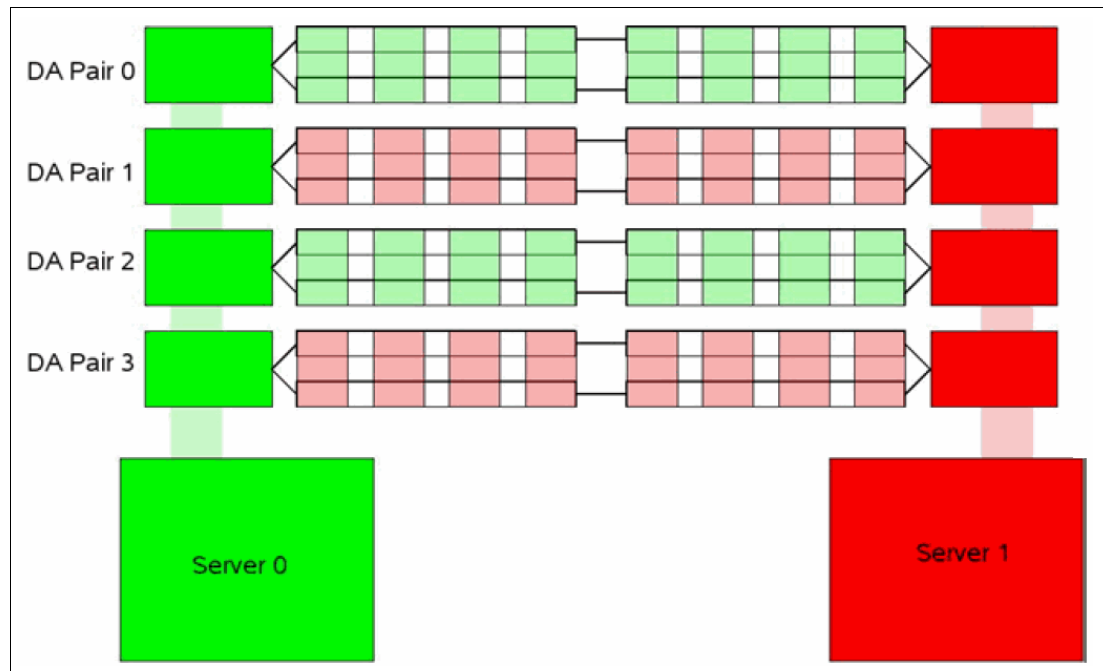


Figure 3-5 DA pair reduced bandwidth configuration

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.

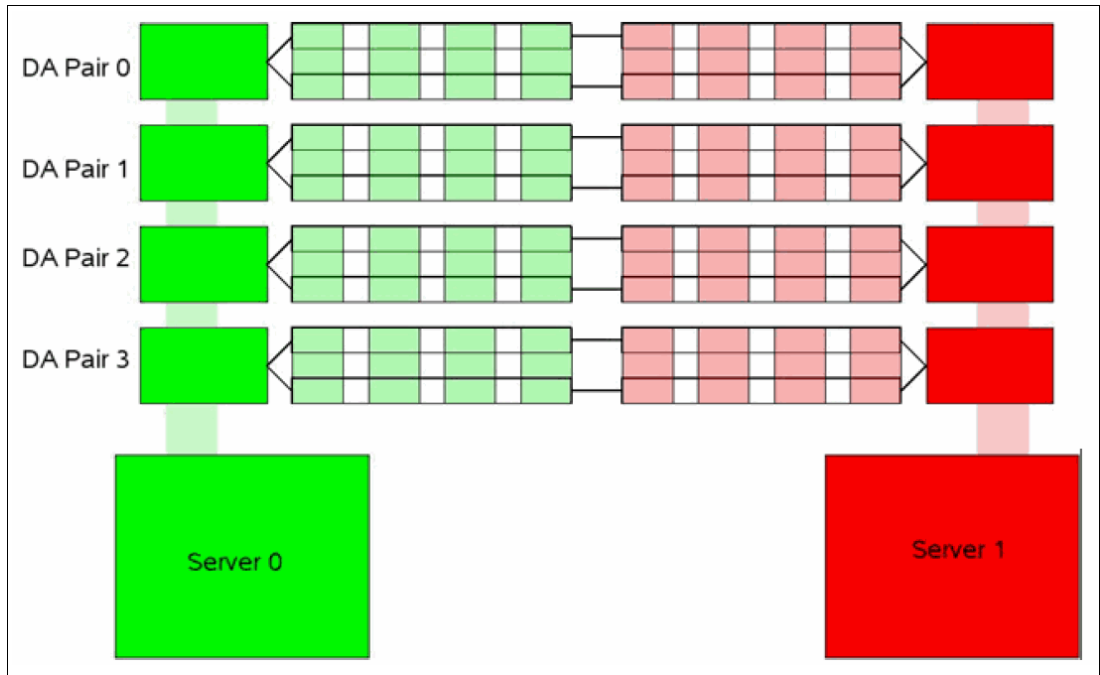


Figure 3-6 DA pair correct configuration

Figure 3-7 shows a correct configuration, as depicted in the CLI output of the `lsarray` and `lsrank` commands. Notice that 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 FlashSystem storage pool. IBM FlashSystem stripes across all of 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) containing MDisks of all the 6+P RAID 5 ranks of the DS8000 storage system
- One (or more) containing the slightly larger 7+P RAID 5 ranks

This approach maintains equal load balancing across all ranks when the IBM FlashSystem striping occurs because each MDisk in a storage pool is the same size.

The IBM FlashSystem 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 justifications. However, it also has a few minor drawbacks.

- There can be 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 FlashSystem 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 either DS8000 storage pool striping or automated EasyTier 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 EasyTier is activated on the DS8000 pools. However, when virtualized, tier management has more advantages when handled by the IBM FlashSystem. For information on choosing the level on which to run EasyTier, see “External controller tiering considerations” on page 177.

You need only one volume size with this multi-rank approach because plenty of space is available in each large DS8000 extent pool. As mentioned previously, the maximum number of back-end storage ports to be presented to the IBM FlashSystem is 16. Each port represents a path to the IBM FlashSystem. Therefore, when sizing the number of LUN/MDisks to be presented to the IBM FlashSystem, the suggestion is to present least between two and four volumes per path. So using the maximum of 16 paths, create 32, 48, or 64 DS8000 volumes, and for this configuration IBM FlashSystem 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 General Parallel File System (GPFS), use the single-rank extent pools.

LUN masking

For a storage controller, all IBM FlashSystem nodes must detect the same set of LUs from all target ports that logged in. If target ports are visible to the nodes or canisters that do not have the same set of LUs assigned, IBM FlashSystem 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 the IBM FlashSystem.

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 being presented to a two-node IBM FlashSystem cluster.

Example 3-7 Output of the showvolgrp command

```
dscli> showvolgrp V0
Date/Time: Oct 20, 2016 10:33:23 AM BRT IBM DSCLI Version: 7.8.1.62 DS: IBM.2107-75FPX81
Name ITSO_SVC
ID V0
Type SCSI Mask
```

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) and, therefore, 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 DSCLI 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
    
```

From Example 3-8 you can see that only the IBM FlashSystem WWPNs are assigned to V0.

Attention: Data corruption can occur if the same LUN is assigned to IBM FlashSystem nodes and other devices, such as hosts attached to DS8000.

Next, you see how the IBM FlashSystem detects these LUNs if the zoning is properly configured. The Managed Disk Link Count (`mdisk_link_count`) represents the total number of MDisks that are presented to the IBM FlashSystem 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

```

IBM_FlashSystem:FS9100-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 FlashSystem MDisks and storage pool considerations

Recommended practice is to create a single IBM FlashSystem storage pool per DS8900F system. This provides simplicity of management, and best overall performance.

An example of the preferred configuration is shown in Figure 3-8 on page 91. Four storage pools or extent pools (one even and one odd) of DS8900F are joined into one IBM FlashSystem storage pool.

To determine how many logical volumes need to be created to present to IBM FlashSystem as MDisks, see 3.3.2, “Guidelines for creating optimal backend configuration” on page 80.

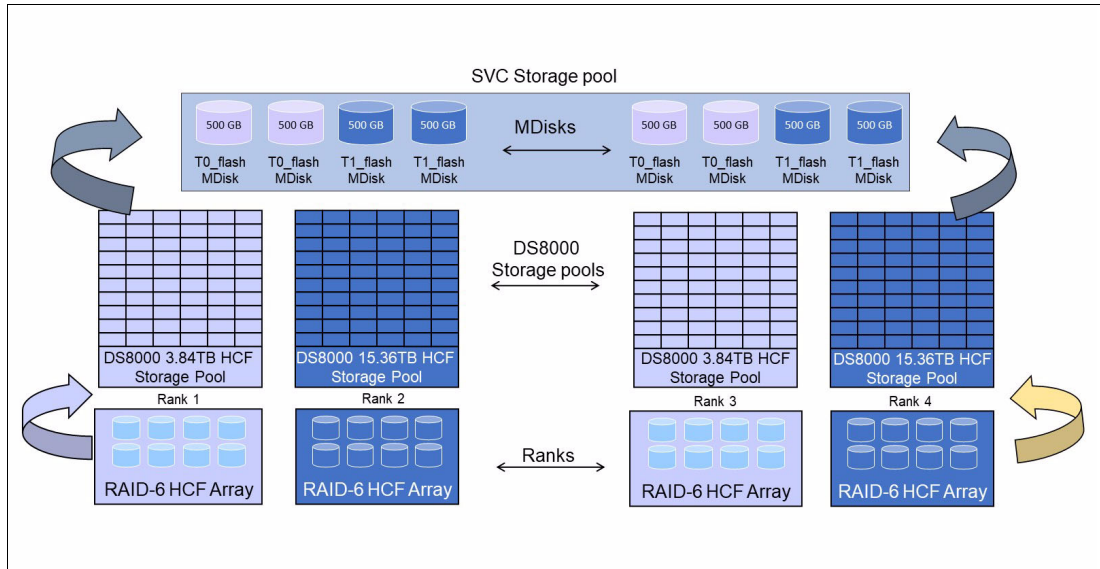


Figure 3-8 Four DS8900F extent pools as one IBM FlashSystem storage pool

3.4.2 Considerations for IBM XIV Storage System

The XIV Gen3 volumes can be provisioned to IBM FlashSystem via iSCSI and via FC. However, it is preferred 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 FlashSystem systems as hosts 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 System must be associated with the same LUN map.

From an IBM FlashSystem perspective, an XIV Type Number 281x controller can consist of more than one WWPN. However, all are placed under one worldwide node number (WWNN) that identifies the entire XIV system.

Creating a host object for IBM FlashSystem for an IBM XIV

A single host object with all WWPNs of IBM FlashSystem nodes can be created when implementing IBM XIV. This technique makes the host configuration easier to configure. However, the ideal host definition is to consider each node IBM FlashSystem 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 IBM FlashSystem node level.

A detailed procedure to create a host on XIV is described 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 currently supports the following configurations:

- ▶ 28 - 81 TB when using 1 TB drives
- ▶ 55 - 161 TB when using 2 TB disks
- ▶ 84 - 243 TB when using 3 TB disks
- ▶ 112 - 325 TB when using 4 TB disks
- ▶ 169 - 489 TB when using 6 TB disks

Figure 3-9 details how XIV configuration varies according to the number of modules present 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	Enabled
Interface module 5 state	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled
Interface module 4 state	Enabled	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	18	18	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	528 GB	576 GB	624 GB	672 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 described in 3.3.2, “Guidelines for creating optimal backend configuration” on page 80 are preferred when you virtualize XIV with IBM Spectrum Virtualize.

Table 3-5 shows the suggested volume sizes and quantities for IBM FlashSystem on the XIV systems with different drive capacities.

Table 3-5 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

Additional considerations

Consider the following restrictions when using the XIV system as back-end storage for the IBM FlashSystem:

- ▶ Volume mapping

When mapping a volume, you must use the same LUN ID to all IBM FlashSystem 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 zero (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 IBM FlashSystem level.

As all LUNs on a single XIV system share performance and capacity characteristics, use a single IBM FlashSystem storage pool for a single XIV system.

- ▶ Thin provisioning

XIV thin provisioning pools are not supported by IBM FlashSystem. 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 the IBM FlashSystem.

For more information about 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/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/A9000R can easily yield an effective data capacity of five times the original usable physical capacity.

Deduplication can be implemented on the IBM FlashSystem by attaching an IBM FlashSystem A9000/A9000R as external storage instead of using IBM Spectrum Virtualize DRP-level deduplication.

There are several considerations when you are attaching an IBM FlashSystem A9000/A9000R system as a back-end controller.

Volume considerations

IBM FlashSystem A9000/A9000R designates resources to data reduction, and as this designation is always on, it is strongly advised that data reduction be done only in the IBM FlashSystem A9000/A9000R and not in the Spectrum Virtualize cluster. Otherwise, as IBM FlashSystem A9000/A9000R tries to reduce the data, unnecessary additional latency occurs.

Estimated data reduction is important because that helps determine volume size. Always try to use a conservative data-reduction ratio when attaching A9000/A9000R, because the storage pool will go offline if the back-end storage runs out of capacity.

To determine the controller volume size:

- ▶ Calculate effective capacity: reduce measured-data reduction ratio (for example, if the data reduction estimation tool 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 3-6 and Table 3-7.
- ▶ Volume size will be 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 3-6 Host connections for A9000

Number of controllers	Total FC ports available	Total ports that are connected to SAN Volume Controller	Actual ports that are connected
3	12	6	All controllers, ports 1 and 3

Table 3-7 Host connections for A9000R

Grid Element	Number of controllers	Total FC ports available	Total ports that are connected to SAN Volume Controller	Actual ports that are connected
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 that takes the storage pool offline. It is important to closely monitor the FlashSystem A9000/A9000R. If you start to run out of space, you can use the migration functions of Spectrum Virtualize to move data to another storage system. The following cases are examples:

Example #1:

FlashSystem A9000 with 57 TB of usable capacity, or 300 TB of effective capacity, at the standard 5.26:1 data efficiency ratio.

We were able to run the data reduction tool on a good representative sample of the volumes that we will be virtualizing, we know that we have a data reduction ratio of 4.2:1 and for extra safety will use 4:1 for further calculations. 4×57 gives you 228 TB. Divide this by 12 (six paths \times 2), and you get 19 TB per volume.

Example #2:

A five grid element FlashSystem A9000R, using 29 TB Flash enclosures, has a total usable capacity of 145 TB.

We are using 10 paths and have not run any of the estimation tools on the data. However, we know that the host is not compressing the data. We assume a compression ratio of 2:1, 2×145 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 more volumes and make them available to Spectrum Virtualize.

The biggest concern about the number of volumes is to ensure there is adequate queue depth. Given that the maximum volume size on the FlashSystem A9000/A9000R is 1 PB and you are ensuring two volumes per path, you should be able to create a small number of larger volumes and still have good queue depth and not have numerous volumes to manage.

Additional considerations

Spectrum Virtualize is able to detect that the IBM FlashSystem A9000 controller is using deduplication technology and show that the **Deduplication** attribute of the managed disk is **Active**.

Deduplication status is important because it allows IBM Spectrum Virtualize to enforce the following restrictions:

- ▶ Storage pools with deduplicated MDisks should only contain MDisks from the same IBM FlashSystem A9000 or IBM FlashSystem A9000R storage controller.
- ▶ Deduplicated MDisks cannot be mixed in an EasyTier enabled storage pool.

3.4.4 Considerations for FlashSystem 5000, 5100, 7200, 9100, and 9200

Recommendations that are listed in this section apply to a solution when IBM FlashSystem family system is virtualized by another IBM FlashSystem family system.

Connectivity considerations

It is expected that 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 additional SAN and zoning preferred practices, see Chapter 2, “Storage area network” on page 33.

System layers

Spectrum Virtualize systems have a concept of system layers. There are two layers - *storage* and *replication*. Systems that are configured into storage layer can work as a back-end storage. Systems that are configured into replication layer, can virtualize another IBM FlashSystem 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.

The system layer on IBM FlashSystem can be switched. For instructions and limitations, see [IBM FlashSystem 9200 8.4.0 Documentation - System layers](#).

Automatic configuration

IBM FlashSystem family systems that run code version 8.3x and above can be automatically configured for optimal performance as a back-end storage behind IBM SVC.

An automatic configuration wizard must be used on a system where volumes, pools, and host objects are not configured. Wizard will configure internal storage devices, create volumes and map the to the host object, representing IBM SVC.

Array and disk pool considerations

The back-end IBM FlashSystem family system can have a hybrid configuration, containing FlashCore Modules and SSD drives, or SSDs and spinning drives.

Internal storage attached to the back-end system needs to be joined into RAID arrays. You might need one or more DRAID6 arrays, depending on the number and the type of available drives. For RAID recommendations, see 3.2.2, “Array considerations” on page 73.

Consider creating a separate disk pool for each type (tier) of storage and use the EasyTier function on a front-end system. Front-end FlashSystem family systems cannot monitor EasyTier activity of the back-end storage. If EasyTier is enabled on both front-end 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, EasyTier must be enabled only on one level, preferably the front-end. For more recommendations on EasyTier with external storage, see Chapter 2, “Storage area network” on page 33.

For most use cases, standard pools are preferred to data-reduction pools on the back-end storage. If planned, the front-end will perform reduction. Data reduction on both levels is not recommended as it adds processing overhead and does not result in capacity savings.

If EasyTier is disabled on the back-end, as advised above, back-end FlashSystem pool extent size is not a performance concern.

Volume considerations

Volumes in IBM FlashSystem can be created as *striped* or *sequential*. The general rule is to create striped volumes. Volumes on back-end system must be fully allocated.

To determine a number of volumes to create on back-end IBM FlashSystem to provide a virtualizer as MDisks, see the general rules provided in 3.3.2, “Guidelines for creating optimal backend configuration” on page 80. When virtualizing back-end with spinning drives, 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 very small arrays (for example, if you have 16 or less 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 are recommended.

When choosing volume size, take into account which system (front-end or back-end) will perform compression. If data is compressed and deduplicated on the front-end SVC, FCMs will not be able to compress it further, which will result in a 1:1 compression ratio. So, the back-end volume size should be calculated from the pool physical capacity that is divided by the number of volumes (16 or more).

Example:

FlashSystem 9100 with 24 x 19.2 TB modules. This configuration will provide raw disk capacity of 460 TB, with 10+P+Q DRAID6 and one distributed spare, physical array capacity will be 365 TB or 332 TiB. As it is not recommended to provision more than 85% of a physical flash, we have 282 TiB. As we do not expect any compression on FCM (back-end is getting data that is already compressed by upper levels), we provision storage to upper level assuming 1:1 compression, which means we create 32 volumes $282\text{TiB} / 32 = 8.8$ TiB each.

If the front-end system is not compressing data, space savings will be achieved with FCM hardware compression. Use compression-estimation tools to determine the expected compression ratio 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 using the calculated effective pool capacity.

Example:

Storwize V7000 Gen3 with 12 x 9.6 TB modules. This configuration will provide raw disk capacity of 115 TB, with 9+P+Q DRAID6 and one distributed spare, physical capacity will be 85 TB or 78 TiB. As it is not recommended to provision more than 85% of a physical flash, we have 66 TiB. Compresstimator has shown that we can achieve 3.2:1 compression ratio, decreasing in and assuming 3:1, we have $66\text{ TiB} \times 3 = 198$ TiB of effective capacity. Create 16 volumes, $198\text{TiB} / 16 = 12.4$ TiB each. If compression ratio will be higher than expected, we can create and provision to front end more volumes.

3.4.5 Considerations for IBM FlashSystem 900

The main advantage of integrating FlashSystem 900 with IBM Spectrum Virtualize is to combine the extreme performance of IBM FlashSystem 900 with the Spectrum Virtualize enterprise-class solution such as tiering, volume mirroring, deduplication, and copy services.

When you configure the IBM FlashSystem 900 as a backend for Spectrum Virtualize family systems, you must 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 the 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 of the 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 on either side until all 12 slots are populated.

The array configuration is performed during system setup. The system automatically creates MDisk/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 Spectrum Virtualize system resources, create 32 volumes (or 16 volumes if 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 FlashSystem 900 MDisk storage pool. FlashSystem 900 can process I/Os much faster than traditional storage. Sometimes they are 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 the FlashSystem 900 can take 100 - 200 microseconds. So, in some rare use-case, it might be recommended to disable Spectrum Virtualize cache to optimize for maximum IOPS.

You must keep the cache *enabled* in the following situations:

- ▶ If volumes from FlashSystem 900 pool are compressed
- ▶ If volumes from FlashSystem 900 pool are in a Metro/Global Mirror relationship
- ▶ If volumes from FlashSystem 900 pool are in a FlashCopy relationship (either source or target)
- ▶ If the same pool has MDisk from FlashSystem 900 contains also MDisk 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 and Hitachi Data Systems

Many third-party storage options are available and supported, This section describes the pathing considerations for EMC VMAX and Hitachi Data Systems (HDS).

Most storage controllers, when presented to the IBM FlashSystem, are recognized as a single WWNN per controller. However, for some EMC VMAX and HDS storage controller types, the system recognizes each port as a different WWNN. For this reason, each storage port, when zoned to an IBM FlashSystem, appears as a different external storage controller.

IBM Spectrum Virtualize supports a maximum of 16 WWNNs per storage system, so it is preferred to connect up to 16 storage ports.

To determine a number of logical volumes or LUNs to be configured on third-party storage, see 3.3.2, “Guidelines for creating optimal backend configuration” on page 80.

3.5 Quorum disks

Note: This section does not cover IP-attached quorum. For information on P-attached quorum, see Chapter 7, “Business continuity” on page 339.

A system uses a quorum disk for two purposes:

- ▶ To break a tie when a SAN fault occurs, when exactly half of 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 its capacity is reserved for quorum data. Its 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 actually designated as quorums, capacity for quorum data is reserved on each of them, as the designation might change (for example, if 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 has 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 were able to 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:

- ▶ Internal drive or module must follow these rules:
 - Be a member of an array or a “Candidate”.
 - Be in “Unused” state cannot be quorums.
 - MDisk must be in “Managed” state. “Unmanaged” or “Image” MDisks cannot be quorums.
- ▶ External MDisks can be provisioned over only FC and not iSCSI.
- ▶ An MDisk must be presented by a disk subsystem, LUNs from which are supported to be quorum disks.

The system uses the following rules when selecting quorum devices:

- ▶ Fully connected candidates are preferred over partially connected candidates.

This means that in a multiple enclosure environment, MDisks will be 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 to be preferred over a second drive from the same enclosure.

If IBM FlashSystem 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 drive in control enclosure will be chosen rather than SAS expansion drive.

To become an active quorum device (tie-break device), it must be visible to all nodes in a cluster.

In practice, these rules mean:

- ▶ For IBM FlashSystem with a single control enclosure, quorums including active quorum disk are assigned out of its internal drives automatically. No actions required.
- ▶ For IBM FlashSystem with two or more I/O groups and with external storage virtualized, the active quorum will be 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 IBM FlashSystem with two or more I/O groups and without external storage, there will be no active quorum selected automatically. However, a standard topology cluster in most use cases will operate without any issues. For HyperSwap topology, IP quorum or FC-attached quorum needs to be deployed on the third site.

To list IBM FlashSystem quorum devices, run the **lsquorum** command as shown in Example 3-10.

Example 3-10 The lsquorum command

```
IBM_FlashSystem:FS9100-ITS0:superuser>lsquorum
quorum_index status id name controller_id controller_name active object_type
0             online 4
1             online 1             yes  drive
2             online 2             no   drive
```

To move quorum assignment, use the **chquorum** command. Note that it is not supported on NVMe drives, so you can move it only *from* NVMe drive, but not *to* NVMe drive.



Storage pools

This chapter describes considerations for planning storage pools for an IBM FlashSystem implementation. It explains various pool configuration options, including Easy Tier and data reduction pools (DRP). It provides and provides best practices on implementation and an overview of some typical operations with MDisks.

This chapter includes the following sections:

- ▶ 4.1, “Introduction to pools” on page 102
- ▶ 4.2, “Storage pool planning considerations” on page 123
- ▶ 4.3, “Data reduction pools best practices” on page 130
- ▶ 4.4, “Operations with storage pools” on page 137
- ▶ 4.5, “Considerations when using encryption” on page 148
- ▶ 4.6, “Easy Tier, tiered and balanced storage pools” on page 159

4.1 Introduction to pools

In general, a storage pool or pool, sometimes referred to by its familiar name of “*managed disk group*”, is a grouping of storage capacity that is used to provision volumes and logical units (LUNs) that can subsequently be made visible to hosts.

IBM FlashSystem supports the following types of pools:

- ▶ Standard pools: Parent pools and child pools
- ▶ Data reduction pools (DRPs): parent pools and Quotaless child pools

Standard pools were available since the initial release of IBM Spectrum Virtualize in 2003 and can include fully allocated or thin-provisioned volumes.

Real-time Compression (RTC) is allowed only with standard pools on some older IBM SAN Volume Controller (SVC) hardware models and should not be implemented in new configurations.

Note: The latest node hardware does not support RTC.

SA2 and SV2 SVC node hardware do not support the use of RTC volumes. To migrate a system to use these node types, all RTC volumes must be removed (migrated) to uncompressed standard pool volumes, or into a DRP.

IBM FlashSystems using standard pools cannot be configured with Real-time Compression.

DRPs represent a significant enhancement to the storage pool concept because the virtualization layer is primarily a simple layer that runs the task of lookups between virtual and physical extents. With the introduction of data reduction technology, compression, and deduplication, it has become more of a requirement to have an uncomplicated way to stay thin.

DRPs increase existing infrastructure capacity usage by employing new efficiency functions and reducing storage costs. The pools enable you to automatically de-allocate (not to be confused with deduplicate) and reclaim capacity of thin-provisioned volumes containing deleted data. In addition, for the first time, the pools enable this reclaimed capacity to be reused by other volumes.

Either pool type can be made up of different tiers. A tier defines a performance characteristic of that subset of capacity in the pool. Often, no more than three tier types are defined in a pool (fastest, average, and slowest). The tiers and their usage are managed automatically by the Easy Tier function.

4.1.1 Standard pool

Standard pools (also referred to as traditional storage pools), provide a way of providing storage in IBM FlashSystem. They use a fixed allocation unit of an extent. Standard pools are still a valid method to providing capacity to hosts. For more information about guidelines for implementing standard pools, see 4.2, “Storage pool planning considerations” on page 123.

IBM FlashSystem can define parent and child pools. A *parent* pool has all the capabilities and functions of a normal IBM FlashSystem pool. A *child* pool is a logical subdivision of a storage pool or managed disk group. Like a parent pool, a child pool supports volume creation and migration.

When you create a child pool in a standard parent pool the user 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 detracts from the available capacity in the parent pool. This process is different than the method with which child pools are implemented in a DRP. For more information, see “Quotaless data reduction child pool” on page 108.

A child pool inherits its tier setting from the parent pool. Changes to a parent’s tier setting are inherited by child pools.

A child pool supports the Easy Tier function if Easy Tier is enabled on the parent pool. The child pool also inherits Easy Tier status, pool status, capacity information, and back-end storage information. The I/O activity of 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 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.

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.

Choose extent size wisely according to your future needs. A small extent size limit 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 only create a 6 GiB volume, one entire extent is allocated to this volume (8 GiB) and hence 2 GiB are unused.

When you create or manage a parent pool, consider the following general guidelines:

- ▶ Ensure that all MDisks that are allocated to the same tier of a parent pool are the same RAID type. 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.
- ▶ An MDisk can be associated with only one parent pool.
- ▶ 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 for the duration of any migration between parent pools.
- ▶ Volumes that are allocated from a parent pool are by default striped across all the storage that is placed into that parent pool. Wide striping can provide performance benefits.

- ▶ You can only add MDisks to a parent pool that is in unmanaged mode. When MDisks are added to a parent pool, their mode changes from unmanaged to managed.
- ▶ You can delete MDisks from a parent pool under the following conditions:
 - 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.
 - The system ensures that all extents that are used by volumes in the child pool are migrated to other MDisks in the parent pool to ensure that data is not lost.

Important: Before you remove MDisks from a parent pool, ensure that the parent pool has enough capacity for child pools that are associated with the parent pool.

- ▶ 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.

Important: 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 the command line tools. See the `rmmdiskgrp` command-help for details.

- ▶ 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 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 code versions 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 similar to 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 if 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 general guidelines:

- ▶ Child pools are created automatically by IBM Spectrum Connect VASA client to implement VMware vVols.
- ▶ 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.
- ▶ 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. You can then use volume mirroring to migrate the volumes from the non-encrypted parent pool to the encrypted child pool.
- ▶ 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 data is not lost.
- ▶ You cannot shrink the capacity of a child pool to less than its real capacity. The system uses reserved extents from the parent pool that use multiple extents. 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.
- ▶ 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. The volume copy can then 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.
- ▶ When migrating a volume between parent and child pool (with the same encryption key or no encryption), the result is a *nocopy migration*. That is, the data does not move. Instead, the extents are re-allocated to the child or parent pool and the accounting of the used space is corrected. That is, the free extents are reallocated to the child or parent to ensure the total capacity allocated to the child pool remains unchanged.
- ▶ A special form of *quotaless* data reduction child pool can be created from a data reduction parent pool. For more information, see “Quotaless data reduction child pool” on page 108

Small Computer System Interface unmap in a standard pool

A standard pool can use Small Computer System Interface (SCSI) unmap space reclamation, but not as efficiently as a DRP.

When a host submits a SCSI **unmap** command to a volume in a standard pool, the system changes the **unmap** command into a **write_same** command of zeros. This **unmap** command becomes an internal special command and can be handled accordingly by different layers in the system.

For example, the cache does not mirror the data; instead, it passes the special reference to zeros. The RTC functions reclaim those areas (assuming 32 KB or larger) and shrink the volume allocation.

The back-end layers also submit the `write_same` command of zeros to the internal or external MDisk devices. For a Flash or SSD-based MDisk this process results in the device freeing the capacity back to its available space. Therefore, it shrinks the used capacity on Flash or SSD, which helps to improve the efficiency of *garbage collection* on the device and performance. The process of reclaiming space is called *garbage collection*.

For Nearline SAS drives, the `write_same` of zeros commands can overload the drives themselves, this can result in performance problems.

Important: A standard pool does shrink its used space as the result of a SCSI `unmap` command. The backend capacity might shrink its used space, but the pool used capacity will not shrink.

The exception is with RTC volumes where the reused capacity of the volume might shrink; however, the pool allocation to that RTC volume remains unchanged. It means that an RTC volume can reuse that unmapped space first before requesting more capacity from the thin provisioning code..

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. IBM FlashSystem supports thin-provisioned volumes in standard pools.

Note: While DRPs fundamentally support thin-provisioned volumes, they are used in conjunction with compression and deduplication. With DRPs, you should avoid the use of thin-provisioned volumes without additional data reduction.

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 significantly 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 using any real capacity. For more information about storage system, pool, and volume capacity metrics, see Chapter 9, “Monitoring” on page 363.

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. You can then 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 you have physically available in your IBM FlashSystem. 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 Spectrum Control and IBM Storage Insights” on page 400.

Thin provisioning on top of Flash Core Modules

If you use the compression functions that are provided by the IBM Flash Core Modules (FCM) in your FlashSystem 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 theoretical maximum capacity, which can be as large as 4:1. 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.

It is recommended that you start conservatively, especially if you are allocating this capacity to IBM SAN Volume Controller or another IBM FlashSystem (the virtualizer).

You must first understand your expected compression ratio. In an initial deployment, allocate approximately 50% fewer savings. You can easily add “volumes” to the back-end storage system to present as new external “MDisk” capacity to the virtualizer later if your compression ratio is met or bettered.

For example, you have 100 TiB of physical usable capacity in an FCM RAID array before compression. Your compressor 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 as external MDisks to the virtualizer. Monitor the real compression rates and usage and over time add in the other 50 TiB of MDisk capacity to the same virtualizer pool. Be sure to leave spare space for unexpected growth, and consider the guidelines that are outlined in 3.2, “Arrays” on page 72

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 space. You now need to monitor the backend controller pool usage and the virtualizer pool usage in terms of volume thin provisioning over-allocation. 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 FlashSystem leverages innovative DRPs that incorporate deduplication and hardware-accelerated compression technology, plus SCSI **unmap** support. It also uses all of the thin provisioning and data efficiency features that you expect from IBM Spectrum Virtualize-based storage to potentially reduce your CAPEX and OPEX. Also, all of 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 allocation (re-direct on all overwrites)
- ▶ Garbage collection to free whole extents
- ▶ Fine-grained (8 KB) chunk allocation/de-allocation within an extent.
- ▶ SCSI unmap and write same (Host) with automatic space reclamation
- ▶ Support for “back-end” unmap and write same
- ▶ Support compression
- ▶ Support deduplication
- ▶ Support for traditional fully allocated volumes

Data reduction can 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 providing the ability to reclaim capacity that is no longer in use.

The potential capacity savings that compression alone can provide are shown directly in the GUI interfaces by way of the included “comprestimator” functions. Since version 8.4 of the Spectrum Virtualize software, comprestimator is always on and you can see the overall expected savings in the dashboard summary view. The specific savings per volume in the volumes views also are available.

To estimate potential total capacity savings that data reduction technologies (compression and deduplication) can provide on the system, use the Data Reduction Estimation Tool (DRET). This tool is a command line, host-based utility that analyzes user workloads that are to be migrated to a new system. The 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.

You download DRET and its readme file to a Windows client and follow the installation instructions in the readme. The readme file also describes how to use DRET on a variety of host servers.

See [FixCentral](#) and search under IBM SAN Volume Controller to find the tool and its readme file.

To use data reduction technologies on the system, you must create a DRP, and create compressed or compressed and deduplicated volumes.

For more information, see 4.1.4, “Data reduction estimation tools” on page 115.

Quotaless data reduction child pool

From version 8.4, DRP added support for a special type of child pool, known as a *quotaless child pool*.

The concepts and high-level description of parent-child pools are the same as for standard pools with a few major exceptions.

- ▶ You cannot define a capacity or 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 that belong to the child pool.
- ▶ Child and parent pools share the same data volume; therefore, data is de-duplicated between parent and child volumes.
- ▶ A DRP child pool can consume 100% of the capacity of the parent pool
- ▶ The **migratevdisk** commands can now be used between parent and child pools. Because they share the encryption key, this operation becomes a “nocopy” operation.

To create a DRP child pool, use the new pool type of “child_quotaless”

Because a DRP share capacity between volumes (when deduplication is used), it is virtually impossible to attribute capacity ownership of a specific grain to a specific volume because it

might be used by two more volumes, which is the 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 now be applied to DRP child pools or volumes as OBAC requires a child pool to function.

VMware vVols for DRP is not yet supported or certified at the time of writing; however, it is now technically possible because vVols support requires child pools.

SCSI unmap

DRPs support end-to-end unmap functionality. Space that is freed from the hosts by means of a SCSI **unmap** command results in the reduction of the used space in the volume and pool.

For example, a user deletes a small file on a host, which the operating system turns into a SCSI **unmap** for the blocks that made up the file. Similarly, a large amount of capacity can be freed if the user deletes (or Vmotions) a volume that is part of a data store on a host. This process might result in many contiguous blocks being freed. Each of these contiguous blocks results in a SCSI **unmap** command being sent to the storage device.

In a DRP, when the IBM FlashSystem receives a SCSI **unmap** command, the result is that the capacity is freed that is allocated within that contiguous chunk. The deletion is asynchronous, and the unmapped capacity is first added to the “reclaimable” capacity, which is later physically freed by the garbage collection code. For more information, see 4.1.5, “Understanding capacity use in a data reduction pool” on page 120.

Similarly, deleting a volume at the DRP level frees all of the capacity back to the pool. The DRP also marks those blocks as “reclaimable” capacity, which the garbage collector later frees back to unused space. After the garbage collection frees an entire extent, a new SCSI **unmap** command is issued to the backend MDisk device.

Unmapping can help ensure good MDisk performance; for example, Flash drives can reuse the space for wear-leveling and to maintain a healthy capacity of “pre-erased” (ready to be used) blocks.

Virtualization devices like IBM FlashSystem with external storage can also forward unmap information (such as when extents are deleted or migrated) to other storage systems.

Enabling, monitoring, throttling and disabling SCSI unmap

By default, host-based unmap support is disabled on all product other than the FlashSystem 9000 series. Backend unmap is enabled by default on all products.

To enable or disable host-based unmap, run the following command:

```
chsystem -hostunmap on|off
```

To enable or disable backend unmap run the following command:

```
chsystem -backendunmap on|off
```

You can check how much SCSI unmap processing is occurring on a per volume or per-pool basis by using the performance statistics. This information can be viewed with Spectrum Control or Storage Insights.

Note: SCSI `unmap` might add more workload to the backend storage.

Performance monitoring helps to notice possible effects and if SCSI `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 just deleted many volumes.

You can throttle the amount of “offload” operations (such as the SCSI `unmap` command) using the per-node settings for offload throttle. For example:

```
mkthrottle -type offload -bandwidth 500
```

This setting limits each node to 500MiBps of offload commands.

You can also stop the IBM FlashSystem from processing SCSI `unmap` operations for one or more host systems. You might find an over-zealous host, or not have the ability to configure the settings on some of your hosts. To modify a host to disable `unmap`, change the host type:

```
chhost -type generic_no_unmap <host_id_or_name>
```

If you experience severe performance problems as a result of SCSI `unmap` operations, you can disable SCSI `unmap` on the entire IBM FlashSystem for the front end (host), backend, or both.

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. That is, when created that space is reserved (used) from the DRP and is not available for other volumes in the DRP.

Data will not be deduplicated or compressed in a fully allocated volume. Similarly, because it does not use the internal fine-grained allocation functions, the allocation and performance are the same or better than 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 by its nature thin-provisioned. A compressed volume uses only its compressed data size in the pool. The volume grows only as you write data to it.

It is possible, but *not* recommended that you create a deduplicated-only volume in a DRP. A deduplicated volume is thin-provisioned in nature. The additional processing that is required to also compress the de-duplicated block is minimal; therefore, it is recommended that you create a compressed and de-duplicated volume rather than only a de-duplicated volume.

The DRP will first look for deduplication matches; then, it compress the data before writing to the storage.

Thin-provisioned only volumes in a DRP

It is not recommended that you create a thin-provisioned only volume in a DRP.

Thin-provisioned volumes use the fine-grained allocation functions of DRP. The main benefit of DRP is in the data reduction functions (compression and deduplication). Therefore, if you want to create a thin-provisioned volume in a DRP, create a compressed volume.

Note: In some cases, when the backend storage is thin-provisioned or data reduced, the GUI might not offer the option to create only thin-provisioned volumes in a DRP. This issue occurs because it is highly recommended that you don't use this option because it can cause extreme capacity-monitoring problems with a high probability of running out of space.

DRP internal details

DRPs consists of various internal metadata volumes and it is important to understand how these metadata volumes are used and mapped to user volumes. Each user volume has a corresponding journal, forward lookup, and directory volume.

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. However, internally it defines a Directory Volume for each user volume that is created within the pool. The directory points to grains of data that are stored in the Customer Data Volume. All volumes in a single DRP use the same Customer Data Volume to actually store their data. Therefore, deduplication is possible across volumes in a single DRP.

Other internal volumes are created, one per DRP. There is one Journal Volume per I/O group that can be used for recovery purposes, 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 denotes the difference between DRP volumes and volumes in standard pools.

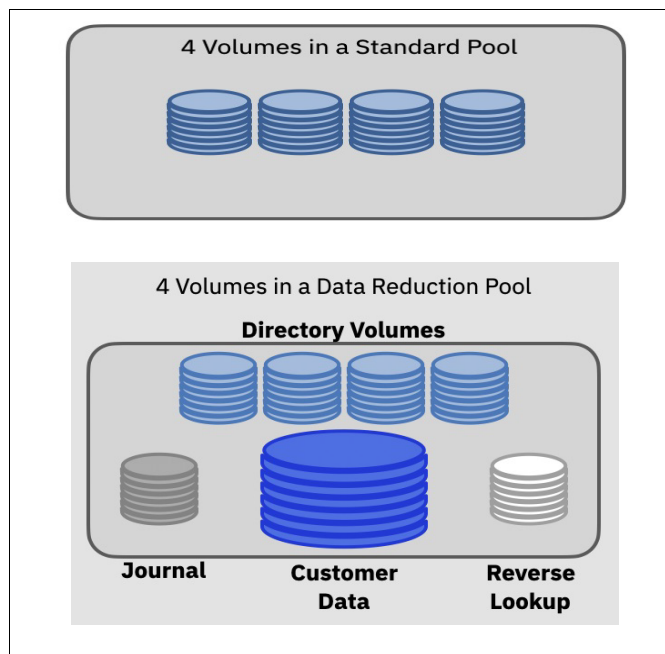


Figure 4-1 Standard and data reduction pool - volumes

The Customer Data Volume 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 and 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 is to reclaim space; that is, to track all of the regions that were invalidated, and to 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. This issue leads to the writes always occupying new space on back-end storage while the old data is still in its original location.

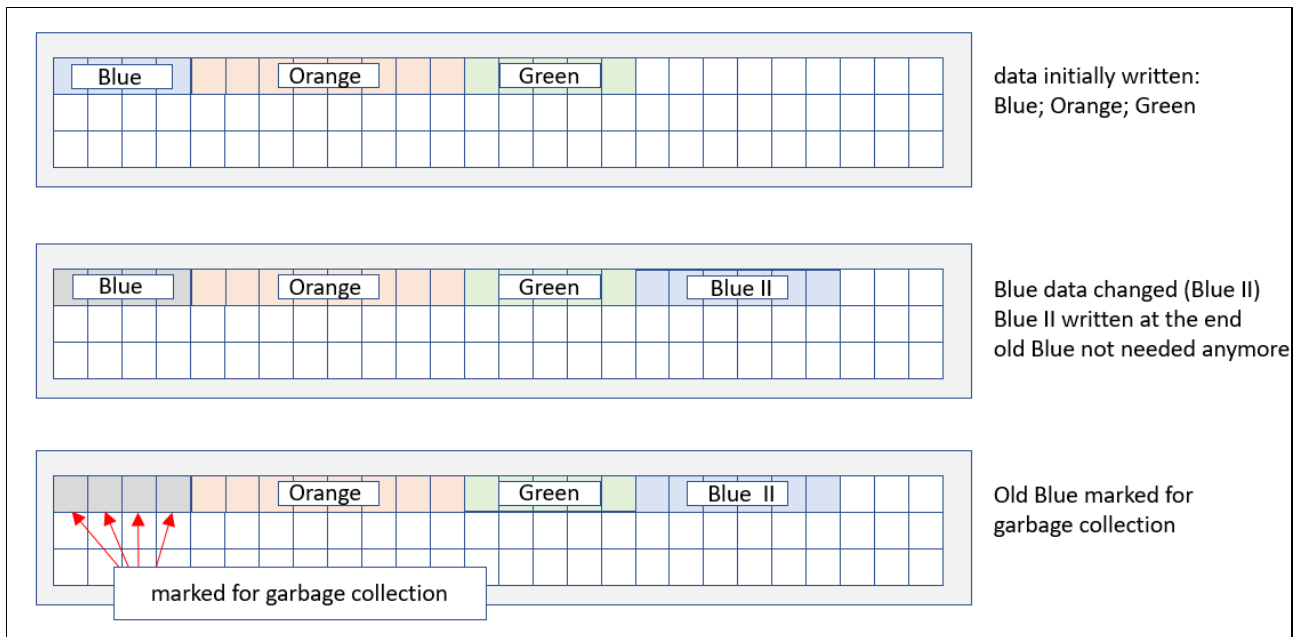


Figure 4-2 Garbage Collection principle

For garbage collection, stored data is divided into regions. As data is overwritten, a record is kept of which areas of those regions have been invalidated. Regions that have many invalidated parts are potential candidates for garbage collection. When the majority of a region has invalidated data, it is fairly inexpensive to move the remaining data to another location, therefore freeing 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 created by overwrites or unmaps. garbage collection looks for extents with the most unused space. After a whole extent has had all valid data moved elsewhere, it can be freed back to the set of unused extents in that pool, or it can be reused for new written data.

Because garbage collection needs to move data to free regions, it is suggested that you size pools to keep a specific amount of free capacity available. This practice ensures that some free space for garbage collection. For more information, see 4.1.5, “Understanding capacity use in a data reduction pool” on page 120.

4.1.3 Standard pools versus data reduction pools

When it comes to designing pools during the planning of an IBM FlashSystem project, it is important to know all requirements, and to understand the upcoming workload of the environment. The IBM FlashSystem is flexible in creating and using pools. This section describes how to figure out which types of pool or setup you can use.

Some of the information that you should be aware of in the planned environment is as follows:

- ▶ Is your data compressible?
- ▶ Is your data deduplicable?
- ▶ What are the workload and performance requirements?
 - Read/write ratio
 - Block size
 - Input/Output Operations per Second (IOPS), MBps, and response time
- ▶ Flexibility for the future
- ▶ Thin provisioning

Determine if your data is compressible

Compression is one option of DRPs. The deduplication algorithm is used to reduce the on-disk footprint of data that is written-to by thin provisioning. In IBM FlashSystem, this compression is an inline compression or a deduplication approach rather than an attempt to compress data as a background task. DRP provides unmap support at the pool and volume level. Out-of-space situations can be managed at the DRP pool level.

Compression can be enabled in DRPs on a per-volume basis, and thin provisioning is a prerequisite. The input size changed to a fixed 8 KB. Compression is suited to Flash workloads (IOPS) and a typical 2:1 compression ratio will result in approximately 4 KB operations and streaming 256 KB chunks of compressed 8 KB blocks for consistent write performance, by allowing the cache to build full stride writes enabling the most efficient RAID throughput.

Data compression techniques depend on the type of data that must be compressed and on the desired performance. Effective compression savings generally rely on the accuracy of your planning and the understanding if the specific data is compressible or not. Several methods are available to help you decide whether your data is compressible, including the following examples:

- ▶ General assumptions
- ▶ Tools

General assumptions

IBM FlashSystem compression is lossless; that is, data is compressed without losing any of the data. The original data can be recovered after the compress or expend cycle. Good compression savings might be achieved in the following environments (and others):

- ▶ Virtualized Infrastructure
- ▶ Database and Data Warehouse
- ▶ Home Directory, Shares, and shared project data
- ▶ CAD/CAM

- ▶ Oil and Gas data
- ▶ Log data
- ▶ SW 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 describes the compression ratio of common data types and applications that provide high compression ratios

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%

Also, do not compress encrypted data (for example, compression on host or application). Compressing already encrypted data does not result in many savings, because the data contains pseudo random data. The compression algorithm relies on patterns in order to gain efficient size reduction. Because encryption destroys such patterns, the compression algorithm would be unable to provide much data reduction.

For more information on compression, see 4.1.4, “Data reduction estimation tools” on page 115.

Note: Saving assumptions that are based on the type of data are imprecise. Therefore, you should determine compression savings with the proper tools.

Determine if your data is a deduplication candidate

Deduplication is done by using hash tables to identify previously written copies of data. If duplicate data is 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.
- ▶ Data reduction pools build 256 KiB chunks of data consisting of multiple de-duplicated and compressed 8 KiB grains.
- ▶ Data reduction pools will write contiguous 256 KiB chunks allowing for efficient write streaming with the capability for cache and RAID to operate on full stride writes.
- ▶ Data reduction pools provide deduplication then compress capability.
- ▶ The scope of deduplication is within a DRP within an I/O Group.

General assumptions

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 environments. Therefore, these environments might be good candidates for deduplication.

IBM provides the Data Reduction Estimate Tool (DRET) to help determine the deduplication capacity-saving benefits.

4.1.4 Data reduction estimation tools

IBM provides two tools to estimate the savings when you use data reduction technologies.

- ▶ **Comprestimator**

This tool is built into the IBM FlashSystem. It reports the expected compression savings on a per-volume basis in the GUI and command line.

- ▶ **Data Reduction Estimation Tool (DRET)**

The DRET tool must be installed on and used to scan the volumes that are mapped to a host and 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 of your volumes to provide an accurate summary.

Comprestimator

Comprestimator is provided in the following ways:

- ▶ As a stand-alone, host-based command-line utility. It can be used to estimate the expected compression for block volumes where you do not have an IBM Spectrum Virtualize product providing those volumes.
- ▶ Integrated into the IBM FlashSystem. In software versions before 8.4, triggering a volume sampling (or all volumes) was done manually.
- ▶ Integrated into the IBM FlashSystem and always on, in versions 8.4 and later.

Host-based Comprestimator

The tool can be downloaded at [IBM FlashSystem Comprestimator](#).

IBM FlashSystem Comprestimator is a command-line and host-based utility that can be used to estimate an expected compression rate for block devices.

Integrated Comprestimator - software levels before 8.4.0

IBM FlashSystem also features an integrated Comprestimator tool that is available through the management GUI and CLI. If you are considering to apply compression on existing non-compressed volumes in an IBM FlashSystem, you can use this tool to evaluate if compression will generate capacity savings.

To access the Comprestimator tool in management GUI, select **Volumes** → **Volumes**.

If you want to analyze all the volumes in the system, click **Actions** → **Capacity Savings** → **Estimate Compression Savings**.

If you want to select a list of volumes and click **Actions** → **Capacity Savings** → **Analyze** to evaluate only the capacity savings of the selected volumes, as shown in Figure 4-3 on page 116.

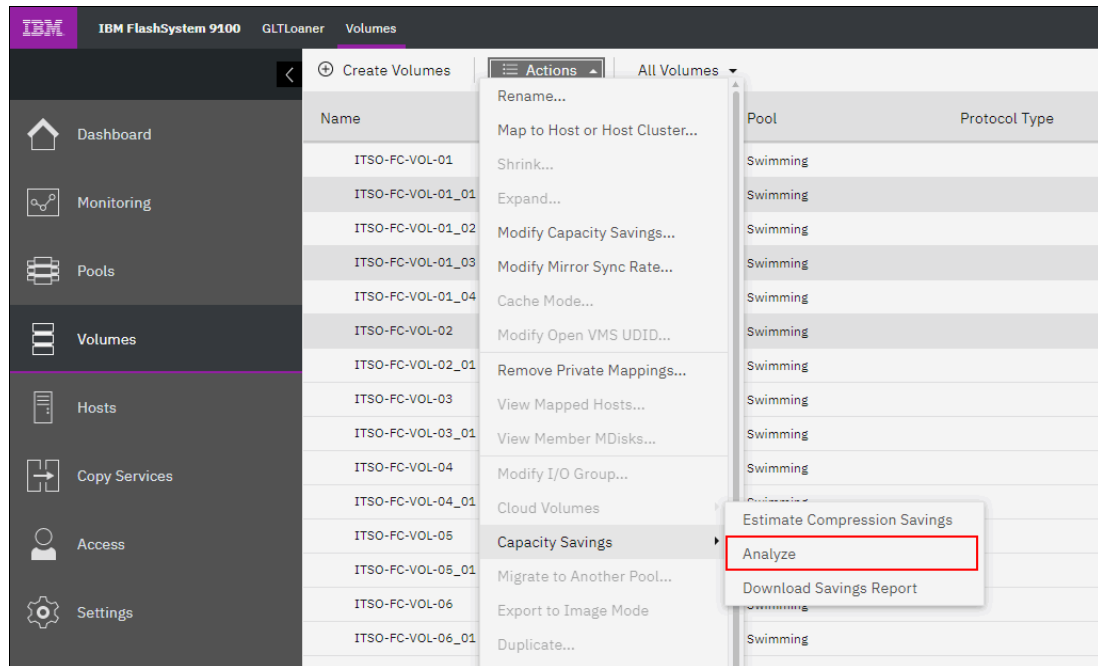


Figure 4-3 Capacity savings analysis

To display the results of the capacity savings analysis, click **Actions** → **Capacity Savings** → **Download Savings Report**, as shown in Figure 4-3, or enter the command `lsvdiskanalysis` in the command line, as shown in Example 4-1.

Example 4-1 Results of capacity savings analysis

```

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
IBM_FlashSystem:superuser>

```

The following actions are preferred practices:

- ▶ After you run Comprestimator, consider applying compression only on those volumes that show greater than or equal to 25% capacity savings. For volumes that show less than 25% savings, the trade-off between space saving and hardware resource consumption to compress your data might not make sense. With DRPs, the penalty for the data that cannot be compressed is no longer seen. However, the DRP includes overhead in grain management.
- ▶ After you compress your selected volumes, review which volumes have the most space-saving benefits from thin provisioning rather than compression. Consider moving these volumes to thin provisioning only. This configuration requires some effort, but saves hardware resources that are then available to give better performance to those volumes, which achieves more benefit from compression than thin provisioning.

You can customize the Volume view to view the metrics you might need to help make your decision, as shown in Figure 4-4.

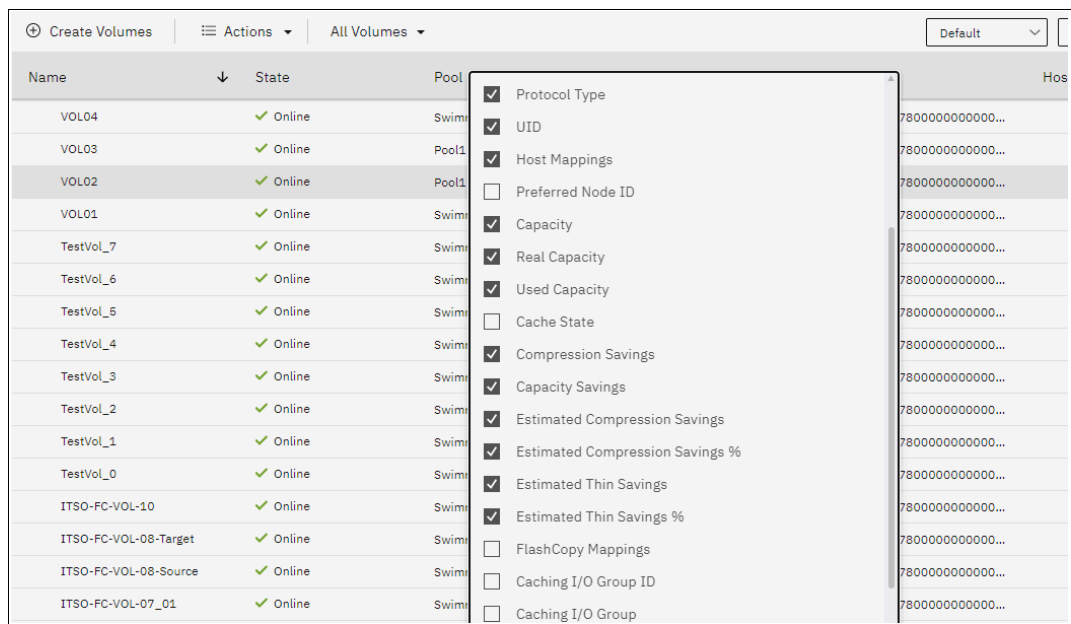


Figure 4-4 Customized view

Integrated comprestimator - software version 8.4 onwards

Because the newer code levels include an always-on comprestimator, you can view the expected capacity savings in the main dashboard view, pool views, volume views. You do not need to first trigger the “estimate” or “analyze” tasks; these are performed automatically as background tasks.

Data Reduction Estimation Tool (DRET)

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 IBM FlashSystem planning.

The 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 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 additional 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 actually removed, but rather the file system index and pointers were updated to reflect this change.

When the DRET is used to analyze a block device 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, then delete all the files in the file system making it 0% used. When scanning the block device used for storing the file system in this example, the 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 the DRET to analyze volumes that contain as much active data as possible rather than volumes that are mostly empty of data. The use increases the accuracy level and reduces the risk of analyzing old data that is deleted, but might still have traces on the device.

The DRET can be downloaded from: [FixCentral](#).

Example 4-2 shows the DRET command line.

Example 4-2 DRET command line

```
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]
```

The DRET can be used on the following client operating systems:

- ▶ Windows 2008 Server, Windows 2012
- ▶ Red Hat Enterprise Linux Version 5.x, 6.x, 7.x (64-bit)
- ▶ UBUNTU 12.04
- ▶ ESX 5.0, 5.5, 6.0
- ▶ AIX 6.1, 7.1
- ▶ Solaris 10

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.

Determining the workload and performance requirements

An important factor of sizing and planning for an IBM FlashSystem 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 will affect 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, keep in mind that 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 for 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 sent to the IBM FlashSystem to no more than 256 KiB. This limitation is general 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 FlashSystem environment, capacity (physical) must be sized accordingly. Compression and deduplication might save space, but metadata uses little space. For optimal performance, our 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 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 on performance as well for robust application performance.

To size an IBM FlashSystem environment, your IBM account team or IBM Business Partner must access to *IBM Storage Modeller (StorM)*. The tool can be used to determine if DRPs can provide suitable bandwidth and latency. If the data does not deduplicate (according to the DRET), the volume can be either fully allocated or compressed only.

Flexibility for the future

During the planning and configuration of storage pools, you must decide pools to create. Because the IBM FlashSystem enables you to create standard pools or DRPs, you must decide which type best fits the requirements.

Verify if performance requirements meet the capabilities of the specific pool type. For more information, see “Determining the workload and performance requirements” on page 118. We will cover the dependencies with child pools with respect to vVols in 4.3.3, “Data reduction pools and VMware vVols” on page 132 and “DRP restrictions” on page 133.

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.

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 that it is not possible to convert a pool. If you must change the pool type (from standard pool to DRP, or vice versa), it will be an offline process and you will have to migrate your data as described in 4.3.7, “Data migration with DRP” on page 135.

Note: We recommend the use of DRPs pools with fully allocated volumes if the restrictions and capacity do not affect your environment. For more details about the restrictions, see “DRP restrictions” on page 133.

4.1.5 Understanding capacity use in a data reduction pool

This section describes capacity terms associated with DRPs.

After a reasonable period of time, the DRP will have approximately 15-20% of overall free space. 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 will dictate how much “reclaimable space” is present at any given time. The capacity in a DRP consists of the components that are listed in Table 4-2 on page 121.

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 allocated to fully allocated volumes (assumed to be 100% written)
Free	The amount of free space, not in use by any volume
Reclaimable Data	The amount of garbage in the pool. This is either old (overwritten) yet to be freed data or data that has is unmapped but not yet freed or associated with recently deleted volumes
Metadata	Approximately 1-3% overhead for DRP metadata volumes

Table 4-2 *DRP Capacity Uses*

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.

Figure 4-5 shows an example of steady state DRP.

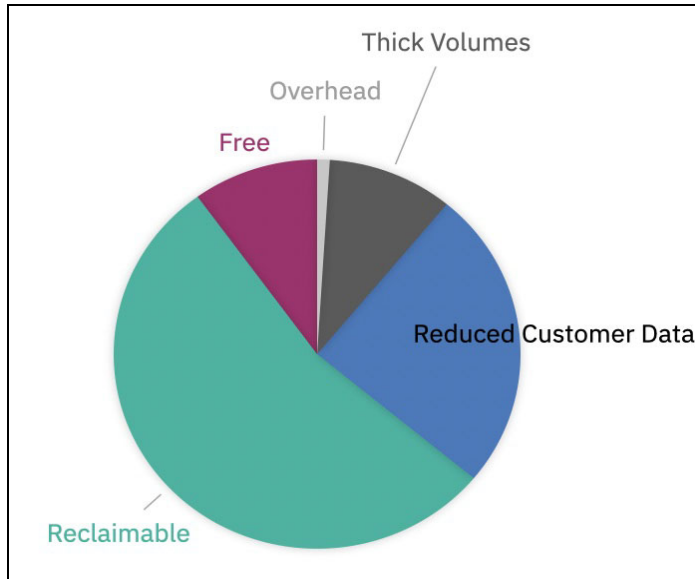


Figure 4-5 *Data reduction pool capacity use example*

Consider these points:

- ▶ If you create a large capacity of fully allocated volumes in a DRP, you are taking this capacity directly from free space only. This could result in triggering heavy garbage collection if there is little free space remaining and a large amount of reclaimable space, as shown in Figure 4-5.
- ▶ If you create a large number of 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 (compressed or deduplicated), the following is a two-phase approach can be used:
 - a. The grain must be inspected to determine if this was the last volume that referenced this grain (deduplicated):
 - If so, the grains can be freed.
 - If not, the grain references need to be updated and the grain might need to be re-homed 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. It is the responsibility of garbage collection to convert them to free space.
 - The garbage-collection process runs in the background, attempting to maintain a sensible amount of free space. If there is little free space and you delete a large number of volumes, the garbage-collection code might trigger a large amount of backend data movement and could result in performance issues.

- ▶ Deleting a volume might not immediately create free space.
- ▶ If you are at risk of running out of space, but a lot of 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 extreme caution when using up all or most of the free space with fully-allocated volumes. garbage collection requires free space in order to coalesce data blocks into whole extents and hence free capacity. If there is very little free space, then the garbage collector has to work even harder to free space.

- ▶ It might be worth creating some “get out of jail free” 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 in an efficient manner.

Consider these points:

- This type of volume should not be mapped to hosts.
- This type of volume should be labeled accordingly. For example, “RESERVED_CAPACITY_DO_NOT_USE”

4.2 Storage pool planning considerations

The implementation of storage pools in an IBM FlashSystem requires an holistic approach that involves application availability and performance considerations. Usually a trade-off between these two aspects must be taken into account.

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 additional specific best practices for DRPs, see 4.6, “Easy Tier, tiered and balanced storage pools” on page 159. For more information, see specific practices for high-availability solutions.

4.2.1 Planning for availability

By design, IBM Spectrum Virtualize based storage systems take the entire storage pool offline if a single MDisk in that storage pool goes offline. This means that the storage pool's 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.

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. For instance, in IBM FlashSystems configurations that do not include expansion enclosures, the number of physical drives is limited (up to 24), and creating more arrays reduces the usable space because of spare and protection capacity.

Consider, for instance, a single I/O group FlashSystem configuration with 24 7.68 TB NVMe drives. In a case of a single array DRAID 6 creation, the available physical capacity would be 146.3 TB, while creating two arrays DRAID 6 would provide 137.2 TB of available physical capacity with a reduction of 9.1 TB.

When virtualizing external storage, remember that the failure domain is defined by the external storage itself, rather than by the pool definition on the front-end system. For instance, 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 preferred practices when planning storage pools for availability:

- ▶ Create separate pools for internal storage and external storage, unless you are creating a hybrid pool managed by Easy Tier (see 4.2.5, “External pools” on page 128).
- ▶ Create a storage pool for each external virtualized storage subsystem, unless you are creating a hybrid pool managed by Easy Tier (see 4.2.5, “External pools” on page 128).
- ▶ Use dedicated pools for image mode volumes.

Limitation: Image Mode volumes are not supported with DRPs.

When you are selecting storage subsystems, the decision often comes down to the ability of the storage subsystem to be more reliable and resilient, and meet application requirements. While IBM Spectrum Virtualize does not provide any physical level-data redundancy for virtualized external storages, the availability characteristics of the storage subsystems' controllers have the most impact on the overall availability of the data that is virtualized by IBM Spectrum Virtualize.

4.2.2 Planning for performance

When planning storage pools for performance the capability to stripe across disk arrays is one of the most important advantages IBM Spectrum Virtualize provides. To implement performance-oriented pools, create large pools with many 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.

Adding more arrays to a pool, rather than creating a new pool, can be a way to improve the overall performance if the added arrays have the same or better performance characteristics than the existing ones.

Note that in IBM FlashSystem configurations arrays built from FCM and SAS SSD drives have different characteristic, both in terms of performance and data reduction capabilities. Therefore, when using FCM and SAS SSD arrays in the same pool, follow these recommendations:

- ▶ Enable the Easy Tier function (see 4.6, “Easy Tier, tiered and balanced storage pools” on page 159). The Easy Tier treats the two-array technologies as different tiers (`tier0_flash` for FCM arrays and `tier1_flash` for SAS-SSD arrays), so the resulting pool is a multi-tiered pool with inter-tier balancing enabled.
- ▶ Strictly monitor the FCM physical usage. As Easy Tier moves the data between the tiers, the compression ratio can vary frequently and an out-of-space condition can be reached without changing the data contents.

The number of arrays that are required in terms of performance must be defined in the pre-sales or solution design phase, but when sizing the environment remember that adding too many arrays to a single storage pool increases the failure domain, and therefore it is important to find the trade-off between the performance, availability, and scalability cost of the solution.

Using the following external virtualization capabilities, you can boost the performance of the back-end storage systems:

- ▶ Using wide-striping across multiple arrays
- ▶ Adding additional read/write cache capability

It is typically understood that wide-striping can add approximately 10% additional Input/Output Processor (IOP) performance to the backend-system by using these mechanisms.

Another factor is the ability of the virtualized-storage subsystems to 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 as to whether to scale out back-end system, or add them as individual systems behind IBM FlashSystem. Either case is valid. However, adding

individual controllers is likely to allow IBM FlashSystem to generate more I/O, based on queuing and port-usage algorithms. It is recommended that you add each controller (I/O Group) of an IBM FlashSystem back-end as its own controller; that is, do not cluster the IBM FlashSystem when it acts as an external storage controller behind another Spectrum Virtualize product, such as IBM SAN Volume Controller. Adding each controller (I/O Group) of an IBM FlashSystem backend as its own controller adds additional management IP addresses and configuration. However, it provides the best scalability in terms of IBM FlashSystem 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 additional hardware that is required lowers the availability characteristics of the IBM FlashSystem cluster.

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.

The following actions are the starting preferred 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 a FlashSystem clustered environment, create storage pools with IOgrp or Control Enclosure affinity. That means you have to use only arrays or MDisks 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 SAN and consuming Fibre Channel bandwidth.
- ▶ Use dedicated pools for image mode volumes.

Limitation: Image Mode volumes are not supported with DRPs.

- ▶ For Easy Tier-enabled storage pools, always allow free capacity for Easy Tier to deliver better performance.
- ▶ Consider implementing child pools when you need to have a logical division of your volumes for each application set. There are often cases where you want to subdivide a storage pool but maintain a larger number of MDisks in that pool. Child pools are logically similar to 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

The system automatically defines a logical cache partition per storage pool. Child pools do not count towards cache partitioning. The cache partition number matches the storage pool ID.

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 kick in and essentially slow down write operations in the pool to the same speed that the backend can handle. *Overloaded* means that the front-end write throughput is greater than back-end storage that the pool can sustain. This situation should be avoided.

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 workload should be moved from that pool, or additional storage capacity should be added to that pool.

4.2.3 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 used to report usable, used, and free capacity.

The terminology and its reporting in the GUI has changed in recent versions and is listed in Table 4-3.

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.
N/A	Written Capacity	The total capacity that is written to the volumes in the system. This is shown as a percentage of the provisioned capacity and is reported before any data reduction.

Table 4-3 Capacity Terminology in 8.4.0

The *usable capacity* describes the amount of capacity that can be written-to on the system and includes any backend data reduction (that is, the “virtual” capacity is reported to the system).

Note: In DRP, the `rsize` parameter, used capacity, and tier capacity are not reported per volume. These items are reported only at the parent pool level because of the complexities of deduplication capacity reporting.

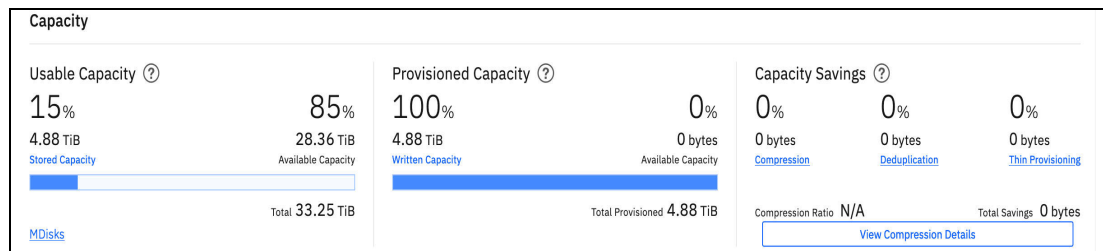


Figure 4-6 Example dashboard capacity view

For FlashCore Modules (FCM), this will be the maximum capacity that can be written to the system. However, for the smaller capacity drives (4.8 TB), this will report 20 TiB as usable. The actual usable capacity may be lower because of the actual data reduction 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 FCM, 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 respectively.

The *provisioned capacity* shows the total provisioned capacity in terms of the volume allocations. This is the “virtual” capacity that is allocated to fully-allocated, and thin-provisioned volumes. Therefore, it is in theory that the capacity could be written to if all volumes were filled 100% by the using system.

The *written capacity* is the actual amount of data that has been written into the provisioned capacity.

- ▶ For fully-allocated volumes, the written capacity is always 100% of the provisioned capacity.
- ▶ For thin-provisioned (including data reduced volumes), the written capacity is the actual amount of data the host writes to the volumes.

The final set of capacity numbers relates to the data reduction. This is reported in two ways:

- ▶ As the savings from DRP (compression and deduplication) provided at the DRP level, as shown in Figure 4-7.
- ▶ As the FCM compression

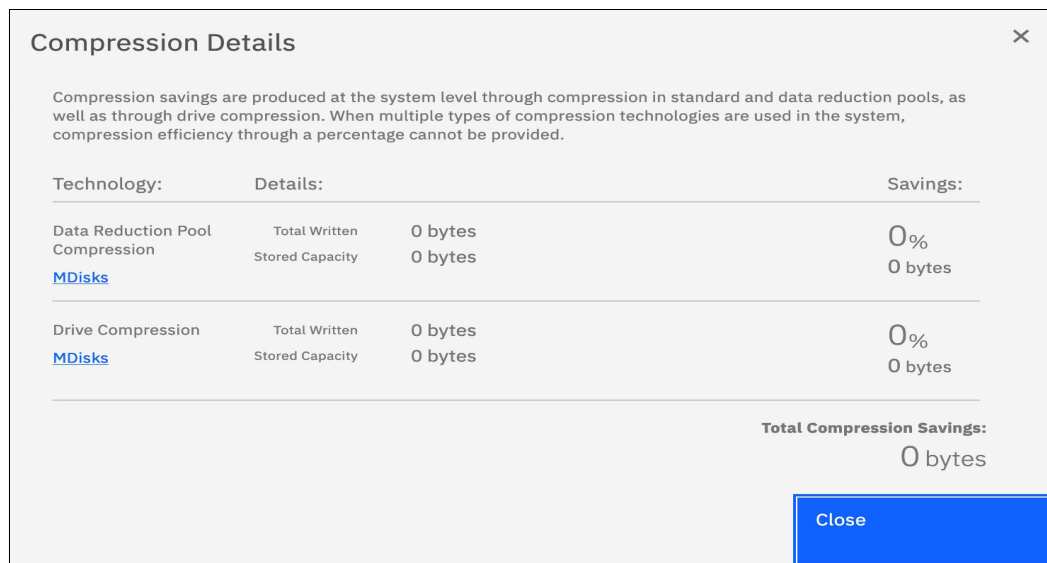


Figure 4-7 Compression Savings dashboard report

4.2.4 Extent size considerations

When adding MDisks to a pool they are logically divided into chunks of equal size. These chunks are called *extents* and are indexed internally. Extent sizes can be 16, 32, 64, 128, 256, 512, 1024, 2048, 4096, or 8192 MB. IBM Spectrum Virtualize architecture can manage 2²² extents for a system, and therefore the choice of extent size affects the total amount of storage that can be addressed. For the capacity limits per extent, see [V8.4.0.x Configuration Limits and Restrictions for IBM FlashSystem 9200](#).

When planning for the extent size of a pool, remember that you cannot change the extent size later, it must remain constant throughout the lifetime of the pool.

For pool-extent size planning, consider the following recommendations:

- ▶ For standard pools, usually 1 GB is suitable.
- ▶ For DRPs, use 4 GB (see 4.6, “Easy Tier, tiered and balanced storage pools” on page 159 for further considerations on extent size on DRP).
- ▶ With Easy Tier enabled hybrid pools, consider smaller extent sizes to better utilize the higher tier resources and therefore provide better performance.
- ▶ Keep the same extent size for all pools if possible. The extent-based migration function is not supported between pools with different extent sizes. However, you can use volume mirroring to create copies between storage pools with different extent sizes.

Limitation: Extent-based migrations from standard pools to DRPs are not supported unless the volume is fully allocated.

4.2.5 External pools

IBM FlashSystem-based storage systems have the ability to virtualize external storage systems. This section describes special considerations when configuring storage pools with external storage.

Availability considerations

IBM FlashSystem 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 the IBM FlashSystem cluster.

IBM Spectrum 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.

Remember that IBM FlashSystem must take 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 preferred practices:

- ▶ It is recommended that each storage pool must contain only MDisks from a single storage subsystem. An exception exists when you are working Easy Tier hybrid pools. For more information, see 4.6, “Easy Tier, tiered and balanced storage pools” on page 159.

- ▶ 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 4.6, “Easy Tier, tiered and balanced storage pools” on page 159.

When you are selecting storage subsystems, the decision often comes down to the ability of the storage subsystem to be more reliable and resilient, and meet application requirements.

IBM Spectrum Virtualize does not provide any physical-level data redundancy for virtualized external storages. The availability characteristics of the storage subsystems’ controllers have the most impact on the overall availability of the data that is virtualized by IBM Spectrum Virtualize.

Performance considerations

Performance is a determining factor, where adding IBM FlashSystem as a front-end results in considerable gains. Another factor is the ability of your virtualized storage subsystems to 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.
- ▶ IBM FlashSystem series can be scaled out with enough units to deliver the same 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 additional hardware that is required lowers the availability characteristics of the IBM FlashSystem cluster.

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 can also have effects on availability and performance.

The backend storage access is controlled through MDisks where the IBM FlashSystem acts like a host to the backend controller systems. Just as you have to 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 information on the queue depth algorithm, and the rules about how many MDisks to present for an external pool, see “Volume considerations” on page 91. This section details how many “volumes” to create on the backend controller (that are seen as MDisks by the virtualizing controller) based on the type and number of drives (such as HDD and SSD).

4.3 Data reduction pools best practices

This section describes the DRP planning and implementation best practices.

4.3.1 Data reduction pools with IBM FlashSystem NVMe attached drives

Note that the compression ratio reported in Table 3-1 on page 66 is the maximum achievable considering the effective capacity available. Depending on the pool filling, the actual compression ratio can be higher. To have an estimation of the compression ratio for a specific workload, see “Determine if your data is a deduplication candidate” on page 114.

Important: If you plan to use DRP with FCM storage, assume zero additional compression from the FCM. That is, use the reported physical or usable capacity from the RAID array as the actual usable capacity in the pool and ignore the above maximum effective capacity.

Assuming zero extra compression from the FCMs is the result of the DRP function sending compressed data to the FCM and thus 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.

When providing industry standard NVMe-attached flash drives capacity for the DRP, some considerations must be taken into account.

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 “Determine if your data is a deduplication candidate” on page 114.

Consider DRP configurations with IBM **FCM drives**:

- ▶ **Data is deduplicable.** In this case, the recommendation is to use compressed and deduplicated volume type. The double compression, first from DRP and then from FCMs, will not affect the performance and the overall compression ratio.
- ▶ **Data is not deduplicable.** In this case, you might use standard pools (instead of DRP with FCM), and let the FCM hardware do the compression, because the overall achievable throughput will be higher.

With standard off-the-shelf **NVMe drives**, which do not support 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 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:

1. Use FCM compression, unless your data deduplicates well with IBM FlashSystem Family products that support FCMs
2. With SSD and HDD, use DRP, deduplicate if applicable with IBM FlashSystem 5100, 7000, and 9000 family.
3. Use of a small amount (1-3%) of SCM capacity in a DRP will significantly improve 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 will very quickly promote the metadata extents to the fastest available tier.
4. Never create a DRP with only Nearline (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.
5. In general, DRP is avoided on FlashSystem 5030 unless you have very little performance expectations or requirements. The FlashSystem 5030 does not have additional offload hardware and uses the internal CPU to provide the compress and decompress the engine. This has very limited throughput capability and is only suitable for extremely low throughput workloads. Latency will also be adversely impacted in most cases.
6. Do not compress encrypted data. That is, if the application or operating system provides encryption, do not attempt to use DRP volumes. Data at rest encryption, provided by IBM FlashSystem is still possible because the encryption is performed after the data is reduced. If host-based encryption is unavoidable, assume data reduction is not be possible. That is, ensure there is a 1:1 mapping of physical-to-effective capacity.
7. While 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 standard pools or fully allocated volumes if no data reduction benefits are realized.
8. 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. See “Determine if your data is compressible” on page 113 for more details.
9. Avoid the use of any client, file system, or application based-compression with the system compression. If this is not possible, use a standard pool for these volumes.
10. Never use DRP on the IBM FlashSystem and virtualized external storage at same time (DRP over DRP). In all cases, use DRP at the virtualizer level rather than the backend storage as this simplifies capacity management and reporting.

4.3.2 DRP and external storage considerations

Avoid configurations that attempt to perform data reduction at two levels.

The recommended configuration is to run DRP at only the IBM FlashSystem that is acting as the virtualizer. For storage behind the virtualizer, you should provision fully-allocated volumes to the virtualizer.

By running in this configuration, you ensure that:

- ▶ The virtualizer understands the real physical capacity available and can warn and avoid out-of-space situations (where access is lost due to no space).

- ▶ Capacity monitoring can be wholly performed on the virtualizer level as 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 resource than the backend storage systems as it does not need to deal with RAID and so forth.

If you cannot avoid backend data reduction (for example the backend storage controller cannot disable its data reduction features), ensure that:

- ▶ You do not excessively over-provision the physical capacity on the backend.
 - For example, you have 100 TiB of real capacity. Start by presenting just 100 TiB of volumes to the IBM FlashSystem. Monitor the actual data reduction on the backend controller. If your data is reducing well over time, increase the capacity that is provisioned to the IBM FlashSystem.
 - This ensures you can monitor and validate your data reduction rates and avoids panic if you do not achieve the expected rates and have presented too much capacity to IBM FlashSystem.
- ▶ Do not run DRP on top of the backend device. Since the backend device is going to attempt to reduce the data, use a standard pool or fully-allocated volumes in the IBM FlashSystem DRP.
- ▶ Understand that IBM FlashSystem does not know the real capacity usage. You have to monitor and watch for out-of-space at the backend storage controller and the IBM FlashSystem.

Important: Never run DRP on top of DRP. This is wasteful and causes performance problems without additional capacity savings.

4.3.3 Data reduction pools and VMware vVols

At the time of writing, DRPs are not supported or certified with VMware vVols. Therefore, choose standard pools instead.

4.3.4 Data reduction pool configuration limits

Table 4-4 describes the limitations of DRPs (IBM FlashSystem version 8.4.0) at the time of writing. Keep in mind that since version 8.2.0, the software does not support 2145-CG8 or earlier node types. Only 2145-DH8 or later nodes support versions since 8.2.0.

Table 4-4 Data reduction pool properties

Property	Maximum Number	Comments
Data reduction pools per system	4	
MDisks per DRP	128	
Volumes per DRP	10,000 - (Number of DRPs x 12)	
Extents per I/O group per DRP	128 K	
Compressed volume copies in DRPs per system	-	No limit is imposed here beyond the volume copy limit per DRP

Property	Maximum Number	Comments
Compressed volume copies in DRPs per I/O group	-	No limit is imposed here beyond the volume copy limit per DRP
Deduplicated volume copies in DRPs per system	-	No limit is imposed here beyond the volume copy limit per DRP
Deduplicated volume copies in DRPs per I/O group	-	No limit is imposed here beyond the volume copy limit per DRP
Maximum volume capacity (fully allocated or data reduced)	256 TB	<ul style="list-style-type: none"> ▶ Maximum size for an individual fully allocated volume. ▶ Maximum size depends on the extent size of the storage pool. ▶ Comparison Table: Maximum Volume, MDisk and System capacity for each extent size. See V8.4.0.x Configuration Limits and Restrictions for IBM FlashSystem 9200

For more information, see: [IBM Support - Support Information for FlashSystem 9200](#).

V8.4.0.x Configuration Limits and Restrictions for IBM FlashSystem 9200

4.3.5 DRP provisioning considerations

This section describes practices to consider during DRP implementation.

DRP restrictions

Consider the following important restrictions when planning for a DRP implementation:

- ▶ Maximum number of supported DRPs is four.
- ▶ vVols is not currently supported in DRP.
- ▶ Volume shrinking is not supported in DRP with thin/compressed volumes.
- ▶ Non-Disruptive Volume Move (NDVM) is not supported with DRP volumes.
- ▶ The volume copy split of a Volume Mirror in a different I/O Group is not supported for DRP thinly-provisioned or compressed volumes.
- ▶ Image and Sequential mode VDisk are not supported in DRP.
- ▶ Extent level migration is not allowed between DRP unless volumes are fully allocated.
- ▶ Volume migrate for any volume type is permitted between a quotaless child and its parent DRP pool.
- ▶ A 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 4 GB extent size or above.
- ▶ Recommended pool size is at least 20 TB.
- ▶ Lower 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 capacity are not reported per volume for DRP volumes. Instead, only information 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 extent 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 107. Currently, the maximum number of extent supported for a data volume is 128 K. Note that according to Figure 4-1 on page 111, there is one data volume per pool. Table 4-5 shows the maximum size per pool, by extent size and I/O group number.

Table 4-5 Pool size by extent size and IO 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. Table 4-6 shows the minimum capacity that is required by extent size and I/O group number.

Table 4-6 Minimum recommended pool size by extent size and IO 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

Note that the values reported in Table 4-6 represent the minimum required capacity for a DRP to create a single volume.

When sizing a DRP, it is important to remember that the garbage-collection process is constantly running to reclaim the unused space, which optimizes the extents usage. For more information on the garbage-collection process, see “DRP internal details” on page 111.

This garbage-collection process then requires a certain amount of free space to work efficiently. For this reason it is recommended to keep approximately 15% free space in a DRP pool. For more information, see [Do not provision 100% of the physical flash](#).

4.3.6 Standard and DRP pools coexistence

While homogeneous configurations in terms of pool type are preferable, there is no technical reason to avoid using standard and DRP pools in the same system. In some circumstances, this coexistence is unavoidable. Consider the following scenarios:

- ▶ IBM FlashSystem installation that require VMware vVols support and data reduction capabilities for other environments. This scenario requires the definition of both standard and DRP pools because of the restriction of DRP regarding the vVols. For more information, see “DRP restrictions” on page 133.
- ▶ In this case, the standard pool will be used for vVols environments only, while the DRP will be used for the other environments. Note that some data-reduction capability can be achieved for the vVols standard pool by using the inline data compression provided by the IBM FCMs on FlashSystem.
- ▶ IBM FlashSystem installation that require an external pool for image mode volumes and data reduction capabilities for other environments. Also, this scenario requires the definition of both standard and DRP pools because of the restriction of DRP regarding the Image mode volumes. For more information, see “DRP restrictions” on page 133.

In this case, the standard pool will be used for Image mode volumes only, 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 229. DRP is used for all the other environments.

- ▶ IBM FlashSystem installation that includes a FlashSystem system with DRP capabilities as an external pool. In this scenario, the external pool must be a standard pool, as recommended in 4.3.2, “DRP and external storage considerations” on page 131. In this case, the internal storage can be defined in a separate DRP enabling the data reduction capabilities if needed.
- ▶ IBM FlashSystem installation that requires more than four pools.

4.3.7 Data migration with DRP

As mentioned in “DRP restrictions” on page 133, extent-level migration to and from a DRP (such as migrate-volume or migrate-extent functions) is not supported. For an existing IBM FlashSystem configuration, where you plan to move data to or from a DRP and use of data reduced volumes, there are two options: host-based migrations and volume mirroring based migrations.

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 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 can be summarized as follows:

1. Create the target volumes of the migration in the target pool. Note that, depending on the migration technique, the size and the amount of the volumes can be different from the original ones. For instance, you can migrate two 2 TB VMware datastore volumes in a single 4 TB datastore volume.
2. Map the target volumes to the host.
3. Rescan the 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 IBM FlashSystem.

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. Consider that the target volumes are subject to very high write-workload, which can consume many resources because of the compression and deduplication tasks. To avoid a potential performance impact on the existing workload, try to limit the migration throughput at the host level or, if this is not possible, 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 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 and at the same time, change 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 wanted 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 DRP thin-provisioned or compressed volumes.

4.4 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 type.

4.4.1 Creating data reduction pools

This section describes how to create DRPs.

Using the management GUI

To create DRPs by using the management GUI, complete the following steps:

1. Create a DRP, as shown in Figure 4-8:
 - a. In the management GUI, select **Pools** → **Pools**.
 - b. On the **Pools** page, click **Create**.
 - c. On the **Create Pool** page, enter a name for the pool and select **Data Reduction**.
 - d. Click **Create**.

The screenshot shows a 'Create Pool' dialog box with the following fields and options:

- Pool Name:** Pool0
- Extent Size:** 4.00 GiB
- Maximum Capacity:** 512.00 TiB
- Additional Options:** Data Reduction
- Warning:** If the usable capacity usage of a data reduction pool exceeds more than 85%, I/O performance can be affected. The system needs 15% of usable capacity available in data reduction pools to ensure that capacity reclamation can be performed efficiently.
- Buttons:** Cancel, Create
- Link:** Need Help

Figure 4-8 Create pool page

2. Create a Data Reduction child pool, as shown in Figure 4-10 on page 138:
 - a. In the management GUI, select **Pools** → **Pools**.
 - b. Right-click on the parent pool you want to create the child pool in, as shown in Figure 4-9 on page 138.

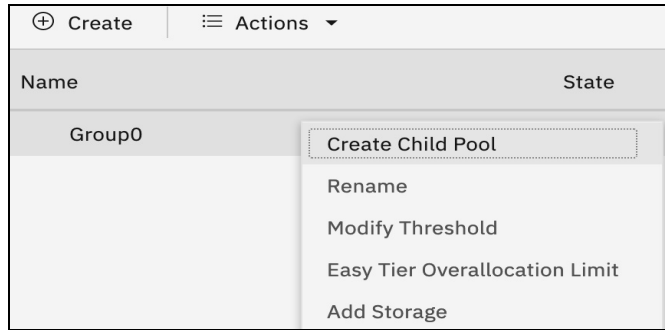


Figure 4-9 Right-click parent pool actions menu

- c. Select **Create Child Pool**.
- d. Enter a name for the child pool, as shown in Figure 4-10.
- e. Click **Create**.

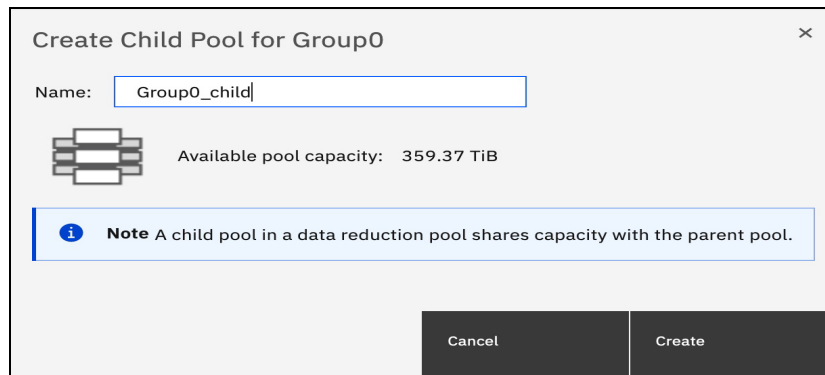


Figure 4-10 Create child pool page

3. Add storage to a parent DRP by completing these steps:
 - a. In the management GUI, select **Pools** → **Pools**.
 - b. Right-click the DRP that you created and select **Add Storage**.
 - c. Select from the available storage and allocate capacity to the pool. Click **Assign**.
4. Create fully-allocated, compressed, deduplicated, or a combination of compressed and deduplicated volumes in the DRP and map them to hosts by completing the following steps:
 - a. In the management GUI, select **Volumes** → **Volumes**.
 - b. On the **Volumes** page, click **Create Volumes**.
 - c. On the **Create Volume** page, select the type of volume that you want to create.
 - d. Enter the following information for the volume:
 - **Pool**
Select a DRP from the list. Compressed, thin-provisioned, and deduplicated volumes, and copies, must be in DRPs.
 - **Volume details**
Enter the quantity, capacity, and name for the volume or volumes that you are creating.

- **Capacity savings**

Select either **None** (fully-allocated), or **Compressed**. When compressed is selected, you can also select to use deduplication for the volume that you create.

Note: If your system contains self-compressed drives, ensure that the volume is created with Compression enabled. If not, the system cannot calculate accurate available physical capacity.

e. Click **Create and Map**, as shown in Figure 4-11.

Figure 4-11 Create Volume page

Note: Select **Create** to create the volumes in the DRP without mapping to hosts. If you want to map volumes to hosts later, select **Hosts** → **Hosts** → **Add Hosts**.

- f. On the **Create Mapping** page, select **Host** to display all hosts that are available for mapping. Hosts must support SCSI **unmap** commands. Verify that the selected host type supports SCSI **unmap** commands. Click **Next**.
- g. Starting with version 8.3.1 the system will try to map the SCSI LUN ID the same on all Host clusters, if you want to assign specific IDs then select the **Self Assign** checkbox.
- h. Verify the volume, and then click **Map Volumes**. See Figure 4-12 on page 140.

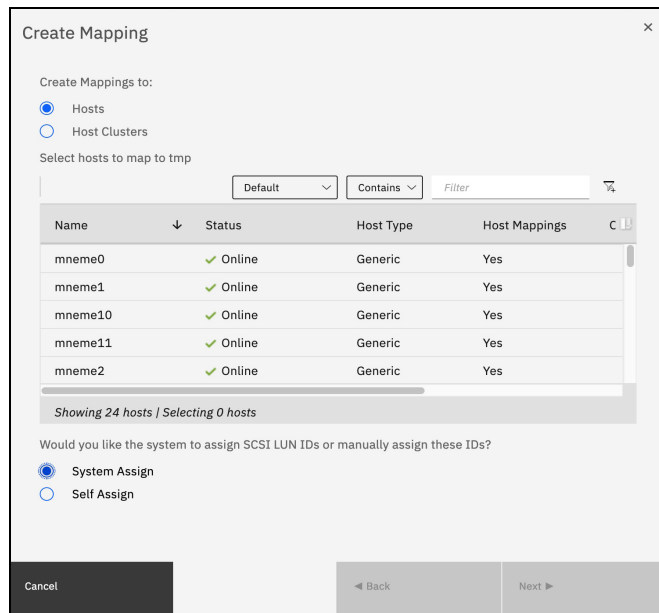


Figure 4-12 The Create Mapping page

Using the command line interface

To create DRPs by using the CLI, complete the following steps:

1. To create a DRP, enter the following command:

```
mkmdiskgrp -name pool_name -ext extent_size -datareduction yes
```

Where *pool_name* is the name of the pool and *extent_size* is the extent size of the pool. You can create DRPs only as parent pools, not child pools.

2. To create a compressed volume within a DRP, enter the following command:

```
mkvolume -name name -pool storage_pool_name -size disk_size -compressed
```

3. Where *name* is the name of the new volume, *storage_pool_name* is the name of the DRP, and *disk_size* is the capacity of the volume.

4. To map the volume to a host, enter the following command:

```
mkvdiskhostmap -host host_name vdisk_name
```

Where *host_name* is the name of the host and *vdisk_name* is the name of the volume.

For more information, see [IBM FlashSystem 9200 8.4.0 Documentation - Creating data reduction pools](#).

Monitor the physical capacity of DRPs in the management GUI by selecting **Pools** → **Pools**. In the command line interface, use the `lsmdiskgrp` command to display the physical capacity of a DRP.

4.4.2 Adding external MDisks to existing storage pools

If MDisks are being added to an IBM FlashSystem cluster, it is likely because you want to provide more capacity. In Easy Tier enabled pools, the storage-pool balancing feature guarantees 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. You must 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.

The tier_type can be changed using the CLI:

```
chmdisk -tier <new_tier> <mdisk>
```

For more information see 4.6.9, “Easy Tier settings” on page 174

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 Spectrum 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 Spectrum 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 preferred 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, 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 them back to the host that is holding the reserve. Alternatively, map them to another host that can remove the reserve by using a utility such as the Microsoft Windows SDD Persistent Reserve Tool.

4.4.3 Renaming MDisks

After you discover MDisks, rename them from their IBM FlashSystem default name. This 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 IBM FlashSystem cluster, 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.

Preferred 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/id>_<volume id>.

4.4.4 Removing MDisks from existing storage pools

You might want to remove MDisks from a storage pool (for example, when you decommission a storage controller). 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.

Sufficient space: 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 on an IBM FlashSystem, 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.

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 IBM FlashSystem to automatically move all used extents on the MDisk to the remaining MDisks in the storage 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 of 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 IBM FlashSystem cluster as unmanaged before their controller LUN mapping is removed. Unmapping LUNs from IBM FlashSystem that are still part of a storage pool results in the storage pool that goes offline and affects all hosts with mappings to volumes in that storage pool.

If the MDisk was named using the preferred 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 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 FlashSystem host object.

Tip: Renaming your back-end storage controllers as recommended also helps you with MDisk identification.

For more information about how to correlate back-end volumes (LUNs) to MDisks, see “Correlating the back-end volume with the MDisk” on page 143.

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 FlashSystem V7000 storage controllers.

DS8000 LUN

The LUN ID only uniquely identifies LUNs within the same storage controller. If multiple storage devices are attached to the same IBM FlashSystem cluster, the LUN ID must be combined with the worldwide node name (WWNN) attribute to uniquely identify LUNs within the IBM FlashSystem cluster.

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 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 the Example 4-4.

Example 4-4 DS8000 UID example

```
6005076305ffc74c00000000000010070000000000000000000000000000000000000000
```

In Example 4-4, you can identify the MDisk 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 on page 144 shows the Volume Properties dialog box that opens.

2. To identify your LUN, in the volumes by Hosts view, expand your IBM FlashSystem host group and then review the LUN column, as shown in Example 4-5.
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, as shown in Example 4-5.

Example 4-5 The lscontroller command

```
IBM_2145:tpcsvc62:admin>lscontroller 10
id 10
controller_name controller10
WWNN 5001738002860000
...
```

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, as shown in Example 4-5.
 - Bits 14 - 16 are the XIV volume serial number (897) in hexadecimal format (resulting in 381 hex).
 - The translation is
001738000286038100,
where:
 - The controller WWNN (bits 2 - 13) is 0017380002860
 - The XIV volume serial number that is converted in hex is 381
5. To correlate the IBM FlashSystem `ctr1_LUN_#`:
 - a. Convert the XIV volume number in hexadecimal format.
 - b. Check the last three bits from the IBM FlashSystem `ctrl_LUN_#`.

In this example, the number is 0000000000000002, as shown in Figure 4-13 on page 145.

FlashSystem volumes

The IBM FlashSystem solution is built upon the IBM Spectrum Virtualize technology base and uses similar terminology.

Complete the following steps to correlate the IBM FlashSystem volumes with the external MDisks that are seen by the virtualizer:

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 4-13 on page 145.

Properties for Volume

Volume Overview Copy 0

Name:	VD_SVC_4	Cache mode:	Enabled
Volume ID:	5	Cache state:	Not empty
State:	✓ Online	UDID (OpenVMS):	N/A
Capacity:	9.70 TiB	Volume UID:	6005076810810026D800000000000008
IOPS limit:	Disabled	I/O group:	Caching: io_grp0
Bandwidth limit:	Disabled		Accessible: io_grp0
Encrypted:	No	Preferred node:	node1
FlashCopy mappings:	0	Protocol type:	SCSI
Mirror sync rate:	50		

Figure 4-13 FlashSystem volume details

- On the Host Maps tab, check the SCSI ID number for the specific volume, as shown in Figure 4-14. This value is used to match the virtualizer ctrl_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 4-14 FlashSystem volume details for host maps

3. At the virtualizer, review the MDisk details and compare the MDisk UID field with the FlashSystem Volume UID, as shown in Figure 4-15. The first 32 bits should be the same.

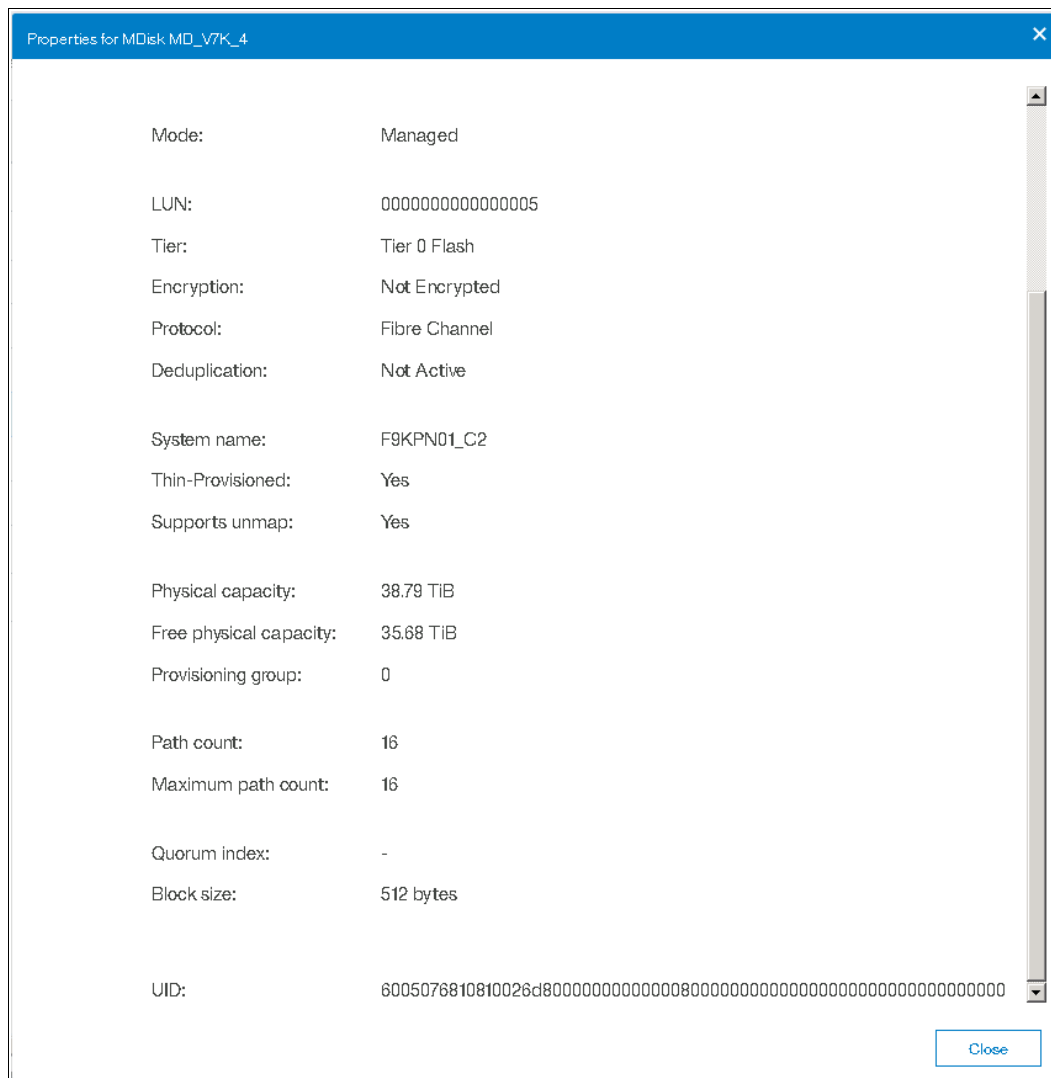


Figure 4-15 IBM SAN Volume Controller MDisk details for IBM FlashSystem volumes

4. Double-check that the virtualizer ctrl_LUN_# is the IBM FlashSystem SCSI ID number in hexadecimal format. In this example, the number is 0000000000000005.

4.4.5 Remapping managed MDisks

Generally, you do not unmap managed external MDisks from IBM FlashSystem because this process causes the storage pool to go offline. However, if managed MDisks were unmapped from IBM FlashSystem for a specific reason, the LUN must present the same attributes to IBM FlashSystem before it is mapped back. Such attributes include UID, subsystem identifier (SSID), and LUN_ID.

If the LUN is mapped back with different attributes, IBM FlashSystem recognizes this MDisk as a new MDisk. In this case, the associated storage pool does *not* come back online. Consider this situation for storage controllers that support LUN selection because selecting a different LUN ID changes the UID. If the LUN was mapped back with a different LUN ID, it must be mapped again by using the previous LUN ID.

4.4.6 Controlling extent allocation order for volume creation

When creating a new volume on a standard pool, the allocation of extents is performed 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, that this allocation pattern is unlikely to remain for long as Easy Tier balancing begins to move the extents to balance the load evenly across all MDisk in the tier. The hot and cold extents are also moved between tiers.

In a multi-tier 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 will be used if it exists. If the cold tier does not exist, the hot tier will be used. For more information on Easy Tier, see 4.6, “Easy Tier, tiered and balanced storage pools” on page 159.

DRP restriction: With compressed and deduplicated volumes on DRP, it is not possible to check the extent distribution across the MDisks. Initially there are a minimal number of extents allocated to the volume, based on the `rsize` parameter.

4.5 Considerations when using encryption

IBM SAN Volume Controller (since 2145-DH8) and all IBM FlashSystem support 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 IBM FlashSystem I/O Group that support encryption.

Note: Check if you have the required IBM Security™ Key Lifecycle Manager licenses at hand. Consider redundancy and high-availability regarding Key Lifecycle Manager servers.

4.5.1 General considerations

USB encryption, key server encryption, or both can be enabled on the system. The system supports IBM Security Key Lifecycle Manager version 2.6.0 or later for enabling encryption with a key server. To encrypt data that is stored on drives, the IBM FlashSystem I/O Groups 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. 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. 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.

Note: Hot Spare Nodes also need encryption licenses if they are to be used to replace the failed nodes that support encryption.

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
- ▶ Availability of key data
- ▶ Encryption license is purchased, activated, and enabled on the system
- ▶ Using Security Key Lifecycle Manager clones

Note: It is suggested that you use IBM Security Key Lifecycle Manager version 2.7.0 or later for new clone end points created on the system.

For configuration details about IBM FlashSystem encryption, see the following publications:

- ▶ *Implementing the IBM FlashSystem 5010 and FlashSystem 5030 with IBM Spectrum Virtualize V8.3.1, SG24-8467*
- ▶ *Implementing the IBM SAN Volume Controller with IBM Spectrum Virtualize V8.3.1, SG24-8465*

4.5.2 Hardware and software encryption

There are two ways to perform encryption IBM FlashSystem devices: hardware encryption and software encryption. Both methods of encryption protect against the potential exposure of sensitive user data that are stored on discarded, lost, or stolen media. Both can also facilitate the warranty return or disposal of hardware. The method to be used for encryption is chosen automatically by the system based on the placement of the data.

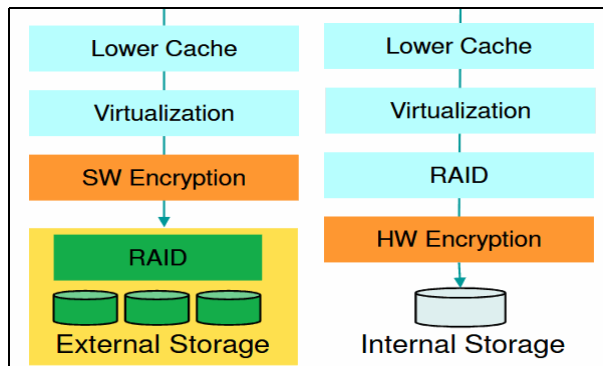


Figure 4-16 Encryption placement in lower layers of the IBM FlashSystem software stack

Hardware encryption only storage pool

Hardware encryption features the following characteristics:

- ▶ Algorithm is built in SAS chip for all SAS attached drives, or built into the drive itself for NVMe attached drives (FCM, IS NVMe and SCM).
- ▶ No system overhead.
- ▶ Only available to direct attached SAS disks.
- ▶ Can only be enabled 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 only use the same encryption key as their parent.

Software encryption only storage pool

Software encryption features the following characteristics:

- ▶ The algorithm is running at the interface device driver.
- ▶ Uses 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.
- ▶ Only available to virtualized external storage.

- ▶ Can only be enabled when you create storage pools and child pools 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 4-17.

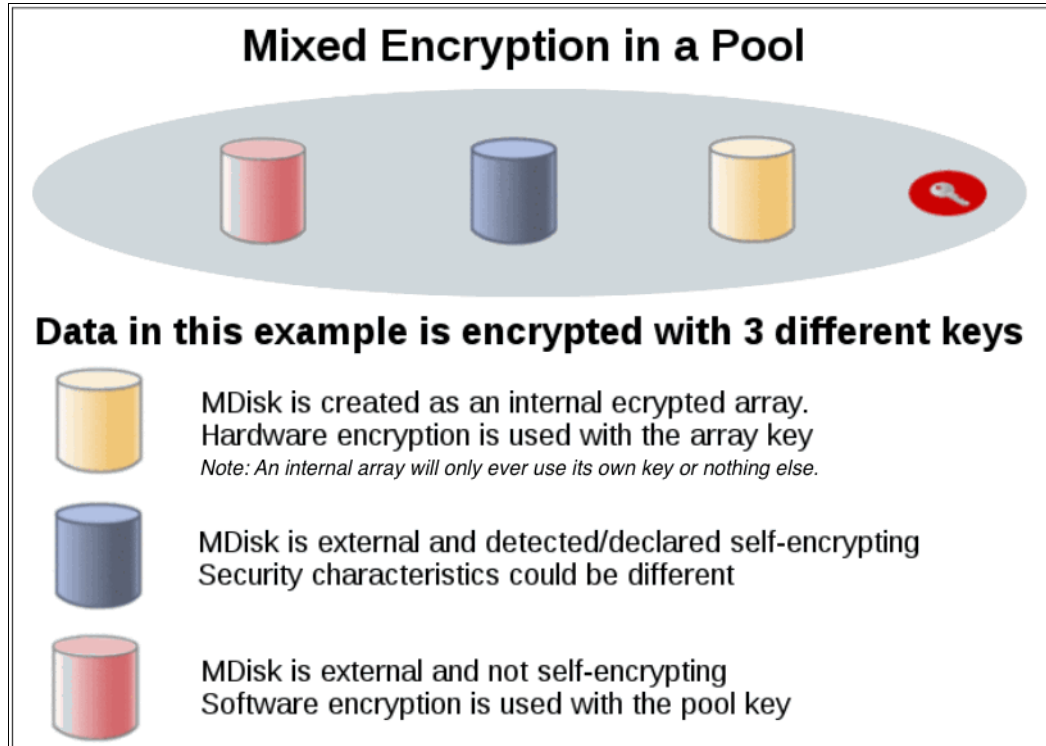


Figure 4-17 Mixed encryption in a storage pool

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 IBM FlashSystem nodes in the system must support software encryption and have an activated encryption license.

Note: An encrypted child pool 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 it is intended 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: Always implement encryption on the self-encryption capable back-end storage, such as IBM FlashSystem, IBM Storwize, IBM XIV, IBM FlashSystem A9000, and IBM DS8000, to avoid potential system overhead.

Declare or identify the self-encrypted virtualized external MDisks as encrypted on IBM FlashSystem by setting the **-encrypt** option to **yes** in the **chmdisk** command, as shown in Example 4-6. This configuration is important to avoid IBM FlashSystem trying to encrypt them again.

Example 4-6 Command to declare or identify a self-encrypted MDisk from a virtualized external storage

```
IBM_2145:ITS0_DH8_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 FlashSystem.

4.5.3 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 operations or 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 at all times.

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, all nodes then 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 in 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.

During operations that the system requires an encryption key to be present, the USB flash drives must be inserted manually into each node so data can be accessed. After the system completes unlocking the drives, the USB flash drives must be removed and stored securely to prevent theft or loss.

4.5.4 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.

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 not to use USB ports.
- ▶ Strong key generation.
- ▶ Key self-replication and automatic backups.
- ▶ Implementations follow an open standard that aids in interoperability.
- ▶ Audit detail.
- ▶ 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 a large number of 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 4 key servers (one master and three clones) that connect to the system over both a public network or a separate private network.

The system supports using an IBM Security Key Lifecycle Manager key server to enable encryption. All key servers must be configured on the IBM Security Key Lifecycle Manager before defining the key servers in the management GUI. IBM Security Key Lifecycle Manager supports Key Management Interoperability Protocol (KMIP), which is a standard for encryption of stored data and management of cryptographic keys.

IBM Security Key Lifecycle Manager can be used to create managed keys for the system and provide access to these keys through a certificate. If you are configuring multiple key servers, use IBM Security Key Lifecycle Manager 2.6.0.2 or later. The additional key servers (clones) support more paths when delivering keys to the system; however, during rekeying only the path to the primary key server is used. When the system is rekeyed, secondary key servers are unavailable until the primary has replicated the new keys to these secondary key servers.

Replication must complete before keys can be used on the system. You can either schedule automatic replication or complete it manually with IBM Security Key Lifecycle Manager. During replication, key servers are not available to distribute keys or accept new keys. The time a replication completes on the IBM Security Key Lifecycle Manager depends on the number of key servers that are configured as clones, and the amount of key and certificate information that is being replicated.

The IBM Security Key Lifecycle Manager issues a completion message when the replication completes. Verify that all key servers contain replicated key and certificate information before keys are used on the system.

Recommendations for key server configuration

The following section provides some essential recommendations for key server configuration with IBM FlashSystem.

Transport Layer Security

Define the IBM Security Key Lifecycle Manager to use Transport Layer Security version 2 (TLSv2).

The default setting on IBM Security Key Lifecycle Manager since version 3.0.1 is TLSv1.2, but the IBM FlashSystem only supports version 2. On the IBM Security Key Lifecycle Manager, set the value to SSL_TLSv2, which is a set of protocols that includes TLSv1.2.

For more information on the protocols, see [IBM SDK, Java Technology Edition Documentation - Protocols](#).

Example 4-7 shows the example of a SKLMConfig.properties configuration file. The default path on a Linux based server is /opt/IBM/WebSphere/AppServer/products/sklm/config/SKLMConfig.properties.

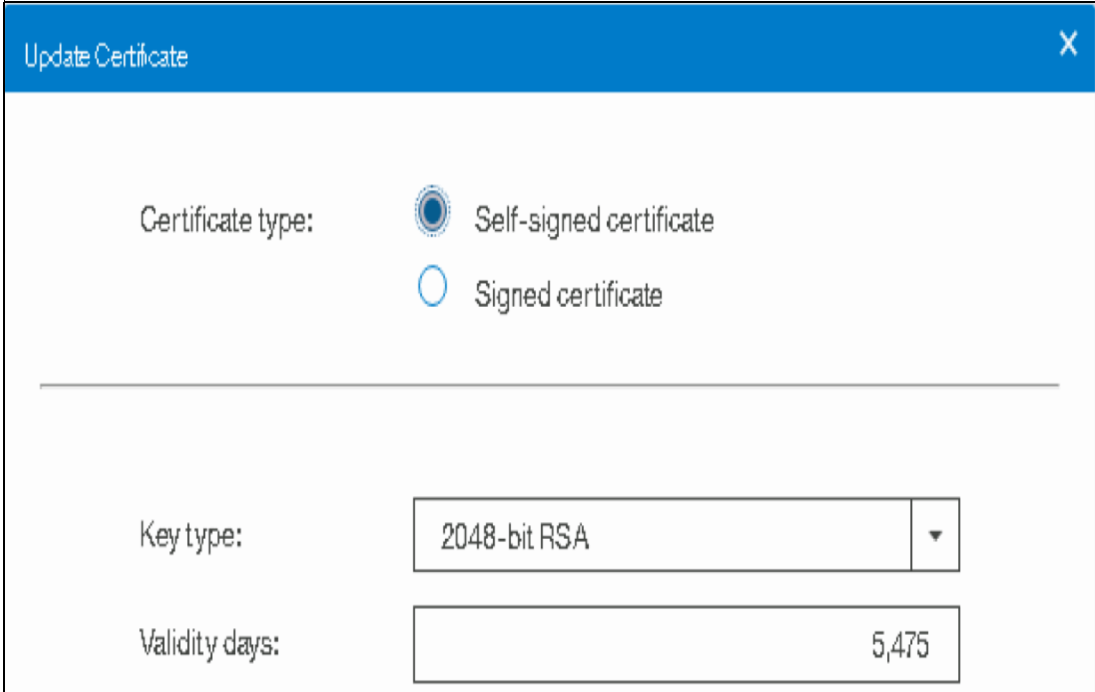
Example 4-7 Example of a SKLMConfig.properties configuration file

```
#Mon Nov 20 18:37:01 EST 2017
KMIPListener.ssl.port=5696
Audit.isSyslog=false
Audit.syslog.server.host=
TransportListener.ssl.timeout=10
Audit.handler.file.size=10000
user.gui.init.config=true
config.keystore.name=defaultKeyStore
tklm.encryption.password=D1181E14054B1E1526491F152A4A1F3B16491E3B160520151206
Audit.event.types=runtime,authorization,authentication,authorization_terminate,resource_management,key_management
tklm.lockout.enable=true
enableKeyRelease=false
TransportListener.tcp.port=3801
Audit.handler.file.name=logs/audit/sklm_audit.log
config.keystore.batchUpdateTimer=60000
Audit.eventQueue.max=0
enableClientCertPush=true
debug=none
tklm.encryption.keysize=256
TransportListener.tcp.timeout=10
backup.keycert.before.serving=false
TransportListener.ssl.protocols=SSL_TLSv2
Audit.syslog.isSSL=false
cert.validate=false
config.keystore.batchUpdateSize=10000
useSKIDefaultLabels=false
maximum.keycert.expiration.period.in.years=50
config.keystore.ssl.certalias=sklm
TransportListener.ssl.port=441
Transport.ssl.vulnerableciphers.patterns=_RC4_,RSA_EXPORT,_DES_
Audit.syslog.server.port=
tklm.lockout.attempts=3
fips=off
Audit.event.outcome=failure
```

Self-signed certificate type and validity period

The default certificate type on IBM Security Key Lifecycle Manager server and IBM FlashSystem is RSA. If you use different certificate type, make sure you match the certificate type on both end. The default certificate validity period is 1095 days on IBM Security Key Lifecycle Manager server and 5475 days on IBM FlashSystem.

You can adjust the validity period to comply with specific security policies and always match the certificate validity period on IBM FlashSystem and IBM Security Key Lifecycle Manager server. A mismatch will cause certificate authorization error and lead to unnecessary certificate exchange. Figure 4-18 shows the default certificate type and validity period on IBM FlashSystem.



The screenshot shows a dialog box titled "Update Certificate" with a close button (X) in the top right corner. The dialog contains the following fields:

- Certificate type:** Two radio button options are shown. The first is "Self-signed certificate" with a selected radio button. The second is "Signed certificate" with an unselected radio button.
- Key type:** A dropdown menu showing "2048-bit RSA" with a downward arrow on the right side.
- Validity days:** A text input field containing the value "5,475".

Figure 4-18 Update certificate on IBM FlashSystem

Figure 4-19 shows the default certificate type and validity period on 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 4-19 Create self-signed certificate on IBM Security Key Lifecycle Manager server

Device group configuration

The SPECTRUM_VIRT device group is not pre-defined on IBM Security Key Lifecycle Manager, it must be created based on a GPFS device family as shown in Figure 4-20.

Create Device Group

*Device family:

- Many asymmetric keys to many devices (3592)
- Many devices to many keys with access via certificate (GPFS)
- Many symmetric keys to many devices (LTO)
- Symmetric Keys directly tied to a single device (DS5000) Enable machine affinity

*Device group name:

Figure 4-20 Create device group for IBM FlashSystem

By default, IBM FlashSystem the SPECTRUM_VIRT group name is predefined in the encryption configuration wizard. SPECTRUM_VIRT contains all the keys for the managed IBM FlashSystem. However, It is possible to use different device groups as long as they are GPFS device family based. For example, one device group for each environment (Production or Disaster Recover (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 Key Lifecycle Manager is one hour, as shown in Figure 4-21. 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 new configuration on the IBM Security Key Lifecycle Manager server, and it is important not to wait for the next replication window but to manually synchronize the configuration to the additional key servers (clones). Otherwise, an error message is generated by the IBM FlashSystem system, which indicates that the key is missing on the clones.

Figure 4-21 shows the replication interval.

The screenshot displays the configuration interface for the replication schedule. It features two tabs: 'Basic Properties' and 'Advance Properties', with the latter being active. Under 'Advance Properties', there are several configuration fields:

- Replication backup destination directory:** A text box containing '/opt/IBM/WebSphere/AppServ' and a 'Browse...' button.
- Maximum number of replication files to keep before rollover:** A spinner box set to '2'.
- Replication Scheduler Section:** A collapsed section that is expanded in the image, containing:
 - Replication frequency (in hours):** A radio button selected next to a spinner box set to '24'.
 - Daily replication time (in HH:MM format):** A radio button selected next to a dropdown menu showing '00:00'.
- Replication Log Section:** A collapsed section that is expanded in the image, containing:
 - Replication log file name:** A text box containing 'replicationMaster.log'.
 - Maximum log file size (in KB):** A spinner box set to '1000'.
 - Maximum number of log files to keep:** A spinner box set to '30'.

Figure 4-21 SKLM Replication Schedule

Example 4-8 shows an example of manually triggered replication.

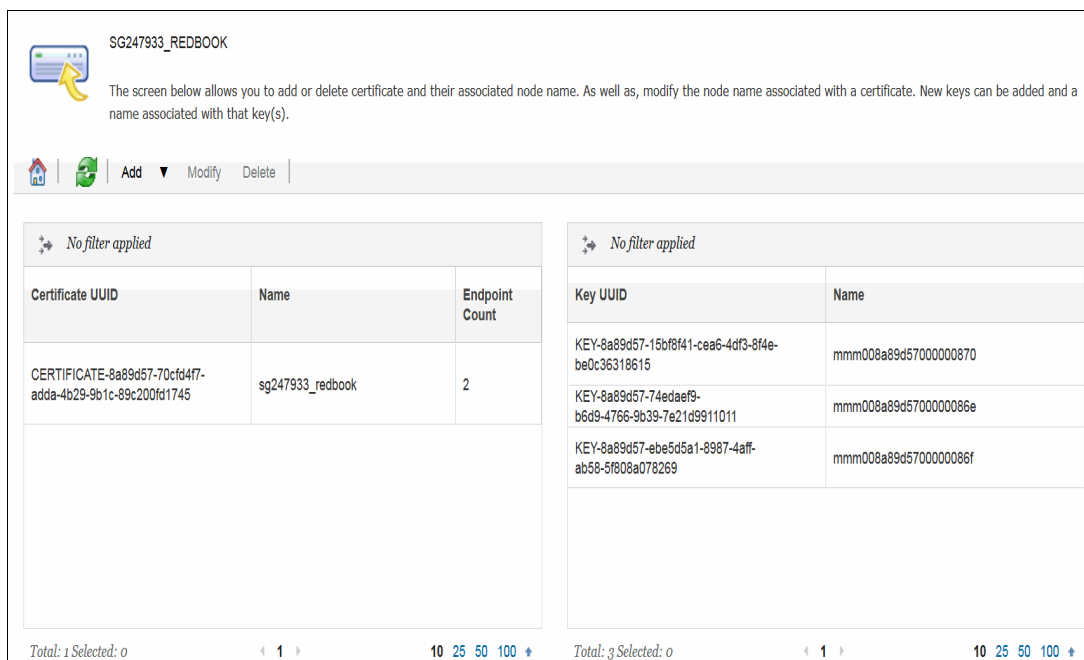
Example 4-8 Manually triggered replication

```
/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 IBM FlashSystem 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 4-22 shows the keys associated with a device group. In this example, the SG247933_REDBOOK device group contains one encryption-enabled IBM FlashSystem, and it has three associated keys. Only one of the keys is activated, and the other two were deactivated after the rekey process.



The screenshot displays the management interface for the device group SG247933_REDBOOK. It includes a navigation bar with 'Add', 'Modify', and 'Delete' options. Two tables are shown, both with 'No filter applied'.

Certificate UUID	Name	Endpoint Count
CERTIFICATE-8a89d57-70cfd4f7-adda-4b29-9b1c-89c200fd1745	sg247933_redbook	2

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

Figure 4-22 Keys associated to a device group

Example 4-9 shows an example to check the state of the keys.

Example 4-9 Verify 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 Transport Layer Security configuration, is only required on the master IBM Security Key Lifecycle Manager server. The restore or replication process duplicates all of the required configurations to the clone servers.

If encryption was enabled on a pre-V7.8.0 code level system and the system is updated to V7.8.x or above, 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 4-10.

Example 4-10 Commands to enable key server encryption option on a system upgraded from pre-7.8.0

```
chencryption -usb newkey -key prepare
chencryption -usb newkey -key commit
```

For more information about Encryption with Key Server, see [IBM Security Guardium Key Lifecycle Manager 4.1.0](#).

4.6 Easy Tier, tiered and balanced storage pools

Easy Tier was originally developed to provide the maximum performance benefit from a few SSDs or flash drives. Because of their low response times, high throughput, and IOPS-energy-efficient characteristics, SSDs and flash arrays were a welcome addition to the storage system, but initially their acquisition cost per Gigabyte (GB) was more than for HDDs.

By implementing an evolving almost 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 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 15 K RPM and 10 K 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 “flash and trash” approaches, with 50% 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. 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, Storage Class Memory (SCM) technologies, which were introduced to FlashSystem in 2020, now provide lower latency than even flash, but as with flash when first introduced, at a considerably higher cost of acquisition per GB.

Choosing the correct mix of drives and the 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 also can be enabled for internal and external storage to achieve optimal performance.

Also, the Easy Tier feature that is called storage pool balancing (introduced in V7.3) automatically moves extents within the same storage tier from overloaded to less loaded managed disks (MDisks). Storage pool balancing ensures that your data is optimally placed among all disks within storage pools.

4.6.1 Easy Tier concepts

IBM FlashSystem products implement Easy Tier enterprise storage functions, which were originally designed in conjunction with the development of Easy Tier on IBM DS8000 enterprise class storage systems. It enables automated subvolume 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, flash 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. Because Easy Tier monitors I/O performance from the device end (after cache), it can pick up the performance issues that cache cannot solve, and complement the overall storage system performance.

The primary benefit of Easy Tier is to reduce latency for hot spots; however, this benefit also 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). In addition,

Easy Tier can be used in a single tier pool to balance the workload across storage MDisks and 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 4-23) or the same storage tier (see Figure 4-25 on page 167). 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 is 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 is balanced between MDisks in the same tier.

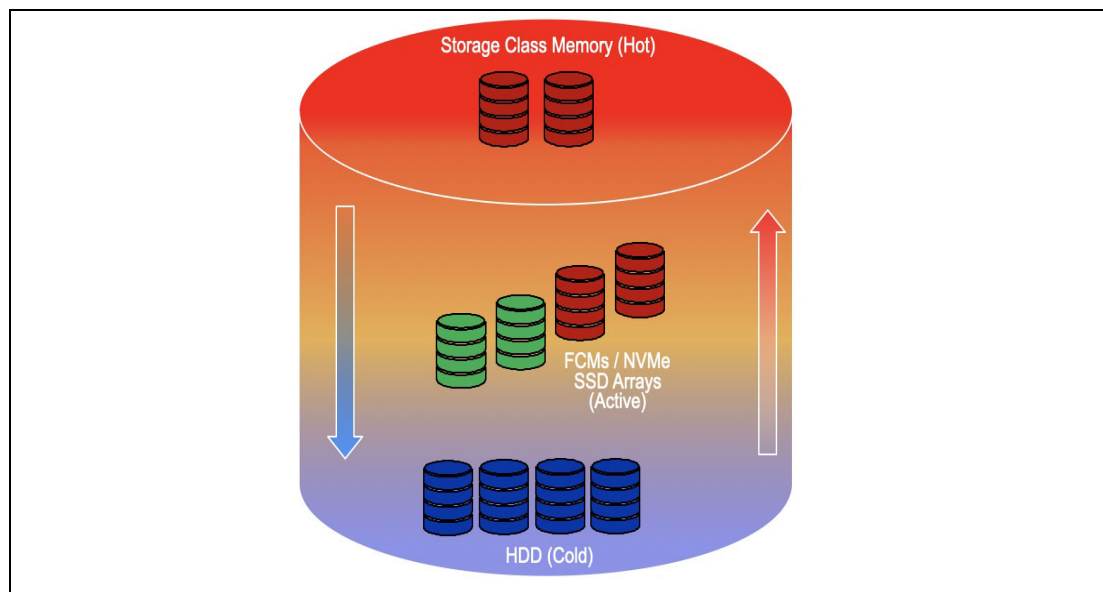


Figure 4-23 Easy Tier single volume, multiple tiers

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 MDisk back to a lower tier MDisk. When Easy Tier runs in a storage pool rebalance mode, it moves extents from busy MDisk to less busy MDisk of the same type.

Note: Image mode and sequential volumes are not candidates for Easy Tier automatic data placement since all extents for those types of volumes must be on one specific MDisk, and cannot be moved.

4.6.2 Easy Tier definitions

Easy Tier measures and classifies each extent into one of its three tiers. It performs this classification process by looking for the 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 come from MDisk 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 MDisk that are designated as the cold tier. 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 MDisk 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 adds negligible processing resource use function to the IBM FlashSystem nodes.

► Data Placement Advisor (DPA)

The DPA uses workload statistics to make a cost-benefit decision as to which extents are to be candidates for migration to a higher performance tier.

This process also 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 five minutes is migrated, which equates to approximately 3.4TiB per day that is migrated between disk tiers.

Note: You can increase the target migration rate to 48 GiB every five minutes by temporarily enabling accelerated mode. See “Easy Tier acceleration” on page 175.

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 cold extent in an upper tier with 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 bandwidth or IOPS exceeds predefined threshold. If you see these operations, it is a trigger to suggest you should add additional capacity to the higher tier.

▶ **Warm promote**

Introduced with version 7.8, this feature addresses the situation where a lower tier suddenly becomes very 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 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. In that way 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 used MDisks to low used MDisks, and exchanges extents between high-use MDisks and low-use MDisks.

Easy Tier attempts to migrate the most active volume extents up to 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 instance, in a three-tiered storage pool, Easy Tier will not move extents from the flash tier directly to the nearline tier and vice versa without moving them first to the enterprise tier.

Easy Tier extent migration types are shown in Figure 4-24.

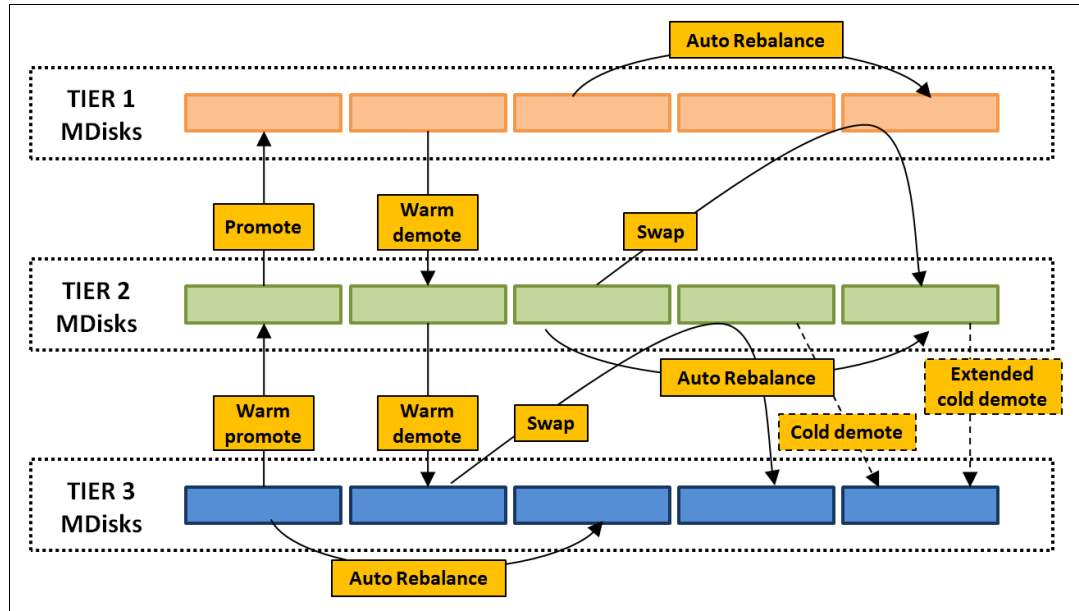


Figure 4-24 Easy Tier extent migration types

4.6.3 Easy Tier operating modes

Easy Tier includes the following main operating modes:

- ▶ Off
- ▶ On
- ▶ Automatic
- ▶ Measure

Easy Tier is a licensed feature on some FlashSystem 50x0 systems. 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 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 additional 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 under the **Monitoring --> Easy Tier Reports**. See 4.6.10, “Monitoring Easy Tier using the GUI” on page 180

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.

- ▶ If the pool only contains 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 173.

There might be situations where the Easy Tier setting is “auto” however the system is running in monitoring mode only. For example, with unsupported tier types or if you have not enabled the Easy Tier license. See Table 4-9 on page 169

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 triggered by Easy Tier. These migrations are reported in terms of the migration types, as described in “Internal processing” on page 161.

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 169, using On mode will force Easy Tier to the active state and it will perform to the best of its ability. The system raises an alert and 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 “buckets”, you might end up with Easy Tier running only in measure mode.

These configurations are unsupported because they can cause a performance problem in the longer term; for example, disparate performance within a single tier.

The ability to override the automatic mode is provided to enable temporary migration from an older set of tiers to new tiers and 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 given tier.

Balancing is when you maintain equivalent latency across all MDisks in a given tier, this 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 only contains a single type of drive.

Balancing is automatically active on all storage pools, no matter the Easy Tier setting. For a single tier pool the Easy Tier state will report as balancing.

Note: Storage pool balancing can be used to balance extents when mixing different size 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 take advantage of the additional 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 shows possible combinations of Easy Tier setting.

Table 4-7 Easy Tier settings

Storage pool Easy Tier setting	Number of tiers in the storage pool	Volume copy Easy Tier setting	Volume copy Easy Tier status
Off	One or more	off	inactive (see note 2)
		on	inactive (see note 2)
Measure	One or More	off	measured (see note 3)
		on	measured (see note 3)
Auto	One	off	measured (see note 3)
		on	balanced (see note 4)
	Two - four	off	measured (see note 3)
		on	active (see note 5 & 6)
	Five	any	measured (see note 3)
	On	One	off
on			balanced (see note 4)
Two - four		off	measured (see note 3)
		on	active (see note 5)
Five		off	measured (see note 3)
		on	active (see note 6)

Table notes:

1. If the volume copy is in image or sequential mode, or is being migrated, the volume copy Easy Tier status is measured rather than active.
2. When the volume copy status is inactive, no Easy Tier functions are enabled for that volume copy.
3. When the volume copy status is measured, the Easy Tier function collects usage statistics for the volume, but automatic data placement is not active.
4. When the volume copy status is balanced, the Easy Tier function enables performance-based pool balancing for that volume copy.
5. When the volume copy status is active, the Easy Tier function operates in automatic data placement mode for that volume.
6. When five tiers (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. See Table 4-9 on page 169

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 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.

4.6.4 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 used by internal or external MDisks defines their expected performance characteristics. These characteristics are used to help define a *tier type* for each MDisk in the system.

Five tier types that can be assigned. The tables in this section use the numbers from this list as a shorthand for the tier name:

1. *tier_scm* that represents Storage Class Memory MDisks
2. *tier0_flash* that represents enterprise flash technology, including FCM
3. *tier1_flash* that represents lower performing tier1 flash technology (lower DWPD)
4. *tier_enterprise* that represents enterprise HDD technology (both 10 K and 15 K RPM)
5. *tier_nearline* that represents nearline HDD technology (7.2 K RPM)

Consider the following points:

- ▶ Easy Tier is designed to operate with up to 3 tiers of storage, “hot”, “medium”, “cold”
- ▶ An MDisk can only belong to one tier type.
- ▶ Today, five MDisk tier-types exist.
- ▶ Internal MDisks have their tier type set automatically.
- ▶ External MDisks default to the “enterprise” tier and may need to be changed by the user.
- ▶ The number of MDisk tier-types found in a pool will determine if the pool is a single-tier pool or a multi-tier pool.

Attention: As mentioned in 4.6.5, “Changing the tier type of an MDisk” on page 170, IBM FlashSystem do 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 4-25 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 can benefit from *storage pool balancing* mode, which moves extents between MDisks of the same type.

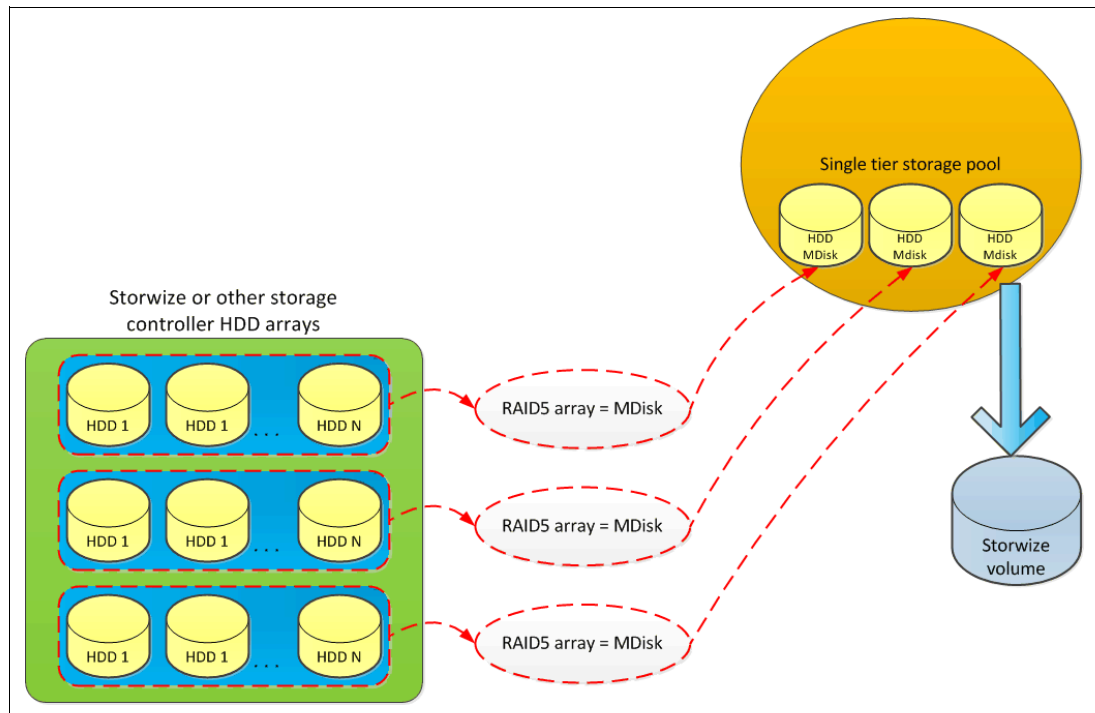


Figure 4-25 Single tier storage pool with striped volume

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 is not possible, you can modify the Easy Tier load setting to manually balance the workload; however, care must be taken. For more information, see “MDisk Easy Tier load” on page 175.

For internal MDisks, the system can cope with different geometries as the number of drives will be reported to Easy Tier, which then uses the Overload Protection information to balance the workload appropriately. See 4.6.6, “Easy Tier overload protection” on page 172.

Multi-tier storage pools

A multi-tier 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 4-26 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 4-26 shows RAID 5 arrays, other RAID types can also be used.

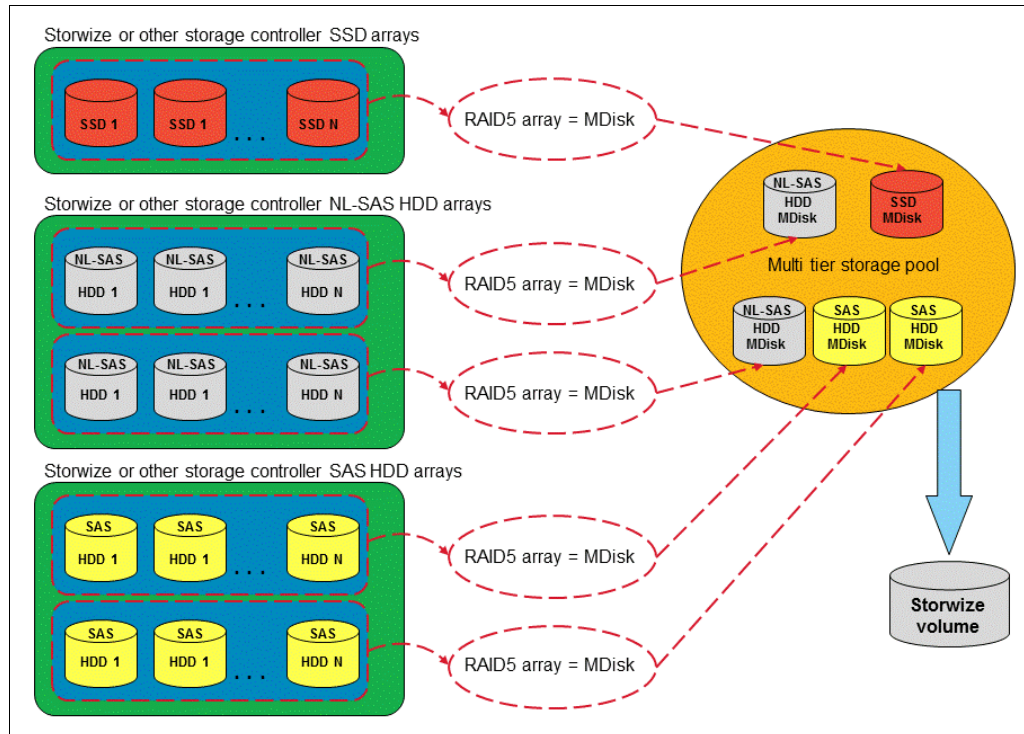


Figure 4-26 Multitier storage pool with striped volume

Note: If you add MDisks to a pool and they have (or you assign) more than three tier types, Easy Tier will try to group two or more of the tier types together into a single “bucket” and use them both as either the “middle” or “cold” tier. The groupings are described in table Table 4-9 on page 169

However, overload protection and pool balancing might result in a bias on the load being placed on those MDisks 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

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
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

Four- and Five-Tier pools

In general, Easy Tier will try to place **tier_enterprise (4)**- and **tier1_flash (3)**-based tiers into the one bucket to reduce the number of tiers defined in a pool to 3. See Table 4-9.

Table 4-9 4 and 5 Tier mapping policy4 and 5 Tier mapping policy

Tier Mix	1+2+3+4, 1+2+3+5, 1+2+4+5	1+3+4+5, 2+3+4+5	1+2+3+4+5
Hot Tier	<i>not supported: measure only</i>	1 or 2	<i>not supported: measure only</i>
Middle Tier		3 & 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 4 or 5 tier mapping

If you need to temporarily define four or five 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: Extreme caution should be deployed and a full understanding of the implications should be made before forcing Easy Tier to run in this mode.

This setting is provided to allow temporary migrations where it is unavoidable to create 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, the following mapping in Table 4-10 will be used by Easy Tier.

Table 4-10 Unsupported temporary 4 and 5 Tier mapping policy

Tier Mix	1+2+3+4, 1+2+3+5	1+2+4+5	1+2+3+4+5
Hot Tier	1	1	1

Tier Mix	1+2+3+4, 1+2+3+5	1+2+4+5	1+2+3+4+5
Middle Tier	2 & 3	2	2 & 3
Cold Tier	4 or 5	4 & 5	4 & 5
comment	See Note 1	See Note 2	See Note 1 & 2

Note:

- ▶ **Note 1:** In these configurations, Enterprise HDD and Nearline HDD are placed into the cold tier. These two drive types feature different latency characteristics and the difference can skew the metrics that are measured by Easy Tier for the cold tier.
- ▶ **Note 2:** In these configurations, Tier0 and Tier 1 flash devices are placed in the middle tier. The different drive writes per day does not make the most efficient use of the Tier0 flash.

4.6.5 Changing the tier type of an MDisk

By default, IBM FlashSystem adds external MDisks to a pool with the tier type “enterprise”. This addition is made because it cannot determine the technology type of the MDisk without further information.

Attention: When adding external MDisks to a pool, be sure to validate the **tier_type** setting is correct. Incorrect **tier_type** settings can cause performance problems; for example, if you inadvertently create a multi-tier pool.

IBM FlashSystem internal MDisks should automatically be created with the correct **tier_type** as the IBM FlashSystem 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 subsequently change the tier of an MDisk by using the CLI, use the **chmdisk** command as in Example 4-11.

Example 4-11 Changing 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
6005076802880102c0000000000020000000000000000000000000000000000000000 tier_enterprise
no 2 SITE_B no no
2 mdisk2 online managed 1 POOL_V7K_SITEB 250.0GB 0000000000000002 V7K_SITEB_C2
6005076802880102c000000000002100000000000000000000000000000000000000 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
6005076802880102c000000000002000000000000000000000000000000000000000 tier_nearline no
2 SITE_B no no
```

```
2 mdisk2 online managed 1 POOL_V7K_SITEB 250.0GB 0000000000000002 V7K_SITEB_C2
6005076802880102c0000000000002100000000000000000000000000000000 tier_enterprise
no 2 SITE_B no no
```

It is also possible to change the MDisk tier from the GUI, but this applies only to external MDisks. To change the tier, complete the following steps:

1. Click **Pools** → **External Storage** and click the **Plus** sign (+) next to the controller that owns the MDisks for which you want to change the tier.
2. Right-click the wanted MDisk and select **Modify Tier** (Figure 4-27).

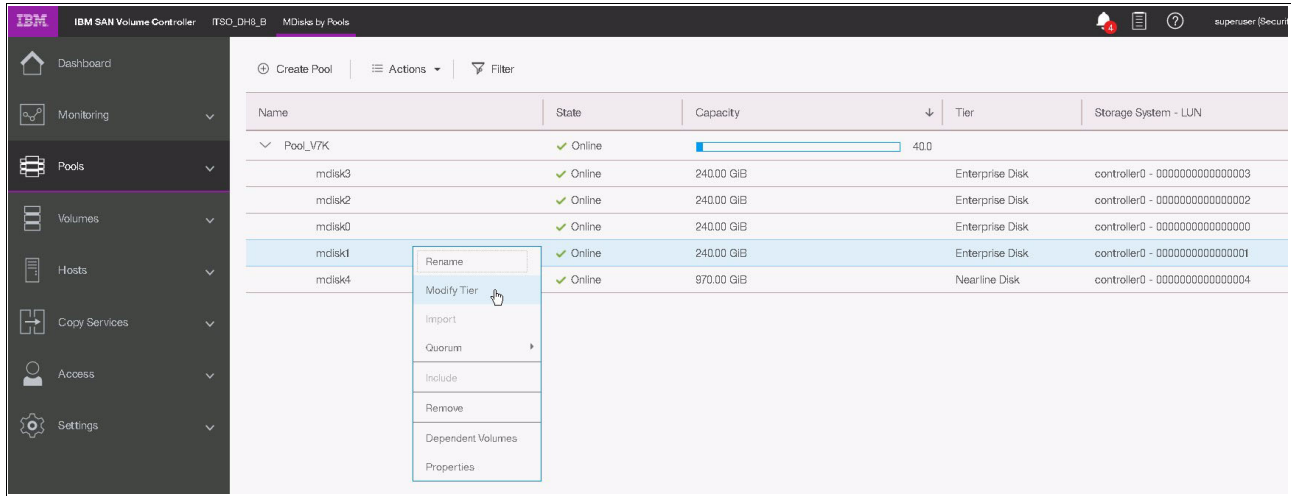


Figure 4-27 Change the MDisk tier

3. The new window opens with options to change the tier (Figure 4-28).

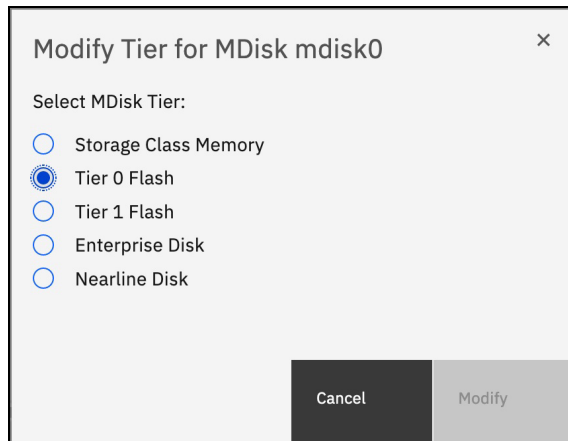


Figure 4-28 Select wanted MDisk tier

This change happens online and has no effect on hosts or availability of the volumes.

- If you do not see the *Tier* column, right-click the blue title row and select the **Tier** check box, as shown in Figure 4-29.

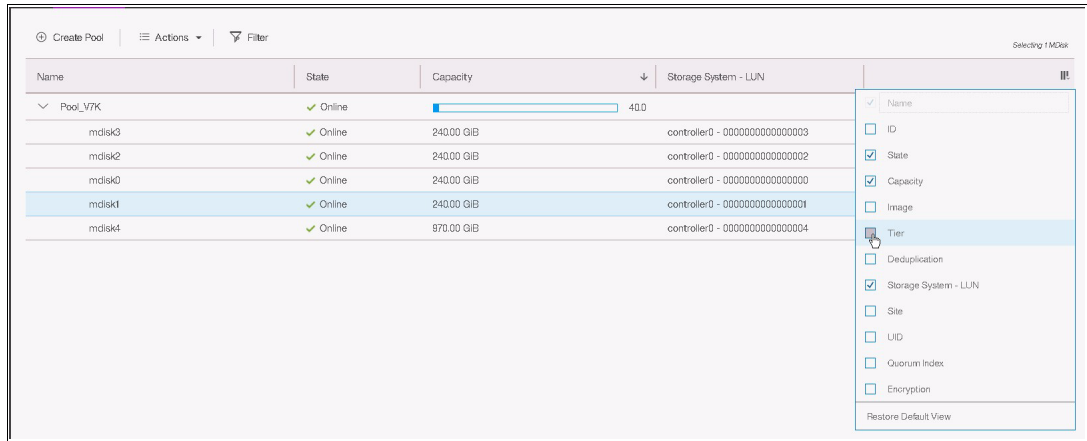


Figure 4-29 Customizing the title row to show the tier column

4.6.6 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; therefore, 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 as to 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 we have is the storage controller type. Therefore, we know if the controller is an Enterprise, Midrange, or Entry level system and can make some assumptions about the load it can handle.

However, external MDisks cannot automatically have their MDisk tier type or “Easy Tier Load” defined. You must set the tier type manually and (if wanted), modify the load setting. For more information about Easy Tier load, see “MDisk Easy Tier load” on page 175.

Overload Protection is also used by the “warm promote” functionality. 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.

4.6.7 Removing an MDisk from an Easy Tier pool

When you remove an MDisk from a pool that still include defined volumes, and that pool is an Easy Tier pool, the extents that are still in use on the MDisk 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 for the typical migration cases, for example replacing the Enterprise HDD tier with Tier1 Flash arrays or replacing Nearline HDD with Tier1 Flash arrays.

4.6.8 Easy Tier implementation considerations

Easy Tier is part of the IBM Spectrum 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 HDD). With single tier (homogeneous) pools, Easy Tier uses storage pool balancing only.

Important: Easy Tier uses the extent migration capabilities of IBM Spectrum Virtualize. These migrations require free capacity, as 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 will 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 consideration: 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 IBM FlashSystem system creates volumes or expands volumes by using extents from MDisk from the HDD tier. However, if necessary, it uses extents from MDisk from the SSD tier.
- ▶ Do not provision the 100% of an 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 also 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 on the IBM FlashSystem system, 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 MDisk in the same tier as the MDisk that is being removed, if possible. If insufficient extents exist in that tier, extents from the other 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 IBM FlashSystem 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 between resumes for the moved volume, if appropriate.

When the system migrates a volume from one storage pool to another, it 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 same priority rules outlined in 4.6.7, “Removing an MDisk from an Easy Tier pool” on page 172.
- ▶ 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.

4.6.9 Easy Tier settings

The Easy Tier setting for storage pools and volumes can only be changed from the CLI. All the changes are done online without any effect on hosts or data availability.

Turning Easy Tier on and off

Use the **chvdisk** command to turn off or turn on Easy Tier on selected volumes. Use the **chmdiskgrp** command to change status of Easy Tier on selected storage pools, as shown in Example 4-12 on page 175.

Example 4-12 Changing Easy Tier setting

```
IBM_FlashSystem:V7000 Gen 2:superuser>chvdisk -easytier on test_vol_2  
IBM_FlashSystem:V7000 Gen 2: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 is a system-wide setting, and is disabled by default. Turning on this setting makes Easy Tier move extents up to four times faster than when in default setting. In accelerate mode, Easy Tier can move up to 48 GiB every five 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:

- ▶ When installing a new system, accelerating Easy Tier will quickly reach a steady state and reduce the time needed to reach an optimal configuration. This applies to single-tier and multi-tier pools alike. In a single-tier pool this will allow balancing to spread the workload quickly and in a multi-tier pool it will allow 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 via pool balancing. It can also help if you added more capacity to stop warm demote operations. In this case, Easy Tier knows that certain extents are hot and were only demoted 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 any effect on host or data availability. To turn Easy Tier acceleration mode on or off, run the following command:

```
chsystem -easytieracceleration <on/off>
```

Important: Do not leave accelerated mode on indefinitely. It is recommended 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 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 indicates how much load Easy Tier can put on that particular 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 MDisk (array) because the system is able to determine the exact load that an internal MDisk can handle, based on the drive technology type, 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 assigned to the MDisk. These

performance profiles are generic, which means that they do not take into account the actual backend configuration. For instance, 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 concurrent I/O and expected latency from the MDisk. Setting this value incorrectly, or 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 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 particular 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 SDD and Flash MDisks.

This setting can be changed online, without any effect on the hosts or data availability.

To change this setting, run the following command:

```
chmdisk -easytierload high mdisk0
```

Important: Consider the following points:

- ▶ When IBM SAN Volume Controller is used with FlashSystem backend storage, it is recommended to set the Easy Tier load to “very high” for FlashSystem MDisks other than FlashSystem 50x0 where the default is recommended.

The same would be recommended for modern high-performance all-flash storage controllers from other vendors.

- ▶ After changing the load setting, make a note of the old and new settings and record the date and time of the change. Use 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 at each change that you are seeing an increase in throughput without a corresponding detrimental increase in latency (and vice versa if you are decreasing the load setting).

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 therefore the same temperature. So, moving a hot extent will probably also move data that is not actually hot. The overall Easy Tier efficiency to put hot data in the proper tier is then inversely proportional to the extent size.

Consider the following practical aspects:

- ▶ Easy Tier efficiency is affecting 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 Easy Tier background process. The smaller the extent size, the lower that the bandwidth consumption is.

However, Easy Tier efficiency should not be the only factor considered when choosing the extent size. Manageability and capacity requirement considerations must also be taken into account.

As a general rule, use the default 1GB (standard pool) or 4GB (DRP) extent size for Easy Tier enabled configurations.

External controller tiering considerations

IBM Easy Tier is an algorithm that has been developed by IBM Almaden Research and made available to many members of the IBM storage family, such as the DS8000, IBM SAN Volume Controller, and FlashSystem products. The DS8000 is the most advanced in Easy Tier implementation and currently 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 is recommended; that is, the virtualizer and any backend controller tiering functions should be disabled.

Important: Never run tiering at two levels. This will cause thrashing and unexpected heat/cold jumps to be seen at both levels.

Consider the following two options:

- ▶ **Easy Tier is done at the virtualizer level:**
 - a. In this case, complete these steps at the backend level:
 - i. Set up homogeneous pools according to the tier technology available.
 - ii. Create volumes to present to the virtualizer from the homogeneous pool.
 - iii. Disable tiering functions.
 - b. At a virtualizer level, you need to complete the following actions:
 - i. Discover the MDisks provided by the backend storage and set the tier properly.
 - ii. Create hybrid pools that aggregate the MDisks.
 - iii. Enable the Easy Tier function.
- ▶ **Easy Tier is done at the backend level.**
 - a. In this case, complete these actions at the back-end level:
 - i. Set up hybrid pools according to the tier technology available.
 - ii. Create volumes to present to the virtualizer from the hybrid pools.
 - iii. Enable the tiering functions.

- b. At virtualizer level, you need to complete the following actions:
 - i. Discover the MDisk provided by the backend storage and set the same tier for all.
 - ii. Create standard pools that aggregate the MDisk.
 - iii. Disable the Easy Tier function.

Even though both of these options provide benefits in terms of performance, they have different characteristics.

Option 1 provides some advantages when compared to option 2, including:

- ▶ With option 1, Easy Tier can be enabled or disabled at volume level. This feature allows users to decide which volumes will benefit from Easy Tier and which will not. With option 2, this goal cannot be achieved.
- ▶ With option 1, the volume heat map matches directly to the host workload profile using the volumes. This option also allows you to use Easy Tier across different storage controllers, so using lower performance and cost systems to implement the middle or cold tiers.

With option 2, the volume heat map on the backend storage is based on the IBM FlashSystem workload. It therefore does not exactly represent the host workload profile because of the effects of the IBM FlashSystem caching.
- ▶ With option 1, you have the chance to change the extent size to improve the overall Easy Tier efficiency (as described in “Extent size considerations” on page 176).

Option 2, especially with DS8000 as the backend, offers some advantages when compared to option 1. For example, when using external storage, 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 175. These profiles might not exactly match the actual backend capabilities, which can lead to a resource utilization that is not optimized. With option 2, this problem rarely happens because the performance profiles are based on the real back-end configuration.

Easy Tier and thin-provisioned backend considerations

When using a data reduction-capable backend in Easy Tier-enabled pools, it is important to note that the data-reduction ratio on the physical backend might vary over time because of Easy Tier data-moving. Easy Tier continuously moves extents across the tiers, as well as within the same tier, trying to optimize performance. As a result, the amount of data written to the backend, and therefore the compression ratio, can unpredictably fluctuate over time even though the data itself is not modified by the user.

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 the secondary system can differ. Easy Tier at the primary system sees a normal workload, and at the secondary system, it sees only the write workloads.

This situation means that the optimized extent distribution on the primary system can differ considerably from the one 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 at this time 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.

It 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 175.

IBM FlashSystem Remote Copy configurations that use NearLine tier at the secondary system must be carefully planned, especially when practicing disaster recovery using FlashCopy. In these scenarios, FlashCopy is usually started just before the beginning of the disaster recovery test. It is likely that the FlashCopy target volumes are in the NearLine tier because of prolonged inactivity.

As soon as the FlashCopy is initiated, an intensive workload is usually added to the FlashCopy target volumes due to 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 of Log Structured Array (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 111.

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 collected. This process provides the following advantages:

- ▶ Writes to a DRP volume are always treated as sequential: so all the 8 KB chunks can be built into a larger 256 KB chunk and destage the writes from cache, either 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 we 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 not only read data, but write data, when data reduction is in use.

Before DRP, write operations to compressed volumes held lower value to the Easy Tier algorithms because writes were always to a new extent; therefore, 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 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 configurations that address some or most common customer requirements. The same benefits can be achieved by adding Storage Class Memory 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 be shifted down to drop the coldest tier.

- ▶ 10-20% Flash, 80-90% Enterprise
This configuration provides Flash-like performance with reduced costs.
- ▶ 5% Tier 0 Flash, 15% Tier 1 Flash, 80% Nearline
This configuration provides Flash-like performance with reduced costs.
- ▶ 3-5% Flash, 95-97% Enterprise
This configuration provides improved performance compared to a single tier solution, and all data is guaranteed to have at least enterprise performance. It also removes the requirement for over provisioning for high access density environments.
- ▶ 3-5% Flash, 25-50% Enterprise, 40-70% Nearline
This configuration provides improved performance and density compared to a single tier solution. It also provides significant reduction in environmental costs.
- ▶ 20-50% Enterprise, 50-80% Nearline
This configuration provides reduced costs and comparable performance to a single-tier Enterprise solution.

4.6.10 Monitoring Easy Tier using the GUI

Since software version 8.3.1, the GUI includes various reports and statistical analysis that can be used to understand which Easy Tier movement, activity, and skew is present in a storage pool. These panels 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 to the STAT tool, from version 8.3.1 onwards, the system continually reports the Easy Tier information. Therefore, the GUI always displays the most up-to-date information.

Accessing Easy Tier reports

In the GUI, select **Monitoring** → **Easy Tier Reports** to show the Easy Tier Reports page.

If the system or Easy Tier has been running for less than 24 hours there might not be any data to display.

The reports page has three views, which can be accessed via the tabs at the top of the page.

- ▶ Data Movement
- ▶ Tier Composition
- ▶ Workload Skew

Data movement report

The data movement report shows the amount of data that has been moved in a given time period. You can change the time period using the drop-down selection on the right side. See Figure 4-30 on page 181.



Figure 4-30 Easy Tier Data Movement page

The report breaks down the type of movement and these are described in terms of the internal Easy Tier extent movement types, as detailed in 4.6.2, “Easy Tier definitions” on page 161.

To aid your understanding and remind you of the definitions, click the **Movement Description** button to view the information panel as shown in Figure 4-31 on page 182.

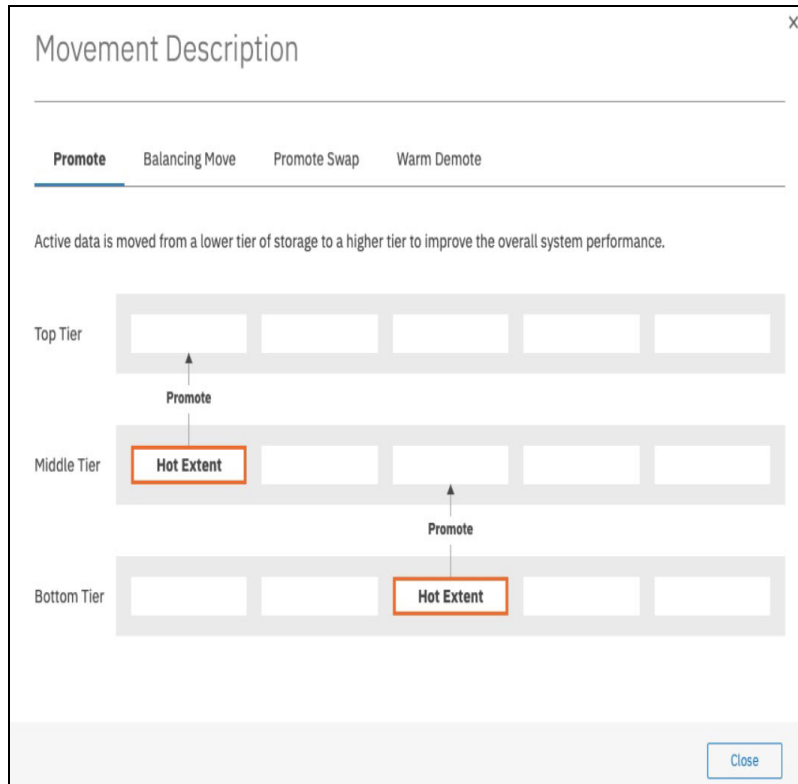


Figure 4-31 Easy Tier Movement description page

Important: If you regularly see *warm demote* in the movement data, you should consider increasing the amount of hot tier that you have. A warm demote suggests that an extent is hot, but there is either not enough capacity or Overload Protection has been triggered in the hot tier.

Tier composition report

The tier composition window shows how much data in each tier is active vs inactive, as shown in Figure 4-32 on page 183. In an ideal case the majority of your active data should reside in the hot tier alone. In most cases the active data set will not be able to fit in only the hot tier. Therefore, you would expect to see active data in the middle tier also.

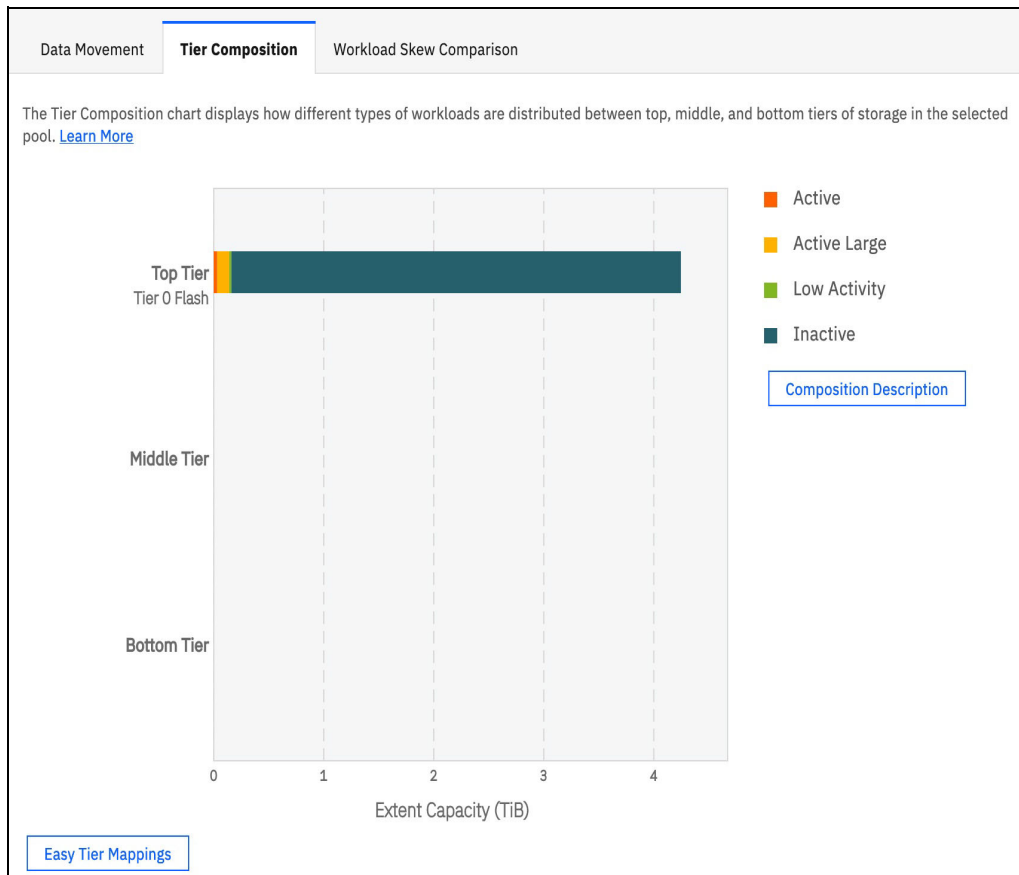


Figure 4-32 Easy Tier - single tier pool - composition report page

If all active data can fit in the hot tier alone, 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 84 KiB for which Easy Tier is optimized. This data is still monitored and can contribute to “expanded cold demote” operations.

The presence of 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 movement page, you can click the **Composition Description** to view the information regarding each composition type. See Figure 4-33 on page 184.

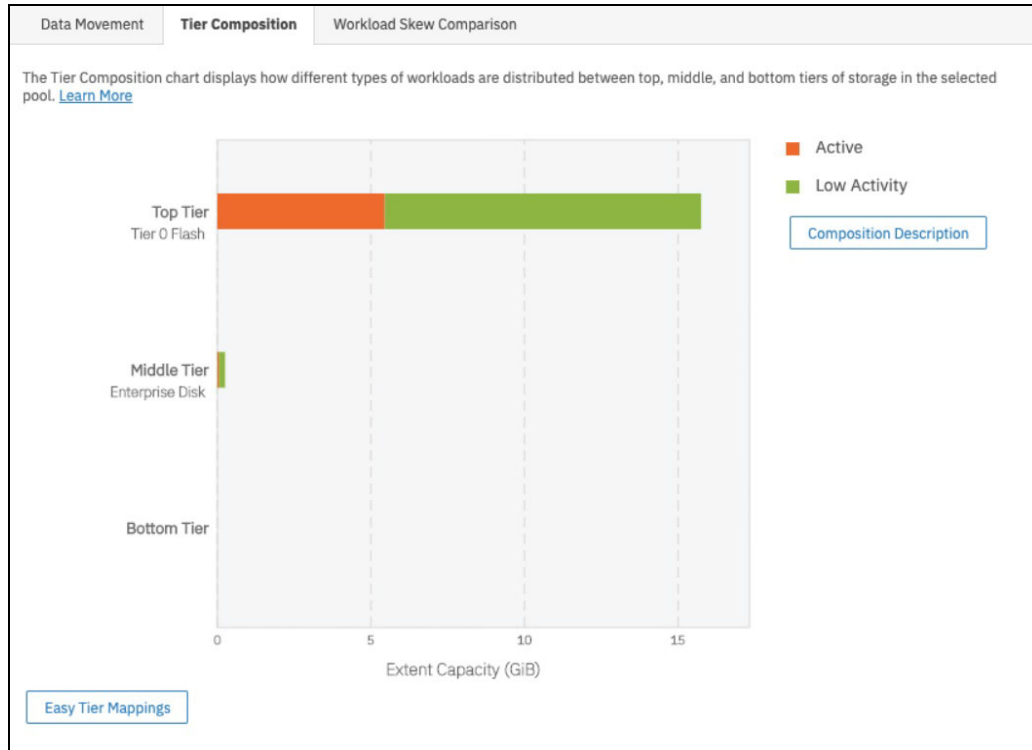


Figure 4-33 Easy tier - multi-tier pool - composition page

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 have the most optimal configuration 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 few percentages 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, and consideration to a good performing middle tier when you cannot configure enough hot tier capacity. See Figure 4-34 on page 185.

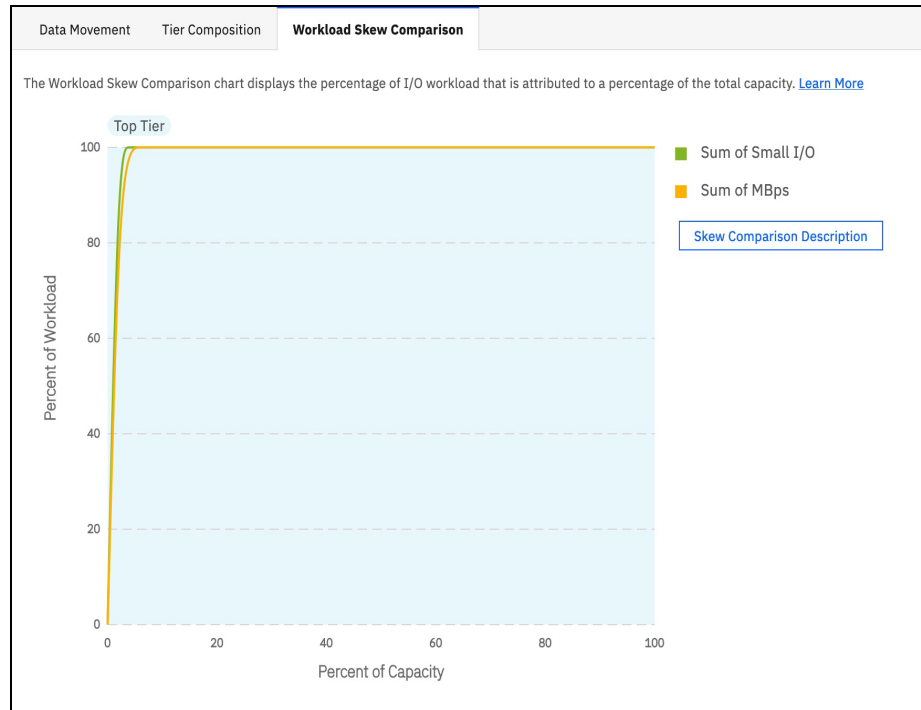


Figure 4-34 Workload skew - single tier pool

In the first example, shown in Figure 4-34, you can clearly see that the workload is highly-skewed. This single-tier pool uses less than 5% of the capacity, but is performing 99% of the workload, both in terms of IOPS and MBps.

This result is a prime example of adding a small amount of faster storage to create a “hot” tier and improve overall pool performance, as shown in Figure 4-35.

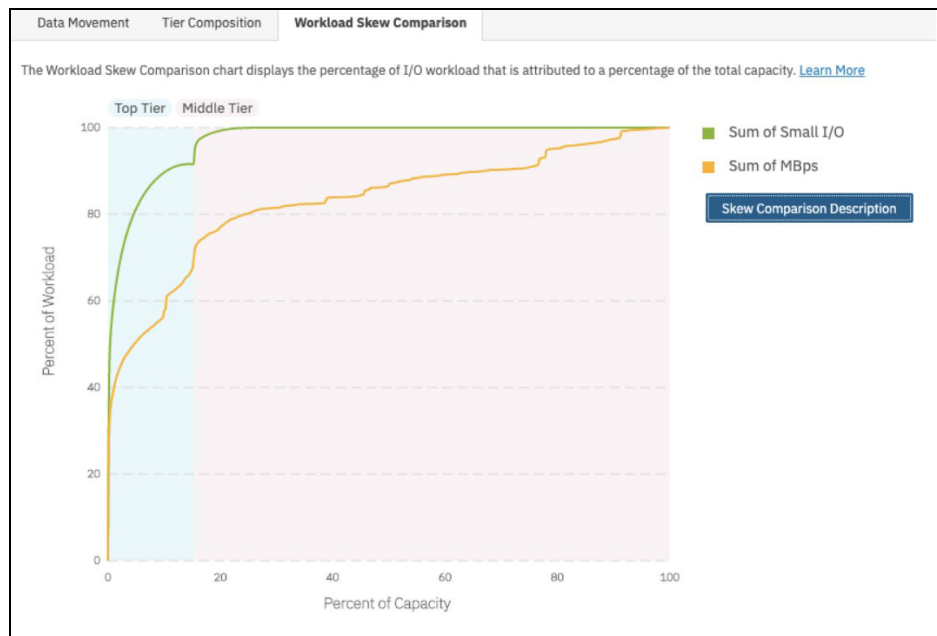


Figure 4-35 Workload skew - multi-tier configuration

In the second example, shown in Figure 4-35 on page 185, the system is already configured as a multi-tier pool and Easy Tier has been optimizing 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 as about 10% of the IOP workload is coming from the middle tier and could be further optimized to reduce latency.

This graph in Figure 4-35 on page 185 also shows the split between IOPS and MBps. Although the middle tier is not doing much of the IOP workload it is providing a reasonably large proportion of the MBps workload.

In these cases, ensure that the middle tier can manage large-block throughput. A case might be made for further improving the 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 used to provide the existing tiers.

Any new configuration with three tiers would need to comply with the configuration rules regarding the different types of storage supported in three tier configurations as discussed in “Easy Tier mapping to MDisk tier types” on page 168.

If you implemented a new system and the majority of the workload is coming from a middle or cold tier, it might take a day or two for Easy Tier to complete the migrations after it has initially analyzed the system.

If after a few days, a distinct bias to the lower tiers still exists, you might want to consider enabling “Accelerated Mode” for a week or so. However, remember to turn it back off once the system reaches a steady state. See “Easy Tier acceleration” on page 175.



Volumes

In IBM FlashSystem, a volume is a logical disk that the system presents to attached hosts. This chapter describes the various types of volumes and guidance for managing the properties.

This chapter includes the following sections:

- ▶ 5.1, “Overview of volumes” on page 188
- ▶ 5.2, “Guidance for creating volumes” on page 188
- ▶ 5.3, “Thin-provisioned volumes” on page 192
- ▶ 5.4, “Mirrored volumes” on page 200
- ▶ 5.5, “HyperSwap volumes” on page 205
- ▶ 5.6, “VMware virtual volumes” on page 206
- ▶ 5.7, “Cloud volumes” on page 209
- ▶ 5.8, “Volume migration” on page 212
- ▶ 5.9, “Preferred paths to a volume” on page 216
- ▶ 5.10, “Moving a volume between I/O groups and nodes” on page 217
- ▶ 5.11, “Volume throttling” on page 218
- ▶ 5.12, “Volume cache mode” on page 221
- ▶ 5.13, “Additional considerations” on page 224

5.1 Overview of volumes

IBM FlashSystem includes the following types of volumes:

- ▶ **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, as each I/O to the volume is spread across a larger number of disk drives comparing to a sequential volume.
- ▶ **Sequential**
A volume that has extents that are allocated sequentially from one MDisk. This type of volume is rarely used, as striped volume is better suited to most of the cases.
- ▶ **Image**
A volume that has a direct relationship with one MDisk. The extents on the volume are directly mapped to the extents on the MDisk. This is commonly used for data migration from a storage subsystem to an IBM FlashSystem.

Volumes can be created with various attributes:

- ▶ **Standard provisioned volumes**
Volumes with no special attributes.
- ▶ **Thin-provisioned volumes**
Volumes that present a larger capacity to the host than their real capacity.
- ▶ **Compressed volumes**
Volumes where data is compressed and optionally deduplicated.
- ▶ **Mirrored volumes**
A volume might contain a duplicate copy of the data in another volume. Two copies are called a mirrored volume.
- ▶ **HyperSwap volumes**
Volumes that participate in a HyperSwap relationship.
- ▶ **VMware Virtual Volumes (vVols)**
Volumes that are managed remotely by VMware vCenter.
- ▶ **Cloud volumes**
Volume that are enabled for transparent cloud tiering.

5.2 Guidance for creating volumes

More information about the guidelines can be found in the next sections of this chapter.

When you create volumes, consider the following guidelines:

- ▶ Consider the naming rules before you create volumes. It is easier to assign the correct name when the volume is created, rather than to modify it afterwards.
- ▶ Choose which kind of volume you will create. First, decide whether fully allocated (standard volumes) or thin-provisioned. And then, if you decide to create a thin-provisioned volume, analyze if you need compression and deduplication enabled.

A fully-allocated volume will be automatically formatted, which can be a time-consuming process. However, this is a background process that does not impede the immediate use of the volume.

Actions like moving, expanding, shrinking, or adding a volume copy are disabled when the specified volume is formatting. It is hard to perform one of those actions after the volume is created. However, if you prefer, you can disable the format option on the Custom tab of the volume creation window by clearing the **Format volumes** box shown in Figure 5-1.

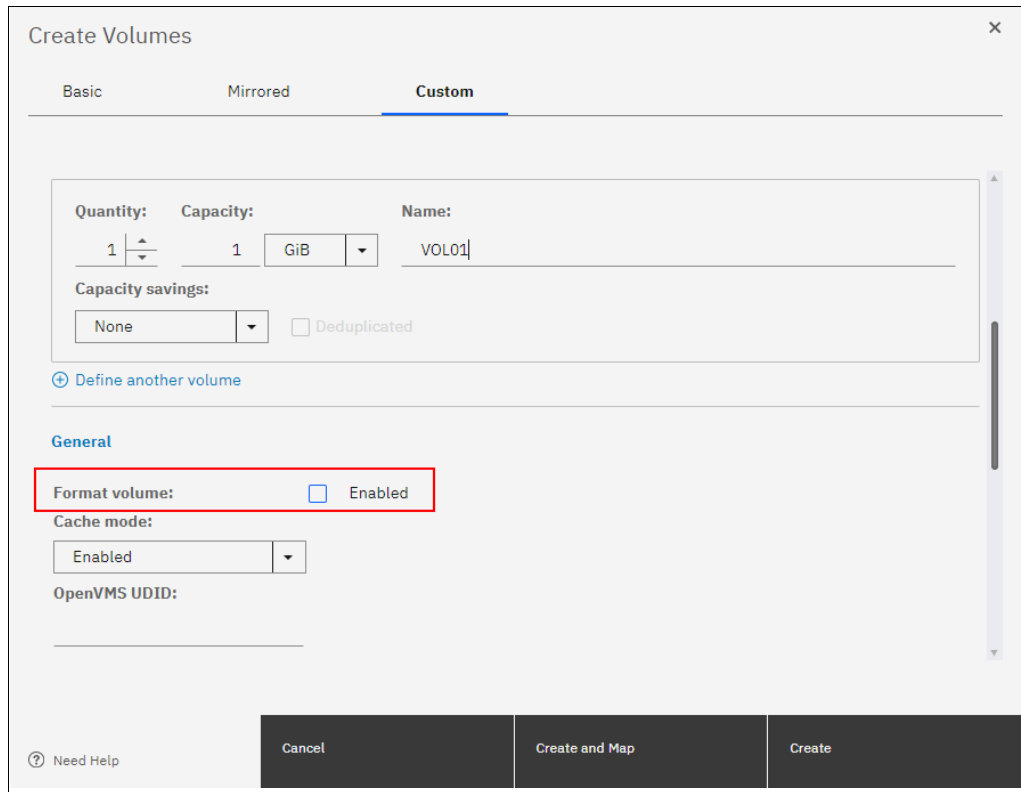


Figure 5-1 Volumes format option

You can also create volumes by using the command line interface (CLI). Example 5-1 shows the command to disable auto formatting option with the **-nofmtdisk** parameter.

Example 5-1 Volume creation without 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

Remember that when you create a volume, it takes some time to completely format it (depending on the volume size). The *syncrate* parameter of the volume that specifies the volume copy synchronization rate, 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 four 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 192.

- ▶ Each volume has an I/O group and an associated preferred node. When creating a new volume, you should consider balancing volumes across the I/O groups to balance the load across the cluster.

In configurations where it is not possible to zone a host to multiple I/O groups so the host has access to 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 recommended that a volume is accessible to hosts by only the caching 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 being 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 *load balancing* basis. Although it is not easy to estimate the workload when the volume is created, you should distribute the workload evenly on each node within an I/O group.
- ▶ Except for a few cases, the cache mode of a volume should be set to read/write. For more information, see 5.12, “Volume cache mode” on page 221.
- ▶ The maximum number of volumes per I/O group and system is described in Table 5-1.

Table 5-1 Maximum number of volumes in IBM FlashSystem

Volume type	Maximum number	Comments
Volumes (VDisks) per system	FS9x00: 10,000 FS7200: 10,000 FS5x00: 8,192	Each basic volume uses 1 VDisk, each with one copy.
HyperSwap volumes per system	FS9x00: 2,000 FS7200: 2,000 FS5100: 2,000 FS5030: 1,250	Each HyperSwap volume uses 4 VDIs, each with one copy, 1 active-active Remote Copy relationship and 4 FlashCopy mappings.
Volumes per I/O group (volumes per caching I/O group)	FS9x00: 10,000 FS7200: 10,000 FS5x00: 8,192	

Volume type	Maximum number	Comments
Volumes accessible per I/O group	FS9x00: 10,000 FS7200: 10,000 FS5x00: 8,192	
Thin-provisioned (space-efficient) volume copies in regular pools per system	8,192	
Compressed volume copies in data reduction pools per system	-	No limit is imposed here beyond the volume copy limit per data reduction pool
Compressed volume copies in data reduction pools per I/O group	-	No limit is imposed here beyond the volume copy limit per data reduction pool group
Deduplicated volume copies in data reduction pools per system	-	No limit is imposed here beyond the volume copy limit per data reduction pool
Deduplicated volume copies in data reduction pools per I/O group	-	No limit is imposed here beyond the volume copy limit per data reduction pool group
Volumes per storage pool	-	No limit is imposed beyond the volumes per system limit
Fully-allocated volume capacity	256 TB	<ul style="list-style-type: none"> ▶ Maximum size for an individual fully allocated volume. ▶ Maximum size depends on the extent size of the Storage Pool. ▶ Check 5.3.4, "Limits on virtual capacity of thin-provisioned volumes" on page 200.
Thin-provisioned (space-efficient) per-volume capacity for volumes copies in regular and data reduction pools	256 TB	<ul style="list-style-type: none"> ▶ Maximum size for an individual thin-provisioned volume. ▶ Maximum size depends on the extent size of the Storage Pool. ▶ Check 5.3.4, "Limits on virtual capacity of thin-provisioned volumes" on page 200.
Compressed volume capacity in data reduction pools	256 TB	<ul style="list-style-type: none"> ▶ Maximum size for an individual compressed volume ▶ Maximum size depends on the extent size of the Storage Pool. ▶ Check 5.3.4, "Limits on virtual capacity of thin-provisioned volumes" on page 200.

- ▶ The pool extent size does not affect the overall storage performance. A volume occupies an integer number of extents, but its length does not need to be an integer multiple of the extent size. And the length does need to be an integer multiple of the block size. Any space left over between the last logical block in the volume and the end of the last extent in the volume is unused.

A small extent size is used to minimize this unused space, and also to have a finer granularity of the volume space that is occupied on the underlying storage controller. On

the other hand, you have to consider the best extent size for your storage pools considering the back-end storage.

Important: Volume migration (using the `migratevdisk` command) between storage pools requires that both (source and destination) pools have the same extent size.

5.3 Thin-provisioned volumes

A thin-provisioned volume presents a different capacity to mapped hosts than the capacity that the volume consumes in the storage pool. The system supports thin-provisioned volumes in both *standard pools* and *data reduction pools (DRPs)*.

Figure 5-2 shows the basic concept of a thin-provisioned volume.

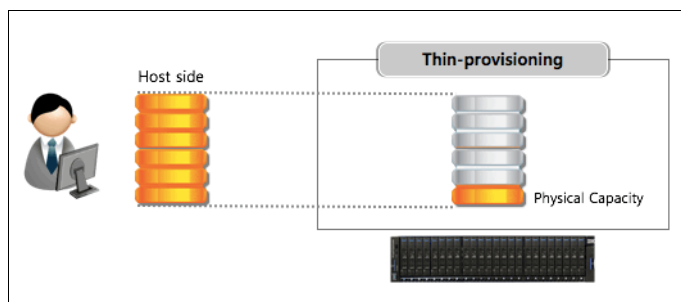


Figure 5-2 Thin-provisioned volume

Figure 5-3 shows the various types of volumes in DRP.

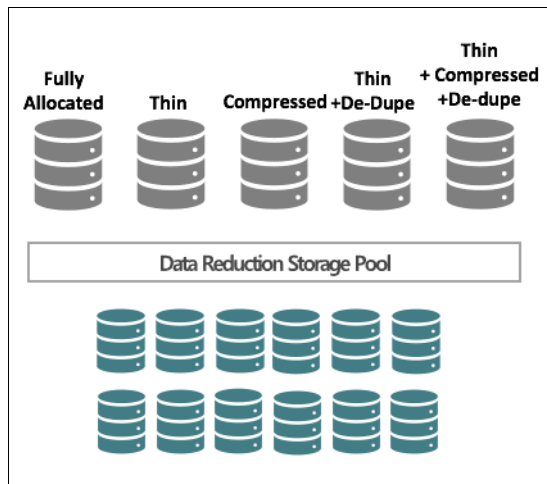


Figure 5-3 Different kinds of volumes in DRP

In standard pools, thin-provisioned volumes are created based on capacity savings criteria. These properties are managed at the volume level. However, in DRP, all the benefits of thin-provisioning are available to all the volumes that are assigned to the pool. For the thin-provisioned volumes in DRP, you can configure compression and data deduplication on these volumes, increasing the capacity savings for the entire pool.

You can enhance capacity efficiency for thin-provisioned volumes by monitoring the hosts use of capacity. When the host indicates that the capacity is no longer needed, the space is

released and can be reclaimed by the DRP. It is redistributed automatically. Standard pools do not have these functions.

Figure 5-4 shows a diagram with concepts of thin-provisioned volumes.

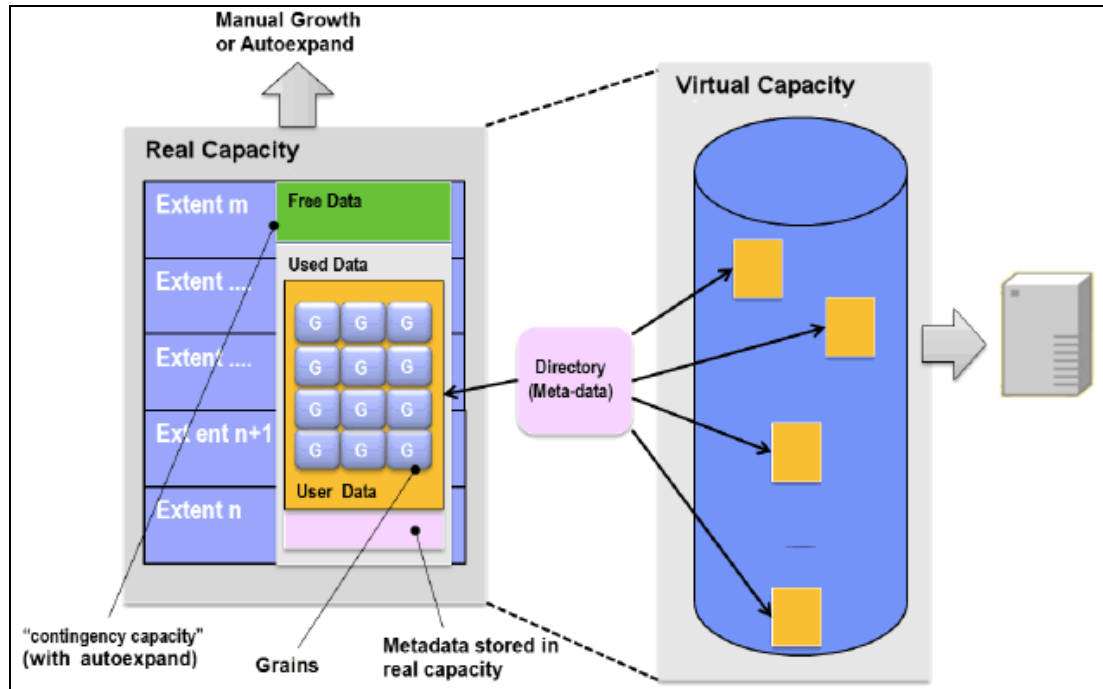


Figure 5-4 Conceptual diagram of thin-provisioned volume

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 will be 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 use of any real capacity.

Thin-provisioned volumes are available in two operating modes: *autoexpand* and *noautoexpand*. You can switch the mode at any time. If you select the autoexpand feature, IBM FlashSystem automatically adds a fixed amount of extra 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 using autoexpand, by default, to avoid volume-offline issues.

The amount of unused capacity for the volume 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.

Regardless of whether compression and deduplication are enabled, when you create a thin-provisioned volume, you must be careful for out-of-space issues in the volume and pool where the volume is created. You should set the warning threshold in the pools containing 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 an SNMP trap) in the storage pool. If you are not using the autoexpand feature, you 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 situation.

If you do not want to be concerned with monitoring the 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 which is initially allocated to it (**-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 autoexpand enabled and warning disabled options will work in most scenarios. In some cases, you might consider using different values to suit your environment.

Example 5-2 shows the command to create a volume.

Example 5-2 Thin-provisioned volume creation

```
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 5-5 on page 195 shows how to modify capacity savings of a volume. You can right-click the volume and select **Modify Capacity Savings**.

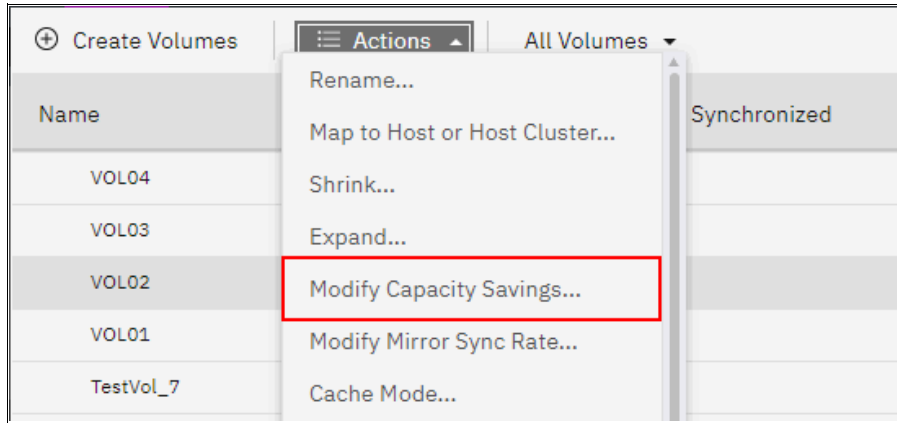


Figure 5-5 Modifying capacity savings of a volume nondisruptively

The fully-allocated to thin-provisioned migration procedure uses a zero-detection algorithm so 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, saving more space. When data is read to hosts, the data is decompressed.

Compression is provided via DRPs. If you want to use compressed volumes, they have to be created in DRPs.

DRPs also reclaim capacity that is not used by hosts that support Small Computer System Interface (SCSI) **unmap** commands. When these hosts issue SCSI **unmap** commands, the DRP reclaims the released capacity for redistribution.

For compressed volumes in DRPs, the used capacity before compression indicates the total amount of data that is written to volume copies in the storage pool before data reduction occurs.

This compression solution provides nondisruptive conversion between compressed and uncompressed volumes and eliminates the need for special procedures to deal with compressed volumes.

If you are planning to virtualize volumes that are connected to your hosts directly from any storage subsystems, and you want to know the space saving you will achieve, run one of the following tools:

- ▶ *IBM FlashSystem Comprestimator* is an integrated Comprestimator tool, which is available through the management GUI and command line interface. If you are considering to apply compression on existing non-compressed volumes in an IBM FlashSystem, you can use this tool to evaluate if compression will generate capacity savings.

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

- ▶ *IBM Data Reduction Estimator Tool (DRET)* is a host-based utility that can estimate capacity savings achieved by using deduplication.

For more information, see: [IBM Data Reduction Estimator Tool \(DRET\) for SVC, Storwize and FlashSystem products](#).

Both tools can be used to estimate an expected compression rate for block devices.

For more information, see 4.1.4, “Data reduction estimation tools” on page 115.

As shown in Figure 5-6, customize the Volume view to see the compression savings for a compressed volume, and estimated compression savings for a non-compressed volume that you are planning to migrate.

The screenshot shows a storage management interface with a table of volumes and a custom view menu. The table has columns for Name, State, Pool, and Host ID. The custom view menu is open, showing various columns that can be selected or deselected.

Name	State	Pool	Host ID
VOL04	Online	Swim	7800000000000...
VOL03	Online	Pool1	7800000000000...
VOL02	Online	Pool1	7800000000000...
VOL01	Online	Swim	7800000000000...
TestVol_7	Online	Swim	7800000000000...
TestVol_6	Online	Swim	7800000000000...
TestVol_5	Online	Swim	7800000000000...
TestVol_4	Online	Swim	7800000000000...
TestVol_3	Online	Swim	7800000000000...
TestVol_2	Online	Swim	7800000000000...
TestVol_1	Online	Swim	7800000000000...
TestVol_0	Online	Swim	7800000000000...
ITSO-FC-VOL-10	Online	Swim	7800000000000...
ITSO-FC-VOL-08-Target	Online	Swim	7800000000000...
ITSO-FC-VOL-08-Source	Online	Swim	7800000000000...
ITSO-FC-VOL-07_01	Online	Swim	7800000000000...

Column Name	Selected
Protocol Type	<input checked="" type="checkbox"/>
UID	<input checked="" type="checkbox"/>
Host Mappings	<input checked="" type="checkbox"/>
Preferred Node ID	<input type="checkbox"/>
Capacity	<input checked="" type="checkbox"/>
Real Capacity	<input checked="" type="checkbox"/>
Used Capacity	<input checked="" type="checkbox"/>
Cache State	<input type="checkbox"/>
Compression Savings	<input checked="" type="checkbox"/>
Capacity Savings	<input checked="" type="checkbox"/>
Estimated Compression Savings	<input checked="" type="checkbox"/>
Estimated Compression Savings %	<input checked="" type="checkbox"/>
Estimated Thin Savings	<input checked="" type="checkbox"/>
Estimated Thin Savings %	<input checked="" type="checkbox"/>
FlashCopy Mappings	<input type="checkbox"/>
Caching I/O Group ID	<input type="checkbox"/>
Caching I/O Group	<input type="checkbox"/>

Figure 5-6 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 DRP for saving capacity.

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, resulting in the amount of data that must be stored being greatly reduced.

If a volume is configured with both deduplication and compression, data is deduplicated first, and then compressed. Therefore, deduplication references are created on the compressed data stored on the physical domain.

Figure 5-7 on page 197 shows the settings to create a compressed and deduplicated volume.

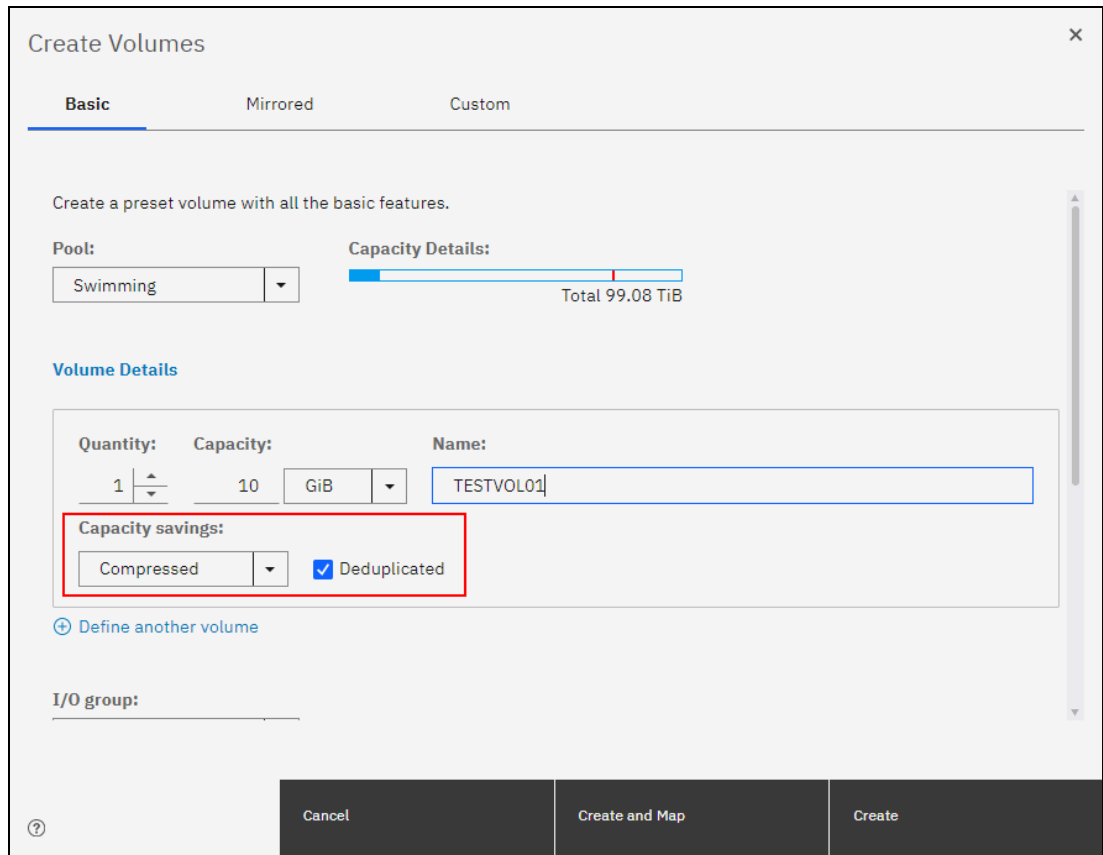


Figure 5-7 Creating deduplicated volumes

To create a thin-provisioned volume that uses deduplication, enter the following command in the CLI, see Example 5-3.

Example 5-3 Creating thin-provisioned volume with 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, enter the following command in Example 5-4.

Example 5-4 Creating compressed volume with 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 stored on the volume, the capacity savings can be significant. Examples of use cases that typically benefit from deduplication are virtual environments with multiple virtual machines running the same operating system and backup servers.

In both cases it is expected that there will be multiple copies of identical files, such as components of the standard operating system or applications used in the given organization.

Conversely, data encrypted or compressed at the file-system level does not benefit from deduplication because these operations already removed redundancy.

If you want to evaluate if savings will be realized by migrating a set of volumes to deduplicated volumes, you can use DRET, which is a command line host-based utility for estimating the data reduction saving on block devices. More information about DRET can be found in 4.1.4, “Data reduction estimation tools” on page 115.

5.3.3 Thin-provisioning considerations

Thin-provisioning works only if the host limits the writes to areas of the volume that store data. If the host, for instance, performs a low-level format of the entire volume, the host writes to the whole volume and there will be no advantage of using a thin-provisioning volume over a fully-allocated volume.

Consider the following properties of thin-provisioned volumes:

- ▶ When the used capacity first exceeds the volume *warning threshold*, an event is raised, indicating additional real capacity is required. The default warning threshold value is 80% of the volume capacity. To disable warnings, specify 0%.
- ▶ Compressed volumes have an *uncompressed used capacity* attribute (for standard pools) and a *used capacity before reduction* attribute (for DRP). These are the used capacities before compression or data reduction and are used to calculate the compression ratio.

Thin-provisioning and over-allocation

Because thin-provisioned volumes do not store the zero blocks, a storage pool is over-allocated only after the sum of all volume capacities exceeds the size of the storage pool.

Storage administrators likely think 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 servers writes 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

It is critical that capacity be monitored when thin-provisioned or compressed volumes are used. Be sure to add more capacity *before* you run 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 its used capacity grows when volumes are created or expanded.

However, in a pool with thin-provisioned volumes, the used capacity can increase at any time if the host file system grows. For this reason, the storage administrator needs to consider capacity-planning carefully. It is critical to put in place volume and pool capacity monitoring.

Tools, such as IBM Spectrum Control and 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 will run out of space.

IBM FlashSystem 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 Storage Insights, IBM now has the ability to monitor systems and flag systems that may be short on capacity. Extreme situations will result in a support ticket being generated and the customer 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 and 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, remember that 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 have run out of space, you have a limited amount of time to react. You need a plan in place and you must understand what to do next.

- ▶ Have unallocated storage on standby

You can always have spare drives or MDisks ready to be added to whichever storage pool runs out of space within only a few minutes. This capacity gives you some breathing room while you take other actions. The more drives or MDisks you have, the more time you have 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 up space. However, data migration on IBM FlashSystem is designed to progress slowly to avoid performance problems. Therefore, it might be impossible to complete this migration before your applications go offline.

- ▶ Policy-based solutions

A policy is not going to solve the problem if you run out of space. However, you can use policies to reduce the likelihood of that ever happening to the point where you feel comfortable doing less of the other options.

You can use these types of policies for thin-provisioning:

Note: The following policies use arbitrary numbers. These arbitrary numbers are designed to make the suggested policies more readable. We do not provide 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 there is always enough free capacity for your ten largest volumes to reach 100% full without running out of free space.
- Never over-allocate more than 200%. That is, if you have 100 TB of capacity in the storage pool, then 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 and 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 granularities can save more space, but they have larger directories. 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 is a predefined value and cannot be set or changed.

If you are planning to use thin provisioning with FlashCopy, remember that 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 using thin-provisioned volume in DRPs as a FlashCopy source or target volume.

Note: The use of thin-provisioned volumes in DRP for FlashCopy is not recommended.

5.3.4 Limits on virtual capacity of thin-provisioned volumes

The extent and grain size factors limit the virtual capacity of thin-provisioned volumes beyond the factors that limit the capacity of regular volumes. For more information about the maximum volume capacity for each extent size on IBM FlashSystem families, see the following links:

- ▶ [V8.4.0.x Configuration Limits and Restrictions for IBM FlashSystem 9200](#)
- ▶ [V8.4.0.x Configuration Limits and Restrictions for IBM FlashSystem 9100](#)
- ▶ [V8.4.0.x Configuration Limits and Restrictions for IBM FlashSystem 7200](#)
- ▶ [V8.4.0.x Configuration Limits and Restrictions for IBM FlashSystem 5x00](#)

5.4 Mirrored volumes

By using volume mirroring, a volume can have two copies. Each volume copy can belong to a different pool, and each copy has the same virtual 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.

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 non-mirrored volume by deleting one copy or by splitting one copy to create a new 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 located in data reduction pools, you can also use deduplication to the volume copies to increase the capacity savings. If you are creating a new volume, the two copies can be of different types, but to use deduplication, both copies must reside in a data reduction pool. You can add a deduplicated volume copy in a data reduction pool to an existing volume with a copy in a standard pool. You can use this method to migrate existing volume copies to data migration pools.

You can use mirrored volumes for the following reasons:

- ▶ Improving 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 both 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 there is a problem 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 data reduction pools to improve capacity savings.

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 is unable to update the state information, a mirrored volume might need to be taken offline to maintain data integrity. To ensure the high availability of the system, ensure that multiple quorum candidate disks are allocated and configured on different storage systems.

When a volume mirror is synchronized, a mirrored copy can become unsynchronized if it goes offline and write I/O requests need to 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-provisioned volumes, the primary volume formats before synchronizing to the volume copies. The **-syncrate** parameter on 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 shown in Figure 5-8.

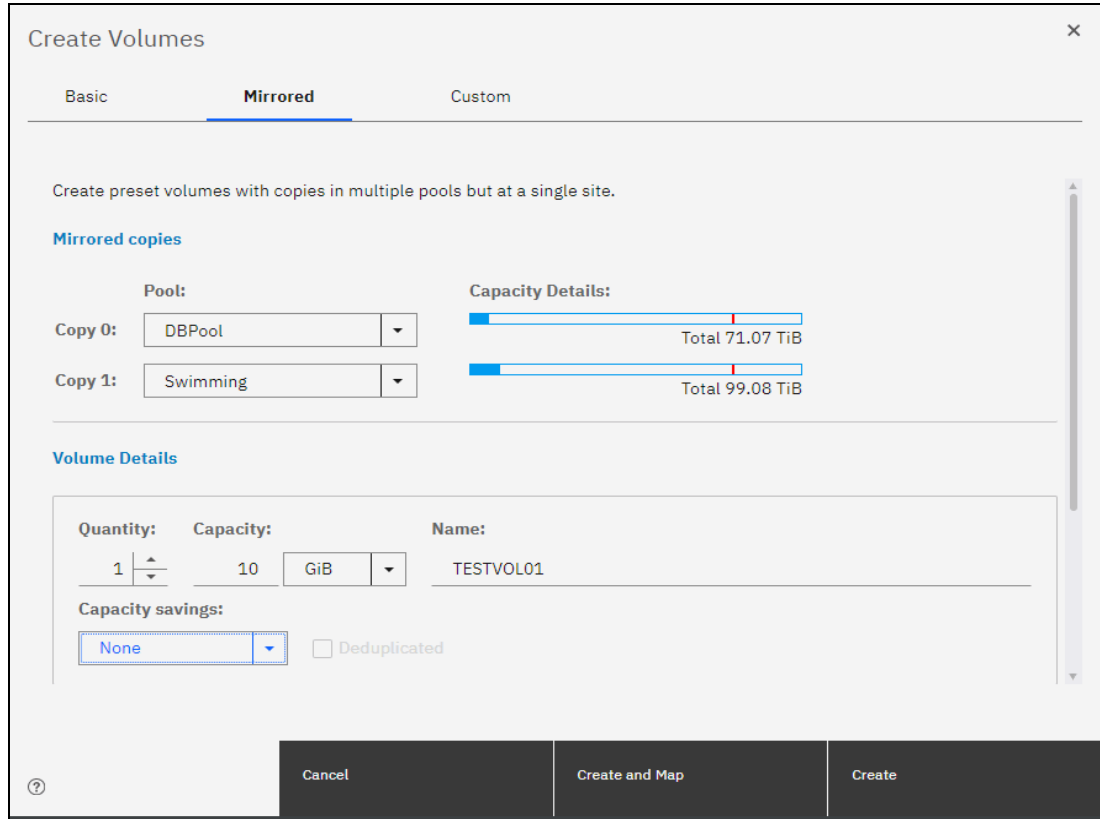


Figure 5-8 Mirrored volume creation

You can convert a non-mirrored volume into a mirrored volume by adding a copy, as shown in Figure 5-9 on page 203.

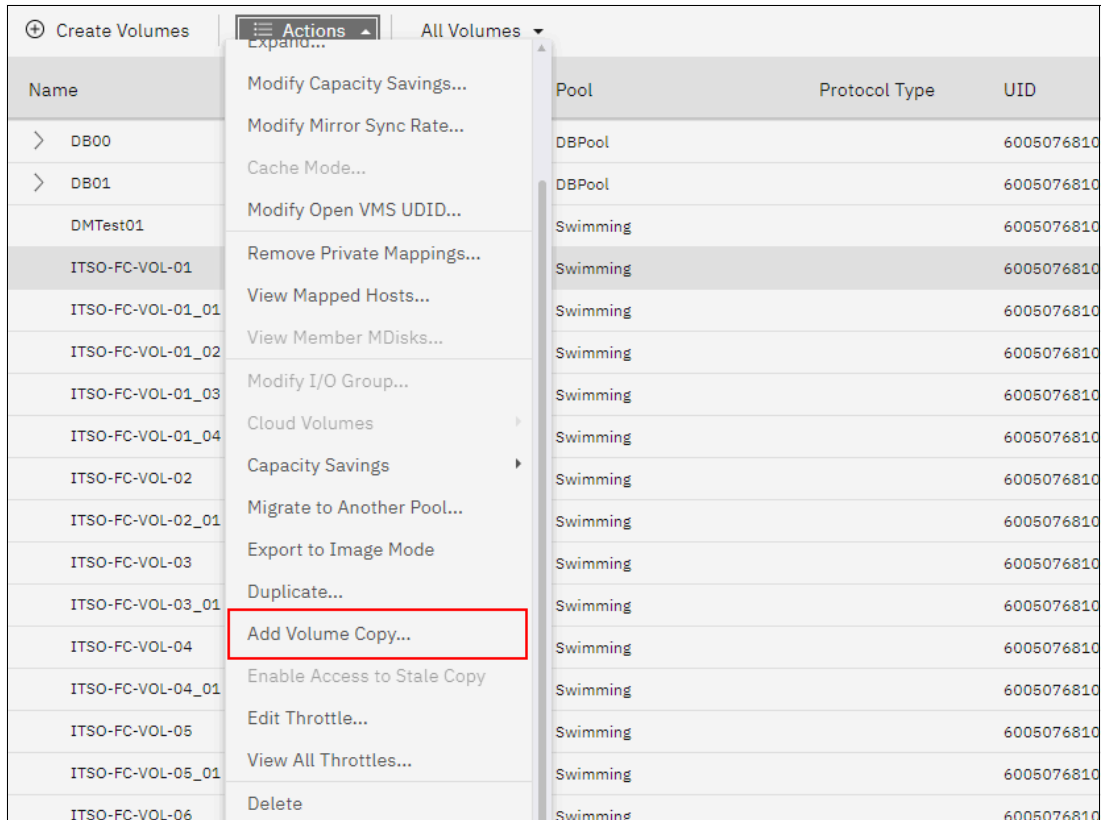


Figure 5-9 Adding a volume copy

5.4.1 Write fast failovers

With write fast failovers, during processing of host write I/O, the system submits writes to both copies. 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 network, the I/O ending sequence typically completes in 10 to 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 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. Additionally, 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 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, there can be an underlying performance problem 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 enter the `mkvdisk` command to create a new 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 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, then 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. This means that the system tries the other copy first for read I/O during the following 4 - 6 minutes. After that, the system reverts to read the original primary-for-read copy.

During this period, if read I/O to the other copy also goes 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 can happen 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 has failed. However, this mirroring does not affect 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 disaster recovery and high availability between I/O groups at different locations. Before you configure HyperSwap volumes, the system topology needs to be configured for HyperSwap and sites must be defined. Figure 5-10 shows an overall diagram of IBM FlashSystem HyperSwap configured with two sites.

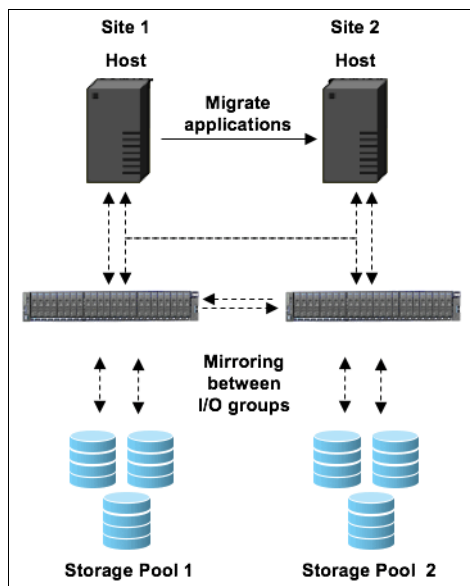


Figure 5-10 Overall HyperSwap diagram

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 either 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 will be 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 data reduction pools on both sites.

The management GUI creates an HyperSwap relationship and change volumes automatically. HyperSwap relationships manage the synchronous replication of data between HyperSwap volume copies at the two sites. If your HyperSwap system supports self-encrypting drives and the base volume is fully allocated in a data reduction pool, then the

corresponding change volume is created with compression enabled. If the base volume is in a standard pool, then the change volume is created as a thin-provisioned volume.

You can specify a consistency group that contains multiple active-active relationships to simplify management of replication and provide consistency across multiple volumes. A consistency group is commonly used when an application spans multiple volumes. Change volumes maintain a consistent copy of data during resynchronization. Change volumes allow an older copy to be used for disaster recovery if a failure occurred on the up-to-date copy before resynchronization completes.

You can also use the `mkvolume` command line to create a HyperSwap volume. The command also defines pools and sites for HyperSwap volume copies and creates the active-active relationship and change volumes automatically. If your HyperSwap system supports self-encrypting drives and the base volume is fully allocated in a data reduction pool, then the corresponding change volume is created with compression enabled. If the base volume is in a standard pool, then the change volume is created as a thin-provisioned volume.

The relationship between Master and Auxiliary volume in a two-site HyperSwap topology is shown in Figure 5-11.

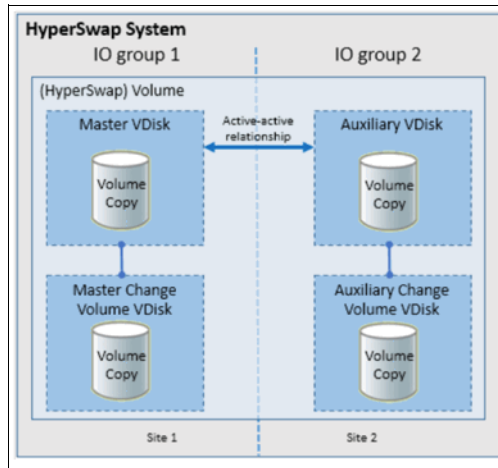


Figure 5-11 Master and Auxiliary volumes

For more details about HyperSwap volumes, see 7.3, “HyperSwap Volumes” on page 346.

5.6 VMware virtual volumes

The IBM FlashSystem supports VMware vVols, which allows VMware vCenter to automate the management of system objects such as volumes and pools.

You can assign ownership of Virtual Volumes to IBM Spectrum Connect by creating a user with the VASA Provider security role. IBM Spectrum Connect provides communication between the VMware vSphere infrastructure and the system. Although you can complete certain actions on volumes and pools that are owned by the VASA Provider security role, IBM Spectrum Connect retains management responsibility for Virtual Volumes.

When virtual volumes are enabled on the system, a utility volume is created to store metadata for the VMware vCenter applications. You can select a pool to provide capacity for the utility volume. With each new volume created by the VASA provider, VMware vCenter defines a few kilobytes of metadata that are stored on the utility volume.

The utility volume can be mirrored to a second storage pool to ensure that the failure of a storage pool does not result in loss of access to the metadata. Utility volumes are exclusively used by the VASA provider and cannot be deleted or mapped to other host objects.

Note: The utility volume cannot be created in a DRP.

Figure 5-12 provides a high-level overview of the key components that enable the vVols management framework.

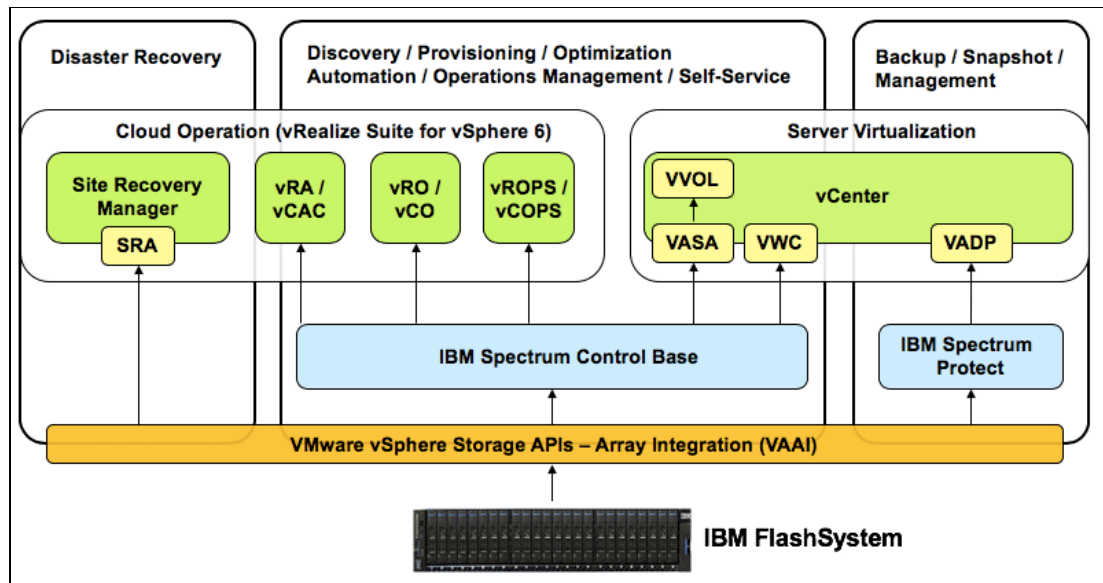


Figure 5-12 Overview of the key components of VMware environment

You can also use data copy through VMware vSphere Storage APIs Array Integration (VAAI) in Figure 5-13.

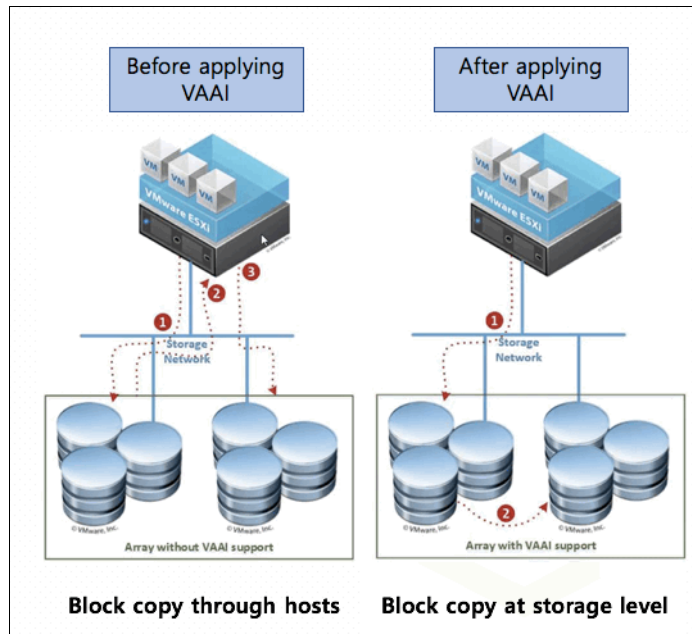


Figure 5-13 VMware vSphere Storage APIs Array Integration (VAAI)

Before configuring vVols, the following prerequisites must be met:

- ▶ An existing IBM Spectrum Connect
- ▶ VMware vSphere ESXi hosts and vCenter running version 6.0 or later
- ▶ Network Time Protocol (NTP) server configured on both IBM FlashSystem and IBM Spectrum Connect

To start using vVols, complete the following tasks on the IBM FlashSystem before you configure any settings within the IBM Spectrum Connect server:

1. Enable vVols on the IBM FlashSystem.
 - a. In the management GUI, click **Settings** → **System** → **VVOL** and select **On**.
 - b. Select the pool where to store the utility volume. If possible, store a mirrored copy of the utility volume in a second storage pool that is in a separate failure domain. The utility volume cannot be created in a data reduction pool.
 - c. Create a user for IBM Spectrum Connect to communicate with the IBM FlashSystem, as shown in Figure 5-14.

The screenshot shows the 'Enable VVOL' configuration window. It includes the following elements:

- Title:** Enable VVOL
- Store utility volume here:** Pool1
- Store mirrored copy here (optional):** Select Pool
- Create a user for IBM Spectrum Control Base to connect to this system.**
- User name:** SpectrumUser
- Password:** [Masked]
- Verify password:** [Masked]
- Password requirements:**
 - ✓ Minimum 6 characters long
 - ✓ Must not include problematic characters (ex: control characters), or start or end with a space
- Buttons:** Cancel, Enable

Figure 5-14 Enable VVOL window

2. Create the user account for the IBM Spectrum Connect and the user group with VMware vSphere API for Storage Awareness (VASA) provider role, if they were not set in the previous step.
 - a. Create user group by clicking **Access** → **Users by Group** → **Create User Group**. Enter the user group name, select **VASA Provider** for the role and click **Create**.

- b. Create the user account by clicking **Access** → **Users by Group**, select the user group created in the previous step, and click **Create User**. Enter the name of the user account, select the user group with VASA Provider role, enter a valid password for the user and click **Create**.
3. For each ESXi host server to use vVols, create a host object.
 - a. In the management GUI, select **Hosts** → **Hosts** → **Add Host**.
 - b. Enter the name of the ESXi host server, enter connection information, select **VVOL** for the host type and click **Add Host**.
 - c. If the ESXi host was previously configured, the host type can be changed by modifying the ESXi host type.

Note: The user account with VASA Provider role is used by only the IBM Spectrum Connect server to access the IBM FlashSystem and to run the automated tasks that are required for Virtual Volumes. Users must not directly log in to the management GUI or CLI with this type of account and complete system tasks, unless they are directed to by support.

5.7 Cloud volumes

A cloud volume is any volume that is enabled for transparent cloud tiering. After transparent cloud tiering 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. These snapshots can be restored to the system for disaster recovery purposes. Before you create cloud volumes, a valid connection to a supported cloud service provider must be configured.

With transparent cloud tiering, the system supports connections to cloud service providers 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 cloud service provider.

A cloud account defines the connection between the system and a supported cloud service provider, and must be configured before data can be transferred to or restored from the cloud storage. After a cloud account is configured with the cloud service provider, you determine which volumes you want to create cloud snapshots of and enable transparent cloud tiering on those volumes.

Figure 5-15 shows a sample diagram of IBM FlashSystem Transparent Cloud Tying.

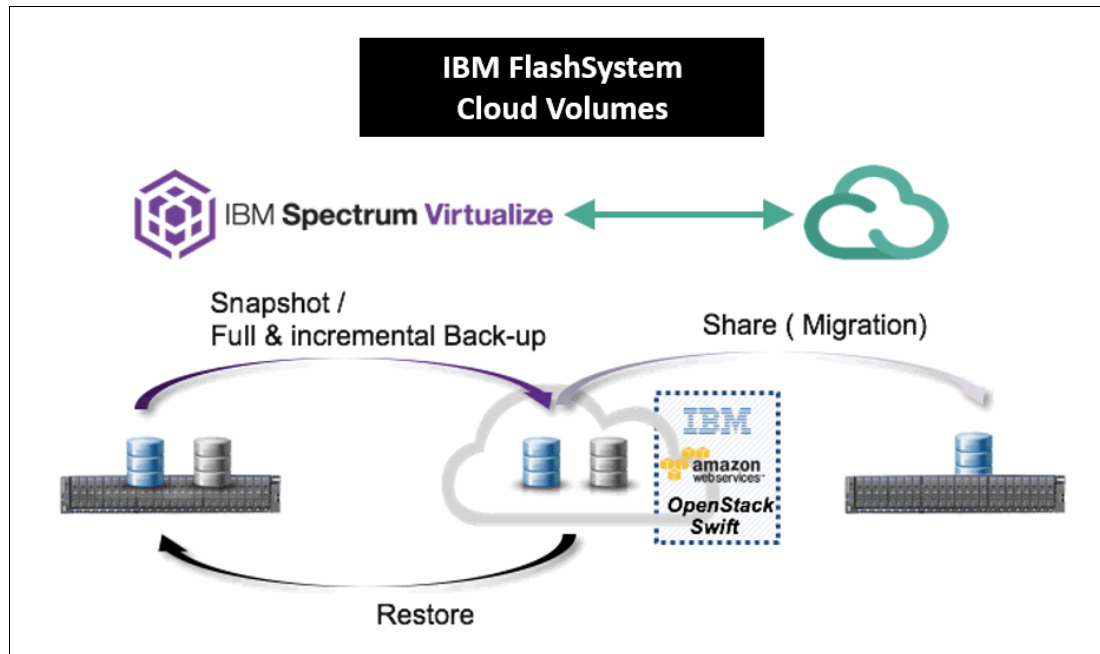


Figure 5-15 Cloud volumes - Transparent Cloud Tying

A cloud account is an object on the system that represents a connection to a cloud service provider by using a particular set of credentials. These credentials differ depending on the type of cloud service provider that is being specified. Most cloud service providers require the host name of the cloud service provider and an associated password, and some cloud service providers also require certificates to authenticate users of the cloud storage. Public clouds use certificates that are signed by well-known certificate authorities.

Private cloud service providers can use either a self-signed certificate or a certificate that is signed by a trusted certificate authority. These credentials are defined on the cloud service provider and passed to the system through the administrators of the cloud service provider. A cloud account defines whether the system can successfully communicate and authenticate with the cloud service provider by using the account credentials.

If the system is authenticated, then it can access cloud storage to either 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 cloud service provider. Migration between providers is not supported.

The system supports IBM Cloud, OpenStack Swift, and Amazon S3 cloud service providers.

Note: Transparent Cloud Tying is not supported on the IBM FlashSystem 5000 family.

5.7.1 Transparent cloud tiering configuration limitations and rules

Consider the following limitations and rules regarding transparent cloud tiering:

- ▶ 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 expanded or shrunk.
- ▶ A volume cannot be configured for a cloud snapshot if any of the following conditions are valid:
 - The volume is part of a Remote Copy relationship (Metro Mirror, Global Mirror, active-active) master, auxiliary, or change volume. This configuration prevents the cloud snapshot from being used with HyperSwap volumes.
 - The volume is a VMware vVol, including IBM FlashCopy owned volumes that are used internally for vVols 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 an existing 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 vVol, including FlashCopy volumes that are used internally for vVol restoration functions
 - A file system volume
 - Part of a Remote Copy relationship (Metro Mirror, Global Mirror, active-active) master, auxiliary, or change volume
- ▶ 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 Restore to the production volume

This is a process where snapshot version is restored 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 exists on 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 Restore 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, then the restore operation uses the unique identifier (UID) of the new volume. If the new volume does not exist on the system, you need to choose whether to use the UID from the original volume or create a new UID. If you plan on using 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. A volume can also be migrated to a different type of storage pool.

The command varies when you migrate from image-mode to managed-mode, or vice versa, as shown in Table 5-2.

Table 5-2 Migration types and associated commands

Storage pool-to-storage pool	Command
Managed-to-managed or Image-to-managed	<code>migratevdisk</code>
Managed-to-image or Image-to-image	<code>migratetoimage</code>

Migrating a volume from one storage pool to another is nondisruptive to the host application using the volume. Depending on the workload of IBM FlashSystem, there might be a slight performance impact.

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 have the same extent site. If you need to migrate a volume from one storage pool to another storage pool with different extent sizes, you can use volume mirroring.

The next sections contain guidance for migrating volumes by using the methods mentioned above.

5.8.1 Image-type to striped-type volume migration

When you migrate existing storage into the IBM FlashSystem, the existing storage is brought in as *image-type volumes*, which means that the volume is based on a single MDisk. The CLI command that can be used is `migratevdisk`.

Example 5-5 shows the `migratevdisk` command that can be used to migrate an *image-type volume* to a *striped-type volume*, and can be used to migrate a *striped-type volume* to a *striped-type volume* as well.

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 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).

It is only represented 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 existing data to an IBM FlashSystem and to 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 managed disk 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 managed disk group must be in the same site as the source volume. If the volume is in an active-active relationship, the new managed disk group must be located in the same site as the source volume. Additionally, the site information for the MDisk being added must be well-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 storage pool.

The **migratetoimage** command is useful when you want to use your system as a *data mover*. To better understand all requirements and specifications for that command, see [IBM FlashSystem 9200 8.4.0 Documentation - migratetoimage](#).

5.8.3 Migrating with volume mirroring

Volume mirroring also offers the ability to migrate volumes between storage pools with different extent sizes.

Complete the following steps to migrate volumes between storage pools:

1. Add a copy to the target storage pool.
2. Wait until the synchronization is complete.
3. Remove the copy in 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 preferred practice on this type of migration is to try not to overload the systems with a high *syncrate*, and not overload the system with too many migrations at the same time.

Note: You cannot use the data migration function to move a volume between storage pools that have different extent sizes. Migration commands fail if the target or source volume is offline, a quorum disk is not defined, or the defined quorum disks are unavailable. Correct the offline or quorum disk condition and rerun the command.

The **syncrate** parameter specifies the copy synchronization rate. A value of zero (0) prevents synchronization. The default value is 50. The supported **-syncrate** values and their corresponding rates are listed in Table 5-3.

Table 5-3 Sample syncrate values

User-specified syncrate attribute value	Data copied/sec
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

User-specified syncrate attribute value	Data copied/sec
81 - 90	32 MB
91 - 100	64 MB

We recommend modifying the **syncrate** value after you monitor overall bandwidth and latency. Then, if the performance is not impacted on migration, increase the **syncrate** value to complete within allotted time.

You can also use volume mirroring when you migrate a volume from an existing and non-virtualized storage device to IBM FlashSystem. As you can see in Figure 5-16, you must first attach the existing storage to IBM FlashSystem by using the virtualization solution, which requires some downtime because hosts will start to access the volumes through IBM FlashSystem.

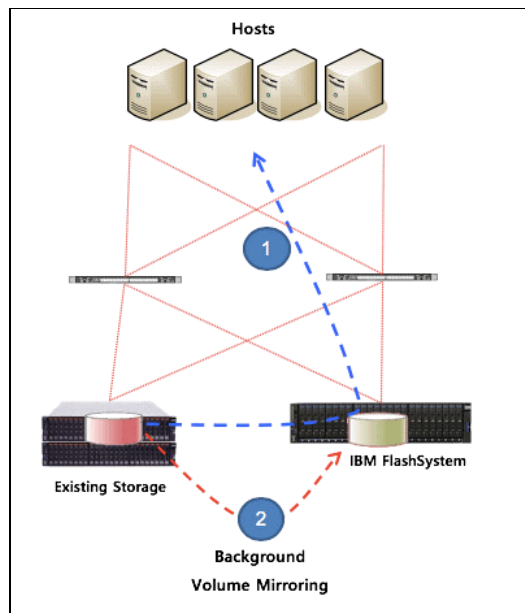


Figure 5-16 Migration with Volume Mirroring

After the storage is properly attached to IBM FlashSystem, map the image-type volumes to the hosts, so it will recognize volumes as if they were being 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 then starts in background.

For more information, see [IBM FlashSystem 9200 8.4.0 Documentation - Migrating volumes between pools using the CLI](#).

5.8.4 Migration from Standard Pool to Data Reduction Pool

If you want to migrate volumes to DRP, you can move them with volume mirroring between a standard pool and DRP. Hosts I/O operations are not disrupted during migration. Figure 5-17 on page 216 shows two examples of how you can use volume mirroring to convert volumes to a different type or to migrate volumes to a different type of pool.

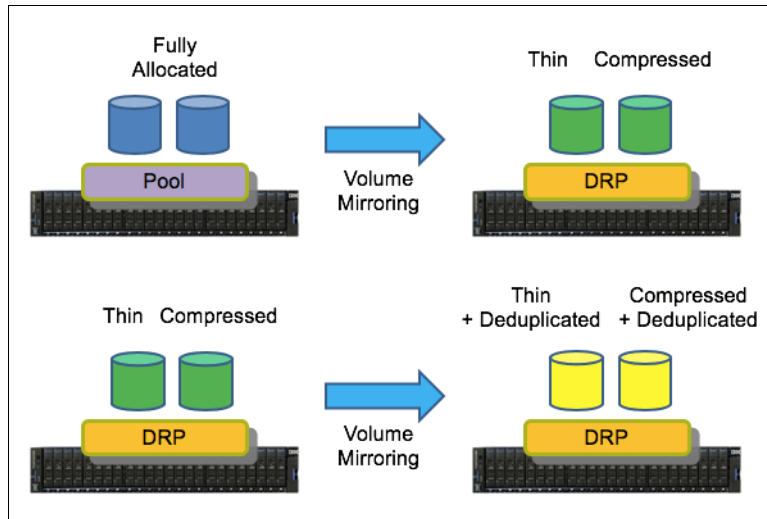


Figure 5-17 Converting volumes with Volume Mirroring

You can also move compressed or thin-provisioned volumes in standard pools to data reduction pools to simplify management of reclaimed capacity. The data reduction pool 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 data reduction pool. This method creates a copy of the volume in a new DRP and does not disrupt host operations.

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 to determine which node destages writes to the backend storage.

Preferred node assignment is normally automatic. The system selects the node in the I/O group that has the fewest volumes. However, the preferred node can be specified or changed, if needed.

All modern multipathing drivers support Asymmetric Logical Unit Access (ALUA). ALUA allows the storage to mark certain paths as preferred (paths to the preferred node). ALUA multipathing drivers honor preferred pathing and only send I/O to the other node if the preferred node is not accessible.

Figure 5-18 on page 217 shows write operations from a host to two volumes with different preferred nodes.

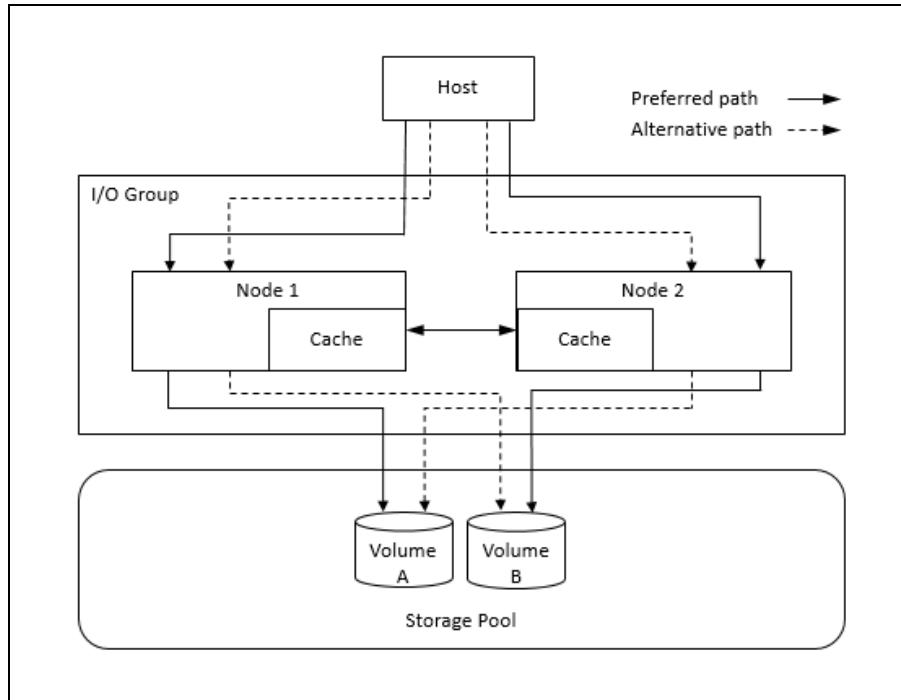


Figure 5-18 Write operations from a host

When debugging performance problems, it can be useful to look at 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 and can be identified on these tools.

For more information about this performance metric and more in IBM Spectrum Control, see [IBM Spectrum Control 5.4.2 Documentation - 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.

The change of preferred node of a volume either 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 case of some specific operating systems or applications, they could detect some time outs.

This operation can be done by using both CLI and GUI, but if you have only one I/O group, this is not possible using the GUI. To change the preferred node within an I/O group using CLI, use the `movevdisk -node <node_id or node_name> <vdisk_id or vdisk_name>` command.

5.10.2 Moving a volume between I/O groups

When moving a volume between I/O groups, it is recommended you let the system choose the volume preferred node in the new I/O group. But, it is possible to manually set the preferred node during this operation using both GUI and CLI.

Non-Disruptive Volume Move (NDVM) occurs when a volume is moved across I/O groups. Some limitations exist with NDVM, mostly in Host Cluster environments. You can check the compatibility at the [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 must first 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 non-disruptive 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 has changed and the ports by which the volume is accessed has changed. This can be done in the situation where one pair of nodes becomes over used.

If there are any host mappings 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 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 is complete.

5.11 Volume throttling

Volume throttling effectively throttles the number of I/O 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. Volume throttle can be used to limit I/Os for these volumes so that I/O operations for production volumes are not affected. Figure 5-19 shows an example of volume throttling.



Figure 5-19 Volume throttling for each LUN

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 large amounts of 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. It would be better to set an IOPS limit.

On the other hand, a streaming video application often issues a small amount of I/O, but it transfers large amounts of data. In contrast to the database example, defining an I/O throttle based in IOPS does not achieve a good result. For a streaming video application, it would be better to set an MBps limit.

You can edit the throttling value in the menu, as shown in Figure 5-20.

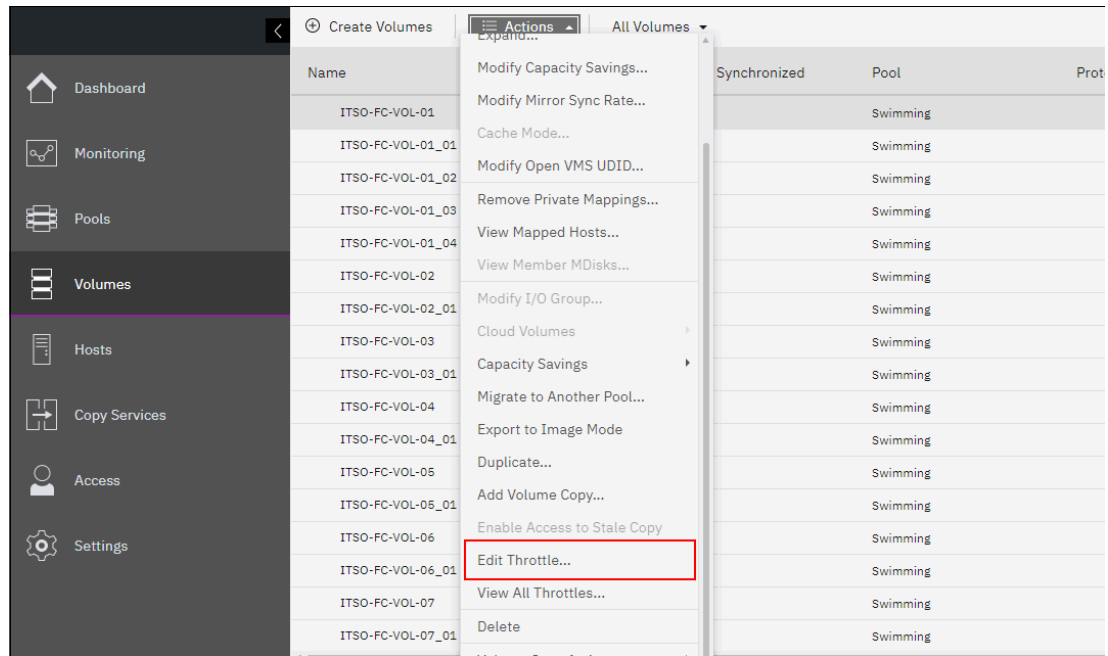


Figure 5-20 Volume Throttling

Figure 5-21 shows both bandwidth and IOPS parameter that can be set.

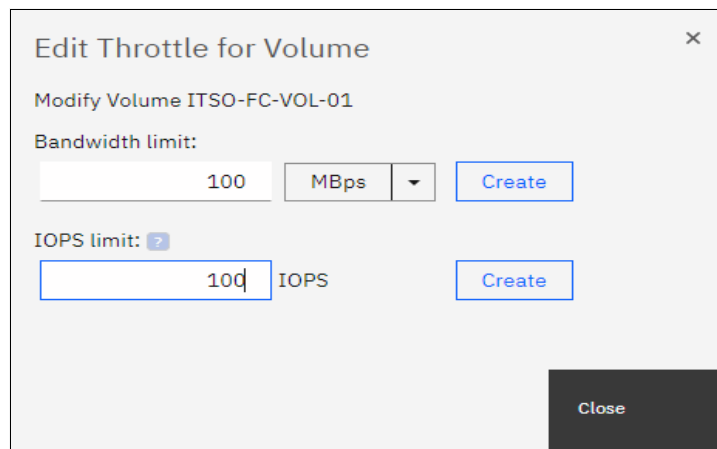


Figure 5-21 Edit bandwidth and IOPS limit

Throttling at a volume level can be set by using the following two commands:

► **mkthrottle**

This command is used to set I/O throttles for volumes using this command. It must be used with **-type vdisk** parameter, followed by **-bandwidth bandwidth_limit_in_mbdisk** or **-iops iops_limit** to define MBps and IOPS limits.

► **chvdisk**

When used with the **-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 megabits per second (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 is used to not only create throttles for volumes, but also for hosts, host clusters, pools, and system offload.

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 throttle is set to a value smaller than 100 IOPS, the actual throttling occurs at 100 IOPS.

After you use any of the commands 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
```

For more information, and the procedure to set volume throttling, see [IBM FlashSystem 9200 8.4.0 Documentation - Managing throttles for volumes](#).

5.12 Volume cache mode

Cache mode in IBM FlashSystem determines if read and write operations are stored in cache. By default, when a volume is created, the cache mode is set to *readwrite* (Enabled). Disabling cache can impact performance, which can increase read and write response time.

For each volume, one of the following cache modes can be used:

- ▶ **readwrite** (enabled)

All read and write I/O operations that are performed by the volume are stored in cache. This is the default cache mode 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 it will decrease performance. In a DRP, it is not possible to create a thin-provisioned or compressed0-volume copy with a cache-mode setting 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 and write I/O operations that are performed by the volume are not stored in cache.

Figure 5-22 shows write-operation behavior when volume cache is activated (**readwrite**).

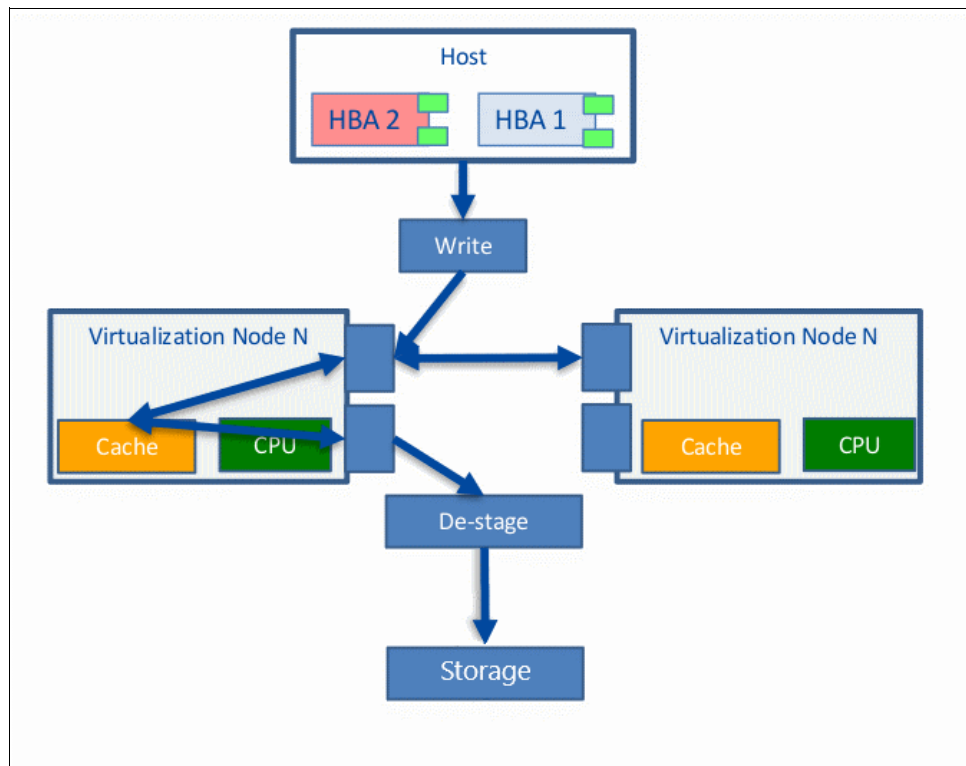


Figure 5-22 Cache activated

Figure 5-23 shows a write operation behavior when volume cache is deactivated (*none*).

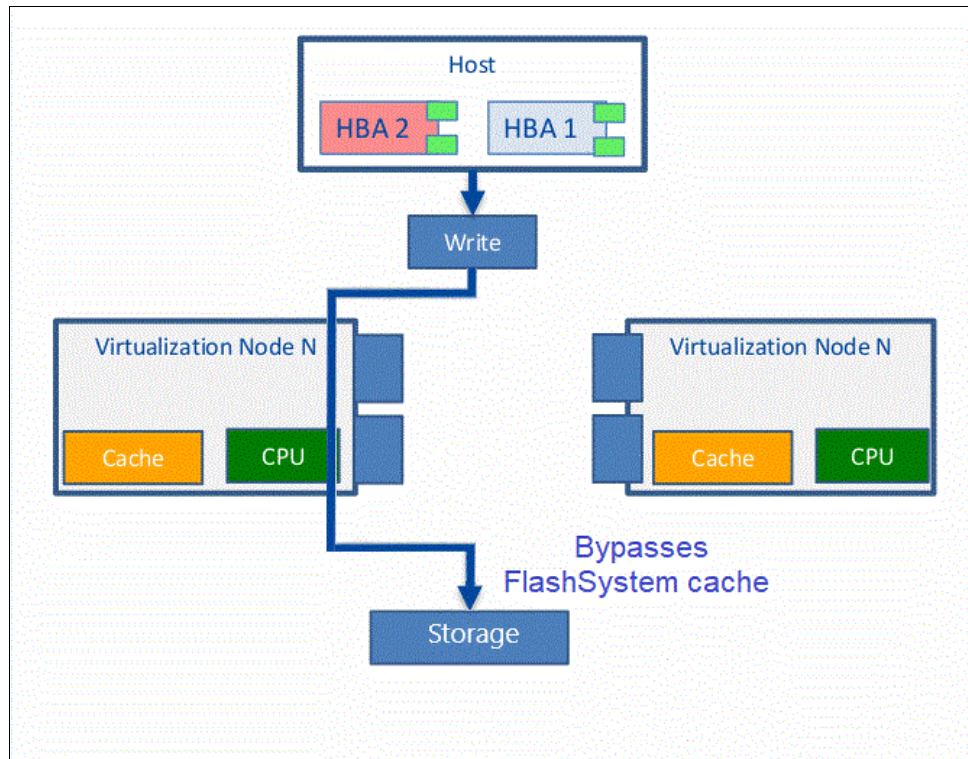


Figure 5-23 Cache deactivated

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 backend storage controller, and these volumes are virtualized in IBM FlashSystem devices as image VDisks. Another possible use of cache-disabled volumes is when intellectual capital is in existing copy services automation scripts. Keep the use 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 certain volumes, more cache resources are available to cache I/Os to other volumes in the same I/O group. For example, a non-critical application that uses volumes in MDisks from all-flash storage.

Even if this application generates many IOPS and requires low response time, you can disable the cache of the volumes in IBM FlashSystem. This action does not greatly impact on the application and results in a consumption of less resources in the storage system.

Note: Volumes with readwrite cache enabled are recommended.

By default, volumes are created with cache mode enabled (**readwrite**), but 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 GUI, selecting **Volumes** → **Volumes** → **Actions** → **Cache Mode**. Figure 5-24 on page 223 shows editing cache mode for a volume.

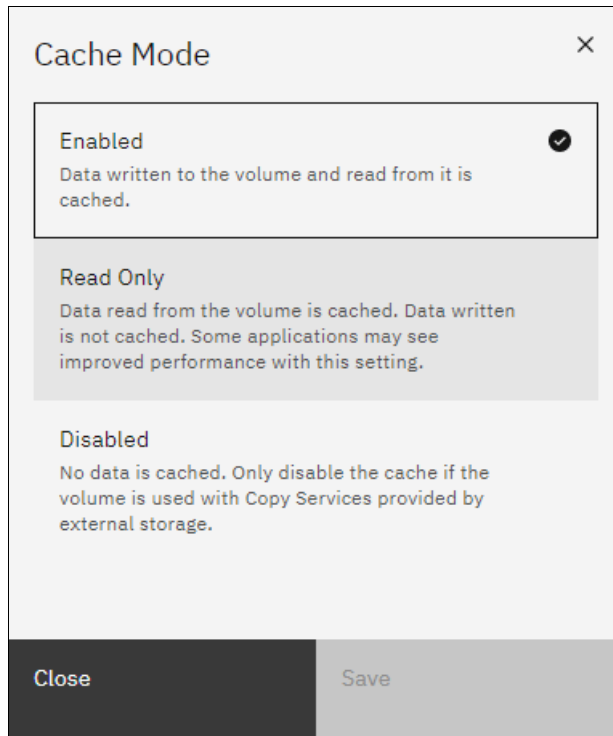


Figure 5-24 Edit cache mode

The command line will 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 will 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, Metro Mirror, Global Mirror, and Volume Mirroring) and typical workloads, disabling IBM FlashSystem cache is detrimental to overall performance.

Attention: Carefully evaluate the impact to the entire system with quantitative analysis before and after making this change

5.13 Additional considerations

The following section describes additional and brief considerations regarding volumes.

5.13.1 Volume protection

You can protect volumes to prevent active volumes or host mappings from being deleted. IBM FlashSystem has a global setting enabled by default that prevents these objects from being deleted if the system detects recent I/O activity.

To protect volumes, you can do one of the following:

- ▶ Set the system-level volume protection value to apply to all volumes that are configured on your system.
- ▶ Control whether the system-level volume protection is enabled or disabled on specific pools.

To prevent an active volume from being deleted unintentionally, administrators should enable volume protection. They can also specify a time period that the volume must be idle before it can be deleted. If volume protection is enabled and the time 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 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.

To enable volume protection, use the following command:

```
chsystem vdiskprotectionenabled yes -vdiskprotectiontime <value_in_minutes>
```

To enable volume-protection in your system, but disable volume-protection in a specific storage pool, use the following command:

```
chmdiskgrp -vdiskprotectionenabled no <pool_name_or_ID>
```

You can also manage volume protection in GUI, navigating through **Settings** → **System** → **Volume Protection**, as shown in Figure 5-25 on page 225.

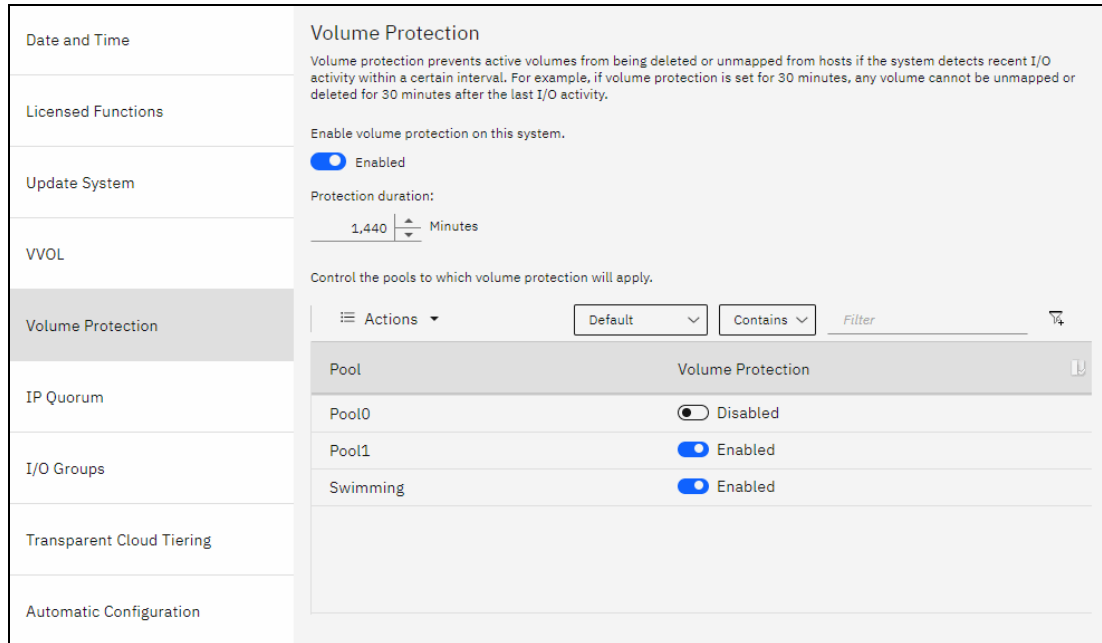


Figure 5-25 Volume Protection

5.13.2 Volume resizing

Fully-allocated and thin-provisioned volumes can have their sizes increased or decreased. A volume can be expanded with concurrent I/Os for some operating systems, but you should never attempt to shrink a volume in use that contains data, because volume capacity is removed from the end of the disk, whether or not that capacity is in use by a server. Remember that 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 particular volume that is already mapped to a host.
- ▶ To increase the size of a volume aiming to make it match the size of the source or master volume so that it can be used in a FlashCopy mapping or Metro Mirror relationship.

Figure 5-26 shows the Expand Volume window.

Expand Volume - ITSO-FC-VOL-01

i You selected to expand the provisioned capacity of volume ITSO-FC-VOL-01. This action increases the available capacity in the volume.

Current size: GiB

Expand by: GiB ▾

Final size: GiB

Format: Enabled

Maximum final size: 106,578.00 GiB

Cancel Expand

Figure 5-26 Expanding a volume

Shrinking a volume

Volumes can be reduced in size if necessary. If a volume does not contain any data, there should be no issues to shrink its size. However, if a volume is in use and contains data, do not shrink its size, because IBM Spectrum Virtualize will not be aware if it is removing used or non-used capacity.

Attention: When you shrink a volume, capacity is removed from the end of the disk, whether or not that capacity is in use. Even if a volume has free capacity, do not assume that only unused capacity is removed when you shrink a volume.

Figure 5-27 shows shrinking volumes.

Shrink Volume - ITSO-FC-VOL-02

! If this volume contains data that is being used, do not attempt to shrink a volume under any circumstances without first backing up your data. You selected to shrink the capacity of volume ITSO-FC-VOL-02. The system arbitrarily reduces the capacity of the volume by removing one or more extents that are assigned to the volume. You cannot control which extents are removed and cannot ensure that unused capacity is removed.

Current size: GiB

Shrink by: GiB ▾

Final size: GiB

Cancel Shrink

Figure 5-27 Shrinking a volume

5.13.3 Migrating from Fibre Channel connections to RDMA over Ethernet connections between nodes

IBM FlashSystem 9100, 9200, 7200, and 5100 support node-to-node connections that use Ethernet protocols that support remote direct memory access (RDMA) technology, such as RDMA over Converged Ethernet (RoCE) or iWARP. To use these protocols, the system requires that an RDMA-capable adapter is installed on each node and dedicated RDMA-capable Ethernet ports are only configured for node-to-node communication. If your system currently uses Fibre Channel ports, you can migrate to RDMA-capable Ethernet ports for node-to-node communications.

RDMA technologies, such as RoCE and iWARP, enable the RDMA-capable adapter to transfer data directly between nodes, and bypass CPU and cache, which makes transfers faster. RDMA technologies provide faster connection and processing time than traditional iSCSI connections.

The following prerequisites must be met for all RDMA-capable Ethernet ports that are used between nodes:

- ▶ All node hardware is installed.
- ▶ The 25-Gbps Ethernet adapter that supports RDMA technology is installed on each node. If you are using RDMA-technology for node-to-node communications, ensure that the RDMA-capable adapters use the same technology, such as RoCE or iWARP. These RDMA-capable adapters must be installed in the same slots across all the nodes of the system. These installation requirements ensure that port identifiers are the same across all nodes in the system.
- ▶ Ethernet cables between each node are connected correctly.
- ▶ The protocol technology on the source and destination adapters is the same.
- ▶ The local and remote IP addresses can be reached.
- ▶ Each IP address for RDMA-capable Ethernet ports and their associated subnet masks are unique on each node.
- ▶ The router is not placed between nodes that use RDMA-capable Ethernet ports for node-to-node communication.
- ▶ The negotiated speeds on the local and remote adapters are the same.
- ▶ The local and remote port virtual LAN identifiers are the same. Use virtual LAN to create physical separation of networks for unrelated systems, wherever possible. All the ports that are used for node-to node communication must be assigned with one VLAN ID and ports that are used for host attachment must have a different VLAN ID.

If you plan to use VLAN to create this separation, you must configure VLAN support on the all the Ethernet switches in your network before you define the RDMA-capable Ethernet ports on nodes in the system. On each switch in your network, set VLAN to Trunk mode and specify the VLAN ID for the RDMA-ports that will be in the same VLAN.

In addition, if VLAN settings for a RDMA-capable Ethernet port need to be updated, these settings cannot be updated independently of other configuration settings. Before you update VLAN settings on specific RDMA-capable Ethernet ports, you must unconfigure the port, make necessary changes to the switch configuration, then reconfigure RDMA-capable Ethernet ports on each of the nodes in the system.

- ▶ A minimum of two dedicated RDMA-capable Ethernet ports is required for node-to-node communications to ensure best performance and reliability. 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.

- ▶ A maximum of four RDMA-capable Ethernet ports per node is allowed for node-to-node communications.



Copy services

Copy services are a collection of functions that provide capabilities for disaster recovery, data migration, and data duplication solutions. This chapter provides an overview and the preferred practices of IBM FlashSystem copy services capabilities, including FlashCopy, Metro Mirror and Global Mirror, and Volume Mirroring.

This chapter includes the following sections:

- ▶ 6.1, “Introduction to copy services” on page 230
- ▶ 6.2, “FlashCopy” on page 231
- ▶ 6.3, “Remote Copy services” on page 257
- ▶ 6.4, “Native IP replication” on page 314
- ▶ 6.5, “Volume Mirroring” on page 330

6.1 Introduction to copy services

IBM Spectrum Virtualize based systems, including the IBM FlashSystem family, offer a complete set of copy services functions that provide capabilities for disaster recovery, business continuity, data movement, and data duplication solutions.

6.1.1 FlashCopy

FlashCopy is a function that allows you to create a point-in-time copy of one of your volumes. This function might be helpful when performing backups or application testing. These copies can be cascaded on one another, read from, written to, and even reversed. These copies are able to conserve storage, if needed, by being space-efficient copies that only record items that have changed from the originals instead of full copies.

6.1.2 Metro Mirror and Global Mirror

Metro Mirror and Global Mirror are technologies that enable you to keep a real-time copy of a volume at a remote site that contains another IBM Spectrum Virtualize based system:

- ▶ Metro Mirror makes synchronous copies of your volumes. This means that the original writes are not considered complete until the write to the destination volume has been confirmed. The distance between your two sites is usually determined by the amount of latency your applications can handle.
- ▶ Global Mirror makes *asynchronous* copies of your volumes. This means that the write is considered complete after it is complete at the local volume. It does not wait for the write to be confirmed at the remote system as Metro Mirror does. This requirement greatly reduces the latency experienced by your 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 committed to the local volume.

IBM Spectrum Virtualize provides two types of asynchronous mirroring technology:

- The standard Global Mirror (referred to as Global Mirror)
- The Global Mirror with Change Volume (GMCV)

6.1.3 Volume Mirroring

Volume Mirroring is a function that is designed to increase high availability of the storage infrastructure. It provides the ability to 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 also 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 migrate from a thin-provisioned volume to a non-thin-provisioned volume, and to migrate data between storage pools of different extent sizes.

6.2 FlashCopy

By using the IBM FlashCopy function of the IBM FlashSystem, you can perform a *point-in-time copy* of one or more volumes. This section describes the inner workings of FlashCopy, and provides some preferred practices for its use.

You can use FlashCopy to help you 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 FlashCopies.

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. FlashCopy can also transfer the resource usage of performing intensive backups from production systems.

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 use of these methods puts less load on your infrastructure.

When FlashCopy is used for backup purposes, the target data usually is 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.

Restore 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.

Preferred practices: Although restoring from a FlashCopy is quicker than a traditional tape media restore, do not use restoring from a FlashCopy as a substitute for good archiving practices. Instead, keep one to several iterations of your FlashCopies so that you can near-instantly recover your data from the most recent history. Keep your long-term archive as appropriate 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, 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 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 that for your testing. This copy is a duplicate of your production data down to the block level so that even physical disk identifiers are copied. Therefore, it is impossible for your applications to tell the difference.

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 FlashSystem systems support 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. You can perform the actions of preparing, starting, or stopping FlashCopy on either a stand-alone mapping or a Consistency Group.

Figure 6-1 shows the concept of FlashCopy mapping.

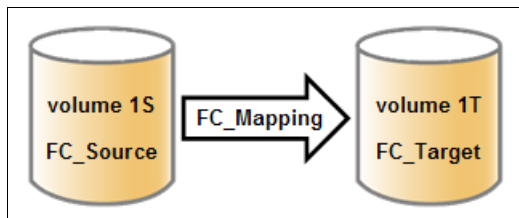


Figure 6-1 FlashCopy mapping

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 allows 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 has 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 is complete. 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 shows 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 Spectrum Virtualize features, it is advised not to exceed a copyrate value of 130.

FlashCopy Consistency Groups

Consistency Groups 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 Consistency Groups are used, the FlashCopy commands are issued to the Consistency Groups. The groups perform the operation on all FlashCopy mappings contained within the Consistency Groups at the same time.

Figure 6-2 illustrates a Consistency Group consisting of two volume mappings.

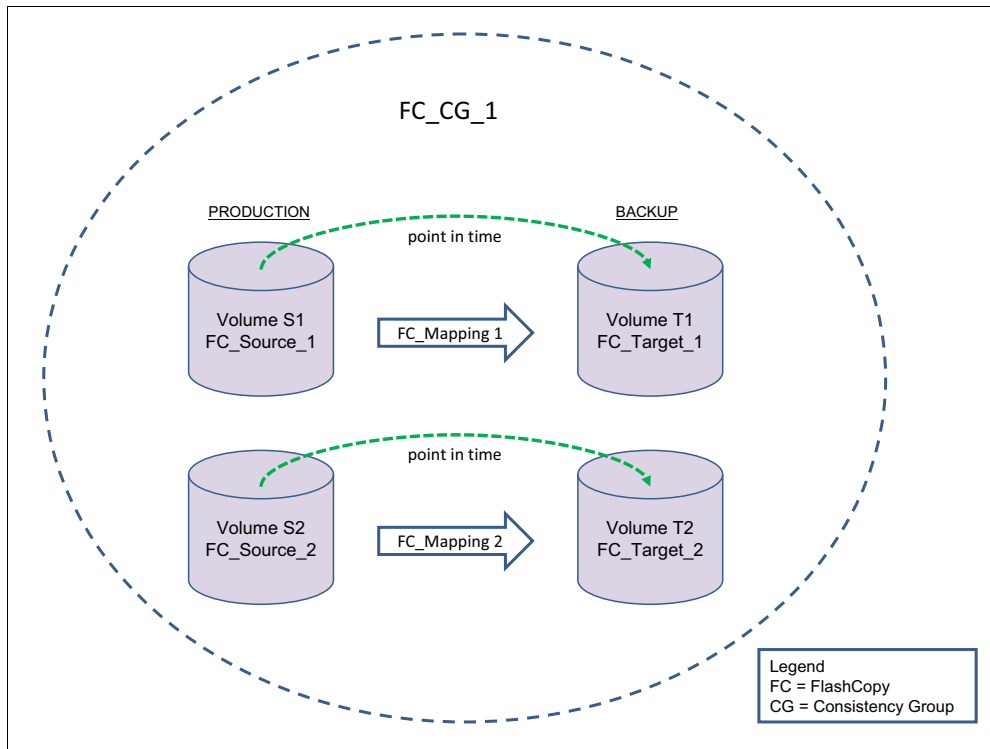


Figure 6-2 Multiple volumes mapping in a Consistency Group

FlashCopy mapping considerations: If the FlashCopy mapping has been added to a Consistency Group, it can only be managed as part of the group. This limitation means that FlashCopy operations are no longer allowed on the individual FlashCopy mappings.

Incremental FlashCopy

Using Incremental FlashCopy, you can reduce the required time of copy. Also, because less data must be copied, the workload put on the system and the back-end storage is reduced.

Basically, Incremental FlashCopy does not require that you copy an entire disk source volume every time the FlashCopy mapping is started. It means that only the changed regions on source volumes are copied to target volumes, as shown in Figure 6-3.

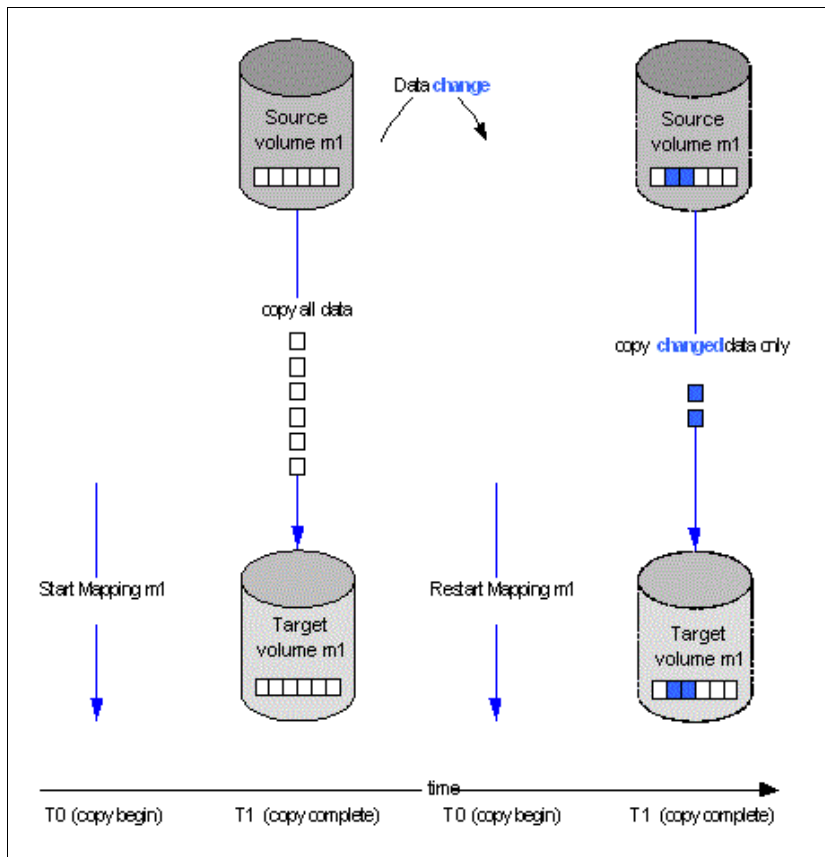


Figure 6-3 Incremental FlashCopy

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 was 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 has 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 has been 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 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.

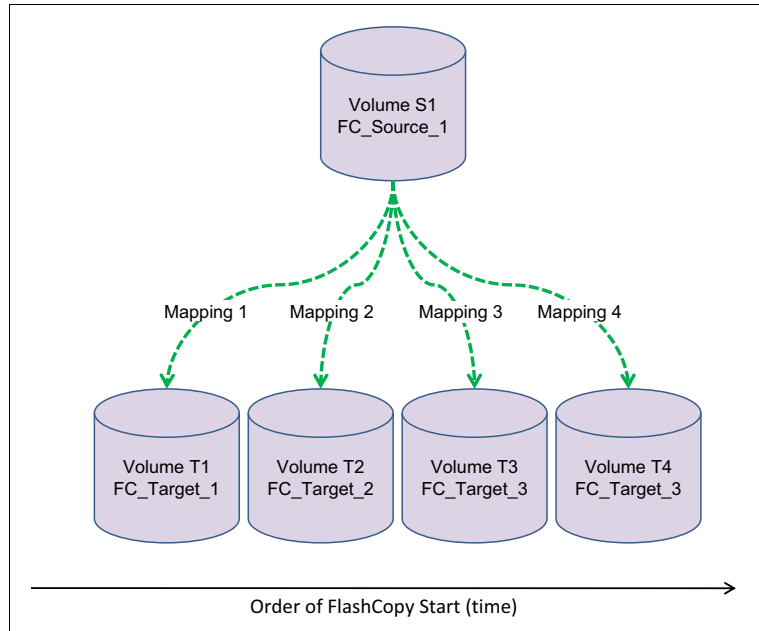


Figure 6-4 Multiple Target FlashCopy

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 Consistency Groups. In cases where all the mappings are in the same Consistency Group, the result of starting the Consistency Group 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; this is referred to as a *Cascaded FlashCopy*. This function is illustrated in Figure 6-5.

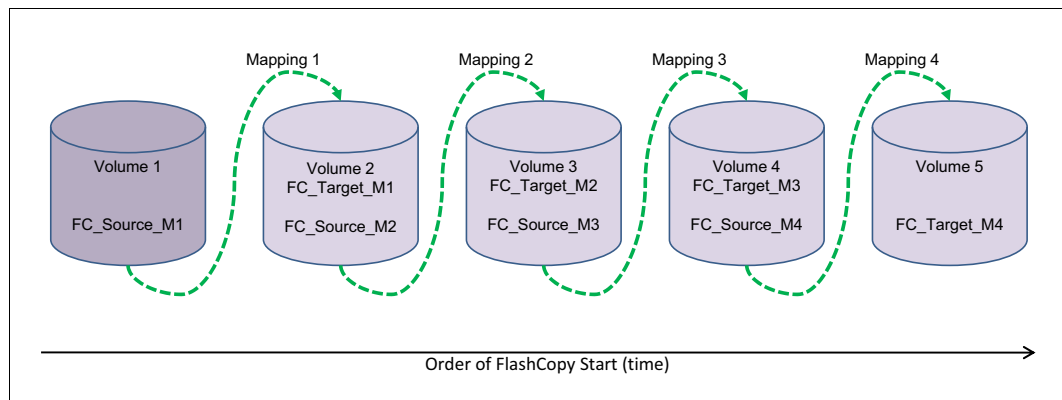


Figure 6-5 Cascaded FlashCopy

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 in combination 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 FlashSystem systems also allow 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 Spectrum Virtualize based 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. You can then increase or shrink the real capacity as the application 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. It can be used for incremental FlashCopy too, but using thin-provisioned volumes for incremental FlashCopy only makes sense if the source and target are thin-provisioned.

When using thin provisioned volumes on Data Reduction Pools (DRPs), consider also implementing compression because it provides several benefits:

- ▶ Reduced amount of I/O operation to the back-end as the amount of data to be actually written to the back-end reduces with compressed data. This is particularly relevant with a

poorly performing back-end, but less of an issue with the high performing back-end on IBM FlashSystem systems.

- ▶ Space efficiency as the compressed data provides more capacity savings.
- ▶ Better back-end capacity monitoring, as 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 use 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 the FlashCopy is performed, only data that has been modified is recopied to the target. Note that 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 using FlashCopy to another fully allocated volume at the same time as 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. At the same time, 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 Consistency Group. You can perform the actions of preparing, starting, or stopping FlashCopy on either a stand-alone mapping or a Consistency Group.

A FlashCopy mapping has an attribute that represents the state of the mapping. The FlashCopy states are the following:

Idle_or_copied

Read and 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 242) 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 and writes are executed on the target as though the contents of the source were instantaneously copied to the target during

the **startfcmaporstartfcconsistgrp** command. The source and target can be independently updated. Internally, the target depends on the source for certain tracks. Read and 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 will be lost) or the FlashCopy mapping must be deleted. The source volume is accessible, and read and write caching is enabled for the source. In the Stopped state, a mapping can either be prepared again or deleted.

Stopping

The mapping is in the process of transferring data to a dependent mapping. The behavior of the target volume depends on whether the background copy process had completed while the mapping was in the Copying state. If the copy process had completed, the target volume remains online while the stopping copy process completes. If the copy process had not completed, 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 has been copied, a stop complete asynchronous event notification is issued. The mapping will move to the Idle/Copied state if the background copy has completed or to the Stopped state if the background copy has not completed. 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 has been halted. When the metadata becomes available again, the FlashCopy mapping will return to the Copying or Stopping state. Access to the source and target volumes will be restored, and the background copy or stopping process will resume. 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 in the process of preparing the mapping. While in this state, data from cache is destaged to disk and a consistent copy of the source exists on disk. At this time, cache is operating in write-through mode and therefore writes to the source volume will experience additional latency. The target volume is reported as online, but it will not perform reads or writes. These reads and writes are failed by the 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 also 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 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. Flushing any modified write data associated with the source volume from the cache. Read data for the source will be left in the cache.
2. Placing the cache for the source volume into write-through mode, so that subsequent writes wait until data has been written to disk before completing the write command that is received from the host.
3. Discarding 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 additional 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.

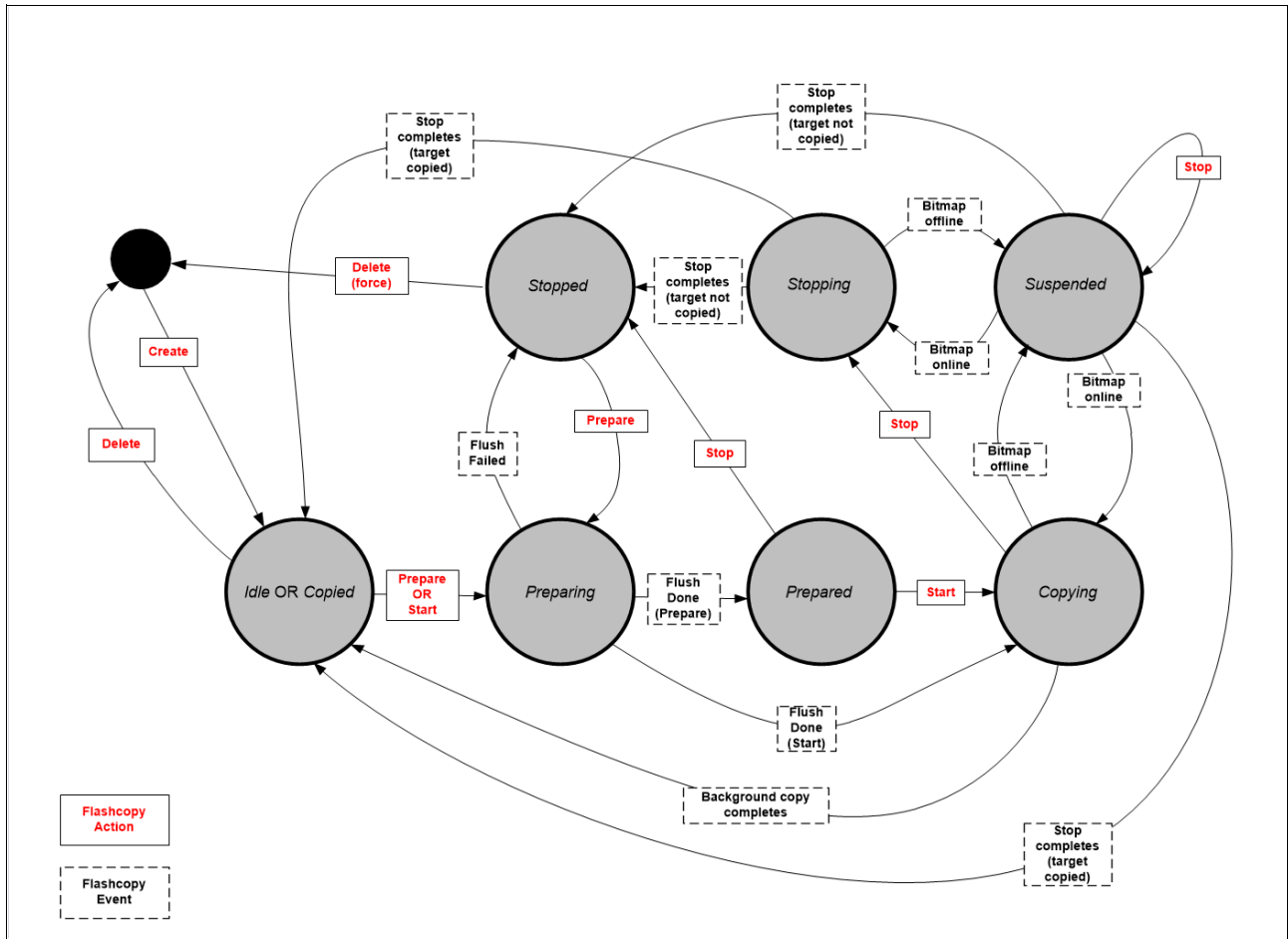


Figure 6-6 FlashCopy mapping states diagram

FlashCopy bitmaps and grains

A *bitmap* is an internal data structure stored in a particular I/O Group that is used to track which data in FlashCopy mappings has been copied from the source volume to the target volume. *Grains* are units of data grouped together to optimize the use of the bitmap. One bit in each bitmap represents the state of one grain. 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 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 and 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 Consistency Group.
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 Consistency Group (creating the FlashCopy bitmap).
5. Ensure that the indirection layer governs all of 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 Consistency Group. The indirection layer then determines how each of the I/O is to be 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 the I/O to go through the underlying volume preserving the point-in-time copy. In order to do that, the Spectrum 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.

Spectrum 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 IOs for the write operations, while a significant capacity-saving can be achieved by improving the overall deduplication ratio.

The RoW has been introduced with Spectrum Virtualize version 8.4 and it is used in the following conditions:

- ▶ Source and target volumes in the same pool
- ▶ Source and target volumes in the same IO group
- ▶ The pool containing 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 summarizes the indirection layer algorithm in case of 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 grain to the most recently started target for this source, 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 has already been 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 has been 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.

Interaction with cache

The Spectrum Virtualize technology provides a two-layer cache, as follows:

- ▶ *Upper cache* serves mostly as write cache and hides the write latency from the hosts and application.
- ▶ *Lower cache* is a read/write cache and optimizes I/O to and from disks.

Figure 6-7 shows the IBM Spectrum Virtualize cache architecture.

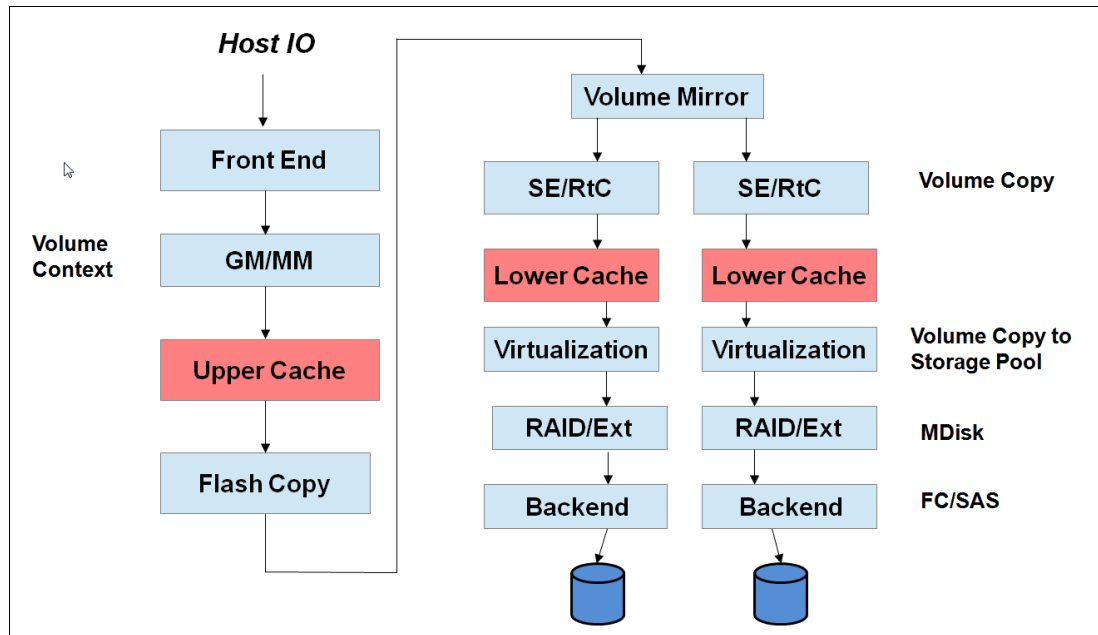


Figure 6-7 New cache architecture

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.

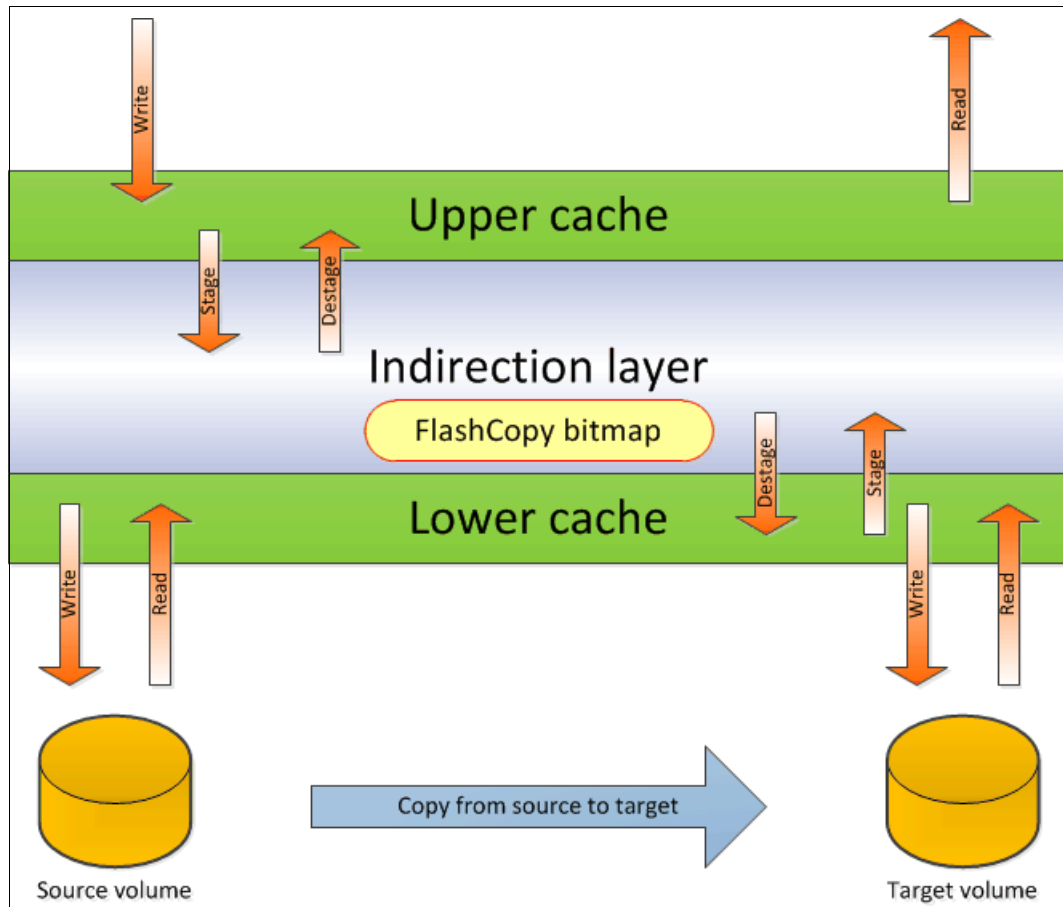


Figure 6-8 Logical placement of the FlashCopy indirection layer

The introduction of the two-level cache provides additional performance improvements to the FlashCopy mechanism. Because the FlashCopy layer is now above the lower cache in the IBM Spectrum Virtualize software stack, it can benefit from read pre-fetching and coalescing writes to back-end storage. Also, preparing FlashCopy is much faster because upper cache write data does not have to go directly to back-end storage, but just to the lower cache layer.

Additionally, in multi-target FlashCopy, the target volumes of the same image share cache data. This design is opposite to previous IBM Spectrum Virtualize code versions, where each volume had its own copy of cached data.

Interaction and dependency between Multiple Target FlashCopy mappings

Figure 6-9 on page 246 represents a set of three FlashCopy mappings that share a common source. The FlashCopy mappings target target volumes Target 1, Target 2, and Target 3.

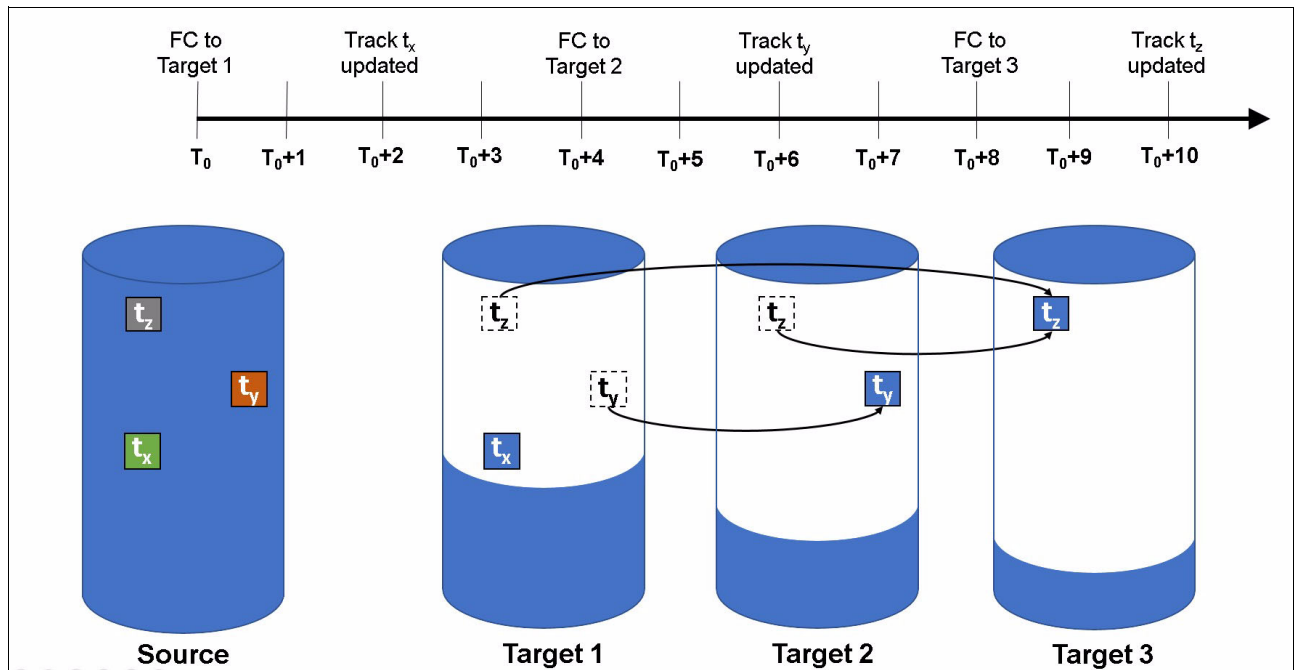


Figure 6-9 Interaction between Multiple Target FlashCopy mappings

Consider the following events timeline:

- ▶ At time T_0 a Flashcopy mapping is started between the source and the Target 1.
- ▶ At time T_0+2 the track t_x is updated in the source. Since this track has not yet been copied in background on Target 1, the copy-on-write process copies this track to the Target 1 before being updated on the source.
- ▶ At time T_0+4 a Flashcopy mapping is started between the source and the Target 2.
- ▶ At time T_0+6 the track t_y is updated in the source. Because this track has not yet been copied in background on Target 2, the copy-on-write process copies this track to the Target 2 only before being updated on the source.
- ▶ At time T_0+8 a Flashcopy mapping is started between the source and the Target 3.
- ▶ At time T_0+10 the track t_z is updated in the source. Because this track has not yet been copied in background on Target 3, the copy-on-write process copies this track to the 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 is dependent upon Target 2 and Target 3. It remains dependent until all of Target 1 has been copied. No target depends on Target 1, so the mapping can be stopped without need to copy any data to maintain the consistency in the other targets.
- ▶ Target 2 depends on Target 3, and will remain dependent until all of Target 2 has been 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 is not dependent 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 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 yet been copied, 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 being read has already been copied from the source to the target, the read simply returns data from the target being read. If the grain has not been copied, each of the newer mappings is examined in turn and the read is performed from the first copy 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 FlashSystem products features, offers useful capabilities. However, some basic planning considerations are to 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 shows the system limits that apply to the latest version at the time of writing this book.

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	10000	This maximum is the maximum number of FlashCopy mappings per system.
FlashCopy Consistency Groups 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 Consistency Group	512	This limit is due to the time that is taken to prepare a Consistency Group with many mappings.

Configuration Limits: The configuration limits always change with the introduction of new hardware and software capabilities. Check the IBM FlashSystem online documentation for the latest configuration limits.

The total amount of cache memory reserved for the FlashCopy bitmaps limits the amount of capacity that can be used as a FlashCopy target. Table 6-4 illustrates 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 specified I/O Group

Copy Service	Grain size in 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, for a mapping with a 256 KB grain size, 8 KB of memory allows two mappings between one 8 GB source volume and two 8 GB target volumes.

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 towards the specified I/O Group, not towards 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 `chlogrp` 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 towards the specified I/O Group, not towards 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 in a FlashCopy mapping must be the same.
- ▶ 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 no multiple FlashCopy can be active to the same target volume.
- ▶ Expansion or shrinking of volumes defined in a FlashCopy mapping is not allowed. To modify the size of a source or target volume, first remove the FlashCopy mapping.
- ▶ 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 consistency group while the FlashCopy mapping status is *Copying*.
- ▶ No FlashCopy mapping can be added to a consistency group while the consistency group status is *Copying*.
- ▶ The use of Consistency Groups is restricted when using Cascading FlashCopy. A Consistency Group serves the purpose of starting FlashCopy mappings at the same point in time. Within the *same* Consistency Group, 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 Consistency Group, mappings cannot be established in a certain order. This limitation renders the content of the target volume undefined. For instance, 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 Consistency Group, the result is equal to Multi 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 FlashCopies), rather than at the same time in an undefined order (from within one single Consistency Group).

- ▶ Both source and target volumes can be used as primary in a Remote Copy relationship. For more details about the FlashCopy and the Remote Copy possible interactions, see “Interaction between Remote Copy and FlashCopy” on page 290.

FlashCopy presets

The IBM FlashSystem GUI interface provides three FlashCopy presets (Snapshot, Clone, and Backup) to simplify the more common FlashCopy operations. Figure 6-10 on page 250 shows the preset selection panel in the GUI.

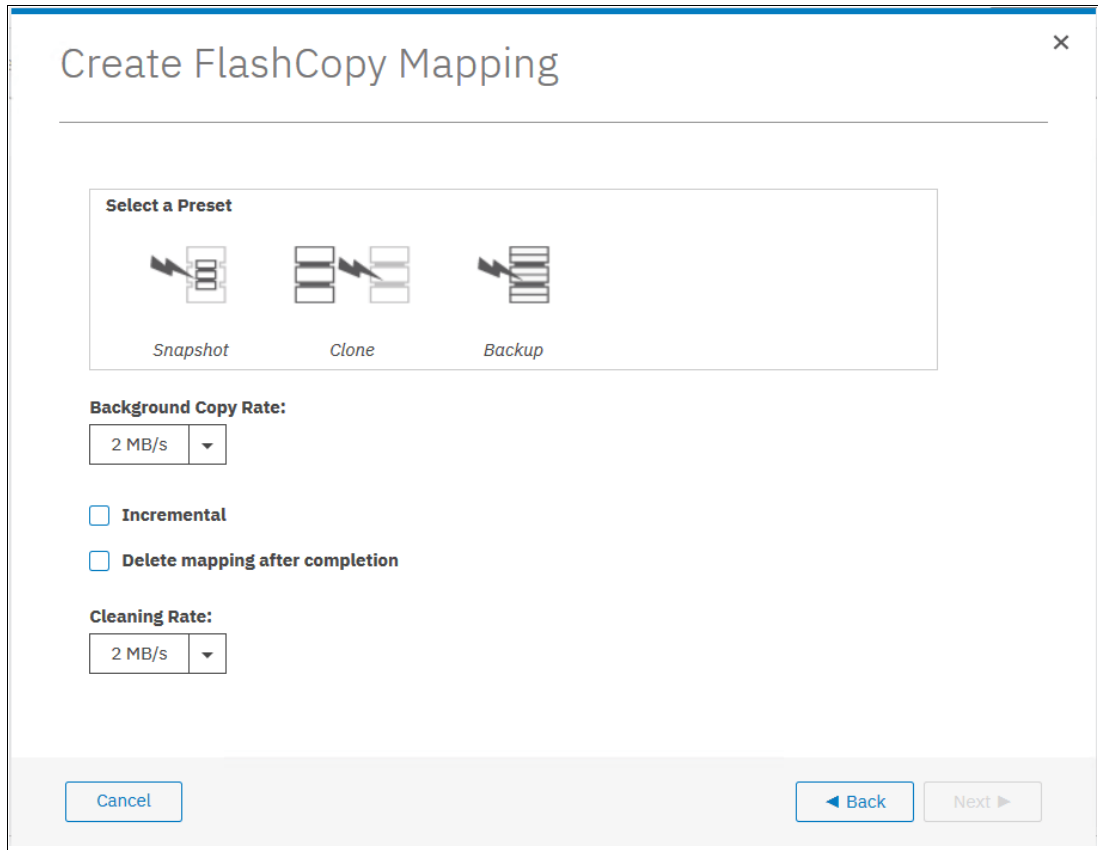


Figure 6-10 GUI Flashcopy Presets

Although these presets meet most FlashCopy requirements, they do not provide support for all possible FlashCopy options. If more specialized options are required that are not supported by the presets, the options must be performed by using CLI commands.

This section describes the three preset options and their use cases.

Snapshot

This preset creates a copy-on-write point-in-time copy. 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
- ▶ Primary copy source pool: Target 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
- ▶ Primary copy source pool: Target 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
- ▶ Primary copy source pool: Target 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 new 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 in conjunction with thin provisioned target volumes, usually the no-copy option is 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 from 0 to 100% of the virtual capacity.

In case of thin provisioned volumes used as FlashCopy target, it is important to provide a non-zero real size. This is because when the FlashCopy is initiated, the copy-on-write process requires to allocate capacity on the target volumes. If some capacity is not yet allocated, as with thin provisioned volumes with zero real size, the write IO can be delayed until the capacity is made available. Usually the write caching hides this impact, but in case of heavy write workload, the performance can be affected.

Real size: The recommendation with thin provisioned target volumes is to assign at least 2GB of real capacity.

DRP Optimized Snapshots: Thin provisioned FlashCopy can greatly benefit from the Redirect-on-Write capability introduced with Spectrum Virtualize version 8.4. For further details and restrictions, see “Indirection layer” on page 242.

Grain size considerations

When creating a mapping a grain size of 64 KB can be specified as compared to the default 256 KB. This smaller grain size has been introduced specifically for the incremental FlashCopy, even though its use is not restricted to the incremental mappings.

In an 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.

The following are the preferred settings for thin-provisioned FlashCopy:

- ▶ Thin-provisioned volume grain size should be equal to the FlashCopy grain size. Anyway 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.
- ▶ Thin-provisioned volume grain size must be 64 KB for the best performance and the best space efficiency.

The exception is where the thin target volume is going to become a production volume (and is 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 very large numbers of FlashCopy/Remote Copy relationships might be forced to choose a 256 KB grain size for FlashCopy to avoid constraints on the amount of bitmap memory.

Note that Cascading FlashCopy and Multi Target FlashCopy require all the mappings participating to the FlashCopy chain to have the same grain size. For more information, see “FlashCopy general restrictions” on page 248.

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 condition (that is with all the nodes/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

	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 source volume's I/O group	Preferred node in target volume's I/O group	Preferred node in source volume's I/O group	Preferred node in target volume's I/O group
Node that performs the back-end I/O if the grain is not yet copied	Preferred node in source volume's I/O group	Preferred node in source volume's I/O group	The preferred node in source volume's I/O group will read and write, and the preferred node in target volume's I/O group will write	The preferred node in source volume's I/O group will read, and the preferred node in target volume's I/O group will write

Note that 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:

1. Source and target volumes use the same preferred node.

In this scenario, the node that is acting as preferred for both source and target volume manages all the read and write FlashCopy operations. Only resources from this node are consumed for the FlashCopy operations, and no node-to-node bandwidth is used.

2. Source and target volumes use the different preferred node.

In this scenario, both nodes that are acting as preferred nodes manage read and write FlashCopy operations according to the schemes described above. The data that is transferred between the two preferred nodes goes through the node-to-node network.

Both alternatives 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 of the source and target volumes on the back-end storage. For example, if the source volume is on site A and the target volume is on site B, then the target volumes preferred node must be in site B. Placing the target volumes preferred node in site A will cause 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 should follow the location of source and target volumes on the internal storage. For example, if the source volume is on the internal storage attached to control enclosure A and the target volume is on internal storage 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 will cause the re-direction 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 exploit the Redirect-on-Write capability introduced with Spectrum Virtualize version 8.4, check the volume placement restrictions described in “Indirection layer” on page 242

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 (as shown in Table 6-1 on page 234).

If the copy process is unable to achieve these goals, it starts contending resources to the foreground I/O (that is the I/O coming from the hosts). As result, both background copy and foreground I/O will tend to see an increase in latency and therefore reduction in throughput compared to the situation when the bandwidth not been limited. Degradation is graceful. Both background copy and foreground I/O continue to make progress, and will 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, taking in account the overall workload running in the systems. The background copy basically reads and writes data to managed disks. 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 instance, starting 20 FlashCopy with a background copy rate of 70 each 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 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 is already started. The background copy process can even be completely 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. Note that the background copy rate setting follows an exponential scale, so changing, for instance, 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 FlashCopies such as Cascaded and Multi 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 Multi Target FlashCopy that is not the oldest in the FlashCopy chain. In this case to avoid the stopping state lasting for a long time, the cleaning rate can be adjusted accordingly.

There is an interaction 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 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 is started. When the mapping is stopped, it goes into 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 is started. When the mapping is stopped, it goes into 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 is 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 FlashSystem systems, and therefore 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. FlashCopies that are crash consistent often can be used following file system and application recovery procedures.

Note: Although the best way to perform FlashCopy is to flush host cache first, some companies, such as Oracle, support using snapshots without it, as stated in Metalink note 604683.1.

Various operating systems and applications provide facilities to stop I/O operations and ensure that all data is flushed from 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.

Preferred 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 Remote Copy services

IBM FlashSystem technology offers various Remote Copy services functions that address Disaster Recovery and Business Continuity needs.

Metro Mirror is designed for metropolitan distances with a zero recovery point objective (RPO), which is zero data loss. This objective is achieved with a synchronous copy of volumes. Writes are not acknowledged 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, and includes round-trip network latencies. Metro Mirror has the following characteristics:

- ▶ Zero RPO
- ▶ Synchronous
- ▶ Production application performance that is affected by round-trip latency

Global Mirror technologies are designed to minimize the network latency effects by replicating asynchronously. Spectrum Virtualize provides two types of asynchronous mirroring technology:

- ▶ The standard Global Mirror (referred to as Global Mirror)
- ▶ The Global Mirror with Change Volume (GMCV)

With the Global Mirror, writes are acknowledged as soon as they can be committed to the local storage system, sequence-tagged, and passed on to the replication network. This technique allows Global Mirror to be used over longer distances. By definition, any vendors' asynchronous replication results in an RPO greater than zero. However, for Global Mirror, the RPO is quite small, typically anywhere from several milliseconds to some number of seconds.

Although Global Mirror is asynchronous, the RPO is still small, and thus the network and the remote storage system must both still be able to cope with peaks in traffic. Global Mirror has the following characteristics:

- ▶ Near-zero RPO
- ▶ Asynchronous
- ▶ Production application performance that is affected by I/O sequencing preparation time

GMCV provides an option to 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 Global Mirror with Change Volumes is higher than traditional Global Mirror. Global Mirror with Change Volumes has the following characteristics:

- ▶ Larger RPO

- ▶ Point-in-time copies
- ▶ Asynchronous
- ▶ Possible system performance effect because point-in-time copies are created locally

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 SAN configurations, local and remote system and storage configuration, and the intersystem network.

6.3.1 Remote Copy use cases

Data replication techniques are the foundations of Disaster Recovery and Business Continuity solutions. Besides these common use cases, Remote Copy technologies can be used in other data movement scenarios, as 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 a Spectrum Virtualize-based system to another, while applications remain online. After the volumes are fully copied and synchronized, the application can be stopped and then immediately started on the new storage system.

Data center moving

Remote Copy functions can be used to move data between Spectrum Virtualize-based systems in order to facilitate data centers 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.

6.3.2 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*
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 nonrestricted access. Mapped hosts can read and write to the volume.
- ▶ *Intersystem link or intersystem network*
The network that provides connectivity between the local and the remote site. It can be a Fibre Channel network (SAN), an IP network, or a combination of the two.
- ▶ *Remote system or auxiliary 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 a Metro Mirror or Global Mirror 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. However, having remote copies is not a prerequisite. 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. Metro Mirror 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. Global Mirror 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 intersystem network and remote SAN.

Figure 6-11 on page 260 shows some of the concepts of remote copy.

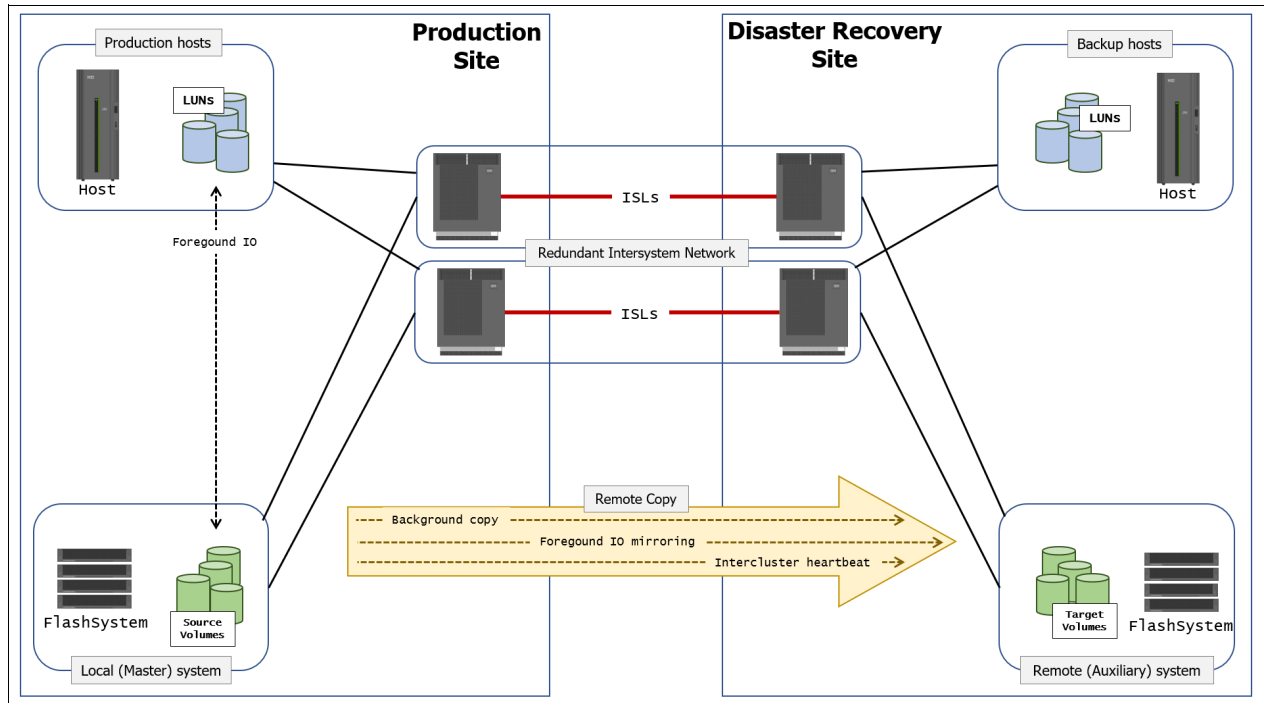


Figure 6-11 Remote Copy components and applications

A successful implementation of intersystem Remote Copy services significantly depends on the quality and configuration of the intersystem 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 6-12 on page 261.

Partnerships are established between two systems by issuing the **mkfcpartnership** or **mkippartnership** command once from each end of the partnership. The parameters that need to be specified are:

- ▶ 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 determines the maximum speed of the initial synchronization and resynchronization of the relationships.

Tip: To establish a fully functional Metro Mirror or Global Mirror partnership, issue the **mkfcpartnership** or **mkippartnership** command from both systems.

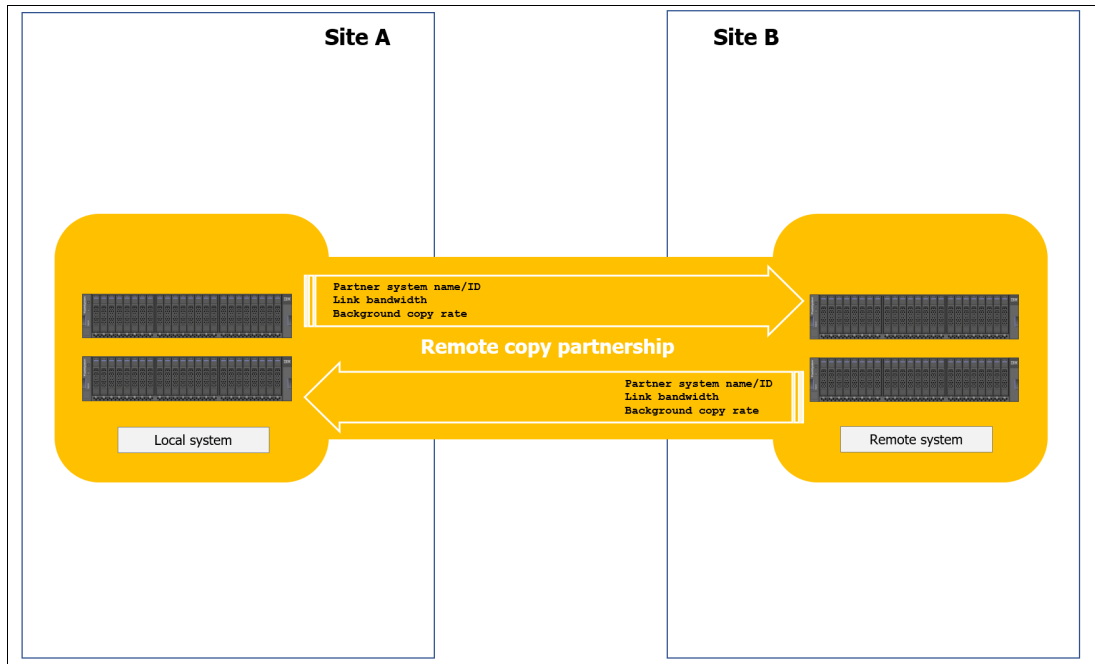


Figure 6-12 Remote Copy partnership

In addition to the background copy rate setting, the initial synchronization can be adjusted at relationship level with the `relationship_bandwidth_limit` parameter. The `relationship_bandwidth_limit` 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 became *secondary* volumes.

After the initial synchronization is complete, you can change the copy direction (see Figure 6-13 on page 262) by switching the roles of primary and secondary. The ability to change roles is used to facilitate disaster recovery.

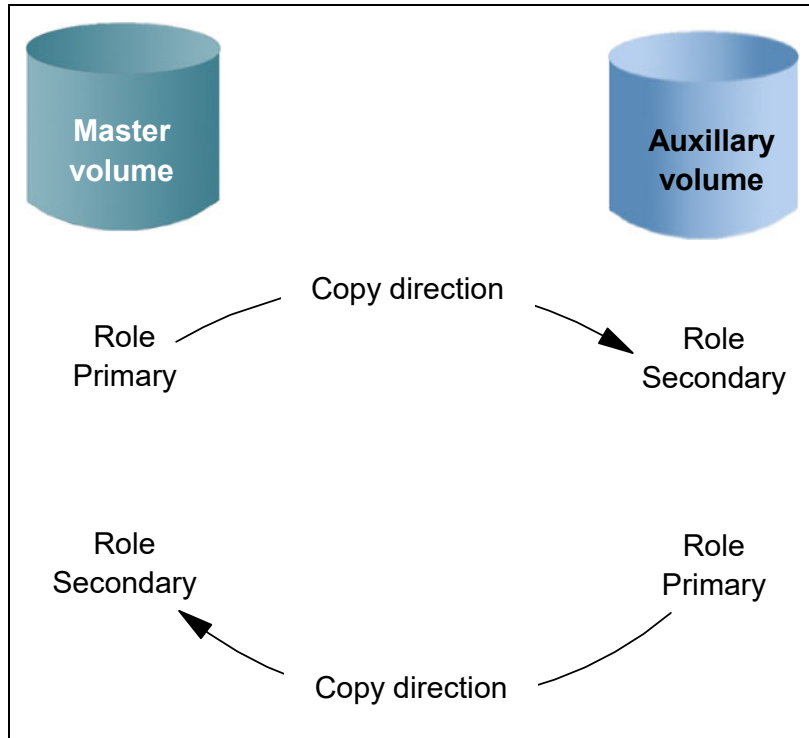


Figure 6-13 Role and direction changes

Attention: When the direction of the relationship is changed, the roles primary/secondary of the volumes are altered. A consequence is that the read/write properties are also changed, meaning that the master volume takes on a secondary role and becomes read-only.

Consistency Groups

A Consistency Group (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 Consistency Group, you cannot manage the relationship in isolation from the Consistency Group. So, for example, issuing a **stoprelationship** command on the stand-alone volume would fail because the system knows that the relationship is part of a Consistency Group.

Similarly to the Remote Copy relationships, also a Consistency Group, when created, assigns the role of *master* to the source storage system and *auxiliary* to the target storage system.

Note the following points regarding Consistency Groups:

- ▶ Each volume relationship can belong to only one Consistency Group.
- ▶ Volume relationships can also be stand-alone, that is, not in any Consistency Group.
- ▶ Consistency Groups can also be created and left empty, or can contain one or many relationships.
- ▶ You can create up to 256 Consistency Groups on a system.
- ▶ All volume relationships in a Consistency Group must have matching primary and secondary systems, but they do not need to share I/O groups.

- ▶ All relationships in a Consistency Group have the same copy direction and state.
- ▶ Each Consistency Group is either for Metro Mirror or for Global Mirror relationships, but not both. This choice is determined by the first volume relationship that is added to the Consistency Group.

Consistency Group consideration: A Consistency Group relationship does not have to be in a directly matching I/O group number at each site. A Consistency Group 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

Consistency Groups can also be used as a way to spread replication workload across multiple streams within a partnership.

The Metro or Global Mirror partnership architecture allocates traffic from each Consistency Group 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 Consistency Group also goes into stream0. You might want to consider creating an empty Consistency Group 0 so that stand-alone volumes do not share a stream with active Consistency Group volumes.

It can also pay to optimize your streams by creating more Consistency Groups. 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 would want to create at least as many Consistency Groups as you have IBM FlashSystem canisters, and, ideally, perfect multiples of the node count.

Layer concept

The *layer* is an attribute of Spectrum Virtualize-based systems which allows you to create partnerships among different Spectrum Virtualize products. The key points concerning layers are listed here:

- ▶ IBM SAN Volume Controller is always in the *Replication* layer.
- ▶ By default, IBM FlashSystem products are in the *Storage* layer.
- ▶ A system can only form partnerships with systems in the same layer.
- ▶ An IBM SAN Volume Controller can virtualize an IBM FlashSystem system only if the FlashSystem is in Storage layer.
- ▶ An IBM FlashSystem system in the Replication layer can virtualize an IBM FlashSystem system in the Storage layer.

Figure 6-14 illustrates the concept of layers.

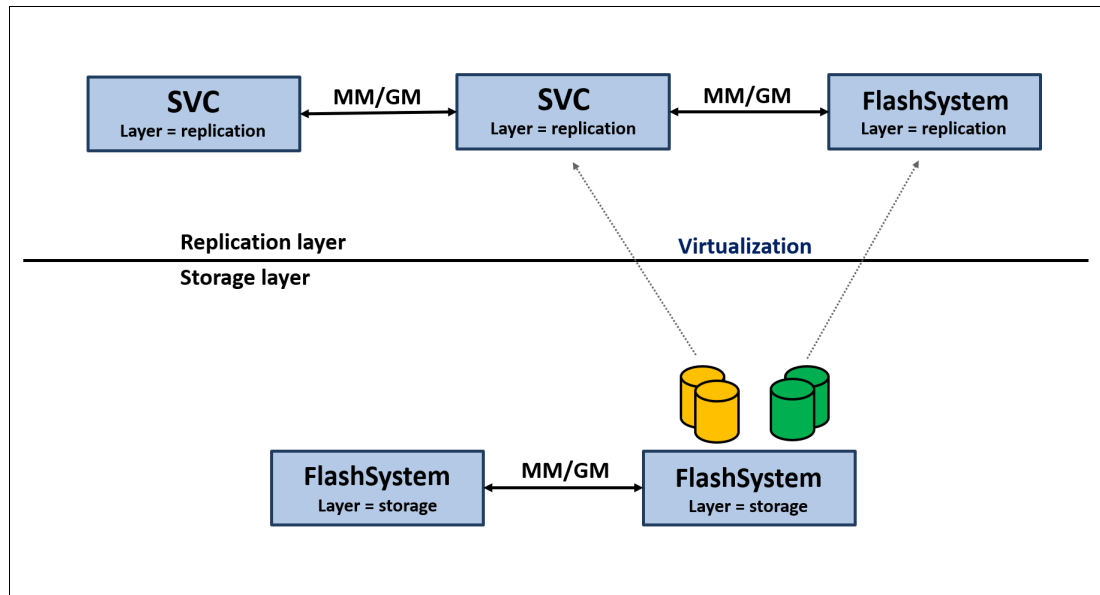


Figure 6-14 Conceptualization of layers

Generally, changing the layer is only performed at initial setup time or as part of a major reconfiguration. To change the layer of an IBM FlashSystem system, the system must meet the following pre-conditions:

- ▶ The IBM FlashSystem system must not have IBM Spectrum Virtualize, Storwize, or FlashSystem host objects defined, and must not be virtualizing any other IBM FlashSystem/Storwize controllers.
- ▶ The IBM FlashSystem system must not be visible to any other IBM Spectrum Virtualize, Storwize, or FlashSystem system in the SAN fabric, which might require SAN zoning changes.
- ▶ The IBM FlashSystem system must not have any system partnerships defined. If it is already using Metro Mirror or Global Mirror, the existing partnerships and relationships must be removed first.

Changing an IBM FlashSystem system from Storage layer to Replication layer can only be performed by using the CLI. After you are certain that all of the pre-conditions have been met, issue the following command:

```
chsystem -layer replication
```

Partnership topologies

Spectrum Virtualize allows various partnership topologies, as shown in Figure 6-15 on page 265. Each box represents an IBM Spectrum Virtualize based system.

The set of systems directly or indirectly connected form the *connected set*. A system may be partnered with up to three remote systems. No more than four systems may be in the same connected set is allowed.

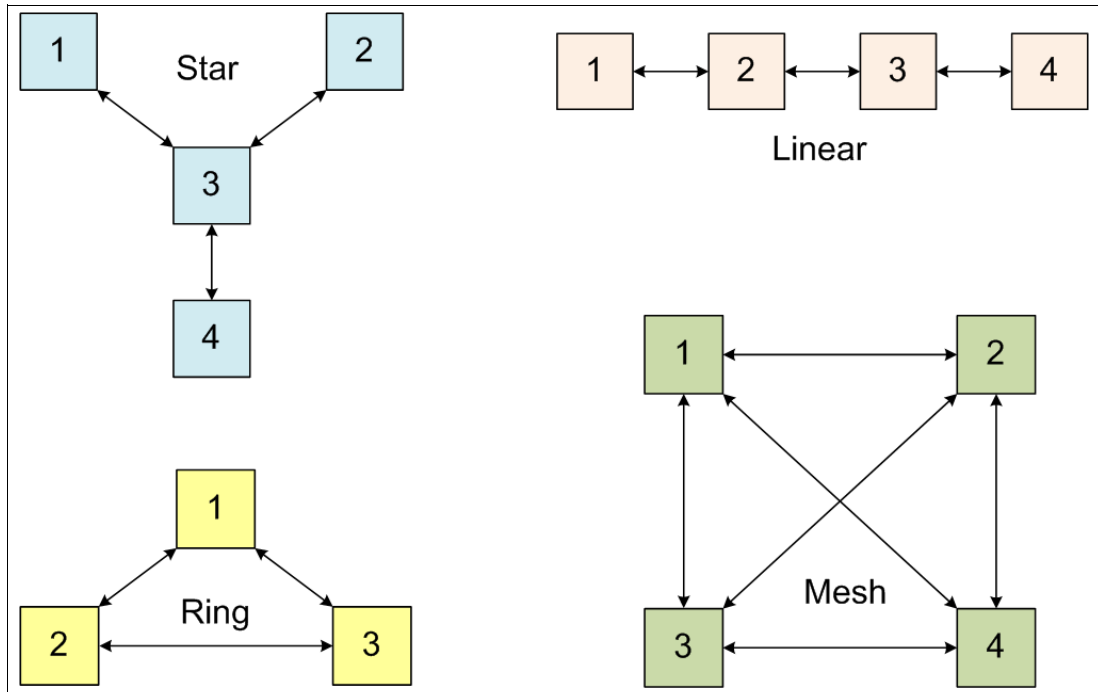


Figure 6-15 Supported topologies for Remote Copy partnerships

Star topology

A star topology can be used, for example, to share a centralized disaster recovery system (3, in this example) with up to three other systems, for example replicating 1 → 3, 2 → 3, and 4 → 3.

Ring topology

A ring topology (3 or more systems) can be used to establish a one-in, one-out implementation. For example, the implementation can be 1 → 2, 2 → 3, 3 → 1 to spread replication loads evenly among three systems.

Linear topology

A linear topology of two or more sites is also possible. However, it would generally be 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. That is, you cannot establish a multi-target Remote Copy relationship for a given volume. However, three-site replication is possible with the introduction of the *Spectrum 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 versus intersystem

Although Remote Copy services are available for intrasystem, it has no functional value for production use. Intrasystem Metro Mirror provides the same capability with less overhead. However, leaving this function in place simplifies testing and allows for experimentation and testing. For example, you can validate server failover on a single test system.

Intrasystem remote copy: Intrasystem Global Mirror is not supported on IBM Spectrum Virtualize based systems that run V6 or later.

Metro Mirror functional overview

Metro Mirror 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, Metro Mirror is designed to ensure that the secondary volume holds the same data as the primary did immediately before the failure.

Metro Mirror 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.

Metro Mirror supports relationships between volumes that are up to 300 kilometers (km) apart. Latency is an important consideration for any Metro Mirror 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 6-16 shows the order of Metro Mirror write operations.

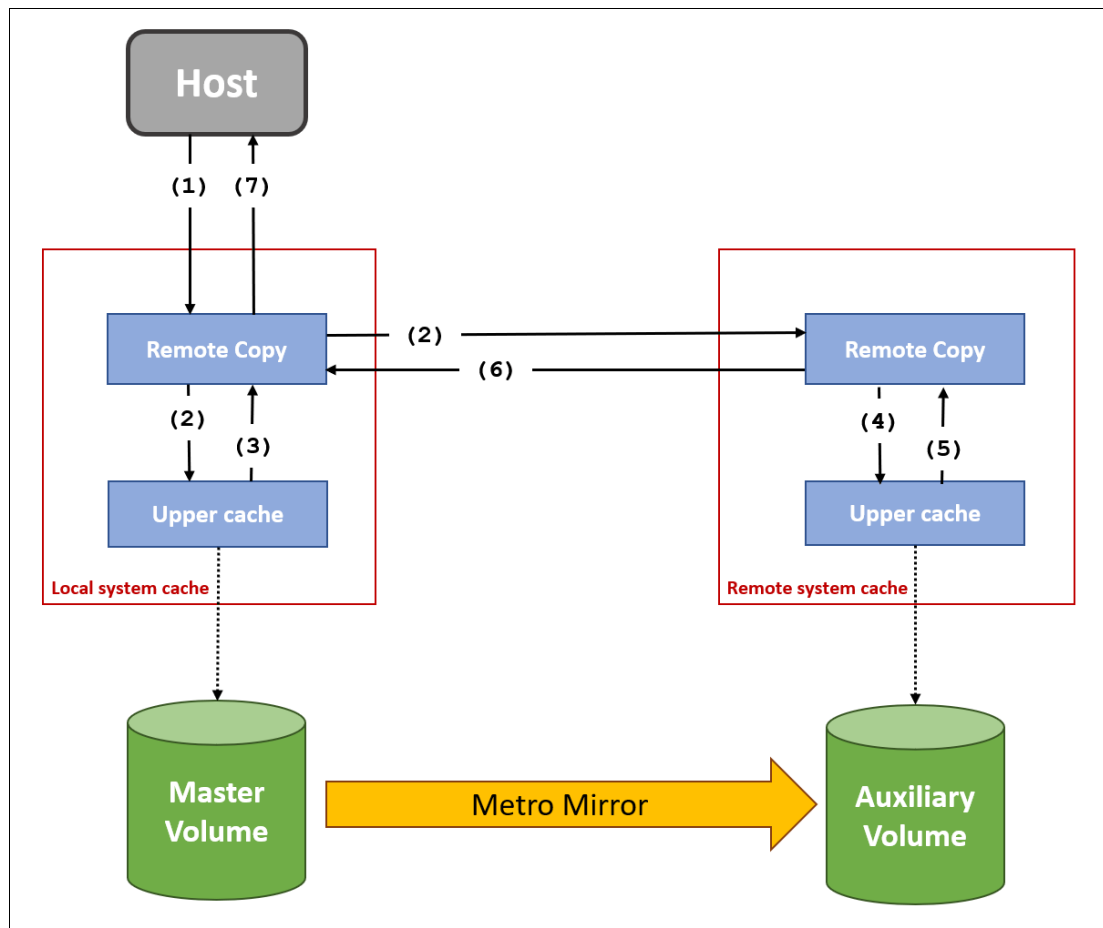


Figure 6-16 Metro Mirror write sequence

The write operation sequence is the following:

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 Remote Copy component on local system.
4. The write operation is written in the upper cache component of the remote system. This operation is initiated as soon as the data arrives from the local system and do not depend on operation ongoing in the local system.
5. The write operation on remote system upper cache is acknowledged back to Remote Copy component on remote system.
6. The remote write operation is acknowledged back to Remote Copy component on local system.
7. The write operation is acknowledged back to the host.

For a write to be considered as committed, it is required that the data is 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 a Metro Mirror write acknowledgment.

Global Mirror functional overview

Global Mirror 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 completely for any reason, Global Mirror is designed to ensure 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 between 10 ms and 10 seconds, but varies according to individual circumstances.

Global Mirror provides a way to maintain a write-order-consistent copy of data at a secondary site only slightly behind the primary. Global Mirror has minimal impact on the performance of the primary volume.

Although Global Mirror is an asynchronous Remote Copy technique, foreground writes at the local system and mirrored foreground writes at the remote system are not wholly independent of one another. IBM Spectrum Virtualize implementation of Global Mirror uses algorithms to maintain a consistent image at the target volume always.

They achieve this image 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 6-17 on page 269 shows that a write operation to the master volume is acknowledged back to the host that issues the write before the write operation is mirrored to the cache for the auxiliary volume.

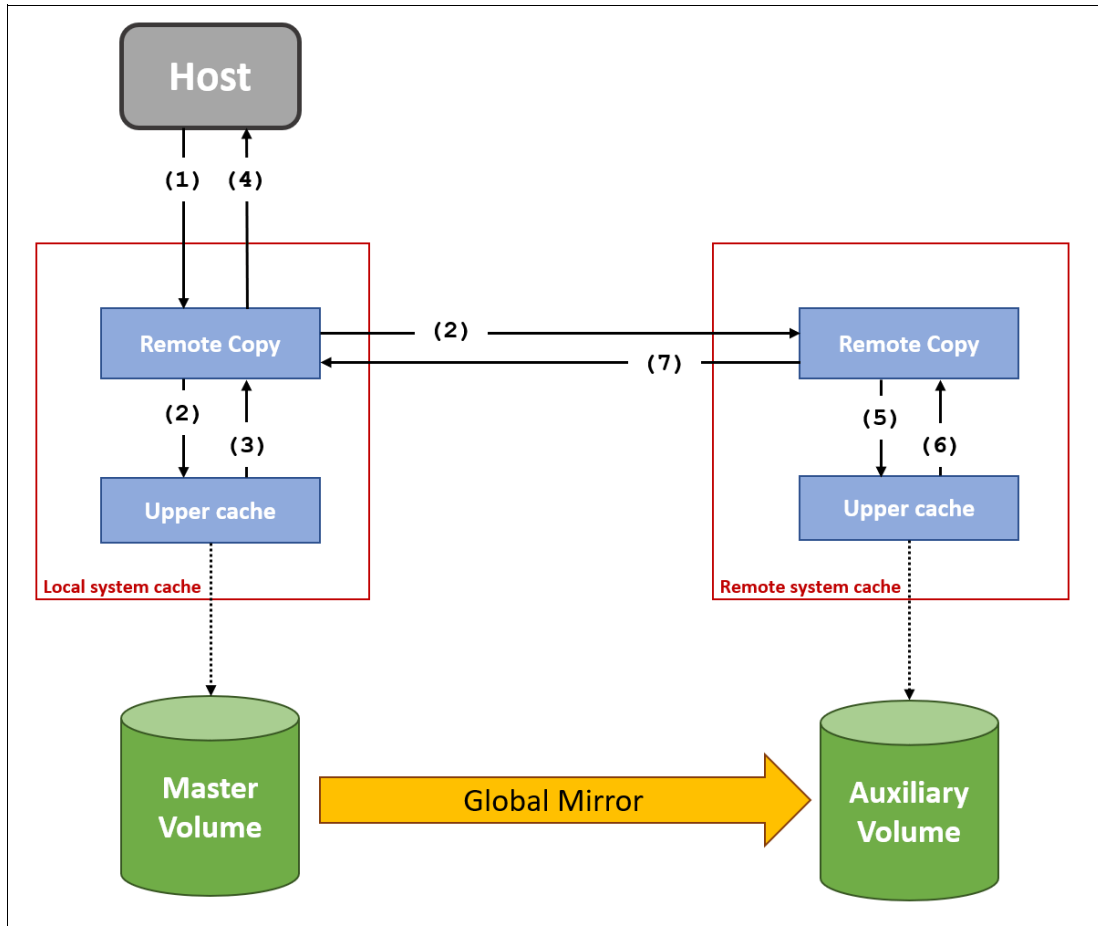


Figure 6-17 Global Mirror relationship write operation

The write operation sequence is the following:

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 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 local system upper cache is acknowledged back to Remote Copy component on local system
4. The write operation is acknowledged back to the host
5. The Remote Copy component on remote system initiates the write operation to the upper cache component according with the sequence number. This operation is initiated as soon as the data arrives from the local system and do not depend on operation ongoing in the local system
6. The write operation on remote system upper cache is acknowledged back to Remote Copy component on remote system
7. The remote write operation is acknowledged back to Remote Copy component on local system

With Global Mirror, 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. It then 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. Finally, note that any delay in step 2 is reflected in write-delay on primary volume.

Write ordering

Many applications that use block storage are required to survive failures, such as a loss of power or a software crash. They are also required to 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 can lead to problems, such as detected or undetected data corruption.

Colliding writes

Colliding writes are defined as new write I/Os that overlap existing active write I/Os.

The original Global Mirror 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 must be retried (which is known as *reconstruction*). Conceptually, the data for reconstruction comes from the master volume.

If multiple writes were allowed to 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 was inconsistent.

Applications that deliver such write activity do not achieve the performance that Global Mirror is intended to support. A volume statistic is maintained about the frequency of these collisions. The original Global Mirror implementation has been modified to allow multiple writes to a single location to be outstanding in the Global Mirror algorithm.

A need still exists for master writes to be serialized. The intermediate states of the master data must be kept in a non-volatile 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 6-18 on page 271 shows a colliding write sequence.

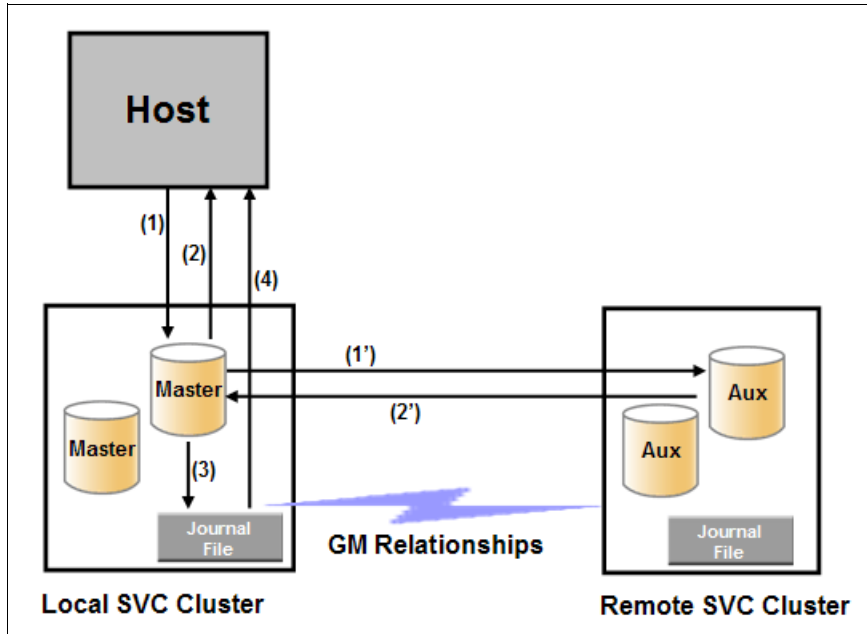


Figure 6-18 Colliding writes

The following numbers correspond to the numbers that are shown in Figure 6-18:

1. A first write is performed from the host to LBA X.
2. A host is provided acknowledgment that the write is complete, even though the mirrored write to the auxiliary volume is not yet completed.

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 is provided acknowledgment that the second write is complete.

Global Mirror Change Volumes functional overview

Global Mirror with Change Volumes (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 to reduce any impact on production hosts.

Metro Mirror and Global Mirror both require the bandwidth to be sized to meet the peak workload. Global Mirror with Change Volumes must only be sized to meet the average workload across a cycle period.

Figure 6-19 shows a high-level conceptual view of Global Mirror with Change Volumes. GMCV uses FlashCopy to maintain image consistency and to isolate host volumes from the replication process.

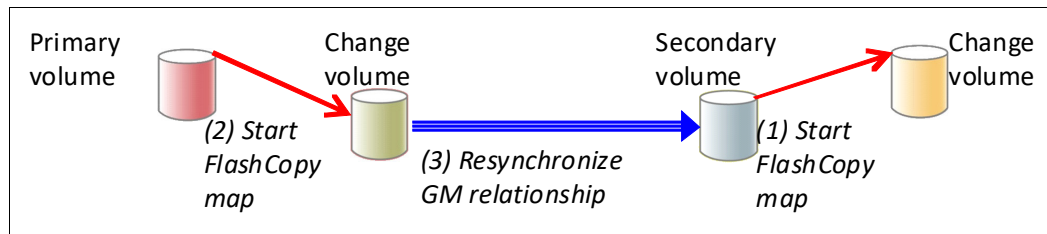


Figure 6-19 Global Mirror with Change Volumes

Global Mirror with Change Volumes also only sends 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 is designed to ensure that the secondary volume holds the same data as the primary did at a specific point in time. That period of data loss is typically between 5 minutes and 24 hours, but varies according to the design choices that you make.

Change Volumes 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 change volume to preserve its contents. Change Volumes 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.

Primary and Change Volumes are always in the same I/O group and the Change Volumes are always thin-provisioned. Change Volumes cannot be mapped to hosts and used for host I/O, and they cannot be used as a source for any other FlashCopy or Global Mirror operations.

Figure 6-20 shows how a Change Volume 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 Change Volume until the next replication cycle has completed.

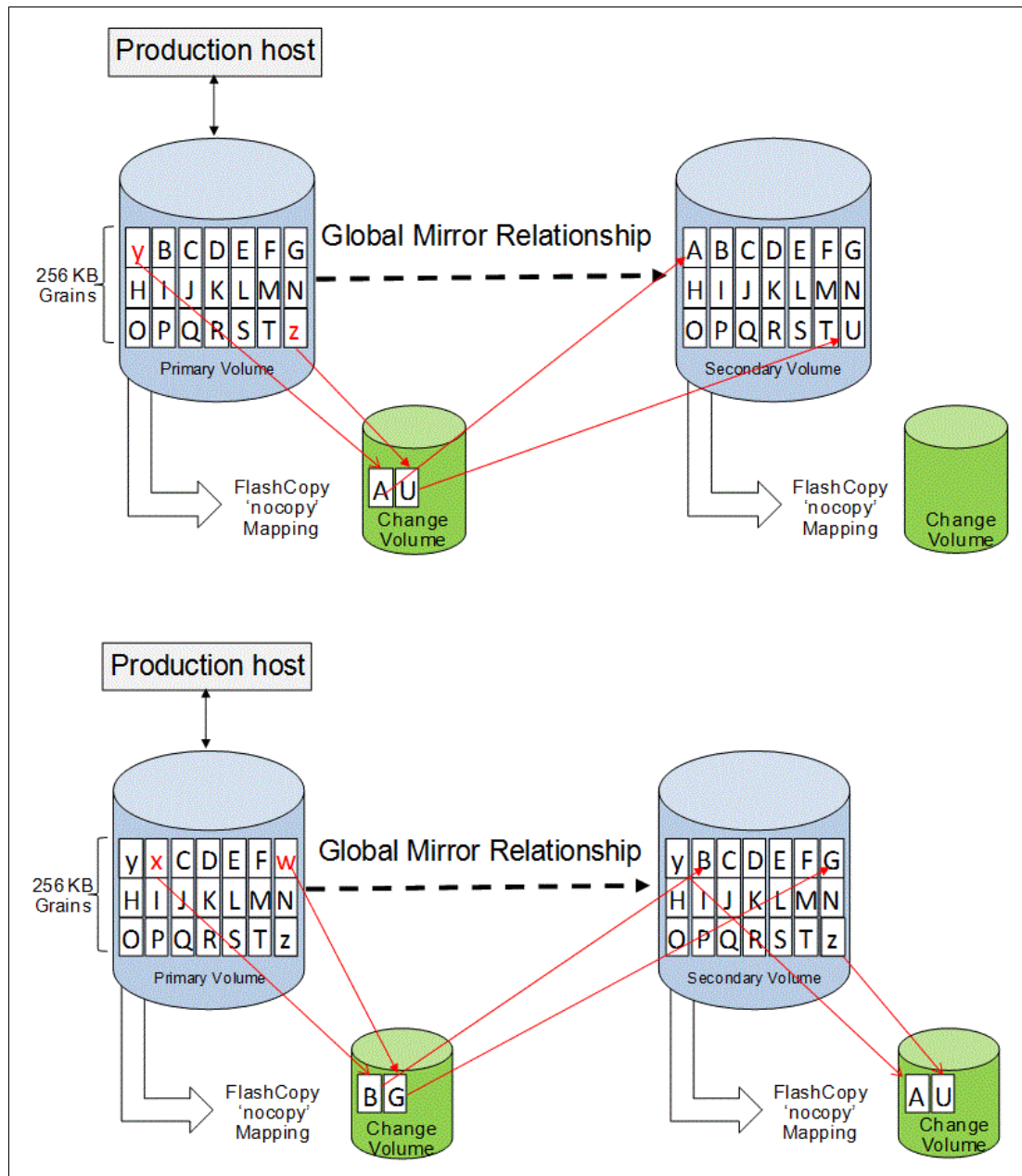


Figure 6-20 Global Mirror with Change Volumes uses FlashCopy point-in-time copy technology

FlashCopy mapping note: These FlashCopy mappings are not standard FlashCopy volumes and are not accessible for general use. They are internal structures that are dedicated to supporting Global Mirror with Change Volumes.

The options for `-cyclingmode` are `none` and `multi`.

Specifying or taking the default `none` means that Global Mirror acts in its traditional mode without Change Volumes.

Specifying `multi` means that Global Mirror starts cycling based on the cycle period, which defaults to 300 seconds. The valid range is from 60 seconds to 24*60*60 seconds (86,400 seconds = one day).

If all of the changed grains cannot be copied to the secondary site within the specified time, then the replication is designed to take as long as it needs and to start the next replication as soon as the earlier one completes. You can choose to implement this approach by deliberately setting the cycle period to a short amount of time, which is a perfectly valid approach. However, remember that 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 only be changed when the Global Mirror relationship is in a stopped state.

Recovery point objective 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 `1srcrelationship` freeze time attribute. The freeze time is the time stamp of the last primary Change Volume that has completed copying to the secondary site. Note the following example:

1. The cycle period is the default of 5 minutes and a cycle is triggered at 6:00 AM. At 6:03 AM, the cycle completes. 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 has been exceeded, the cycle immediately starts again.

6.3.3 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 intersystem network, giving an overview of the remote system connectivity options.

Terminology

The intersystem network is specified in terms of *latency* and *bandwidth*. These parameters define the capabilities of the link regarding the traffic that is on it. They must be chosen so that they support all forms of traffic, including mirrored foreground writes, background copy writes, and intersystem 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 and is measured in milliseconds. The latency measures the time 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 (Mbps) or billions of bits 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).

Intersystem 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 intersystem partnerships.

Fibre Channel connectivity is the standard connectivity that is used for the Remote Copy intersystem networks. It uses the Fibre Channel protocol and SAN infrastructures to interconnect the systems.

Native IP connectivity is a connectivity option based on standard TCP/IP infrastructures provided by IBM Spectrum Virtualize technology.

Standard SCSI operations and latency

A single SCSI read operation over a Fiber Channel network is shown in Figure 6-21.

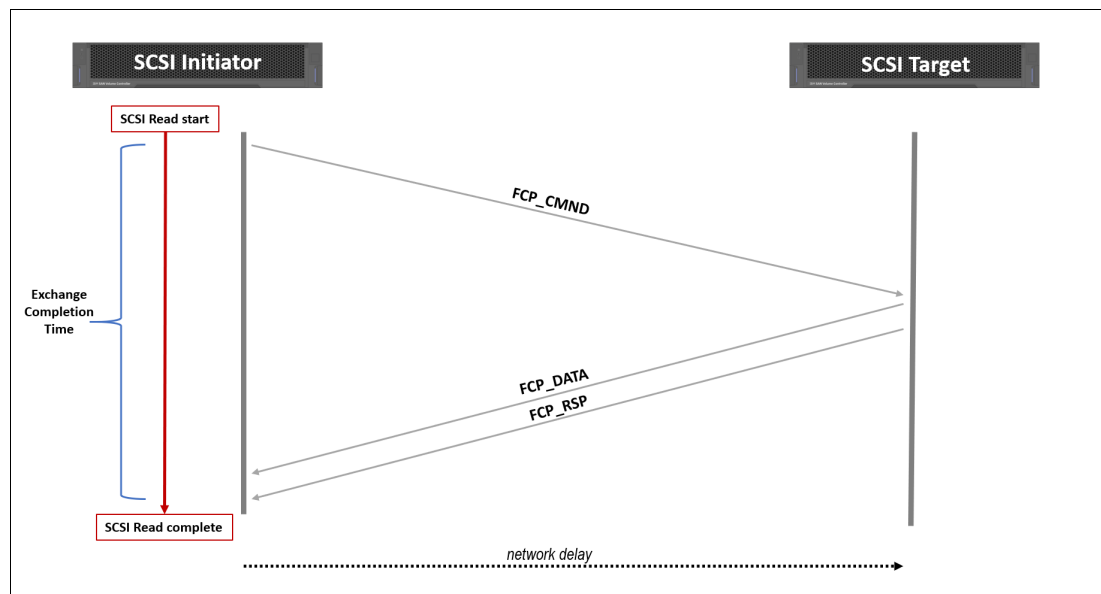


Figure 6-21 Standard SCSI read operation

The initiator starts by sending a read command (FCP_CMND) across the network to the target. The target is responsible to retrieve the data and to respond sending the data (FCP_DATA_OUT) to the initiator. Finally, the target completes the operation sending the command completed response (FCP_RSP). Note that FCP_DATA_OUT and FCP_RSP are sent to the initiator in sequence. Overall, one round trip is required to complete the read; therefore, the read takes at least one RTT, plus the time for the data out.

Typical SCSI behavior for a write is shown in Figure 6-22 on page 276.

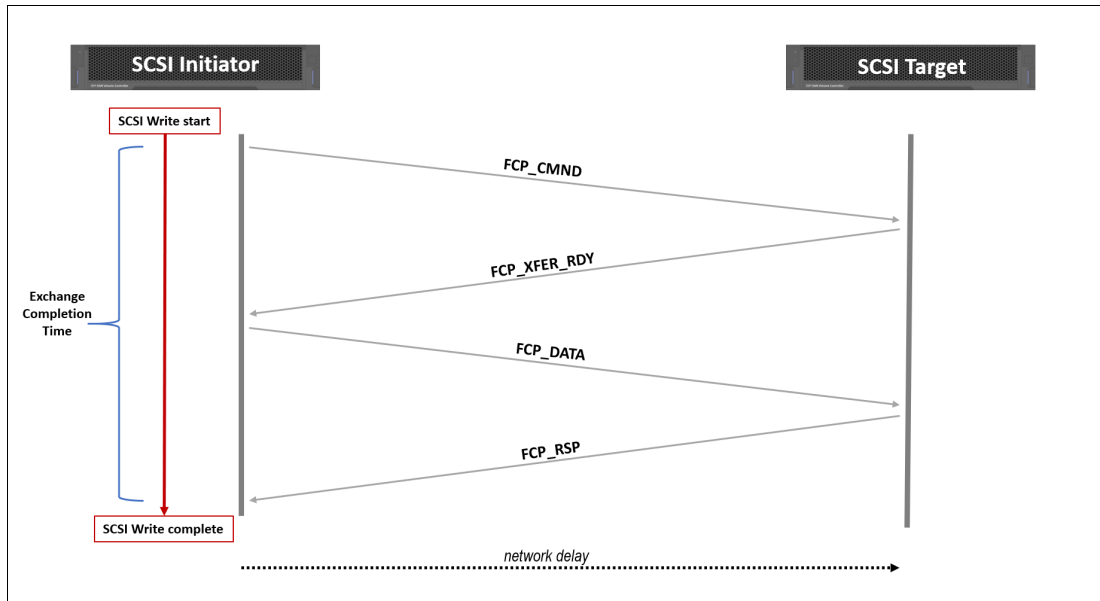


Figure 6-22 Standard SCSI write operation

A standard-based SCSI write is a two-step process. First, 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). Finally, the target completes the operation sending the command completed response (FCP_RSP). Overall, two round trips are required to complete the write; therefore, 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 μsec (10 μsec RTT), the resulting minimum service time per every km of distance for a SCSI operation is 10 μsec and 20 μsec for reads and writes respectively; for example, a SCSI write over 50 km has a minimum service time of 1000 μsec (that is, 1 ms).

Spectrum Virtualize remote write operations

With the standard SCSI operations, the writes are particularly affected by the latency. Spectrum Virtualize implements a proprietary protocol in order to mitigate the effects of the latency in the write operations over a Fibre Channel network.

Figure 6-23 on page 277 summarize how a remote copy write operation is performed over a Fibre Channel network.

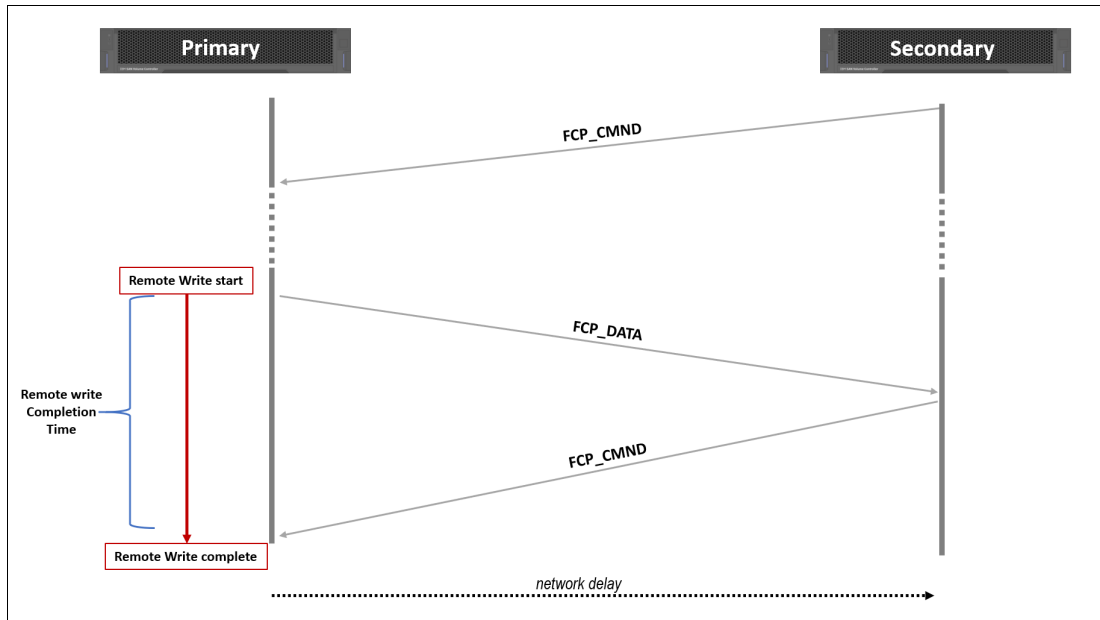


Figure 6-23 Spectrum Virtualize remote copy write

As soon as 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 initiated, the data is sent to the target as response of the dummy read command (FCP_DATA_OUT). Finally, the target completes the operation sending a new dummy read command (FCP_CMND).

Overall, one round trip is required to complete the remote write using this protocol; therefore, to replicate a write it 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 round-trip time is greater than 80 ms. If you use remote mirroring between systems with 80 - 250 ms round-trip latency, you must meet the following additional 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 -rcbuffer size 512` command on each system.

Important: Changing this setting is disruptive to Metro Mirror and Global Mirror operations. Use this command only before partnerships are created between systems, or when all partnerships with the system are stopped.

- ▶ Two Fibre Channel 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.
- ▶ 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 zoning guidelines, see “Remote system ports and zoning considerations” on page 283.

Link bandwidth that is used by internode communication

IBM Spectrum Virtualize uses part of the bandwidth for its internal intersystem 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 megabits per second) that is generated by different sizes of systems.

Table 6-7 IBM Spectrum Virtualize intersystem heartbeat traffic (megabits per second)

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 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 intersystem bandwidth should be capable of supporting 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 Global Mirror bandwidth parameter
- ▶ Intersystem 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

With the Metro Mirror, due to its synchronous nature, the amount of replication bandwidth required to mirror a given foreground write-data throughput is not less than the foreground write-data throughput itself.

The Global Mirror, not having write buffering resources, tends to mirror the foreground write as soon as it is committed in cache and therefore the bandwidth requirements are very similar to the Metro Mirror.

For a proper bandwidth sizing with Metro or Global Mirror, you must know your peak write workload to at least a five-minute interval. This information can be easily gained from tools like IBM Spectrum Control. Finally, you need to allow for the 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 a Metro or Global Mirror 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.

Metro Mirror or Global Mirror require bandwidth on the instantaneous peak of 4500 writes per second as follows:

$4500 \times 0.0625 = 282 \text{ Mbps} + 20\% \text{ resync allowance} + 5 \text{ Mbps heartbeat} = 343 \text{ Mbps}$
dedicated plus any safety margin plus growth

GMCV network sizing

The GMCV is typically less demanding in terms of bandwidth requirements for a number of reasons.

First, 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.

Another factor that can reduce the bandwidth that is required for GMCV is that it only sends one copy of a changed grain, which might have been rewritten many times within the cycle period.

The GMCV network sizing is basically a trade-off between RPO, journal capacity, and network bandwidth. A direct relationship exists between the RPO and the physical occupancy of the change volumes: the lower the RPO, the less capacity is used by change volumes. However, higher RPO requires usually less network bandwidth.

For a proper bandwidth sizing with GMCV, you need to know your average write workload during the cycle time. This information can be easily gained from tools like IBM Spectrum

Control. Finally, you need to allow for the background resync workload, intercluster communication traffic, and a safe margin for unexpected peaks and workload growth.

Consider the following sizing exercises:

► **GMCV peak 30-minute cycle time**

If we look at this time 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 then 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:

$$(4500 + 5 \times 1500) \text{ replicating until 8 AM at the latest would probably require at least the following bandwidth:}$$

$$(9000 + 70 \times 1500) / 72 = 1584 \times 0.0625 = 99 \text{ Mbps} + 100\% + 5 \text{ Mbps heartbeat} = 203 \text{ Mbps}$$

at night plus any safety margin plus growth, non-dedicated, time-shared with daytime traffic

The central principle of sizing is that you need to know your write workload:

- For Metro Mirror and Global Mirror, you need to know the peak write workload.
- For GMCV, you need to know the average write workload.

GMCV bandwidth: In the above samples, the bandwidth estimation for the GMCV is based on the assumption that the write operations occurs in such a way that a change volume 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 the edge case, can be even higher than the one required for a standard Global Mirror.

Global Mirror and GMCV coexistence considerations

Global Mirror 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 Global Mirror technologies, as previously described, use the available bandwidth in different ways:

- Regular Global Mirror uses the amount of bandwidth needed to sustain the write workload of the replication set.
- The 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 Global Mirror. To avoid bandwidth contention, which could lead to a 1920 error (see 6.3.6, “1920 error” on page 302) or delayed GMCV cycle creation, the bandwidth must be sized to take into account both requirements.

Ideally, in these cases the bandwidth should be enough to accommodate the peak write workload for the Global Mirror replication set plus the estimated bandwidth needed to fulfill the RPO of GMCV. If these requirements cannot be met due to bandwidth restrictions, the least impacting option is to increase the GMCV cycle period and then reduce the background copy rate to minimize the chance of a 1920 error.

Note that these considerations also apply to configurations where multiple IBM Spectrum Virtualize based systems are sharing the same bandwidth resources.

Fibre Channel connectivity

When you use Fibre Channel (FC) technology for the intersystem network, consider the following items:

- ▶ Redundancy
- ▶ Basic topology and problems
- ▶ Distance extensions options
- ▶ Hops
- ▶ Buffer credits
- ▶ Remote system ports and zoning considerations

Redundancy

The intersystem network must adopt the same policy toward redundancy as for the local and remote systems to which it is connecting. The ISLs must have redundancy, and the individual ISLs must provide the necessary bandwidth in isolation.

Basic topology and problems

Because of the nature of Fibre Channel, you must avoid ISL congestion whether within individual SANs or across the intersystem network. Although FC (and IBM FlashSystem system) 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 are inherent to FC. They are not unique to the IBM Spectrum Virtualize products.

When an FC network becomes congested, the FC switches stop accepting more frames until the congestion clears. They can also drop frames. Congestion can quickly move upstream in the fabric and clog the end devices from communicating anywhere.

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 IBM FlashSystem canisters that cannot mirror their 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 Fibre Channel, 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 will start to see errors in your frames.

- ▶ Long-distance Small Form-factor Pluggable (SFP) transceivers and 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.

- ▶ Fibre Channel-to-IP conversion boxes. Fibre Channel over IP (FCIP) is, by far, the most common and least expensive form of distance extension. It is also 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, 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* as opposed to *megabits per second*. In the storage world, bandwidth is often specified in megabytes per second (MBps), and network engineers specify bandwidth in megabits per second (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 about supported SAN routers and FC extenders, see [IBM SAN Volume Controller \(2145 and 2147\) 8.4.0 Documentation - 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 is 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. It is considered that every extra hop adds about 1.2 microseconds of latency to the transmission.

Currently, IBM FlashSystem 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*. Fibre Channel 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**

During login, N_Ports and F_Ports at both ends of a link establish its Buffer to Buffer Credit (BB_Credit).

► **End_to_End Credit**

In the same way during login, all N_Ports establish End-to-End Credit (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 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. Therefore, 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 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 have changed over time.

The current preferred configuration is based on the following: [Nodes in Metro or Global Mirror Inter-cluster Partnerships May Reboot if the Inter-cluster Link Becomes Overloaded](#).

The preferred practice for the IBM FlashSystem systems is to provision dedicated node ports for local node-to-node traffic (by using port masking) and isolate Global Mirror 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 non-clustered 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 not participating to the remote partnership will use the node/nodes 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. Note that this configuration, even though supported, allows the replication traffic to go through the node-to-node connectivity and this 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 intersystem Fibre Channel 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 Global Mirror 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 Fibre Channel ports on each node in the local system to two Fibre Channel 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 6-24.

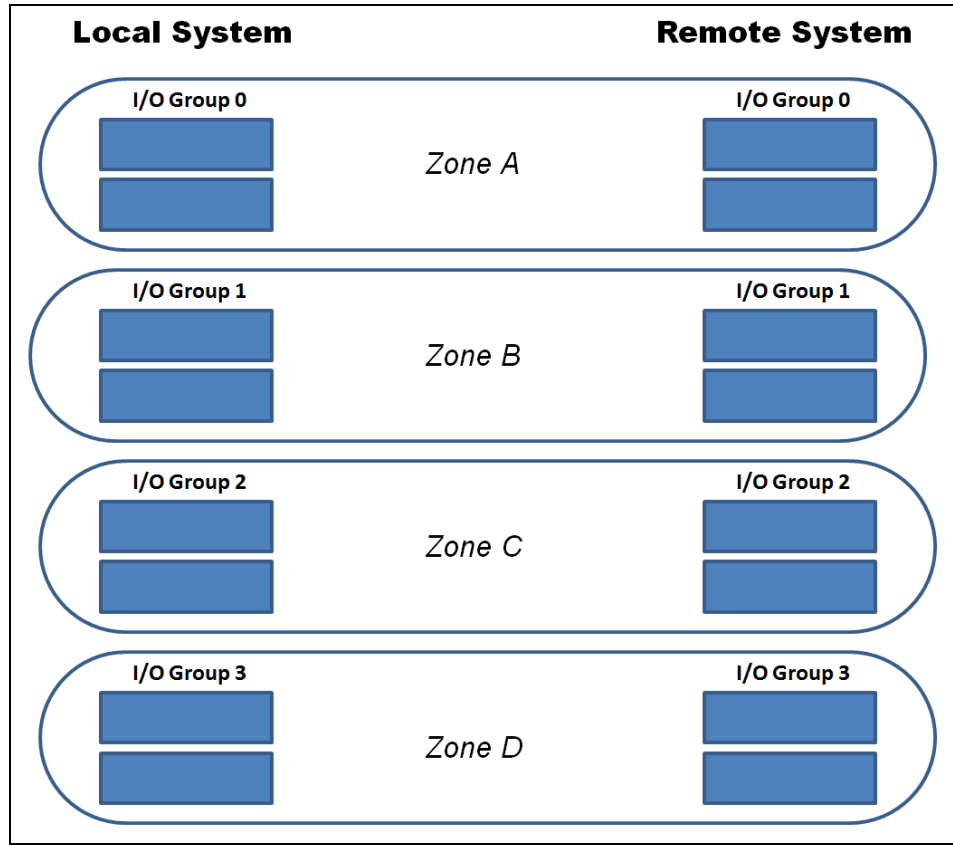


Figure 6-24 Zoning scheme for >80 ms Remote Copy partnerships

NPIV: IBM FlashSystem systems with the NPIV feature enabled provide virtual WWPN for the host zoning. Those WWPNs are intended for host zoning only and cannot be used for the Remote Copy partnership.

SAN Extension design considerations

Disaster Recovery solutions based on Remote Copy technologies require reliable SAN extensions over geographical links. In order to avoid single points of failure, multiple physical links are usually implemented. When implementing these solutions, particular attention must be paid in the Remote Copy network connectivity set up.

Consider a typical implementation of a Remote Copy connectivity using ISLs, as shown in Figure 6-25 on page 286.

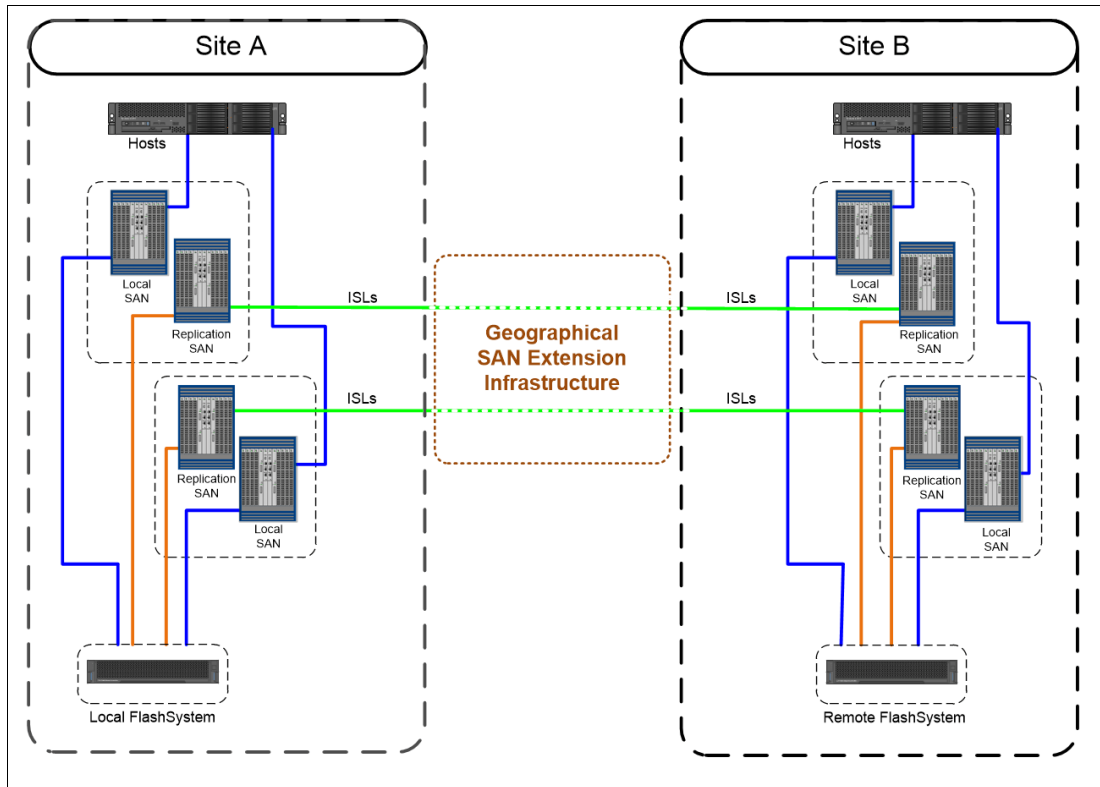


Figure 6-25 Typical Remote Copy network configuration

In the configuration in Figure 6-25, the Remote Copy network is isolated in a Replication SAN that interconnects Site A and Site B through a SAN extension infrastructure using of 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, shown in Figure 6-26 on page 287, 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.

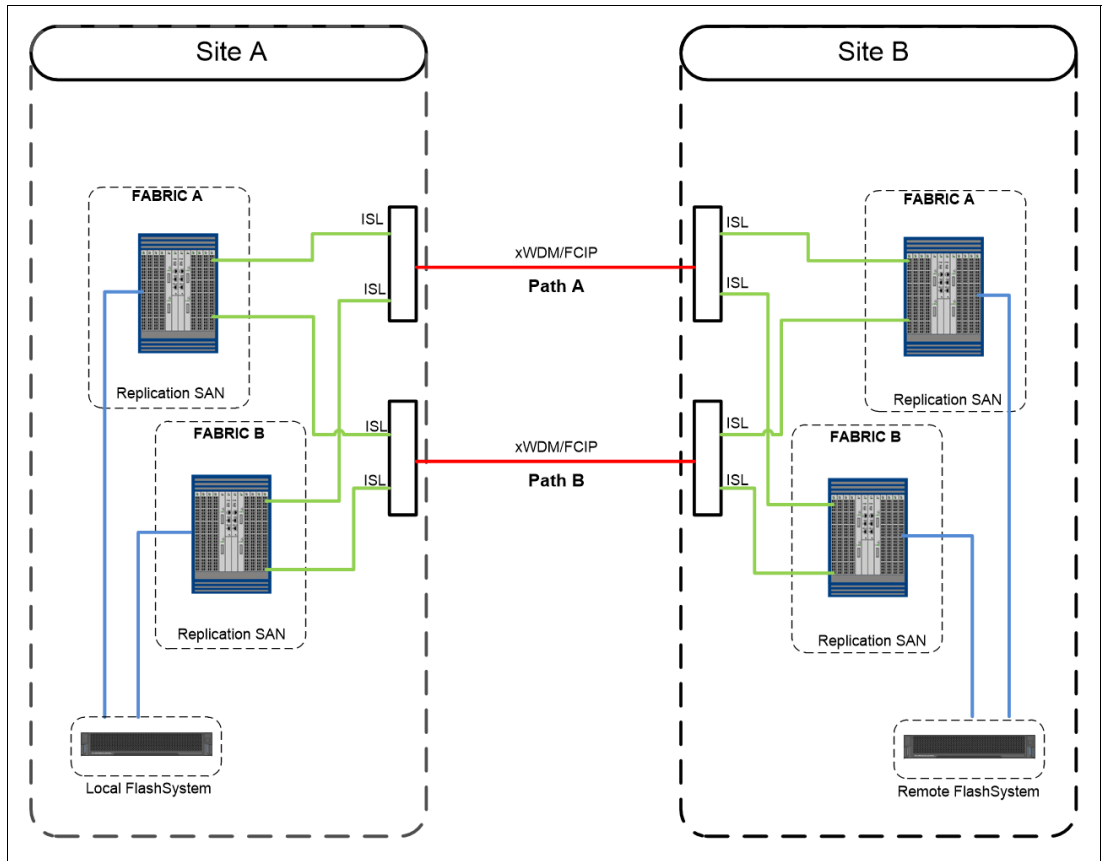


Figure 6-26 Configuration 1: physical paths shared among the fabrics

In Configuration 2, shown in Figure 6-27, 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.

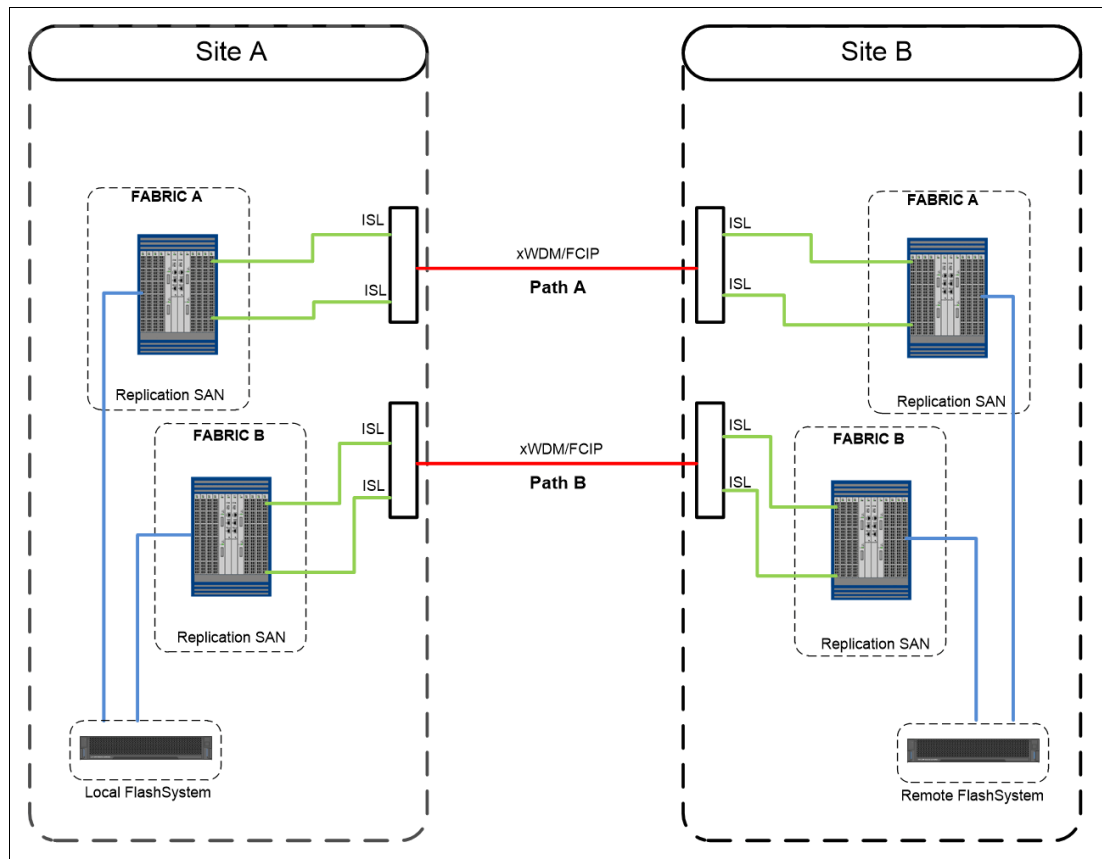


Figure 6-27 Configuration 2: physical paths not shared among the fabrics

With Configuration 1, in case of failure of one of the physical paths, both fabrics are simultaneously affected and a fabric reconfiguration occurs because of an ISL loss. This situation could 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 would be guaranteed through the unaffected fabric.

Summarizing, the recommendation is to fully understand the implication of a physical path or xWDM/FCIP router loss in the SAN extension infrastructure and to implement the appropriate architecture to avoid a simultaneous impact.

6.3.4 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 V8.4. Check the online documentation as these limits can change over time.

Table 6-8 Remote Copy maximum limits

Remote copy property	Maximum	Comment
Remote Copy (Metro Mirror and Global Mirror) relationships per system	10000	This configuration can be any mix of Metro Mirror and Global Mirror relationships.
Active-Active Relationships	2000	This is the limit for the number of HyperSwap volumes in a system.
Remote Copy relationships per consistency group	None	No limit is imposed beyond the Remote Copy relationships per system limit. Apply to Global Mirror and Metro Mirror.
GMCV relationships per consistency group	200	
Remote Copy consistency groups per system	256	
Total Metro Mirror and Global Mirror volume capacity per I/O group	1024 TB	This limit is the total capacity for all master and auxiliary volumes in the I/O group.
Total number of Global Mirror with Change Volumes relationships per system	256	60s cycle time.
	2500	300s cycle time.

Similar to FlashCopy, the Remote Copy services require memory to allocate the bitmap structures used to track the updates while volume 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.

Remote Copy general restrictions

To use Metro Mirror and Global Mirror, you must adhere to the following rules:

- ▶ You must have the same size for source and target volume 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 Metro Mirror and Global Mirror only. GMCV is not supported.
 - The Remote Copy Consistency Protection feature is not allowed and must be removed before resizing the volumes.
 - No active FlashCopy allowed.
 - The Remote Copy relationship must be in synchronized status.
 - The resize order must guarantee the target volume to be always larger than the source volume.

Note: The volume expansion for Metro and Global Mirror volumes was introduced with Spectrum Virtualize version 7.8.1 with some restrictions:

- ▶ In the first implementation (up to version 8.2.1), only thin provisioned or compressed volumes were supported.
- ▶ With version 8.2.1 also non-mirrored fully allocated volumes were supported.
- ▶ With version 8.4 all the restrictions on volume type have been removed.

- ▶ You can mirror intrasystem Metro Mirror or Global Mirror only between volumes in the same I/O group.

Intrasystem remote copy: The intrasystem Global Mirror is not supported on IBM Spectrum Virtualize based systems running version 6 or later.

- ▶ Global Mirror is not recommended for cache-disabled volumes that are participating in a Global Mirror 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 change volumes in case of change from Metro or Global Mirror to GMCVs.

Interaction between Remote Copy and FlashCopy

Remote Copy functions can be used in conjunction with the FlashCopy function so that you can have both operating concurrently on the same volume. The possible combinations between Remote Copy and FlashCopy follow:

- ▶ 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 a Metro or Global Mirror 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 with the following restriction: A FlashCopy mapping must be in the `idle_copied` state when its target volume is the target volume of an active Metro Mirror or Global Mirror relationship.

When implementing Flashcopy functions for volumes in GMCV relationships, remember that Flashcopy multi-target mappings will be created. As described in “Interaction and dependency between Multiple Target FlashCopy mappings” on page 245, this 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 255.

With such configurations, it is recommended to set the Cleaning Rate accordingly. This recommendation applies also to Consistency Protection volumes and HyperSwap configurations.

Native back-end controller copy functions considerations

As previously discussed, the IBM FlashSystem technology provides a widespread set of copy services functions that cover most of the clients requirements.

However, some storage controllers can provide specific copy services capabilities not available with the current version of IBM Spectrum Virtualize software. The IBM FlashSystem technology addresses these situations by using cache-disabled image mode volumes that virtualize LUN participating to the native back-end controller's copy services relationships.

Keeping the cache disabled guarantees 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 FlashSystem 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 FlashSystem does not recognize why this happens. Therefore, it might log errors when these events occur. For this reason, if the IBM FlashSystem 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 the IBM FlashSystem. Note that accidental incorrect configurations of the back-end controller copy services involving IBM FlashSystem attached LUN can produce unpredictable results.

For example, if you accidentally use a LUN with IBM FlashSystem data on it as a point-in-time target LUN, you can corrupt that data. Moreover, if that LUN was a managed disk in a managed-disk group with striped or sequential volumes on it, the managed disk group might be brought offline. This situation, in turn, makes all of 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 intersystem relationships, upgrade only one cluster at a time. That is, 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 one system before it is started 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 Spectrum Virtualize versions. For more information, see [Spectrum 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 consequently the background copy rate, is evenly divided among the cluster's I/O groups. The available bandwidth for the background copy can be used by either canister, or shared by both canisters within the I/O Group.

This bandwidth allocation is independent from the number of volumes for which a canister is responsible. Each node, in turn, 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 resource 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. That is, less data can be outstanding from any one local node to any other remote node. Therefore, to optimize performance, Global Mirror 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 FlashSystem canister 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 also responsible for sending that write to the preferred node of the target volume. The primary preferred node is also 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. That is, 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 6-28 on page 293 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.

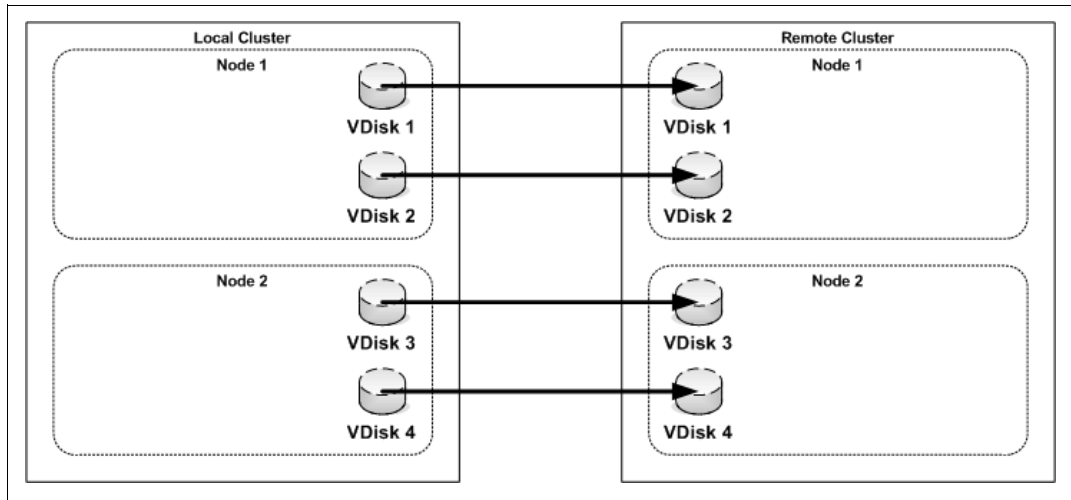


Figure 6-28 Remote Copy resources that are not optimized

With the configuration shown in Figure 6-28, 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 shown in Figure 6-28 changes to the configuration shown in Figure 6-29, all Remote Copy resources for each node are used, and Remote Copy operates with better performance.

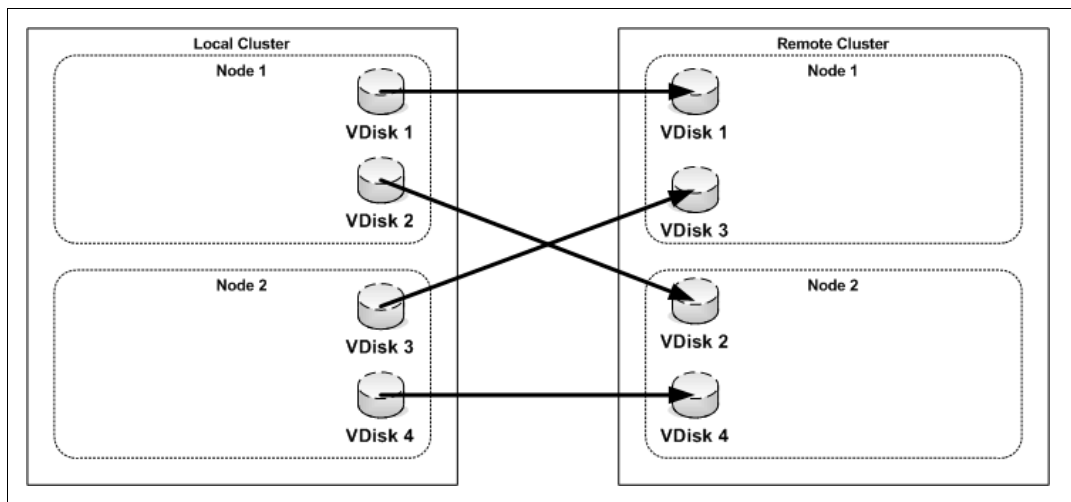


Figure 6-29 Optimized Global Mirror resources

GMCV Change Volumes placement considerations

The Change Volumes in a GMCV configuration are basically thin-provisioned volumes used as FlashCopy targets. For this reason the same considerations described in “Volume placement considerations” on page 253 apply. The Change Volumes can be compressed to reduce the amount of space used, however it is important to note that the Change Volumes might be subject to heavy write workload both in the primary and secondary system.

Therefore, the placement on the back-end is critical to provide adequate performances. Consider to use DRP for the change volumes only if very beneficial in terms of space savings.

Trick: The internal FlashCopy used by the GMCV is with 256KB grain size. However, it is possible to force a 64KB grain size by creating a FlashCopy with 64KB grain size from the GMCV volume and a dummy target volume before assigning the change volume to the relationship. This can be done to both 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 intersystem network capability, the background copy resynchronization writes use too much of the intersystem network. It starves 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 and 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. It does so by pushing more mirrored foreground write traffic along the intersystem 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
- ▶ Intersystem 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 is not able to sustain an acceptable background copy rate, consider alternatives to the initial copy as described in “Initial synchronization options and Offline Synchronization” on page 295.

Changes in the environment, or loading of it, can affect the foreground I/O. IBM FlashSystem technology provides a means to monitor, and a parameter to control, how foreground I/O is affected by running Remote Copy processes. IBM Spectrum Virtualize software 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 6.3.6, “1920 error” on page 302.

Finally, note that with Global Mirror Change Volume, 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 280. For this reason, the background copy rate, and the `relationship_bandwidth_limit` setting, affects the available bandwidth not only during the initial synchronization, but also during the normal cycling process.

Background copy bandwidth allocation: As already mentioned in “Volume placement considerations” on page 292, 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 accordingly to reach the desired background copy rate.

Consider, for example, a 4-I/O groups 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 then 250MB/s. Having the available bandwidth evenly divided among the I/O groups, every I/O group in this cluster can theoretically synchronize data at a maximum rate of about 62 MBps (50% of 1,000 Mbps). Now in an edge case where only volumes from one I/O group are being replicated, in order to reach the full background copy rate (250 MBps) the partnership bandwidth should be adjusted to 16000 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 is started, 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 FlashSystem 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 is started. 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:
 - Create both volumes with the security delete feature 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. The administrator must then complete the following actions:

- Create the relationship with the `already_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 a Metro Mirror and Global Mirror 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 and 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 Metro Mirror relationship or the Global Mirror relationship target before it uses Remote Copy services to maintain the Metro Mirror or Global Mirror. This example does not require downtime for the hosts that use the source volumes.

Before you set up Global Mirror relationships and save bandwidth, complete the following steps:

1. Ensure that the hosts are up and running and are using their volumes normally. The Metro Mirror relationship nor Global Mirror relationship is not yet defined.

Identify all volumes that become the source volumes in a Metro Mirror relationship or in a Global Mirror relationship.

2. Establish the Remote Copy partnership with the target IBM Spectrum Virtualize-based system.

To set up Global Mirror relationships and save bandwidth, complete the following steps:

1. Define a Metro Mirror relationship or a Global Mirror 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 of these steps are redundant because the IBM Spectrum Virtualize system will perform 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 being modified while you are copying the image, the IBM FlashSystem is tracking those changes. The image that you create might have some of the changes and is likely to also miss some of the changes.

When the relationship is restarted, the IBM FlashSystem applies all of 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 Metro Mirror or Global Mirror relationship. You can mount the Metro Mirror and Global Mirror 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 you can get the media to the remote site and load it, the quicker IBM FlashSystem system starts running and maintaining the Metro Mirror and Global Mirror.

5. Unmount the target volumes from your host. When you start the Metro Mirror and Global Mirror relationship later, the IBM FlashSystem stops write-access to the volume while the mirror relationship is running.
6. Start your Metro Mirror and Global Mirror relationships. The relationships must be started with the `-clean` parameter. This way, changes that are made on the secondary volume are ignored. Only changes 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 `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 attitude can be risky especially when using the Global Mirror technology where the application performances at the primary system can indeed be limited by the performance of the remote system.

The preferred 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 Global Mirror or Metro Mirror 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` is an optional parameter that specifies the new background copy bandwidth in the range 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` also applies to Metro Mirror relationships.

gmlinktolerance and gmmaxhostdelay

The **gmlinktolerance** and **gmmaxhostdelay** parameters are critical in the system for deciding internally whether to terminate a relationship due to a performance problem. In most cases, these two parameters need to be considered in tandem. The defaults would not normally be changed unless you had 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 Global Mirror 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 Global Mirror. That is, how long would that local I/O take with Global Mirror turned off, and how long does it take with Global Mirror turned on. The difference is the host delay due to Global Mirror tag and forward processing.

Although the default settings are adequate for most situations, increasing one parameter while reducing another might deliver a tuned performance environment for a particular circumstance.

Example 6-1 shows how to change **gmlinktolerance** and the **gmmaxhostdelay** parameters 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 such as these before putting them into a live production environment.

A detailed description and settings considerations about the **gmlinktolerance** and the **gmmaxhostdelay** parameters are described in 6.3.6, “1920 error” on page 302.

rcbuffersize

rcbuffersize was introduced with the version 6.2 code level so that systems with intense and bursty write I/O would not fill the internal buffer while Global Mirror writes were 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 is 512 MB.

Example 6-2 Changing rcbuffersize to 64 MB

```
chsystem -rcbuffersize 64
```

Remember that any additional buffers 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 Global Mirror logic. If a write is hung for the specified amount of time, for example due to a rebuilding array on the secondary system,

Global Mirror stops the relationship (and any containing consistency group), which triggers a 1920 error.

The **partnershipexclusionthreshold** parameter was introduced to allow users to set the timeout for an I/O that triggers a temporarily dropping of the link to the remote cluster. The value must be a number from 30 - 315.

Important: Do not change the **partnershipexclusionthreshold** parameter, except under the direction of IBM Support.

A detailed description and settings considerations about the **maxreplicationdelay** parameter are described in 6.3.6, “1920 error” on page 302.

Link delay simulation parameters

Even though Global Mirror is an asynchronous replication method, there can be an impact to server applications due to Global Mirror managing transactions and maintaining write order consistency over a network. To mitigate this impact, as a testing and planning feature, Global Mirror allows you to simulate the effect of the round-trip delay between sites by using the following parameters:

- ▶ The **gminterclusterdelaysimulation** parameter
This optional parameter specifies the intersystem delay simulation, which simulates the Global Mirror round-trip delay between two systems in milliseconds. The default is 0. The valid range is 0 - 100 milliseconds.
- ▶ The **gmintraclusterdelaysimulation** parameter
This optional parameter specifies the intrasystem delay simulation, which simulates the Global Mirror round-trip delay in milliseconds. The default is 0. The valid range is 0 - 100 milliseconds.

6.3.5 Multiple site remote copy

The most common use cases for the Remote Copy functions are obviously Disaster Recovery solutions. Code level 8.3.1 introduced further Disaster Recovery capabilities such as the *Spectrum Virtualize 3-site replication* that provides a solution for coordinated disaster recovery across three sites in various topologies. A complete discussion about the Disaster Recovery solutions based on IBM Spectrum Virtualize technology is beyond the intended scope for this book. For an overview of the Disaster Recovery solutions with the IBM Spectrum 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 the data movement among distant locations as required, for instance, for data center relocation and consolidation projects. In these scenarios, the IBM Spectrum Virtualize Remote Copy technology is particularly effective when combined with the image copy feature that allows data movement among storage systems of different technology or vendor.

Mirroring scenarios that involve multiple sites can be implemented using a combination of Spectrum Virtualize capabilities as described in the following sections.

Performing cascading copy service functions

Cascading copy service functions that use IBM FlashSystem 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 FlashSystem copy services and by combining IBM FlashSystem copy services (with image mode cache-disabled volumes) and native storage controller copy services.

DRP limitation: Currently, the image mode VDisk is not supported with DRP.

Cascading with native storage controller copy services

Figure 6-30 describes the configuration for 3-site cascading by using the native storage controller copy services in combination with IBM FlashSystem Remote Copy functions.

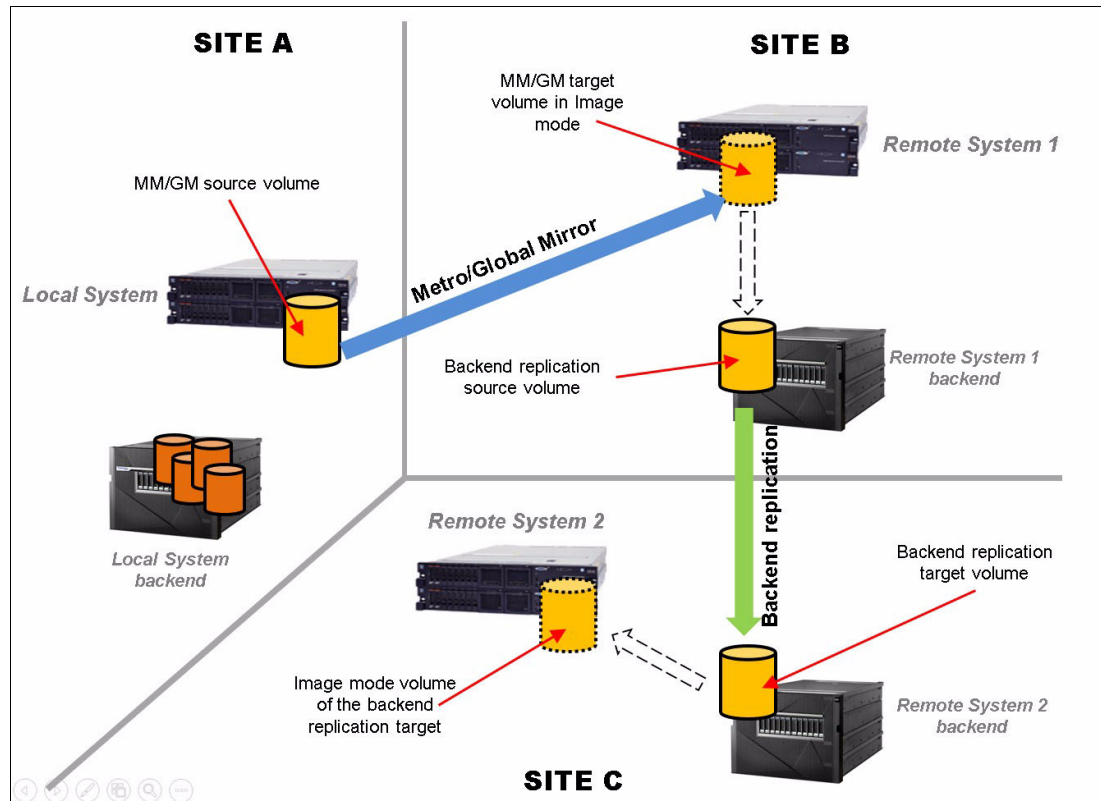


Figure 6-30 Using three-way copy services

In Figure 6-30, the primary site uses IBM FlashSystem Remote Copy functions (Global Mirror or Metro Mirror) 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 FlashSystem. For more information, see “Native back-end controller copy functions considerations” on page 291.
- ▶ The source and target volumes used by the back-end controller native copy services must be imported to the IBM FlashSystem system as image-mode volumes with the cache disabled.

Cascading with IBM FlashSystem systems copy services

Remote Copy services cascading is allowed with the Spectrum Virtualize 3-site replication capability. See *IBM Spectrum Virtualize 3-Site Replication*, SG24-8504. However, a cascading-like solution is also possible by combining the IBM FlashSystem copy services. These Remote Copy services implementations are useful in 3-site disaster recovery solutions and data center moving scenarios.

In the configuration described in Figure 6-31, a Global Mirror (Metro Mirror can also be used) solution is implemented between the Local System in Site A, the production site, and the Remote System 1 located in Site B, the primary disaster recovery site. A third system, Remote System 2, is located in Site C, the secondary disaster recovery site. Connectivity is provided between Site A and Site B, between Site B and Site C, and optionally between Site A and Site C.

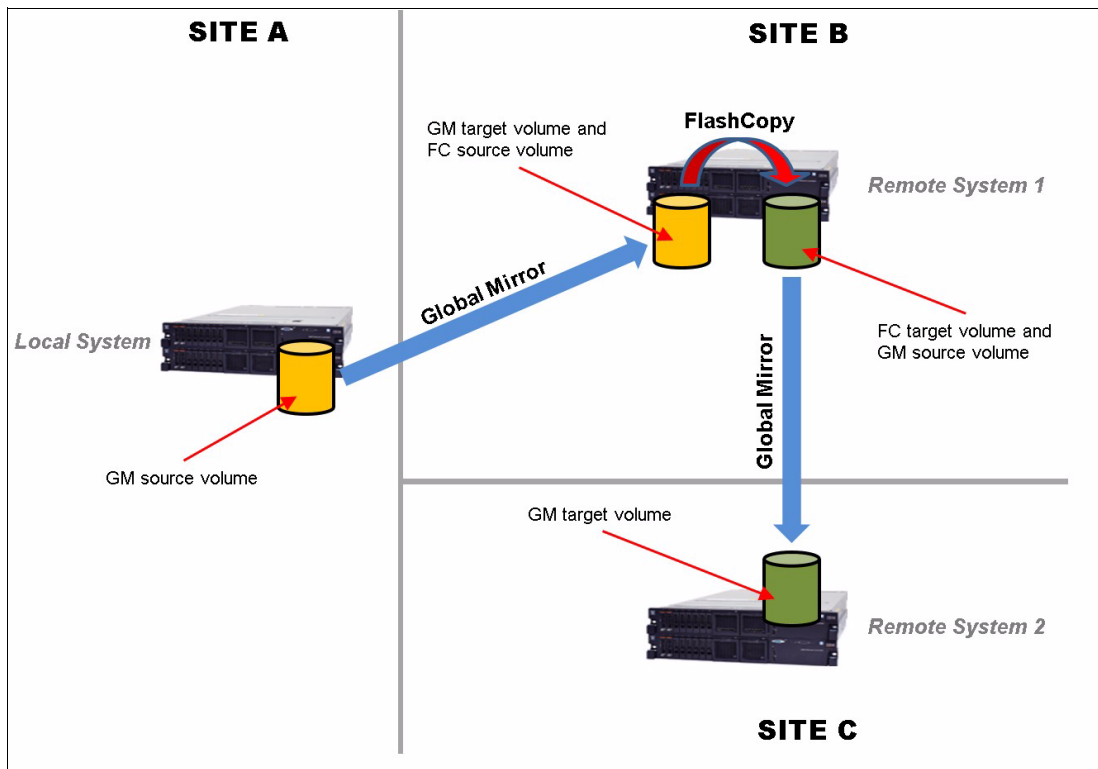


Figure 6-31 Cascading-like infrastructure

To implement a cascading-like solution, the following steps must be completed:

1. Set up phase. Perform the following actions to initially set up the environment:
 - a. Create the Global Mirror relationships between the Local System and Remote System 1.
 - b. Create the FlashCopy mappings in the Remote System 1 using the target Global Mirror volumes as FlashCopy source volumes. The FlashCopy must be incremental.
 - c. Create the Global Mirror relationships between Remote System 1 and Remote System 2 using the FlashCopy target volumes as Global Mirror source volumes.
 - d. Start the Global Mirror from Local System to Remote System 1.

After the Global Mirror is in ConsistentSynchronized state, you are ready to create the cascading.

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 Global Mirror between Remote System 1 and Remote System 2 is in stopped or idle status, if it is not, stop the Global Mirror.
 - b. Stop the Global Mirror between the Local System to Remote System 1.
 - c. Start the FlashCopy in Remote Site 1.
 - d. Resume the Global Mirror between the Local System and Remote System 1.
 - e. Start/resume the Global Mirror 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 executions of these operations perform incremental resynchronization instead. After the Global Mirror between Remote System 1 and Remote System 2 is in Consistent Synchronized state, the consistency point in Site C is created. The Global Mirror between Remote System 1 and Remote System 2 can now be stopped to be ready for the next consistency point creation.

6.3.6 1920 error

An IBM Spectrum Virtualize based system generates a 1920 error message whenever a Metro Mirror or Global Mirror relationship stops because of adverse conditions. The adverse conditions, if left unresolved, might affect performance of foreground I/O.

A 1920 error can result for many reasons. The condition might be the result of a temporary failure, such as maintenance on the intersystem connectivity, 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 could be triggered both for Metro Mirror and Global Mirror relationships. However, in Metro Mirror configurations the 1920 error is associated only with a permanent I/O error condition. For this reason, the main focus of this section is 1920 errors in a Global Mirror configuration.

Internal Global Mirror control policy and raising 1920 errors

Although Global Mirror 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. The IBM FlashSystem code implements many control mechanisms that mitigate the impacts of the Global Mirror to the foreground I/Os.

gmmxhostdelay and gmlinktolerance

The **gmlinktolerance** parameter helps 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. The system configuration, in terms of nodes and backend characteristics, must be provisioned so that when combined, they can support the maximum throughput that is delivered by the applications at the primary that is using Global Mirror.

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 Global Mirror 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.

Global Mirror attempts to detect and differentiate between backlogs that occur because of the operation of the Global Mirror 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 Global Mirror were disabled.

Global Mirror uses the **gmmaxhostdelay** and **gmlinktolerance** parameters to monitor Global Mirror 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, Global Mirror samples all of the Global Mirror writes and determines how much of a delay it 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. That is for instance, ten bad periods, followed by five good periods, followed by ten bad periods, results in a bad period count of 15.
- ▶ The **gmlinktolerance** parameter is defined in seconds. Since bad periods are assessed at intervals of ten seconds, the maximum bad period count is the **gmlinktolerance** parameter value that is divided by 10. For instance, 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.

Note that according to the above description, under a light I/O load, a single bad write can become significant. For example, if only one write I/O is performed for every ten and this write is considered slow, the bad period count is incremented.

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 the 20 s. With unlucky timing, a single bad I/O results (that is, a write I/O that took over 1 ms in remote copy), and it spans the boundary of two, ten-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.

maxreplicationdelay and partnershipexclusionthreshold

maxreplicationdelay is a system-wide attribute that configures how long a single write can be outstanding from the host before the relationship is stopped, triggering a 1920 error. It can protect the hosts from seeing timeouts due to 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 instance, 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, triggering a 1920 error. This 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.

The **partnershipexclusionthreshold** is a system-wide parameter that sets the timeout for an I/O that triggers a temporarily dropping of the link to the remote system. Similar to **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.

In an IBM FlashSystem system, a node assert (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 gets to 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 15 seconds (315 seconds). This action often removes hang conditions 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 in the range 30 - 315.

If an I/O takes longer the **partnershipexclusionthreshold** value, a 1720 error is triggered (with an event ID 987301) and any regular Global Mirror or Metro Mirror 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, the **chsystem** command must be used, as shown in Example 6-3.

Example 6-3 maxreplicationdelay and partnershipexclusionthreshold setting

```
IBM_FlashSystem:ITS0:superuser>chsystem -maxreplicationdelay 30
IBM_FlashSystem:ITS0:superuser>chsystem -partnershipexclusionthreshold 180
```

The **maxreplicationdelay** and **partnershipexclusionthreshold** parameters do not interact with the **gmlinktolerance** and **gmmaxhostdelay** 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 also 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 FlashSystem. 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 are the main performance statistics to investigate 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 submitted per second by hosts on average over the sample period, and the bandwidth of those writes. For secondary volumes in relationships, this is the average number of replicated writes that are received per second, and the bandwidth that these writes consume. Summing the rate over the volumes you intend to replicate gives a coarse estimate of the replication link bandwidth required.

► *Write Response Time and Peak Write Response Time*

On primary volumes, these are the average time (in milliseconds) and 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 the write response time leap from 1 ms to 20 ms, the two are most likely linked. However, some applications with high queue depths and low to moderate workloads will not be affected by increased response times. Note that this being high is an effect of some other problem. The peak is less useful, as it is very sensitive to individual glitches in performance, but it can show more detail of 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 those 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, the (regular) replication feature is processing for this volume. It applies to both types of Global Mirror and to Metro Mirror, but in each case 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 will capture a different set of writes).

► *Global Mirror Overlapping Write I/O Rate*

This statistic monitors the amount of overlapping I/O that the Global Mirror feature is handling for regular Global Mirror relationships. That is where an LBA is written again after the primary volume has been updated, but before the secondary volume has been updated for an earlier write to that LBA. To mitigate the effects of the overlapping I/Os, a journaling feature has been implemented, as discussed in “Colliding writes” on page 270.

► *Global Mirror secondary write lag*

This statistic is valid for regular Global Mirror 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 non-volatile 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 round-trip time 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 could not be instantly accepted into the system cache because cache was full. It is 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 Global Mirror to provide a Write Response Time below 1 ms.

This requirement is necessary because up to three round-trip messages within the local system will 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 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 very close to the low-level ping time between your sites. If this starts going 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 (say, 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 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 IO Percentage*

This is the fraction of messages that this node attempted to send through Fibre Channel 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 it is attached to, there might be benefit in getting 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 Fibre Channel performs flow control. Normally, this value is well under 1%. From 1 - 10% is a concerning level of congestion, but you might find the performance acceptable. Over 10% indicates severe congestion. This amount is also 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 used for messages to nodes in the same system are much more concerning than those 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 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 intersystem connectivity network
- Primary storage and remote storage
- IBM FlashSystem node canisters
- Storage area network

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 running (if possible)
- IBM Spectrum Control performance statistics data (if possible)
- The following information and logs from other components:
 - Intersystem network and switch details:
 - Technology
 - Bandwidth
 - Typical measured latency on the Intersystem 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 dedicated or shared; if so, the resource and amount of those resources 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

Intersystem network

For diagnostic purposes, ask the following questions about the intersystem network:

- ▶ Was network maintenance being performed?

Consider the hardware or software maintenance that is associated with intersystem network, such as updating firmware or adding more capacity.

- ▶ Is the intersystem 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 the storage metrics for the Global Mirror 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 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 used.

- Look at the port statistics for high response time in the 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 Global Mirror 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 IO 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 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 Global Mirror partnership to bring the combined mirrored foreground I/O and background copy I/O rates back within the remote links bandwidth.

- Look at the volume statistics (after the 1920 error is posted):
 - Source volume write throughput after the Global Mirror relationships were stopped.

If write throughput increases greatly (by 30% or more) after the Global Mirror relationships are stopped, the application host was attempting to perform more I/O than the remote link can sustain.

When the Global Mirror relationships are active, the overloaded remote link causes higher response times to the application host. This overload, in turn, decreases the throughput of application host I/O at the source volume. After the Global Mirror 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 Global Mirror 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 Global Mirror 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. In case of 1920 error triggered by the “max replication delay exceeded” condition, check the peek back-end write response time to see if it has exceeded the **maxreplicationdelay** 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 Global Mirror 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 back-end storage at the primary site regardless of Global Mirror. In case of 1920 error triggered by the “max replication delay exceeded” condition, check the peek back-end write response time to see if it has exceeded the **maxreplicationdelay** value around the 1920 occurrence.

However, if bad performance continues for a prolonged period, a false 1920 error might be flagged.

Node canister

For IBM FlashSystem node canister 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.

Global Mirror needs to synchronize its I/O processing across all nodes in the system to ensure data consistency. If any node is running out of CPU, it can affect all relationships. So check the CPU cores usage statistic. If it 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 has 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 CPU cores overloading.

In a primary system, if it is sufficiently busy, the write ordering detection in Global Mirror 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/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 on. It captures the round-trip time between nodes in the same system. Anything over 1.0 ms is surprisingly high, and will cause high secondary volume write response time. Values greater than 1 ms on primary system will cause an impact on write latency to Global Mirror primary volumes of 3 ms or more.

If you have 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 Fibre Channel device.

A good indicator of SAN congestion is the *Zero Buffer Credit Percentage* and *Port Send Delay IO Percentage* on the port statistics. For more information on buffer credit, see “Buffer credits” on page 282. If a port has more than 10% zero buffer credits, that will definitely cause a problem for all I/O, not just Global Mirror writes. Values from 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, this only really becomes an option 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 Global Mirror 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 Global Mirror relationship.

Consider also that FlashCopy can add significant workload to the back-end storage, especially when the background copy is active (see “Background copy considerations” on page 255). In cases where the remote system is used to create golden or practice copies for Disaster Recovery testing, the workload 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 is less of an issue because of high-performing flash back-end.

Careful planning of the back-end resources is particularly important with these kinds of scenarios. Reducing the FlashCopy background copy rate can also help to mitigate this situation. Furthermore, note that the FlashCopy copy-on-write process adds some latency by delaying the write operations on the primary volumes until the data is written to the FlashCopy target.

This process doesn't directly affect the Remote Copy operations since it is logically placed below the Remote Copy processing in the I/O stack, as shown in Figure 6-7 on page 244. Nevertheless, in some circumstances, especially with write intensive environments, the copy-on-write process tends to stress some of the internal resources of the system, such as CPU and memory. This condition can also 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 6-32 on page 312, 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). Allow 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 6-32 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 6-32 Effect of packet size (in bytes) versus the link size

In Figure 6-32, 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.

Recovery after 1920 errors

After a 1920 error occurs, the Global Mirror 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 Metro Mirror or Global Mirror 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 Global Mirror (or the Metro Mirror) relationships are synchronized again and back in a consistent state.

IBM Spectrum 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 once restarted. Instead, the system uses a secondary *change volume* 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 further details about the Consistency Protection feature, see *Implementing the IBM System Storage SAN Volume Controller with IBM Spectrum Virtualize V8.2.1*, SG24-7933.

To ensure that the system can handle the background copy load, delay restarting the Metro Mirror or Global Mirror relationship until a quiet period occurs. If the required link capacity is unavailable, you might experience another 1920 error, and the Metro Mirror or Global Mirror relationship might stop in an inconsistent state.

Copy services tools, like IBM Copy Services Manager (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 case of a persistent problem. CSM attempts an automatic restart for every occurrence of 1720/1920 error a certain number of times (determined by the `gmlinktolerance` value) within a 30 minute time period.

If the number of allowable automatic restarts is exceeded within the time period, CSM will not automatically restart GM on the next 1720/1920 error. Furthermore, with CSM it is possible to specify the amount of time, in seconds, in which the tool will wait after an 1720/1920 error before automatically restarting the GM. For more information, see [IBM Copy Services Manager](#).

Tip: When implementing automatic restart functions, it is advised to preserve the data consistency on GM target volumes during the resynchronization using features like Flashcopy or Consistency Protection.

Adjusting the Global Mirror settings

Although the default values are valid in most configurations, the settings of the `gmlinktolerance` and `gmmaxhostdelay` can be adjusted to accommodate particular environment or workload conditions.

For example, Global Mirror is designed to 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 reboot. Given that it is better to terminate a Global Mirror relationship than it is to reboot 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 “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 “If more than one third of the I/Os are extremely slow and that happens repeatedly for 30 seconds, then terminate the busiest relationship in the stream.”

So one approach is designed to pick up general slowness, and the other approach is designed to pick 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 Global Mirror 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 Global Mirror volumes.
- ▶ During periods when application hosts can tolerate extended response times and it is expected that the **gmLinkTolerance** feature might stop the Global Mirror relationships. For example, you are testing 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 Global Mirror relationships. Disabling the **gmLinkTolerance** parameter stops the Global Mirror relationships at the risk of exposing the test host to extended response times.

Another tunable parameter that interacts with the GM is the **maxReplicationDelay**. Note that the **maxReplicationDelay** settings do not mitigate the 1920 error occurrence because it actually adds a trigger to the 1920 error itself. However, the **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 above VMware example, an alternative option is to set the **maxReplicationDelay** to 30 seconds and leave the **gmLinkTolerance** and **gmMaxHostDelay** settings to their default. With these settings, the **maxReplicationDelay** timeout effectively handles the hung I/Os conditions, while the **gmLinkTolerance** and **gmMaxHostDelay** settings still provide an adequate mechanism to protect from ongoing performance issues.

6.4 Native IP replication

The native IP replication feature enables replication between any IBM Spectrum Virtualize products by using the built-in networking ports or optional 1/10 Gb adapter.

Native IP replication uses SANslide technology 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 significantly 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.

6.4.1 Native IP replication technology

Remote Mirroring over IP communication is supported on the IBM FlashSystem systems by using Ethernet communication links. The IBM Spectrum Virtualize Software IP replication uses innovative *Bridgeworks SANslide* technology to optimize network bandwidth and utilization. This new function enables the use of a lower-speed and lower-cost networking infrastructure for data replication.

Bridgeworks' SANSlide technology, which is integrated into the IBM Spectrum Virtualize Software, uses artificial intelligence to help optimize network bandwidth use 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 disaster recovery 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 6-33.

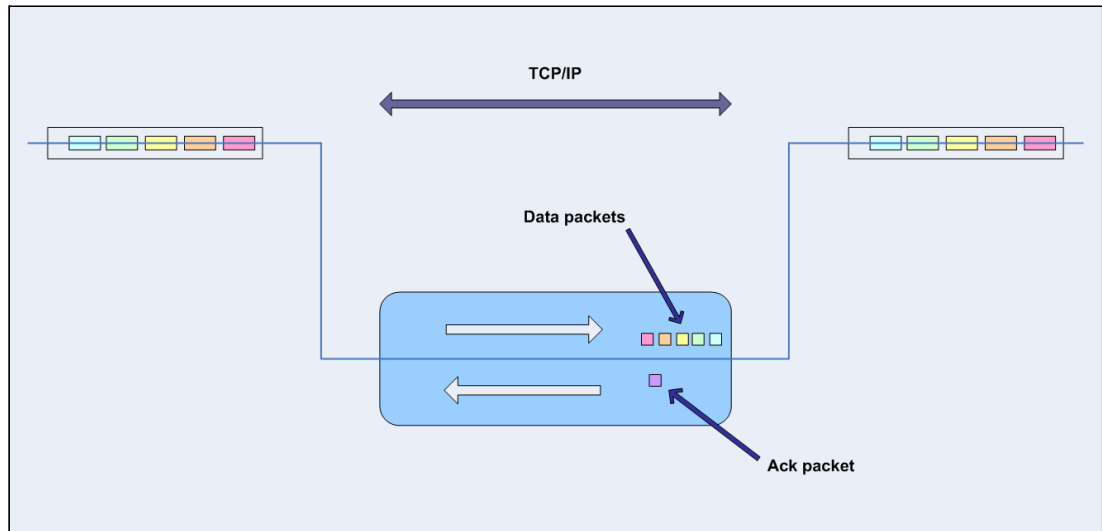


Figure 6-33 Typical Ethernet network data flow

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 artificial intelligence engine can dynamically adjust the number of VCs, receive window size, and packet size as appropriate 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 6-34.

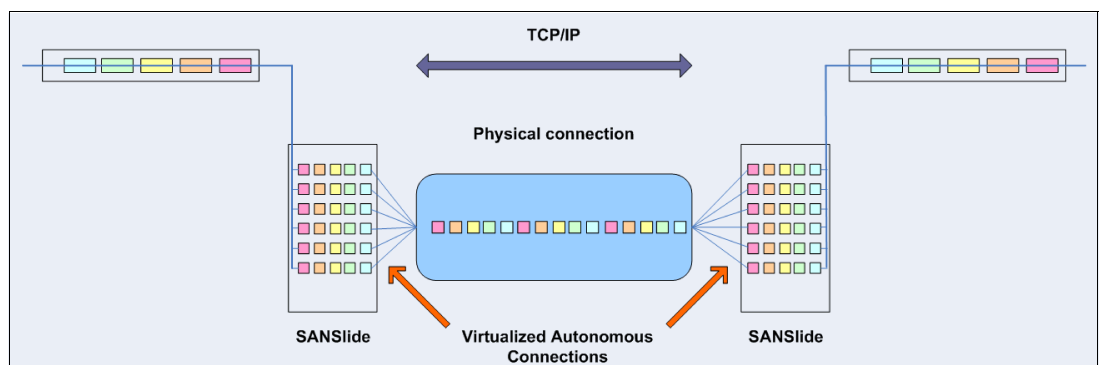


Figure 6-34 Optimized network data flow by using Bridgeworks SANSlide technology

For more information about this technology, see *IBM SAN Volume Controller and Storwize Family Native IP Replication*, REDP-5103.

Metro Mirror, Global Mirror, and Global Mirror Change Volume are supported with native IP partnership.

6.4.2 IP partnership limitations

The following prerequisites and assumptions must be considered before IP partnership between two IBM Spectrum Virtualize based systems can be established:

- ▶ The systems have 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 is 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 both IP and Fibre Channel (FC) partnerships, is limited to the current maximum that is supported, which is three partnerships (four systems total).
- ▶ 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 SAN Volume Controller.
- ▶ IP partnerships are supported with 25, 10, and 1 Gbps links. However, the intermix /on a single link is not supported.
- ▶ The maximum supported round-trip time is 80 ms for 1 Gbps links.
- ▶ The maximum supported round-trip time 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.
- ▶ Only nodes from two I/O Groups can have ports that are configured for an IP partnership.
- ▶ 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 LAN (VLAN) tagging of the IP addresses that are configured for Remote Copy is supported.
- ▶ Management IP and Internet SCSI (iSCSI) IP 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 with the following restrictions:
 - Only two direct-attach link are allowed.
 - The direct-attach links must be on the same I/O group.

- Use two port groups, where a port group contains only the two ports that are directly linked.
- ▶ Transmission Control Protocol (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 single Remote Copy data session per physical link can be established. It is intended that only one connection (for sending/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 an RC data session with an IP address on cluster B.

- ▶ The maximum throughput is restricted based on the use of 1 Gbps or 10 Gbps Ethernet ports. The output varies based on distance (for example, round-trip latency) and quality of communication link (for example, packet loss). The maximum achievable throughput is the following:
 - One 1 Gbps port can transfer up to 110 MB
 - One 10 Gbps port can transfer up to 500 MB

Table 6-9 shows a summary of the IP replication limits.

Table 6-9 IP replication limits

Remote Copy property	Maximum	Apply to	Comment	Remote Copy property
Inter-system IP partnerships per system	1	All models	A system can be partnered with up to three remote systems. A maximum of one of those can be IP and the other two FC.	Inter-system IP partnerships per system
I/O groups per system in IP partnerships	2	All models	The nodes from a maximum of two I/O groups per system can be used for IP partnership.	I/O groups per system in IP partnerships
Inter-site links per IP partnership	2	All models	A maximum of two inter-site links can be used between two IP partnership sites.	Inter-site links per IP partnership
Ports per node	1	All models	A maximum of one port per node can be used for IP partnership.	Ports per node
IP partnership Software Compression Limit	140 MBps	All models		IP partnership Software Compression Limit

6.4.3 VLAN support

VLAN tagging is supported for both 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 IP connections. 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 either iSCSI host attach or IP replication, the appropriate 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 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, there are no paths to IBM FlashSystem system during a node failure. Therefore, 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 the layer 2 level 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 IPs 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, appropriate 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 again when the configuration is complete.

6.4.4 IP Compression

IBM FlashSystem can leverage 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 similar to those experienced with existing Real-time Compression implementations.

No License: The IP compression feature does not require an RtC software license.

The data compression is made within the IP replication component of the IBM Spectrum Virtualize code. It can be used with all the Remote Copy technology (Metro Mirror, Global Mirror, and GMCV). The IP compression feature provides two kinds of compression mechanisms: the hardware compression and software compression. The 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 makes use of 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, the Comprestimator tool can be used to estimate the compression ratio 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 enabled in both directions, the IP compression can be enabled in only one direction. IP compression is supported for IPv4 and IPv6 partnerships.

6.4.5 Remote Copy groups

This section describes Remote Copy groups (or Remote Copy port groups) and different ways to configure the links between the two remote systems. The two systems can be connected to each other over one link or, at most, two links. The concept of Remote Copy port groups was introduced to address the requirement to enable the systems to know about the physical links between the two sites.

Remote Copy port group ID is a numerical tag that is associated with an IP port of system that indicates to which physical IP link it is connected. Multiple IBM FlashSystem canisters can be connected to the same physical long-distance link, and must therefore share a Remote Copy port group ID.

In scenarios with two physical links between the local and remote clusters, two Remote Copy port group IDs must be used to designate which IP addresses are connected to which physical link. This configuration must be done by the system administrator by using the GUI or the `cfgportip` CLI command. Note that the relationship between the physical links and the Remote Copy group IDs is not policed by the IBM Spectrum Virtualize code. This means that two different Remote Copy group can be used with a single physical link and vice versa.

Remember: IP ports on both partners must have been configured with identical Remote Copy port group IDs for the partnership to be established correctly.

The system IP addresses that are connected to the same physical link should be designated with identical Remote Copy port groups. The IBM FlashSystem systems support three Remote Copy groups: 0, 1, and 2.

The IP addresses are, by default, in Remote Copy port group 0. Ports in port group 0 are not considered for creating Remote Copy data paths between two systems. For partnerships to be established over IP links directly, IP ports must be configured in Remote Copy group 1 if a single inter-site link exists, or in Remote Copy groups 1 and 2 if two inter-site links exist.

You can assign one IPv4 address and one IPv6 address to each Ethernet port on the IBM FlashSystem systems. Each of these IP addresses can be shared between iSCSI host attach and the IP partnership. The user must configure the required IP address (IPv4 or IPv6) on an Ethernet port with a Remote Copy port group.

The administrator might want to use IPv6 addresses for Remote Copy operations and use IPv4 addresses on that same port for iSCSI host attach. This configuration also implies that for two systems to establish an IP partnership, both systems must have IPv6 addresses that are configured.

Administrators can choose to dedicate an Ethernet port for IP partnership only. In this case, host access must be explicitly disabled for that IP address and any other IP address that is configured on that Ethernet port.

Failover operations within and between port groups

Within one remote-copy port group, only one port from each system is selected for sending and receiving Remote Copy data at any one time. Therefore, on each system, at most one port for each remote-copy port group is reported as used.

If the IP partnership becomes unable to continue over an IP port, the system fails over to another port within that remote-copy port group. Some reasons this might occur are 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 remote-copy port group. If only one link is configured between the two systems, configure two ports (one per node) within the remote-copy port group. You can configure these two ports on two nodes within the same I/O group or within separate I/O groups.

While failover is in progress, no connections in that remote-copy port group 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 remote-copy port groups, the failover process within each port group 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 intersystem links, you must configure two remote-copy port groups.

When a node fails in this scenario, the IP partnership can continue over the other link until the node failure is rectified. Failback then happens 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 if there is more than one I/O Group and the node in the other I/O group has IP ports pre-configured within the remote-copy port group, the discovery is triggered. The discovery makes the active IP partnership data path available from the new IP address.

6.4.6 Supported configurations examples

Multiple IP partnership configurations are available depending on the number of physical links and the number of nodes. In the following sections, some example configurations are described.

Single inter-site link configurations

Consider two 2-node systems in IP partnership over a single inter-site link (with failover ports configured), as shown in Figure 6-35 on page 321.

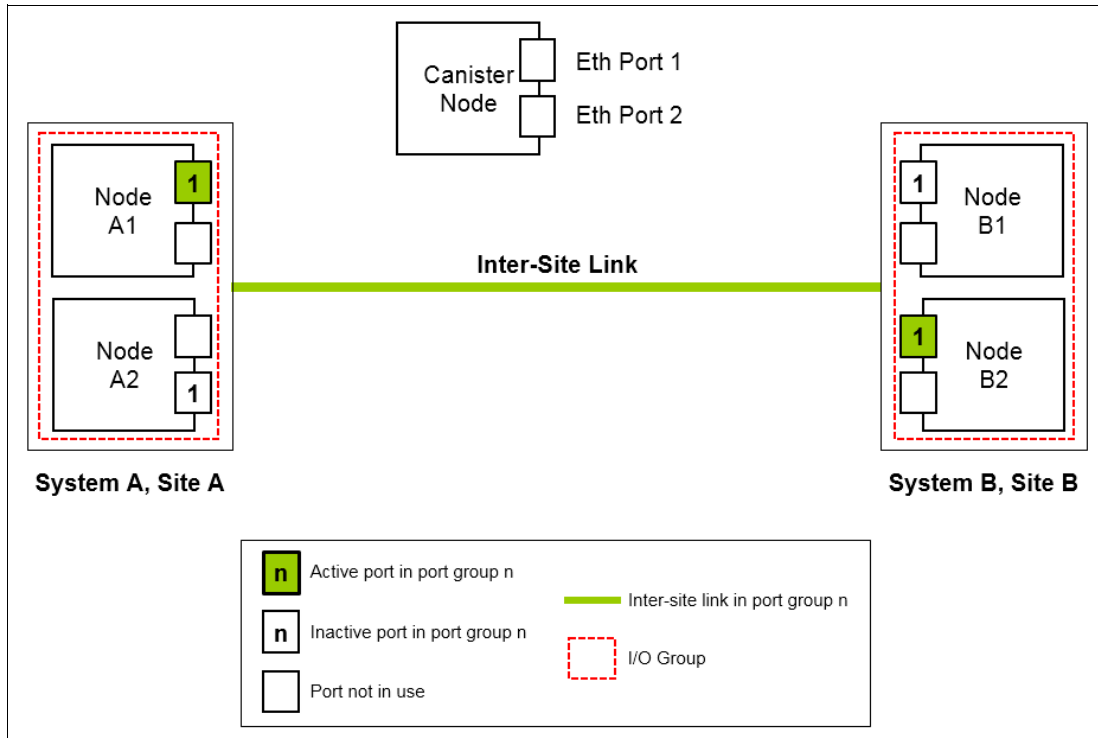


Figure 6-35 Only one Remote Copy group on each system and nodes with failover ports configured

Figure 6-35 shows two systems: System A and System B. A single Remote Copy port group 1 is configured on two Ethernet ports, one each on Node A1 and Node A2 on System A. Similarly, a single Remote Copy port group is configured on two Ethernet ports on Node B1 and Node B2 on System B.

Although two ports on each system are configured for Remote Copy port group 1, 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 node in the I/O Group behaves as a standby port that is used during a node failure. If Node A1 fails in System A, IP partnership continues servicing replication I/O from Ethernet Port 2 because a failover port is configured on 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 `1sport ip` output (reported as used).

This configuration has the following characteristics:

- ▶ Each node in the I/O group has the same Remote Copy port group that is configured. However, only one port in that Remote Copy port group is active at any time on each system.
- ▶ If Node A1 in System A or Node B2 in System B fails in the respective systems, IP partnership 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 then recover.

Figure 6-36 shows a configuration with two 4-node systems in IP partnership over a single inter-site link (with failover ports configured).

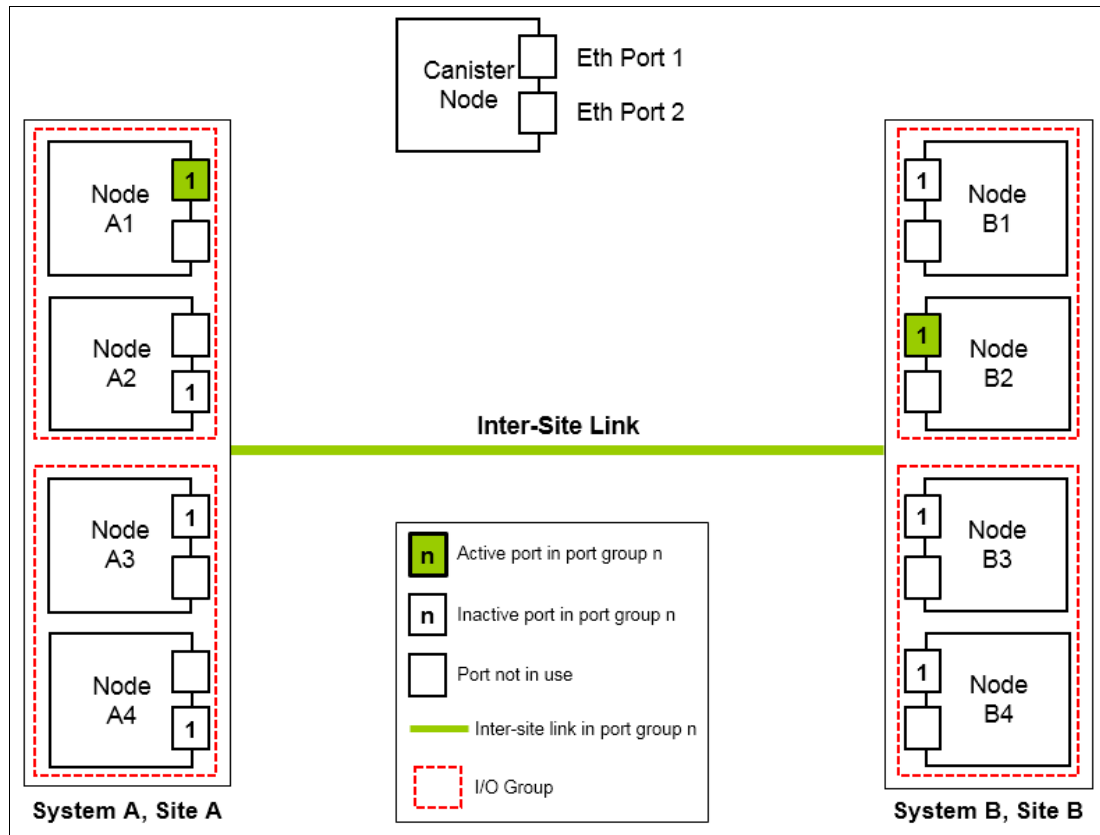


Figure 6-36 Multinode systems single inter-site link with only one Remote Copy port group

Figure 6-36 shows two 4-node systems: System A and System B. A single Remote Copy port group 1 is configured on nodes A1, A2, A3, and A4 on System A, Site A, and on nodes B1, B2, B3, and B4 on System B, Site B.

Although four ports are configured for Remote Copy group 1, only one Ethernet port in each Remote Copy port group on each system actively participates in the IP partnership process. Port selection is determined by a path configuration algorithm. The other ports play the role of standby ports.

If Node A1 fails in System A, the IP partnership selects one of the remaining ports that is configured with Remote Copy port group 1 from any of the nodes from either of the two I/O groups in System A. However, it might take some time (generally seconds) for discovery and path configuration logic to reestablish paths post failover. This process can cause partnerships to change to the Not_Present state.

This result causes Remote Copy relationships to stop. The administrator might need to manually verify the issues in the event log and start the relationships or Remote Copy consistency groups, if they do not automatically recover. The details of the particular IP port actively participating in the IP partnership process is provided in the `lspport ip` view (reported as used). This configuration has the following characteristics:

- ▶ Each node has the Remote Copy port group that is configured in both I/O groups. However, only one port in that Remote Copy port group remains active and participates in IP partnership on each system.

- ▶ If Node A1 in System A or Node B2 in System B encounter some failure in the system, IP partnerships discovery 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 then recover.
- ▶ The bandwidth of the single link is used completely.

An eight-node system in IP partnership with four-node system over single inter-site link is shown in Figure 6-37.

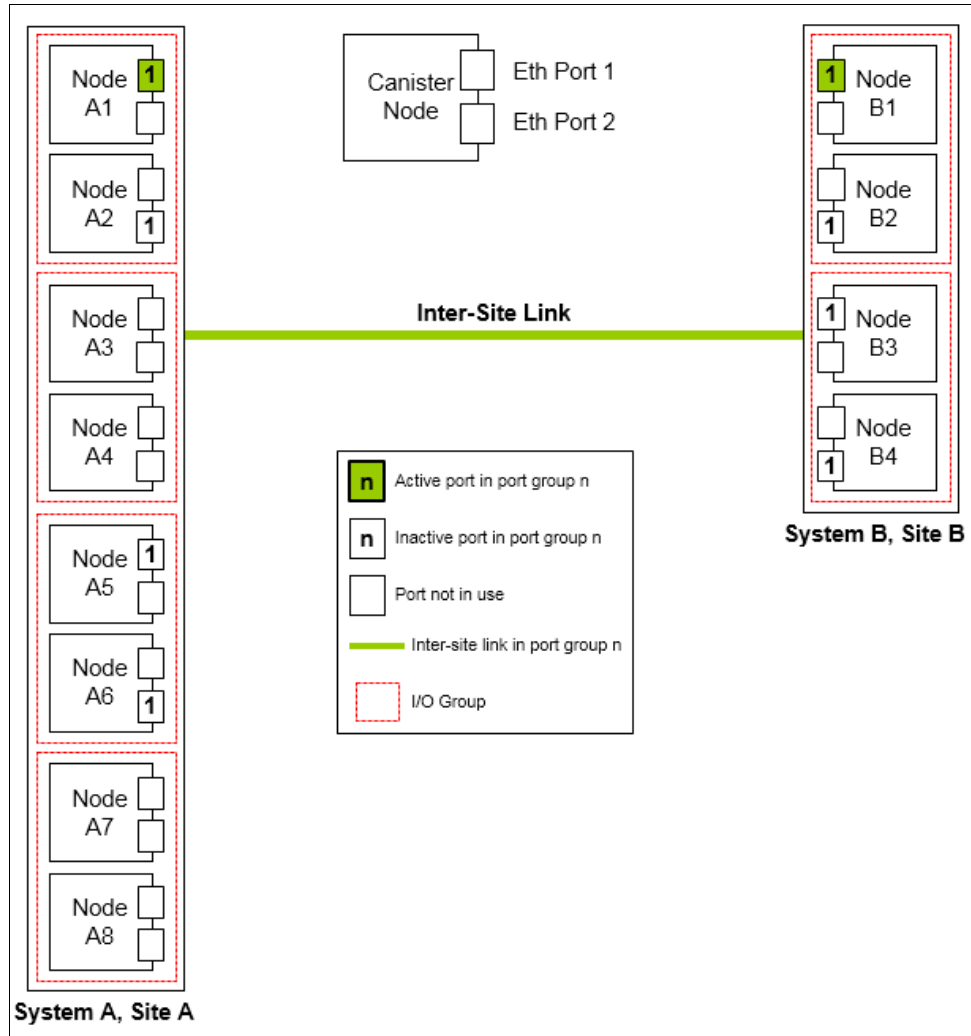


Figure 6-37 Multinode systems single inter-site link with only one Remote Copy port group

Figure 6-37 shows an eight-node system (System A in Site A) and a four-node system (System B in Site B). A single Remote Copy port group 1 is configured on nodes A1, A2, A5, and A6 on System A in Site A. Similarly, a single Remote Copy port group 1 is configured on nodes B1, B2, B3, and B4 on System B in Site B.

Although there are four I/O groups (eight nodes) in System A, any two I/O groups at maximum are supported to be configured for IP partnerships. If Node A1 fails in System A, the IP partnership continues to use one of the ports that is configured in Remote Copy port group from any of the nodes from either of the two I/O groups 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 the relationships if they do not auto-recover. The details of which particular IP port is actively participating in IP partnership process is provided in `lspport ip` output (reported as used).

This configuration has the following characteristics:

- ▶ Each node has the Remote Copy port group that is configured in both the I/O groups that are identified for participating in IP Replication. However, only one port in that Remote Copy port group remains active on each system and participates in IP Replication.
- ▶ If the Node A1 in System A or the 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 two 2-node systems with two inter-site links configuration is depicted in Figure 6-38.

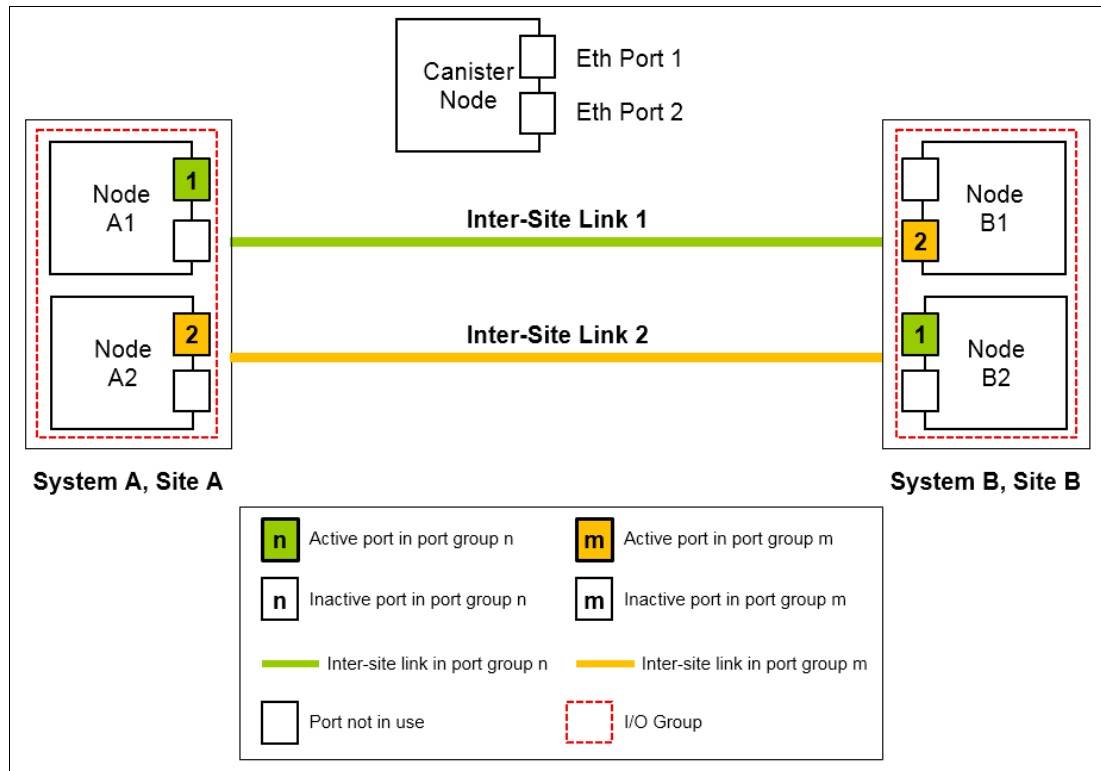


Figure 6-38 Dual links with two Remote Copy groups on each system configured

As shown in Figure 6-38, Remote Copy port groups 1 and 2 are configured on the nodes in System A and System B because two inter-site links are available. In this configuration, the failover ports are not configured on partner nodes in the I/O group. Rather, the ports are maintained in different Remote Copy port groups on both of the nodes. They can remain active and participate in IP partnership by using both of the links.

However, if either of the nodes in the I/O group fail (that is, if Node A1 on System A fails), the IP partnership continues only from the available IP port that is configured in Remote Copy port group 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 has the following characteristics:

- ▶ There are two inter-site links, and two Remote Copy port groups are configured.
- ▶ Each node has only one IP port in Remote Copy port group 1 or 2.
- ▶ Both the IP ports in the two Remote Copy port groups participate simultaneously in IP partnerships. Therefore, both of the links are used.
- ▶ During node failure or link failure, the IP partnership traffic continues from the other available link and the port group. 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 node failure or link failure is resolved and failback happens, the entire bandwidth of both of the links is available as before.

A configuration with two 4-node systems in IP partnership with dual inter-site links is shown in Figure 6-39 on page 326.

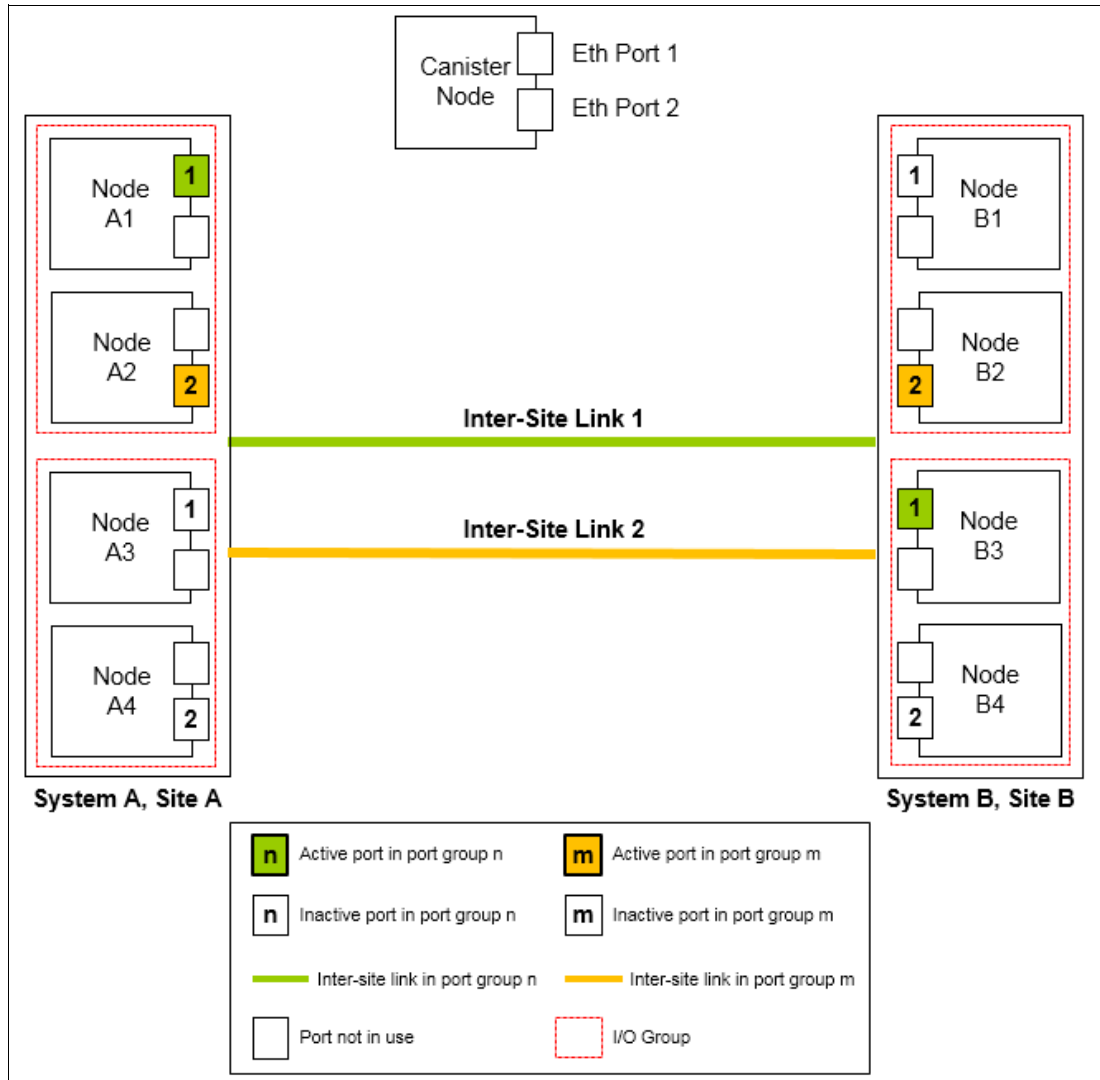


Figure 6-39 Multinode systems with dual inter-site links between the two systems

Figure 6-39 shows two 4-node systems: System A and System B. This configuration is an extension of configuration depicted in Figure 6-38 on page 324 to a multinode multi-I/O group environment.

As seen in this configuration, there are two I/O groups. Each node in the I/O group has a single port that is configured in Remote Copy port groups 1 or 2.

Although two ports are configured in Remote Copy port groups 1 and 2 on each system, only one IP port in each Remote Copy port group on each system actively participates in IP partnership. The other ports that are configured in the same Remote Copy port group act as standby ports during a failure. A path-configuration algorithm determines which port in a configured Remote Copy port group will participate in IP partnership at any moment.

In this configuration, if Node A1 fails in System A, IP partnership traffic continues from Node A2 (that is, Remote Copy port group 2). At the same time, the failover also causes discovery in Remote Copy port group 1. Therefore, the IP partnership traffic continues from Node A3, on which Remote Copy port group 1 is configured. The details of the particular IP port that is actively participating in IP partnership process is provided in the `1sport ip` output (reported as used).

This configuration has the following characteristics:

- ▶ Each node has the Remote Copy port group that is configured in the I/O groups 1 or 2. However, only one port per system in both Remote Copy port groups remains active and participates in IP partnership.
- ▶ Only a single port per system from each configured Remote Copy port group participates simultaneously in IP partnership. Therefore, both of the links are used.
- ▶ During node failure or port failure of a node that is actively participating in IP partnership, IP partnership continues from the alternative port because another port is in the system in the same Remote Copy port group, but in a different I/O Group.
- ▶ The pathing algorithm can start discovery of available port in the affected Remote Copy port group in the second I/O group and pathing is reestablished. This process restores the total bandwidth, so both of the links are available to support IP partnership.

Finally, an eight-node system in IP partnership with a four-node system over dual inter-site links is depicted in Figure 6-40.

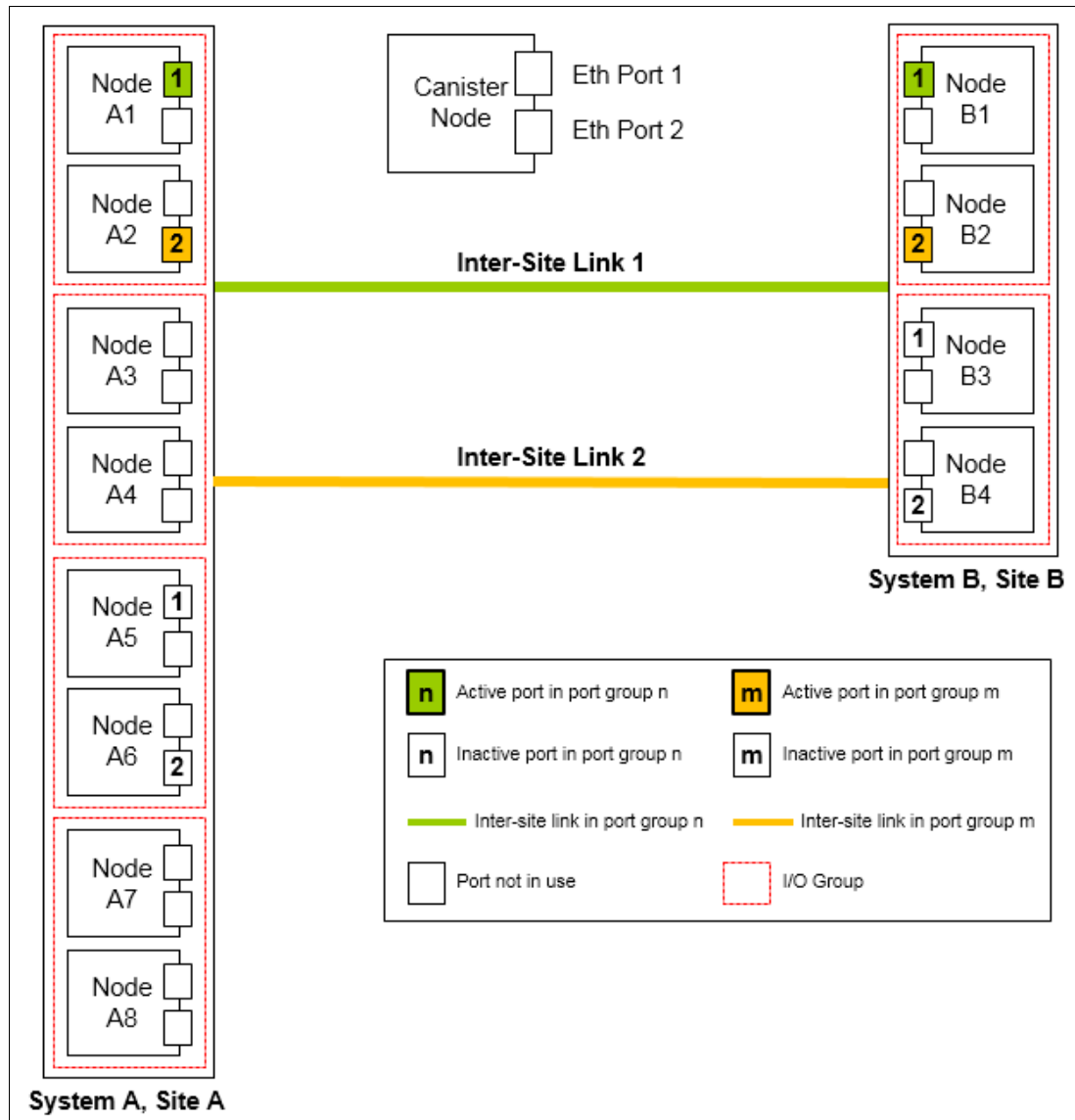


Figure 6-40 Multinode systems with dual inter-site links between the two systems

Figure 6-40 shows an eight-node System A in Site A and a four-node System B in Site B. Because a maximum of two I/O groups in IP partnership is supported in a system, although there are four I/O groups (eight nodes), nodes from only two I/O groups are configured with Remote Copy port groups in System A. The remaining or all the I/O groups can be configured to be Remote Copy partnerships over FC.

In this configuration, two links and two I/O groups are configured with Remote Copy port groups 1 and 2. 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 partnership. Even if Node A5 and Node A6 are configured with Remote Copy port groups properly, active IP partnership traffic on both of the links can be driven from Node A1 and Node A2 only.

If Node A1 fails in System A, IP partnership traffic continues from Node A2 (that is, Remote Copy port group 2). The failover also causes IP partnership traffic to continue from Node A5 on which Remote Copy port group 1 is configured. The details of the particular IP port actively participating in IP partnership process is provided in the `lspport ip` output (reported as used).

This configuration has the following characteristics:

- ▶ There are two I/O Groups with nodes in those I/O groups that are configured in two Remote Copy port groups because there are two inter-site links for participating in IP partnership. However, only one port per system in a particular Remote Copy port group remains active and participates in IP partnership.
- ▶ One port per system from each Remote Copy port group participates in IP partnership simultaneously. Therefore, both of the links are used.
- ▶ If a node or port on the node that is actively participating in IP partnership fails, the Remote Copy (RC) data path is established from that port because another port is available on an alternative node in the system with the same Remote Copy port group.
- ▶ The path selection algorithm starts discovery of available ports in the affected Remote Copy port group in the alternative I/O groups and paths are reestablished. This process restores the total bandwidth across both links.
- ▶ The remaining or all the I/O groups can be in Remote Copy partnerships with other systems.

Remote Copy port groups: As described in this section, configuring two Remote Copy port group provides more bandwidth and more resilient configurations, in case of a link failure. Two Remote Copy port groups can be configured with a single physical link. This configuration makes sense only if the total link bandwidth exceeds the aggregate bandwidth of two Remote Copy port groups together. Creating two Remote Copy port groups when the link bandwidth does not provide the aggregate throughput can lead to network resources contention and bad link performance.

6.4.7 Native IP replication performance consideration

A number of 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, many improvements have been made to make the IP replication better performing and more reliable.

Nevertheless, in presence of poor quality networks that have significant packet loss and high latency, the actual usable bandwidth might decrease considerably.

Figure 6-41 shows the throughput trend for a 1 Gbps port in respect of the packet loss ratio and the latency.

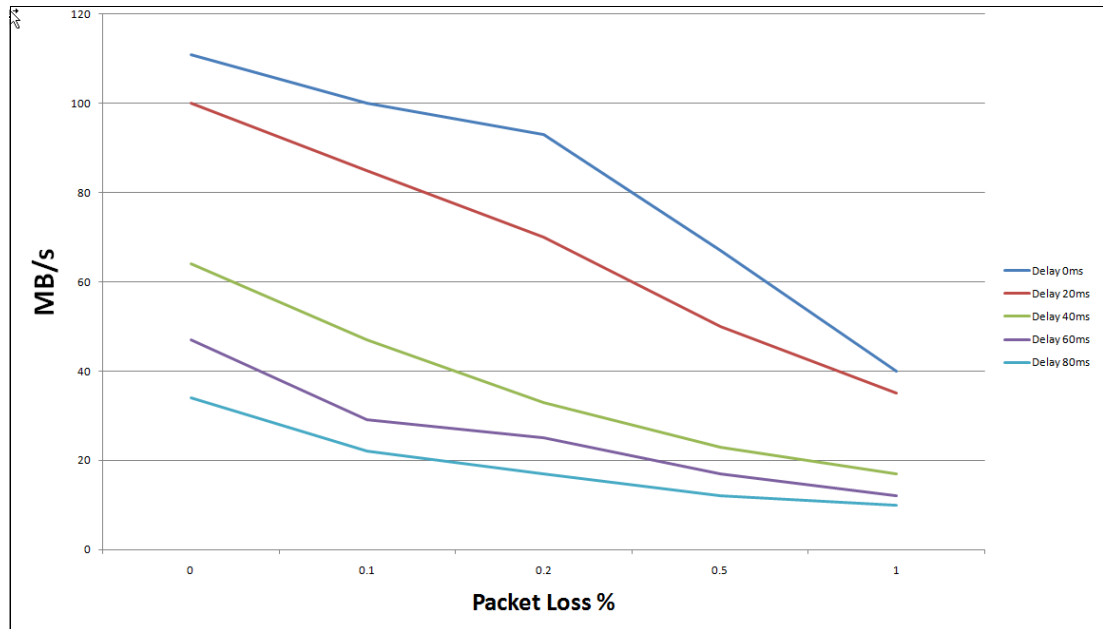


Figure 6-41 1 Gbps port throughput trend

The chart in Figure 6-41 shows how the combined effect of the packet loss and the latency can lead to a throughput reduction of more than 85%. For these reasons, the IP replication option should be considered only for the replication configuration that is not affected by poor quality and poor performing networks. Due to its characteristic of low-bandwidth requirement, the GMCV is the preferred solution with the IP replication.

To improve the performance when using compression and IP partnership in the same system, it is advised that you 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.

6.5 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 has been 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 of 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, just as 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 based on Remote Copy (Global or Metro Mirror 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 6-42 provides an overview of Volume Mirroring.

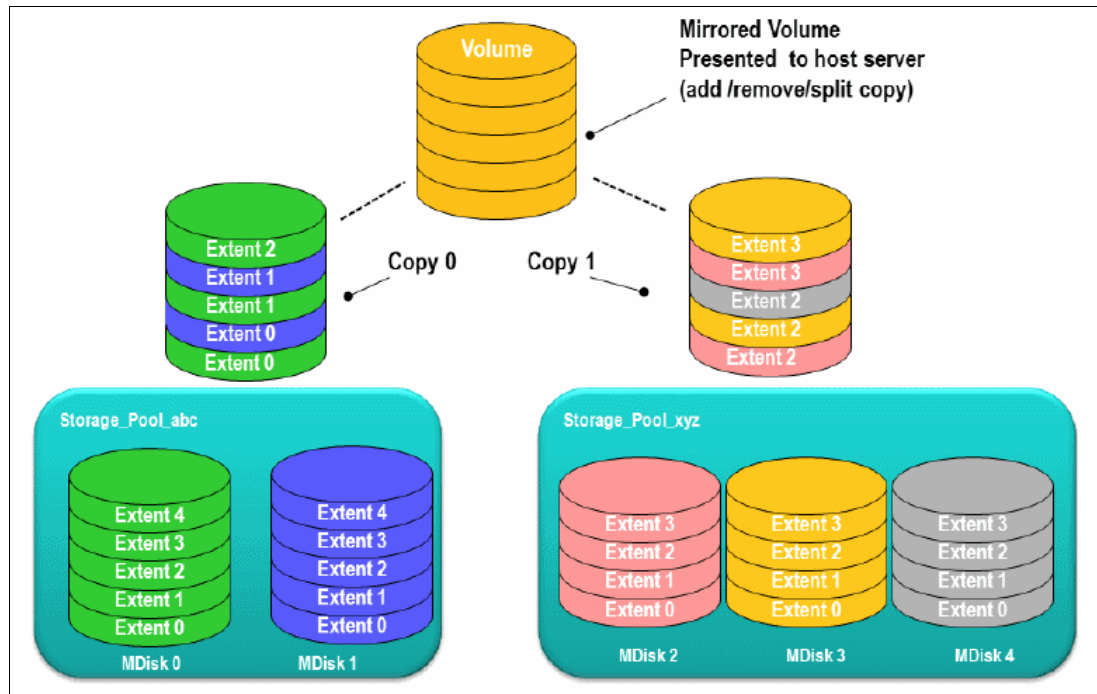


Figure 6-42 Volume Mirroring overview

6.5.1 Read and write operations

Read and 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 Enhanced Stretched Cluster topology, which applies to SAN Volume Controller system types only.

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 remembers which areas of the volume are written and resynchronizes these areas when both copies are available.
- ▶ The remaining copy can service read I/O when the failing one is offline, without user intervention.

6.5.2 Volume mirroring use cases

Volume Mirroring offers the capability to provide extra copies of the data that can be used for High Availability solutions and data migration scenarios. You can convert a non-mirrored volume into a mirrored volume by adding a copy. When a copy is added 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 non-mirrored volume by deleting one copy or by splitting one copy to create a new non-mirrored volume.

Access: Servers can access the volume during the synchronization processes described.

You can use mirrored volumes to provide extra protection for your environment or to perform a migration. This solution offers several options:

- ▶ **Export to Image mode**

This option allows you to move storage from *managed mode* to *image mode*. This option is useful if you are using IBM FlashSystem 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.

Using *Export to image mode* allows you to migrate data by using the Copy Services functions and then return control to the native array, while maintaining access to the hosts.

- ▶ **Import to Image mode**

This option allows you to import an existing storage MDisk or logical unit number (LUN) with its existing data from an external storage system, without putting metadata on it. The existing data remains intact. After you import it, the volume mirroring function can be used to migrate the storage to the other locations, while the data remains accessible to your hosts.

- ▶ **Volume cloning using Volume Mirroring and then using the Split into New Volume option**

This option allows any volume to be cloned without any interruption to the host access. You have to create two mirrored copies of data and then break the mirroring with the split option to make two independent copies of data. This option doesn't apply to already mirrored volumes.

- ▶ **Volume pool migration using the volume mirroring option**

This option allows any volume to be moved between storage pools without any interruption to the host access. You might use this option to move volumes as an alternative to 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 also it allows you to move volumes among pools with different extent sizes. This option doesn't 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. See 4.3.7, "Data migration with DRP" on page 135.

- ▶ **Volume capacity saving change**

This option allows you to modify the capacity saving 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 specifying a

different capacity saving for the newly created copy. This option doesn't 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 is unable to update the state information, a mirrored volume might need to be taken offline to maintain data integrity. To ensure the high availability 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 available. This behavior occurs because 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.

The following are other Volume Mirroring usage cases and characteristics:

- ▶ Creating a mirrored volume:
 - The maximum number of copies is two.
 - Both copies are created with the same virtualization policy, by default.
To have a volume mirrored using different policies, you need to add a volume copy with a different policy to a volume that has only one copy.
 - Both copies can be located 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 to be 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 new volume with the split copy:
 - This function is only allowed when the volume copies are synchronized. Otherwise, use the **-force** command.
 - It is not possible to recombine the two volumes after they have been 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. It can iterate by starting at a specific LBA by using the **-startlba** parameter.
 - Create virtual medium errors where there are differences. This is very useful in case of back-end data corruption.
 - Correct the differences that are found (reads from primary copy and writes to secondary copy).

- ▶ View to list volumes affected by a back-end disk subsystem being offline:
 - Assumes that a standard use is for mirror between disk subsystems.
 - Verifies that mirrored volumes remain accessible if a disk system is being shut down.
 - Reports 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/compressed volumes shrinking.

- ▶ Delete a volume. When a volume gets deleted, all copies get deleted.
- ▶ Migration commands apply to a specific volume copy.
- ▶ Out-of-sync bitmaps share the bitmap space with FlashCopy and Metro Mirror/Global Mirror. Creating, expanding, and changing I/O groups might fail if there is insufficient memory.
- ▶ GUI views contain volume copy identifiers.

6.5.3 Mirrored volume components

Note the following points regarding mirrored volume components:

- ▶ A mirrored volume is always composed of two copies (copy 0 and copy 1).
- ▶ 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 non-mirrored volume. That is, upper layers within the cluster software, such as FlashCopy and Metro Mirror/Global Mirror, 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.

6.5.4 Volume Mirroring synchronization options

As soon as a volume is created with two copies, copies are in the *out-of-sync* state. The primary volume copy (located in the first specified storage pool) is defined as in sync and the secondary volume copy 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 6-10 on page 335), or at the defined rate while creating or modifying the volume. For more information on the effect of the copy rate setting, see 6.5.5, “Volume Mirroring performance considerations” on page 335. When the synchronization process is completed, the volume mirroring copies are *in-sync* state.

By default, when a mirrored volume is created a format process is also initiated. This process guarantees that the volume data is zeroed, avoiding access to data that is still present on reused extents.

This format process runs in background at the defined synchronization rate, as shown in Table 6-10 on page 335. Before Spectrum Virtualize version 8.4, the format processing

overwrite with zeros only the Copy 0 and then synchronize the Copy 1. With version 8.4 or later, the format process is initiated concurrently to both volume mirroring copies and thus avoiding the second synchronization step.

You can specify that a volume is synchronized (`-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 has been 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 has been written to disk and unwritten areas return zeros (0). If the synchronization between the volume copies has been lost, the resynchronization process is incremental. This term means that only grains that have been written to need to be copied, and then get synchronized volume copies again.

The progress of the volume mirror synchronization can be obtained from the GUI or by using the `lsvdi sksyncprogress` command.

6.5.5 Volume Mirroring performance considerations

Because the writes of mirrored volumes always occur to both copies, mirrored volumes put more workload on the cluster, the back-end disk subsystems, and the connectivity infrastructure. The mirroring is symmetrical, and writes are only acknowledged 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: Locate volume copies of one volume on storage pools of 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 always only read from one volume copy, reads are not faster than without Volume Mirroring.

However, be aware that this 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 - 100. A value of 0 disables synchronization.

Table 6-10 shows the relationship between the *rate value* and the *data copied per second*.

Table 6-10 Relationship between the rate value and the data copied per second

User-specified rate attribute value per volume	Data copied/sec
0	Synchronization is disabled
1 - 10	128 KB
11 - 20	256 KB
21 - 30	512 KB

User-specified rate attribute value per volume	Data copied/sec
31 - 40	1 MB
41 - 50	2 MB ** 50% is the default value
51 - 60	4 MB
61 - 70	8 MB
71 - 80	16 MB
81 - 90	32 MB
91 - 100	64 MB

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 FlashSystem configurations, the settings of the copy rate can considerably affect the performance in scenarios where a back-end storage failure occurs. For instance, 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 the 300 mirrored volumes starts at the same time. With a copy rate of 100 for each volume, this process would add a theoretical workload of 18.75 GBps, which will considerably overload the system.

The general suggestion for the copy rate settings is then to evaluate the impact of massive resynchronization and set the parameter accordingly. Consider setting the copy rate to high values for initial synchronization only, and with a limited number of volumes at a time. Alternatively, consider defining a volume provisioning process that allows the safe creation of already synchronized mirrored volumes, as described in 6.5.4, “Volume Mirroring synchronization options” on page 334.

Volume mirroring I/O Time-out configuration

A mirrored volume has pointers to the two copies of data, usually in different storage pools, and each write completes on both copies before the host receives I/O completion status. For a synchronized mirrored volume, if a write I/O to a copy has failed or a long timeout has expired, then system has completed all available controller level Error Recovery Procedures (ERPs). In this case, that copy is taken offline and goes out of sync. The volume remains online and continues to service I/O requests from the remaining copy.

The *Fast Failover* feature isolates hosts from temporarily 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 Fibre Channel abort sequence can take around 25 seconds.

When the stop is completed, 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 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 the 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 that will, typically, quickly complete 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 Enhanced Stretched Cluster configurations is *latency*.

6.5.6 Bitmap space for out-of-sync volume copies

The grain size for the synchronization of volume copies is 256 KB. One grain takes up one 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 Metro Mirror) and standard FlashCopy with a grain size of 256 KB (Table 6-11).

Table 6-11 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 Metro Mirror, Global Mirror, 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 allocated and available on each I/O Group with the CLI, see the Example 6-4.

Example 6-4 A `lsiogrp` and `chiogrp` command example

```
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_FlashSystem:ITS0:superuser>lsiogrp io_grp0
id 0
.
lines removed for brevity
.
remote_copy_free_memory 19.9MB
mirroring_total_memory 20.0MB
mirroring_free_memory 20.0MB
raid_total_memory 40.0MB
raid_free_memory 40.0MB
.
lines removed for brevity
.
IBM_FlashSystem:ITS0:superuser>chiogrp -feature mirror -size 64 io_grp0
IBM_FlashSystem:ITS0:superuser>lsiogrp io_grp0
id 0
.
lines removed for brevity
.
remote_copy_free_memory 19.9MB
mirroring_total_memory 64.0MB
mirroring_free_memory 64.0MB
.
lines removed for brevity
.

```



Business continuity

Business continuity (BC) 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 can now manage both planned and unplanned outages, and provide the flexibility and cost efficiencies that are available from cloud-computing models.

This chapter briefly describes the HyperSwap solutions for IBM Spectrum Virtualize systems. Technical details or implementation guidelines are not presented in this chapter because they are described in separate publications.

Important: This book was written specifically for IBM FlashSystems products. Therefore, it does not cover Stretched Cluster and Enhanced Stretched Cluster topologies. For IBM SAN Volume Controller detailed information, refer to the IBM Redbooks publication *IBM SAN Volume Controller Best Practices and Performance Guidelines*, SG24-8502.

This book does not cover the 3-site replication solutions, available with the IBM Spectrum Virtualize code version 8.3.1 or later. It is covered by the *IBM Spectrum Virtualize 3-Site Replication*, SG24-8504 publication.

This chapter includes the following sections:

- ▶ 7.1, “Business continuity with HyperSwap” on page 340
- ▶ 7.2, “Third site and IP quorum” on page 344
- ▶ 7.3, “HyperSwap Volumes” on page 346
- ▶ 7.4, “Other considerations and general recommendations” on page 348

7.1 Business continuity with HyperSwap

The *HyperSwap* high-availability feature in the IBM Spectrum Virtualize and FlashSystems products enables business continuity during a hardware failure, power outage, connectivity problem, or other disasters, such as fire or flooding.

It provides highly available volumes accessible through two sites located at 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 is completed. HyperSwap automatically optimizes itself to minimize data that is transmitted between sites, and to minimize host read and write latency. For information on the optimization algorithm, see 7.3, “HyperSwap Volumes” on page 346.

HyperSwap has the following key features:

- ▶ Works with all IBM Spectrum Virtualize products except for IBM FlashSystem 5010.
- ▶ Uses intra-cluster synchronous Remote Copy (Active-Active Metro Mirror) capability, with change volumes and access I/O group technologies.
- ▶ Makes a host’s volumes accessible across two IBM Spectrum Virtualize I/O groups in a clustered system by using the Active-Active Metro Mirror relationship. The volumes are presented as a single volume to the host.
- ▶ Works with the standard multipathing drivers that are available on various host types. Additional host support is not required to access the highly available volumes.

The IBM Spectrum Virtualize HyperSwap configuration requires that at least one control enclosure is implemented in each location. Therefore, a minimum of two control enclosures for each cluster is needed to implement HyperSwap. Configuration with three or four control enclosures is also supported for the HyperSwap.

The typical IBM FlashSystems HyperSwap implementation is depicted in Figure 7-1.

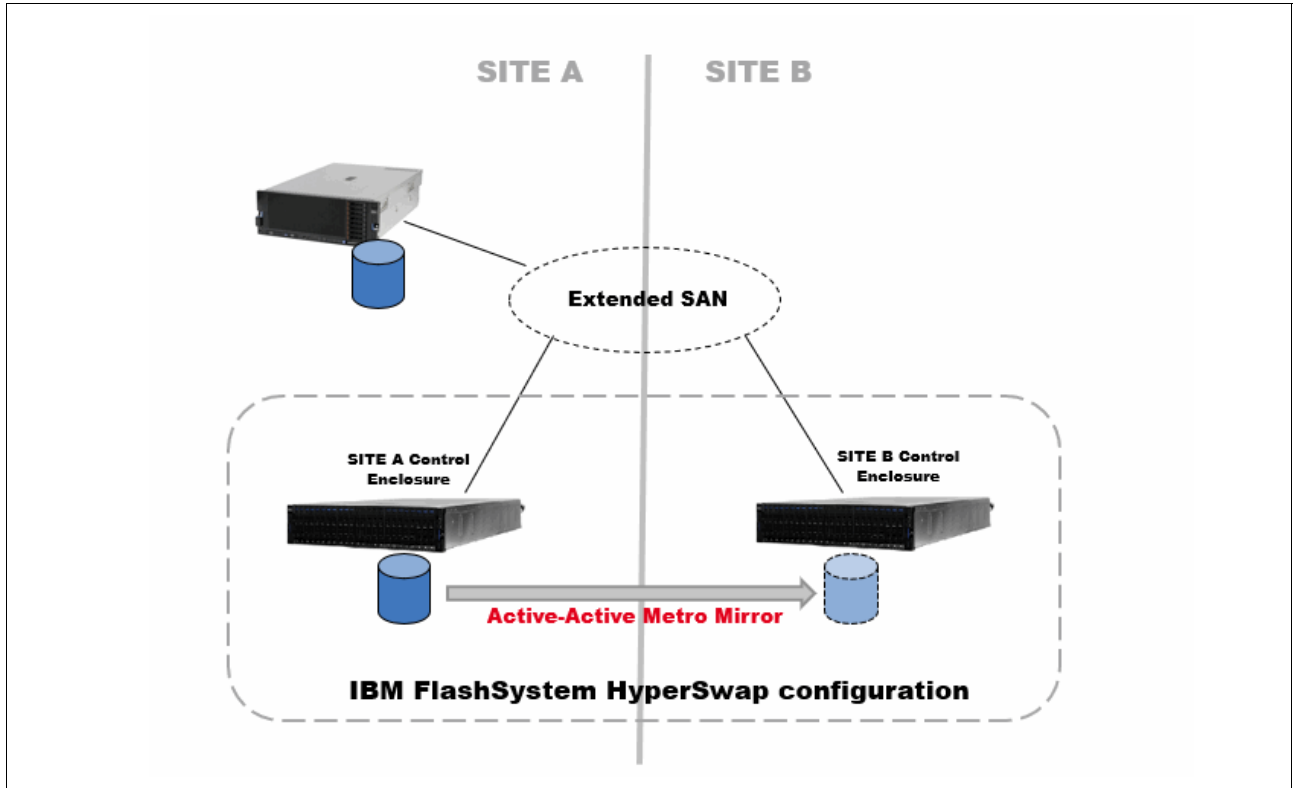


Figure 7-1 Typical HyperSwap configuration with IBM FlashSystem

With a copy of the data that is stored at each location, HyperSwap configurations can handle different failure scenarios.

Figure 7-2 shows how HyperSwap operates in a storage failure in one location. In this case, after the storage failure was detected in Site A, the HyperSwap function provides access to the data through the copy in the surviving site.

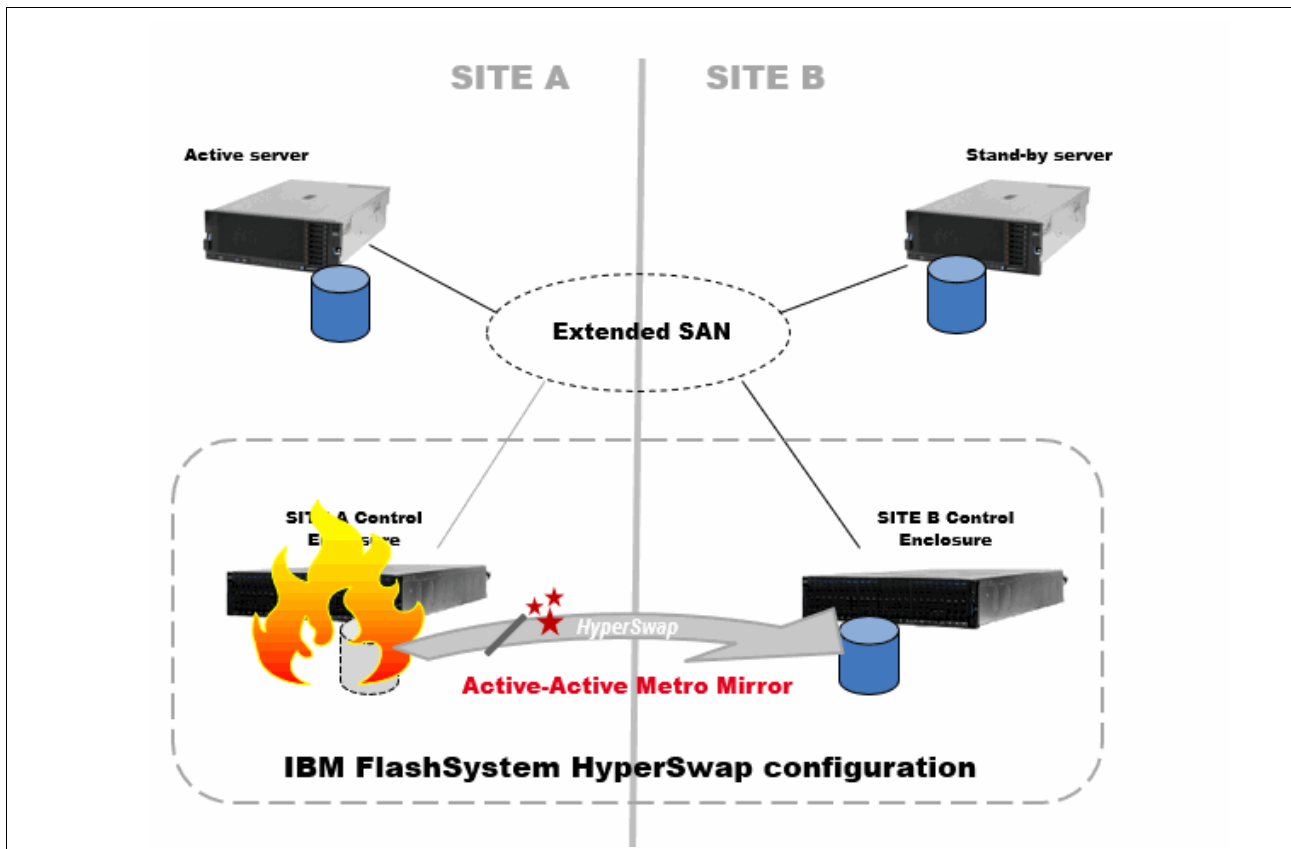


Figure 7-2 IBM FlashSystem HyperSwap in a storage failure scenario

You can lose an entire location, and access to the disks remains available at the alternate location. The use of this behavior requires clustering software at the application and server layer to fail over to a server at the alternate location and resume access to the disks.

The active-active synchronous mirroring feature, depicted in Figure 7-3, provides the capability to keep both copies of the storage in synchronization. Therefore, the loss of one location causes no disruption to the alternate location.

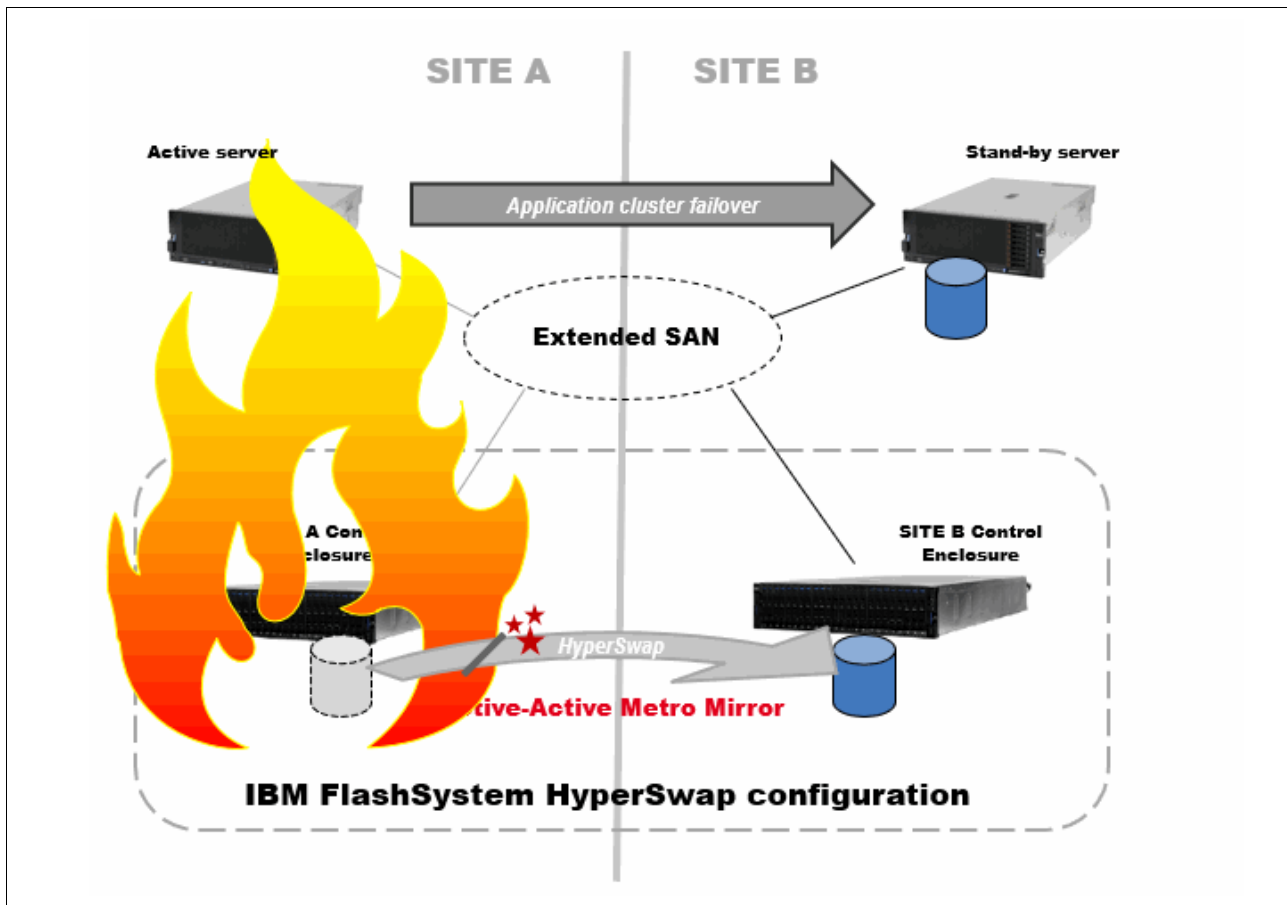


Figure 7-3 IBM FlashSystem HyperSwap in a site failure scenario

In addition to the active-active Metro Mirror feature, the HyperSwap feature also introduced the *site awareness* concept for node canisters, internal and external storage, and hosts. Finally, with the HyperSwap *DR feature*, you can manage rolling-disaster scenarios effectively.

7.2 Third site and IP quorum

In HyperSwap configurations, you can use a third, independent site to house a quorum device to act as the tie-breaker in case of 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.

To use a quorum disk as the quorum device, this third site must have Fibre Channel or iSCSI connectivity between an external storage system and the IBM Spectrum 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 based on a Java application, known as the IP quorum application, is introduced with the release V7.6.

Initially, IP quorum was used only as a tie-breaker solution. However, with the release V8.2.1, it was expanded to be able to store cluster configuration metadata, fully serving as an alternative for quorum disk devices. To use an IP quorum application as the quorum device for the third site, Fibre Channel connectivity is not used. An IP quorum application can be run on any host at the third site, as shown in the Figure 7-4.

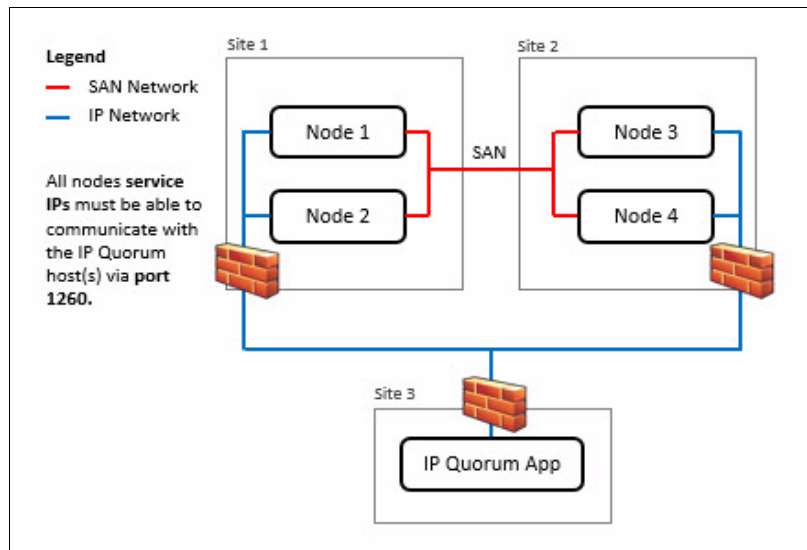


Figure 7-4 IP Quorum network layout

However, there are strict requirements on the IP network when using an IP quorum application, as follows:

- ▶ Connectivity from the servers that are running an IP quorum application to the service IP addresses of all nodes or node canisters. The network must also handle the possible security implications of exposing the service IP addresses, as this connectivity can also 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.
- ▶ 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 two megabytes per second is guaranteed 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 megabytes per second is guaranteed 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 megabytes 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 provided by the system it is protecting, as during a tie-break situation this storage is paused for I/O.

For more information about IP Quorum requirements and installation, including supported Operating Systems and Java runtime environments (JREs), see [IBM FlashSystem 9200 8.4.0 Documentation - Configuring quorum](#).

For more information related to quorum disk devices, see 3.5, “Quorum disks” on page 99.

Note: The IP Quorum configuration process has been integrated into the IBM Spectrum Virtualize GUI and can be found at **Settings** → **Systems** → **IP Quorum**.

7.2.1 Quorum modes

Quorum mode is a new configuration option that was added to the IP Quorum functionality with the release of IBM Spectrum Virtualize V8.3. By default, the IP quorum mode is set to **Standard**. In HyperSwap clusters, this mode can be changed to **Preferred** or **Winner**.

This configuration allows you to 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-break 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-break operations on other sites that are not specified as “preferred”. In other words, 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 non-preferred sites. Thus, the likelihood of reaching the quorum device first is higher. If the preferred site is damaged or is unable to reach the quorum device, the other sites have the chance to win the tie-break 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 configured as the winner will always be the one to continue processing I/O, regardless of the failure and

its condition. The nodes at the non-winner site always loses the tie-break and stops processing I/O requests until the fault is fixed.

7.3 HyperSwap Volumes

HyperSwap Volumes is one type of volume. It consists of a Master Volume and a Master Change Volume (CV) in one system site, and an Auxiliary Volume and Auxiliary Change Volume (CV) in the other system site. An active-active synchronous mirroring relationship exists between the two sites. As with a regular Metro Mirror relationship, the active-active relationship keeps the Master Volume and Auxiliary Volume synchronized.

The relationship uses the CVs as journaling volumes during any resynchronization process. The Master CV must be in the same I/O Group as the Master Volume, and it is recommended that it is in the same pool as the Master Volume. A similar practice applies to the Auxiliary CV and the Auxiliary Volume. For other considerations regarding the Change Volume, see “Global Mirror Change Volumes functional overview” on page 271.

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 guaranteed by the HyperSwap function.

Figure 7-5 shows how the HyperSwap Volume is implemented.

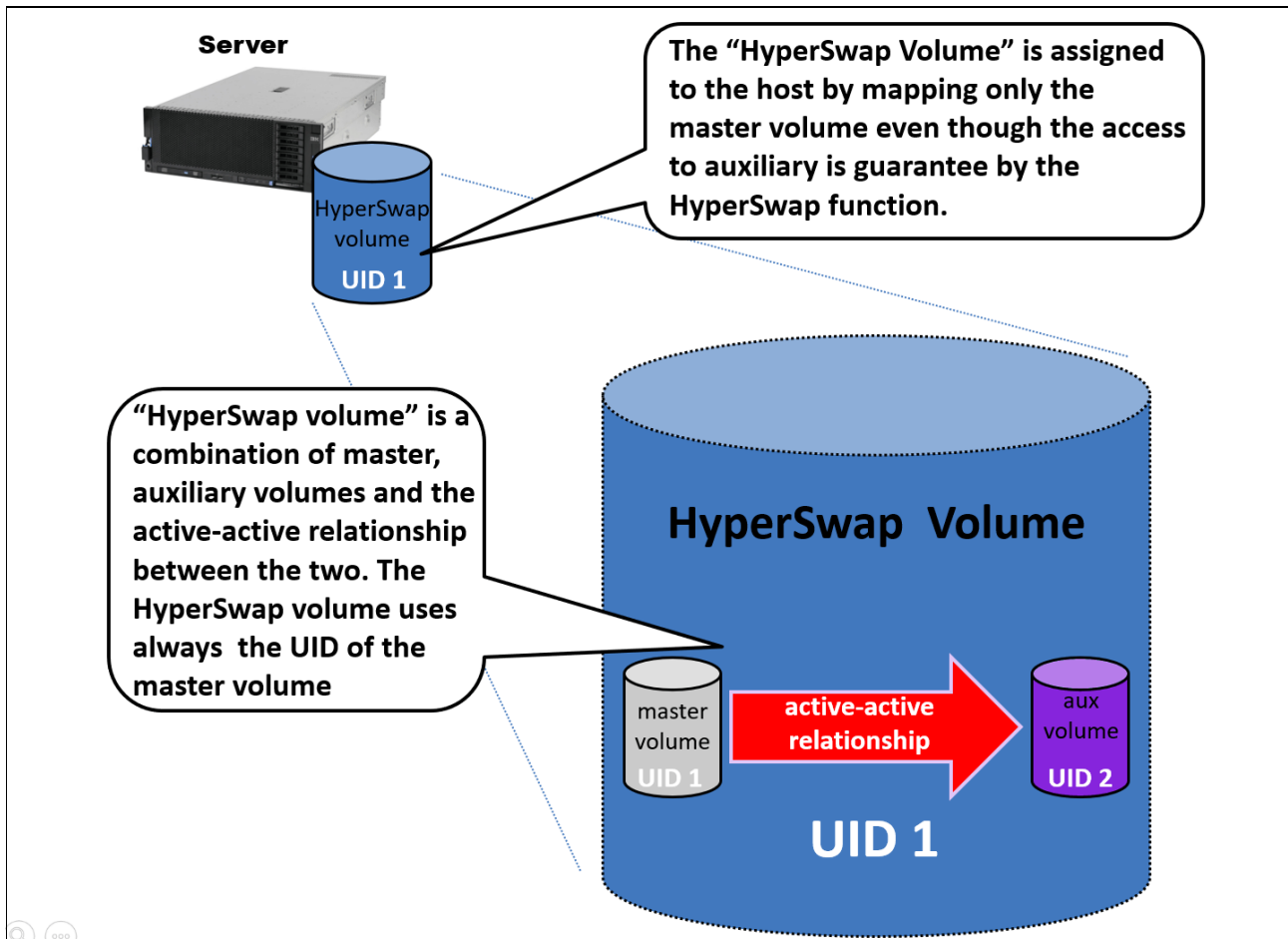


Figure 7-5 HyperSwap Volume

The active-active synchronous replication workload traverses the SAN by using the node-to-node communication. Master and Auxiliary Volumes also have a specific role of Primary or Secondary. Master or Auxiliary Volumes are Primary or Secondary based on the Metro Mirror active-active relationship direction.

Starting with the IBM Spectrum Virtualize 8.3.1 code level, reads are always done in the local copy of the volume. Write operations are always routed to the Primary copy. Therefore, hosts that access the Secondary copy for writes might experience an increased latency in the I/O operations. As a mitigation of this behavior, if sustained workload (that is, more than 75% of I/O operations for at least 20 minutes) is running over Secondary volumes, the HyperSwap function switches the direction of the active-active relationships, swapping the Secondary volume to Primary and vice versa.

Note: Frequent or continuous primary to secondary volume swap can lead to performance degradation. Avoid constantly switching the workload between sites at the host level.

7.4 Other considerations and general recommendations

Business continuity solutions implementation requires special considerations in the infrastructure and network setup. In HyperSwap topologies, the communication between the IBM Spectrum Virtualize controllers must be optimal and free of errors for best performance, as the internode messaging and cache mirroring is done across the sites. Have a dedicated private SAN for internode communication so that it is not impacted by regular SAN activities.

One other important recommendation is to review the site attribute of all the components, to make sure they are accurate. With the site awareness algorithm present in the IBM Spectrum Virtualize code, optimizations are done to reduce the cross-site workload. If this attribute is missing or not accurate, there might be an increased unnecessary cross-site traffic, which might lead to higher response time to the applications.

Throughout this book, many recommendations have been made regarding specific topics, but for a complete coverage of the implementation guidelines, see *IBM Spectrum Virtualize HyperSwap SAN Implementation and Design Best Practices*, REDP-5597 for HyperSwap configurations.

For a detailed step-by-step configuration, see [IBM FlashSystem 9200 8.4.0 Documentation - HyperSwap system configuration details](#).



Hosts

This chapter provides general guidelines and best practices around configuring host systems. The primary reference for host configuration is at [IBM FlashSystem 9200 8.4.0 Documentation](#).

For more information about host attachment, see [IBM FlashSystem 9200 8.4.0 Documentation - Host attachment](#).

For more information on hosts that are connected via Fibre Channel, see Chapter 2, “Storage area network” on page 33.

Before attaching a new host, confirm the host is supported by the Spectrum Virtualize storage. For more information, see [IBM System Storage Interoperation Center \(SSIC\)](#).

The host configuration guidelines apply equally to all Spectrum Virtualize systems. As such, the product name will often be referred to as a Spectrum Virtualize system.

This chapter includes the following sections:

- ▶ 8.1, “General configuration guidelines” on page 350
- ▶ 8.2, “Host pathing” on page 353
- ▶ 8.3, “I/O queues” on page 353
- ▶ 8.4, “Host clusters” on page 354
- ▶ 8.5, “AIX hosts” on page 357
- ▶ 8.6, “Virtual I/O server hosts” on page 358
- ▶ 8.7, “Windows hosts” on page 358
- ▶ 8.8, “Linux hosts” on page 359
- ▶ 8.9, “Oracle Solaris hosts” on page 359
- ▶ 8.10, “VMware ESXi server hosts” on page 361

8.1 General configuration guidelines

The following subsections complement the content in Chapter 2, “Storage area network” on page 33.

8.1.1 Number of paths

It is generally recommended that the total number of Fibre Channel paths per volume be limited to four paths. Adding more paths does not significantly increase redundancy and it tends to bog down the host with path management. Too many paths might increase failover time.

8.1.2 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.

Preferred practice: Keep Fibre Channel tape (including Virtual Tape Libraries) and Fibre Channel disks on separate HBAs. These devices have two different data patterns when operating in their optimum mode. The switching between applications can cause unwanted processor usage and performance slowdown for the applications.

8.1.3 Port masking

In general, Fibre Channel ports should be dedicated to certain functions. Hosts should be zoned only to ports designated for host I/O.

For more details on port masking, see Chapter 2, “Storage area network” on page 33.

8.1.4 N-port ID virtualization (NPIV)

Spectrum Virtualize uses NPIV by default. This reduces failover time and allows for features such as hot spare nodes.

For more details on configuring NPIV, see Chapter 2, “Storage area network” on page 33.

8.1.5 Host to I/O group mapping

An *I/O group* consists of two nodes or node canisters that share management of volumes within the cluster. Use a single I/O group (iogrp) for all volumes that are allocated to a particular host. This guideline has the following benefits:

- ▶ Minimizes port fan-outs within the SAN fabric
- ▶ Maximizes the potential host attachments to IBM Spectrum Virtualize because maximums are based on I/O groups
- ▶ Fewer target ports to manage within the host

8.1.6 Volume size as opposed to quantity

In general, host resources, such as memory and processing time, are used up by each storage LUN that is mapped to the host. For each extra path, more memory can be used, and a portion of more processing time is also required. 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.

If a host does not have tunable parameters, such as on the Windows operating system, the host does not benefit from large volume sizes. AIX greatly benefits from larger volumes with a smaller number of volumes and paths that are presented to it.

8.1.7 Host volume mapping

Host mapping is the process of controlling which hosts have access to specific volumes within the system. Spectrum Virtualize always present a specific volume with the same Small Computer System Interface (SCSI) ID on all host ports. When a volume is mapped, IBM Spectrum Virtualize software automatically assigns the next available SCSI ID if none is specified. In addition, a unique identifier, called the *UID*, is on each volume.

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 Spectrum Virtualize updates and even node reboots, mostly for ESX operating systems.

If you are using image mode to migrate a host to IBM Spectrum 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 VDisk (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
```

Example 8-2 shows the results of using the `lshostvdiskmap` command.

Example 8-2 Output of using the lshostvdiskmap command

```
svcinfo lsvdiskhostmap -delim : EEXCLS_HBin01
id:name:SCSI_id:host_id:host_name:wwpn:vdisk_UID
950:EEXCLS_HBin01:14:109:HDMCENTEX1N1:10000000C938CFDF:600507680191011D480000000000466
950:EEXCLS_HBin01:14:109:HDMCENTEX1N1:10000000C938D01F:600507680191011D480000000000466
950:EEXCLS_HBin01:13:110:HDMCENTEX1N2:10000000C938D65B:600507680191011D480000000000466
950:EEXCLS_HBin01:13:110:HDMCENTEX1N2:10000000C938D3D3:600507680191011D480000000000466
950:EEXCLS_HBin01:14:111:HDMCENTEX1N3:10000000C938D615:600507680191011D480000000000466
950:EEXCLS_HBin01:14:111:HDMCENTEX1N3:10000000C938D612:600507680191011D480000000000466
950:EEXCLS_HBin01:14:112:HDMCENTEX1N4:10000000C938CFBD:600507680191011D480000000000466
950:EEXCLS_HBin01:14:112:HDMCENTEX1N4:10000000C938CE29:600507680191011D480000000000466
950:EEXCLS_HBin01:14:113:HDMCENTEX1N5:10000000C92EE1D8:600507680191011D480000000000466
950:EEXCLS_HBin01:14:113:HDMCENTEX1N5:10000000C92EDFFE:600507680191011D480000000000466
```

Note: Example 8-2 shows the same volume mapped to five different hosts, but host 110 has a different SCSI ID than the other four hosts. This example is a non-recommended practice that can lead to loss of access in some situations due to SCSI ID mismatch.

8.1.8 Server adapter layout

If your host system has multiple internal I/O buses, place the two adapters that are used for IBM Spectrum 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.

8.1.9 Host status improvements

As of 8.3, Spectrum 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. There are cases where this might be normal and can cause confusion.

At the host level, there is a new **status_policy** setting. It can be set to **complete** or **redundant**. The **complete** setting uses the original host status definitions. With the **redundant** setting, a host will not be reported as *degraded* unless there are insufficient ports for redundancy.

8.1.10 Considerations for NVMe over Fibre Channel host attachments

As of 8.3, Spectrum Virtualize now supports a single host initiator port using both SCSI and NVMe connections to the storage.

Asymmetric Namespace Access has been added to the FC-NVMe protocol standard, which gives it functionality similar to Asymmetric Logical Unit Access (ALUA). As a result, FC-NVMe can now be used in stretched clusters.

IBM Spectrum Virtualize code 8.4.0 allows a maximum of 32 NVMe hosts, if no other types of hosts are attached. IBM Spectrum Virtualize code does not monitor or enforce these limits. If you are planning to use NVMe hosts with IBM FlashSystem, see [IBM Support: V8.4.0.x Configuration Limits and Restrictions for IBM FlashSystem 9200](#).

8.1.11 Considerations for iSER host attachments

On IBM FlashSystem, Internet SCSI (iSCSI) Extensions for RDMA (iSER) hosts with different operating systems can be attached to the system. iSER is a network protocol that extends the iSCSI to use Remote Direct Memory Access (RDMA).

If you are planning to use iSER hosts in your IBM SVC, see following links when planning your environment:

- ▶ [IBM FlashSystem 9200 8.4.0 Documentation - iSER Ethernet host attachment](#)
- ▶ [IBM Support: V8.4.0.x Configuration Limits and Restrictions for IBM FlashSystem 9200](#)

8.2 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 Spectrum 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 and presents a single storage device to the operating system and provide failover in the case of a lost path. If your Spectrum Virtualize system uses NPIV, the multipathing driver will not need to do the failover.

8.2.1 Path selection

I/O for a particular 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. This is called the *preferred node*. The primary purposes of using a preferred node are to have load balancing and to determine which node will destage writes to the backend storage.

When a volume is created, an I/O group and preferred node are defined and can optionally be set by the administrator. The owner node for a volume is the preferred node when both nodes are available.

Spectrum Virtualize uses ALUA, as do most multipathing drivers. This means that the multipathing driver will give preference to paths to the preferred node. Most modern storage systems use ALUA.

Note: Some competitors try to claim that ALUA means that Spectrum Virtualize is effectively an active-passive cluster. This claim is not true. Both nodes in Spectrum Virtualize can service I/O concurrently.

In the small chance that an I/O goes to the non-preferred node, that node will service the I/O without an issue.

8.3 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 unique method to control I/O queue can be one of the following:

- ▶ Host adapter-based
- ▶ Memory and thread resources-based
- ▶ Based on the number of commands that are outstanding for a device

8.3.1 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, at the disk or flash level, at the storage controller level and at the per volume and host bus adapter (HBA) level on the host. For example, each Spectrum Virtualize node has a queue depth of 10,000. A typical disk drive will operate efficiently at a queue depth of 8. Most host volume queue depth defaults will be around 32.

Guidance for limiting queue depths in large SANs that was described in previous documentation has been replaced with calculations for overall I/O group based queue depth considerations.

There isn't a set rule for setting a queue-depth value per host HBA or per volume. The requirements for your environment will be driven by the intensity of each workload. You should ensure that one application or host cannot run away and use the entire controller queue. However, if you have a specific host application that requires the lowest latency and highest throughput, then you should consider giving it a proportionally larger share than others.

- ▶ A single Spectrum Virtualize Fibre Channel port will accept a maximum concurrent queue depth of 2028. A single Spectrum Virtualize node will accept a maximum concurrent queue depth of 10,000. After this it will report queue full status.
- ▶ Host HBA queue depths should be set to the maximum - typically 1024
- ▶ Host queue depth should be controlled through the per volume value.
 - A typical random workload volume should use around 32
 - To limit the workload of a volume, use 4 or less
 - To maximize throughput and give a higher share to a volume, use 64

Remember that the total workload capability can be calculated by multiplying the number of volumes by their respective queue depths and summing. As a pointer, with very low latency storage a workload of over 1 million input output processors (IOPs) can be achieved with a concurrency on a single IO Group of 1000.

For further guidance, see:

- ▶ [IBM FlashSystem 9200 8.4.0 Documentation - Queue Depth for FC hosts](#)
- ▶ [IBM FlashSystem 9200 8.4.0 Documentation - Queue Depth for iSCSI hosts](#)
- ▶ [IBM FlashSystem 9200 8.4.0 Documentation - Queue Depth for iSER hosts](#)

8.4 Host clusters

IBM Spectrum Virtualize supports host clusters. This feature allows multiple hosts to have access to 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 world wide port names (WWPNs) that belong to the hosts 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 handle host clusters:

- ▶ **lshostcluster**
- ▶ **lshostclustermember**
- ▶ **lshostclustervolumemap**
- ▶ **addhostclustermember**
- ▶ **chostcluster**
- ▶ **mkhost** (with parameter **-hostcluster** to create the host in one existing cluster)
- ▶ **rmhostclustermember**
- ▶ **rmhostcluster**
- ▶ **rmvolumehostclustermap**

Starting with IBM Spectrum Virtualize 8.1, host clusters are added by using the GUI, which allows you to let the system assign the SCSI IDs for the volumes or you can manually assign them. For ease of management purposes, it is suggested that you 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 above 100 for the cluster host volumes. When you choose the option **System Assign**, 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, and on the right part of the screen it shows the SCSI IDs that are already used by the selected host/host cluster, as shown in Figure 8-1.

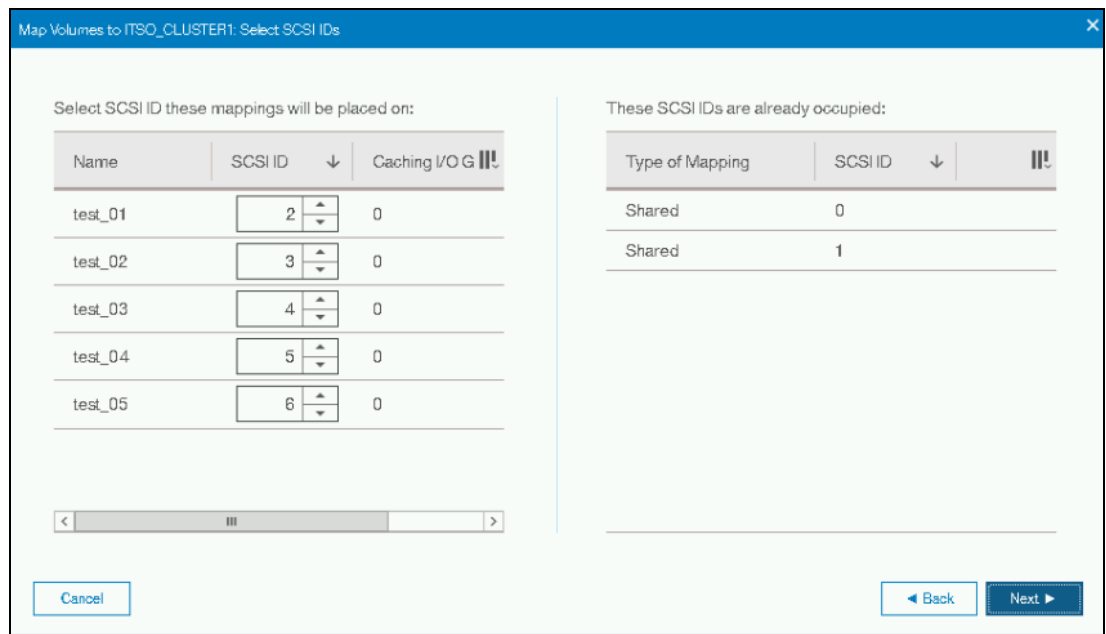


Figure 8-1 SCSI ID assignment on volume mappings

Note: Although extra care is always recommended when dealing with hosts, IBM Spectrum Virtualize does not allow you to join a host into a host cluster if it already has 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 ID's for their  
private mappings.
```

```
IBM_2145:ITS0-SVCLab:superuser>
```

8.4.1 Persistent reservations

To prevent hosts from sharing storage inadvertently, establish a storage reservation mechanism. The mechanisms for restricting access to IBM Spectrum 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 Spectrum Virtualize between hosts. To share storage between hosts, maintain control over accessing the volumes. Some clustering software uses software locking methods.

You can choose other methods of control by the clustering software or by 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`: Reservations are not 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`. The driver then opens the device. For persistent

reserve, each host that is attached to the shared disk must use a unique registration key value.

8.4.2 Clearing reserves

It is possible to accidentally leave a reserve on the IBM Spectrum Virtualize volume or on the IBM Spectrum Virtualize MDisk during migration into IBM Spectrum Virtualize, or when disks are reused for another purpose. Several tools are available from the hosts to clear these reserves. You can also clear the IBM Spectrum Virtualize volume reserves by removing all the IBM Spectrum Virtualize MDisk reserves.

There are instances in which a host image mode migration appears to succeed, but problems occur when the volume is opened for read or write I/O. The problems can result from not removing the reserve on the MDisk before image mode migration is used in IBM Spectrum Virtualize. You cannot clear a leftover reserve on an IBM Spectrum Virtualize MDisk from IBM Spectrum 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.

8.5 AIX hosts

This section describes various topics that are specific to AIX hosts.

8.5.1 Multipathing support

SDD PCM is no longer supported. Use the default AIX PCM.

For more details, see [The Recommended Multi-path Driver to use on IBM AIX and VIOS When Attached to SVC and Storwize storage](#).

The Recommended Multi-path Driver to use on IBM AIX and VIOS When Attached to SVC and Storwize storage

8.5.2 Configuration recommendations for AIX

The following device settings can be changed with the **chdev** AIX command.

▶ **reserve_policy=no_reserve**

The default reserve policy is **single_path** (SCSI-2 reserve). Unless there is a specific need for reservations, use **no_reserve**.

▶ **algorithm=shortest_queue**

If coming from SDD PCM, AIX defaults this to **fail_over**. You cannot set algorithm to **shortest_queue** unless reservation policy is **no_reserve**.

▶ **queue_depth=32**

The default queue depth is **20**. IBM recommends **32**.

▶ **rw_timeout=30**

IBM recommends **30**.

Note: **60** was the default for the SDD PCM and **30** is the default for AIX PCM.

8.6 Virtual I/O server hosts

8.6.1 Multipathing support

SDD PCM is no longer supported. Use the default AIX PCM.

For more details, see [The Recommended Multi-path Driver to use on IBM AIX and VIOS When Attached to SVC and Storwize storage](#).

8.6.2 Physical and logical volumes

Virtual SCSI is based on a client/server relationship. The Virtual I/O Server (VIOS) owns the physical resources and acts as the server or target device. Physical adapters with attached disks (in this case, volumes on IBM Spectrum Virtualize) on the VIOS partition can be shared by one or more partitions. These partitions contain a virtual SCSI client adapter that detects these virtual devices as standard SCSI-compliant devices and LUNs.

You can create the following types of volumes on a VIOS:

- ▶ Physical volume (PV) virtual SCSI (VSCSI) hdisks
- ▶ Logical volume (LV) virtual SCSI (VSCSI) hdisks

PV VSCSI hdisks are entire LUNs from the VIOS perspective. If you are concerned about failure of a VIOS and have configured redundant VIOSs for that reason, you must use PV VSCSI hdisks. Therefore, PV VSCSI hdisks are entire LUNs that are volumes from the virtual I/O client perspective. An LV VSCSI hdisk cannot be served up from multiple VIOSs.

LV VSCSI hdisks are in LVM volume groups on the VIOS, and cannot span PVs in that volume group or be striped LVs. Because of these restrictions, use PV VSCSI hdisks.

8.6.3 Methods to identify a disk for use as a virtual SCSI disk

The VIOS uses the following methods to uniquely identify a disk for use as a virtual SCSI disk:

- ▶ Unique device identifier (UDID)
- ▶ IEEE volume identifier
- ▶ Physical volume identifier (PVID)

Each of these methods can result in different data formats on the disk. The preferred disk identification method for volumes is the use of UDIDs.

8.7 Windows hosts

This section describes various topics that are specific to Microsoft Windows hosts.

8.7.1 Multipathing 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.

8.7.2 Windows configuration

For more information about configuring Windows hosts, see: [IBM FlashSystem 9200 8.4.0 Documentation - Windows hosts](#).

8.8 Linux hosts

IBM Spectrum Virtualize supports Linux hosts using native Device Mapper-Multipathing (DM-MP) multipathing. Veritas Dynamic Multi-pathing (DMP) is also available for certain kernels.

For more information about configuring Linux hosts, see [IBM FlashSystem 9200 8.4.0 Documentation - Linux hosts](#).

Note: Occasionally we see storage admins modify parameters in the `multipath.conf` file to address some perceived shortcoming in the DM-MP configuration. This can create unintended and unexpected behaviors. The recommendations provided in IBM Documentation are optimal for a vast majority of configurations.

8.9 Oracle Solaris hosts

Two options are available for multipathing support on Solaris hosts:

- ▶ Symantec Veritas Volume Manager
- ▶ Solaris MPxIO

The option that you choose depends on your file system requirements and the operating system levels in the latest interoperability matrix. For more information, see [IBM System Storage Interoperation Center \(SSIC\)](#).

IBM SDD is no longer supported because its features are now available natively in the multipathing driver for Solaris MPxIO.

For more information about host attachment for Solaris hosts, see [IBM FlashSystem 9200 Documentation - Solaris hosts](#).

8.9.1 Solaris MPxIO

SAN boot and clustering support is available for V5.9, V5.10, and 5.11, depending on the multipathing driver and HBA choices. Support for load balancing of the MPxIO software is included in IBM SAN Volume Controller (SVC) running IBM Spectrum Virtualize 8.2. If you want to run MPxIO on your Sun SPARC host, configure your IBM Spectrum Virtualize host object with the type attribute set to `tpgs`, as shown in the following example:

```
svctask mkhost -name new_name_arg -hbawwn wwpn_list -type tpgs
```

In this command, `-type` specifies the type of host. Valid entries are `hpux`, `tpgs`, `generic`, `openvms`, `adminlun`, and `hide_secondary`. The `tpgs` option enables an extra target port unit. The default is `generic`.

8.9.2 Symantec Veritas Volume Manager

When you are managing IBM Spectrum Virtualize storage in Symantec volume manager products, you must install an ASL on the host so that the volume manager is aware of the storage subsystem properties (active/active or active/passive). If the appropriate Array Support Library (ASL) is not installed, the volume manager did not claim the LUNs. Usage of the ASL is required to enable the special failover or failback multipathing that IBM Spectrum Virtualize requires for error recovery.

Use the commands that are shown in Example 8-3 to determine the basic configuration of a Symantec Veritas server.

Example 8-3 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 8-4 and Example 8-5 determine whether the IBM Spectrum Virtualize is properly connected. They show at a glance which ASL is used (native DMP ASL or SDD ASL). Example 8-4 shows what you see when Symantec Volume Manager correctly accesses IBM Spectrum Virtualize by using the SDD pass-through mode ASL.

Example 8-4 Symantec Volume Manager using 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 8-5 shows what you see when IBM Spectrum Virtualize is configured by using native DMP ASL.

Example 8-5 IBM Spectrum Virtualize that is configured by using native ASL

```
# vxdmpadm listenclosure all
ENCLR_NAME ENCLR_TYPE ENCLR_SNO STATUS
=====
OTHER_DISKS OTHER_DSKSI OTHER_DISKS CONNECTED
SAN_VCO SAN_VC 0200628002faXX00 CONNECTED
```

8.9.3 DMP multipathing

For the latest ASL levels to use native DMP, see the array-specific module table at [Veritas Services and Operations Readiness Tools \(SORT\)](#).

To check the installed Symantec Veritas version, enter the following command:

```
showrev -p |grep vxvm
```

To check which IBM ASLs are configured into the Volume Manager, enter 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 **vxdt1 enable** command. To list the ASLs that are active, enter the following command:

```
vxddladm listsupport
```

8.10 VMware ESXi server hosts

To determine the various VMware vSphere/ESXi levels that are supported, see [IBM System Storage Interoperation Center \(SSIC\)](#).

8.10.1 Configuring VMware

For information about specific configuration best practices for VMware, see [IBM FlashSystem 9200 8.4.0 Documentation - Configuring the ESXi operating system](#).

VMware has a built-in multipathing driver which supports Spectrum Virtualize ALUA preferred path algorithms.

Consider the following points:

- ▶ The storage array type should be ALUA (VMW_SATP_ALUA).
- ▶ Path selection policy should be RoundRobin.
- ▶ The Round Robin Input/Output Operations Per Second (IOPS) should be changed from 1000 to 1, to evenly distribute I/Os across as many ports on the system as possible. For more information about how to change this setting, see [VMware Adjusting Round Robin IOPS limit from default 1000 to 1](#).

8.10.2 Multipathing configuration maximums

The VMware multipathing software supports the following maximum configuration:

- ▶ A total of 256 SCSI devices
- ▶ Up to 32 paths to each volume
- ▶ Up to 4096 paths per server

Tip: Each path to a volume equates to a single SCSI device.

For a complete list of maximums, see [VMware Configuration Maximums](#).

For IBM i-related considerations, see Appendix A, “IBM i considerations” on page 525.



Monitoring

Monitoring in a storage environment is crucial and it is part of what usually is called *storage governance*.

With a robust and reliable storage monitoring system, you can save significant money and minimize pain in your operation, by monitoring and predicting utilization bottlenecks in your storage environment.

This chapter provides suggestions and the basic concepts of how to implement a storage monitoring system for IBM FlashSystem, using specific functions or external IBM Tools.

This chapter includes the following sections:

- ▶ 9.1, “Generic monitoring” on page 364
- ▶ 9.2, “Performance monitoring” on page 367
- ▶ 9.3, “Capacity metrics for block storage systems” on page 387
- ▶ 9.4, “Creating alerts for IBM Spectrum Control and IBM Storage Insights” on page 400
- ▶ 9.5, “Error condition example with IBM Spectrum Control: FC port” on page 409
- ▶ 9.6, “Important metrics” on page 411
- ▶ 9.7, “Performance support package” on page 412
- ▶ 9.8, “Metro and Global Mirror monitoring with IBM Copy Services Manager and scripts” on page 414
- ▶ 9.9, “Monitoring Tier1 SSD” on page 415

9.1 Generic monitoring

With IBM FlashSystem, you can implement generic monitoring using IBM FlashSystem-specific functions that are integrated with the product without adding external tools or cost.

9.1.1 Monitoring with the GUI

The management GUI is the primary tool that is used to service your system. 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.

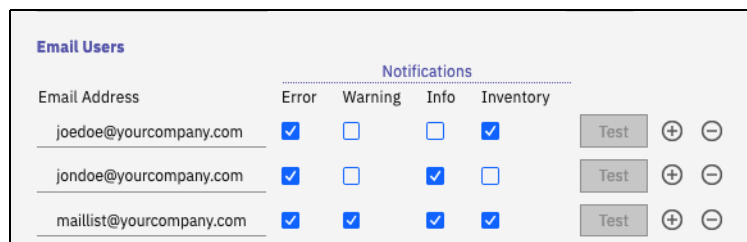
Use the views that are available in the management GUI to verify the status of the system, the hardware devices, the physical storage, and the available volumes. The **Monitoring** → **Events** window provides access to all problems that exist on the system. Use the **Recommended Actions** filter to display the most important events that need to be resolved.

If there is a service error code for the alert, you can run a fix procedure that assists you in resolving the problem. The fix procedures analyze the system and provide more information about the problem. They suggest actions to take and step you through the actions that automatically manage the system when necessary. After the fix procedure finishes and the problem is resolved, the alert will be closed.

If an error is reported, always use the fix procedures within the management GUI to resolve the problem, even if the error is configuration or hardware problem. The fix procedures analyze the system to ensure that the required changes do not cause volumes to be inaccessible to the hosts. The fix procedures automatically perform configuration changes that are required to return the system to its optimum state.

Email notification

It is possible to add multiple e-mail addresses, or you can use your e-mail distribution list, to receive notifications from the storage. For each e-mail box that you added, you can set notification options with different sets of information, as shown in Figure 8-1.



Email Address	Notifications				Test	+	-
	Error	Warning	Info	Inventory			
joedoe@yourcompany.com	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Test	+	-
jondoe@yourcompany.com	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Test	+	-
maillist@yourcompany.com	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Test	+	-

Figure 9-1 Email notification options

E-mail notification is one of the most common and important tools for monitoring that you can use and set up. From the notification events you can validate whether your system is running under normal status or needs attention.

Call Home

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. When configured,

this function alerts IBM service personnel about hardware failures and potentially serious configuration or environment issues.

You can configure the Cloud Call Home option that will alert IBM support for any problem reported by the system. These steps can be found at [Implementing IBM FlashSystem with IBM Spectrum Virtualize V8.4](#).

SNMP 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 FlashSystem system.

The Management Information Base (MIB) file describes the format of the SNMP messages that are sent by IBM FlashSystem. Use the MIB file to configure a network management program to receive SNMP event notifications that are sent from an IBM FlashSystem system. This MIB file is suitable for use with SNMP messages from all versions of IBM FlashSystem.

The latest IBM FlashSystem MIB file can be downloaded at [Management Information Base \(MIB\) file for SNMP](#).

Syslog notification

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 IPv4 or IPv6. The system can send Syslog messages that notify personnel about an event. You can configure a syslog server to receive log messages from various systems and store them in a central repository.

Figure 9-2 on page 366 demonstrate the new syslog grid layout from the IBM FlashSystem GUI. You can configure multiple syslog servers and monitor the communication between IBM FlashSystem to the syslog server from the syslog panel.

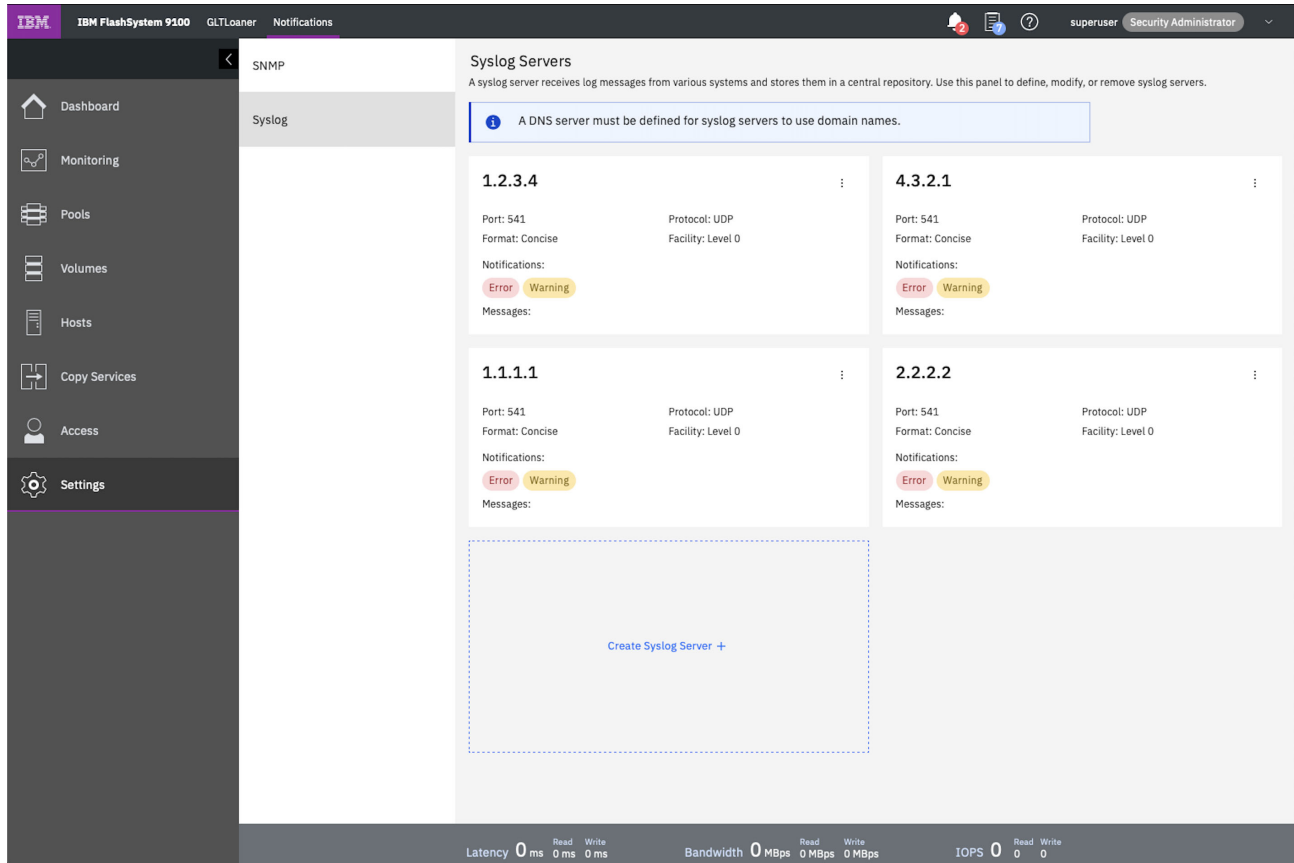


Figure 9-2 Syslog Configuration

Note: Starting with version 8.4, it is possible to use FQDN for services such as Syslog, LDAP, and NTP.

9.1.2 Monitoring using quotas and alert

In an IBM FlashSystem system, the space usage of storage pools, thin provisioned volumes or Compressed Volumes can be monitored by setting some specific quota alerts.

Storage Pool

During storage pool configuration, you can set a warning such that when the pool capacity reaches this quota setting, an alert is issued. This setting generates a warning when the used capacity in the storage pool first exceeds the specified threshold. You can specify a `disk_size` integer, which defaults to megabytes (MB) unless the `-unit` parameter is specified. Or you can specify a `disk_size%`, which is a percentage of the storage pool size. To disable warnings, specify `0` or `0%`. The default value is `0`.

Volumes

Thin-provisioned and compressed volumes near their size limits are monitored at specified thresholds to preserve data integrity. If a volume can be shrunk to below the recommended new limit, you are advised to do so. If volume capacity cannot be reduced to meet the recommended limit, you are advised to create a non-compressed mirror of the data (if one does not exist) and delete the primary copy.

9.2 Performance monitoring

The ability to collect historical performance metrics is essential to properly monitor and manage storage subsystems and IBM FlashSystem. During troubleshooting and performance tuning, the historical data can be used as a parameter for changes and fixes.

The next sections show which performance analysis tools are integrated with IBM FlashSystem and which IBM external tools are available to collect performance statistics for historical retention.

Remember that performance statistics are useful to not only debug or prevent some potential bottlenecks, but also to make capacity planning for future growth easier, as shown in Figure 9-3 on page 368.

9.2.1 Performance monitoring with the GUI

In IBM FlashSystem, 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 bandwidth of all the volumes, interfaces, and MDisks that are being used on your system. You can also monitor the overall CPUs utilization for the system. These statistics summarize the overall performance health of the system and can be used to monitor trends in bandwidth and CPU utilization.

You can monitor changes to stable values or differences between related statistics, such as the latency between volumes and MDisks. These differences can then be further evaluated by performance diagnostic tools.

Additionally, with system-level statistics, you can quickly view 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 use and to evaluate whether a specific data point might represent performance impacts. For example, you can monitor the interfaces, such as for Fibre Channel or SAS interfaces, to determine whether the host data-transfer rate is different from the expected rate.

You can also select canister-level statistics, which can help you determine the performance impact of a specific canister. As with system statistics, canister statistics help you to evaluate whether the canister is operating within normal 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 Interfaces graph displays data points for Fibre Channel (FC), Internet Small Computer Systems Interface (iSCSI), serial-attached SCSI (SAS), and IP Remote Copy interfaces. You can use this information to help determine connectivity issues that might affect performance.

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, or connectivity issues.

Each graph represents five minutes of collected statistics, updated every five seconds, and provides a means of assessing the overall performance of your system, as shown in Figure 9-3.

Figure 9-3 denotes an example of monitoring the IBM FlashSystem subsystem GUI.

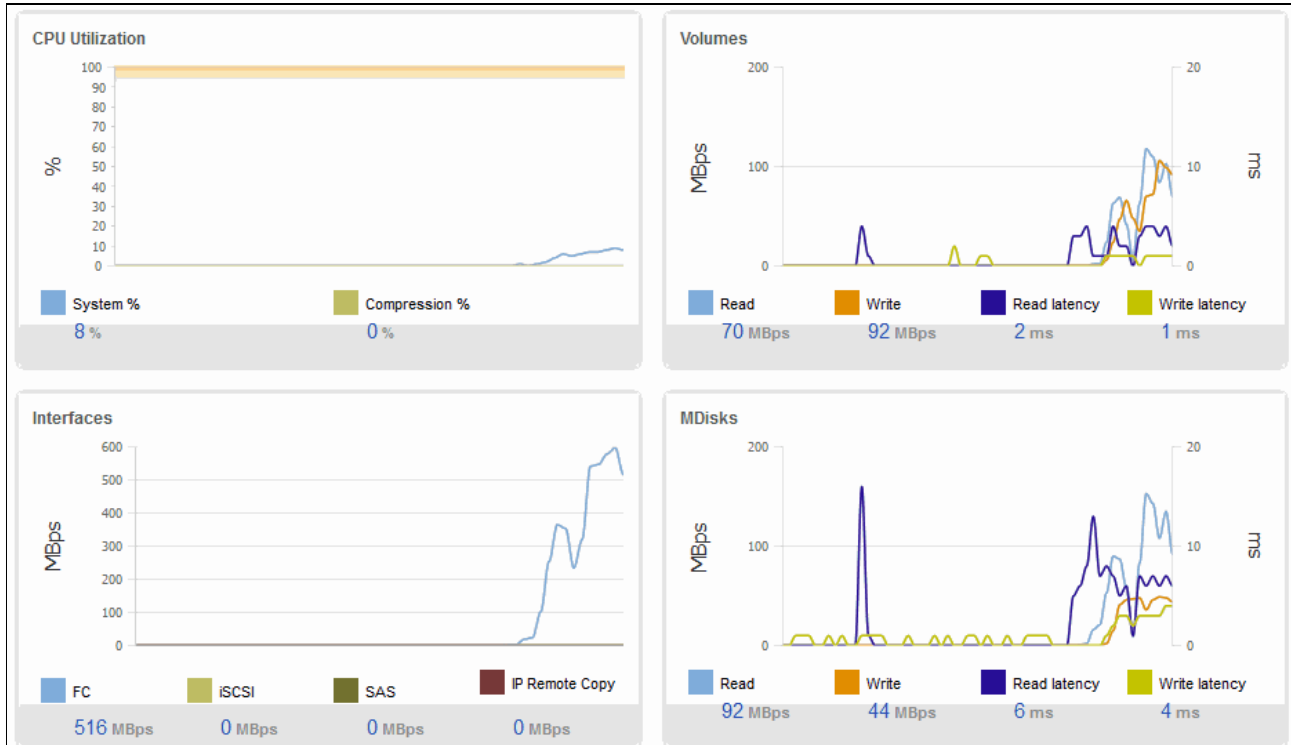


Figure 9-3 Monitoring GUI example

You can then choose the metrics that you want to be displayed, as shown in Figure 9-4.

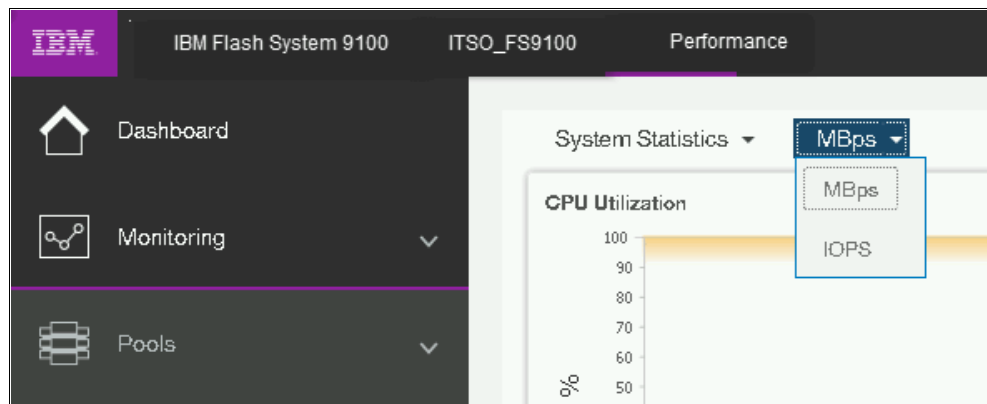


Figure 9-4 Selecting metrics

You can also obtain a quick overview by using the GUI option **System** → **Dashboard**, as shown in Figure 9-5.

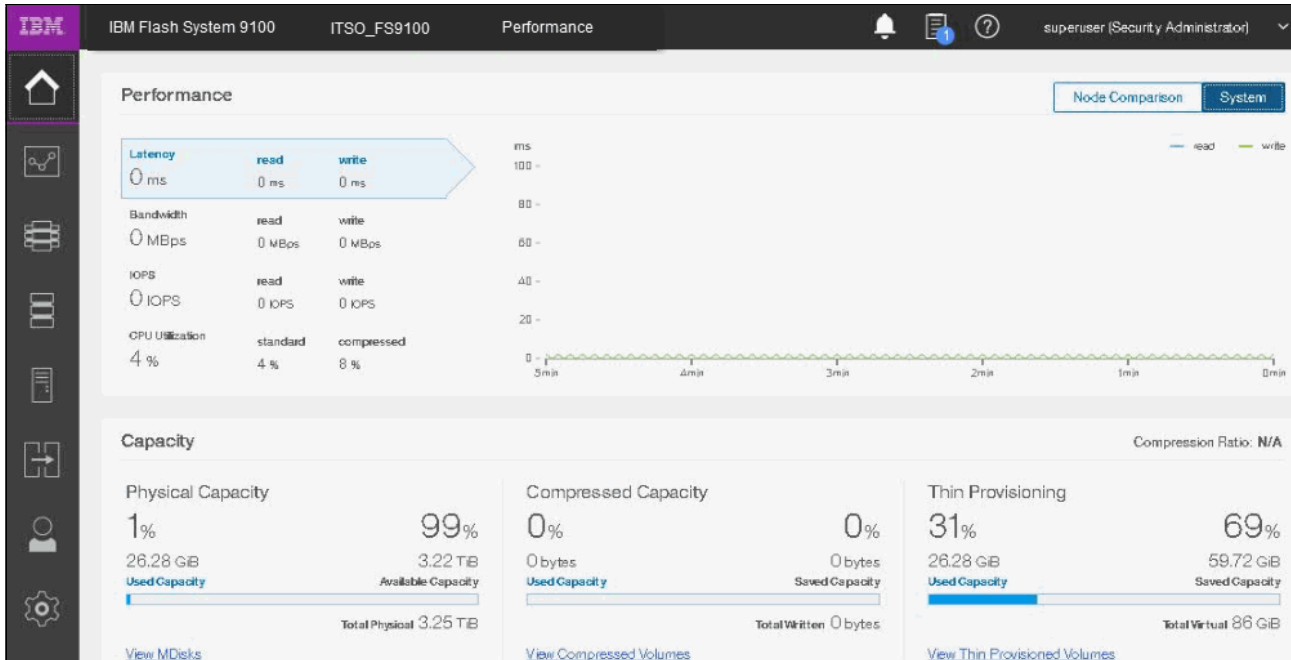


Figure 9-5 System -> Dashboard

9.2.2 Performance monitoring with IBM Spectrum Control

IBM Spectrum Control is an on-premises storage management, monitoring, and reporting solution. It leverages 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, but IBM Spectrum Control also monitors hypervisors, fabrics, and switches to provide you with unique analytics and insights into the topology of your storage network. It also provides more granular collection of performance data, with one-minute intervals rather than the five-minute intervals in IBM Storage Insights or IBM Storage Insights Pro. For more information on IBM Storage Insights, see 9.2.3, "Performance monitoring with IBM Storage Insights" on page 373.

Because IBM Spectrum Control is an on-premises tool, it does not send the metadata about monitored devices offsite, which is ideal for dark shops and sites that don't want to open ports to the cloud.

For more information on the capabilities of IBM Spectrum Control, see [IBM Spectrum Control documentation](#).

For pricing and other purchasing information, see [IBM Spectrum Control](#).

Note: If you currently have IBM Spectrum Control or manage IBM block storage systems, you already have access to IBM Storage Insights (free version). To get started, see [Getting Started with IBM Storage Insights](#).

IBM Spectrum Control offers several reports that you can use to monitor IBM FlashSystem 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 FlashSystem. The next sections provide basic suggestions about the metrics that need to be monitored and analyzed to debug potential bottleneck problems. In addition, which alerts need to be set to be notified when some specific metrics exceed limits that are considered important for this specific environment.

For more information about the installation, configuration, and administration of IBM Spectrum Control (including how to add a storage system), see:

- ▶ [5.4.0 Limitations and known issues for IBM Spectrum Control](#)
- ▶ [IBM Spectrum Control 5.4.2 Documentation - Installing](#)

Note: IBM Spectrum Control 5.3.0 or higher is recommended for monitoring IBM FlashSystem.

IBM Spectrum Control dashboard

The performance dashboard provides Key Performance Indicators (in prior releases, Best Practice Performance Guidelines) for the critical monitoring metrics. These guidelines do not represent the maximum operating limits of the related components. They represent suggested limits that are selected with an emphasis on maintaining a stable and predictable performance profile.

The dashboard displays 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 prior. 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.

These dashboards provide two critical functions:

- ▶ Provides an “at-a-glance” view of all the critical IBM Flash System monitoring metrics
- ▶ Provides a historical comparison of the metric profile of the current day versus the previous day that enables rapid detection of anomalous workloads and behaviors.

Figure 9-6 shows how to change the day of reference.

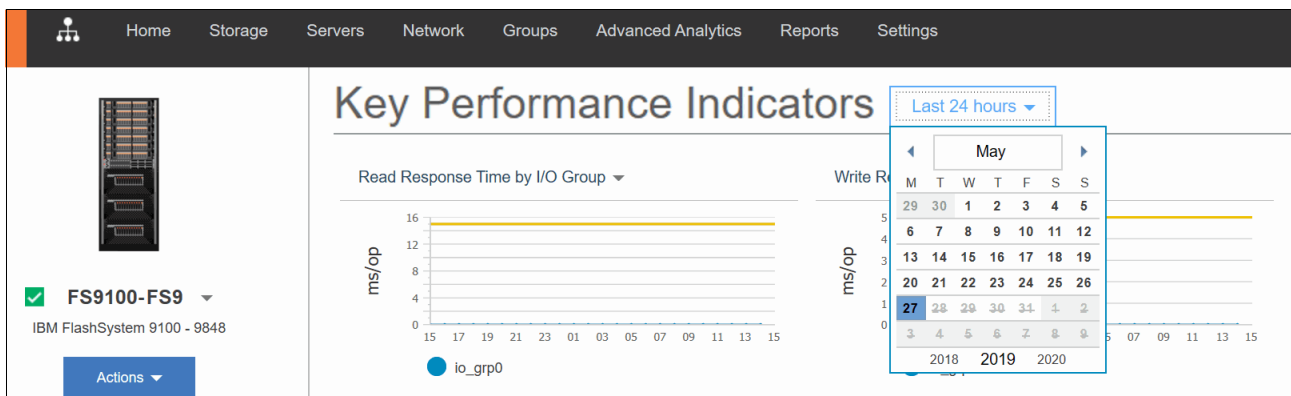


Figure 9-6 Change day of reference

Figure 9-7 shows a metric that is exceeding the best practice limit (orange line).

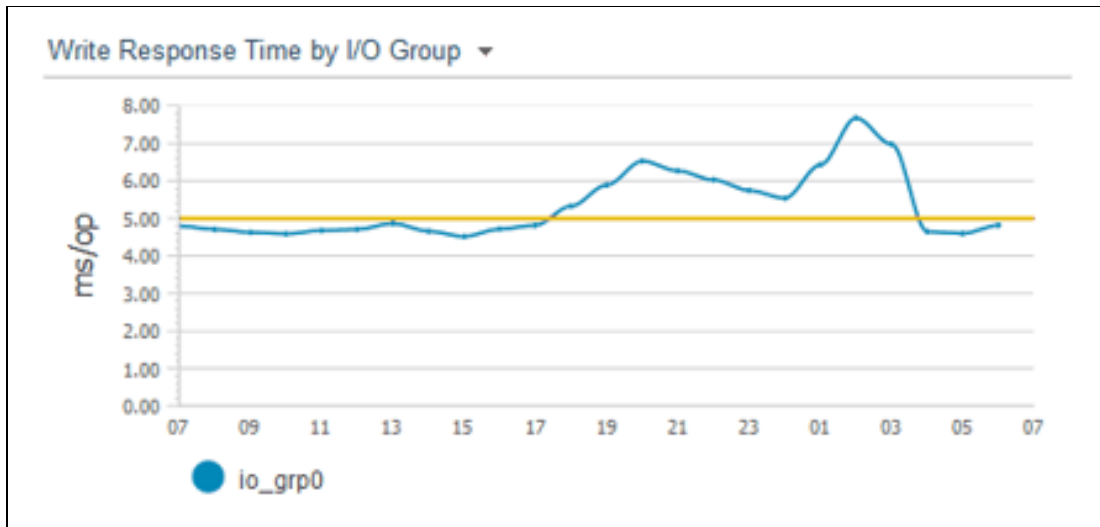


Figure 9-7 Metric exceeding best practice

Figure 9-8 shows the same chart as in Figure 9-7 with `io_grp0` selected, which overlays the previous day and 7 days prior.

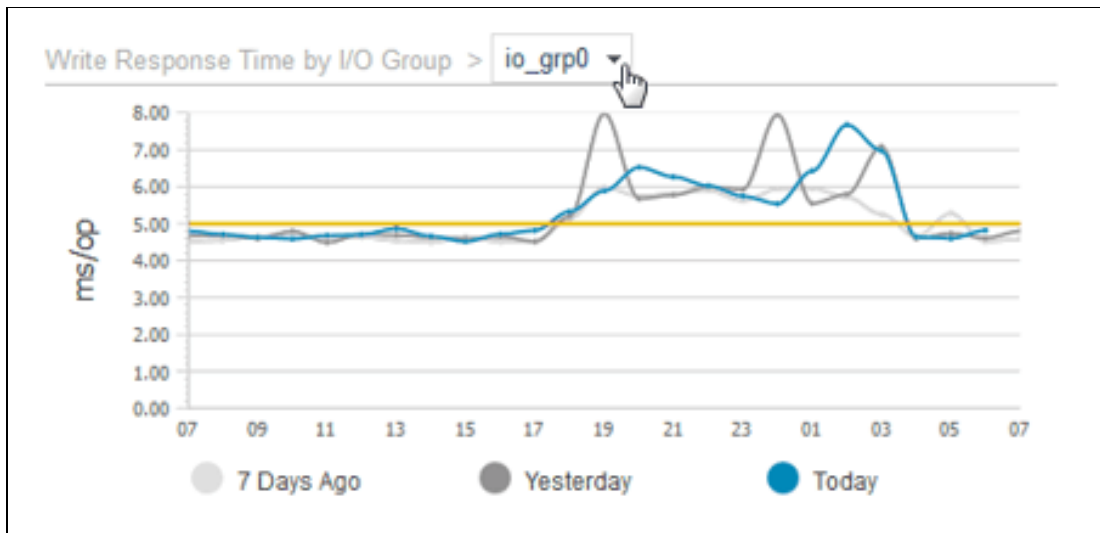


Figure 9-8 Changed chart due to `io_grp` selection

From this information, you can conclude that this exception occurs every day at the same time and is not a new phenomenon.

The yellow line is displayed if a component is constantly breaching the limit, which means that this component might be overly utilized. Therefore, an occasional peak doesn't matter. This is more to help you understand how the hardware is working and is not a service indicator. (Response times > 10 milliseconds (ms) are not acceptable.)

Note: The Best Practices Guidelines panel has recently been renamed *Key Performance Indicators*.

Key Performance Indicators (Best Practice Performance Guidelines)

You can view the key metrics that are outside of a standard range for storage systems that run IBM FlashSystem by using the performance guidelines. The guidelines were established by a historical analysis of storage environments.

Most of the performance charts show an orange line that indicates the best practice value for the metric. These guidelines are established as 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 canisters.

You can use the lines to compare how close to potentially becoming overloaded your resources are. If your storage system is responding poorly and the charts indicate overloaded resources, you might have to better balance the workload. You can balance the workload between the canisters of the cluster, potentially adding more canisters to the cluster, or move some workload to other storage systems.

Figure 9-9 shows the Key Performance Indicators in the Dashboard.

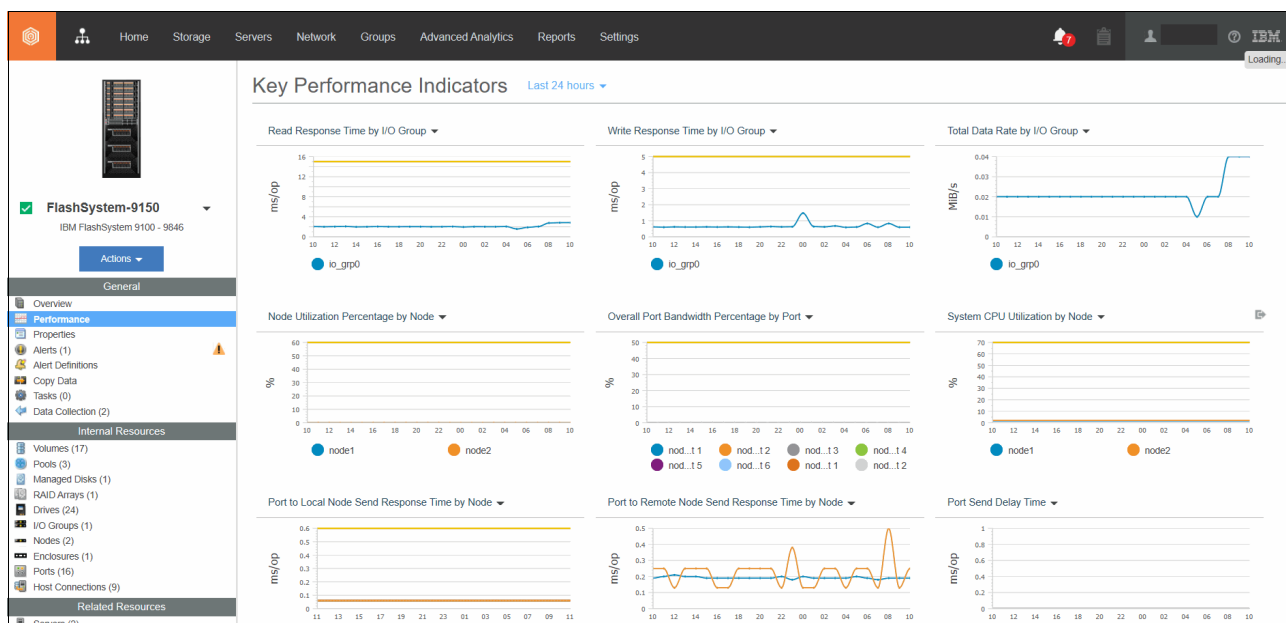


Figure 9-9 Dashboard- Key Performance Indicators

The charts show the hourly performance data measured for each resource on the selected day. Use the following charts to compare the workloads on your storage system with the best practice guidelines:

- ▶ **Node Utilization Percentage by Node:** Compare the guideline value for this metric, for example, 60% utilization, with the measured value from your system. 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. For clusters without FC ports this chart is empty (or when no host I/O is going on).
- ▶ **Overall Port Bandwidth Percentage by Port:** Compare the guideline value for this metric, for example, 50%, with the measured value from your system. Because a cluster can have many ports, the chart shows only the eight ports with the highest average bandwidth over the selected day.

- ▶ **Port-to-Local Node Send Response Time by Node:** Compare the guideline value for this metric, for example, 0.6 ms/op, with the measured value from your system. This is a very important metric for a good performing cluster.
- ▶ **Port-to-Remote Node Send Response Time by Node:** Because latencies for copy-services operations can vary widely, a guideline is not established for this metric. Use this chart to identify any discrepancies between the data rates of different nodes.
- ▶ **Read Response Time by I/O Group:** Compare the guideline value for this metric, for example, 15 ms/op, with the measured value from your system. It means, when you see this constantly being exceeded, something might be wrong with the hardware.
- ▶ **System CPU Utilization by Node:** Compare the guideline value for this metric, for example, 70% utilization, with the measured value from your system.
- ▶ **Total Data Rate by I/O Group:** Because data rates can vary widely, a guideline is not established for this metric. Use this chart to identify any significant discrepancies between the data rates of different I/O groups because these discrepancies indicate that the workload is not balanced.
- ▶ **Write Response Time by I/O Group:** 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:** Compare the guideline value for this metric, for example, 20%, with the measured value from your system. Keep in mind, that this will only work with 8 Gbps adapters, for 16 Gbps or 32 Gbps adapters using the port delay metrics (not in this overview).

Note: The guidelines are not thresholds and they are not related to the alerting feature in IBM Spectrum Control. To create performance alerts that use the guidelines as thresholds, go to a resource detail window in the web-based GUI, click **Alerts** in the General section and then click **Definitions**.

9.2.3 Performance monitoring with IBM Storage Insights

IBM Storage Insights (ISI) is an off-premises, IBM Cloud service that provides cognitive support capabilities, monitoring, and reporting for storage systems. Because it is an IBM Cloud service, getting started is simple and upgrades are handled automatically.

By leveraging 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 wraparound 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 SaaS (Software as a Service) 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. It 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, it enables you to deploy quickly and save storage administration time while optimizing your storage. It also helps automate aspects of the support process to enable faster resolution of issues. ISI optimizes storage infrastructure using cloud-based storage management and support platform with predictive analytics.

It allows you to optimize performance and to tier your data and storage systems for the right combination of speed, capacity and economy. IBM Storage Insights provides comprehensive storage management, helps to keep costs low, and can prevent downtime and loss of data or revenue. IBM Storage Insights Key features are:

- ▶ Rapid results when you need them
- ▶ Single pane view across your storage environment
- ▶ Performance analyses at your fingertips
- ▶ Valuable insight from predictive analytics
- ▶ Two editions that meet your needs
- ▶ Simplified, comprehensive, and proactive product support

Figure 9-10 shows an IBM Storage Insight® example screen.

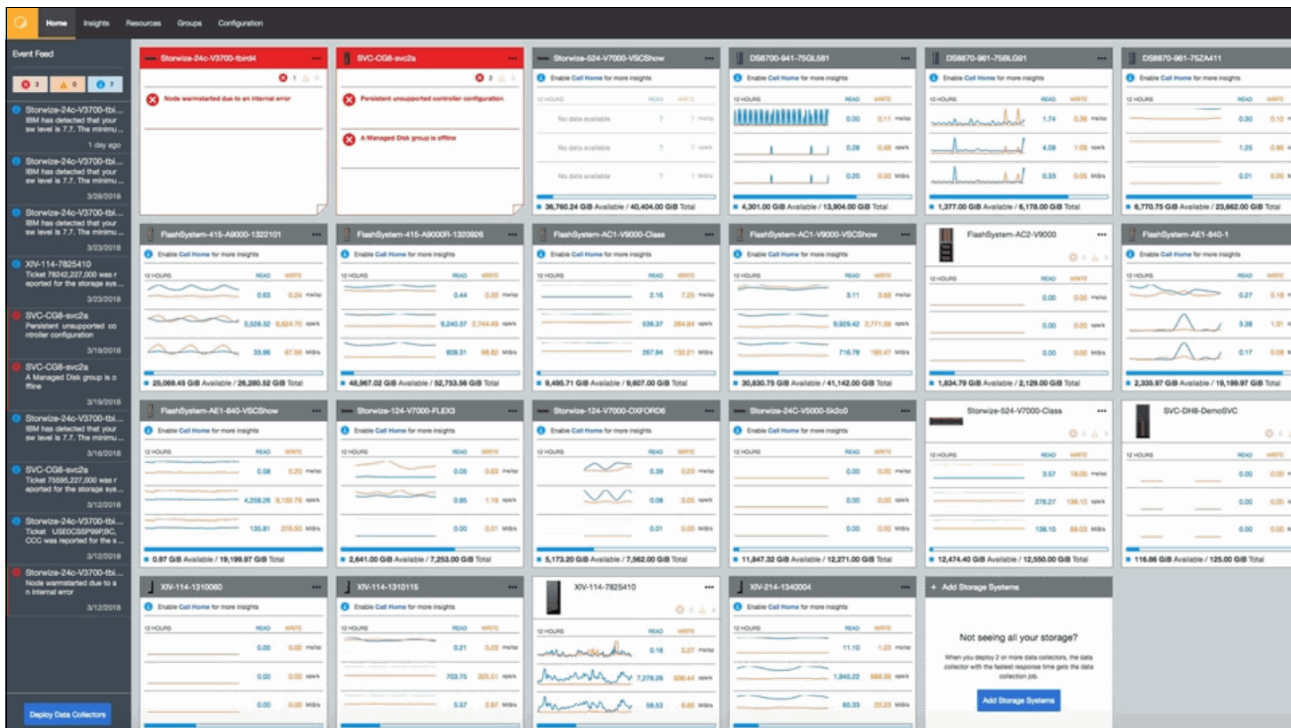


Figure 9-10 IBM Storage Insight

Understanding the security and data collection features of IBM Storage Insights Pro and IBM Storage Insights can help address the concerns of administrators and IT professionals who deploy the products in their environments and want to learn more about security and data collection. For more information, see [IBM Storage Insights Documentation - Security](#).

Note: IBM strongly recommends the use of IBM Storage Insights or IBM Spectrum Control for a better user experience. IBM Storage Insights requires the use of data collectors. The method of data collection has recently changed to improve security and ease of management. It is no longer required that you have a user with admin privileges for data collectors. A simple monitor user can get status information from the management node.

Licensing and editions of IBM Storage Insights

Several editions of IBM Storage insights enable you to select the capabilities that serve your needs best. Licensing is implemented through different subscription levels.

- ▶ The **free** version is called **IBM Storage Insights** and provides a unified view of a storage environment with a diagnostic events feed, an integrated support experience, and key capacity and 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. IBM Storage Insights provides an environment overview, integration in support processes and will you show IBM analysis results.
- ▶ 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. 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 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.

Table 9-1 on page 376 shows the different features of both versions.

Table 9-1 Features in IBM Storage Insights and IBM Storage Insights Pro

Resource Management	Functions	IBM Storage Insights (free)	IBM Storage Insights Pro (subscription)
Monitoring	Inventory management	IBM block storage	IBM and non-IBM block storage, file storage, and object storage
	Logical configuration	Basic	Advanced
	Health	Call Home events	Call Home events
	Performance	Basic (3 metrics: I/O rate, data rate, and response times aggregated for storage systems)	Advanced (100+ metrics for storage systems and their components)
	Capacity	Basic (4 metrics: allocated space, available space, total space, and compression savings aggregated for storage systems)	Advanced (25+ metrics for storage systems and their components)
	Drill down performance workflows to enable deep troubleshooting		✓
	Explore virtualization relationships		✓
	Explore replication relationships		✓
	Retention of configuration and capacity data	Only the last 24 hours is shown	2 years
	Retention of performance data	Only the last 24 hours is shown	1 year
Reporting		✓	
Service	Filter events to quickly isolate trouble spots	✓*	✓
	Hassle-free log collection	✓	✓
	Simplified ticketing	✓	✓
	Show active PMRs and ticket history	✓*	✓

Resource Management	Functions	IBM Storage Insights (free)	IBM Storage Insights Pro (subscription)
Analytics and optimization	Predictive Alerts	✓	✓
	Customizable, multi-conditional alerting, including alert policies		✓
	Performance planning		✓
	Capacity planning		✓
	Business impact analysis (applications, departments, and groups)		✓
	Optimize data placement with tiering		✓
	Optimize capacity with reclamation		✓
Security	ISO/IEC 27001 Information Security Management standards certified	✓	✓
Entitlements		Free	Capacity-based subscription

Restriction: *If you have access to IBM Storage Insights but are not an IBM Storage Insights Pro subscriber, you must have a current warranty or maintenance agreement for an IBM block storage system to open tickets and send log packages.

The IBM FlashSystem (V8.2 and higher) is supported in conjunction with IBM Storage Insights and IBM Storage Insights Pro.

Note: The reporting feature is not available in IBM Storage Insights (free). In order to use the reporting feature, you must subscribe to IBM Storage Insights Pro or you can use IBM Spectrum Control.

For information about how to try to buy the IBM Storage Insights Pro version, see [IBM Support](#).

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 is available at 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.

As an on-premises application, IBM Spectrum Control doesn't send the metadata about monitored devices offsite, which is ideal for dark shops and sites that don't 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 hand in hand to monitor your storage environment. IBM Storage Insights for IBM Spectrum Control can transform your monitoring and support experience, as follows:

- ▶ 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 you don't have to.
- ▶ Use Call Home to monitor devices, get best practice recommendations, and filter events to quickly isolate trouble spots.
- ▶ Leverage IBM Support's ability 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 for as long as you have an active license with a current subscription and support agreement for IBM Spectrum Control license. If your subscription and support lapses, you're no longer eligible for IBM Storage Insights for IBM Spectrum Control. To continue using IBM Storage Insights for IBM Spectrum Control, simply renew your IBM Spectrum Control license. You can also choose to subscribe to IBM Storage Insights Pro.

Feature comparison of IBM Spectrum Control and IBM Storage Insights for IBM Spectrum Control

To understand the usability of IBM Spectrum Control and IBM Storage Insights for IBM Spectrum Control for your environment, we compare the features of IBM Spectrum Control and IBM Storage Insights for IBM Spectrum Control.

Table 9-2 on page 379 shows the features in IBM Spectrum Control and IBM Storage Insights for IBM Spectrum Control.

Table 9-2 Feature comparison

Resource Management	Features	IBM Spectrum Control (Advanced edition)	IBM Storage Insights for IBM Spectrum Control
Monitoring	Inventory	IBM and non-IBM block storage, file storage, object storage, hypervisors, fabrics, switches	IBM and non-IBM block storage, file storage, and object storage
	Call Home events		✓
	Performance	✓ (1-minute intervals)	✓ (5-minute intervals)
	Capacity	✓	✓
	Drill down performance workflow to troubleshoot bottlenecks	✓	✓
	Explore virtualization relationships		
	Explore replication relationships	✓	✓
	Retain performance data		
Service	Deployment method		
	Filter Call Home events to quickly isolate trouble spots		✓
	Hassle-free log collection		✓
	Simplified ticketing		✓
	Show active PMRs and ticket history		✓
	Active directory and LDAP integration for managing users	✓	
Reporting	Inventory, capacity, performance, and storage consumption reports	✓	✓
	Rollup reporting	✓	
	REST API	✓	
Alerting	Predictive Alerts	✓	✓
	Customizable, multi-conditional alerting, including alert policies	✓	✓

Resource Management	Features	IBM Spectrum Control (Advanced edition)	IBM Storage Insights for IBM Spectrum Control
Analytics	Performance planning	✓	✓
	Capacity planning	✓	✓
	Business impact analysis (applications, departments, and groups)	✓	✓
	Provisioning with service classes and capacity pools	✓	
	Balance workload across pools	✓	
	Optimize data placement with tiering	✓	✓
	Optimize capacity with reclamation	✓	✓
	Transform and convert volumes	✓	
Pricing		On-premises licensing	No charge for IBM Spectrum Control customers

You can upgrade IBM Storage Insights to IBM Storage Insights for IBM Spectrum Control, if you have an active license of IBM Spectrum Control. Details can be found at [IBM Storage Insights registration](#). Choose the option for IBM Spectrum Control, and follow the prompts.

IBM Storage Insights for IBM Spectrum Control doesn't include the service level agreement 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 there are differences:

- ▶ **IBM Storage Insights** is an off-premises, IBM Cloud service that is available free of charge if you own IBM block storage systems. It provides a unified dashboard for IBM block storage systems 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 software-defined storage (SDS) systems, and non-IBM block and file storage systems such as Dell/EMC storage systems.

It 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 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 gives you 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 just starting out with software-defined storage as well as those with established infrastructures who need to expand their capabilities.

- ▶ IBM Spectrum Control. Analytics-driven hybrid cloud data management to reduce costs.
- ▶ IBM Spectrum Protect. Optimized hybrid cloud data protection to reduce backup costs.
- ▶ IBM Spectrum Protect Plus. Complete VM protection and availability that's easy to set up and manage yet scalable for the enterprise.
- ▶ IBM Spectrum Archive. Fast data retention that reduces total cost of ownership for active archive data.
- ▶ IBM Spectrum Virtualize. Virtualization of mixed block environments to increase data storage.
- ▶ IBM Spectrum Accelerate. Enterprise block storage for hybrid cloud.
- ▶ IBM Spectrum Scale. High-performance, highly scalable hybrid cloud storage for unstructured data driving cognitive applications.
- ▶ IBM Cloud Object Storage. Flexible, scalable and simple object storage with geo-dispersed enterprise availability and security for hybrid cloud workloads.

As IBM Spectrum Storage Suite it contains IBM Spectrum Control, you can deploy IBM Storage Insight for IBM Spectrum Control.

Tip: Alerts are a good way to be notified of conditions and potential problems that are detected on your storage. If you use IBM Spectrum Control and IBM Storage Insights for IBM Spectrum Control together to enhance your monitoring capabilities, it is recommended that you 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

The following sections describe the steps to implement and set up IBM Storage Insights.

Sign up process

To use IBM Storage Insights with the IBM FlashSystem, first you have to sign up at [IBM Storage Insights registration](#):

- ▶ For the sign-up process, you need an IBM ID. If you don't have one, 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'll receive a Welcome email when IBM Storage Insights is ready. The email contains a direct link to your dashboard.

Figure 9-11 shows the IBM Storage Insight registration screen.

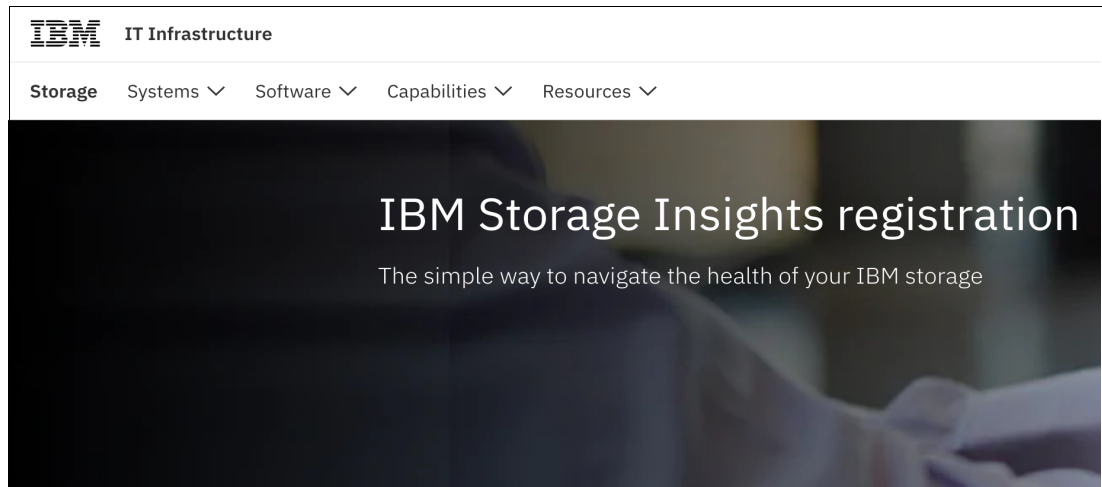


Figure 9-11 IBM Storage Insight registration screen

Figure 9-12 denotes the registration website when you scroll down. You can select here whether you want to register for IBM Storage Insights or IBM Storage Insights for Spectrum Control. For more details on the different editions of the IBM Storage Insights software, see “Licensing and editions of IBM Storage Insights” on page 374.

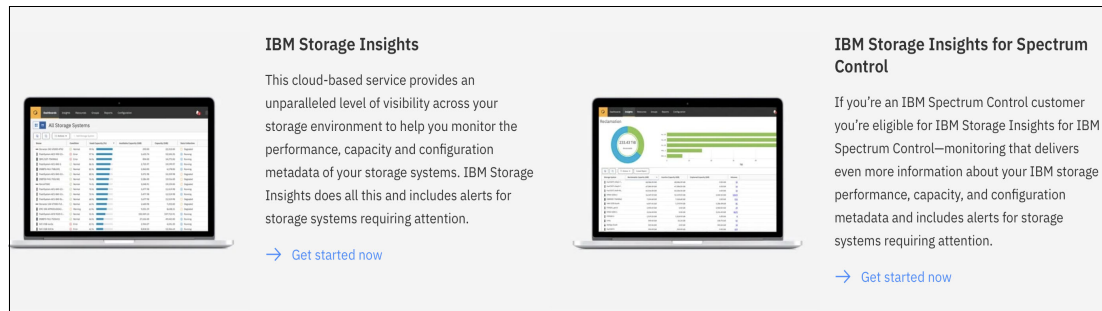


Figure 9-12 Choose IBM Storage Insights or IBM Storage Insights for Spectrum Control

Figure 9-13 shows the Log-in screen in the registration process. If you already have your credentials, type in your ID and click **Continue** to proceed to the next screen. If you do not have an ID, click **Create an IBMid**.

Log in to IBM

IBMid [Forgot IBMid?](#)

Remember me [i](#)

[Continue](#)

Don't have an account? [Create an IBMid](#)

Figure 9-13 Registration login screen

If you want to create an IBMid, see Figure 9-14 on page 384 and provide the following information

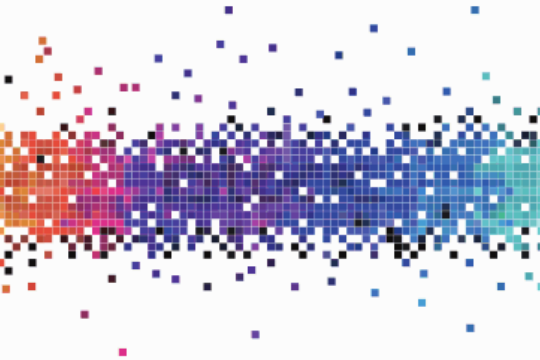
- ▶ Email
- ▶ First name
- ▶ Last name
- ▶ Country or region
- ▶ Password

Select the box if you want to receive Information from IBM to keep you informed of 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 an email.

For more information, see [IBM Privacy Statement](#).

Create your IBM account

Access to trials, demos, starter kits, services and APIs



Sign up for an IBMid

Already have an IBM account? [Log in](#)

Email *

First name *

Last name *

Country or region * (?) ✓

United States ▼

Set a password *

 👁

- 8 characters minimum
- One uppercase character
- One lowercase character
- One number

IBM may use my contact data to keep me informed of products, services and offerings:

by email.

You can withdraw your marketing consent at any time by sending an email to netsupp@us.ibm.com. Also you may unsubscribe from receiving marketing emails by clicking the unsubscribe link in each such email.


More information on our processing can be found in the [IBM Privacy Statement](#). By submitting this form, I acknowledge that I have read and understand the IBM Privacy Statement.

I accept the product [Terms and Conditions](#) of this registration form.

[Continue](#)

Figure 9-14 Create an IBM account

Figure 9-15 shows the login screen prompt for ID and password.



Sign in with your w3id

Remember my email address

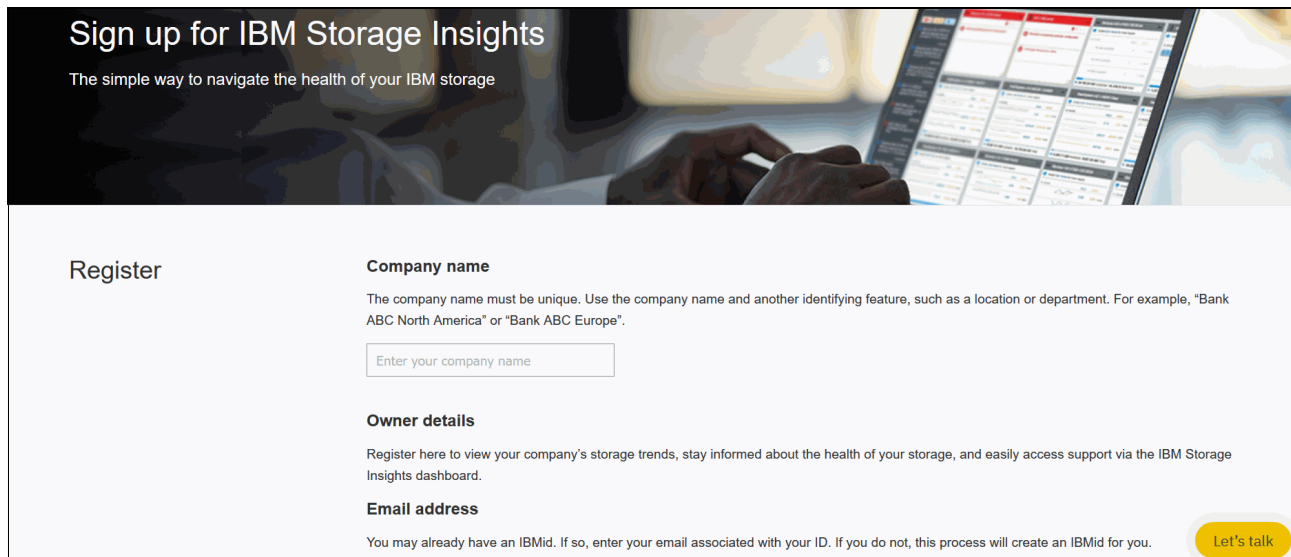
[Forgot password?](#)

[Sign In](#)

Figure 9-15 Registration - ID and password

Figure 9-16 shows the registration form. Complete the necessary information:

- ▶ Company name (must be unique)
- ▶ You might fill in other identifying features, such as a location or department.
 - Owner details
 - Email address / ID
 - The person who registered for IBM Storage Insights
 - Access granted for storage trends, health of storage and access to support
 - First and last name



Sign up for IBM Storage Insights
The simple way to navigate the health of your IBM storage

Register

Company name
The company name must be unique. Use the company name and another identifying feature, such as a location or department. For example, "Bank ABC North America" or "Bank ABC Europe".

Enter your company name

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 your email associated with your ID. If you do not, this process will create an IBMid for you.

Let's talk

Figure 9-16 IBM Storage Insights registration form

After registration for Storage Insights is complete, download and install the data collector for your system. Extract the data collector, run the data collector installer script, and ensure that your server (or virtual machine) can access the `host_name:port` that is specific to your instance of Storage Insights. After the data collector is installed on the system, you can add your storage devices to a Storage Insights dashboard.

Note: To connect to your instance of Storage Insights, you must configure your firewall to allow outbound communication on the default HTTPS port 443 using Transmission Control Protocol (TCP). User Datagram Protocol (UDP) is not supported.

Deploy a data collector

To deploy a lightweight data collector in your data center to stream performance, capacity, and configuration metadata to IBM Storage Insights:

1. Log in to IBM Storage Insights (the link is in your Welcome email).
2. From the Configuration > Data Collector page, download the data collector for your operating system (Windows, Linux, or AIX).
3. Extract the contents of the data collector file on the virtual machine or physical server where you want it to run.
4. For Windows, run `installDataCollectorService.bat`.
For Linux or AIX, run `installDataCollectorService.sh`.

After the data collector is deployed, it attempts to establish a connection to IBM Storage Insights. When the connection is complete, you're ready to start adding your storage systems for monitoring.

Requirements: 1 GB RAM, 1 GB disk space, and Windows, AIX, or Linux (x86-64 systems only).

Learn more at [IBM Storage Insights Documentation - Downloading and installing data collectors](#).

Note: To avoid potential problems, ensure that the operating system on the server or virtual machine where you install the data collector has general or extended support for maintenance and security.

Storage system metadata is sent to IBM Storage Insights such as:

- ▶ Information about the configuration of the storage system, such as name, firmware, and capacity.
- ▶ Information about the internal resources of the storage system, such as volumes, pools, nodes, ports, and disks. This includes the names and the configuration and capacity information for each internal resource.
- ▶ Information about the performance of storage system resources and internal resources such as pools and volumes.

For more information on how the metadata is collected and used, see:

- ▶ [IBM Storage Insights Fact Sheet PDF Download](#)
- ▶ [IBM Storage Insights Security Guide PDF Download](#)

Add storage system

Connect IBM Storage Insights to the storage systems that you want to monitor.

1. On the Operations dashboard in IBM Storage Insights, look for the button to add storage systems.
2. Click Add Storage Systems and follow the prompts. You can add one or more storage systems at a time.

Read more at [IBM Storage Insights Documentation - Adding storage systems](#).

View your dashboard

On the Operations dashboard, view:

- Storage systems that are being monitored.
- A dynamic diagnostic feed that tells you which storage systems require attention.
- Key capacity metrics so you know whether you've got enough capacity to meet your storage demands.
- Key performance metrics so you know whether the performance of your storage systems meets operational requirements.

Read more at [IBM Storage Insights Documentation - NOC dashboard](#).

Enable 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 you can act quickly to resolve incidents before they affect critical storage operations.

Read more at [IBM Storage Insights Documentation - Monitoring resources with Call Home](#).

Add 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.

1. In IBM Storage Insights, click your user name in the upper-right corner of the dashboard.
2. Click Manage Users.
3. On your MYIBM page, ensure that IBM Storage Insights is selected.
4. Click Add new user.

For more information, see [IBM Storage Insights Documentation - Adding and removing users](#).

9.3 Capacity metrics for block storage systems

Effective and exact capacity management is based on fundamental knowledge of capacity metrics in the IBM FlashSystem system. Data reduction pools (DRPs), thin provisioning, compression, and deduplication add several metrics to the IBM FlashSystem GUI, IBM Spectrum Control, and IBM Storage Insights. The capacity metrics in this section are based on IBM Spectrum Control V5.3.3.

The Capacity section on the Dashboard provides an overall view of system capacity. This section displays physical capacity, volume capacity, and capacity savings.

Physical capacity indicates the total capacity in all storage on the system. Physical capacity includes all the storage the system can virtualize and assign to pools. Physical capacity is displayed in a bar graph and divided into three categories: Stored Capacity, Available Capacity, and Total Physical. If the system supports self-compressing drives, certain system configurations make determining accurate physical capacity on the system difficult.

For example, if the system contains self-compressed drives and data reduction pools without compression enabled, the system cannot determine the accurate amount of physical capacity that is used on the system. In this case, over provisioning and losing access to write operations is possible. If this condition is detected by the system, the Physical Capacity section of the Dashboard page displays a message instead of capacity information.

To recover from this condition, you need to ensure that all thin-provisioned volumes and thin-provisioned volumes that are deduplicated are migrated to volumes with compression enabled in the data reduction pools. Alternatively, you can migrate the volumes to fully allocated volumes and use the drive compression to save capacity.

We will discuss the different values and how they help to determine the capacity utilization, trends in capacity and space usage, and more importantly can prevent an out of space situation in the environment. In order to not run out of space you have to understand the different level of components, such as arrays, pools, system, and so on.

You should understand which limits exist for each of them so that you understand while one pool might be fine, another pool(s) could run out of storage and just monitoring at the system level is not appropriate if you have two or more pools.

Figure 9-17 shows how to interpret the capacity and savings in a storage environment.

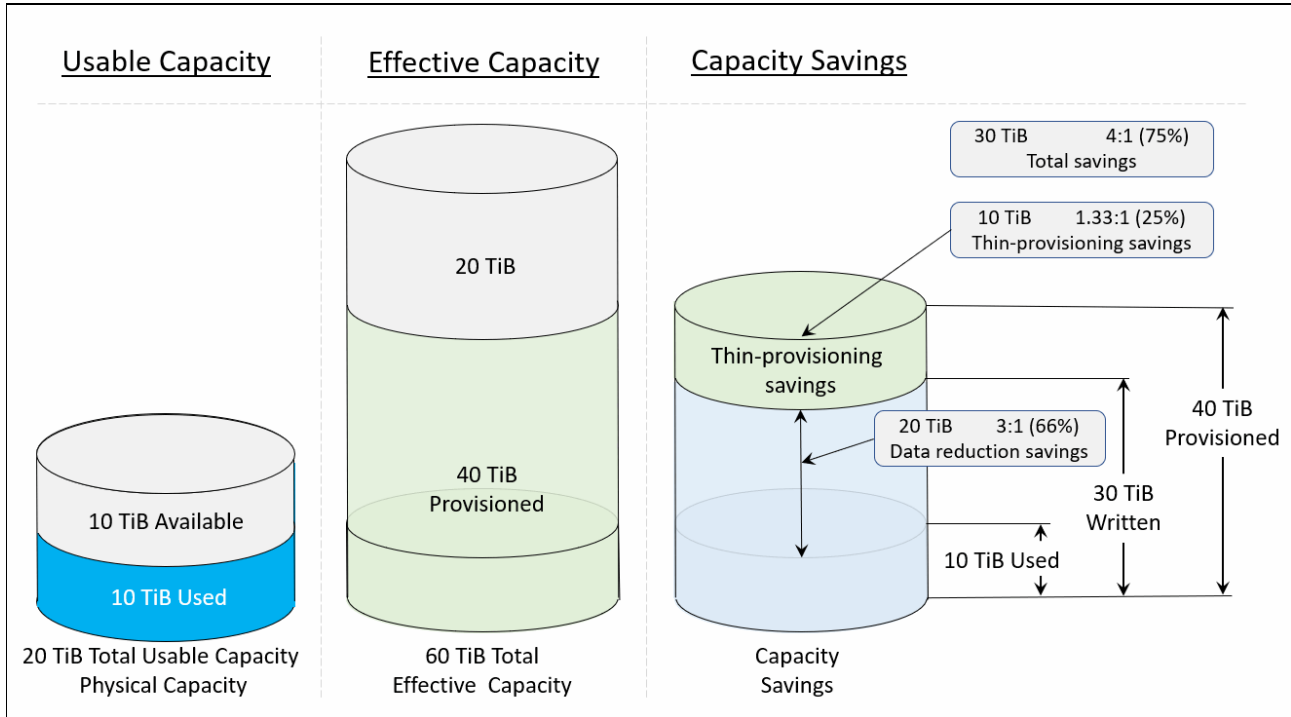


Figure 9-17 Understanding capacity information

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.

Alphabetical lists of the capacity and space usage metrics that you can add to charts are provided in the following sections:

- ▶ Storage system capacity metrics
- ▶ Pool capacity metrics
- ▶ Volume capacity metrics

9.3.1 Storage system capacity metrics

The following are the system capacity metrics.

Allocated Space (GiB)

The amount of space that is allocated to the regular and thin-provisioned volumes in the pools. If the pool is a parent pool, the amount of space that is allocated to the volumes in the child pools is also included.

The space that is allocated for thin-provisioned volumes is less than their virtual capacity, which is shown in the **Total Volume Capacity (GiB)** column. If a pool doesn't have thin-provisioned volumes, the value for allocated space is the same as the value for total volume capacity.

Allocated space is the same as used space on all storage systems with the following exceptions:

- ▶ IBM FlashSystem
- ▶ SAN Volume Controller
- ▶ Storwize that are thin-provisioned

Figure 9-18 shows the Allocated Space (61.45 GiB) of an IBM FlashSystem 9100.

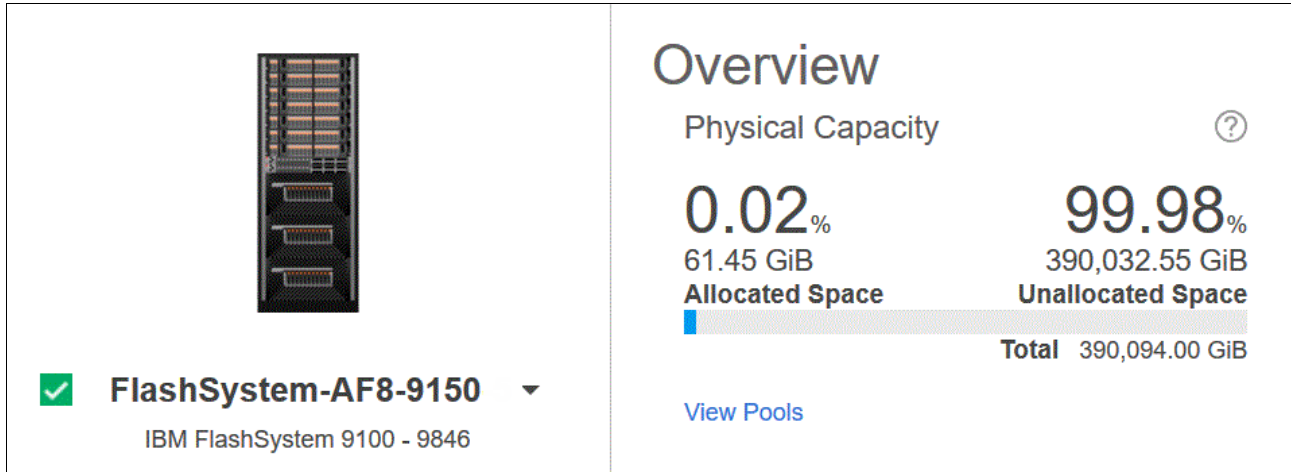


Figure 9-18 Allocated Space

Assigned Volume Space (GiB)

The total volume space in the storage system that is mapped or assigned to host systems, including child pool capacity. It is the sum of all volumes virtual size, so this value can exceed the physical capacity of your system, when you over provision storage. Complete the following steps:

1. Select **IBM Spectrum Control** → **choose your IBM Flash System** (for example, FlashSystem AF8-9150 - FS9100).
2. Select **Actions** → **View Capacity** and scroll down.
3. Verify whether the **Assigned Volume Space** check box has been selected (right-click and determine if the **Assigned Volume Space** is marked).

Figure 9-19 shows the Assigned Volume Space (GiB) in an IBM FlashSystem. The figure shows the total volume space for this system.

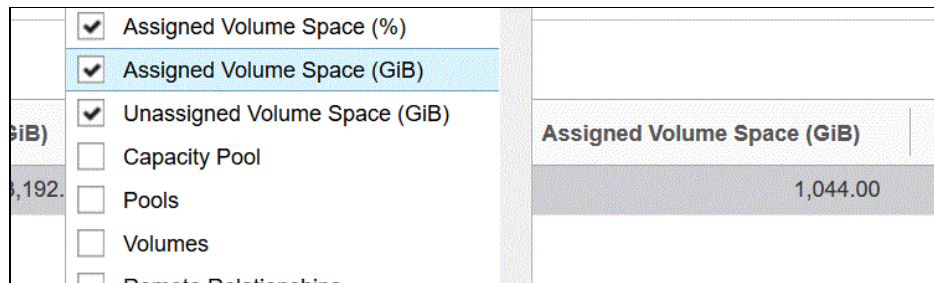


Figure 9-19 Assigned Volume Space (GiB)

Figure 9-20 shows a view of the current and historic Assigned Volume Space in an IBM FlashSystem.

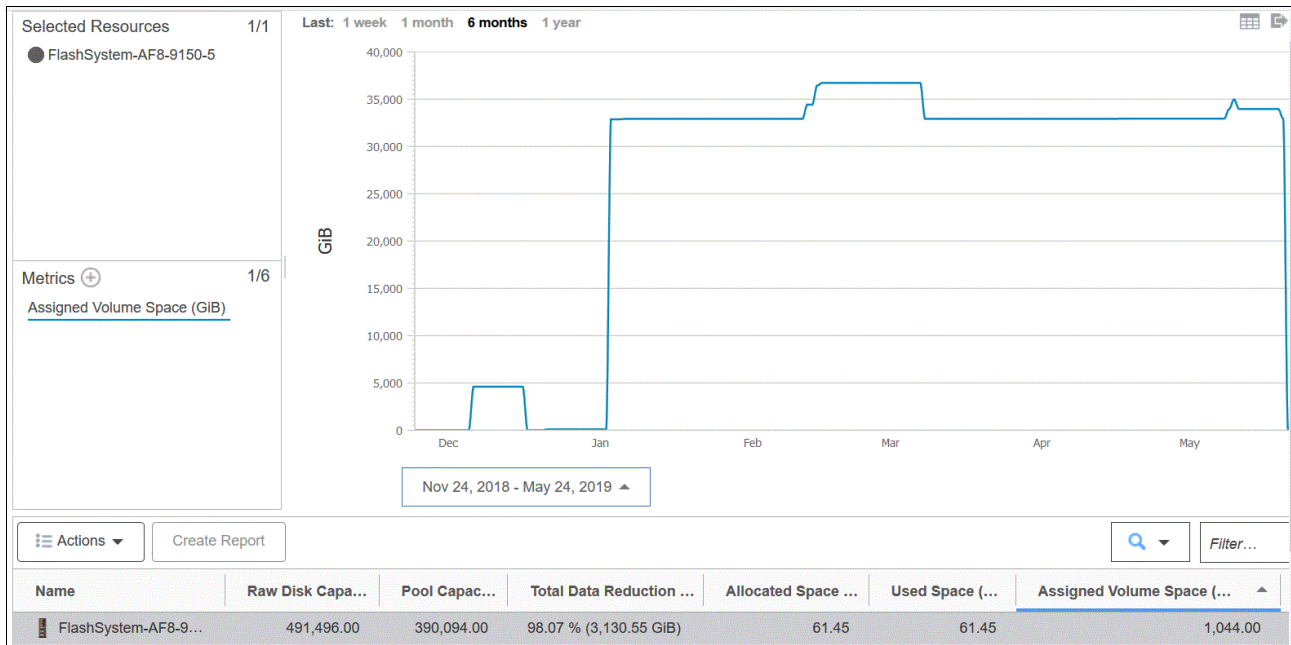


Figure 9-20 Assigned Volume Space - graph

To view other information, click the plus sign next to **Metrics** on the left.

Available Pool Space (GiB)

The total amount of the space in the pools that is not allocated to the volumes in the pools. To calculate available space, the following formula is used:

$$(\text{pool capacity} - \text{allocated space})$$

For some storage systems, the pool space is limited by the physical capacity after data reduction, so the more you can compress data the more you can store in such a pool. For other systems there is a limit in the address space before compression. This means that even if you can compress the data extremely highly, you might not be able to use all the physical space, because the address range before compression is exceeded. See 3.1.3, “Internal storage considerations” on page 68 for more details.

Compression savings (%)

The estimated amount and percentage of capacity that is saved by using data compression, across all pools in 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. Drive level compression occurs after striping across MDisk and the RAID distribution on the disks/FCM level. There is no information available about which volume a block of data belongs to that was just compressed, so the information about compression within the drives is only available at the array level.

The following formula is used to calculate the amount of storage space that is saved:

$$(\text{written space} \cdot \text{compressed size})$$

The following formula is used to calculate the percentage of capacity that is saved:

$$((\text{written space} \cdot \text{compressed size}) \div \text{written space}) \times 100$$

For example, the written space, 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 just 10 GiB. Therefore, the compression savings percentage across all compressed volumes is 75%.

Deduplication Savings (%)

The estimated amount and percentage of capacity that is saved by using data deduplication, across all data reduction pools on the storage system. The percentage is calculated across all deduplicated volumes in the pools and 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:

$$(\text{written space} \cdot \text{deduplicated size})$$

The following formula is used to calculate the percentage of capacity that is saved:

$$((\text{written space} \cdot \text{deduplicated size}) \div \text{written space}) \times 100$$

For example, the written space, 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%.

Physical Allocation (%)

The percentage of physical capacity in the pools that is allocated to the regular volumes, the thin-provisioned volumes, and the volumes in child pools. Check the value for physical allocation to see:

- ▶ Whether the physical capacity of the pools is fully allocated. That is, the value for physical allocation is 100%.
- ▶ Whether you have sufficient capacity to provision new volumes with storage
- ▶ Whether you have sufficient capacity to allocate to the compressed and thin-provisioned

Figure 9-21 shows a brief description of Physical Allocation (%).

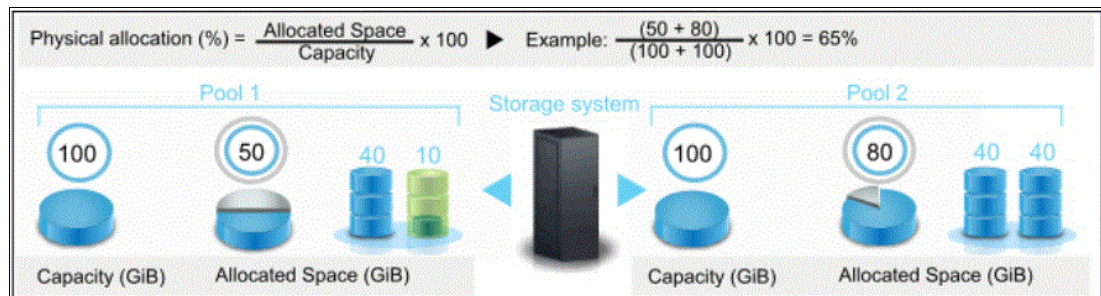


Figure 9-21 Physical Allocation (%)

Pool Capacity (GiB)

The total amount of storage space in the pools, which might include overhead space if the disks for the pools aren't formatted.

Pool Shortfall (%)

The percentage of space that is over-committed to the pools with thin-provisioned volumes. For example, you commit 100 GiB of space to a thin-provisioned volume in a pool with a capacity of 50 GiB. As the space is allocated to the thin-provisioned volume in increments of 10 GiB, the space available for allocation decreases and the shortfall in capacity becomes more acute.

To calculate the shortfall, the following formula is used:

$$[(\text{unallocatable space} \div \text{committed but unallocated space}) \times 100]$$

A pool shortfall occurs when you commit more space to the volumes in the pools than is physically available to the pools. If the physical space available to the pools is less than the committed virtual space, then the pools do not have enough space to fulfill the commitment to the virtual space.

For example, the physical capacity of the pools is 70 GiB, but 150 GiB of virtual space was committed to the thin-provisioned volumes. If the volumes are using 50 GiB, then 100 GiB is still committed to those volumes (150 GiB - 50 GiB) with only 20 GiB of available pool space (70 GiB - 50 GiB). Because only 20 GiB of the pool space is available, 80 GiB of the committed space cannot be allocated (100 GiB - 20 GiB). In this case, the percentage of committed space that is unavailable is 80% [(80 GiB ÷ 100 GiB) × 100].

The advantage of using shortfall rather than a simple overprovisioning factor, is that shortfall always has values between 0 and 100% and therefore is well suited for simple alerting that can be applied to multiple pools.

Figure 9-22 explains the shortfall.

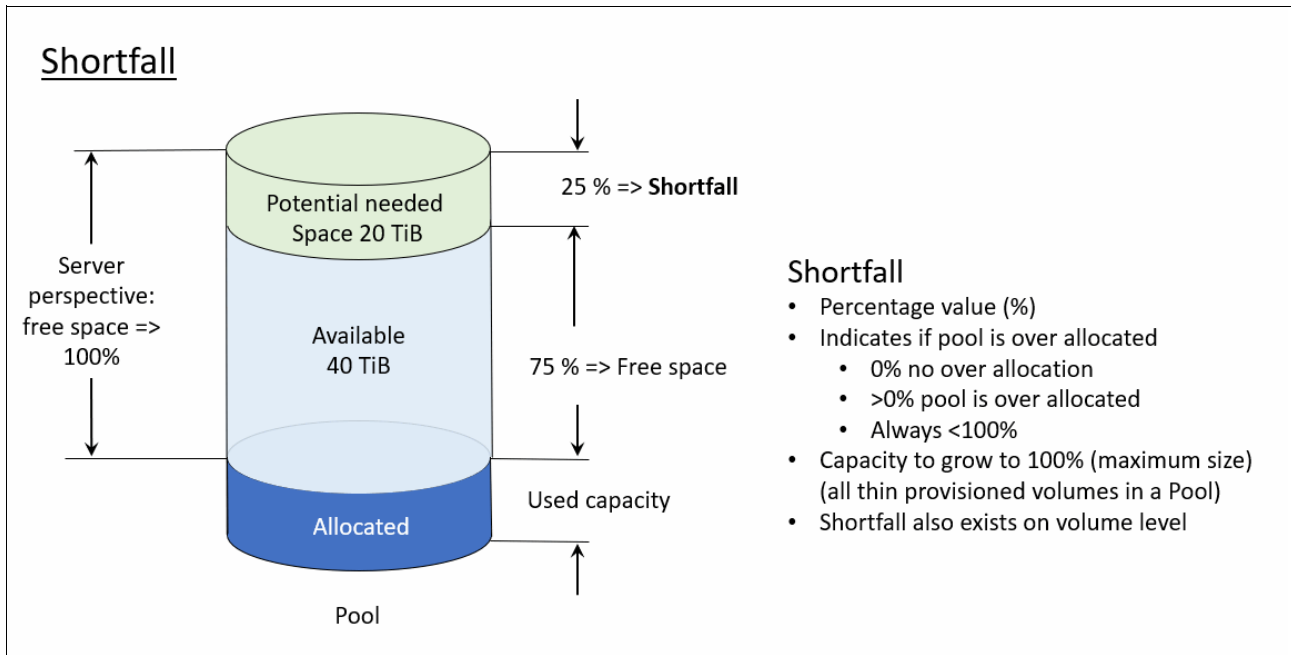


Figure 9-22 Shortfall

Total Data Reduction Savings (%)

The estimated amount and percentage of capacity that is saved by using data deduplication, data compression, and thin provisioning.

The following formula is used to calculate the amount of storage space that is saved:

$$(\text{Total Volume Capacity} \cdot \text{Allocated Space})$$

The following formula is used to calculate the percentage of capacity that is saved:

$$((\text{Total Volume Capacity} \cdot \text{Allocated Space}) \div \text{Total Volume Capacity}) \times 100$$

Total Volume Capacity (GiB)

The total amount of storage space that can be made available to the regular and thin-provisioned volumes in the pools. 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. In other words this is the capacity that a server will see.

Unallocatable Volume Space (GiB)

The amount of space that cannot be allocated to volumes because the physical capacity of the pools cannot meet the demands for virtual space. The following formula is used to calculate this value:

$$[\text{Total Volume Capacity} \cdot \text{Pool Capacity}]$$

Unallocated Volume Space (GiB)

The total amount of remaining space that can be allocated to the volumes in the pools. The following formula is used to calculate this value:

$$[\text{Total Volume Capacity} - \text{Allocated Space}]$$

The space that is allocated for thin-provisioned volumes is typically less than their virtual capacity. Therefore, the unallocated space represents the difference between the virtual capacity and the allocated space for all the volumes in the pools.

Virtual Allocation (%)

The percentage of the physical capacity that is committed to the virtual capacity of the volumes in the pool. If the value exceeds 100%, the physical capacity doesn't meet the demands for virtual capacity. The following formula is used to calculate this value:

$$[(\text{Total Volume Capacity} \div \text{Pool Capacity}) \times 100]$$

Example: If the allocation percentage is 200% for a total storage pool size of 15 GiB, then the virtual capacity that is committed to the volumes in the pool is 30 GiB. This configuration means that twice as much space is committed than is physically contained in the pool. If the allocation percentage is 100% for the same pool, then the virtual capacity that is committed to the pool is 15 GiB. This configuration means that all the physical capacity of the pool is already allocated to volumes.

An allocation percentage that is higher than 100% is considered aggressive because insufficient physical capacity is available in the pool to satisfy the maximum allocation for all the thin-provisioned volumes in the pool. In such cases, you can use the value for Shortfall (%) to estimate how critical the shortage of space is for a pool.

9.3.2 Pool capacity metrics

If sufficient data is collected about the pools in your data center, you can view charts that compare the capacity of the pools with the space that is allocated to the pools and the space that is still available in the pools. In the **Zero Capacity** column on the Pools page, you can see the date, based on the space usage trends for the pool, when the pool runs out of available space.

Tip: To order the pools in the table by the amount of space available to the pools, click **Filter by column**, and then click **Zero Capacity**.

Figure 9-23 shows an example of Zero Capacity trend.

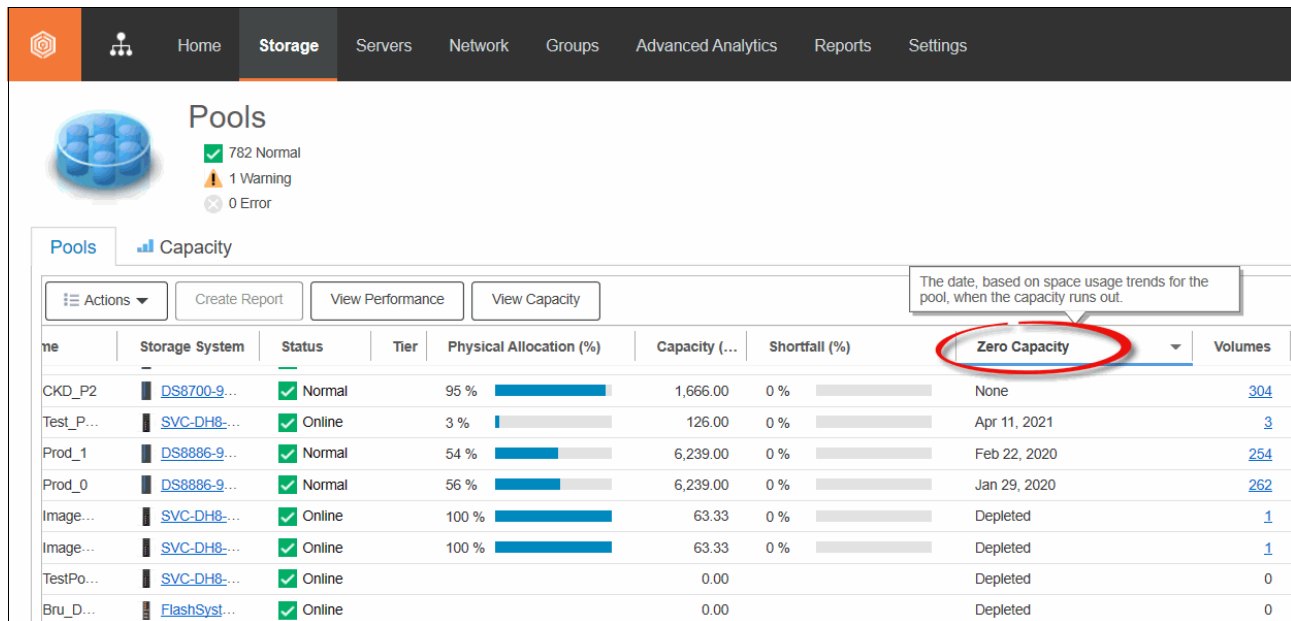


Figure 9-23 Zero Capacity trend

The values that can be shown in the Zero Capacity column:

- ▶ **A date**
The data 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 when the pool will be filled, for example if the trend is negative, as data is moved out of the pool
- ▶ **Depleted**
the pool is already 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.

Allocated Space (GiB)

The amount of space that is allocated to the regular and thin-provisioned volumes in the pool. If the pool is a parent pool, the amount of space that is allocated to the volumes in the child pools is also included.

The space that is allocated for thin-provisioned volumes is less than their virtual capacity, which is shown in the **Total Volume Capacity** column. If a pool does not contain thin-provisioned volumes, this value is the same as Total Volume Capacity.

Allocated space is the same as used space on all storage systems, except for resources that run IBM Spectrum Virtualize. These resources might have more allocated space than used space if the storage administrator pre-allocated some space for thin-provisioned volumes when the volumes were created.

Assigned Volume Space (GiB)

The space on all of the volumes in a pool that are mapped or assigned to host systems. For a thin-provisioning pool, this value includes the virtual capacity of thin-provisioned volumes, which might exceed the total space in the pool.

Available Pool Space (GiB)

The amount of space that is available to create new volumes in the pool. If the pool is a parent pool, the amount of space that is allocated to the volumes in the child pools is also included.

Available Soft Space (GiB)

The amount of virtual storage space that is available to allocate to volumes in a storage pool.

Capacity (GiB)

The total amount of storage space in the pool, which might include overhead space if the disks for the pool aren't formatted.

Compression Savings (%)

The estimated amount and percentage of capacity that is saved by using data compression. The percentage is calculated across all compressed volumes in the pool 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:

$(\text{written space} \cdot \text{compressed size})$

The following formula is used to calculate the percentage of capacity that is saved:

$((\text{written space} \cdot \text{compressed size}) \div \text{written space}) \times 100$

For example, the written space, 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 just 10 GiB. Therefore, the compression savings percentage across all compressed volumes is 75%.

Deduplication Savings (%)

The estimated amount and percentage of capacity that is saved by using data deduplication. The percentage is calculated across all deduplicated volumes in the pool and 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:

$(\text{written space} \cdot \text{deduplicated size})$

The following formula is used to calculate the percentage of capacity that is saved:

$$((\text{written space} \cdot \text{deduplicated size}) \div \text{written space}) \times 100$$

For example, the written space, 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%.

Enterprise HDD Available Space (GiB)

The amount of storage space that is available on the Enterprise hard disk drives that can be used by Easy Tier for re-tiering the volume extents in the pool.

Enterprise HDD Capacity (GiB)

The total amount of storage space on the Enterprise hard disk drives that can be used by Easy Tier for re-tiering the volume extents in the pool.

Nearline HDD Available Space (GiB)

The amount of storage space that is available on the Nearline hard disk drives that can be used by Easy Tier for re-tiering the volume extents in the pool.

Nearline HDD Capacity (GiB)

The total amount of storage space on the Nearline hard disk drives that can be used by Easy Tier for re-tiering the volume extents in the pool.

Physical Allocation (%)

The percentage of physical capacity in the pool that is allocated to the regular volumes, the thin-provisioned volumes, and the volumes in child pools. This value is always less than or equal to 100% because you cannot allocate more physical space than is available in a pool. Check the value for physical allocation to see:

- ▶ Whether the physical capacity of the pool is fully allocated. That is, the value for physical allocation is 100%.
- ▶ Whether you have sufficient capacity to provision new volumes with storage.
- ▶ Whether you have sufficient capacity to allocate to the compressed and thin-provisioned volumes in the pool.

Figure 9-24 shows the Physical Allocation.

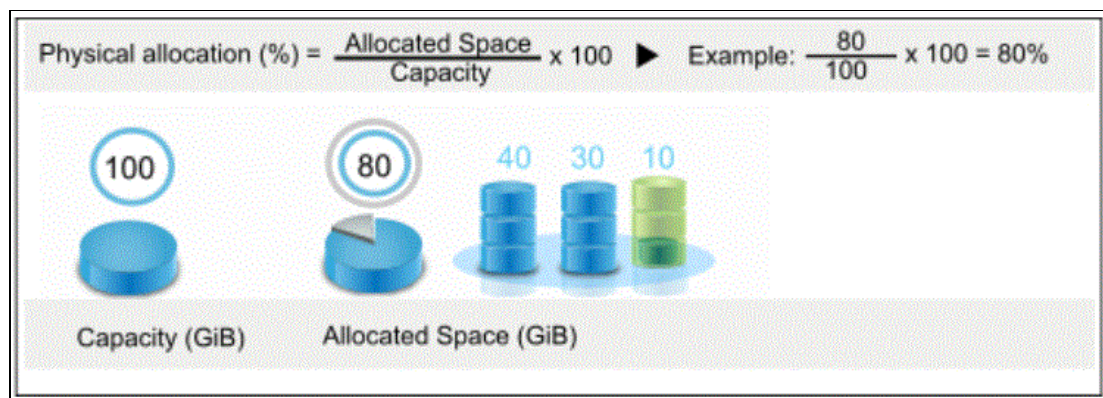


Figure 9-24 Physical Allocation

Shortfall (%)

The percentage of space that is over committed to pools with thin-provisioned volumes. For example, you commit 100 GiB of space to a thin-provisioned volume in a pool with a capacity of 50 GiB. As the space is allocated to the thin-provisioned volume in increments of 10 GiB, the space available for allocation decreases and the shortfall in capacity becomes more acute.

If the pool is not thin-provisioned, the shortfall percentage equals zero. If shortfall percentage isn't calculated for the storage system, the field is left blank.

To calculate shortfall, the following formula is used:

$$[(\text{Unallocatable Space} \div \text{Committed but Unallocated Space}) \times 100]$$

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 virtual space, then the pool does not have enough space to fulfill the commitment to virtual space. This value represents the percentage of the committed virtual space 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.

Example: The remaining physical capacity of a pool is 70 GiB, but 150 GiB of virtual space 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 space in the pool, the shortfall is 30% when the following calculation is used:

$$[(100 \text{ GiB unallocated volume space} - 70 \text{ GiB remaining pool space}) \div 100 \text{ GiB unallocated volume space}] \times 100$$

Soft Space (GiB)

The amount of virtual storage space that is configured for the pool.

SSD Available Space (GiB)

The amount of storage space that is available on the solid-state drives that can be used by Easy Tier for re-tiering the volume extents in the pool.

SSD Capacity (GiB)

The total amount of storage space on the solid-state drives that can be used by Easy Tier for re-tiering the volume extents in the pool.

Tier 0 Flash Available Space (GiB)

The amount of storage space that is available on the Tier 0 flash solid-state drives that can be used by Easy Tier for re-tiering the volume extents in the pool.

Tier 0 Flash Capacity (GiB)

The total amount of storage space on the Tier 0 flash solid-state drives that can be used by Easy Tier for re-tiering the volume extents in the pool.

Tier 1 Flash Available Space (GiB)

The amount of storage space that is available on the Tier 1 flash, read-intensive solid-state drives that can be used by Easy Tier for re-tiering the volume extents in the pool.

Tier 1 Flash Capacity (GiB)

The total amount of storage space on the Tier 1 flash, read-intensive solid-state drives that can be used by Easy Tier for re-tiering the volume extents in the pool.

Total Data Reduction Savings (%)

The estimated amount and percentage of capacity that is saved by using data deduplication, data compression, and thin provisioning, across all volumes in the pool.

The following formula is used to calculate the amount of storage space that is saved:

Total Volume Capacity · Allocated Space

The following formula is used to calculate the percentage of capacity that is saved:

$((\text{Total Volume Capacity} \cdot \text{Allocated Space}) \div \text{Total Volume Capacity}) \times 100$

Total Volume Capacity (GiB)

The total amount of storage space that can be made available to the regular and thin-provisioned volumes in 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.

Unallocatable Volume Space (GiB)

The amount of space that cannot be allocated to volumes because the physical capacity of the pools cannot meet the demands for virtual space. The following formula is used to calculate this value:

$[\text{Total Volume Capacity} \cdot \text{Pool Capacity}]$

Unallocated Volume Space (GiB)

The total amount of remaining space that can be allocated to the volumes in the pools. The following formula is used to calculate this value:

$[\text{Total Volume Capacity} - \text{Allocated Space}]$

The space that is allocated for thin-provisioned volumes is typically less than their virtual capacity. Therefore, the unallocated space represents the difference between the virtual capacity and the allocated space for all the volumes in the pools.

Unassigned Volume Space (GiB)

The total amount of space in the volumes that are not assigned to hosts.

Virtual Allocation (%)

The percentage of the physical capacity that is committed to the virtual capacity of the volumes in the pool. If the value exceeds 100%, the physical capacity doesn't meet the demands for virtual capacity. The following formula is used to calculate this value:

$[(\text{Total Volume Capacity} \div \text{Pool Capacity}) \times 100]$

Example: If the allocation percentage is 200% for a total storage pool size of 15 GiB, then the virtual capacity that is committed to the volumes in the pool is 30 GiB. This configuration means that twice as much space is committed than is physically contained in the pool. If the allocation percentage is 100% for the same pool, then the virtual capacity that is committed to the pool is 15 GiB. This configuration means that all the physical capacity of the pool is already allocated to volumes.

An allocation percentage that is higher than 100% is considered aggressive because insufficient physical capacity is available in the pool to satisfy the maximum allocation for all the thin-provisioned volumes in the pool. In such cases, you can use the value for Shortfall (%) to estimate how critical the shortage of space is for a pool.

9.3.3 Volume capacity metrics

You use the capacity chart to detect capacity shortages for the following types of volumes:

- ▶ Space-efficient volumes, such as compressed volumes and thin-provisioned volumes
- ▶ Regular volumes that use Easy Tier to re-tier volume extents

You can review the allocation of space to space-efficient volumes to detect capacity shortfalls. You can also review the space usage of volumes that use Easy Tier to distribute volume extents across Enterprise HDD, Nearline HDD, and SSD storage.

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.

Allocated Space (GiB)

The amount of space that is allocated to the compressed, 5-provisioned, or the Easy Tier volume. Typically, the space that is allocated to the compressed or thin-provisioned volume is less than the capacity of the volume. For Easy Tier volumes, allocated space is the space that is allocated to the volume's extents on the Enterprise HDD, Nearline HDD, or SSD drives.

Capacity (GiB)

The capacity of the compressed or the thin-provisioned volume, which comprises the sum of the allocated and the unallocated space. If the disks for the pool aren't formatted, the capacity of the volume might include the overhead space.

Compression Savings (%)

The estimated amount and percentage of capacity that is saved by using data compression. The following formula is used to calculate the amount of storage space that is saved:

(written space · compressed size)

The following formula is used to calculate the percentage of capacity that is saved:

$((\text{written space} \cdot \text{compressed size}) \div \text{written space}) \times 100$

Exception: For compressed volumes that are deduplicated on storage systems that run IBM Spectrum Virtualize (for example, IBM FlashSystem), the Compression Savings (%) column is blank.

Enterprise HDD Capacity (GiB)

The total amount of storage space on the Enterprise hard disk drive that the Easy Tier volume uses for re-tiering the volume extents.

Nearline HDD Capacity (GiB)

The total amount of storage space on the Nearline hard disk drive that the Easy Tier volume uses for re-tiering the volume extents.

SSD Capacity (GiB)

The total amount of storage space on the solid-state drive that the Easy Tier volume uses for re-tiering the volume extents.

Tier 0 Flash Capacity (GiB)

The total amount of storage space on the Tier 0 flash solid-state drives that can be used by Easy Tier for re-tiering the volume extents in the pool.

Tier 1 Flash Capacity (GiB)

The total amount of storage space on the Tier 1 flash, read-intensive solid-state drives that can be used by Easy Tier for re-tiering the volume extents in the pool.

Used Space (GiB)

The amount of allocated space that is used by the compressed, thin-provisioned, or Easy Tier volume.

For compressed and thin-provisioned volumes, used space might not be the same as allocated space (for example, IBM FlashSystem, Storwize or IBM SAN Volume Controller).

For thin-provisioned volumes, used space might not be the same as allocated space because you can preallocate space to the thin-provisioned volumes when the volumes are created. For compressed volumes, used space might not be the same as allocated space because more space is used to read the data than is used to write the data to the disk. For regular volumes that use Easy Tier on the storage systems that are listed, used space is the same as allocated space.

Written Space (GiB)

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 space 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 just 10 GiB.

9.4 Creating alerts for IBM Spectrum Control and IBM Storage Insights

This section describes alerting with IBM Spectrum Control and IBM Storage Insights Pro. Note that the free version of Storage Insights does not support alerting.

New data reduction technologies add more intelligence and capacity savings to your environment. If you use data reduction on different layers, such as hardware compression in the IBM FlashSystem Flash Core Modules and additionally in the DRPs, or if you virtualize an IBM FlashSystem system, pay more attention to preventing insufficient space from remaining in the back-end storage device.

First it is important to distinguish between thin provisioning and over-allocation (over-provisioning). Thin provisioning is a method for optimizing the use of available storage. It relies on allocation of blocks of data on demand versus the traditional method of allocating all of the blocks up front. This methodology 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, 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. This will result in an out-of-space condition.

Note: You must constantly monitor your environment to avoid over-provisioning situations, which can be harmful to the environment and can cause data loss.

It is also important to keep free space for garbage collection in the background. For more information, see “DRP internal details” on page 111 and Chapter 4, “Storage pools” on page 101.

Data Reduction technologies give back some space. If the space that’s used for the data can be reduced, the saved-up space can be used for other data. But keep in mind that, depending on the type of data, deleting the data might not result in freeing up much space.

If you have three identical or almost identical files on a file system which have been deduplicated, the compression ratio would be good (three files - but stored only once). If you now delete one file, you will not gain more space, because the deduplicated data must stay on the storage since the two other versions of the file still see the data. The same is true when you use several FlashCopies of one source.

9.4.1 Alert examples

Table 9-3 shows alerts for IBM FlashSystem systems based on Array or Pool level.

Table 9-3 Event examples for IBM Flash System

System	Entity	Resource Type	Event
Flash System with FCM	Array	Usable capacity	Available Physical Space <= nn% (* Example shown in 9.4.2, “Alert to monitor back-end capacity: Available Physical Space (%)” on page 401)
	Pool	Efficient Capacity	Physical allocation >= nn%
Flash System other media	Pool	Usable Capacity	Physical allocation >= nn%

Other alerts are possible as well, but generally % alerts are best suited, as the alert definition applies to all pool in a storage system.

9.4.2 Alert to monitor back-end capacity: Available Physical Space (%)

This section describes how to deploy IBM Spectrum Control or IBM Storage Insights Pro to monitor storage capacity and set up thresholds to notify and prevent your system from running out of space.

The following example shows how to create an alert to get status Information about the remaining physical space on an IBM FlashSystem.

First, assign a severity to an alert. Assigning a severity can help you 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 9-4 shows the possible alert severities.

Table 9-4 Alert severities

Option	Description
Critical	Alert is critical and needs to 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 three thresholds:

- ▶ Critical (15% space in the RAID Array left)
- ▶ Warning (20% space in the RAID Array left)
- ▶ Information (30% space in the RAID Array left)

Adjust the percentage levels to the required levels as needed. Keep in mind that the process to extend storage might take its time (ordering process, installation, provisioning, and so on).

The advantage of this method of setting up an **Alert Policy** is, that you can add various IBM FlashSystem (or even other storage subsystem such as the IBM FlashSystem 900) to this customized alert.

Figure 9-25 shows how to start creating a new Alert Policy which monitors the remaining free capacity in the RAID Array. The customized Alert is not tied to the Data Reduction Pool, as it works regardless of the pool type.

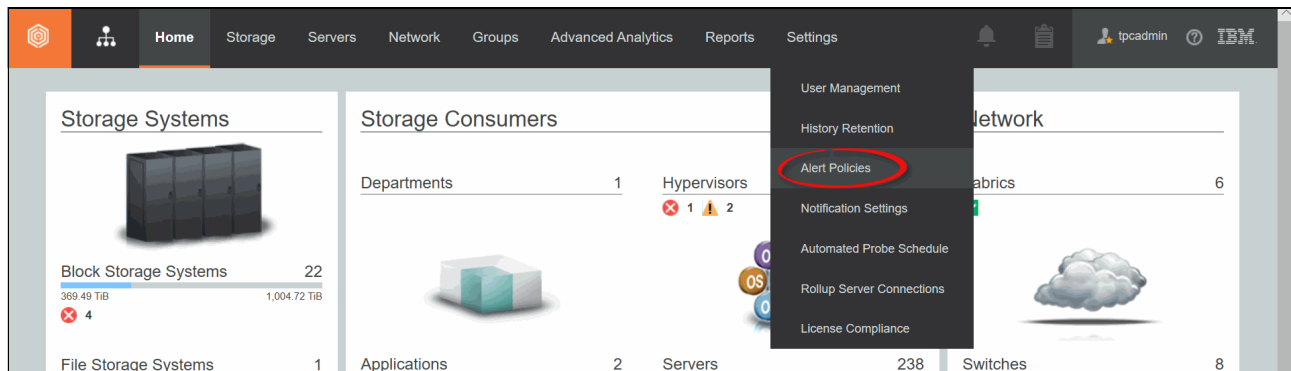


Figure 9-25 Create new Alert Policy

In the following example we create a new Alert Policy by copying the existing one. You might change an existing Alert Policy (in our example the Default Policy) as well. Keep in mind, a storage subsystem can be active in only one Alert Policy.

Figure 9-26 shows the Default Policy of the IBM FlashSystem 9100.

All Policies				
Alert Policies		Resources by Policy		
☰ Actions ▾ Create Policy				
Name	Resource Type	Resources	Alert Definitions	Email Addresses
Agentless AIX Server	Agentless AIX	2	0	
Agentless Linux Server	Agentless Linux	2	0	
asd	Storwize	0	23	
Custom Windows policy	Windows	1	10	
Custom DS8000 policy	DS8000	1	25	
Custom DS8000 policy Copy	DS8000	0	25	
Custom SAN Volume Controller policy	SAN Volume Controller	4	22	
Custom Switch policy	Switch	0	13	
Default Cloud Object Storage policy	Cloud Object Storage	0	2	
Default DS8000 policy	DS8000	0	25	
Default Fabric policy	Fabric	0	2	
Default FlashSystem 840 or 900 policy	FlashSystem 840 or 940	0	17	
Default FlashSystem 9100 policy	FlashSystem 9100	0	22	
Default FlashSystem A9000 or A9000R policy	FlashSystem A9000 or A90...	0	16	

Figure 9-26 Default Alert Policy

Figure 9-27 describes how to copy an existing Policy in order to create a new one. Hover the mouse pointer over the existing Policy to be copied, then click it, and choose **Copy Policy**.

Default DS8000 policy	DS8000	0
Default Fabric policy	Fabric	0
Default FlashSystem 840 or 900 policy	FlashSystem 840 or 940	0
Default FlashSystem 9100 policy	System 9100	0
Default FlashSystem A9000 or A9000R policy	System A9000 or A90...	0
Default FlashSystem V840 or V9000 policy	FlashSystem V840 or V9000	0

Figure 9-27 Copy existing Policy and create a new one

Figure 9-28 on page 404 shows how to rename the previously copied Policy. The new Policy will be stored as an additional Policy. Keep in mind that one IBM FlashSystem can only be added to a single policy. You can add the system (choose **Optional**, select the resource, and select the check box) later if you are not sure at this time.

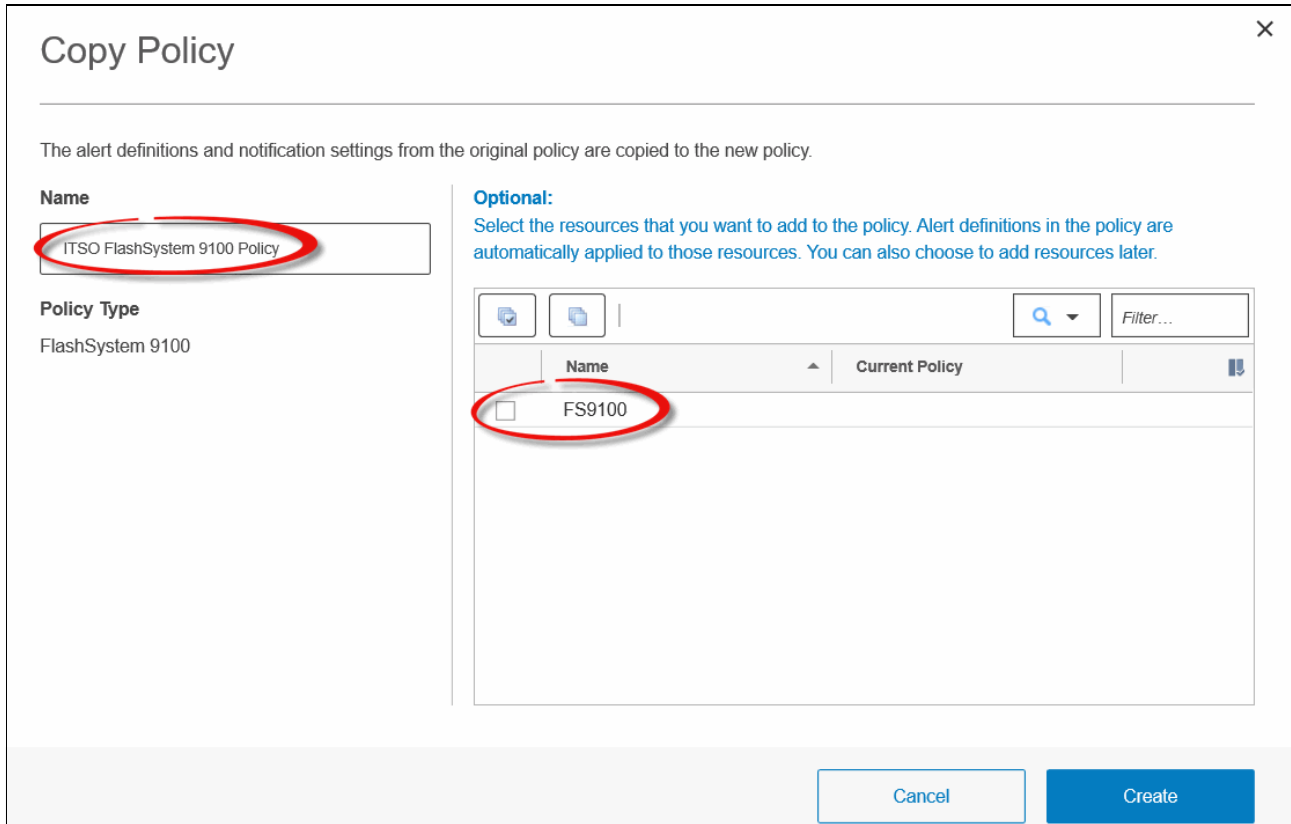


Figure 9-28 Store copied policy

Figure 9-29 shows the newly created Alert Policy “ITSO FlashSystem 9100 Policy” with all the existing alerts inherited Alert Definitions from the Default Policy.

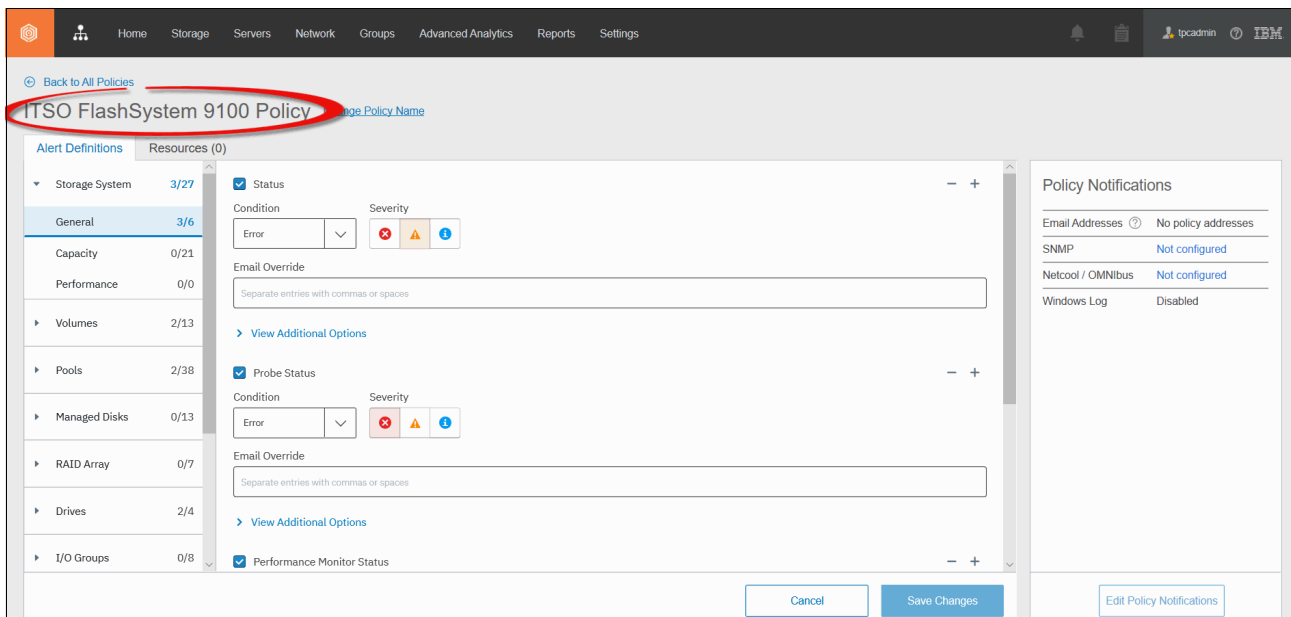


Figure 9-29 New Policy with inherited Alert Definitions

Figure 9-30 shows how to choose the required Alert Definitions **RAID Array** → **Capacity** in the screen.

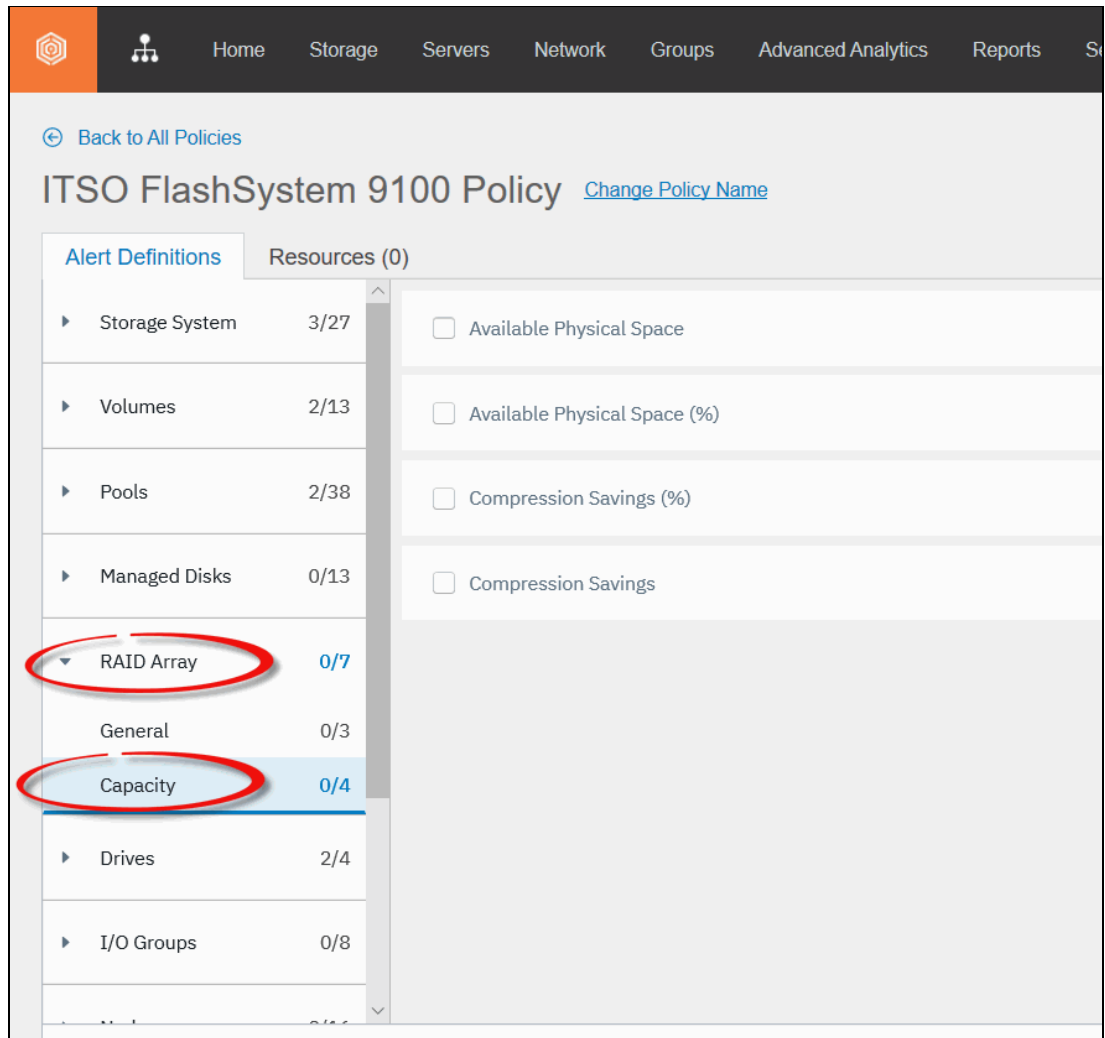


Figure 9-30 Alert Definition RAID Array > Capacity

Figure 9-31 on page 406 denotes the tasks for setting up the Critical definition by monitoring the Available Physical Space (%) and releasing Policy Notifications at 15%. This example implies that when 85% or more physical space is taken a critical Notification via the predefined method will be sent.

Predefined methods can be:

- ▶ Email Addresses
- ▶ SNMP
- ▶ IBM Netcool® / OMNibus
- ▶ Windows Event Log or UNIX syslog

You must define the methods before you can choose them. If your environment does not have predefined methods, see Figure 9-31.

Note: With IBM Storage Insights, you can only send emails.

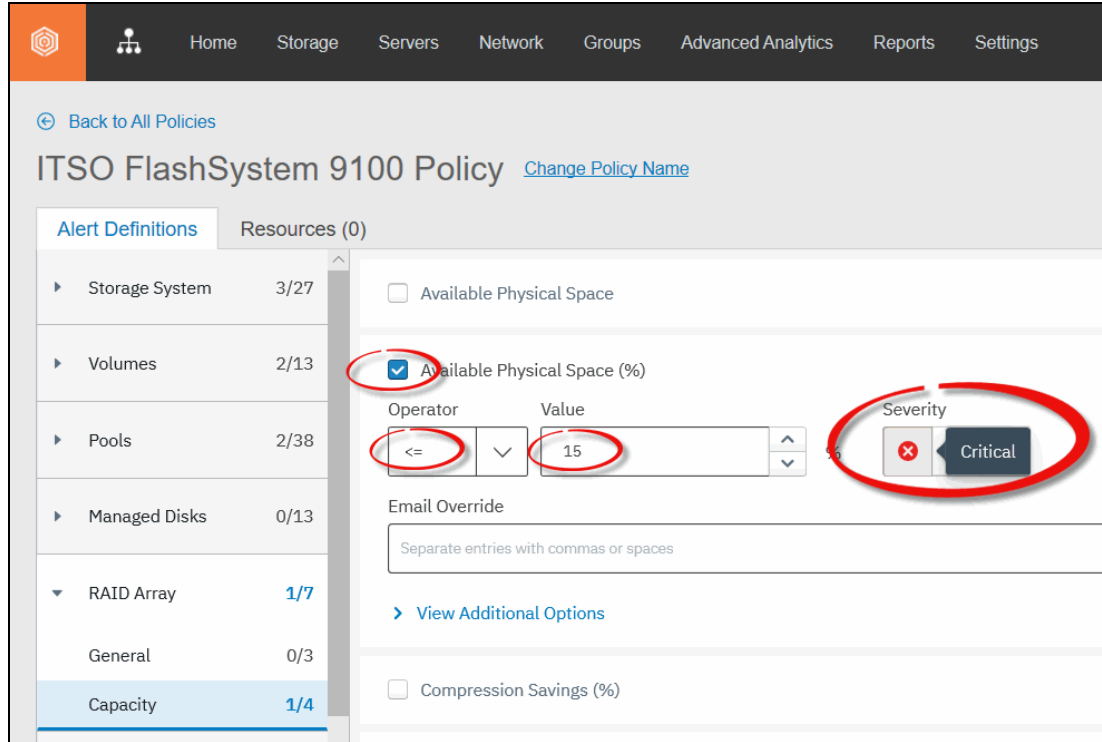


Figure 9-31 Alert Definition 15% or less Available Physical Space - Critical

Figure 9-32 shows how to change the Frequency of the notification. You can choose here to get more frequent notification for the Critical Threshold “15% Available Physical Space”. In this example we choose to set the frequency to “every day”.

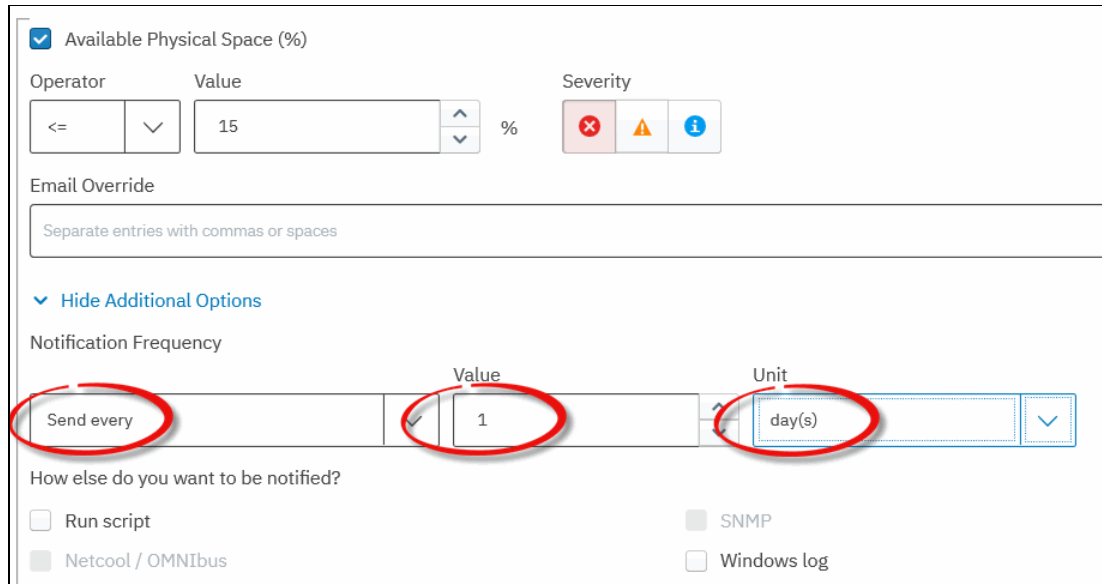


Figure 9-32 Alert Definition - Critical - change frequency

Figure 9-33 shows how to set up the Warning level at 20% or less Available Physical Space. To proceed, choose the plus sign at the previously defined Definition (Critical) and make the following selections, (Operator: \leq , Value: 20%, and Severity: **Warning**).

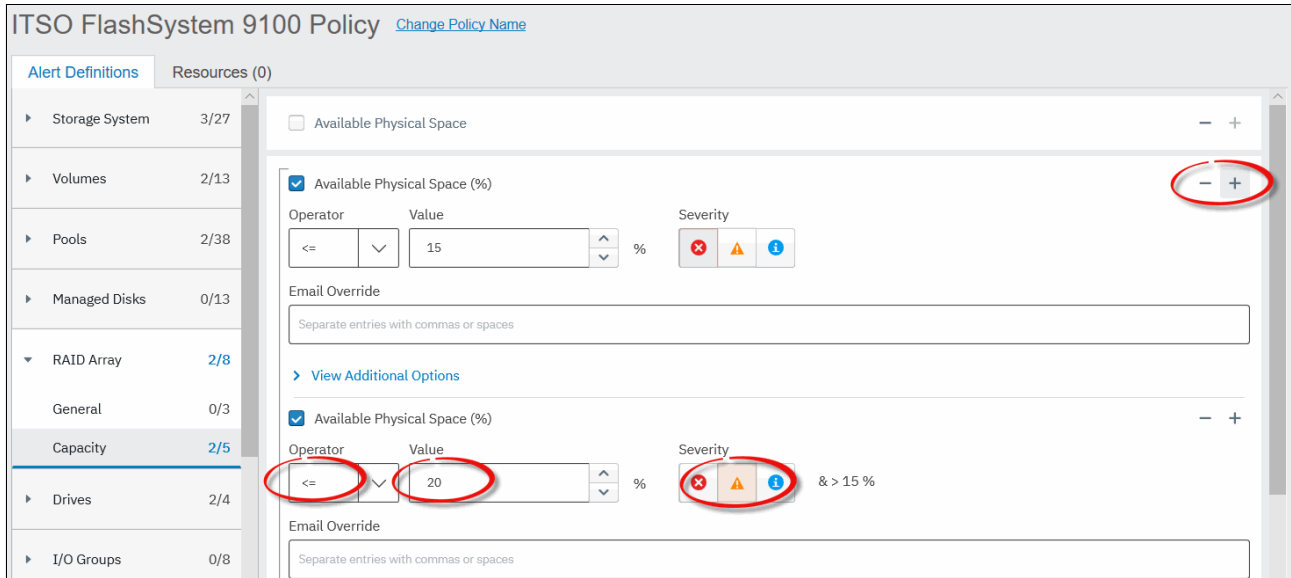


Figure 9-33 Alert Definition 20% or less Available Physical Space - Warning

Figure 9-34 depicts how to set up the Notification Threshold at 30%.

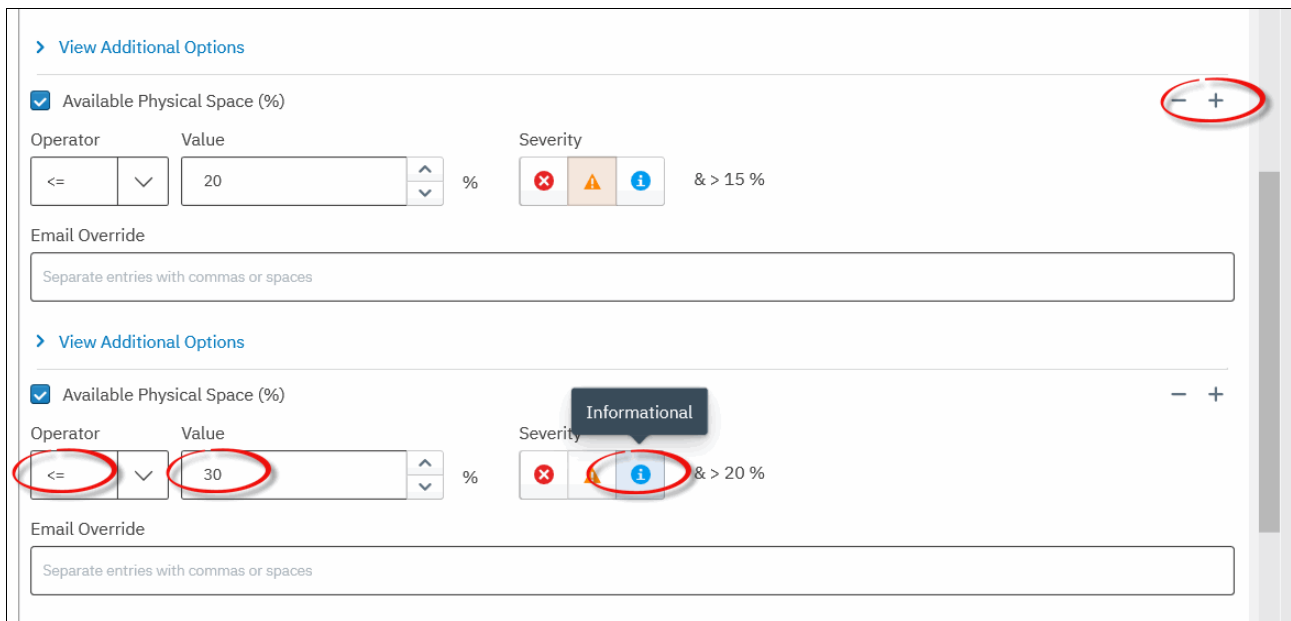


Figure 9-34 Alert Definition 30% or less Available Physical Space - Notification

Figure 9-35 shows how to configure the Notification Settings in your monitoring environment.

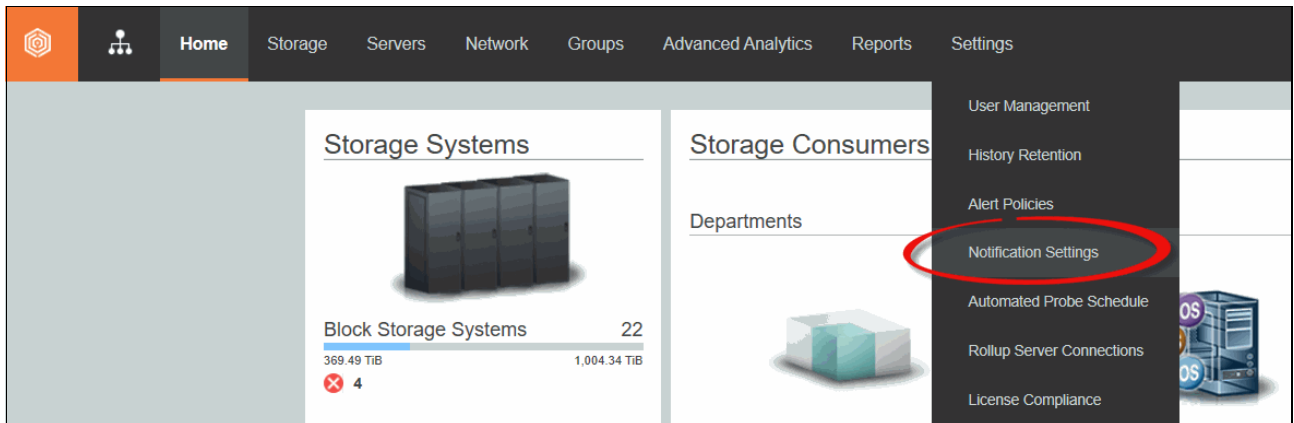


Figure 9-35 Change Notification Settings

Figure 9-36 shows how to set up e-mail Notification Settings.

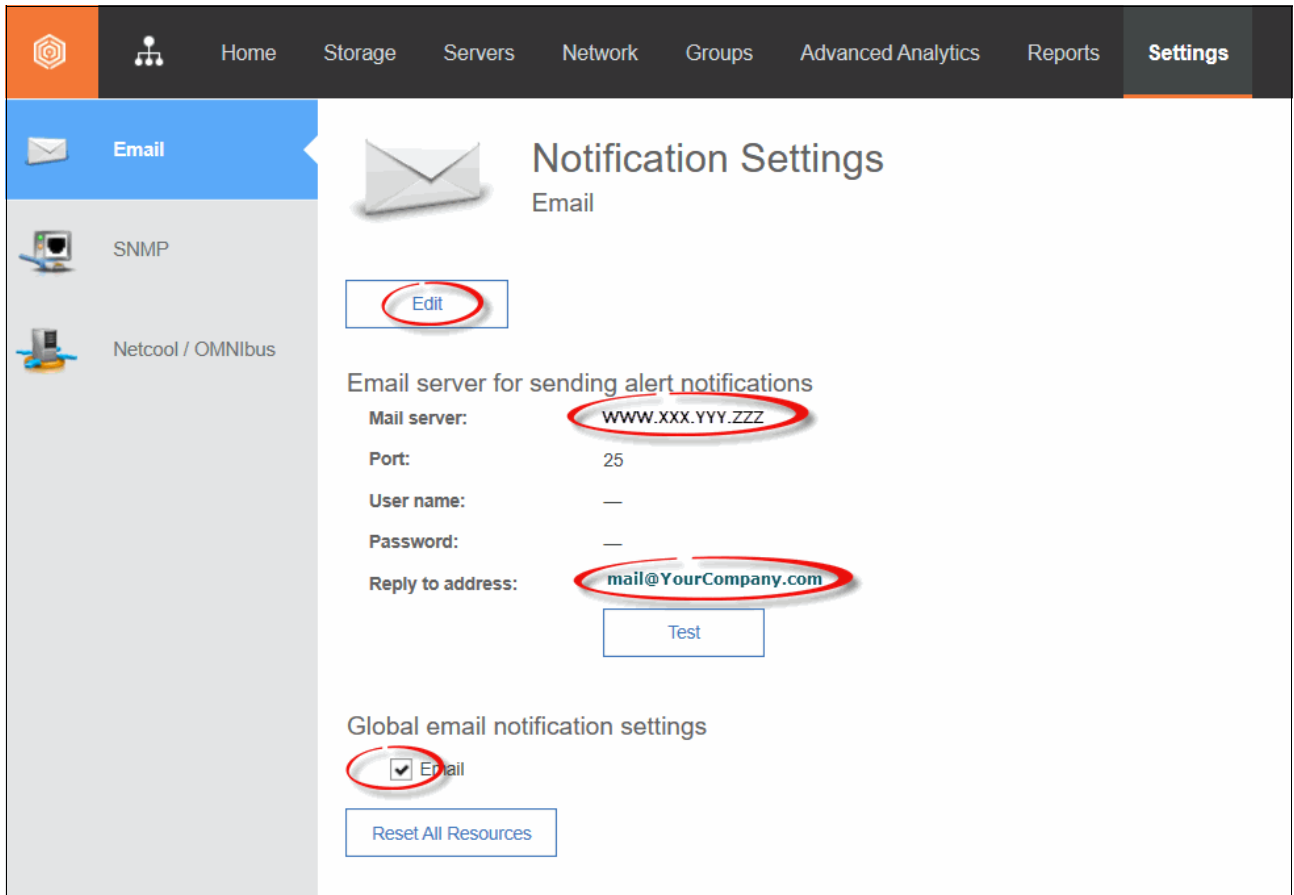


Figure 9-36 Notification Settings: Details

9.5 Error condition example with IBM Spectrum Control: FC port

The following guidance shows an example of an FC port problem. It shows how you can spot an error with IBM Spectrum Control and shows how to drill down into the details.

Figure 9-37 denotes a testing environment with several storage subsystems. Three errors can be spotted in the GUI. In this example, we show how you can get more details about the first highlighted error.

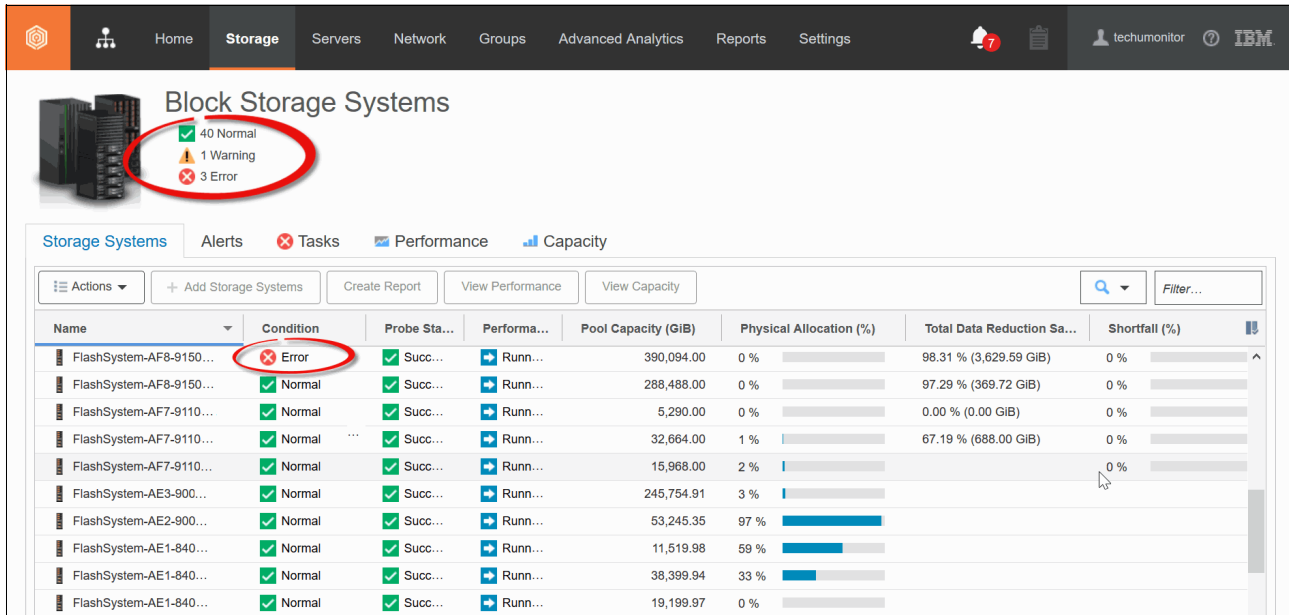


Figure 9-37 Error condition spotted on a storage subsystem with IBM Spectrum Control

Figure 9-38 shows details of the storage subsystem and which entity is affected. In this case it is related to internal resources: ports. Two ports have been stopped and have caused this condition in the environment.

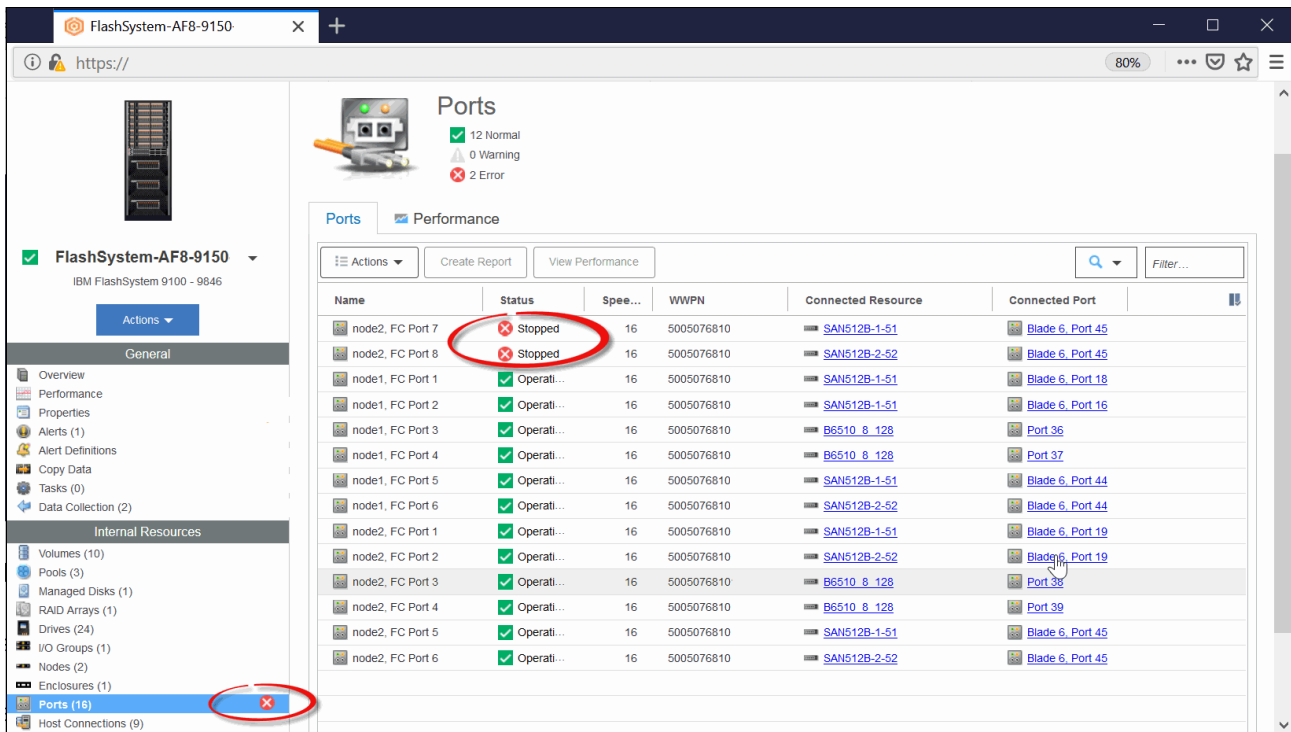


Figure 9-38 FlashSystem error condition - internal resources - ports

Figure 9-39 shows the details of one of the stopped ports.

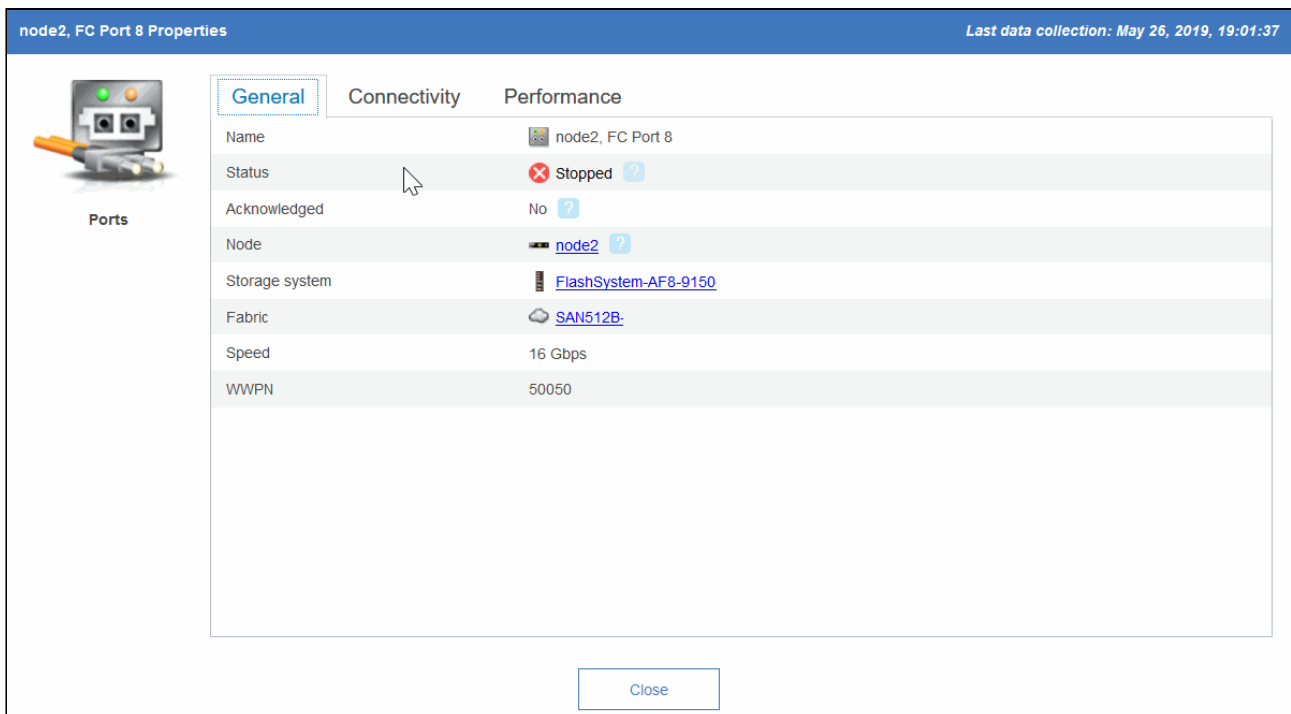


Figure 9-39 FC ports stopped: detail view

The ports are probably stopped for a reason, so from the panel in Figure 9-38 on page 410 select both, click **Actions** and select **Acknowledge Status**. The ports will still be shown with the red icon, but now the icon is overlaid with a check mark. After a short time, this change will be propagated so that the storage system is shown as being in a green status again.

In other cases, you might have to replace hardware, after you have opened a ticket in your internal system with the vendor, you should still acknowledge the status, so that any other errors will make the storage system go from green to red again and you see that a second event has happened.

9.6 Important metrics

The following metrics are some of the most important metrics that must be analyzed to understand a performance problem in IBM FlashSystem systems. Those metrics are valid to analyze the front end (by node, by host, or by volume) or the back-end (by MDisk or by Storage Pool):

Terminology: R/W stands for Read and Write operations.

- ▶ **I/O Rate R/W:** The term *I/O* is used to describe any program, operation, or device that transfers data to or from a computer, and to or from a peripheral device. Every transfer is an output from one device and an input into another. Typically measured in IOPS.
- ▶ **Data Rate R/W:** The data transfer rate (DTR) is the amount of digital data that is moved from one place to another in a specific time. In case of Disk or Storage Subsystem, this metric is the amount of data moved from a host to a specific storage device. Typically measured in MB per second.
- ▶ **Response time R/W:** This is the time taken for a circuit or measuring device, when subjected to a change in input signal, to change its state by a specified fraction of its total response to that change. In case of Disk or Storage Subsystem, this is the time used to complete an I/O operation. Typically measured in ms.
- ▶ **Cache Hit R/W:** This is the percentage of times that read data or write data can be found in cache or can find cache free space that it can be written to.
- ▶ **Average Data Block Size R/W:** The block size is the unit of work for the file system. Every read and write is done in full multiples of the block size. The block size is also the smallest size on disk that a file can have.
- ▶ **Port-to-Local Node Queue Time (Send):** 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. A good scenario has less than 1 ms on average.
- ▶ **Port Protocol Errors (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. The amount of time value is measured from the last time that the node was reset. In Fibre Channel technology, buffer-to-buffer credit is used to control the flow of frames between ports. In our experience less is better than more. However, in the real life this metric can be from 5% on average up to 20% peak without affecting performance.
- ▶ **Port data rate (send and receive):** The average amount of data in MBps for operations in which the port receives or sends data.

- ▶ **Port Protocol Errors (Zero Buffer Credit Timer):** The number of microseconds that the port is not able to send frames between ports because there is insufficient buffer-to-buffer credit. In Fibre Channel technology, buffer-to-buffer credit is used to control the flow of frames between ports. Buffer-to-buffer credit is measured from the last time that the node was reset. This value is related to the data collection sample interval.
- ▶ **Port Congestion Index:** The estimated degree to which frame transmission was delayed due to a lack of buffer credits. This value is generally 0 - 100. The value 0 means there was no congestion. The value can exceed 100 if the buffer credit exhaustion persisted for an extended amount of time. When you troubleshoot a SAN, use this metric to help identify port conditions that might slow the performance of the resources to which those ports are connected.
- ▶ **Global Mirror (Overlapping Write Percentage):** The percentage of overlapping write operations that are issued by the Global Mirror primary site. Some overlapping writes are processed in parallel, and so they are excluded from this value.
- ▶ **Global Mirror (Write I/O Rate):** The average number of write operations per second that are issued to the Global Mirror secondary site. Keep in mind that IBM FlashSystem systems have a limited number of GM I/Os that can be delivered.
- ▶ **Global Mirror (Secondary Write Lag):** The average number of extra milliseconds that it takes to service each secondary write operation for Global Mirror. This value does not include the time to service the primary write operations. Monitor the value of Global Mirror Secondary Write Lag to identify delays that occurred during the process of writing data to the secondary site.

Note: The host attributed response time is also a very important metric, which should be used in conjunction with IBM Spectrum Control V5.3.3 or higher. Previous versions had a calculation error.

V5.2.x version is not supported after September 30th 2019.

Several other metrics are supplied to IBM Spectrum Control from IBM FlashSystem. For more information about all metrics, see [IBM Spectrum Control 5.4.2 Documentation - Performance metrics for resources that run IBM Spectrum Virtualize](#).

9.7 Performance support package

If you have performance issues on your system at any level (for example, host, volume, node, or pools), consult IBM Support. You need to provide detailed performance data about the IBM FlashSystem so that the problem can be diagnosed. Generate a performance support package with detailed data by using IBM Spectrum Control.

In this scenario, you export performance data for a IBM FlashSystem to a compressed package, and you then send the package to IBM Support, as shown in Figure 9-40.

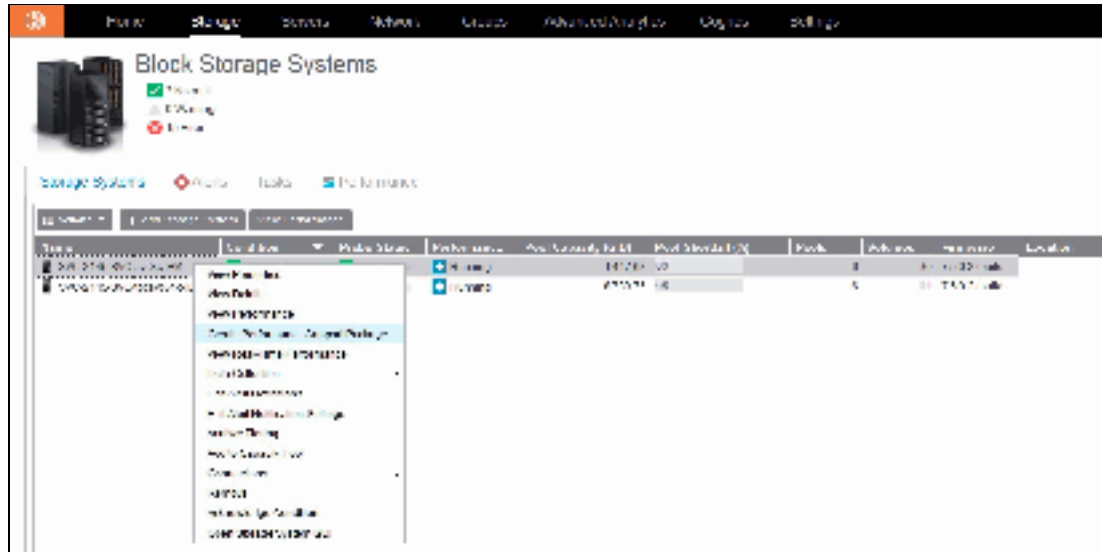


Figure 9-40 Performance support package creation

When the package has been created, you are requested to download it in .zip format. The package includes different reports in .csv format, as shown in Figure 9-41.

File Name	File Type	Size
log.txt	Text Document	1 KB
PerfReport_ITSO_SVC_ESC_Disks_20161017-233400_12hrs0mins.csv	CSV File	1 KB
PerfReport_ITSO_SVC_ESC_HostConnections_20161017-233400_12hrs0mins.csv	CSV File	1 KB
PerfReport_ITSO_SVC_ESC_IOGroups_20161017-233400_12hrs0mins.csv	CSV File	12 KB
PerfReport_ITSO_SVC_ESC_ManagedDisks_20161017-233400_12hrs0mins.csv	CSV File	18 KB
PerfReport_ITSO_SVC_ESC_Nodes_20161017-233400_12hrs0mins.csv	CSV File	18 KB
PerfReport_ITSO_SVC_ESC_Pools_20161017-233400_12hrs0mins.csv	CSV File	16 KB
PerfReport_ITSO_SVC_ESC_StoragePorts_20161017-233400_12hrs0mins.csv	CSV File	36 KB
PerfReport_ITSO_SVC_ESC_StorageSystem_20161017-233400_12hrs0mins.csv	CSV File	12 KB
PerfReport_ITSO_SVC_ESC_Volumes_20161017-233400_12hrs0mins.csv	CSV File	12 KB

Figure 9-41 Package files example

For more information about how to create a performance support package, see [IBM Spectrum Control 5.4.2 Documentation - Exporting performance data for storage systems and fabrics](#).

Note: The performance data might be large, especially if the data is for storage systems that have 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 is automatically excluded from the performance data. To include volume data and 1-minute sample data, select the **Advanced package** option on the Create Performance Support Package wizard.

9.8 Metro and Global Mirror monitoring with IBM Copy Services Manager and scripts

Copy Services Manager is part of IBM Spectrum Control and controls copy services in storage environments. Copy services are features that are used by storage systems, such as IBM FlashSystem systems, to configure, manage, and monitor data-copy functions. Copy services include IBM FlashCopy, Metro Mirror, Global Mirror, and Global Mirror Change Volumes (GMCV).

You can use Copy Services Manager 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

One of the most important events that needs to be monitored when IBM FlashSystem systems are implemented in a disaster recovery (DR) solution with Metro Mirror (MM) or Global Mirror (GM) functions, is to check whether MM or GM has been suspended because of a 1920 or 1720 error.

With IBM FlashSystem systems, you can suspend the MM or GM relationship to protect the performance on the primary site when MM or GM starts to affect write response time. That suspension can be caused by several factors. IBM FlashSystem systems *do not restart MM or GM automatically*. They must be restarted manually.

For information on setting IBM FlashSystem systems alert monitoring, see 9.1.1, “Monitoring with the GUI” on page 364. When MM or GM is managed by IBM CSM and if a 1920 error occurs, IBM CSM can automatically restart MM or GM sessions, and can set the delay time on the automatic restart option. This delay allows some time for the situation to correct itself.

Alternatively, if you have several sessions, you can stagger them 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 Copy Services Manager processes the 1720/1920 event and when the automatic restart is issued.

CSM is also able to automatically restart unexpected suspends. When you select this option, the Copy Services Manager server automatically restarts the session when it unexpectedly suspends due to reason code 1720 or 1920. 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 time period.

The number of times that a restart is attempted is determined by the storage server **gmlinktolerance** value. If the number of allowable automatic restarts is exceeded within the time 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 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 copy services management server waits after an unexpected suspend before automatically restarting the session. The range of possible values is 0 - 43200. The default is 0, which specifies that the session is restarted immediately following an unexpected suspend.

9.8.1 Monitoring MM and GM with scripts

The IBM FlashSystem system provides a complete command line interface (CLI), which allows you to interact with your systems by using scripts. The scripts can run in the IBM FlashSystem shell, but with a limited script command set available, or they can run out of the shell using any preferred scripting language.

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 have occurred, or to react to an SNMP or email alert received. The script can then start some specific recovery action based on your recovery plan and environment.

Customers who do not use IBM Copy Service Manager have created 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 you want to implement with scripts, and together try to find if one exists in the IBM Intellectual Capital Management repository that can be reused.

9.9 Monitoring Tier1 SSD

Monitoring Tier1 Solid State Drive (SSD) requires that special attention must be paid to the endurance events that can be triggered. For monitoring purposes, make note of the new fields that are listed in Table 9-5.

Table 9-5 Field changes to drive and array devices

Field	Description
write_endurance_used	Metric pulled from within drive (SAS spec) relating to the amount of data written across the life of the drive that is divided by the anticipated amount (2.42 PB for the 15.36 TB drive) Starts at 0, and can continue > 100
write_endurance_usage_rate	Measuring / Low / Marginal / High Takes 160 Days to get initial measurement; Low: Approximately 5.5 Years or more Marginal: Approximately 4.5 – 5.5 Years High: Approximately < 4.5 years High triggers event SS_EID_VL_ER_SSD_WRITE_ENDURANCE_USAGE_RATE_HIGH
replacement_date	The Current Date + Endurance Rate * Remaining Endurance Triggers event SS_EID_VL_ER_SSD_DRIVE_WRITE_ENDURANCE_LIMITED at 6 Months before limit

If you see either of these triggered events, contact your IBM service representative to put an action plan in place:

SS_EID_VL_ER_SSD_WRITE_ENDURANCE_USAGE_RATE_HI4GH

SS_EID_VL_ER_SSD_DRIVE_WRITE_ENDURANCE_LIMITED



Maintenance

As an IT environment grows and is renewed, so must the storage infrastructure. One of the many benefits that the IBM FlashSystem family software (IBM Spectrum Virtualize) provides, is to greatly simplify the storage management tasks that system administrators need to perform.

This chapter highlights guidance for the maintenance activities of storage administration by using the IBM FlashSystem family software installed on the product. This guidance can help you to maintain your storage infrastructure with the levels of availability, reliability, and resiliency demanded by today's applications, and to keep up with storage growth needs.

This chapter concentrates on the most important topics to consider in IBM FlashSystem administration so that you can use it as a checklist. It also provides best practice tips and guidance. To simplify the Storage Area Network (SAN) storage administration tasks that you use often, such as adding new users, storage allocation and removal, or adding and removing a host from the SAN, create step-by-step, standard procedures for them.

The discussion in this chapter focuses on the IBM FlashSystem 9200 for the sake of simplicity, using screenshots and command outputs from this model. But the recommendations and practices discussed in this chapter are applicable to all the following models:

- ▶ IBM FlashSystem 5010
- ▶ IBM FlashSystem 5030
- ▶ IBM FlashSystem 5100
- ▶ IBM FlashSystem 7200
- ▶ IBM FlashSystem 9100
- ▶ IBM FlashSystem 9200

Note: The practices that are described here were effective in many installations of different models of the IBM FlashSystem family. These installations were performed in various business sectors for a variety of international organizations. They all had one common need, which was to manage their storage environment easily, effectively, and reliably.

This chapter includes the following sections:

- ▶ 10.1, "User interfaces" on page 419

- ▶ 10.2, “Users and groups” on page 421
- ▶ 10.3, “Volumes” on page 423
- ▶ 10.4, “Hosts” on page 424
- ▶ 10.5, “Software updates” on page 425
- ▶ 10.6, “Drive firmware updates” on page 435
- ▶ 10.7, “SAN modifications” on page 437
- ▶ 10.8, “Server HBA replacement” on page 439
- ▶ 10.9, “Hardware upgrades” on page 440
- ▶ 10.10, “I/O Throttling” on page 459
- ▶ 10.11, “Automation” on page 465
- ▶ 10.12, “Documenting IBM FlashSystem and SAN environment” on page 467

These sections provide guidance for keeping your IBM FlashSystem family environment working correctly and reliably.

10.1 User interfaces

The IBM FlashSystem family provides several user interfaces to allow you to maintain your system. The interfaces provide different sets of facilities to help resolve situations that you might encounter. The interfaces for servicing your system connect through the 1 Gbps Ethernet ports that are accessible from port 1 of each canister.

- ▶ Use the management graphical user interface (GUI) to monitor and maintain the configuration of storage that is associated with your clustered systems.
- ▶ Use the service assistant tool GUI to complete service procedures.
- ▶ Use the command line interface (CLI) to manage your system.

The best practice recommendation is to use the interface most appropriate to the task you are attempting to complete. For example, a manual software update is best performed using the service assistant GUI or the CLI. Running fix procedures to resolve problems or configuring expansion enclosures can only be performed using the management GUI. The creation of a large number of volumes with customized names is best performed using the CLI using a script. To ensure efficient storage administration you should 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. Regularly monitor the status of the system 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.

To access the Management GUI, start a supported web browser and go to `https://<flashsystem_ip_address>`, where the `<flashsystem_ip_address>` is the management IP address set when the clustered system is created.

For a more detailed explanation of the task menus and functionality of the Management GUI, see Chapter 4 in the Redbooks publication *Implementing the IBM System Storage SAN Volume Controller with IBM Spectrum Virtualize V8.4*, SG24-8491.

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 inappropriately, the service actions that are available through the service assistant can cause loss of access to data or even data loss.

You connect to the service assistant on one node canister through 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 can also access the service assistant interface by using the technician port.

The service assistant only provides facilities to help you service control enclosures. Always service the expansion enclosures by using the management GUI.

You can also complete the following actions 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 SSH key if a key is not installed and CLI access is required.
- ▶ Restart the services used by the system.

To access the Service Assistant Tool GUI, start a supported web browser and go to: https://<flashsystem_ip_address>/service, where <flashsystem_ip_address> is the service IP address for the node canister or the management IP address for the system on which you want work.

10.1.3 Command line interface

The system CLI is intended for use by advanced users who are confident using a CLI. Up to 32 simultaneous interactive Secure Shell (SSH) sessions to the management IP address are supported.

Nearly all the functionality that is offered by the CLI is available through the management GUI. However, the CLI does not provide the fix procedures that are available in the management GUI. On the other hand, use the CLI when you require a configuration setting that is unavailable in the management GUI.

Typing **he1p** in a CLI will display 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

UNIX command	Description
tr	Translates characters
wc	Counts lines, words and characters in the output
history	Display command history
scp	Secure copy protocol

For more detailed information on command reference and syntax, see:

- ▶ [IBM FlashSystem 9200 Documentation](#)
- ▶ [IBM Spectrum Virtualize for SAN Volume Controller, FlashSystem, and Storwize Family - Command-Line Interface User's Guide](#)

Service command line interface

You also have the option of running service CLI commands on a specific node. To do this, log in to the service IP address of the node that requires servicing.

For more information on using the service command line, see [IBM FlashSystem 9200 Documentation](#).

USB command interface

When a Universal Serial Bus (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 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 on using the USB port, see [IBM FlashSystem 9200 Documentation](#).

Technician port

The technician port is an Ethernet port on the back panel of the IBM FlashSystem product that you can use to configure the node. You can use the technician port to do most of the system configuration operations, which includes the following tasks:

- ▶ Defining a management IP address
- ▶ Initializing a new system
- ▶ Servicing the system

For more information on using the Technician port, see [IBM FlashSystem 9200 Documentation](#).

10.2 Users and groups

Almost all organizations have IT security policies that enforce the use 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 one 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.

IBM FlashSystem supports the following authentication methods:

- ▶ Local authentication using a password
- ▶ Local authentication using SSH keys
- ▶ Remote authentication using Lightweight Directory Access Protocol (LDAP) (Microsoft Active Directory or IBM Security Directory Server)

Local authentication is appropriate for small, single enclosure environments whereas larger environments with multiple clusters and multiple enclosures would benefit from the ease of maintenance achieved by using single sign on (SSO) using remote authentication using LDAP, for example.

By default, the following user groups are defined:

- ▶ **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 is the role that should be assigned to the IBM Storage Insights user. For further detailed information on IBM Storage Insights, see Chapter 9, "Monitoring" on page 363.
- ▶ **Copy Operator:** Users with this role have monitor role privileges and can create, change, and manage all Copy Services functions.
- ▶ **Service:** Users can set the time and date on the system, 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.
- ▶ **Administrator:** Users with this role can access all functions on the system except those that deal with managing users, user groups, and authentication.
- ▶ **Security Administrator:** Users with this role can access all functions on the system, including managing users, user groups, and user authentication.
- ▶ **Restricted Administrator:** Users with this role can complete some tasks, 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.
- ▶ **vStorage Application Programming Interface (API) for Storage Awareness (VASA) Provider:** Users with this role can manage virtual volumes (vVols) that are used by VMware vSphere and managed through Spectrum Control software.

In addition to standard groups, you can also 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. Users within an ownership group can only view or change 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 used by their databases. Their views and privileges in the management GUI are automatically restricted, as shown in Figure 10-1.

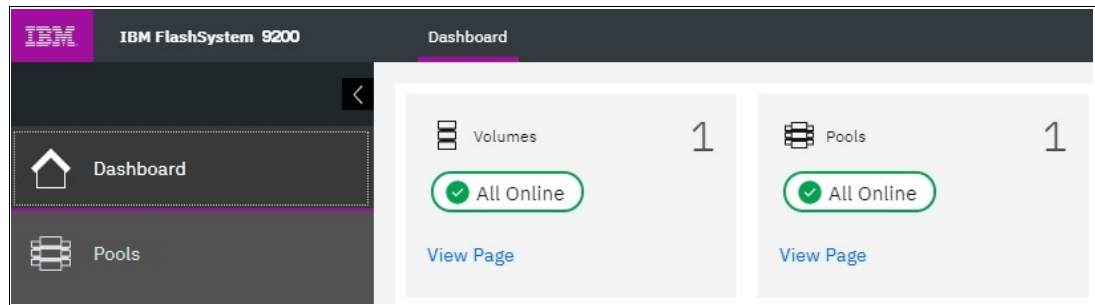


Figure 10-1 Restricted Dashboard view for a user in an ownership group

Regardless of the authentication method 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 only require Monitor group access to perform their daily tasks, whereas 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.

10.3 Volumes

A volume is a logical disk presented to a host by an I/O group (pair of nodes), and within that group there is a preferred node which will serve 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 (VDisks), determine whether your server's operating system supports the particular volume to be extended online. AIX V6.1 TL2 and lower, 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 the IBM FlashSystem. 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 the IBM FlashSystem.
- ▶ Ensure that you delete any volume or LUN definition in the server *before* you unmap it in the IBM FlashSystem. For example, in AIX, remove the hdisk from the volume group (reducevg) and delete the associated hdisk device (rmdev).

- ▶ Consider enabling volume protection by using `chsystem vdiskprotectionenabled yes -vdiskprotectiontime <value_in_minutes>`. Volume protection ensures that some CLI actions (most of those 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*; the system has detected I/O activity to the volume from any host within a specified time period (15 - 1440 minutes).

Note: Volume protection cannot be overridden using the `-force` flag in the affected CLI commands. Volume protection must be disabled to carry on an activity that is currently 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: You must avoid the use of the `-force` parameter in `rmvdisk`.

- ▶ If you issue the `svctask rmvdisk` command and it still has pending mappings, the IBM FlashSystem prompts you to confirm the action and this 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`). The IBM internal Storage Technical Quality Review Process (STQRP) asks for a minimum of a 48-hour period, and having at least a one business day interval so that you can perform a quick backout if you later realize you still need some data on that volume.

For further detailed information on volumes, see Chapter 5, “Volumes” on page 187.

10.4 Hosts

A host is a computer that is connected to the SAN switch through Fibre Channel (FC), iSCSI, and other protocols.

When you add and remove hosts in the IBM FlashSystem, consider the following guidelines:

- ▶ Before you map new servers to the IBM FlashSystem, verify that they are all error free. Fix errors that you find in your server and IBM FlashSystem before you proceed. In the IBM FlashSystem, 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 the IBM FlashSystem, you should list its World Wide Port Name (WWPN) by using the `lshbaportcandidate` command. Use the `lsfabric` command to certify that it was detected by the IBM FlashSystem nodes and ports that you expected. When you create the host definition in the IBM FlashSystem (`mkhost`), try to avoid the `-force` parameter. If you do not see the host’s WWPNs, it might be necessary to scan fabric from the host. For example, use the `cfgmgr` command in AIX.

For further detailed information on hosts, see Chapter 8, “Hosts” on page 349.

10.5 Software updates

Because the IBM FlashSystem might be at the core of your disk and SAN storage environment, its update requires planning, preparation, and verification. However, with the appropriate precautions, an update can be conducted easily and transparently to your servers and applications. This section highlights applicable guidelines for the IBM FlashSystem update.

Most of the following sections explain how to prepare for the software update. These sections also present version-independent guidelines on how to update the IBM FlashSystem family systems and flash drives.

Before you update the system, ensure that the following requirements are met:

- ▶ The latest update test utility was downloaded from IBM Fix Central to your management workstation.
- ▶ The latest system update package was downloaded from IBM Fix Central to your management workstation.
- ▶ All node canisters are online.
- ▶ All errors in the system event log are addressed and marked as fixed.
- ▶ There are no volumes, MDisks, or storage systems with Degraded or Offline status.
- ▶ The service assistant IP is configured on every node in the system.
- ▶ The system superuser password is known.
- ▶ The current system configuration is backed up and saved (preferably off-site). Use the steps described in Example 10-11 on page 474.
- ▶ You have physical access to the hardware.

The following actions are not required, but are suggestions to reduce unnecessary load on the system during the update:

- ▶ Stop all Metro Mirror, Global Mirror, 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.

Note: For customers who have purchased the IBM FlashSystem 9200 with a three-year warranty (9848 Models AG8 and UG8), this comes with Enterprise Class Support (ECS) and this entitles the customer to two code upgrades per year performed by IBM (total of six across the three years of warranty). 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 Systems Service Representative (SSR). A similar optional service is available for the IBM FlashSystem 7200.

For more information on ECS, see [IBM FlashSystem 9200 8.4.0 Documentation - Enterprise Class Support \(ECS\)](#).

10.5.1 Deciding the target software level

The first step is to determine your current and your target IBM FlashSystem software level.

Using the example of an IBM FlashSystem 9200, log in to the web-based GUI and find the current version. Either from the right-hand side of the top menu drop-down line, click on the question mark symbol (?) and select **About IBM FlashSystem 9200** to display the current version or select **Settings** → **System** → **Update System** to display both current and target levels.

Figure 10-2 shows the Update System output panel and displays the code levels. In this example the current software level is 8.4.0.0.

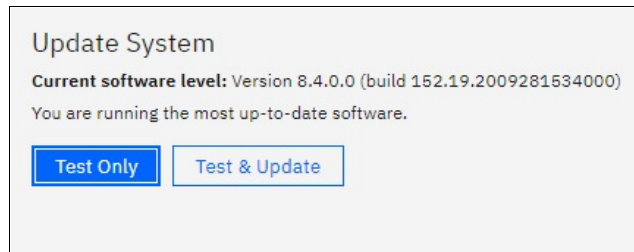


Figure 10-2 Update System output panel

Alternatively, if you are using the CLI, run the `svcinfo 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 lssystem command

```
IBM_FlashSystem:IBM Redbook FS:superuser>lssystem|grep code
code_level 8.4.0.0 (build 152.19.2009281534000)
```

IBM FlashSystem software levels are specified by four digits in the following format:

- ▶ In our example V.R.M.F = 8.4.0.0
 - V is the major version number
 - R is the release level
 - M is the modification level
 - F is the fix level

Use the latest IBM FlashSystem release unless you have a specific reason not to update, such as:

- ▶ The specific version of an application or other component of your SAN Storage environment has a known problem or limitation.
- ▶ The latest IBM FlashSystem 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 use of the “latest release minus 1” or requiring “seasoning” in the field before implementation in a production environment.
- ▶ For more information, see [Spectrum Virtualize Family of Products Upgrade Planning](#).

Obtaining the software packages

To obtain a new release of software for a system update, go to [IBM Fix Central](#) and follow these steps:

1. From the **Product selector** list, type **IBM FlashSystem 9200** (or whatever model is appropriate in your environment).
2. From the **Installed Version** list, select the current software version level determined in the section 10.5.1, “Deciding the target software level” on page 426.
3. Select **Continue**.
4. In the **Product Software** section select the following three items as shown in Figure 10-3:

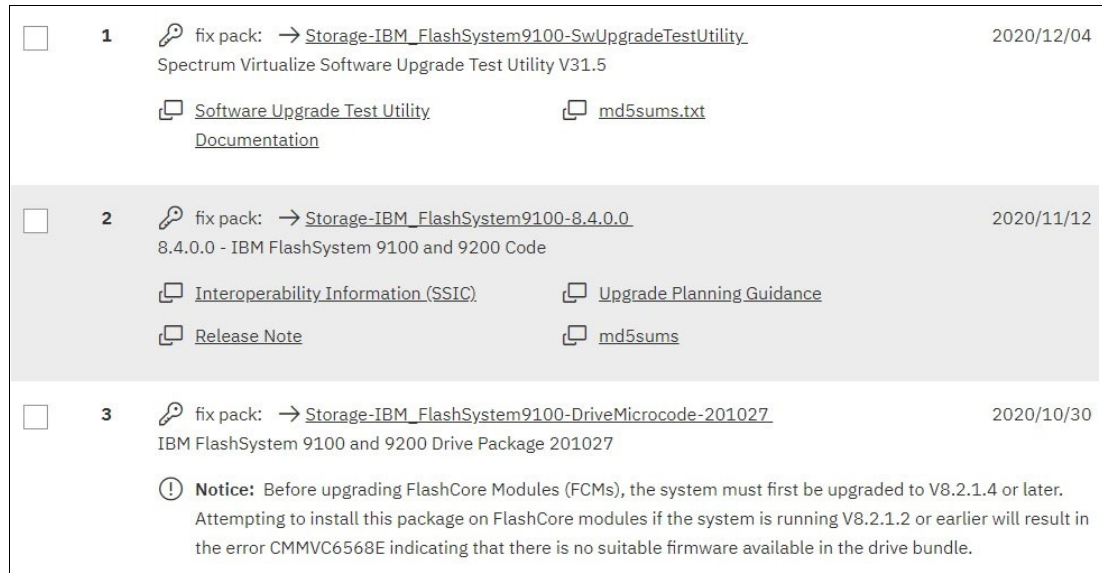


Figure 10-3 Fix Central software packages

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. Select **Continue**.
9. Read and then select **I Agree** to the terms and conditions.
10. Select **Download Now** and save the 3 files onto your management computer.

10.5.2 Hardware considerations

Before you start the update process, always check whether your IBM FlashSystem 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 FlashSystem code first.

10.5.3 Update sequence

Check the compatibility of your target IBM FlashSystem code level with all components of your SAN storage environment (SAN switches, storage controllers, server HBAs) and its attached servers (operating systems and eventually, applications).

Applications often certify only the operating system that they run under and leave to the operating system provider the task of certifying its compatibility with attached components (such as SAN storage). 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 FlashSystem target code level is compatible.

The IBM FlashSystem Supported Hardware List provides the complete information for using your IBM FlashSystem SAN storage environment components with the current and target code level. For links to the Supported Hardware List, Device Driver, Firmware, and Recommended Software Levels for different products and different code levels, see [Support Information for FlashSystem 9200 family](#).

By cross-checking the version of IBM FlashSystem 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.

If you are not making major version or multistep updates in any components, the following update order is less prone to eventual problems:

1. SAN switches or directors
2. Storage controllers
3. Servers HBAs microcode and multipath software
4. IBM FlashSystem
5. IBM FlashSystem internal Non-Volatile Memory express (NVMe) drives
6. IBM FlashSystem Serial Attached SCSI (SAS) attached solid-state drive (SSD)

Attention: Do *not* update two components of your IBM FlashSystem SAN storage environment simultaneously, such as an IBM FlashSystem 9200 and one storage controller. This caution is true even if you intend to do it with your system offline. An update of this type can lead to unpredictable results, and an unexpected problem is more difficult to debug.

10.5.4 SAN fabrics preparation

If you are using symmetrical, redundant, independent SAN fabrics, preparing these fabrics for an IBM FlashSystem 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 FlashSystem environment will continue working through the other fabric and your applications will remain unaffected.

Because you are updating your IBM FlashSystem, 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, independent SAN fabrics, fix this problem as a high priority because it represents a single point of failure.

10.5.5 Storage controllers preparation

As critical as with the attached hosts, the attached storage controllers must correctly handle the failover of MDisk paths. Therefore, they must be running supported microcode versions and their own SAN paths to IBM FlashSystem must be free of errors.

10.5.6 Hosts preparation

If the appropriate precautions are taken, the IBM FlashSystem update is not apparent to the attached servers and their applications. The automated update procedure updates one IBM FlashSystem 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 properly. This capability can be mitigated by enabling NPIV if your current code level supports this function. For more information about N_Port ID Virtualization (NPIV), see Chapter 8, “Hosts” on page 349.

Before you start IBM FlashSystem update preparation, check the following items for every server that is attached to IBM FlashSystem that you update:

- ▶ The operating system type, version, and maintenance or fix level
- ▶ The make, model, and microcode version of the HBAs
- ▶ The multipath software type, version, and error log

For information about troubleshooting, see: [IBM FlashSystem 9200 8.4.0 Documentation - Troubleshooting](#) (requires an IBM ID).

Fix every problem or “suspect” that you find with the disk path failover capability. Because a typical IBM FlashSystem environment can have hundreds of servers 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.

10.5.7 Copy services considerations

When you update an IBM FlashSystem 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 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.

10.5.8 Running the Upgrade Test Utility

It is a requirement that you install and run the latest IBM FlashSystem Upgrade Test Utility before you update the IBM FlashSystem software. For more details, see [Software Upgrade Test Utility](#).

This tool verifies the health of your IBM FlashSystem storage array for the update process. It also checks for unfixed errors, degraded MDisks, inactive fabric connections, configuration conflicts, hardware compatibility, drive firmware, and many other issues that might otherwise require cross-checking a series of command outputs.

Note: The Upgrade Test Utility does not log in to storage controllers or SAN switches. Instead, it reports the status of the connections of the IBM FlashSystem to these devices. It is the users' responsibility to check these components for internal errors.

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 these steps:

1. In the management GUI, select **Settings** → **System** → **Update System**.
2. Click **Test Only**.
3. Select the test utility that you downloaded from the Fix Central support site. Upload the Test utility file and enter the code level you are planning to update to. Figure 10-4 shows the IBM FlashSystem management GUI window, that is used to install and run the Upgrade Test Utility.

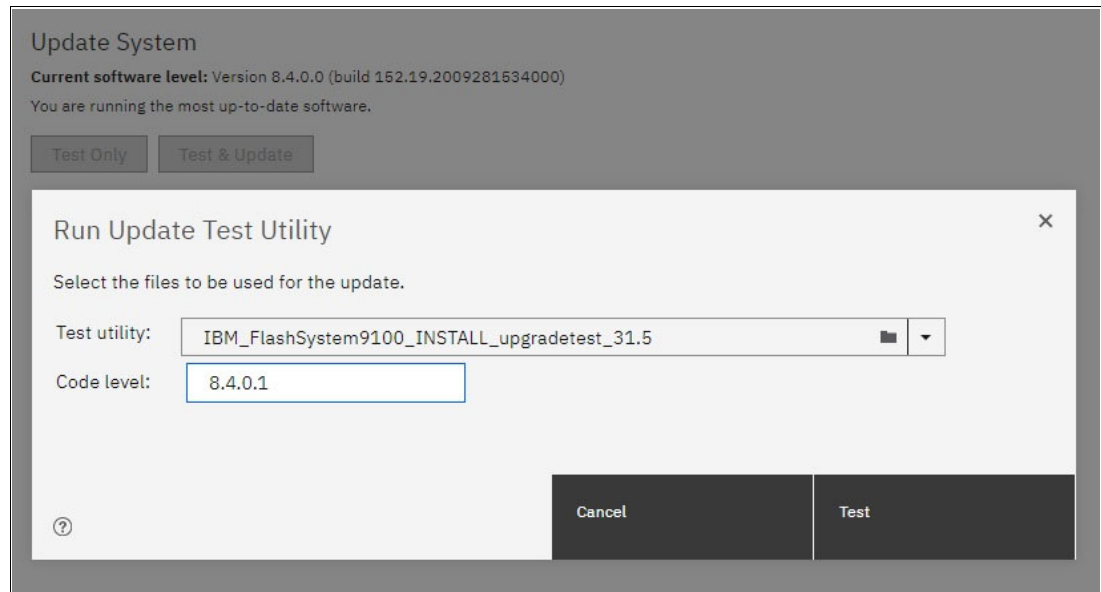


Figure 10-4 IBM FlashSystem Upgrade Test Utility using the GUI

4. Click **Test**. The test utility verifies that the system is ready to be updated. After the Update Test Utility has completed, you will be presented with the results. The results will either state that there have been no warnings or problems found, or will direct you to details about known issues which have been discovered on the system. Figure 10-5 on page 431 shows a successful completion of the update test utility.

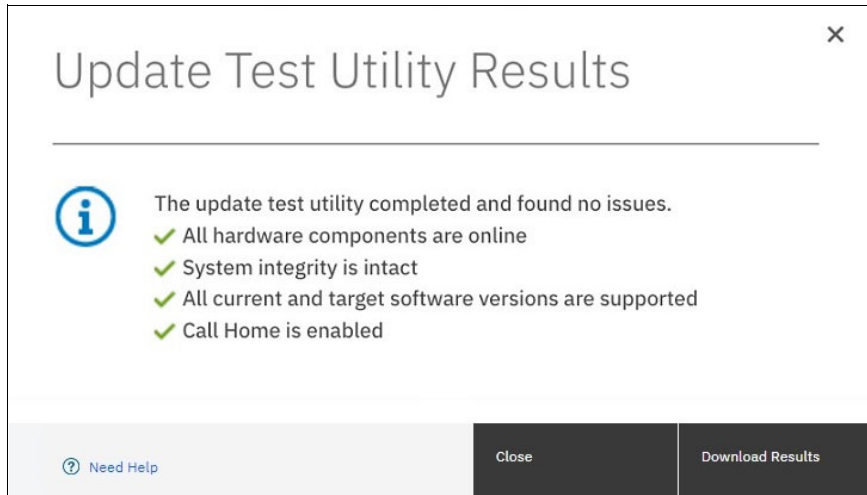


Figure 10-5 IBM FlashSystem Upgrade Test Utility completion panel

5. Click **Download Results** to save the results to a file.
6. Click **Close**.

Using the command line

To test the software on the system, complete these steps:

1. Using OpenSSH scp or PuTTY pscp, copy the software update file and the Software Update Test Utility package to the /home/admin/upgrade directory by using the management IP address of the IBM FlashSystem. Documentation and online help might refer to the /home/admin/update directory, which points to the same location on the system.

An example for the IBM FlashSystem 9200 is shown in Example 10-2.

Example 10-2 Copying the upgrade test utility to IBM FlashSystem 9200

```
C:\>pscp -v -P 22 IBM_FlashSystem9100_INSTALL_upgradetest_31.5
superuser@9.10.11.12:/home/admin/upgrade
Looking up host "9.10.11.12" for SSH connection
Connecting to 9.10.11.12 port 22
We claim version: SSH-2.0-PuTTY_Release_0.74
Remote version: SSH-2.0-OpenSSH_8.0
Using SSH protocol version 2
No GSSAPI security context available
Doing ECDH key exchange with curve Curve25519 and hash SHA-256 (unaccelerated)
Server also has ssh-rsa host key, but we don't know it
Host key fingerprint is:
ecdsa-sha2-nistp521 521 a8:f0:de:cf:eb:fd:b4:74:9e:95:c7:bd:5c:f1:3b:b5
Initialised AES-256 SDCTR (AES-NI accelerated) outbound encryption
Initialised HMAC-SHA-256 (unaccelerated) outbound MAC algorithm
Initialised AES-256 SDCTR (AES-NI accelerated) inbound encryption
Initialised HMAC-SHA-256 (unaccelerated) inbound MAC algorithm
Using username "superuser".
Attempting keyboard-interactive authentication
Keyboard-interactive authentication prompts from server:
| Password:
End of keyboard-interactive prompts from server
Access granted
```

```
Opening main session channel
Opened main channel
Primary command failed; attempting fallback
Started a shell/command
Using SCP1
Connected to 9.10.11.12
Sending file IBM_FlashSystem9100_INSTALL_upgradetest_31.5, size=333865
Sink: C0644 333865 IBM_FlashSystem9100_INSTALL_upgradetest_31.5
IBM_FlashSystem9100_INSTA | 326 kB | 326.0 kB/s | ETA: 00:00:00 | 100%
Session sent command exit status 0
Main session channel closed
All channels closed
C:\>
```

2. Ensure that the update file was successfully copied as shown by the `exit status 0` return code or you can use the `lsdumps -prefix /home/admin/upgrade` command.
3. Install and run Upgrade Test Utility in the CLI, as shown in Example 10-3. In this case, the Upgrade Test Utility found no errors and completed successfully.

Example 10-3 Upgrade test using the CLI

```
IBM_FlashSystem:IBM Redbook FS:superuser>svctask applysoftware -file
IBM_FlashSystem9100_INSTALL_upgradetest_31.5
```

```
CMMVC9001I The package installed successfully.
```

```
IBM_FlashSystem:IBM Redbook FS:superuser>svcupgradetest -v 8.4.0.1
```

```
svcupgradetest version 31.5
```

```
Please wait, the test may take several minutes to complete.
```

```
Results of running svcupgradetest:
```

```
=====
```

```
The tool has found 0 errors and 0 warnings.
```

```
The tool has not found any problems with the cluster.
```

Note: The return code for the `applysoftware` command will always be 1, whether the installation has succeeded or failed. However, the message which is returned when the command completes reports the correct installation result.

Review the output to check whether there have been any problems found by the utility. The output from the command either states that there have been no problems found, or directs you to details about known issues that have been discovered on the system.

10.5.9 Updating the software

There are three methods of updating FlashSystem software:

- ▶ GUI: During a standard update procedure in the management GUI, the system updates each of the nodes systematically. This is the recommended method for updating software on nodes.

- ▶ **CLI:** The command line interface gives you more control over the automatic upgrade process. You have the ability to 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 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. At this point, 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 will become unavailable during the update.

Whichever method (GUI, CLI, or manual) that you choose to perform the update make sure you adhere to the following guidelines for your IBM FlashSystem software update:

- ▶ Schedule the IBM FlashSystem 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, reboot, or reset an IBM FlashSystem node during 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. Bear in mind that the update process can take one hour per node with a further, optional, 30-minute mid-point delay.
- ▶ If you are planning for a major IBM FlashSystem version update, update your current version to its latest fix level *before* you run the major update.
- ▶ Check whether you are running a web browser type and version that is supported by the IBM FlashSystem target software level on every computer that you intend to use to manage your IBM FlashSystem.

This section describes the steps required to update the software.

Using the management GUI

To update the software on the system automatically, complete these 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. 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.
 - Fully 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

To update the software on the system automatically, complete these steps:

1. You must run the latest version of the test utility to verify that no issues exist with the current system. See Example 10-3 on page 432.
2. Copy the software package to the IBM FlashSystem using the same method as described in Example 10-2 on page 431.
3. Before you begin the update, you must be aware of the following situations:
 - The installation process will fail under the following conditions:
 - If the software that is installed on the remote system is not compatible with the new software or if an intersystem 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 rebooting the nodes as part of the update process. You can find details about which volumes would be affected by using the `lsdependentvdisks` 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 thirty-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 as soon as 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 off.
 - 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 filename of the software update file. If the system identifies any volumes that would go offline as a result of rebooting 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 that the update has successfully completed, issue the **lnodecanistervpd** command for each node in the system. The code_level field displays the new code level for each node.

10.6 Drive firmware updates

The updating of drive firmware is concurrent process that can be performed online while the drive is in use, whether it is NVMe or SCM drives in the control enclosure or the SSD drives in any SAS-attached expansion enclosures.

When used on an array member drive the update checks for volumes that are dependent on the drive and refuses to run if any are found. Drive dependent volumes are usually caused by non-redundant or degraded RAID arrays. Where possible you should restore redundancy to the system by replacing any failed drives before upgrading drive firmware. When this 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 Spectrum Virtualize code levels. Therefore, you will find three different NVMe firmware files available for download depending on the size of the drives you have installed.

Using the management GUI

To update the drive firmware automatically, complete these steps:

1. Select Pools → Internal Storage → Actions → Upgrade All.
2. As shown in Figure 10-6 on page 436, in the Upgrade Package text box browse to the drive firmware package you downloaded in section , “Obtaining the software packages” on page 426.
3. Click **Upgrade**. Each drive upgrade will take approximately 6 minutes.

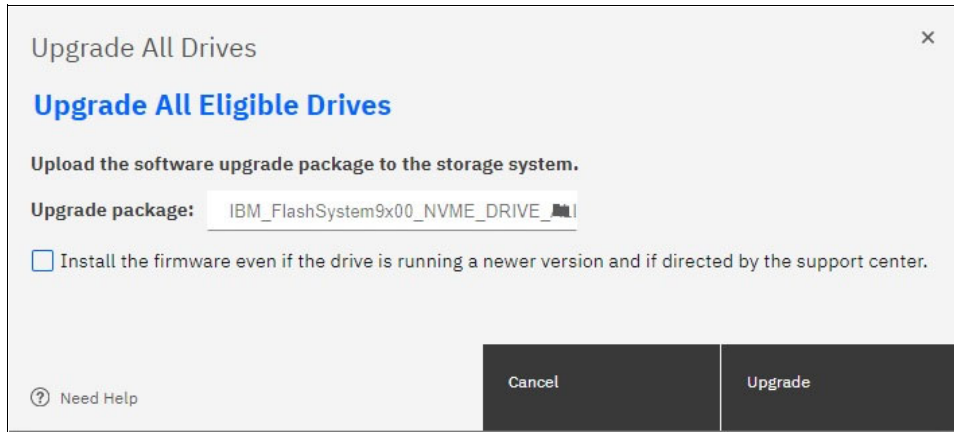


Figure 10-6 Drive firmware upgrade

4. You can also update individual drives by right-clicking on a single drive and selecting **Upgrade**.
5. To monitor the progress of the upgrade, select **Monitoring** → **Background Tasks**.

Using the command line

To update the software on the system manually, complete these steps:

1. Copy the drive firmware package to the IBM FlashSystem using the same method as described in Example 10-2 on page 431.
2. Issue this CLI command to start the update process for all drives:

```
applydrivesoftware -file <software_update_file> -type firmware -all
```

where <software_update_file> is the filename of the software update file. The use of the **-all** option will update 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 the quorum is moved using the **lsquorum** and **chquorum** commands in between **applydrivesoftware** invocations.

Note: The maximum number of drive IDs that can be specified on a command line 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:


```
lsdriveupgradeprogress
```

This command displays success when the update is complete.
4. To verify that the update has successfully completed, issue the **lsdrive** command for each drive in the system. The **firmware_level** field displays the new code level for each drive. Example 10-4 demonstrates how to list the firmware level for four specific drives:

Example 10-4 List firmware level for drives 0, 1, 2 and 3

```
IBM_FlashSystem:GLTL-FS9K:superuser>for i in 0 1 2 3; do echo "Drive $i = `lsdrive $i|grep firmware`"; done
Drive 0 = firmware_level 1_2_11
Drive 1 = firmware_level 1_2_11
Drive 2 = firmware_level 1_2_11
Drive 3 = firmware_level 1_2_11
```

For more detailed information, see: [IBM FlashSystem 9200 8.4.0 Documentation - Updating firmware drive](#).

10.7 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. That error can then affect other servers or applications because appropriate precautions were not taken.

Human error can include 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.

10.7.1 Cross-referencing WWPN

With the WWPN of an HBA, you can uniquely identify one server in the SAN. If a server's name is changed at the operating system level and not at the IBM FlashSystem host definitions, it continues to access its previously mapped volumes exactly because the WWPN of the HBA did not change.

Alternatively, if the HBA of a server is removed and installed in a second server and the first server's SAN zones and IBM FlashSystem host definitions are not updated, the second server can access volumes that it probably should not access.

Complete the following steps to cross-reference HBA WWPNs:

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 SDDPCM, run the `pcmpath query WWPN` command to see output similar to what is shown in Example 10-5.

Example 10-5 Output of the `pcmpath query WWPN` command

```
[root@Server127]> pcmpath query wwpn
Adapter Name PortWWN
fscsi0       10000090FA021A13
fscsi1       10000090FA021A12
```

If you are using server virtualization, verify the World Wide Port Names (WWPNs) in the server that is attached to the SAN, such as AIX Virtual Input/Output (VIO) or VMware ESX. Cross-reference with the output of the IBM FlashSystem `lshost <hostname>` command, as shown in Example 10-6.

Example 10-6 Output of the `lshost <hostname>` command

```
IBM_FlashSystem:IBM Redbook FS:superuser>svcinfo lshost Server127
id 0
name Server127
port_count 2
type generic
mask 1111111111111111111111111111111111111111111111111111111111111111
iogrp_count 4
status active
site_id
site_name
host_cluster_id
host_cluster_name
protocol scsi
WWPN 1000090FA021A13
node_logged_in_count 1
state active
WWPN 1000090FA021A12
node_logged_in_count 1
state active
```

2. If necessary, cross-reference information with your SAN switches, as shown in Example 10-7. In Brocade switches use the `nodefind <WWPN>` command.

Example 10-7 Cross-referencing information with SAN switches

```
blg32sw1_B64:admin> nodefind 10:00:00:90:FA:02:1A:13
Local:
  Type Pid    COS     PortName                                     NodeName                                     SCR
  N   401000;    2,3;10:00:00:90:FA:02:1A:13;20:00:00:90:FA:02:1A:13; 3
      Fabric Port Name: 20:10:00:05:1e:04:16:a9
      Permanent Port Name: 10:00:00:90:FA:02:1A:13
      Device type: Physical Unknown(initiator/target)
      Port Index: 16
      Share Area: No
      Device Shared in Other AD: No
      Redirect: No
      Partial: No
      Aliases: nybixtdb02_fcs0
b32sw1_B64:admin>
```

For storage allocation requests that are submitted by the server support team or application support team to the storage administration team, always include the server's HBA WWPNs to which the new LUNs or volumes are supposed to be mapped. For example, a server might use separate HBAs for disk and tape access or distribute its mapped LUNs across different HBAs for performance. You cannot assume that any new volume is supposed to be mapped to every WWPN that server logged in the SAN.

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.

10.7.2 Cross-referencing LUN ID

Always cross-reference the IBM FlashSystem `vdisk_UID` with the server logical unit number (LUN) ID before you perform any modifications that involve IBM FlashSystem volumes. Example 10-8 shows an AIX server that is running Subsystem Device Driver Path Control Module (SDDPCM). The IBM FlashSystem `vdisk_name` has no relation to the AIX device name. Also, the first SAN LUN mapped to the server (`SCSI_id=0`) shows up as `hdisk4` in the server because it had four internal disks (`hdisk0 - hdisk3`).

Example 10-8 Results of running the `lshostvdiskmap` command

```
IBM_FlashSystem:IBM Redbook FS:superuser>lshostvdiskmap NYBIXTDB03
id name          SCSI_id vdisk_id vdisk_name      vdisk_UID
0 NYBIXTDB03 0          0          NYBIXTDB03_T01 60050768018205E12000000000000000
```

```
root@nybixtdb03:~/> pcmpath query device
Total Dual Active and Active/Asymmetric Devices : 1
DEV#: 4 DEVICE NAME: hdisk4 TYPE: 2145 ALGORITHM: Load Balance
SERIAL: 60050768018205E12000000000000000
```

```
=====
Path#      Adapter/Path Name      State   Mode     Select   Errors
0*         fscsi0/path0           OPEN    NORMAL    7         0
1          fscsi0/path1           OPEN    NORMAL   5597      0
2*         fscsi2/path2           OPEN    NORMAL    8         0
3          fscsi2/path3           OPEN    NORMAL   5890      0
=====
```

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 reclaim.

Note: Because a host can have many volumes with the same `scsi_id`, always cross-reference the IBM FlashSystem volume UID with the host volume UID and record the `scsi_id` and LUN ID of that volume.

10.8 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).

Complete the following steps to replace a failed HBA and retain the working HBA:

1. In your server, identify the failed HBA and record its WWPNs. For more information, see 10.7.1, “Cross-referencing WWPN” on page 437. Then, place this HBA and its associated paths offline, gracefully if possible. This approach is important so that the multipath software stops trying to recover it. 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 into 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 you noted in step 1, and make sure that you did not record the wrong 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 the IBM FlashSystem by using the **lsfcportcandidate** 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 FlashSystem host definition by using the **addhostport** command. It is important that you do not remove the old one yet. Run the **lshost <servername>** command. Then, verify that the working HBA shows as *active*, while the failed HBA should show 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 the IBM FlashSystem 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.

10.9 Hardware upgrades

The IBM FlashSystem scalability features allow significant flexibility in its configuration. As discussed in previous chapters, the IBM FlashSystem family has two different types of enclosures: control enclosures and expansion enclosures.

- ▶ Control Enclosures manage your storage systems, communicate with the host, and manage interfaces. In addition, they can also house up to 24 NVMe-capable flash drives.
- ▶ Expansion Enclosures increase the available capacity of an IBM FlashSystem 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 of flash (solid-state drive (SSD)) SAS type drives,

A basic configuration of an IBM FlashSystem storage platform consists of one IBM FlashSystem control enclosure. For a balanced increase of performance and scale, up to four (depending on model) IBM FlashSystem 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. Consequently, several scenarios are possible for its growth. The following sections describe these processes in more detail:

- ▶ “Adding control enclosures” on page 441
- ▶ “Upgrading nodes in an existing cluster” on page 444
- ▶ “Upgrading NVMe drives” on page 450
- ▶ “Moving to a new IBM FlashSystem cluster” on page 450
- ▶ “Splitting an IBM FlashSystem cluster” on page 451

10.9.1 Adding control enclosures

If your existing IBM FlashSystem 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 an existing cluster of IBM Storwize V7000 nodes that you want to add the IBM FlashSystem enclosures to, since the latter are more powerful than your existing ones. Therefore, your cluster will have different node models in different I/O groups.

To install these control enclosures, determine whether you need to upgrade your IBM FlashSystem first (or Storwize V7000 code level if you are merging an existing Storwize V7000 Gen2 cluster with a IBM FlashSystem 9200 for example). For more information, see 10.5.2, “Hardware considerations” on page 427.

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.

IBM FlashSystem 9200

To add a control enclosure to an existing FlashSystem 9200 system, the IBM SSR engineer must first install the new control enclosure in the rack and cable it to SAN or Ethernet switches or directly to the existing control enclosure. You are then able to add it to the system using the management GUI where it should automatically appear if cabled correctly. For more details, see [IBM FlashSystem 9200 8.4.0 Documentation - Adding a control enclosure to an existing system](#).

IBM FlashSystem 9100

To add a control enclosure to an existing FlashSystem 9100 system, the IBM SSR engineer must first install the new control enclosure in the rack and cable it to SAN or Ethernet switches or directly to the existing control enclosure. You are then able to add it to the system using the management GUI where it should automatically appear if cabled correctly. For more details, see [IBM FlashSystem 9100 8.4.0 Documentation - Adding a control enclosure to an existing system](#).

IBM FlashSystem 7200

To add a control enclosure to an existing FlashSystem 7200 system, you must first install it in the rack. Then, you must connect it to the system through a zone in the SAN or by using RDMA over Ethernet. Finally, you can add it to the system using the management GUI where it should automatically appear if cabled correctly. For more details, see [IBM FlashSystem 7200 8.4.0 Documentation - Adding an expansion enclosure to an existing system](#).

state inactive

```
IBM_FlashSystem:IBM Redbook FS:superuser>lsiogrp
id name          node_count vdisk_count host_count site_id site_name
0 io_grp0        2          11          2          2
1 io_grp1        0          0           2          2
2 io_grp2        0          0           2          2
3 io_grp3        0          0           2          2
4 recovery_io_grp 0          0           0          0
```

```
?IBM_FlashSystem:IBM Redbook FS:superuser>lshostiogrp Win2012srv1
```

```
id name
0 io_grp0
1 io_grp1
2 io_grp2
3 io_grp3
```

```
IBM_FlashSystem:IBM Redbook FS:superuser>rmhostiogrp -iogrp 3 Win2012srv1
IBM_FlashSystem:IBM Redbook FS:superuser>
```

```
IBM_FlashSystem:IBM Redbook FS:superuser>lshostiogrp Win2012srv1
```

```
id name
0 io_grp0
1 io_grp1
2 io_grp2
```

```
IBM_FlashSystem:IBM Redbook FS:superuser>
```

```
IBM_FlashSystem:IBM Redbook FS:superuser>addhostiogrp -iogrp io_grp3
Win2012srv1
```

```
IBM_FlashSystem:IBM Redbook FS:superuser>
```

```
IBM_FlashSystem:IBM Redbook FS:superuser>lshostiogrp Win2012srv1
```

```
id name
0 io_grp0
1 io_grp1
2 io_grp2
3 io_grp3
```

```
IBM_FlashSystem:IBM Redbook FS:superuser>lsiogrp
id name          node_count vdisk_count host_count site_id site_name
0 io_grp0        2          11          2          2
1 io_grp1        0          0           2          2
2 io_grp2        0          0           2          2
3 io_grp3        0          0           2          2
4 recovery_io_grp 0          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 of time. 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.

10.9.2 Upgrading nodes in an existing cluster

If you want to upgrade the nodes or canisters of your existing IBM FlashSystem, there is the option to increase the cache memory size and/or the adapter cards in each node. This can be done, one node at a time, and so as to be non-disruptive to the systems operations. For further details, see [IBM FlashSystem 9200 8.4.0 Documentation - Removing and replacing a memory modul](#).

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 switched virtual circuit (SVC)), a large cache can accelerate older controllers with smaller caches.
- ▶ If you're using DRP, then maximize the cache size and consider adding SCM drives with Easy Tier for the best performance.
- ▶ If you're making heavy use of copy services, consider increasing the cache beyond just your working set requirements.
- ▶ Finally, bear in mind that a truly random working set may 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 a node from the system so that it is not showing in the management GUI or in the output from the `l snode` 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 card upgrades consider the following points:

- ▶ A single 32 Gb Fibre Channel port can deliver over 3 GB/s (allowing for overheads).
- ▶ A 32 Gb FC card in each canister, with 8 ports can deliver more than 24 GB/s.
- ▶ An FCM NVMe device can perform at over 1 GB/s.
- ▶ A single 32 Gb Fibre Channel port can deliver 80,000 to 125,000 IOPS with a 4k block size.
- ▶ A 32 Gb FC card in each canister, with 8 ports can deliver up to 1,000,000 IOPS.
- ▶ A FlashSystem 9200 can deliver 1,200,000 4k read miss Input/Output Operations Per Second (IOPS) and up to 4,500,000 4k read hit IOPS.
- ▶ If you have more than 12 NVMe devices, consider 2 Fibre Channel cards per canister. A third Fibre Channel card will allow you to achieve up to 45 GB/s.
- ▶ If you want to achieve more than 800,000 IOPS, use at least 2 Fibre Channel cards per canister.
- ▶ If the FlashSystem is performing Remote Copy or clustering, consider using separate ports to ensure there is no conflict with host traffic.

- ▶ iSER via 25 GbE ports has similar capabilities as 16 Gb FC ports, but with less overall ports available. If you're planning on using 10 Gb iSCSI, ensure it can service your expected workloads.

Real-time performance statistics are available in the management GUI from the **Monitoring** → **Performance** menu, as shown in Figure 10-7.

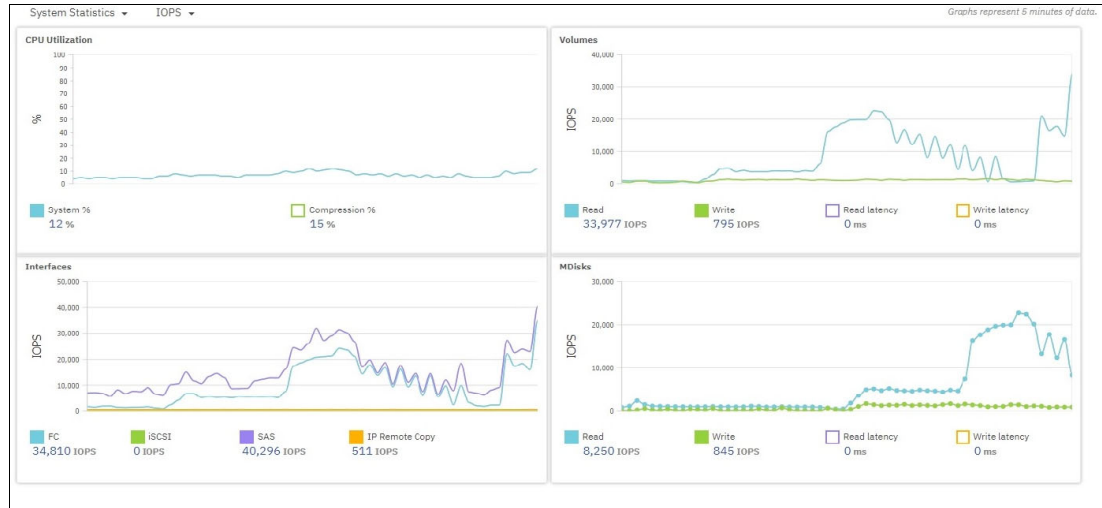


Figure 10-7 IBM FlashSystem performance statistics (IOPS)

Memory options for an IBM FlashSystem 9200 control enclosure

A CPU processor has six memory channels, which are labeled A-F. Each memory channel has two dual in-line memory module (DIMM) slots, numbered 0-1. For example, DIMM slots A0 and A1 are in memory channel A.

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 shows the available memory configuration for each FlashSystem 9200 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 configuration for one node in a control enclosure

DIMM Slot	Total enclosure memory 256 GB	Total enclosure memory 768 GB	Total enclosure memory 1536 GB
F0 (CPU0)	Blank	32 GB	32 GB
F1 (CPU0)	Blank	Blank	32 GB
E0 (CPU0)	Blank	32 GB	32 GB

DIMM Slot	Total enclosure memory 256 GB	Total enclosure memory 768 GB	Total enclosure memory 1536 GB
E1 (CPU0)	Blank	Blank	32 GB
D0 (CPU0)	32 GB	32 GB	32 GB
D1 (CPU0)	Blank	Blank	32 GB
A1 (CPU0)	Blank	Blank	32 GB
A0 (CPU0)	32 GB	32 GB	32 GB
B1 (CPU0)	Blank	Blank	32 GB
B0 (CPU0)	Blank	32 GB	32 GB
C1 (CPU0)	Blank	Blank	32 GB
C0 (CPU0)	Blank	32 GB	32 GB
C0 (CPU1)	Blank	32 GB	32 GB
C1 (CPU1)	Blank	Blank	32 GB
B0 (CPU1)	Blank	32 GB	32 GB
B1 (CPU1)	Blank	Blank	32 GB
A0 (CPU1)	32 GB	32 GB	32 GB
A1 (CPU1)	Blank	Blank	32 GB
D1 (CPU1)	Blank	Blank	32 GB
D0 (CPU1)	32 GB	32 GB	32 GB
E1 (CPU1)	Blank	Blank	32 GB
E0 (CPU1)	Blank	32 GB	32 GB
F1 (CPU1)	Blank	Blank	32 GB
F0 (CPU1)	Blank	32 GB	32 GB

Memory options for an IBM FlashSystem 9100 control enclosure

Each of the six memory channels in each CPU has two DIMM slots, for a total of 12 DIMM slots per CPU, which means 24 DIMM slots per node canister and 48 DIMM slots per enclosure. You can install six distinct memory configurations in those 24 DIMM slots in each node canister. (Each canister must have the same amount of memory and the same configuration).

Initially, each control enclosure ships with one of the following features, depending on what has been ordered, as shown in Table 10-3.

Table 10-3 Base memory features

Feature	Memory per enclosure	Maximum per enclosure
ACG0	128 GB base cache memory (eight 16 GB DIMMs - 2 per CPU)	1

Feature	Memory per enclosure	Maximum per enclosure
ACG1	768 GB base cache memory (twenty-four 32 GB DIMMs - 6 per CPU)	1

You can order the following features to upgrade to more memory at any time. Table 10-4 shows the various options.

Table 10-4 Additional memory features

Feature	Memory per enclosure	Maximum per enclosure
ACGA	128 GB memory upgrade (eight 16 GB DIMMs)	3
ACGB	768 GB memory upgrade (twenty-four 32 GB DIMMs)	2

Memory options for an IBM FlashSystem 7200 control enclosure

Table 10-5 lists the various memory options available for the IBM FlashSystem 7200 by feature code.

Table 10-5 IBM FlashSystem 7200 memory options

Base memory (GB)	Field Upgrade ACGJ (GB)	Field Upgrade ACGB (GB)	Total memory (GB)
256	N/A	N/A	256
256	512	N/A	768
256	512	768	1536

Memory options for an IBM FlashSystem 5000 control enclosure

The IBM FlashSystem 5000 family consists of different models, and each model type provides a different set of features. Table 10-6 shows the memory feature on the FlashSystem 5000 and 5100 models.

Table 10-6 Memory options

Platform	FS5010	FS5030	FS5100
Option 1 per node	1 x 8 GB	1 x 16 GB	2 x 16 GB
Option 2 per node	1 x 16 GB	2 x 16 GB	6 x 16 GB
Option 3 per node	2 x 16 GB	N/A	6 x 16 GB + 6 x 32 GB
Maximum per I/O Group	64 GB	64 GB	576 GB

Adapter card options for an IBM FlashSystem 9200 control enclosure

You can also add new adapter cards to the IBM FlashSystem 9200 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 9200 control enclosure. Each canister has three PCIe adapter slots 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 16 Gb FC ports or up to twelve 25 Gb Ethernet (iSCSI or iSCSI Extensions for Remote Direct Memory Access (RDMA) (iSER) capable) ports. The control enclosure also includes eight 10 Gb Ethernet ports as standard for iSCSI connectivity and two 1 Gb Ethernet 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. The options for the features available are shown in Table 10-7.

Table 10-7 IBM FlashSystem 9200 control enclosure adapter card options

Number of control enclosures	16 Gbps/32 Gbps FC	Onboard iSCSI	25 Gbps iSCSI (RoCE)	25 Gbps iSCSI (iWARP)
1	24	8	12	12
2	48	16	24	24
3	72	24	36	36
4	96	32	48	48

For more information on the feature codes, memory options, and functions of each adapter, see: *IBM FlashSystem 9200 Product Guide*, REDP-5586.

Adapter card options for an IBM FlashSystem 9100 control enclosure

You can also add new adapter cards to the IBM FlashSystem 9100 nodes. These adapters are added as a pair (one card in each node) and the options for the features available are shown in Table 10-8.

Table 10-8 IBM FlashSystem 9100 control enclosure adapter card options

Number of Cards	Ports	Protocol	Possible Slots	Comments
0 - 3	4	16 Gb Fibre Channel	1, 2, 3	
0 - 3	2	25 Gb Ethernet (iWarp)	1, 2, 3	
0 - 3	2	25 Gb Ethernet (RoCE)	1, 2, 3	
0 - 1	2 - see comment	12 Gb SAS Expansion	1, 2, 3	Card is 4 port with only 2 ports active (ports 1 and 3)

Further details of the feature codes, memory options, and functions of each adapter can be found in the Planning chapter of the following IBM Redbooks publication: “*IBM FlashSystem 9100 Architecture, Performance and Implementation*”, SG24-8425.

Adapter card options for an IBM FlashSystem 7200 control enclosure

Six PCIe slots are available for port expansions in the IBM FlashSystem 7200 control enclosure. Each canister has three PCIe adapter slots 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 IBM FlashSystem 7200 control enclosure can be configured with three I/O adapter features to provide up to twenty-four 16 Gb FC ports or up to twelve 25 Gb Ethernet (iSCSI or iSER-capable) ports. The control enclosure also includes eight 10 Gb Ethernet ports as standard for iSCSI connectivity and two 1 Gb Ethernet ports for system management. A feature code also is available to include the SAS Expansion adapter if the user wants to implement the optional expansion enclosures.

Adapter card options for an IBM FlashSystem 5000 control enclosure

All of the FlashSystem 5000 control enclosures include 1 Gb Ethernet (GbE) or 10 GbE ports as standard for iSCSI connectivity. The standard connectivity can be extended with additional ports or enhanced with additional connectivity through an optional I/O adapter card feature. Table 10-9 shows which configurations are available for the FlashSystem 5000 and 5100.

Table 10-9 IBM FlashSystem 5000 family configurations

Platform	FS5010	FS5030	FS5100
iSCSI	1 x 1 GbE tech port + iSCSI 1 x 1 GbE iSCSI only	1 x 1 GbE dedicated tech port	1 x 1 GbE dedicated tech port
iSCSI	N/A	2 x 10 GbE (iSCSI only)	4 x 10 GbE (iSCSI only)
SAS	1 x 12 Gb SAS expansion	2 x 12 Gb SAS expansion	N/A

Table 10-10 shows the possible adapter card installation for the FlashSystem 5000 and 5100. Only one interface card can be installed per canister and the interface card must be the same in both canisters.

Table 10-10 IBM FlashSystem 5000 family adapter cards

Platform	FS5010	FS5030	FS5100
FC	4-port 16 Gb Fibre Channel or	4-port 16 Gb Fibre Channel or	4-port 16 Gb Fibre Channel, FC NVMeoF or
iSCSI	4-port 10 GbE iSCSI or	4-port 10 GbE iSCSI or	2-port 25 GbE ROCE ISER, iSCSI or
iSCSI	2-port 25 GbE iSCSI or	2-port 25 GbE iSCSI or	2-port 25 GbE iWARP ISER, iSCSI and
SAS	4-port 12 Gb SAS host attach	4-port 12 Gb SAS host attach	2-port 12 Gb SAS to allow SAS expansions

10.9.3 Upgrading NVMe drives

To provide ultra-low latency for performance sensitive but less cache-friendly workloads, storage-class memory (SCM) drives from Intel and Samsung are available as a persistent storage tier for IBM FlashSystem 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. This means 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 using Easy Tier to accelerate much larger workloads. Like the FlashCore Modules, SCM drives are also available as upgrades for the previous generation of all flash arrays.

Spectrum Virtualize V8.4 supports up to 12 SCM drives in a control enclosure for IBM FlashSystem 9000, 7000 and 5100 families.

For more information, see [IBM FlashSystem 9200 8.4.0 Documentation - Removing and replacing a memory module](#).

10.9.4 Moving to a new IBM FlashSystem cluster

You might have a highly populated, intensively used IBM FlashSystem cluster that you want to upgrade. You might also want to use the opportunity to refresh your IBM FlashSystem and SAN storage environment.

Complete the following steps to replace your cluster entirely with a newer, more powerful one:

1. Install your new IBM FlashSystem cluster.
2. Create a replica of your data in your new cluster.
3. Migrate your servers to the new IBM FlashSystem cluster when convenient.

If your servers can tolerate a brief, scheduled outage to switch from one IBM FlashSystem cluster to another, you can use the IBM FlashSystem Remote Copy services (Metro Mirror or Global Mirror) to create your data replicas, following these steps:

1. Select a host that you want to move to the new IBM FlashSystem cluster and find all the old volumes you must move.
2. Zone your host to the new IBM FlashSystem cluster.
3. Create Remote Copy relationships from the old volumes in the old IBM FlashSystem cluster to new volumes in the new IBM FlashSystem cluster.
4. Map the new volumes from the new IBM FlashSystem cluster to the host.
5. Discover new volumes on the host.
6. Stop all I/O from the host to the old volumes from the old IBM FlashSystem cluster.
7. Disconnect and remove the old volumes on the host from the old IBM FlashSystem cluster.
8. Unmap the old volumes from the old IBM FlashSystem cluster to the host.
9. Make sure Remote Copy relationships between old and new volumes in the old and new IBM FlashSystem cluster are synced.
10. Stop and remove Remote Copy relationships between old and new volumes so that the target volumes in the new IBM FlashSystem cluster receive read/write access.
11. Import data from the new volumes and start your applications on the host.

If you must migrate a server online, instead, you must use host-based mirroring by completing these steps:

1. Select a host that you want to move to the new IBM FlashSystem cluster and find all the old volumes that you must move.
2. Zone your host to the new IBM FlashSystem cluster.
3. Create volumes in the new IBM FlashSystem cluster of the same size as the old volumes in the old IBM FlashSystem cluster.
4. Map the new volumes from the new IBM FlashSystem cluster to the host.
5. Discover new volumes on the host.
6. For each old volume, use host-based mirroring (such as AIX `mirrorvg`) to move your data to the corresponding new volume.
7. For each old volume, after the mirroring is complete, remove the old volume from the mirroring group.
8. Disconnect and remove the old volumes on the host from the old IBM FlashSystem cluster.
9. Unmap the old volumes from the old IBM FlashSystem cluster to the host.

This approach uses the server's computing resources (CPU, memory, and I/O) to replicate the data. It can be done online if properly planned. Before you begin, make sure it has enough spare resources.

The biggest benefit to using either approach is that they easily accommodate (if necessary) the replacement of your SAN switches or your back-end storage controllers. You can upgrade the capacity of your back-end storage controllers or replace them entirely, as you can replace your SAN switches with bigger or faster ones. However, you do need to have spare resources, such as floor space, power, cables, and storage capacity, available during the migration.

10.9.5 Splitting an IBM FlashSystem cluster

Splitting an IBM FlashSystem 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 FlashSystem that handles all I/O operations, hosts, and subsystem storage attachments. The goal is to create a second IBM FlashSystem 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 FlashSystem 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 workload on the new system.
- ▶ Migrate the workload onto the new system by using the approach described in 10.9.4, "Moving to a new IBM FlashSystem cluster" on page 450.

10.9.6 Adding expansion enclosures

As time passes and your environment grows, you will need to add more storage to your system. Depending on the IBM FlashSystem family product and the code level that you have installed, you can add different numbers of expansion enclosures to your system. Before you add an enclosure to a system, check that the licensed functions of the system support the additional enclosure.

Because all IBM FlashSystem models were designed to make managing and maintaining them as simple as possible, adding an expansion enclosure is an easy task.

IBM FlashSystem 9200

Currently, IBM offers the following SAS expansion enclosures that can be attached to the IBM FlashSystem 9200. Each node can support 10 SAS Connections thus a control enclosure can support up to 20 expansion enclosures.

Note: To support SAS expansion enclosures, an AHBA - SAS Enclosure Attach adapter card must be installed in each node canister of the IBM FlashSystem 9200 control enclosure.

The following types of expansion enclosures are available:

- ▶ IBM FlashSystem 9000 LFF Expansion Enclosure Model A9F
- ▶ IBM FlashSystem 9000 SFF Expansion Enclosure Model AFF

The new IBM FlashSystem 9200 SFF expansion enclosure Model AFF offers new tiering options with solid-state drive (SSD flash drives). Up to 480 drives of serial-attached SCSI (SAS) expansions are supported per IBM FlashSystem 9200 control enclosure. The expansion enclosure is 2U high.

The new IBM FlashSystem 9200 LFF expansion enclosure Model A9F offers new tiering options with solid-state drive (SSD flash drives). Up to 736 drives of serial-attached SCSI (SAS) expansions are supported per IBM FlashSystem 9200 control enclosure. The expansion enclosure is 5U high.

The best practice recommendation is to balance equally the expansion enclosures between chains. So, if you have two additional expansion enclosures one should be installed on the first SAS chain and one on the second SAS chain. In addition, when you add a single expansion enclosure to an existing system, it is preferable to add the enclosure directly below the control enclosure. When you add a second expansion enclosure, it is preferable to add the enclosure directly above the control enclosure. As more expansion enclosures are added, alternate adding them above and below.

The IBM FlashSystem 9200 system supports up to four control enclosures and up to two chains of SAS expansion enclosures per control enclosure. To limit contention for bandwidth on a chain of SAS enclosures, no more than ten expansion enclosures can be chained to SAS port 1 of a node canister and no more than ten expansion enclosures can be chained to SAS port 3 of a node canister. On each SAS chain, the systems can support up to a SAS chain weight of ten, where:

- ▶ Each 9846-A9F or 9848-A9F expansion enclosure adds a value of 2.5 to the SAS chain weight.
- ▶ Each 9846-AFF or 9848-AFF expansion enclosure adds a value of 1 to the SAS chain weight.

For example, each of the following expansion enclosure configurations has a total SAS weight of ten:

- ▶ Four 9848-A9F enclosures per SAS chain
- ▶ Two 9846-A9F enclosures and five 9848-AFF enclosures per SAS chain

Figure 10-8 shows the cabling for adding two A9F expansion enclosures and two AFF expansion enclosures to a single control enclosure (in the center of Figure 10-8).

For more detailed information, see [IBM FlashSystem 9200 8.4.0 Documentation - Adding a node or enclosure to increase the size of the system.](#)

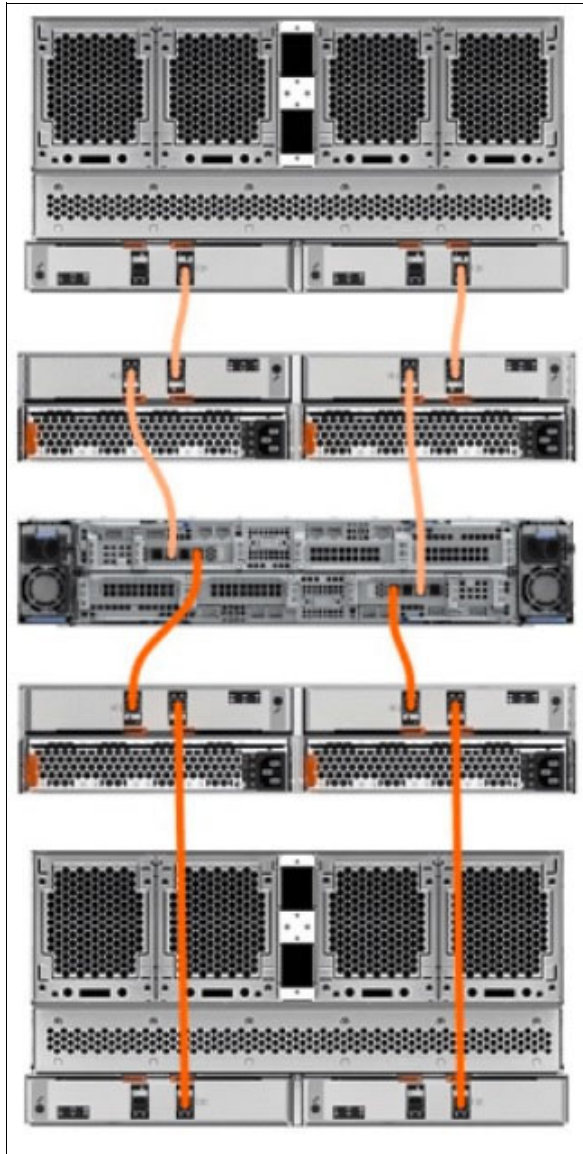


Figure 10-8 Cabling for adding four expansion enclosures in two SAS chains

Adding expansion enclosures is simplified because IBM FlashSystem 9200 can automatically discover new expansion enclosures after the SAS cables are connected. It is possible to manage and use the new drives without managing the new expansion enclosures. However, unmanaged expansion enclosures are not monitored properly. This issue can lead to more

difficult troubleshooting and can make problem resolution take longer. To avoid this situation, always manage newly added expansion enclosures and follow these guidelines:

- ▶ FlashSystem 9200 systems support 4-port SAS interface adapters. However, only ports 1 and 3 are used for SAS connections.
- ▶ Connect SAS port 1 of the upper node canister in the control enclosure to SAS port 1 of the left expansion canister in the first expansion enclosure.
- ▶ Connect SAS port 1 of the lower node canister in the control enclosure to SAS port 1 of the right expansion canister in the first expansion enclosure.
- ▶ In general, the SAS interface adapter must be installed in PCIe slot 3 of the node canister.
- ▶ No cable can be connected between a port on a left canister and a port on a right canister.
- ▶ A cable must not be connected between ports in the same enclosure.
- ▶ A connected port on the node canister must connect to a single port on an expansion canister. Cables that split the connector out into separate physical connections are not supported.
- ▶ Attach cables serially between enclosures; do not skip an enclosure.
- ▶ The last expansion enclosure in a chain must not have cables in port 2 of canister 1 or port 2 of canister 2.
- ▶ Ensure that cables are installed in an orderly way to reduce the risk of cable damage when replaceable units are removed or inserted.

IBM FlashSystem 9100

The procedure for adding expansion enclosures to an IBM FlashSystem 9100 control enclosure is similar to that described in section “IBM FlashSystem 9200” on page 452.

For more detailed information, see [IBM FlashSystem 9100 8.4.0 Documentation](#).

IBM FlashSystem 7200

The following types of expansion enclosures are available:

- ▶ IBM FlashSystem 7200 LFF Expansion Enclosure Model 12G
- ▶ IBM FlashSystem 7200 SFF Expansion Enclosure Model 24G
- ▶ IBM FlashSystem 7200 Dense Expansion Enclosure Model 92G

When attaching expansion enclosures to the control enclosure, you are not limited by the type of the enclosure (if it meets all generation level restrictions). The only limitation for each SAS chain is its chain weight. Each type of enclosure defines its own chain weight, as follows:

- ▶ Enclosures 12G and 24G have a chain weight of 1.
- ▶ Enclosure 92G has a chain weight of 2.5.
- ▶ The maximum chain weight for any SAS chain is 10.
- ▶ The maximum number of SAS chains per control enclosure is 2.

For example, you can combine seven 24G and one 92G expansion enclosures ($7 \times 1 + 1 \times 2.5 = 9.5$ chain weight), or two 92G enclosures, one 12G, and four 24G ($2 \times 2.5 + 1 \times 1 + 4 \times 1 = 10$ chain weight).

You can use either the **addcontrolenclosure** command or the Add Enclosure wizard in the management GUI to add the new expansion enclosure to the system.

To access the Add Enclosure wizard, select **Monitoring** → **System Hardware**. On the **System Hardware - Overview** page, select **Add Enclosure** to start the wizard. Complete the wizard and verify the new enclosure. If **Add Enclosure** is not displayed, it indicates a potential cabling issue. Check the installation information to ensure that the enclosure was cabled correctly.

Complete these steps to add an enclosure to the system by using the command line interface:

1. Using the `sainfo lsservicestatus` command (on the service CLI of the new enclosure), record the WWNN of the new enclosure.
2. Record the serial number of the enclosure, which is needed in later steps.
3. Enter the following command to verify that the enclosure is detected on the fabric:

svcinfolcontrolenclosurecandidate

4. Enter the `lsiogrp` command to determine the I/O group, where the enclosure must be added:
5. Record the name or ID of the first I/O group that has a node count of zero. You will need the ID for the next step.
6. Enter the following command to add the enclosure to the system:

addcontrolenclosure -iogrp iogrp_name | iogrp_id -sernum enclosureserialnumber

where:

- `iogrp_name | iogrp_id` is the name or ID of the I/O group
- `enclosureserialnumber` is the serial number of the enclosure

7. Record this information for future reference:
 - Serial number.
 - Worldwide node name of both node canisters.
 - All of the worldwide port names.
 - The name or ID of the I/O group that contains the enclosure.
8. Enter the `lsnodecanister` command to verify that the node canisters in the enclosure are online.

For more detailed information, see [IBM FlashSystem 7200 8.4.0 Documentation](#).

IBM FlashSystem 5000

Similar to the IBM FlashSystem 7200, the following types of expansion enclosures are available for the 5000 family:

- ▶ IBM FlashSystem 5000 LFF Expansion Enclosure Model 12G
- ▶ IBM FlashSystem 5000 SFF Expansion Enclosure Model 24G
- ▶ IBM FlashSystem 5000 High Density Expansion Enclosure Model 92G

The IBM FlashSystem 5010 supports only one control enclosure and only one SAS expansion chain.

The procedure for adding expansion enclosures to an IBM FlashSystem 5000 control enclosure is similar to that described in section “IBM FlashSystem 7200” on page 454.

For more detailed information, see [IBM FlashSystem 5000 8.4.0 Documentation](#).

10.9.7 Removing expansion enclosures

As storage environments change and grow it is sometimes necessary to move expansion enclosures between control enclosures. Removing an expansion enclosure is a straightforward task.

To remove an expansion enclosure from a control enclosure, complete the following steps:

1. If the expansion enclosure that you want to move is not at the end of a SAS chain, you might need a longer pair of SAS cables to complete the procedure. In that case, ensure that you have two SAS cables of suitable length before you start this procedure.
2. Delete any volumes that are no longer needed and that depend on the enclosure that you plan to remove.
3. Delete any remaining arrays that are formed from drives in the expansion enclosure. Any data in those arrays is automatically migrated to other managed disks in the pool if there is enough capacity.
4. Wait for data migration to complete.
5. Mark all the drives (including any configured as spare or candidate drives) in the enclosures to be removed as unused.
6. Unmanage and remove the expansion enclosure by using the management GUI. Select **Monitoring** → **System Hardware**. On the **System Hardware - Overview** page, select the directional arrow next to the enclosure that you are removing to open the Enclosure Details page. Select **Enclosure Actions** → **Remove**.

Important: Do not proceed until the enclosure removal process completes successfully.

7. On the I/O group that contains the expansion enclosure that you want to remove, enter the following command to put the I/O group into maintenance mode:

```
chiogrp -maintenance yes <iogroup_name_or_id>
```
8. If the expansion enclosure that you want to move is at the end of a SAS chain, complete the following steps to remove the enclosure from the SAS chain:
 - a. Disconnect the SAS cable from port 1 of canister 1 and canister 2. The enclosure is now disconnected from the system.
 - b. Disconnect the other ends of the SAS cables from the previous enclosure in the SAS chain. The previous enclosure is now the end of the SAS chain. Proceed to step 10.
9. If the expansion enclosure is not at the end of a SAS chain, complete the following steps to remove the enclosure from the SAS chain:
 - a. Disconnect the SAS cable from port 2 of canister 1 of the expansion enclosure that you want to move.
 - b. Disconnect the other end of the same SAS cable from port 1 of canister 1 of the next expansion enclosure in the SAS chain.
 - c. Disconnect the SAS cable from port 1 of canister 1 of the expansion enclosure that you want to move.
 - d. Reroute the cable that was disconnected in the previous step and connect it to port 1 of canister 1 of the next expansion enclosure in the SAS chain.

Important: Do not continue until you complete this cable connection step.

- e. Disconnect the SAS cable from port 2 of canister 2 of the expansion enclosure that you want to move.
 - f. Disconnect the other end of the same SAS cable from port 1 of canister 2 of the next expansion enclosure in the SAS chain.
 - g. Disconnect the SAS cable from port 1 of canister 2 of the expansion enclosure that you want to move.
 - h. Reroute the cable that was disconnected in the previous step and connect it to port 1 of canister 2 of the next expansion enclosure in the SAS chain.
10. Take the I/O group out of maintenance mode by entering the following command:
- ```
chiogrp -maintenance no <iogroup_name_or_id>
```
11. Check the event log for any errors and fix those errors as needed.
12. Disconnect the power from the expansion enclosure that you want to remove.
13. Remove the expansion enclosure from the rack along with its two power cables and two SAS cables.

Note that the IBM FlashSystem products provide methods to securely erase data from a drive when an enclosure is decommissioned or before a drive is removed from the system during a repair activity. For more information about the CLI commands that are used to run this secure erase function, see [IBM FlashSystem 9200 8.4.0 Documentation - Secure data deletion](#).

## 10.9.8 IBM FlashWatch

IBM FlashWatch is an offering from IBM to complement the purchase of the IBM FlashSystem product. It provides the following features that are included in the purchase of the product:

- ▶ IBM Flash Momentum
 

Flash Momentum is a storage upgrade program which allows you to replace your controller and storage every 3 years with full flexibility. Prior to the expiration of the agreement period, you decide whether to keep your FlashSystem, refresh it or simply walk away. You can refresh your FlashSystem for the same monthly price or less, or upsize or downsize your system to meet your needs.
- ▶ High Availability guarantee
 

Robust Spectrum 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 2:1 data reduction is guaranteed and you will need to self-certify that the data you're writing is able to be reduced (e.g. not encrypted, not already compressed). Up to 5:1 data reduction can be guaranteed 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 seven years of 24x7 support, with three years of IBM Technical Advisor support, enhanced response times of 30 minutes for severity 1 incidents, and six managed code upgrades over three years. However, this feature is not available for all IBM FlashSystem models; refer to the product matrix in Table 10-11 on page 458.
- ▶ Storage Insights
 

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 three years of your planned capacity needs on day one. To predict and control your future needs, IBM utilizes IBM Storage Insights to help you easily meet your capacity needs without interrupting your data center. The IBM FlashSystem 9200 (9848-UG8) 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. If storage needs to 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 non-IBM) to your FlashSystem product using an approach of your choice, without having to pay any additional external licensing.

Table 10-11 provides a summary product matrix for IBM FlashSystem products.

Table 10-11 IBM FlashWatch product matrix for IBM FlashSystem products

| IBM FlashWatch feature                   | FS5000                                                                                                   | FS5100                                                                           | FS7200                                                                           | FS9200                                                               | FS9200R                                |
|------------------------------------------|----------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------|----------------------------------------------------------------------------------|----------------------------------------------------------------------|----------------------------------------|
| High Availability guarantee              | FS5030 only                                                                                              | included with all models                                                         | included with all models                                                         | included with all models                                             | included with all models               |
| Data reduction guarantee                 | FS5030 only                                                                                              | included with all models                                                         | included with all models                                                         | included with all models                                             | included with all models               |
| All-inclusive Licensing                  | N/A                                                                                                      | N/A                                                                              | included with all models                                                         | included with all models                                             | included with all models               |
| Comprehensive Care (24x7 and 3-year ECS) | alternative optional services available; 9x5 next business day warranty                                  | alternative optional services available; 9x5 next business day warranty          | optional service; 9x5 next business day warranty                                 | included with models 9848-AG8 9848-UG8                               | included with models 9848-AG8 9848-UG8 |
| Storage Insights                         | included with all models                                                                                 | included with all models                                                         | included with all models                                                         | included with all models                                             | included with all models               |
| Flash Endurance guarantee                | included with all models                                                                                 | included with all models                                                         | included with all models                                                         | included with all models                                             | included with all models               |
| IBM Flash Momentum                       | included with models<br>2072-2H2<br>2072-2H4<br>2072-3H2<br>2072-3H4<br>2072-12G<br>2072-24G<br>2072-92G | included with models<br>2078-4H4<br>2078-UHB<br>2078-12G<br>2078-24G<br>2078-92G | included with models<br>2076-824<br>2076-U7C<br>2076-12G<br>2076-24G<br>2076-92G | included with models<br>9848-AG8<br>9848-UG8<br>9848-AFF<br>9848-A9F | included with model<br>9848-AG8        |

| IBM FlashWatch feature      | FS5000                   | FS5100                       | FS7200                       | FS9200                       | FS9200R                  |
|-----------------------------|--------------------------|------------------------------|------------------------------|------------------------------|--------------------------|
| IBM Storage Utility pricing | N/A                      | included with model 2078-UHB | included with model 2076-U7C | included with model 9848-UG8 | N/A                      |
| No Cost Migration           | included with all models | included with all models     | included with all models     | included with all models     | included with all models |

For more information about the IBM FlashWatch offering, see [IBM FlashWatch FAQ](#).

## 10.10 I/O Throttling

I/O throttling is a mechanism that allows you to limit the volume of I/O 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 FlashSystem V8.3 and later code brings the possibility to set the throttling at a volume level, host, host cluster, storage pool, 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.

### 10.10.1 General information on I/O throttling

These are characteristics of I/O throttling:

- ▶ Both 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 managed disk groups can be throttled.
- ▶ Queuing occurs at microsecond granularity.
- ▶ Internal I/O operations (such as 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.
- ▶ Throttle limit is a per node value.

### 10.10.2 I/O throttling on front-end I/O control

You can use throttling for a 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 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 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 throttle defined.

For example, a volume used for backups can be configured to use less bandwidth than a volume 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.

### 10.10.3 I/O Throttling on back-end I/O control

You can also use throttling to control the back-end I/O by throttling the storage pools, which can be useful in the following scenarios:

- ▶ Each storage pool can have a throttle defined.
- ▶ Allows control of back-end I/Os from the IBM FlashSystem.
- ▶ Useful to avoid overwhelming any external back-end storage.
- ▶ Useful in case of VVOLS since a VVOL gets 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.

### 10.10.4 Overall benefits of using I/O throttling

The overall benefits of using I/O throttling is a better distribution all system resources:

- ▶ Avoids overwhelming the controller objects.
- ▶ Avoids starving the external entities, like *hosts*, from their share of controller.
- ▶ A scheme of distribution of controller resources that, in turn, results in better utilization of external resources such as host capacities.

With throttling not enabled, there is a scenario where Host1 dominates the bandwidth, and after enabling the throttle, there is a much better distribution of the bandwidth among the hosts, as shown in Figure 10-9 on page 461.

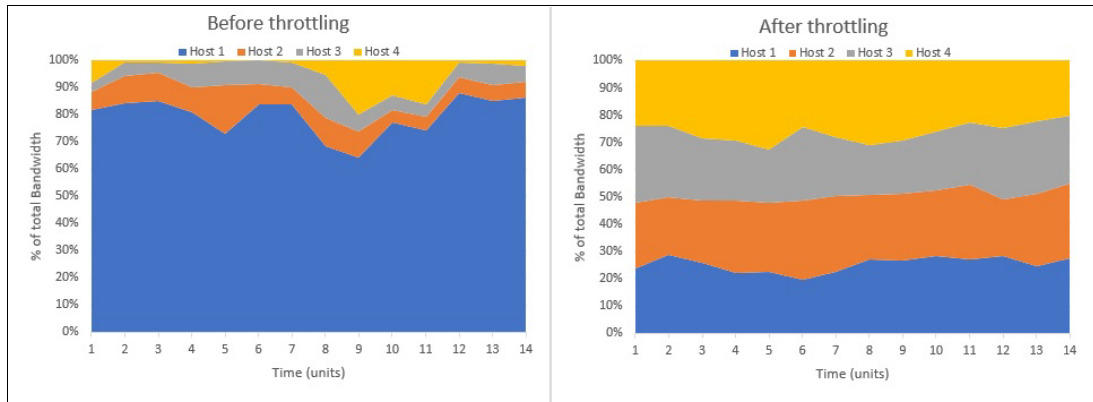


Figure 10-9 Distribution of controller resources before and after I/O throttling

### 10.10.5 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 hostcluster which already has a hostcluster throttle.
- ▶ If the hostcluster does not have a throttle defined, its member hosts can have their individual host throttles defined.
- ▶ If a volume has multiple copies, then throttling would be done for the storage pool serving the primary copy. The throttling will not be applicable on the secondary pool for mirrored volumes and stretched cluster implementations.
- ▶ A host cannot be added to a hostcluster if both have their individual throttles defined. If just one of the host or hostcluster throttles is present, the command will succeed.
- ▶ A seeding host used for creating a hostcluster cannot have a host throttle defined for it.

**Note:** Throttling is only applicable at the I/Os that an IBM FlashSystem receives from hosts and hostclusters. The I/Os generated internally, such as mirrored volume I/Os, cannot be throttled.

### 10.10.6 Configuring I/O throttling using the CLI

To create a throttle using the CLI, use the `mkthrottle` command, as shown in Example 10-10. The bandwidth limit is the maximum amount of bandwidth the system can process before the system delays I/O processing. Similarly, the `iops_limit` is the maximum amount of IOPS the system can process before the system delays I/O processing.

*Example 10-10 Creating a throttle 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:IBM Redbook FS:superuser>mkthrottle -type host -bandwidth 100
```

```
-host ITS0_HOST3
```

```
IBM_FlashSystem:IBM Redbook FS:superuser>mkthrottle -type hostcluster -iops 30000
```

```
-hostcluster ITS0_HOSTCLUSTER1
```

```
IBM_FlashSystem:IBM Redbook FS:superuser>mkthrottle -type mdiskgrp -iops 40000
```

```
-mdiskgrp 0
```

```
IBM_FlashSystem:IBM Redbook FS:superuser>mkthrottle -type offload -bandwidth 50
```

```
IBM_FlashSystem:IBM Redbook FS:superuser>mkthrottle -type vdisk -bandwidth 25
```

```
-vdisk volume1
```

```
IBM_FlashSystem:IBM Redbook FS: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 hostcluster
```

```
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.

## 10.10.7 Configuring I/O throttling using the GUI

The following sections show how to configure the throttle by using the management GUI.

## 10.10.8 Creating a volume throttle

To create a volume throttle, go to **Volumes** → **Volumes**, then select the desired volume, right-click on it and chose **Edit Throttle**, as shown in Figure 10-10. The bandwidth can be set from 1 MBps - 256 TBps and IOPS can be set from 1 to 33,254,432.

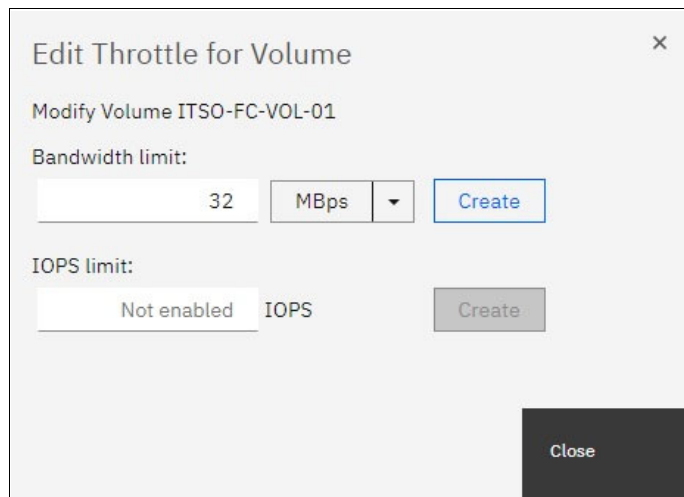


Figure 10-10 Creating a volume throttle in the GUI



If a throttle already exists, the dialog box that is shown in Figure 10-10 on page 462 also shows a **Remove** button that is used to delete the throttle.

### 10.10.9 Creating a host throttle

To create a host throttle, go to **Hosts** → **Hosts**, select the desired host, then right-click it and chose **Edit Throttle**, as shown in Figure 10-11.

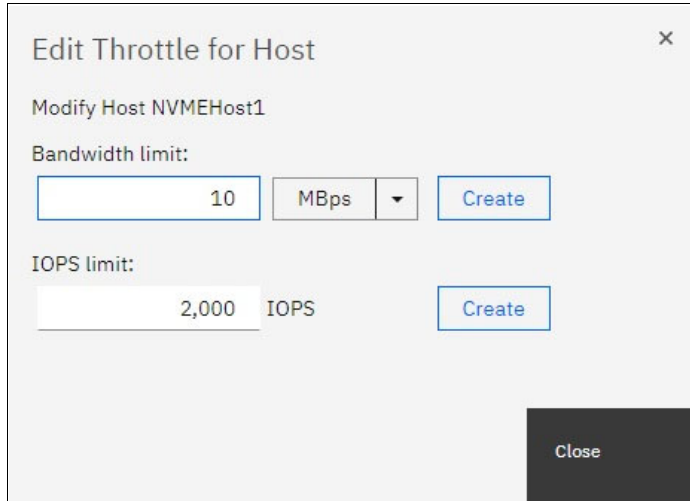


Figure 10-11 Creating a host throttle in the GUI

### 10.10.10 Creating a host cluster throttle

To create a host cluster throttle, go to **Hosts** → **Host Clusters**, select the desired host cluster, then right-click it and chose **Edit Throttle**, as shown in Figure 10-12.

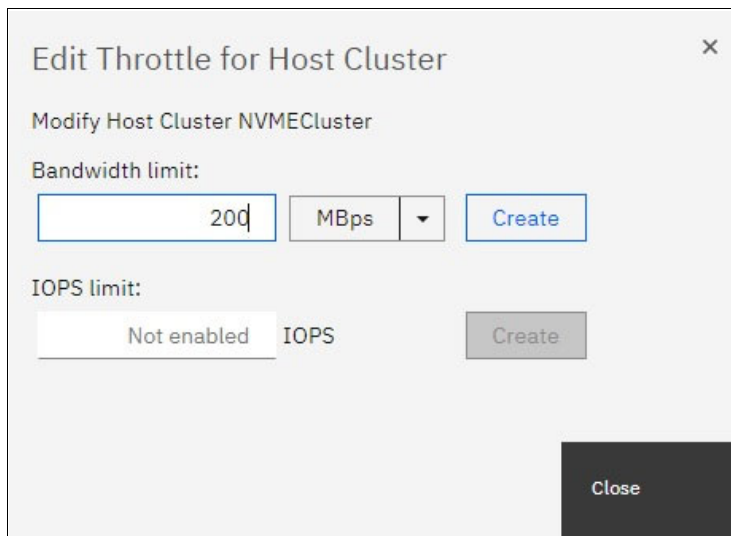


Figure 10-12 Creating a host cluster throttle in the GUI

### 10.10.11 Creating a storage pool throttle

To create a storage pool throttle, go to **Pools** → **Pools**, select the desired storage pool, then right-click on it and choose **Edit Throttle**, as shown in Figure 10-13.

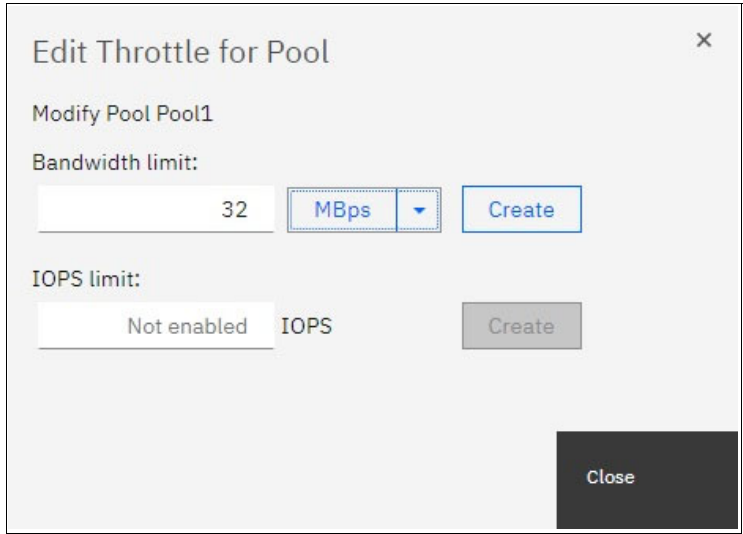


Figure 10-13 Creating a storage pool throttle in the GUI

### 10.10.12 Creating an offload throttle

To create an offload throttle, go to **Monitoring** → **System Hardware** → **Actions**, then select **Edit System Offload Throttle**, as shown in Figure 10-14.

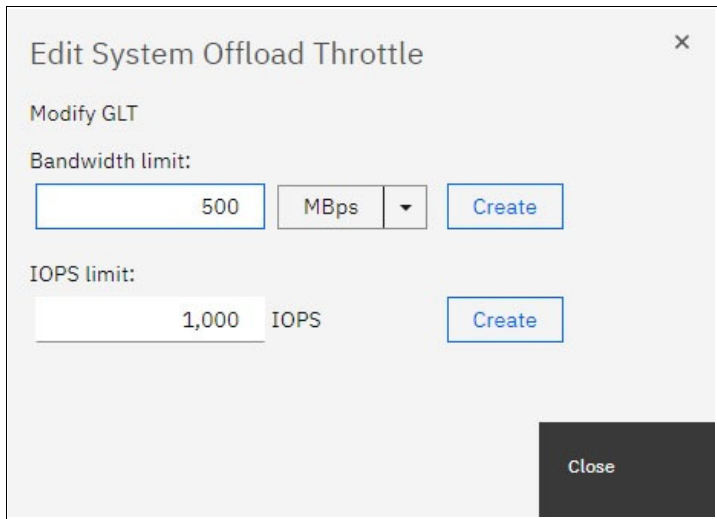


Figure 10-14 Creating system offload throttle in the GUI

## 10.11 Automation

Automation has become a priority for maintaining today's busy storage environments. Automation software allows the creation of repeatable sets of instructions and processes to reduce the need for human interaction with computer systems. Red Hat Ansible and other third-party automation tools are becoming increasingly used across the enterprise IT environments and it is not unexpected that their use in storage environments is becoming more popular.

### 10.11.1 Red Hat Ansible

IBM FlashSystem family includes integration with Red Hat Ansible Automation Platform, allowing IT to create an Ansible playbook that automates repetitive tasks across an organization in a consistent way, helping improve outcomes and reduce errors.

Ansible is an agentless automation management tool that uses the SSH protocol. Currently, Ansible can be run from any machine with Python 2 (version 2.7) or Python 3 (versions 3.5 and higher) installed. This includes Red Hat, Debian, CentOS, macOS, any of the BSDs. Windows is not supported for the Ansible control node.

IBM is a Red Hat certified support module vendor, providing simple management for the following commands used in the IBM Spectrum Virtualize Ansible Collection:

- ▶ **Collect facts:** Collect basic information including hosts, host groups, snapshots, consistency groups, and volumes.
- ▶ **Manage hosts:** Create, delete, or modify hosts.
- ▶ **Manage volumes:** Create, delete, or extend the capacity of volumes.
- ▶ **Manage mdisk:** Create or delete a managed disk.
- ▶ **Manage pool:** Create or delete a pool (managed disk group).
- ▶ **Manage volume map:** Create or delete a volume map.
- ▶ **Manage consistency group snapshot:** Create or delete consistency group snapshots.
- ▶ **Manage snapshot:** Create or delete snapshots.
- ▶ **Manage volume clones:** Create or delete volume clones.

This collection provides a series of Ansible modules and plugins for interacting with the IBM Spectrum Virtualize Family storage systems. The modules in the IBM Spectrum Virtualize Ansible collection leverage the Representational State Transfer (REST) application programming interface (API) to connect to the IBM Spectrum Virtualize storage system. These storage systems include the IBM SAN Volume Controller, IBM FlashSystem family including FlashSystem 5010, 5030, 5100, 7200, 9100, 9200, 9200R, and IBM Spectrum Virtualize for Public Cloud.

More information can be found in the following IBM Redpaper *Automate and Orchestrate® Your IBM FlashSystem Hybrid Cloud with Red Hat Ansible*, REDP-5598.

For IBM Spectrum Virtualize modules, Ansible version 2.9 or higher is required. For more information about IBM Spectrum Virtualize modules, see [Ansible Collections for IBM Spectrum Virtualize](#).

## 10.11.2 RESTful API

The Spectrum Virtualize 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 Spectrum Virtualize command line interface, which handles parsing parameter specifications for validity and error reporting. Hypertext Transfer Protocol Secure (HTTPS) is used to communicate with the RESTful API server.

To interact with the storage system by using the RESTful API, use the curl utility (see <https://curl.se>) to make an HTTPS command request with a valid configuration node URL destination. Open TCP port 7443 and include the keyword `rest` followed by the Spectrum Virtualize target command you want to run:

Each `curl` command takes the following form:

```
curl -k -X POST -H <header_1> -H <header_2> ... -d <JSON input>
https://<flashsystem_ip_address>:7443/rest/<target>
```

where:

- ▶ `POST` is the only HTTPS method that the Spectrum Virtualize RESTful API supports.
- ▶ Headers `<header_1>` and `<header_2>` are individually-specified HTTP headers (e.g. Content-Type and X-AuthUsername).
- ▶ `-d` is followed by the JSON input e.g. `{“raid_level”: “raid5”}`.
- ▶ `<flashsystem_ip_address>` is the IP address of the IBM FlashSystem that you are sending requests to.
- ▶ `<target>` is the target object of commands, which includes any object IDs, names, and parameters.

### Authentication

Aside from data encryption, the HTTPS server requires authentication of a valid username and password for each API session. Use two authentication header fields to specify your credentials: `X-Auth-Username` and `X-Auth-Password`.

Initial authentication requires that you `POST` the authentication target (`/auth`) with the username and password. The RESTful API server returns a hexadecimal token. A single session lasts a maximum of two active hours or thirty inactive minutes, whichever occurs first. When your session ends due to inactivity, or if you reach the maximum time that is allotted, error code 403 indicates the loss of authorization. Use the `/auth` command target to reauthenticate with the user name and password.

The following is an example of the correct procedure for authenticating. You authenticate by first producing an authentication token and then using that token in all future commands until the session ends.

For example, the following command passes the authentication command to IBM FlashSystem node IP address 192.168.10.20 at port 7443:

```
curl -k -X POST -H 'Content-Type: application/json' -H 'X-Auth-Username:
superuser' -H 'X-Auth-Password: passwOrd' https://192.168.10.20:7443/rest/auth
```

**Note:** Make sure you format the request correctly using spaces after each colon in each header otherwise the command will fail.

This request yields an authentication token, which can be used for all subsequent commands. For example:

```
{"token": "38823f60c758dca26f3eaac0ffee42aac4664964905a6f058ae2ec92e0f0b63"}
```

### Example command

Most actions must be taken only after authentication. The following example of creating an array demonstrates the use of the previously generated token in place of the authentication headers used in the authentication process.

```
curl -k -X POST -H 'Content-Type: application/json' -H 'X-Auth-Token: 38823f60c758dca26f3eaac0ffee42aac4664964905a6f058ae2ec92e0f0b63' -d '{"level": "raid5", "drive": "6:7:8:9:10", "raid6grp"}' https://192.168.10.20:7443/rest/mkarray
```

For more information about using the API, see [IBM Spectrum Virtualize as Software Only 8.3.1 Documentation - Spectrum Virtualize RESTful API](#).

For more information about other examples, see the following web pages:

- ▶ [IBM Spectrum Virtualize Interfacing Using the RESTful API](#)
- ▶ [Implementing a RESTful API to the IBM Storwize Family](#)
- ▶ [Tips and tricks using the Spectrum Virtualize REST API](#)

## 10.12 Documenting IBM FlashSystem and SAN environment

This section focuses on the challenge of automating the documentation that is needed for an IBM FlashSystem 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.
- ▶ Remember to also include photographs in the documentation where appropriate.

**Storing documentation:** Avoid storing IBM FlashSystem and SAN environment documentation only in the SAN. If your organization has a disaster recovery plan, include this storage documentation in it. Follow its guidelines about how to update and store this data. If no disaster recovery plan exists and you have the proper security authorization, it might be helpful to store an updated copy offsite.

In theory, this IBM FlashSystem 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 disaster recovery scenario, which is also why it is so important to run periodic disaster recovery 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 FlashSystem solution, use these forms to help you document how your IBM FlashSystem solution was first configured. Minimum documentation is needed for an IBM FlashSystem solution. Because you might have more business requirements that require other data to be tracked, remember that the following sections do not address every situation.

## 10.12.1 Naming conventions

Whether you are creating your IBM FlashSystem 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 FlashSystem and SAN environment. System administrators can then 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 FlashSystem and SAN component names limit the number and type of characters you can use. For example, IBM FlashSystem names are limited to 63 characters, which makes creating a naming convention a bit easier

Many names in IBM FlashSystem 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 FlashSystem names its internal canisters or nodes as nodeX, with X being a sequential decimal number. These will range from two up to eight, in a four IBM FlashSystem 9200 system cluster.

If multiple additional external controllers are attached to your IBM FlashSystem solution, these 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 FlashSystem to know which storage controller to check.

**Note:** An IBM FlashSystem detects external controllers based on their worldwide node name (WWNN). If you have an external storage controller that has one WWNN for each worldwide port name (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 an IBM FlashSystem 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 (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 avoid the use 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 it to include the following information:

- ▶ For internal MDisk refer to the IBM FlashSystem 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 it has in the storage controller.

Consider the following examples of MDisk names with this convention:

- ▶ FS9200CL01-MD03, where FS9200CL01 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 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:

- ▶ FS9200-P00L01: where FS9200 is the system or cluster name, and POOL01 is the pool.
- ▶ P05XYZ1\_3GR5: Pool 05 from serial 75VXYZ1, LUNs with 300 GB FC DDMs and RAID 5.
- ▶ P16XYZ1\_EX01: Pool 16 from serial 75VXYZ1, pool 01 dedicated to Exchange Mail servers.
- ▶ XIV01\_F9H02\_ET: Pool with disks from XIV named XIV01 and FlashSystem 900 F9H02, both 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 FlashSystem 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 datastore composed of a single volume
  - esx01-sessions-001a and esx01-sessions-001b: For a datastore composed of 2 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 FlashSystem is as simple as changing any other IBM FlashSystem object name. However, changing the name on the operating system of a server might have implications for application configuration, DNS and may require a server reboot. 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, 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 it. 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 ESX server NYVMCLUSTER01.
- ▶ IBM-NYFS9200-N1P1\_HOST: Port 1 of Node 1 from FS9200 Cluster NYFS9200 dedicated for hosts.
- ▶ IBM-NYFS9200-N1P5\_INTRA: Port 5 of Node 1 from FS9200 Cluster NYFS9200 dedicated to intracluster traffic.
- ▶ IBM-NYFS9200-N1P7\_REPL: Port 7 of Node 1 from FS9200 Cluster NYFS9200 dedicated to replication.

Be mindful of the IBM FlashSystem 9200 port aliases. There are mappings between the last digits of the port WWPN and the node FC port.

- ▶ IBM\_D88870\_75XY131\_I0301: DS8870 serial number 75XY131, 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 interoperations modes, use WWPNs in your zones. However, remember to update every zone that uses a WWPN, if you change it.



Your SAN zone name should reflect the devices in the SAN it includes (normally in a one-to-one relationship), as shown in the following examples:

- ▶ SERVERALIAS\_T1\_FS9200CLUSTERNAME (from a server to the IBM FlashSystem 9200, where you use T1 as an identifier to zones that uses for example, node ports P1 on Fabric A, and P2 on Fabric B).
- ▶ SERVERALIAS\_T2\_FS9200CLUSTERNAME (from a server to the IBM FlashSystem 9200, where you use T2 as an identifier to zones that uses for example, node ports P3 on Fabric A, and P4 on Fabric B).
- ▶ IBM\_DS8870\_75XY131\_FS9200CLUSTERNAME (zone between an external back-end storage and the IBM FlashSystem 9200).
- ▶ NYC\_FS9200\_P0K\_FS9200\_REPLICATION (for Remote Copy services).

## 10.12.2 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 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 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. The tool then creates a compressed file with the data collection. This file is sent to a Brocade automated machine for processing by secure web or e-mail.

After some time (typically a few hours), you will receive an e-mail 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 provided to learn how to create a well-organized and useful diagram.

Figure 10-15 on page 472 shows an example of a poorly formatted diagram.

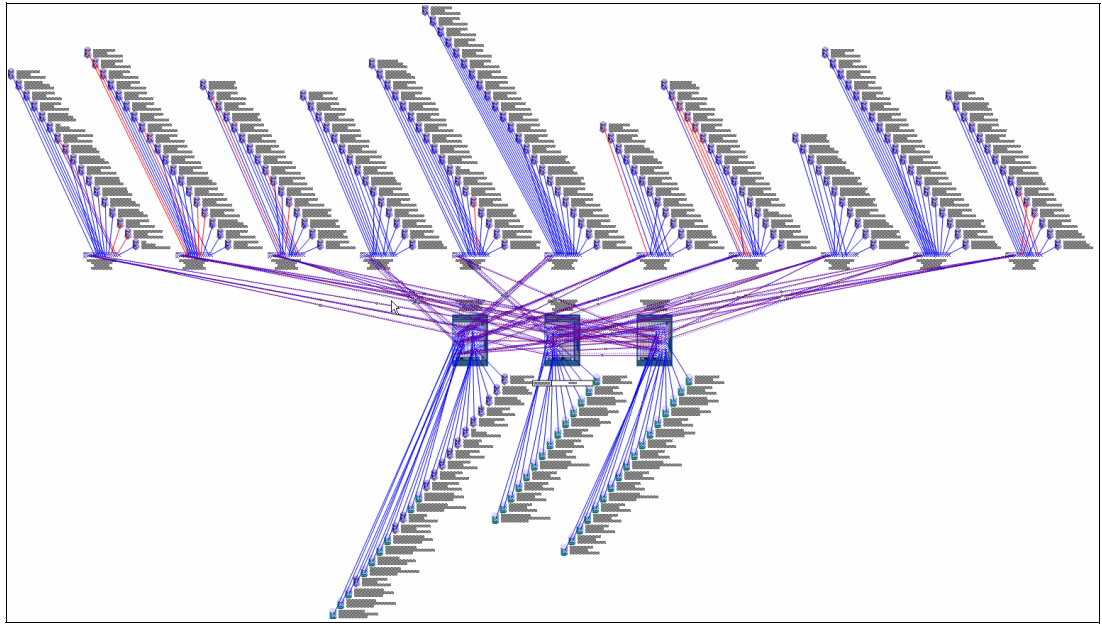


Figure 10-15 A poorly formatted SAN diagram

Figure 10-16 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.

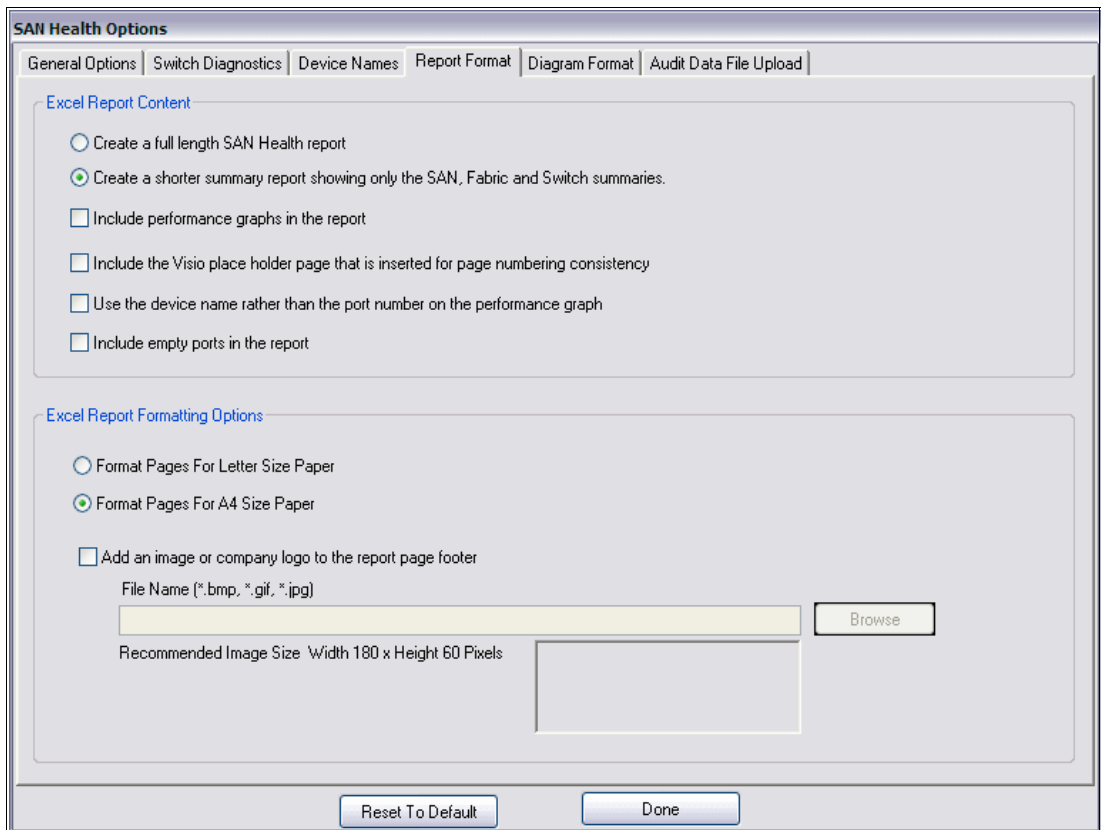


Figure 10-16 Brocade SAN Health Options window

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 [Broadcom Support Download SAN Health Diagnostics Capture](#).

**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 5.4.2 Documentation](#).

For more information on how to configure and set-up Spectrum Control, see Chapter 9, “Monitoring” on page 363.

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.

## 10.12.3 IBM FlashSystem documentation

You can back up the configuration data for an IBM FlashSystem system after preliminary tasks are completed. Configuration data for the system provides information about your system and the objects that are defined in it. It 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 is copied to `/dumps/svc.config.cron.xml_<serial#>`.

Use these instructions to generate a manual backup at any time:

1. Issue the **svconfig backup** command to back up your configuration. The command displays messages similar to the ones in Example 10-11 on page 474.

*Example 10-11 Sample svcconfig backup command output*

---

```
IBM_FlashSystem:IBM Redbook FS:superuser>svcconfig backup
.....
...
.....
...
.....
CMMVC6155I SVCCONFIG processing completed successfully
```

---

The **svcconfig 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 10-12 describes the three files that are created by the backup process.

*Table 10-12 Files created by the backup process*

| File name                       | Description                                                                             |
|---------------------------------|-----------------------------------------------------------------------------------------|
| svc.config.backup.xml_<serial#> | Contains your configuration data.                                                       |
| svc.config.backup.sh_<serial#>  | Contains the names of the commands that were issued to create the backup of the system. |
| svc.config.backup.log_<serial#> | Contains details about the backup, including any reported errors or warnings.           |

2. Check that the **svcconfig 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 10-12 shows the output of the **pscp** command.

*Example 10-12 Saving the config 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 FlashSystem documentation spreadsheet. The configuration backup file might be quite 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 FlashSystem 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 lsmdisk`
- ▶ `svcinfo lsmdiskgrp`
- ▶ `svcinfo lsvdisk`
- ▶ `svcinfo lshost`
- ▶ `svcinfo lshostvdiskmap`

**Note:** Most CLI commands that are shown above will work without the `svcinfo` prefix, however there might be some that do not work with just the short-name, and so 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 FlashSystem documentation spreadsheet.

When you are gathering IBM FlashSystem information, consider the following preferred 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 values (CSV) format, specifying the same delimiter.

**Note:** It is important to 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 pre-process the temporary files and remove any “garbage” or undesired 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 run automatically on a regular basis. 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 10.12.7, “Automated support data collection” on page 478.

## 10.12.4 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 FlashSystem internal drives and enclosures.

Any portion of your external storage controllers that is used outside the IBM FlashSystem solution might have its configuration changed frequently. In this case, see your back-end storage controller documentation for more information about how to gather and store the information that you need.

Fully allocate all of the available space in any of the optional external storage controllers that you might use as additional back-end to the IBM FlashSystem solution. This way, you can perform all your disk storage management tasks by using the IBM FlashSystem user interface.

## 10.12.5 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, serial number (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.

## 10.12.6 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 SNMP traps to this incident management tool, you should manually open incidents whenever an error is detected.
- ▶ The IBM FlashSystem has the ability to be managed by the IBM Storage Insights (SI) tool, that is available free of charge to owners of IBM storage systems. The SI tool allows you to monitor all the IBM storage devices information on SI. For more information, see Chapter 9, “Monitoring” on page 363.
- ▶ 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 FlashSystem 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 e-mail 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 is to have. These activities are typically safe and non-disruptive to other services and applications when properly handled.

However, they have the potential to cause collateral damage if a 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, 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 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 need to 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.

## 10.12.7 Automated support data collection

In addition to the easier-to-use documentation of your IBM FlashSystem 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 FlashSystem, this information includes **snap** data. For other equipment, see the related documentation for more information about how to gather and store the support data that you might need.

You can create procedures that automatically create and store this data on scheduled dates, delete old data, or transfer the data to tape.

There is also the possibility to use IBM Storage Insights to create support tickets and then attach the snap data to this record from within the SI GUI. For more information, see Chapter 11, “Troubleshooting and diagnostics” on page 479.

## 10.12.8 Subscribing to IBM FlashSystem support

Subscribing to IBM FlashSystem 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, see [Stay up to date with My Notifications](#).

If you do not have an IBM ID, create an ID.

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 you wants to receive. For example, type **IBM FlashSystem 9200** in the Product lookup text box and then click **Subscribe** to subscribe to FlashSystem 9200 notifications, as shown in Figure 10-17.



The screenshot shows a web interface titled "Subscribe to notifications". It features a "Product lookup:" label followed by a text input field containing "IBM FlashSystem 9200". Below this, there is a "Product subscri" label and a button labeled "IBM FlashSystem 9200 + Subscribe".

Figure 10-17 Creating a subscription to IBM FlashSystem 9200 notifications





# Troubleshooting and diagnostics

This chapter provides information to start troubleshooting common problems that can occur in IBM FlashSystem environment. It describes situations that are related to IBM FlashSystem, the SAN environment, optional external storage subsystems and hosts. It also explains how to collect the necessary problem determination data.

This chapter includes the following sections:

- ▶ 11.1, “Starting troubleshooting” on page 480
- ▶ 11.2, “Diagnostic data collection” on page 485
- ▶ 11.3, “Common problems and isolation techniques” on page 489
- ▶ 11.4, “Remote Support Assistance” on page 506
- ▶ 11.5, “Call Home Connect Cloud and Health Checker feature” on page 507
- ▶ 11.6, “IBM Storage Insights” on page 509

## 11.1 Starting troubleshooting

Troubleshooting is a systematic approach to solving a problem. The goal of troubleshooting or problem determination is to understand why something does not work as expected and find a resolution. Hence, the first step is to describe the problem as accurately as possible, then perform log collection from all the involved products of the solution as soon as the problem is reported. An effective problem report ideally should describe the expected behavior, the actual behavior, and, if possible, how to reproduce the behavior.

The following questions help define the problem for effective troubleshooting.

- ▶ What are the symptoms of the problem?
  - What is reporting the problem?
  - What are the error codes and messages?
  - What is the business impact of the problem?
  - Where does the problem occur?
  - Is the problem specific to one or multiple host, one or both nodes?
  - Is the current environment and configuration supported?
- ▶ When does the problem occur?
  - Does the problem happen only at a certain time of day or night?
  - How often does the problem happen?
  - What sequence of events leads up to the time that the problem is reported?
  - Does the problem happen after an environment change, such as upgrading 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 on a test system?
  - Are multiple users or applications encountering the same type of problem?
  - Can the problem be recreated by running a single command, a set of commands, or a particular application, or a stand-alone application?

Log files collection close to the time of the incident and accurate timeline is absolutely critical for effective troubleshooting.

IBM FlashSystem Graphical User Interface (GUI) is a good starting point for your troubleshooting. It has two icons at the top, which can be accessed from any panel of the GUI. As shown in Figure 11-1 on page 481, the first icon shows IBM FlashSystem events, such as an error or a warning, and the second icon shows suggested tasks and background tasks that are running, or that were recently completed.

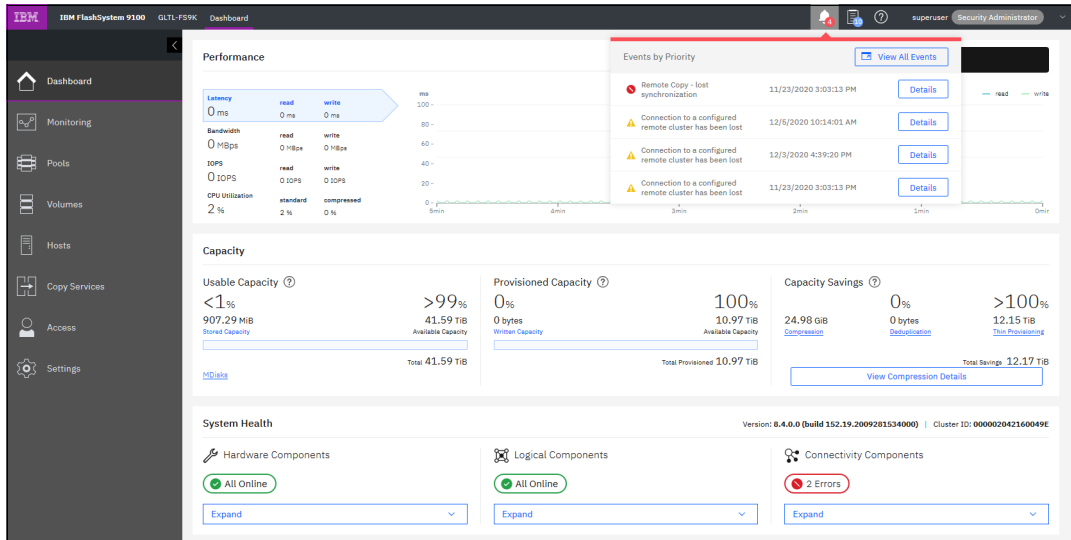


Figure 11-1 Events icon in GUI

The Dashboard provides an at-a-glance look into the condition of the system and notification of any critical issues that require immediate action. It contains sections for performance, capacity, and system health that provide an overall understanding of what is happening on the system.

Figure 11-2 shows the Dashboard panel displaying the system health panels.

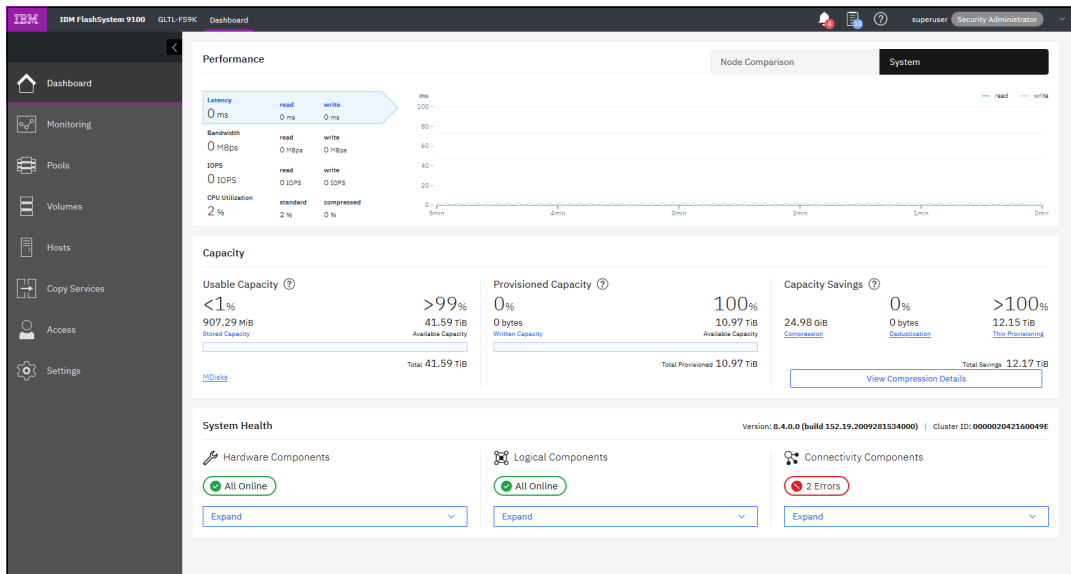


Figure 11-2 Dashboard showing system health

The System Health section in the bottom part of the Dashboard provides information on the health status of hardware, and logical and connectivity components. If you click **Expand** in each of these categories, the status of individual components is shown, as shown in the example in Figure 11-3. You can also go further and click **More Details**, which will take you to the panel related to that specific component, or will show you more information about it.

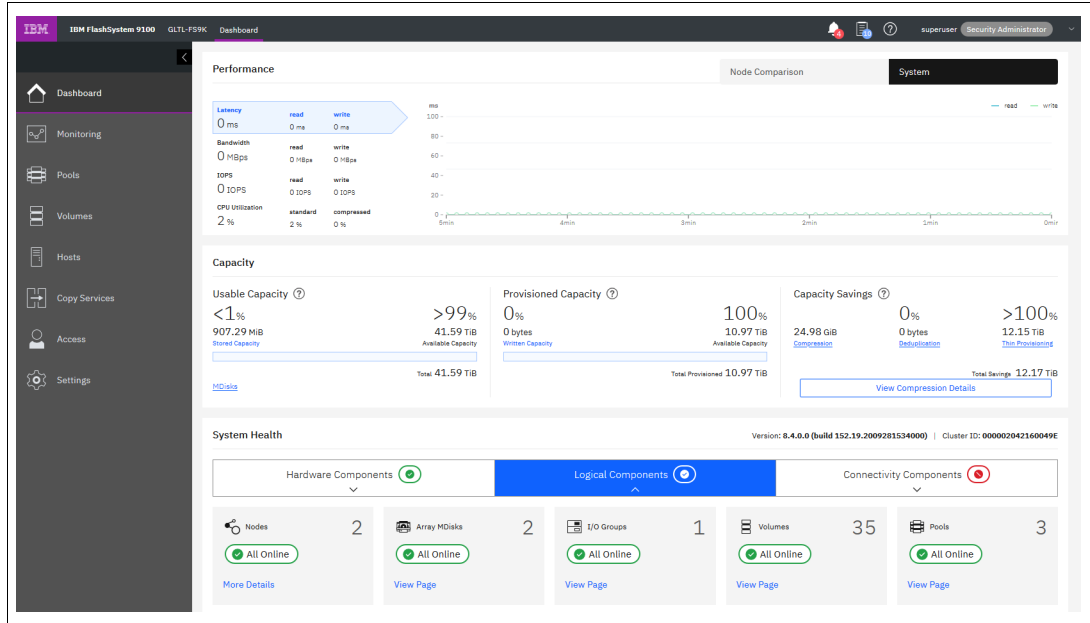


Figure 11-3 System Health expanded section in dashboard

For the entire list of components in each category and information about IBM FlashSystem troubleshooting, see: [IBM FlashSystem 9200 8.4.0 Documentation - Troubleshooting](#) (requires an IBM ID).

### 11.1.1 Recommended actions and fix procedure

The fix procedures have been carefully designed to assist users to fix the problem without doing harm. When there are multiple unfixed error codes in the event log, the management GUI provides a way to run the next recommended fix procedure. Hence the first step in troubleshooting is to run the fix procedures on the error codes in the event log.

These messages and codes provide reference information about informational configuration events, error event codes when a service action is required. Cluster Error Code (CEC) is visible in the cluster event log whereas Node Error Code (NEC) is visible in node status in the service assistant GUI. A cluster might encounter the following types of failure recoveries due to various conditions:

- ▶ Node assert (warmstart or Tier1/T1 recovery) is reported as CEC 2030
- ▶ Cluster recovery (Tier2/T2 recovery) is reported as CEC 1001
- ▶ System recovery (Tier3/T3 recovery) is required when all nodes of the clustered system report NEC 550/578

For a full listing of the messages and codes, see: [IBM FlashSystem 9200 8.4.0 Documentation - Messages and codes](#).

The **Monitoring** → **Events** panel shows information messages, warnings, and issues on the IBM FlashSystem. Therefore, this is a good place to check the current problems in the system.

Use the **Recommended Actions** filter to display the most important events that need to be fixed.

If there is an important issue that needs to be fixed, the **Run Fix** button will be available in the top-left corner with an error message, indicating which event should be fixed as soon as possible. This fix procedure assists you to resolve problems in IBM FlashSystem. It analyzes the system, provides more information on the problem, suggest actions to be taken with steps to be followed, and finally checks to see if the problem is resolved.

**Note:** IBM FlashSystem detects and reports error messages, however many events could potentially be triggered by external storage subsystems or the SAN.

Always use the fix procedures to resolve errors that are reported by the system, such as system configuration problems or hardware failures.

Figure 11-4 shows **Monitoring** → **Events** panel with the **Run Fix** button.

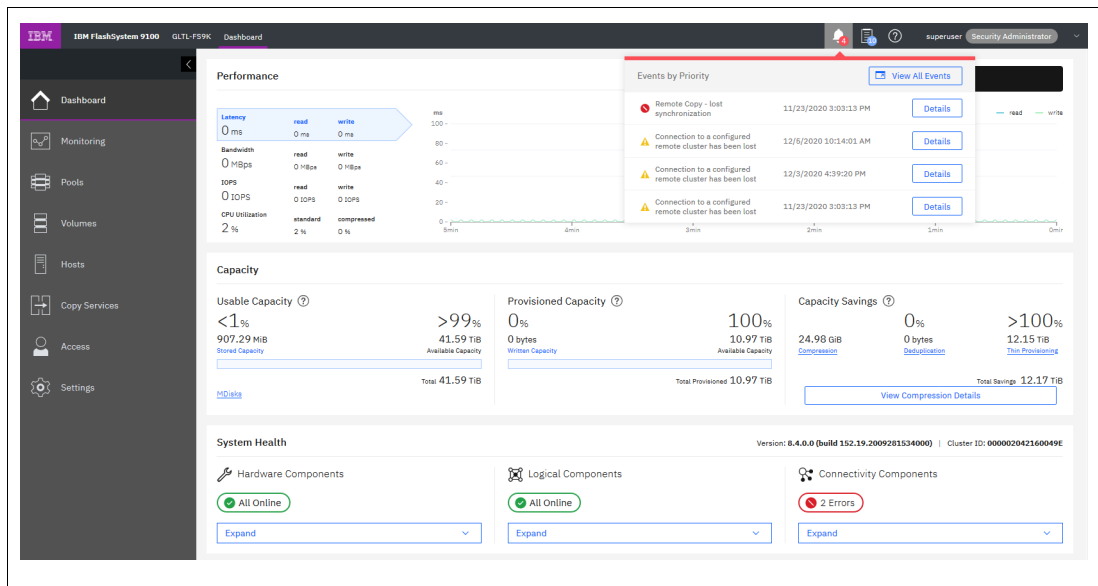


Figure 11-4 Monitoring > Events panel

**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, select an event in the table, and click **Properties** in the **Actions** menu. You can also get access to the **Run Fix Procedure** and properties by right-clicking an event.

Additional information about it is displayed in the *Properties and Sense Data* window for the specific event, as shown in Figure 11-5 on page 484. You can review and also click **Run Fix** to run the fix procedure.

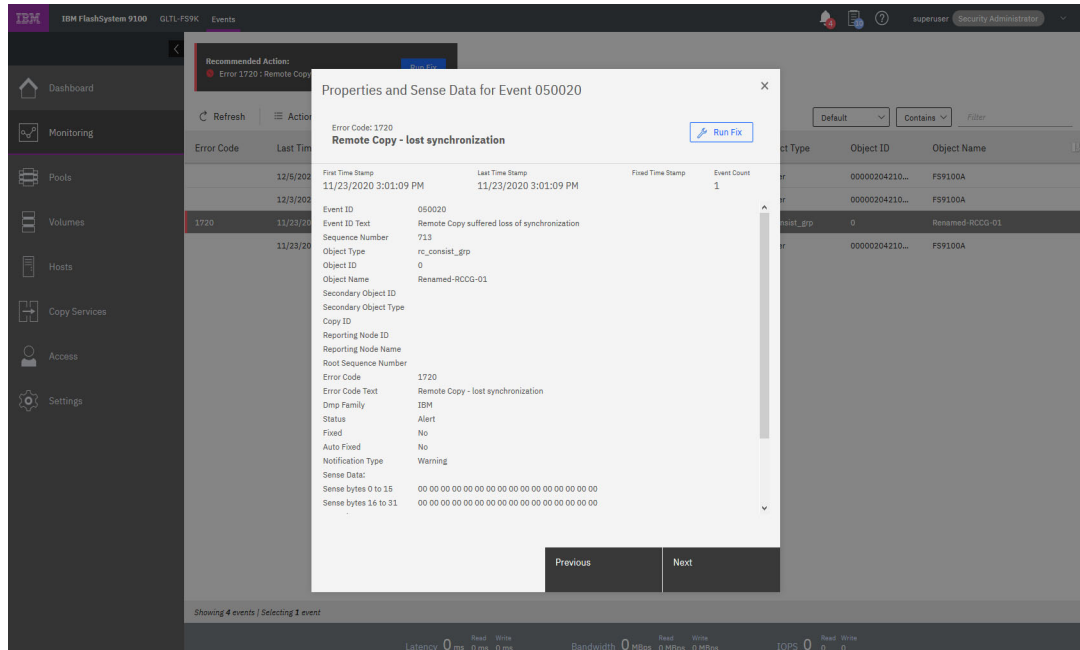


Figure 11-5 Properties and sense data for event window

**Tip:** On the Properties and Sense Data for Event Window, use the **Previous** and **Next** buttons to move between events.

Another common practice is to use the IBM Spectrum Virtualize command line interface (CLI) to find issues and resolve them. You can use the IBM Spectrum Virtualize CLI to perform common error recovery steps. Although the maintenance procedures perform these steps, it is sometimes faster to run these commands directly through the CLI.

Run the following commands when you have the following issues:

- ▶ You experience a back-end storage issue (for example, error code 1370 or error code 1630).
- ▶ You performed maintenance on the back-end storage subsystems.

**Important:** Run the following commands when any type of change related to the communication between IBM Spectrum Virtualize and back-end storage subsystem occurs (such as back-end storage is just configured or a zoning change occurs). This ensures that IBM Spectrum Virtualize recognizes the changes.

Common error-recovery involves the following IBM Spectrum Virtualize CLI commands:

- ▶ **lscontroller** and **lsmdisk**  
Provides current status of all controllers and MDisks.
- ▶ **detectmdisk**  
Discovers the changes in the back-end.
- ▶ **lscontroller <controller\_id\_or\_name>**

Checks the controller that was causing the issue and verifies that all the WWPNs are listed as you expect. It also checks that the path\_counts are distributed evenly across the WWPNs.

► **lsmdisk**

Determines whether all MDisks are online.

**Note:** When an issue is resolved using the CLI, verify that the error disappears from **Monitoring** → **Events** panel. If not, make sure the error has truly been fixed, and if so, manually mark the error as fixed.

## 11.2 Diagnostic data collection

Data collection and problem isolation in an IT environment are sometimes difficult tasks. In the following section, the essential steps that are needed to collect debug data to find and isolate problems in an IBM FlashSystem environment are described.

### 11.2.1 IBM FlashSystem data collection

When there is a problem with an IBM FlashSystem and you have to open a case with IBM support, you need to provide the support packages for the device. To collect and upload the support packages to IBM support center, you can do it automatically using IBM FlashSystem, or download the package from the device and manually upload to IBM. The easiest way is to automatically upload the support packages from IBM FlashSystem. It can be done using either the GUI or the CLI.

You can also use the new IBM Storage Insights application to do the log data upload and this is described in 11.6.4, “Updating support tickets” on page 520.

#### Data collection using the GUI

To perform data collection using the GUI, complete the following steps:

1. In the panel **Settings** → **Support** → **Support Package**, both options to collect and upload support packages are available
2. To automatically the support packages, click **Upload Support Package** button.
3. In the pop-up screen, enter the PMR number and the type of support package to upload to the IBM support center. The **Snap Type 4** can be used to collect standard logs and generate a new statesave on each node of the system

The *Upload Support Package* panel is shown in Figure 11-6 on page 486.

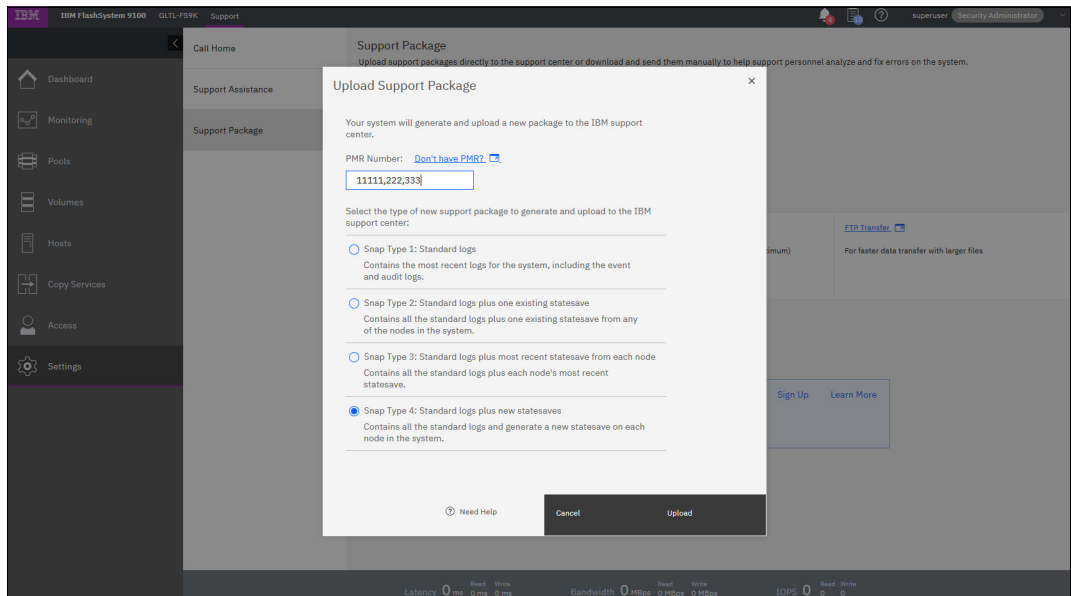


Figure 11-6 Upload Support Package panel

To learn more about the required support package which is most appropriate to diagnose different type of issues, see [What Data Should You Collect for a Problem on SVC or Storwize Systems](#).

In summary:

- ▶ For issues related to interoperability with hosts or storage, use **Snap Type 4**.
- ▶ For critical performance issues, collect Option 1 and then collect **Snap Type 4**.
- ▶ For general performance issues, collect **Snap Type 4**.
- ▶ For issues related to replication, including 1920 errors, collect **Snap Type 4** from both systems.
- ▶ For issues related to compressed volumes, collect **Snap Type 4**.
- ▶ For 2030, 1196 or 1195 errors collect **Snap Type 4**.
- ▶ For all other issues, collect **Snap Type 4**.

### Data collection using the CLI

To use the CLI to collect the same type of support packages mentioned in “Data collection using the GUI” on page 485, you have to first generate a new `livedump` of the system using the `svc_livedump` command, and then upload the log files and newly-generated dumps using the `svc_snap` command, as shown in Example 11-1. To verify if the support package was successfully uploaded, use the `sa info lscmdstatus` command (XXXXX,YYY,ZZZ is the PMR number).

Example 11-1 The `svc_livedump` command

```
IBM_FlashSystem:FLASHPFE95:superuser>svc_livedump -nodes all -yes
Livedump - Fetching Node Configuration
Livedump - Checking for dependent vdisks
Livedump - Check Node status
Livedump - Prepare 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 2
Livedump - Successfully captured livedumps on nodes 1,2
```

```
IBM_FlashSystem:FLASHPFE95:superuser>svc_snap upload pmr=XXXXX,YYY,ZZZ gui3
Collecting data
Packaging files
Snap data collected in /dumps/snap.ABCDEFG.171128.223133.tgz
```

```
IBM_FlashSystem:FLASHPFE95:superuser>sainfo lscmdstatus
last_command satask supportupload -pmr ppppp,bbb,ccc -filename
/dumps/snap.ABCDEFG.171128.223133.tgz
last_command_status CMMVC8044E Command completed successfully.
T3_status
T3_status_data
cpfiles_status Complete
cpfiles_status_data Copied 1 of 1
snap_status Complete
snap_filename /dumps/snap.ABCDEFG.171128.223133.tgz
installcanistersoftware_status
supportupload_status Complete
supportupload_status_data [PMR=ppppp,bbb,ccc] Upload complete
supportupload_progress_percent 0
supportupload_throughput_KBps 0
supportupload_filename /dumps/snap.ABCDEFG.171128.223133.tgz
downloadsoftware_status
downloadsoftware_status_data
downloadsoftware_progress_percent 0
downloadsoftware_throughput_KBps 0
downloadsoftware_size
IBM_FlashSystem:FLASHPFE95:superuser>
```

---

## 11.2.2 Host multipath software data collection

If a problem occurs that is related to host communication with IBM FlashSystem, collecting data from hosts and multipath software is very useful.

For more details, see the operating system-specific multipathing documentation.

The Subsystem Device Driver Device Specific Module (SDDDSM) and the Subsystem Device Driver Path Control Module (SDDPCM) have reached End of Service (EOS). Hence you should migrate SDDDSM to Microsoft Device Specific Module (MSDSM) on the Windows platform and SDDPCM to AIXPCM on the AIX/VIOS platforms respectively. The following websites provide more details:

- ▶ [Migrating from SDDDSM to Microsoft's MSDSM - SVC/Storwize](#)
- ▶ [How to Migrate SDDPCM to AIXPCM](#)

If the systems are still using SDD multipathing modules, note SDDPCM for AIX provides the **sddpcmgetdata** script to collect information used for problem determination. This script creates a tar file in the current directory with the current date and time as a part of the file name. When you suspect you have an issue with SDDPCM, it is essential to run this script and send this tar file to IBM support.

SDDDSM for Windows hosts also contains a utility to collect information for problem determination. The **sddgetdata.bat** tool creates a CAB file in the installation directory with the current date and time as part of the file name. The CAB file includes the following information:

- ▶ SystemInfo
- ▶ HKLM \SYSTEM\CurrentControlSet, HKLM\HARDWARE\DEVICEMAP, and HKLM\Cluster output from the registry
- ▶ SDDDSM directory contents
- ▶ HBA details
- ▶ Datapath outputs
- ▶ Pathtest trace
- ▶ SDDSRV logs
- ▶ Cluster logs
- ▶ System disks and paths

The execution and output of **sddgetdata.bat** tool is shown in Example 11-2.

*Example 11-2 The sddgetdata.bat tool*

---

```
C:\Program Files\IBM\SDDDSM>sddgetdata.bat
Collecting SDD trace Data

Flushing SDD kernel logs

SDD logs flushed

Collecting datapath command outputs

Collecting System Information

Collecting SDD and SDDsrV logs

Collecting Most current driver trace

Please wait for 30 secs... Writing DETAILED driver trace to trace.out

Generating a CAB file for all the Logs

sdddata_WIN-IWG6VLJN3U3_20171129_151423.cab file generated

C:\Program Files\IBM\SDDDSM>
```

---

For more information about diagnostics for IBM SDD, see [Latest Multipath Subsystem Device Driver User's Guide](#).

### 11.2.3 Additional 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 have to collect data from several devices in the SAN.

The following basic information should be collected for each type of device:

- ▶ Hosts
  - Operating system: Version and level
  - HBA: Driver and firmware level
  - Multipathing driver level
- ▶ SAN switches
  - Hardware model
  - Software version
- ▶ Storage subsystems
  - Hardware model
  - Software version

## 11.3 Common problems and isolation techniques

SANs, storage subsystems and host systems can be complicated. They often consist of hundreds or thousands of disks, multiple redundant subsystem controllers, virtualization engines, and different types of SAN switches. All of these components must be configured, monitored, and managed properly. If issues occur, administrators must know what to look for and where to look.

IBM FlashSystem has useful error logging mechanisms. It keeps track of 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. So, with these functions, administrators can easily locate any issue areas and take the necessary steps to fix any events.

In many cases, IBM FlashSystem and its service and maintenance features guide administrators directly, provide help, and suggest remedial action. Furthermore, IBM FlashSystem determines whether the problem still persists or not.

Another feature that helps administrators to isolate and identify issues that might be related to IBM FlashSystem is the ability of their nodes to maintain a database of other devices that communicate with the IBM FlashSystem device. Devices, such as hosts and optional back-end storages, are added or removed from the database as they start or stop communicating to IBM FlashSystem.

Although IBM FlashSystem 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. They need to have a troubleshooting performed outside of IBM FlashSystem, too. As an example, a misconfiguration in the SAN zoning might lead to the IBM FlashSystem cluster not working properly. This problem occurs because the IBM FlashSystem cluster nodes communicate with each other by using the Fibre Channel SAN fabrics.

In this case, check the following areas from an IBM FlashSystem perspective:

- ▶ The attached hosts. For more information, see 11.3.1, “Host problems” on page 490.
- ▶ The SAN. For more information, see 11.3.2, “SAN problems” on page 494.
- ▶ The optional attached storage subsystem. For more information, see 11.3.3, “Storage subsystem problems” on page 495.
- ▶ The local FC port masking. For more information, see 8.1.3, “Port masking” on page 350.



```
host_cluster_name
WWPN 100000051E0F81CD
node_logged_in_count 2
state active
WWPN 100000051E0F81CC
node_logged_in_count 0
state offline
```

---

► **lshostvdiskmap**

Check that all volumes are mapped to the correct hosts. If a volume is not mapped correctly, create the necessary host mapping.

► **lsfabric -host <host\_id\_or\_name>**

Use this command with parameter **-host <host\_id\_or\_name>** to display Fibre Channel (FC) connectivity between nodes and hosts. Example 11-5 shows the **lsfabric -host <host\_id\_or\_name>** command output.

*Example 11-5 lsfabric -host <host\_id\_or\_name> command*

---

```
IBM_FlashSystem:FLASHPFE95:superuser>lsfabric -host Win2K8
remote_wwpn remote_nportid id node_name local_wwpn local_port
local_nportid state name cluster_name type
10000090FAB386A3 502100 3 node1 5005076810120230 2 540200
inactive Win2K8 host
10000090FAB386A3 502100 1 node2 5005076810120242 2 540000
inactive Win2K8 host
```

---

To perform troubleshooting on the host side, check the following:

- Special software that you are using
- Recent changes in the OS, such as patching the OS, an upgrade, and so on
- Operating system version and maintenance or service pack level
- Multipathing type and driver level
- Host bus adapter model, firmware, and driver level
- Host bus adapter connectivity issues

Based on this list, the host administrator must check and correct any problems.

Hosts with higher queue depth can potentially overload shared storage ports. Hence it is recommended that you verify that the sum total of the queue depth of all hosts sharing a single target Fibre Channel port is limited to 2048. If any of the hosts have a queue depth of more than 128, that should be reviewed, as queue-full conditions could lead to I/O errors and extended error recoveries.

For more information about managing hosts on IBM FlashSystem, see Chapter 8, “Hosts” on page 349.

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, for a volume path issue reported by SDD output on the host by using the **datapath query adapter** and **datapath query device** commands. The adapter in degraded

state means that specific HBA on the server side can't reach all the nodes in the I/O group which the volumes are associated.

**Note:** SDDDSM and SDDPCM have reached EOS. Hence migrate SDDDSM to MSDSM on Windows platform and SDDPCM to AIXPCM on AIX/VIOS platforms, respectively.

For more information, see:

- ▶ [Migrating from SDDDSM to Microsoft's MSDSM - SVC/Storwize](#)
- ▶ [How to Migrate SDDPCM to AIXPCM](#)

Faulty paths can be caused by hardware and software problems, such as the following examples:

- ▶ Hardware
  - Faulty Small Form-factor Pluggable transceiver (SFP) on the host or SAN switch
  - Faulty fiber optic cables
  - Faulty HBAs
- ▶ Software
  - A back-level multipathing driver
  - Obsolete HBA firmware or driver
  - Wrong zoning
  - Incorrect host-to-VDisk mapping

Based on field experience, it is recommended that you complete the following hardware checks first:

- ▶ Check whether connection error indicators are lit on the host or SAN switch.
- ▶ Check 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 broken. If possible, swap the cables with cables that are known to work.

After the hardware check, continue to check the following aspects of software setup:

- ▶ Check that the HBA driver level and firmware level are at the preferred and supported levels.
- ▶ Check the multipathing driver level, and make sure that it is at the preferred and supported level.
- ▶ Check for link layer errors that are reported by the host or the SAN switch, which can indicate a cabling or SFP failure.
- ▶ Verify your SAN zoning configuration.
- ▶ Check the general SAN switch status and health for all switches in the fabric.

## **iSCSI/iSER configuration and performance issues**

This section describes the Internet Small Computer Systems Interface (iSCSI) and iSCSI Extensions for RDMA (iSER) configuration and performance issues.

### ***Link issues***

If the Ethernet port link does not come online, then check if the SFP/Cables and check if the port supports auto-negotiation with the switch. This is especially true for SFPs which support 25G and higher, as there could be potential mismatch in Forward Error Correction (FEC) which may prevent a port to auto-negotiate.

Longer cables get exposed to more noise or interference (high Bit Error Ratio [BER]), hence they require more powerful error correction codes.

There are two IEEE 802.3 FEC specifications. For an auto-negotiation issue, please verify if 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 i.e. RS-FEC (22.4 dB loss specification)

### ***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 protocol (DCBX) supported switches, verify that 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 in the range 0 to 7.

A DCBX-enabled switch and a storage adapter exchange parameters that describe traffic classes and PFC capabilities.

In the IBM FlashSystem, Ethernet traffic is divided into three Classes of Service based on feature use case:

- ▶ Host attachment (iSCSI/iSER)
- ▶ Backend Storage (iSCSI)
- ▶ Node-to-node communication (Remote Direct Memory Access (RDMA) clustering)

If there are challenges while configuring PFC, verify the following attributes to determine the issue.

- ▶ Configure IP/VLAN using **cfgport ip**.
- ▶ Configure COS (class of service) using **chsytsemethernet**.
- ▶ Ensure that the priority tag is enabled on the switch.
- ▶ Ensure that **lspport ip** output shows: `dcbx_state`, `pfc_enabled_tags`.
- ▶ Enhanced Transmission Selection (ETS) settings is recommended if a port is shared.

### ***Standard network connectivity check***

Verify that the required TCP/UDP ports are allowed in the network firewall. A list of ports for various host attachments follows:

- ▶ Software iSCSI requires TCP Port 3260.
- ▶ iSER/RoCE host requires 3260.
- ▶ iSER/iWRAP host requires TCP Port 860.

Verify the IP addresses are reachable and the TCP ports are open.

### ***iSCSI performance issues***

In certain situations, the TCP/IP layer may try to combine several ACK responses together into a single response to improve performance, but that 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 very low queue depth.

It is recommended 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 if a smaller “Max I/O Transfer Size” with more concurrency (queue

depth >16) improves overall latency and bandwidth utilization for the specific host workload. In most Linux distributions this is controlled by the `max_sectors_kb` parameter with a suggested transfer size of 32kB.

In addition, you should review network switch diagnostic data to evaluate packet drop or retransmission in the network. Therefore, it is advisable to enable flow control or PFC to enhance the reliability of the network delivery system, to avoid packet loss, which enhances storage performance.

### 11.3.2 SAN problems

It is not a difficult task to introduce IBM FlashSystem into your SAN environment and to use its virtualization functions. However, before you can use IBM FlashSystem 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.

Two types of SAN zones are needed to run IBM FlashSystem in your environment: A *host zone* and a *storage zone* for optional external-attached storage. In addition, you must have an IBM FlashSystem zone that contains all the IBM FlashSystem node ports of the IBM FlashSystem cluster. This IBM FlashSystem zone enables intra-cluster communication. For more information and important points about setting up IBM FlashSystem in a SAN fabric environment, see Chapter 2, “Storage area network” on page 33.

Because IBM FlashSystem is a major component of the SAN and connects the host to the storage subsystem, you should check and monitor the SAN fabrics.

Some situations of performance degradation and buffer-to-buffer credit exhaustion can be caused by incorrect local FC port masking and remote FC port masking. To ensure healthy operation of your IBM FlashSystem, configure both your local FC port masking and your remote FC port masking accordingly.

The ports intended to have only intracluster/node-to-node communication traffic must not have replication data or host/back-end data running on it. The ports intended to have only replication traffic must not have intracluster/node-to-node communication data or host/back-end data running on it.

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. But before you take any other steps, check the installed code level for any known issues. Also, check whether a new code level is available that resolves the problem that you are experiencing.

The most common SAN issues often are related to zoning. For example, perhaps you chose the wrong WWPN for a host zone, such as when two IBM FlashSystem node ports must be zoned to one HBA with one port from each IBM FlashSystem node. As shown in Example 11-6, two ports are zoned that belong to the same node. Therefore, the result is that the host and its multipathing driver do not see all of the necessary paths.

*Example 11-6 Incorrect WWPN zoning*

---

```
zone: Senegal_Win2k3_itsosvcc11_iogrp0_Zone
 50:05:07:68:10:20:37:dc
 50:05:07:68:10:40:37:dc
 20:00:00:e0:8b:89:cc:c2
```

---

The correct zoning must look like the zoning that is shown in Example 11-7.

*Example 11-7 Correct WWPN zoning*

---

```
zone: Senegal_Win2k3_itsosvcc11_iogrp0_Zone
 50:05:07:68:10:40:37:e5
 50:05:07:68:10:40:37: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 related to SAN switches. The bottleneck can be present in a port where a host, storage subsystem or IBM Spectrum Virtualize device is connected, or in Inter-Switch Link (ISL) ports. The bottleneck can occur in some cases, such as when a device connected to the fabric is slow to process received frames or if a SAN switch port is unable to transmit frames at a rate that is required by a device connected to the fabric.

These cases can slow down communication between devices in your SAN. To resolve this type of issue, refer to the SAN switch documentation or open a case with the vendor to investigate and identify what is causing the bottleneck and fix it.

If you cannot fix the issue with these actions, use the method that is described in 11.2, “Diagnostic data collection” on page 485, collect the SAN switch debugging data, and then contact the vendor for assistance.

### 11.3.3 Storage subsystem problems

Today, various heterogeneous storage subsystems are available. All of 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, following the respective preferred practices and with no existing issues.

Check the following areas if you experience a storage-subsystem-related issue:

- ▶ Storage subsystem configuration. Ensure that a valid configuration and preferred practices are applied to the subsystem.
- ▶ Storage subsystem node controllers. Check the health and configurable settings on the node controllers.
- ▶ Storage subsystem array. Check the state of the hardware, such as a FCM's, SSD's failures or enclosure alerts.
- ▶ Storage volumes. Ensure that the logical unit number (LUN) masking is correct.
- ▶ Host attachment ports. Check the status, configuration and connectivity to SAN switches.
- ▶ Layout and size of RAID arrays and LUNs. Performance and redundancy are contributing factors.

IBM FlashSystem has several CLI commands that you can use to check the status of the system and attached optional storage subsystems too. Before you start a complete data collection or problem isolation on the SAN or subsystem level, use the following commands first and check the status from the IBM FlashSystem perspective:

▶ **lsmdisk**

Check that all MDisks are online (not degraded or offline).

▶ **lsmdisk <MDisk\_id\_or\_name>**

Check several of the MDisks from each storage subsystem controller. Are they online? See Example 11-8 for an example of the output from this command.

*Example 11-8 Issuing an lsmdisk command*

---

```
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_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 11-8 on page 496 shows that for MDisk 1, the external storage controller has eight ports zoned to IBM FlashSystem, and IBM FlashSystem has two nodes, so 8 x 2= 16.

► **lsvdisk**

Check that all volumes are online (not degraded or offline). If the volumes are degraded, are there stopped FlashCopy jobs? Restart stopped FlashCopy jobs or seek IBM FlashSystem support guidance.

► **Isfabric**

Use this command with the various options, such as **-controller controllerid**. Also, check different parts of the IBM FlashSystem configuration to ensure that multiple paths are available from each IBM FlashSystem node port to an attached host or controller. Confirm that IBM FlashSystem node port WWPNs are also consistently connected to the optional external back-end storage.

## Determining the number of paths to an external storage subsystem

By using IBM FlashSystem CLI commands, it is possible to determine the total number of paths to an optional external storage subsystem. To determine the proper value of the available paths, use the following formula:

Number of MDisks x Number of FlashSystem nodes per Cluster = Number of paths  
MDisk\_link\_count x Number of FlashSystem nodes per Cluster = Sum of path\_count

Example 11-9 shows how to obtain this information by using the **lscontroller <controllerid>** and **svcinfo lsnode** commands.

*Example 11-9 Output of the svcinfo lscontroller command*

---

```
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_FlashSystem:FLASHPFE95:superuser>svcinfolnode
```

```

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_FlashSystem:FLASHPFE95:superuser>

```

---

Example 11-9 on page 498 shows that 16 MDisks are present for the external storage subsystem controller with ID 1, and two IBM FlashSystem nodes are in the cluster. In this example, the path\_count is 16 x 2 = 32.

Further information about FC and iSCSI configurations can be found in the IBM Documentation.

IBM FlashSystem has useful tools for finding and analyzing optional back-end storage subsystem issues because it has a monitoring and logging mechanism.

Typical events for storage subsystem controllers include incorrect configuration, which results in a *1625 - Incorrect disk controller configuration* error code. Other issues related to the storage subsystem include failures pointing to the managed disk 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 explicit reason for being issued. Therefore, you must check multiple areas for issues, not just the storage subsystem.

To determine the root cause of a problem, complete the following tasks:

1. Check the Recommended Actions panel by clicking **Monitoring** → **Events**.
2. Check the attached storage subsystem for misconfigurations or failures:
  - a. Independent of the type of storage subsystem, first check whether the system has any unfixed errors. Use the service or maintenance features that are provided with the storage subsystem to fix these issues.
  - b. Check if volume mapping is correct. The storage subsystem LUNs should be mapped to a host object with IBM FlashSystem ports. For the IBM FlashSystem restrictions for optional back-end storage subsystems, see [V8.4.0.x Configuration Limits and Restrictions for IBM FlashSystem 9200](#).

If you need to identify which of the externally attached MDisk has which corresponding LUN ID, run the IBM FlashSystem `lsmdisk` CLI command as shown in Example 11-10. This command also shows to which storage subsystem a specific MDisk belongs (the controller ID).

*Example 11-10 Determining the ID for the MDisk*

---

```

IBM_FlashSystem:FLASHPFE95:superuser>lsmdisk
id name status mode MDisk_grp_id MDisk_grp_name capacity ctrl_LUN_#
controller_name UID tier encrypt site_id site_name distributed dedupe
over_provisioned supports_unmap
0 MDisk0 online array 0 Pool0 198.2TB
tier0_flash no yes no yes
yes
0 MDisk1 online managed 0 MDG-1
600.0GB 0000000000000000 controller0
600a0b800017423300000059469cf84500000000000000000000000000000000
2 MDisk2 online managed 0 MDG-1
70.9GB 00000000000000002 controller0
600a0b800017443100000096469cf0e800000000000000000000000000000000

```

---

3. Check the SAN environment for switch problems or zoning failures.

Make sure the zones are properly configured, and the zoneset is activated. The zones that allow communication between the storage subsystem and the IBM FlashSystem device should contain WWPNs of the storage subsystem and WWPNs of IBM FlashSystem.

4. Collect all support data and contact IBM Support.

Collect the support data for the involved SAN, IBM FlashSystem, or optional external storage systems as described in 11.2, “Diagnostic data collection” on page 485.

### 11.3.4 Native IP replication problems

The native IP replication feature uses the following TCP/IP ports for remote cluster path discovery and data transfer:

- ▶ 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.

If the IP addresses are reachable and TCP ports are open, verify if 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. Once the connection is made, the error will auto-correct.

Network quality of service largely influences the effective bandwidth utilization of the dedicated link between the cluster. Bandwidth utilization is inversely proportional to round trip time (RTT) and rate of packet drop/retransmission in the network. For standard block traffic, a packet drop/retransmission of 0.5% or more may lead to unacceptable utilization of the available bandwidth. Work with network team to investigate either over-subscription or other

quality-of-service of the link, with an objective of having the lowest possible (less than 0.1%) packet-drop percentage.

### 11.3.5 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 the CPU intervention while processing the I/O leading to lower latency and a faster rate of data transfer.

IBM FlashSystem Cluster can be formed using RDMA-capable NICs which use RoCE or iWARP technology.

- ▶ Inter-node Ethernet connectivity can be done only over identical ports; such ports must be connected within the same switching fabric.
- ▶ If the cluster is to be created without any ISL (up to 300 meters) then deploy Independent (isolated) switches.
- ▶ If the cluster is to be created on short-distance ISL (upto 10km) then provision as many ISLs between switches as RDMA-capable cluster ports.
- ▶ For long-distance ISL (up to 100km) DWDM and CWDM methods are applicable for L2 networks. Packet switched or VXLAN methods are deployed for L3 network as this equipment comes with deeper buffer “pockets”.

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

The first step to review if the node IP address is reachable and verify the required TCP/UDP ports are accessible in both directions. The following CLI output could be helpful to find the reason for connectivity error:

```
sainfo lsnodeipconnectivity
```

### 11.3.6 Advanced Copy services problems

Performance of a specific storage feature or overall storage subsystem is generally interlinked, meaning a bottleneck in one SW/HW layer can potentially propagate to other layers. Therefore, problem isolation is a critical part of performance analysis.

The first thing to check is if 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/estimation of advanced copy services (replication, FlashCopy, volume mirroring), or I/O load change, due to hardware component failure.

#### Remote Copy

Disturbances in the SAN or Wide Area Network (WAN) can cause congestion and packet drop, which can impact Metro Mirror (MM) or Global Mirror (GM) traffic. As host I/O latency is dependent on MM or GM I/O completion to the remote cluster, a host will potentially

experience high latency. Based on various parameters replication could be operatively stopped to protect host. The following conditions can affect GM/MM:

- ▶ Network congestion or fluctuation. Fix the network. Additionally, verify that port masking is enabled, so that the congestion in replication ports does not affect clustering or host/storage ports.
- ▶ Overload of secondary/primary cluster. Monitor and throttle the host which causes the condition.
- ▶ High background copy rate which leaves less bandwidth to replicate foreground host I/O. Adjust the background copy rate so that the link does not get oversubscribed.
- ▶ A large Global Mirror with Change Volumes (GMCV) consistency group could potentially introduce hundreds of milliseconds of pause when the replication cycle starts. Reduce the number of relationships in a consistency group if the observed I/O pause is not acceptable.

## HyperSwap

Verify that the link between the sites is stable and has enough bandwidth to replicate the peak workload. Additionally, check if a volume has to frequently change the replication direction from one site to other. This happens when a specific volume is being written by hosts from both the sites. Evaluate if this can be avoided to reduce frequent direction change. You could ignore it if the solution is designed taking active/active access into account.

If a single volume resynchronization between the sites takes a very long time, review the partnership `link_bandwidth_mbits` and per relationship `bandwidth_limit` parameters.

## FlashCopy

Consider the following points for FlashCopy troubleshooting:

- ▶ Verify that the preferred node of FlashCopy source and target volumes is the same to avoid excessive internode communications.
- ▶ High background copy rate and clean rate of FlashCopy relations could 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. The larger the number of relationships, the higher the I/O pause time (Peak I/O Latency) when the CG starts.
- ▶ If the host I/O pattern is small and random, then evaluate if reducing the FlashCopy grain size to 64 KB provides any improvement in latency compared to the default grain size of 256 KB.

## Compression

Compress a volume if the data is compressible. There is no benefit to compressing a volume where compression saving is less than 25% as that could potentially reduce the overall performance of the Random Access Compression Engine (RACE). If the I/O access pattern is sequential, that may not be suitable candidate for RACE. Use the Comprestimator or Data Reduction Estimation Tool to size the workload.

## Volume mirroring

Write-performance of the mirrored volumes will be dictated by the slowest copy. Reads are served from the Copy0 of the volume (in the case of a stretched cluster topology, both the copies could serve reads, dictated by the host site attribute). Therefore, you should size the solution accordingly. Note that the mirroring layer maintains a bitmap copy on the



quorum device therefore any unavailability of the quorum will take the mirroring volumes offline. Similarly, slow access to the quorum could also impact the performance of mirroring volumes.

### **Data reduction pools**

Data reduction pools (DRPs) internally implement a log structured array (LSA) which means that writes (new or over-writes/updates) always allocate newer storage blocks. The older blocks (with invalid data) are marked for garbage collection at a later time. 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 only having to move a little 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 needs to speed up and to move valid data more aggressively to make space available sooner. This might lead to increased latency due to increased CPU utilization and load on the back-end. Therefore, it is recommended to manage storage provisioning to avoid such scenarios.

Users are encouraged to pay particular attention to any GUI notifications and employ best practices for managing physical space. Use data reduction only at one layer (at the virtualization layer or the back-end storage or drives) as there is no benefit to compress and deduplicate the same data twice.

Encrypted data cannot be compressed, hence data reduction needs to be done before the data is encrypted. Proper sizing is very important to get best of performance from data reduction, hence you should use data reduction estimation tools to evaluate system performance and space saving.

Note that IBM FlashSystem uses two types of data reduction techniques:

- IBM FlashSystem using the FCM NVMe drives have built-in hardware compression.
- IBM FlashSystem using industry standard NVMe drives rely on the Spectrum Virtualize software and DRP pools to deliver data reduction.

### **11.3.7 Health status during upgrade**

It is important to understand that during the software upgrade process, alerts that indicate the system is not healthy are reported. This is a normal behavior because the IBM FlashSystem node canisters go offline during this process, so the system triggers these alerts.

While trying to upgrade an IBM FlashSystem, you might also get a message such as an error in verifying the signature of the update package.

This message does not mean that you have an issue in your system. Sometimes this happens because there is not enough space on the system to copy the file, or the package is incomplete or contains errors. In this case, open a PMR with IBM support and follow their instructions.

### **11.3.8 Managing physical capacity of over provisioned storage controllers**

Drives and back-end controllers exist that have built-in hardware compression and other data reduction technologies which allows capacity to be provisioned over and above the available real physical capacity. Different data sets lead to different capacity savings and some data, such as encrypted data or already compressed data, will not even compress. When the physical capacity savings does not match the expected or provisioned capacity, the storage may run out of physical space leading to a write-protected drive/array.

To avoid running out of space on the system, the usable capacity should be carefully monitored on the GUI of the IBM FlashSystem. The IBM FlashSystem GUI is the only capacity dashboard that shows the physical capacity.

Monitoring is especially important when migrating substantial amounts of data onto the IBM FlashSystem, which typically happens during the first part of the workload life cycle as data is on-boarded, or initially populated into the storage system. IBM strongly encourages users to configure Call Home on the IBM FlashSystem. Call Home monitors the physical free space on the system and will automatically open a service call for systems that reach 99% of their usable capacity.

IBM Storage Insights also has the ability to monitor and report on any potential out of space conditions and the new Advisor function will warn when the IBM FlashSystem almost at full capacity. For more information, see 11.6.5, “SI Advisor” on page 523

When IBM FlashSystem reaches an out of space condition, the device will drop into a read-only state. An assessment of the data compression ratio and the re-planned capacity estimation should be done to determine how much actual outstanding storage demand might exist. This additional capacity will need to be prepared and presented to the host so that recovery can begin.

The approaches that can be taken to reclaim space on the IBM FlashSystem in this scenario vary by the capabilities of the system, optional external back-end controllers, the system configuration, and pre-planned capacity overhead needs.

Generally speaking, the following options are available:

- ▶ Add additional capacity to the IBM FlashSystem. Customers should have a plan that allows them to add additional capacity to the system when needed.
- ▶ Reserve a set of space in the IBM FlashSystem that makes it “seem” fuller than it really is, and that you can free up in an emergency situation. IBM FlashSystem has the ability to create a volume which isn't compressed, de-duped or thin provisioned (a fully allocated volume). Simply create some of these volumes to reserve an amount of physical space. You can name them something like “emergency buffer space”. If you are reaching the limits for physical capacity, you can simply delete one or more of these volumes to give yourself a temporary reprieve.

**Important:** Running completely out of space can be a very serious situation. Recovery can be extremely complicated and time-consuming. For this reason, it is imperative that proper planning and monitoring be done to avoid reaching this condition.

The following sections describe the process for recovering from an out of space condition.

## Reclaiming and unlocking

After you've assessed and accounted for storage capacity, the first step is to contact IBM Support who can aid in unlocking the read-only mode and restoring operations. The reclamation task can take a long time to run, and larger flash arrays will of course take longer to recover than smaller ones.

## Freeing up space

You can reduce the amount of consumed space once the IBM FlashSystem has been unlocked by IBM support, in several ways.

To recover from Out of Space conditions on Standard Pools, these are the steps for the user:

1. Add more storage to the system if possible.
2. Migrate extents from the write protected array to other non-write protected MDisks with enough extents. This could be an external back-end storage array.
3. Migrate volumes with extents on the write protected array to another pool. If possible, moving volumes from the IBM FlashSystem pool to another external pool can free up space in the IBM FlashSystem pool to allow for space reclamation. As this volume moves into the new pool, its previously occupied flash extends will be freed up (using SCSI unmap), which then goes to provide more free space to the IBM FlashSystem enclosure to be configured to a proper provisioning to support the compression ratio.
4. Delete dispensable volumes to free-up space. If possible, within the pool (managed disk group) on the IBM FlashSystem, delete unnecessary volumes. The IBM FlashSystem supports SCSI unmap so deleting volumes will have space reclamation benefits using this method.
5. Bring the volumes in the pool back online using a Directed Maintenance Procedure.

Further information on types of recovery can be found in the IBM Support Technote at [Handling out of physical space conditions](#).

### 11.3.9 Replacing a failed flash drive

When IBM FlashSystem detects a failed NVMe drive or optional external attached flash drive, it automatically generates an error in the *Events* panel. To replace the failed drive, always run **Fix Procedure** for this event in the **Monitoring** → **Events** panel.

The **Fix Procedure** will help you to identify the enclosure and slot where the bad drive is located, and will guide you to the correct steps to follow in order to replace it. When a flash drive fails, it is removed from the array. If a suitable spare drive is available, it is taken into the array and the rebuild process starts on this drive.

After the failed flash drive is replaced and the system detects the replacement, it reconfigures the new drive as spare. So, the failed flash drive is removed from the configuration, and the new drive is then used to fulfill the array membership goals of the system.

### 11.3.10 Recovering from common events

You can recover from several of the more common events that you might encounter. In all cases, you must read and understand the current product limitations to verify the configuration and to determine whether you need to upgrade any components or install the latest fixes or patches.

To obtain support for any IBM product, see the [IBM Support Homepage](#).

**c**If the problem is caused by IBM Spectrum Virtualize and you are unable to fix it by using the Recommended Action feature or by examining the event log, collect the IBM Spectrum Virtualize support package as described in 11.2.1, “IBM FlashSystem data collection” on page 485. To identify and fix other issues outside of IBM Spectrum Virtualize, consider the guidance in the other sections in this chapter that are not related to IBM Spectrum Virtualize.

## 11.4 Remote Support Assistance

Remote Support Assistance (RSA) enables IBM support to access the IBM FlashSystem device to perform troubleshooting and maintenance tasks. Support assistance can be configured to support personnel work on-site only, or to access the system both on-site 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 in the GUI, select **Settings** → **Support** → **Support Assistance** → **Reconfigure Settings**, as shown in Figure 11-7.

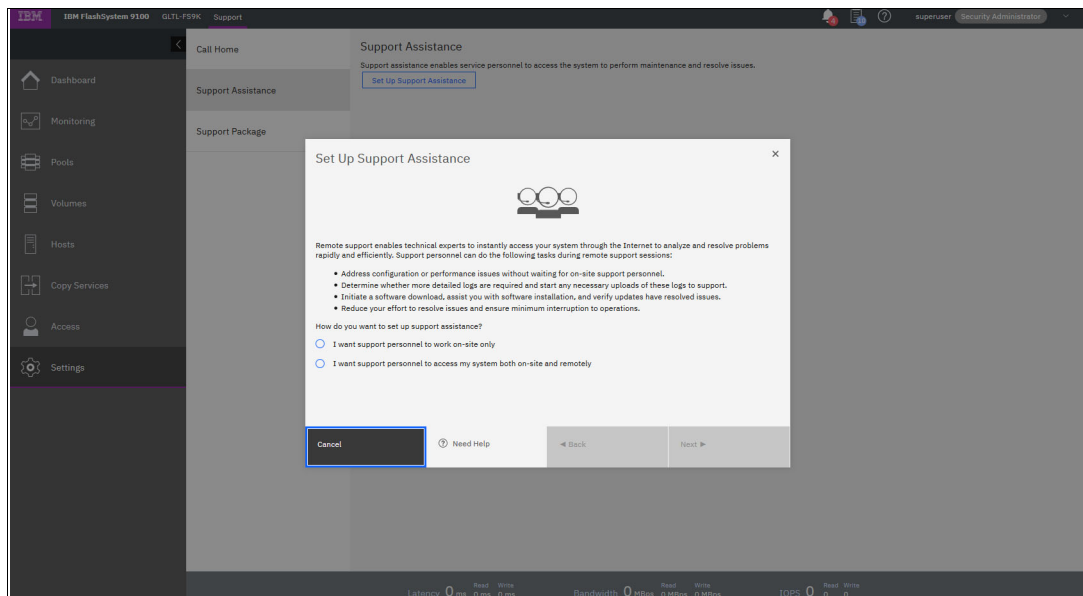


Figure 11-7 Remote Support options

You can use just local support assistance if you have security restrictions that don't allow support to connect remotely to your systems. With RSA, support personnel can work both on-site and remotely through 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 remote support assistance in the GUI, local support assistance is also enabled.

The following access types are in the remote support assistance method:

- ▶ At any time  
Support center can start remote support sessions at any time.
- ▶ By permission only  
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 purchased with a three-year warranty include Enterprise Class Support (ECS) and are entitled to IBM support using Remote Support Assistance to quickly connect and diagnose problems. However, IBM support might choose to utilize this feature on non-ECS systems at their discretion, therefore we recommend configuring and testing the connection on all systems.

To configure remote support assistance, the following prerequisites should be met:

- ▶ Ensure that Call Home is configured with a valid email server.
- ▶ Ensure that a valid service IP address is configured on each node on the system.
- ▶ If your system is behind a firewall or if you want to route traffic from multiple storage systems to the same place, you must configure a Remote Support Proxy server. Before you configure remote support assistance, the proxy server must be installed and configured separately. The IP address and the port number for the proxy server needs to be set-up on when enabling remote support centers.
- ▶ For more information about setting up the Remote Proxy Server, see [IBM FlashSystem 9200 8.4.0 Documentation - Configuring Remote Support Proxy](#).
- ▶ If you do not have firewall restrictions and the storage nodes are directly connected to the Internet, request your network administrator to allow connections to 129.33.206.139 and 204.146.30.139 on port 22.
- ▶ Both 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. The Remote Proxy Server cannot be used to download files.
- ▶ To ensure that support packages are uploaded correctly, configure the firewall to allow connections to the following IP addresses on port 443: 129.42.56.189, 129.42.54.189, and 129.42.60.189.
- ▶ To ensure that software is downloaded correctly, configure the firewall to allow connections to the following IP addresses on port 22: 170.225.15.105, 170.225.15.104, 170.225.15.107, 129.35.224.105, 129.35.224.104, and 129.35.224.107.

Remote support assistance can be configured using both GUI and CLI. The detailed steps to configure it can be found in *IBM FlashSystem 9100 Architecture, Performance, and Implementation*, SG24-8425

## 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 feature to view Call Home information on the web.

Call Home is a functionality present in several IBM systems, including IBM FlashSystem, which allows them to automatically report problems and status to IBM.

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

To access the Call Home Connect Cloud (Call Home Web), go to [IBM Support Homepage](#).

In the IBM support website, Call Home Web is available at **My support** → **Call Home Web** as shown in Figure 11-8 on page 508.

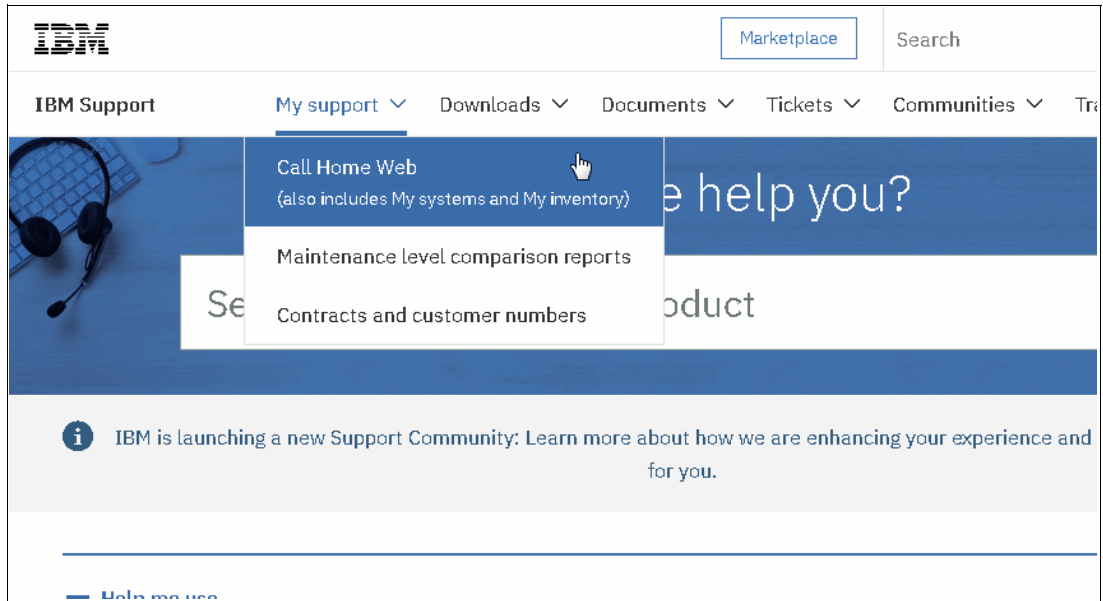


Figure 11-8 Call Home Web

Call Home Web has been replaced by the new Call Home Connect Cloud web application. The new cloud-based application provides an enhanced, live view of your assets, including the status of cases, warranties, maintenance contracts, service levels, end of service information, and other online tools.

To allow Call Home Connect Cloud analyze data of IBM FlashSystem systems and provide useful information about them, the devices need to be added to the tool. The machine type, model and serial number are required to register the product in Call Home Web. Also, it is required that IBM FlashSystem have Call Home and inventory notification enabled and operational.

Figure 11-9 on page 509 shows the Call Home Connect Cloud details panel of IBM FlashSystem.

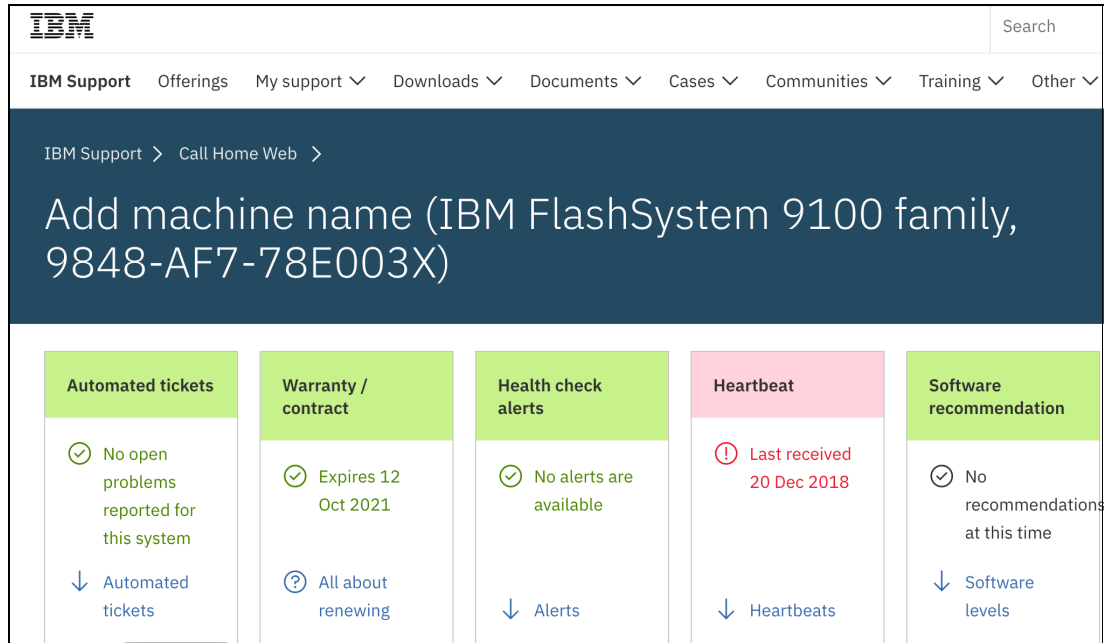


Figure 11-9 Call Home Web details panel

For a guide on how to setup 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, a tool that runs in the IBM Cloud.

It analyzes Call Home and inventory data of systems registered in Call Home Connect Cloud and validates their configuration. Then, it displays alerts and provide 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 related to them.

Some of the function of the IBM Call Home Connect Cloud and Health Checker have been ported to IBM Storage Insights, which is explained in detail in 11.6, “IBM Storage Insights” on page 509.

## 11.6 IBM Storage Insights

IBM Storage Insights is an integral part of the monitoring and ensuring continued availability of the IBM FlashSystem.

Available at no charge, cloud-based IBM Storage Insights provides a single dashboard that gives you a clear view of all of your IBM block storage. You’ll be able to make better decisions by seeing trends in performance and capacity. Storage health information enables you to focus on areas needing attention.

In addition, when IBM support is needed, 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 some of those included:

- ▶ A unified view of IBM systems:
  - Provides a single pane to see all of your system's characteristics
  - See all of your IBM storage inventory
  - Provides a live event feed so you know, up to the second, what is going on with your storage, which enables you to act fast.
- ▶ IBM Storage Insight 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
  - Monitor the configuration to see if it meets the best practices
  - System resource management: determine if the system is being overly taxed and provide proactive recommendations to fix it
- ▶ Storage Insights provides advanced customer service with an event filter that enables the following functions:
  - The ability for you and support to view support tickets, open and close them, and track trends
  - Auto log collection capability to enable you to collect the logs and send them to IBM before support starts looking into the problem. This can save as much as 50% of the time to resolve the case

In addition to the free Storage Insights, there is also the option of Storage Insights Pro, which is a subscription service that provides longer historical views of data, offers more reporting and optimization options, and supports IBM file and block storage together with EMC VNX and VMAX.

Figure 11-10 on page 511 shows the comparison of Storage Insights and Storage Insights Pro.



| Product Comparison      |                                                                 | IBM Storage Insights<br>(Free) | IBM Storage Insights Pro<br>(Subscription) |
|-------------------------|-----------------------------------------------------------------|--------------------------------|--------------------------------------------|
|                         | Capability                                                      |                                |                                            |
| <b>Monitoring</b>       | 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                         |                                | ✓                                          |
| <b>Support Services</b> | Simplified ticketing / log workflows and ticket history         | ✓                              | ✓                                          |
|                         | Proactive notification of risks (select systems)                | ✓                              | ✓                                          |
| <b>Device Analytics</b> | Part failure prediction                                         | ✓                              | ✓                                          |
|                         | Configuration best practice                                     | ✓                              | ✓                                          |
|                         | Customized upgrade recommendation                               | ✓                              | ✓                                          |
| <b>TCO Analytics</b>    | Capacity planning                                               |                                | ✓                                          |
|                         | Performance planning                                            |                                | ✓                                          |
|                         | Application / server storage consumption                        |                                | ✓                                          |
|                         | Capacity optimization with reclamation planning                 |                                | ✓                                          |
|                         | Data optimization with tier planning                            |                                | ✓                                          |

Figure 11-10 Storage Insights versus Storage Insights Pro comparison

Storage Insights provides a very lightweight data collector that is deployed on a customer supplied server. This can be either a Linux, Windows, or AIX server, or a guest in a virtual machine (for example, a VMware guest).

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.

### What metadata is collected

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 compression ratio.
- ▶ Performance metrics, such as read and write data rates, I/O rates, and response times.
- ▶ The actual application data that is stored on the storage systems cannot be accessed by the data collector

### Who can access 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 Storage Insights Customer Dashboard

Figure 11-11 shows a view of the Storage Insights (SI) main dashboard and the systems that it is monitoring.

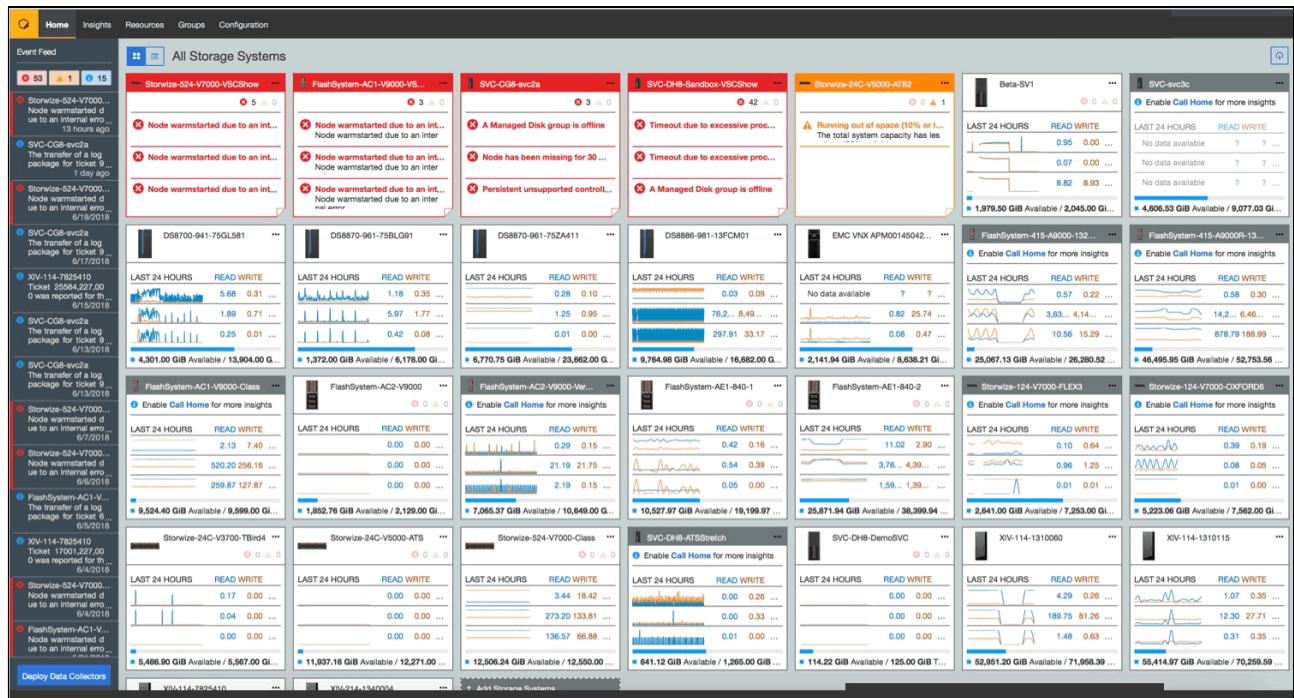


Figure 11-11 Storage Insights Main Dashboard

### 11.6.2 Customized dashboards to monitor your storage

With the latest release of IBM Storage Insights (SI) you are able to customize the dashboard to only show a subset of the systems monitored. This is useful for customers that might be Cloud Service Providers (CSP) and only want a specific end-user to see those machines that they are paying for.

For further details on setting up the customized dashboard, see:

[IBM Storage Insights Documentation - Creating customized dashboards to monitor your storage](#)

### 11.6.3 Creating support tickets

IBM Storage Insights also has the ability to update, from the Dashboard GUI, support tickets for any of the systems it reports about.

To do this, go to the SI dashboard and then choose the system you want to update the ticket for. From this screen select **Actions** → **Create/Update Ticket**.

Figure 11-12 shows how to create or update a support ticket from the SI dashboard.

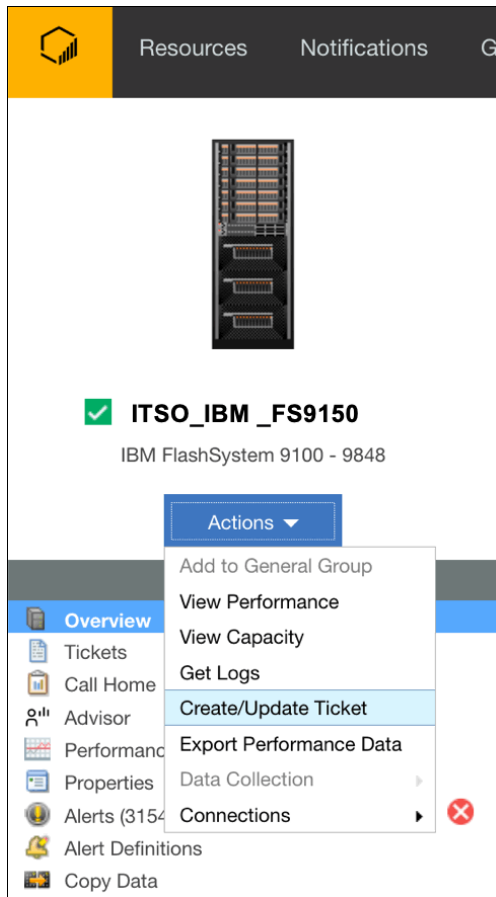


Figure 11-12 SI Create / Update a support Ticket

6. Figure 11-13 on page 514 shows you the panel where you can either create a new ticket or update a previously created one.

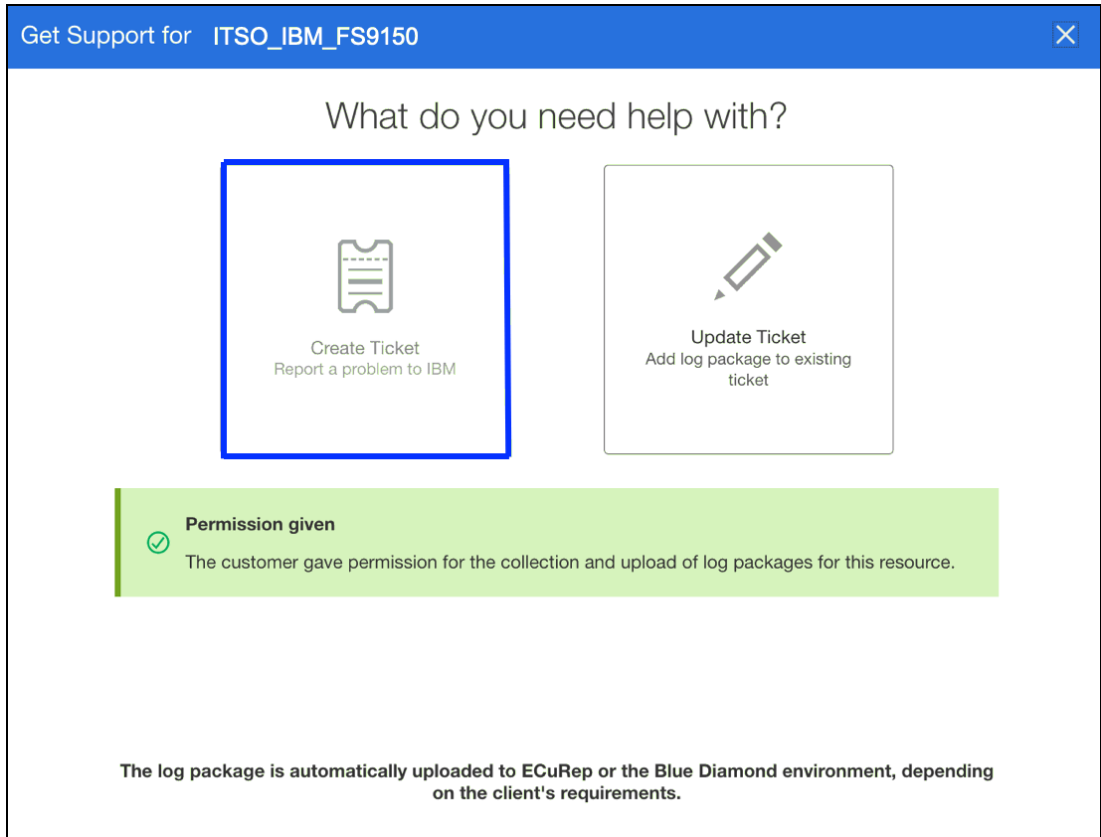


Figure 11-13 Create ticket

**Note:** The *Permission given* information box, shown in Figure 11-13, is an option the customer needs to enable on in the IBM FlashSystem GUI. See the 11.4, "Remote Support Assistance" on page 506 to enable this function.

7. Select the **Create Ticket** option and you will be presented with the following screens to complete with the machine details, problem description and the option to upload logs.

Figure 11-14 shows the ticket data collection done by the SI application.

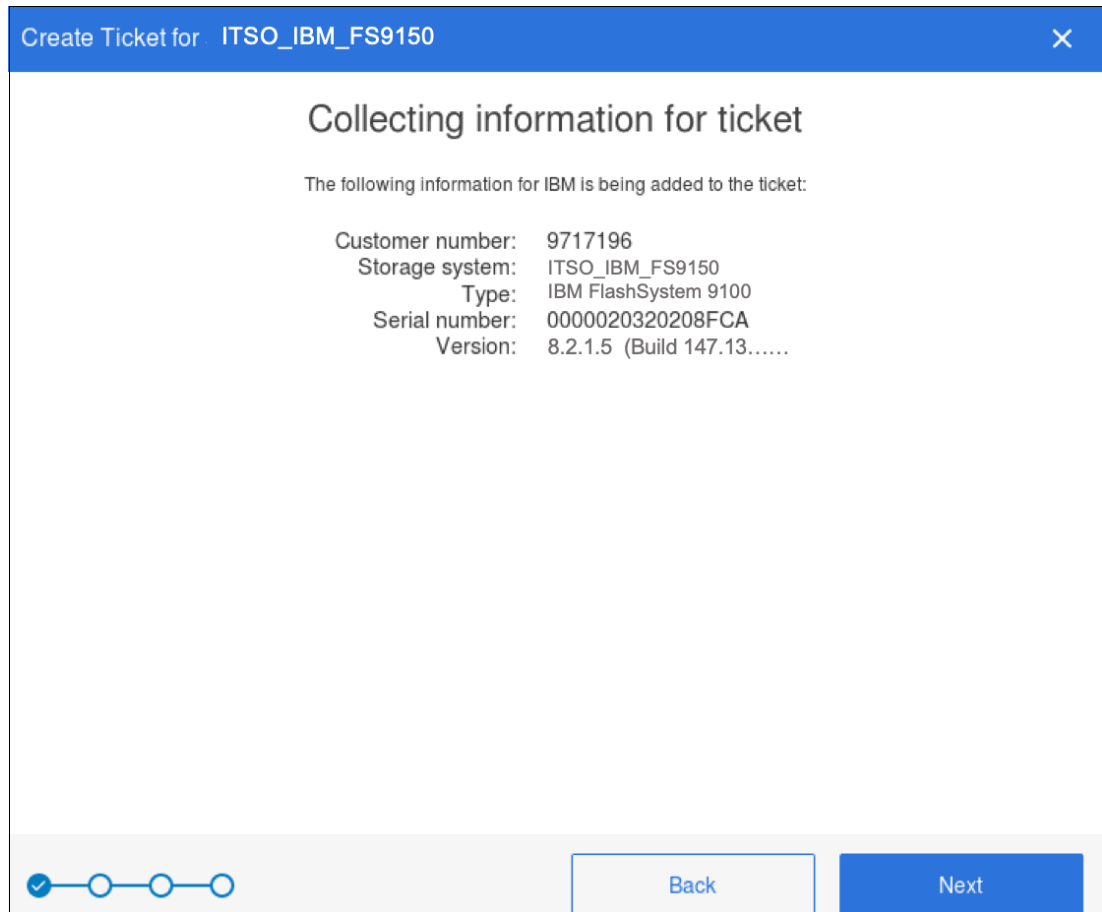


Figure 11-14 Collecting ticket information

8. Figure 11-15 on page 516 allows you to add a problem description and also attach additional files, such as error logs or screen grabs of error messages, and so on.

Create Ticket for ITSO\_IBM\_FS9150 ×

### Add a note or attachment

38

Device is showing a failure alert.

Hint: Include what happened and the error code, if any.

3000

(Optional) Type a note to add to the ticket

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.

**Attach Image or File:**

Browse

OR

↑  
Drag file here

✓
✓
○
○

Back

Next

Figure 11-15 Adding problem description and any additional information

9. prompts you to Set a severity level for the ticket, as shown in Figure 11-16 on page 517. Severity levels range from severity 1 (for a system down or extreme business impact) to severity 4 (for non-critical issues).

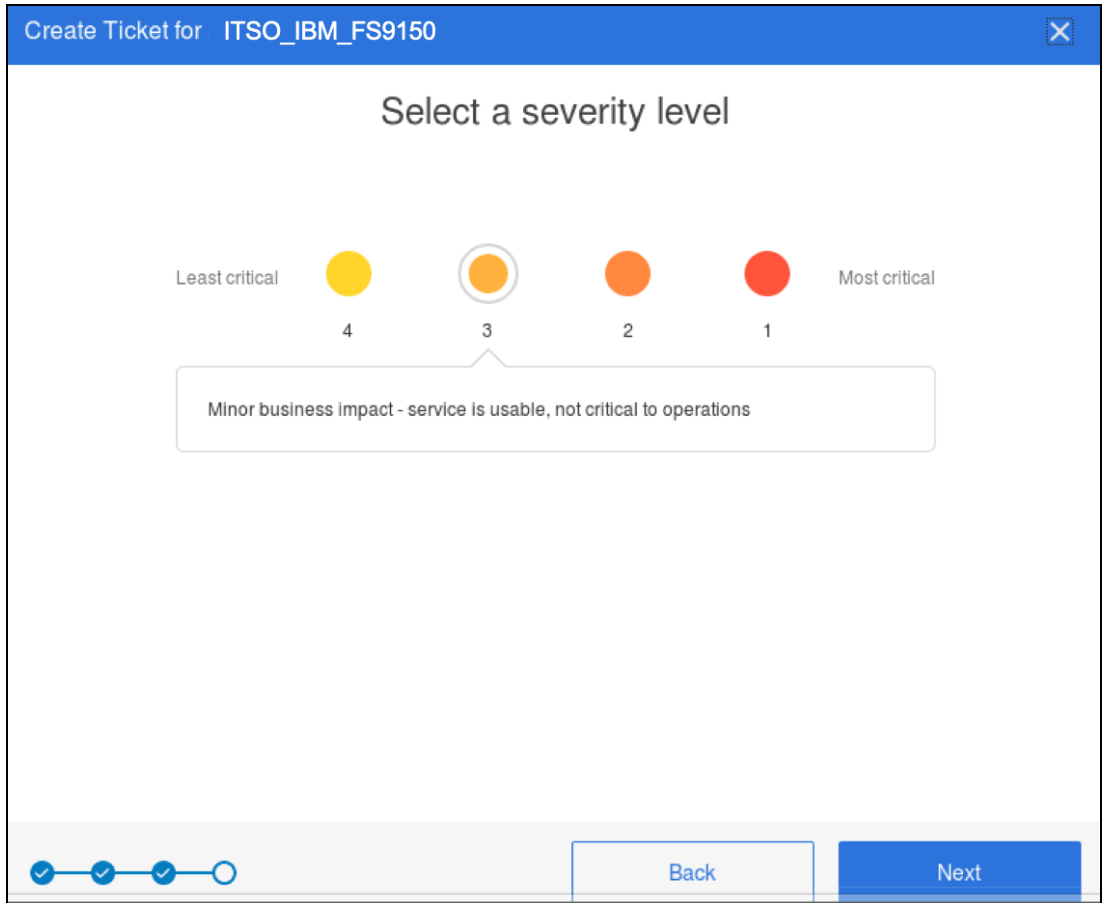


Figure 11-16 Set severity level

10. Figure 11-17 on page 518 gives you a summary of the data that will be used to create the ticket.

Create Ticket for ITSO\_IBM\_FS9150
✕

## Review the ticket

Problem summary: Device is showing a failure alert.

Severity level: 3 Minor business impact - service is usable, not critical to opera...

Type of problem: Hardware

Contact name:

Contact email:

Contact phone:

Customer number:  United States ▼

Storage system: ITSO\_IBM\_FS9150  
 Type: IBM FlashSystem 9100  
 Serial number: 0000020320208FCA  
 Version: 8.2.1.5 (Build 147.13.194231708000)

i **Did you know?**

We can add a log package automatically for you if you monitor your storage system with a data collector and turn on Call Home.

[Learn more about monitoring resources with a data collector](#)

Figure 11-17 Review the ticket information

11. Figure 11-18 on page 519 shows the final summary panel, and the option to add logs to the ticket. When completed, click the **Create Ticket** button to create the support ticket and send it to IBM. The ticket number is created by the IBM Support system and sent back to your SI instance.



Create Ticket for ITSO\_IBM\_FS9150
✕

## Review the ticket

Problem summary: **aaaa**

Severity level: 3 Minor business impact - service is usable, not critical to opera...

Type of problem: **Hardware**

Contact name:

Contact email:

Contact phone:

Customer number:  United States ▼

Storage system: ITSO\_IBM\_FS9150  
 Type: IBM FlashSystem 9100

Serial number: 0000020320208FCA  
 Version: 8.2.1.5 (Build 147.13.194231708000)

Log package: Type 1: Standard logs ▼

✓
✓
✓
✓

Back
Create Ticket

Figure 11-18 Final summary before ticket creation

12. Figure 11-19 on page 520 shows how to view the summary of the open and closed ticket numbers for the system selected, using the **Action** menu option.

The screenshot displays the IBM Storage Insights dashboard for a specific system. At the top, there is a navigation bar with a logo on the left and menu items: Resources, Notifications, Groups, and Configuration. Below the navigation bar, the system information is shown, including a server rack image, a green checkmark, the system name **ITSO\_IBM\_FS9150**, and the model **IBM FlashSystem 9100 - 9848**. A blue button labeled **Actions** with a dropdown arrow is visible. Below this, a tabbed interface shows **General** as the active tab, with **Overview** and **Tickets** as other options. The **Tickets** section is expanded, showing a summary of ticket updates and a list of tickets. It indicates **Open Tickets (2)** and **Closed Tickets (9)**. The open tickets list includes:

- Ticket ID: 018XP2M,866, Opened on May 21, 2019.
- Ticket ID: 33031,019,866, Opened on Apr 14, 2019, with the subject **PMR FOR CAPACITY INVESTIGATION - DO NOT**.

The closed tickets list shows one ticket with the subject **Higher latency after migrating**.

Figure 11-19 Ticket summary

### 11.6.4 Updating support tickets

IBM Storage Insights also has the ability to update, from the Dashboard GUI, support tickets for any of the systems it reports about.

1. To do this go to the SI dashboard and then chose the system you want to update the ticket for. From this screen select **Actions** → **Create/Update Ticket**.
2. Select the option to update an existing ticket, as shown in Figure 11-20 on page 521.

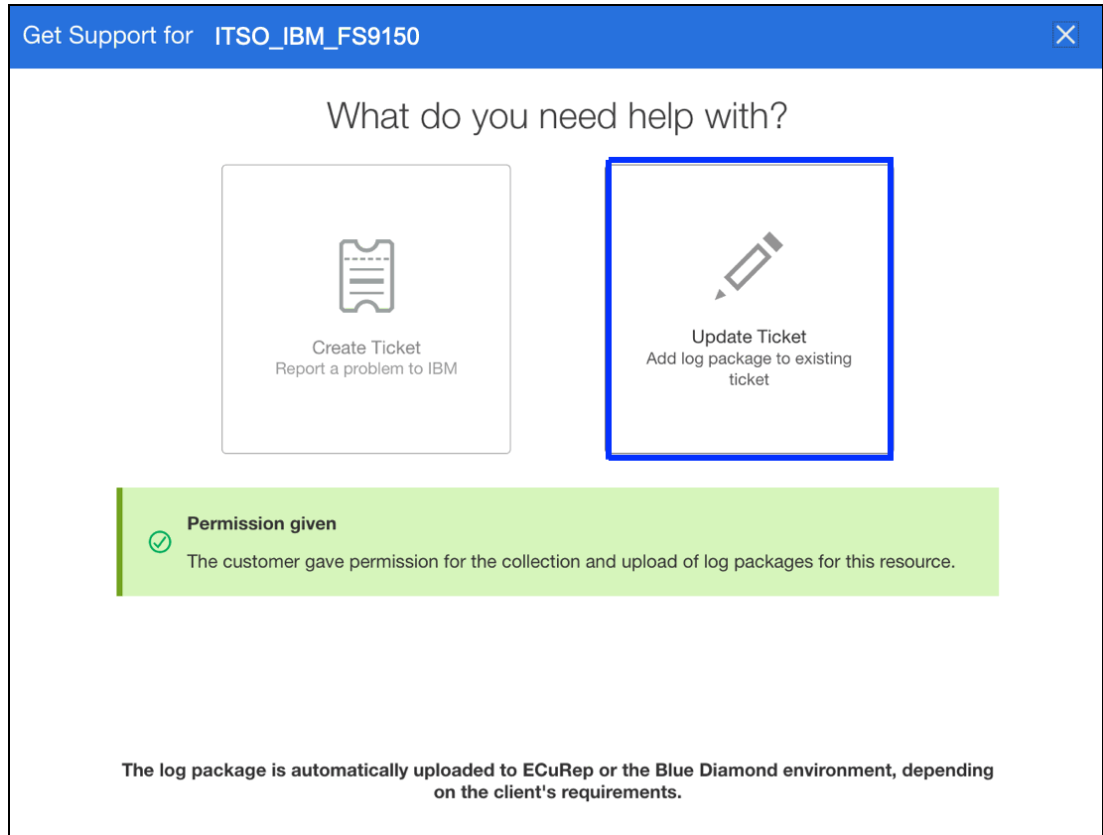


Figure 11-20 SI Update Ticket

Enter the PMR number then press **Next**, as shown in Figure 11-21 on page 522. The PMR input is in the format **XXXXX,YYY,ZZZ**, where:

- ▶ **XXXXX** is the PMR record number
- ▶ **YYY** is the IBM Branch office number
- ▶ **ZZZ** in the IBM country number

These details are supplied either when you created the ticket or by IBM support in the event of the PMR being created by a problem Call Home event (assuming that Call Home is enabled).

Update Ticket for ITSO\_IBM\_FS9150

Select or type the ticket to which you want to add a log package

Ticket: 33031,019,866

Back Next

Figure 11-21 Entering the PMR ticket number

Press **Next** to display the screen where choose the log type to upload. Figure 11-22 on page 523 shows the log selection screen and the options.

The options are as follows:

▶ **Type 1 - Standard logs**

For general problems, including simple hardware and simple performance problems

▶ **Type 2 - Standard logs and the most recent state save log**

▶ **Type 3 - Standard logs and the most recent state save log from each node**

For 1195 and 1196 node errors and 2030 software restart errors

▶ **Type 4 - Standard logs and new state save logs**

For complex performance problems, and problems with interoperability of hosts or storage systems, compressed volumes, and Remote Copy operations including 1920 errors

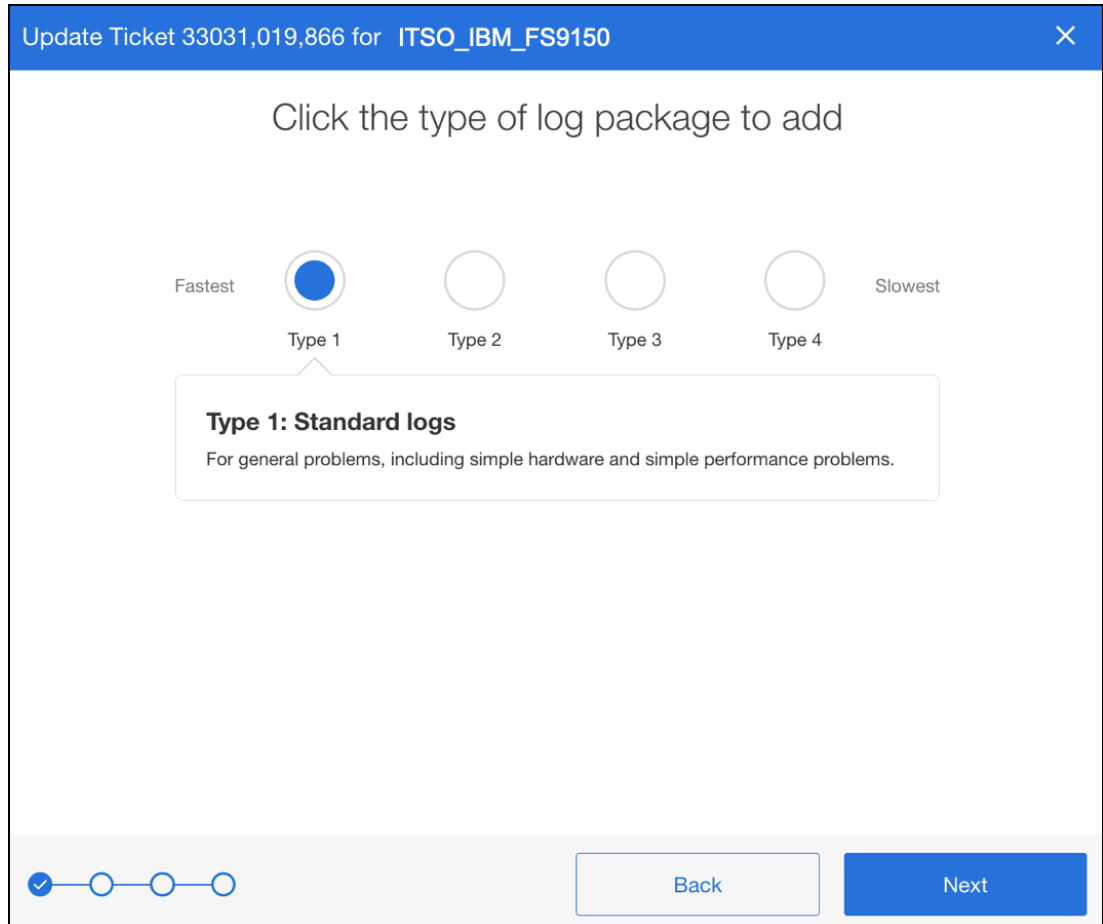


Figure 11-22 Log type selection

If you are unsure which log type to upload, then please ask IBM Support for guidance. The most common type to use is type 1, so this is the default. The other types are more detailed logs and for issues in order of complexity.

After selecting the type of logs and pressing Next, the log collection and upload will start. When completed you will be presented with the log completion screen.

## 11.6.5 SI Advisor

IBM Storage Insights continually evolves and the latest addition is a new option from the action menu 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. Select the system from the dashboard and then click the **Advisor** option to view these recommendations. To see details of a recommendation or to acknowledge it, double-click the recommendation.

Figure 11-23 shows the initial SI advisor menu.

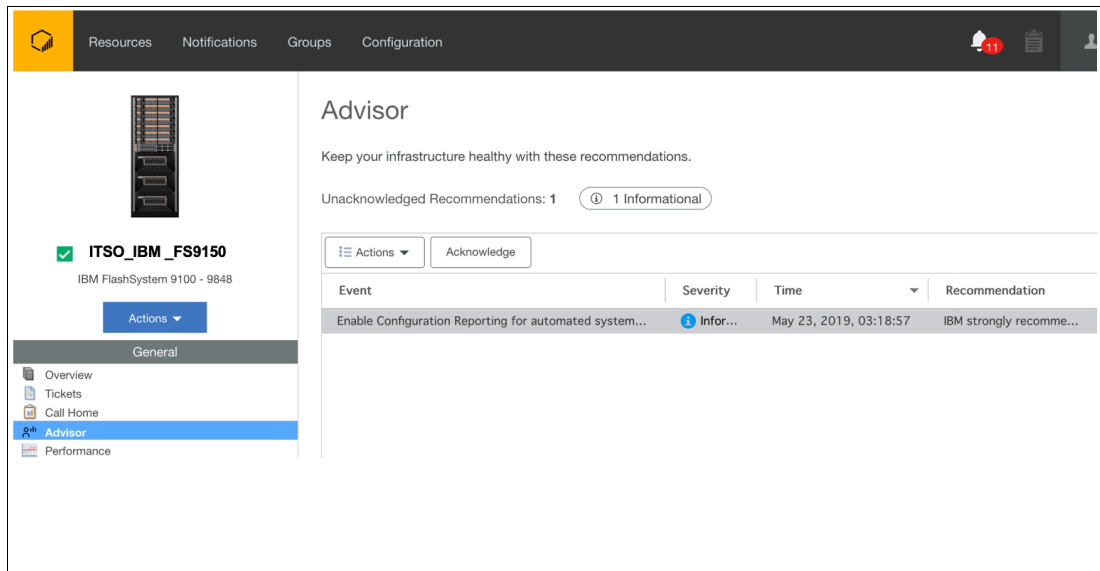


Figure 11-23 SI Advisor menu

Figure 11-24 shows an example of the detailed SI Advisor recommendations.

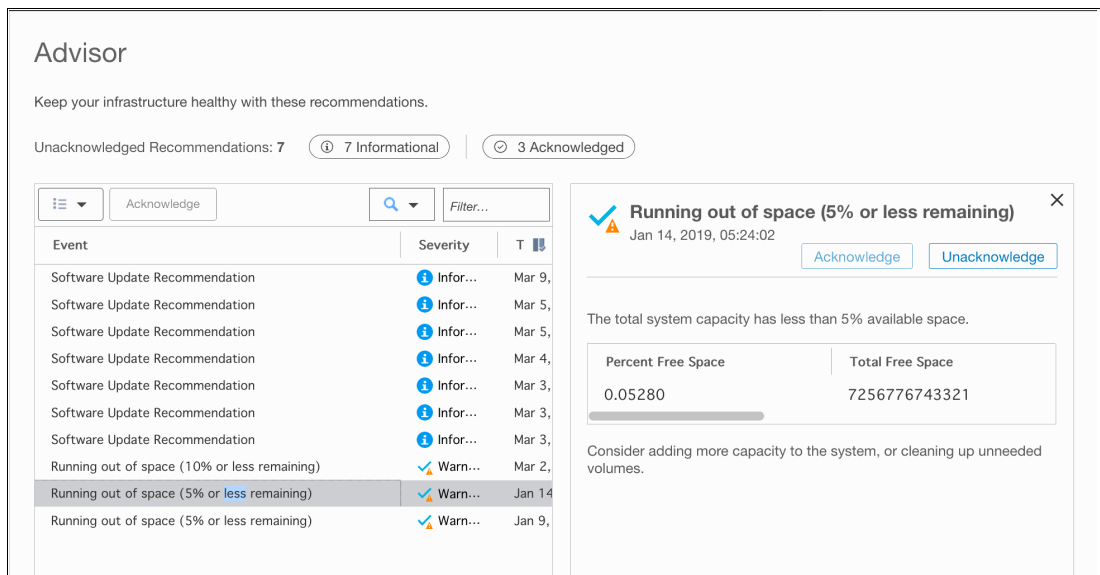


Figure 11-24 Advisor detailed summary of recommendations

The image shows the details of a “Running out of space” recommendation on the Advisor page. In this scenario, the user clicked the **Warning** tag to focus only on recommendations that have a severity of “Warning.” For more information about setting and configuring the Advisor options, see:

[IBM Storage Insights Documentation - Monitoring recommended actions](#)



## IBM i considerations

The IBM Spectrum Virtualize family of block storage systems including the IBM Flash System 5000 series, IBM FlashSystem 7200, and IBM FlashSystem 9200/9200R provides a broad range of flexible and scalable SAN storage solutions capable of meeting demands of IBM i customers for entry to high-end storage infrastructure solutions.

All family members based on IBM Spectrum Virtualize software use a common management interface, and based on their model provide a comprehensive set of advanced functions and technologies like advanced Copy Services functions, encryption, compression, storage tiering, NVMe flash and storage class memory (SCM) devices, and external storage virtualization. Many of these advanced functions and technologies are also of interest to IBM i customers looking for a flexible, high-performing and highly-available storage area network (SAN) storage solution.

This appendix provides important considerations and guidelines for successfully implementing the IBM Spectrum Virtualize family and its advanced functions with IBM i. Unless otherwise stated the considerations also apply to previous generations of products like the IBM Storwize family, the IBM Flash System 9100 series and IBM Flash System V9000.

This appendix includes the following sections:

- ▶ “IBM i Storage management”
- ▶ “Single-level storage” on page 527
- ▶ “IBM i response time” on page 529
- ▶ “Planning for IBM i storage capacity” on page 532
- ▶ “Storage connection to IBM i” on page 533
- ▶ “Setting of attributes in VIOS” on page 539
- ▶ “Disk drives for IBM i” on page 540
- ▶ “Defining LUNs for IBM i” on page 543
- ▶ “Data layout” on page 544
- ▶ “Fibre Channel adapters in IBM i and VIOS” on page 545
- ▶ “Zoning SAN switches” on page 546
- ▶ “IBM i Multipath” on page 546
- ▶ “Boot from SAN” on page 547
- ▶ “IBM i mirroring” on page 547
- ▶ “Copy services considerations” on page 548
- ▶ “Db2 mirroring for IBM i” on page 552

# IBM i Storage management

Due to the unique IBM i storage architecture, special considerations for planning and implementing a SAN storage solution are required also with IBM Spectrum Virtualize based storage. This section provides a short description of 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 must also 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 that the system itself takes over many of the storage management functions, which on other platforms are the responsibility of a system administrator. 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. Figure A-1 shows an example of IBM i storage.

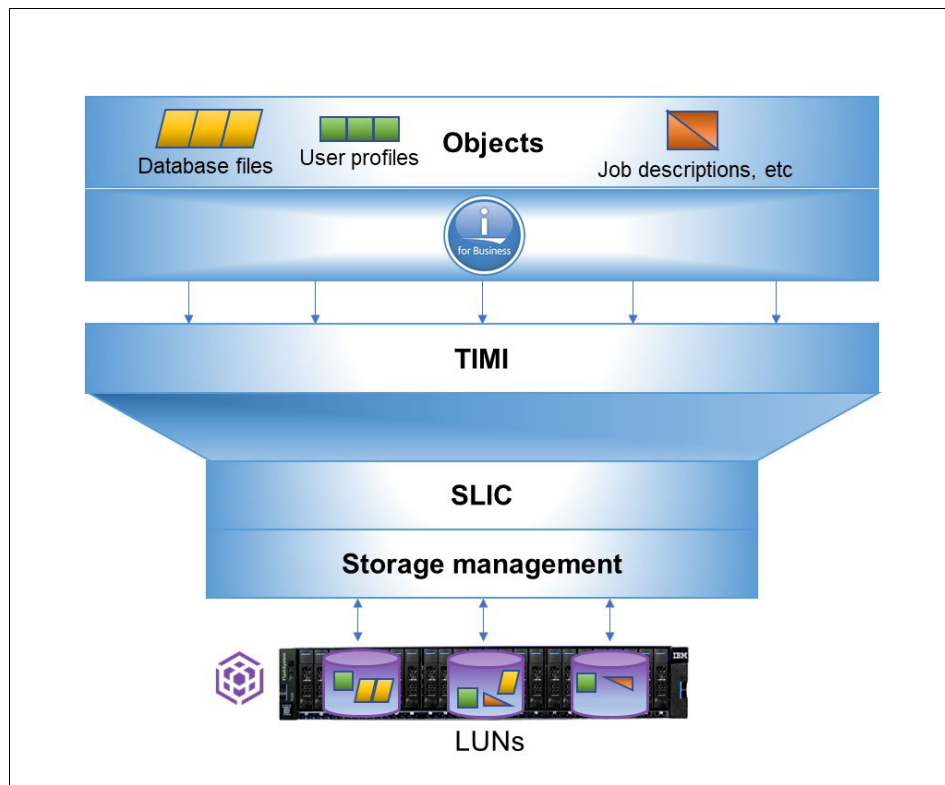


Figure A-1 IBM i storage management spreads objects across LUNs.

**Note:** When a program presents instructions to the machine interface for execution, it assumes that the interface is the system hardware, but is not. The instructions presented to TIMI pass through a layer of microcode before they are understood by the hardware itself. Therefore, Technology Independent Machine Interface (TIMI) and System Licensed Internal Code (SLIC) allow IBM Power Systems with IBM i to take technology in stride.



IBM i storage management, as a component of the SLIC, normally spreads the data in the file across multiple disk units (LUNs when external storage is used). When you add more 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 and/or main storage as one address space, and uses the same set of virtual addresses to cover main memory and disk space. Paging of the objects in this virtual address space is performed in 4 KB pages as shown in Figure A-2. Once a page gets written to disk, it is stored together with metadata including its unique virtual address. For this purpose IBM i originally used a proprietary 520 bytes per sector disk format.

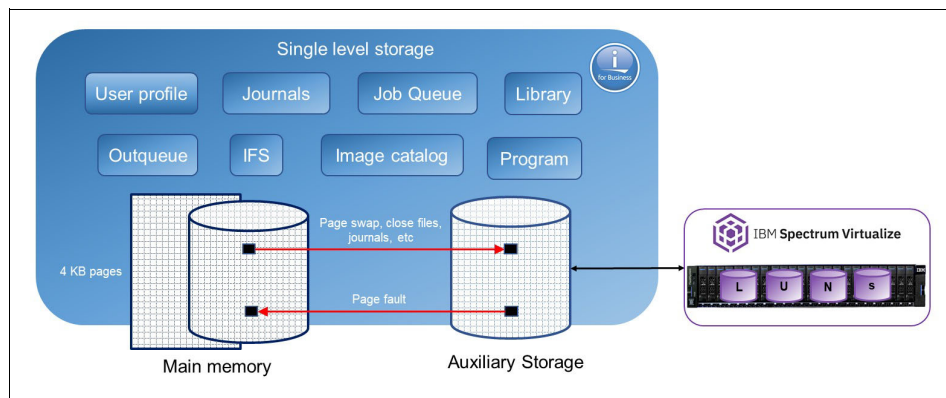


Figure A-2 Virtual address space.

**Note:** The system storage that is being conform with main storage and/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, 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 using auxiliary storage pools. 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 (ASP 2-33). The system ASP and the user ASPs are designated as SYSBAS and constitute the system database. The single-level storage with its unique virtual addresses also implies that the disk storage configured in SYSBAS of an IBM i system must be available in its entirety for the system to remain operational and that 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 auxiliary storage pools architecture is shown in Figure A-3 on page 528.

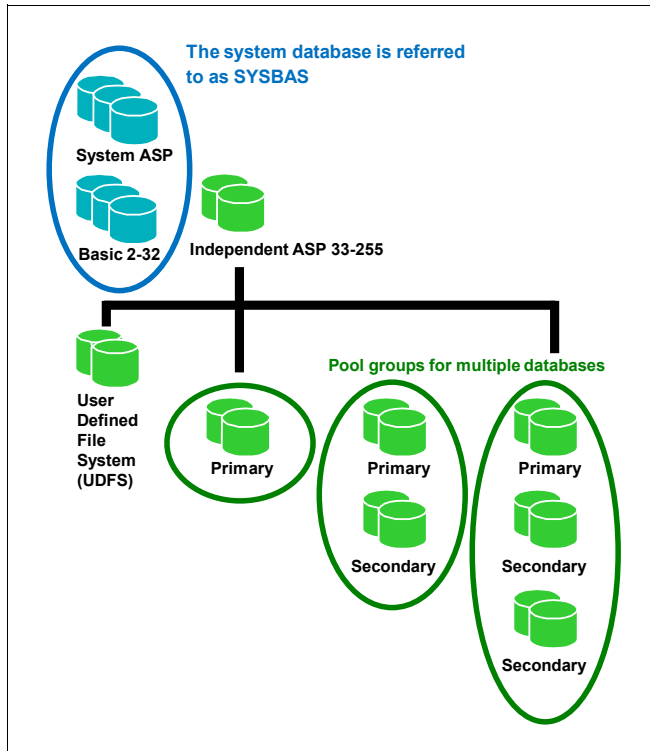


Figure A-3 IBM i auxiliary storage pools architecture

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. Writes are done to main memory and/or main storage, and write operations to disk are performed as a result of swap, file close, or forced write. Application response time depends not only on disk response time, but on 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 as follows in Figure A-4 on page 529.

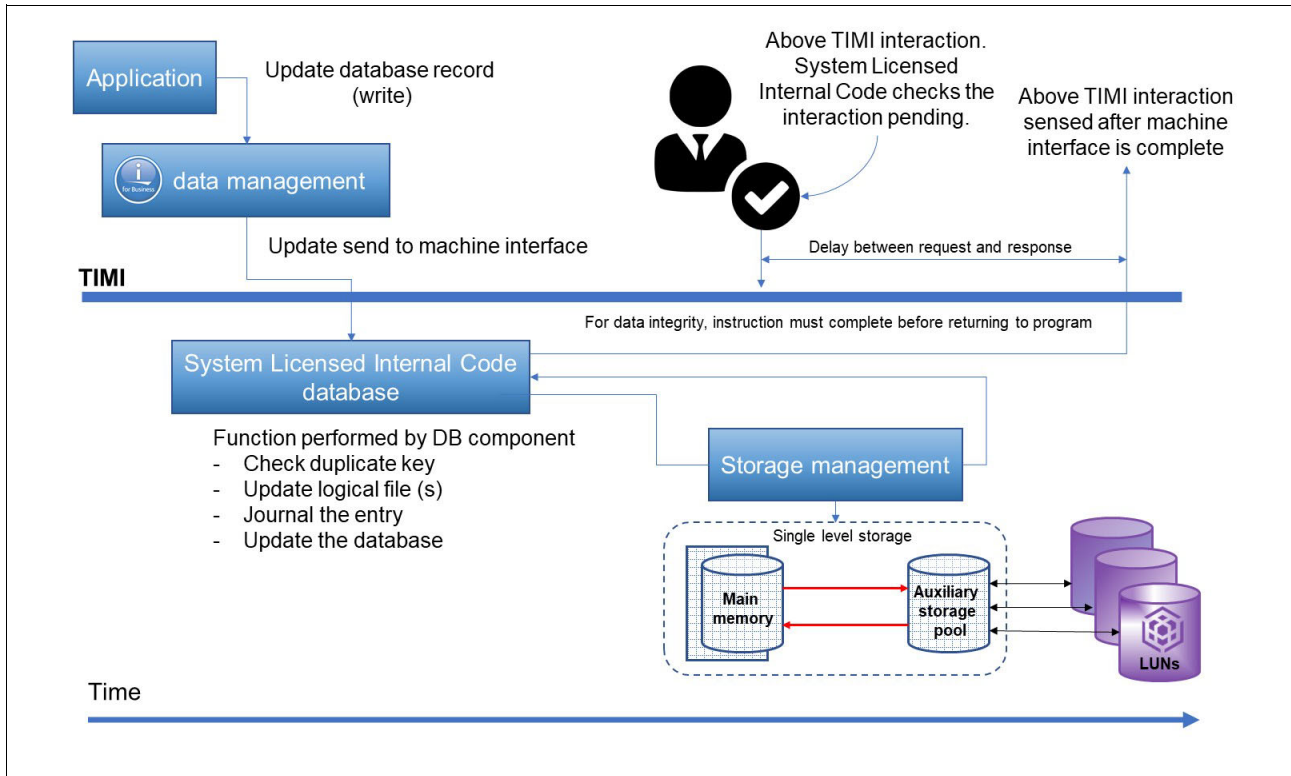


Figure A-4 TIMI atomicity

**Note:** In Figure A-4 auxiliary storage pool (ASP) is conformed and assigned LUNs from IBM Spectrum Virtualize to the IBM i. It also shows an application requests and updates a database record. Throughout the time TIMI task is in progress, an interaction above TIMI can ensue. Of course, the interaction does not carry out upon until TIMI task concludes.

## IBM i response time

IBM i customers are usually concerned about the following types of performance:

- ▶ **Application response time:** The response time of an application transaction. This time is usually critical for the customer.
- ▶ **Duration of batch job:** Batch jobs are usually 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 actual I/O processing and the wait time for potential I/O queuing on the IBM i host. Disk response time can significantly influence both application response time and the duration of a batch job. In this context, due performance of the disk subsystem has a significant impact on overall system performance, which is highlighted in greater detail in the following sections.

### Disk performance considerations

Performance of the disk subsystem has a significant impact on overall IBM i system performance-particularly in a commercial data processing environment where there is usually

a large volume of data to be processed. Disk drives or LUNs' response times contribute to a major portion of the overall response time (OLTP) or runtime (batch).

Definitely, performance of a disk subsystem is affected by the type of protection (RAID, DRAID, or mirroring)

The amount of free space (GB) on the drives and the extent of fragmentation also has an impact. The reason for the impact is the need to find suitable contiguous space on the disks to create new objects or extend existing objects. Disk space is usually allocated in extents of 32 KB. If a 32 KB contiguous extent is not available, two extents of 16 KB are used.

The following sections describe disk performance considerations such as:

- ▶ 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***

Greater sources of disk request are faults that arise from a request for information not being satisfied by what is in memory. Request to bring information in to memory also result in disk I/O. Memory pages can also be purged from time to time, resulting 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 main storage resident by selecting a pool for the object that has available space and does not have jobs associated with it. For further details see, [IBM i Documentation - 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 on page 531.

A request information (data or instructions) from the CPU based on user interactions are submitted to the disk subsystem if it cannot be satisfied from the contents of memory. Whether 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. Operating system attempts to satisfy the request by writing to the controller cache.

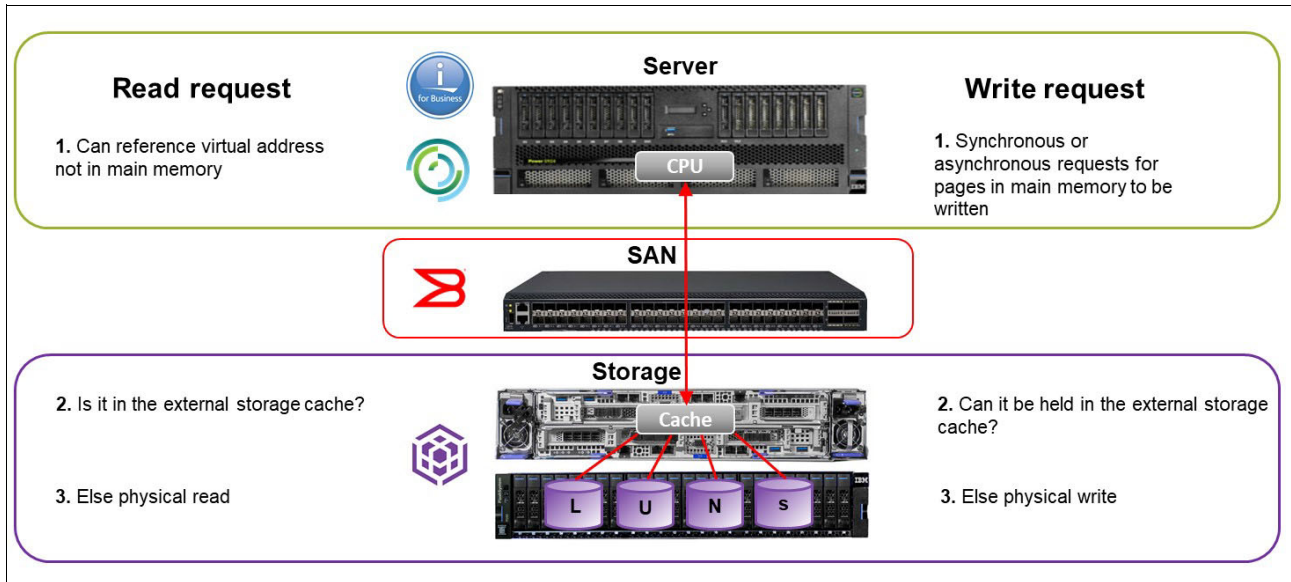


Figure A-5 Disk subsystem

**Note:** The QAPMDISKRB from collections services data files in IBM i, includes disk file response bucket entries and contains one record for each device resource name. It is intended to be used in conjunction with the QAPMDISK file. For more details, see [IBM i 7.4 Documentation - 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.

The time that is taken to respond to synchronous disk I/Os contributes to the Online transaction processing (OLTP), response time or batch runtime. With asynchronous I/O, the progress of a request does not wait for the completion of I/O.

Usually, write requests are asynchronous, including journal deposits with commitment control. However, if journaling is active without commitment control, the writes become synchronous.

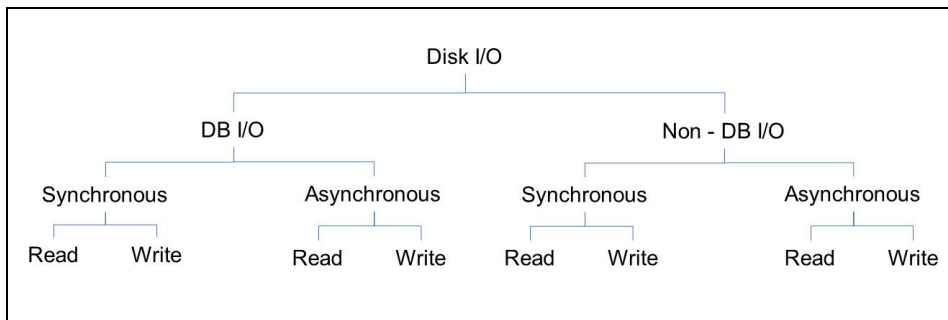


Figure A-6 Disk I/O on IBM i

**Asynchronous I/O wait**

On IBM i, at times jobs may have to wait 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 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 the process explicitly waited for outstanding *asynchronous* I/O operations to complete. For more information, see:

[IBM i 7.4 Documentation - Collection Services data files: QAPMJOBS and QAPMJOBL](#)

This issue might be caused by faster processors running with relatively poor disk subsystems performance. Disk subsystem performance can be impacted by busy or slow disk, small I/O cache.

### ***Disk protection***

For more information about external storage consideration to set up your RAID protection, see Chapter 3, “Storage pools” on page 128.

**Note:** If you need extremely high I/O performance on your IBM i workload, you can create a DRAID 1 on your supported storage system, such as IBM Flash System 7200 and 9200 with IBM Spectrum Virtualize 8.4. 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 buffers memory is available for use by any job/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 they are available in memory. If so, it 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 user program and the partition buffer. This is a count of the number of buffer movements and not a reflection of the records processed.

For more information, see the following web pages:

- ▶ [IBM i 7.4 Documentation - Sharing an Open Data Path](#)
- ▶ [IBM i 7.4 Documentation - Searching for a Perspective: Metric Finder](#)

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 7.4 Documentation - Performance](#).

## **Planning for IBM i storage capacity**

To correctly plan the storage capacity provided by IBM Spectrum Virtualize family systems for IBM i, you must be aware of IBM i block translation for external storage formatted in 512-byte blocks. IBM i internal disks use a block size of either 520 or 4160 bytes.

IBM Spectrum Virtualize storage for hosts is formatted with a block size of 512 bytes, so 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 previously stored in eight

sectors is now stored by using nine sectors, so the required disk capacity on IBM Spectrum Virtualize based systems is 9/8 of the IBM i usable capacity. Similarly, the usable capacity in IBM i is 8/9 of the allocated capacity in these storage systems.

When attaching IBM Spectrum Virtualize family storage to IBM i, plan for the 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 byte sectors for IBM i.



Figure A-7 IBM i with different sector sizes

## Storage connection to IBM i

IBM Spectrum Virtualize storage can be attached to IBM i in the following ways:

- ▶ Native connection without the use of the IBM PowerVM Virtual I/O Server (VIOS)
- ▶ Connection with VIOS in NPIV mode
- ▶ Connection with VIOS in virtual SCSI mode

The decision for IBM i native storage attachment or a VIOS attachment is based on the customer's requirements. Native attachment has its strength in terms of simplicity and 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 IBM PowerVM advanced functions, such as Live Partition Mobility or remote restart.

Table A-1 on page 534 lists key criteria to help you with the decision for 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  |
|-----------------------------------------------------------------------------------------------------------------------|-------------------|------------------|
| <i>Simplicity</i><br>(configuration, maintenance, failure analysis)                                                   | ✓                 | more complex     |
| <i>Performance</i>                                                                                                    | ✓                 | ✓<br>(with NPIV) |
| <i>Consolidation</i><br>(storage / network adapters)                                                                  | more limited      | ✓                |
| <i>PowerVM advanced functions</i><br>(partition mobility, suspend / resume, remote restart, private cloud deployment) | not available     | ✓                |
| <i>Hardware support</i><br>(storage / network adapters, entry level servers)                                          | more limited      | ✓                |

The following sections describe the guidelines and preferred 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

Native connection support for IBM i with IBM Spectrum Virtualize storage is available with IBM Power Systems POWER7® or later server technology. It requires IBM i 7.1, Technology Refresh (TR) 7 or later for POWER7, and IBM i 7.1 TR 8 or later for POWER8®.

Native connection *with* SAN switches can be done by using the following adapters:

- ▶ 32 Gb PCIe3 2-port FC adapters feature number #EN1A or #EN1B (IBM POWER9™ only)
- ▶ 16 Gb PCIe3 4-port FC adapters feature number #EN1C or #EN1D (POWER9 only)
- ▶ 16 Gb PCIe3 2-port FC adapters feature number #EN0A or #EN0B
- ▶ 8 Gb PCIe 2-port FC adapters feature number #5735 or #5273
- ▶ 4 Gb PCIe 2-port Fibre Channel (FC) adapters feature number #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 Spectrum Virtualize V7.5 or later based storage with non-NPIV target ports
- ▶ 4 Gb FC adapters in IBM i connected to 8 Gb adapters in IBM Spectrum Virtualize based storage with non-NPIV target ports

For resiliency and performance reasons, connect IBM Spectrum Virtualize storage to IBM i with multipath using two or more FC adapters:



- ▶ 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 Technology Refresh (TR) 7 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 i 7.2 TR7 and 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 Spectrum 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 Spectrum Virtualize storage:

- ▶ 32 Gb PCIe3 2-port FC adapter feature number #EN1A or #EN1B (POWER9 only)
- ▶ 16 Gb PCIe3 4-port FC adapter feature number #EN1C or #EN1D (POWER9 only)
- ▶ 16 Gb PCIe3 2-port FC adapter feature number #EN0A or #EN0B
- ▶ 8 Gb PCIe 2-port FC adapter feature number #5735 or #5273
- ▶ 8 Gb PCIe2 2-port FC adapter feature number #EN0G or #EN0F
- ▶ 8 Gb PCIe2 4-port FC adapter feature number #5729
- ▶ 8 Gb PCIe2 4-port FC adapter feature number #EN12 or #EN0Y

**Important:** 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](#)

## Connection with VIOS NPIV

IBM i storage attachment support that uses IBM PowerVM Virtual I/O Server N\_Port ID Virtualization (NPIV) was introduced with POWER6 server technology. With NPIV, volumes (LUNs) from the IBM Spectrum 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 Spectrum Virtualize-based storage. NPIV attachment requires 8 Gb or newer generation FC adapter technology and SAN switches that must be NPIV enabled, as shown in Figure A-8 on page 536.

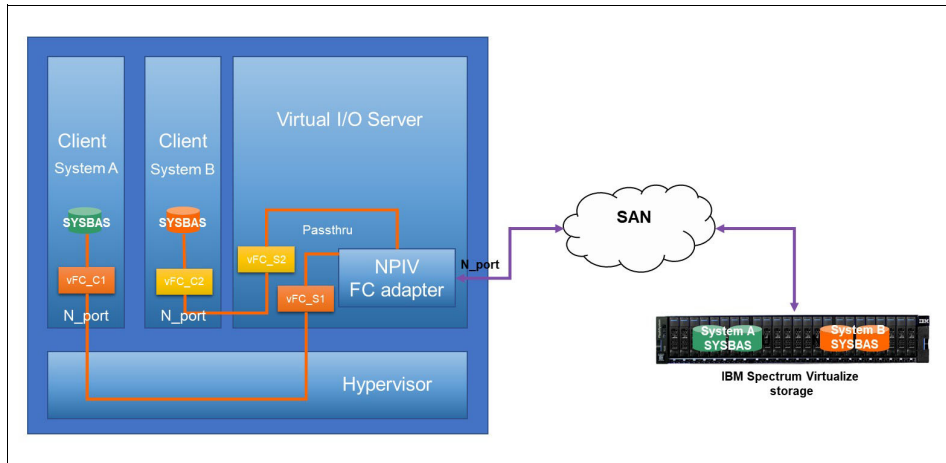


Figure A-8 IBM i SAN access using NPIV

### **Redundant VIOS with NPIV**

For both resiliency and performance reasons, connect IBM Spectrum Virtualize storage to IBM i using multipathing across two or more VIOS servers.

Observe the following rules for mapping IBM i server virtual FC client adapters to the physical FC ports in VIOS when implementing NPIV connection:

- ▶ You can map up to 64 virtual FC adapters to the same physical FC adapter port in VIOS. With VIOS 3.1 and later this limit got increased for support of mapping of up to 255 virtual FC adapters to a 32 Gb physical FC adapter port.
- ▶ Mapping of more than one NPIV client virtual FC 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 using PowerVM partition mobility only a single virtual FC adapter is allowed to be mapped from the *same* IBM i system to a VIOS physical FC adapter port.
- ▶ You can use the same port in VIOS for both NPIV mapping and connection with VIOS virtual SCSI (VSCSI).
- ▶ If PowerHA solutions with IBM i independent auxiliary storage pools (IASPs) are implemented, you need to use different virtual FC adapters for attaching the IASP LUNs, and not share the same adapter between SYSBAS and IASP LUNs.

You can configure a maximum of 127 LUNs (up to 127 active + 127 passive paths) to a virtual FC adapter with IBM i 7.2 TR7 or later, and with IBM i 7.3 TR3 or later.

You can configure a maximum of 64 LUNs (up to 64 active + 64 passive paths) to a virtual FC adapter with IBM i release and TR lower than i 7.2 TR7 and i 7.3 TR3.

IBM i enables SCSI command tag queuing for LUNs from VIOS NPIV connected to IBM Spectrum 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/Virtual FC of IBM i that is attached to an IBM Spectrum Virtualize, such as missing paths and missing disk units, consider the following suggestions:

- ▶ System Snapshot (SYSSNAP), for the SYSSNAP. Be sure to include LIC LOGs, QHST, and PALs. Change the date range to include the date range of the problem. For more information, see [QMGTOOLS: System Snapshot \(SYSSNAP\)](#).
- ▶ VIOS SNAPS 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 Save, SNAP, and so on, from the IBM Spectrum Virtualize at the time the problem is occurring. This information is needed by the storage support team.

If you experience a performance problem with poor disk response time and the IBM Spectrum Virtualize is connected with NPIV, see [Must Gather for Performance Problems with NPIV Connected Storage \(Virtual Fiber Channel\)](#).

### ***NPIV acceleration***

Virtual I/O Server version 3.1.2 or later strengthened FC N\_Port ID Virtualization (NPIV) to provide multiqueue support. This enhanced performance, including more throughput, reduced latency, and higher 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 number #EN1A or #EN1B (POWER9 only)
- ▶ 16 Gb PCIe3 4-port FC adapters feature number #EN1C or #EN1D (POWER9 only)
- ▶ 16 Gb PCIe3 2-port FC adapters feature number #EN2A or #EN2B

**Note:** NPIV acceleration is supported by IBM i 7.2 or later, and by the firmware level for IBM Power Systems 9 is FW940 or later.

### **Connection with VIOS virtual SCSI**

IBM i storage attachment via the IBM PowerVM Virtual I/O Server Connection using virtual SCSI was introduced with IBM Power Systems POWER6 technology.

When deciding on an IBM PowerVM Virtual I/O Server storage attachment for IBM i, NPIV attachment is often preferred over virtual SCSI attachment for the following reasons:

- ▶ With virtual SCSI 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.
- ▶ Virtual SCSI provides much lower scalability in terms of maximum supported LUNs per virtual adapter than NPIV, and requires additional storage management such as multipath configuration and customization at the VIOS layer which adds additional complexity.
- ▶ Due to the virtual SCSI emulation unique device characteristics of the storage device such as device type or in case of tape devices media type and other device attributes are not presented anymore to the IBM i client.

Virtual SCSI attachment is not supported for PowerHA LUN level switching technology which is required for IASP HyperSwap solutions with IBM Spectrum Virtualize. Similar considerations as for NPIV apply with regards to using IBM i multipathing across two or more

VIOS to improve resiliency and performance. However, since with virtual SCSI multipathing is also implemented at the VIOS layer additional considerations apply:

- ▶ IBM i multipathing is performed with two or more VSCSI client adapters, each of them assigned to a VSCSI server adapter in different VIOS. With virtual SCSI, volumes (LUNs) from the IBM Spectrum Virtualize storage system are not mapped directly to an IBM i host but to the two or more VIOS servers. These LUNs which are detected as hdisks on each VIOS need to 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 virtual SCSI, multipathing should also be 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 Spectrum Virtualize based storage running microcode V7.6.1 or later is the VIOS built-in AIXPCM multipath driver which replaces the previously recommended 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](#).

IBM i storage attachment by using the IBM PowerVM Virtual I/O Server Connection that uses virtual SCSI was introduced with IBM Power Systems POWER6 technology.

When deciding on an IBM PowerVM Virtual I/O Server storage attachment for IBM i, NPIV attachment is often preferred over virtual SCSI attachment for the following reasons:

- ▶ With virtual SCSI, 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.
- ▶ Virtual SCSI 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 virtual SCSI 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.
- ▶ Virtual SCSI attachment is not supported for PowerHA LUN level switching technology, which is required for IASP HyperSwap solutions with IBM Spectrum Virtualize.

Similar considerations as for NPIV apply with regards to the use of IBM i multipathing across two or more VIOS to improve resiliency and performance. However, because with virtual SCSI multipathing is also 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 different VIOS. With virtual SCSI, volumes (LUNs) from the IBM Spectrum Virtualize storage system are not mapped directly to an IBM i host but to the two or more VIOS servers. These LUNs that are detected as 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 virtual SCSI, 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 Spectrum Virtualize-based storage running microcode V7.6.1 or later is the VIOS built-in AIXPCM multipath driver, which replaces the previously recommended 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](#).

It is possible to connect up to 4095 LUNs per target, and up to 510 targets per port in a physical adapter in VIOS. With IBM i 7.2 and later, you can attach a maximum of 32 disk LUNs to a virtual SCSI adapter in IBM i. With IBM i releases before i 7.2, a maximum of 16 disk LUNs can be attached to a virtual SCSI 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 VIOS VSCSI connected to IBM Spectrum Virtualize storage. The queue depth on a LUN with this type of connection is 32.

## Setting of attributes in VIOS

This section describes the values of certain device attributes in VIOS which should be configured for resiliency and performance.

### FC adapter attributes

With either VIOS virtual SCSI 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 Spectrum Virtualize storage LUN for 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 certain 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, and will cause unnecessary retries or manual actions.

**Note:** The above attributes are also set to these recommended values when applying the *default rules set* available with VIOS 2.2.4.x or later.

### Disk device attributes

With VIOS virtual SCSI connection, use the VIOS **chdev** command to specify the following attributes for each hdisk device that represents an IBM Spectrum Virtualize storage LUN connected to IBM i:

- ▶ If IBM i multipathing across two or more VIOS servers is used, the attribute `reserve_policy` should be set to `no_reserve`.
- ▶ The attribute `queue_depth` should be set to 32.
- ▶ The attribute `algorithm` should be set to `shortest_queue`.

Setting `reserve_policy` to `no_reserve` is required to be set in each VIOS if multipath with two or more VIOS is implemented, to prevent SCSI reservations on the hdisk device.

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 a dynamic load balancing instead of the default path failover algorithm for distributing the I/O across the available paths to IBM Spectrum Virtualize storage.

Setting a physical volume identifier (PVID) for HDD devices that are used for virtual SCSI 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 mix both regarding the paths to the same LUN. However, VIOS supports NPIV and SCSI concomitantly; that is, some LUNs can be attached to the virtual WWPNs of the NPIV FC adapter. At the same time, the VIOS can also provide access to LUNs that are mapped to virtual target devices and exported as VSCSI devices.

You can have one or more Virtual I/O Servers providing the pass-through function for NPIV. Also, you can have one or more Virtual I/O Servers hosting VSCSI storage. Therefore, the physical HBA in the Virtual I/O Server 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 Virtual FC) and hardware page tables to support the maximum memory value for each partition.

The suggestion is to use the IBM Workload Estimator tool to estimate the needed Virtual I/O Server resources. However, as a starting point in context of CPU and memory for Virtual I/O Server, 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 Spectrum Virtualize storage or externally virtualized backend 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 due to 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 multi-tier storage configuration with either using enterprise class (`tier0_flash`) and high-capacity (`tier1_flash`) flash drives or even enterprise hard disk drives (`tier2_HDD`) is preferred, make sure that a sufficiently large part of disk capacity resides on flash drives. As a rule of thumb for a multi-tier configuration considering 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 usually achieves the best performance when using disk capacity entirely from enterprise class flash (tier0\_flash) storage. High capacity or read-intensive flash drives are typically not the best choice for IBM i performance critical workload, especially for the top storage tier, considering a usually high IBM i write percentage of often 50% and higher, the disk write amplification by using RAID 6 and the significant lower random write performance of tier1 compared to tier0 flash drives.

The use of a multitier storage configuration by IBM Spectrum Virtualize storage is achieved through Easy Tier. For more information, see *Implementing the IBM System Storage SAN Volume Controller with IBM Spectrum Virtualize Version 8.4*, SG24-8491.

Even if you do not plan to install 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 you might gain by adding flash technology in the future.

### Compression considerations

If compression is desired, the preferred choice for using compression at the IBM Spectrum Virtualize storage system layer for performance critical IBM i workload is by using IBM FlashCore module (FCM) hardware compression technology at the disk drive level within IBM Spectrum 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 Real-Time Compression 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 okay.

However, if a physical page on disk is corrupted, potentially have 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 strongly recommend that you test 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 Business Partners, which are recommended to be used for performance modeling and sizing before implementing a wanted IBM Spectrum 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 existing 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 on page 542.



The modeling helps assure an adequate solution sizing by providing predictions for the modeled IBM Spectrum Virtualize storage resource of system usage, the predicted disk response time for IBM i, and the usage and response times at workload growth.



Figure A-9 Diagram of sizing and modeling for IBM i using Disk Magic.

**Note:** Contact your IBM representative or IBM Business Partner to discuss a performance modeling and sizing for a planned IBM Spectrum Virtualize storage solution for IBM i.

### IBM i Unmap support

To better use IBM Spectrum Virtualize storage flash technology with an efficient storage space allocation and deallocation, IBM i supports the space of storage system unmap capabilities by corresponding host unmap functions.

Initially, IBM i unmap support that is implemented by way of the SCSI Write Same command was introduced with i 7.2 TR8 and i 7.3 TR4 for LUN initialization only; that is, for the add disk units to ASP function.

With i 7.3 TR9 and 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 i 7.3 TR7 and i 7.4 TR1, which implements an asynchronous periodic free-space cleaning.

IBM Spectrum Virtualize V8.1.1 and later storage systems can use the unmap function for efficiently deallocate space, such as for volume deletion, on their back-end storage by sending SCSI unmap commands to specific supported internal SSDs and FCMs, and selected virtualized external flash storage. Space reclamation that is triggered by host unmap commands is supported with IBM Spectrum Virtualize V8.1.2 and later for DRP thinly-thinly provisioned volumes, which can increase the free capacity in the storage pool so that it becomes available also for use by other volumes in the pool.



For more information about IBM Spectrum Virtualize storage SCSI unmap support, see 4.1.2, “Data reduction pools” on page 107, and [SCSI Unmap support in Spectrum Virtualize systems](#).

## Defining LUNs for IBM i

LUNs for an IBM i host are defined from IBM Spectrum Virtualize block-based storage. They are created from available extents within a storage pool, the same way as for open system hosts.

Even though IBM i supports a usable LUN size of up to 2 TB - 1 byte for IBM Spectrum Virtualize storage, using 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. The following are the reasons for this:

- ▶ If more LUNs are attached to IBM i, storage management uses more threads and therefore enables better performance.
- ▶ More LUNs provide a higher I/O concurrency which reduces the likelihood of I/O queuing and therefore the wait time component of the disk response time resulting in lower latency of disk I/O operations.

For planning, consider that a higher number of LUNs may also require more physical or/and virtual FC adapters on IBM i based on the maximum number of LUNs supported by IBM i per FC adapter port.

The sizing process helps to determine a reasonable number of LUNs required to access the needed capacity, while meeting performance objectives. Regarding both these aspects and the preferred practices, our guidelines are as follows:

- ▶ 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 a fixed rule 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 a significantly fewer number of LUNs being configured with possibly detrimental effects on performance.

- ▶ A minimum of 8 LUNs for each ASP is preferred for small IBM i partitions and typically a couple of dozen LUNs for medium and up to a few hundreds for large systems.

When defining LUNs for IBM i, consider the following required minimum capacities for the load source (boot disk) LUN:

- ▶ With IBM i release 7.1, the minimum capacity is 20 GB
- ▶ With IBM i release 7.2 before TR1, the minimum capacity is 80 GB in IBM i
- ▶ With IBM i release 7.2 TR1 and later, the minimum capacity is 40 GB in IBM i

IBM Spectrum Virtualize dynamic volume expansion is supported for IBM i with IBM i 7.3 TR4 and later. An IBM i IPL is required to use the extra volume capacity.

**Tip:** For more information about cross-referencing IBM i disks units with IBM Spectrum Virtualize LUNs by using N-Port ID Virtualization (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 related to disk arms and LUNs sizes have been increased in IBM i 7.4, as listed in Table A-2.

Table A-2 Limits increased for Max Disk Arms and LUN size

| System Limits                                                                                   | IBM i 7.2 | IBM i 7.3 | IBM i 7.4 |
|-------------------------------------------------------------------------------------------------|-----------|-----------|-----------|
| Disk arms in all basic auxiliary storage pools (ASPs 1 - 32), per LPAR                          | 2047      | 2047      | 3999      |
| Disk arms in all independent auxiliary storage pools (IASPs 33 - 255) in all nodes in a cluster | 2047      | 2047      | 5999      |
| Maximum combined number of disk arms and redundant connections to disk units                    | 35.600    | 35.600    | 35.600    |
| 512 byte block size LUNs <sup>a</sup>                                                           | 2 TB      | 2 TB      | 2 TB      |
| 4096 byte block size LUNs <sup>b</sup>                                                          | 2 TB      | 2 TB      | 16 TB     |

a. Actual limit is one block short of the max shown in Table A-2. Note that for all 512 block LUNs, the maximum is still up to 2TB. This includes IBM Storwize LUNs, and SAN Volume Controller LUNs.

b. This includes IBM FlashSystems LUNs, and 4 K block SAS disks (VSCSI attached).

**Note:** For more information about these limits, and others, see [IBM i 7.4 Documentation - Miscellaneous limits](#).

## Data layout

Spreading workloads across all IBM Spectrum Virtualize storage components maximizes the use of the hardware resources in the storage subsystem. I/O activity should be balanced between the two nodes or controllers of the IBM Spectrum Virtualize storage system I/O group, which is usually taken care of by the alternating preferred node volume assignments at LUN creation.

However, especially with improper sizing or unanticipated workload increases, it is possible when sharing resources that performance problems might arise due to resource contention.

Some isolation of workloads, at least regarding a shared back-end storage, can be accomplished by a configuration where each IBM i ASP or LPAR has its own managed storage pool. Such a configuration with dedicated storage pools results in a tradeoff between accomplishing savings from storage consolidation and isolating workloads for performance protection. This is 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 layout:

- ▶ For all-flash storage configurations, assuming a properly sized storage backend, there is typically no reason 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 may also be shared among IBM i workloads. Only very large performance critical workloads should be 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 using Easy Tier on IBM Spectrum Virtualize for managing a multi-tier storage pool, there is also an option to create a separate storage pool for different storage tiers on IBM Spectrum Virtualize storage and create different IBM i ASPs for each tier. IBM i applications that have their data located in an ASP of a higher storage tier will experience a performance boost compared to those using 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 Spectrum 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, take into account the maximum I/O rate (IOPS) and data rate (MBps) that a port in a particular 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, take into account 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 specified in Table A-3 that shows the throughput of a port in a particular adapter at 70% utilization.

*Table A-3 Throughput of Fibre Channel 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 using separate FC adapters for IBM i disk and tape attachment. This separation is recommended due to the required IBM i virtual I/O 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 Spectrum Virtualize storage target, each port from one node canister of the I/O group, as shown in Figure A-10. This provides resiliency for the I/O to and from a LUN assigned to the IBM i FC initiator port. If the preferred node for that LUN fails, the I/O continues using the non-preferred node.

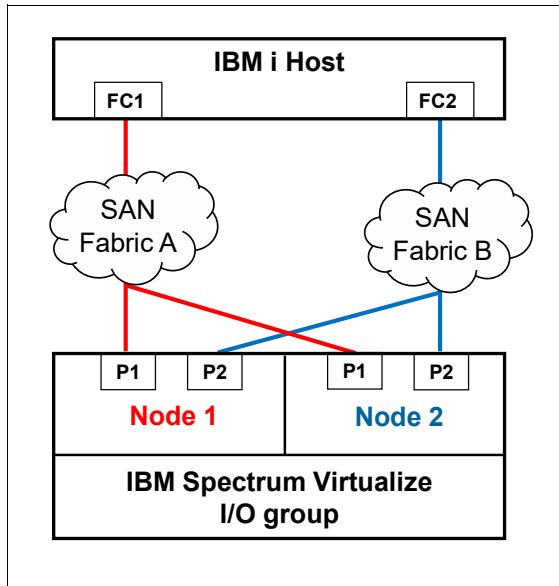


Figure A-10 SAN switch zoning for IBM i with IBM Spectrum Virtualize storage

For VIOS virtual SCSI attachment, zone one physical port in VIOS with one or more available FC ports from each of both node canisters of the IBM Spectrum Virtualize storage I/O group. SVC or Storwize ports that are zoned with one VIOS port should be evenly spread between both node canisters. Keep in mind that a maximum of eight host paths is supported from VIOS to IBM Spectrum Virtualize storage.

## IBM i Multipath

Multipath provides greater resiliency for SAN-attached storage. IBM i supports up to eight active paths and up to eight passive paths to each LUN. In addition to the availability considerations, lab performance testing has shown that two or more paths provide performance improvements when compared to a single path.

Typically two active paths to a LUN is a good balance of price and performance. The scenario shown in Figure A-10 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 very high I/O rates are expected to the LUNs respectively where a high I/O access density is expected.

It is important to understand that IBM i multipath 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 Spectrum Virtualize storage node canister to a SAN switch zone with an IBM i FC initiator port does not provide more active paths since an IBM i FC initiator port, by design, will log in into one target port of a node only.

With IBM i native attachment, the ports for multipath must be from different physical FC adapters in IBM i. With VIOS NPIV, the virtual Fibre Channel adapters for multipath must be assigned to different VIOS for redundancy. However, if more than two active paths are used, you can use two VIOS and split the paths among them. With VIOS virtual SCSI attachment, the virtual SCSI adapters for IBM i multipath must be assigned to different VIOS.

IBM Spectrum Virtualize storage uses a redundant dual active controller design that implements SCSI asymmetrical logical unit access (ALUA). That is, 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 preferred node. The non-optimized paths, which often are associated with the non-preferred node, are not actively used.

In the case of an IBM Spectrum Virtualize storage topology, such as HyperSwap or IBM SAN Volume Controller Enhanced Stretched Cluster 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 attribute 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 of 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 using the passive paths if all active paths failed.

## Boot from SAN

All IBM i storage attachment options that is native, VIOS NPIV, and VIOS virtual SCSI, support IBM i boot from SAN. The IBM i load source is on an IBM Spectrum Virtualize storage LUN that is connected the same way as the other LUNs. Apart from the required minimum size there are not any special requirements for the load source LUN. The FC or SCSI I/O adapter for the load source needs to be *tagged* that is to say specified, by the user in the IBM i partition profile on the IBM Power Systems Hardware Management Console (HMC). When installing the IBM SLIC with disk capacity on IBM Spectrum Virtualize storage, the installation prompts you to select one of the available LUNs for the load source.

## IBM i mirroring

Some clients prefer to use IBM i mirroring functions for resiliency. For example, they use IBM i mirroring between two IBM Spectrum Virtualize storage systems, each connected with one VIOS.

When setting up IBM i mirroring with VIOS connected IBM Spectrum Virtualize storage, you should add the LUNs to the mirrored ASP in steps:

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 Spectrum Virtualize storage systems and not among the LUNs from the same storage system.

## Copy services considerations

This section covers IBM Spectrum Virtualize Copy Services considerations for usage with IBM i.

### Remote replication

The IBM Spectrum Virtualize family products support both Metro Mirror synchronous remote replication and Global Mirror asynchronous remote replication. For Global Mirror there are two options: *Standard* Global Mirror, and Global Mirror with *change volumes*, 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 also during resynchronization.

Regarding the IBM Spectrum Virtualize Copy Services functions, the IBM i single-level storage architecture requires that the disk storage of an IBM i system needs to be treated as a single entity, i.e. the scope of copying or replicating an IBM i disk space needs to include either 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. When a planned or unplanned outage occurs on the IBM i production server, the IBM i standby server can be started (IPLed) from the replicated SYSBAS volumes after they have been switched on IBM Spectrum Virtualize to a primary role to become accessible for the IBM i standby host.

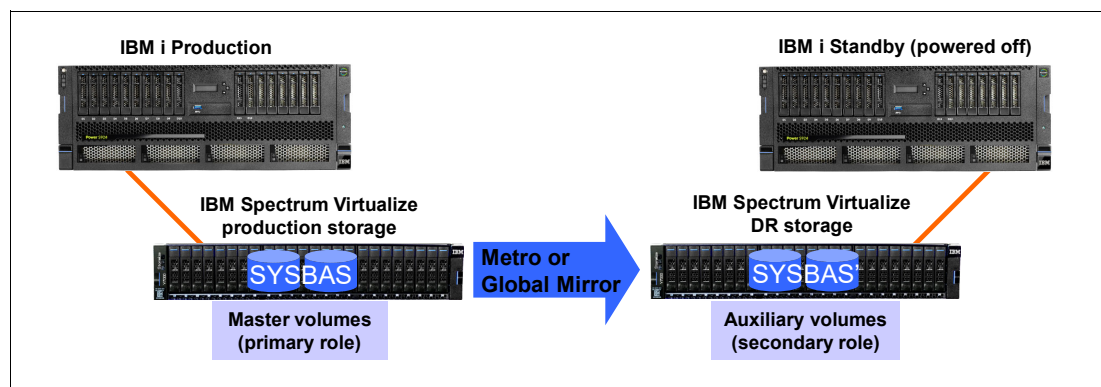


Figure A-11 IBM i full system replication with IBM Spectrum Virtualize

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 Spectrum Virtualize remote replication is switchable between the two cluster nodes as shown in Figure A-12 on page 549. 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) for managing both the IBM i cluster node switch- and failovers as well as the IBM Spectrum Virtualize storage remote replication switching.

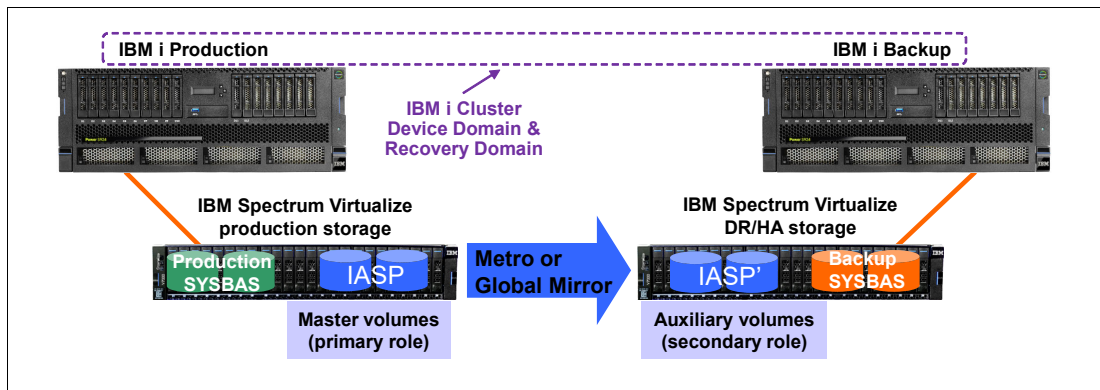


Figure A-12 IBM i IASP replication with IBM Spectrum Virtualize

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) for managing the IBM i cluster node switch and fail overs and the IBM Spectrum Virtualize storage remote replication switching.

For more information about IBM i high availability solutions with IBM Spectrum Virtualize Copy Services see *PowerHA SystemMirror for IBM i Cookbook*, SG24-7994.

The sizing of the required replication link bandwidth for Metro Mirror or Global Mirror must be based on the peak write data rate of the IBM i workload to avoid affecting production performance. Refer to “SAN Extension design considerations” on page 285.

For more information about current IBM Spectrum Virtualize storage zoning guidelines, see 2.3.2, “Port naming and distribution” on page 45

For environments that use remote replication, a minimum of two FC ports is suggested on each IBM Spectrum Virtualize storage node that is used for remote mirroring. The remaining ports on the node should not have any visibility to any other IBM Spectrum Virtualize cluster. Following these zoning guidelines helps to 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 both FlashCopy with background copy and without background copy.

When using FlashCopy with thinly provisioned target LUNs, make sure that there is sufficient capacity available in the storage pool to be dynamically allocated when needed for the copy-on-write operations. The required thin target LUN capacity depends on the amount of write operations to both 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 source system and 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.
- ▶ Source IASP must be varied off before performing a FlashCopy.
- ▶ This method is only method to ensure all write are sent out to disk and included.

### FlashCopy warm

The following considerations apply to FlashCopy warm:

- ▶ Memory is not 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 auxiliary storage pools (ASPs) or independent ASPs (IASPs).

The following considerations apply to FlashCopy quiesced:

- ▶ Some memory flushed to disk.
- ▶ Attempt to flush writes to disk and suspend DB I/O and to reach commitment control boundaries.
- ▶ Minimal disruption to source, is the preferred practice, and better quality than warm

## HyperSwap

IBM Spectrum Virtualize storage 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 for VIOS NPIV attachment.

HyperSwap for IBM i IASP replication is supported by IBM i 7.2 TR5 or later and by IBM i 7.3 TR1 or later. With this solution you need to install IBM PowerHA SystemMirror for i Standard Edition (5770-HAS \*BASE and option 2) that enables LUN level switching to site 2. It is supported for native and VIOS NPIV attachment.

IBM Spectrum Virtualize HyperSwap relies on the SCSI ALUA aware IBM i host multipath driver to manage the paths to the local and remote IBM Spectrum 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 Spectrum Virtualize nodes of the I/O group on each site. For a balanced configuration the SAN switches from a dual fabric configuration should be evenly used.

An example of the SAN fabric connections for IBM i HyperSwap with VIOS NPIV attachment is shown in Figure A-13 on page 551. This configuration example results in four active paths and twelve passive paths presented on IBM i for each HyperSwap LUN.



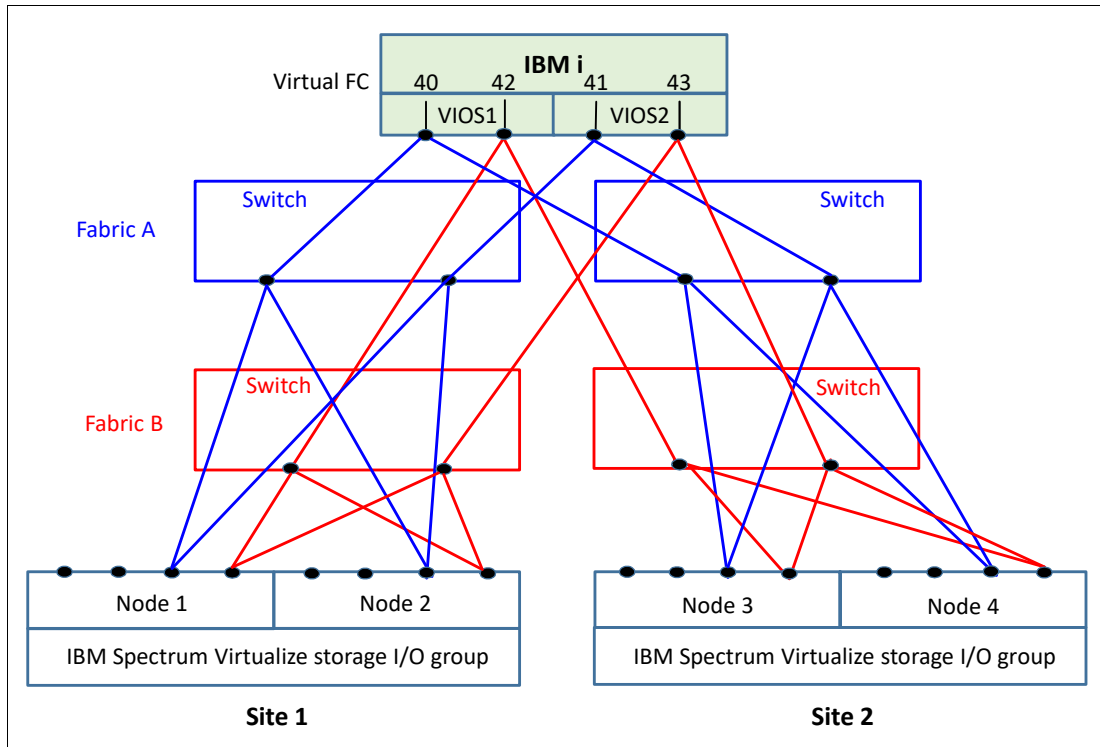


Figure A-13 IBM i HyperSwap SAN fabric connection example

In the following we briefly describe some high availability scenarios with using HyperSwap for IBM i.

### Outage of Spectrum Virtualize I/O group at site 1

In this scenario the entire IBM i storage capacity resides on HyperSwap LUNs.

After the outage of I/O group at site 1 occurs the I/O rate automatically transfers to the IBM Spectrum Virtualize nodes at site 2. The IBM i workload keeps running, and there are no relevant messages in IBM i message queues.

When the outage has finished, 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 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 Spectrum Virtualize storage cluster: one host with the WWPNs of IBM i at site 1, and one with WWPNs of site 2.

After a failure of site 1, including a failure of the IBM i production system as well as the storage at site 1, the IBM i LUNs are still available from the IBM Spectrum 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 IPL the IBM i standby host at site 2. After the IPL is finished we can resume the workload on site 2.

Once the outage of site 1 is finished, power-down IBM i at site 2, unmap the IBM i LUNs from the host at site 2 and map them to the host at site 1. IPL IBM i at site 1 and resume the workload. The I/O rate will be transferred to the IBM Spectrum Virtualize storage nodes at site 1.

## Disaster at site 1 with IASP HyperSwap

This scenario requires IBM PowerHA SystemMirror for 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 PowerHA for i setup refer to *IBM PowerHA SystemMirror for i: Preparation (Volume 1 of 4)*, SG24-8400.

In this scenario, ensure that all IBM i LUNs, not just the IASP LUNs, are HyperSwap volumes.

If there is a disaster 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

IBM PowerVM® Live Partition Mobility (LPM) allows you to move a running logical partition, including its operating system and running applications, from one system to another without any shutdown or without disrupting the operation of that logical partition.

In this scenario Live Partition Mobility is combined 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 Spectrum Virtualize with the WWPNs of the second port of the virtual FC adapters. We recommend creating a separate host object definition for the secondary ports to specify site 2 for this host object. Then, you 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 completed, 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:** Live Partition Mobility now supports multiple client virtual FC (vFC) adapter ports being mapped to a single physical FC port. Each client virtual FC must be mapped to a separate physical port in advance, whether LPM with FC N\_Port ID Virtualization is used. That restriction was removed for the use of Virtual I/O Server version 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 use of the adapter.

## Db2 mirroring for IBM i

The Db2 Mirror base configuration consists of two systems that are in the same data center. This configuration cannot span locations because it is active-active read/write which means that by definition all write operations are synchronous (using Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE) network) to the application state. The write operations between two systems necessitates that the distance between the systems be limited so as to not impact performance.

The following two broad approaches are used to deploy active-active solutions:

1. Distributed lock management, where multiple application servers can access the common or shared database but are prevented from performing simultaneous updates by the distributed lock management which locks the other users out while you do an update.
2. The replication approach, where each update of any type is totally synchronous to the application state. When an application has an update, it does not proceed to the next application step until the current write operations has completed on both the primary and secondary objects. This is referred to as a two-phase commit between two systems.

**Note:** Applications can be deployed in an active-active manner, where 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 goes down, the application servers continue performing I/O operations to the other system in the mirrored pair. This configuration has the additional benefit of enabling *workload balancing*.

However, applications can also be deployed in an active-passive manner, where application servers will conduct write operations to one of the two systems in the two-system complex and, in the event that the primary goes away, the application groups will be switched to the secondary system. The active-active case will necessitate that the application servers be hosted separately from the database servers and be connected through a client/server construct such as Java Database Connectivity (JDBC).

**Note:** IBM i JDBC drivers now contain alternate server fail-over support to automatically transition the JDBC request between systems when one connection is no longer available. For many IBM i application workloads, deployment is through the traditional 5250 emulation screen and contained in the same LPAR as the operating system and database. In this case, if the primary goes down, the database has been continuously replicated to the secondary system synchronously and is immediately available. The application will need to be restarted on the secondary system prior to the resumption of the workload processing.

When one of the systems in the Db2 Mirror configuration is not available, Db2 Mirror will track all update, change, and delete operations to the database table and all other mirror-eligible objects. When the pair is reconnected, changes are synchronized between the systems. This includes databases that reside in either an Independent Auxiliary Storage Pool (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.

In addition, Db2 Mirror supports applications that use either 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 solutions requires the following software:

- ▶ IBM Power Systems POWER8 or later
- ▶ IBM i 7.4 with IBM Db2 Mirror for i V7.4 (5770-DBM)
- ▶ IBM i Option 48, Db2 Data Mirroring, is required for Db2 Mirror for i. Therefore, entitlement for Option 48 is automatically included with Db2 Mirror for i orders. Make sure that IBM i Option 48 is installed and a key is applied with the Db2 Mirror for i Licensed Program Product.

For more information on the software requirements for Db2 Mirror see [IBM i 7.4 Documentation - Software requirements](#).

Disaster Recovery can be performed with various options, such as:

- ▶ The IBM PowerHA SystemMirror for i Enterprise Edition
- ▶ Full system replication
- ▶ Logical replication

**Important:** Consider the following points when Db2 Mirror local continuous availability is combined with existing high availability (HA) and disaster recovery (DR) replication technologies:

- ▶ Remote replication for DR can be implemented either by storage-based replication, this means using the *Copy Services of IBM Spectrum Virtualize software*.
- ▶ Integrated File System (IFS) IASP must remain switchable between both local Db2 Mirror nodes by choosing a DR topology that is supported by IBM PowerHA SystemMirror for i.
- ▶ 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 four-node Db2 Mirror PowerHA Cluster configuration. Both 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 RCL).

For more details, see *IBM Db2 Mirror for i Getting Started*, REDP-5575.

## Overview of the setup process

The following three node types are part of the setup and configuration of Db2 Mirror:

- ▶ Managing node
- ▶ Setup source node
- ▶ Setup copy mode

For more information about the nodes, setup, and configuration, see [IBM i 7.4 Documentation - Overview of the setup process](#).

Db2 Mirror is initially configured on a single partition, which is 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 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 both external and internal storage. For the external storage used during the cloning process, IBM storage systems are recommended rather than non-IBM external storage because Db2 Mirror automates the cloning for *IBM Spectrum Virtualize family*.

The cloning technologies used for IBM storage systems are as follows:

- ▶ *FlashCopy* (cold and warm) is used when both Db2 Mirror nodes connect to the same IBM Spectrum Virtualize storage system
  - A cold clone requires the *setup source node* to be 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 Spectrum Virtualize storage*.

For more information, see [IBM i 7.4 Documentation - Software requirements](#).

**Note:** Volume mirroring supported in IBM FlashSystem 9200 and IBM SAN Volume Controller is a valid cloning method for DB2 Mirror on the category of *manual copy*, but it is not automated as using FlashCopy, Metro Mirror or Global Mirror.

## IBM Spectrum Virtualize and Db2 Mirror

IBM Spectrum Virtualize storage systems establish communication with Db2 Mirror utilizing Secure Shell (SSH) to manage Copy Services functions. By the way, IBM Spectrum Virtualize user ID must have the user role of administrator. In that context, for *managing node*, the following product are mandatory:

- ▶ 5733SC1 \*BASE IBM Portable Utilities for i
- ▶ 5733SC1 Option 1 OpenSSH, OpenSSL, zlib
- ▶ 5770SS1 Option 33 Portable Application Solutions Environment

**Note:** To create an SSH key pair, see [IBM i 7.4 Documentation - Creating an SSH key file for accessing IBM Spectrum Virtualize storage](#). After the creation of SSH key pair, attach the SSH public key to a use on the IBM Spectrum Virtualize storage system. The corresponding private key file must be uploaded to the *managing node* so it can be used during DB2 Mirror setup.

### Virtual I/O Server and native attachment

The Db2 Mirror storage cloning process for IBM Spectrum Virtualize requires Fibre Channel adapters with native attachment or attachment with Virtual I/O Server N\_Port ID Virtualization.

### Host object definition and volume planning

Before you setup Db2 Mirror, you must define the host object and assign volumes to the hosts to be used by the *setup copy node*. It is mandatory to have the following:

- ▶ The same number of host objects and volumes
- ▶ The same size volumes defined for the *setup source node* and *setup copy node*

Afterwards, the Db2 Mirror cloning process pairs storage volumes between the *setup source node* and *setup copy node*. The cloning process applies to SYSBAS and IASPs, as follows:

- ▶ The *setup source node* and *setup 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 beginning to add a database IASP to DB2 Mirror.

### Remote Copy cloning

Db2 Mirror Remote Copy cloning uses the following IBM Spectrum Virtualize Copy Services operations to copy the *setup source node* volumes to the *setup copy nodes* volumes:

- ▶ Global Mirror for cold clone
- ▶ GMCV for warm clone

Regardless of whether you plan to perform the Remote Copy during a planned outage window, you need to make sure that your bandwidth between storage systems is sufficient to

complete the Remote Copy during that period of time. The Db2 Mirror cloning process does not provide the capability of pausing the cloning and then resuming it later. Thus, you need to plan enough time for the Remote Copy to complete.

**Important:** For IBM Spectrum Virtualize, the Copy Services partnership between storage systems must be manually created before beginning to configure Db2 Mirror.

## Architectures and considerations for DB2 Mirror

Due to the synchronous design of Db2 Mirror, the distance between the nodes is limited to within a data center in most cases. Multiple configurations are supported for both a data center Db2 Mirror implementation and for the addition of a DR solution. Several options are shown in this section as examples with IBM Spectrum Virtualize storage systems. A specific implementation depends on your business resilience requirement.

**Note:** Db2 Mirror supports IBM SAN Volume Controller topologies such as Enhanced Stretched Cluster or HyperSwap.

### ***Db2 Mirror environment with one IBM Spectrum Virtualize storage system***

In this example, one IBM Spectrum 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. You can consider this solution as a DR strategy, to provide storage resiliency.

Figure A-14 shows two IBM Power System servers are used (at least one RoCE adapter per server). However, you can reduce this scenario in terms of cost of decreased resiliency by implementing Db2 Mirror across two IBM i LPARs on the same IBM Power Systems. For this example, a SYSBAS is cloned; however, IASP also can be added by using another set of volumes.

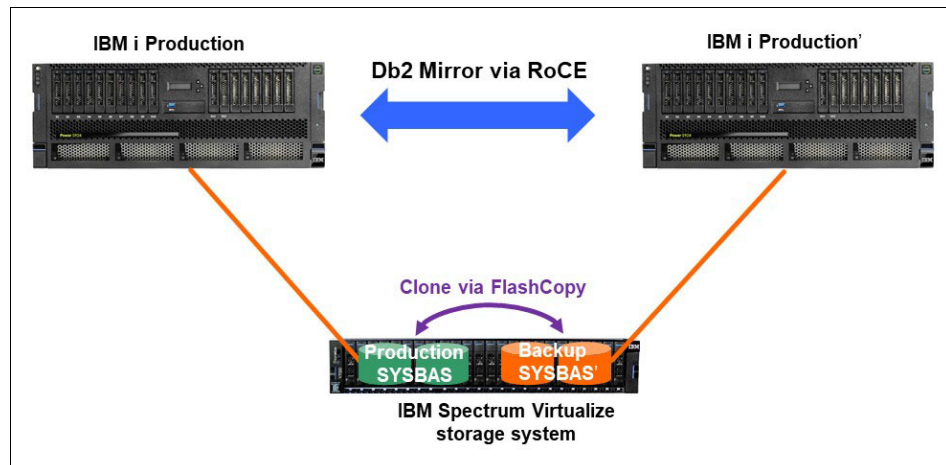


Figure A-14 Db2 Mirror environment with one IBM Spectrum Virtualize storage

### ***DB2 Mirror environment with two IBM Spectrum Virtualize storage systems***

The use of two IBM Spectrum 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 Systems servers and IBM Spectrum Virtualize storage systems are used. Also, Remote Copy is used to set up DB2 Mirror.

As shown in Figure A-15, the set of volumes for SYSBAS and the set of volumes for IASP are replicated. Global Mirror can also be used.

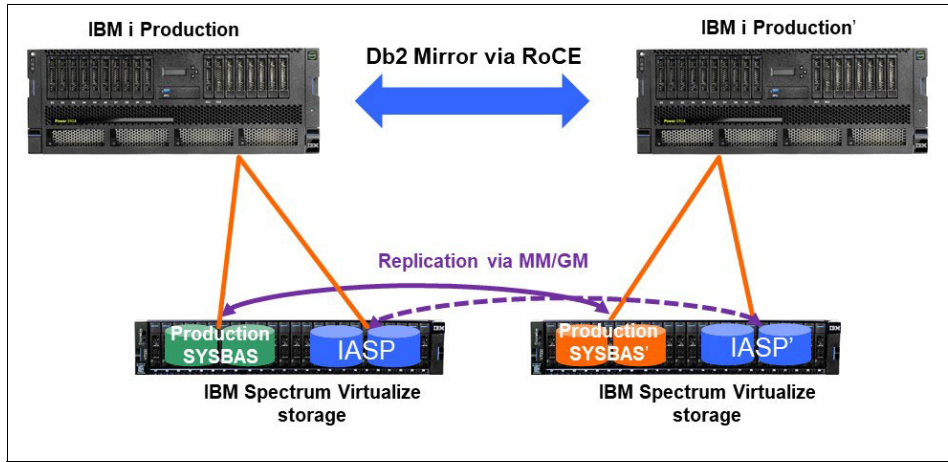


Figure A-15 Db2 Mirror environment with two IBM Spectrum Virtualize storages

### **Db2 Mirror and DR considerations**

Db2 Mirror is a continuous availability solution, but it is *not* considered a DR solution. However, Db2 Mirror can be used within your DR strategy to improve your availability, even within a disaster situation.

The Db2 Mirror nodes must be close to each other because the maximum distance between IBM Power Systems servers is 200 meters (656 feet).

- ▶ At Site 1 (the continuous-availability location), Db2 Mirror nodes are used.
- ▶ At Site 2 (the DR location), you can have a single server or multiple servers with Db2 Mirror nodes, and a unique or multiple IBM Spectrum 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 i using Metro Mirror
- ▶ Global Mirror with IASPs
- ▶ Full system replication
- ▶ Logical replication from third-party vendor

### **Db2 Mirror and full system replication**

The use 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 in DR.

A topology with multiple IBM Spectrum Virtualize storage systems and multiple IBM Power Systems servers is shown in Figure A-16 on page 558.



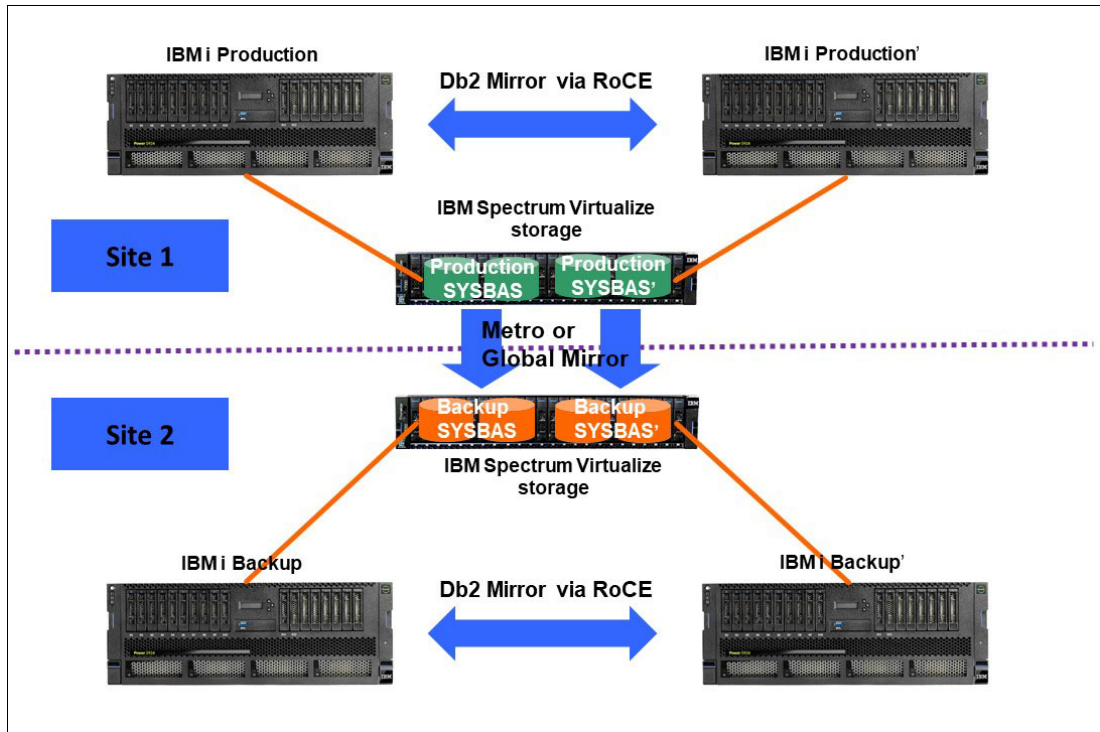


Figure A-16 DB2 Mirror and full system replication

Full system replication is fully supported. If you are not using IASP, you can do this type of replication for IBM i at IBM Spectrum Virtualize storage level.

In Figure A-16, two sites are configured as follows:

- ▶ Site 1: Db2 Mirror with DR and Db2 Mirror production
- ▶ Site 2: Db2 Mirror with DR

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 storage to be able 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.

- ▶ *IBM FlashSystem 9200 Product Guide*, REDP-5586
- ▶ *IBM FlashSystem 9100 Product Guide*, REDP-5524
- ▶ *IBM FlashSystem 7200 Product Guide*, REDP-5587
- ▶ *IBM FlashSystem 5000 and 5100 for Mid-Market*, REDP-5594
- ▶ *Automate and Orchestrate Your IBM FlashSystem Hybrid Cloud with Red Hat Ansible Version 1 Release 1*, REDP-5598
- ▶ *IBM FlashSystem 9100 Architecture, Performance, and Implementation*, SG24-8425
- ▶ *Implementing the IBM FlashSystem 5010 and FlashSystem 5030 with IBM Spectrum Virtualize V8.3.1*, SG24-8467
- ▶ *Implementing the IBM SAN Volume Controller with IBM Spectrum Virtualize V8.3.1*, SG24-8465
- ▶ *IBM SAN Volume Controller Best Practices and Performance Guidelines*, SG24-8502
- ▶ *IBM Spectrum Virtualize 3-Site Replication*, SG24-8474

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](https://ibm.com/redbooks)

## Online resources

These websites are also relevant as further information sources:

- ▶ IBM FlashSystem 9200 documentation:  
[https://www.ibm.com/support/knowledgecenter/en/STSLR9\\_8.4.0/com.ibm.fs9200\\_840.doc/fs9200\\_ichome.html](https://www.ibm.com/support/knowledgecenter/en/STSLR9_8.4.0/com.ibm.fs9200_840.doc/fs9200_ichome.html)
- ▶ IBM System Storage Interoperability Center (SSIC):  
<https://www-03.ibm.com/systems/support/storage/ssic/interoperability.wss>
- ▶ Configuration Limits and Restrictions for IBM Storwize V5100:  
<https://www.ibm.com/support/pages/node/6361755>
- ▶ Configuration Limits and Restrictions for IBM Storwize V7000:  
<https://www.ibm.com/support/pages/node/6361679>

## Help from IBM

IBM Support and downloads

[ibm.com/support](https://ibm.com/support)

IBM Global Services

[ibm.com/services](https://ibm.com/services)



**Redbooks**

# IBM FlashSystem Best Practices and Performance Guidelines

SG24-8503-00

ISBN 0738459704

(1.5" spine)  
1.5" <-> 1.998"  
789 <-> 1051 pages



**Redbooks**

# IBM FlashSystem Best Practices and Performance Guidelines

SG24-8503-00

ISBN 0738459704

(1.0" spine)  
0.875" <-> 1.498"  
460 <-> 788 pages



**Redbooks**

# IBM FlashSystem Best Practices and Performance Guidelines

SG24-8503-00

ISBN 0738459704

(0.5" spine)  
0.475" <-> 0.873"  
250 <-> 459 pages



**Redbooks**

# IBM FlashSystem Best Practices and Performance Guidelines

(0.2" spine)

0.17" <-> 0.473"

90 <-> 249 pages

(0.1" spine)

0.1" <-> 0.169"

53 <-> 89 pages



# IBM FlashSystem Best Practices and

SG24-8503-00

ISBN 0738459704

(2.5" spine)  
2.5" <-> mmm.n"  
1315 <-> mmm pages



# IBM FlashSystem Best Practices and Performance Guidelines

SG24-8503-00

ISBN 0738459704

(2.0" spine)  
2.0" <-> 2.498"  
1052 <-> 1314 pages







SG24-8503-00

ISBN 0738459704

Printed in U.S.A.

Get connected

