

IBM z13s Technical Guide

Octavian Lascu
Barbara Sannerud
Cecilia A. De Leon
Edzard Hoogerbrug
Ewerson Palacio
Franco Pinto
Jin J. Yang
John P. Troy
Martin Soellig



z Systems

In partnership with
IBM Academy of Technology



International Technical Support Organization

IBM z13s Technical Guide

June 2016

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

First Edition (June 2016)

This edition applies to IBM z13s servers.

© Copyright International Business Machines Corporation 2016. 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	xv
Notices	xix
Trademarks	xx
IBM Redbooks promotions	xxi
Preface	xxiii
Authors	xxiii
Now you can become a published author, too!	xxv
Comments welcome	xxv
Stay connected to IBM Redbooks	xxvi
Chapter 1. Introducing IBM z13s servers	1
1.1 Overview of IBM z13s servers	2
1.2 z13s servers highlights	3
1.2.1 Processor and memory	3
1.2.2 Capacity and performance	4
1.2.3 I/O subsystem and I/O features	5
1.2.4 Virtualization	8
1.2.5 Reliability, availability, and serviceability design	11
1.3 z13s technical overview	11
1.3.1 Models	12
1.3.2 Model upgrade paths	13
1.3.3 Frame	13
1.3.4 CPC drawer	14
1.3.5 I/O connectivity: PCIe and InfiniBand	16
1.3.6 I/O subsystem	16
1.3.7 Parallel Sysplex Coupling and Server Time Protocol connectivity	20
1.3.8 Special-purpose features	23
1.3.9 Reliability, availability, and serviceability	26
1.4 Hardware Management Consoles and Support Elements	27
1.5 IBM z BladeCenter Extension (zBX) Model 004	27
1.5.1 Blades	28
1.5.2 IBM WebSphere DataPower Integration Appliance XI50 for zEnterprise	28
1.6 IBM z Unified Resource Manager	29
1.7 IBM Dynamic Partition Manager	29
1.8 Operating systems and software	30
1.8.1 Supported operating systems	30
1.8.2 IBM compilers	33
Chapter 2. Central processor complex hardware components	35
2.1 Frame and drawers	36
2.1.1 The z13s frame	36
2.1.2 PCIe I/O drawer and I/O drawer features	38
2.2 Processor drawer concept	39
2.2.1 CPC drawer interconnect topology	43
2.2.2 Oscillator	43
2.2.3 System control	44

2.2.4	CPC drawer power	46
2.3	Single chip modules	46
2.3.1	Processor units and system control chips	47
2.3.2	Processor unit (PU) chip	47
2.3.3	Processor unit (core)	49
2.3.4	PU characterization	50
2.3.5	System control chip	51
2.3.6	Cache level structure	52
2.4	Memory	53
2.4.1	Memory subsystem topology	54
2.4.2	Redundant array of independent memory	55
2.4.3	Memory configurations	56
2.4.4	Memory upgrades	63
2.4.5	Preplanned memory	63
2.5	Reliability, availability, and serviceability	64
2.5.1	RAS in the CPC memory subsystem	64
2.5.2	General z13s RAS features	65
2.6	Connectivity	65
2.6.1	Redundant I/O interconnect	67
2.7	Model configurations	69
2.7.1	Upgrades	71
2.7.2	Concurrent PU conversions	72
2.7.3	Model capacity identifier	72
2.7.4	Model capacity identifier and MSU values	73
2.7.5	Capacity BackUp	74
2.7.6	On/Off Capacity on Demand and CPs	76
2.8	Power and cooling	77
2.8.1	Power considerations	77
2.8.2	High-voltage DC power	78
2.8.3	Internal Battery Feature	78
2.8.4	Power capping	78
2.8.5	Power Estimation tool	78
2.8.6	Cooling requirements	79
2.9	Summary of z13s structure	79
Chapter 3. Central processor complex system design		81
3.1	Overview	82
3.2	Design highlights	82
3.3	CPC drawer design	84
3.3.1	Cache levels and memory structure	84
3.3.2	CPC drawer interconnect topology	87
3.4	Processor unit design	88
3.4.1	Simultaneous multithreading	89
3.4.2	Single-instruction multiple-data	90
3.4.3	Out-of-order execution	92
3.4.4	Superscalar processor	95
3.4.5	Compression and cryptography accelerators on a chip	95
3.4.6	Decimal floating point accelerator	96
3.4.7	IEEE floating point	97
3.4.8	Processor error detection and recovery	97
3.4.9	Branch prediction	98
3.4.10	Wild branch	98
3.4.11	Translation lookaside buffer	99

3.4.12	Instruction fetching, decoding, and grouping	99
3.4.13	Extended Translation Facility	99
3.4.14	Instruction set extensions	100
3.4.15	Transactional Execution	100
3.4.16	Runtime Instrumentation	100
3.5	Processor unit functions	100
3.5.1	Overview	100
3.5.2	Central processors	102
3.5.3	Integrated Facility for Linux	103
3.5.4	Internal Coupling Facility	103
3.5.5	IBM z Integrated Information Processor	105
3.5.6	System assist processors	109
3.5.7	Reserved processors	110
3.5.8	Integrated firmware processor	110
3.5.9	Processor unit assignment	110
3.5.10	Sparing rules	111
3.5.11	Increased flexibility with z/VM mode partitions	112
3.6	Memory design	112
3.6.1	Overview	112
3.6.2	Main storage	115
3.6.3	Expanded storage	115
3.6.4	Hardware system area (HSA)	116
3.7	Logical partitioning	116
3.7.1	Overview	117
3.7.2	Storage operations	124
3.7.3	Reserved storage	127
3.7.4	Logical partition storage granularity	128
3.7.5	LPAR dynamic storage reconfiguration	129
3.8	Intelligent Resource Director	129
3.9	Clustering technology	131
3.9.1	Coupling Facility Control Code	132
3.9.2	Coupling Thin Interrupts	133
3.9.3	Dynamic CF dispatching	133
3.9.4	CFCC and Flash Express use	135
Chapter 4. Central processor complex I/O system structure		137
4.1	Introduction to the InfiniBand and PCIe for I/O infrastructure	138
4.1.1	InfiniBand I/O infrastructure	138
4.1.2	PCIe I/O infrastructure	138
4.1.3	InfiniBand specifications	139
4.1.4	PCIe Generation 3	139
4.2	I/O system overview	140
4.2.1	Characteristics	140
4.2.2	Summary of supported I/O features	141
4.3	I/O drawer	142
4.4	PCIe I/O drawer	143
4.5	PCIe I/O drawer and I/O drawer offerings	146
4.6	Fanouts	147
4.6.1	PCIe Generation 3 fanout (FC 0173)	149
4.6.2	HCA2-C fanout (FC 0162)	149
4.6.3	Integrated Coupling Adapter (FC 0172)	149
4.6.4	HCA3-O (12x IFB) fanout (FC 0171)	150
4.6.5	HCA3-O LR (1x IFB) fanout (FC 0170)	151

4.6.6 Fanout considerations	152
4.7 I/O features (cards)	154
4.7.1 I/O feature card ordering information	155
4.7.2 Physical channel report	156
4.8 Connectivity	157
4.8.1 I/O feature support and configuration rules	158
4.8.2 FICON channels	160
4.8.3 OSA-Express5S	166
4.8.4 OSA-Express4S features	168
4.8.5 OSA-Express for ensemble connectivity	171
4.8.6 HiperSockets	172
4.9 Parallel Sysplex connectivity	174
4.9.1 Coupling links	174
4.9.2 Migration considerations	178
4.9.3 Pulse per second input	183
4.10 Cryptographic functions	183
4.10.1 CPACF functions (FC 3863)	183
4.10.2 Crypto Express5S feature (FC 0890)	183
4.11 Integrated firmware processor	183
4.12 Flash Express	184
4.12.1 IBM Flash Express read/write cache	185
4.13 10GbE RoCE Express	185
4.14 zEDC Express	186
Chapter 5. Central processor complex channel subsystem	187
5.1 Channel subsystem	188
5.1.1 Multiple logical channel subsystems	189
5.1.2 Multiple subchannel sets	190
5.1.3 Channel path spanning	193
5.2 I/O configuration management	196
5.3 Channel subsystem summary	196
Chapter 6. Cryptography	199
6.1 Cryptography in IBM z13 and z13s servers	200
6.2 Some fundamentals on cryptography	200
6.2.1 Modern cryptography	200
6.2.2 Kerckhoffs' principle	201
6.2.3 Keys	202
6.2.4 Algorithms	203
6.3 Cryptography on IBM z13s servers	204
6.4 CP Assist for Cryptographic Functions	207
6.4.1 Cryptographic synchronous functions	208
6.4.2 CPACF protected key	209
6.5 Crypto Express5S	211
6.5.1 Cryptographic asynchronous functions	213
6.5.2 Crypto Express5S as a CCA coprocessor	214
6.5.3 Crypto Express5S as an EP11 coprocessor	218
6.5.4 Crypto Express5S as an accelerator	218
6.5.5 Management of Crypto Express5S	219
6.6 TKE workstation	222
6.6.1 Logical partition, TKE host, and TKE target	222
6.6.2 Optional smart card reader	222
6.6.3 TKE workstation with Licensed Internal Code 8.0	223

6.6.4	TKE workstation with Licensed Internal Code 8.1	223
6.6.5	TKE hardware support and migration information	224
6.7	Cryptographic functions comparison	225
6.8	Cryptographic software support	227
Chapter 7.	Software support	229
7.1	Operating systems summary	230
7.2	Support by operating system	230
7.2.1	z/OS	231
7.2.2	z/VM	231
7.2.3	z/VSE	232
7.2.4	z/TPF	232
7.2.5	Linux on z Systems	232
7.2.6	KVM for IBM z Systems	233
7.2.7	z13s function support summary	233
7.3	Support by function	243
7.3.1	Single system image	243
7.3.2	zIIP support	245
7.3.3	Transactional Execution	246
7.3.4	Maximum main storage size	247
7.3.5	Flash Express	247
7.3.6	z Enterprise Data Compression Express	249
7.3.7	10GbE RoCE Express	249
7.3.8	Large page support	250
7.3.9	Hardware decimal floating point	251
7.3.10	Up to 40 LPARs	251
7.3.11	Separate LPAR management of PUs	252
7.3.12	Dynamic LPAR memory upgrade	252
7.3.13	LPAR physical capacity limit enforcement	252
7.3.14	Capacity Provisioning Manager	253
7.3.15	Dynamic PU add	253
7.3.16	HiperDispatch	254
7.3.17	The 63.75-K subchannels	255
7.3.18	Multiple Subchannel Sets	256
7.3.19	Three subchannel sets	256
7.3.20	IPL from an alternative subchannel set	257
7.3.21	Modified Indirect Data Address Word facility	257
7.3.22	HiperSockets Completion Queue	257
7.3.23	HiperSockets integration with the intraensemble data network	258
7.3.24	HiperSockets Virtual Switch Bridge	258
7.3.25	HiperSockets Multiple Write Facility	259
7.3.26	HiperSockets IPv6	259
7.3.27	HiperSockets Layer 2 support	259
7.3.28	HiperSockets network traffic analyzer for Linux on z Systems	260
7.3.29	FICON Express16S	260
7.3.30	FICON Express8S	261
7.3.31	FICON Express8	262
7.3.32	z/OS Discovery and Auto-Configuration	263
7.3.33	High-performance FICON	264
7.3.34	Request node identification data	265
7.3.35	32 K subchannels for the FICON Express16S	266
7.3.36	Extended distance FICON	266
7.3.37	Platform and name server registration in FICON channel	266

7.3.38	FICON link incident reporting	267
7.3.39	FCP provides increased performance.	267
7.3.40	N_Port ID Virtualization.	267
7.3.41	OSA-Express5S 10-Gigabit Ethernet LR and SR	268
7.3.42	OSA-Express5S Gigabit Ethernet LX and SX.	268
7.3.43	OSA-Express5S 1000BASE-T Ethernet	269
7.3.44	OSA-Express4S 10-Gigabit Ethernet LR and SR	270
7.3.45	OSA-Express4S Gigabit Ethernet LX and SX.	271
7.3.46	OSA-Express4S 1000BASE-T Ethernet	272
7.3.47	Open Systems Adapter for IBM zAware	273
7.3.48	Open Systems Adapter for Ensemble.	273
7.3.49	Intranode management network	274
7.3.50	Intraensemble data network	274
7.3.51	OSA-Express5S and OSA-Express4S NCP support	274
7.3.52	Integrated Console Controller	275
7.3.53	VLAN management enhancements	276
7.3.54	GARP VLAN Registration Protocol	276
7.3.55	Inbound workload queuing for OSA-Express5S and OSA-Express4S	276
7.3.56	Inbound workload queuing for Enterprise Extender	277
7.3.57	Querying and displaying an OSA configuration	277
7.3.58	Link aggregation support for z/VM	278
7.3.59	Multi-VSwitch Link Aggregation	278
7.3.60	QDIO data connection isolation for z/VM	278
7.3.61	QDIO interface isolation for z/OS	278
7.3.62	QDIO optimized latency mode	279
7.3.63	Large send for IPv6 packets	279
7.3.64	OSA-Express5S and OSA-Express4S checksum offload.	280
7.3.65	Checksum offload for IPv4and IPv6 packets when in QDIO mode.	280
7.3.66	Adapter interruptions for QDIO	280
7.3.67	OSA Dynamic LAN idle.	281
7.3.68	OSA Layer 3 virtual MAC for z/OS environments	281
7.3.69	QDIO Diagnostic Synchronization.	281
7.3.70	Network Traffic Analyzer.	281
7.3.71	Program-directed re-IPL	282
7.3.72	Coupling over InfiniBand and Integrated Coupling Adapter	282
7.3.73	Dynamic I/O support for InfiniBand and ICA CHPIDs	283
7.3.74	Simultaneous multithreading.	283
7.3.75	Single Instruction Multiple Data.	284
7.3.76	Shared Memory Communication - Direct Memory Access	284
7.4	Cryptographic support.	285
7.4.1	CP Assist for Cryptographic Function	285
7.4.2	Crypto Express5S.	286
7.4.3	Web deliverables	286
7.4.4	z/OS Integrated Cryptographic Service Facility FMIDs.	286
7.4.5	ICSF migration considerations	289
7.5	GDPS Virtual Appliance	289
7.6	z/OS migration considerations	289
7.6.1	General guidelines	290
7.6.2	Hardware configuration definition	290
7.6.3	Coupling links	290
7.6.4	Large page support.	290
7.6.5	Capacity Provisioning Manager	290
7.6.6	Decimal floating point and z/OS XL C/C++ considerations.	291

7.7 IBM z Advanced Workload Analysis Reporter (zAware)	291
7.7.1 z Appliance Container Infrastructure mode LPAR	292
7.8 Coupling facility and CFCC considerations	292
7.8.1 CFCC Level 21	293
7.8.2 Flash Express exploitation by CFCC	294
7.8.3 CFCC Coupling Thin Interrupts	295
7.9 Simultaneous multithreading	295
7.10 Single-instruction multiple-data	297
7.11 The MIDAW facility	298
7.11.1 MIDAW technical description	299
7.11.2 Extended format data sets	301
7.11.3 Performance benefits	301
7.12 IOCP	302
7.13 Worldwide port name tool	303
7.14 ICKDSF	303
7.15 IBM z BladeCenter Extension (zBX) Model 004 software support	304
7.15.1 IBM Blades	304
7.15.2 IBM WebSphere DataPower Integration Appliance XI50 for zEnterprise	304
7.16 Software licensing	305
7.16.1 Software licensing considerations	305
7.16.2 Monthly license charge pricing metrics	306
7.16.3 Advanced Workload License Charges	307
7.16.4 Advanced Entry Workload License Charge	308
7.16.5 System z New Application License Charges	308
7.16.6 Midrange workload license charges	308
7.16.7 Parallel Sysplex License Charges	309
7.16.8 z Systems International Program License Agreement	309
7.16.9 zBX licensed software	310
7.17 References	310
Chapter 8. System upgrades	311
8.1 Upgrade types	312
8.1.1 Overview of upgrade types	312
8.1.2 Terminology related to CoD for z13s systems	313
8.1.3 Permanent upgrades	315
8.1.4 Temporary upgrades	316
8.2 Concurrent upgrades	317
8.2.1 Model upgrades	317
8.2.2 Customer Initiated Upgrade facility	319
8.2.3 Summary of concurrent upgrade functions	322
8.3 Miscellaneous equipment specification upgrades	323
8.3.1 MES upgrade for processors	324
8.3.2 MES upgrade for memory	324
8.3.3 Preplanned Memory feature	325
8.3.4 MES upgrades for the zBX	327
8.4 Permanent upgrade through the CIU facility	328
8.4.1 Ordering	330
8.4.2 Retrieval and activation	331
8.5 On/Off Capacity on Demand	332
8.5.1 Overview	333
8.5.2 Ordering	334
8.5.3 On/Off CoD testing	337
8.5.4 Activation and deactivation	338

8.5.5 Termination	338
8.5.6 IBM z/OS capacity provisioning	339
8.6 Capacity for Planned Event.	343
8.7 Capacity BackUp.	345
8.7.1 Ordering	345
8.7.2 CBU activation and deactivation	347
8.7.3 Automatic CBU for Geographically Dispersed Parallel Sysplex	349
8.8 Nondisruptive upgrades	349
8.8.1 Processors	350
8.8.2 Memory	350
8.8.3 I/O	350
8.8.4 Cryptographic adapters.	350
8.8.5 Special features	350
8.8.6 Concurrent upgrade considerations	350
8.9 Summary of capacity on-demand offerings.	353
Chapter 9. Reliability, availability, and serviceability.	355
9.1 The RAS strategy	356
9.2 Availability characteristics	356
9.3 RAS functions	358
9.3.1 Unscheduled outages	358
9.3.2 Scheduled outages	359
9.4 Enhanced Driver Maintenance	360
9.5 RAS capability for the HMC and SE	362
9.6 RAS capability for zBX Model 004	363
BladeCenter components	363
zBX firmware.	364
9.6.1 zBX RAS and the IBM z Unified Resource Manager	364
9.6.2 zBX Model 004: 2458-004	364
9.7 Considerations for PowerHA in zBX environment.	366
9.8 IBM z Advanced Workload Analysis Reporter	367
9.9 RAS capability for Flash Express	368
Chapter 10. Environmental requirements	369
10.1 IBM z13s power and cooling.	370
10.1.1 Power and I/O cabling.	370
10.1.2 Power consumption	371
10.1.3 Internal Battery Feature	372
10.1.4 Balanced Power Plan Ahead	373
10.1.5 Emergency power-off	373
10.1.6 Cooling requirements	373
10.1.7 New environmental class for IBM z13s servers: ASHREA Class A3	374
10.2 IBM z13s physical specifications.	374
10.2.1 Weights and dimensions.	375
10.2.2 Four-in-one (4-in-1) bolt-down kit	375
10.3 IBM zBX environmental components	375
10.3.1 IBM zBX configurations.	376
10.3.2 IBM zBX power components.	376
10.3.3 IBM zBX cooling	377
10.3.4 IBM zBX physical specifications	379
10.4 Energy management.	380
10.4.1 Power estimation tool	381
10.4.2 Query maximum potential power	381

10.4.3	System Activity Display and Monitors Dashboard	382
10.4.4	IBM Systems Director Active Energy Manager	383
10.4.5	Unified Resource Manager: Energy management	384
Chapter 11.	Hardware Management Console and Support Elements	387
11.1	Introduction to the HMC and SE	388
11.2	HMC and SE enhancements and changes	388
11.2.1	Driver Level 27 HMC and SE enhancements and changes	389
11.2.2	Rack-mounted HMC	392
11.2.3	New Support Elements	393
11.2.4	New backup options for HMCs and primary SEs	393
11.2.5	SE driver support with the HMC driver	396
11.2.6	HMC feature codes	397
11.2.7	Tree Style User Interface and Classic Style User Interface	398
11.3	HMC and SE connectivity	398
11.3.1	Network planning for the HMC and SE	400
11.3.2	Hardware prerequisite changes	401
11.3.3	RSF is broadband-only	402
11.3.4	TCP/IP Version 6 on the HMC and SE	402
11.3.5	Assigning addresses to the HMC and SE	402
11.4	Remote Support Facility	403
11.4.1	Security characteristics	403
11.4.2	RSF connections to IBM and Enhanced IBM Service Support System	404
11.4.3	HMC and SE remote operations	404
11.5	HMC and SE key capabilities	405
11.5.1	Central processor complex management	405
11.5.2	Logical partition management	406
11.5.3	Operating system communication	407
11.5.4	HMC and SE microcode	409
11.5.5	Monitoring	411
11.5.6	Capacity on demand support	415
11.5.7	Features on Demand support	416
11.5.8	Server Time Protocol support	416
11.5.9	NTP client and server support on the HMC	420
11.5.10	Security and user ID management	422
11.5.11	System Input/Output Configuration Analyzer on the SE and HMC	424
11.5.12	Automated operations	424
11.5.13	Cryptographic support	425
11.5.14	Installation support for z/VM using the HMC	426
11.5.15	Dynamic Partition Manager	426
11.6	HMC in an ensemble	428
11.6.1	Unified Resource Manager	428
11.6.2	Ensemble definition and management	432
11.6.3	HMC availability	434
11.6.4	Considerations for multiple HMCs	434
11.6.5	HMC browser session to a primary HMC	434
11.6.6	HMC ensemble topology	434
Chapter 12.	Performance	437
12.1	Performance information	438
12.2	LSPR workload suite	439
12.3	Fundamental components of workload capacity performance	439
12.3.1	Instruction path length	439

12.3.2	Instruction complexity	440
12.3.3	Memory hierarchy and memory nest.	440
12.4	Relative nest intensity	441
12.5	LSPR workload categories based on relative nest intensity	443
12.6	Relating production workloads to LSPR workloads	444
12.7	Workload performance variation	445
12.7.1	Main performance improvement drivers with z13s	446
Appendix A. IBM z Appliance Container Infrastructure		449
A.1	What is zACI?	450
A.2	Why use zACI?	450
A.3	IBM z Systems servers and zACI	450
12.7.2	Example: Deploying IBM zAware	451
Appendix B. IBM z Systems Advanced Workload Analysis Reporter (IBM zAware).		453
B.1	Troubleshooting in complex IT environments	454
B.2	Introducing IBM zAware	455
B.2.1	Hardware requirements overview	456
B.2.2	Value of IBM zAware	456
B.2.3	IBM z/OS Solutions to improve problem diagnostic procedures.	457
B.3	Understanding IBM zAware technology	458
B.3.1	Training period	463
B.3.2	Priming IBM zAware	463
B.3.3	IBM zAware ignore message support.	463
B.3.4	IBM zAware graphical user interface	464
B.3.5	IBM zAware complements your existing tools	464
B.4	IBM zAware prerequisites	464
B.4.1	IBM zAware features and ordering	464
12.7.3	Feature on Demand (FoD)	465
B.4.2	IBM zAware operating requirements	467
B.5	Configuring and using IBM zAware virtual appliance	468
Appendix C. Channel options		471
C.1	Channel options supported on z13s servers	472
C.2	Maximum unrepeated distance for FICON SX features	473
Appendix D. Shared Memory Communications		475
D.1	Shared Memory Communications overview	476
D.2	Shared Memory Communication over RDMA.	476
D.2.1	RDMA technology overview	476
D.2.2	Shared Memory Communications over RDMA.	477
D.2.3	Single Root I/O Virtualization	478
D.2.4	Hardware	479
D.2.5	10GbE RoCE Express feature	479
D.2.6	10GbE RoCE Express configuration example	481
D.2.7	Hardware configuration definitions	483
D.2.8	Software exploitation of SMC-R	484
D.2.9	SMC-R support overview	485
D.2.10	SMC-R use cases for z/OS to z/OS	486
D.2.11	Enabling SMC-R support in z/OS Communications Server	488
D.3	Shared Memory Communications - Direct Memory Access	489
D.3.1	Internal Shared Memory technology overview	490
D.3.2	SMC-D over Internal Shared Memory	490
D.3.3	Internal Shared Memory - Introduction.	492

D.3.4 Virtual PCI Function (vPCI Adapter)	492
D.3.5 Planning considerations	495
D.3.6 Hardware configuration definitions	495
D.3.7 Sample IOCP FUNCTION statements	496
D.3.8 Software exploitation of ISM	498
D.3.9 SMC-D over ISM prerequisites	498
D.3.10 Enabling SMC-D support in z/OS Communications Server	499
D.3.11 SMC-D support overview	500
Appendix E. IBM Dynamic Partition Manager	501
E.1 What is IBM Dynamic Partition Manager?	502
E.2 Why use DPM?	502
E.3 IBM z Systems servers and DPM	503
E.4 Setting up the DPM environment	504
E.4.1 Defining partitions in DPM mode	507
E.4.2 Summary	510
Appendix F. KVM for IBM z Systems	511
F.1 Why KVM for IBM z Systems	512
F.1.1 Advantages of using KVM for IBM z Systems	512
F.2 IBM z Systems servers and KVM	513
F.2.1 Storage connectivity	514
F.2.2 Network connectivity	514
F.2.3 Hardware Management Console	515
F.2.4 Open source virtualization	515
F.2.5 What comes with KVM for IBM z Systems	516
F.3 Managing the KVM for IBM z Systems environment	518
F.3.1 Hypervisor Performance Manager	520
F.4 Using IBM Cloud Manager with OpenStack	520
Appendix G. Native Peripheral Component Interconnect Express	521
G.1 Design of native PCIe I/O adapter management	522
G.1.1 Native PCIe adapter	522
G.1.2 Integrated firmware processor	522
G.1.3 Resource groups	523
G.1.4 Management tasks	524
G.2 Native PCIe feature plugging rules	524
G.3 Native PCIe feature definitions	525
G.3.1 FUNCTION identifier	527
G.3.2 Virtual function number	527
G.3.3 Physical network identifier	527
Appendix H. Flash Express	529
H.1 Flash Express overview	530
H.2 Using Flash Express	532
H.3 Security on Flash Express	535
H.3.1 Integrated Key Controller	535
H.3.2 Key serving topology	537
H.3.3 Error recovery scenarios	538
Appendix I. GDPS Virtual Appliance	541
I.1 GDPS overview	542
I.2 Overview of GDPS Virtual Appliance	544
I.3 GDPS Virtual Appliance recovery scenarios	547

I.3.1 Planned disk outage	547
I.3.2 Unplanned disk outage	548
I.3.3 Disaster recovery	549
Appendix J. IBM zEnterprise Data Compression Express	551
J.1 Overview	552
J.2 zEDC Express	552
J.3 Software support	553
J.3.1 z/VM V6R3 support with PTFs	554
J.3.2 IBM SDK 7 for z/OS Java support.	554
J.3.3 IBM z Systems Batch Network Analyzer.	554
Related publications	555
IBM Redbooks	555
Other publications	555
Online resources	555
Help from IBM	555

Figures

2-1 z13s frame: Rear and front view	36
2-2 z13s (one CPC drawer) I/O drawer configurations	38
2-3 z13s (two CPC Drawers) I/O Drawer Configurations	38
2-4 Model N10 components (top view)	40
2-5 Model N20 components (top view)	41
2-6 CPC drawer (front view)	41
2-7 CPC drawer logical structure	42
2-8 Drawer to drawer communication	43
2-9 Oscillators cards	44
2-10 Conceptual overview of system control elements	45
2-11 Redundant DCAs and blowers for CPC drawers	46
2-12 Single chip modules (PU SCM and SC SCM) N20 CPC Drawer	47
2-13 PU Chip Floorplan	48
2-14 PU Core floorplan	50
2-15 SC chip diagram	51
2-16 Cache level structure	52
2-17 CPC drawer memory topology	54
2-18 RAIM configuration per node	55
2-19 Model N10 memory plug locations	56
2-20 Model N20 one drawer memory plug locations	58
2-21 Model N20 two drawer memory plug locations	60
2-22 Memory allocation diagram	62
2-23 Model N10 drawer: Location of the PCIe and IFB fanouts	66
2-24 Model N20 two CPC drawer: Locations of the PCIe and IFB fanouts	66
2-25 Redundant I/O interconnect for I/O drawer	68
2-26 Redundant I/O interconnect for PCIe I/O drawer	69
2-27 z13s upgrade paths	71
3-1 z13s cache levels and memory hierarchy	84
3-2 z13s cache topology	85
3-3 z13s and zBC12 cache level comparison	86
3-4 z13s CPC drawer communication topology	87
3-5 Point-to-point topology z13s two CPC drawers communication	88
3-6 Two threads running simultaneously on the same processor core	90
3-7 Schematic representation of add SIMD instruction with 16 elements in each vector	91
3-8 Floating point registers overlaid by vector registers	91
3-9 z13s PU core logical diagram	93
3-10 z13s in-order and out-of-order core execution improvements	94
3-11 Compression and cryptography accelerators on a core in the chip	96
3-12 PU error detection and recovery	98
3-13 ICF options: Shared ICFs	105
3-14 Logical flow of Java code execution on a zIIP	106
3-15 IRD LPAR cluster example	129
3-16 Sysplex hardware overview	131
3-17 Dynamic CF dispatching (shared CPs or shared ICF PUs)	134
4-1 I/O drawer	142
4-2 I/O domains of an I/O drawer	143
4-3 PCIe I/O drawer	144
4-4 z13s (N20) connectivity to PCIe I/O drawers	145

4-5	PCIe I/O drawer with 32 PCIe slots and four I/O domains	146
4-6	Infrastructure for PCIe and InfiniBand coupling links	148
4-7	OM3 50/125 μ m multimode fiber cable with MPO connectors	150
4-8	z13 Parallel Sysplex coupling connectivity	176
4-9	CPC drawer front view: Coupling links	179
4-10	HCA3-O Fanouts: z13s versus z114 / zBC12 servers	180
4-11	PCIe I/O drawer that is fully populated with Flash Express cards.	185
5-1	Multiple channel subsystems and multiple subchannel sets.	188
5-2	Output for display ios,config(all) command.	193
5-3	z Systems CSS: Channel subsystems with channel spanning	194
5-4	CSS, LPAR, and identifier example	195
6-1	Three levels of protection with three levels of speed.	203
6-2	Cryptographic hardware supported on IBM z13s servers.	204
6-3	z13s cryptographic support in z/OS	206
6-4	The cryptographic coprocessor CPACF	207
6-5	CPACF key wrapping	210
6-6	Customize Image Profiles: Crypto	220
6-7	SE: View LPAR Cryptographic Controls	221
7-1	Result of the display core command.	296
7-2	Simultaneous Multithreading.	297
7-3	IDAW usage	299
7-4	MIDAW format	300
7-5	MIDAW usage.	300
8-1	The provisioning architecture	320
8-2	Example of temporary upgrade activation sequence	321
8-3	Memory sizes and upgrades for the N10	325
8-4	Memory sizes and upgrades for the single drawer N20	326
8-5	Memory sizes and upgrades for the two drawer N20	326
8-6	Feature on Demand window for zBX Blades features HWMs.	327
8-7	Permanent upgrade order example	329
8-8	CIU-eligible order activation example	330
8-9	Machine profile	331
8-10	IBM z13s Perform Model Conversion window	332
8-11	Customer Initiated Upgrade Order Activation Number window.	332
8-12	Order On/Off CoD record window.	336
8-13	On/Off CoD order example	337
8-14	The capacity provisioning infrastructure	339
8-15	A Capacity Provisioning Domain.	341
8-16	Example of C02 with three CBU features	348
8-17	STSI output on z13s server	351
9-1	Typical PowerHA cluster diagram.	367
9-2	Flash Express RAS components	368
10-1	IBM z13s cabling options	370
10-2	Top Exit I/O cabling feature	371
10-3	Recommended environmental conditions.	374
10-4	Rear Door Heat eXchanger (left) and functional diagram.	379
10-5	Top Exit Support for the zBX	380
10-6	Maximum potential power.	382
10-7	Power usage on the System Activity Display	382
11-1	Diagnostic sampling authorization control	390
11-2	Change LPAR security	391
11-3	Rack-mounted HMC	392
11-4	SEs location	393

11-5	Configure backup FTP server	394
11-6	Backup Critical Console Data destinations	394
11-7	Backup Critical Data destinations of SEs	395
11-8	Scheduled Operation for HMC backup	396
11-9	Scheduled Operation for SEs backup	396
11-10	HMC and SE connectivity	399
11-11	SE Physical connection	399
11-12	HMC connectivity examples	401
11-13	Change LPAR Group Controls - Group absolute capping	407
11-14	Manage Trusted Signing Certificates	408
11-15	Import Remote Certificate example	408
11-16	Configure 3270 Emulators	408
11-17	Microcode terms and interaction	410
11-18	System Information: Bundle level	411
11-19	HMC Monitor task group	412
11-20	Monitors Dashboard task	413
11-21	Display the activity for an LPAR by processor type	413
11-22	Display the SMT usage	413
11-23	Monitors Dashboard: Crypto function integration	414
11-24	Monitors Dashboard - Flash Express function integration	414
11-25	Environmental Efficiency Statistics	415
11-26	Customize Console Date and Time	418
11-27	Timing Network window with Scheduled DST and Scheduled leap second offset	419
11-28	HMC NTP broadband authentication support	421
11-29	Time coordination for zBX components	422
11-30	Cryptographic Configuration window	426
11-31	Enabling Dynamic Partition Manager	427
11-32	HMC welcome window (CPC in DPM Mode)	428
11-33	Unified Resource Manager functions and suites	429
11-34	Ensemble example with primary and alternate HMCs	435
12-1	z13s to zBC12, z114, z10 BC, and z9 BC performance comparison	438
12-2	Memory hierarchy on the z13s one CPC drawer system (two nodes)	441
12-3	Relative Nest Intensity	442
12-4	The traditional factors that were used to categorize workloads	442
12-5	New z/OS workload categories defined	444
A-1	zACI Framework basic outline	451
A-2	IBM zAware Image Profile based on zACI	452
A-3	zACI icon in the HMC interface	452
B-1	IBM zAware complements an existing environment	456
B-2	IBM zAware shortens the business impact of a problem	457
B-3	Basic components of the IBM zAware environment	459
B-4	HMC Image Profile for an IBM zAware LPAR	460
B-5	HMC Image Profile for an IBM zAware LPAR	460
B-6	IBM zAware Heat Map view analysis	461
B-7	IBM zAware bar score with intervals	462
B-8	Feature on Demand window for IBM zAware feature	466
D-1	RDMA technology overview	477
D-2	Dynamic transition from TCP to SMC-R	478
D-3	Shared RoCE mode concepts	479
D-4	10GbE RoCE Express	480
D-5	10GbE RoCE Express sample configuration	481
D-6	RNIC and OSD interaction	482
D-7	Physical network ID example	484

D-8	Reduced latency and improved wall clock time with SMC-R	485
D-9	Sysplex Distributor before RoCE	487
D-10	Sysplex Distributor after RoCE	488
D-11	Connecting two LPARs on the same CPC using SMC-D	490
D-12	Clustered systems: Multitier application solution. RDMA, and DMA	490
D-13	Dynamic transition from TCP to SMC-D by using two OSA-Express adapters	491
D-14	Concept of vPCI adapter implementation	493
D-15	SMC-D configuration that uses Ethernet to provide connectivity	494
D-16	SMC-D configuration that uses HiperSockets to provide connectivity	495
D-17	ISM adapters shared between LPARs	496
D-18	Multiple LPARs connected through multiple VLANs	497
D-19	SMCD parameter in GLOBALCONFIG	499
E-1	High-level view of DPM implementation	503
E-2	Enabling DPM mode of operation from the SE CPC configuration options	504
E-3	Entering the OSA ports that will be used by the management network	505
E-4	DPM mode welcome window	506
E-5	Traditional HMC Welcome window when no CPCs are running in DPM mode	507
E-6	User Options when the HMC presents the DPM welcome window	507
E-7	DPM wizard welcome window options	509
F-1	KVM running in z Systems LPARs	514
F-2	Open source virtualization (KVM for IBM z Systems)	516
F-3	KVM for IBM z Systems management interface	518
F-4	KVM management by using the libvirt API layers	519
G-1	I/O domains and resource groups managed by the IFP	523
G-2	Sample output of AO data or PCHID report	525
G-3	Example of IOCP statements for zEDC Express and 10GbE RoCE Express	527
H-1	z Systems storage hierarchy	530
H-2	Flash Express PCIe adapter	531
H-3	PCIe I/O drawer that is fully populated with Flash Express cards	531
H-4	Sample SE/HMC window for Flash Express allocation to LPAR	533
H-5	Flash Express allocation in z/OS LPARs	534
H-6	Integrated Key Controller	536
H-7	Key serving topology	538
I-1	GDPS offerings	542
I-2	Positioning a virtual appliance	544
I-3	GDPS virtual appliance architecture overview	546
I-4	GDPS Storage failover	548
J-1	Relationships between the PCIe I/O drawer card slots, I/O domains, and resource groups	553

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](http://www.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 z Systems™	Redpapers™
CICS®	IBM z13™	Redbooks (logo)  ®
Cognos®	IBM®	Resource Link®
DataPower®	IMS™	Resource Measurement Facility™
DB2 Connect™	Language Environment®	RMF™
DB2®	Lotus®	System Storage®
developerWorks®	MVS™	System z10®
Distributed Relational Database Architecture™	NetView®	System z9®
Domino®	OMEGAMON®	System z®
DRDA®	Parallel Sysplex®	SystemMirror®
DS8000®	Passport Advantage®	Tivoli®
ECKD™	Power Systems™	VIA®
FICON®	POWER6®	VTAM®
FlashCopy®	POWER7®	WebSphere®
GDPS®	PowerHA®	z Systems™
Geographically Dispersed Parallel Sysplex™	PowerPC®	z/Architecture®
HACMP™	PowerVM®	z/OS®
HiperSockets™	PR/SM™	z/VM®
HyperSwap®	Processor Resource/Systems Manager™	z/VSE®
IBM Systems Director Active Energy Manager™	RACF®	z10™
	Redbooks®	z13™
	Redpaper™	z9®
		zEnterprise®

The following terms are trademarks of other companies:

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

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.

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

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

Find and read thousands of IBM Redbooks publications

- ▶ Search, bookmark, save and organize favorites
- ▶ Get personalized notifications of new content
- ▶ Link to the latest Redbooks blogs and videos

Get the latest version of the **Redbooks Mobile App**



Promote your business in an IBM Redbooks publication

Place a Sponsorship Promotion in an IBM® Redbooks® publication, featuring your business or solution with a link to your web site.

Qualified IBM Business Partners may place a full page promotion in the most popular Redbooks publications. Imagine the power of being seen by users who download millions of Redbooks publications each year!



ibm.com/Redbooks
About Redbooks → Business Partner Programs

THIS PAGE INTENTIONALLY LEFT BLANK

Preface

Digital business has been driving the transformation of underlying information technology (IT) infrastructure to be more efficient, secure, adaptive, and integrated. IT must be able to handle the explosive growth of mobile clients and employees. It also must be able to process enormous amounts of data to provide deep and real-time insights to help achieve the greatest business impact.

This IBM® Redbooks® publication addresses the new IBM z Systems™ single frame, the IBM z13s server. IBM z Systems servers are the trusted enterprise platform for integrating data, transactions, and insight. A data-centric infrastructure must always be available with a 99.999% or better availability, have flawless data integrity, and be secured from misuse. It needs to be an integrated infrastructure that can support new applications. It also needs to have integrated capabilities that can provide new mobile capabilities with real-time analytics delivered by a secure cloud infrastructure.

IBM z13s servers are designed with improved scalability, performance, security, resiliency, availability, and virtualization. The superscalar design allows z13s servers to deliver a record level of capacity over the prior single frame z Systems server. In its maximum configuration, the z13s server is powered by up to 20 client characterizable microprocessors (cores) running at 4.3 GHz. This configuration can run more than 18,000 millions of instructions per second (MIPS) and up to 4 TB of client memory. The IBM z13s Model N20 is estimated to provide up to 100% more total system capacity than the IBM zEnterprise® BC12 Model H13.

This book provides information about the IBM z13s server and its functions, features, and associated software support. Greater detail is offered in areas relevant to technical planning. It is intended for systems engineers, consultants, planners, and anyone who wants to understand the IBM z Systems™ functions and plan for their usage. It is not intended as an introduction to mainframes. Readers are expected to be generally familiar with existing IBM z Systems technology and terminology.

Authors

This book was produced by a team working at the International Technical Support Organization, Poughkeepsie Center.

Octavian Lascu is a Senior IT Consultant for IBM Romania with over 25 years of IT experience. He specializes in designing, implementing, and supporting complex IT infrastructure environments (cloud, systems, storage, and networking), including high availability and disaster recovery solutions and high-performance computing deployments. He has developed materials for and taught over 50 workshops for technical audiences around the world. He has written several IBM Redbook and IBM Redpaper™ publications.

Barbara Sannerud is a Worldwide Technical Enablement Manager for the IBM z Systems platform. She has 30 years of experience in services, strategy, marketing, and product management. Before her current role, she was Offering Manager for IBM z/OS®, and also held competitive marketing roles. She holds math and MBA degrees and joined IBM from the software and professional services industries where she specialized in performance, systems management, and security.

Cecilia A. De Leon is a Certified IT Specialist in the Philippines. She has 15 years of experience in the z Systems field. She has worked at IBM for 7 years. She holds a degree in Computer Engineering from Mapua Institute of Technology. Her areas of expertise include z Systems servers and operating system. In her current role as Client Technical Specialist, she supports mainframe clients and IBM sales representatives on technical sales engagements. She has also worked as a systems programmer for large banks in the Philippines.

Edzard Hoogerbrug is a System Support Representative in The Netherlands. During the past 26 years, he has worked in various roles within IBM, mainly with mainframes. He has 14 years experience working for EMEA L2 support for Z series after a 2.5 years assignment in Montpellier France. He holds a degree in electrotechnology. His areas of expertise include failure analysis on for z Systems hardware.

Ewerson Palacio is an IBM Distinguished Engineer and a Certified Consulting IT Specialist for Large Systems in Brazil. He has more than 40 years of experience in IBM large systems. Ewerson holds a Computer Science degree from Sao Paulo University. His areas of expertise include z Systems client technical support, mainframe architecture, infrastructure implementation, and design. He is an ITSO z Systems hardware official speaker who has presented technical ITSO seminars, workshops, and private sessions to IBM clients, IBM IT Architects, IBM IT Specialists, and IBM Business Partners around the globe. He has also been a z Systems Hardware Top Gun training designer, developer, and instructor for the last generations of the IBM high-end servers. Ewerson leads the Mainframe Specialty Services Area (MF-SSA), which is part of GTS Delivery, Technology and Engineering (DT&E). He is a member of the IBM Academy of Technology.

Franco Pinto is a Client Technical Specialist in IBM Switzerland. He has 20 years of experience in the mainframe and z/OS fields. His areas of expertise include z Systems technical pre-sales covering mainframe sizing and installation planning, and providing support on existing and new z Systems functions.

Jin J. Yang is a Senior System Service Representative in China. He joined IBM in 1999 to support z Systems products maintenance for clients in China. He has been working in the Technical Support Group to provide second-level support for z Systems clients as a country Top Gun since 2009. His areas of expertise include z Systems hardware, channel connectivity, IBM z/VM®, and Linux on z Systems.

John P. Troy is a z Systems and Storage hardware National Top Gun in the northeast area of the United States. He has 35 years of experience in the service field. His areas of expertise include z Systems server and high-end storage systems technical and customer support. John has been a z Systems hardware technical support course designer, developer, and instructor for the last six generations of IBM high-end servers.

Martin Soellig is a Consulting IT Specialist in Germany. He has 26 years of experience working in the z Systems field. He holds a degree in mathematics from University of Hamburg. His areas of expertise include z/OS and z Systems hardware, specifically in IBM Parallel Sysplex® and GDPS® environments.

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

William G. White
International Technical Support Organization, Poughkeepsie Center

Robert Haimowitz
Development Support Team Poughkeepsie

Patty Driever
Diana Henderson
Anuja Deedwaniya
Lisa Schloemer
Christine Smith
Martin Ziskind
Romney White
Jerry Stevens
Anthony Sofia
IBM Poughkeepsie

Leslie Geer III
IBM Endicott

Monika Zimmerman
Carl Mayer
Angel Nunez Mencias
IBM Germany

Parwez Hamid
IBM Poughkeepsie

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 ITSO 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, International Technical Support Organization
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>



Introducing IBM z13s servers

Digital business has been driving the transformation of IT to be more efficient, secure, adaptive, and intelligent. Today's technology foundation must be capable of increasing the value of data, drawing on the wealth of applications on the z Systems platform, and delivering new opportunities made possible by the API economy. z Systems is designed to harvest information from huge volumes of data to unlock real-time insights, and enable smarter decision making to deliver the greatest business advantage. The hybrid cloud infrastructure requires a secured and resilient platform that is capable of in-transaction analytics, and capable of providing customers with the agility and insight they need to succeed in today's marketplace.

The IT infrastructure must always be on to serve customers in a 24 x 7 global business, have flawless data integrity, and be capable of thwarting cyber security threats. It needs to be resilient to support the introduction of new workloads and handle spikes in processing while meeting the most demanding service levels. It must support both traditional delivery models and new "as a service" cloud delivery models. It also must support the needs of traditional and mobile knowledge workers.

This chapter introduces the basic concepts of IBM z13s servers.

This chapter includes the following sections:

- ▶ Overview of IBM z13s servers
- ▶ z13s servers highlights
- ▶ z13s technical overview
- ▶ Hardware Management Consoles and Support Elements
- ▶ IBM z BladeCenter Extension (zBX) Model 004
- ▶ IBM z Unified Resource Manager
- ▶ IBM Dynamic Partition Manager
- ▶ Operating systems and software

1.1 Overview of IBM z13s servers

The IBM z13s server, like its predecessors, is designed from the chip level up for intense data serving and transaction processing, the core of business. IBM z13s servers are designed to provide unmatched support for data processing by providing these features:

- ▶ A strong, powerful I/O infrastructure
- ▶ Larger cache sizes on the chip to bring data close to processing power
- ▶ Enhanced cryptographic support and compression capabilities of the co-processors and I/O features
- ▶ Simultaneous multithreading to drive throughput
- ▶ The 99.999%¹ application availability design of the IBM z Systems clustering technologies

z13s servers feature a newly designed eight-core² processor chip introduced and shared with the IBM z13™. This chip features innovations designed to deliver maximum performance, and up to 4 TB of available memory, representing an eight-fold increase in memory over the prior processor generation. More memory can help improve response time, minimize application constraints, speed up batch processing, and reduce processor cycles that are consumed for a specific workload. z13s servers also offer a 2.75x boost in the system I/O bandwidth, which when combined with new key I/O enhancements can reduce transfer time for the global enterprise. z13s servers also have more throughput per core (1.3 times the performance of a zBC12 based on workload and model), and more processing cores (50% over zBC12) to drive higher volumes of workloads.

Terminology: The remainder of the book uses the designation *CPC* to refer to the *central processor complex*.

The IBM operating systems participate in providing new advantages for the digital era. For instance, z/OS V2.2 running on z13s servers sets the groundwork for digital business by providing support for demanding workloads such as operational analytics and cloud along with traditional mission-critical applications. z/OS V2.2 continues to support the IBM z Integrated Information Processor (zIIP)³ that can take advantage of the simultaneous multithreading (SMT) feature implemented in the IBM z Systems processor unit (PU) for greater throughput. Applications running under z/OS V2.2 can use vector processing and single-instruction, multiple-data (SIMD) for new analytics applications, Java processing and so on.

Operating system enhancements drive simplification and enables cost-saving consolidation opportunities. With enhancements to extend the reach of your skills, z/OS V2.2 and the entitled z/OS Management Facility can help extend the reach of administrators and other personnel, enabling them to handle configuration tasks with ease. Mobile Workload Pricing for z/OS can help reduce the cost of growth for mobile transactions that are processed by programs such as IBM CICS®, IBM IMS™, and IBM DB2® for z/OS.

IBM z/VM 6.3 has been enhanced to use the SMT and SIMD features offered on the new processor chip⁴, and also supports 64 threads⁵ for Linux workloads. With support for sharing

¹ A z/OS Parallel Sysplex is designed to provide 99.999% application availability. See: <http://www.ibm.com/systems/z/advantages/resiliency/datadriven/application.html>

² Only 6 or 7 active cores per PU Single Chip Module (SCM) in z13s servers.

³ zAAP workloads are now run on zIIP.

⁴ z/VM 6.3 SMT support is provided on Integrated Facility for Linux (IFL) processors only.

⁵ z/VM supports 20 processors on z13s servers and up to 40 threads when SMT enabled.

Open Systems Adapters (OSAs) across z/VM systems, z/VM 6.3 delivers enhanced availability and reduced cost of ownership in network environments.

IBM z13s brings a new approach for Enterprise-grade Linux with offerings and capabilities for availability, virtualization with z/VM, and a focus on open standards and architecture. The new support of kernel-based virtual machine (KVM) on the mainframe provides a new modern, virtualization platform to accelerate the introduction of Linux applications and new open source-based technologies.

KVM: In addition to continued investment in z/VM, IBM has made available a KVM for z Systems that can host Linux guest virtual machines. KVM for IBM z Systems can coexist with z/VM virtualization environments, z/OS, IBM z/VSE®, and z/TPF. This modern, open-source-based hypervisor enables enterprises to capitalize on virtualization capabilities by using common Linux administration skills, while enjoying the robustness of z Systems' scalability, performance, security, and resilience. KVM for IBM z Systems is optimized for z Systems architecture and provides standard Linux and KVM interfaces for operational control of the environment. In addition, it integrates with standard OpenStack virtualization management tools, enabling enterprises to easily integrate Linux servers into their existing traditional infrastructure and cloud offerings.

z13s servers continue to provide heterogeneous platform investment protection with the updated IBM z BladeCenter Extension (zBX) Model 004 and IBM z Unified Resource Manager (zManager). Enhancements to the zBX include the uncoupling of the zBX from the server and installing a Support Element (SE) into the zBX. The zBX Model 002 and Model 003 can be upgraded to the zBX Model 004.

1.2 z13s servers highlights

This section reviews some of the most important features and functions of z13s servers:

- ▶ Processor and memory
- ▶ Capacity and performance
- ▶ I/O subsystem and I/O features
- ▶ Virtualization
- ▶ Increased flexibility with z/VM mode logical partition
- ▶ IBM zAware logical partition
- ▶ 10GbE RoCE Express
- ▶ Flash Express
- ▶ Reliability, availability, and serviceability design
- ▶ IBM z Appliance Container Infrastructure (zACI)
- ▶ Dynamic Partition Manager (DPM)

1.2.1 Processor and memory

IBM continues its technology leadership at a consumable entry point with the z13s servers. z13s servers are built using the IBM modular multi-drawer design that supports 1 or 2 CPC drawers per CPC. Each CPC drawer contains eight-core SCMs with either 6 or 7 cores enabled for use. The SCMs host the redesigned complementary metal-oxide semiconductor (CMOS) 14S0⁶ processor units, storage control chips, and connectors for I/O. The superscalar processor has enhanced out-of-order (OOO) instruction execution, redesigned caches, and an expanded instruction set that includes a Transactional Execution facility. It

⁶ CMOS 14S0 is a 22-nanometer CMOS logic fabrication process.

also includes innovative SMT capability, and provides 139 SIMD vector instruction subset for better performance.

Depending on the model, z13s servers can support from 64 GB to a maximum of 4 TB of usable memory, with up to 2 TB of usable memory per CPC drawer. In addition, a fixed amount of 40 GB is reserved for the hardware system area (HSA) and is not part of customer-purchased memory. Memory is implemented as a redundant array of independent memory (RAIM) and uses extra physical memory as spare memory. The RAIM function uses 20% of the physical installed memory in each CPC drawer.

1.2.2 Capacity and performance

The z13s server provides increased processing and I/O capacity over its predecessor, the zBC12 system. This capacity is achieved both by increasing the performance of the individual processor units, which increases the number of PUs per system, redesigning the system cache and increasing the amount of memory. The increased performance and the total system capacity available, with possible energy savings, allow consolidating diverse applications on a single platform, with significant financial savings. The introduction of new technologies and instruction set ensure that the z13s server is a high performance, reliable, and secure platform. z13s servers are designed to maximize resource exploitation and utilization, and allows you to integrate and consolidate applications and data across your enterprise IT infrastructure.

z13s servers come in two models: The N10 (10 configurable PUs) and N20 (20 configurable PUs). An N20 can have one or two CPC drawers. Each drawer has two nodes (N10 only one). Each node has 13 PUs available for characterization. The second drawer for an N20 is added when more than 2 TB of memory or more I/O features are needed.

IBM z13s Model Capacity Identifier A01 is estimated to provide 60% more total system capacity than the zBC12 Model capacity Identifier A01, with the same amount of memory and power requirements. With up to 4 TB of main storage and SMT, the performance of the z13s processors provide considerable improvement. Uniprocessor performance has also increased significantly. A z13s Model z01 offers, on average, performance improvements of 34% over the zBC12 Model z01. However, variations on the observed performance increasingly depend on the workload type.

The IFL and zIIP processor units on z13s servers can run two simultaneous threads per clock cycle in a single processor, increasing the capacity of the IFL and zIIP processors up to 1.2 times and 1.25 times over the zBC12. However, the observed performance increase varies depending on the workload type.

z13s servers offer 26 capacity levels for six processors that can be characterized as CP. This configuration gives a total of 156 distinct capacity settings and provides a range of 80 - 7123 MIPS. z13s servers deliver scalability and granularity to meet the needs of medium-sized enterprises, while also satisfying the requirements of large enterprises that have demanding, mission-critical transaction and data processing requirements.

This comparison is based on the Large System Performance Reference (LSPR) mixed workload analysis. For a description about performance and workload variation on z13s servers, see Chapter 12, "Performance" on page 437. For more information about LSPR, see:

<https://www.ibm.com/servers/resourceLink/lib03060.nsf/pages/lsprindex>

z13s servers continue to offer all the specialty engines that are available on previous z Systems except for IBM System z® Application Assist Processors (zAAPs). A zAAP qualified

workload runs on a zIIP processor, thus reducing the complexity of the IBM z/Architecture®. zAAPs are no longer supported beginning with the z13 and z13s servers.

Workload variability

Consult the LSPR when considering performance on z13s servers. Individual logical partitions (LPARs) have more performance variability of when an increased number of partitions and more PUs are available. For more information, see Chapter 12, “Performance” on page 437.

For detailed performance information, see the LSPR website at:

<https://www.ibm.com/servers/resourceLink/lib03060.nsf/pages/lsprindex>

The millions of service units (MSUs) ratings are available from the following website:

<http://www.ibm.com/systems/z/resources/swprice/reference/exhibits/>

Capacity on demand (CoD)

CoD enhancements enable clients to have more flexibility in managing and administering their temporary capacity requirements. z13s servers support the same architectural approach for CoD offerings as the zBC12 (temporary or permanent). Within z13s servers, one or more flexible configuration definitions can be available to solve multiple temporary situations and multiple capacity configurations can be active simultaneously.

The customer can stage records to prepare for many scenarios. Up to eight of these records can be installed on the server at any time. After the records are installed, the activation of the records can be done manually, or the z/OS Capacity Provisioning Manager can automatically start the activation when Workload Manager (WLM) policy thresholds are reached. Tokens are available that can be purchased for On/Off CoD either before or after workload execution (pre- or post-paid).

LPAR group absolute capping

IBM PR/SM™ and the Hardware Management tool have been enhanced to limit the amount of physical processor capacity that is consumed by a group of LPARs when a processor unit is defined as a general-purpose processor or as an IFL shared across a set of LPARs. Currently, a user can define the LPAR group capacity limits that allow one or more groups of LPARs to each have their own capacity limit. This new feature adds the ability to define an absolute capping value for an entire LPAR group. This group physical capacity limit is enforced as an absolute limit, so it is not affected by changes to the logical or physical configuration of the system.

1.2.3 I/O subsystem and I/O features

The z13s servers support both PCIe and InfiniBand I/O infrastructure. PCIe Gen3 features are installed in PCIe I/O drawers. Up to two PCIe I/O drawers per z13s server are supported, providing space for up to 64 PCIe features. When upgrading a zBC12 or a z114 to a z13s server, one I/O drawer is also supported as carry forward. I/O drawers were introduced with the IBM z10™ BC.

The z13s Model N20 single CPC drawer can have up to eight PCIe and four InfiniBand (IFB) fanouts, a Model N20 with two CPC drawers has up to 16 PCIe and eight IFB fanouts, whereas the N10 can have up to four PCIe and two IFB fanouts used for the I/O infrastructure and coupling.

For I/O constraint relief, the IBM z13s server has improved scalability with support for three logical channel subsystems (LCSSs), three subchannel sets (to support more devices per logical channel subsystem), and up to 32K devices per IBM FICON® channel up from 24K channels in the previous generation.

For improved device connectivity for parallel access volumes (PAVs), Peer-to-Peer Remote Copy (PPRC) secondary devices, and IBM FlashCopy® devices, the third subchannel set allows extending the amount of addressable external storage. In addition to performing an IPL from subchannel set 0, z13s servers allow you to also perform an IPL from subchannel set 1 (SS1), or subchannel set 2 (SS2).

The system I/O buses take advantage of the PCIe technology and the InfiniBand technology, which are also used in coupling links.

z13s connectivity supports the following I/O or special purpose features:

- ▶ Storage connectivity:
 - Fibre Channel connection (FICON):
 - FICON Express16S 10 KM long wavelength (LX) and short wavelength (SX)
 - FICON Express8S 10 KM long wavelength (LX) and short wavelength (SX)
 - FICON Express8 10 KM LX and SX (carry forward only)
- ▶ Networking connectivity:
 - Open Systems Adapter (OSA):
 - OSA-Express5S 10 GbE LR and SR
 - OSA-Express5S GbE LX and SX
 - OSA-Express5S 1000BASE-T Ethernet
 - OSA-Express4S 10 GbE LR and SR (carry forward only)
 - OSA-Express4S GbE LX and SX (carry forward only)
 - OSA-Express4S 1000BASE-T Ethernet (carry forward only)
 - IBM HiperSockets™
 - Shared Memory Communications - Direct Memory Access (SMC-D) over Internal Shared Memory (ISM)
 - 10 GbE Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE) using Shared Memory Communications over RDMA (SMC-R)
- ▶ Coupling and Server Time Protocol (STP) connectivity:
 - Integrated Coupling Adapter (ICA SR)
 - Parallel Sysplex InfiniBand coupling links (IFB)
 - Internal Coupling links (IC)

In addition, z13s servers support the following special function features, which are installed on the PCIe I/O drawers:

- ▶ Crypto Express5S
- ▶ Flash Express
- ▶ zEnterprise Data Compression (zEDC) Express

Flash Express

Flash Express is an innovative optional feature that was first introduced with the zEC12 and has been enhanced with more cache memory to benefit availability during memory dump

processing for z13s and z13 GA2 servers. It is intended to provide performance improvements and better availability for critical business workloads that cannot afford any impact to service levels. Flash Express is easy to configure, and provides rapid time to value.

Flash Express implements storage-class memory (SCM) through an internal NAND Flash solid-state drive (SSD), in a PCIe card form factor. The Flash Express feature is designed to allow each LPAR to be configured with its own SCM address space.

Flash Express is used in these situations:

- ▶ By z/OS V1R13 (or later), for handling z/OS paging activity such as start of day processing.
- ▶ Coupling Facility Control Code (CFCC) Level 20 or later (CFCC Level 21 on z13s servers), to use Flash Express as an overflow device for shared queue data. This configuration provides emergency capacity to handle IBM WebSphere® MQ shared queue buildups during abnormal situations, such as when “putters” are putting to the shared queue, but “getters” are transiently not retrieving data from the shared queue.
- ▶ Linux for z Systems (Red Hat Enterprise Linux (RHEL) and SUSE Enterprise Linux (SLES)), for use as temporary storage.
- ▶ Stand alone memory dumps and Supervisor Call (SAN Volume Controller) dumps.
- ▶ Read/write cache, which greatly increases performance by allowing customer data to be stored temporarily in a cache in the system's HSA RAM.

For more information, see Appendix H, “Flash Express” on page 529.

10GbE RoCE Express

The 10 Gigabit Ethernet (10GbE) RoCE Express is an optional feature that uses RDMA over Converged Ethernet, and is designed to provide fast memory-to-memory communications between two z Systems CPCs. It is transparent to applications.

Use of the 10GbE RoCE Express feature helps reduce CPU consumption for applications that use the TCP/IP stack such as IBM WebSphere Application Server accessing a DB2 database on z/OS). It can also help reduce network latency with memory-to-memory transfers using SMC-R in z/OS V2R1 and later.

The 10GbE RoCE Express feature on z13s servers can now be shared among up to 31 LPARs running z/OS, and uses both ports on the feature. z/OS V2.1 with PTF or z/OS V2.2 supports the new sharing capability of the RoCE Express features on z13s processors. Also, the z/OS Communications Server has been enhanced to support automatic selection between TCP/IP and RoCE transport layer protocols based on traffic characteristics. The 10GbE RoCE Express feature is supported on z13, z13s, zEC12, and zBC12 servers, and is installed in the PCIe I/O drawer. A maximum of 16 features can be installed. On zEC12 and zBC12, only one port can be used, whereas on the z13 and z13s servers, both RoCE Express ports can be used.

Shared Memory Communications - Direct Memory Access

In addition to supporting SMC-R, z13 servers support a new feature called Shared Memory Communications - Direct Memory Access (SMC-D). Unlike SMC-R, SMC-D does not depend on the RoCE Express feature. SMC-R affords low latency communications within a CEC by using an RDMA connection.

IBM z Systems servers (z13 and z13s) now support a new ISM virtual PCIe (vPCIe) device to enable optimized cross-LPAR TCP communications that use a new “sockets-based DMA”, the SMC-D. SMC-D maintains the socket-API transparency aspect of SMC-R so that applications

that use TCP/IP communications can benefit immediately without requiring any application software or IP topology changes.

With its lightweight design, SMC-D is designed to improve throughput, latency, and CPU consumption over other alternatives without sacrificing quality of service. SMC-D extends the benefits of SMC-R to the same CPC operating system instances without requiring physical resources such as RoCE adapters, PCI bandwidth, ports, I/O slots, network resources, 10 GbE switches, and so on. SMC-D requires either OSA connections or HiperSockets to establish the initial TCP connection, and can coexist with them.

SMC-D uses a virtual PCIe adapter and is configured like a physical PCIe device. There are up to 32 ISM adapters, each with a unique Physical Network ID per CPC.

Notes:

- ▶ SMC-D does not support Coupling facilities or access to the IEDN network.
- ▶ Shared Memory Communication protocols (SMC-R and SMC-D) do not currently support multiple IP subnets.

zEDC Express

The growth of data that needs to be captured, transferred, and stored for large periods of time is unrelenting. The large amounts of data that needs to be handled require ever-increasing bandwidth and storage space. Software-implemented compression algorithms are costly in terms of processor resources, and storage costs are not negligible.

Beginning with zEC12, bandwidth and storage space requirements are addressed by providing hardware-based acceleration for data compression and decompression. zEDC, an optional feature that is available for z13s, provides data compression with lower CPU consumption than previously existing compression technology on z Systems. zEDC compression provides these advantages, among others:

- ▶ QSAM/BSAM for better disk utilization and batch elapsed time improvements
- ▶ SMF for increased availability and online storage reduction
- ▶ DFSMSdss for better disk and tape utilization for backup data
- ▶ Java for high throughput standard compression through `java.util.zip`
- ▶ Encryption Facility for z/OS for better industry data exchange
- ▶ IBM Sterling Connect: Direct for z/OS for better throughput and link utilization
- ▶ ISV support for increased client value
- ▶ DFSMSHsm for improved throughput and MIPS reduction

For more information, see Appendix J, “IBM zEnterprise Data Compression Express” on page 551.

1.2.4 Virtualization

The IBM Processor Resource/Systems Manager™ (PR/SM) is Licensed Internal Code (LIC) that manages and virtualizes all the installed and enabled system resources as a single large symmetric multiprocessor (SMP) system. This virtualization enables full sharing of the installed resources with high security and efficiency. It does so by configuring up to 40 LPARs, each of which has logical processors, memory, and I/O resources, and are assigned from the installed CPC drawers and features. For more information about PR/SM functions, see 3.7, “Logical partitioning” on page 116.

LPAR configurations can be dynamically adjusted to optimize the virtual servers’ workloads. On z13s servers, PR/SM supports an option to limit the amount of physical processor

capacity that is consumed by an individual LPAR or by a group of LPARs. This limit is still valid when a PU defined as a CP or an IFL is shared across a set of LPARs. This feature is designed to provide and enforce a physical capacity limit as an absolute (versus a relative) limit. Physical capacity limit enforcement is not affected by changes to the logical or physical configuration of the system. This physical capacity limit can be specified to a fine level of granularity, to hundredths of units of a processor.

z13s servers provide improvements to the PR/SM HiperDispatch function. *HiperDispatch* provides work alignment to logical processors, and alignment of logical processors to physical processors. This alignment optimizes cache utilization, minimizes inter-CPC drawer communication, and optimizes z/OS work dispatching, with the result of increasing throughput. For more information, see “HiperDispatch” on page 87

z13s servers support the definition of up to 32 IBM HiperSockets. *HiperSockets* provide for memory to memory communication across LPARs without the need for any I/O adapters, and have virtual LAN (VLAN) capability. HiperSockets have been extended to bridge to the intraensemble data network (IEDN).

Increased flexibility with z/VM mode logical partition

z13s servers provide for the definition of a z/VM mode LPAR containing a mix of processor types. These types include CPs and specialty processors, such as IFLs, zIIPs, and ICFs. z/VM V6R2 and later support this capability, which increases flexibility and simplifies system management. In a single LPAR, z/VM can perform the following tasks:

- ▶ Manage guests that use Linux on z Systems on IFLs or CPs, and manage IBM z/VSE, z/TPF, and z/OS guests on CPs.
- ▶ Run designated z/OS workloads, such as parts of IBM DB2 Distributed Relational Database Architecture™ (DRDA®) processing and XML, on zIIPs.

Coupling Facility mode logical partition

Parallel Sysplex is the clustering technology used with z13s servers. To use this technology, a special LIC is used. This code is called CFCC. To activate the CFCC, a special logical partition must be defined. Only PUs characterized as CPs or Internal Coupling Facilities (ICFs) can be used for Coupling Facility (CF) partitions. For a production CF workload, use dedicated ICFs.

IBM z Appliance Container Infrastructure

z Appliance Container Infrastructure (zACI) is a new partition type that, along with an appliance installer, enables the secure deployment of software appliances. Typically, appliances can be implemented as firmware or software, depending on the environment where the appliance runs. To support the execution of software applications, base infrastructure is needed. This partition with its infrastructure is the z Appliance Container Infrastructure. zACI is designed to shorten the deployment and implementation of building and deploying appliances. zACI will be delivered as part of the base code on z13s and z13 (Driver 27) servers.

zACI provides a standardized framework for deploying products as software or firmware. An appliance is an integration of operating system, middleware, and software components that work autonomously and provide core services and infrastructure that focus on consumability and security.

zACI reduces the work that is needed to create and maintain a product, and enforces common functions that appliances need. The zACI framework provides a consistent set of utilities to implement these common functions such as first failure data capture (FFDC),

network setup, and appliance configuration. The design of zACI allows a simpler product (function) deployment model.

Several exploiters are planned for delivery through zACI. For instance, z/VSE Network Appliance provides network access for TCP/IP socket applications running on z/VSE in an LPAR. The z/VSE Network Appliance is an example of a product that is intended to be managed by using the zACI infrastructure. IBM zAware is also designed to be implemented by using the zACI partition, which replaces the zAware partition infrastructure used previously. For more information about zACI, see Appendix A, “IBM z Appliance Container Infrastructure” on page 449.

IBM z/VSE Network Appliance

The z/VSE Network Appliance builds on the z/VSE Linux Fast Path (LFP) function and provides TCP/IP network access without requiring a TCP/IP stack in z/VSE. The appliance uses zACI, which was introduced on z13 and z13s servers. Compared to a TCP/IP stack in z/VSE, this configuration can support higher TCP/IP traffic throughput while reducing the processing resource consumption in z/VSE. The z/VSE Network Appliance is an extension of the IBM z/VSE - z/VM IP Assist (VIA®) function introduced on z114 and z196 servers. VIA provides network access for TCP/IP socket applications running on z/VSE as a z/VM guest. With the new z/VSE Network Appliance, this configuration is available for z/VSE systems running in an LPAR. When available, the z/VSE Network Appliance is provided as a downloadable package. It can then be deployed with the appliance installer.

In summary, the VIA function is available for z/VSE systems running as z/VM guests. The z/VSE Network Appliance is available for z/VSE systems running without z/VM in LPARs.

Both provide network access for TCP/IP socket applications that use the Linux Fast Path. However, no TCP/IP stack is required on the z/VSE system, and no Linux on z Systems needs to be installed.

IBM zAware

IBM zAware is a feature that was introduced with the zEC12 that provides the next generation of system monitoring. IBM zAware is designed to offer a near real-time, continuous-learning diagnostic and monitoring capability. This function helps pinpoint and resolve potential problems quickly enough to minimize their effects on your business.

The ability to tolerate service disruptions is diminishing. In a continuously available environment, any disruption can have grave consequences. This negative effect is especially true when the disruption lasts days or even hours. But increased system complexity makes it more probable that errors occur, and those errors are also increasingly complex. Some incidents' early symptoms go undetected for long periods of time and can become large problems. Systems often experience “soft failures” (sick but not dead), which are much more difficult and unusual to detect. IBM zAware is designed to help in those circumstances. For more information, see Appendix B, “IBM z Systems Advanced Workload Analysis Reporter (IBM zAware)” on page 453.

IBM zAware also now offers support for Linux running on z Systems and supports Linux message log analysis. IBM zAware enables processing of message streams including those without message IDs. It provides increased flexibility for analysis with the ability to group multiple systems for modeling and analysis purposes together, which is especially helpful for Linux workloads.

For use across the sysplex, IBM zAware now features an aggregated Sysplex view for z/OS and system views. Visualization and usability are enhanced with an enhanced Heat Map display, enhanced filtering and visualization capabilities, and improvements in time zone display.

Beginning with z13s servers, IBM zAware runs in a *zACI* mode logical partition. Either CPs or IFLs can be configured to the IBM zAware partition. This special partition is defined for the exclusive use of the IBM z Systems Advanced Workload Analysis Reporter (IBM zAware) offering. IBM zAware requires a special license, IBM z Advanced Workload Analysis Reporter (zAware)

For customers now considering IBM zAware, z13 GA2 introduces zACI, which supports IBM zAware. IBM zAware functions are the same whether it uses zACI or runs as a stand-alone zAware partition. Existing zAware instances are automatically converted to use the zACI partition. For new zACI instances that will use IBM zAware, use the zACI web interface to select IBM zAware.

1.2.5 Reliability, availability, and serviceability design

System reliability, availability, and serviceability (RAS) is an area of continuous IBM focus. The RAS objective is to reduce, or eliminate if possible, all sources of planned and unplanned outages, while providing adequate service information in case something happens. Adequate service information is required for determining the cause of an issue without the need to reproduce the context of an event. With a properly configured z13s server, further reduction of outages can be attained through improved, nondisruptive replace, repair, and upgrade functions for memory, drawers, and I/O adapters. In addition, z13s servers have extended nondisruptive capability to download and install Licensed Internal Code (LIC) updates.

Enhancements include removing pre-planning requirements with the fixed 40 GB HSA. Client-purchased memory is *not* used for traditional I/O configurations, and you no longer need to reserve capacity to avoid disruption when adding new features. With a fixed amount of 40 GB for the HSA, maximums are configured and an IPL is performed so that later insertion can be dynamic, which eliminates the need for a power-on reset of the server.

IBM z13s RAS features provide many high-availability and nondisruptive operational capabilities that differentiate the z Systems in the marketplace.

The ability to cluster multiple systems in a Parallel Sysplex takes the commercial strengths of the z/OS platform to higher levels of system management, scalable growth, and continuous availability.

1.3 z13s technical overview

This section briefly reviews the major elements of z13s servers:

- ▶ Models
- ▶ Model upgrade paths
- ▶ Frame
- ▶ CPC drawer
- ▶ I/O Connectivity
- ▶ I/O subsystem
- ▶ Parallel Sysplex Coupling and Server Time Protocol connectivity

- ▶ Special-purpose features:
 - Cryptography
 - Flash Express
 - zEDC Express
- ▶ Reliability, availability, and serviceability

1.3.1 Models

z13s servers have a machine type of 2965. Two models are offered: N10 and N20. The model name indicates the maximum number of PUs available for purchase. A PU is the generic term for the IBM z/Architecture processor unit (processor core) on the SCM.

On z13s servers, some PUs are part of the system base. That is, they are not part of the PUs that can be purchased by clients. They are characterized by default as follows:

- ▶ System assist processor (SAP) that is used by the channel subsystem. There are three standard SAPs for the N20, and two for the N10.
- ▶ One integrated firmware processor (IFP), which is used in support of select features, such as zEDC and 10GbE RoCE.
- ▶ Two spare PUs for the N20 only that can transparently assume any characterization during a permanent failure of another PU.

The PUs that clients can purchase can assume any of the following characterizations:

- ▶ Central processor (CP) for general-purpose use.
- ▶ Integrated Facility for Linux (IFL) for the use of Linux on z Systems.
- ▶ z Systems Integrated Information Processor (zIIP⁷). One CP must be installed with or before the installation of any zIIPs.

zIIPs: At least one CP must be purchased with, or before, a zIIP can be purchased. Clients can purchase up to two zIIPs for each purchased CP (assigned or unassigned) on the system (2:1 ratio). However, for migrations from zBC12 with zAAPs, the ratio (CP:zIIP) can go up to 4:1.

- ▶ ICF is used by the CFCC.
- ▶ Extra SAPs are used by the channel subsystem.

A PU that is not characterized cannot be used, but is available as a spare. The following rules apply:

- ▶ At least one CP, ICF, or IFL must be purchased and activated for either the N10 or N20 model.
- ▶ PUs can be purchased in single PU increments and are orderable by feature code.
- ▶ The total number of PUs purchased cannot exceed the total number that are available for that model.
- ▶ The number of installed zIIPs cannot exceed twice the number of installed CPs.

⁷ zAAPs are not available on z13s servers. The zAAP workload is run on zIIPs.

- ▶ The z13s CPC drawer system design provides the capability to increase the capacity of the system. You can add capacity by activating more CPs, IFLs, ICFs, or zIIPs on an existing CPC drawer. The second CPC drawer can be used to provide more memory, or one or more adapters to support a greater number of I/O features.
- ▶ In z13s servers, the addition of a CPC drawer or memory is disruptive

1.3.2 Model upgrade paths

A z13s Model N10 can be upgraded to a z13s Model N20 hardware model. The upgrades to N20 are disruptive (that is, the system is not available during the upgrade). Any zBC12 or z114 model can be upgraded to a z13s N10 or N20 model, which is also disruptive. For more information see also 2.7.1, “Upgrades” on page 71.

Consideration: Note that the z13s servers are air cooled only.

z114 upgrade to z13s server

When a z114 is upgraded to a z13s server, the z114 Driver level must be at least 93. If a zBX is involved, the Driver 93 must be at bundle 27 or later. Family to family (z114 or zBC12) upgrades are frame rolls, but all z13s upgrade paths are disruptive.

zBC12 upgrade to z13s server

When a BC12 is upgraded to a z13s server, the zBC12 must be at least at Driver level 15. If a zBX is involved, Driver 15 must be at Bundle 22 or later.

The following processes are not supported:

- ▶ Reverting to an earlier level within the z13 or z13s models
- ▶ Upgrades from IBM System z10® or earlier systems
- ▶ Attachment of a zBX Model 002 or model 003 to a z13s server

zBX upgrade

The zBX Model 004 is available as an upgrade from an existing zBX Model 002 or Model 003. The upgrade decouples the zBX from its controlling CPC. With the addition of redundant SEs, it becomes a stand-alone node within an ensemble.

1.3.3 Frame

The z13s server has one frame that contains the following CPC components:

- ▶ Up to two CPC drawers, up to two PCIe I/O drawers, and up to one I/O drawer that hold I/O features and special purpose features
- ▶ Power supplies
- ▶ An optional internal battery feature (IBF)
- ▶ Air cooling
- ▶ Two System Control Hubs (SCHs) to interconnect the CPC components through Ethernet.
- ▶ Two new 1U, rack-mounted SEs, each with one keyboard, pointing device, and display mounted on a tray.

1.3.4 CPC drawer

Up to two CPC drawers (minimum one) can be installed in the z13s frame. Each CPC drawer houses the SCMs, memory, and fanouts. The CPC drawer supports up to 8 PCIe fanouts and 4 IFB fanouts for I/O and coupling connectivity.

In the N20 two CPC drawer model, CPC drawers are connected through cables that are also field-replaceable units (FRUs) and can be replaced concurrently, unlike the disruptive bus repair on the zBC12.

Important: Concurrent drawer repair, concurrent drawer add, and concurrent memory add are not available on z13s servers.

SCM technology

z13s servers are built on the superscalar microprocessor architecture of its predecessor, and provides several enhancements over the zBC12. Each CPC drawer is physically divided into two nodes. Each node has three SCMs: Two PU SCMs, and one storage control (SC) SCM.

A fully configured CPC drawer has four PU SCMs and two SC SCMs. The PU SCM has eight cores (six or seven active), which can be characterized as CPs, IFLs, ICFs, zIIPs, SAPs, or IFPs. Two CPC drawers sizes are offered: 10 cores (N10) and 20 cores (N20).

On the N20, the PU configuration includes two designated spare PUs. The N10 has no dedicated spares. Two standard SAPs are installed with the N10 Model and up to three SAPs for the N20 Model. In addition, one PU is used as an IFP and is not available for client use. The remaining PUs can be characterized as CPs, IFL processors, zIIPs, ICF processors, or extra SAPs. For more information, see 3.3, “CPC drawer design” on page 84 and 3.4, “Processor unit design” on page 88.

Processor features

The processor chip runs at 4.3 GHz. Depending on the model, either 13 PUs (N10) or 26 PUs (N20) are available. Each core on the PU chip includes an enhanced dedicated coprocessor for data compression and cryptographic functions, which are known as the Central Processor Assist for Cryptographic Function (CPACF)⁸.

Hardware data compression can play a significant role in improving performance and saving costs over performing compression in software (thus consuming CPU cycles). In addition, the zEDC Express feature offers more performance and savings. It is designed to provide compression capabilities that compliment those provided by the data compression on the coprocessor.

The micro-architecture of the core has been enhanced to increase parallelism and improve pipeline efficiency. The core has a new branch prediction and instruction fetch front end to support simultaneous multithreading in a single core and to improve the branch prediction throughput, a wider instruction decode (six instructions per cycle), and 10 arithmetic logical execution units that offer double instruction bandwidth over the zBC12.

Each core has two hardware decimal floating point units that are designed according to a standardized, open algorithm. Two on-core hardware decimal floating point units meet the requirements of today’s business and user applications, and provide greater floating point execution throughput with improved performance and precision.

⁸ No charge feature code (FC) 3863 must be ordered to enable CPACF

In the unlikely case of a permanent core failure, each core can be individually replaced by one of the available spares. Core sparing is transparent to the operating system and applications. The N20 has two spares, but the N10 does not offer dedicated spares (the firmware can determine cores available for sparing).

Simultaneous multithreading

The micro-architecture of the core of z13s servers allows simultaneous execution of two threads (SMT) in the same zIIP or IFL core, dynamically sharing processor resources such as execution units and caches. This feature allows a more efficient utilization of the core and increased capacity. While one of the threads is waiting for a storage access (cache miss), the other thread that is running simultaneously in the core can use the shared resources rather than remain idle.

z/OS and z/VM control programs use SMT on z13s servers to optimize their workloads while providing repeatable metrics for capacity planning and chargeback.

Single instruction multiple data instruction set

The z13s instruction set architecture includes a subset of 139 new instructions for SIMD execution, which was added to improve efficiency of complex mathematical models and vector processing. These new instructions allow a larger number of operands to be processed with a single instruction. The SIMD instructions use the superscalar core to process operands in parallel, which enables more processor throughput.

Transactional Execution facility

The z13s server, like its predecessor zBC12, has a set of instructions that allows defining groups of instructions that are run atomically, that is, either all the results are committed or none are. The facility provides for faster and more scalable multi-threaded execution, and is known as *hardware transactional memory*.

Out-of-order execution

As with its predecessor zBC12, a z13s server has an enhanced superscalar microprocessor with OOO execution to achieve faster throughput. With OOO, instructions might not run in the original program order, although results are presented in the original order. For example, OOO allows a few instructions to complete while another instruction is waiting. Up to six instructions can be decoded per system cycle, and up to 10 instructions can be in execution.

Concurrent processor unit conversions

z13s servers support concurrent conversion between various PU types, which provides the flexibility to meet the requirements of changing business environments. CPs, IFLs, zIIPs, ICFs, and optional SAPs can be converted to CPs, IFLs, zIIPs, ICFs, and optional SAPs.

Memory subsystem and topology

z13s servers use a new buffered dual inline memory module (DIMM) technology. For this purpose, IBM has developed a chip that controls communication with the PU, and drives address and control from DIMM to DIMM. z13s servers use the new DIMM technology, and carry forward is not supported. The memory subsystem supports 20 DIMMs per drawer and the DIMM capacities are 16 GB, 32 GB, 64 GB, and 128 GB.

Memory topology provides the following benefits:

- ▶ A RAIM for protection at the dynamic random access memory (DRAM), DIMM, and memory channel levels
- ▶ A maximum of 4 TB of user configurable memory with a maximum of 5.1 TB of physical memory (with a maximum of 4 TB configurable to a single LPAR)

- ▶ One memory port for each PU chip.
- ▶ Increased bandwidth between memory and I/O
- ▶ Asymmetrical memory size and DRAM technology across CPC drawers
- ▶ Large memory pages (1 MB and 2 GB)
- ▶ Key storage
- ▶ Storage protection key array that is kept in physical memory
- ▶ Storage protection (memory) key that is also kept in every L2 and L3 cache directory entry
- ▶ A larger (40 GB) fixed-size HSA that eliminates having to plan for HSA

PCIe fanout hot-plug

The *PCIe fanout* provides the path for data between memory and the PCIe features through the PCIe 16 GBps bus and cables. The PCIe fanout is hot-pluggable. During an outage, a redundant I/O interconnect allows a PCIe fanout to be concurrently repaired without loss of access to its associated I/O domains. Up to four PCIe fanouts per drawer are available on the N10 and up to eight PCIe fanouts per drawer on the N20. The PCIe fanout can also be used for the ICA SR. If redundancy in coupling link connectivity is ensured, the PCIe fanout can be concurrently repaired.

Host channel adapter fanout hot-plug

The HCA fanout provides the path for data between memory and the I/O cards in an I/O drawer through 6 GBps IFB cables. The HCA fanout is hot-pluggable. During an outage, an HCA fanout can be concurrently repaired without the loss of access to its associated I/O features by using redundant I/O interconnect to the I/O drawer. Up to four HCA fanouts are available per CPC drawer.

1.3.5 I/O connectivity: PCIe and InfiniBand

z13s servers offer various improved I/O features and uses technologies such as PCIe and InfiniBand. This section briefly reviews the most relevant I/O capabilities.

z13s servers take advantage of PCIe Generation 3 to implement the following features:

- ▶ PCIe Generation 3 (Gen3) fanouts that provide 16 GBps connections to the PCIe I/O features in the PCIe I/O drawers.
- ▶ PCIe Gen3 fanouts that provide 8 GBps coupling link connections through the new IBM ICA SR. Up to eight ICA SR fanouts and up to 16 ICA SR ports are supported on an IBM z13s server, enabling improved connectivity for short distance coupling.

z13s servers take advantage of InfiniBand to implement the following features:

- ▶ A 6 GBps I/O bus that includes the InfiniBand infrastructure (HCA2-C) for the I/O drawer for non-PCIe I/O features.
- ▶ Parallel Sysplex coupling links using IFB include 12x InfiniBand coupling links (HCA3-O) for local connections and 1x InfiniBand coupling links (HCA3-O LR) for extended distance connections between any two zEnterprise CPCs. The 12x IFB link (HCA3-O) has a bandwidth of 6 GBps and the HCA3-O LR 1X InfiniBand links have a bandwidth of 5 Gbps.

1.3.6 I/O subsystem

The z13s I/O subsystem is similar to z13 servers and includes a new PCIe Gen3 infrastructure. The I/O subsystem is supported by both a PCIe bus and an I/O bus similar to

that of zEC12. This infrastructure is designed to reduce processor usage and latency, and provide increased throughput.

z13s servers offer two I/O infrastructure elements for driving the I/O features: PCIe I/O drawers for PCIe features, and up to one I/O drawer for non-PCIe features carried forward only.

PCIe I/O drawer

The *PCIe I/O drawer*, together with the PCIe features, offers finer granularity and capacity over previous I/O infrastructures. It can be concurrently added and removed in the field, easing planning. Only PCIe cards (features) are supported, in any combination. Up to two PCIe I/O drawers can be installed on a z13s server with up to 32 PCIe I/O features per drawer (64 in total).

I/O drawer

On z13s servers, a maximum of one I/O drawer is supported only when carried forward on upgrades from zBC12 or z114 to z13s servers. For a z13s new order, the I/O drawer is not available.

Native PCIe and Integrated Firmware Processor

Native PCIe was introduced with the zEDC and RoCE Express features, which are managed differently from the traditional PCIe features. The device drivers for these adapters are available in the operating system.

The diagnostic tests for the adapter layer functions of the native PCIe features are managed by LIC that is designated as a resource group partition, which runs on the IFP. For availability, two resource groups are present and share the IFP.

During the ordering process of the native PCIe features, features of the same type are evenly spread across the two resource groups (RG1 and RG2) for availability and serviceability reasons. Resource groups are automatically activated when these features are present in the CPC.

I/O and special purpose features

z13s servers support the following PCIe features on a new build system, which can be installed only in the PCIe I/O drawers:

- ▶ FICON Express16S Short Wave (SX) and Long Wave (LX)
- ▶ FICON Express8S Short Wave (SX) and Long Wave (LX)
- ▶ OSA-Express5S 10 GbE Long Reach (LR) and Short Reach (SR)
- ▶ OSA Express5S GbE LX and SX
- ▶ OSA Express5S 1000BASE-T
- ▶ 10GbE RoCE Express
- ▶ Crypto Express5S
- ▶ Flash Express (FC 0403)
- ▶ zEDC Express

When carried forward on an upgrade, z13s servers also support the following features in the PCIe I/O drawers:

- ▶ FICON Express8S SX and LX
- ▶ OSA-Express5S (all)
- ▶ OSA-Express4S (all)
- ▶ 10GbE RoCE Express
- ▶ Flash Express (FC 0402)
- ▶ zEDC Express

Also, when carried forward on an upgrade, the z13s servers support one I/O drawer on which the FICON Express8 SX and LX (10 km) feature can be installed.

In addition, InfiniBand coupling links HCA3-O and HCA3-O LR, which attach directly to the CPC drawers, are supported.

Tip: FICON Express8 and 8S cards should only be ordered or carried forward to support attachments to 2 Gbps devices and for older I/O optics not supported for 16 Gbps capable features.

FICON channels

Up to 64 features of any combination of FICON Express16s or FICON Express8S are supported in PCIe I/O Drawer. The FICON Express8S features support link data rates of 2, 4, or 8 Gbps, and the FICON Express16S support 4, 8, or 16 Gbps.

Up to eight features with up to 32 FICON Express8 channels (carry forward) are supported in one I/O drawer (also carry forward). The FICON Express8 features support link data rates of 2, 4, or 8 Gbps.

The z13s FICON features support the following protocols:

- ▶ FICON (FC) and High Performance FICON for z Systems (zHPF). zHPF offers improved performance for data access, which is of special importance to OLTP applications.
- ▶ FICON channel-to-channel (CTC).
- ▶ Fibre Channel Protocol (FCP).

FICON also offers the following capabilities:

- ▶ Modified Indirect Data Address Word (MIDAW) facility that provides more capacity over native FICON channels for programs that process data sets that use striping and compression, such as DB2, VSAM, partitioned data set extended (PDSE), hierarchical file system (HFS), and z/OS file system (zFS). It does so by reducing channel, director, and control unit processor usage.
- ▶ Enhanced problem determination, analysis, and manageability of the storage area network (SAN) by providing registration information to the fabric name server for both FICON and FCP.

Read Diagnostic Parameter

A new command called Read Diagnostic Parameter (RDP) allows z Systems to obtain extra diagnostics from the Small Form Factor Pluggable (SFP) optics located throughout the SAN fabric. RDP is designed to help improve the accuracy of identifying a failed/failing component in the SAN fabric.

Open Systems Adapter

z13s servers allow any mix of the supported Open Systems Adapter (OSA) Ethernet features. Up to 48 OSA-Express5S or OSA-Express4S features, with a maximum of 96 ports, are supported. OSA-Express5S and OSA-Express4S features are plugged into the PCIe I/O drawer.

The maximum number of combined OSA-Express5S and OSA-Express4S features cannot exceed 48.

The N10 supports 64 OSA ports and the N20 supports 96 OSA ports.

OSM and OSX channel path identifier types

The z13s provides OSA-Express5S, OSA-Express4S, and channel-path identifier (CHPID) types OSA-Express for Unified Resource Manager (OSM) and OSA-Express for zBX (OSX) connections:

- ▶ OSA-Express for Unified Resource Manager (OSM)
Connectivity to the intranode management network (INMN) Top of Rack (ToR) switch in the zBX is not supported on z13s servers. When the zBX model 002 or 003 is upgraded to a model 004, it becomes an independent node that can be configured to work with the ensemble. The zBX model 004 is equipped with two 1U rack-mounted SEs to manage and control itself, and is independent of the CPC SEs.
- ▶ OSA-Express for zBX (OSX)
Connectivity to the IEDN provides a data connection from the z13s to the zBX. OSX can use OSA-Express5S 10 GbE (preferred) and the OSA-Express4S 10 GbE feature.

OSA-Express5S, OSA-Express4S feature highlights

z13s servers support five different types each of OSA-Express5S and OSA-Express4S features. OSA-Express5S features are a technology refresh of the OSA-Express4S features:

- ▶ OSA-Express5S 10 GbE Long Reach (LR)
- ▶ OSA-Express5S 10 GbE Short Reach (SR)
- ▶ OSA-Express5S GbE Long Wave (LX)
- ▶ OSA-Express5S GbE Short Wave (SX)
- ▶ OSA-Express5S 1000BASE-T Ethernet
- ▶ OSA-Express4S 10 GbE Long Reach
- ▶ OSA-Express4S 10 GbE Short Reach
- ▶ OSA-Express4S GbE Long Wave
- ▶ OSA-Express4S GbE Short Wave
- ▶ OSA-Express4S 1000BASE-T Ethernet

OSA-Express features provide important benefits for TCP/IP traffic, which are reduced latency and improved throughput for standard and jumbo frames. Performance enhancements are the result of the data router function being present in all OSA-Express features.

For functions that were previously performed in firmware, the OSA Express5S and OSA-Express4S now perform those functions in hardware. Additional logic in the IBM application-specific integrated circuit (ASIC) that is included with the feature handles packet construction, inspection, and routing. This process allows packets to flow between host memory and the LAN at line speed without firmware intervention.

With the data router, the *store and forward* technique in direct memory access (DMA) is no longer used. The data router enables a direct host memory-to-LAN flow. This configuration avoids a *hop*, and is designed to reduce latency and increase throughput for standard frames (1492 bytes) and jumbo frames (8992 bytes).

HiperSockets

The HiperSockets function is also known as *internal queued direct input/output (internal QDIO or iQDIO)*. It is an integrated function of the z13s servers that provides users with attachments to up to 32 high-speed virtual LANs with minimal system and network processor usage.

HiperSockets can be customized to accommodate varying traffic sizes. Because the HiperSockets function does not use an external network, it can free system and network resources, eliminating equipment costs while improving availability and performance.

For communications between LPARs in the same z13s server, HiperSockets eliminates the need to use I/O subsystem features and to traverse an external network. Connection to HiperSockets offers significant value in server consolidation by connecting many virtual servers. It can be used instead of certain coupling link configurations in a Parallel Sysplex.

HiperSockets is extended to allow integration with IEDN, which extends the reach of the HiperSockets network outside the CPC to the entire ensemble, and displays it as a single Layer 2 network.

10GbE RoCE Express

The 10 Gigabit Ethernet (10GbE) RoCE Express feature is a RDMA-capable network interface card. The 10GbE RoCE Express feature is supported on z13, z13s, zEC12, and zBC12 servers, and is used in the PCIe I/O drawer. Each feature has one PCIe adapter. A maximum of 16 features can be installed.

The 10GbE RoCE Express feature uses a short reach (SR) laser as the optical transceiver, and supports the use of a multimode fiber optic cable that terminates with an LC Duplex connector. Both a point-to-point connection and a switched connection with an enterprise-class 10 GbE switch are supported.

Support is provided by z/OS, which supports one port per feature, dedicated to one partition in zEC12 and zBC12. With z13s servers, both ports, shared by up to 31 partitions, are supported.

For more information, see Appendix D, “Shared Memory Communications” on page 475.

Shared Memory Communications - Direct Memory Access

In addition to supporting SMC-R, z13s support is a new protocol called SMC-D over ISM. Unlike SMC-R, SMC-D has no dependency on the RoCE Express feature. SMC-D provides low latency communications within a CPC by using shared memory.

SMC-R, by using RDMA, can vastly improve the transaction, throughput, and CPU consumption rates for unchanged TCP/IP applications. Direct memory access (DMA) can provide comparable benefits on an internal LAN.

z Systems now supports a new ISM virtual PCIe (vPCIe) device to enable optimized cross-LPAR TCP communications using SMC-D. SMC-D maintains the socket-API transparency aspect of SMC-R so that applications that use TCP/IP communications can benefit immediately without requiring any application software or IP topology changes. For more information, see D.3, “Shared Memory Communications - Direct Memory Access” on page 489.

Note: SMC-D does not support Coupling facilities or access to the IEDN network. In addition, SMC protocols do not currently support multiple IP subnets.

1.3.7 Parallel Sysplex Coupling and Server Time Protocol connectivity

Support for Parallel Sysplex includes the Coupling Facility Control Code and coupling links.

Coupling links support

The z13s CPC drawer supports up to 8 PCIe Gen3 fanouts and up to four IFB fanouts for Parallel Sysplex coupling links. A z13s Model N20 with optional second CPC drawer supports a total of 16 PCIe Gen3 fanouts and 8 IFB fanout slots (four per CPC drawer). Coupling

connectivity in support of Parallel Sysplex environments is provided on z13s servers by the following features:

- ▶ Integrated Coupling Adapter Short Reach - ICA SR (PCIe Coupling Links): The ICA SR PCIe fanout has two 8 GBps ports that use short reach optics (up to 150m between CPCs). The ICA SR can only be used for coupling connectivity between IBM z13 or IBM z13s servers, and the ICA SR can only connect to another ICA SR. Generally, order ICA SR on the IBM z13s processors used in a Parallel Sysplex to help ensure coupling connectivity with future processor generations
- ▶ HCA3-O: 12x InfiniBand coupling links offering up to 6 GBps of bandwidth between z13, zBC12, z196, and z114 systems, for a distance of up to 150 m (492 feet). These links have improved service times over the HCA2-O links that were used on prior z Systems families.
- ▶ HCA3-O LR: 1x InfiniBand (up to 5 Gbps connection bandwidth) between z13, zEC12, zBC12, z196, and z114 systems for a distance of up to 10 km (6.2 miles). The HCA3-O LR (1xIFB) type has twice the number of links per fanout card as compared to type HCA2-O LR (1xIFB) that was used in the previous z Systems generations.
- ▶ Internal Coupling Channels (ICs): Operate at memory speed.

Attention: IBM z13s servers do not support ISC-3 connectivity.

CFCC Level 21

CFCC level 21 is delivered on z13s servers with driver level 27. CFCC Level 21 introduces the following enhancements:

- ▶ Support for up to 20 ICF processors per z Systems CPC:
 - The maximum number of logical processors in a Coupling Facility Partition remains 16
- ▶ Large memory support:
 - Improves availability/scalability for larger CF cache structures and data sharing performance with larger DB2 group buffer pools (GBPs).
 - This support removes inhibitors to using large CF structures, enabling the use of Large Memory to scale to larger DB2 local buffer pools (LBPs) and GBPs in data sharing environments.
 - The CF structure size remains at a maximum of 1 TB.
- ▶ See Table 1-1 for the maximum number of Coupling Links on N20 models.

Table 1-1 Coupling Links on N20

Coupling Links	Features	Ports	Speed
HCA3-O LR	8	32	5 Gbps
HCA-O 12X	8	16	6 GBps
ICA SR	8	16	8 GBps

z13s systems with CFCC Level 21 require z/OS V1R13 or later, and z/VM V6R2 or later for virtual guest coupling.

To support an upgrade from one CFCC level to the next, different levels of CFCC can coexist in the same sysplex while the coupling facility LPARs are running on different servers. CF LPARs that run on the same server share the CFCC level.

A CF running on a z13s server (CFCC level 21) can coexist in a sysplex with CFCC levels 17 and 19. Review the CF LPAR size by using the CFSizer tool:

<http://www.ibm.com/systems/z/cfsizer>

Server Time Protocol facility

If you require time synchronization across multiple IBM z Systems servers, you can implement Server Time Protocol (STP). STP is a server-wide facility that is implemented in the LIC of z Systems CPCs (including CPCs running as stand-alone coupling facilities). STP presents a single view of time to PR/SM and provides the capability for multiple servers to maintain time synchronization with each other.

Any z Systems CPC can be enabled for STP by installing the STP feature. Each server that must be configured in a Coordinated Timing Network (CTN) must be STP-enabled.

The STP feature is the supported method for maintaining time synchronization between z Systems images and coupling facilities. The STP design uses the CTN concept, which is a collection of servers and coupling facilities that are time-synchronized to a time value called Coordinated Server Time (CST).

Network Time Protocol (NTP) client support is available to the STP code on the z13, z13s, zEC12, zBC12, z196, and z114 servers. With this function, the z13, z13s, zEC12, zBC12, z196, and z114 servers can be configured to use an NTP server as an external time source (ETS).

This implementation answers the need for a single time source across the heterogeneous platforms in the enterprise. An NTP server becomes the single time source for the z13, z13s, zEC12, zBC12, z196, and z114 servers, as well as other servers that have NTP clients, such as UNIX and Microsoft Windows systems.

The time accuracy of an STP-only CTN is improved by using, as the ETS device, an NTP server with the pulse per second (PPS) output signal. This type of ETS is available from various vendors that offer network timing solutions.

Improved security can be obtained by providing NTP server support on the HMC for the SE. The HMC is normally attached to the private dedicated LAN for z Systems maintenance and support. For z13s and zBC12 server, authentication support is added to the HMC NTP communication with NTP time servers. In addition, TLS/SSL with Certificate Authentication is added to the HMC/SE support to provide a secure method for connecting clients to z Systems.

Attention: A z13s server cannot be connected to a Sysplex Timer and cannot join a Mixed CTN.

If a current configuration consists of a Mixed CTN, the configuration must be changed to an STP-only CTN before z13s integration. z13s servers can coexist only with z Systems CPCs that do not have the External Time Reference (ETR) port capability.

Support has been added to enable STP communications to occur by using the ICA SR (new for z13s servers).

Enhanced Console Assisted Recovery

Console Assisted Recovery (CAR) support is designed to help a Backup Time Server (BTS) determine whether the Primary Time Server is still up and running if Coupling traffic ceases. The CAR process is initiated by the BTS if there is no communication between the Primary and Backup Time Servers. The BTS queries the state of the Primary Time Server (PTS)/ Central Time Server (CTS) SE by using the SE and HMC of the BTS. If the PTS is down, the BTS initiates takeover.

With the new Enhanced Console Assisted Recovery (ECAR), the process of BTS takeover is faster. When the PTS encounters a check-stop condition, the CEC informs the SE and HMC of the condition. The PTS SE recognizes the pending check-stop condition, and an ECAR request is sent directly from the HMC to the BTS SE to start the takeover. The new ECAR support is faster than the original support because there is almost no delay between the system check-stop and the start of CAR processing. ECAR is only available on z13 GA2 and z13s servers. In a mixed environment with previous generation machines, you should define a z13 or z13s server as the PTS and CTS.

1.3.8 Special-purpose features

This section overviews several features that, although installed in the PCIe I/O drawer or in the I/O drawer, provide specialized functions without performing I/O operations. No data is moved between the CPC and externally attached devices.

Cryptography

Integrated cryptographic features provide industry leading cryptographic performance and functions. The cryptographic solution that is implemented in z Systems has received the highest standardized security certification (FIPS 140-2 Level 4⁹). In addition to the integrated cryptographic features, the cryptographic features (Crypto Express5S, the only crypto-card that is supported on z13s servers) allows adding or moving crypto-coprocessors to LPARs without pre-planning.

z13s servers implement PKCS#11, one of the industry-accepted standards that are called Public Key Cryptographic Standards (PKCS), which are provided by RSA Laboratories of RSA, the security division of EMC Corporation. It also implements the IBM Common Cryptographic Architecture (CCA) in its cryptographic features.

CP Assist for Cryptographic Function

The CP Assist for Cryptographic Function (CPACF) offers the full complement of the Advanced Encryption Standard (AES) algorithm and Secure Hash Algorithm (SHA) with the Data Encryption Standard (DES) algorithm. Support for CPACF is available through a group of instructions that are known as the Message-Security Assist (MSA). z/OS Integrated Cryptographic Service Facility (ICSF) callable services, and the *z90crypt* device driver running on Linux on z Systems also start CPACF functions. ICSF is a base element of z/OS. It uses the available cryptographic functions, CPACF, or PCIe cryptographic features to balance the workload and help address the bandwidth requirements of your applications.

CPACF must be explicitly enabled by using a no-charge enablement feature (FC 3863), except for the SHAs, which are included enabled with each server.

The enhancements to CPACF are exclusive to the z Systems servers, and are supported by z/OS, z/VM, z/VSE, z/TPF, and Linux on z Systems.

⁹ Federal Information Processing Standard (FIPS) 140-2 Security Requirements for Cryptographic Modules

Configurable Crypto Express5S feature

Crypto Express5S represents the newest generation of cryptographic features. It is designed to complement the cryptographic capabilities of the CPACF. It is an optional feature of the z13 and z13s server generation. The Crypto Express5S feature is designed to provide granularity for increased flexibility with one PCIe adapter per feature. For availability reasons, a minimum of two features are required.

With z13 and z13s servers, a cryptographic coprocessor can be shared across more than 16 domains, up to the maximum number of LPARs on the system.(up to 85 domains for z13 servers and 40 domains for z13s servers).

The Crypto Express5S is a state-of-the-art, tamper-sensing, and tamper-responding programmable cryptographic feature that provides a secure cryptographic environment. Each adapter contains a tamper-resistant hardware security module (HSM). The HSM can be configured as a Secure IBM CCA coprocessor, as a Secure IBM Enterprise PKCS #11 (EP11) coprocessor, or as an accelerator:

- ▶ A Secure IBM CCA coprocessor is for secure key encrypted transactions that use CCA callable services (default).
A Secure IBM Enterprise PKCS #11 (EP11) coprocessor implements an industry standardized set of services that adhere to the PKCS #11 specification v2.20 and more recent amendments. This new cryptographic coprocessor mode introduced the PKCS #11 secure key function.
- ▶ An accelerator for public key and private key cryptographic operations is used with Secure Sockets Layer/Transport Layer Security (SSL/TLS) acceleration.

The Crypto Express5S is designed to meet these cryptographic standards, among others:

- FIPS 140-2 Level 4
- ANSI 9.97
- Payment Card Industry (PCI) HSM
- Deutsche Kreditwirtschaft (DK)

FIPS 140-2 certification is supported only when Crypto Express5S is configured as a CCA or an EP11 coprocessor.

Crypto Express5S supports a number of ciphers and standards that include those in this section. For more information about cryptographic algorithms and standards, see Chapter 6, “Cryptography” on page 199.

TKE workstation and support for smart card readers

The TKE feature is an integrated solution that is composed of workstation firmware, hardware, and software to manage cryptographic keys in a secure environment. The TKE is either network-connected or isolated in which case smart cards are used.

The Trusted Key Entry (TKE) workstation and the most recent TKE 8.1 LIC are optional features on the z13s. The TKE 8.1 requires the crypto adapter FC 4767. You can use TKE 8.0 to collect data from previous generations of Cryptographic modules and apply the data to Crypto Express5S coprocessors.

The TKE workstation offers a security-rich solution for basic local and remote key management. It provides to authorized personnel a method for key identification, exchange, separation, update, and backup, and a secure hardware-based key loading mechanism for

operational and master keys. TKE also provides secure management of host cryptographic module and host capabilities.

Support for an optional smart card reader that is attached to the TKE workstation allows the use of smart cards that contain an embedded microprocessor and associated memory for data storage. Access to and the use of confidential data on the smart cards are protected by a user-defined personal identification number (PIN). A FIPS-certified smart card, part number 00JA710, is now included in the smart card reader and additional smart cards optional features.

When Crypto Express5S is configured as a Secure IBM Enterprise PKCS #11 (EP11) coprocessor, the TKE workstation is required to manage the Crypto Express5S feature. The TKE is recommended for CCA mode processing as well. If the smart card reader feature is installed in the TKE workstation, the new smart card part 00JA710 is required for EP11 mode. If EP11 is to be defined, smart cards that are used must have FIPS certification.

For more information about the Cryptographic features, see Chapter 6, “Cryptography” on page 199. Also, see the Web Deliverables download site for the most current ICSF updates available (currently HCR77B0 Web Deliverable 14 and HCR77B1 Web Deliverable 15):

<http://www.ibm.com/systems/z/os/zos/tools/downloads/>

Flash Express

The *Flash Express* optional feature is intended to provide performance improvements and better availability for critical business workloads that cannot afford any impact to service levels. Flash Express is easy to configure, and provides rapid time to value.

Flash Express implements SCM in a PCIe card form factor. Each Flash Express card implements an internal NAND Flash SSD, and has a capacity of 1.4 TB of usable storage. Cards are installed in pairs, which provide mirrored data to ensure a high level of availability and redundancy. A maximum of four pairs of cards (four features) can be installed on a z13s server, for a maximum capacity of 5.6 TB of storage.

The Flash Express feature, recently enhanced, is designed to allow each LPAR to be configured with its own SCM address space. It is used for paging and enables the use of pageable 1 MB pages.

Encryption is included to improve data security. Data security is ensured through a unique key that is stored on the SE hard disk drive (HDD). It is mirrored for redundancy. Data on the Flash Express feature is protected with this key, and is usable only on the system with the key that encrypted it. The Secure Keystore is implemented by using a smart card that is installed in the SE. The smart card (one pair, one for each SE) contains the following items:

- ▶ A unique key that is personalized for each system
- ▶ A small cryptographic engine that can run a limited set of security functions within the smart card

Flash Express is supported by z/OS V1R13 (or later) for handling z/OS paging activity, and has support for 1 MB pageable pages and SAN Volume Controller memory dumps. Support was added to the CFCC to use Flash Express as an overflow device for shared queue data to provide emergency capacity to handle WebSphere MQ shared queue buildups during abnormal situations. Abnormal situations include when “putters” are putting to the shared queue, but “getters” cannot keep up and are not getting from the shared queue.

Flash memory is assigned to a CF image through HMC windows. The coupling facility resource management (CFRM) policy definition allows the correct amount of SCM to be used by a particular structure, on a structure-by-structure basis. Additionally, Linux (RHEL and

SUSE) can use Flash Express for temporary storage. Recent enhancements of Flash Express include 2 GB page support and dynamic reconfiguration for Flash Express.

For more information, see Appendix H, “Flash Express” on page 529.

zEDC Express

zEDC Express, an optional feature that is available to z13, z13s, zEC12, and zBC12 servers, provides hardware-based acceleration for data compression and decompression with lower CPU consumption than the previous compression technology on z Systems.

Use of the zEDC Express feature by the z/OS V2R1 or later zEnterprise Data Compression acceleration capability delivers an integrated solution to help reduce CPU consumption, optimize performance of compression-related tasks, and enable more efficient use of storage resources. It also provides a lower cost of computing and helps to optimize the cross-platform exchange of data.

One to eight features can be installed on the system. There is one PCIe adapter/compression coprocessor per feature, which implements compression as defined by RFC1951 (DEFLATE).

A zEDC Express feature can be shared by up to 15 LPARs.

See the IBM System z Batch Network Analyzer 1.4.2 tool, which reports on potential zEDC usage for QSAM/BSAM data sets:

<http://www.ibm.com/support/techdocs/atmastr.nsf/WebIndex/PRS5132>

For more information, see Appendix J, “IBM zEnterprise Data Compression Express” on page 551.

1.3.9 Reliability, availability, and serviceability

The z13 and z13s RAS strategy employs a building-block approach, which is developed to meet the client's stringent requirements for achieving continuous reliable operation. Those building blocks are error prevention, error detection, recovery, problem determination, service structure, change management, measurement, and analysis.

The initial focus is on preventing failures from occurring. This goal is accomplished by using *Hi-Rel* (highest reliability) components that use screening, sorting, burn-in, and run-in, and by taking advantage of technology integration. For LICC and hardware design, failures are reduced through rigorous design rules; design walk-through; peer reviews; element, subsystem, and system simulation; and extensive engineering and manufacturing testing.

The RAS strategy is focused on a recovery design to mask errors and make them transparent to client operations. An extensive hardware recovery design is implemented to detect and correct memory array faults. In cases where transparency cannot be achieved, you can restart the server with the maximum capacity possible.

z13s servers include the following RAS improvements, among others:

- ▶ Cables for SMP fabric
- ▶ CP and SC SCMs are FRUs
- ▶ Point of load (POL) replaces the Voltage Transformation Module, and is a FRU
- ▶ z13s servers have both raised floor and non-raised-floor options
- ▶ The redundant oscillators are isolated on their own backplane
- ▶ The CPC drawer is a FRU (empty)
- ▶ A built-in Time Domain Reflectometry (TDR) isolates failures
- ▶ CPC drawer level degrade

- ▶ FICON (better recovery on fiber)
- ▶ N+1 SCH and power supplies
- ▶ N+1 SEs
- ▶ ASHRAE A3 support

For more information, see Chapter 9, “Reliability, availability, and serviceability” on page 355.

1.4 Hardware Management Consoles and Support Elements

The HMCs and SEs are appliances that together provide platform management for z Systems.

The HMC is a workstation designed to provide a single point of control for managing local or remote hardware elements. In addition to the HMC console tower, a new 1U Rack Mounted HMC with 19-inch rack-mounted components can be placed in customer supplied 19-inch racks.

The HMCs and SEs are appliances that together provide platform management for z Systems. For z13s servers, a new driver level 27 is required. HMC (V2.13.1) plus Microcode Change Levels (MCLs) and the Support Element (V2.13.1) are required to be installed.

HMCs and SEs also provide platform management for zBX Model 004 and for the ensemble nodes when the z Systems CPCs and the zBX Model 004 nodes are members of an ensemble. In an ensemble, the HMC is used to manage, monitor, and operate one or more z Systems CPCs and their associated LPARs, and the zBX Model 004 machines. Also, when the z Systems and a zBX Model 004 are members of an ensemble, the HMC¹⁰ has a global (ensemble) management scope, which is compared to the SEs on the zBX Model 004 and on the CPCs, which have local (node) management responsibility.

When tasks are performed on the HMC, the commands are sent to one or more CPC SEs or zBX Model 004 SEs, which then issue commands to their CPCs and zBXs. To provide high availability, an ensemble configuration requires a pair of HMCs: A primary and an alternate.

For more information, see Chapter 11, “Hardware Management Console and Support Elements” on page 387.

1.5 IBM z BladeCenter Extension (zBX) Model 004

The IBM z BladeCenter Extension (zBX) Model 004 improves infrastructure reliability by extending the mainframe systems management and service across a set of heterogeneous compute elements in an ensemble.

The zBX Model 004 is only available as an optional upgrade from a zBX Model 003 or a zBX Model 002, through MES, in an ensemble that contains at least one z13s CPC and consists of these components:

- ▶ Two internal 1U rack-mounted Support Elements providing zBX monitoring and control functions.
- ▶ Up to four IBM 42U Enterprise racks.
- ▶ Up to eight BladeCenter chassis with up to 14 blades each, with up to two chassis per rack.

¹⁰ From Version 2.11. For more information, see 11.6, “HMC in an ensemble” on page 428.

- ▶ Up to 112¹¹ blades.
- ▶ INMN ToR switches. On the zBX Model 004, the new local zBX Support Elements directly connect to the INMN within the zBX for management purposes. Because zBX Model 004 is an independent node, there is no INMN connection to any z Systems CPC.
- ▶ IEDN ToR switches. The IEDN is used for data paths between the zEnterprise ensemble members and the zBX Model 004, and also for customer data access. The IEDN point-to-point connections use message authentication code (MAC) addresses, not IP addresses (Layer 2 connectivity).
- ▶ 8 Gbps Fibre Channel switch modules for connectivity to customer supplied storage (through SAN).
- ▶ Advanced management modules (AMMs) for monitoring and management functions for all the components in the BladeCenter.
- ▶ Power Distribution Units (PDUs) and cooling fans.
- ▶ Optional acoustic rear door or optional rear door heat exchanger.

The zBX is configured with redundant hardware infrastructure to provide qualities of service similar to those of z Systems, such as the capability for concurrent upgrades and repairs.

Geographically Dispersed Parallel Sysplex Peer-to-Peer Remote Copy (GDPS/PPRC) and GDPS Global Mirror (GDPS/GM) support zBX hardware components, providing workload failover for automated multi-site recovery. These capabilities facilitate the management of planned and unplanned outages.

1.5.1 Blades

Two types of blades can be installed and operated in the IBM z BladeCenter Extension (zBX):

- ▶ Optimizer Blades: IBM WebSphere DataPower® Integration Appliance XI50 for zBX blades
- ▶ IBM Blades:
 - A selected subset of IBM POWER7® blades
 - A selected subset of IBM BladeCenter HX5 blades
- ▶ IBM BladeCenter HX5 blades are virtualized by using an integrated hypervisor.
- ▶ IBM POWER7 blades are virtualized by PowerVM® Enterprise Edition, and the virtual servers run the IBM AIX® operating system.

Enablement for blades is specified with an entitlement feature code that is configured on the ensemble HMC.

1.5.2 IBM WebSphere DataPower Integration Appliance XI50 for zEnterprise

The IBM WebSphere DataPower Integration Appliance XI50 for zEnterprise (DataPower XI50z) is a multifunctional appliance that can help provide multiple levels of XML optimization.

This configuration streamlines and secures valuable service-oriented architecture (SOA) applications. It also provides drop-in integration for heterogeneous environments by enabling core enterprise service bus (ESB) functions, including routing, bridging, transformation, and event handling. It can help to simplify, govern, and enhance the network security for XML and web services.

¹¹ The maximum number of blades varies according to the blade type and blade function.

When the DataPower XI50z is installed in the zBX, the Unified Resource Manager provides integrated management for the appliance. This configuration simplifies control and operations, including change management, energy monitoring, problem detection, problem reporting, and dispatching of an IBM service support representative (IBM SSR), as needed.

Important: The zBX Model 004 uses the blades carried forward in an upgrade from a previous model. Customers can install more entitlements up to the full zBX installed blade capacity if the existing blade centers chassis have available slots. After the entitlements are acquired from IBM, clients must procure and purchase the additional zBX supported blades to be added, up to the full installed entitlement LIC record, from another source or vendor.

1.6 IBM z Unified Resource Manager

The IBM z Unified Resource Manager is the integrated management software that runs on the Ensemble HMC and on the zBX model 004 SEs. The Unified Resource Manager consists of six management areas (for more information, see 11.6.1, “Unified Resource Manager” on page 428):

- ▶ Operational controls (Operations)
Includes extensive operational controls for various management functions.
- ▶ Virtual server lifecycle management (Virtual servers)
Enables directed and dynamic virtual server provisioning across hypervisors from a single point of control.
- ▶ Hypervisor management (Hypervisors)
Enables the management of hypervisors and support for application deployment.
- ▶ Energy management (Energy)
Provides energy monitoring and management capabilities that can be used to better understand the power and cooling demands of the zEnterprise System.
- ▶ Network management (Networks)
Creates and manages virtual networks, including access control, which allows virtual servers to be connected.
- ▶ Workload Awareness and platform performance management (Performance)
Manages CPU resource across virtual servers that are hosted in the same hypervisor instance to achieve workload performance policy objectives.

The Unified Resource Manager provides energy monitoring and management, goal-oriented policy management, increased security, virtual networking, and storage configuration management for the physical and logical resources of an ensemble.

1.7 IBM Dynamic Partition Manager

A new administrative mode for the z Systems CPC is being introduced for Linux only CPCs with FCP attached storage. The IBM DPM provides simplified z Systems hardware and virtual infrastructure management including integrated dynamic I/O management for users that intend to run KVM for IBM z Systems as hypervisor or Linux on z Systems running in LPAR mode.

DPM provides simplified, consumable, and enhanced partition lifecycle and dynamic I/O management capabilities by using the Hardware Management Console and is designed to perform these tasks:

- ▶ Create and provision an environment:
 - Create new partitions
 - Assignment of processors and memory
 - Configuration of I/O adapters (Network, FCP Storage, Crypto, and Accelerators)
- ▶ Manage the environment: Modify system resources without disrupting running workloads
- ▶ Monitor and troubleshoot the environment: Source identification of system failures, conditions, states, or events that can lead to workload degradation.
- ▶ A CPC can be in either the DPM mode or the standard PR/SM mode.
- ▶ DPM mode is enabled before the CPC power-on reset.
- ▶ Operating the CPC in DPM mode requires two OSA-Express 1000BASE-T Ethernet ports (recommended that these be on separate features) for primary and backup SE connectivity.

1.8 Operating systems and software

IBM z13s servers are supported by a large set of software, including independent software vendor (ISV) applications. This section lists only the supported operating systems. Use of various features might require the latest releases. For more information, see Chapter 7, “Software support” on page 229.

1.8.1 Supported operating systems

The following operating systems are supported for z13s servers:

- ▶ z/OS Version 2 Release 2 with program temporary fixes (PTFs)
- ▶ z/OS Version 2 Release 1 with PTFs
- ▶ z/OS Version 1 Release 13 with PTFs
- ▶ z/OS V1R12 with required maintenance (compatibility support only) and extended support agreement
- ▶ z/VM Version 6 Release 4 with PTFs (preview)
- ▶ z/VM Version 6 Release 3 with PTFs
- ▶ z/VM Version 6 Release 2 with PTFs
- ▶ KVM for IBM z Systems Version 1.1.1
- ▶ z/VSE Version 6 Release 1 with PTFs
- ▶ z/VSE Version 5 Release 2 with PTFs
- ▶ z/VSE Version 5 Release 1 with PTFs
- ▶ z/TPF Version 1 Release 1 with PTFs

- ▶ Linux on z Systems distributions:
 - SUSE Linux Enterprise Server (SLES): SLES 12 and SLES 11.
 - Red Hat Enterprise Linux (RHEL): RHEL 7 and RHEL 6.
 - Customers should monitor for new distribution releases supported

For recommended and new Linux on z Systems distribution levels, see the following website:

<http://www.ibm.com/systems/z/os/linux/resources/testedplatforms.html>

The following operating systems will be supported on zBX Model 004:

- ▶ An AIX (on POWER7) blade in IBM BladeCenter Extension Mod 004): AIX 5.3, AIX 6.1, AIX 7.1 and later releases, and PowerVM Enterprise Edition
- ▶ Linux (on IBM BladeCenter HX5 blade installed in zBX Mod 004):
 - Red Hat RHEL 5.5 and later, 6.0 and later, 7.0 and later
 - SLES 10 (SP4) and later, 11 SP1 and later, SLES 12 and later - 64 bit only
- ▶ Microsoft Windows (on IBM BladeCenter HX5 blades installed in zBX Mod 004):
 - Microsoft Windows Server 2012, Microsoft Windows Server 2012 R2
 - Microsoft Windows Server 2008 R2 and Microsoft Windows Server 2008 (SP2) (Datacenter Edition recommended) 64 bit only

Together with support for IBM WebSphere software, IBM z13s servers provide full support for SOA, web services, Java Platform, Enterprise Edition, Linux, and Open Standards. The z13s server is intended to be a platform of choice for the integration of the newest generations of applications with existing applications and data.

z Systems software is also designed to take advantage of the many enhancements on z13 and z13s servers. Several platform enhancements have been announced with the z13:

- ▶ KVM for IBM z Systems 1.1.1
- ▶ z/VM Support
- ▶ z/OS Support

KVM for IBM z Systems 1.1.1

KVM on z offers a number of enhancements announced with z13s servers for availability. Because KVM support is evolving on z Systems, monitor for new enhancements to KVM and Linux on z Systems distributions.

These enhancements are intended to use many of the innovations of the server. KVM on z provides the following enhancements, among others:

- ▶ Simultaneous multithreading (SMT) exploitation
- ▶ Guest use of the Vector Facility for z/Architecture (SIMD)
- ▶ Hypervisor enhancements that include support for iSCSI and NFS
- ▶ Hypervisor Crypto use
- ▶ Enhanced RAS capabilities such as improved first-failure data capture (FFDC)
- ▶ Improved high availability configuration
- ▶ Unattended installation of hypervisor

z/VM Support

z/VM Support for the z13 servers includes but is not limited to these enhancements:

- ▶ z/VM 6.2 and 6.3 provide these enhancements:
 - Guest support for Crypto Express5S, and support for 85 Crypto Express domains.
 - Absolute Capping of an LPAR group, enabling each LPAR to consume capacity up to its individual limit.
- ▶ In addition, z/VM 6.3 will provide support for the following enhancements with service:
 - Guest exploitation support of SMC-D.
 - Dynamic Memory Management, which provides improved efficiency for memory upgrades by using only a portion of the reserved main storage for the partition by initializing and clearing just the amount of storage requested.
 - Guest exploitation of Vector Facility (SIMD).
- ▶ z/VM SMT exploitation for IFL processors in a Linux only mode or for a z/VM mode LPAR. Note that z/VM SMT exploitation support does not virtualize threads for Linux guests, but z/VM itself is designed to achieve higher throughput through SMT.
- ▶ z/VM CPU pools for limiting the CPU resources consumed by a group of virtual machines to a specific capacity.
- ▶ z/VM Multi-VSwitch Link Aggregation allows a port group of OSA-Express features to span multiple virtual switches within a single z/VM or between multiple z/VM systems to increase optimization and utilization of OSA-Express when handling larger traffic loads.

z/OS Support

z/OS uses many of the new functions and features of IBM z13 servers that include but are not limited to the following:

- ▶ z/OS V2.2 supports zIIP processors in SMT mode to help improve throughput for zIIP workloads. This support is also available for z/OS V2.1 with PTFs.
- ▶ z/OS V2.2 supports up to 141 processors per LPAR or up to 128 physical processors per LPAR in SMT mode.
- ▶ z/OS V2.2 also supports up to 4 TB of real memory per LPAR. This support is also available on z/OS V2.1 with PTFs.
- ▶ z/OS V2.2 supports the vector extension facility (SIMD) instructions. This support is also available for z/OS V2.1 with PTFs.
- ▶ z/OS V2.2 supports up to four subchannel sets on IBM z13 and z13s servers to help relieve subchannel constraints, and can allow you to define larger I/O configurations that support multi-target Metro Mirror (PPRC) sets. This support is also available for z/OS V1.13 and z/OS V2.1 with service.
- ▶ z/OS V1.13 and later releases support z13 and z13s FICON function to allow cascading up to 4 FICON switches. Dynamic cache management (DCM) support is provided for cascaded switches on z/OS V2.1 and later.
- ▶ z/OS V2.1 and later with PTFs is designed to use the Read Diagnostic Parameters (RDP) Extended Link Service (ELS) on z13 and z13s processors to retrieve and display information about the status of FICON fiber optic connections, and to provide health checks for diagnosing FICON error conditions that might help with early detection of deterioration of connection quality.
- ▶ z/OS V2.2 running on IBM z13 and z13s servers with IBM System Storage® DS8000® devices and a minimum MCL supports a new health check for FICON dynamic routing that is designed to check all components of a dynamic routing fabric. This support, which is

also available for z/OS V1.13 and z/OS V2.1 with PTFs, can help you identify configuration errors that can result in data integrity exposures.

- ▶ z/OS V2.2, and z/OS V2.1 with PTFs support the new LPAR absolute group capping function.
- ▶ z/OS V2.2 Communications Server supports the virtualization capability of 10GbE RoCE Express on IBM z13 and z13s processors. z/OS V2.2 is designed to support the SMC-D protocol for low-latency, high-bandwidth, cross-LPAR connections for applications.
- ▶ Exploitation of Crypto Express5S features is provided for z/OS V2.2 and with the Enhanced Cryptographic Support for z/OS V1.13 - z/OS V2.1 web deliverable.
- ▶ z/OS V2.2 XL C/C++ supports z13 and z13s processors with ARCH(11) andTUNE(11) parameters that are designed to take advantage of the new instructions.
- ▶ XL C/C++ support for SIMD instructions with the vector programming language extensions, and the IBM MASS and ATLAS libraries. This function is also available for z/OS V2.1 XL C/C++ with a web deliverable available at:
<http://www.ibm.com/systems/z/os/zos/tools/downloads/#webdees>
- ▶ New functions are available for ICSF in a new Cryptographic Support for z/OS V1R13 - z/OS V2R2 This web deliverable is available for download from:
<http://www.ibm.com/systems/z/os/zos/tools/downloads/>
- ▶ More support for the z13 processor family is planned with the ICSF Cryptographic Support for z/OS V1R13 - z/OS V2R2 web deliverable in PTFs in the first quarter of 2016.
- ▶ Support for the new TKE 8.1 workstation.

1.8.2 IBM compilers

The following IBM compilers for z Systems can use z13s servers:

- ▶ Enterprise COBOL for z/OS
- ▶ Enterprise PL/I for z/OS
- ▶ XL C/C++ for Linux on z Systems
- ▶ z/OS XL C/C++

The compilers increase the return on your investment in z Systems hardware by maximizing application performance by using the compilers' advanced optimization technology for z/Architecture. Compilers like COBOL and C, C++ and Java all use SIMD. Through their support of web services, XML, and Java, they allow for the modernization of existing assets into web-based applications. They support the latest IBM middleware products (CICS, DB2, and IMS), allowing applications to use their latest capabilities.

To fully use the capabilities of z13s servers, you must compile by using the minimum level of each compiler as specified in Table 1-2.

Table 1-2 Supported IBM compiler levels

Compiler	Minimum level
Enterprise COBOL for z/OS	V5.2
Enterprise PL/I for z/OS	V4.5
XL C/C++ for Linux on z Systems	V1.1
z/OS XL C/C++	V 2.1 ^a

a. Web update required

To obtain the best performance, you must specify an architecture level of 11 by using the **-qarch=arch11** option for the XL C/C++ for Linux on z Systems compiler or the **ARCH(11)** option for the other compilers. This option grants the compiler permission to use machine instructions that are available only on z13s servers.

Because specifying the architecture level of 11 results in a generated application that uses instructions that are available only on the z13s or z13 servers, the application will not run on earlier versions of the hardware. If the application must run on z13s servers and on older hardware, specify the architecture level corresponding to the oldest hardware on which the application needs to run. For more information, see the documentation for the **ARCH** or **-qarch** options in the guide for the corresponding compiler product.



Central processor complex hardware components

This chapter introduces the IBM z13s hardware components, significant features and functions, and their characteristics and options. The main objective of this chapter is to explain the z13s hardware building blocks, and how these components interconnect from a physical point of view. This information can be useful for planning purposes, and can help to define configurations that fit your requirements.

This chapter provides information about the following topics:

- ▶ Frame and drawers
- ▶ Processor drawer concept
- ▶ Single chip modules
- ▶ Memory
- ▶ Reliability, availability, and serviceability
- ▶ Connectivity
- ▶ Model configurations
- ▶ Power considerations
- ▶ Summary of z13s structure

2.1 Frame and drawers

The z Systems frames are enclosures that are built to Electronic Industries Alliance (EIA) standards. The z13s central processor complex (CPC) has one 42U EIA frame, which is shown in Figure 2-1. The frame has positions for one or two CPC drawers, and a combination of up to two PCIe drawers and up to one I/O drawer.

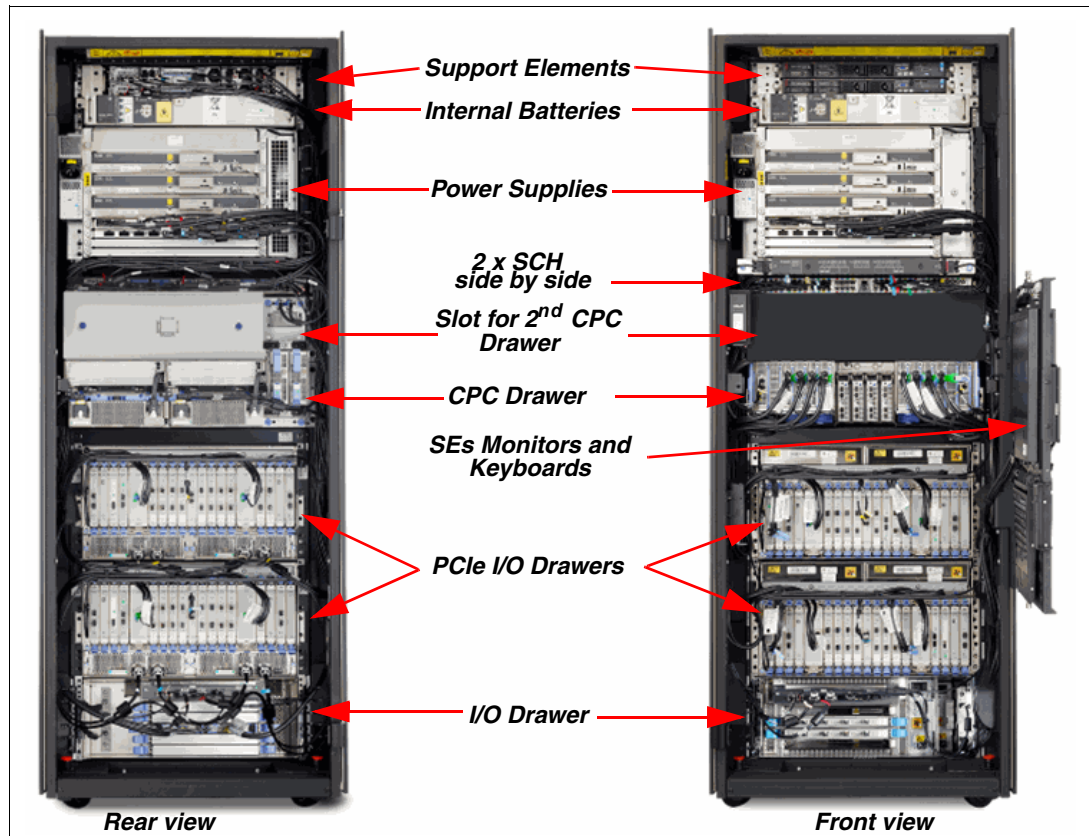


Figure 2-1 z13s frame: Rear and front view

2.1.1 The z13s frame

The frame includes the following elements, which are shown in Figure 2-1 from top to bottom:

- ▶ Two Support Element (SE) servers that are installed at the top of the A frame. In previous z Systems, the SEs were notebooks installed on the swing tray. For z13s servers, the SEs are 1U servers that are mounted at the top of the 42U EIA frame. The external LAN interface connections are now directly connected to the SEs at the rear of the system.
- ▶ Optional Internal Battery Features (IBFs) that provide the function of a local uninterrupted power source.

The IBF further enhances the robustness of the power design, increasing power line disturbance immunity. It provides battery power to preserve processor data in a loss of power on both power feeds from the utility company. The IBF provides battery power to preserve full system function despite the loss of power. It enables continuous operation through intermittent losses, brownouts, and power source switching, or can provide time for an orderly shutdown in a longer outage.

Table 10-2 on page 372 lists the IBF holdup times for various configurations.

- ▶ Two Bulk Power Assemblies (BPAs) that house the power components, with one in the front and the other in the rear of the system. Each BPA is equipped with up to three Bulk Power Regulators (BPRs), a controller, and two distributor cards. The number of BPRs varies depending on the configuration of the z13s servers. For more information see , “The Top Exit I/O Cabling feature adds 15 cm (6 in.) to the width of the frame and about 60 lbs (27.3 kg) to the weight.” on page 371.
- ▶ Two side by side System Control Hubs (SCHs) that are the replacement for the Bulk Power Hubs that were used in previous z Systems and provide the internal communication among the various system elements.
- ▶ Up to two CPC drawers. At least one CPC drawer must be installed, referenced as drawer 1. The additional one can be installed above the first CPC drawer, and is referenced as drawer 0.
 - A Model N10 CPC drawer contains one node, consisting of two processor unit Single Chip Modules (PU SCMs) and one storage controller SCM (SC SCM) and the associated memory slots.
 - The N10 model is always the single CPC drawer.
 - Model N20 CPC drawer contains two nodes, four PU SCMs and two SC SCMs.
 - The N20 model can be a one or two CPC drawers system.
 - Memory dual inline memory modules (DIMMs), point of load regulators, and fanout cards for internal and external connectivity
- ▶ A newly designed swing tray that is in front of the I/O drawer slots, and contains the keyboards and displays that connect to the two SEs.
- ▶ Combinations of PCIe and I/O drawers are shown in Table 2-1. The older I/O drawer can only be carried forward by way of upgrade from previous systems z114 or zBC12. The number of CPC drawers and PCIe and InfiniBand fanouts are also shown. The general rule is that the number of CPC drawers plus the number of PCIe and I/O drawers cannot be greater than four due to power restrictions.

Table 2-1 PCIe drawer and I/O drawer on z13s servers

N10		N20			
1 CPC drawer 4 PCIe fanouts 2 InfiniBand fanouts		1 CPC drawer 8 PCIe fanouts 4 InfiniBand fanouts		2 CPC drawers 16 PCIe fanouts 8 InfiniBand fanouts	
I/O Drawer 0-1	PCIe Drawer 0-1	I/O Drawer 0-1	PCIe Drawer 0-2	I/O Drawer 0-1	PCIe Drawer 0-2
0	0	0	0	0	0
0	1	0	1	0	1
		0	2	0	2
1	0	1	0	1	0
1	1	1	1	1	1
		1	2		

In Figure 2-2, the various I/O drawer configurations are displayed when one CPC drawer is installed for both Model N10 and N20. The view is from the rear of the A frame.

- ▶ PCIe drawers are installed from the top down
- ▶ An I/O drawer (legacy) is only present during a MES carry forward and is always at the bottom of the frame.

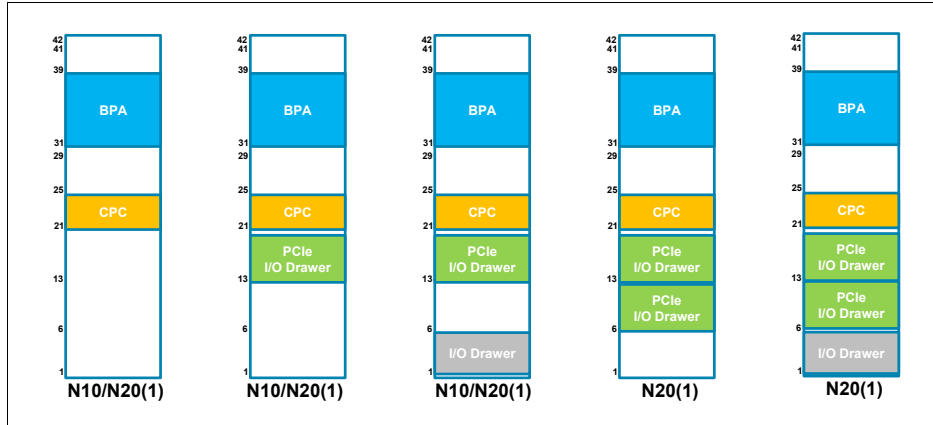


Figure 2-2 z13s (one CPC drawer) I/O drawer configurations

In Figure 2-3, the various I/O drawer configurations are displayed when two CPC drawers are installed for the Model N20 only. The view is from the rear of the A frame.

- ▶ PCIe drawers are installed from the top down
- ▶ I/O drawer (legacy) is only present during a MES carry forward, and is always at the bottom of the frame.

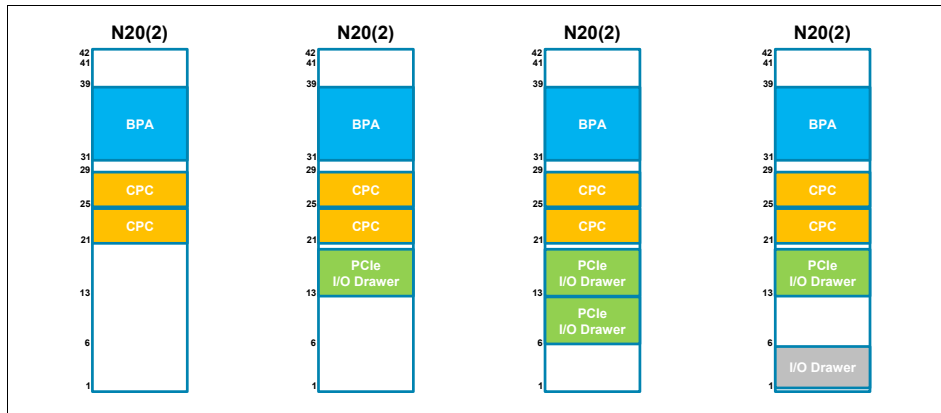


Figure 2-3 z13s (two CPC Drawers) I/O Drawer Configurations

2.1.2 PCIe I/O drawer and I/O drawer features

Each CPC drawer has PCIe Generation3 fanout slots and InfiniBand fanout slots to support two types of I/O infrastructure for data transfer:

- ▶ PCIe I/O infrastructure with a bandwidth of 16 GBps
- ▶ InfiniBand I/O infrastructure with a bandwidth of 6 GBps

PCIe I/O infrastructure

The PCIe I/O infrastructure uses the PCIe fanout to connect to the PCIe I/O drawer, which can contain the following features:

- ▶ FICON Express16S, a two port card (long wavelength (LX) or short wavelength (SX)) and two physical channel IDs (PCHIDs)
- ▶ FICON Express8S (two port card, LX or SX, and two PCHIDs)
- ▶ Open Systems Adapter (OSA)-Express5S features:
 - OSA-Express5S 10 Gb Ethernet (one port card, Long Reach (LR) or Short Reach (SR), and one PCHID)
 - OSA-Express5S Gb Ethernet (two port card, LX or SX, and one PCHID)
 - OSA-Express5S 1000BASE-T Ethernet (two port card, RJ-45, and one PCHID)
- ▶ OSA-Express4S features (only for a carry-forward MES):
 - OSA-Express4S 10 Gb Ethernet (one port card, LR or SR, and one PCHID)
 - OSA-Express4S Gb Ethernet (two port card, LX or SX, and one PCHID)
 - OSA-Express4S 1000BASE-T Ethernet (two port card, RJ-45, and one PCHID)
- ▶ Crypto Express5S feature. Each feature holds one PCI Express cryptographic adapter. Each adapter can be configured during installation as a Secure IBM Common Cryptographic Architecture (CCA) coprocessor, as a Secure IBM Enterprise Public Key Cryptography Standards (PKCS) #11 (EP11) coprocessor, or as an accelerator.
- ▶ Flash Express. Each Flash Express feature occupies two slots in the PCIe I/O Drawer, but does not have a CHPID type. Logical partitions (LPARs) in all channel subsystems (CSSs) have access to the features.
- ▶ 10 GbE Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE) Express. It has a 2-port card, and up to 31 LPARS can share a physical adapter.
- ▶ zEnterprise Data Compression (zEDC) Express. The zEDC Express feature occupies one slot in the PCIe I/O Drawer, and one PCHID. Up to 15 partitions can share the feature concurrently.

InfiniBand I/O infrastructure

InfiniBand I/O infrastructure uses the HCA2-C fanout to connect to I/O drawers. The I/O drawers can contain only these features:

- ▶ FICON Express8 cards, maximum quantity eight. Each card has four ports, LX or SX, and four PCHIDs.

2.2 Processor drawer concept

The z13s CPC drawer contains up to six SCMs, memory, symmetric multiprocessor (SMP) connectivity, and connectors to support:

- ▶ PCIe I/O drawers (through PCIe fanout hubs)
- ▶ I/O drawers through InfiniBand fanout features
- ▶ Coupling links fanouts

z13s servers can have up to two CPC drawers installed (the minimum is one CPC drawer). The contents of the CPC drawer and its components are model dependent. The Model N10 CPC drawer that is shown in Figure 2-4 has a single node, half the structure of the Model N20 CPC drawer that has two nodes.

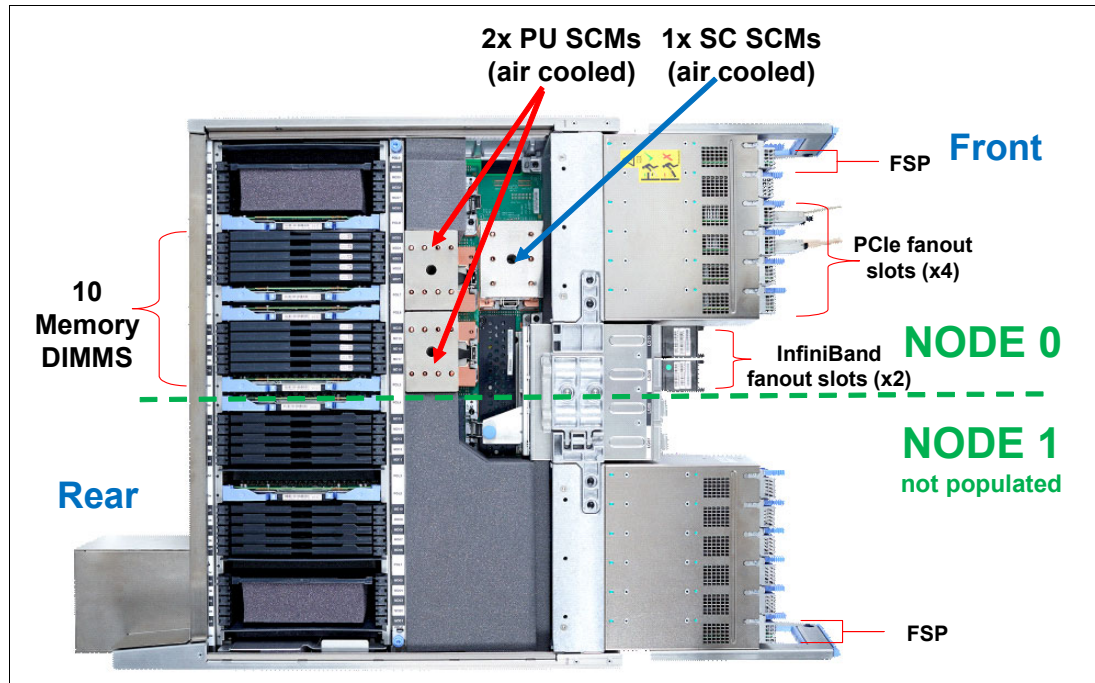


Figure 2-4 Model N10 components (top view)

- ▶ Single node drawer contains (Figure 2-4):
 - Two PU chips up to 13 active cores.
 - One SC chip (480 MB L4 cache).
 - Four PCIe fanout slots.
 - Two InfiniBand fanout slots.
 - One memory controller per PU chip.
 - Five DDR3 DIMM slots per memory controller.
 - Two Flexible Support Processors (FSPs).
 - Node 1 resources (SCMs, memory, and fanouts) are not installed in the Model N10.
- ▶ Two node drawer contains (Figure 2-5 on page 41):
 - Four PU chips, up to 26 active PUs
 - Two SC chips (960 MB L4 cache)
 - Eight PCIe fanout slots
 - Four InfiniBand fanout slots
 - One memory controller per PU chip
 - Five DDR3 DIMM slots per memory controller
 - Two FSPs

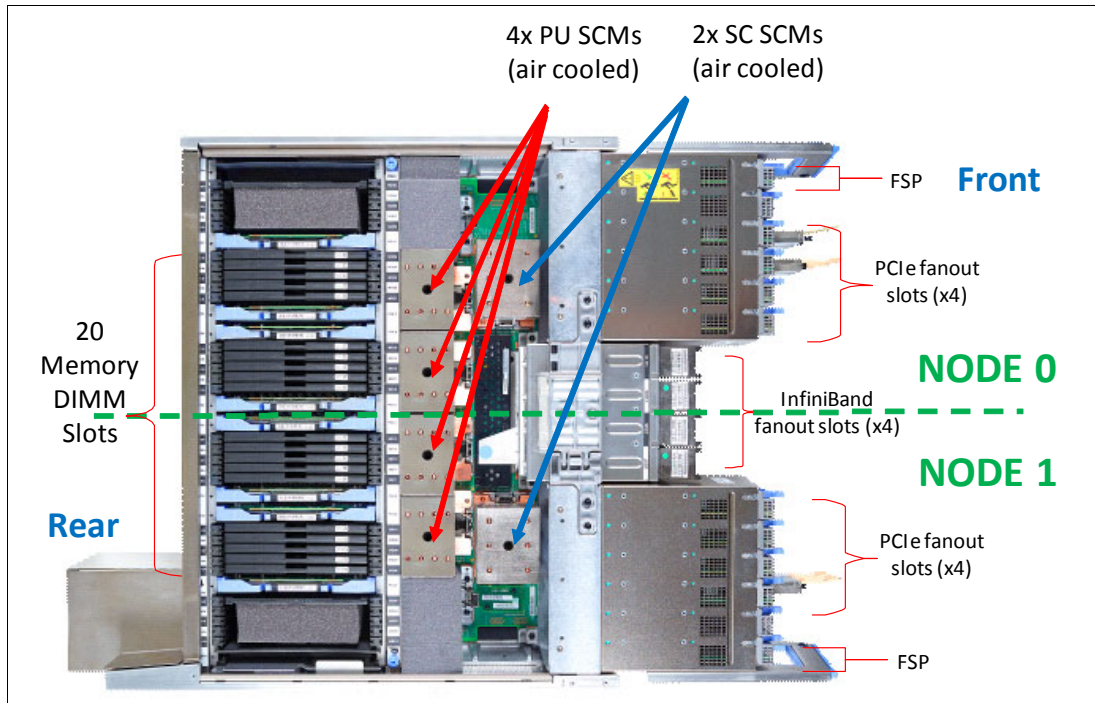


Figure 2-5 Model N20 components (top view)

Figure 2-6 shows the front view of the CPC drawer with fanout slots and FSP slots for both a Model N20 and a Model N10.

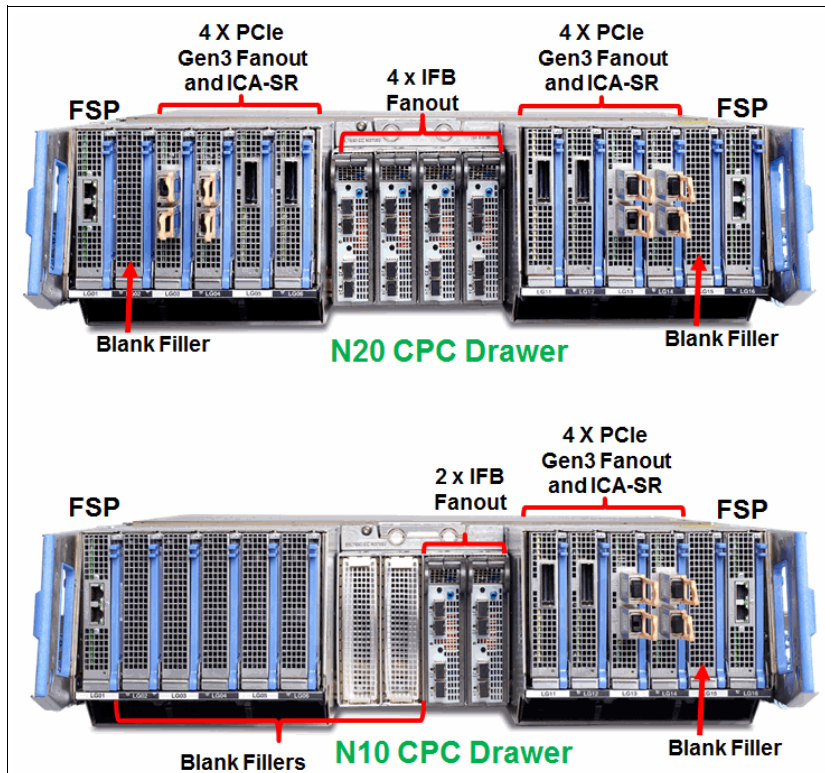


Figure 2-6 CPC drawer (front view)

Figure 2-7 shows the memory DIMM to PU SCM relationship and the various buses on the nodes. Memory is connected to the SCMs through memory control units (MCUs). There is one MCU per PU SCM that provide the interface to controller on memory DIMM. A memory controller drives five DIMM slots.

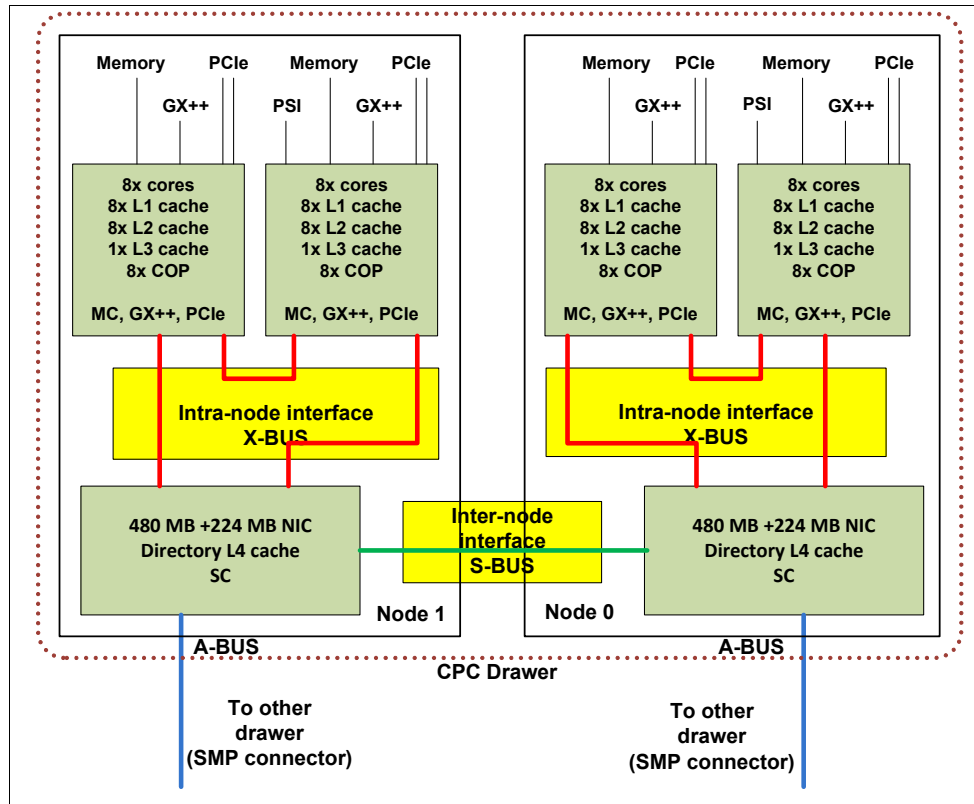


Figure 2-7 CPC drawer logical structure¹

The buses are organized as follows:

- ▶ The GX++ I/O bus slots provide connectivity for host channel adapters (HCAs). They are fully buffered, and can sustain up to 6 GBps data transfer per bus direction. GX++ I/O slots provide support for InfiniBand and non-PCIe I/O features (FICON Express 8).
- ▶ The PCIe I/O buses provide connectivity for PCIe fanouts and can sustain up to 18 GBps data traffic per bus direction.
- ▶ The X-bus provides interconnects between SC chip and PUs chips to each other, in the same node.
- ▶ The S-bus provides interconnects between SC chips in the same drawer.
- ▶ The A-bus provides interconnects between SC chips in different drawers (through SMP cables).
- ▶ Processor support interfaces (PSIs) are used to communicate with FSP cards for system control.

¹ The drawing shows eight cores per PU chip. This is by design. The PU chip is a FRU shared with the z13. When installed in a z13s server, each PU chip can have either six or seven active cores,

2.2.1 CPC drawer interconnect topology

Figure 2-8 shows the point-to-point topology for CPC drawers and node communication through the SMP cables. Each CPC drawer communicates directly to all other processor drawers in the CPC through two ways.

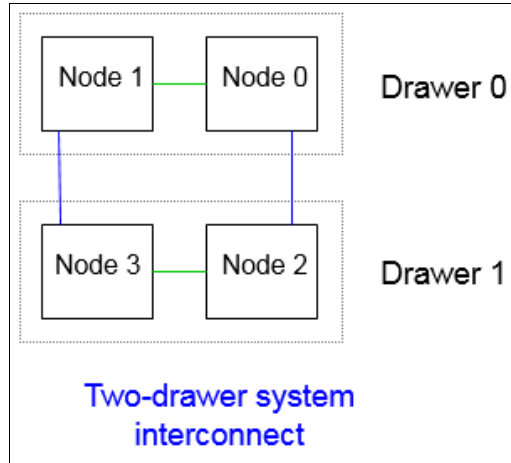


Figure 2-8 Drawer to drawer communication

The CPC drawers are populated from bottom to top. Table 2-2 indicates where the CPC drawers are installed.

Table 2-2 CPC drawers installation order and position in Frame A

CPC drawer	CPC drawer 1	CPC drawer 0
Position in Frame A	A21A	A25A

2.2.2 Oscillator

z13s servers have two oscillator cards (OSCs): One primary and one backup. If the primary OSC fails, the secondary detects the failure, takes over transparently, and continues to provide the clock signal to the CPC. The two oscillators have Bayonet Neill-Concelman (BNC) connectors that provide pulse per second signal (PPS) synchronization to an external time source with PPS output.

The SEs provide the Simple Network Time Protocol (SNTP) client function. When Server Time Protocol (STP) is used, the time of an STP-only Coordinated Timing Network (CTN) can be synchronized to the time that is provided by an NTP server. This configuration allows time-of-day (TOD) synchronization in a heterogeneous platform environment.

The accuracy of an STP-only CTN is improved by using an NTP server with the PPS output signal as the external time source (ETS). NTP server devices with PPS output are available from several vendors that offer network timing solutions. A cable connection from the PPS port on the OSC to the PPS output of the NTP server is required when the z13s server is using STP and is configured in an STP-only CTN using NTP with PPS as the external time source. z13s servers cannot participate in a mixed CTN. They can participate only in an STP only CTN.

STP tracks the highly stable and accurate PPS signal from the NTP server and maintains an accuracy of 10 μ s to the ETS, as measured at the PPS input of IBM z13s servers.

If STP uses an NTP server without PPS, a time accuracy of 100 ms to the ETS is maintained.

Although not part of the CPC drawer design, the OSCs cards are located beside the drawers, and are connected to the same backplane to which the drawers are connected. Both drawers connect to the OSC.

Figure 2-9 shows the location of the two OSC cards with BNC connectors for PPS on the CPC, which is beside the drawer 1 and drawer 0 locations.

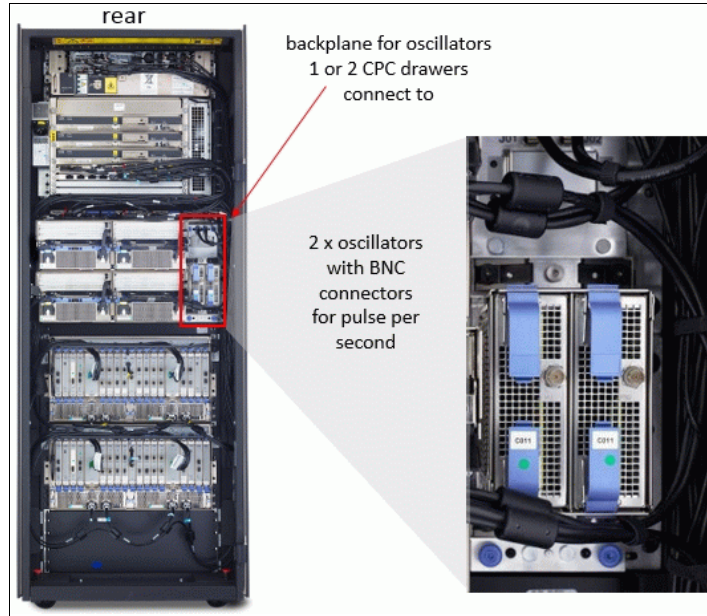


Figure 2-9 Oscillators cards

Tip: STP is available as FC 1021. It is implemented in the Licensed Internal Code (LIC), and allows multiple servers to maintain time synchronization with each other and synchronization to an ETS. For more information, see the following publications:

- ▶ *Server Time Protocol Planning Guide*, SG24-7280
- ▶ *Server Time Protocol Implementation Guide*, SG24-7281
- ▶ *Server Time Protocol Recovery Guide*, SG24-7380

2.2.3 System control

Various system elements are managed through the flexible service processors (FSPs). An FSP is based on the IBM PowerPC® microprocessor technology. Each FSP card has two ports that connect to two internal Ethernet LANs, through system control hubs (SCH1 and SCH2). The FSPs communicate with the SEs and provide a subsystem interface (SSI) for controlling components.

Figure 2-10 depicts a logical diagram of the system control design.

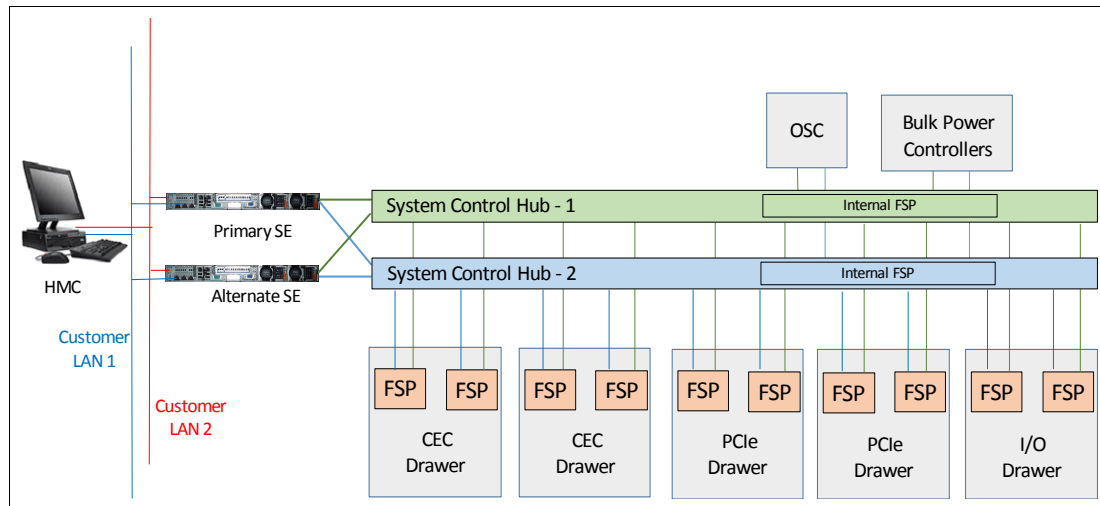


Figure 2-10 Conceptual overview of system control elements

Note: The maximum number of drawers (CEC and I/O) is four for z13s servers. The diagram in Figure 2-10 references the various supported FSP connections

A typical FSP operation is to control a power supply. An SE sends a command to the FSP to start the power supply. The FSP (through SSI connections) cycles the various components of the power supply, monitors the success of each step and the resulting voltages, and reports this status to the SE.

Most system elements are duplexed ($n+1$), and each element has at least one FSP. Two *internal* Ethernet LANs are used by two SEs, for redundancy. Crossover capability between the LANs is available so that both SEs can operate on both LANs.

The HMCs and SEs are connected directly to one or two Ethernet Customer LANs. One or more HMCs can be used.

2.2.4 CPC drawer power

Each CPC drawer gets its power from two distributed converter assemblies (DCAs) in the CPC drawer (see Figure 2-11). The DCAs provide the power for the CPC drawer. Loss of one DCA leaves enough power to satisfy CPC drawer power requirements. The DCAs can be concurrently maintained, and are accessed from the rear of the frame. During a blower failure, the adjacent blower increases speed to satisfy cooling requirements until the failed blower is replaced.



Figure 2-11 Redundant DCAs and blowers for CPC drawers

2.3 Single chip modules

The SCM is a multi-layer metal substrate module that holds either one PU chip or a SC chip. Its size is 678.76 mm² (28.4 mm x 23.9 mm). Each node of a CPC drawer has three SCMs, two PU SCMs, and one SC SCM. The Model N20 CPC drawer has six SCMs, four PU SCMs, and two SC SCMs, with more than 20 billion transistors in total.

2.3.1 Processor units and system control chips

The two types of SCMs (PU and SC) are shown in Figure 2-12.

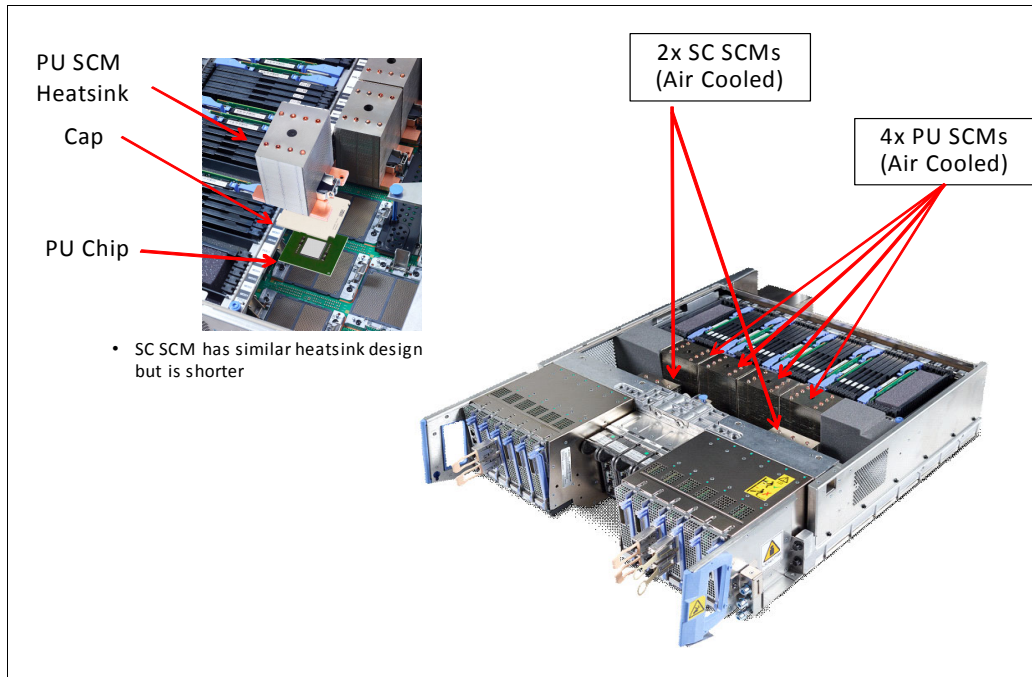


Figure 2-12 Single chip modules (PU SCM and SC SCM) N20 CPC Drawer

Both PU and SC chips use CMOS 14S0 process state-of-the-art semiconductor technology, which is implemented with 17-layer (PU chip) or 15-layer (SC chip) copper interconnections and Silicon-On-Insulator (SOI) technologies. The chip lithography line width is 22 nm.

The SCMs are plugged into a card that is part of the CPC drawer packaging. The interconnectivity between the CPC drawers is accomplished through SMP connectors and cables. One inter-drawer connection per node of the CPC drawer is used for the two drawer Model N20 configuration. This configuration allows a multi drawer system to be displayed as a symmetric multiprocessor (SMP) system.

Each node has three SCMs: Two PU SCMs and one SC SCM.

2.3.2 Processor unit (PU) chip

The z13s PU chip (installed as a PU SCM) is an evolution of the zBC12 core design. It uses CMOS 14S0 technology, out-of-order instruction processing, pipeline enhancements, dynamic simultaneous multithreading (SMT), single-instruction multiple-data (SIMD), and redesigned, larger caches.

By design, each PU chip has eight cores. Core frequency is 4.3 GHz with a cycle time of 0.233 ns. When installed in a z13s server, the PU chips have either six or seven active cores. This limit means that a Model N10 has 13 active cores, and the Model N20 has 26 active cores. Model N10 has 10 customizable cores, whereas Model N20 has 20 customizable cores. A schematic representation of the PU chip is shown in Figure 2-13.

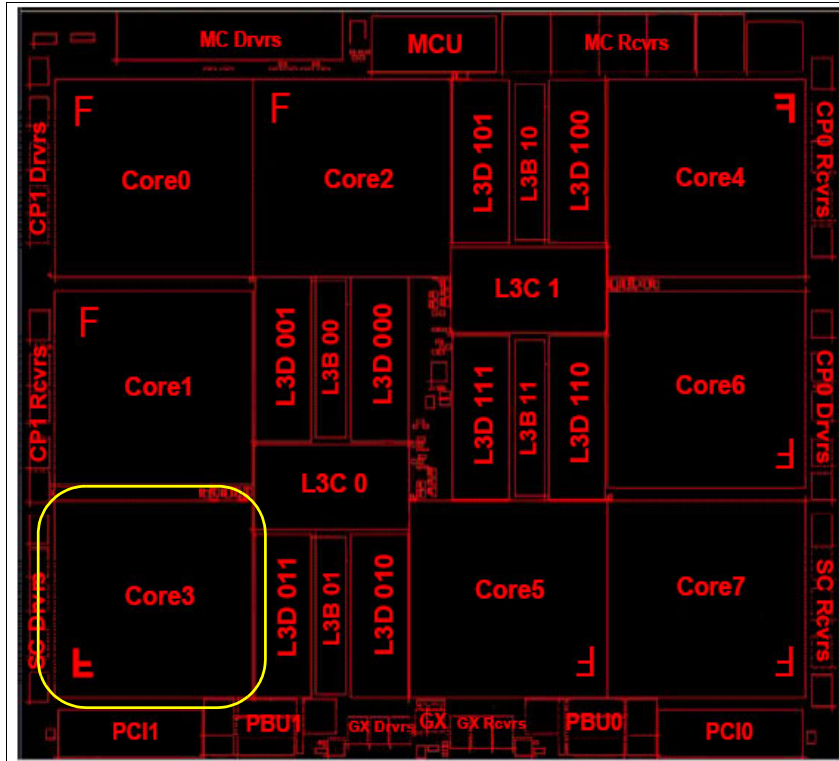


Figure 2-13 PU Chip Floorplan

Each PU chip has 3.99 billion transistors. Each one of the eight cores has its own L1 cache with 96 KB for instructions and 128 KB for data. Next to each core is its private L2 cache, with 2 MB for instructions and 2 MB for data.

Each PU chip has one L3 cache, with 64 MB. This 64 MB L3 cache is a store-in shared cache across all cores in the PU chip. It has 192 x 512 KB eDRAM macros, dual address-sliced and dual store pipe support, an integrated on-chip coherency manager, cache, and cross-bar switch. The L3 directory filters queries from the local L4. Both L3 slices can deliver up to 16 GBps bandwidth to each core simultaneously. The L3 cache interconnects the eight cores, GX++ I/O buses, PCIe I/O buses, and memory controllers (MCs) with SC chips.

The MC function controls access to memory. The GX++ I/O bus controls the interface to the InfiniBand fanouts, while the PCIe bus controls the interface to PCIe fanouts. The chip controls traffic between the cores, memory, I/O, and the L4 cache on the SC chips.

One coprocessor is dedicated for data compression and encryption functions for each core. The compression unit is integrated with the CP Assist for Cryptographic Function (CPACF), benefiting from combining (or sharing) the use of buffers and interfaces. The assist provides high-performance hardware encrypting and decrypting support for clear key operations.

For more information, see 3.4.5, “Compression and cryptography accelerators on a chip” on page 95.

2.3.3 Processor unit (core)

Each processor unit, or core, is a superscalar and out-of-order processor that has 10 execution units and two load/store units, which are divided into two symmetric pipelines as follows:

- ▶ Four fixed-point units (FXUs) (integer)
- ▶ Two load/store units (LSUs)
- ▶ Two binary floating-point units (BFUs)
- ▶ Two binary coded decimal floating-point units (DFUs)
- ▶ Two vector floating point units (vector execution units (VXUs))

Up to six instructions can be decoded per cycle, and up to 10 instructions/operations can be initiated to run per clock cycle. The running of the instructions can occur out of program order, and memory address generation and memory accesses can also occur out of program order. Each core has special circuitry to display execution and memory accesses in order to the software. Not all instructions are directly run by the hardware, which is the case for several complex instructions. Some are run by millicode, and some are broken into multiple operations that are then run by the hardware.

Each core has the following functional areas, which are also shown in Figure 2-14 on page 50:

- ▶ Instruction sequence unit (ISU): This unit enables the out-of-order (OOO) pipeline. It tracks register names, OOO instruction dependency, and handling of instruction resource dispatch.

This unit is also central to performance measurement through a function called *instrumentation*.

- ▶ Instruction fetch and branch (IFB) (prediction) and instruction cache and merge (ICM): These two subunits (IFB and ICM) contain the instruction cache, branch prediction logic, instruction fetching controls, and buffers. The relative size of these subunits is the result of the elaborate branch prediction design, which is described in 3.4.4, “Superscalar processor” on page 95.
- ▶ Instruction decode unit (IDU): The IDU is fed from the IFB buffers, and is responsible for the parsing and decoding of all z/Architecture operation codes.
- ▶ Load/store unit (LSU): The LSU contains the data cache. It is responsible for handling all types of operand accesses of all lengths, modes, and formats that are defined in the z/Architecture.
- ▶ Translation unit (XU): The XU has a large translation look aside buffer (TLB) and the dynamic address translation (DAT) function that handles the dynamic translation of logical to physical addresses.
- ▶ Core pervasive unit - PC: Used for instrumentation and error collection.
- ▶ Vector and floating point units:
 - Fixed-point unit (FXU): The FXU handles fixed-point arithmetic.
 - Binary floating-point unit (BFU): The BFU handles all binary and hexadecimal floating-point and fixed-point multiplication operations.
 - Decimal floating-point unit (DFU): The DFU runs both floating-point and decimal fixed-point operations and fixed-point division operations.
 - Vector execution unit (VXU)

- ▶ Recovery unit (RU): The RU keeps a copy of the complete state of the system that includes all registers, collects hardware fault signals, and manages the hardware recovery actions.
- ▶ Dedicated Coprocessor (COP): The dedicated coprocessor is responsible for data compression and encryption functions for each core.

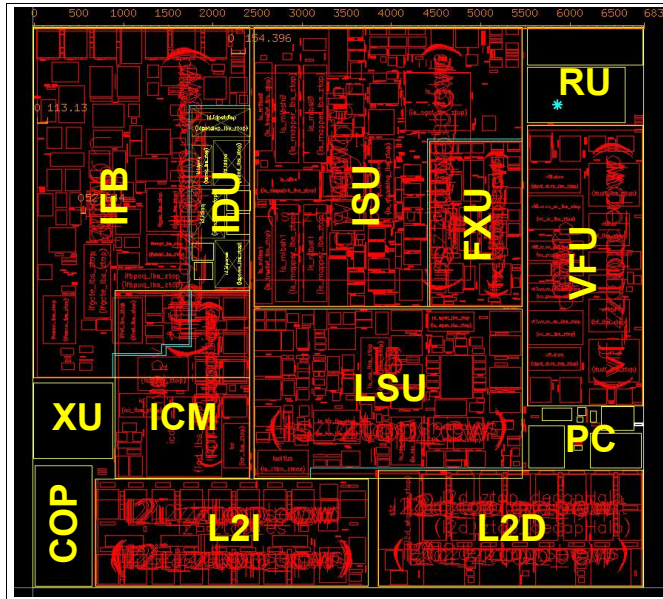


Figure 2-14 PU Core floorplan

2.3.4 PU characterization

In each CPC drawer, PUs can be characterized for client use. The characterized PUs can be used in general to run supported operating systems, such as z/OS, z/VM, and Linux on z Systems. They can also run specific workloads, such as Java, XML services, IPsec, and some DB2 workloads, or functions such as Coupling Facility Control Code (CFCC). For more information about PU characterization, see 3.5, “Processor unit functions” on page 100.

The maximum number of characterized PUs depends on the z13s model. Some PUs are characterized by the system as standard system assist processors (SAPs) to run the I/O processing. By default, there are at least two spare PUs per Model N20 system that are used to assume the function of a failed PU. The remaining installed PUs can be characterized for client use. The Model N10 uses unassigned PUs as spares when available. A z13s model nomenclature includes a number that represents the maximum number of PUs that can be characterized for client use, as shown in Table 2-3.

Table 2-3 Number of PUs per z13s model

Model	CPC Drawers / PUs	PUs	IFLs uIFLs	zIIPs	ICFs	Standard SAPs	Optional SAPs	Standard Spares	IFP
N10	1 / 13	0-6	0-10	0-6	0-10	2	0-2	0	1
N20	1 / 26	0-6	0-20	0-12	0-20	3	0-3	2	1
N20	2 / 26	0-6	0-20	0-12	0-20	3	0-3	2	1

2.3.5 System control chip

The SC chip uses the CMOS 14S0 22 nm SOI technology, with 15 layers of metal. It measures 28.4 x 23.9 mm, has 7.1 billion transistors, and has 2.1 billion cells of eDRAM. Each node of the CPC drawer has one SC chip. The L4 cache on each SC chip has 480 MB of non-inclusive cache and a 224 MB non-data inclusive coherent (NIC) directory, which results in 960 MB of on-inclusive L4 cache and 448 MB in a NIC directory that is shared per CPC drawer.

Figure 2-15 shows a schematic representation of the SC chip.

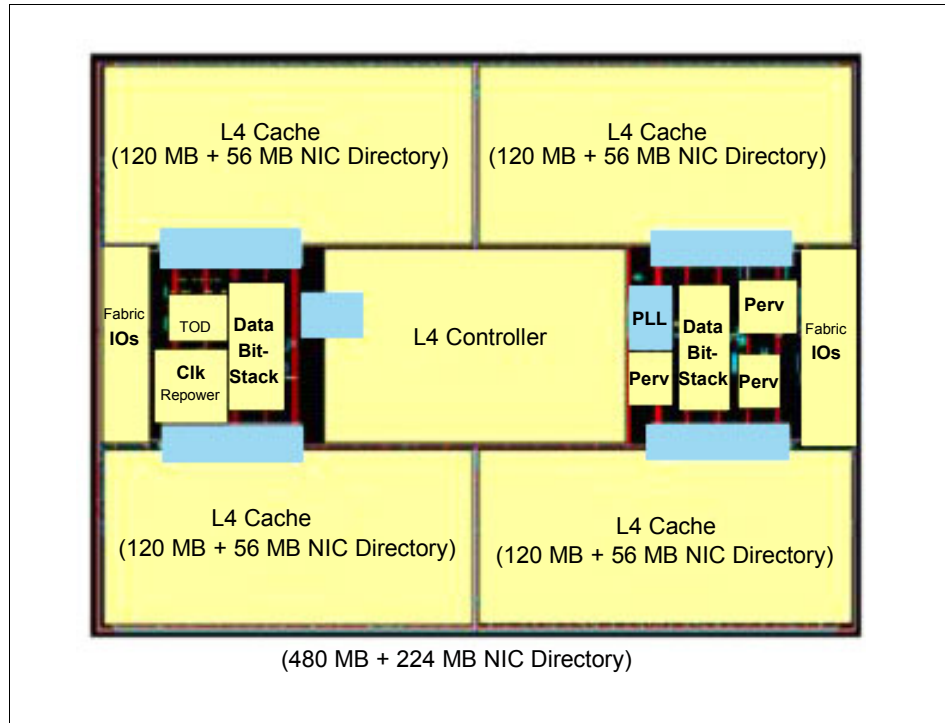


Figure 2-15 SC chip diagram

Most of the SC chip space is taken by the L4 controller and the 480 MB L4 cache. The cache consists of four 120 MB quadrants with 256 x 1.5 MB eDRAM macros per quadrant. The L4 cache is logically organized as 16 address-sliced banks, with 30-way set associative. The L4 cache controller is a single pipeline with multiple individual controllers, which is sufficient to handle 125 simultaneous cache transactions per chip.

The L3 caches on PU chips communicate with the L4 caches through the attached SC chip by using unidirectional buses. L3 is divided into two logical slices. Each slice is 32 MB, and consists of two 16 MB banks. L3 is 16-way set associative. Each bank has 4 K sets, and the cache line size is 256 bytes.

The bus/clock ratio (2:1) between the L4 cache and the PU is controlled by the storage controller on the SC chip.

The SC chip also acts as an L4 cache cross-point switch for L4-to-L4 traffic to up to three remote CPC drawers through three bidirectional data buses. The SMP cables and system coherency manager use the L4 directory to filter snoop traffic from remote CPC drawers. This process uses an enhanced synchronous fabric protocol for improved latency and cache management. There are six clock domains, and the clock function is distributed between both SC chips.

2.3.6 Cache level structure

z13s implements a four level cache structure, as shown in Figure 2-16.

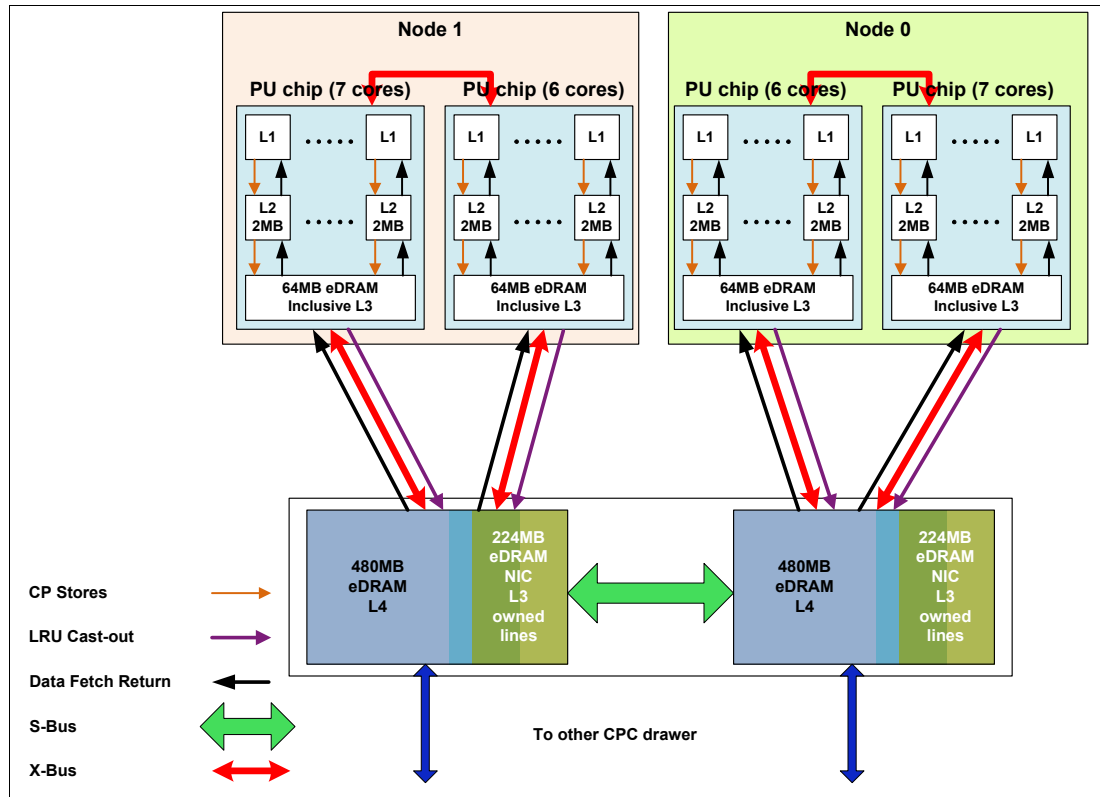


Figure 2-16 Cache level structure

Each core has its own 224-KB Level 1 (L1) cache, split into 96 KB for instructions (I-cache) and 128 KB for data (D-cache). The L1 cache is designed as a store-through cache, meaning that altered data is also stored in the next level of memory.

The next level is the Level 2 (L2) private cache on each core. This cache has 4 MB, split into a 2 MB D-cache and 2 MB I-cache. It is designed as a store-through cache.

The Level 3 (L3) cache is also on the PU chip. It is shared by the active cores, has 64 MB, and is designed as a store-in cache.

Cache levels L2 and L3 are implemented on the PU chip to reduce the latency between the processor and the L4 large shared cache, which is on the two SC chips. Each SC chip has 480 MB, which is shared by PU chips on the node. The S-bus provide the inter-node interface between the two L4 caches (SC chips) in each node. The L4 cache uses a store-in design.

2.4 Memory

The maximum physical memory size is directly related to the number of CPC drawers in the system. The Model N10 CPC drawer can contain up to 1,280 GB of physical memory. The Model N20 single drawer system can contain up to 2,560 GB, and the Model N20 with two drawers up to 5,120 GB of installed (physical) memory per system.

A z13s has more memory installed than the amount that is ordered. Part of the physical installed memory is used to implement the redundant array of independent memory (RAIM) design. As a result, for customer use, Model N10 can have up to 984 GB, Model N20 single CPC drawer up to 2008 GB, and Model N20 with two CPC drawers up to 4056 GB.

Table 2-4 shows the maximum and minimum memory sizes that you can order for each z13s model.

Table 2-4 z13s memory sizes

Model	Increment (GB)	Standard (GB)	Plan Ahead ^a (GB)
N10 and N20	8	64 - 88	88-88
N10 and N20	32	120-344	120-344
N10 and N20	64	408-600	408-600
N10 only	128	728-984	728-984
N20 only	128	1112-2008	1112-2008
N20 (2) only ^b	256	2264-4056	2264-4056

a. The maximum amount of Preplanned Memory capacity is 2 TB.

b. N20 with two CPC drawers (FC 1045)

The minimum physical installed memory is 160 GB per CPC drawer. The minimum initial amount of customer memory that can be ordered is 64 GB for all z13 models. The maximum customer memory size is based on the physical installed memory minus the RAIM (20% of physical) and the hardware system area (HSA) memory, which has a fixed amount of 40 GB.

Table 2-5 shows the memory granularity that is based on the installed customer memory. For more information, see the *z Systems Processor Resource/Systems Manager Planning Guide*, SB10-7162.

Table 2-5 Central storage granularity for z13s servers

Largest Central Storage Amount (LCSA)	LPAR Storage Granularity
LCSA <= 256 GB	512 MB
256 GB < LCSA <= 512 GB	1 GB
512 GB < LCSA <= 1024 GB	2 GB
1024 GB < LCSA <= 2048 GB	4 GB
2048 GB < LCSA <= 4096 GB	8 GB

Note: The required granularity for all main storage fields of an LPAR for which an origin has been specified (for example, initial main storage amount, reserved main storage amount, and main storage origin) is fixed at 2 GB. This configuration helps to simplify customer management of absolute storage regarding 2 GB large page support for these partitions. In support of 2 GB large pages, all logical partition origin MBs and limit MBs must be on a 2 GB boundary.

2.4.1 Memory subsystem topology

The z13s memory subsystem uses high speed, differential-ended communications memory channels to link a host memory to the main memory storage devices.

Figure 2-17 shows an overview of the CPC drawer memory topology of a z13s server.

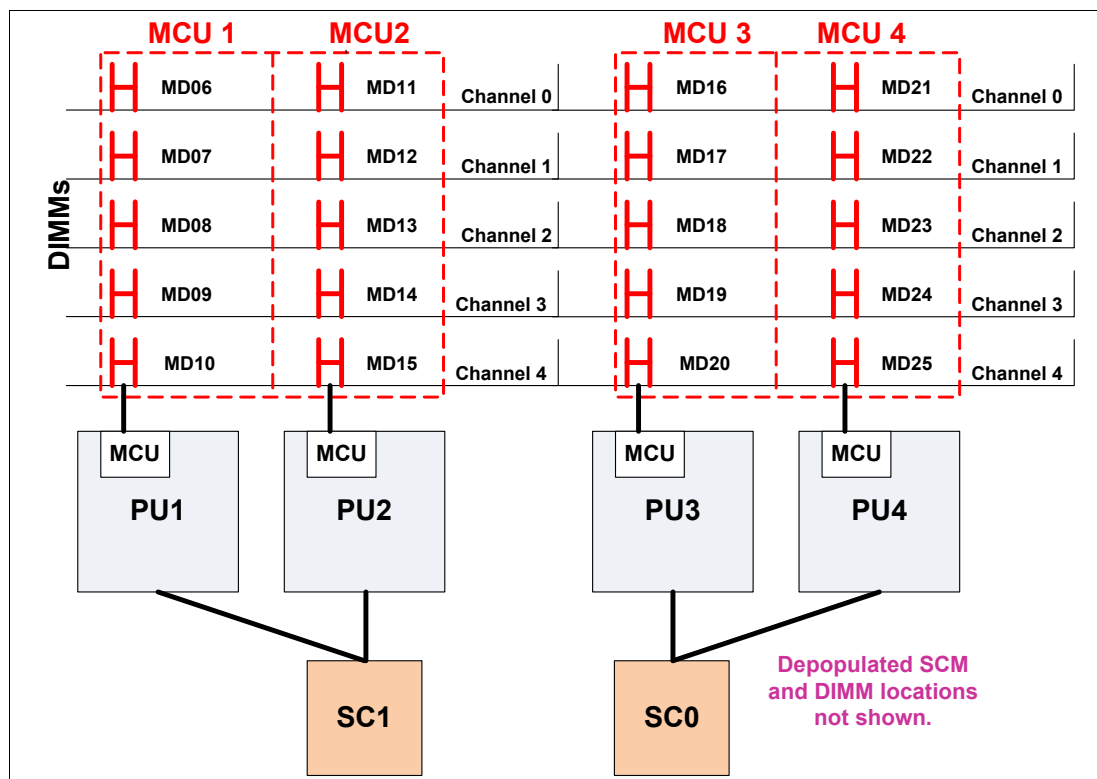


Figure 2-17 CPC drawer memory topology

Each CPC drawer has up to 20 DIMMs. DIMMs are connected to each PU chip through the MCUs. Each PU chip has one MCU, which uses five channels, one for each DIMM and one for RAIM implementation, in a 4 +1 design.

Each DIMM can be 16 GB, 32 GB, 64 GB, or 128 GB. DIMM sizes cannot be mixed in the same CPC drawer, but a two CPC drawer Model N20 can have different (but not mixed) DIMM sizes in each drawer.

2.4.2 Redundant array of independent memory

z13s servers use the RAIM technology. The RAIM design detects and recovers from failures of DRAMs, sockets, memory channels, or DIMMs.

The RAIM design requires the addition of one memory channel that is dedicated for reliability, availability, and serviceability (RAS), as shown in Figure 2-18.

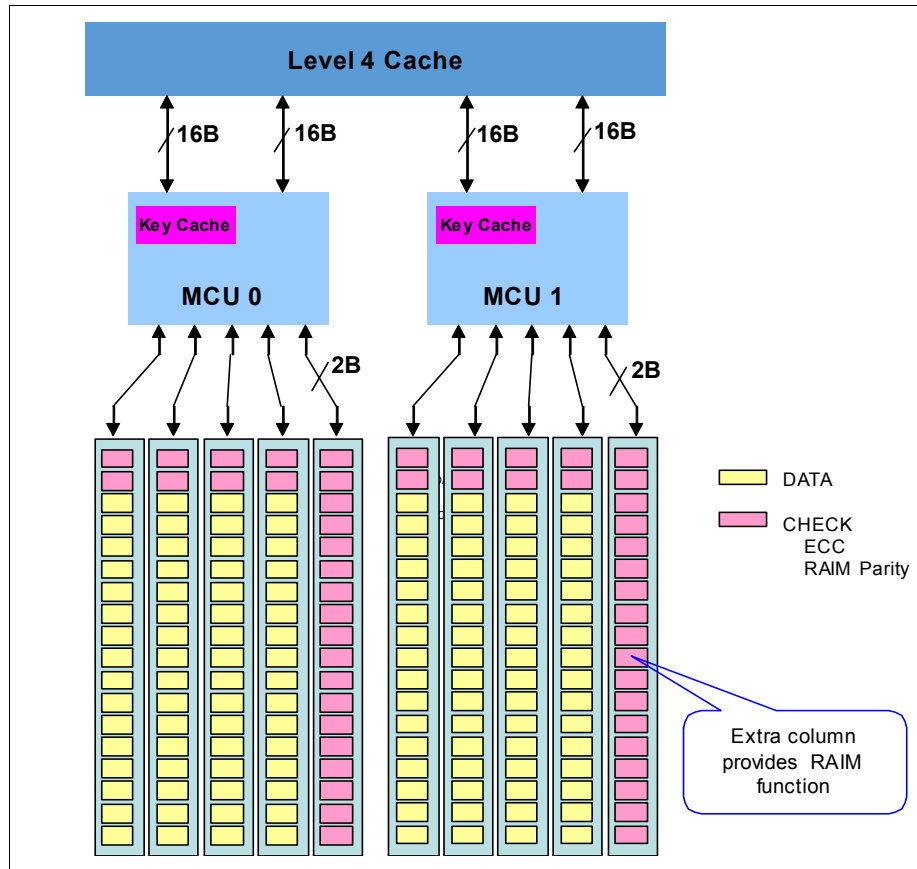


Figure 2-18 RAIM configuration per node

The parity of the four “data” DIMMs is stored in the DIMMs that are attached to the fifth memory channel. Any failure in a memory component can be detected and corrected dynamically. This system simplifies the memory subsystem design, while maintaining a fully fault-tolerant RAIM design.

The RAIM design provides the following layers of memory recovery:

- ▶ ECC with 90B/64B Reed Solomon code.
- ▶ DRAM failure, with marking technology in which two DRAMs can be marked and no half sparing is needed. A call for replacement occurs on the third DRAM failure.
- ▶ Lane failure with CRC retry, data-lane sparing, and clock-RAIM with lane sparing.
- ▶ DIMM failure with CRC retry, data-lane sparing, and clock-RAIM with lane sparing.
- ▶ DIMM controller ASIC failure.
- ▶ Channel failure.

2.4.3 Memory configurations

Physically, memory is organized in the following manner:

- ▶ Within a drawer, all slots must be populated with the same size DIMM.
- ▶ Each CPC drawer in a two drawer Model N20 can contain different amounts of memory.
- ▶ A CPC drawer can have available unused memory, which can be ordered as a memory upgrade and enabled by LIC without DIMM changes. The amount of memory that can be enabled by the client is the total physical installed memory minus the RAIM amount and minus the 40 GB of HSA memory.
- ▶ The memory controller on the adjacent PU chip manages the five memory slots.
- ▶ DIMM changes require a disruptive IML on z13s.

Model N10 memory configurations

The memory in Model N10 can be configured in the following manner:

- ▶ Ten DIMM slots per N10 drawer supported by two memory controllers on PU chips with five slots each.
- ▶ Any remaining unpopulated DIMM banks, MD06-MD10 and MD11-MD15, must be plugged with memory DIMM airflows
- ▶ Drawer memory configuration based on DIMM sizes of 16, 32, 64, 128 supporting up to 984 GB customer available memory in the drawer.
- ▶ HSA size (40 GB) is fixed and is not taken from customer purchased memory.

Figure 2-19 shows the physical view of the N10 CPC drawer memory DIMM plug locations.

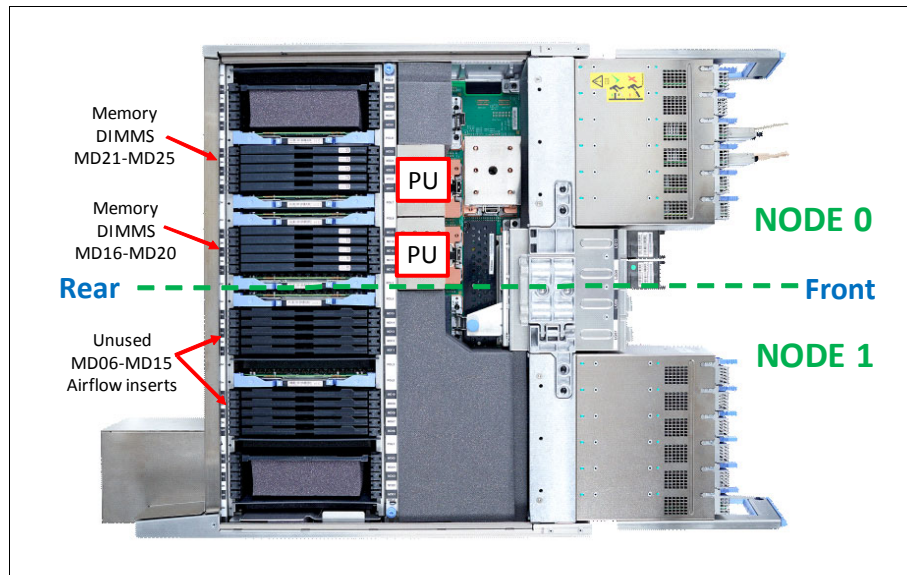


Figure 2-19 Model N10 memory plug locations

Table 2-6 shows the standard memory plug summary by node for new build systems.

Table 2-6 Model N10 physically installed memory

Customer Memory (GB)	Total Physical GB	Increment GB	Node 0 DIMM location / size GB		
			MD16-MD20	MD21-MD25	Dial Max
64	160	8	16	16	88
72			16	16	
80			16	16	
88			16	16	
120	320	32	32	32	216
152			32	32	
184			32	32	
216			32	32	
248	640	32	64	64	472
280			64	64	
312			64	64	
344			64	64	
408		64	64	64	
472			64	64	
536	1280	32	128	128	984
600			128	128	
728		128	128	128	
856			128	128	
984			128	128	

Model N20 single drawer memory configurations

The memory in Model N20 with a single drawer can be configured in the following manner:

- ▶ Ten or twenty DIMM slots per N20 drawer supported by two or four memory controllers with five slots each.
- ▶ The remaining unpopulated DIMM banks, MD06-MD10 and MD21-MD25, must be plugged with memory DIMM airflows.
- ▶ Drawer memory configurations based on DIMM sizes of 16, 32, 64, 128 supporting up to 2008 GB customer available memory in the drawer.
- ▶ HSA size (40 GB) is fixed and is not taken from customer purchased memory.

Figure 2-20 shows physical view of the N20 one CPC drawer memory DIMM plug locations

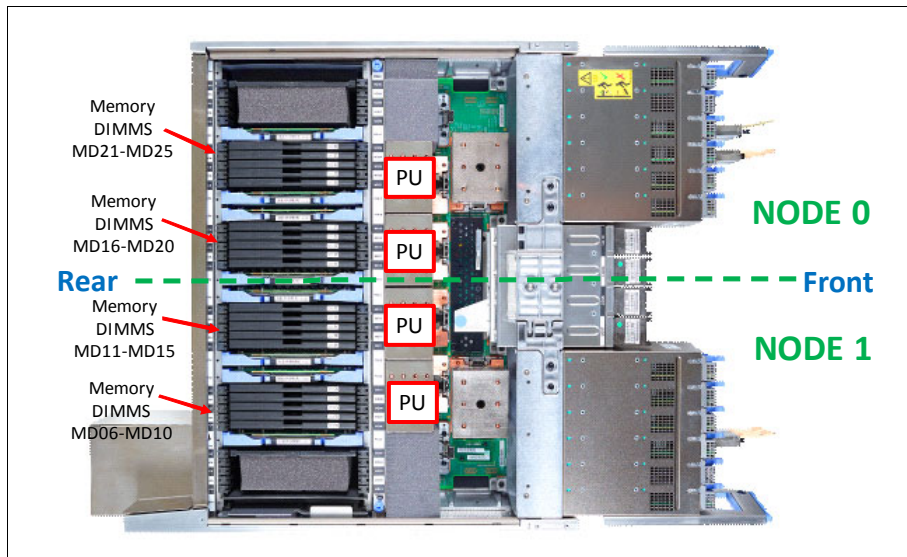


Figure 2-20 Model N20 one drawer memory plug locations

Table 2-7 shows the standard memory plug summary by node for new build systems.

Table 2-7 Model N20 single CPC drawer - physically installed memory

Cust Mem (GB)	Total Phys (GB)	Increment (GB)	Node 1 DIMM location / size GB		Node 0 DIMM location / size GB		Dial Max
			MD06-MD10	MD11-MD15	MD16-MD20	MD21-MD25	
64	160	8		16	16		88
72				16	16		
80				16	16		
88				16	16		

Cust Mem (GB)	Total Phys GB	Increment GB	Node 1 DIMM location / size GB		Node 0 DIMM location / size GB		Dial Max
			MD06-MD10	MD11-MD15	MD16-MD20	MD21-MD25	
120	320	32	16	16	16	16	216
152			16	16	16	16	
184			16	16	16	16	
216			16	16	16	16	
248	640	32	32	32	32	32	472
280			32	32	32	32	
312			32	32	32	32	
344			32	32	32	32	
408	640	64	32	32	32	32	472
472			32	32	32	32	
536	1280	64	64	64	64	64	984
600			64	64	64	64	
728		128	64	64	64	64	
856			64	64	64	64	
984	64		64	64	64		
1112	2560	128	128	128	128	128	2008
1240			128	128	128	128	
1368			128	128	128	128	
1496			128	128	128	128	
1624			128	128	128	128	
1752			128	128	128	128	
1880			128	128	128	128	
2008			128	128	128	128	

Model N20 two drawer memory configurations

The memory in Model N20 with two drawers can be configured in the following manner:

- ▶ Ten or twenty DIMM slots per N20 drawer supported by two or four memory controllers with five slots each.
- ▶ The remaining unpopulated DIMM banks, MD06-MD10 and MD21-MD25, must be plugged with memory DIMM airflows.
- ▶ Drawer memory configurations based on DIMM sizes of 16, 32, 64, 128 supporting up to 2008 GB in the drawer and up to 4056 GB customer available memory per system.
- ▶ HSA size (40 GB) is fixed and is not taken from customer purchased memory.

Figure 2-21 shows the physical view of the N20 two CPC drawer memory DIMM plug locations.

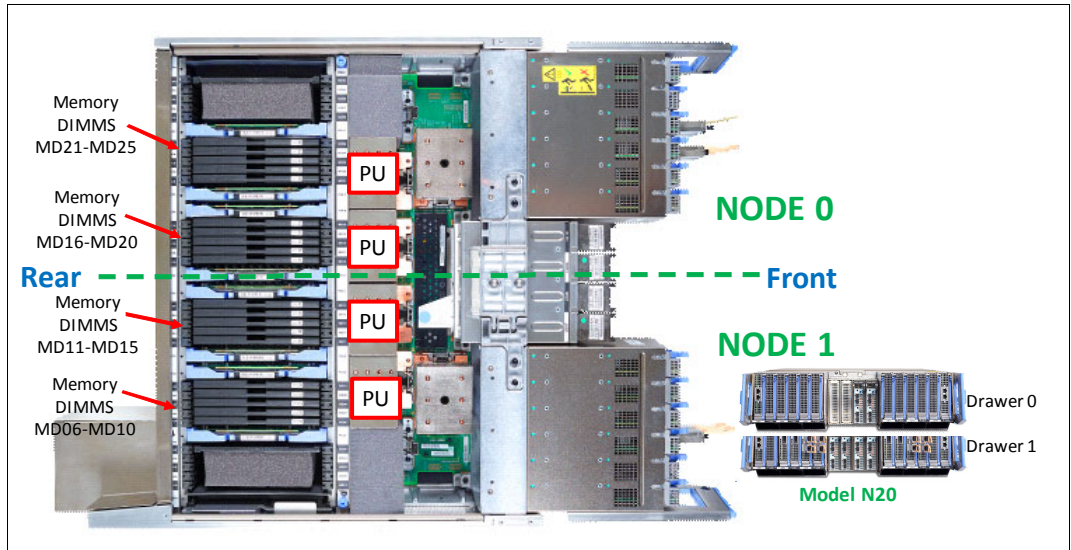


Figure 2-21 Model N20 two drawer memory plug locations

Table 2-8 shows the Standard Memory Plug summary by node for new build systems.

Table 2-8 Model N20 with two CPC drawers - physically installed memory

Cust. Mem (GB)	Physical GB	Increment GB	Drawer 1				Drawer 0 (Second drawer)				Dial Max
			Node 1 DIMM loc / size		Node 0 DIMM loc/size		Node 1 DIMM loc /size		Node 0 DIMM loc/size		
			MD06-MD10	MD11-MD15	MD16-MD20	MD21-MD25	MD06-MD10	MD11-MD15	MD16-MD20	MD21-MD25	
64	320	8		16	16			16	16		88
72				16	16			16	16		
80				16	16			16	16		
88				16	16			16	16		

Cust. Mem (GB)	Physical GB	Increment GB	Drawer 1				Drawer 0 (Second drawer)				Dial Max
			Node 1 DIMM loc / size		Node 0 DIMM loc/size		Node 1 DIMM loc /size		Node 0 DIMM loc/size		
			MD06- MD10	MD11- MD15	MD16- MD20	MD21- MD25	MD06- MD10	MD11- MD15	MD16- MD20	MD21- MD25	
120	480	32	16	16	16	16		16	16		216
152			16	16	16	16		16	16		
184			16	16	16	16		16	16		
216			16	16	16	16		16	16		
248	640	32	32	32	32	32		16	16		472
280			32	32	32	32		16	16		
312			32	32	32	32		16	16		
344			32	32	32	32		16	16		
408	640	64	32	32	32	32		16	16		472
472			32	32	32	32		16	16		
536	1280	64	64	64	64	64		16	16		984
600			64	64	64	64		16	16		
728		1280	128	64	64	64	64		16	16	
856	64			64	64	64		16	16		
984	64			64	64	64		16	16		
1112	2720			128	128	128	128	128		16	16
1240		128	128		128	128		16	16		
1368		128	128		128	128		16	16		
1496		128	128		128	128		16	16		
1624		128	128		128	128		16	16		
1752		128	128		128	128		16	16		
1880		128	128		128	128		16	16		
2008	128	128	128	128		16	16				

Cust. Mem (GB)	Physical (GB)	Increment (GB)	Drawer 1				Drawer 0 (Second drawer)				Dial Max
			Node 1 DIMM loc / size		Node 0 DIMM loc/size		Node 1 DIMM loc /size		Node 0 DIMM loc/size		
			MD06-MD10	MD11-MD15	MD16-MD20	MD21-MD25	MD06-MD10	MD11-MD15	MD16-MD20	MD21-MD25	
2264	2880	256	128	128	128	128	16	16	16	16	2264
2520	3200		128	128	128	128	32	32	32	32	2520
2776	3840		128	128	128	128	64	64	64	64	3032
3032			128	128	128	128	64	64	64	64	
3288	5120		128	128	128	128	128	128	128	128	4056
3544			128	128	128	128	128	128	128	128	
3800			128	128	128	128	128	128	128	128	
4056			128	128	128	128	128	128	128	128	

Figure 2-22 illustrates how the physical installed memory is allocated on a z13s server, showing HSA memory, RAIM, customer memory, and the remaining available unused memory that can be enabled by using LIC when required.

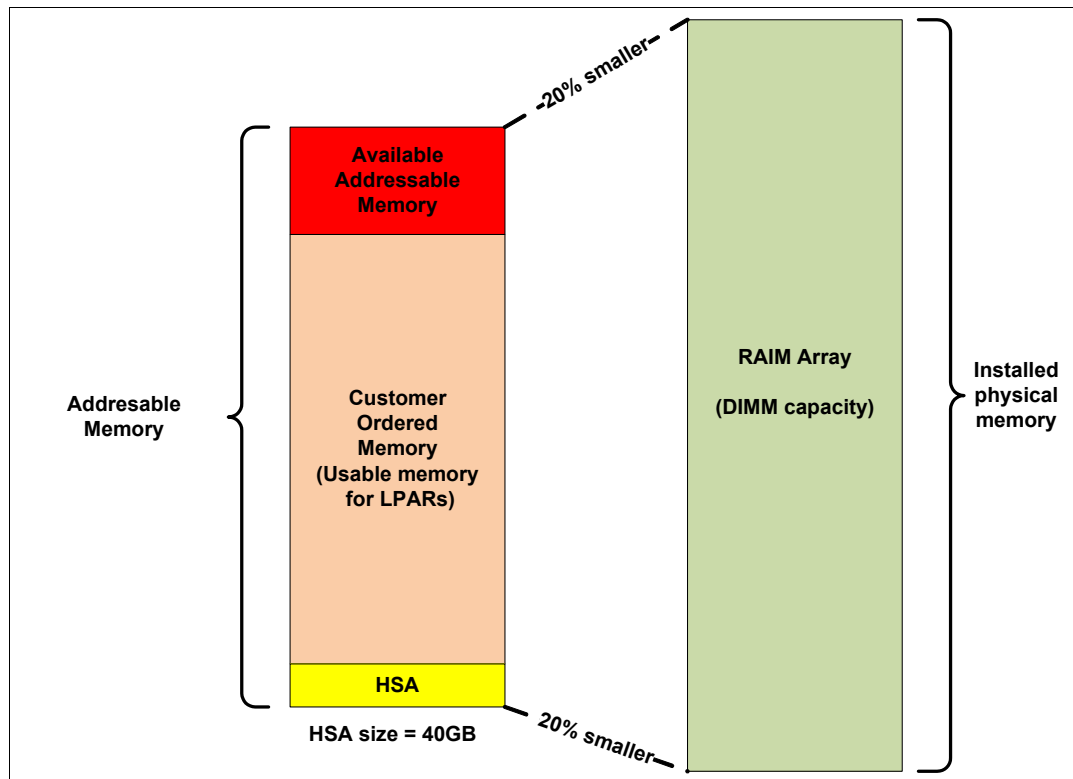


Figure 2-22 Memory allocation diagram

As an example, a z13s Model N20 (one CPC drawer) that is ordered with 1496 GB of memory has the following memory sizes:

- Physical installed memory is 2560 GB: 1280 GB on Node 0 and 1280 GB on Node 1.

- ▶ CPC drawer 1 has 40 GB of HSA memory and up to 2008 GB for customer memory.
- ▶ Because the customer ordered 1496 GB, provided the granularity rules are met, 512 GB are available for future nondisruptive upgrades by activating LIC.

Memory upgrades are satisfied from already installed unused memory capacity until it is exhausted. When no more unused memory is available from the installed memory cards (DIMMs), one of the following additions must occur:

- ▶ Memory cards must be upgraded to a higher capacity.
- ▶ A CPC drawer with more memory must be added.
- ▶ Memory cards (DIMMs) must be added.

Memory upgrades are concurrent when it requires no change of the physical memory cards. A memory card change is always disruptive.

When activated, an LPAR can use memory resources that are in any CPC drawer. No matter where the memory is, an LPAR has access to that memory up to a maximum of 4 TB. This access is possible because despite the CPC drawer structure, the z13s is still an SMP system. The existence of an I/O drawer in the CPC limits the memory LPAR to 1 TB. For more information, see 3.7, “Logical partitioning” on page 116.

2.4.4 Memory upgrades

Memory upgrades can be ordered and enabled using Licensed Internal Code Configuration Control (LICCC) by upgrading the DIMM cards, by adding new DIMM cards, or by adding a CPC drawer.

For a model upgrade that results in the addition of a CPC drawer, the minimum memory increment is added to the system. Each CPC drawer has a minimum physical memory size of 160 GB.

During a model upgrade, adding a CPC drawer is a disruptive operation. Adding physical memory to any drawer is also disruptive.

2.4.5 Preplanned memory

The idea of preplanned memory is to allow for nondisruptive memory upgrades. Any hardware that is required would be pre-plugged based on a target capacity specified by the customer. The maximum amount of Preplanned Memory capacity is 2 TB. Pre-plugged hardware can be enabled by using an LICCC order placed by the customer when additional memory capacity is needed.

You can order this LICCC through these channels:

- ▶ The IBM Resource Link® (a login is required):
<http://www.ibm.com/servers/resourceLink/>
- ▶ Your IBM representative

The installation and activation of any preplanned memory requires the purchase of the required feature codes (FC), which are described in Table 2-9. FCs 1996 and 1993 are used to track the quantity of 16 GB and 8 GB physical increments of plan ahead memory capacity.

The payment for plan-ahead memory is a two-phase process. One charge takes place when the plan-ahead memory is ordered, and another charge takes place when the prepaid

memory is activated for actual use. For the exact terms and conditions, contact your IBM representative.

Table 2-9 Feature codes for plan-ahead memory

Memory	z13s Feature Code / Increment
Preplanned memory Charged when physical memory is installed. Used for tracking the quantity of physical increments of plan-ahead memory capacity.	FC 1993 / 8 GB FC 1996 / 16 GB
Preplanned memory activation Charged when plan-ahead memory is enabled. Used for tracking the quantity of increments of plan-ahead memory being activated.	FC 1903

You install preplanned memory by ordering FC 1993 / 1996. The ordered amount of plan-ahead memory is charged with a reduced price compared to the normal price for memory. One FC 1993 is needed for each 8 GB physical increment, and one FC 1996 for each 16 GB physical increment.

The activation of installed pre-planned memory is achieved by ordering FC 1903, which causes the other portion of the previously contracted charge price to be invoiced. FC 1903 indicates 8 GB, and FC 1996 for 16 GB of LICCC increments of memory capacity.

Memory upgrades: Normal memory upgrades use up the plan-ahead memory first.

2.5 Reliability, availability, and serviceability

IBM z Systems continue to deliver enterprise class RAS with IBM z13s servers. The main intent behind RAS is about preventing or tolerating (masking) outages and providing the necessary instrumentation (in hardware, LIC/microcode, and software) to capture (collect) the relevant failure information to help identify an issue without requiring a reproduction of the event. These outages can be planned or unplanned. Planned and unplanned outages can include the following situations (examples are not related to the RAS features of z Systems servers):

- ▶ A planned outage because of the addition of extra processor capacity
- ▶ A planned outage because of the addition of extra I/O cards
- ▶ An unplanned outage because of a failure of a power supply
- ▶ An unplanned outage because of a memory failure

The z Systems hardware has decades of intense engineering behind it, which has resulted in a robust and reliable platform. The hardware has many RAS features built into it.

2.5.1 RAS in the CPC memory subsystem

Patented error correction technology in the memory subsystem continues to provide the most robust error correction from IBM to date. Two full DRAM failures per rank can be spared and a third full DRAM failure can be corrected. DIMM level failures, including components such as the memory controller application-specific integrated circuit (ASIC), the power regulators, the clocks, and the system board can be corrected. Memory channel failures, such as signal lines, control lines, and drivers/receivers on the SCM, can be corrected. Upstream and

downstream data signals can be spared by using two spare wires on both the upstream and downstream paths. One of these signals can be used to spare a clock signal line (one upstream and one downstream). The following improvements were also added in the z13s servers:

- ▶ No cascading of memory DIMMs
- ▶ Independent channel recovery
- ▶ Double tabs for clock lanes
- ▶ Separate replay buffer per channel
- ▶ Hardware driven lane soft error rate (SER) and sparing.

2.5.2 General z13s RAS features

z13s servers have the following RAS features:

- ▶ z13s servers provide a true $N+1$ (fully redundant) cooling function for the CPC drawers and the PCIe drawers by using the blowers. If a blower fails, the second blower increases the speed of the fans to provide necessary cooling until the failed blower is replaced. The power supplies for the z13s servers are also based on the $N+1$ design. The second power supply can maintain operations and avoid an unplanned outage of the system.
- ▶ The z Systems processors have improved chip packaging (encapsulated chip connectors) and use SER hardened latches throughout the design.
- ▶ z13s servers have $N+2$ point of load (POL) power conversion. This redundancy protects the processor from the loss of voltage because of POL failures.
- ▶ z13s servers have $N+2$ redundancy on the environmental sensors (ambient temperature, relative humidity, air density², and corrosion).
- ▶ Enhanced bus structure using integrated time-domain reflectometry (TDR) technology.
- ▶ z13s servers have these Peripheral Component Interconnect Express (PCIe) service enhancements:
 - Mandatory end-to-end cyclic redundancy check (ECRC)
 - Customer operation code separate from maintenance code
 - Native PCIe firmware stack running on the integrated firmware processor (IFP) to manage isolation and recovery

IBM z13s servers continue to deliver robust server designs through exciting new technologies, hardening both new and classic redundancy.

For more information, see Chapter 9, “Reliability, availability, and serviceability” on page 355.

2.6 Connectivity

Connections to PCIe I/O drawers, I/O drawers, Parallel Sysplex InfiniBand (PSIFB) coupling, and Integrated Coupling Adapters (ICA) are driven from the CPC drawer fanout cards. These fanouts are on the front of the CPC drawer.

² The air density sensor measures air pressure and is used to control blower speed.

Figure 2-23 shows the location of the fanouts for a Model N10. Model N10 has four PCIe fanout slots and two IFB fanout slots. The CPC drawer has two FSPs for system control. *LGXX* is the location code.

Up to 4 PCIe fanouts (LG11 - LG14) and two IFB fanouts (LG09 - LG10) can be installed in the CPC drawer.

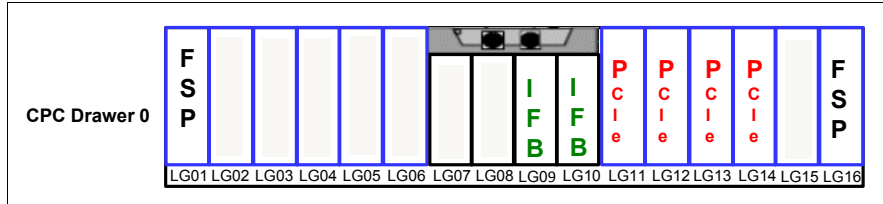


Figure 2-23 Model N10 drawer: Location of the PCIe and IFB fanouts

Figure 2-24 shows the location of the fanouts for a Model N20 with two CPC drawers. Each CPC drawer has two FSPs for system control. *LGXX* is the location code. If the model is a N20 single CPC drawer system, only *CPC drawer 1* is installed.

Up to 8 PCIe fanouts (LG03 - LG06 and LG11 - LG14) and four IFB fanouts (LG07 - LG10) can be installed in each CPC drawer.

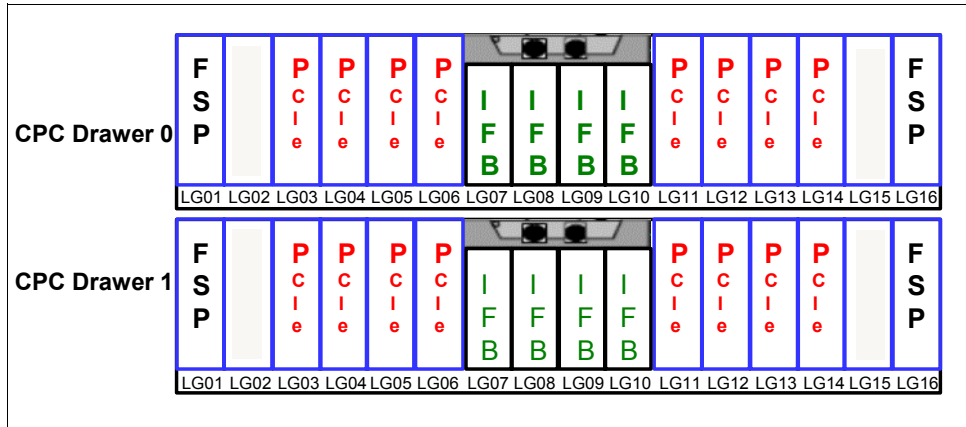


Figure 2-24 Model N20 two CPC drawer: Locations of the PCIe and IFB fanouts

A PCIe Generation 3 fanout connected to a PCIe I/O drawer can be repaired concurrently with the use of a redundant I/O interconnect. For more information, see 2.6.1, “Redundant I/O interconnect” on page 67.

Five types of fanouts are available:

- ▶ PCIe Generation 3 fanout card: This copper fanout provides connectivity to the PCIe switch cards in the PCIe I/O drawer.
- ▶ Host Channel Adapter (HCA2-C): This copper fanout provides connectivity to the IFB-MP cards in the I/O drawer.
- ▶ Integrated Coupling Adapter (ICA SR): This adapter provides coupling connectivity between z13s and z13/z13s servers.
- ▶ Host Channel Adapter (HCA3-O (12xIFB)): This optical fanout provides 12x InfiniBand coupling link connectivity up to 150 meters (492 ft.) distance to z13s, z13, zEC12, zBC12, z196, and z114 servers.

- ▶ Host Channel Adapter (HCA3-O LR (1xIFB)): This optical long reach fanout provides 1x InfiniBand coupling link connectivity up to a 10 km (6.2 miles) unrepeated, or 100 km (62 miles) when extended by using z Systems qualified DWDM equipment) distance to z13s, z13, zEC12, zBC12, z196, and z114 servers.

When you are configuring for availability, balance the channels, coupling links, and OSAs across drawers. In a system that is configured for maximum availability, alternative paths maintain access to critical I/O devices, such as disks and networks.

Note: Fanout plugging rules for z13s servers are different than previous EC12 and z114 servers.

- ▶ Preferred plugging for PCIe Generation 3 fanouts will always be in CPC Drawer 1 (bottom drawer) alternating between the two nodes. Previous systems EC12 and z114 were plugged across two drawers. This configuration is for performance reasons and to maintain RAS characteristics.
- ▶ If the configuration contains two CPC drawers, two PCIe I/O drawers, and PCIe ICA Coupling adapters. All PCIe Generation 3 fanouts are plugged in CPC Drawer 1, and all PCIe ICA coupling adapters are plugged in CPC Drawer 0 (top), alternating between the internal node affinity.

A fanout can be repaired concurrently with the use of redundant I/O interconnect.

2.6.1 Redundant I/O interconnect

Redundancy is provided for both InfiniBand I/O and for PCIe I/O interconnects.

InfiniBand I/O connection

Redundant I/O interconnect is accomplished by the facilities of the InfiniBand I/O connections to the InfiniBand Multiplexer (IFB-MP) card. Each IFB-MP card is connected to a jack in the InfiniBand fanout of a CPC drawer. IFB-MP cards are half-high cards and are interconnected through the I/O drawer backplane. This configuration allows redundant I/O interconnect if the connection coming from a CPC drawer ceases to function. This situation can happen when, for example, a CPC drawer connection to the I/O Drawer is broken for maintenance.

A conceptual view of how redundant I/O interconnect is accomplished is shown in Figure 2-25.

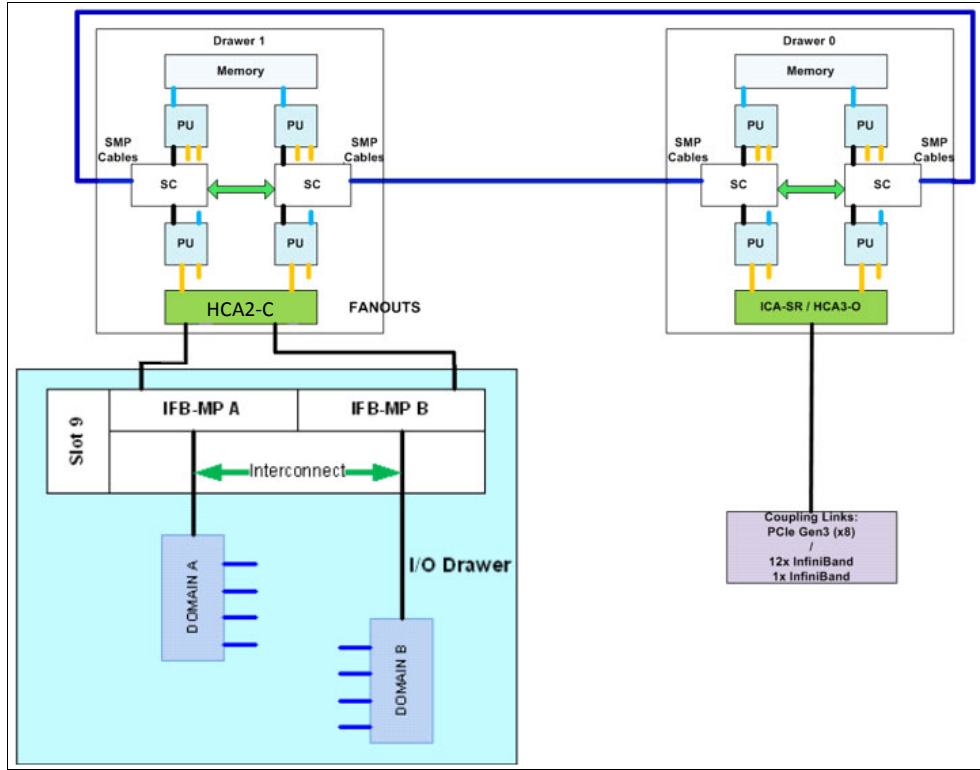


Figure 2-25 Redundant I/O interconnect for I/O drawer

Normally, the HCA2-C fanout in a CPC drawer connects to the IFB-MP (A) card and services domain A in an I/O drawer. In the same fashion, a second HCA2-C fanout of a CPC drawer connects to the IFB-MP (B) card and services domain B in an I/O drawer. If the connection from the CPC drawer IFB-MP (A) to the I/O drawer is removed, connectivity to domain B is maintained. The I/O is guided to domain B through the interconnect between IFB-MP (A) and IFB-MP (B).

Note: Both IFB-MP cards must be installed in the I/O Drawer to maintain the interconnect across I/O domains. If one of the IFB-MP cards is removed, then the I/O cards in that domain (up to four) become unavailable.

PCIe I/O connection

The PCIe I/O drawer supports up to 32 PCIe features. They are organized in four hardware domains per drawer, as shown in Figure 2-26.

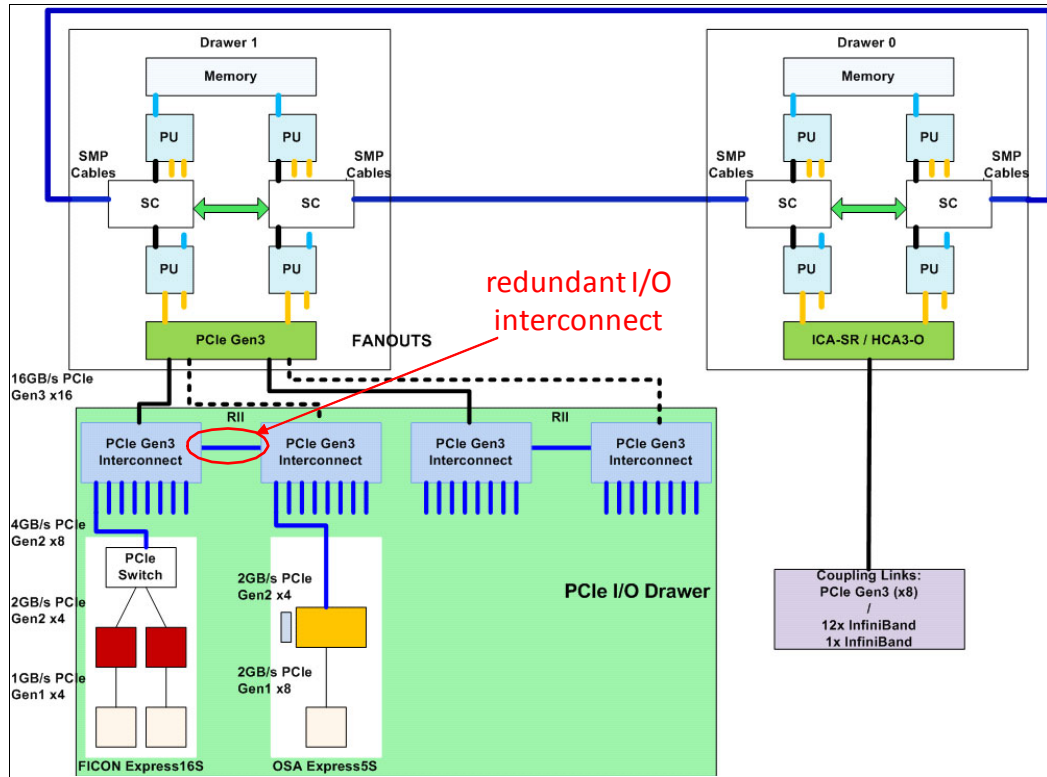


Figure 2-26 Redundant I/O interconnect for PCIe I/O drawer

Each domain is driven through a PCIe switch card. The two PCIe switch cards provide a backup path for each other through the passive connection in the PCIe I/O drawer backplane. During a PCIe fanout or cable failure, all 16 PCIe features in the two domains can be driven through a single PCIe switch card.

To support redundant I/O interconnect (RII) between front to back domain pairs 0,1 and 2,3, the two interconnects to each pair must be driven from two different PCIe fanouts. Normally, each PCIe interconnect in a pair supports the eight features in its domain. In backup operation mode, one PCIe interconnect supports all 16 features in the domain pair.

Note: The PCIe Gen3 Interconnect adapter must be installed in the PCIe Drawer to maintain the interconnect across I/O domains. If the PCIe Gen3 Interconnect adapter is removed, then the I/O cards in that domain (up to eight) become unavailable.

2.7 Model configurations

When a z13s order is configured, PUs are characterized according to their intended usage. They can be ordered as any of the following items:

- ▶ CP: The processor is purchased and activated. CP supports the z/OS, z/VSE, z/VM, z/TPF, and Linux on z Systems operating systems. It can also run Coupling Facility Control Code and IBM zAware code.

- ▶ Capacity marked CP: A processor that is purchased for future use as a CP is marked as available capacity. It is offline and not available for use until an upgrade for the CP is installed. It does not affect software licenses or maintenance charges.
- ▶ IFL: The Integrated Facility for Linux is a processor that is purchased and activated for use by z/VM for Linux guests and Linux on z Systems operating systems. It can also run the IBM zAware code.
- ▶ Unassigned IFL: A processor that is purchased for future use as an IFL. It is offline and cannot be used until an upgrade for the IFL is installed. It does not affect software licenses or maintenance charges.
- ▶ ICF: An internal coupling facility (ICF) processor that is purchased and activated for use by the Coupling Facility Control Code.
- ▶ zIIP: An IBM System z Integrated Information Processor (zIIP) that is purchased and activated to run eligible workloads, such as Java code under the control of a z/OS Java virtual machine (JVM) or z/OS XML System Services, DB2 Distributed Relational Database Architecture (DRDA), or z/OS Communication Server IPsec.
- ▶ Additional SAP: An optional processor that is purchased and activated for use as an SAP.

A minimum of one PU that is characterized as a CP, IFL, or ICF is required per system. The maximum number of CPs is six for IFLs, and 20 for ICFs. The maximum number of zIIPs is always up to twice the number of PUs that are characterized as CPs.

Not all PUs on a model must be characterized.

The following items are present in z13s servers, but they are not part of the PUs that clients purchase and require no characterization:

- ▶ An SAP to be used by the channel subsystem. The number of predefined SAPs depends on the z13s model.
- ▶ One IFP that is used in the support of designated features, such as zEDC and 10GbE RoCE.
- ▶ Two spare PUs, which can transparently assume any characterization during a permanent failure of another PU.

The z13s model is based on the number of PUs that are available for client use in each configuration. The models are summarized in Table 2-10.

Table 2-10 z13s configurations

Model	Drawers / PUs	CPs	IFLs/uIFL	ICFs	zIIPs	Optional SAPs	Std. SAPs	Spares	IFP
N10	1/13	0 - 6	0 - 10	0 - 10	0 - 6	0 - 2	2	0	1
N20	1/26	0 - 6	0 - 20	0 - 20	0 - 12	0 - 3	3	2	1
N20	2/26	0 - 6	0 - 20	0 - 20	0 - 12	0 - 3	3	2	1

A *capacity marker* identifies the number of CPs that have been purchased. This number of purchased CPs is higher than or equal to the number of CPs that is actively used. The capacity marker marks the availability of purchased but unused capacity that is intended to be used as CPs in the future. They usually have this status for software-charging reasons. Unused CPs are not a factor when establishing the millions of service units (MSU) value that is used for charging monthly license charge (MLC) software, or when charged on a per-processor basis.

2.7.1 Upgrades

Concurrent CP, IFL, ICF, zIIP, or SAP upgrades are done within a z13s server. Concurrent upgrades require available PUs, and that extra PUs be installed previously, but not activated.

Spare PUs are used to replace defective PUs. On Model N10, unassigned PUs are used as spares. A fully configured Model N10 does not have any spares. Model N20 always has two dedicated spares.

If an upgrade request cannot be accomplished within the N10 configuration, a hardware upgrade to Model N20 is required. The upgrade involves extra hardware to add the second node in the first drawer, and might force the addition of a second drawer. The upgrade from N10 to N20 is disruptive.

Supported upgrade paths (see Figure 2-27 on page 71):

- ▶ Upgrade to z13s processors from earlier processors:
 - From both the z114 and zBC12 servers
- ▶ Upgrade from z13s servers:
 - Model N10 to N20
 - Model N20 to z13 N30 Radiator-based (air-cooled) only

Note: Memory downgrades are not offered. Model downgrade (removal of a CPC drawer) is not offered, nor supported.

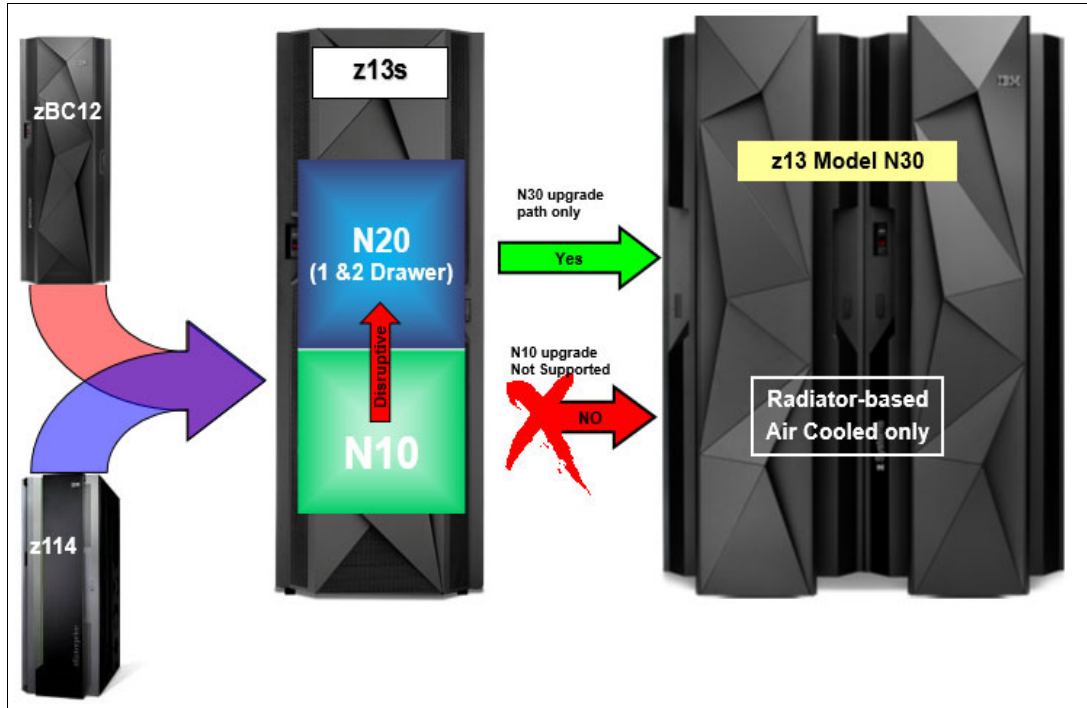


Figure 2-27 z13s upgrade paths

You can upgrade an IBM z114 or a zBC12 (frame roll) preserving the CPC serial number. Features that are common between systems are candidates to move.

Important: Upgrades from z114 and zBC12 are disruptive.

2.7.2 Concurrent PU conversions

Assigned CPs, assigned IFLs, and unassigned IFLs, ICFs, zIIPs, and SAPs can be converted to other assigned or unassigned feature codes.

Most PU conversions are nondisruptive. In exceptional cases, the conversion can be disruptive, such as when a Model N20 with six CPs is converted to an all IFL system. In addition, an LPAR might be disrupted when PUs must be freed before they can be converted. Conversion information is summarized in Table 2-11.

Table 2-11 Concurrent PU conversions

From	To	CP	IFL	Unassigned IFL	ICF	zAAP	zIIP	SAP
CP	-	Yes	Yes	Yes	Yes	Yes	Yes	Yes
IFL	Yes	-	Yes	Yes	Yes	Yes	Yes	Yes
Unassigned IFL	Yes	Yes	-	Yes	Yes	Yes	Yes	Yes
ICF	Yes	Yes	Yes	Yes	-	Yes	Yes	Yes
zAAP	Yes	Yes	Yes	Yes	Yes	-	Yes	Yes
zIIP	Yes	Yes	Yes	Yes	Yes	Yes	-	Yes
SAP	Yes	Yes	Yes	Yes	Yes	Yes	Yes	-

2.7.3 Model capacity identifier

To recognize how many PUs are characterized as CPs, the store system information (STSI) instruction returns a model capacity identifier (MCI). The MCI determines the number and speed of characterized CPs. Characterization of a PU as an IFL, an ICF, or a zIIP is not reflected in the output of the STSI instruction because characterization has no effect on software charging. For more information about STSI output, see “Processor identification” on page 351.

Capacity identifiers: Within a z13s server, all CPs have the same capacity identifier. Specialty engines (IFLs, zAAPs, zIIPs, and ICFs) operate at full speed.

2.7.4 Model capacity identifier and MSU values

All model capacity identifiers have a related MSU value that is used to determine the software license charge for MLC software, as shown in Table 2-12.

Table 2-12 Model capacity identifier and MSU values

Model cap ID	MSU	Model cap ID	MSU	Model cap ID	MSU	Model cap ID	MSU	Model cap ID	MSU	Model cap ID	MSU
A01	10	B01	11	C01	12	D01	14	E01	16	F01	19
A02	19	B02	21	C02	23	D02	26	E02	30	F02	35
A03	27	B03	30	C03	33	D03	38	E03	44	F03	51
A04	35	B04	39	C04	43	D04	48	E04	57	F04	66
A05	42	B05	47	C05	52	D05	59	E05	69	F05	80
A06	50	B06	55	C06	61	D06	68	E06	81	F06	93
G01	21	H01	24	I01	27	J01	30	K01	34	L01	38
G02	40	H02	45	I02	51	J02	57	K02	64	L02	71
G03	58	H03	65	I03	74	J03	82	K03	92	L03	103
G04	75	H04	84	I04	95	J04	206	K04	119	L04	133
G05	92	H05	103	I05	115	J05	128	K05	144	L05	161
G06	107	H06	120	I06	135	J06	150	K06	168	L06	188
M01	42	N01	48	O01	53	P01	60	Q01	67	R01	75
M02	80	N02	90	O02	100	P02	112	Q02	125	R02	140
M03	115	N03	129	O03	144	P03	161	Q03	180	R03	202
M04	148	N04	166	O04	187	P04	208	Q04	232	R04	260
M05	180	N05	202	O05	225	P05	252	Q05	282	R05	315
M06	210	N06	235	O06	264	P06	295	Q06	329	R06	369
S01	83	T01	93	U01	104	V01	117	W01	130	X01	145
S02	156	T02	175	U02	195	V02	218	W02	243	X02	272
S03	224	T03	251	U03	281	V03	316	W03	351	X03	392
S04	290	T04	324	U04	362	V04	406	W04	453	X04	507
S05	352	T05	394	U05	441	V05	493	W05	551	X05	616
S06	412	T06	460	U06	515	V06	576	W06	645	X06	721

Model cap ID	MSU	Model cap ID	MSU	Model cap ID	MSU	Model cap ID	MSU	Model cap ID	MSU	Model cap ID	MSU
Y01	162	Z01	178								
Y02	304	Z02	333								
Y03	439	Z03	481								
Y04	567	Z04	622								
Y05	690	Z05	756								
Y06	806	Z06	884								

A00: Model capacity identifier A00 is used for IFL-only or ICF-only configurations.

2.7.5 Capacity BackUp

The Capacity BackUp (CBU) feature delivers temporary backup capacity in addition to what an installation might have already installed in numbers of assigned CPs, IFLs, ICFs, zIIPs, and optional SAPs.

The following CBU types are available:

- ▶ CBU for CP
- ▶ CBU for IFL
- ▶ CBU for ICF
- ▶ CBU for zIIP
- ▶ Optional SAPs

When CBU for CP is added within the same capacity setting range (indicated by the model capacity indicator) as the currently assigned PUs, the total number of active PUs (the sum of all assigned CPs, IFLs, ICFs, zIIPs, and optional SAPs) plus the number of CBUs cannot exceed the total number of PUs available in the system.

When CBU for CP capacity is acquired by switching from one capacity setting to another, no more CBU can be requested than the total number of PUs available for that capacity setting.

CBU and granular capacity

When CBU for CP is ordered, it replaces lost capacity for disaster recovery. Specialty engines (ICFs, IFLs, and zIIPs) always run at full capacity, and also when running as CBU to replace lost capacity for disaster recovery.

When you order CBU, specify the maximum number of CPs, ICFs, IFLs, zIIPs, and SAPs to be activated for disaster recovery. If disaster strikes, you decide how many of each of the contracted CBUs of each type must be activated. The CBU rights are registered in one or more records on the server. Up to eight records can be active, which can contain several CBU activation variations that apply to the installation.

You can test the CBU. The number of CBU test activations available for no additional fee in each CBU record is determined by the number of years that are purchased with the CBU record. For example, a 3-year CBU record has three test activations, and a 1-year CBU record has one test activation. You can increase the number of tests up to a maximum of 15 for each CBU record.

The real activation of CBU lasts up to 90 days with a grace period of two days to prevent sudden deactivation when the 90-day period expires. The contract duration can be set from one to five years.

The CBU record describes the following properties related to the CBU:

- ▶ Number of CP CBUs that it is possible to activate
- ▶ Number of IFL CBUs that it is possible to activate
- ▶ Number of ICF CBUs that it is possible to activate
- ▶ Number of zIIP CBUs that it is possible to activate
- ▶ Number of SAP CBUs that it is possible to activate
- ▶ Number of additional CBU tests possible for this CBU record
- ▶ Number of total CBU years ordered (duration of the contract)
- ▶ Expiration date of the CBU contract

The record content of the CBU configuration is documented in the IBM configurator output, as shown in Example 2-1. In the example, one CBU record is made for a 5-year CBU contract without additional CBU tests for the activation of one CP CBU.

Example 2-1 Simple CBU record and related configuration features

On Demand Capacity Selections:
 NEW00001 - CBU - CP(1) - Years(5) - Tests(0)
 Expiration(09/10/2013)

Resulting feature numbers in configuration:

6817	Total CBU Years Ordered	5
6818	CBU Records Ordered	1
6820	Single CBU CP-Year	5

In Example 2-2, a second CBU record is added to the same configuration for two CP CBUs, two IFL CBUs, two zAAP CBUs, and two zIIP CBUs, with five additional tests and a 5-year CBU contract. The result is now a total number of 10 years of CBU ordered, which is the standard five years in the first record and an extra five years in the second record.

Two CBU records from which to choose are in the system. Five additional CBU tests have been requested. Because five years have been contracted for a total of 3 CP CBUs, two IFL CBUs, and two zIIP CBUs, they are shown as 15, 10, and 10 CBU years for their respective types.

Example 2-2 Second CBU record and resulting configuration features

NEW00002 - CBU - CP(2) - IFL(2) - zIIP(2)
 Tests(5) - Years(5)

Resulting cumulative feature numbers in configuration:

6817	Total CBU Years Ordered	10
6818	CBU Records Ordered	2
6819	5 Additional CBU Tests	1
6820	Single CBU CP-Year	15
6822	Single CBU IFL-Year	10
6828	Single CBU zIIP-Year	10

CBU for CP rules

Consider the following guidelines when planning for CBU for CP capacity:

- ▶ The total CBU CP capacity features are equal to the number of added CPs plus the number of permanent CPs changing capacity level. For example, if two CBU CPs are added to the current model D03, and the capacity level does not change, the D03 becomes D05: (D03 + 2 = D05).

If the capacity level changes to an E06, the numbers of additional CPs (3) are added to the three CPs of the D03, resulting in a total number of CBU CP capacity features of six: (3 + 3 = 6).

- ▶ The CBU cannot decrease the number of CPs.
- ▶ The CBU cannot lower the capacity setting.

On/Off Capacity on Demand: Activation of CBU for CPs, IFLs, ICFs, zIIPs, and SAPs can be activated together with On/Off Capacity on Demand (CoD) temporary upgrades. Both facilities can be implemented on one system, and can be activated simultaneously.

CBU for specialty engines

Specialty engines (ICFs, IFLs, and zIIPs) run at full capacity for all capacity settings. This characteristic also applies to CBU for specialty engines. Table 2-13 shows the minimum and maximum (min-max) numbers of all types of CBUs that can be activated on each of the models. The CBU record can contain larger numbers of CBUs than can fit in the current model.

Table 2-13 Capacity Backup matrix

Model	Total PUs available	CBU CPs min - max	CBU IFLs min - max	CBU ICFs min - max	CBU zIIPs min - max	CBU SAPs min - max
Model N10	13	0 - 6	0 - 10	0 - 10	0 - 6	0 - 2
Model N20	26	0 - 6	0 - 20	0 - 20	0 - 12	0 - 3

2.7.6 On/Off Capacity on Demand and CPs

On/Off CoD provides temporary capacity for all types of characterized PUs. Relative to granular capacity, On/Off CoD for CPs is treated similarly to the way CBU is handled.

On/Off CoD and granular capacity

When temporary capacity requested by On/Off CoD for CPs matches the model capacity identifier range of the permanent CP feature, the total number of active CP equals the sum of the number of permanent CPs plus the number of temporary CPs ordered. For example, when a model capacity identifier D03 has two CPs added temporarily, it becomes a model capacity identifier D05.

When the addition of temporary capacity requested by On/Off CoD for CPs results in a cross-over from one capacity identifier range to another, the total number of CPs active when the temporary CPs are activated is equal to the number of temporary CPs ordered. For example, when a configuration with model capacity identifier D03 specifies four temporary CPs through On/Off CoD, the result is a server with model capacity identifier E05.

A cross-over does not necessarily mean that the CP count for the additional temporary capacity increases. The same D03 can temporarily be upgraded to a server with model capacity identifier F03. In this case, the number of CPs does not increase, but more temporary capacity is achieved.

On/Off CoD guidelines

When you request temporary capacity, consider the following guidelines:

- ▶ Temporary capacity must be greater than permanent capacity.
- ▶ Temporary capacity cannot be more than double the purchased capacity.
- ▶ On/Off CoD cannot decrease the number of engines on the server.
- ▶ Adding more engines than are currently installed is not possible.

For more information about temporary capacity increases, see Chapter 8, “System upgrades” on page 311.

2.8 Power and cooling

As environmental concerns increase the focus on energy consumption, z13s servers offer a holistic focus on the environment. New efficiencies and functions, such as energy optimizer advisor, enable a dramatic reduction of energy usage and floor space when consolidating workloads from distributed servers.

The power service specifications for the zEnterprise CPCs are the same as their particular predecessors, but the power consumption is more efficient. A fully loaded z13s CPC maximum power consumption is nearly the same as a zBC12. However, with a maximum performance ratio of 1.58:1, it has a much higher exploitation on the same footprint.

2.8.1 Power considerations

The z Systems CPCs operate with two redundant power supplies. Each power supply has an individual power cord for the z13s server.

For redundancy, the servers have two power feeds. Power cords attach either 50/60 hertz (Hz), alternating current (AC) power single-phase³ 200 - 415 volt (v) or three-phase 200 - 480 V AC power, or 380 - 520 V direct current (DC) power. The total loss of one power feed has no effect on system operation.

A Balanced Power Plan Ahead feature is available for future growth that helps ensure adequate and balanced power with AC power cord selection by using three-phase power. With this feature, downtimes for upgrading a server are eliminated by including the maximum power requirements in terms of bulk power regulators (BPRs) and power cords to your installation. For ancillary equipment, such as the HMC, its display, and its modem, extra single-phase outlets are required.

The power requirements depend on the number of processor and I/O drawers in the z13s server. Table 10-1 on page 372 shows the maximum power consumption tables for the various configurations and environments.

z13s servers can operate in raised floor and non-raised-floor environments. For both types of installation, an overhead power cable option for the top exit of the cables is available. For a

³ Only available on select z13s configurations, see Table 10-1 on page 372.

non-raised floor environment, the *Top Exit Power* and *Top Exit I/O Cabling* features are mandatory.

2.8.2 High-voltage DC power

In data centers today, many businesses pay increasingly expensive electric bills, and are running out of power. The IBM z Systems CPC High Voltage Direct Current power feature adds nominal 380 - 520 volt DC input power capability to the existing AC power capability. It enables CPCs to directly use the high voltage (HV) DC distribution in new, green data centers. A direct HV DC data center power design can improve data center energy efficiency by removing the need for a DC to AC inversion step.

The IBM z Systems CPCs bulk power supplies have been modified to support HV DC, so the only difference in shipped hardware to implement the option is the DC power cords. Because HV DC is a new technology, there are multiple proposed standards.

The z13s CPC supports both ground-referenced and dual-polarity HV DC supplies, such as +/- 190 V or +/- 260 V, or +380 V, and other supplies. Beyond the data center, uninterruptible power supply (UPS), and power distribution energy savings, a zEnterprise CPC run on HV DC power draws 1 - 3% less input power. HV DC does not change the number of power cords that a system requires.

2.8.3 Internal Battery Feature

IBF is an optional feature on the z13s CPC server. See Figure 2-1 on page 36 for a pictorial view of the location of this feature. This optional IBF provides the function of a local uninterrupted power source (UPS).

The IBF further enhances the robustness of the power design, increasing power line disturbance immunity. It provides battery power to preserve processor data during a loss of power on all four AC feeds from the utility company. The IBF can hold power briefly during a brownout, or for orderly shutdown during a longer outage. The values for the holdup time depend on the I/O configuration, as shown in Table 10-2 on page 372.

2.8.4 Power capping

Power capping limits the maximum power consumption and reduces the cooling requirements for the zBX. z13s servers do not support power capping.

The z13s system has a mechanism to vary the frequency and voltage that it supplies its chips. The mechanism can be used to reduce the energy consumption of the system in periods of low usage or for systems that are designed mainly for disaster recovery. The mechanism is under the full control of the client. The client controls are implemented in the HMC, SE, and Active Energy Manager where you can choose between High Performance (default) or Low Power (power saving). The expectation is that the frequency change is 20%, the voltage change is 9%, and the total power savings is 6% - 16%, depending on the configuration.

2.8.5 Power Estimation tool

The Power Estimation tool for the z Systems CPCs enables you to enter your precise server configuration to produce an estimate of power consumption:

1. Log in to Resource Link with any user ID.

2. Click **Planning** → **Tools** → **Power Estimation Tools**.
3. Specify the quantity for the features that are installed in your system.

This tool estimates the power consumption for the specified configuration. The tool does *not* verify that the specified configuration can be physically built.

Power consumption: The exact power consumption for your system will vary. The object of the tool is to produce an *estimation* of the power requirements to aid you in planning for your system installation. Actual power consumption after installation can be confirmed on the HMC System Activity Display.

2.8.6 Cooling requirements

The z13s is an air-cooled system. It requires chilled air, ideally coming from under a raised floor, to fulfill the air cooling requirements. The chilled air is usually provided through perforated floor tiles. The amount of chilled air that is required for various temperatures under the floor of the computer room is indicated in the *z13s Installation Manual for Physical Planning*, GC28-6953.

2.9 Summary of z13s structure

Table 2-14 summarizes all aspects of the z13s structure.

Table 2-14 System structure summary

Description	Model N10	Model N20 - 1	Model N20 - 2
Number of PU SCMs	2	4	8
Number of SC SCMs	1	2	4
Total number of PUs	13	26	26
Maximum number of characterized PUs	10	20	20
Number of CPs	0 - 6	0 - 6	0-6
Number of IFLs	0 - 10	0 - 20	0-20
Number of ICFs	0 - 10	0 - 20	0-20
Number of zIIPs	0 - 6	0 - 12	0-12
Standard SAPs	2	3	3
Additional SAPs	0-2	0-3	0-3
Standard spare PUs	0	2	2
Enabled memory sizes	64 - 984 GB	64 - 2008 GB	64 - 4056 GB
L1 cache per PU	96-I / 128-D KB	96-I / 128-D KB	96-I / 128-D KB
L2 cache per PU	2-I / 2-D MB	2-I / 2-D MB	2-I / 2-D MB
L3 shared cache per PU chip	64 MB	64 MB	64 MB
L4 shared cache	480 MB	960 MB	1920 MB
Cycle time (ns)	0.233	0.233	0.233

Description	Model N10	Model N20 - 1	Model N20 - 2
Clock frequency	4.3 GHz	4.3 GHz	4.3 GHz
Maximum number of PCIe GEN3 fanouts	4	8	16
Maximum number of InfiniBand fanouts	2	4	8
Interconnect Bus - InfiniBand cable	6 GBps	6 GBps	6 GBps
Interconnect Bus - PCIe cable	16 GBps	16 GBps	16 GBps
Number of Support Elements	2	2	2
External AC power	1 phase, 3 phase	1 phase, 3 phase	1 phase, 3 phase
Optional external DC	520 V / 380 V DC	520 V / 380 V DC	520 V / 380 V DC
Internal Battery Feature	Optional	Optional	Optional



Central processor complex system design

This chapter explains some of the design considerations for the IBM z13s processor unit. This information can be used to understand the functions that make the z13s server a system that accommodates a broad mix of workloads for large enterprises.

This chapter includes the following sections:

- ▶ Overview
- ▶ Design highlights
- ▶ CPC drawer design
- ▶ Processor unit design
- ▶ Processor unit functions
- ▶ Memory design
- ▶ Logical partitioning
- ▶ Intelligent Resource Director
- ▶ Clustering technology

3.1 Overview

The z13s symmetric multiprocessor (SMP) system is the next step in an evolutionary trajectory that began with the introduction of the IBM System/360 in 1964. Over time, the design was adapted to the changing requirements that were dictated by the shift toward new types of applications that clients depend on.

z13s servers offer high levels of reliability, availability, serviceability, resilience, and security. It fits into the IBM strategy in which mainframes play a central role in creating secure cloud-enabled, cognitive IT solutions. z13s servers are designed so that everything around them, such as operating systems, middleware, storage, security, and network technologies that support open standards, help you achieve your business goals.

The modular CPC drawer design aims to reduce, or in some cases even eliminate, planned and unplanned outages. The design does so by offering advanced self-healing, concurrent repair, replace, and upgrade functions. For more information about the z13s reliability, availability, and serviceability (RAS) features, see Chapter 9, “Reliability, availability, and serviceability” on page 355.

z13s servers have ultra-high frequency, large high-speed buffers (caches) and memory, superscalar processor design, out-of-order core execution, simultaneous multithreading (SMT), single instruction multiple data (SIMD), and flexible configuration options. It is the next implementation of z Systems to address the ever-changing IT environment.

The CPC drawer packaging is derived from the z13. The PU single chip modules (SCMs) are air cooled (rather than water cooled like the z13) because of lower power and cooling requirements for the z13s SCMs.

z13s servers have several microprocessor improvements over zBC12 servers:

- ▶ Higher instruction execution parallelism (improved out-of-order (OOO) instruction handling)
- ▶ Two vector execution units (SIMD)
- ▶ SMT architecture that supports concurrent execution of two threads
- ▶ Eight-core processor chip (6 or 7 cores per SCM enabled)
- ▶ A robust cache hierarchy

3.2 Design highlights

The physical packaging of the z Systems z13s server is different from zEnterprise zBC12 (zBC12) systems. Its modular CPC drawer and SCM design address the increasing costs that are related to building systems with ever-increasing capacities. The modular CPC drawer design is flexible and expandable, offering unprecedented capacity to meet consolidation needs.

z13s servers continue the line of mainframe processors that are compatible with an earlier version. It introduces more complex instructions that are run by millicode, and more complex instructions that are broken down into multiple operations. It uses 24-bit, 31-bit, and 64-bit addressing modes, multiple arithmetic formats, and multiple address spaces for robust interprocess security.

The z13s system design has the following main objectives:

- ▶ Offer a flexible infrastructure to concurrently accommodate a wide range of operating systems and applications, which range from the traditional systems (for example, z/OS and z/VM) to the world of Linux, cloud, analytics, and mobile computing.
- ▶ Offer state-of-the-art integration capability for server consolidation by using virtualization capabilities in a highly secure environment:
 - Logical partitioning, which allows for up to 40 independent logical servers.
 - z/VM, which can virtualize hundreds to thousands of servers as independently running virtual machines (guests).
 - HiperSockets, which implement virtual LANs between logical partitions (LPARs) within the system.
 - The z Systems PR/SM is designed for Common Criteria Evaluation Assurance Level 5+ (EAL 5+) certification for security. Therefore, an application running on one partition (LPAR) cannot access another application on a different partition, providing essentially the same security as an air-gapped system.

This configuration allows for a logical and virtual server coexistence and maximizes system utilization and efficiency by sharing hardware resources.

- ▶ Offer high performance computing to achieve the outstanding response times that are required by new workload-type applications. This performance is achieved by high frequency, enhanced superscalar processor technology, out-of-order core execution, large high-speed buffers (cache) and memory, an architecture with multiple complex instructions, and high-bandwidth channels.
- ▶ Offer the high capacity and scalability that are required by the most demanding applications, both from the single-system and clustered-systems points of view.
- ▶ Offer the capability of concurrent upgrades for processors, memory, and I/O connectivity, avoiding system outages in planned situations.
- ▶ Implement a system with high availability and reliability. These goals are achieved with the redundancy of critical elements and sparing components of a single system, and the clustering technology of the Parallel Sysplex environment.
- ▶ Have internal and external connectivity offerings, supporting open standards such as 10 Gigabit Ethernet (10 GbE) and Fibre Channel Protocol (FCP).
- ▶ Provide leading cryptographic performance. Every processor unit (PU) has a dedicated and optimized CP Assist for Cryptographic Function (CPACF). Optional Crypto Express features with cryptographic coprocessors provide the highest standardized security certification.¹ These optional features can also be configured as Cryptographic Accelerators to enhance the performance of Secure Sockets Layer/Transport Layer Security (SSL/TLS) transactions.
- ▶ Be self-managing and self-optimizing, adjusting itself when the workload changes to achieve the best system throughput. This process can be done through the Intelligent Resource Director or the Workload Manager functions, which are assisted by HiperDispatch.
- ▶ Have a balanced system design, providing large data rate bandwidths for high performance connectivity along with processor and system capacity.

The remaining sections describe the z13s system structure, showing a logical representation of the data flow from PUs, caches, memory cards, and various interconnect capabilities.

¹ Federal Information Processing Standard (FIPS) 140-2 Security Requirements for Cryptographic Modules

3.3 CPC drawer design

A z13s system can have up to two CPC drawers in a full configuration, up to 20 PUs can be characterized, and up to 4 TB of customer usable memory capacity can be ordered. Each CPC drawer is physically divided into two nodes to improve the processor and memory affinity and availability. The CPC drawer topology is shown in Figure 3-5 on page 88.

Memory has up to four memory controllers that use 5-channel redundant array of independent memory (RAIM) protection, with dual inline memory modules (DIMM) bus cyclic redundancy check (CRC) error retry. The 4-level cache hierarchy is implemented with eDRAM (embedded) caches. Until recently, eDRAM was considered to be too slow for this use. However, a breakthrough in technology by IBM has removed that limitation. In addition, eDRAM offers higher density, less power utilization, fewer soft errors, and better performance.

z13s servers use CMOS 14S0 Silicon-On-Insulator (SOI) 22 nm chip technology, with advanced low latency pipeline design, creating high-speed yet power-efficient circuit designs. The PU SCM has a dense packaging, but due to lower cycle frequency can be air cooled.

3.3.1 Cache levels and memory structure

The z13s memory subsystem focuses on keeping data “closer” to the PU core. With the current processor configuration, all cache levels have increased, and the second-level private cache (L2) and the total node-level shared cache (L4 including NIC directory) in a central processor complex (CPC) drawer have doubled in size.

Figure 3-1 shows the z13s cache levels and memory hierarchy.

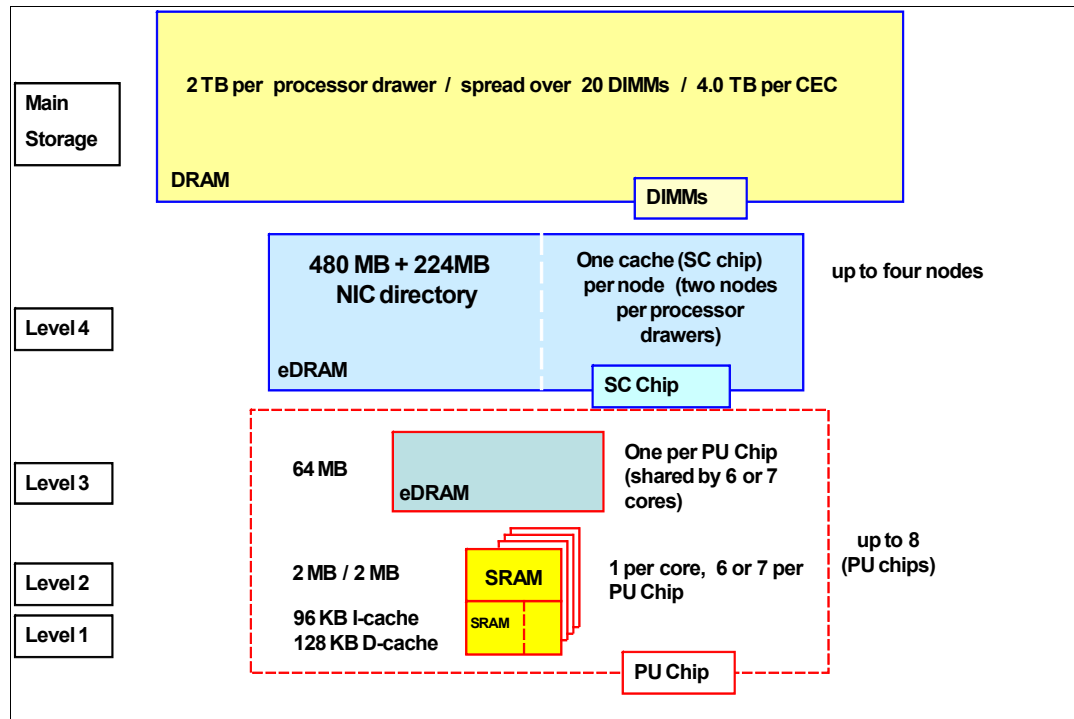


Figure 3-1 z13s cache levels and memory hierarchy

The 4-level cache structure is implemented within the PU and Storage Control (SC) SCMs. Each node of the CPC drawer has one L4 cache. The first three levels (L1, L2, and L3) are on each PU chip (PU SCM), and L4 is on the SC SCMs:

- ▶ L1 and L2 caches use static random-access memory (SRAM), and are private for each core.
- ▶ The L3 cache uses embedded dynamic SRAM (eDRAM) and is shared by all active cores (six or seven) within the PU chip. Each node in the CPC drawer has two L3 caches. A z13s Model N20 with two CPC drawers therefore has eight L3 caches, resulting in 512 MB (8 x 64 MB) of shared PU chip-level cache.
- ▶ L4 cache also uses eDRAM, and is shared by all PU chips on the node of a CPC drawer. Each L4 cache has 480 MB for previously owned and some L3-owned lines least recently used (LRU) and 224 MB for a non-data inclusive coherent directory that points to L3 owned lines that have not been included in L4 cache. A z13s Model N20 with two CPC drawers has 1920 MB (2 x 2 x 384 MB) of shared L4 cache and 896 MB (2 x 2 x 224 MB) of NIC directory.
- ▶ Main storage has up to 2.0 TB addressable memory per CPC drawer, using 20 DIMMs (total of 5 per feature). A z13s Model N20 with two CPC drawers can have up to 4 TB of addressable main storage.

Considerations

Cache sizes are being limited by ever-diminishing cycle times because they must respond quickly without creating bottlenecks. Access to large caches costs more cycles. Instruction and data cache (L1) sizes must be limited because larger distances must be traveled to reach long cache lines. This L1 access time generally occurs in one cycle, which prevents increased latency.

Also, the distance to remote caches as seen from the microprocessor becomes a significant factor. An example is the L4 cache that is not on the microprocessor (and might not even be in the same CPC drawer). Although the L4 cache is rather large, several cycles are needed to travel the distance to the cache. Figure 3-2 shows the node-cache topology of z13s servers.

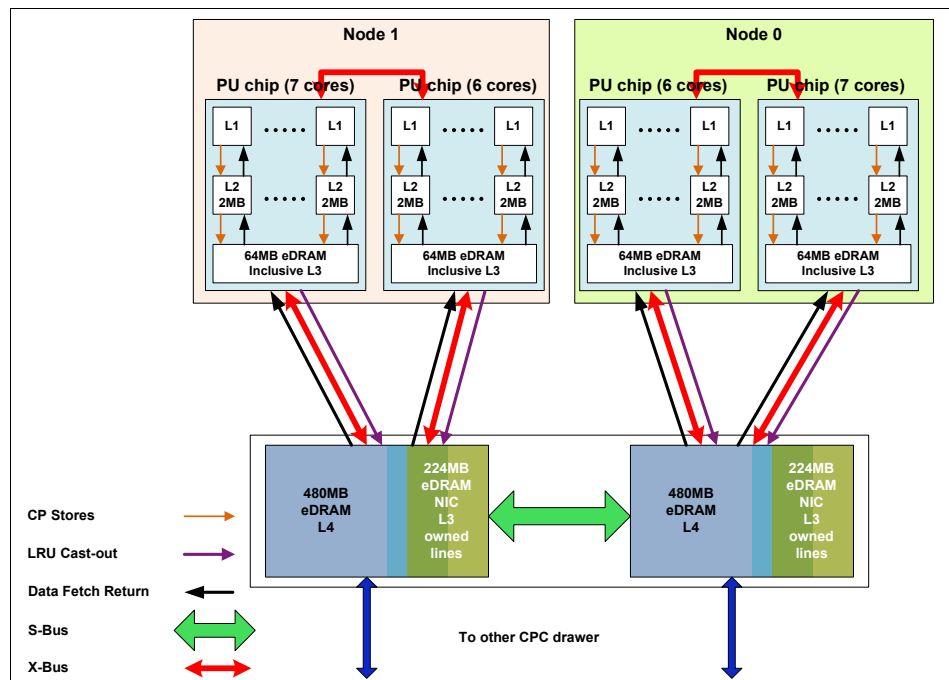


Figure 3-2 z13s cache topology

Although large caches mean increased access latency, the new technology of CMOS 14S0 (22 nm chip lithography) and the lower cycle time allow z13s servers to increase the size of cache levels (L1, L2, and L3) within the PU chip by using denser packaging. This design reduces traffic to and from the shared L4 cache, which is on another chip (SC chip). Only when a cache miss in L1, L2, or L3 occurs is a request sent to L4. L4 is the coherence manager, which means that all memory fetches must be in the L4 cache before that data can be used by the processor.

However, in the z13s cache design, some lines of the L3 cache are not included in the L4 cache. The L4 cache has a NIC directory that has entries that point to the non-inclusive lines of L3 cache. This design ensures that L3 locally owned lines (same node) can be accessed over the X-bus by using the intra-node snoop interface without being included in L4. Inter-node snoop traffic to L4 can still be handled effectively.

Another approach is available for avoiding L4 cache access delays (latency). The L4 cache straddles up to two CPC drawers and up to four nodes. This configuration means that relatively long distances exist between the higher-level caches in the processors and the L4 cache content. To overcome the delays that are inherent in the SMP CPC drawer design and save cycles to access the remote L4 content, keep instructions and data as close to the processors as possible. You can do so by directing as much work of a particular LPAR workload to the processors in the same CPC drawer as the L4 cache. This configuration is achieved by having the IBM Processor Resource/Systems Manager (PR/SM) scheduler and the z/OS Workload Manager (WLM) and dispatcher work together. Have them keep as much work as possible within the boundaries of as few processors and L4 cache space (which is best within a node of a CPC drawer boundary) without affecting throughput and response times.

Figure 3-3 compares the cache structures of z13s servers with the previous generation of z Systems, zBC12 servers.

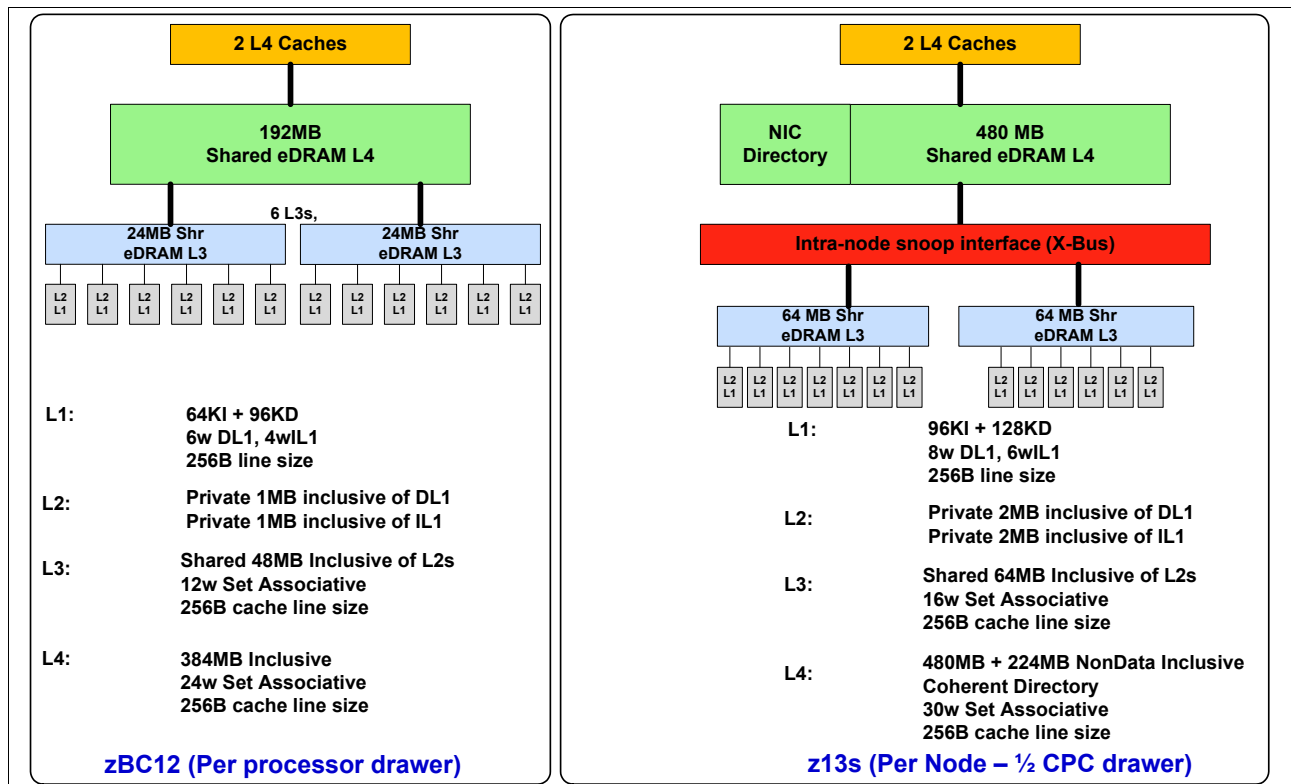


Figure 3-3 z13s and zBC12 cache level comparison

Compared to zBC12, the z13s cache design has much larger cache level sizes. z13s servers have more affinity between the memory of a partition, the L4 cache in the SC SCM, and the cores in the PU SCMs of a node. The access time of the private cache usually occurs in one cycle. The z13 cache level structure is focused on keeping more data closer to the PU. This design can improve system performance on many production workloads.

HiperDispatch

To help avoid latency in a high-frequency processor design, such as z13s servers, prevent PR/SM and the dispatcher from scheduling and dispatching a workload on any processor available. Also, keep the workload in as small a portion of the system as possible. The cooperation between z/OS and PR/SM is bundled in a function called *HiperDispatch*. HiperDispatch uses the z13s cache topology, which has reduced cross-node “help” and better locality for multi-task address spaces.

PR/SM can use dynamic PU reassignment to move processors (CPs, zIIPs, IFLs, ICFs, SAPs, and spares) to a different chip, node, and drawer to improve the reuse of shared caches by processors of the same partition. It can use dynamic memory relocation (DMR) to move a running partition’s memory to different physical memory to improve the affinity and reduce the distance between the memory of a partition and the processors of the partition. For more information about HiperDispatch, see 3.7, “Logical partitioning” on page 116.

3.3.2 CPC drawer interconnect topology

CPC drawers are interconnected in a point-to-point topology, allowing a node in a CPC drawer to communicate with every other node (four nodes in two CPC drawers). Data transfer does not always have to go through another node or CPC drawer (cache) to address the requested data or control information.

Figure 3-4 shows the z13s CPC drawer communication structure (intra- and inter- drawers).

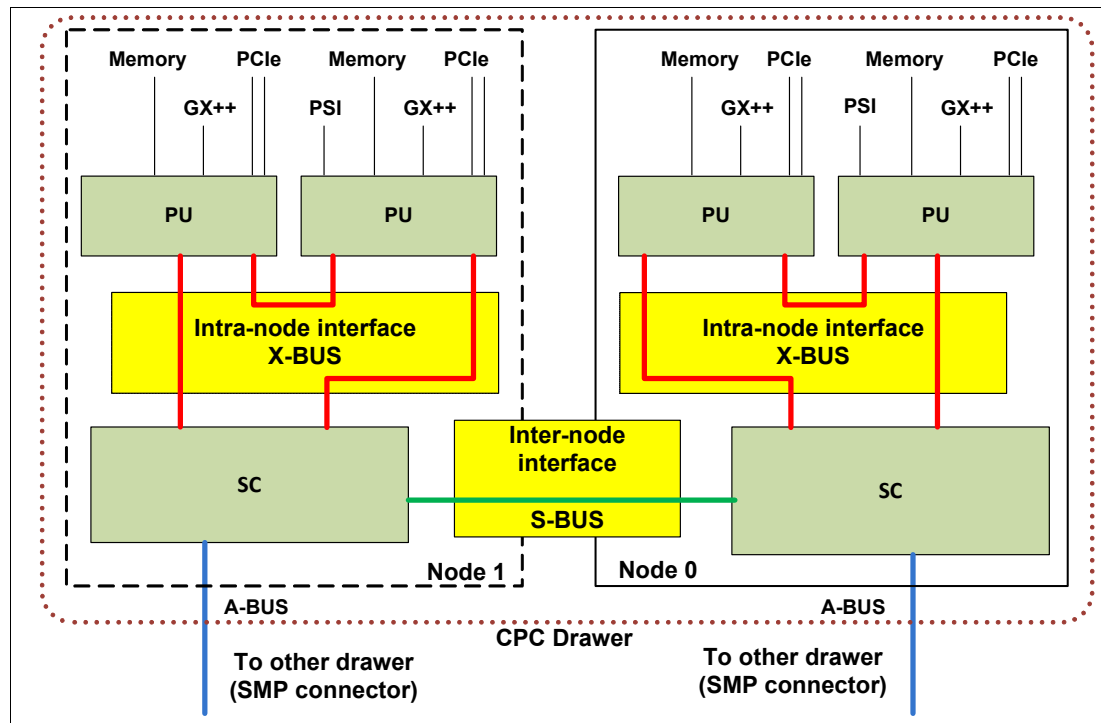


Figure 3-4 z13s CPC drawer communication topology

Figure 3-5 shows a simplified topology of a z13s two CPC drawer system.

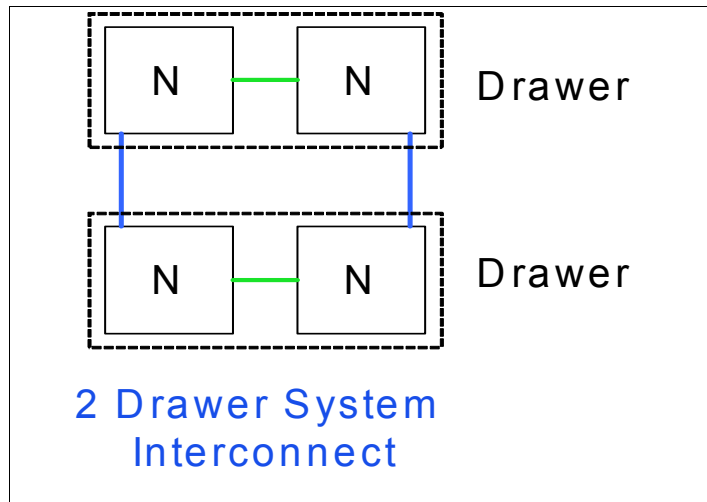


Figure 3-5 Point-to-point topology z13s two CPC drawers communication

Inter-CPC drawer communication takes place at the L4 cache level, which is implemented on SC SCMs in each node. The SC function regulates coherent node-to-node traffic.

3.4 Processor unit design

Processor cycle time is especially important for processor-intensive applications. Current systems design is driven by processor cycle time, although improved cycle time does not automatically mean that the performance characteristics of the system improve. The System z10 BC introduced a dramatic PU cycle time improvement. Its succeeding generations, the z114 and the zBC12, reduced the cycle time even further, reaching 0.263 ns (3.8 GHz) and 0.238 ns (4.2 GHz).

z13s servers have a cycle time of 0.232 ns (4.3 GHz), and an improved design that allows the increased number of processors that share larger caches to have quick access times and improved capacity and performance. Although the cycle time of the z13s processor has only slightly decreased compared to zBC12 (2.4%), the processor performance was increased much more through improved processor design, such as redesigned pipeline and out-of-order execution, branch prediction, time of access to high-speed buffers (caches redesign), and the relative nest intensity (RNI). For more information about RNI, see 12.4, “Relative nest intensity” on page 441.

The z13s processor unit core is a superscalar, OOO, SMT processor with 10 execution units (compared to six for zBC12). For instructions that are not directly run by the hardware, some are run by millicode, and others are split into multiple operations.

The z13s system introduces architectural extensions with instructions that are designed to allow reduced processor quiesce effects, reduced cache misses, reduced pipeline disruption, and increased parallelism with instructions that process several operands in a single instruction (SIMD). The new z13s architecture includes the following features:

- ▶ SMT
- ▶ SIMD instructions set
- ▶ Improved OOO core execution
- ▶ Improvements in branch prediction and handling

- ▶ Performance per watt improvements when compared to the zBC12 system
- ▶ Enhanced instruction dispatch and grouping efficiency
- ▶ Enhanced branch prediction structure and sequential instruction fetching
- ▶ Millicode improvements
- ▶ Decimal floating-point (DFP) improvements

The z13s enhanced Instruction Set Architecture (ISA) includes a set of instructions that are added to improve compiled code efficiency. These instructions optimize PUs to meet the demands of a wide variety of business and analytics workload types without compromising the performance characteristics of traditional workloads.

3.4.1 Simultaneous multithreading

z13s servers, aligned with industry directions, can process up to two simultaneous threads in a single core while sharing certain resources of the processor, such as execution units, translation lookaside buffers (TLBs), and caches. When one thread in the core is waiting for other hardware resources, the second thread in the core can use the shared resources rather than remaining idle. This capability is known as SMT.

SMT is only supported by Integrated Facility for Linux (IFL) and IBM z Systems Integrated Information Processor (zIIP) speciality engines on z13s servers, and it requires operating system support. An operating system with SMT support can be configured to dispatch work to a thread on a zIIP (for eligible workloads in z/OS) or an IFL (for z/VM) core in single thread or SMT mode so that HiperDispatch cache optimization can be considered. For more information about operating system support, see Chapter 7, “Software support” on page 229.

To support SMT, z13s servers have a double symmetric instruction pipeline width and full architectural state per thread. Beyond this, the CPU address changes and the 16-bit CPU ID consist of 15-bit core ID and a 1-bit thread ID. For example, the CPU ID 6 (b'0000000000000110') means core 3 thread 0 and the CPU ID 7 (b'0000000000000111') means core 3 thread 1. For CPs, only thread 0 is used in each core.

SMT technology allows instructions from more than one thread to run in any pipeline stage at a time. Each thread has its own unique state information, such as PSW and registers. The simultaneous threads cannot necessarily run instructions instantly and must at times compete to use certain core resources that are shared between the threads. In some cases, threads can use shared resources that are not experiencing competition.

Figure 3-6 shows two threads (A and B) running on the same processor core on different pipeline stages, sharing the core resources.

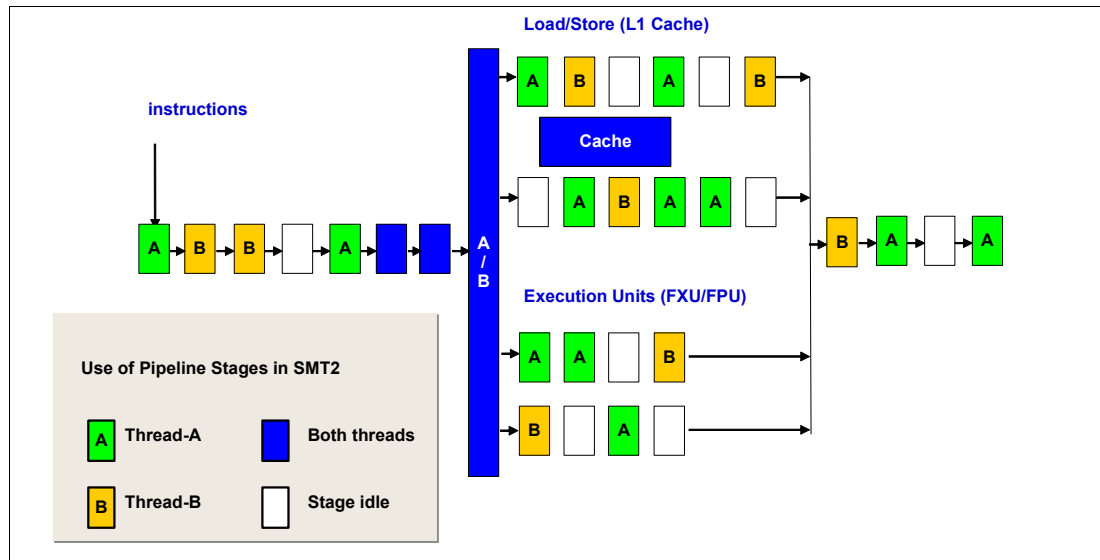


Figure 3-6 Two threads running simultaneously on the same processor core

The use of SMT provides more efficient use of the processors' resources and helps address memory latency, resulting in overall throughput gains. The active thread shares core resources in space, such as data and instruction caches, TLBs, branch history tables, and, in time, pipeline slots, execution units, and address translators.

Although SMT increases the processing capacity, the performance in some cases might be superior if you use a single thread. Enhanced hardware monitoring supports measurement through CPU Measurement Facility (CPUMF) for thread usage and capacity.

For workloads that need maximum thread speed, the partition's SMT mode can be turned off. For workloads that need more throughput to decrease the dispatch queue size, the partition's SMT mode can be turned on.

The SMT exploitation is functionally transparent to middleware and applications, and no changes are required to run them in an SMT-enabled partition.

3.4.2 Single-instruction multiple-data

The z13s superscalar processor has 32 vector registers and an instruction set architecture that includes a subset of 139 new instructions, known as SIMD, added to improve the efficiency of complex mathematical models and vector processing. These new instructions allow a larger number of operands to be processed with a single instruction. The SIMD instructions use the superscalar core to process operands in parallel.

SIMD provides the next phase of enhancements of z Systems analytics capability. The set of SIMD instructions are a type of data parallel computing and vector processing that can decrease the amount of code and accelerate code that handles integer, string, character, and floating point data types. The SIMD instructions improve performance of complex mathematical models and allow integration of business transactions and analytic workloads on z Systems.

The 32 new vector registers have 128 bits. The 139 new instructions include string operations, vector integer, and vector floating point operations. Each register contains multiple data elements of a fixed size. The instructions code specifies which data format to use and the size of the elements:

- ▶ Byte (sixteen 8-bit operands)
- ▶ Halfword (eight 16-bit operands)
- ▶ Word (four 32-bit operands)
- ▶ Doubleword (two 64-bit operands)
- ▶ Quadword (one 128-bit operand)

The collection of elements in a register is called a *vector*. A single instruction operates on all of the elements in the register. Instructions have a non-destructive operand encoding that allows the addition of the register vector A and register vector B and stores the result in the register vector A ($A = A + B$).

Figure 3-7 shows a schematic representation of a SIMD instruction with 16-byte size elements in each vector operand.

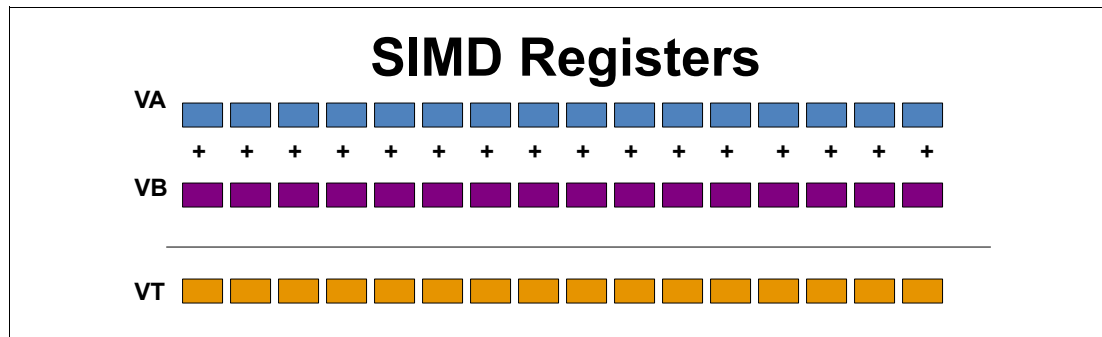


Figure 3-7 Schematic representation of add SIMD instruction with 16 elements in each vector

The vector register file overlays the floating-point registers (FPRs), as shown in Figure 3-8. The FPRs use the first 64 bits of the first 16 vector registers, which saves hardware area and power, and makes it easier to mix scalar and SIMD codes. Effectively, the core gets 64 FPRs, which can further improve FP code efficiency.

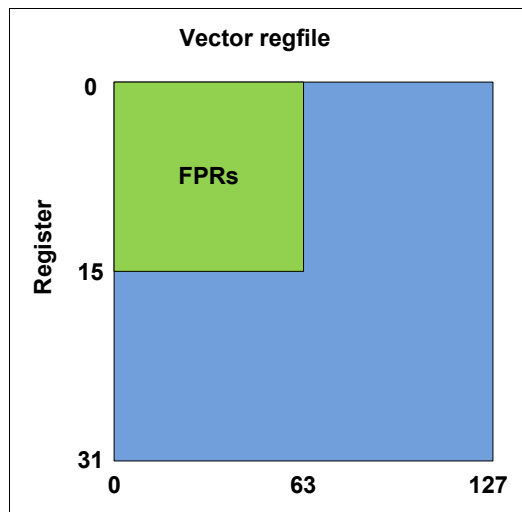


Figure 3-8 Floating point registers overlaid by vector registers

Here are some examples of SIMD instructions:

- ▶ Integer byte to quadword add, sub, and compare
- ▶ Integer byte to doubleword min, max, and average
- ▶ Integer byte to word multiply
- ▶ String find 8-bits, 16-bits, and 32-bits
- ▶ String range compare
- ▶ String find any equal
- ▶ String load to block boundaries and load/store with length

For most operations, the condition code is not set. A summary condition code is used only for a few instructions.

3.4.3 Out-of-order execution

z13s servers have an OOO core, much like the previous z114 and zBC12 systems. OOO yields significant performance benefits for compute-intensive applications. It does so by reordering instruction execution, allowing later (younger) instructions to be run ahead of a stalled instruction, and reordering storage accesses and parallel storage accesses. OOO maintains good performance growth for traditional applications. Out-of-order execution can improve performance in the following ways:

- ▶ Reordering instruction execution: Instructions stall in a pipeline because they are waiting for results from a previous instruction or the execution resource that they require is busy. In an in-order core, this stalled instruction stalls all later instructions in the code stream. In an out-of-order core, later instructions are allowed to run ahead of the stalled instruction.
- ▶ Reordering storage accesses: Instructions that access storage can stall because they are waiting on results that are needed to compute the storage address. In an in-order core, later instructions are stalled. In an out-of-order core, later storage-accessing instructions that can compute their storage address are allowed to run.
- ▶ Hiding storage access latency: Many instructions access data from storage. Storage accesses can miss the L1 and require 7 - 50 more clock cycles to retrieve the storage data. In an in-order core, later instructions in the code stream are stalled. In an out-of-order core, later instructions that are not dependent on this storage data are allowed to run.

The z13s processor has pipeline enhancements that benefit OOO execution. The z Systems processor design has advanced micro-architectural innovations that provide these benefits:

- ▶ Maximized instruction-level parallelism (ILP) for a better cycles per instruction (CPI) design by reviewing every part of the z114 design
- ▶ Maximized performance per watt
- ▶ Enhanced instruction dispatch and grouping efficiency
- ▶ Increased OOO resources (Global Completion Table entries, physical General Purpose Register (GPR) entries, and physical FPR entries)
- ▶ Improved completion rate
- ▶ Reduced cache/TLB miss penalty
- ▶ Improved execution of D-Cache store and reload, and new fixed-point divide
- ▶ New oscillator card (OSC) (load-hit-store conflict) avoidance scheme
- ▶ Enhanced branch prediction structure and sequential instruction fetching

Program results

The OoO execution does not change any program results. Execution can occur out of (program) order, but all program dependencies are honored, ending up with the same results of the in-order (program) execution.

This implementation requires special circuitry to make execution and memory accesses display in order to the software. The logical diagram of a z13s core is shown in Figure 3-9.

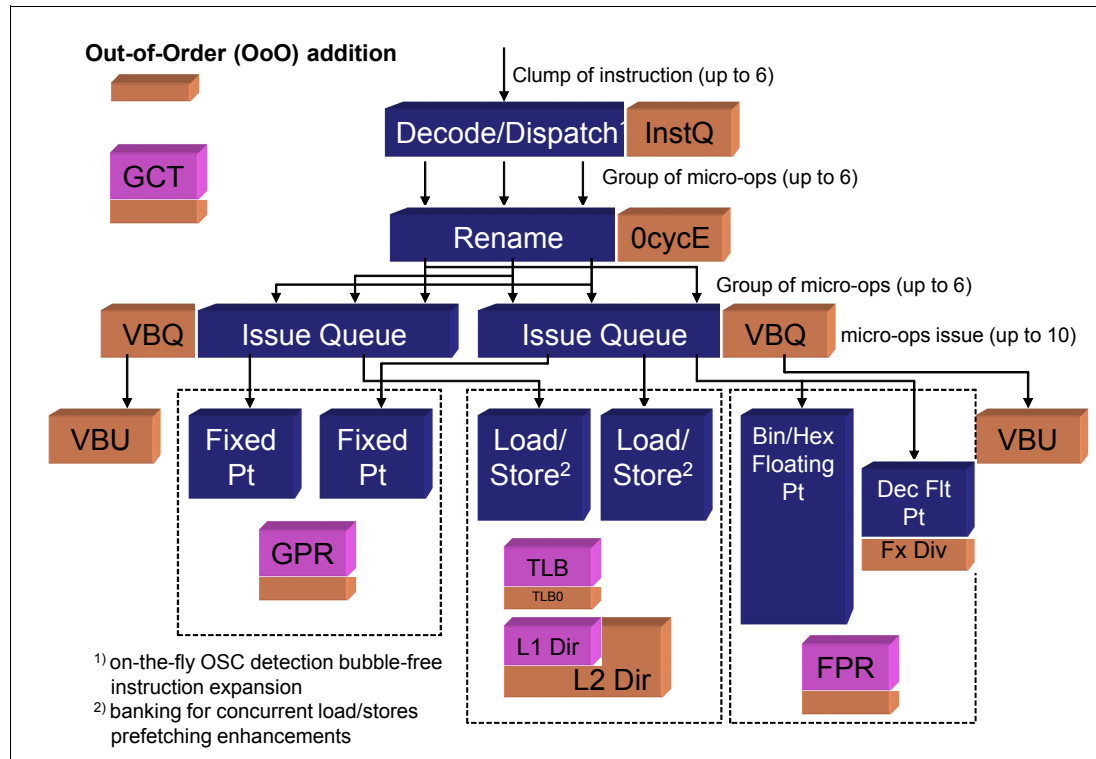


Figure 3-9 z13s PU core logical diagram

Memory address generation and memory accesses can occur out of (program) order. This capability can provide a greater use of the z13s superscalar core, and can improve system performance.

Figure 3-10 shows how OOO core execution can reduce the run time of a program.

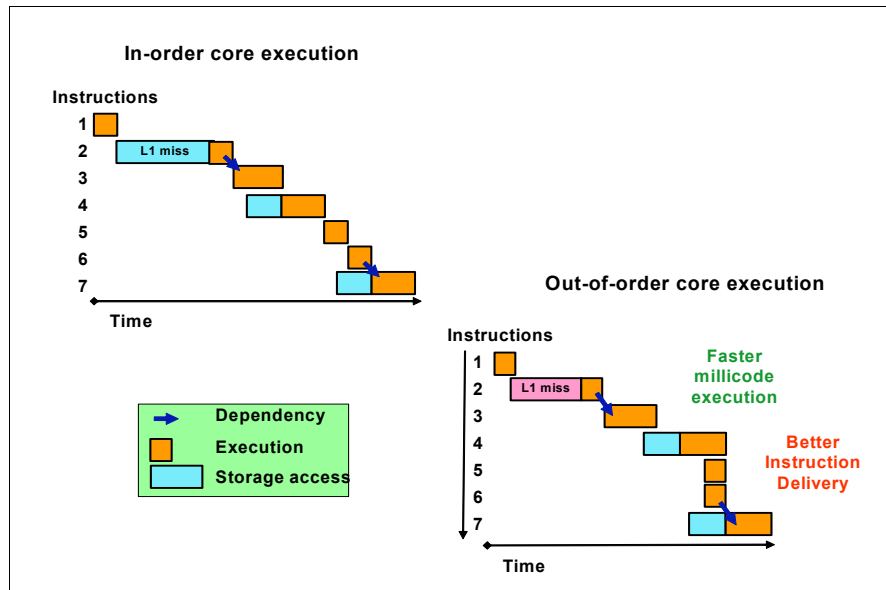


Figure 3-10 z13s in-order and out-of-order core execution improvements

The left side of Figure 3-10 shows an in-order core execution. Instruction 2 has a large delay because of an L1 cache miss, and the next instructions wait until Instruction 2 finishes. In the usual in-order execution, the next instruction waits until the previous instruction finishes. Using OOO core execution, which is shown on the right side of Figure 3-10, Instruction 4 can start its storage access and run while instruction 2 is waiting for data. This situation occurs only if no dependencies exist between the two instructions. When the L1 cache miss is solved, Instruction 2 can also start its run while Instruction 4 is running. Instruction 5 might need the same storage data that is required by Instruction 4. As soon as this data is on L1 cache, Instruction 5 starts running at the same time. The z13 superscalar PU core can have up to 10 instructions/operations running per cycle. This technology results in a shorter run time.

Branch prediction

If the branch prediction logic of the microprocessor makes the wrong prediction, removing all instructions in the parallel pipelines might be necessary. The wrong branch prediction is expensive in a high-frequency processor design. Therefore, the branch prediction techniques that are used are important to prevent as many wrong branches as possible.

For this reason, various history-based branch prediction mechanisms are used, as shown on the in-order part of the z13s PU core logical diagram in Figure 3-9 on page 93. The branch target buffer (BTB) runs ahead of instruction cache pre-fetches to prevent branch misses in an early stage. Furthermore, a branch history table (BHT) in combination with a pattern history table (PHT) and the use of tagged multi-target prediction technology branch prediction offer a high branch prediction success rate.

The z13s microprocessor improves the branch prediction throughput by using the new branch prediction and instruction fetch front end.

3.4.4 Superscalar processor

A *scalar processor* is a processor that is based on a single-issue architecture, which means that only a single instruction is run at a time. A *superscalar processor* allows concurrent (parallel) execution of instructions by adding more resources to the microprocessor in multiple pipelines, each working on its own set of instructions to create parallelism.

A superscalar processor is based on a multi-issue architecture. However, when multiple instructions can be run during each cycle, the level of complexity is increased because an operation in one pipeline stage might depend on data in another pipeline stage. Therefore, a superscalar design demands careful consideration of which instruction sequences can successfully operate in a long pipeline environment.

On z13s servers, up to six instructions can be decoded per cycle and up to 10 instructions or operations can be in execution per cycle. Execution can occur out of (program) order. These improvements also make the simultaneous execution of two threads in the same processor possible.

Many challenges exist in creating an efficient superscalar processor. The superscalar design of the PU has made significant strides in avoiding address generation interlock (AGI) situations. Instructions that require information from memory locations can suffer multi-cycle delays while getting the needed memory content. Because high-frequency processors wait “faster” (spend processor cycles more quickly while idle), the cost of getting the information might become prohibitive.

3.4.5 Compression and cryptography accelerators on a chip

This section describes the compression and cryptography features.

Coprocessor units

Each core in the chip has one coprocessor (COP) unit for compression and cryptography. The compression engine uses static dictionary compression and expansion. The compression dictionary uses the L1-cache (instruction cache).

The cryptography engine is used for CPACF, which offers a set of symmetric cryptographic functions for encrypting and decrypting of clear key operations.

These are some of the characteristics of the z13s coprocessors:

- ▶ Each core has an independent compression and cryptographic engine.
- ▶ COP has been redesigned from scratch to support SMT operation and to increase throughput.
- ▶ It is available to any processor type.
- ▶ The owning processor is busy when its coprocessor is busy.

Figure 3-11 shows the location of the coprocessor on the chip.

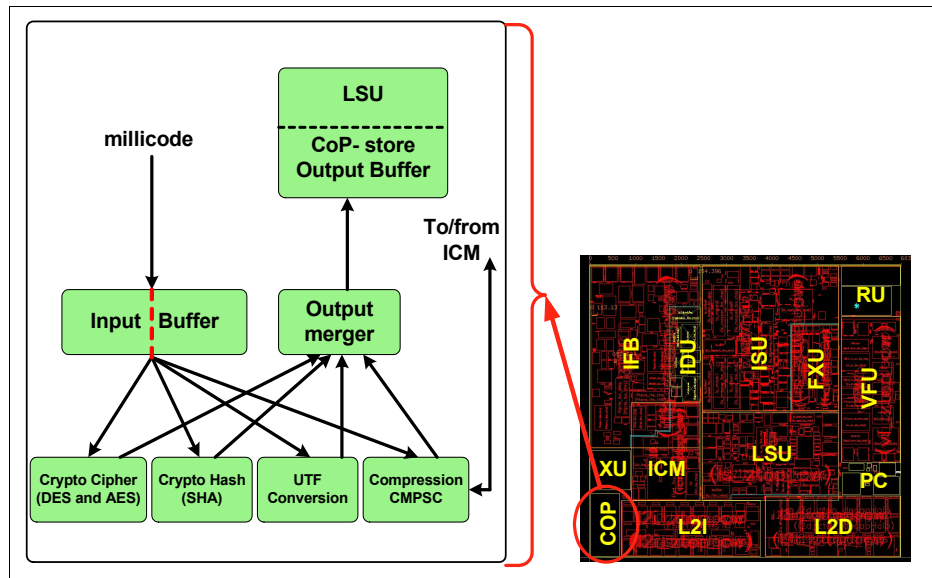


Figure 3-11 Compression and cryptography accelerators on a core in the chip

CP Assist for Cryptographic Function

CPACF accelerates the encrypting and decrypting of SSL/TLS transactions, virtual private network (VPN)-encrypted data transfers, and data-storing applications that do not require FIPS² 140-2 level 4 security. The assist function uses a special instruction set for symmetrical clear key cryptographic encryption and decryption, and for hash operations. This group of instructions is known as the *Message-Security Assist* (MSA). For more information about these instructions, see *z/Architecture Principles of Operation*, SA22-7832.

For more information about cryptographic functions on z13s servers, see Chapter 6, “Cryptography” on page 199.

3.4.6 Decimal floating point accelerator

The DFP accelerator function is present on each of the microprocessors (cores) on the 8-core chip. Its implementation meets business application requirements for better performance, precision, and function.

Base 10 arithmetic is used for most business and financial computation. Floating point computation that is used for work that is typically done in decimal arithmetic involves frequent data conversions and approximation to represent decimal numbers. This process has made floating point arithmetic complex and error-prone for programmers who use it for applications in which the data is typically decimal.

Hardware decimal floating point computational instructions provide the following features:

- ▶ Data formats of 4 bytes, 8 bytes, and 16 bytes
- ▶ An encoded decimal (base 10) representation for data
- ▶ Instructions for running decimal floating point computations
- ▶ An instruction that runs data conversions to and from the decimal floating point representation

² Federal Information Processing Standard (FIPS) 140-2 Security Requirements for Cryptographic Modules

Benefits of the DFP accelerator

The DFP accelerator offers the following benefits:

- ▶ Avoids rounding issues, such as those that happen with binary-to-decimal conversions.
- ▶ Controls existing binary-coded decimal (BCD) operations better.
- ▶ Follows the dominant decimal data and decimal operations in commercial computing industry standardization (IEEE 745R) of decimal floating point operations. Instructions are added in support of the Draft Standard for Floating-Point Arithmetic - IEEE 754-2008, which is intended to supersede the ANSI/IEEE Standard 754-1985.
- ▶ Allows COBOL programs that use zoned-decimal operations to take advantage of the z/Architecture DFP instructions.

z13s servers have two DFP accelerator units per core, which improve the decimal floating point execution bandwidth. The floating point instructions operate on newly designed vector registers (32 new 128-bit registers).

z13s servers have new decimal floating point packed conversion facility support with the following benefits:

- ▶ Reduces code path length because extra instructions to format conversion are not needed anymore.
- ▶ Packed data is operated in memory by all decimal instructions without general-purpose registers, which were required only to prepare for decimal floating point packed conversion instruction.
- ▶ Converting from packed can now force the input packed value to positive instead of requiring a separate OI, OILL, or load positive instruction.
- ▶ Converting to packed can now force a positive zero result instead of requiring ZAP instruction.

Software support

Decimal floating point is supported in the following programming languages and products:

- ▶ Release 4 and later of the High Level Assembler
- ▶ C/C++ (requires z/OS 1.10 with program temporary fixes (PTFs) for full support or later)
- ▶ Enterprise PL/I Release 3.7 and Debug Tool Release 8.1 or later
- ▶ Java Applications using the BigDecimal Class Library
- ▶ SQL support as of DB2 Version 9 or later

3.4.7 IEEE floating point

Binary and hexadecimal floating-point instructions are implemented in z13s servers. They incorporate IEEE standards into the system.

The key point is that Java and C/C++ applications tend to use IEEE BFP operations more frequently than earlier applications. Therefore, improving the hardware implementation of this set of instructions also improves the performance of applications.

3.4.8 Processor error detection and recovery

The PU uses a process called *transient recovery* as an error recovery mechanism. When an error is detected, the instruction unit tries the instruction again and attempts to recover the error. If the second attempt is unsuccessful (that is, a permanent fault exists), a relocation process is started that restores the full capacity by moving work to another PU. Relocation

under hardware control is possible because the R-unit has the full designed state in its buffer. The principle is shown in Figure 3-12.

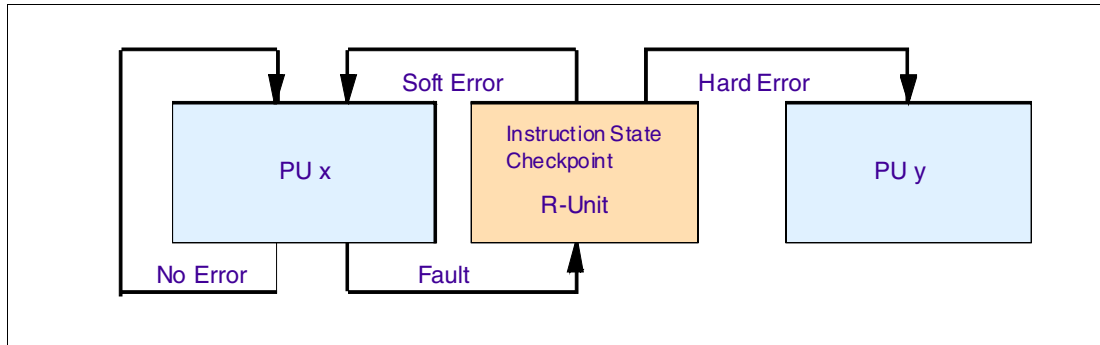


Figure 3-12 PU error detection and recovery

3.4.9 Branch prediction

Because of the ultra-high frequency of the PUs, the penalty for a wrongly predicted branch is high. Therefore, a multi-pronged strategy for branch prediction, based on gathered branch history that is combined with other prediction mechanisms, is implemented on each microprocessor.

The BHT implementation on processors provides a large performance improvement. Originally introduced on the IBM ES/9000 9021 in 1990, the BHT is continuously improved.

The BHT offers significant branch performance benefits. The BHT allows each PU to take instruction branches based on a stored BHT, which improves processing times for calculation routines. In addition to the BHT, z13s servers use various techniques to improve the prediction of the correct branch to be run. The following techniques are used:

- ▶ Branch history table (BHT)
- ▶ Branch target buffer (BTB)
- ▶ Pattern history table (PHT)
- ▶ BTB data compression

The success rate of branch prediction contributes significantly to the superscalar aspects of z13s servers because the architecture rules prescribe that, for successful parallel execution of an instruction stream, the correctly predicted result of the branch is essential.

3.4.10 Wild branch

When a bad pointer is used or when code overlays a data area that contains a pointer to code, a random branch is the result. This process causes a 0C1 or 0C4 abend. Random branches are hard to diagnose because clues about how the system got there are not evident.

With the wild branch hardware facility, the last address from which a successful branch instruction was run is kept. z/OS uses this information with debugging aids, such as the **SLIP** command, to determine where a wild branch came from. It might also collect data from that storage location. This approach decreases the number of debugging steps that are necessary when you want to know from where the branch came from.

3.4.11 Translation lookaside buffer

The TLB in the instruction and data L1 caches use a secondary TLB to enhance performance. In addition, a translator unit is added to translate misses in the secondary TLB.

The size of the TLB is kept as small as possible because of its short access time requirements and hardware space limitations. Because memory sizes have recently increased significantly as a result of the introduction of 64-bit addressing, a smaller working set is represented by the TLB. To increase the working set representation in the TLB without enlarging the TLB, large (1 MB) page and giant page (2 GB) support is available and can be used when appropriate. For more information, see “Large page support” on page 114.

With the enhanced DAT-2 (EDAT-2) improvements, the z Systems introduce architecture enhancements to allow support for 2 GB page frames.

3.4.12 Instruction fetching, decoding, and grouping

The superscalar design of the microprocessor allows for the decoding of up to six instructions per cycle and the execution of up to 10 instructions per cycle. Both execution and storage accesses for instruction and operand fetching can occur out of sequence.

Instruction fetching

Instruction fetching normally tries to get as far ahead of instruction decoding and execution as possible because of the relatively large instruction buffers available. In the microprocessor, smaller instruction buffers are used. The operation code is fetched from the I-cache and put in instruction buffers that hold prefetched data that is awaiting decoding.

Instruction decoding

The processor can decode up to six instructions per cycle. The result of the decoding process is queued and later used to form a group.

Instruction grouping

From the instruction queue, up to 10 instructions can be completed on every cycle. A complete description of the rules is beyond the scope of this book.

The compilers and JVMs are responsible for selecting instructions that best fit with the superscalar microprocessor. They abide by the rules to create code that best uses the superscalar implementation. All the z Systems compilers and the JVMs are constantly updated to benefit from new instructions and advances in microprocessor designs.

3.4.13 Extended Translation Facility

Instructions have been added to the z/Architecture instruction set in support of the Extended Translation Facility. They are used in data conversion operations for Unicode data, causing applications that are enabled for Unicode or globalization to be more efficient. These data-encoding formats are used in web services, grid, and on-demand environments where XML and SOAP technologies are used. The High Level Assembler supports Extended Translation Facility instructions.

3.4.14 Instruction set extensions

The processor supports many instructions to support functions:

- ▶ Hexadecimal floating point instructions for various unnormalized multiply and multiply-add instructions.
- ▶ Immediate instructions, including various add, compare, OR, exclusive-OR, subtract, load, and insert formats. Use of these instructions improves performance.
- ▶ Load instructions for handling unsigned halfwords, such as those used for Unicode.
- ▶ Cryptographic instructions, which are known as the MSA, offer the full complement of the AES, SHA-1, SHA-2, and DES algorithms. They also include functions for random number generation.
- ▶ Extended Translate Facility-3 instructions, enhanced to conform with the current Unicode 4.0 standard.
- ▶ Assist instructions that help eliminate hypervisor processor usage.
- ▶ SIMD instructions, which allow the parallel processing of multiple elements in a single instruction.

3.4.15 Transactional Execution

The Transactional Execution (TX) capability, which is known in the industry as hardware transactional memory, runs a group of instructions atomically. That is, either all their results are committed or no result is committed. The execution is optimistic. The instructions are run, but previous state values are saved in a “transactional memory”. If the transaction succeeds, the saved values are discarded. Otherwise, they are used to restore the original values.

The Transaction Execution Facility provides instructions, including declaring the beginning and end of a transaction, and canceling the transaction. TX is expected to provide significant performance benefits and scalability by avoiding most locks. This benefit is especially important for heavily threaded applications, such as Java.

3.4.16 Runtime Instrumentation

Runtime Instrumentation (RI) is a hardware facility for managed run times, such as the Java runtime environment (JRE). RI allows dynamic optimization of code generation as it is being run. It requires fewer system resources than the current software-only profiling, and provides information about hardware and program characteristics. It enhances JRE in making the correct decision by providing real-time feedback.

3.5 Processor unit functions

This section describes the PU functions.

3.5.1 Overview

All PUs on a z13s server are physically identical. When the system is initialized, one integrated firmware processor (IFP) is allocated from the pool of PUs that is available for the whole system. The other PUs can be characterized to specific functions (CP, IFL, ICF, zIIP, or SAP).

The function that is assigned to a PU is set by the Licensed Internal Code (LIC). The LIC is loaded when the system is initialized (at power-on reset (POR)) and the PUs are *characterized*.

Only characterized PUs have a designated function. Non-characterized PUs are considered spares. Order at least one CP, IFL, or ICF on a z13s server.

This design brings outstanding flexibility to z13s servers because any PU can assume any available characterization. The design also plays an essential role in system availability because PU characterization can be done dynamically, with no system outage.

For more information about software level support of functions and features, see Chapter 7, “Software support” on page 229.

Concurrent upgrades

Except on fully configured models, concurrent upgrades can be done by the LIC, which assigns a PU function to a previously non-characterized PU. Within the CPC drawer boundary or boundary of multiple CPC drawers, no hardware changes are required. The upgrade can be done concurrently through the following facilities:

- ▶ Customer Initiated Upgrade (CIU) for permanent upgrades
- ▶ On/Off capacity on demand (On/Off CoD) for temporary upgrades
- ▶ Capacity BackUp (CBU) for temporary upgrades
- ▶ Capacity for Planned Event (CPE) for temporary upgrades

If the PU SCMs in the installed CPC drawer node have no available remaining PUs, an upgrade results in a model upgrade (upgrade Model N10 to N20, same drawer). If the upgrade is for more than 2048 GB of memory or additional I/O features, then installation of an extra CPC drawer is necessary. However, there is a limit of two CPC drawers. CPC drawer installation is disruptive, and takes more time than a simple LIC upgrade.

For more information about Capacity on Demand, see Chapter 8, “System upgrades” on page 311.

PU sparing

In the rare event of a PU failure, the failed PU’s characterization can dynamically and transparently be reassigned to a spare PU if the physical configuration allows for it. The z13s Model N20 has two spare PUs. PUs that are not characterized on a CPC configuration can also be used as spare PUs. For more information about PU sparing, see 3.5.10, “Sparing rules” on page 111.

PU pools

PUs that are defined as CPs, IFLs, ICFs, and zIIPs are grouped in their own pools, from where they can be managed separately. This configuration significantly simplifies capacity planning and management for LPARs. The separation also affects weight management because CP and zIIP weights can be managed separately. For more information, see “PU weighting” on page 102.

All assigned PUs are grouped in the PU pool. These PUs are dispatched to online logical PUs.

As an example, consider a z13s server with 6 CPs, two IFLs, five zIIPs, and one ICF. This system has a PU pool of 14 PUs, called the *pool width*. Subdivision defines these pools:

- ▶ A CP pool of 6 CPs
- ▶ An ICF pool of one ICF

- ▶ An IFL pool of two IFLs
- ▶ A zIIP pool of five zIIPs

POs are placed in the pools in the following circumstances:

- ▶ When the system undergoes POR
- ▶ At the time of a concurrent upgrade
- ▶ As a result of the addition of POs during a CBU
- ▶ Following a capacity on-demand upgrade through On/Off CoD or CIU

POs are removed from their pools when a concurrent downgrade takes place as the result of the removal of a CBU. They are also removed through On/Off CoD process and the conversion of a PO. When a dedicated LPAR is activated, its POs are taken from the correct pools. This is also the case when an LPAR logically configures a PO on, if the width of the pool allows for it.

For an LPAR, logical POs are dispatched from the supporting pool only. The logical CPs are dispatched from the CP pool, logical zIIPs from the zIIP pool, logical IFLs from the IFL pool, and the logical ICFs from the ICF pool.

PO weighting

Because CPs, zIIPs, IFLs, and ICFs have their own pools from where they are dispatched, they can be given their own weights. For more information about PO pools and processing weights, see the *z Systems Processor Resource/Systems Manager Planning Guide*, SB10-7162.

3.5.2 Central processors

A central processor (CP) is a PO that uses the full z/Architecture instruction set. It can run z/Architecture-based operating systems (z/OS, z/VM, TPF, z/TPF, z/VSE, and Linux), the Coupling Facility Control Code (CFCC), and IBM zAware. Up to 6 POs can be characterized as CPs, depending on the configuration.

z13s servers can be initialized either in LPAR mode or in Elastic PR/SM mode. For more information, see Appendix E, “IBM Dynamic Partition Manager” on page 501. CPs are defined as either dedicated or shared. Reserved CPs can be defined to an LPAR to allow for nondisruptive image upgrades. If the operating system in the LPAR supports the logical processor add function, reserved processors are no longer needed. Regardless of the installed model, an LPAR can have up to 141 logical processors defined (the sum of active and reserved logical processors). In practice, define no more CPs than the operating system supports. For example, the z/OS V1R13 LPAR supports a maximum of 100 logical CPs and z/OS V2R1 LPAR supports a maximum of 141 logical CPs. See the *z Systems Processor Resource/Systems Manager Planning Guide*, SB10-7162.

All POs that are characterized as CPs within a configuration are grouped into the CP pool. The CP pool can be seen on the HMC workplace. Any z/Architecture operating systems, CFCCs, and IBM zAware can run on CPs that are assigned from the CP pool.

z13s servers have 26 capacity levels, which are named from A to Z. Within each capacity level, a one, two, three, four, five, or six-way model is offered. Each model is identified by its capacity level indicator (A through Z) followed by an indication of the number of CPs available (01 - 06). Therefore, z13s servers offer 156 capacity settings. All models have a related millions of service units (MSU) value that is used to determine the software license charge for monthly license charge (MLC) software (see Table 2-12 on page 73).

A00: Model capacity identifier A00 is used for IFL or ICF configurations only.

For more information about capacity settings, see 2.7, “Model configurations” on page 69.

3.5.3 Integrated Facility for Linux

An Integrated Facility for Linux (IFL) is a PU that can be used to run Linux, Linux guests on z/VM operating systems, kernel-based virtual machine (KVM) for IBM z, and IBM zAware. Up to 20 PUs can be characterized as IFLs, depending on the configuration. IFLs can be dedicated to a Linux, a z/VM, or an IBM zAware LPAR, or they can be shared by multiple Linux guests, z/VM LPARs, or IBM zAware running on the same z13s server. Only z/VM, Linux on z Systems operating systems, KVM for z, IBM zAware, and designated software products can run on IFLs. IFLs are orderable by using FC 1924.

IFL pool

All PUs that are characterized as IFLs within a configuration are grouped into the IFL pool. The IFL pool can be seen on the HMC workplace.

IFLs do not change the model capacity identifier of z13s servers. Software product license charges that are based on the model capacity identifier are not affected by the addition of IFLs.

Unassigned IFLs

An IFL that is purchased but not activated is registered as an unassigned IFL (FC 1928). When the system is later upgraded with an additional IFL, the system recognizes that an IFL already was purchased and is present.

3.5.4 Internal Coupling Facility

An Internal Coupling Facility (ICF) is a PU that is used to run the CFCC for Parallel Sysplex environments. Within the capacity of the sum of all unassigned PUs in two CPC drawer configuration, up to 20 ICFs can be characterized, depending on the model. However, the maximum number of ICFs that can be defined on a coupling facility LPAR is limited to 16. ICFs are orderable by using FC 1925.

ICFs exclusively run CFCC. ICFs do not change the model capacity identifier of z13s servers. Software product license charges that are based on the model capacity identifier are not affected by the addition of ICFs.

All ICFs within a configuration are grouped into the ICF pool. The ICF pool can be seen on the Hardware Management Console (HMC) workplace.

The ICFs can only be used by coupling facility LPARs. ICFs are either dedicated or shared. ICFs can be dedicated to a CF LPAR, or shared by multiple CF LPARs that run on the same system. However, having an LPAR with dedicated and shared ICFs at the same time is not possible.

Coupling Thin Interrupts

With the introduction of Driver 15F (zEC12 and IBM zEnterprise BC12 (zBC12)), the z Systems architecture provides a new thin interrupt class called *Coupling Thin Interrupts*. The capabilities that are provided by hardware, firmware, and software support the generation of coupling-related “thin interrupts” when the following situations occur:

- ▶ On the CF side, a CF command or a CF signal (arrival of a CF-to-CF duplexing signal) is received by a shared-engine CF image, or when the completion of a CF signal that was previously sent by the CF occurs (completion of a CF-to-CF duplexing signal).
- ▶ On the z/OS side, a CF signal is received by a shared-engine z/OS image (arrival of a List Notification signal) or an asynchronous CF operation completes.

The interrupt causes the receiving partition to be dispatched by an LPAR, if it is not already dispatched. This action allows the request, signal, or request completion to be recognized and processed in a more timely manner.

After the image is dispatched, existing “poll for work” logic in both CFCC and z/OS can be used largely as is to locate and process the work. The new interrupt speeds up the redispaching of the partition.

LPAR presents these Coupling Thin Interrupts to the guest partition, so CFCC and z/OS both require interrupt handler support that can deal with them. CFCC also changes to give up control of the processor as soon as all available pending work is exhausted, or when the LPAR undispatches it off the shared processor, whichever comes first.

Coupling facility combinations

A coupling facility image can have one of the following combinations that are defined in the image profile:

- ▶ Dedicated ICFs
- ▶ Shared ICFs
- ▶ Dedicated CPs
- ▶ Shared CPs

Shared ICFs add flexibility. However, running only with shared coupling facility PUs (either ICFs or CPs) is not a preferable production configuration. A production CF should generally operate by using dedicated ICFs. With CFCC Level 19 (and later; z13s servers run CFCC level 21), Coupling Thin Interrupts are available, and dedicated engines continue to be recommended to obtain the best coupling facility performance.

In Figure 3-13, the CPC on the left has two environments that are defined (production and test), each of which has one z/OS and one coupling facility image. The coupling facility images share an ICF. The coupling facility images share an ICF.

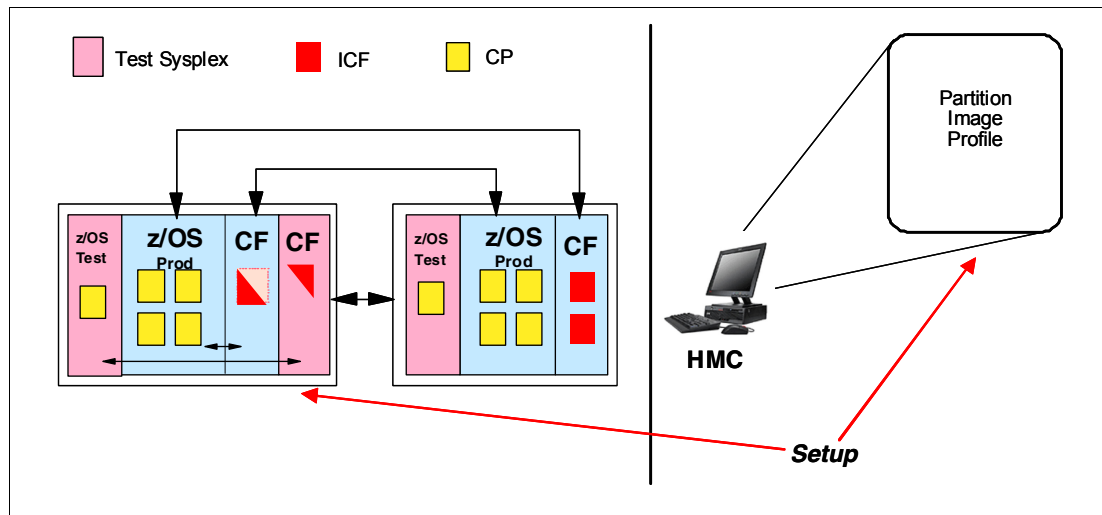


Figure 3-13 ICF options: Shared ICFs

The LPAR processing weights are used to define how much processor capacity each coupling facility image can have. The capped option can also be set for a test coupling facility (CF) image to protect the production environment.

Connections between these z/OS and coupling facility images can use internal coupling links to avoid the use of real (external) coupling links, and get the best link bandwidth available.

Dynamic coupling facility dispatching

The *dynamic coupling facility dispatching* function has a dispatching algorithm that you can use to define a backup coupling facility in an LPAR on the system. When this LPAR is in backup mode, it uses few processor resources. When the backup CF becomes active, only the resources that are necessary to provide coupling are allocated.

CFCC Level 19 introduced Coupling Thin Interrupts and the new DYNDISP specification. It allows more environments with multiple CF images to coexist in a server, and to share CF engines with reasonable performance. For more information, see 3.9.3, “Dynamic CF dispatching” on page 133.

3.5.5 IBM z Integrated Information Processor

A zIIP³ reduces the standard processor (CP) capacity requirements for z/OS Java, XML system services applications, and a portion of work of z/OS Communications Server and DB2 UDB for z/OS Version 8 or later, freeing up capacity for other workload requirements.

A zIIP enables eligible z/OS workloads to have a portion of them directed to zIIP. The zIIPs do not increase the MSU value of the processor, and so do not affect the IBM software license changes.

The z13s server is the first business class z Systems Processor to support SMT. z13s servers support two threads per core on IFLs and zIIPs only. SMT must be enabled at the LPAR level

³ z Systems Application Assist Processors (zAAPs) are not available on z13s servers. A zAAP workload is dispatched to available zIIPs (zAAP on zIIP capability).

and supported by the z/OS operating system. SMT enables continued scaling of per-processor capacity.

How zIIPs work

zIIPs are designed to support designated z/OS workloads. One of the workloads is Java code execution. When Java code must be run (for example, under control of IBM WebSphere), the z/OS JVM calls the function of the zIIP. The z/OS dispatcher then suspends the JVM task on the CP that it is running on and dispatches it to an available zIIP. After the Java application code execution is finished, z/OS redispaches the JVM task to an available CP. After this process occurs, normal processing is resumed.

This process reduces the CP time that is needed to run Java WebSphere applications, freeing that capacity for other workloads.

Figure 3-14 shows the logical flow of Java code running on a z13s server that has a zIIP available. When JVM starts the execution of a Java program, it passes control to the z/OS dispatcher that verifies the availability of a zIIP.

The availability is treated in the following manner:

- ▶ If a zIIP is available (not busy), the dispatcher suspends the JVM task on the CP and assigns the Java task to the zIIP. When the task returns control to the JVM, it passes control back to the dispatcher. The dispatcher then reassigns the JVM code execution to a CP.
- ▶ If no zIIP is available (all busy), the z/OS dispatcher allows the Java task to run on a standard CP. This process depends on the option that is used in the OPT statement in the IEAOPTxx member of SYS1.PARMLIB.

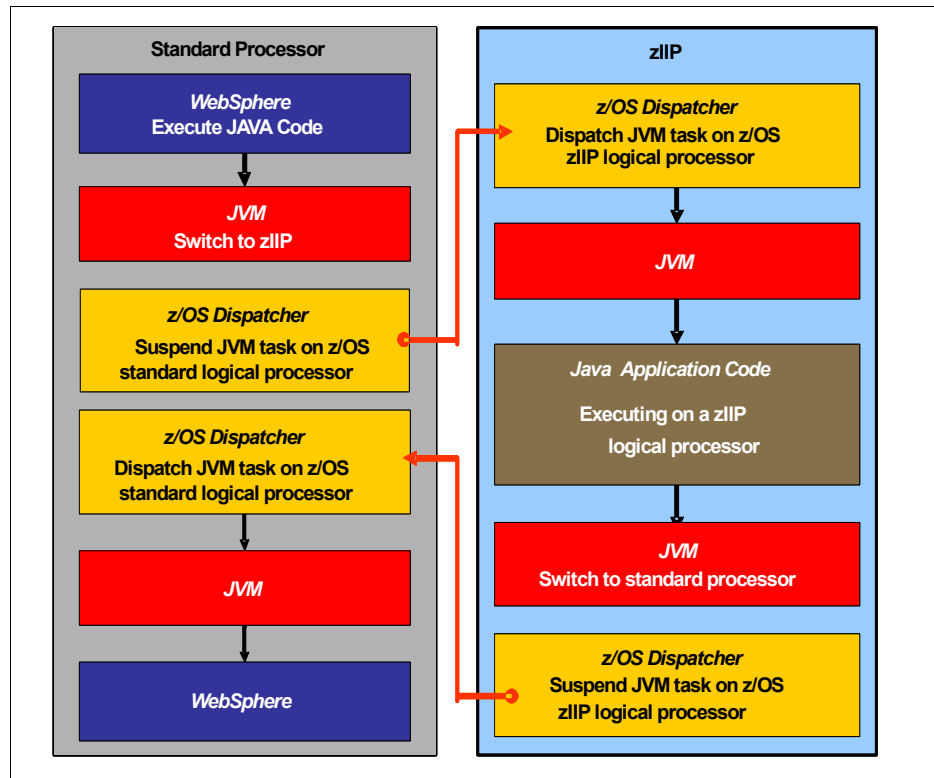


Figure 3-14 Logical flow of Java code execution on a zIIP

A zIIP runs only IBM authorized code. This IBM authorized code includes the z/OS JVM in association with parts of system code, such as the z/OS dispatcher and supervisor services. A zIIP cannot process I/O or clock comparator interruptions, and it does not support operator controls, such as IPL.

Java application code can either run on a CP or a zIIP. The installation can manage the use of CPs so that Java application code runs only on CPs, only on zIIPs, or on both.

Two execution options for zIIP-eligible code execution are available. These options are user-specified in IEAOPTxx, and can be dynamically altered by the **SET OPT** command. The following options are currently supported for z/OS V1R10 and later releases:

- ▶ Option 1: Java dispatching by priority (IIPHONORPRIORITY=YES): This is the default option, and specifies that CPs must not automatically consider zIIP-eligible work for dispatching on them. The zIIP-eligible work is dispatched on the zIIP engines until WLM determines that the zIIPs are overcommitted. WLM then requests help from the CPs. When help is requested, the CPs consider dispatching zIIP-eligible work on the CPs themselves based on the dispatching priority relative to other workloads. When the zIIP engines are no longer overcommitted, the CPs stop considering zIIP-eligible work for dispatch.

This option runs as much zIIP-eligible work on zIIPs as possible, and allows it to spill over onto the CPs only when the zIIPs are overcommitted.

- ▶ Option 2: Java dispatching by priority (IIPHONORPRIORITY=NO): zIIP-eligible work runs on zIIPs only while at least one zIIP engine is online. zIIP-eligible work is not normally dispatched on a CP, even if the zIIPs are overcommitted and CPs are unused. The exception is that zIIP-eligible work can sometimes run on a CP to resolve resource conflicts.

Therefore, zIIP-eligible work does not affect the CP utilization that is used for reporting through the subcapacity reporting tool (SCRT), no matter how busy the zIIPs are.

If zIIPs are defined to the LPAR but are not online, the zIIP-eligible work units are processed by CPs in order of priority. The system ignores the IIPHONORPRIORITY parameter in this case and handles the work as though it had no eligibility to zIIPs.

zIIPs provide the following benefits:

- ▶ Potential cost savings.
- ▶ Simplification of infrastructure as a result of the collocation and integration of new applications with their associated database systems and transaction middleware, such as DB2, IMS, or CICS. Simplification can happen, for example, by introducing a uniform security environment, and by reducing the number of TCP/IP programming stacks and system interconnect links.
- ▶ Prevention of processing latencies that occur if Java application servers and their database servers are deployed on separate server platforms.

The following DB2 UDB for z/OS V8 or later workloads are eligible to run in SRB mode:

- ▶ Query processing of network-connected applications that access the DB2 database over a TCP/IP connection by using IBM Distributed Relational Database Architecture (DRDA). DRDA enables relational data to be distributed among multiple systems. It is native to DB2 for z/OS, and so reduces the need for more gateway products that can affect performance and availability. The application uses the DRDA requester or server to access a remote database. IBM DB2 Connect™ is an example of a DRDA application requester.
- ▶ Star schema query processing, mostly used in business intelligence (BI) work. A *star schema* is a relational database schema for representing multidimensional data. It stores

data in a central fact table and is surrounded by more dimension tables that hold information about each perspective of the data. A star schema query, for example, joins various dimensions of a star schema data set.

- ▶ DB2 utilities that are used for index maintenance, such as LOAD, REORG, and REBUILD. Indexes allow quick access to table rows, but over time, as data in large databases is manipulated, the databases become less efficient and must be maintained.

The zIIP runs portions of eligible database workloads, and so helps to free computer capacity and lower software costs. Not all DB2 workloads are eligible for zIIP processing. DB2 UDB for z/OS V8 and later gives z/OS the information to direct portions of the work to the zIIP. The result is that in every user situation, different variables determine how much work is redirected to the zIIP.

On a z13s server, the following workloads can also benefit from zIIPs:

- ▶ z/OS Communications Server uses the zIIP for eligible Internet Protocol Security (IPSec) network encryption workloads. This configuration requires z/OS V1R10 or later releases. Portions of IPSec processing take advantage of the zIIPs, specifically end-to-end encryption with IPSec. The IPSec function moves a portion of the processing from the general-purpose processors to the zIIPs. In addition, to run the encryption processing, the zIIP also handles the cryptographic validation of message integrity and IPSec header processing.
- ▶ z/OS Global Mirror, formerly known as Extended Remote Copy (XRC), uses the zIIP as well. Most z/OS Data Facility Storage Management Subsystem (DFSMS) system data mover (SDM) processing that is associated with z/OS Global Mirror can run on the zIIP. This configuration requires z/OS V1R10 or later releases.
- ▶ The first IBM user of z/OS XML system services is DB2 V9. For DB2 V9 before the z/OS XML System Services enhancement, z/OS XML System Services non-validating parsing was partially directed to zIIPs when used as part of a distributed DB2 request through DRDA. This enhancement benefits DB2 V9 by making all z/OS XML System Services non-validating parsing eligible to zIIPs. This configuration is possible when processing is used as part of any workload that is running in enclave SRB mode.
- ▶ z/OS Communications Server also allows the HiperSockets Multiple Write operation for outbound large messages (originating from z/OS) to be run by a zIIP. Application workloads that are based on XML, HTTP, SOAP, and Java, and traditional file transfer, can benefit.
- ▶ For business intelligence, IBM Scalable Architecture for Financial Reporting provides a high-volume, high performance reporting solution by running many diverse queries in z/OS batch. It can also be eligible for zIIP.

For more information about zIIP and eligible workloads, see the IBM zIIP website:

<http://www.ibm.com/systems/z/hardware/features/ziip/index.html>

zIIP installation information

One CP must be installed with or before any zIIP is installed. In z13s servers, the zIIP to CP ratio is 2:1, which means that up to 12 zIIPs on any specific model x06 can be characterized. Table 3-1 shows the allowed number of zIIPs for each model.

Table 3-1 Number of zIIPs per model

Model	N10	N20
Maximum zIIPs	0 - 6	0 - 12

zIIPs are orderable by using FC 1927. Up to two zIIP can be ordered for each CP or marked CP configured in the system.

PUs characterized as zIIPs within a configuration are grouped into the zIIP pool. This configuration allows zIIPs to have their own processing weights, independent of the weight of parent CPs. The zIIP pool can be seen on the hardware console.

The number of permanent zIIPs plus temporary zIIPs cannot exceed twice the number of purchased CPs plus temporary CPs. Also, the number of temporary zIIPs cannot exceed the number of permanent zIIPs.

zIIPs and logical partition definitions

zIIPs are either dedicated or shared depending on whether they are part of an LPAR with dedicated or shared CPs. In an LPAR, at least one CP must be defined before zIIPs for that partition can be defined. The number of zIIPs that are available in the system is the number of zIIPs that can be defined to an LPAR.

Logical partition: In a logical partition, as many zIIPs as are available can be defined together with at least one CP.

3.5.6 System assist processors

A system assist processor (SAP) is a PU that runs the channel subsystem LIC to control I/O operations. All SAPs run I/O operations for all LPARs. All models have standard SAPs configured. The number of standard SAPs depends on the z13s model, as shown in Table 3-2.

Table 3-2 SAPs per model

Model	N10	N20
Standard SAPs	2	3

SAP configuration

A standard SAP configuration provides a well-balanced system for most environments. However, some application environments have high I/O rates, which are typically various Transaction Processing Facility (TPF) environments. In this case, more SAPs can be ordered. Assignment of more SAPs can increase the capability of the channel subsystem to run I/O operations. In z13s systems, the number of SAPs can be greater than the number of CPs. However, additional SAPs plus standard SAPs cannot exceed 6.

Optional additional orderable SAPs

The ability to order more SAPs is an option that is available on all models (FC 1926). These additional SAPs increase the capacity of the channel subsystem to run I/O operation, which is suggested for TPF environments. The maximum number of optional additional orderable SAPs depends on the configuration and the number of available uncharacterized PUs. The number of SAPs is listed in Table 3-3.

Table 3-3 Optional SAPs per model

Model	N10	N20
Optional SAPs	0 - 2	0 - 3

3.5.7 Reserved processors

Reserved processors are defined by the PR/SM to allow for a nondisruptive capacity upgrade. Reserved processors are like spare logical processors, and can be shared or dedicated. Reserved CPs can be defined to an LPAR dynamically to allow for nondisruptive image upgrades.

Reserved processors can be dynamically configured online by an operating system that supports this function if enough unassigned PUs are available to satisfy this request. The PR/SM rules that govern logical processor activation remain unchanged.

By using reserved processors, you can define more logical processors than the number of available CPs, IFLs, ICFs, and zIIPs in the configuration to an LPAR. This definition makes it possible to configure online, nondisruptively, more logical processors after more CPs, IFLs, ICFs, and zIIPs are made available concurrently. They can be made available with one of the Capacity on Demand options.

The maximum number of reserved processors that can be defined to an LPAR depends on the number of logical processors that are already defined. The maximum number of logical processors plus reserved processors is 20. If the operating system in the LPAR supports the logical processor add function, reserved processors are no longer needed.

Do not define more active and reserved processors than the operating system for the LPAR can support. For more information about logical processors and reserved processors and their definitions, see 3.7, “Logical partitioning” on page 116.

3.5.8 Integrated firmware processor

An IFP is allocated from the pool of PUs. This is available for the whole system. Unlike other characterized PUs, the IFP is standard and not defined by the client. It is a single PU dedicated solely to supporting the *native* Peripheral Component Interconnect Express (PCIe) features (10GbE Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE) Express and zEnterprise Data Compression (zEDC) Express) and is initialized at POR. The IFP supports Resource Group LIC to provide native PCIe I/O feature management and virtualization functions. For more information, see Appendix G, “Native Peripheral Component Interconnect Express” on page 521.

3.5.9 Processor unit assignment

The processor unit assignment of characterized PUs is done at POR time, when the system is initialized. The initial assignment rules keep PUs of the same characterization type grouped as much as possible in relation to PU chips and CPC drawer boundaries to optimize shared cache usage.

The PU assignment is based on CPC drawer plug ordering. The CPC drawers are populated from the bottom upward. This process defines the low-order and the high-order CPC drawers:

- ▶ CPC drawer 1: Plug order 1 (low-order CPC drawer)
- ▶ CPC drawer 0: Plug order 2 (high-order CPC drawer)

The assignment rules follow this order:

- ▶ Spare: On CPC drawer 1, no spare is assigned on the high PU SCM. In the Model N10, no spares are assigned.
- ▶ IFP: If only CPC drawer 1 is present (Model N10, one CPC drawer machine, or Model N20, one CPC drawer machine), the IFP is assigned to the high PU SCM. If drawer 0 is present on the machine (Model N20, two CPC drawer machine), then the IFP is in CPC drawer 0 high PU SCM.
- ▶ SAPs: Spread across CPC drawers and high PU SCMs. CPC drawer 1 has two standard SAPs. Start with the highest PU SCM high core, then the next highest PU SCM high core. This configuration prevents all the SAPs from being assigned on one PU SCM.
- ▶ IFLs and ICFs: Assign IFLs and ICFs to cores on SCMs in higher CPC drawers working downward.
- ▶ CPs and zIIPs: Assign CPs and zIIPs to cores on SCMs in lower CPC drawers working upward.

These rules are intended to isolate, as much as possible, on different CPC drawers and even on different PU chips, processors that are used by different operating systems. This configuration ensures that different operating systems do not use the same shared caches. For example, CPs and zIIPs are all used by z/OS, and can benefit by using the same shared caches. However, IFLs are used by z/VM and Linux, and ICFs are used by CFCC. Therefore, for performance reasons, the assignment rules prevent them from sharing L3 and L4 caches with z/OS processors.

This initial PU assignment, which is done at POR, can be dynamically rearranged by an LPAR by swapping an active core with a core in a different PU chip in a different CPC drawer or node to improve system performance. For more information, see “LPAR dynamic PU reassignment” on page 123.

When an extra CPC drawer is added after a POR, processor resources are spread across all available CPC drawers. The processor unit assignment rules consider the newly installed CPC drawer only after the next POR.

Note: The addition of a second CPC drawer is disruptive.

3.5.10 Sparing rules

On a z13s Model N20 system, two PUs are reserved as spares. There are no spare PUs reserved on a z13s Model N10. The reserved spares are available to replace any characterized PUs, whether they are CP, IFL, ICF, zIIP, SAP, or IFP.

Systems with a failed PU for which no spare is available will *call home* for a replacement. A system with a failed PU that is spared and requires an SCM to be replaced (referred to as a *pending repair*) can still be upgraded when sufficient PUs are available.

Transparent CP, IFL, ICF, zIIP, SAP, and IFP sparing

Depending on the model, sparing of CP, IFL, ICF, zIIP, SAP, and IFP is transparent and does not require operating system or operator intervention.

With *transparent sparing*, the status of the application that was running on the failed processor is preserved. The application continues processing on a newly assigned CP, IFL, ICF, zIIP, SAP, or IFP (allocated to one of the spare PUs) without client intervention.

Application preservation

If no spare PU is available, *application preservation* (z/OS only) is started. The state of the failing processor is passed to another active processor that is used by the operating system. Through operating system recovery services, the task is resumed successfully (in most cases, without client intervention).

Dynamic SAP and IFP sparing and reassignment

Dynamic recovery is provided during a failure of the SAP or IFP. If the SAP or IFP fails, and if a spare PU is available, the spare PU is dynamically assigned as a new SAP or IFP. If no spare PU is available, and more than one CP is characterized, a characterized CP is reassigned as an SAP or IFP. In either case, client intervention is not required. This capability eliminates an unplanned outage and allows a service action to be deferred to a more convenient time.

3.5.11 Increased flexibility with z/VM mode partitions

z13s servers provide a capability for the definition of a z/VM mode LPAR that contains a mix of processor types that includes CPs and specialty processors, such as IFLs, zIIPs, and ICFs.

z/VM V5R4 and later support this capability, which increases flexibility and simplifies systems management. In a single LPAR, z/VM can perform these tasks:

- ▶ Manage guests that use Linux on z Systems on IFLs, and those that use z/VSE, or z/OS on CPs
- ▶ Run designated z/OS workloads, such as Java and XML execution and parts of DB2 DRDA processing, on zIIPs

If the only operating system to run under z/VM is Linux, define a Linux only LPAR.

3.6 Memory design

This section describes various considerations of the z13s memory design.

3.6.1 Overview

The z13s memory design also provides flexibility, high availability, and upgrades:

- ▶ Concurrent memory upgrades (if the physically installed capacity is not yet reached): z13s servers can have more memory physically installed than the initial available capacity. Memory upgrades within the physically installed capacity can be done concurrently by LIC, and no hardware changes are required. However, memory upgrades *cannot* be done through CBU or On/Off CoD.

When the total capacity that is installed has more usable memory than required for a configuration, the LIC Configuration Control (LICCC) determines how much memory is used from each CPC drawer. The sum of the LICCC provided memory from each CPC drawer is the amount that is available for use in the system.

Memory allocation

When system is activated by using a POR, PR/SM determines the total installed memory and the customer enabled memory. Later in the process, during LPAR activation, PR/SM assigns and allocates each partition memory according to its image profile.

PR/SM has control over all physical memory, and can so make physical memory available to the configuration when a CPC drawer is added.

In prior z Systems processors, memory allocation was striped across the available CPC drawers because there was relatively fast connectivity between the drawers. Splitting the work between all of the memory controllers allowed a smooth performance variability.

With the z13 and z13s memory design, the memory allocation algorithm has changed. PR/SM tries to allocate memory to a single CPC drawer, striped between the two nodes. Basically, the PR/SM memory and logical processor resources allocation goal is to place all partition resources on a single CPC drawer, if possible.

The resources, like memory and logical processors, are assigned to the logical partitions at the time of their activation. Later on, when all partitions are activated, PR/SM can move memory between CPC drawers to benefit each LPARs performance, without OS knowledge. This process was done on the previous families of z Systems only for PUs that used PR/SM dynamic PU reallocation.

With z13 and z13s servers, this process occurs each time the configuration changes, including these examples:

- ▶ Activating an LPAR
- ▶ Deactivating an LPAR
- ▶ Changing the LPARs' processing weights
- ▶ Upgrading the system through a temporary or permanent record
- ▶ Downgrading the system through a temporary record

PR/SM schedules a global reoptimization of the resources in use. It does so by looking at all the partitions that are active and prioritizing them based on their processing entitlement and weights, creating a high and low priority rank. Then the resources, such as logical processors and memory, can be moved from one CPC drawer to another to address the priority ranks that have been created.

When partitions are activated, PR/SM tries to find a home assignment CPC drawer, home assignment node, and home assignment chip for the logical processors that are defined to them. The PR/SM goal is to allocate all the partition logical processors and memory to a single CPC drawer (the home drawer for that partition). If all logical processors can be assigned to a home drawer and the partition defined memory is greater than what is available in that drawer, the exceeding memory amount is allocated on another CPC drawer. If all the logical processors cannot fit in one CPC drawer, the remaining logical processors spill to another CPC drawer and, when that happens, PR/SM stripes the memory, if possible, across the CPC drawers where the logical processors are assigned.

The process of reallocating memory is based on the *memory copy/reassign* function, which was used to allow enhanced book availability (EBA) and concurrent book replacement (CBR) in previous z Systems families. However, this process has been enhanced to provide more efficiency and speed to the process without affecting system performance. PR/SM controls the reassignment of the content of a specific physical memory array in one CPC drawer to a physical memory array in another CPC drawer. To do that, PR/SM uses all the available physical memory in the system. This memory includes the memory not in use by the system, that is available but not purchased by the client, and the planned memory options, if installed.

Different sized DIMMs are available for use (16 GB, 32 GB, 64 GB, or 128 GB). On a z13s server, the CPC drawer design allows for only same-sized DIMMs to be used in one particular CPC drawer. However, two CPC drawer z13s machines can have DIMMs of a different size installed in the second CPC drawer. If the memory fails, it is technically feasible to run a POR of the system with the remaining working memory resources. After the POR completes, the

memory distribution across the CPC drawers will be different, as will be the total amount of available memory.

z13s CPC drawer memory information

On a z13s model N10 machine, the maximum amount of memory is limited to 984 GB of customer addressable memory. If more memory is needed by the customer, a model upgrade to Model N20 is mandatory. For availability and performance reasons, all available memory slots in both nodes of the same CPC drawer are populated.

On a z13s Model N20 single CPC drawer machine, the maximum amount of memory is limited to 2008 GB of customer addressable memory. If more memory is needed by the customer, a model upgrade to Model N20 two CPC drawer model is mandatory (up to 4056 GB of customer usable memory). See Table 2-4 on page 53 for memory configuration options and memory increments possible.

Note: Physical memory upgrade on a z13s server is disruptive.

Large page support

By default, page frames are allocated with a 4 KB size. z13s servers also support large page sizes of 1 MB or 2 GB. The first z/OS release that supports 1 MB pages is z/OS V1R9. Linux on z Systems support for 1 MB pages is available in SUSE Linux Enterprise Server (SLES) 10 SP2 and Red Hat Enterprise Linux (RHEL) 5.2.

The TLB exists to reduce the amount of time that is required to translate a virtual address to a real address. This translation is done by dynamic address translation (DAT) when it must find the correct page for the correct address space. Each TLB entry represents one page. Like other buffers or caches, lines are discarded from the TLB on a least recently used (LRU) basis. The worst-case translation time occurs during a TLB miss when both the segment table (which is needed to find the page table) and the page table (which is needed to find the entry for the particular page in question) are not in cache. This case involves two complete real memory access delays plus the address translation delay. The duration of a processor cycle is much shorter than the duration of a memory cycle, so a TLB miss is relatively costly.

It is preferable to have addresses in the TLB. With 4 K pages, holding all the addresses for 1 MB of storage takes 256 TLB lines. When you are using 1 MB pages, it takes only one TLB line. Therefore, large page size users have a much smaller TLB footprint.

Large pages allow the TLB to better represent a large working set and suffer fewer TLB misses by allowing a single TLB entry to cover more address translations.

Users of large pages are better represented in the TLB and can expect to see performance improvements in both elapsed time and processor usage. These improvements are because DAT and memory operations are part of processor busy time even though the processor waits for memory operations to complete without processing anything else in the meantime.

To overcome the processor usage that is associated with creating a 1 MB page, a process must run for some time. It must maintain frequent memory access to keep the pertinent addresses in the TLB.

Short-running work does not overcome the processor usage. Short processes with small working sets are expected to receive little or no improvement. Long-running work with high memory-access frequency is the best candidate to benefit from large pages.

Long-running work with low memory-access frequency is less likely to maintain its entries in the TLB. However, when it does run, a smaller number of address translations is required to

resolve all the memory it needs. Therefore, a long-running process can benefit even without frequent memory access. Weigh the benefits of whether something in this category must use large pages as a result of the system-level costs of tying up real storage. You must balance the performance of a process using large pages with the performance of the remaining work on the system.

On z13s servers, 1 MB large pages become pageable if Flash Express is available and enabled. They are available only for 64-bit virtual private storage, such as virtual memory above 2 GB.

It is easy to assume that increasing the TLB size is a feasible option to deal with TLB-miss situations. However, this process is not as straightforward as it seems. As the size of the TLB increases, so does the processor usage that is involved in managing the TLB contents. Correct sizing of the TLB is subject to complex statistical modeling to find the optimal tradeoff between size and performance.

3.6.2 Main storage

Main storage is addressable by programs, and storage that is not directly addressable by programs. Non-addressable storage includes the hardware system area (HSA).

Main storage provides these functions:

- ▶ Data storage and retrieval for PUs and I/O
- ▶ Communication with PUs and I/O
- ▶ Communication with and control of optional expanded storage
- ▶ Error checking and correction

Main storage can be accessed by all processors, but cannot be shared between LPARs. Any system image (LPAR) must have a main storage size defined. This defined main storage is allocated exclusively to the LPAR during partition activation.

3.6.3 Expanded storage

Expanded storage can optionally be defined on z13s servers. Expanded storage is physically a section of processor storage. It is controlled by the operating system, and transfers 4-KB pages to and from main storage.

Storage considerations

Except for z/VM, z/Architecture operating systems do not use expanded storage. Because they operate in 64-bit addressing mode, they can have all the required storage capacity allocated as main storage. z/VM is an exception because even when it operates in 64-bit mode, it can have guest virtual machines that are running in 31-bit addressing mode. The guest systems can use expanded storage. In addition, z/VM uses expanded storage for its own operations.

z/VM 6.3: Starting in z/VM 6.3, it is no longer preferable to configure any of the memory as expanded storage unless it is being attached to a virtual machine for testing or some unique purpose. Generally, configure all the processor memory as main storage.

Defining expanded storage to a coupling facility image is not possible. However, any other image type can have expanded storage that is defined, even if that image runs a 64-bit operating system and does not use expanded storage.

Removal of support for Expanded Storage (XSTORE): z/VM V6.3 is the last z/VM release that supports Expanded Storage (XSTORE) for either host or guest use. The IBM z13 and z13s servers are the last z Systems servers to support Expanded Storage (XSTORE).

All statements regarding IBM plans, directions, and intent are subject to change or withdrawal without notice. Any reliance on these statements of general direction is at the relying party's sole risk and will not create liability or obligation for IBM.

z13s servers can run in LPAR or Dynamic Partition Manager (DPM) mode. For DPM mode, see Appendix E, "IBM Dynamic Partition Manager" on page 501.

When running in LPAR mode, storage is placed into a single storage pool that is called the *LPAR single storage pool*. This pool can be dynamically converted to expanded storage and back to main storage as needed when partitions are activated or deactivated.

LPAR single storage pool

In LPAR mode, storage is not split into main storage and expanded storage at POR. Rather, the storage is placed into a single main storage pool that is dynamically assigned to expanded storage and back to main storage, as needed.

On the HMC, the storage assignment tab of a reset profile shows the customer storage. *Customer storage* is the total installed storage. LPARs are still defined to have main storage and, optionally, expanded storage.

Activation of LPARs and dynamic storage reconfiguration cause the storage to be assigned to the type that is needed (central or expanded). It does not require a POR.

3.6.4 Hardware system area (HSA)

The HSA is a non-addressable storage area that contains system LIC and configuration-dependent control blocks. On z13s servers, the HSA has a fixed size of 40 GB and is not part of the purchased memory that customers order and is installed by IBM.

The fixed size of the HSA eliminates planning for future expansion of the HSA because the hardware configuration definition (HCD) input/output configuration program (IOCP) always reserves space for the following items:

- ▶ Three channel subsystems (CSSs)
- ▶ 15 LPARs in the first two CSSs and 10 LPARs for the third CSS for a total of 40 LPARs
- ▶ Subchannel set 0 with 63.75-K devices in each CSS
- ▶ Subchannel set 1 with 64-K devices in each CSS
- ▶ Subchannel set 2 with 64-K devices in each CSS

The HSA has sufficient reserved space to allow for dynamic I/O reconfiguration changes to the maximum capability of the processor.

3.7 Logical partitioning

This section addresses logical partitioning features.

3.7.1 Overview

Logical partitioning is a function that is implemented by the PR/SM on z13s servers. The z13s server runs either in LPAR-, or Dynamic Partition Manager mode. Therefore, all system aspects are controlled by PR/SM functions.

PR/SM is aware of the CPC drawer structure on the z13s server. However, LPARs do not have this awareness. LPARs have resources that are allocated to them from various physical resources. From a systems standpoint, LPARs have no control over these physical resources, but the PR/SM functions do.

PR/SM manages and optimizes allocation and the dispatching of work on the physical topology. Most physical topology that was previously handled by the operating systems is now the responsibility of PR/SM.

As shown in 3.5.9, “Processor unit assignment” on page 110, the initial PU assignment is done during POR by using rules to optimize cache usage. This is the “physical” step, where CPs, zIIPs, IFLs, ICFs, and SAPs are allocated on the CPC drawers.

When an LPAR is activated, PR/SM builds logical processors and allocates memory for the LPAR.

Memory allocation has changed from the previous z Systems. IBM System z9® memory used to be spread across all books. This optimization was done by using a round-robin algorithm with a number of increments per book to match the number of memory controllers (MCs) per book. This memory allocation design is driven by performance results, with minimized variability for most workloads.

With z13 and z13s servers, memory allocation has changed from the model that was used for the z9. Partition memory is now allocated in a per-CPC-drawer basis and striped across processor nodes. For more information, see “Memory allocation” on page 112.

Logical processors are dispatched by PR/SM on physical processors. The assignment topology that is used by PR/SM to dispatch logical processors on physical PUs is also based on cache usage optimization.

CPC drawers and node level assignments are more important because they optimize L4 cache usage. Therefore, logical processors from a specific LPAR are packed into a CPC drawer as often as possible.

Then, PR/SM optimizes chip assignments within the assigned CPC drawer (or drawers) to maximize L3 cache efficiency. Logical processors from an LPAR are dispatched on physical processors on the same PU chip as much as possible. The number of processors per chip (eight) matches the number of z/OS processor affinity queues that is used by HiperDispatch, achieving optimal cache usage within an affinity node.

PR/SM also tries to redispach a logical processor on the same physical processor to optimize private cache (L1 and L2) usage.

Removal of an option for the way shared logical processors are managed under

PR/SM LPAR: IBM z13 and z13s servers will be the last z Systems servers to support selection of the option to “Do not end the time slice if a partition enters a wait state” when the option to set a processor runtime value has been previously selected in the CPC RESET profile. The CPC RESET profile applies to all shared logical partitions on the system, and is not selectable by logical partition.

All statements regarding IBM plans, directions, and intent are subject to change or withdrawal without notice. Any reliance on these statements of general direction is at the relying party's sole risk and will not create liability or obligation for IBM.

HiperDispatch

PR/SM and z/OS work in tandem to use processor resources more efficiently. HiperDispatch is a function that combines the dispatcher actions and the knowledge that PR/SM has about the topology of the system.

Performance can be optimized by redispersing units of work to the same processor group, keeping processes running near their cached instructions and data, and minimizing transfers of data ownership among processors and CPC drawers.

The nested topology is returned to z/OS by the Store System Information (STSI) instruction. HiperDispatch uses the information to concentrate logical processors around shared caches (L3 at PU chip level, and L4 at CPC drawer level), and dynamically optimizes the assignment of logical processors and units of work.

z/OS dispatcher manages multiple queues, called *affinity queues*, with a target number of eight processors per queue, which fits nicely onto a single PU chip. These queues are used to assign work to as few logical processors as are needed for an LPAR workload. So, even if the LPAR is defined with many logical processors, HiperDispatch optimizes this number of processors to be near the required capacity. The optimal number of processors to be used is kept within a CPC drawer boundary where possible.

Tip: HiperDispatch is now also supported by z/VM V6.3.

Logical partitions (LPARs)

PR/SM enables the z13s servers to be initialized for a logically partitioned operation, supporting up to 40 LPARs. Each LPAR can run its own operating system image in any image mode, independently from the other LPARs.

An LPAR can be added, removed, activated, or deactivated at any time. Changing the number of LPARs is not disruptive and does not require POR. Certain facilities might not be available to all operating systems because the facilities might have software corequisites.

Each LPAR has the same resources as a real CPC:

- ▶ **Processors:** Called *logical processors*, they can be defined as CPs, IFLs, ICFs, or zIIPs. They can be dedicated to an LPAR or shared among LPARs. When shared, a processor weight can be defined to provide the required level of processor resources to an LPAR. Also, the capping option can be turned on, which prevents an LPAR from acquiring more than its defined weight, limiting its processor consumption.

LPARs for z/OS can have CP and zIIP logical processors. The two logical processor types can be defined as either all dedicated or all shared. The zIIP support is available in z/OS.

The weight and number of online logical processors of an LPAR can be dynamically managed by the LPAR CPU Management function of the Intelligent Resource Director (IRD). These processors can then be used to achieve the defined goals of this specific partition and of the overall system. The provisioning architecture of the servers, described in Chapter 8, “System upgrades” on page 311, adds another dimension to the dynamic management of LPARs.

PR/SM is enhanced to support an option to limit the amount of physical processor capacity that is consumed by an individual LPAR when a PU is defined as a general-purpose processor (CP) or an IFL shared across a set of LPARs.

This enhancement is designed to provide a physical capacity limit that is enforced as an absolute (versus relative) limit. It is not affected by changes to the logical or physical configuration of the system. This physical capacity limit can be specified in units of CPs or IFLs. The “Change LPAR Controls” and “Customize Activation Profiles” tasks on the Hardware Management Console have been enhanced to support this new function.

For the z/OS Workload License Charges (WLC) pricing metric, and metrics that are based on it, such as Advanced Workload License Charges (AWLC), an LPAR *defined capacity* can be set. This defined capacity enables the soft capping function. Workload charging introduces the capability to pay software license fees based on the processor utilization of the LPAR on which the product is running, rather than on the total capacity of the system:

- In support of WLC, the user can specify a defined capacity in MSU per hour. The defined capacity sets the capacity of an individual LPAR when soft capping is selected. The defined capacity value is specified on the Options tab in the Customize Image Profiles window.
- WLM keeps a 4-hour rolling average of the processor usage of the LPAR. When the 4-hour average processor consumption exceeds the defined capacity limit, WLM dynamically activates LPAR capping (soft capping). When the rolling 4-hour average returns below the defined capacity, the soft cap is removed.

For more information about WLM, see *System Programmer's Guide to: Workload Manager*, SG24-6472. For a review of software licensing, see 7.16.1, “Software licensing considerations” on page 305.

Weight settings: When defined capacity is used to define an uncapped LPAR's capacity, carefully consider the weight settings of that LPAR. If the weight is much smaller than the defined capacity, PR/SM uses a discontinuous cap pattern to achieve the defined capacity setting. This configuration means PR/SM alternates between capping the LPAR at the MSU value corresponding to the relative weight settings, and no capping at all. It is best to avoid this scenario, and try to establish a defined capacity that is equal or close to the relative weight.

- **Memory:** Memory, either main storage or expanded storage, must be dedicated to an LPAR. The defined storage must be available during the LPAR activation. Otherwise, the LPAR activation fails.

Reserved storage can be defined to an LPAR, enabling nondisruptive memory addition to and removal from an LPAR, by using the LPAR dynamic storage reconfiguration (z/OS and z/VM). For more information, see 3.7.5, “LPAR dynamic storage reconfiguration” on page 129.

- Channels: Channels can be shared between LPARs by including the partition name in the partition list of a channel-path identifier (CHPID). I/O configurations are defined by the IOCP or the HCD with the CHPID mapping tool (CMT). The CMT is an optional tool that is used to map CHPIDs onto physical channel IDs (PCHIDs). PCHIDs represent the physical location of a port on a card in an I/O cage, I/O drawer, or PCIe I/O drawer.

IOCP⁴ is available on the z/OS, z/VM, and z/VSE operating systems, and as a stand-alone program on the hardware console. HCD is available on the z/OS and z/VM operating systems. Consult the appropriate 2965DEVICE Preventive Service Planning (PSP) buckets before implementation.

Fibre Channel connection (FICON) channels can be managed by the Dynamic CHPID Management (DCM) function of the Intelligent Resource Director. DCM enables the system to respond to ever-changing channel requirements by moving channels from lesser-used control units to more heavily used control units, as needed.

Modes of operation

Table 3-4 shows the modes of operation, summarizing all available mode combinations, including their operating modes and processor types, operating systems, and addressing modes. Only the currently supported versions of operating systems are considered.

Table 3-4 z13s modes of operation

Image mode	PU type	Operating system	Addressing mode
ESA/390	CP and zIIP	z/OS z/VM	64-bit
	CP	z/VSE and Linux on z Systems (64-bit)	64-bit
	CP	Linux on z Systems (31-bit)	31-bit
ESA/390 TPF	CP only	z/TPF	64-bit
Coupling facility	ICF or CP	CFCC	64-bit
Linux only	IFL or CP	Linux on z Systems (64-bit)	64-bit
		z/VM	
		Linux on z Systems (31-bit)	31-bit
z/VM	CP, IFL, zIIP, or ICF	z/VM	64-bit
zACI ^a	IFL or CP	IBM zAware	64-bit
		z/VSE Network Appliance ^b	

a. z Appliance Container Infrastructure

b. Planned; will be supported

The 64-bit z/Architecture mode has no special operating mode because the architecture mode is not an attribute of the definable images operating mode. The 64-bit operating systems are in 31-bit mode at IPL and, optionally, can change to 64-bit mode during their initialization. The operating system is responsible for taking advantage of the addressing capabilities that are provided by the architectural mode.

For more information about operating system support, see Chapter 7, “Software support” on page 229.

⁴ For more information, see *z Systems Input/Output Configuration Program User's Guide for ICP IOCP*, SB10- 7163.

Logically partitioned mode

If the z13s server runs in LPAR mode, each of the 40 LPARs can be defined to operate in one of the following image modes:

- ▶ ESA/390 mode to run the following systems:
 - A z/Architecture operating system, on dedicated or shared CPs
 - An ESA/390 operating system, on dedicated or shared CPs
 - A Linux on z Systems operating system, on dedicated or shared CPs
 - z/OS, on any of the following processor units:
 - Dedicated or shared CPs
 - Dedicated CPs and dedicated zIIPs
 - Shared CPs and shared zIIPs

zIIP usage: zIIPs can be defined to an ESA/390 mode or z/VM mode image, as shown in Table 3-4 on page 120. However, zIIPs are used only by z/OS. Other operating systems cannot use zIIPs even if they are defined to the LPAR. z/VM V5R4 and later provide support for real and virtual zIIPs to guest z/OS systems. However, z/VM V5R4 is not supported on IBM z13s servers.

- ▶ ESA/390 TPF mode to run the z/TPF operating system, on dedicated or shared CPs
- ▶ Coupling facility mode, by loading the CFCC code into the LPAR that is defined as one of these types:
 - Dedicated or shared CPs
 - Dedicated or shared ICFs
- ▶ Linux only mode to run these systems:
 - A Linux on z Systems operating system, on either of these types:
 - Dedicated or shared IFLs
 - Dedicated or shared CPs
 - A z/VM operating system, on either of these types:
 - Dedicated or shared IFLs
 - Dedicated or shared CPs
- ▶ z/VM mode to run z/VM on dedicated or shared CPs or IFLs, plus zIIPs and ICFs
- ▶ z Appliance Container Infrastructure (zACI) mode can run on these types:
 - Dedicated or shared CPs
 - Dedicated or shared IFLs

Table 3-5 shows all LPAR modes, required characterized PUs, operating systems, and the PU characterizations that can be configured to an LPAR image. The available combinations of dedicated (DED) and shared (SHR) processors are also shown. For all combinations, an LPAR also can have reserved processors that are defined, allowing nondisruptive LPAR upgrades.

Table 3-5 LPAR mode and PU usage

LPAR mode	PU type	Operating systems	PUs usage
ESA/390	CPs	z/Architecture operating systems ESA/390 operating systems Linux on z Systems	CPs DED or CPs SHR
	CPs <i>and</i> zIIPs	z/OS z/VM (V6R2 and later for guest exploitation)	CPs DED <i>and (or)</i> zIIPs DED <i>or</i> CPs SHR <i>and (or)</i> zIIPs SHR
ESA/390 TPF	CPs	z/TPF	CPs DED or CPs SHR
Coupling facility	ICFs <i>or</i> CPs	CFCC	ICFs DED <i>or</i> ICFs SHR, <i>or</i> CPs DED <i>or</i> CPs SHR
Linux only	IFLs <i>or</i> CPs	Linux on z Systems z/VM	IFLs DED <i>or</i> IFLs SHR, <i>or</i> CPs DED <i>or</i> CPs SHR
z/VM	CPs, IFLs, zIIPs, <i>or</i> ICFs	z/VM (V6R2 and later)	All PUs must be SHR or DED
zACI ^a	IFLs, <i>or</i> CPs	IBM zAware ^b IBM z/VSE Network Appliance ^c	IFLs DED <i>or</i> IFLs SHR, <i>or</i> CPs DED <i>or</i> CPs SHR

- a. z Appliance Container Infrastructure
- b. Encapsulated as a firmware appliance.
- c. Planned

Dynamically adding or deleting a logical partition name

Dynamically adding or deleting an LPAR name is the ability to add or delete LPARs and their associated I/O resources to or from the configuration without a POR.

The extra channel subsystem and multiple image facility (MIF) image ID pairs (CSSID/MIFID) can be later assigned to an LPAR for use (or later removed). This process can be done through dynamic I/O commands by using HCD. At the same time, required channels must be defined for the new LPAR.

Partition profile: Cryptographic coprocessors are not tied to partition numbers or MIF IDs. They are set up with Adjunct Processor (AP) numbers and domain indexes. These are assigned to a partition profile of a given name. The client assigns these AP numbers and domains to the partitions and continues to have the responsibility to clear them out when their profiles change.

Adding logical processors to a logical partition

Logical processors can be concurrently added to an LPAR by defining them as reserved in the image profile and later configuring them online to the operating system by using the appropriate console commands. Logical processors can also be concurrently added to a logical partition dynamically by using the Support Element “Logical Processor Add” function under the “CPC Operational Customization” task. This SE function allows the initial and reserved processor values to be dynamically changed. The operating system must support

the dynamic addition of these resources. In z/OS, this support is available since Version 1 Release 10 (z/OS V1.10).

Adding a crypto feature to a logical partition

You can plan the addition of Crypto Express5S features to an LPAR on the crypto page in the image profile by defining the Cryptographic Candidate List, and the Usage and Control Domain indexes, in the partition profile. By using the “Change LPAR Cryptographic Controls” task, you can add crypto adapters dynamically to an LPAR without an outage of the LPAR. Also, dynamic deletion or moving of these features does not require pre-planning. Support is provided in z/OS, z/VM, z/VSE, and Linux on z Systems.

LPAR dynamic PU reassignment

The system configuration is enhanced to optimize the PU-to-CPC drawer assignment of physical processors dynamically. The initial assignment of client-usable physical processors to physical CPC drawers can change dynamically to better suit the actual LPAR configurations that are in use. For more information, see 3.5.9, “Processor unit assignment” on page 110. Swapping of specialty engines and general processors with each other, with spare PUs, or with both, can occur as the system attempts to change LPAR configurations into physical configurations that span the least number of CPC drawers.

LPAR dynamic PU reassignment can swap client processors of different types between CPC drawers. For example, reassignment can swap an IFL on CPC drawer 0 with a CP on CPC drawer 1. Swaps can also occur between PU chips within a CPC drawer or a node, and can include spare PUs. The goals are to pack the LPAR on fewer CPC drawers and also on fewer PU chips, based on the z13s CPC drawers’ topology. The effect of this process is evident in dedicated and shared LPARs that use HiperDispatch.

LPAR dynamic PU reassignment is transparent to operating systems.

LPAR group capacity limit

The group capacity limit feature allows the definition of a group of LPARs on a z13s system, and limits the combined capacity usage by those LPARs. This process allows the system to manage the group so that the group capacity limits in MSUs per hour are not exceeded. To take advantage of this feature, you must be running z/OS V1.10 or later in the all LPARs in the group.

PR/SM and WLM work together to enforce the capacity that is defined for the group and the capacity that is optionally defined for each individual LPAR.

LPAR absolute capping

Absolute capping is a new logical partition control that was made available with zEC12 and is supported on z13s servers. With this support, PR/SM and the HMC are enhanced to support a new option to limit the amount of physical processor capacity that is consumed by an individual LPAR when a PU is defined as a general-purpose processor (CP), zIIP, or an IFL processor that is shared across a set of LPARs.

Unlike traditional LPAR capping, absolute capping is designed to provide a physical capacity limit that is enforced as an absolute (versus relative) value that is not affected by changes to the virtual or physical configuration of the system.

Absolute capping provides an optional maximum capacity setting for logical partitions that is specified in terms of the absolute processor capacity (for example, 5.00 CPs or 2.75 IFLs). This setting is specified independently by processor type (namely CPs, zIIPs, and IFLs) and

provides an enforceable upper limit on the amount of the specified processor type that can be used in a partition.

Absolute capping is ideal for processor types and operating systems that the z/OS WLM cannot control. Absolute capping is not intended as a replacement for defined capacity or group capacity for z/OS, which are managed by WLM.

Absolute capping can be used with any z/OS, z/VM, or Linux on z LPAR running on zEnterprise. If specified for a z/OS LPAR, it can be used concurrently with defined capacity or group capacity management for z/OS. When used concurrently, the absolute capacity limit becomes effective before other capping controls.

Dynamic Partition Manager mode

DPM is a new z Systems mode of operation that provides a simplified approach to create and manage virtualized environments, reducing the barriers of its adoption for new and existing customers.

The implementation provides built-in integrated capabilities that allow advanced virtualization management on z Systems. With DPM, customers can use their existing Linux and virtualization skills while using the full value of z Systems hardware, robustness, and security in a workload-optimized environment.

DPM provides facilities to define and run virtualized computing systems, using a firmware-managed environment, that coordinates the physical system resources shared by the partitions. The partitions' resources include processors, memory, network, storage, crypto, and accelerators.

DPM provides a new mode of operation for z Systems that provides these capabilities:

- ▶ Facilitates defining, configuring, and operating PR/SM LPARs in a similar way to how someone performs these tasks on another platform.
- ▶ Lays the foundation for a general z Systems new user experience.

DPM is not an extra hypervisor for z Systems. DPM uses the existing PR/SM hypervisor infrastructure and on top of it, provides an intelligent interface that allows customers to define, use, and operate the platform virtualization without prior z Systems experience or skills. For more information about DPM, see Appendix E, "IBM Dynamic Partition Manager" on page 501.

3.7.2 Storage operations

In z13s servers, memory can be assigned as a combination of main storage and expanded storage, supporting up to 40 LPARs. Expanded storage is used only by the z/VM operating system.

Removal of support for Expanded Storage (XSTORE): z/VM V6.3 is the last z/VM release that supports Expanded Storage (XSTORE) for either host or guest use. The IBM z13 and z13s servers are the last z Systems servers to support Expanded Storage (XSTORE).

All statements regarding IBM plans, directions, and intent are subject to change or withdrawal without notice. Any reliance on these statements of general direction is at the relying party's sole risk and will not create liability or obligation for IBM.

Before you activate an LPAR, main storage (and optionally, expanded storage) must be defined to the LPAR. All installed storage can be configured as main storage. Each z/OS individual LPAR can be defined with a maximum of 4 TB of main storage. z/VM supports 1 TB of main storage.

Main storage can be dynamically assigned to expanded storage and back to main storage as needed without a POR. For more information, see “LPAR single storage pool” on page 116.

Memory *cannot* be shared between system images. It is possible to reallocate dynamically storage resources for z/Architecture LPARs that run operating systems that support dynamic storage reconfiguration (DSR). This process is supported by z/OS, and z/VM V5R4 and later releases. z/VM, in turn, virtualizes this support to its guests. For more information, see 3.7.5, “LPAR dynamic storage reconfiguration” on page 129.

Operating systems that run as guests of z/VM can use the z/VM capability of implementing virtual memory to guest virtual machines. The z/VM dedicated real storage can be shared between guest operating systems.

Table 3-6 shows the z13 storage allocation and usage possibilities, depending on the image mode.

Table 3-6 Main storage definition and usage possibilities

Image mode	Architecture mode (addressability)	Maximum main storage		Expanded storage	
		Architecture	z13s definition	z13s definable	Operating system usage ^a
ESA/390	z/Architecture (64-bit)	16 EB	4 TB	Yes	Yes
	ESA/390 (31-bit)	2 GB	128 GB	Yes	Yes
ESA/390 TPF	ESA/390 (31-bit)	2 GB	2 GB	Yes	No
Coupling facility	CFCC (64-bit)	1.5 TB	1 TB	No	No
Linux only	z/Architecture (64-bit)	16 EB	1 TB	Yes	<i>Only by z/VM</i>
	ESA/390 (31-bit)	2 GB	2 GB	Yes	<i>Only by z/VM</i>
z/VM ^b	z/Architecture (64-bit)	16 EB	1 TB	Yes	Yes
zACI	z/Architecture (64-bit)	16 EB	1 TB	Yes	No

a. z/VM supports the use of expanded storage, but expanded storage is not recommended for z/VM V6.3.

b. z/VM mode is supported by z/VM V6R2 and later.

Note: When an I/O drawer (carry forward only) is installed, the maximum amount of LPAR memory is limited to 1 TB.

ESA/390 mode

In ESA/390 mode, storage addressing can be 31 or 64 bits, depending on the operating system architecture and the operating system configuration.

An ESA/390 mode image is always initiated in 31-bit addressing mode. During its initialization, a z/Architecture operating system can change it to 64-bit addressing mode and operate in the z/Architecture mode.

Certain z/Architecture operating systems, such as z/OS, *always* change the 31-bit addressing mode and operate in 64-bit mode. Other z/Architecture operating systems, such as z/VM, can be configured to change to 64-bit mode or to stay in 31-bit mode and operate in the ESA/390 architecture mode.

The following modes are provided:

- ▶ **z/Architecture mode:** In z/Architecture mode, storage addressing is 64-bit, allowing for virtual addresses up to 16 exabytes (16 EB). The 64-bit architecture theoretically allows a maximum of 16 EB to be used as main storage. However, the current main storage limit for LPARs is 10 TB of main storage on a z13s server. The operating system that runs in z/Architecture mode must be able to support the real storage. Currently, z/OS supports up to 4 TB⁵ of real storage (z/OS V1R10 and later releases).

Expanded storage also can be configured to an image that runs an operating system in z/Architecture mode. However, only z/VM can use expanded storage. Any other operating system that runs in z/Architecture mode (such as a z/OS or a Linux on z Systems image) *does not* address the configured expanded storage. This expanded storage remains configured to this image and is not used.

- ▶ **ESA/390 architecture mode:** In ESA/390 architecture mode, storage addressing is 31 bit, allowing for virtual addresses up to 2 GB. A maximum of 2 GB can be used for main storage. Because the processor storage can be configured as central and expanded storage, memory above 2 GB can be configured as expanded storage. In addition, this mode allows the use of either 24-bit or 31-bit addressing, under program control.

Because an ESA/390 mode image can be defined with up to 128 GB of main storage, the main storage above 2 GB is *not* used. Instead, it remains configured to this image.

Storage resources: Either a z/Architecture mode or an ESA/390 architecture mode operating system can run in an ESA/390 image on a z13s server. Any ESA/390 image can be defined with more than 2 GB of main storage, and can have expanded storage. These options allow you to configure more storage resources than the operating system can address.

ESA/390 TPF mode

In ESA/390 TPF mode, storage addressing follows the ESA/390 architecture mode. The z/TPF operating system runs in 64-bit addressing mode.

Coupling facility mode

In coupling facility mode, storage addressing is 64 bit for a coupling facility image that runs at CFCC Level 12 or later. This configuration allows for an addressing range up to 16 EB. However, the current z13s definition limit for LPARs is 1 TB of storage.

CFCC Level 21, which is available for z13s servers with driver level 27, introduces several enhancements in the performance, reporting, and serviceability areas.

CFCC Level 19, which is available for the zEC12 with driver level 15F, introduces several improvements in the performance and resiliency areas, including Coupling Thin Interrupts and Flash Express.

For more information, see 3.9.1, “Coupling Facility Control Code” on page 132. Expanded storage cannot be defined for a coupling facility image. Only IBM CFCC can run in the coupling facility mode.

⁵ 1 TB if an I/O drawer is installed in the system (carry forward only).

Linux only mode

In Linux only mode, storage addressing can be 31 bit or 64 bit, depending on the operating system architecture and the operating system configuration, in the same way as in ESA/390 mode.

Only Linux and z/VM operating systems can run in Linux only mode. Linux on z Systems 64-bit distributions (SUSE Linux Enterprise Server 10 and later, and Red Hat RHEL 5 and later) use 64-bit addressing and operate in z/Architecture mode. z/VM also uses 64-bit addressing and operates in z/Architecture mode.

z/VM mode

In z/VM mode, certain types of processor units can be defined within one LPAR. This increases flexibility and simplifies systems management by allowing z/VM to run the following tasks in the same z/VM LPAR:

- ▶ Manage guests to operate Linux on z Systems on IFLs
- ▶ Operate z/VSE and z/OS on CPs
- ▶ Offload z/OS system software processor usage, such as DB2 workloads on zIIPs
- ▶ Provide an economical Java execution environment under z/OS on zIIPs

zACI mode

In zACI mode, storage addressing is 64 bit for an IBM zAware image running IBM z Advanced Workload Analysis Reporter firmware. This configuration allows for an addressing range up to 16 EB. However, the current z13s definition limit for LPARs is 4 TB of storage.

Currently, the IBM zAware Version 2 feature, which is available on z13 and z13s servers, runs in a zACI LPAR.

Important: The IBM zAware LPAR mode has been replaced for z13s and z13 servers at Driver Level 27 with zACI. Existing IBM zAware LPARs will be automatically converted during Enhanced Driver Maintenance from Driver 22 to Driver 27. No reconfiguration of IBM zAware is required.

For more information, see Appendix B, “IBM z Systems Advanced Workload Analysis Reporter (IBM zAware)” on page 453.

3.7.3 Reserved storage

Reserved storage can be optionally defined to an LPAR, allowing a nondisruptive image memory upgrade for this partition. Reserved storage can be defined to both central and expanded storage, and to any image mode except coupling facility mode.

An LPAR must define an amount of main storage and, optionally (if not a coupling facility image), an amount of expanded storage. Both main storage and expanded storage can have two storage sizes defined:

- ▶ The initial value is the storage size that is allocated to the partition when it is activated.
- ▶ The reserved value is an additional storage capacity beyond its initial storage size that an LPAR can acquire dynamically. The reserved storage sizes that are defined to an LPAR do not have to be available when the partition is activated. They are predefined storage sizes to allow a storage increase, from an LPAR point of view.

Without the reserved storage definition, an LPAR storage upgrade is a disruptive process that requires the following actions:

1. Partition deactivation
2. An initial storage size definition change
3. Partition activation

The additional storage capacity for an LPAR upgrade can come from these sources:

- ▶ Any unused available storage
- ▶ Another partition that has released storage
- ▶ A memory upgrade

A concurrent LPAR storage upgrade uses DSR. z/OS uses the reconfigurable storage unit (RSU) definition to add or remove storage units in a nondisruptive way.

z/VM V6R2 and later releases support the dynamic addition of memory to a running LPAR by using reserved storage. It also virtualizes this support to its guests. Removal of storage from the guests or z/VM is disruptive.

SUSE Linux Enterprise Server 11 supports both concurrent add and remove.

3.7.4 Logical partition storage granularity

Granularity of main storage for an LPAR depends on the largest main storage amount that is defined for either initial or reserved main storage, as shown in Table 3-7.

Table 3-7 Logical partition main storage granularity (z13s servers)

Logical partition: Largest main storage amount	Logical partition: Main storage granularity
Main storage amount <= 256 GB	512 MB
256 GB < main storage amount <= 512 GB	1 GB
512 GB < main storage amount <= 1 TB	2 GB
1 TB < main storage amount <= 2 TB	4 GB
2 TB < main storage amount <= 4 TB	8 GB

The granularity applies across all main storage that is defined, both initial and reserved. For example, for an LPAR with an initial storage amount of 30 GB and a reserved storage amount of 48 GB, the main storage granularity of both initial and reserved main storage is 512 MB. Expanded storage granularity is fixed at 512 MB.

Removal of support for Expanded Storage (XSTORE): The IBM z13 and z13s servers are the last z Systems servers to support Expanded Storage (XSTORE).

All statements regarding IBM plans, directions, and intent are subject to change or withdrawal without notice. Any reliance on these statements of general direction is at the relying party's sole risk and will not create liability or obligation for IBM.

LPAR storage granularity information is required for LPAR image setup and for z/OS RSU definition. LPARs are limited to a maximum size of 10 TB of main storage (4 TB on z13s servers). For z/VM V6R2, the limit is 256 GB, whereas for z/VM V6.3, the limit is 1 TB.

3.7.5 LPAR dynamic storage reconfiguration

Dynamic storage reconfiguration on z13s servers allows an operating system that is running on an LPAR to add (nondisruptively) its reserved storage amount to its configuration. This process can occur only if unused storage exists. This unused storage can be obtained when another LPAR releases storage, or when a concurrent memory upgrade takes place. With dynamic storage reconfiguration, the unused storage does not have to be continuous.

When an operating system running on an LPAR assigns a storage increment to its configuration, PR/SM determines whether any free storage increments are available. PR/SM then dynamically brings the storage online. PR/SM dynamically takes offline a storage increment and makes it available to other partitions when an operating system running on an LPAR releases a storage increment.

3.8 Intelligent Resource Director

IRD is a z13, z13s, and z Systems capability that is used only by z/OS. IRD is a function that optimizes processor and channel resource utilization across LPARs within a single z Systems server.

This feature extends the concept of goal-oriented resource management. It does so by grouping system images that are on the same z13, z13s or z Systems servers running in LPAR mode, and in the same Parallel Sysplex, into an *LPAR cluster*. This configuration allows WLM to manage resources, both processor and I/O, not just in one single image, but across the entire cluster of system images.

Figure 3-15 shows an LPAR cluster. It contains three z/OS images, and one Linux image that are managed by the cluster. Included as part of the entire Parallel Sysplex is another z/OS image and a coupling facility image. In this example, the scope over which IRD has control is the defined LPAR cluster.

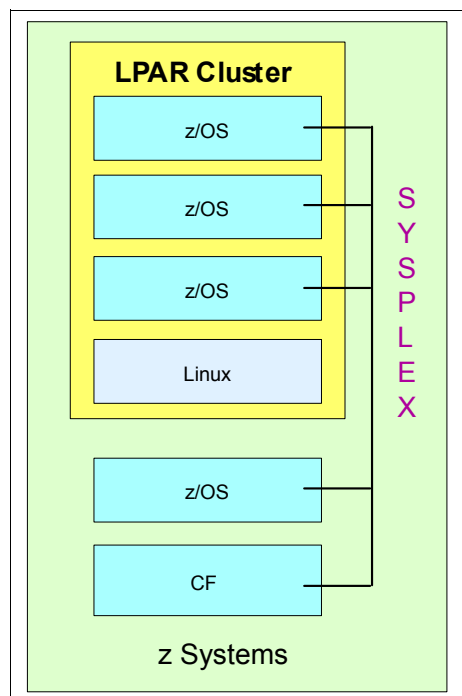


Figure 3-15 IRD LPAR cluster example

IRD has the following characteristics:

- ▶ IRD processor management: WLM dynamically adjusts the number of logical processors within an LPAR and the processor weight based on the WLM policy. The ability to move the processor weights across an LPAR cluster provides processing power where it is most needed, based on WLM goal mode policy.

The processor management function is automatically deactivated when HiperDispatch is active. However, the LPAR weight management function remains active with IRD with HiperDispatch. For more information about HiperDispatch, see 3.7, “Logical partitioning” on page 116.

HiperDispatch manages the number of logical CPs in use. It adjusts the number of logical processors within an LPAR to achieve the optimal balance between CP resources and the requirements of the workload.

HiperDispatch also adjusts the number of logical processors. The goal is to map the logical processor to as few physical processors as possible. This configuration uses the processor resources more efficiently by trying to stay within the local cache structure. Doing so makes efficient use of the advantages of the high-frequency microprocessors, and improves throughput and response times.

- ▶ Dynamic channel path management (DCM): DCM moves FICON channel bandwidth between disk control units to address current processing needs. z13s servers support DCM within a channel subsystem.
- ▶ Channel subsystem priority queuing: This function on the z13, z13s, and z Systems servers allow the priority queuing of I/O requests in the channel subsystem and the specification of relative priority among LPARs. When running in goal mode, WLM sets the priority for an LPAR and coordinates this activity among clustered LPARs.

For more information about implementing LPAR processor management under IRD, see *z/OS Intelligent Resource Director*, SG24-5952.

3.9 Clustering technology

Parallel Sysplex is the clustering technology that is used with z13 and z13s servers. Figure 3-16 illustrates the components of a Parallel Sysplex as implemented within the z Systems architecture. The figure shows one of many possible Parallel Sysplex configurations.

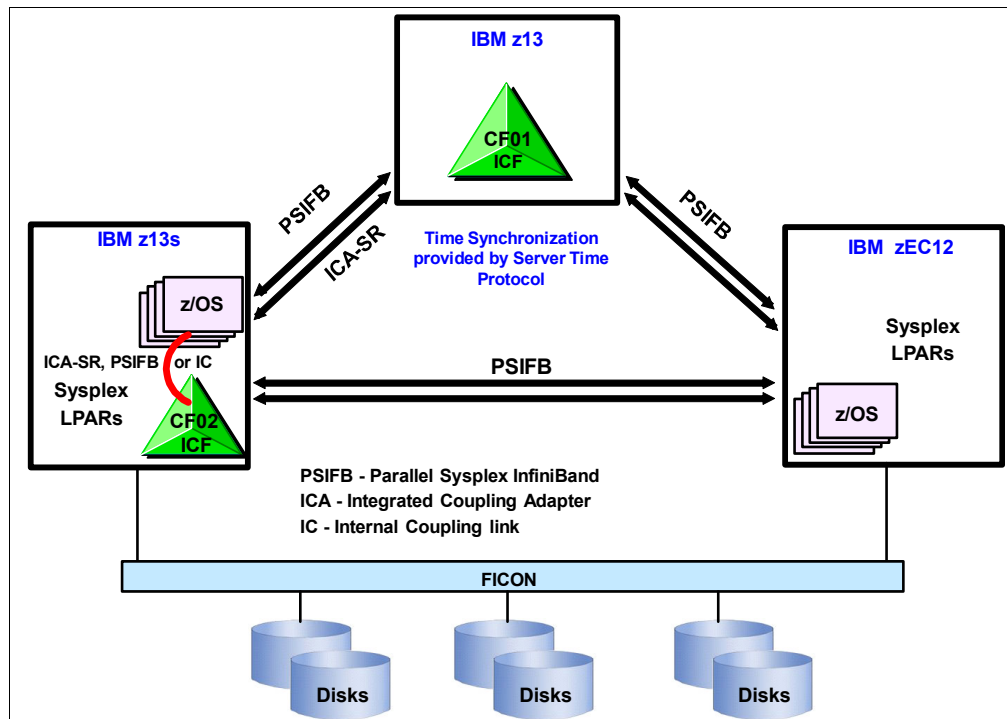


Figure 3-16 Sysplex hardware overview

Figure 3-16 shows a z13s system that contains multiple z/OS sysplex partitions. It contains an internal coupling facility (CF02), a z13 system that contains a stand-alone CF (CF01), and a zEC12 that contains multiple z/OS sysplex partitions. STP over coupling links provides time synchronization to all systems. The appropriate CF link technology (ICA SR, 1x IFB, or 12x IFB) selection depends on the system configuration and how distant they are physically. The ISC-3 coupling link is not supported on z13 and z13s servers. For more information about link technologies, see 4.9.1, “Coupling links” on page 174.

Parallel Sysplex technology is an enabling technology, allowing highly reliable, redundant, and robust z Systems technology to achieve near-continuous availability. A Parallel Sysplex makes up one or more (z/OS) operating system images that are coupled through one or more Coupling Facilities. The images can be combined together to form clusters.

A correctly configured Parallel Sysplex cluster maximizes availability in these ways:

- ▶ Continuous (application) availability: Changes can be introduced, such as software upgrades, one image at a time, while the remaining images continue to process work. For more information, see *Parallel Sysplex Application Considerations*, SG24-6523.
- ▶ High capacity: Scales can be 2 - 32 images.
- ▶ Dynamic workload balancing: Because it is viewed as a single logical resource, work can be directed to any similar operating system image in a Parallel Sysplex cluster that has available capacity.

- ▶ **Systems management:** The architecture provides the infrastructure to satisfy client requirements for continuous availability, and provides techniques for achieving simplified systems management consistent with this requirement.
- ▶ **Resource sharing:** A number of base (z/OS) components use the coupling facility shared storage. This configuration enables sharing of physical resources with significant improvements in cost, performance, and simplified systems management.
- ▶ **Single system image:** The collection of system images in the Parallel Sysplex is displayed as a single entity to the operator, the user, and the database administrator. A single system image ensures reduced complexity from both operational and definition perspectives.
- ▶ **N-2 support:** Multiple hardware generations (normally three) are supported in the same Parallel Sysplex. This configuration provides for a gradual evolution of the systems in the Parallel Sysplex without having to change all of them simultaneously. Similarly, software support for multiple releases or versions is supported.

Through state-of-the-art cluster technology, the power of multiple images can be harnessed to work in concert on common workloads. The z Systems Parallel Sysplex cluster takes the commercial strengths of the platform to greater levels of system management, competitive price for performance, scalable growth, and continuous availability.

3.9.1 Coupling Facility Control Code

The LPAR running the CFCC can be on z13, z13s, zEC12, zBC12, z196, or z114 systems. For more information about CFCC requirements for supported systems, see 7.8, “Coupling facility and CFCC considerations” on page 292.

Consideration: z13 and z13s systems cannot coexist in the same sysplex with System z10 and previous systems. The introduction of z13s servers into existing installations might require more planning.

CFCC Level 21

CFCC level 21 is delivered on z13s servers with driver level 27. CFCC Level 21 introduces the following enhancements:

- ▶ Asynchronous CF Duplexing for lock structures.
 - z/OS V2.2 SPE with PTFs for APAR OA47796, DB2 V12 with PTFs
- ▶ Enable systems management applications to collect valid CF LPAR information by using z/OS BCPII
 - System Type (CFCC), System Level (CFCC LEVEL)
 - Dynamic Dispatch settings to indicate CF state (dedicated, shared, thin interrupt), which is useful when investigating functional performance problems

z13 and z13s systems with CFCC Level 21 require z/OS V1R12 with PTFs or later, and z/VM V6R2 or later for guest virtual coupling.

To support an upgrade from one CFCC level to the next, different levels of CFCC can be run concurrently while the coupling facility LPARs are running on different servers. CF LPARs that run on the same server share the CFCC level. The CFCC level for z13s servers is CFCC Level 21.

z13s servers (CFCC level 21) can coexist in a sysplex with CFCC levels 17 and 19.

The CFCC is implemented by using the active wait technique. This technique means that the CFCC is always running (processing or searching for service) and never enters a wait state. It also means that the CF Control Code uses all the processor capacity (cycles) that are available for the coupling facility LPAR. If the LPAR running the CFCC has only dedicated processors (CPs or ICFs), using all processor capacity (cycles) is not a problem. However, this configuration can be an issue if the LPAR that is running the CFCC also has shared processors. Therefore, enable dynamic dispatching on the CF LPAR. With CFCC Level 19 and Coupling Thin Interrupts, shared-processor CF can provide more consistent CF service time and acceptable usage in a broader range of configurations. For more information, see 3.9.3, “Dynamic CF dispatching” on page 133.

Performance consideration: Dedicated processor CF still provides the best CF image performance for production environments.

CF structure sizing changes are expected when going from CFCC Level 17 (or earlier) to CFCC Level 20. Review the CF structure size by using the CFSizer tool, which is available at this website:

<http://www.ibm.com/systems/z/cfsizer/>

For latest recommended levels, see the current exception letter on the Resource Link at the following website:

<https://www.ibm.com/servers/resourceLink/lib03020.nsf/pages/exceptionLetters>

3.9.2 Coupling Thin Interrupts

CFCC Level 19 introduced Coupling Thin Interrupts to improve performance in environments that share Coupling Facility engines. Although dedicated engines are preferable to obtain the best Coupling Facility performance, Coupling Thin Interrupts can help facilitate the use of a shared pool of engines, helping to lower hardware acquisition costs.

- ▶ The interrupt causes a shared logical processor coupling facility partition that has not already been dispatched to be dispatched by PR/SM, allowing the request or signal to be processed in a more timely manner. The coupling facility gives up control when work is exhausted or when PR/SM takes the physical processor away from the logical processor.
- ▶ The use of Coupling Thin Interrupts is controlled by the new DYNDISP specification.

You can experience CF response time improvements or more consistent CF response time when using Coupling Facilities with shared engines. This configuration can also allow more environments with multiple CF images to coexist in a server, and share CF engines with reasonable performance. The response time for asynchronous CF requests can also be improved as a result of using Coupling Thin Interrupts on the z/OS host system, regardless of whether the CF is using shared or dedicated engines.

3.9.3 Dynamic CF dispatching

Dynamic CF dispatching provides the following functions on a coupling facility:

1. If there is no work to do, CF enters a wait state (by time).
2. After an elapsed time, CF wakes up to see whether there is any new work to do (that is, there are requests in the CF Receiver buffer).

3. If there is no work, CF sleeps again for a longer period.
4. If there is new work, CF enters the normal active wait until the work is completed. After all work is complete, the process starts again.

With the introduction of the Coupling Thin Interrupt support, which is used only when the CF partition is using shared engines and the new **DYNDISP=THININTERRUPT** parameter, the CFCC code is changed to handle these interrupts correctly. CFCC also was changed to give up voluntarily control of the processor whenever it runs out of work to do. It relies on Coupling Thin Interrupts to dispatch the image again in a timely fashion when new work (or new signals) arrives at the CF to be processed.

This capability allows ICF engines to be shared by several CF images. In this environment, it provides faster and far more consistent CF service times. It also can provide performance that is reasonably close to dedicated-engine CF performance if the CF engines are not Coupling Facility Control Code thin interrupts. The introduction of thin interrupts allows a CF to run by using a shared processor while maintaining good performance. The shared engine is allowed to be undispached when there is no more work, as in the past. The new thin interrupt now gets the shared processor that is dispatched when a command or duplexing signal is presented to the shared engine.

This function saves processor cycles and is an excellent option to be used by a production backup CF or a testing environment CF. This function is activated by the CFCC command **DYNDISP ON**.

The CPs can run z/OS operating system images and CF images. For software charging reasons, generally use only ICF processors to run coupling facility images.

Figure 3-17 shows dynamic CF dispatching.

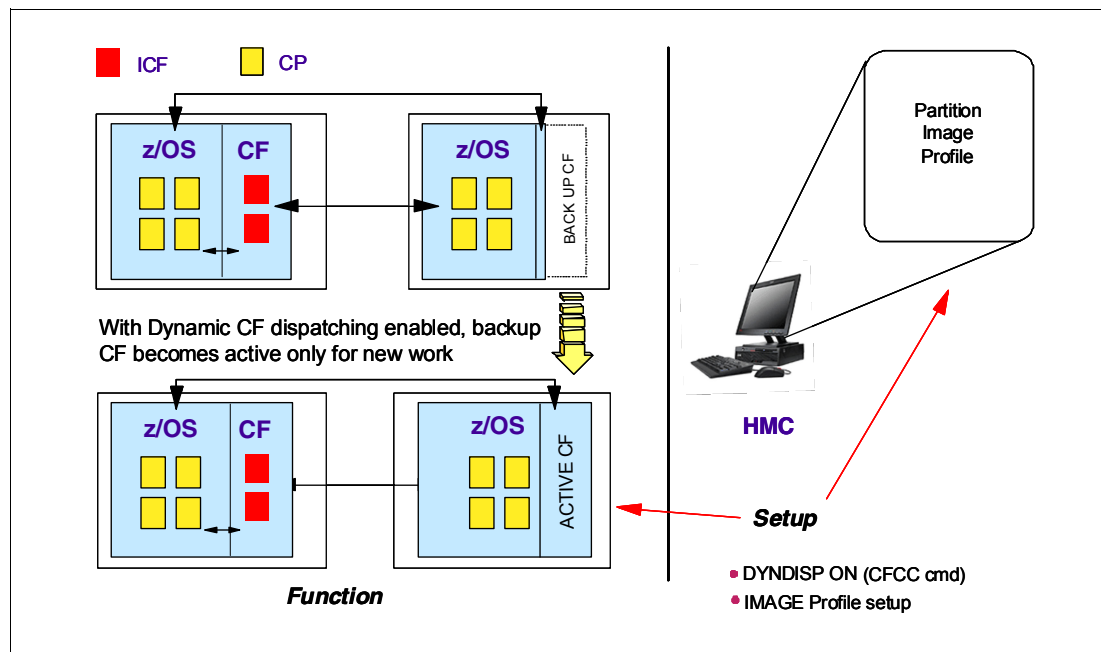


Figure 3-17 Dynamic CF dispatching (shared CPs or shared ICF PUs)

For more information about CF configurations, see *Coupling Facility Configuration Options*, GF22-5042, which is also available at the Parallel Sysplex website:

<http://www.ibm.com/systems/z/advantages/ps0/index.html>

3.9.4 CFCC and Flash Express use

From CFCC Level 19 and later, Flash Express can be used. It improves resilience while providing cost-effective standby capacity to help manage the potential overflow of IBM WebSphere MQ shared queues. Structures can now be allocated with a combination of real memory and SCM that is provided by the Flash Express feature.

For more information about Flash Express and CF Flash use, see Appendix H, “Flash Express” on page 529.



Central processor complex I/O system structure

This chapter describes the I/O system structure and connectivity options that are available on z13s servers.

This chapter includes the following sections:

- ▶ Introduction to the InfiniBand and PCIe for I/O infrastructure
- ▶ I/O system overview
- ▶ I/O drawer
- ▶ PCIe I/O drawer
- ▶ PCIe I/O drawer and I/O drawer offerings
- ▶ Fanouts
- ▶ I/O features (cards)
- ▶ Connectivity
- ▶ Parallel Sysplex connectivity
- ▶ Cryptographic functions
- ▶ Integrated firmware processor
- ▶ Flash Express
- ▶ 10GbE RoCE Express
- ▶ zEDC Express

4.1 Introduction to the InfiniBand and PCIe for I/O infrastructure

z13s servers support two types of internal I/O infrastructure:

- ▶ InfiniBand infrastructure for I/O drawers
- ▶ Peripheral Component Interconnect Express (PCIe)-based infrastructure for PCIe I/O drawers

4.1.1 InfiniBand I/O infrastructure

The InfiniBand I/O infrastructure was first made available on System z10 and is supported on z13s servers. It consists of the following components:

- ▶ InfiniBand fanouts that support the current 6 GBps InfiniBand I/O interconnect
- ▶ InfiniBand I/O card domain multiplexers with redundant I/O interconnect to a 5U, eight-slot, and two-domain I/O drawer.

4.1.2 PCIe I/O infrastructure

IBM continues the use of industry standards on the z Systems platform by offering a Peripheral Component Interconnect Express Generation 3 (PCIe Gen3) I/O infrastructure. The PCIe I/O infrastructure that is provided by the central processor complex (CPC) improves I/O capability and flexibility, while allowing for the future integration of PCIe adapters and accelerators.

The z13s PCIe I/O infrastructure consists of the following components:

- ▶ PCIe fanouts that support 16 GBps I/O bus interconnection for CPC drawer connectivity to the PCIe I/O drawers
- ▶ The 7U, 32-slot, and 4-domain PCIe I/O drawer for PCIe I/O features

The z13s PCIe I/O infrastructure provides these benefits:

- ▶ Increased bandwidth from the CPC drawer to the I/O domain in the PCIe I/O drawer through a 16 GBps bus.
- ▶ The PCIe I/O drawer supports double the number of I/O ports compared to an I/O drawer. Up to 64 channels (32 PCIe I/O features) are supported versus the 32 channels (8 I/O features) that are offered with the I/O drawer.
- ▶ Better granularity for the storage area network (SAN) and the local area network (LAN): For Fibre Channel connection (FICON), High Performance FICON on z Systems (zHPF), and Fibre Channel Protocol (FCP) storage area networks, the FICON Express16S has two channels per feature. For LAN connectivity, the Open System Adapter (OSA)-Express5S GbE and the OSA-Express5S 1000BASE-T features have two ports each, and the OSA-Express5S 10 GbE features have one port each.
- ▶ Newly designed native PCIe features can be plugged into the PCIe I/O drawer, such as Flash Express, zEnterprise Data Compression (zEDC) Express, and 10 GbE Remote Direct Memory Access over Converged Ethernet (RoCE) Express.

4.1.3 InfiniBand specifications

The InfiniBand specification defines a point-to-point, bidirectional serial communications between two nodes. It is intended for the connection of processors with high-speed peripheral devices, such as disks. InfiniBand supports various signaling rates on a physical lane, and as with PCI Express, lanes can be bonded together to create a wider link. Connection with only one physical lane pair is defined as a 1x link width connection. The specification also supports 4x and 12x link width connections by aggregating 4 or 12 pairs of physical lanes.

The base signaling rate on one physical lane is 2.5 Gbps, which is called the single data rate (SDR). InfiniBand can also support enhanced signaling rates like 5.0 Gbps for the double data rate (DDR) and 10 Gbps for the quad data rate (QDR)¹.

While operating at SDR, DDR or QDR, data transmission of physical lane is 8b/10b encoding, which means every 10 bits carry 8 bits of data. The following are the data rates for different link widths operating at different signaling rates:

- ▶ 250 MB/s for 1x SDR
- ▶ 500 MB/s for 1x DDR
- ▶ 1 GB/s for 4x SDR²
- ▶ 2 GB/s for 4x DDR
- ▶ 3 GB/s for 12x SDR
- ▶ 6 GB/s for 12x DDR

Link performance: The link speeds do not represent the actual performance of the link. The actual performance is dependent upon many factors that include latency through the adapters, cable lengths, and the type of workload.

z13s servers use InfiniBand 12x DDR connections between the CPC fanouts and I/O drawers.

For details and the standard for InfiniBand, see the InfiniBand website at:

<http://www.infinibandta.org>

4.1.4 PCIe Generation 3

The z13 and z13s servers are the first generation of z Systems servers to support the PCIe Generation 3 (Gen3) protocol.

PCIe Generation 3 uses 128b/130b encoding for data transmission. This encoding reduces the encoding effort by approximately 1.54% when compared to the PCIe Generation 2, which has an encoding effort of 20% using 8b/10b encoding.

The PCIe standard uses a low voltage differential serial bus. Two wires are used for signal transmission, and a total of four wires (two for transmit and two for receive) become a lane of a PCIe link, which is full duplex. Multiple lanes can be aggregated for larger link width. PCIe supports link widths of one lane (x1), x2, x4, x8, x12, x16, and x32.

The data transmission rate of a PCIe link is determined by the link width (numbers of lanes), the signaling rate of each lane, and the signal encoding rule. The signaling rate of a PCIe Generation 3 lane is 8 giga transmits per second (GT/s), which means 8 giga bits are transmitted per second (Gbps).

¹ Higher link speeds might be available for other applications. For more information, see the InfiniBand Trade Association website at: <http://www.infinibandta.org>

² z Systems does not support 4x link width

A PCIe Gen3 x16 link has these data transmission rates:

- ▶ Data transmission rate per lane: $8 \text{ Gb/s} * 128/130 \text{ bit (encoding)} = 7.87 \text{ Gb/s} = 984.6 \text{ MB/s}$
- ▶ Data transmission rate per link: $984.6 \text{ MB/s} * 16 \text{ (lanes)} = 15.75 \text{ GB/s}$

Considering the PCIe link is full-duplex mode, the data throughput rate of a PCIe Gen3 x16 link is 31.5 GBps (15.75 GB/s in both directions).

Link performance: The link speeds do not represent the actual performance of the link. The actual performance is dependent upon many factors that include latency through the adapters, cable lengths, and the type of workload.

PCIe Gen3 x16 links are used in z13s servers for driving the PCIe I/O drawers, and coupling links for CPC to CPC communications.

Note: Unless otherwise specified, when PCIe is mentioned in remainder of this chapter, it refers to PCIe Generation 3.

4.2 I/O system overview

This section lists the z13s I/O characteristics and a summary of supported features.

4.2.1 Characteristics

The z13s I/O subsystem is designed to provide great flexibility, high availability, and excellent performance characteristics:

- ▶ High bandwidth

Link performance: The link speeds do not represent the actual performance of the link. The actual performance is dependent upon many factors that include latency through the adapters, cable lengths, and the type of workload.

- z13s servers use PCIe as an internal interconnect protocol to drive PCIe I/O drawers and CPC to CPC connections. The I/O bus infrastructure data rate increases up to 16 GBps.

For more information about coupling links connectivity, see 4.9, “Parallel Sysplex connectivity” on page 174.

- z13s servers use InfiniBand as the internal interconnect protocol to drive I/O drawers and CPC to CPC connections. InfiniBand supports I/O bus infrastructure data rates up to 6 GBps.
- ▶ Connectivity options:
 - z13s servers can be connected to an extensive range of interfaces, such as FICON/FCP for SAN connectivity, 10 Gigabit Ethernet, Gigabit Ethernet, and 1000BASE-T Ethernet for LAN connectivity.
 - For CPC to CPC connections, z13s servers use Integrated Coupling Adapter (ICA SR) or Parallel Sysplex InfiniBand (PSIFB).
 - The 10GbE RoCE Express feature provides high-speed memory-to-memory data exchange to a remote CPC by using the Shared Memory Communications over Remote Direct Memory Access (SMC-R) protocol over TCP.

- ▶ Concurrent I/O upgrade
 - You can concurrently add I/O features to z13s servers if unused I/O slot positions are available.
- ▶ Concurrent PCIe I/O drawer upgrade
 - Extra PCIe I/O drawers can be installed concurrently if free frame slots for PCIe I/O drawers are available.
- ▶ Dynamic I/O configuration
 - Dynamic I/O configuration supports the dynamic addition, removal, or modification of the channel path, control units, and I/O devices without a planned outage.
- ▶ Pluggable optics
 - The FICON Express16S, FICON Express8S and FICON Express8, OSA Express5S, and 10GbE RoCE Express features have Small Form-Factor Pluggable (SFP) optics. These optics allow each channel to be individually serviced during a fiber optic module failure. The traffic on the other channels on the same feature can continue to flow if a channel requires servicing.
- ▶ Concurrent I/O card maintenance
 - Every I/O card that is plugged in an I/O drawer or PCIe I/O drawer supports concurrent card replacement during a repair action.

4.2.2 Summary of supported I/O features

The following I/O features are supported:

- ▶ Up to 128 FICON Express16S channels
- ▶ Up to 128 FICON Express8S channels
- ▶ Up to 32 FICON Express8 channels
- ▶ Up to 96 OSA-Express5S ports
- ▶ Up to 96 OSA-Express4S ports
- ▶ Up to 8 ICA SR links with up to 16 coupling links at 8 GBps
- ▶ Up to 8 InfiniBand fanouts with one of these options:
 - Up to 16 12x InfiniBand coupling links with HCA3-O fanout
 - Up to 32 1x InfiniBand coupling links with HCA3-O LR (1xIFB) fanout

Coupling links: ICA SR links are independent from any IFB link count because they occupy a different adapter slot in the CPC drawer.

The maximum number of ICA SR links on z13s servers is 16, and the maximum number of IFB links (any combination of HCA3-O and HCA3-O LR) is 32. Therefore, the maximum number of coupling links on a z13s server is 48.

In addition to I/O features, the following features are supported exclusively in the PCIe I/O drawer:

- ▶ Up to four zFlash Express features
- ▶ Up to eight zEDC Express features
- ▶ Up to sixteen 10GbE RoCE Express features

4.3 I/O drawer

The I/O drawer is five EIA units high, and supports up to eight I/O feature cards. Each I/O drawer supports two I/O domains (A and B) for a total of eight I/O card slots. Each I/O domain uses an InfiniBand Multiplexer (IFB-MP) card in the I/O drawer and a copper cable to connect to a host channel adapter (HCA) fanout in the CPC drawer.

The link between the HCA in the CPC drawer and the IFB-MP in the I/O drawer supports a link rate of up to 6 GBps. All cards in the I/O drawer are installed horizontally. Two distributed converter assemblies (DCAs) distribute power to the I/O drawer.

The locations of the DCAs, I/O feature cards, and IFB-MP cards in the I/O drawer are shown in Figure 4-1.

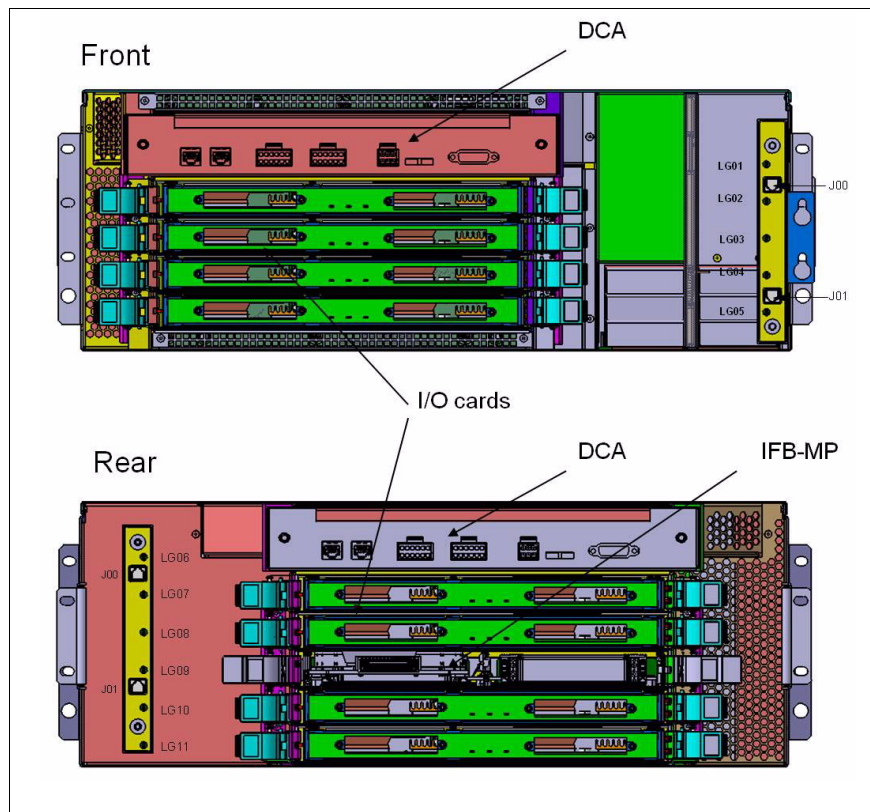


Figure 4-1 I/O drawer

The I/O structure in a z13s CPC is illustrated in Figure 4-2 on page 143. An IFB cable connects the HCA fanout card in the CPC drawer to an IFB-MP card in the I/O drawer. The passive connection between two IFB-MP cards is a design of redundant I/O interconnection (RII). This design provides connectivity between an HCA fanout card and I/O cards during concurrent fanout card or IFB cable replacement. The IFB cable between an HCA fanout card and each IFB-MP card supports a 6 GBps link rate.

Carry-forward only: Only one I/O drawer is supported in z13s servers on a carry-forward basis and can host only FICON Express8 cards (up to eight).

The I/O drawer domains and their related I/O slots are shown in Figure 4-2. The IFB-MP cards are installed at slot 09 at the rear side of the I/O drawer. The I/O features are installed from the front and rear side of the I/O drawer. Two I/O domains (A and B) are supported. Each I/O domain has up to four I/O features for FICON Express8 only (carry-forward). The I/O cards are connected to the IFB-MP card through the backplane board.

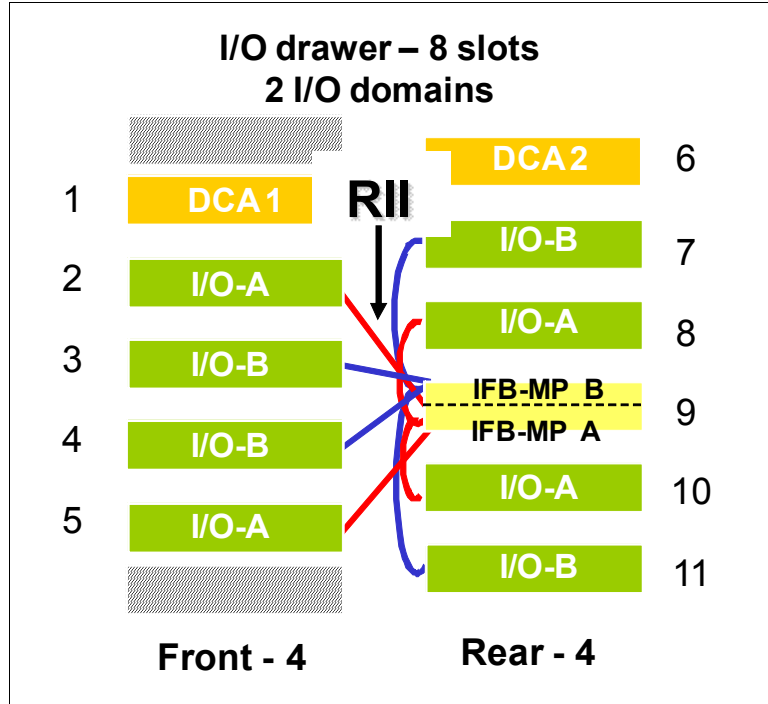


Figure 4-2 I/O domains of an I/O drawer

Each I/O domain supports four I/O card slots. Balancing I/O cards across both I/O domains is automatically done when the order is placed. Table 4-1 lists the I/O domains and their related I/O slots.

Table 4-1 I/O domains of I/O drawer

Domain	I/O slot in domain
A	02, 05, 08, and 10
B	03, 04, 07, and 11

4.4 PCIe I/O drawer

The PCIe I/O drawers are attached to the CPC drawer through a PCIe bus and use PCIe as the infrastructure bus within the drawer. The PCIe I/O bus infrastructure data rate is up to 16 GBps. PCIe switch application-specific integrated circuits (ASICs) are used to fan out the host bus from the CPC drawers to the individual I/O features. Up to 64 channels (32 PCIe I/O features) are supported versus the 32 channels (eight I/O features) that are offered with the I/O drawer. The PCIe I/O drawer is a two-sided drawer (I/O cards on both sides) that is 7U high. The drawer contains 32 I/O slots for feature cards, four switch cards (two in front, two in the back), two DCAs to provide redundant power, and two air-moving devices (AMDs) for redundant cooling.

The locations of the DCAs, AMDs, PCIe switch cards, and I/O feature cards in the PCIe I/O drawer are shown in Figure 4-3.

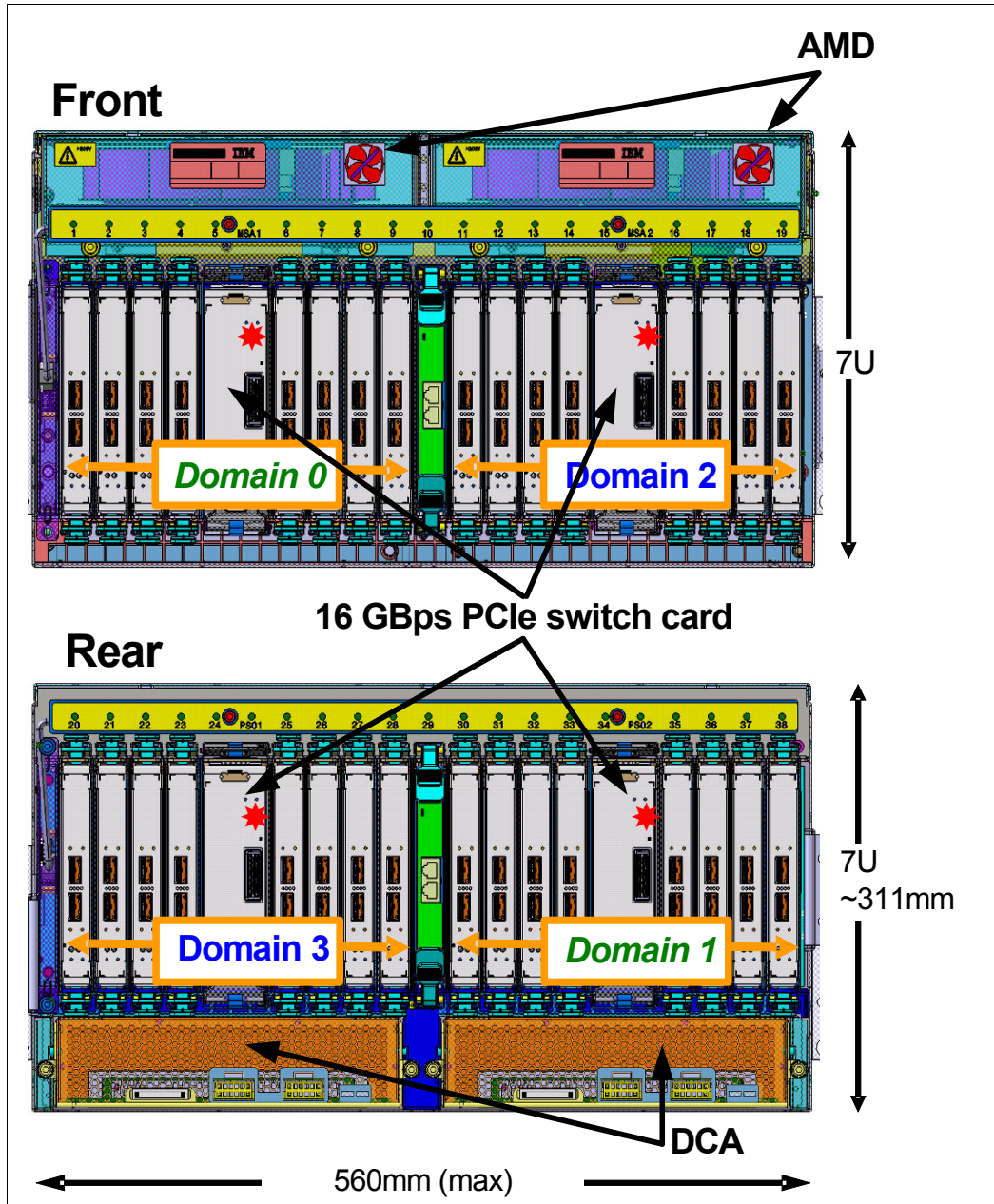


Figure 4-3 PCIe I/O drawer

The I/O structure in a z13s CPC is shown in Figure 4-4. The PCIe switch card provides the fanout from the high-speed x16 PCIe host bus to eight individual card slots. The PCIe switch card is connected to the drawers through a single x16 PCIe Gen3 bus from a PCIe fanout card.

In the PCIe I/O drawer, the eight I/O feature cards that directly attach to the switch card constitute an I/O domain. The PCIe I/O drawer supports concurrent add and replace I/O features to enable you to increase I/O capability as needed without having to plan ahead.

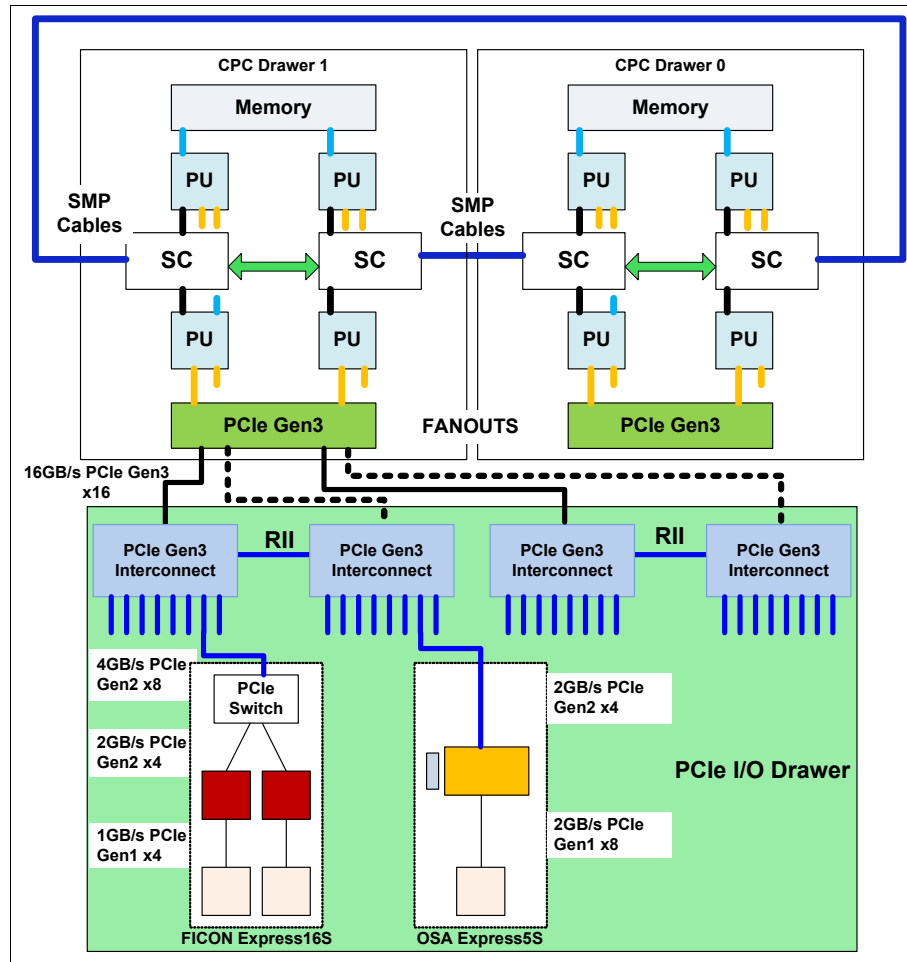


Figure 4-4 z13s (N20) connectivity to PCIe I/O drawers

The PCIe I/O Drawer supports up to 32 I/O features. They are organized into four hardware I/O domains per drawer. Each I/O domain supports up to eight features and is driven through a PCIe switch card. Two PCIe switch cards always provide a backup path for each other through the passive connection in the PCIe I/O drawer backplane. During a PCIe fanout card or cable failure, all 16 I/O cards in the two domains can be driven through a single PCIe switch card.

As a RII design, a switch card in the front is connected to a switch card in the rear through the PCIe I/O drawer board. Switch cards in same PCIe I/O drawer are connected to PCIe fanouts across nodes, and CPC drawers for higher availability. This RII design provides a failover capability during a PCIe fanout card failure or CPC drawer upgrade. All four domains in one of these PCIe I/O drawers can be activated with four fanouts. The flexible service processors (FSPs) are used for system control.

The PCIe I/O drawer domains and their related I/O slots are shown in Figure 4-5.

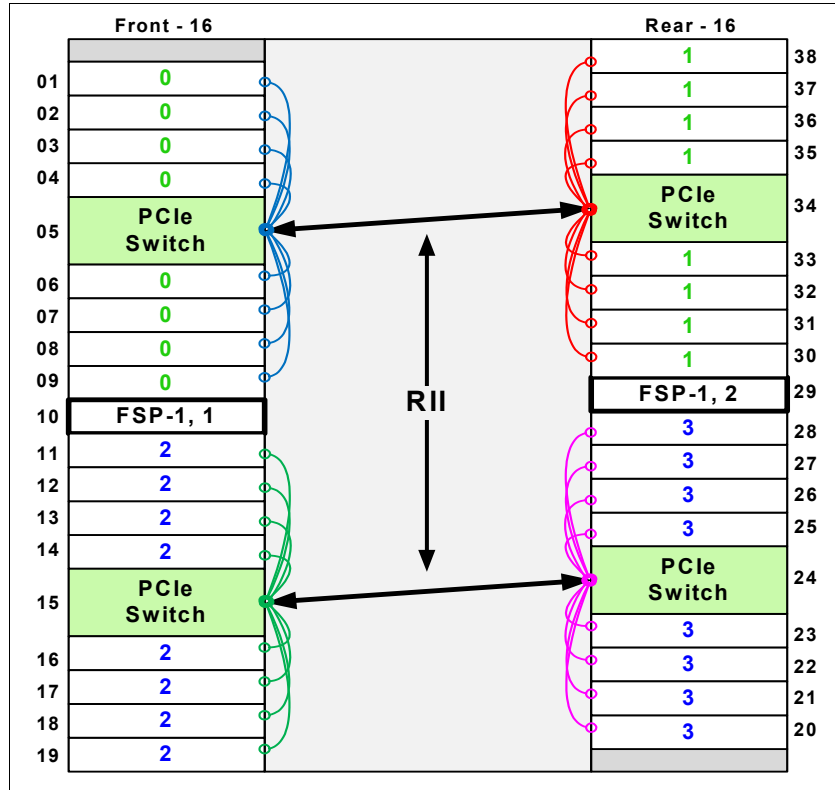


Figure 4-5 PCIe I/O drawer with 32 PCIe slots and four I/O domains

Each I/O domain supports up to eight features (FICON, OSA, and Crypto), and each drawer supports up to two Flash Express features (4 adapters), installed as pairs in slot 1 and 14, 25 and 33. All I/O cards connect to the PCIe switch card through the backplane board.

Note: The limitation of up to two native PCIe features (Flash Express, zEDC Express, and 10GbE RoCE Express) per I/O domain is eliminated on z13 and z13s servers.

Table 4-2 lists the I/O domains and slots.

Table 4-2 I/O domains of PCIe I/O drawer

Domain	I/O slot in domain
0	01, 02, 03, 04, 06, 07, 08, and 09
1	30, 31, 32, 33, 35, 36, 37, and 38
2	11, 12, 13, 14, 16, 17, 18, and 19
3	20, 21, 22, 23, 25, 26, 27, and 28

4.5 PCIe I/O drawer and I/O drawer offerings

A maximum of two PCIe I/O drawers can be installed that support up to 64 PCIe I/O features. Only PCIe I/O features can be ordered for a new system, and the total number of PCIe I/O

features determines how many PCIe I/O drawers are required. During upgrade to a z13s server, the correct mix of I/O drawers and PCIe I/O drawers is determined based on the type and number of I/O features (new and carry-forward) that are required. I/O drawers cannot be ordered.

Consideration: On a *new build* z13s server, only PCIe I/O drawers are supported. A mixture of I/O drawers, and PCIe I/O drawers are available only on upgrades to a z13s server.

The PCIe I/O drawers support the following PCIe features:

- ▶ FICON Express16S
- ▶ FICON Express8S
- ▶ OSA-Express5S
- ▶ OSA-Express4S
- ▶ 10GbE RoCE Express
- ▶ Crypto Express5S
- ▶ Flash Express
- ▶ zEDC Express

A maximum of one I/O drawer can be carried forward only for FICON Express8 cards (also carry-forward). As such, a maximum of 8 FICON Express8 cards (a total of 32 FICON Express8 ports) can be carried forward.

Consideration: FICON Express8 cards are supported as carry-forward only for z13s servers. They are installed in an I/O drawer. The installation of an I/O drawer limits the LPAR storage size to 1 TB (independent from the IOCP definition). Moreover, *each* I/O drawer affects the total fanout slots for PSIFB.

4.6 Fanouts

The z13s server uses fanout cards to connect the I/O subsystem to the CPC drawer. The fanout cards also provide the ICA SR and InfiniBand coupling links for Parallel Sysplex. All fanout cards support concurrent add, replace, and move.

The internal z13s I/O infrastructure consists of PCIe fanout cards and InfiniBand fanout cards:

- ▶ The PCIe Generation3 fanout card is a one port card and connects to a PCIe I/O drawer supporting an eight-slot I/O domain. This card is always installed in pairs to support I/O domain redundant connectivity.
- ▶ The InfiniBand HCA2-C fanout card is available for carry-forward of FICON Express8 features that are installed in I/O drawers.

The PCIe and InfiniBand fanouts are installed in the front of each CPC drawer. A two node CPC drawer has eight PCIe Gen3 fanout slots and four InfiniBand fanout slots. The PCIe fanout slots are named LG02 - LG06 and LG11 - LG15, left to right. The InfiniBand fanout slots are in the middle of the CPC drawer and are named LG07 - LG10, left to right. Slots LG01 and LG16 are not used for I/O infrastructure. Different numbers of fanouts slots are supported according to models of z13s servers:

- ▶ N10, maximum of four PCIe fanouts and two IFB fanouts supported
- ▶ N20 single CPC drawer, maximum of eight PCIe fanouts and four IFB fanouts supported
- ▶ N20 with two CPC drawers, maximum of 16 PCIe fanouts and eight IFB fanouts supported

Five types of fanout cards are supported by z13s servers. Each slot can hold one of the following five fanouts:

- ▶ PCIe Gen3 fanout card: This copper fanout provides connectivity to the PCIe switch card in the PCIe I/O drawer.
- ▶ ICA SR: This adapter provides 8x PCIe optical coupling connectivity between z13, z13s servers, up to 150-meter (492 ft.) distance, 8 GB/s link rate.
- ▶ Host Channel Adapter (HCA2-C): This copper fanout provides connectivity to the IFB-MP card in the I/O drawer.
- ▶ HCA3-O (12xIFB): This optical fanout provides 12x InfiniBand coupling link connectivity up to 150-meter (492 ft.) distance to a z13, z13s, zEC12, zBC12, z196, and z114 servers.
- ▶ HCA3-O LR (1xIFB): This optical long range fanout provides 1x InfiniBand coupling link connectivity up to a 10 km (6.2 miles) unrepeated or 100 km (62 miles) when extended by using z Systems qualified dense wavelength division multiplexing (DWDM) equipment distance to z13, z13s, zEC12, zBC12, z196 and z114 servers.

The PCIe Gen3 fanout card comes with one port, the HCA3-O LR (1xIFB) fanout comes with four ports, and other fanouts come with two ports.

Figure 4-6 illustrates the following PCIe and IFB connections from the CPC drawer:

- ▶ PCIe Gen3 for connectivity to PCIe I/O drawer (not represented here, see Figure 4.4 on page 143)
- ▶ Another CPC through InfiniBand (either 12x or 1x HCA3-O)
- ▶ Another z13s server connected through a dedicated PCIe ICA SR

Important: Figure 4-6 shows an I/O connection scheme that is not tied to a particular CPC drawer. In a real configuration, I/O connectivity is spread across multiple CPC drawers, if available, for I/O connection redundancy.

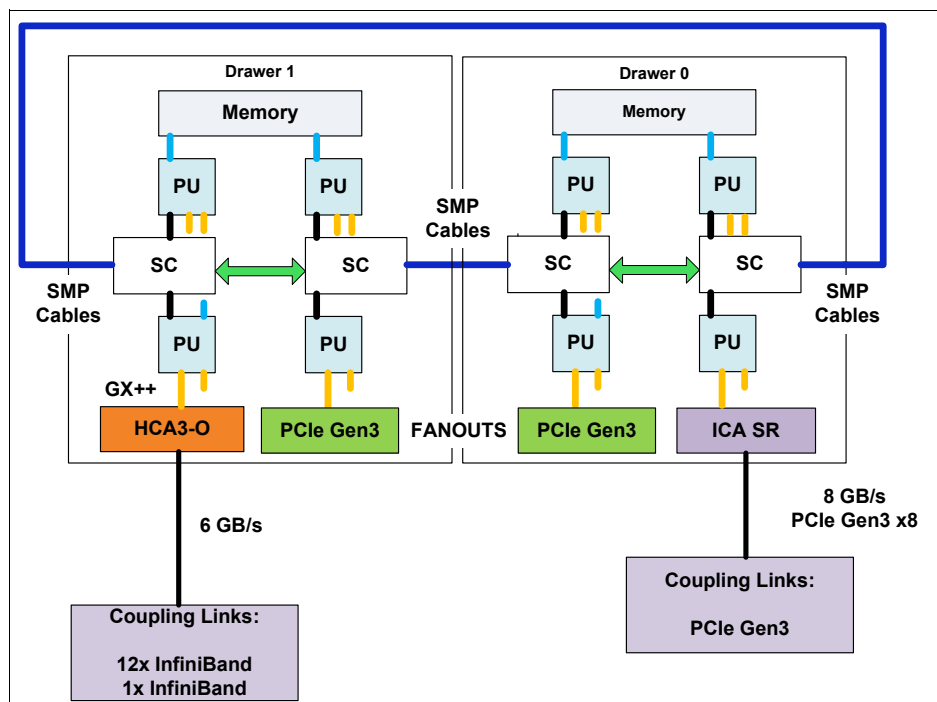


Figure 4-6 Infrastructure for PCIe and InfiniBand coupling links

4.6.1 PCIe Generation 3 fanout (FC 0173)

The PCIe Generation 3 fanout card provides connectivity to an PCIe I/O drawer by using a copper cable. One port on the fanout card is dedicated for PCIe I/O. The bandwidth of this PCIe fanout card supports a link rate of 16 GBps.

A 16x PCIe copper cable of 1.5 meters (4.92 ft) to 4.0 meters (13.1 ft) is used for connection to the PCIe switch card in the PCIe I/O drawer. PCIe fanout cards are always plugged in pairs and provide redundancy for I/O domains within the PCIe I/O drawer.

The pairs of PCIe fanout cards of a z13s Model N20 are split across the two logical nodes within a CPC drawer (LG02 - LG06 and LG11 - LG15), or are split across two CPC drawers for redundancy purposes.

PCIe fanout: The PCIe fanout is used exclusively for I/O and cannot be shared for any other purpose.

4.6.2 HCA2-C fanout (FC 0162)

The HCA2-C fanout is used to connect to an I/O drawer by using a copper cable. The two ports on the fanout are dedicated to I/O. Each port on the HCA2-C fanout supports a link rate of up to 6 GBps.

A 12x InfiniBand copper cable of 1.5 meters (4.92 ft.) to 3.5 meters (11.4 ft.) is used for connection to the IFB-MP card in the I/O drawer. An HCA2-C fanout is supported only if carried forward with a miscellaneous equipment specification (MES) from z114, or zBC12 to z13s server. For a new build z13s server, HCA2-C cannot be ordered.'

HCA2-C fanout: The HCA2-C fanout is used exclusively for a connection to the I/O drawer. It cannot be shared for any other purpose.

4.6.3 Integrated Coupling Adapter (FC 0172)

The IBM ICA SR, introduced on the z13 and z13s platform, is a two-port fanout that is used for short distance coupling connectivity with a new coupling channel type, the CS5. The ICA SR uses PCIe Gen3 technology, with x16 lanes that are bifurcated into x8 lanes for coupling. No performance degradation is expected compared to the Coupling over the InfiniBand 12x IFB3 protocol.

The ICA SR is designed to drive distances up to 150 m and supports a link data rate of 8 GBps. It is also designed to support up to four CHPIDs per port and seven subchannels (devices) per CHPID. The coupling links can be defined as shared between images within a channel subsystem (CSS). They also can be spanned across multiple CSSs in a CPC. Unlike the HCA3-O 12x InfiniBand links, the ICA SR cannot define more than four CHPIDS per port. When Server Time Protocol (STP) is enabled, ICA SR coupling links can be defined as timing-only links to other z13 and z13s CPCs.

The ICA SR fanout is housed in the PCIe I/O fanout slot on the z13s CPC drawer, which supports 4-16 PCIe I/O slots. Up to 8 ICA SR fanouts and up to 16 ICA SR ports are supported on a z13s CPC drawer, enabling greater connectivity for short distance coupling on a single processor node compared to previous generations. The maximum number of ICA SR fanout features is 8 per system.

The ICA SR can be used only for coupling connectivity between z13 and z13s servers. It does not support connectivity to zEC12, zBC12, z196, or z114 servers, and it cannot be connected to HCA3-O or HCA3-O LR coupling fanouts.

The ICA SR fanout requires new cabling that is different from the 12x IFB cables. For distances up to 100 m, clients can choose the OM3 fiber optic cable. For distances up to 150 m, clients must choose the OM4 fiber optic cable. For more information, see *z Systems Planning for Fiber Optics*, GA23-1407 and *z13s Installation Manual for Physical Planning*, GC28-6953, which can be found in the Library section of Resource Link at the following website:

<http://www.ibm.com/servers/resourcelink>

4.6.4 HCA3-O (12x IFB) fanout (FC 0171)

The HCA3-O fanout for 12x InfiniBand provides an optical interface that is used for coupling links. The two ports on the fanout are dedicated to coupling links that connect to z13, z13s, zEC12, zBC12, z196, and z114 CPCs. Up to 8 HCA3-O (12x IFB) fanouts are supported, and provide up to 16 ports for coupling links.

The fiber optic cables are industry standard OM3 (2000 MHz-km) 50- μ m multimode optical cables with MPO connectors. The maximum cable length is 150 meters (492 ft.). There are 12 pairs of fibers: 12 fibers for transmitting and 12 fibers for receiving. The HCA3-O (12xIFB) fanout supports a link data rate of 6 GBps.

Important: The HCA3-O fanout has two ports (1 and 2). Each port has one connector for transmitting (TX) and one connector for receiving (RX). Ensure that you use the correct cables. An example is shown in Figure 4-7.

For more information, see *Planning for Fiber Optic Links*, GA23-1407, and *IBM z13 Installation Manual for Physical Planning*, GC28-6938, which can be found in the Library section of Resource Link at the following website:

<http://www.ibm.com/servers/resourcelink>

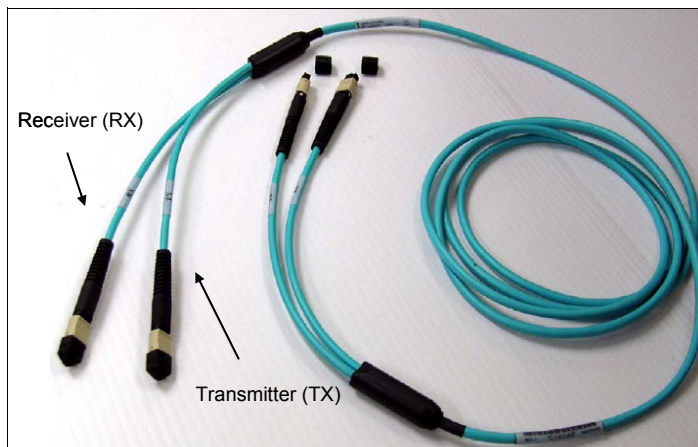


Figure 4-7 OM3 50/125 μ m multimode fiber cable with MPO connectors

A fanout has two ports for optical link connections, and supports up to 16 CHPIDs across both ports. These CHPIDs are defined as channel type CIB in the input/output configuration

data set (IOCDS). The coupling links can be defined as shared between images within a CSS. They also can be spanned across multiple CSSs in a CPC.

Each HCA3-O (12x IFB) fanout has an assigned Adapter ID (AID) number. This number must be used for definitions in IOCDS to create a relationship between the physical fanout location and the channel-path identifier (CHPID) number. For more information about AID numbering, see “Adapter ID number assignment” on page 152.

For more information about how the AID is used and referenced in the hardware configuration definition (HCD), see *Implementing and Managing InfiniBand Coupling Links on IBM System z*, SG24-7539.

When STP is enabled, IFB coupling links can be defined as timing-only links to other z13, zEC12, zBC12, z196, and z114 CPCs.

12x IFB and 12x IFB3 protocols

These protocols are supported by the HCA3-O for 12x IFB feature:

- ▶ 12x IFB3 protocol: When HCA3-O (12xIFB) fanouts are communicating with HCA3-O (12x IFB) fanouts and are defined with four or fewer CHPIDs per port, the 12x IFB3 protocol is used.
- ▶ 12x IFB protocol: If more than four CHPIDs are defined per HCA3-O (12xIFB) port, or HCA3-O (12x IFB) features are communicating with HCA2-O (12x IFB) features on zEnterprise or System z10 CPCs, links run with the 12x IFB protocol.

The HCA3-O feature that supports 12x InfiniBand coupling links is designed to deliver improved service times. When no more than four CHPIDs are defined per HCA3-O (12xIFB) port, the 12x IFB3 protocol is used. When you use the 12x IFB3 protocol, synchronous service times are up to 40% faster than when you use the 12x IFB protocol. Unlike the HCA3-O 12x InfiniBand links, the ICA SR cannot define more than four CHPIDS per port.

4.6.5 HCA3-O LR (1x IFB) fanout (FC 0170)

The HCA3-O LR fanout for 1x InfiniBand provides an optical interface that is used for coupling links. The four ports on the fanout are dedicated to coupling links to connect to z13, z13s, zEC12, zBC12, z196, and z114 servers. Up to 8 HCA3-O LR (1xIFB) fanouts are supported by z13s servers, and provide up to 32 ports for coupling links.

The HCA-O LR fanout supports InfiniBand 1x optical links that offer long-distance coupling links. The cable has one lane that contains two fibers. One fiber is used for transmitting, and the other fiber is used for receiving data.

Each connection supports a link rate of up to 5 Gbps if connected to a z13, z13s, zEC12, zBC12, z196, or z114 server. It supports a link rate of 2.5 Gbps when connected to a z Systems qualified DWDM. The link rate is auto-negotiated to the highest common rate.

The fiber optic cables are 9- μ m SM optical cables that terminate with an LC Duplex connector. With direct connection, the supported unrepeated distance³ is up to 10 km (6.2 miles), and up to 100 km (62 miles) with z Systems qualified DWDM.

A fanout has four ports for optical link connections, and supports up to 16 CHPIDs across all four ports. These CHPIDs are defined as channel type CIB in the IOCDS. The coupling links can be defined as shared between images within a channel subsystem, and also can be spanned across multiple CSSs in a server.

³ On special request, see <http://www.ibm.com/systems/z/advantages/psocflinks.html>

Each HCA3-O LR (1xIFB) fanout has an assigned AID number. This number must be used for definitions in IOCDs to create a relationship between the physical fanout location and the CHPID number. For more information about AID numbering, see “Adapter ID number assignment” on page 152.

When STP is enabled, InfiniBand long range (IFB LR) coupling links can be defined as timing-only links to other z13, z13s, zEC12, zBC12, z196, and z114 CPCs.

4.6.6 Fanout considerations

Fanout slots in each CPC drawer can be used to plug different fanouts. One CPC drawer can hold up to eight PCIe fanouts and four InfiniBand fanout cards.

For migration purposes, the number of available HCA3-O or HCA3-O LR cards in the z13s servers can be important (for coupling connectivity to zEC12, zBC12, z196, and z114). The carry-forward of FICON Express8 features reduces the number of available InfiniBand fanout slots by two. This limitation can be important for coupling connectivity planning, especially for a z13s Model N10, as shown in Table 4-3. No more fanout adapter slots are available for an N10 model after a FICON Express8 feature is carried forward. Table 4-3 shows for the z13s Models N10 and N20 where the limitation of HCA3-O and HCA3-O LR fanouts that are caused by carry-forward of FICON Express8 features might require attention during migration.

For migration considerations, see 4.9.2, “Migration considerations” on page 178.

Table 4-3 CPC drawer InfiniBand fanout availability limitation with carry-forward FICON Express8 features

CPC Model	Carry forward FICON Express8 features	Resulting number of I/O drawers	Max. number of HCA3-O or HCA3-O LR fanout cards	Max. number of PCIe I/O drawers	Max. number of PCIe I/O features
N10	0	0	2	1	32
	1 - 8	1	0	1	32
N20	0	0	8	2	64
	1 - 8	1	6	2 ^a	64 ^b

a. If one I/O drawer is carried forward for a Model N20 with two CPC drawers, only one PCIe I/O drawer can be installed.

b. The number of FICON Express8 carried forward also affects the maximum number of PCIe I/O features.

Adapter ID number assignment

PCIe and IFB fanouts and ports are identified by an AID that is initially dependent on their physical locations, unlike channels that are installed in a PCIe I/O drawer or I/O drawer. Those channels are identified by a physical channel ID (PCHID) number that is related to their physical location. This AID must be used to assign a CHPID to the fanout in the IOCDs definition. The CHPID assignment is done by associating the CHPID to an AID port.

Table 4-4 illustrates the AID assignment for each fanout slot relative to the drawer location on a new build system.

Table 4-4 AID number assignment

Drawer	Location	Fanout slot	AIDs
First	A21A	LG03-LG06 (PCIe) ^a	1B-1E
		LG07-LG10 (IFB) ^a	04-07
		LG11-LG14 (PCIe)	1F-22
Second	A25A	LG03-LG06 (PCIe)	11-14
		LG07-LG10 (IFB)	00-03
		LG11-LG14 (PCIe)	15-18

a. For Model N10, only the first drawer's LG09-LG10 are supported for IFB, and LG11-LG14 are supported for PCIe.

Fanout slots

The fanout slots are numbered LG03 - LG14 left to right, as shown in Table 4-4. All fanout locations and their AIDs for two drawers are shown in Table 4-4 for reference only. Slots LG01 and LG16 never have a fanout installed because they are dedicated for FSPs. Slots LG02 and LG15 are reserved.

Important: The AID numbers in Table 4-4 are valid only for a new build system or if a second CPC drawer is added. If a fanout is moved, the AID follows the fanout to its new physical location.

The AID assignment is listed in the PCHID REPORT that is provided for each new server or for an MES upgrade on existing servers.

Example 4-1 shows part of a PCHID REPORT for a Model N20. In this example, one fanout is installed in the first drawer (location A21A slot LG03) and one fanout is installed in the second drawer (location A25A slot LG11). The assigned AID for the fanout in the first drawer is 1B. The AID that is assigned to the fanout in the second drawer is 15.

Example 4-1 AID assignment in PCHID REPORT

```

CHPIDSTART
12345678                PCHID REPORT                Oct 31,2014
Machine: 2965-N20 SNXXXXXX
-----
Source          Cage Slot F/C   PCHID/Ports or AID          Comment
A21/LG03        A21A LG03 0172  AID=1B
A25/LG11        A25A LG11 0172  AID=15

```

Fanout features that are supported by the z13s server are shown in Table 4-5. The table provides the feature type, feature code, and information about the link that is supported by the fanout feature.

Table 4-5 Fanout summary

Fanout feature	Feature code	Use	Cable type	Connector type	Maximum distance	Link data rate ^a
HCA2-C	0162	Connect to I/O drawer	Copper	N/A	3.5 m (11.48 ft)	6 Gbps
PCIe fanout	0173	Connect to PCIe I/O drawer	Copper	N/A	4 m (13.1 ft)	16 Gbps
HCA3-O (12xIFB)	0171	Coupling link	50- μ m MM OM3 (2000 MHz-km)	MPO	150 m (492 ft)	6 Gbps ^b
HCA3-O LR (1xIFB)	0170	Coupling link	9- μ m SM	LC Duplex	10 km (6.2 miles) 100 km ^c (62 miles)	5.0 Gbps 2.5 Gbps ^d
ICA SR	0172	Coupling link	OM4	MTP	150 m (492 ft)	8 Gbps
			OM3	MTP	100 m (328 ft)	8 Gbps

- a. The link data rates do not represent the actual performance of the link. The actual performance is dependent upon many factors that include latency through the adapters, cable lengths, and the type of workload.
- b. When using the 12x IFB3 protocol, synchronous service times are 40% faster than when you use the 12x IFB protocol.
- c. With z Systems qualified DWDM.
- d. Auto-negotiated, depending on the DWDM equipment.

4.7 I/O features (cards)

I/O features (cards) have ports⁴ to connect z13s servers to external devices, networks, or other servers. I/O features are plugged into the PCIe I/O drawer, based on the configuration rules for the server. Different types of I/O cards are available, one for each channel or link type. I/O cards can be installed or replaced concurrently.

In addition to PCIe I/O features, FICON Express8 features can be installed in I/O drawers. z13s servers support a maximum of one I/O drawers as carry-forward during an MES. This drawer represents a maximum of 8 FICON Express8 features (a total of 32 FICON Express8 ports) that can be carried forward.

Consideration: The installation of an I/O drawer limits the LPAR storage to 1 TB (independent from the IOCP definition). Moreover, *each* I/O drawer affects the configuration as follows:

- ▶ The total fanout slots for PSIFB are reduced by two.

⁴ Certain I/O features do not have external ports, such as Crypto Express5S and zEDC

4.7.1 I/O feature card ordering information

Table 4-6 lists the I/O features that are supported by z13s servers and the ordering information for them.

Table 4-6 I/O features and ordering information

Channel feature	Feature code	New build	Carry-forward
FICON Express16S 10KM LX	0418	Y	N/A
FICON Express16S SX	0419	Y	N/A
FICON Express8S 10KM LX	0409	Y	Y
FICON Express8S SX	0410	Y	Y
FICON Express8 10KM LX	3325	N	Y
FICON Express8 SX	3326	N	Y
OSA-Express5S 10 GbE LR	0415	Y	Y
OSA-Express5S 10 GbE SR	0416	Y	Y
OSA-Express5S GbE LX	0413	Y	Y
OSA-Express5S GbE SX	0414	Y	Y
OSA-Express5S 1000BASE-T Ethernet	0417	Y	Y
OSA-Express4S 10 GbE LR	0406	N	Y
OSA-Express4S 10 GbE SR	0407	N	Y
OSA-Express4S GbE LX	0404	N	Y
OSA-Express4S GbE SX	0405	N	Y
OSA-Express4S 1000BASE-T Ethernet	0408	N	Y
Integrated Coupling Adapter (ICA SR)	0172	Y	N/A
HCA3-O (12xIFB)	0171	Y	Y
HCA3-O LR (1xIFB)	0170	Y	Y
Crypto Express5S	0890	Y	Y
Flash Express	0403	Y	N/A
Flash Express	0402	N	Y
10GbE RoCE Express	0411	Y	Y
zEDC Express	0420	Y	Y

Important: z13s servers do not support the ISC-3, HCA2-O (12x), and HCA2-O LR (1x) features and cannot participate in a Mixed Coordinated Timing Network (CTN).

4.7.2 Physical channel report

A PCHID reflects the physical location of a channel-type interface.

A PCHID number is based on these factors:

- ▶ The I/O drawer, and PCIe I/O drawer location
- ▶ The channel feature slot number
- ▶ The port number of the channel feature

A CHPID does not directly correspond to a hardware channel port, but it is assigned to a PCHID in the HCD or input/output configuration program (IOCP).

A PCHID report is created for each new build server and for upgrades on existing servers. The report lists all I/O features that are installed, the physical slot location, and the assigned PCHID. Example 4-2 shows a portion of a sample PCHID report. For more information about the AID numbering rules for InfiniBand and PCIe fanouts, see Table 4-4 on page 153.

Example 4-2 PCHID REPORT

```
CHPIDSTART
12345678                PCHID REPORT                Nov 17,2015
Machine: 2965-N20  SN1
-----
```

Source	Drwr	Slot	F/C	PCHID/Ports or AID	Comment
A25/LG07	A25A	LG07	0171	AID=00	
A21/LG07	A21A	LG07	0170	AID=04	
A21/LG09	A21A	LG09	0171	AID=06	
A21/LG04	A21A	LG04	0172	AID=1C	
A21/LG03/J01	A13B	01	0403	100/Mem paired 12C/Mem	
A21/LG03/J01	A13B	03	0420	108	RG1
A21/LG03/J01	A13B	06	0413	110/D1D2	
A21/LG03/J01	A13B	07	0418	114/D1 115/D2	
A21/LG05/J01	A13B	14	0403	12C/Mem paired 100/Mem	
A21/LG05/J01	A13B	18	0411	138/D1D2	RG1
A21/LG12/J01	A13B	20	0420	140	RG2
A21/LG14/J01	A13B	37	0411	178/D1D2	RG2

Legend:

```
Source  Book Slot/Fanout Slot/Jack
A25A    CEC Drawer 2 in A frame
A21A    CEC Drawer 1 in A frame
A13B    PCIe Drawer 1 in A frame
0403    Flash Express
RG1     Resource Group One
0420    zEDC Express
0413    OSA Express5S GbE LX 2 Ports
0418    16 Gbps FICON/FCP LX 2 Ports
0411    10GbE RoCE Express
RG2     Resource Group Two
0171    HCA3 0 PSIFB 12x 2 Links
0170    HCA3 0 LR PSIFB 1x 4 Links
0172    ICA SR 2 Links
```

The following list explains the content of this sample PCHID REPORT:

- ▶ Feature code 0170 (HCA3-O LR (1xIFB)) is installed in CPC drawer 1 (location A21A, slot LG07), and has AID 04 assigned.
- ▶ Feature code 0172 (Integrated Coupling Adapter - ICA SR) is installed in CPC drawer 1 (location A21A, slot LG04), and has AID 1C assigned.
- ▶ Feature code 0403 (Flash Express) is installed in PCIe I/O drawer 1 (location A13B, slot 01, and slot 14), with PCHID 100 and 12C assigned. Flash Express adapters are always installed as a pair.
- ▶ One feature code 0420 (zEDC Express) is installed in PCIe I/O drawer 1 (location A13B, slot 03), attached to resource group RG1, with PCHID 108 assigned. Another zEDC Express is installed in PCIe I/O drawer 1 (location A13B, slot 20), attached to RG2, with PCHID 140 assigned.
- ▶ Feature code 0413 (OSA Express5S GbE LX) is installed in PCIe I/O drawer 1 (location A13B, slot 06), with PCHID 110 assigned. The two ports D1 and D2 share a PCHID.
- ▶ Feature code 0418 (FICON Express16S long wavelength (LX) 10 km (6.2 miles)) is installed in PCIe I/O drawer 1 (location A13B, slot 07). PCHID 114 is assigned to port D1 and 115 assigned to port D2.
- ▶ One feature code 0411 (10GbE RoCE Express) is installed in PCIe I/O drawer 1 (location A13B, slot 18), attached to resource group RG1, with PCHID 138 assigned and shared by port D0 and D1. Another 10GbE RoCE Express is installed in PCIe I/O drawer 1 (location A13B, slot 37), attached to resource group RG2, with PCHID 178 assigned.

An RG parameter is shown in the PCHID REPORT for native PCIe features. The native PCIe features have a balanced plugging between two resource groups (RG1 and RG2).

For more information about resource groups, see Appendix G, “Native Peripheral Component Interconnect Express” on page 521.

4.8 Connectivity

I/O channels are part of the CSS. They provide connectivity for data exchange between servers, or between servers and external control units (CUs) and devices, or between networks.

Communication between servers is implemented by using ICA SR, coupling over InfiniBand, or channel-to-channel (CTC) connections.

Communication to LANs is provided by the OSA-Express5S, OSA-Express4S, and 10GbE RoCE Express features.

Connectivity to external I/O subsystems, disks for example, is provided by FICON channels.

4.8.1 I/O feature support and configuration rules

Table 4-7 lists the I/O features that are supported. The table shows the number of ports per card, port increments, the maximum number of feature cards, and the maximum number of channels for each feature type. Also, the CHPID definitions that are used in the IOCDs are listed.

Table 4-7 z13 supported I/O features

I/O feature	Number of		Maximum number of		PCHID	CHPID definition
	Ports per card	Port increments	Ports	I/O slots		
FICON Express16S LX/SX	2	2	128	64	Yes	FC, FCP
FICON Express8S LX/SX	2	2	128	64	Yes	FC, FCP
FICON Express8 LX/SX	4	4	32	8	Yes	FC, FCP
OSA-Express5S 10 GbE LR/SR	1	1	48	48	Yes	OSD, OSX
OSA-Express5S GbE LX/SX	2	2	96	48	Yes	OSD
OSA-Express5S 1000BASE-T	2	2	96	48	Yes	OSC, OSD, OSE, OSM, OSN
OSA- Express4S 10 GbE LR/SR	1	1	48	48	Yes	OSD, OSX
OSA-Express4S GbE LX/SX	2	2	96	48	Yes	OSD
OSA-Express4S 1000BASE-T	2	2	96	48	Yes	OSC, OSD, OSE, OSM, OSN
10GbE RoCE Express	2	2	32	16	Yes	N/A ^a
Integrated Coupling Adapter (ICA SR)	2	2	16	8	N/A ^b	CS5
HCA3-O for 12x IFB and 12x IFB3	2	2	16	8	N/A ^b	CIB
HCA3-O LR for 1x IFB	4	4	16	8	N/A ^b	CIB

a. Defined by PCIe function ID (PFID).

b. identified by AID.

At least one I/O feature (FICON) or one coupling link feature (ICA SR or HCA3-O) must be present in the minimum configuration.

The following features can be shared and spanned:

- ▶ FICON channels that are defined as FC or FCP
- ▶ OSA-Express5S features that are defined as OSC, OSD, OSE, OSM, OSN, or OSX
- ▶ OSA-Express4S features that are defined as OSC, OSD, OSE, OSM, OSN, or OSX

- ▶ Coupling links that are defined as CS5, CIB, or ICP
- ▶ HiperSockets that are defined as IQD⁵

The following features are exclusively plugged into a PCIe I/O drawer and do not require the definition of a CHPID and CHPID type:

- ▶ Each Crypto Express feature occupies one I/O slot, but does not have a CHPID type. However, logical partitions (LPARs) in all CSSs have access to the features. Each Crypto Express adapter can be defined to up to 40 LPARs.
- ▶ Each Flash Express feature occupies two I/O slots, but does not have a CHPID type. However, LPARs in all CSSs have access to the features. The Flash Express feature can be defined to up to 40 LPARs.
- ▶ Each RoCE feature occupies one I/O slot, but does not have a CHPID type. However, LPARs in all CSSs have access to the feature. The RoCE feature can be defined to up to 31 LPARs.
- ▶ Each zEDC feature occupies one I/O slot, but does not have a CHPID type. However, LPARs in all CSSs have access to the feature. The zEDC feature can be defined to up to 15 LPARs.

I/O feature cables and connectors

The IBM Facilities Cabling Services fiber transport system offers a total cable solution service to help with cable ordering requirements. These services can include the requirements for all of the protocols and media types that are supported (for example, FICON, coupling links, and OSA). The services can help whether the focus is the data center, a SAN, a LAN, or the end-to-end enterprise.

Cables: All fiber optic cables, cable planning, labeling, and installation are client responsibilities for new z13s installations and upgrades. Fiber optic conversion kits and mode conditioning patch cables cannot be ordered as features on z13s servers. All other cables must be sourced separately.

The Enterprise Fiber Cabling Services use a proven modular cabling system, the fiber transport system (FTS), which includes trunk cables, zone cabinets, and panels for servers, directors, and storage devices. FTS supports Fiber Quick Connect (FQC), a fiber harness that is integrated in the frame of a z13s server for quick connection. The FQC is offered as a feature on z13s servers for connection to FICON LX channels.

Whether you choose a packaged service or a custom service, high-quality components are used to facilitate moves, additions, and changes in the enterprise to prevent having to extend the maintenance window.

Table 4-8 lists the required connector and cable type for each I/O feature on the z13s server.

Table 4-8 I/O feature connector and cable types

Feature code	Feature name	Connector type	Cable type
0418	FICON Express16S LX 10 km	LC Duplex	9 μm SM ^a
0419	FICON Express16S SX	LC Duplex	50, 62.5 μm MM ^b

⁵ There is a new parameter for HiperSockets IOCP definitions on z13. As such, the z13 IOCP definitions need to be migrated to support the HiperSockets definitions (CHPID type IQD). On z13, the CHPID statement of HiperSockets devices requires the keyword VCHID. VCHID specifies the virtual channel identification number associated with the channel path. Valid range is 7E0 - 7FF. VCHID is not valid on z Systems before z13.

Feature code	Feature name	Connector type	Cable type
0409	FICON Express8S LX 10 km	LC Duplex	9 µm SM
0410	FICON Express8S SX	LC Duplex	50, 62.5 µm MM
3325	FICON Express8 LX 10 km	LC Duplex	9 µm SM
3326	FICON Express8 SX	LC Duplex	50, 62.5 µm MM
0415	OSA-Express5S 10 GbE LR	LC Duplex	9 µm SM
0416	OSA-Express5S 10 GbE SR	LC Duplex	50, 62.5 µm MM
0413	OSA-Express5S GbE LX	LC Duplex	9 µm SM
0414	OSA-Express5S GbE SX	LC Duplex	50, 62.5 µm MM
0417	OSA-Express5S 1000BASE-T	RJ-45	Category 5 UTP ^c
0404	OSA-Express4S GbE LX	LC Duplex	9 µm SM
0405	OSA-Express4S GbE SX	LC Duplex	50, 62.5 µm MM
0406	OSA-Express4S 10 GbE LR	LC Duplex	9 µm SM
0407	OSA-Express4S 10 GbE SR	LC Duplex	50, 62.5 µm MM
0408	OSA-Express4S 1000BASE-T	RJ-45	Category 5 UTP
0411	10GbE RoCE Express	LC Duplex	50, 62.5 µm MM
0172	Integrated Coupling Adapter (ICA SR)	MPO	50 µm MM OM4 (4.7 GHz-km) ^d
0171	HCA3-O (12xIFB)	MPO	50 µm MM OM3 (2 GHz-km)
0170	HCA3-O LR (1xIFB)	LC Duplex	9 µm SM

a. SM is single-mode fiber.

b. MM is multimode fiber.

c. UTP is unshielded twisted pair. Consider using category 6 UTP for 1000 Mbps connections.

d. Or 50 µm MM OM3 (2 GHz-km), but OM4 is highly recommended.

4.8.2 FICON channels

The FICON Express16S, FICON Express8S, and FICON Express8 features conform to the following architectures:

- ▶ Fibre Connection (FICON)
- ▶ High Performance FICON on z Systems (zHPF)
- ▶ Fibre Channel Protocol (FCP)

They provide connectivity between any combination of servers, directors, switches, and devices (control units, disks, tapes, and printers) in a SAN.

Each FICON Express16S or FICON Express8S feature occupies one I/O slot in the PCIe I/O drawer. Each feature has two ports, each supporting an LC Duplex connector, with one PCHID and one CHPID associated with each port.

Each FICON Express8 feature occupies one I/O slot in the I/O drawer. Each feature has four ports, each supporting an LC Duplex connector, with one PCHID and one CHPID associated with each port.

All FICON Express16S, FICON Express8S, and FICON Express8 features use SFP optics that allow for concurrent repairing or replacement for each SFP. The data flow on the unaffected channels on the same feature can continue. A problem with one FICON port no longer requires replacement of a complete feature.

All FICON Express16S, FICON Express8S, and FICON Express8 features also support cascading, which is the connection of two FICON Directors in succession. This configuration minimizes the number of cross-site connections and helps reduce implementation costs for disaster recovery applications, IBM Geographically Dispersed Parallel Sysplex™ (GDPS), and remote copy.

z13s servers now support 32K devices per FICON channel for all FICON features, which were 24K in previous generations.

Each FICON Express16S, FICON Express8S, and FICON Express8 channel can be defined independently for connectivity to servers, switches, directors, disks, tapes, and printers:

- ▶ CHPID type Fibre Channel (FC): FICON, zHPF, and FCTC. All of these protocols are supported simultaneously.
- ▶ CHPID type FCP: Fibre Channel Protocol that supports attachment to SCSI devices directly or through Fibre Channel switches or directors.

FICON channels (CHPID type FC or FCP) can be shared among LPARs and can be defined as spanned. All ports on a FICON feature must be of the same type, either LX or short wavelength (SX). The features are connected to a FICON capable control unit, either point-to-point or switched point-to-point, through a Fibre Channel switch.

Statement of Direction: The z13 and z13s servers are the last z Systems servers to support FICON Express8 features for 2 Gbps connectivity.

All statements regarding IBM plans, directions, and intent are subject to change or withdrawal without notice. Any reliance on these statements of general direction is at the relying party's sole risk and will not create liability or obligation for IBM.

FICON Express16S

The FICON Express16S feature installs exclusively in the PCIe I/O drawer. Each of the two independent ports is capable of 4 Gbps, 8 Gbps, or 16 Gbps. The link speed depends on the capability of the attached switch or device. The link speed is auto-negotiated, point-to-point, and is transparent to users and applications.

The following types of FICON Express16S optical transceivers are supported

- ▶ FICON Express16S 10 km LX feature, FC 0418, with two ports per feature, supporting LC Duplex connectors
- ▶ FICON Express16S SX feature, FC 0419, with two ports per feature, supporting LC Duplex connectors

Each port of the FICON Express16S 10 km LX feature uses an optical transceiver, which supports an unrepeated distance of 10 km (6.2 miles) by using 9 µm single-mode fiber.

Each port of the FICON Express16S SX feature uses an optical transceiver, which supports to up to 125 m (410feet) of distance depending on the fiber that is used.

Consideration: FICON Express16S features do not support auto-negotiation to a data link rate of 2 Gbps.

FICON Express8S

The FICON Express8S feature installs exclusively in the PCIe I/O drawer. Each of the two independent ports is capable of 2 Gbps, 4 Gbps, or 8 Gbps. The link speed depends on the capability of the attached switch or device. The link speed is auto-negotiated, point-to-point, and is transparent to users and applications.

The following types of FICON Express8S optical transceivers are supported:

- ▶ FICON Express8S 10 km LX feature, FC 0409, with two ports per feature, supporting LC Duplex connectors
- ▶ FICON Express8S SX feature, FC 0410, with two ports per feature, supporting LC Duplex connectors

Each port of the FICON Express8S 10 km LX feature uses an optical transceiver, which supports an unrepeated distance of 10 km (6.2 miles) by using 9 μ m single-mode fiber.

Each port of the FICON Express8S SX feature uses an optical transceiver, which supports up to 150 m (492 feet) of distance depending on the fiber used.

Statement of Direction: The IBM z13 and z13s servers will be the last z Systems servers to offer ordering of FICON Express8S channel features. Enterprises that have 2 Gb device connectivity requirements must carry forward these channels.

All statements regarding IBM plans, directions, and intent are subject to change or withdrawal without notice. Any reliance on these statements of general direction is at the relying party's sole risk and will not create liability or obligation for IBM.

FICON Express8

The FICON Express8 feature for z13s servers installs in the I/O drawer (carry-forward only). Each of the four independent ports is capable of 2 Gbps, 4 Gbps, or 8 Gbps. The link speed depends on the capability of the attached switch or device. The link speed is auto-negotiated, point-to-point, and is transparent to users and applications.

The following types of FICON Express8 optical transceivers are supported:

- ▶ FICON Express8 10 km LX feature, FC 3325, with four ports per feature, supporting LC Duplex connectors
- ▶ FICON Express8 SX feature, FC 3326, with four ports per feature, supporting LC Duplex connectors

Each port of FICON Express8 10 km LX feature uses an optical transceiver, which supports an unrepeated distance of 10 km (6.2 miles) by using 9 μ m single-mode fiber.

Each port of FICON Express8 SX feature uses an optical transceiver, which supports to up to 150 m (492 feet) of distance depending on the fiber used.

FICON enhancements

Together with the FICON Express16S, z13s servers have introduced enhancements for FICON in both functional and performance aspects.

Forward Error Correction

Forward Error Correction (FEC) is a technique used for reducing data errors when transmitting over unreliable or noisy communication channels (improving signal to noise ratio). By adding redundancy error-correction code (ECC) to the transmitted information, the receiver can detect and correct a number of errors, without requiring retransmission. This process improves the signal reliability and bandwidth utilization by reducing retransmissions due to bit errors, especially for connections across long distance, like an inter-switch link (ISL) in a GDPS Metro Mirror environment.

The FICON Express16S is designed to support FEC coding on top of its 64b/66b data encoding for 16Gbps connections. This design can correct up to 11 bit errors per 2112 bits transmitted. While connected to devices that support FEC at 16Gbps connections, the FEC design allows FICON Express16S channels to operate at higher speeds, over longer distances, with reduced power and higher throughput, while retaining the same reliability and robustness that FICON channels have traditionally been known for.

With the IBM DS8870, z13s servers can extend the use of FEC to the fabric N_Ports for a complete end-to-end coverage of 16Gbps FC links. For more information, see *IBM DS8884 and z13s: A new cost optimized solution*, REDP-5327.

FICON dynamic routing

With the IBM z13 and IBM z13s servers, FICON channels are no longer restricted to the use of static SAN routing policies for ISLs for cascaded FICON directors. The z Systems now supports dynamic routing in the SAN with the FICON Dynamic Routing (FIDR) feature. It is designed to support the dynamic routing policies provided by the FICON director manufacturers, for example, Brocade's Exchange Based Routing (EBR) and Cisco's Open Exchange ID Routing (OxID). Check with the switch provider for their support statement.

A static SAN routing policy normally assigns the ISL routes according to the incoming port and its destination domain (port based routing), or the source and destination ports pairing (device based routing):

- ▶ Port based routing (PBR) assigns the ISL routes statically based on first come, first served, when a port starts a fabric login (FLOGI) to a destination domain. The ISL is round robin selected for assignment. I/O flow from same incoming port to same destination domain will always be assigned the same ISL route, regardless of the destination port of each I/O. This system can result in some ISLs being overloaded while others are underutilized. The ISL routing table is changed every time a z Systems server undergoes a power-on-reset (POR), so the ISL assignment is somehow unpredictable.
- ▶ Device based routing (DBR) assigns the ISL routes statically based on a hash of the source and destination port. That I/O flow from same incoming port to same destination is assigned to the same ISL route. Compared to PBR, the DBR is more capable of spreading load across ISLs for I/O flow from same incoming port to different destination ports within a same destination domain.

When using a static SAN routing policy, the FICON director is limited to assigning ISL routes based on workload, and ISL can be overloaded or underused.

With dynamic routing, ISL routes are dynamically changed based on the Fibre Channel exchange ID, which is unique for each I/O operation. ISL is assigned at I/O request time, so different I/Os from same incoming port to same destination port are assigned different ISLs.

With FIDR, z13s servers have advantages for performance and management in configurations with ISL and cascaded FICON directors:

- ▶ Support sharing of ISLs between FICON and FCP (PPRC or distributed)
- ▶ I/O traffics is better balanced between all available ISLs

- ▶ Improve utilization of FICON director and ISL
- ▶ Easier to manage with a predictable and repeatable I/O performance.

FICON dynamic routing can be enabled by definition of dynamic routing capable switches and control units in HCD. Also, z/OS has implemented a health check function for FICON dynamic routing.

Improved zHPF I/O execution at distance

By introducing the concept of pre-deposit writes, zHPF reduces the number of round trips of standard FCP I/Os to a single round trip. Originally this benefit is limited to writes that are less than 64 KB. zHPF on z13s and z13 servers has been enhanced to allow all large write operations (> 64 KB) at distances up to 100 km to be run in a single round trip to the control unit. This process does not elongate the I/O service time for these write operations at extended distances.

Read Diagnostic Parameter Extended Link Service Support

To improve the accuracy of identifying a failed component without unnecessarily replacing components in a SAN fabric, a new Extended Link Service (ELS) command called Read Diagnostic Parameter (RDP) has been added to the Fibre Channel T11 standard. The command allows z Systems to obtain extra diagnostic data from the Small Form-Factor Pluggable (SFP) optics located throughout the SAN fabric.

z13s servers now can read this additional diagnostic data for all the ports accessed in the I/O configuration and make the data available to an LPAR. For z/OS LPARs using FICON channels, z/OS displays the data with a new message and display command. For Linux on z Systems, z/VM, z/VSE, and KVM LPARs using FCP channels, this diagnostic data is available in a new window in the SAN Explorer tool.

N_Port ID Virtualization enhancement

N_Port ID Virtualization (NPIV) allows multiple system images (in LPARs or z/VM guests) to use a single FCP channel as though each were the sole user of the channel. This feature, first introduced with z9 EC, can be used with earlier FICON features that have been carried forward from earlier servers.

Using the FICON Express16S as an FCP channel with NPIV enabled, the maximum numbers of these aspects for one FCP physical channel are doubled:

1. Maximum number of NPIV hosts defined: Increased from 32 to 64
2. Maximum number of remote N_Ports communicated: Increased from 512 to 1024
3. Maximum number of addressable LUNs: Increased from 4096 to 8192
4. Concurrent I/O operations: Increased from 764 to 1528

For for information about NPIV, see 7.3.40, “N_Port ID Virtualization” on page 267.

Export / import physical port WWPNs for FCP Channels

z Systems automatically assigns worldwide port names (WWPNs) to the physical ports of an FCP channel based on the PCHID. This WWPN assignment changes when an FCP Channel is moved to a different physical slot position. z13 and z13s servers will now allow for the modification of these default assignments, allowing FCP channels to keep previously assigned WWPNs, even after being moved to a different slot position. This capability can eliminate the need for reconfiguration of the SAN in many situations, and is especially helpful on a system upgrade.

For more information about the FICON enhancement of z13 and z13s servers, see *Get More Out of Your IT Infrastructure with IBM z13 I/O Enhancements*, REDP-5134.

FICON support for multiple-hop cascaded SAN configurations

Before the introduction of the z13 and z13s servers, z Systems FICON SAN configurations only supported a single ISL (or single hop) in a cascaded FICON SAN environment. IBM z13s servers now support up to three hops in a cascaded FICON SAN environment. This support allows you to more easily configure a three or four site disaster recovery solution.

FICON feature summary

Table 4-9 shows the FICON card feature codes, cable type, maximum unrepeated distance, and the link data rate on a z13s server. All FICON features use LC Duplex connectors.

Table 4-9 z13 channel feature support

Channel feature	Feature codes	Bit rate	Cable type	Maximum unrepeated distance^a (MHz - km)
FICON Express16S 10KM LX	0418	4, 8, or 16 Gbps	SM 9 µm	10 km
FICON Express16S SX	0419	16 Gbps	MM 50 µm	35 m (500) 100 m (2000) 125 m (4700)
		8 Gbps	MM 62.5 µm MM 50 µm	21 m (200) 50 m (500) 150 m (2000) 190 m (4700)
		4 Gbps	MM 62.5 µm MM 50 µm	70 m (200) 150 m (500) 380 m (2000) 400 m (4700)
FICON Express8S 10KM LX	0409	2, 4, or 8 Gbps	SM 9 µm	10 km
FICON Express8S SX	0410	8 Gbps	MM 62.5 µm MM 50 µm	21 m (200) 50 m (500) 150 m (2000) 190 m (4700)
		4 Gbps	MM 62.5 µm MM 50 µm	70 m (200) 150 m (500) 380 m (2000) 400 m (4700)
		2 Gbps	MM 62.5 µm MM 50 µm	150 m (200) 300 m (500) 500 m (2000) N/A (4700)
FICON Express8 10KM LX	3325	2, 4, or 8 Gbps	SM 9 µm	10 km

Channel feature	Feature codes	Bit rate	Cable type	Maximum unrepeated distance ^a (MHz - km)
FICON Express8 SX	3326	8 Gbps	MM 62.5 μm MM 50 μm	21 m (200) 50 m (500) 150 m (2000) 190 m (4700)
		4 Gbps	MM 62.5 μm MM 50 μm	70 m (200) 150 m (500) 380 m (2000) 400 m (4700)
		2 Gbps	MM 62.5 μm MM 50 μm	150 m (200) 300 m (500) 500 m (2000) N/A (4700)

a. Minimum fiber bandwidths in MHz/km for multimode fiber optic links are included in parentheses where applicable.

4.8.3 OSA-Express5S

The OSA-Express5S feature is exclusively in the PCIe I/O drawer. The following OSA-Express5S features can be installed on z13s servers:

- ▶ OSA-Express5S 10 Gigabit Ethernet LR, FC 0415
- ▶ OSA-Express5S 10 Gigabit Ethernet SR, FC 0416
- ▶ OSA-Express5S Gigabit Ethernet LX, FC 0413
- ▶ OSA-Express5S Gigabit Ethernet SX, FC 0414
- ▶ OSA-Express5S 1000BASE-T Ethernet, FC 0417

Table 4-10 lists the OSA-Express5S features.

Table 4-10 OSA-Express5S features

I/O feature	Feature code	Number of ports per feature	Port increment	Maximum number of ports	Maximum number of features	CHPID type
OSA-Express5S 10 GbE LR	0415	1	1	48	48	OSD, OSX
OSA-Express5S 10 GbE SR	0416	1	1	48	48	OSD, OSX
OSA-Express5S GbE LX	0413	2	2	96	48	OSD
OSA-Express5S GbE SX	0414	2	2	96	48	OSD
OSA-Express5S 1000BASE-T	0417	2	2	96	48	OSC, OSD, OSE, OSM, OSN

OSA-Express5S 10 Gigabit Ethernet LR (FC 0415)

The OSA-Express5S 10 Gigabit Ethernet (GbE) Long Reach (LR) feature has one PCIe adapter and one port per feature. The port supports CHPID types OSD and OSX. When

defined as CHPID type OSX, the 10 GbE port provides connectivity and access control to the intraensemble data network (IEDN) from z13s servers to z BladeCenter Extension (zBX). The 10 GbE feature is designed to support attachment to a single-mode fiber 10 Gbps Ethernet LAN or Ethernet switch that is capable of 10 Gbps. The port can be defined as a spanned channel and can be shared among LPARs within and across logical channel subsystems.

The OSA-Express5S 10 GbE LR feature supports the use of an industry standard small form factor LC Duplex connector. Ensure that the attaching or downstream device has an LR transceiver. The transceivers at both ends must be the same (LR to LR, which might also be referred to as LW or LX).

The OSA-Express5S 10 GbE LR feature does not support auto-negotiation to any other speed and runs in full duplex mode only.

A 9 μm single-mode fiber optic cable that terminates with an LC Duplex connector is required for connecting this feature to the selected device.

OSA-Express5S 10 Gigabit Ethernet SR (FC 0416)

The OSA-Express5S 10 Gigabit Ethernet (GbE) Short Reach (SR) feature has one PCIe adapter and one port per feature. The port supports CHPID types OSD and OSX. When defined as CHPID type OSX, the 10 GbE port provides connectivity and access control to the IEDN from z13s servers to zBX. The 10 GbE feature is designed to support attachment to a multimode fiber 10 Gbps Ethernet LAN or Ethernet switch that is capable of 10 Gbps. The port can be defined as a spanned channel and can be shared among LPARs within and across logical channel subsystems.

The OSA-Express5S 10 GbE SR feature supports the use of an industry standard small form factor LC Duplex connector. Ensure that the attaching or downstream device has an SR transceiver. The sending and receiving transceivers must be the same (SR to SR, which might also be referred to as SW or SX).

The OSA-Express5S 10 GbE SR feature does not support auto-negotiation to any other speed and runs in full duplex mode only.

A 50 or a 62.5 μm multimode fiber optic cable that terminates with an LC Duplex connector is required for connecting each port on this feature to the selected device.

OSA-Express5S Gigabit Ethernet LX (FC 0413)

The OSA-Express5S Gigabit Ethernet (GbE) long wavelength (LX) feature has one PCIe adapter and two ports. The two ports share a channel path identifier (CHPID type OSD exclusively). The ports support attachment to a 1 Gbps Ethernet LAN. Each port can be defined as a spanned channel and can be shared among LPARs and across logical channel subsystems.

The OSA-Express5S GbE LX feature supports the use of an LC Duplex connector. Ensure that the attaching or downstream device has an LX transceiver. The sending and receiving transceivers must be the same (LX to LX).

A 9 μm single-mode fiber optic cable that terminates with an LC Duplex connector is required for connecting each port on this feature to the selected device. If multimode fiber optic cables are being reused, a pair of Mode Conditioning Patch cables is required, with one cable for each end of the link.

OSA-Express5S Gigabit Ethernet SX (FC 0414)

The OSA-Express5S Gigabit Ethernet (GbE) short wavelength (SX) feature has one PCIe adapter and two ports. The two ports share a channel path identifier (CHPID type OSD exclusively). The ports support attachment to a 1 Gbps Ethernet LAN. Each port can be defined as a spanned channel and can be shared among LPARs and across logical channel subsystems.

The OSA-Express5S GbE SX feature supports the use of an LC Duplex connector. Ensure that the attaching or downstream device has an SX transceiver. The sending and receiving transceivers must be the same (SX to SX).

A multi-mode fiber optic cable that terminates with an LC Duplex connector is required for connecting each port on this feature to the selected device.

OSA-Express5S 1000BASE-T Ethernet feature (FC 0417)

Feature code 0417 occupies one slot in the PCIe I/O drawer. It has two ports that connect to a 1000 Mbps (1 Gbps) or 100 Mbps Ethernet LAN. Each port has an SFP with an RJ-45 receptacle for cabling to an Ethernet switch. The RJ-45 receptacle is required to be attached by using an EIA/TIA Category 5 or Category 6 unshielded twisted pair (UTP) cable with a maximum length of 100 m (328 ft.). The SFP allows a concurrent repair or replace action.

The OSA-Express5S 1000BASE-T Ethernet feature supports auto-negotiation when attached to an Ethernet router or switch. If you allow the LAN speed and duplex mode to default to auto-negotiation, the OSA-Express port and the attached router or switch auto-negotiate the LAN speed and duplex mode settings between them. They then connect at the highest common performance speed and duplex mode of interoperation. If the attached Ethernet router or switch does not support auto-negotiation, the OSA-Express port examines the signal that it is receiving and connects at the speed and duplex mode of the device at the other end of the cable.

The OSA-Express5S 1000BASE-T Ethernet feature can be configured as CHPID type OSC, OSD, OSE, OSM, or OSN. Non-QDIO operation mode requires CHPID type OSE. When defined as CHPID type OSM, the port provides connectivity to the intranode management network (INMN).

The following settings are supported on the OSA-Express5S 1000BASE-T Ethernet feature port:

- ▶ Auto-negotiate
- ▶ 100 Mbps half-duplex or full-duplex
- ▶ 1000 Mbps full-duplex

If you are not using auto-negotiate, the OSA-Express port attempts to join the LAN at the specified speed and duplex mode. If this specified speed and duplex mode do not match the speed and duplex mode of the signal on the cable, the OSA-Express port does not connect.

4.8.4 OSA-Express4S features

This section addresses the characteristics of all OSA-Express4S features that are supported on z13s servers.

The OSA-Express4S feature is exclusively in the PCIe I/O drawer. The following OSA-Express4S features can be installed on z13s servers:

- ▶ OSA-Express4S 10 Gigabit Ethernet LR, FC 0406
- ▶ OSA-Express4S 10 Gigabit Ethernet SR, FC 0407

- ▶ OSA-Express4S Gigabit Ethernet LX, FC 0404
- ▶ OSA-Express4S Gigabit Ethernet SX, FC 0405
- ▶ OSA-Express4S 1000BASE-T Ethernet, FC 0408

Table 4-11 lists the characteristics of the OSA-Express4S features.

Table 4-11 OSA-Express4S features

I/O feature	Feature code	Number of ports per feature	Port increment	Maximum number of ports (CHPIDs)	Maximum number of features	CHPID type
OSA-Express4S 10 GbE LR	0406	1	1	48	48	OSD, OSX
OSA-Express4S 10 GbE SR	0407	1	1	48	48	OSD, OSX
OSA-Express4S GbE LX	0404	2	2	96	48	OSD
OSA-Express4S GbE SX	0405	2	2	96	48	OSD
OSA-Express4S 1000BASE-T	0408	2	2	96	48	OSC, OSD, OSE, OSM, OSN

OSA-Express4S 10 Gigabit Ethernet LR (FC 0406)

The OSA-Express4S 10 Gigabit Ethernet (GbE) LR feature has one PCIe adapter and one port per feature. The port supports CHPID types OSD and OSX. When defined as CHPID type OSX, the 10 GbE port provides connectivity and access control to the IEDN from z13s servers to IBM zBX. The 10 GbE feature is designed to support attachment to a single mode fiber 10-Gbps Ethernet LAN or Ethernet switch that is capable of 10 Gbps. The port can be defined as a spanned channel, and can be shared among LPARs within and across logical channel subsystems.

The OSA-Express4S 10 GbE LR feature supports the use of an industry standard small form factor LC Duplex connector. Ensure that the attaching or downstream device has an LR transceiver. The sending and receiving transceivers must be the same (LR to LR).

The OSA-Express4S 10 GbE LR feature does not support auto-negotiation to any other speed and runs in full duplex mode only. OSA-Express4S 10 GbE LR supports 64B/66B encoding. However, the GbE supports 8B/10 encoding, making auto-negotiation to any other speed impossible.

A 9- μ m single mode fiber optic cable that terminates with an LC Duplex connector is required for connecting this feature to the selected device.

OSA-Express4S 10 Gigabit Ethernet SR (FC 0407)

The OSA-Express4S 10 Gigabit Ethernet (GbE) SR feature has one PCIe adapter and one port per feature. The port supports CHPID types OSD and OSX. When defined as CHPID type OSX, the 10 GbE port provides connectivity and access control to the IEDN from z13s servers to zBX. The 10 GbE feature is designed to support attachment to a multimode fiber 10Gbps Ethernet LAN or Ethernet switch capable of 10 Gbps. The port can be defined as a

spanned channel, and can be shared among LPARs within and across logical channel subsystems.

The OSA-Express4S 10 GbE SR feature supports the use of an industry standard small form factor LC Duplex connector. Ensure that the attaching or downstream device has an SR transceiver. The sending and receiving transceivers must be the same (SR to SR).

The OSA-Express4S 10 GbE SR feature does not support auto-negotiation to any other speed and runs in full duplex mode only. OSA-Express4S 10 GbE SR supports 64B/66B encoding. However, the GbE supports 8B/10 encoding, making auto-negotiation to any other speed impossible.

A 50 or a 62.5- μ m multimode fiber optic cable that terminates with an LC Duplex connector is required for connecting each port on this feature to the selected device.

OSA-Express4S Gigabit Ethernet LX (FC 0404)

The OSA-Express4S GbE LX feature has one PCIe adapter and two ports. The two ports share a channel path identifier (CHPID type OSD exclusively). The ports support attachment to a 1-Gbps Ethernet LAN. Each port can be defined as a spanned channel, and can be shared among LPARs and across logical channel subsystems.

The OSA-Express4S GbE LX feature supports the use of an LC Duplex connector. Ensure that the attaching or downstream device has an LX transceiver. The sending and receiving transceivers must be the same (LX to LX).

A 9- μ m single mode fiber optic cable that terminates with an LC Duplex connector is required for connecting each port on this feature to the selected device. If multimode fiber optic cables are being reused, a pair of Mode Conditioning Patch cables is required, one for each end of the link.

OSA-Express4S Gigabit Ethernet SX (FC 0405)

The OSA-Express4S Gigabit Ethernet (GbE) SX feature has one PCIe adapter and two ports. The two ports share a channel path identifier (CHPID type OSD exclusively). The ports support attachment to a 1 Gbps Ethernet LAN. Each port can be defined as a spanned channel, and can be shared among LPARs and across logical channel subsystems.

The OSA-Express4S GbE SX feature supports the use of an LC Duplex connector. Ensure that the attaching or downstream device has an SX transceiver. The sending and receiving transceivers must be the same (SX to SX).

A 50 or a 62.5- μ m multimode fiber optic cable that terminates with an LC Duplex connector is required for connecting each port on this feature to the selected device.

OSA-Express4S 1000BASE-T Ethernet feature (FC 0408)

The OSA-Express4S 1000BASE-T Ethernet feature occupies one slot in the PCIe I/O drawer. It has two ports that connect to a 1000 Mbps (1 Gbps), 100 Mbps, or 10 Mbps Ethernet LAN. Each port has an RJ-45 receptacle for cabling to an Ethernet switch. The RJ-45 receptacle must be attached by using an EIA/TIA Category 5 or Category 6 UTP cable with a maximum length of 100 meters (328 ft).

The OSA-Express4S 1000BASE-T Ethernet feature supports auto-negotiation when attached to an Ethernet router or switch. If you allow the LAN speed and duplex mode to default to auto-negotiation, the OSA-Express port and the attached router or switch auto-negotiate the LAN speed and duplex mode settings between them. They connect at the highest common performance speed and duplex mode of interoperation. If the attached Ethernet router or switch does not support auto-negotiation, the OSA-Express port examines the signal that it is receiving. It then connects at the speed and duplex mode of the device at the other end of the cable.

The OSA-Express4S 1000BASE-T Ethernet feature can be configured as CHPID type OSC, OSD, OSE, OSN, or OSM. Non-QDIO operation mode requires CHPID type OSE. When defined as CHPID type OSM, the port provides connectivity to the INMN.

The following settings are supported on the OSA-Express4 1000BASE-T Ethernet feature port:

- ▶ Auto-negotiate
- ▶ 10 Mbps half-duplex or full-duplex
- ▶ 100 Mbps half-duplex or full-duplex
- ▶ 1000 Mbps full-duplex

If you are not using auto-negotiate, the OSA-Express port attempts to join the LAN at the specified speed and duplex mode. If this does not match the speed and duplex mode of the signal on the cable, the OSA-Express port does not connect.

4.8.5 OSA-Express for ensemble connectivity

The following OSA-Express features are used to connect z13s servers to IBM z BladeCenter Extension (zBX) Model 004 and other ensemble nodes:

- ▶ OSA-Express5S 10 Gigabit Ethernet (GbE) Long Reach (LR), FC 0415
- ▶ OSA-Express5S 10 Gigabit Ethernet (GbE) Short Reach (SR), FC 0416
- ▶ OSA-Express5S 1000BASE-T Ethernet, FC 0417
- ▶ OSA-Express4S 10 Gigabit Ethernet (GbE) Long Reach (LR), FC 0406
- ▶ OSA-Express4S 10 Gigabit Ethernet (GbE) Short Reach (SR), FC 0407
- ▶ OSA-Express4S 1000BASE-T Ethernet, FC 0408

Intraensemble data network

The IEDN is a private and secure 10-Gbps Ethernet network. It connects all elements of an ensemble, and is access-controlled by using integrated virtual LAN (VLAN) provisioning. No client-managed switches or routers are required. The IEDN is managed by a primary Hardware Management Console (HMC).

The IEDN connection requires two ports. The following features can be used, which are configured as CHPID type OSX:

- ▶ OSA-Express5S 10 GbE
- ▶ OSA-Express4S 10 GbE

For redundancy, one port each from two OSA-Express 10 GbE features must be configured.

The connection is from the z13s server to the IEDN top-of-rack (ToR) switches on the zBX Model 004. With a stand-alone z13s node (no-zBX), the connection is interconnect pairs of OSX ports through LC Duplex directly connected cables, not wrap cables as was previously recommended.

Intranode management network

The INMN is a private and physically isolated 1000BASE-T Ethernet internal management network that operates at 1 Gbps. It connects all resources (z13s and zBX Model 004 components) of an ensemble node for management purposes. It is pre-wired, internally switched, configured, and managed with full redundancy for high availability.

The INMN requires two ports (CHPID port 0 from two OSA-Express5S 1000BASE-T features, or OSA-Express4S 1000BASE-T features; CHPID port 1 is not used at all in this case) that are configured as CHPID type OSM. The connection is through a System Control Hub (SCH) in the z13s server. Because it is a stand-alone node, the INMN ToR switches on zBX Model 004 are not connected to the SCHs.

Ensemble HMC management functions

An HMC can manage multiple z Systems servers and can be at a local or a remote site. If the z13s server is defined as a member of an ensemble, a pair of HMCs (a primary and an alternate) is required, and certain restrictions apply. The primary HMC is required to manage ensemble network connectivity, the INMN, and the IEDN network.

For more information, see 11.6, “HMC in an ensemble” on page 428 and 9.5, “RAS capability for the HMC and SE” on page 362.

4.8.6 HiperSockets

The HiperSockets function of z13s servers provides up to 32 high-speed virtual LAN attachments, just like the IBM zEnterprise zEC12, IBM zEnterprise BC12, IBM zEnterprise 196, and IBM zEnterprise 114 servers. Previous servers provided 16 attachments.

HiperSockets IOCP definitions on z13 and z13s servers: There is a new parameter for HiperSockets IOCP definitions on z13 and z13s servers. As such, the z13 and z13s IOCP definitions need to be migrated to support the HiperSockets definitions (CHPID type IQD).

On z13 and z13s servers, the CHPID statement of HiperSockets devices requires the keyword VCHID. VCHID specifies the virtual channel identification number associated with the channel path. Valid range is 7E0 - 7FF.

VCHID is not valid on z Systems before the z13 and z13s servers.

For more information, see the *z Systems Input/Output Configuration Program User's Guide for ICP IOCP*, SB10-7163.

HiperSockets can be customized to accommodate varying traffic sizes. Because HiperSockets does not use an external network, it can free up system and network resources. This advantage can help eliminate attachment costs, and improve availability and performance.

HiperSockets eliminates having to use I/O subsystem operations and having to traverse an external network connection to communicate between LPARs in the same z13s server. HiperSockets offers significant value in server consolidation when connecting many virtual servers. It can be used instead of certain coupling link configurations in a Parallel Sysplex.

HiperSockets internal networks support two transport modes:

- ▶ Layer 2 (link layer)
- ▶ Layer 3 (network or IP layer)

Traffic can be IPv4 or IPv6, or non-IP, such as AppleTalk, DECnet, IPX, NetBIOS, or SNA.

HiperSockets devices are protocol-independent and Layer 3 independent. Each HiperSockets device (Layer 2 and Layer 3 mode) has its own MAC address. This address allows the use of applications that depend on the existence of Layer 2 addresses, such as Dynamic Host Configuration Protocol (DHCP) servers and firewalls. Layer 2 support helps facilitate server consolidation, and can reduce complexity and simplify network configuration. It also allows LAN administrators to maintain the mainframe network environment similarly to non-mainframe environments.

Packet forwarding decisions are based on Layer 2 information instead of Layer 3. The HiperSockets device can run automatic MAC address generation to create uniqueness within and across LPARs and servers. The use of Group MAC addresses for multicast is supported, and broadcasts to all other Layer 2 devices on the same HiperSockets networks.

Datagram is delivered only between HiperSockets devices that use the same transport mode. A Layer 2 device cannot communicate directly to a Layer 3 device in another LPAR network. A HiperSockets device can filter inbound datagrams by VLAN identification, the destination MAC address, or both.

Analogous to the Layer 3 functions, HiperSockets Layer 2 devices can be configured as primary or secondary connectors, or multicast routers. This configuration enables the creation of high-performance and high-availability link layer switches between the internal HiperSockets network and an external Ethernet network. It also can be used to connect to the HiperSockets Layer 2 networks of different servers.

HiperSockets Layer 2 in the z13s server is supported by Linux on z Systems, and by z/VM for Linux guest use.

z13s, z13, and the other zEnterprise CPCs (zEC12, zBC12, z196, and z114 servers), support the HiperSockets Completion Queue function that is designed to allow HiperSockets to transfer data synchronously if possible, and asynchronously if necessary. This feature combines ultra-low latency with more tolerance for traffic peaks. With the asynchronous support, during high volume situations, data can be temporarily held until the receiver has buffers that are available in its inbound queue. The HiperSockets Completion Queue function requires the following applications at a minimum:

- ▶ z/OS V1.13
- ▶ Linux on z Systems distributions:
 - Red Hat Enterprise Linux (RHEL) 6.2
 - SUSE Linux Enterprise Server (SLES) 11 SP2
- ▶ z/VSE V5.1.1
- ▶ z/VM V6.2 with maintenance

The z13s, z13, and the zEnterprise servers provide the capability to integrate HiperSockets connectivity to the IEDN. This configuration extends the reach of the HiperSockets network outside the CPC to the entire ensemble, which is displayed as a single Layer 2. Because HiperSockets and IEDN are both internal z Systems networks, the combination allows z Systems virtual servers to use the optimal path for communications.

In z/VM V6.2, the virtual switch function is enhanced to transparently bridge a guest virtual machine network connection on a HiperSockets LAN segment. This bridge allows a single HiperSockets guest virtual machine network connection to communicate directly with these systems:

- ▶ Other guest virtual machines on the virtual switch
- ▶ External network hosts through the virtual switch OSA UPLINK port

Shared Memory Communication - Direct Memory Access

With the z13 and z13s servers, IBM introduces Shared Memory Communication - Direct Memory Access (SMC-D). SMC-D maintains the socket-API transparency aspect of SMC-R so that applications using TCP/IP communications can benefit immediately without requiring any application software or IP topology changes. SMC-D completes the overall Shared Memory Communications solution, providing synergy with SMC-R. Both protocols use shared memory architectural concepts, eliminating TCP processing in the data path, yet preserving TCP/IP Qualities of Service for connection management purposes.

Internal Shared Memory

Internal Shared Memory (ISM) is a new function supported by the z13 and z13s machines. It is the firmware that provides the connectivity for shared memory access between multiple operating systems within the same CPC. ISM creates virtual adapters with shared memory allocated for each OS, that OS can use these virtual adapters for SMC-D data exchanges within the same CPC.

ISM is defined by FUNCTION statement with a VCHID in HCD/IOCDS. Identified by the PNETID parameter, each ISM VCHID defines an isolated, internal virtual network for SMC-D communication, without any hardware component required. Virtual adapters are defined by virtual function (VF) statements. Multiple LPARs have share accessing to same virtual network for SMC-D data exchange by associating their VF with same VCHID.

z13s servers support up to 32 ISM VCHIDs per CPC, and each VCHID supports up to 255 VFs, with a total maximum of 8K VFs. For more information about the SMC-D and ISM, see Appendix D, “Shared Memory Communications” on page 475

HiperSockets can have network latency and CPU reduction benefits and performance improvement, by using the SMC-D over ISM.

4.9 Parallel Sysplex connectivity

Coupling links are required in a Parallel Sysplex configuration to provide connectivity from the z/OS images to the coupling facility. A properly configured Parallel Sysplex provides a highly reliable, redundant, and robust z Systems technology solution to achieve near-continuous availability. A Parallel Sysplex is composed of one or more z/OS operating system images that are coupled through one or more coupling facilities.

4.9.1 Coupling links

The type of coupling link that is used to connect a coupling facility (CF) to an operating system LPAR is important. The link performance has a significant effect on response times and coupling processor usage. For configurations that cover large distances, the time that is spent on the link can be the largest part of the response time.

These links are available to connect an operating system LPAR to a coupling facility:

- ▶ Integrated Coupling Adapter (ICA SR) for short distance connectivity and defined as CHPID type CS5. The ICA SR can be used only for coupling connectivity between z13 and z13s servers. It does not support connectivity to zEC12, zBC12, z196, or z114 servers, and it cannot be connected to HCA3-O or HCA3-O LR coupling fanouts. The ICA SR supports distances up to 150 m and a link data rate of 8 GBps. OM3 fiber optic cable is used for distances up to 100 m, and OM4 for distances up to 150 m. ICA SR supports four

CHPIDs per port and seven subchannels (devices) per CHPID. ICA SR supports transmission of STP messages.

- ▶ Parallel Sysplex using IFB connects to z13, z13s, zEC12, zBC12, z196, or z114 servers. 12x InfiniBand coupling links are fiber optic connections that support a maximum distance of up to 150 meters (492 ft.). IFB coupling links are defined as CHPID type CIB. IFB supports transmission of STP messages.

z13s servers support one type of 12x InfiniBand coupling link: FC 0171 HCA3-O (12xIFB) fanout.

- ▶ IFB LR: InfiniBand Long Reach (IFB LR) connects to z13, z13s, zEC12, zBC12, z196, or z114 servers. 1x InfiniBand coupling links are fiber optic connections that support a maximum unrepeated distance of up to 10 km (6.2 miles), and up to 100 km (62 miles) with a z Systems qualified DWDM. IFB LR coupling links are defined as CHPID type CIB, and support 7 or 32 subchannels per CHPID. IFB LR supports transmission of STP messages.

z13s servers support one type of 1x InfiniBand coupling link: FC 0170 HCA3-O LR (1xIFB) fanout.

- ▶ Internal Coupling (IC): CHPIDs (type ICP) that are defined for internal coupling can connect a CF to a z/OS LPAR in the same z13s server. IC connections require two CHPIDs to be defined, which can be defined only in peer mode. A maximum of 32 IC CHPIDs (16 connections) can be defined.

Note: ISC-3, HCA2-O, and HCA2-O LR are not supported on z13s servers. The zEC12 and zBC12 are the last servers that support ISC-3, HCA2-O, and HCA2-O LR coupling links (by carry-forward only). HCA3-O (LR) on z13s servers, however, *can* connect to HCA2-O (LR) on previous generation servers.

IFB and IFB LR links contend for adapter space. Total port counts vary depending upon the number of configured IFB links, IFB LR links, and I/O drawers (I/O drawers that are connected through HCA2-C).

Table 4-12 shows the coupling link options.

Table 4-12 Coupling link options

Type	Description	Use	Link rate ^a	Distance	z13s-N10 Maximum # ports	z13-N20 Maximum # ports
IFB	12x InfiniBand (HCA3-O) ^b	z13s to z13, z13s, zEC12, zBC12, z196, z114	6 Gbps	150 meters (492 feet)	4 ^c	16 ^c
IFB LR	1x IFB (HCA3-O LR)	z13 to z13, z13s, zEC12, zBC12, z196, z114	2.5 Gbps 5.0 Gbps	10 km unrepeated (6.2 miles) 100 km repeated (62 miles)	8 ^c	32 ^c
IC	Internal coupling channel	Internal communication	Internal speeds	N/A	32 ^d	32 ^d
ICA SR	Integrated Coupling Adapter	z13 to z13, z13s	8 Gbps	150 meters (492 feet)	8	16

a. The link data rates do not represent the actual performance of the link. The actual performance depends on many factors including latency through the adapters, cable lengths, and the type of workload.

- b. 12x IFB3 protocol supports a maximum of four CHPIDs and connects to the other HCA3-O port. Otherwise, use the 12x IFB protocol. The protocol is auto-configured when conditions are met for IFB3. For more information, see 4.6.4, “HCA3-O (12x IFB) fanout (FC 0171)” on page 150.
- c. Uses all available fanout slots. Allows no I/O drawer or other IFB coupling option.
- d. One IC connection (or ‘link’) requires the definition of two CHPIDs (‘logical ports’).

The maximum for IFB links is 32, for ICA SR is 16, and for IC is 32. The maximum number of combined external coupling links (active ICA SR links and IFB LR) is 80 per z13s server. And z13s servers support up to 256 coupling CHPIDs per CPC, which provides enhanced connectivity and scalability for a growing number of coupling channel types that is twice the 128 coupling CHPIDs that are supported on zEC12. Unlike the HCA3-O 12x InfiniBand links, the ICA SR cannot define more than four CHPIDs per port.

Coupling CHPIDs: On z13s servers, each CF image continues to support a maximum of 128 coupling CHPIDs.

z13s servers support various connectivity options, depending on the connected servers. Figure 4-8 shows z13s coupling link support for z13, zEC12, zBC12, z196, and z114 servers.

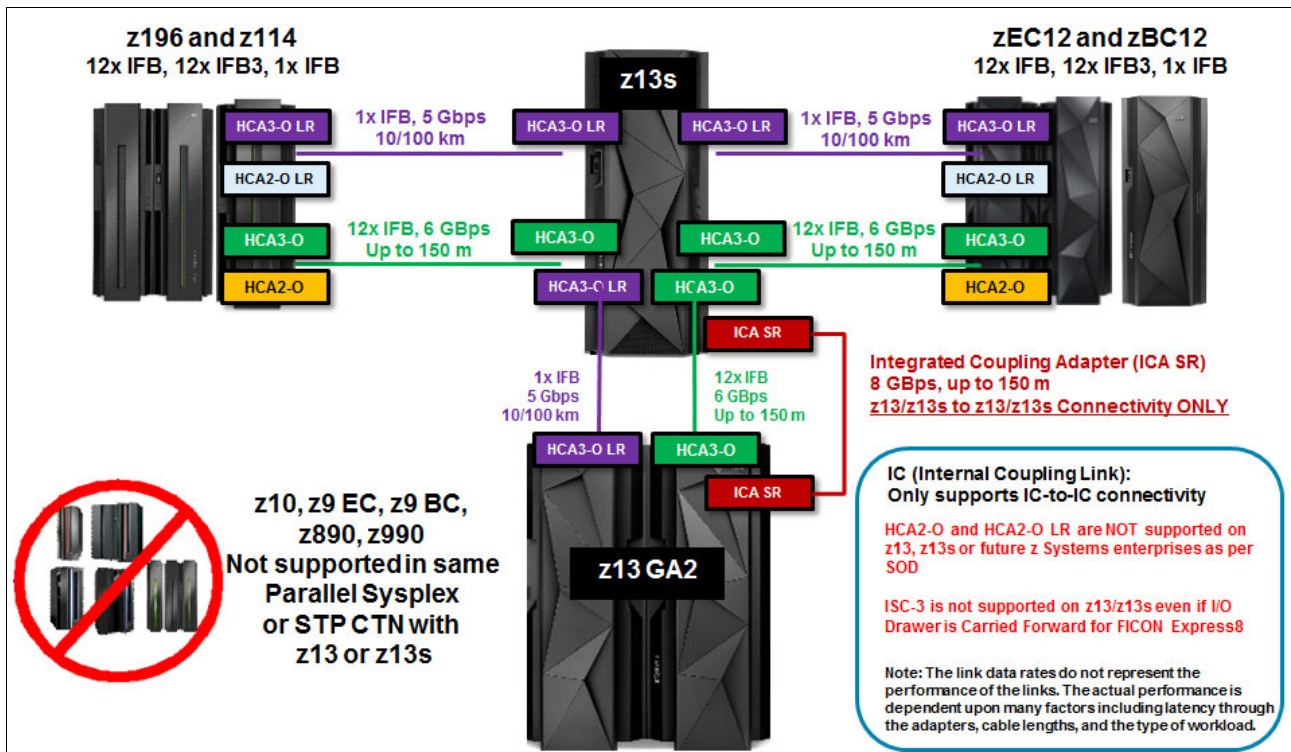


Figure 4-8 z13 Parallel Sysplex coupling connectivity

When defining IFB coupling links (CHPID type CIB), HCD defaults to seven subchannels. Thirty-two subchannels are supported only on HCA2-O LR (1xIFB) and HCA3-O LR (1xIFB) on zEC12 and later when both sides of the connection use 32 subchannels. Otherwise, change the default value from 7 to 32 subchannels on each CIB definition.

In a Parallel Sysplex configuration, z/OS and coupling facility images can run on the same or separate servers. There must be at least one CF that is connected to all z/OS images, although there can be other CFs that are connected only to selected z/OS images. Two coupling facility images are required for system-managed CF structure duplexing. In this case, each z/OS image must be connected to both duplexed CFs.

To eliminate any single points of failure in a Parallel Sysplex configuration, have at least the following components:

- ▶ Two coupling links between the z/OS and coupling facility images.
- ▶ Two coupling facility images not running on the same server.
- ▶ One stand-alone coupling facility. If using system-managed CF structure duplexing or running with *resource sharing* only, a stand-alone coupling facility is not mandatory.

Coupling link features

The z13s servers support three types of coupling link options:

- ▶ HCA3-O fanout for 12x InfiniBand, FC 0171
- ▶ HCA3-O LR fanout for 1x InfiniBand, FC 0170
- ▶ ICA SR fanout, FC0172

Various link options are available to connect a z13s server to other z Systems and zEnterprise servers:

- ▶ ICA SR fanout at 8 GBps to z13 or z13s servers
- ▶ 12x InfiniBand using HCA3-O fanout at 6 GBps to z13 server with HCA3-O fanout, or z13s, zEC12, zBC12, z196, and z114 servers with HCA3-O or HCA2-O fanout
- ▶ 1x InfiniBand using HCA3-O LR fanout at 5.0 or 2.5 Gbps to z13 server with HCA3-O LR, or z13s, zEC12, zBC12, z196, and z114 servers with HCA3-O LR or HCA2-O LR fanout

HCA3-O fanout for 12x InfiniBand (FC 0171)

For more information, see 4.6.4, “HCA3-O (12x IFB) fanout (FC 0171)” on page 150.

HCA3-O LR fanout for 1x InfiniBand (FC 0170)

For more information, see 4.6.5, “HCA3-O LR (1x IFB) fanout (FC 0170)” on page 151.

ICA SR fanout (FC0172)

For more information, see 4.6.3, “Integrated Coupling Adapter (FC 0172)” on page 149.

Extended distance support

For extended distance support, see the *System z End-to-End Extended Distance Guide*, SG24-8047.

Internal coupling links

IC links are LIC-defined links that connect a CF to a z/OS LPAR in the same server. These links are available on all z Systems servers. The IC link is a z Systems server coupling connectivity option that enables high-speed, efficient communication between a CF partition and one or more z/OS LPARs that run on the same server. The IC is a linkless connection (implemented in Licensed Internal Code (LIC)), and so does not require any hardware or cabling.

An IC link is a fast coupling link that uses memory-to-memory data transfers. IC links do not have PCHID numbers, but do require CHPIDs.

IC links require an ICP channel path definition at the z/OS and the CF end of a channel connection to operate in peer mode. They are always defined and connected in pairs. The IC link operates in peer mode, and its existence is defined in HCD/IOCP.

IC links have the following attributes:

- ▶ Operate in peer mode (channel type ICP) on z Systems servers.

- ▶ Provide the fastest connectivity, which is faster than any external link alternatives.
- ▶ Result in better coupling efficiency than with external links, effectively reducing the server cost that is associated with Parallel Sysplex technology.
- ▶ Can be used in test or production configurations, and reduce the cost of moving into Parallel Sysplex technology while also enhancing performance and reliability.
- ▶ Can be defined as spanned channels across multiple CSSs.
- ▶ Are available for no extra fee (no feature code). Employing ICFs with IC channels results in considerable cost savings when you are configuring a cluster.

IC links are enabled by defining channel type ICP. A maximum of 32 IC channels can be defined on a z Systems server.

4.9.2 Migration considerations

Migrating from previous generations of z Systems servers in an existing Parallel Sysplex to z13s servers in that same Parallel Sysplex requires proper planning for coupling connectivity. The change in the supported type of coupling link adapters and the number of available fanout slots of the z13s CPC drawer, as compared to the number of available fanout slots of the previous generation z Systems, z114, and zBC12, need to be planned for.

z13s servers do not support ISC-3 links and HCA2-O (LR).

HCA3-O link compatibility: HCA3-O (LR) links *can* connect to HCA2-O (LR) on z196, z114, zEC12, and zBC12.

The new ICA SR fanout provides connectivity only to another z13 or z13s server. For more information, see Table 4-12 on page 175.

The z13s server fanout slots in the CPC drawer provide coupling links connectivity through the ICA SR and IFB fanout cards. In addition to coupling links for Parallel Sysplex, the fanout cards that the fanout slots provide connectivity for the I/O drawer (HCA2-C fanout) and PCIe I/O drawer (PCIe fanout).

Up to eight PCIe and four IFB fanout cards can be installed in each CPC drawer, as shown in Figure 4-9.

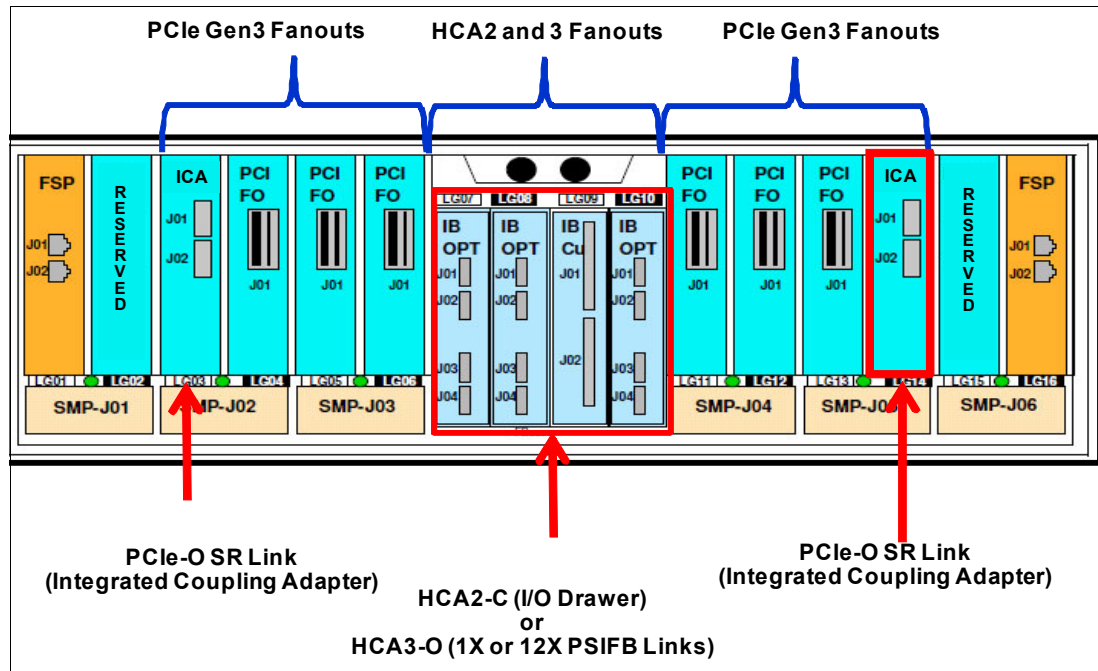


Figure 4-9 CPC drawer front view: Coupling links

Previous generation z Systems machines, the z114 and zBC12, use a different design of CPC drawer that provides connectivity for up to four InfiniBand fanouts per drawer.

As an example of a possible migration case, assume a one-drawer zBC12 used as stand-alone coupling facility with all four fanouts that are used for IFB connectivity. In a 1:1 link migration scenario, this server cannot be migrated to a z13s-N10 because the z13s-N10 cannot accommodate more than two InfiniBand fanouts. Furthermore, if FICON Express8 features are carried forward, the total number of InfiniBand fanout are reduced by two. For more information, see 4.6.6, “Fanout considerations” on page 152.

In this case, a z13s-N20 is needed to fulfill all IFB connectivity, as shown in Figure 4-10.

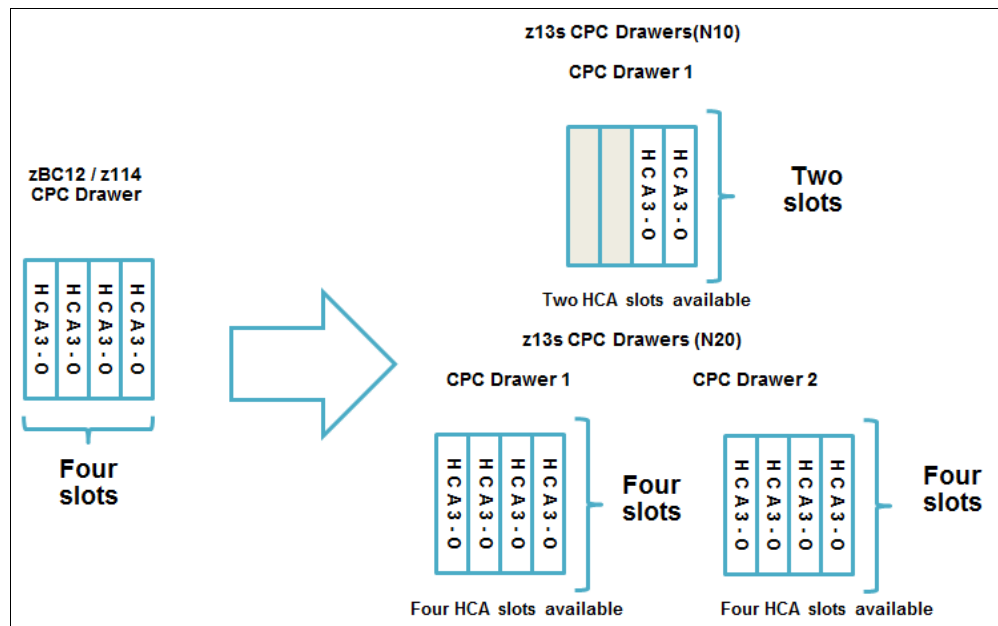


Figure 4-10 HCA3-O Fanouts: z13s versus z114 / zBC12 servers

It is beyond the scope of this book to describe all possible migration scenarios. Always involve subject matter experts (SMEs) to help you to develop your migration strategy.

The following list of considerations can help you assess possible migration scenarios. The objective of this list is to enable migration to z13s servers, supporting legacy coupling where essential, and adopting ICA SR where possible, to avoid the need for more CPC drawers and other possible migration issues.

- ▶ The IBM zEnterprise EC12 and BC12 are the last generation of z Systems to support ISC-3, 12x HCA2-O, and 1x HCA2-O LR. They also are the last z Systems that can be part of a Mixed CTN.
- ▶ Consider Long Distance Coupling requirements first:
 - HCA3-O 1x is the only long-distance coupling link that is available on z13s servers. Keep slots free for HCA3-O 1x (LR) where possible.
 - ICA SR or HCA3-O 12x should be used for short distance coupling requirements.
- ▶ ISC-3 Migration (z13s servers do not support ISC-3):
 - Evaluate current ISC-3 usage (long-, short-distance, coupling data, or timing only) to determine how to fulfill ISC-3 requirements with the links that are available on z13s servers.
 - Your options are to migrate from ISC-3 to ICA SR, InfiniBand 12x, or 1x on z13s servers.
 - 1:1 Mapping of ISC-3 to Coupling over InfiniBand. On previous servers, the HCA2-C fanouts enable ISC-3 coupling in the I/O Drawer. Two HCA2-C fanouts can be replaced by two 1x fanouts (eight 1x links) or two 12x fanouts (four 12x links).
 - ISC-3 supports one CHPID/link. Consolidate ISC-3 CHPIDs into ICA SR or IFB, and use multiple CHPIDs per link.

- ▶ By eliminating FICON Express8 (and thus the need to carry forward the I/O drawers), you can preserve InfiniBand fanout slots for coupling over InfiniBand fanouts. Replace FICON Express8 with FICON Express8S or (depending on your SAN fabric infrastructure) with FICON Express16S.
- ▶ Evaluate configurations for opportunities to eliminate or consolidate InfiniBand links:
 - Eliminate any redundant links. Two physical links between CPCs is the minimum requirement from a reliability, availability, and serviceability (RAS) perspective.
 - Share logical CHPIDs on a physical IFB link connection to reduce the usage of IFB links in z13s servers (even multiple sysplexes can share a single link).
 - Coupling Link Analysis: Capacity Planning tools and services can help.
- ▶ For z13s to z13s, or z13 links, adopt the new ICA SR coupling link. Use ICA SR channel as much as possible to replace existing InfiniBand links for z13s to z13, and z13s connectivity.
- ▶ Install all the ICA SR links that are required to fulfill future short distance coupling requirements:
 - When upgrading a CPC to a z13s system, configure the z13s servers with all the ICA SR coupling links that eventually will be needed (that is, avoid loose piece MES with ICA SR links) in your Parallel Sysplex configuration.
 - Consider a plan-ahead configuration. In a multi-CPC configuration, a final approach can be to establish z13s to z13s, and z13 connectivity by using mainly ICA SR connectivity. Even if the links are not used right away, especially within the first z13s migration, most coupling connectivity uses InfiniBand. However, after the second z13s server is migrated, coupling connectivity can be moved to ICA SR links.
- ▶ Upgrade CPCs with fewer coupling constraints first:
 - Consider upgrading CPCs to have sufficient IFB links in the target z13s configuration first (for example, multi-CPC drawer CPCs).
 - Test the new ICA SR link on the least constrained CPCs (for example, z/OS Host CPCs that have only the lowest number of links). For the CPCs that are involved in this test that do not have a CF, this test might require adding a CF LPAR and ICF engine to one of the CPCs.
 - When migrating CPCs with more coupling links to z13s servers, use enough ICA SR links in place of Coupling over InfiniBand (for example, half ICA SR, half 12x) to maintain the CPC footprint (that is, avoid extra CPC drawers).
- ▶ Consider replacing two servers that are close to each other at the same time. Assess the risk, and, if acceptable, allow immediate replacement of 12x IFB with ICA SR links in z13s peers, thus freeing InfiniBand fanout slots for 1x InfiniBand links.
- ▶ Consider (even temporarily) increased CHPID/link consolidation on 12x InfiniBand links. For less critical sysplexes (that is, test, development, and QA) that use the IFB3 protocol (see 4.6.4, “HCA3-O (12x IFB) fanout (FC 0171)” on page 150), consider temporary performance reduction when consolidating more than four CHPIDs per 12x link. After the peer system has been migrated to z13s servers, the InfiniBand links can be migrated to ICA SR links.

Coupling link considerations

For more information about changing to InfiniBand or ICA SR coupling links, see *Implementing and Managing InfiniBand Coupling Links on IBM System z*, SG24-7539, and the *Coupling Facility Configuration Options* white paper, which can be found at the following website:

<http://www.ibm.com/systems/z/advantages/ps0/whitepaper.html>

Coupling links and Server Time Protocol

All external coupling links can be used to pass time synchronization signals by using STP. STP is a message-based protocol in which timing messages are passed over data links between servers. The same coupling links can be used to exchange time and coupling facility messages in a Parallel Sysplex.

Using the coupling links to exchange STP messages has the following advantages:

- ▶ By using the same links to exchange STP messages and coupling facility messages in a Parallel Sysplex, STP can scale with distance. Servers exchanging messages over short distances, such as IFB or ICA SR links, can meet more stringent synchronization requirements than servers that exchange messages over IFB LR links (distances up to 100 km (62 miles)). This advantage is an enhancement over the IBM Sysplex Timer implementation, which does not scale with distance.
- ▶ Coupling links also provide the connectivity that is necessary in a Parallel Sysplex. Therefore, you can minimize the number of cross-site links that is required in a multi-site Parallel Sysplex.

Between any two servers that are intended to exchange STP messages, configure each server so that at least two coupling links exist for communication between the servers. This configuration prevents the loss of one link from causing the loss of STP communication between the servers. If a server does not have a CF LPAR, timing-only links can be used to provide STP connectivity.

The z13s server does not support attachment to the IBM Sysplex Timer. A z13s server cannot be added into a Mixed CTN and can participate only in an STP-only CTN.

STP recovery enhancement

All coupling host channel adapters (ICA SR, HCA3-O (12xIFB), and HCA3-O LR (1xIFB)) are designed to send a reliable and unambiguous “going away signal.” This signal indicates that the server on which the ICA SR or HCA3 is running is about to enter a failed (check stopped) state. The “going away signal” that is sent by the Current Time Server (CTS) in an STP-only CTN is received by the Backup Time Server (BTS). The BTS can then safely take over as the CTS. The BTS does not have to rely on the previous Offline Signal (OLS) in a two-server CTN, or the Arbiter in a CTN with three or more servers.

This enhancement is exclusive to z Systems and zEnterprise CPCs. It is available only if you have an ICA SR, HCA3-O (12x IFB), or HCA3-O LR (1x IFB) on the CTS communicating with an ICA SR, HCA3-O (12x IFB), or HCA3-O LR (1x IFB) on the BTS. However, the previous STP recovery design is still available for the cases when a “going away signal” is not received or for failures other than a server failure.

A new Enhanced Console Assisted Recovery (ECAR) is introduced in z13s servers that can help to speed up the progress of BTS take over when the CTS encounters a check stopped condition. For more information about the ECAR design, see “Enhanced Console Assisted Recovery” on page 419.

Important: For more information about configuring an STP CTN with three or more servers, see the white paper that is found at the following website:

<http://www.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP101833>

If the guidelines are not followed, it might result in all the servers in the CTN becoming unsynchronized. This condition results in a sysplex-wide outage.

For more information about STP configuration, see these books:

- ▶ *Server Time Protocol Planning Guide*, SG24-7280
- ▶ *Server Time Protocol Implementation Guide*, SG24-7281
- ▶ *Server Time Protocol Recovery Guide*, SG24-7380

4.9.3 Pulse per second input

A pulse per second (PPS) signal can be received from an External Time Source (ETS) device. One PPS port is available on each of the two oscillator cards, which are installed on a small backplane mounted in the rear of the z13s frame, with connection to all the CEC drawers. These cards provide redundancy for continued operation and concurrent maintenance when a single oscillator card fails. Each oscillator card has a Bayonet Neill-Concelman (BNC) connector for PPS connection support, attaching to two different ETSs. Two PPS connections from two different ETSs are preferable for redundancy.

The time accuracy of an STP-only CTN is improved by adding an ETS device with the PPS output signal. STP tracks the highly stable and accurate PPS signal from ETSs. It maintains accuracy of 10 μ s as measured at the PPS input of the z13s server. If STP uses an NTP server without PPS, a time accuracy of 100 ms to the ETS is maintained. ETSs with PPS output are available from various vendors that offer network timing solutions.

4.10 Cryptographic functions

Cryptographic functions are provided by the CP Assist for Cryptographic Function (CPACF) and the PCI Express cryptographic adapters. z13s servers support the Crypto Express5S feature.

4.10.1 CPACF functions (FC 3863)

FC 3863 is required to enable CPACF functions. FC 3863 is subject to export regulations.

4.10.2 Crypto Express5S feature (FC 0890)

Crypto Express5S is a new feature on z13 and z13s servers. On the initial configuration, a minimum of two features are installed. The number of features increases one at a time up to a maximum of 16 features. Each Crypto Express5S feature holds one PCI Express cryptographic adapter. Each adapter can be configured by the installation as a Secure IBM Common Cryptographic Architecture (CCA) coprocessor, as a Secure IBM Enterprise Public Key Cryptography Standards (PKCS) #11 (EP11) coprocessor, or as an accelerator.

Each Crypto Express5S feature occupies one I/O slot in the PCIe I/O drawer, and has no CHPID assigned. However, it has one PCHID.

4.11 Integrated firmware processor

The integrated firmware processor (IFP) was initially introduced with the zEC12 and zBC12 servers. The IFP is dedicated for managing a new generation of PCIe features. These new features are installed exclusively in the PCIe I/O drawer:

- ▶ zEDC Express

- ▶ 10GbE RoCE Express

All native PCIe features should be ordered in pairs for redundancy. According to their physical location in the PCIe I/O drawer, the features are assigned to one of the two resource groups that are running on the IFP, providing management functions and virtualization functions. If two features of same type are installed, one is always managed by resource group 1 (RG 1) and the other feature is managed by resource group 2 (RG 2). This configuration provides redundancy if one of the features or resource groups need maintenance or has a failure.

The IFP and resource groups support the following infrastructure management functions:

- ▶ Firmware update of adapters and resource groups
- ▶ Error recovery and failure data collection
- ▶ Diagnostic and maintenance tasks

For more information about the IFP and resource groups, see Appendix G, “Native Peripheral Component Interconnect Express” on page 521.

4.12 Flash Express

The Flash Express cards are supported in a PCIe I/O drawer with other PCIe I/O cards and can be ordered as new (FC 403) or carry-forward (FC 402). They are plugged into PCIe I/O drawers in pairs for availability. Like the Crypto Express5S cards, each card has a PCHID, and no HCD/IOCP definition is required. Flash Express subchannels are predefined, and are allocated from the 256 subchannels that are reserved in subchannel set 0.

Flash Express cards are internal to the CPC, and are accessible by using the new z Systems extended asynchronous data mover (EADM) facility. EADM is an extension of the asynchronous data mover (ADM) architecture that was used in the past with expanded storage. EADM access is initiated with a Start Subchannel instruction.

z13s servers support a maximum of four pairs of Flash Express cards. Only one Flash Express card is allowed per I/O domain. Each PCIe I/O drawer has four I/O domains, and can install two pairs of Flash Express cards. Each pair is installed either in the front of PCIe I/O drawers at slots 1 and 14, or in the rear at slots 25 and 33. The Flash Express cards are first plugged into the front of the PCIe I/O drawer before being plugged into the rear of drawer. These four slots (1, 14, 25, and 33) are reserved for Flash Express and must not be filled by other types of I/O cards until there is no spare slot.

Figure 4-11 shows a PCIe I/O drawer that is fully populated with Flash Express cards.

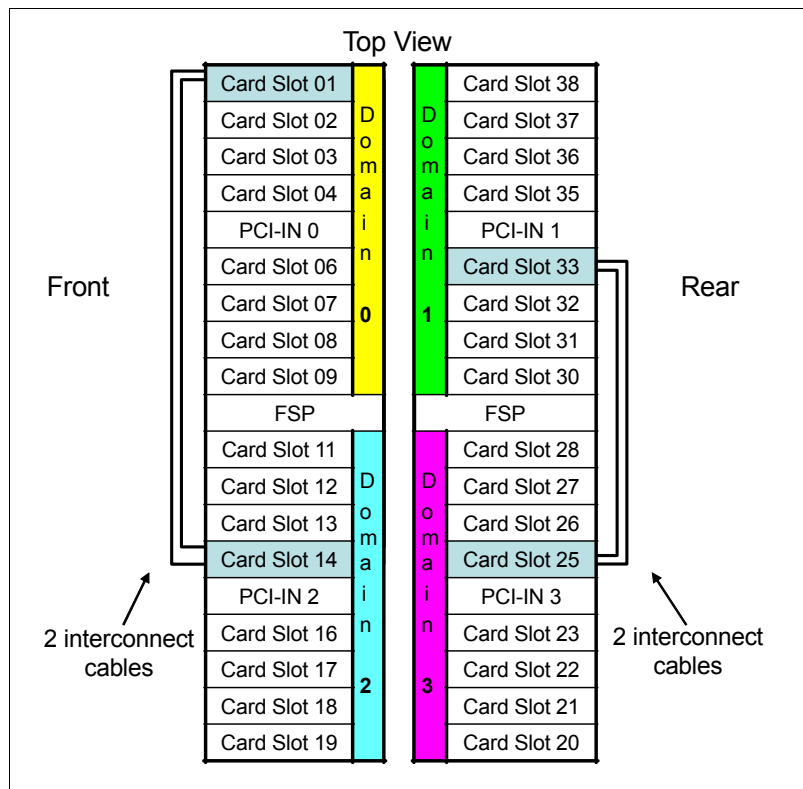


Figure 4-11 PCIe I/O drawer that is fully populated with Flash Express cards

4.12.1 IBM Flash Express read/write cache

On IBM z13s and z13 servers, Driver 27 Flash Express can provide a read/write cache that greatly increases performance by allowing customer data to be stored temporarily in a cache in the system's hardware system area (HSA) RAM. For this purpose, 0.5 GB of HSA space is reserved on z13s servers and 4 GB on z13 Driver 27 servers.

Extended asynchronous data mover (EADM) operations would complete from a partition's perspective immediately after data is moved to/from HSA cache, long before IBM Flash Express operations have completed.

4.13 10GbE RoCE Express

RoCE Express (FC 0411) is exclusively in the PCIe I/O drawer and is supported on z13s, z13, zEC12, and zBC12 servers. The 10GbE RoCE Express feature has one PCIe adapter. It does not use a CHPID. It is defined by using the input/output configuration program (IOCP) **FUNCTION** statement or in the HCD. For zEC12 and zBC12, each feature can only be dedicated to an LPAR, and z/OS can support only one of the two ports. In z13s servers, both two ports are supported by z/OS and can be shared by up to 31 partitions (LPARs). The 10GbE RoCE Express feature uses an SR laser as the optical transceiver, and supports the use of a multimode fiber optic cable that terminates with an LC Duplex connector. Both point-to-point connections and switched connections with an enterprise-class 10 GbE switch are supported.

Switch configuration for RoCE: If the IBM 10GbE RoCE Express features are connected to 10 GbE switches, the switches must meet the following requirements:

- ▶ Global Pause function enabled
- ▶ Priority flow control (PFC) disabled
- ▶ No firewalls, no routing, and no IEDN

The maximum supported unrepeatable distance, point-to-point, is 300 meters (984 ft.). A client-supplied cable is required. Three types of cables can be used for connecting the port to the selected 10 GbE switch or to a 10GbE RoCE Express feature on another server:

- ▶ OM3 50 micron multimode fiber optic cable that is rated at 2000 MHz-km that terminates with an LC Duplex connector (supports 300 meters (984 ft.))
- ▶ OM2 50 micron multimode fiber optic cable that is rated at 500 MHz-km that terminates with an LC Duplex connector (supports 82 meters (269 ft.))
- ▶ OM1 62.5 micron multimode fiber optic cable that is rated at 200 MHz-km that terminates with an LC Duplex connector (supports 33 meters (108 ft.))

For more information about the management and definition of the 10GbE RoCE, see D.2, “Shared Memory Communication over RDMA” on page 476 and Appendix G, “Native Peripheral Component Interconnect Express” on page 521.

4.14 zEDC Express

zEDC Express is an optional feature (FC 0420) that is available on z13s, z13, zEC12, and zBC12 servers. It is designed to provide hardware-based acceleration for data compression and decompression.

The feature installs exclusively on the PCIe I/O drawer. One to eight features can be installed on the system. There is one PCIe adapter/compression coprocessor per feature, which implements compression as defined by RFC1951 (DEFLATE).

A zEDC Express feature can be shared by up to 15 LPARs.

For more information about the management and definition of the zEDC feature, see Appendix J, “IBM zEnterprise Data Compression Express” on page 551 and Appendix G, “Native Peripheral Component Interconnect Express” on page 521.



Central processor complex channel subsystem

This chapter addresses the concepts of the IBM z13s channel subsystem, including multiple channel subsystems and multiple subchannel sets. It also describes the technology, terminology, and implementation aspects of the channel subsystem.

This chapter includes the following sections:

- ▶ Channel subsystem
- ▶ I/O configuration management
- ▶ Channel subsystem summary

5.1 Channel subsystem

Channel subsystem (CSS) is a group of facilities that z Systems uses to control I/O operations.

The channel subsystem directs the flow of information between I/O devices and main storage. It allows data processing to occur at the same time as I/O processing, which relieves data processors (CPs, IFLs) of the task of communicating directly with I/O devices.

The channel subsystem includes subchannels, I/O devices that are attached through control units, and channel paths between the subsystem and control units.

The design of z Systems offers considerable processing power, memory size, and I/O connectivity. In support of the larger I/O capability, the CSS structure has been scaled up correspondingly by introducing the multiple logical channel subsystem (LCSS) with z990, and multiple subchannel sets (MSS) with z9.

z13s servers are designed to support up to three logical channel subsystems (LCSS), and three subchannel sets.

Figure 5-1 summarizes the multiple channel subsystems and multiple subchannel sets.

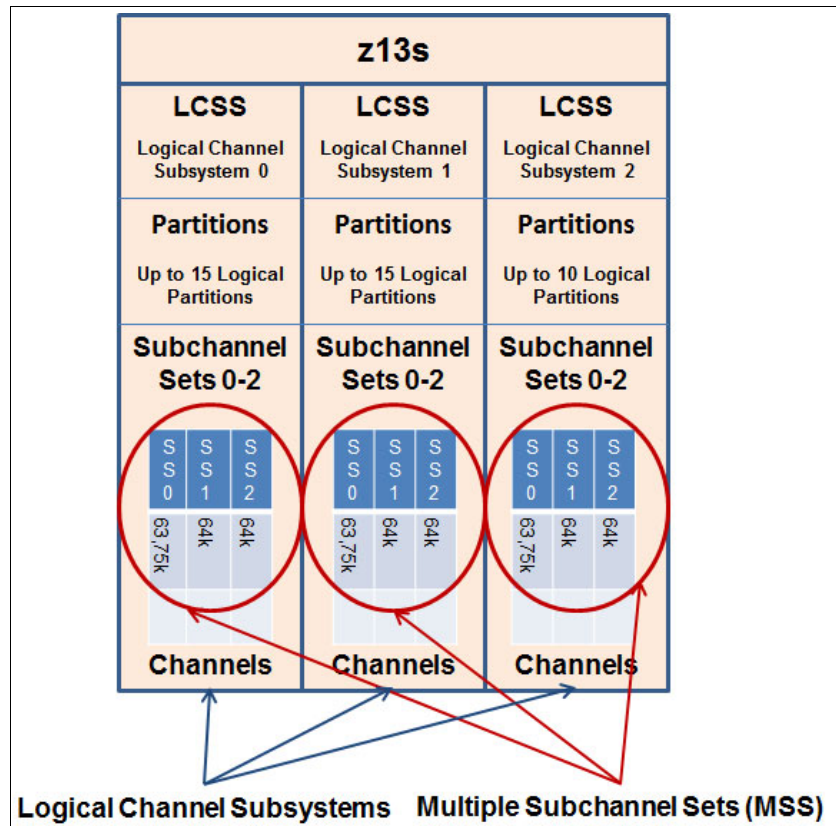


Figure 5-1 Multiple channel subsystems and multiple subchannel sets

All channel subsystems are defined within a single configuration, which is called I/O configuration data set (IOCDS). The IOCDS is loaded into the hardware system area (HSA) during a CPC is power-on reset (POR), to initial all the channel subsystems. On z13s servers, the HSA is pre-allocated in memory with a fixed size of 40 GB, as well as the customer

purchased memory. This fixed size memory for HSA eliminates the requirement of additional planning for initial I/O configuration and pre-planning for future I/O expansions.

CPC drawer repair: For z13s servers, the CPC drawer repair actions are always disruptive.

The objects are always reserved in the z13s HSA during POR, no matter if they are defined in the IOCDS for use or not:

- ▶ Three CSSs
- ▶ Fifteen LPARs in each CSS0 to CSS1
- ▶ Ten LPARs in CSS2
- ▶ Subchannel set 0 with 63.75 K devices in each CSS
- ▶ Subchannel set 1 with 64 K minus one device in each CSS
- ▶ Subchannel set 2 with 64 K minus one device in each CSS

5.1.1 Multiple logical channel subsystems

In z Systems architecture, a single channel subsystem can have up to 256 channel paths defined, which limited the total numbers of I/O connectivities on a z Systems server to within 256.

The introduction of multiple LCSSs has enabled a z Systems server to have more than one channel subsystems (logically), while maintaining the same manner of I/O processing within each LCSS. Also, a logical partition (LPAR) is now attached to a specific logical channel subsystem, which makes the extension of multiple logical channel subsystems transparent to the operating systems and applications. The multiple image facility (MIF) in the structure enables resource sharing across logical partitions within a single LCSS or across the LCSSs.

The multiple LCSS structure has extended z Systems' total number of I/O connections to support a balanced configuration for the growth of processor and I/O capabilities.

A one-digit number ID starting from 0 (CSSID) is assigned to an LCSS, and a one-digit hexadecimal ID (MIF ID) starting from 0 is assigned to an LPAR within the LCSS.

Note: The word “channel subsystem” has same meaning as “logical channel subsystem” in this chapter unless otherwise indicated.

Subchannels

A *subchannel* provides the logical appearance of a device to the program and contains the information required for sustaining a single I/O operation. Each device is accessible by using one subchannel in a channel subsystem to which it is assigned according to the active IOCDS of the z Systems server. A subchannel set (SS) is a collection of subchannels within a channel subsystem, and the maximum number of subchannels in a subchannel set determines how many devices can be accessible to a channel subsystem.

In z Systems architecture, the first subchannel set of an LCSS can have 63.75K subchannels (with 0.25K reserved), with a subchannel set ID (SSID) as 0. By enabling the multiple subchannel sets, which are described in 5.1.2, “Multiple subchannel sets” on page 190, extra subchannel sets are available to increase the device addressability of a channel subsystem.

Channel paths

A *channel path* provides a connection between the channel subsystem and control unites, that the channel subsystem uses to communicate with I/O devices. Depends on the type of

connections, a channel path can be a physical connection to a control unit with I/O devices, like FICON, or an internal logical one, like HiperSockets.

Each channel path in a channel subsystem has a unique 2-digit heximal identifier that is known as a channel-path identifier (CHPID), range from 00 to FF. So a total of 256 CHPIDs are supported by a CSS, and a maximum of 768 CHPIDs are available on z13s servers, with three logical channel subsystems.

By assigning a CHPID to a physical port of an I/O feature card, such as a FICON Express16S, or a fanout adapter port, such as an ICA-SR port, the channel subsystem connect to I/O devices through these physical ports.

A port on an I/O feature card has a unique physical channel ID (PCHID) based on the physical location of this I/O feature card, and the sequence of this port on the card.

Also, a port on a fanout adapter has a unique adapter identifier (AID), according to the physical location of this fanout card, and the sequence of this port on the card.

A CHPID is assigned to a physical port by defining the corresponding PCHID, or AID in the I/O configuration definitions.

A CHPID can be assigned to a physical port of an I/O feature card in an InfiniBand (IFB) I/O drawer or a PCIe I/O drawer, in a multidrawer CPC.

Control units

A *control unit* provides the logical capabilities that are necessary to operate and control an I/O device. It adapts the characteristics of each device so that it can respond to the standard form of control that is provided by the CSS.

A control unit can be housed separately, or can be physically and logically integrated with the I/O device, the channel subsystem, or within the z Systems server itself.

I/O devices

An *I/O device* provides external storage, a means of communication between data-processing systems, or a means of communication between a system and its environment. In the simplest case, an I/O device is attached to one control unit and is accessible through one or more channel paths that connect with the control unit.

5.1.2 Multiple subchannel sets

A subchannel set is a collection of subchannels within a channel subsystem. The maximum number of subchannels in a subchannel set determines how many I/O devices that a channel subsystem can access. This amount determines the number of addressable devices to the program, an operating system for example, that runs in that LPAR.

Each subchannel has a unique 4-digit heximal number, range from 0x0000 to 0xFFFF. Thus, a single subchannel set can address and access up to 64K I/O devices.

MSS was introduced with z9 to extend the maximum number of addressable I/O devices for a channel subsystem.

z13s servers now support three subchannel sets for each logical channel subsystem, and is capable of accessing a maximum of 191.74K devices for a logical channel subsystem, and therefore the logical partition and the program running on it.

Note: Do not confuse the multiple subchannel sets function with multiple channel subsystems.

Subchannel number

The subchannel number is a four-digit heximal number, ranging from 0x0000 to 0xFFFF, assigned to a subchannel within a subchannel set of a channel subsystem. Subchannels in each subchannel set are always assigned subchannel numbers within a single range of contiguous numbers. The lowest-numbered subchannel is subchannel 0, and the highest-numbered subchannel has a subchannel number equal to one less than the maximum numbers of subchannels supported by the subchannel set. Thus, a subchannel number is always unique within a subchannel set of a channel subsystem, and depends on the sequence of assignment.

With the subchannel numbers, a program running on an LPAR, such as an operating system, can specify all I/O functions relative to a specific I/O device, by designating a subchannel assigned to the I/O devices.

Normally, subchannel numbers are only used in communication between the programs and the channel subsystem.

Subchannel set identifier

While introducing the MSS, the channel subsystem is extended to assign a value range from 0 to 2 for each subchannel set. This value is called the SSID. A subchannel is identified by its SSID and subchannel number.

Device number

A device number is an arbitrary number, ranging from 0x0000 to 0xFFFF, that is defined by a system programmer in an I/O configuration for naming an I/O device. The device number must be unique within a subchannel set of a channel subsystem. It is assigned to the corresponding subchannel by channel subsystem while an I/O configuration is activated. Thus a subchannel in a subchannel set of a channel subsystem has a device number together with a subchannel number for designating an I/O operation.

The device number provide a means to identify a device, independent of any limitations imposed by the system model, the configuration, or channel-path protocols.

A device number also can be used to designate an I/O function to a specific I/O device. Because it is an arbitrary number, it is easy to fit it in any configuration management and operating management scenarios. For example, a system administrator can have all the z/OS systems in an environment use a device number 1000 for their system RES volumes.

With the multiple subchannel sets, for a specific I/O device, a subchannel is assigned by the channel subsystem with an automatically assigned subchannel number, and a device number defined by user. That an I/O device can always be identified by an SSID in conjunction with a subchannel number, or a device number. For example, a device with device number AB00 of subchannel set 1, can be designated as 1AB00.

Normally, the subchannel number is used by the programs to communicate with the channel subsystem and I/O device, while the device number is used by system programmers, operators, and administrators.

Device in subchannel set 0 and additional subchannel sets

An LCSS will always have the first subchannel set (SSID 0), which can have up to 63.75K subchannels, with 256 subchannels reserved by the channel subsystem. That user can always define their I/O devices in this subchannel set for general use.

For the additional subchannel sets enabled by the MSS facility, each has 65,535 subchannels (64K minus one) for specific types of devices. These additional subchannel sets are referred to as *alternate subchannel sets* in z/OS. Also, a device that is defined in an alternate subchannel set is considered a *special device*, which normally will have a special device type in the I/O configuration as well.

Currently, a z13s server that runs z/OS supports defining these types of devices in an additional subchannel set, with proper APAR or program temporary fix (PTF) installed:

- ▶ Alias devices of the parallel access volumes (PAV)
- ▶ Secondary devices of GDPS Metro Mirror Copy Service (formerly PPRC)
- ▶ FlashCopy SOURCE and TARGET devices with PTF OA46900
- ▶ DB2 data backup volumes with PTFOA24142

Using an extra subchannel set for these special devices helps to relieve the numbers of devices in the subchannel set 0, which improves growth capability for the system.

IPL from an alternate subchannel set

z13s servers support IPL from alternate subchannel sets in addition to subchannel set 0. Devices that are used early during IPL processing now can be accessed by using subchannel set 1, or subchannel set 2 on a z13s server. This configuration allows the users of Metro Mirror secondary devices that are defined by using the same device number and a new device type in an alternate subchannel set to be used for IPL, an I/O definition file (IODF), and stand-alone memory dump volumes when needed.

IPL from an alternate subchannel set is supported by z/OS V1.13 or later, and Version 1.12 with PTFs.

The display ios,config command

The z/OS **display ios,config(all)** command that is shown in Figure 5-2 includes information about the multiple subchannel sets.

```
D IOS,CONFIG(ALL)
IOS506I 11.32.19 I/O CONFIG DATA 340
ACTIVE IODF DATA SET = SYS6.IODF39
CONFIGURATION ID = L06RMVS1      EDT ID = 01
TOKEN:  PROCESSOR DATE      TIME      DESCRIPTION
SOURCE: SCZP303 14-10-31 08:51:47 SYS6      IODF39
ACTIVE CSS: 0      SUBCHANNEL SETS CONFIGURED: 0, 1, 2
CHANNEL MEASUREMENT BLOCK FACILITY IS ACTIVE
LOCAL SYSTEM NAME (LSYSTEM): SCZP303
HARDWARE SYSTEM AREA AVAILABLE FOR CONFIGURATION CHANGES
PHYSICAL CONTROL UNITS          8099
CSS 0 - LOGICAL CONTROL UNITS    3996
  SS 0  SUBCHANNELS              54689
  SS 1  SUBCHANNELS              58862
  SS 2  SUBCHANNELS              65535
CSS 1 - LOGICAL CONTROL UNITS    4088
  SS 0  SUBCHANNELS              65280
  SS 1  SUBCHANNELS              65535
  SS 2  SUBCHANNELS              65535
CSS 2 - LOGICAL CONTROL UNITS    4088
  SS 0  SUBCHANNELS              65280
  SS 1  SUBCHANNELS              65535
  SS 2  SUBCHANNELS              65535
```

Figure 5-2 Output for display ios,config(all) command

5.1.3 Channel path spanning

With the implementation of multiple LCSSs, a channel path can be available to LPARs as dedicated, shared, and spanned.

Although a shared channel path can be shared by LPARs within a same LCSS, a spanned channel path can be shared by LPARs within and across LCSSs.

By assigning the same CHPID from different LCSSs to the same channel path, a PCHID for example, the channel path can be accessed by any LPAR from these LCSSs at the same time. That CHPID is then spanned across those LCSSs. Using spanned channels paths decreases the number of channels needed in an installation of z Systems servers.

A sample of channel paths that are defined as dedicated, shared, and spanned is shown in Figure 5-3.

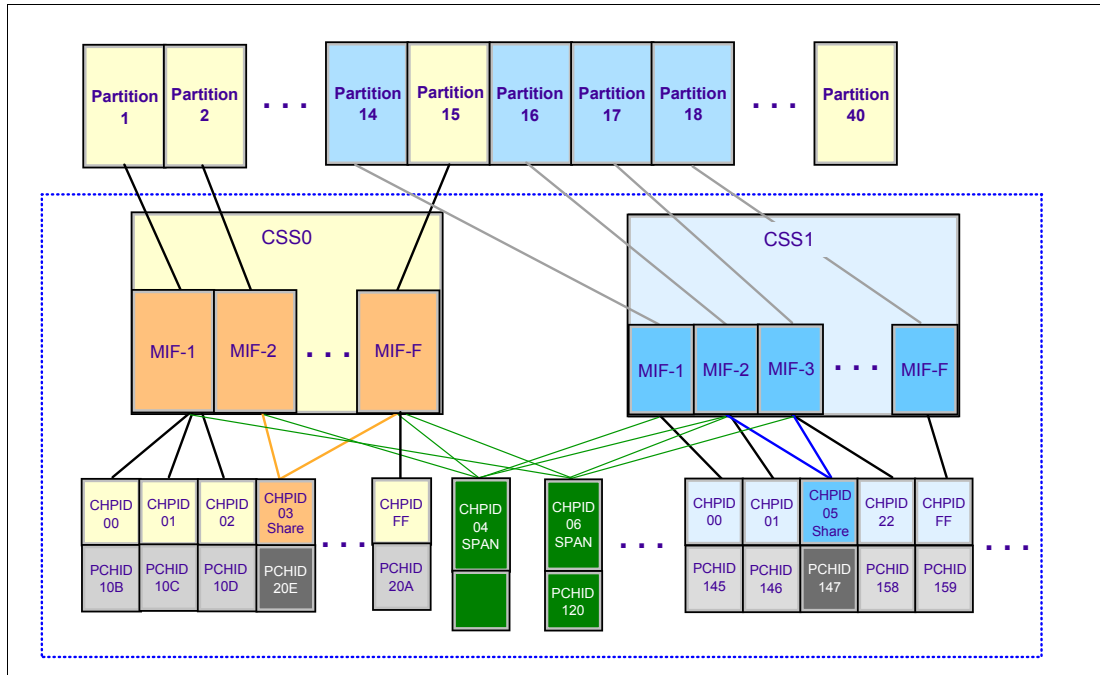


Figure 5-3 z Systems CSS: Channel subsystems with channel spanning

In the sample, three different definitions of a channel path are shown:

- ▶ CHPID FF, assigned to PCHID 20A, is dedicated access for partition 15 of LCSS0, same apply to CHPID 00,01,02 of LCSS0, and CHPID 00,01, FF of LCSS1.
- ▶ CHPID 03, assigned to PCHID 20E is shared access for partition 2, and 15 of LCSS0, same apply to CHPID 05 of LCSS1.
- ▶ CHPID 06, assigned to PCHID 120 is spanned access for partition 1, 15 of LCSS0, and partition 16, 17 of LCSS1, same apply to CHPID 04.

Channel spanning is supported for internal links (HiperSockets and IC links) and for certain types of external links. External links that are supported on z13s servers include FICON Express16S, FICON Express8S, FICON Express8 channels, OSA-Express5S, OSA-Express4S, and Coupling Links.

The definition of LPAR name, MIF image ID, and LPAR ID are used by the channel subsystem to identify an LPAR to identify I/O functions from different LPARs of multiple LCSSs. Doing so supports the implementation of these dedicated, shared, and spanned accessing.

An example that summarizes the definition of these LPAR-related identifications is shown in Figure 5-4.

CSS0	CSS1	CSS2	Specified in HCD / IOCP
Logical Partition Name TST1 PROD1 PROD2	Logical Partition Name TST2 PROD3 PROD4	LPAR Name TST3 TST4	Specified in HCD / IOCP
Logical Partition ID 02 04 0A	Logical Partition ID 14 16 1D	LPAR ID 22 26	Specified in Image Profile
MIF ID 2 4 A	MIF ID 4 6 D	MIF ID 2 6	Specified in HCD / IOCP

Figure 5-4 CSS, LPAR, and identifier example

LPAR name

The LPAR name is defined as partition name parameter in the **RESOURCE** statement of an I/O configuration. The LPAR name must be unique across the server.

MIF image ID

The MIF image ID is defined as a parameter for each LPAR in the **RESOURCE** statement of an I/O configuration. It ranges from 1 to F, and must be unique within an LCSS, and can be duplicate in different LCSSs. If a MIF image ID is not defined, an arbitrarily ID is assigned when the I/O configuration activated. The z13s server supports a maximum of three LCSSs, with a total number of 40 LPARs that can be defined. Each LCSS of a z13s server has these maximum number of LPARs:

- ▶ LCSS0 and LCSS1 support 15 LPARs each, and MIF image ID range from 1 to F within each LCSS.
- ▶ LCSS2 supports 10 LPARs, and MIF image ID range from 1 to A.

LPAR ID

The LPAR ID is defined by user in an image activation profile for each LPAR. It is a 2-digit heximal number from 00 to 7F. The LPAR ID must be unique across the server. Though it is arbitrary defined by user, normally an LPAR ID is the CSS ID concatenated to its MIF image ID, which makes the value more meaningful and easy for system administration. For example, an LPAR with LPAR ID 1A means that the LPAR is defined in LCSS1, with the MIF image ID A.

5.2 I/O configuration management

The following tools are available to help maintain and optimize the I/O configuration:

- ▶ IBM Configurator for e-business (eConfig): The eConfig tool is available from your IBM representative. It is used to create configurations or upgrades of an existing configuration, and it maintains tracking to the installed features of those configurations. eConfig produces helpful reports for understanding the changes that are being made for a new system or a system upgrade, and what the target configuration will look like.
- ▶ Hardware configuration definition (HCD): HCD supplies an interactive dialog to generate the IODF, and later the IOCDs. Generally, use HCD or Hardware Configuration Manager (HCM) to generate the I/O configuration rather than writing IOCP statements. The validation checking that HCD runs against a IODF source file helps minimize the risk of errors before an I/O configuration is activated. HCD support for multiple channel subsystems is available with z/VM and z/OS. HCD provides the capability to make both dynamic hardware and software I/O configuration changes.

Note: Certain functions might require specific levels of an operating system, PTFs, or both.

- ▶ Consult the appropriate 2965DEVICE, 2964DEVICE, 2828DEVICE, 2827DEVICE, 2818DEVICE, and 2817DEVICE Preventive Service Planning (PSP) buckets before implementation.
- ▶ Hardware Configuration Manager (HCM): HCM is a priced optional feature that supplies a graphical interface for HCD. It is installed on a PC and allows you to manage both the physical and the logical aspects of a mainframe's hardware configuration.
- ▶ CHPID Mapping Tool (CMT): The CMT helps to map CHPIDs onto PCHIDs base on a IODF source file and the eConfig configuration file of a mainframe. It provides CHPID to PCHID mapping with high availability for the target I/O configuration. Built-in mechanisms generate a mapping according to customized I/O performance groups. Additional enhancements are implemented in CMT to support the z13 and z13s servers. The CMT is available for download from the IBM Resource Link website:

<http://www.ibm.com/servers/resourceLink>

5.3 Channel subsystem summary

z13s servers support the channel subsystem features of multiple LCSS, multiple SS, and the channel spanning described in this chapter. Table 5-1 lists an overview of z13s channel subsystem with maximum values of different specs.

Table 5-1 z13s CSS overview

Maximum number of CSSs	3
Maximum number of LPARs per CSS	CSS0 - CSS1: 15 CSS2: 10
Maximum number of LPARs per system	40
Maximum number of subchannel sets per CSS	3

Maximum number of subchannels per CSS	191.74 K SS0: 65280 SS1 - SS2: 65535
Maximum number of CHPIDs per CSS	256



Cryptography

This chapter describes the hardware cryptographic functions that are available on IBM z13s servers. The CP Assist for Cryptographic Function (CPACF) with the Peripheral Component Interconnect Express (PCIe) cryptographic coprocessors offer a balanced use of processing resources and unmatched scalability.

This chapter provides a short introduction to the principles of cryptography. It then looks at the implementation of cryptography in the hardware and software architecture of z Systems, followed by a more detailed description of the features the IBM z13s servers are offering. The chapter concludes with a summary of the cryptographic features and the software required.

This chapter includes the following sections:

- ▶ Cryptography in IBM z13 and z13s servers
- ▶ Some fundamentals on cryptography
- ▶ Cryptography on IBM z13s servers
- ▶ CP Assist for Cryptographic Functions
- ▶ Crypto Express5S
- ▶ TKE workstation
- ▶ Cryptographic functions comparison
- ▶ Cryptographic software support

6.1 Cryptography in IBM z13 and z13s servers

IBM z13 and z13s servers introduce the new PCI Crypto Express5S feature, together with a redesigned CPACF Coprocessor, managed by a new Trusted Key Entry (TKE) workstation. Also, the IBM Common Cryptographic Architecture (CCA) and the IBM Enterprise PKCS #11 (EP11) Licensed Internal Code (LIC) have been enhanced. The new functions support new standards and are designed to meet the following compliance requirements:

- ▶ National Institute of Standards and Technology (NIST) through the Federal Information Processing Standard (FIPS) standard to implement guidance requirements
- ▶ Emerging banking standards to strength the cryptographic standards for attack resistance
- ▶ VISA Format Preserving Encryption (VFPE) for credit card numbers
- ▶ Enhanced public key elliptic curve cryptography (ECC) for users such as Chrome, Firefox, and Apple's iMessage

IBM z13s servers include both standard cryptographic hardware and optional cryptographic features for flexibility and growth capability. IBM has a long history of providing hardware cryptographic solutions. This history stretches from the development of the Data Encryption Standard (DES) in the 1970s to the Crypto Express tamper-sensing and tamper-responding programmable features. Crypto Express is designed to meet the US Government's highest security rating, which is FIPS 140-2 Level 4¹. It is also designed to meet several other security ratings like the Common Criteria for Information Technology Security Evaluation, the Payment Card Industry (PCI) hardware security module (HSM) criteria, and the criteria for Deutsche Kreditwirtschaft (DK) evaluation.

The cryptographic functions include the full range of cryptographic operations that are necessary for e-business, e-commerce, and financial institution applications. User Defined Extensions (UDX) allow you to add custom cryptographic functions to the functions that z13s servers offer.

6.2 Some fundamentals on cryptography

From the beginning of human history, there always has been the demand to keep certain messages secret, so that a third person is not able to understand what the sender is telling the receiver. Also, you need to ensure that a message cannot be corrupted, and that the sender and the receiver really are the persons that they seem to be. Over the centuries, several methods have been used to achieve these objectives, with more or less success. Many procedures and algorithms for encrypting and decrypting data have been developed, most of which are complicated and time consuming to handle.

6.2.1 Modern cryptography

With the development of computing technology, encryption and decryption algorithms can be performed by computers, which enables the use of complicated mathematical algorithms, most of them based on the prime factorization of very large numbers. In modern cryptography, these are the main purposes for protecting information:

- ▶ Protection: The protection of data usually is the main concept associated with cryptography. Only authorized persons should be able to read the message or to get information about it. Data is encrypted by using a known algorithm and secret keys such

¹ Federal Information Processing Standards (FIPS) 140-2 Security Requirements for Cryptographic Modules

that the intended party can de-scramble the data but an interloper cannot. This idea is also referred to as confidentiality.

- ▶ **Authentication:** This is the process of determining whether the partners in communication are who they claim to be, which can be done by using certificates and signatures. It must be possible to clearly identify the owner of the data or the sender and the receiver of the message.
- ▶ **Integrity:** The verification of data ensures that what has been received is identical to what was sent. It must be verified that the data is complete and has not been modified.
- ▶ **Non-repudiation:** It must be impossible for the owner of the data or the sender of the message to deny authorship. Non-repudiation ensures that both sides of a communication know that the other side has agreed to what was exchanged. This agreement implies a legal liability and contractual obligation. This is the same as a signature on a contract.

In addition, these goals must be accomplished without unacceptable delay in the communication itself. The goal is to keep the system secure, manageable, and productive.

The basic method for granting the protection of data is encrypting and decrypting it. For authentication, integrity, and non-repudiation, hash algorithms, message authentication codes (MACs), digital signatures, and certificates are used.

When encrypting a message, the sender transforms the clear text into a secret text. Doing this requires two main elements:

- ▶ The algorithm is the mathematical or logical formula that is applied to the key and the clear text to deliver a ciphered result, or to take a ciphered text and deliver the original clear text.
- ▶ The key ensures that the result of the encrypting data transformation by the algorithm is only the same when the same key is used. The decryption of a ciphered message only results in the original clear message when the correct key is used.

Therefore, the receiver of a ciphered message must know which algorithm and which key must be used to decrypt the message.

6.2.2 Kerckhoffs' principle

In modern cryptography, the algorithm is published and known to everyone, while the keys are kept secret. This system adheres to Kerckhoffs' principle, which is named after Auguste Kerckhoffs, a Dutch cryptographer who formulated it in 1883: "A system should not depend on secrecy, and it should be able to fall into the enemy's hands without disadvantage." In other words, the security of a cryptographic system should depend on the security of the key, and the key must be kept secret. Therefore, the secure management of keys is a primary task of modern cryptographic systems.

The reasons for obeying to Kerckhoffs' Principle are obvious:

- ▶ It is much more difficult to keep an algorithm secret than a key.
- ▶ It is harder to exchange a compromised algorithm than to exchange a compromised key.
- ▶ Secret algorithms can be reconstructed by reverse engineering of software or hardware implementations.
- ▶ Errors in public algorithms can generally be found more easily because many experts can study it.
- ▶ In history, most secret encryption methods have proved to be weak and inadequate.

- ▶ When using a secret encryption method, it is possible that a back door has been designed into it.
- ▶ If an algorithm is public, many experts can form an opinion about it, and the method can be more thoroughly investigated for potential weaknesses and vulnerabilities.

6.2.3 Keys

The keys that are used for the cryptographic algorithms usually are sequences of numbers and characters, but can also be any other sequence of bits. The length of a key influences the security of the cryptographic method. The longer the used key, the harder it is to compromise a cryptographic algorithm. The prominent symmetric cryptographic algorithm DES uses keys with a length of 56 bits. Triple-DES (TDES) uses keys with a length of 112 bits, and the Advanced Encryption Standard (AES) uses keys with a length of 128, 192 or 256 bits. The asymmetric RSA algorithm (named after its inventors Rivest, Shamir, and Adleman) uses keys with a length of 1024 - 4096 bits.

As already mentioned, in modern cryptography keys must be kept secret. Depending on the effort that is made to protect the key, keys are classified in three levels:

- ▶ A *clear key* is a key that is transferred from the application in clear to the cryptographic function. The key value is stored in the clear, at least briefly, somewhere in unprotected memory areas, so under certain circumstances the key can be exposed by someone accessing this memory area. This risk must be considered when using clear keys. However, there are many applications where this risk can be accepted. For example, the transaction security for the widely used encryption methods Secure Sockets Layer (SSL) and Transport Layer Security (TLS) is based on clear keys.
- ▶ The value of a *protected key* is only stored in clear in memory areas that cannot be read by applications or users. The key value does not exist outside of the physical hardware, although the hardware might not be tamper-resistant. The principle of protected keys is unique to z Systems and is explained in more detail in 6.4.2, “CPACF protected key” on page 209.
- ▶ For a *secure key*, the key value does not exist in clear outside of a special hardware device, called an HSM, which must be secure and tamper-resistant. A secure key is protected from disclosure and misuse, and can be used for the trusted execution of cryptographic algorithms on highly sensitive data. If used and stored outside of the HSM, a secure key must be encrypted with a *master key*, which is created within the HSM and never leaves the HSM.

Because a secure key must be handled in a special hardware device, the use of secret keys is usually far slower than using clear keys as illustrated in Figure 6-1.

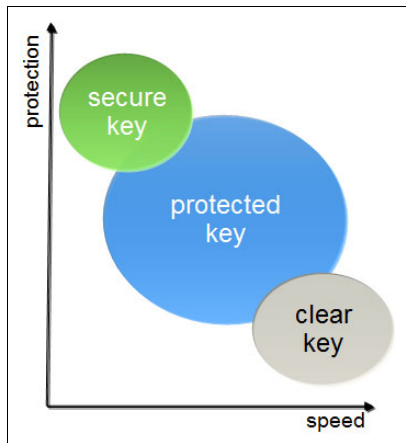


Figure 6-1 Three levels of protection with three levels of speed.

6.2.4 Algorithms

The algorithms of modern cryptography are differentiated by whether they use the same key for the encryption of the message as for the decryption:

- ▶ *Symmetric algorithms* use the same key to encrypt and to decrypt data. The function that is used to decrypt the data is the opposite of the function used to encrypt the data. Because the same key is used on both sides of an operation, it must be negotiated between both parties and kept secret. Symmetric algorithms are also known as *secret key algorithms*.

The main advantage of symmetric algorithms is that they are fast and so can be used for huge amounts of data, even if they are not run on specialized hardware. The disadvantage is that the key must be known by both sender and receiver of the messages. The key therefore must be exchanged between them, and this key exchange is a weak point that can be attacked.

Prominent examples for symmetric algorithms are the already mentioned DES, TDES, and AES.

- ▶ *Asymmetric algorithms* use two distinct but related keys, the *public key* and the *private key*. As the names imply, the private key must be kept secret, while the public key is shown to everyone. However, with asymmetric cryptography, it is not important who sees or knows the public key. Whatever is done with one key can only be undone by the other key. For instance, data encrypted using the public key can only be decrypted by the associated private key, and vice versa. Unlike symmetric algorithms, which use distinct functions for encryption and decryption, only one function is used in asymmetric algorithms. Depending on the values passed to this function, it either encrypts or decrypts the data. Asymmetric algorithms are also known as *public key algorithms*.

Asymmetric algorithms use complex calculations and are rather slow (about 100 - 1000 times slower than symmetric algorithms). As a result, they are not used for the encryption of bulk data. But because the private key is never exchanged, they are less vulnerable than symmetric algorithms. Asymmetric algorithms mainly are used for authentication, digital signatures, and for the encryption and exchange of secret keys (which then are used to encrypt bulk data with a symmetric algorithm).

Examples for asymmetric algorithms are the already mentioned RSA, and elliptic curve algorithms.

- *One-way algorithms* are not strictly speaking cryptographic functions at all. They do not use keys, and can only scramble data. They cannot de-scramble it. These algorithms are used extensively within cryptographic procedures for digital signing and they tend to be developed and governed by the same principles as cryptographic algorithms. One-way algorithms are also known as *hash algorithms*.

The most prominent one-way algorithms are the Secure Hash Algorithms (SHAs).

6.3 Cryptography on IBM z13s servers

In principle, cryptographic algorithms can run on processor hardware. But these workloads are compute intensive, and the handling of secure keys also requires special hardware protection. IBM z Systems offer several cryptographic hardware features that are specialized to meet the requirements for cryptographic workloads. Figure 6-2 shows the cryptographic hardware that is supported on IBM z13s servers, and Table 6-1 on page 205 lists the corresponding feature codes and describes the purpose of these hardware features. All these features are described in more detail later in this chapter.

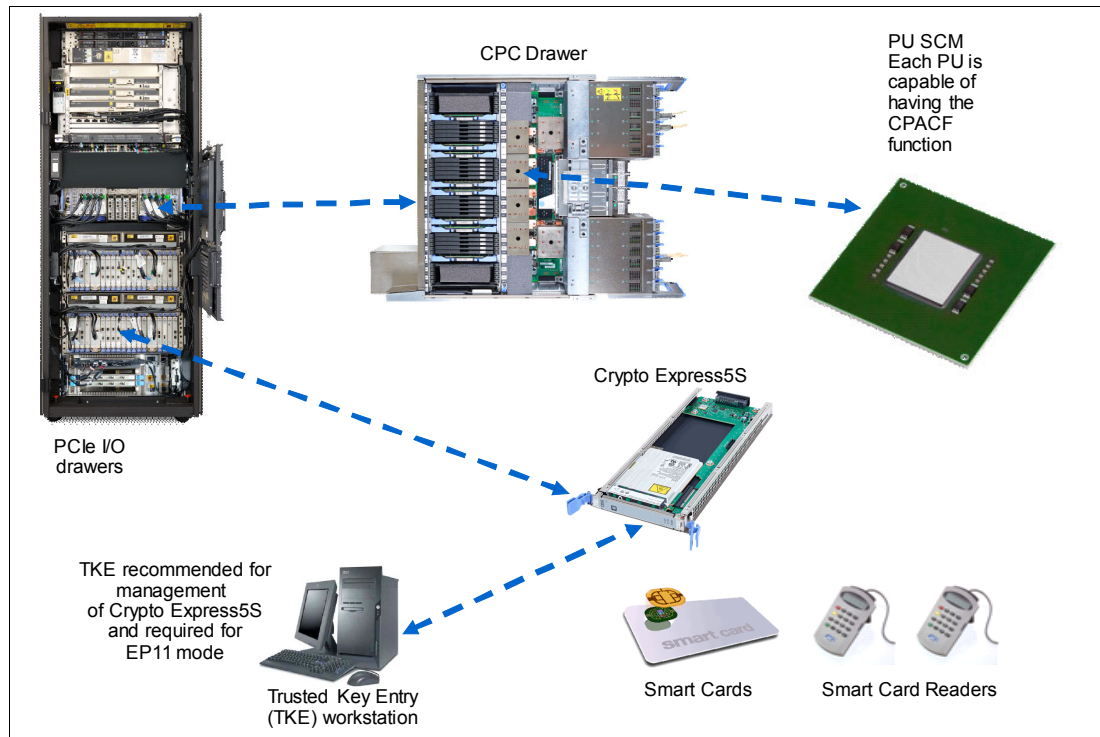


Figure 6-2 Cryptographic hardware supported on IBM z13s servers

Attached to every processor unit (PU) on a single chip module (SCM) in a central processor complex (CPC) is a cryptographic coprocessor that can be used for cryptographic algorithms using clear keys or protected keys. For more information, see 6.4, “CP Assist for Cryptographic Functions” on page 207.

The Crypto Express5S card is an HSM placed in the PCIe I/O drawer of the z13s server. It also supports cryptographic algorithms by using secret keys. This feature is described in more detail in 6.5, “Crypto Express5S” on page 211.

Finally, for entering keys in a secure way into the Crypto Express5S HSM, a TKE is required, usually also equipped with smart card readers. Section 6.6, “TKE workstation” on page 222 provides additional information.

Table 6-1 lists the feature codes and describes the purpose of these hardware features.

Table 6-1 Cryptographic features for IBM z13s servers

Feature code	Description
3863	CP Assist for Cryptographic Function (CPACF) enablement: This feature is a prerequisite to use CPACF (except for SHA-1, SHA-224, SHA-256, SHA-384, and SHA-512) and the Crypto Express5S feature.
0890	Crypto Express5S card: A maximum of 16 features can be ordered (minimum of two adapters). This is an optional feature, and each feature contains one PCI Express cryptographic adapter (adjunct processor). This feature is supported only in z13 and z13s servers.
0847	Trusted Key Entry (TKE) tower workstation: A TKE provides basic key management (key identification, exchange, separation, update, and backup) and security administration. It is optional for running a Crypto Express5S card in CCA mode and required for running it in EP11 mode. The TKE workstation has one Ethernet port, and supports connectivity to an Ethernet local area network (LAN) operating at 10, 100, or 1000 Mbps. Up to 10 features per z13s server can be ordered.
0097	Trusted Key Entry (TKE) rack mounted workstation: The rack-mounted version of the TKE, which needs a customer-provided standard 19-inch rack. It comes with a 1u TKE unit and a 1u console tray (screen, keyboard, and pointing device). When using smart card readers, an extra customer provided tray is needed. Up to 10 features per z13s server can be ordered.
0877	TKE 8.0 Licensed Internal Code (LIC): Shipped with the TKE tower workstation FC 0847 since z13 GA. This LIC is not orderable with a z13s server, but it is able to manage a Crypto Express5S card FC 0890 installed in a z13s server.
0878	TKE 8.1 Licensed Internal Code (LIC): Shipped with the TKE tower workstation FC 0847 and the TKE rack-mounted workstation FC 0097 since z13 GA2 and z13s GA.
0891	TKE Smart Card Reader: Access to information in the smart card is protected by a PIN. One feature code includes two smart card readers, two cables to connect them to the TKE workstation, and 20 smart cards. Smart card part 74Y0551 is required to support CEX5P.
0892	TKE additional smart cards: When one feature code is ordered, 10 smart cards are shipped. The order increment is 1 - 99 (990 blank smart cards). Smart cards 74Y0551 and 54D3338 can be used. A new card 00JA710 will be released because of the end of life of 74Y0551.

A TKE includes support for the AES encryption algorithm with 256-bit master keys and key management functions to load or generate master keys to the cryptographic coprocessor.

If the TKE workstation is chosen to operate the Crypto Express5S features in a z13s server, TKE workstation with the TKE 8.0 LIC or the TKE 8.1 LIC is required. For more information, see 6.6, “TKE workstation” on page 222.

Important: Products that include any of the cryptographic feature codes contain cryptographic functions that are subject to special export licensing requirements by the United States Department of Commerce. It is your responsibility to understand and adhere to these regulations when you are moving, selling, or transferring these products.

To access and use the cryptographic hardware devices that are provided by z13s servers, the application must use an application programming interface (API) provided by the operating system. In z/OS, the Integrated Cryptographic Service Facility (ICSF) provides the APIs and manages access to the cryptographic devices, as shown in Figure 6-3.

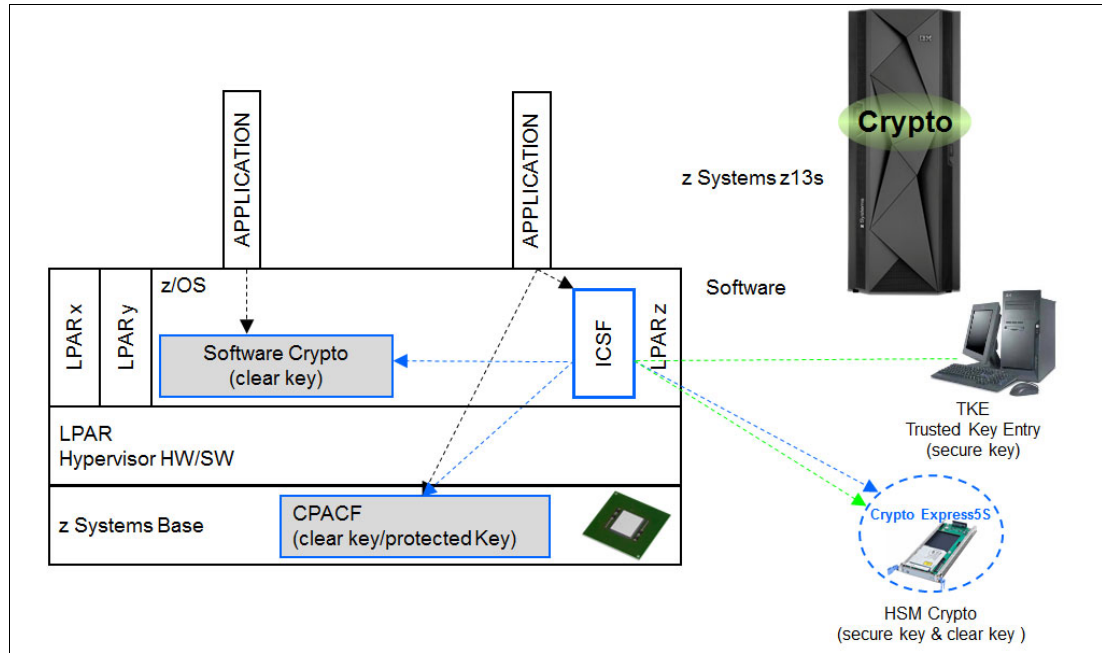


Figure 6-3 z13s cryptographic support in z/OS

ICSF is a software component of z/OS. ICSF works with the hardware cryptographic features and the Security Server (IBM RACF® element) to provide secure, high-speed cryptographic services in the z/OS environment. ICSF provides the APIs by which applications request the cryptographic services, as well from the CPACF and Crypto Express5S features. ICSF transparently routes application requests for cryptographic services to one of the integrated cryptographic engines, either CPACF or a Crypto Express5S card, depending on performance or requested cryptographic function. ICSF is also the means by which the secure Crypto Express5S features are loaded with master key values, allowing the hardware features to be used by applications. The cryptographic hardware installed in the z13s server determines the cryptographic features and services available to the applications.

The users of the cryptographic services call the ICSF API. Some functions are performed by the ICSF software without starting the cryptographic hardware features. Other functions result in ICSF going into routines that contain proprietary z Systems crypto instructions. These instructions are run by a CPU engine and result in a work request being generated for a cryptographic hardware feature.

6.4 CP Assist for Cryptographic Functions

As already mentioned, attached to every PU on an SCM in a CPC of a z13s server are two independent engines, one for compression and one for cryptographic purposes, as shown in Figure 6-4. This cryptographic coprocessor, called the CPACF, is not an HSM and is therefore not suitable for handling algorithms that use secret keys. However, the coprocessor can be used for cryptographic algorithms that use clear keys or protected keys. The CPACF is working synchronously to the PU, which means that the owning processor is busy when its coprocessor is busy. CPACF provides a fast device for cryptographic services.

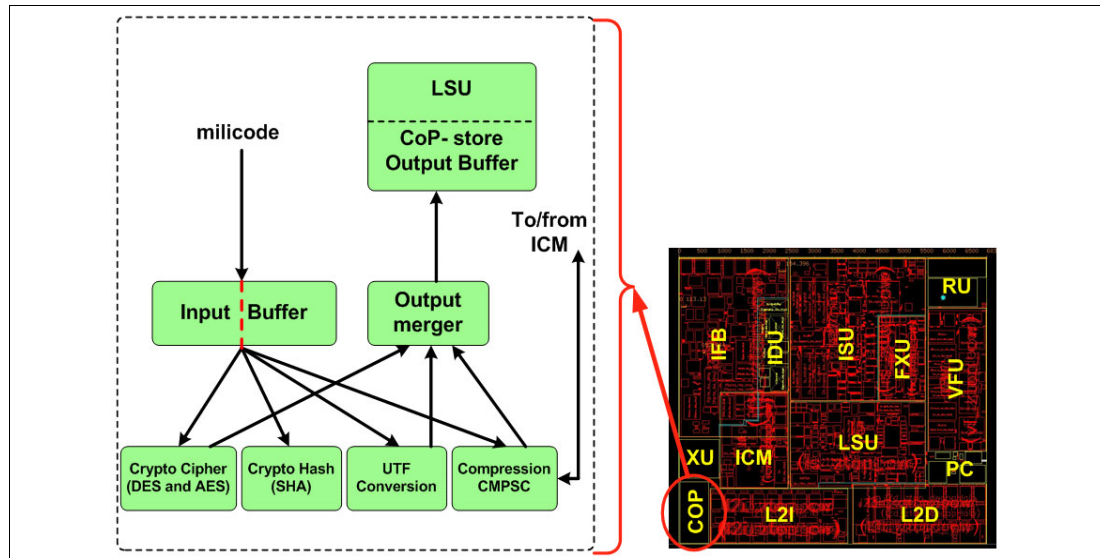


Figure 6-4 The cryptographic coprocessor CPACF

The CPACF offers a set of symmetric cryptographic functions that enhance the encryption and decryption performance of clear key operations. These functions are for SSL, virtual private network (VPN), and data-storing applications that do not require FIPS 140-2 Level 4 security.

CPACF is designed to facilitate the privacy of cryptographic key material when used for data encryption through key wrapping implementation. It ensures that key material is not visible to applications or operating systems during encryption operations. For more information, see 6.4.2, “CPACF protected key” on page 209

The CPACF feature provides hardware acceleration for DES, Triple-DES, AES-128, AES-192, AES-256 (all for clear and protected keys) as well as SHA-1, SHA-256, SHA-384, SHA-512, PRNG and DRNG (only clear key) cryptographic services. It provides high-performance hardware encryption, decryption, hashing, and random number generation support.

The following instructions support the cryptographic assist function:

KMAC	Compute Message Authentic Code
KM	Cipher Message
KMC	Cipher Message with Chaining
KMF	Cipher Message with CFB
KMCTR	Cipher Message with Counter
KMO	Cipher Message with OFB
KIMD	Compute Intermediate Message Digest
KLMD	Compute Last Message Digest
PCKMO	Provide Cryptographic Key Management Operation

These functions are provided as problem-state z/Architecture instructions that are directly available to application programs. These instructions are known as Message-Security Assist (MSA). When enabled, the CPACF runs at processor speed for every CP, IFL, and zIIP. For more information about MSA instructions, see *z/Architecture Principles of Operation*, SA22-7832.

The CPACF must be explicitly enabled by using an enablement feature (feature code 3863) that is available for no additional charge. The exception is the support for the hashing algorithms SHA-1, SHA-256, SHA-384, and SHA-512, which is always enabled.

6.4.1 Cryptographic synchronous functions

As the CPACF is working synchronously to the PU, it provides cryptographic synchronous functions. For IBM and client-written programs, CPACF functions can be started by the MSA instructions. z/OS ICSF callable services on z/OS, in-kernel crypto APIs, and a *libica* cryptographic functions library running on Linux on z Systems can also start CPACF synchronous functions.

The CPACF coprocessor in z13s servers is redesigned for improved performance compared to the zBC12 by more than two times for large block data, depending on the function that is being used. These tools might benefit from the throughput improvements:

- ▶ DB2/IMS encryption tool
- ▶ DB2 built-in encryption
- ▶ z/OS Communication Server: IPsec/IKE/AT-TLS
- ▶ z/OS System SSL
- ▶ z/OS Network Authentication Service (Kerberos)
- ▶ DFDSS Volume encryption
- ▶ z/OS Java SDK
- ▶ z/OS Encryption Facility
- ▶ Linux on z Systems: Kernel, openssl, openCryptoki, and GSKIT

The z13s hardware includes the implementation of algorithms as hardware synchronous operations. This configuration holds the PU processing of the instruction flow until the operation completes. z13s servers offer the following synchronous functions:

- ▶ Data encryption and decryption algorithms for data privacy and confidentiality:
 - Data Encryption Standard (DES):
 - Single-length key DES
 - Double-length key DES
 - Triple-length key DES (also known as Triple-DES)
 - Advanced Encryption Standard (AES) for 128-bit, 192-bit, and 256-bit keys
- ▶ Hashing algorithms for data integrity, such as SHA-1, and SHA-2 support for SHA-224, SHA-256, SHA-384, and SHA-512
- ▶ Message authentication code (MAC):
 - Single-length key MAC
 - Double-length key MAC
- ▶ Pseudo-random number generation (PRNG) and deterministic random number generation (DRNG) for cryptographic key generation.

For the SHA hashing algorithms and the random number generation algorithms, only clear keys are used. For the symmetric encryption/decryption DES and AES algorithms, clear keys

and protected keys can be used. Protected keys require a Crypto Express5S card running in CCA mode. For more information, see 6.5.2, “Crypto Express5S as a CCA coprocessor” on page 214.

The hashing algorithm SHA-1, and SHA-2 support for SHA-224, SHA-256, SHA-384, and SHA-512, are shipped enabled on all servers and do not require the CPACF enablement feature. For all other algorithms, the no-charge CPACF enablement feature (FC 3863) is required.

The CPACF functions are supported by z/OS, z/VM, z/VSE, z/TPF, and Linux on z Systems.

6.4.2 CPACF protected key

z13s server support the protected key implementation. Since IBM 4764 PCI-X cryptographic coprocessor (PCIXCC) deployment, secure keys are processed on the PCI-X and PCIe cards. This process requires an asynchronous operation to move the data and keys from the general-purpose central processor (CP) to the crypto cards. Clear keys process faster than secure keys because the process is done synchronously on the CPACF. Protected keys blend the security of Crypto Express5S coprocessors and the performance characteristics of the CPACF. This process allows it to run closer to the speed of clear keys.

An enhancement to CPACF facilitates the continued privacy of cryptographic key material when used for data encryption. In Crypto Express5S coprocessors, a secure key is encrypted under a master key. However, a protected key is encrypted under a wrapping key that is unique to each LPAR. Because the wrapping key is unique to each LPAR, a protected key cannot be shared with another LPAR. By using key wrapping, CPACF ensures that key material is not visible to applications or operating systems during encryption operations.

CPACF code generates the wrapping key and stores it in the protected area of the hardware system area (HSA). The wrapping key is accessible only by firmware. It cannot be accessed by operating systems or applications. DES/T-DES and AES algorithms are implemented in CPACF code with the support of hardware assist functions. Two variations of wrapping keys are generated: One for DES/T-DES keys and another for AES keys.

Wrapping keys are generated during the clear reset each time an LPAR is activated or reset. No customizable option is available at the Support Element (SE) or Hardware Management Console (HMC) that permits or avoids the wrapping key generation. Figure 6-5 shows this function flow.

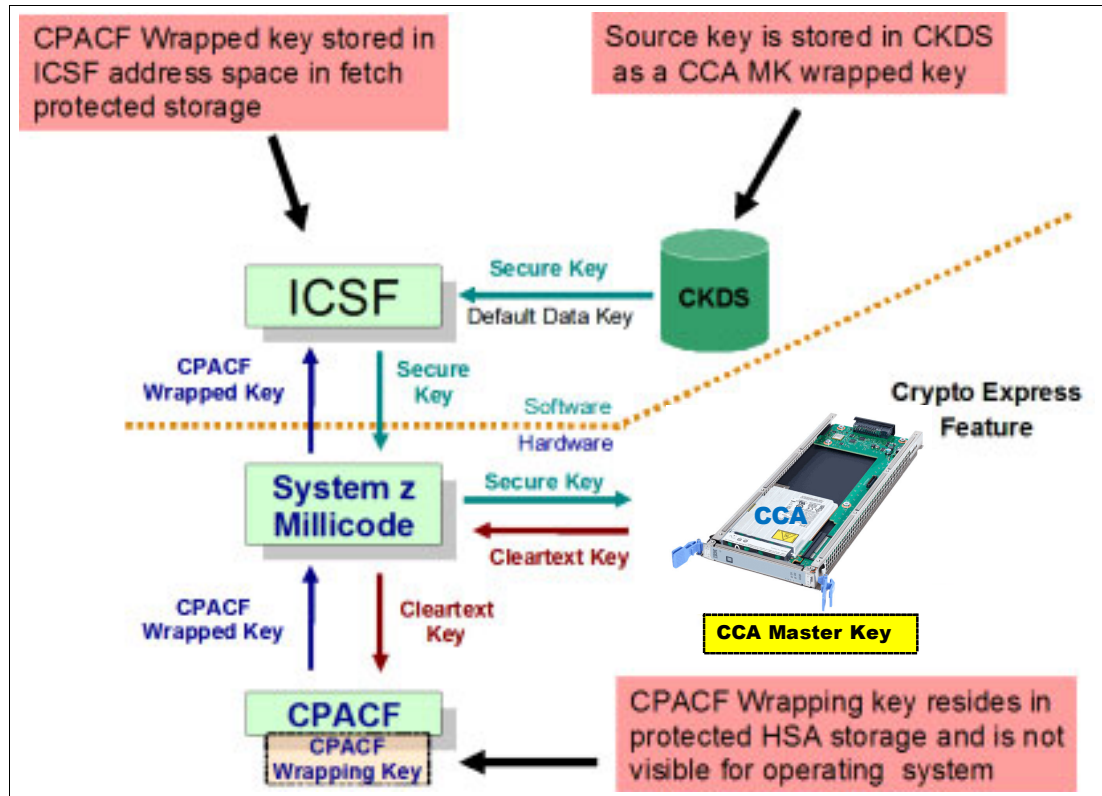


Figure 6-5 CPACF key wrapping

If a Crypto Express5S coprocessor (CEX5C) is available, a protected key can begin its life as a secure key. Otherwise, an application is responsible for creating or loading a clear key value, and then using the PCKMO instruction to wrap the key. ICSF is not called by the application if Crypto Express5S is not available.

A new segment in the profiles of the CSFKEYS class in IBM RACF restricts which secure keys can be used as protected keys. By default, all secure keys are considered not eligible to be used as protected keys. The process that is described in Figure 6-5 considers a secure key as being the source of a protected key.

The source key in this case already is stored in the ICSF cryptographic key data set (CKDS) as a secure key, which has been encrypted under the master key. This secure key is sent to Crypto Express5S to be deciphered, and sent to the CPACF in clear text. At the CPACF, the key is wrapped under the LPAR wrapping key, and is then returned to ICSF. After the key is wrapped, ICSF can keep the protected value in memory. It then passes it to the CPACF, where the key is unwrapped for each encryption/decryption operation.

The protected key is designed to provide substantial throughput improvements for a large volume of data encryption and low latency for encryption of small blocks of data. A high performance secure key solution, also known as a protected key solution, requires the ICSF HCR7770 as a minimum release.

6.5 Crypto Express5S

The Crypto Express5S feature (FC 0890) is an optional feature that is exclusive to z13 and z13s servers. Each feature has one PCIe cryptographic adapter. The Crypto Express5S feature occupies one I/O slot in a z13 or z13s PCIe I/O drawer. This feature is an HSM and provides a secure programming and hardware environment on which crypto processes are run. Each cryptographic coprocessor includes a general-purpose processor, non-volatile storage, and specialized cryptographic electronics. The Crypto Express5S feature provides tamper-sensing and tamper-responding, high-performance cryptographic operations.

Each Crypto Express5S PCI Express adapter can be in one of these configurations:

- ▶ Secure IBM CCA coprocessor (CEX5C) for FIPS 140-2 Level 4 certification. This configuration includes secure key functions. It is optionally programmable to deploy more functions and algorithms by using UDX. See 6.5.2, “Crypto Express5S as a CCA coprocessor” on page 214.
- ▶ Secure IBM Enterprise PKCS #11 (EP11) coprocessor (CEX5P) implements an industry-standardized set of services that adhere to the PKCS #11 specification V2.20 and more recent amendments. It was designed for extended FIPS and Common Criteria evaluations to meet public sector requirements. This new cryptographic coprocessor mode introduced the PKCS #11 secure key function. See 6.5.3, “Crypto Express5S as an EP11 coprocessor” on page 218.

A TKE workstation is required to support the administration of the Crypto Express5S when it is configured in EP11 mode.

- ▶ Accelerator (CEX5A) for acceleration of public key and private key cryptographic operations that are used with SSL/TLS processing. See 6.5.4, “Crypto Express5S as an accelerator” on page 218.

These modes can be configured by using the SE, and the PCIe adapter must be configured offline to change the mode.

Attention: Switching between configuration modes erases all card secrets. The exception is when you are switching from Secure CCA to accelerator, and vice versa.

The Crypto Express5S feature has been released for enhanced cryptographic performance. It is designed to provide more than double the performance of the Crypto Express4S feature. To achieve this performance, L2 Cache has been added, new Crypto ASIC has been implemented, and the internal processor has been upgraded from PowerPC 405 to PowerPC 476.

The Crypto Express5S feature does not have external ports and does not use optical fiber or other cables. It does not use channel-path identifiers (CHPIDs), but requires one slot in the PCIe I/O drawer and one physical channel ID (PCHID) for each PCIe cryptographic adapter. Removal of the feature or card *zeroizes* its content. Access to the PCIe cryptographic adapter is controlled through the setup in the image profiles on the SE.

Adapter: Although PCIe cryptographic adapters have no CHPID type and are not identified as external channels, all logical partitions (LPARs) in all channel subsystems have access to the adapter. In z13s servers, there are up to 40 LPARs per adapter. Having access to the adapter requires a setup in the image profile for each partition. The adapter must be in the candidate list.

Each z13s server supports up to 16 Crypto Express5S features. Table 6-2 shows configuration information for Crypto Express5S.

Table 6-2 *Crypto Express5S features*

Feature	Quantity
Minimum number of orderable features for each server ^a	2
Order increment above two features	1
Maximum number of features for each server	16
Number of PCIe cryptographic adapters for each feature (coprocessor or accelerator)	1
Maximum number of PCIe adapters for each server	16
Number of cryptographic domains for each PCIe adapter ^b	40

- a. The minimum initial order of Crypto Express5S features is two. After the initial order, more Crypto Express5S features can be ordered one feature at a time, up to a maximum of 16.
- b. More than one partition, which is defined to the same CSS or to different CSSs, can use the same domain number when assigned to different PCIe cryptographic adapters.

The concept of *dedicated processor* does not apply to the PCIe cryptographic adapter. Whether configured as a coprocessor or an accelerator, the PCIe cryptographic adapter is made available to an LPAR. It is made available as directed by the domain assignment and the candidate list in the LPAR image profile. This availability is not changed by the shared or dedicated status that is given to the CPs in the partition.

When installed non-concurrently, Crypto Express5S features are assigned PCIe cryptographic adapter numbers sequentially during the power-on reset (POR) that follows the installation. When a Crypto Express5S feature is installed concurrently, the installation can select an out-of-sequence number from the unused range. When a Crypto Express5S feature is removed concurrently, the PCIe adapter numbers are automatically freed.

The definition of domain indexes and PCIe cryptographic adapter numbers in the candidate list for each LPAR must be planned to allow for nondisruptive changes:

- ▶ Operational changes can be made by using the Change LPAR Cryptographic Controls task from the SE, which reflects the cryptographic definitions in the image profile for the partition. With this function, adding and removing the cryptographic feature on a running operating system can be done dynamically.
- ▶ The same usage domain index can be defined more than once across multiple LPARs. However, the PCIe cryptographic adapter number that is coupled with the usage domain index that is specified must be unique across all active LPARs.

The same PCIe cryptographic adapter number and usage domain index combination can be defined for more than one LPAR (up to 40). For example, you might define a configuration for backup situations. However, only one of the LPARs can be active at a time.

For more information, see 6.5.5, “Management of Crypto Express5S” on page 219.

6.5.1 Cryptographic asynchronous functions

The optional PCIe cryptographic coprocessor Crypto Express5S provides asynchronous cryptographic functions to z13s servers. Over 300 cryptographic algorithms and modes are supported, including the following:

- ▶ DES/TDES w DES/TDES MAC/CMAC: The Data Encryption Standard is a widespread symmetrical encryption algorithm. DES and TDES are no longer considered sufficiently secure for many applications. They have been replaced by AES as the official US standard, but are still used in the industry, together with MAC and the Cipher Based Message Authentication Code (CMAC) for verifying the integrity of messages.
- ▶ AES, AESKW, AES GMAC, AES GCM, AES XTS mode, CMAC: AES replaced DES as the official US standard in October 2000. Also, the enhanced standards for AES Key Wrap (AESKW), the AES Galois Message Authentication Code (AES GMAC) and Galois/Counter Mode (AES GCM), as well as the XEX-based tweaked-codebook mode with ciphertext stealing (AES XTS) and CMAC are supported.
- ▶ MD5, SHA-1, SHA-2 (224, 256, 384, 512), HMAC: The Secure Hash Algorithm (SHA-1 and the enhanced SHA-2 for different block sizes) as well as the older message-digest (MD5) algorithm and the advanced Hash-Based Message Authentication Code (HMAC) are used for verifying both the data integrity and the authentication of a message.
- ▶ Visa Format Preserving Encryption (VFPE): A method of encryption where the resulting cipher text has the same form as the input clear text, developed for use with credit cards.
- ▶ RSA (512, 1024, 2048, 4096): RSA has been published in 1977 and was named with the initial letters of the surnames of its authors Ron Rivest, Adi Shamir, and Leonard Adleman. It is a widely used asymmetric public-key algorithm, which means that the encryption key is public while the decryption key is kept secret. It is based on the difficulty of factoring the product of two large prime numbers. The number describes the length of the keys.
- ▶ ECDSA (192, 224, 256, 384, 521 Prime/NIST): Elliptic Curve Cryptography (ECC) is a family of asymmetric cryptographic algorithms based on the algebraic structure of elliptic curves. ECC can be used for encryption, pseudo-random number generation, and digital certificates. The Elliptic Curve Digital Signature Algorithm (ECDSA) Prime/NIST method is used for ECC digital signatures that are recommended for government use by the National Institute of Standards and Technology (NIST).
- ▶ ECDSA (160, 192, 224, 256, 320, 384, 512 BrainPool): ECC Brainpool is a workgroup of companies and institutions that collaborate on developing ECC algorithms. The ECDSA algorithms that are recommended by this group are supported.
- ▶ ECDH (192, 224, 256, 384, 521 Prime/NIST): Elliptic Curve Diffie-Hellman (ECDH) is an asymmetric protocol used for key agreement between two parties that use ECC-based private keys. The recommendations by NIST are supported.
- ▶ ECDH (160, 192, 224, 256, 320, 384, 512 BrainPool): ECDH according to the Brainpool recommendations.
- ▶ Montgomery Modular Math Engine: The Montgomery Modular Math Engine is a method for fast modular multiplication. Many crypto systems like RSA and Diffie-Hellman key Exchange use this method.
- ▶ Random Number Generator (RNG): The generation of random numbers for cryptographic key generation is supported.
- ▶ Prime Number Generator (PNG): The generation of prime numbers is supported.
- ▶ Clear Key Fast Path (Symmetric and Asymmetric): This mode of operations gives a direct hardware path to the cryptographic engine and provides high performance for public-key cryptographic functions.

Several of these algorithms require a secure key and must run on an HSM, some of them can also run with a clear key on the CPACF. Many standards are only supported when the Crypto Express5S card is running in CCA mode, many also when the card is running in EP11 mode. The three modes for the Crypto Express5S card are further described in the following topics. A summary of which algorithms are supported in which modes is shown in 6.7, “Cryptographic functions comparison” on page 225.

6.5.2 Crypto Express5S as a CCA coprocessor

A Crypto Express5S card running in CCA mode supports the IBM CCA. CCA is both an architecture and a set of APIs. It provides cryptographic algorithms and secure key management, especially many special functions required for banking. Over 129 APIs with more than 600 options are provided, with new functions and algorithms are always being added.

The IBM CCA provides functions for these purposes

- ▶ Encryption of data (DES/TDES/AES)
- ▶ Key management
 - Using TDES or AES keys
 - Using RSA or Elliptic Curve keys
- ▶ Message authentication
 - (MAC/HMAC/AES-CMAC)
- ▶ Key generation
- ▶ Digital signatures
- ▶ Random number generation
- ▶ Hashing (SHA, MD5, other)
- ▶ ATM PIN generation and processing
- ▶ Credit card transaction processing
- ▶ Visa Data Secure Platform (DSP) Point to Point Encryption (P2PE)
- ▶ Europay, MasterCard and Visa (EMV) card transaction processing
- ▶ Card personalization
- ▶ Other financial transaction processing
- ▶ Integrated role-based access control system

User Defined Extensions support

UDX allows the user to add customized operations to a cryptographic coprocessor. User-Defined Extensions to the CCA support customized operations that run within the Crypto Express features when defined as a coprocessor.

UDX is supported under a special contract through an IBM or approved third-party service offering. The Crypto Cards website directs your request to an IBM Global Services location for your geographic location. A special contract is negotiated between IBM Global Services and you. The contract is for the development of the UDX code by IBM Global Services according to your specifications and an agreed-upon level of the UDX.

A UDX toolkit for z Systems is tied to specific versions of the CCA card code and the related host code. UDX is available for the Crypto Express5S (Secure IBM CCA coprocessor mode only) features. An UDX migration is no more disruptive than a normal Microcode Change Level (MCL) or ICSF release migration.

In z13s servers, up to four UDX files can be imported. These files can be imported only from a DVD. The UDX configuration window is updated to include a **Reset to IBM Default** button.

Consideration: CCA has a new code level for z13s servers, and the UDX clients require a new UDX.

On z13s servers, the crypto Express5S card is delivered with CCA Level 5.2 firmware. A new set of cryptographic functions and callable services are provided by the IBM CCA LIC to enhance the functions that secure financial transactions and keys:

- ▶ Greater than 16 domains support, up to 40 LPARs on z13s servers and up to 85 LPARs on z13 servers, exclusive to z13 or z13s servers, and to Crypto Express5S
- ▶ VFPE support, exclusive to z13 or z13s servers and to Crypto Express5S
- ▶ AES PIN support for the German banking industry
- ▶ PKA Translate UDX function into CCA
- ▶ Verb Algorithm Currency

Greater than 16 domains support

z13s servers have support for up to 40 LPARs, and z13 servers have support for up to 85 LPARs. The z Systems crypto architecture was designed to support 16 domains (which matched the LPAR maximum at the time). Before z13 and z13s servers, in customer environments where the number of LPARs was larger than 16, crypto workload separation could be complex. These customers had to map a large set of LPARs to a small set of crypto domains.

Now, in z13s and z13 servers, with the Adjunct Processor (AP) Extended Addressing (APXA) facility that is installed, the z Systems crypto architecture can support up to 256 domains in an AP. As such, the Crypto Express cards are enhanced to handle 256 domains, and the z Systems firmware provides up to 40 on z13s and 85 on z13 domains to customers (to match the current LPAR maximum). Customers have the flexibility of mapping individual LPARs to unique crypto domains or continuing to share crypto domains across LPARs.

These are the requirements to support 40 or 85 domains:

- ▶ Hardware requirements:
 - z13s servers and Crypto Express5S with CCA V5.2 firmware
 - z13 servers and Crypto Express5S with CCA V5.0 or later firmware
- ▶ Software requirements:
 - z/OS V2.2
 - z/OS V2.1 and z/OS V1.13 with the Cryptographic Support for z/OS V1R13-z/OS V2R1 web deliverable (FMID HCR77B0)
 - Also, available with HCR7780, HCR7790, HCR77A0, and HCR77A1 (previous WDs with program temporary fixes (PTFs))
 - z/VM V6.2 and Version 6.3 with PTFs for guest use

Visa Format Preserving Encryption

VFPE refers to a method of encryption where the resulting cipher text has the same form as the input clear text. The form of the text can vary according to use and application. One of the classic examples is a 16-digit credit card number. After using VFPE to encrypt a credit card number, the resulting cipher text is another 16-digit number. This helps legacy databases contain encrypted data of sensitive fields without having to restructure the database or applications.

VFPE allows customers to add encryption to their applications in such a way that the encrypted data can flow through their systems without requiring a massive redesign of their application. In the example, if the credit card number is VFPE-encrypted at the point of entry, the cipher text still behaves like a credit card number. It can flow through business logic until it meets that back-end transaction server that can VFPE-decrypt it to get the original credit card number to process the transaction.

Here are the FPE requirements:

- ▶ Hardware requirements:
 - z13s or z13 servers and Crypto Express5S with CCA V5.2 firmware
- ▶ Software requirements:
 - z/OS V2.2
 - z/OS V2.1 and z/OS V1.13 with the Cryptographic Support for z/OS V1R13-z/OS V2R1 web deliverable (FMID HCR77B0)
 - z/VM V6.2 and Version 6.3 with PTFs for guest use

AES PIN support for the German banking industry

The German banking industry organization, DK, has defined a new set of PIN processing functions to be used on the internal systems of banks and their servers. CCA is designed to support the functions that are essential to those parts of the German banking industry that are governed by DK requirements. The functions include key management support for new AES key types, AES key derivation support, and several DK-specific PIN and administrative functions.

This support includes PIN method APIs, PIN administration APIs, new key management verbs, and new access control points support that is needed for DK-defined functions.

Here are the requirements for AES PIN support:

- ▶ Hardware requirements:
 - z13s servers and Crypto Express5S with CCA V5.2 firmware
 - z13 servers and Crypto Express5S with CCA V5.0 or later firmware
 - zEC12 or zBC12 servers and Crypto Express4S with CCA V4.4 firmware
 - zEC12, zBC12, z196, or z114 servers and Crypto Express3 with CCA V4.4 firmware
- ▶ Software requirements:
 - z/OS V2.2
 - z/OS V2.1 or z/OS V1.13 with the Cryptographic Support for z/OS V1R13-z/OS V2R1 web deliverable (FMID HCR77A1) with PTFs
 - z/OS V2.1 (FMID HCR77A0) with PTFs
 - z/OS V1.12 or z/OS V1.13 with the Cryptographic Support for z/OS V1R12-V1R13 web deliverable (FMID HCR77A0) with PTFs
 - z/VM Version 6.2, and Version 6.3 with PTFs for guest use

PKA Translate UDX function into CCA

UDX is custom code that allows the client to add unique operations and extensions to the CCA firmware. Certain UDX functions are integrated into the base CCA code over time to accomplish the following tasks:

- ▶ Removes headaches and pain points that are associated with UDX management and currency.
- ▶ Popular UDX functions are made available to a wider audience to encourage adoption.

UDX is integrated into the base CCA code to support translating an external RSA Chinese Remainder Theorem key into new formats. These new formats use tags to identify key components. Depending on which new rule array keyword is used with the PKA Key Translate callable service, the service TDES encrypts those components in either CBC or ECB mode. In addition, AES CMAC Support is delivered.

This function has the following requirements:

- ▶ Hardware requirements:
 - z13s servers and Crypto Express5S with CCA V5.2 firmware
 - z13 servers and Crypto Express5S with CCA V5.0 or later firmware
 - zEC12 or zBC12 servers and Crypto Express4S with CCA V4.4 firmware
 - zEC12, zBC12, z196 or z114 servers and Crypto Express3 with CCA V4.4 firmware
- ▶ Software requirements:
 - z/OS V2.2
 - z/OS V2.1 or z/OS V1.13 with the Cryptographic Support for z/OS V1R13-z/OS V2R1 web deliverable (FMID HCR77A1) with PTFs
 - z/OS V2.1 (FMID HCR77A0) with PTFs
 - z/OS V1.13 with the Cryptographic Support for z/OS V1R13-z/OS V2R1 web deliverable (FMID HCR77A1) with PTFs
 - z/OS V1.12 or z/OS V1.13 with the Cryptographic Support for z/OS V1R12-V1R13 web deliverable (FMID HCR77A0) with PTFs
 - z/OS V1.12 or V1.13 with the Cryptographic Support for z/OS V1R11-V1R13 web deliverable (FMID HCR7790) with PTFs
 - z/VM Version 6.2, and Version 6.3 with PTFs for guest use

Verb Algorithm Currency

Verb Algorithm Currency is a collection of CCA verb enhancements related to customer requirements, with the intent of maintaining currency with cryptographic algorithms and standards. It is intended for customers wanting to maintain the latest cryptographic capabilities. It provides these functions:

- ▶ Secure key support AES GCM encryption
- ▶ New Key Check Value (KCV) algorithm for service CSNBKYT2 Key Test 2
- ▶ New key derivation options for CSNDEDH EC Diffie-Hellman service

Here are the requirements for this function:

- ▶ Hardware requirements:
 - z13s or z13 servers and Crypto Express5S with CCA V5.2 firmware

- ▶ Software requirements:
 - z/OS V2.2
 - z/OS V2.1 or z/OS V1.13 with the Cryptographic Support for z/OS V1R13-z/OS V2R1 web deliverable (FMID HCR77B1) with PTFs
 - z/VM 5.4, 6.2, and 6.3 with PTFs for guest exploitation

6.5.3 Crypto Express5S as an EP11 coprocessor

A Crypto Express5S card that is configured in Secure IBM Enterprise PKCS #11 (EP11) coprocessor mode provides PKCS #11 secure key support for public sector requirements. Before EP11, the ICSF PKCS #11 implementation supported only clear keys. In EP11, keys can now be generated and securely wrapped under the EP11 Master Key. The secure keys never leave the secure coprocessor boundary unencrypted.

The secure IBM Enterprise PKCS #11 (EP11) coprocessor runs the following tasks:

- ▶ Encrypt and decrypt (AES, DES, TDES, and RSA)
- ▶ Sign and verify (DSA, RSA, and ECDSA)
- ▶ Generate keys and key pairs (DES, AES, DSA, ECC, and RSA)
- ▶ HMAC (SHA1, SHA224, SHA256, SHA384, and SHA512)
- ▶ Digest (SHA1, SHA224, SHA256, SHA384, and SHA512)
- ▶ Wrap and unwrap keys
- ▶ Random number generation
- ▶ Get mechanism list and information
- ▶ Attribute values
- ▶ Key agreement (Diffie-Hellman)

The function extension capability through UDX is not available to the EP11.

When defined in EP11 mode, the TKE workstation is required to manage the Crypto Express5S feature.

6.5.4 Crypto Express5S as an accelerator

A Crypto Express5S card running in accelerator mode supports only RSA clear key and SSL Acceleration. A request is processed fully in hardware. The Crypto Express accelerator is a coprocessor that is reconfigured by the installation process so that it uses only a subset of the coprocessor functions at a higher speed. Reconfiguration is disruptive to coprocessor and accelerator operations. The coprocessor or accelerator must be deactivated before you begin the reconfiguration.

FIPS 140-2 certification is not relevant to the accelerator because it operates with clear keys only. The function extension capability through UDX is not available to the accelerator.

The functions that remain available when the Crypto Express5S feature is configured as an accelerator are used for the acceleration of modular arithmetic operations. That is, the RSA cryptographic operations are used with the SSL/TLS protocol. The following operations are accelerated:

- ▶ PKA Decrypt (CSNDPKD) with PKCS-1.2 formatting
- ▶ PKA Encrypt (CSNDPKE) with zero-pad formatting
- ▶ Digital Signature Verify

The RSA encryption and decryption functions support key lengths of 512 bits to 4,096 bits, in the Modulus-Exponent (ME) and CRT formats.

6.5.5 Management of Crypto Express5S

With zEC12 and older servers, each cryptographic coprocessor has 16 physical sets of registers or queue registers. With z13, this number is raised to 85. This amount corresponds to the maximum number of LPARs running on a z13, which is also 85. Therefore, with z13s servers, the number of register sets for a Crypto Express5S card is 40. Each of these 40 sets belongs to a domain as follows:

- ▶ A cryptographic domain index, in the range of 0 - 39, is allocated to a logical partition in its image profile. The same domain must also be allocated to the ICSF instance that runs in the logical partition that uses the Options data set.
- ▶ Each ICSF instance accesses only the Master Keys or queue registers corresponding to the domain number specified in the logical partition image profile at the SE and in its Options data set. Each ICSF instance sees a logical cryptographic coprocessor that consists of the physical cryptographic engine and the unique set of registers (the domain) allocated to this logical partition

The installation of the CP Assist for Cryptographic Functions (CPACF) DES/TDES enablement, Feature Code #3863, is required to use the Crypto Express5S feature.

Each Crypto Express5S feature contains one PCI-X adapter. The adapter can be in the following configurations:

- ▶ IBM Enterprise Common Cryptographic Architecture (CCA) Coprocessor (CEX5C)
- ▶ IBM Enterprise Public Key Cryptography Standards#11 (PKCS) Coprocessor (CEX5P)
- ▶ IBM Crypto Express5S Accelerator (CEX5A)

During the feature installation, the PCI-X adapter is configured by default as the CCA coprocessor.

The configuration of the Crypto Express5S adapter as EP11 coprocessor requires a Trusted Key Entry (TKE) tower workstation (FC 0847) or a TKE rack mounted workstation (FC 0097) with TKE 8.0 (FC 0877) or 8.1 (FC 0878) Licensed Internal Code.

The Crypto Express5S feature does not use CHPIDs from the channel subsystem pool. However, the Crypto Express5S feature requires one slot in a PCIe I/O drawer, and one PCHID for each PCIe cryptographic adapter.

When enabling an LPAR to use a Crypto Express5S card, define the following cryptographic resources in the image profile for each partition:

- ▶ Usage domain index
- ▶ Control domain index
- ▶ PCI Cryptographic Coprocessor Candidate List
- ▶ PCI Cryptographic Coprocessor Online List

This task is accomplished by using the Customize/Delete Activation Profile task, which is in the Operational Customization Group, either from the HMC or from the SE. Modify the cryptographic initial definition from the Crypto option in the image profile, as shown in Figure 6-6. After this definition is modified, any change to the image profile requires a DEACTIVATE and ACTIVATE of the logical partition for the change to take effect. Therefore, this cryptographic definition is disruptive to a running system.

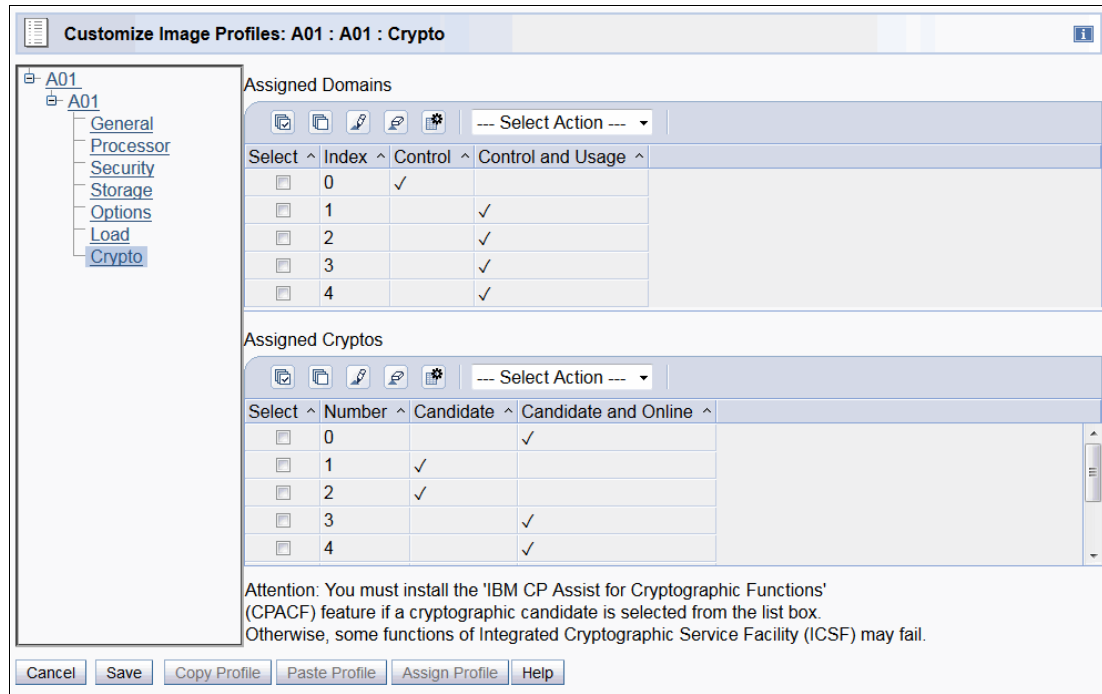


Figure 6-6 Customize Image Profiles: Crypto

The cryptographic resource definitions have the following meanings:

Control Domain

Identifies the cryptographic coprocessor domains that can be administered from this logical partition if it is being set up as the TCP/IP host for the TKE. If you are setting up the host TCP/IP in this logical partition to communicate with the TKE, the partition is used as a path to other domains' Master Keys. Indicate all the control domains that you want to access (including this partition's own control domain) from this partition.

Control and Usage Domain

Identifies the cryptographic coprocessor domains that are assigned to the partition for all cryptographic coprocessors that are configured on the partition. The usage domains cannot be removed if they are online. The numbers that are selected must match the domain numbers entered in the Options data set when you start this partition instance of ICSF. The same usage domain index can be used by multiple partitions regardless to which CSS they are defined. However, the combination of PCI-X adapter number and usage domain index number must be unique across all active partitions.

Cryptographic Candidate list Identifies the cryptographic coprocessor numbers that are eligible to be accessed by this logical partition. From the list, select the coprocessor numbers, in the range 0 - 15, that identify the PCI-X adapters to be accessed by this partition.

Cryptographic Online list Identifies the cryptographic coprocessor numbers that are automatically brought online during logical partition activation. The numbers that are selected in the online list must also be part of the candidate list.

After they are activated, the active partition cryptographic definitions can be viewed from the SE. Select the CPCs, and click **View LPAR Cryptographic Controls** in the CPC Operational Customization window. The resulting window displays the definition of Usage and Control domain indexes, and PCI Cryptographic candidate and online lists, as shown in Figure 6-7. The information is provided only for active logical partitions.

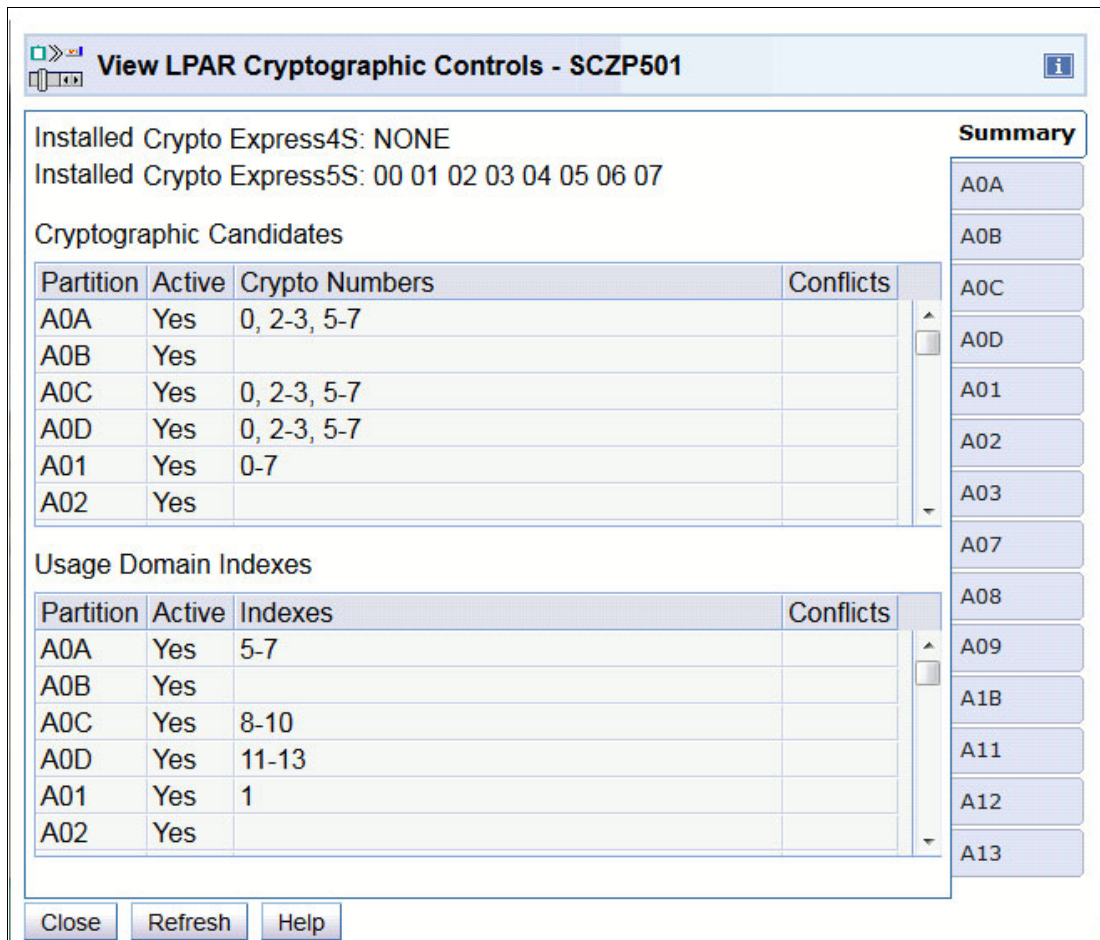


Figure 6-7 SE: View LPAR Cryptographic Controls

Operational changes can be made by using the Change LPAR Cryptographic Controls task from the SE, which reflects the cryptographic definitions in the image profile for the partition. With this function, the cryptographic feature can be added and removed dynamically, without stopping a running operating system.

For more information about the management of Crypto Express5S cards see the corresponding chapter in *IBM z13 Configuration Setup*, SG24-8260.

6.6 TKE workstation

The TKE workstation is an optional feature that offers key management functions. A TKE tower workstation feature (FC 0847) or a TKE rack mounted workstation feature (FC 0097) is required for z13s servers to manage the Crypto Express5S feature.

The TKE contains a combination of hardware and software. A mouse, keyboard, flat panel display, PCIe adapter, and a writable USB media to install the TKE LIC are included with the system unit. The TKE workstation requires an IBM 4767 crypto adapter.

A TKE workstation is part of a customized solution for using the ICSF for z/OS on Linux for z Systems. This program provides a basic key management system for the cryptographic keys of a z13s server that has Crypto Express features installed.

The TKE provides a secure, remote, and flexible method of providing Master Key Part Entry, and to manage remotely PCIe cryptographic coprocessors. The cryptographic functions on the TKE are run by one PCIe cryptographic coprocessor. The TKE workstation communicates with the z Systems server through a TCP/IP connection. The TKE workstation is available with Ethernet LAN connectivity only. Up to 10 TKE workstations can be ordered. TKE FCs 0847 and 0097 can be used to control the Crypto Express5S cards on z13s servers. They can also be used to control the Crypto Express5S on z13, and the older crypto cards on zEC12, zBC12, z196, z114, z10 EC, z10 BC, z9 EC, z9 BC, z990, and z890 servers.

Tip: For handling a TKE, a series of instructive video clips is provided at:

<http://www.youtube.com/user/IBMTKE>

6.6.1 Logical partition, TKE host, and TKE target

If one or more LPARs are configured to use Crypto Express5S coprocessors, the TKE workstation can be used to manage DES, AES, ECC, and PKA master keys. This management can be done for all cryptographic domains of each Crypto Express coprocessor feature that is assigned to the LPARs defined to the TKE workstation.

Each LPAR in the same system that uses a domain that is managed through a TKE workstation connection is either a TKE host or a TKE target. An LPAR with a TCP/IP connection to the TKE is referred to as the *TKE host*. All other partitions are *TKE targets*.

The cryptographic controls that are set for an LPAR through the SE determine whether the workstation is a TKE host or a TKE target.

6.6.2 Optional smart card reader

An optional smart card reader (FC 0885) can be added to the TKE workstation. One FC 0885 includes two smart card readers, two cables to connect them to the TKE workstation, and 20 smart cards. The reader supports the use of smart cards that contain an embedded microprocessor and associated memory for data storage. The memory can contain the keys to be loaded into the Crypto Express features.

Access to and use of confidential data on the smart card are protected by a user-defined PIN. Up to 990 more smart cards can be ordered for backup. The additional smart card feature code is FC 0884. When one feature code is ordered, 10 smart cards are shipped. The order increment is 1 - 99 (10 - 990 blank smart cards).

6.6.3 TKE workstation with Licensed Internal Code 8.0

To control the Crypto Express5S card in a z13s server, a TKE workstation (FC 0847 or 0097) with LIC 8.0 (FC 0877) or LIC 8.1 (FC 0878) is required. LIC 8.0 does not provide the new functions of LIC 8.1. TKE LIC 8.1 is delivered with a z13s server. To control a Crypto Express5S in a z13s server with a TKE workstation running LIC 8.0, delivered with an already installed z13, an MES upgrade to LIC 8.1 is required.

LIC 8.0 has the following enhancements compared to the older LIC 7.x:

- ▶ TKE workstation with LIC 8.0 or higher is required to manage a Crypto Express5S host.
- ▶ Only a TKE workstation with LIC 8.0 or higher can be used to manage domains higher than 16 on a Crypto Express5S feature.
- ▶ The Full Function Migration Wizard is required when data is applied to a Crypto Express5S host. If data is applied to a Crypto Express5S host, the collection must be done by using Key Part Holder Certificates from Key Part Holder (KPH) smart cards that are created on a TKE workstation with LIC 8.0 or higher.

Recommendation: During a migration, if data is applied to a Crypto Express5S, collect the source module from the TKE workstation with LIC 8.0 or later.

6.6.4 TKE workstation with Licensed Internal Code 8.1

The TKE 8.1 LIC (FC 0878) offers the following new features:

- ▶ Domain Cloning: The ability to collect data from one domain and push it to a set of domains. This feature is valuable for deploying new domains.
- ▶ Coordinated Master Key roll: Ability to start Coordinated Master Key roll from the TKE
- ▶ Three new wizard-like features: Create new TKE zone, Create new Migration Zone, and Configure Host Roles and Authorities.
- ▶ Operational Key Option: This feature allows the client to decide whether operational key commands are limited to the master domain or sent to all domains in the group.
- ▶ HMAC key: Support for HMAC key has been added. The key is limited to three specific sizes: 128, 192, and 256.
- ▶ TKE enables Save Customized Data feature: This feature simplifies the way that a client can save and restore client data to a TKE.
- ▶ TKE can be configured to prevent auto-logout: If configured, a password is required to start the Trusted Key Entry Console web application.
- ▶ Binary Key Part File Utility: This feature allows the client to copy a key part from a binary file to a smart card.
- ▶ ACP Usage Information: This feature allows clients to determine which Domain Controls (Access Control Points) are actually “checked/used” on a domain. The utility allows you to activate and deactivate tracking and create reports.
- ▶ Display Crypto Module Settings: This feature allows you to build a report that shows the settings of a crypto module.

6.6.5 TKE hardware support and migration information

The new TKE 8.1 LIC (FC 0878) is shipped with z13 servers after GA2, and with z13s servers.

If a new TKE 8.1 is purchased, two versions are available:

- ▶ TKE 8.1 tower workstation (FC 0847)
- ▶ TKE 8.1 rack-mounted workstation (FC 0097)

The TKE 8.1 LIC requires the 4767 crypto adapter. Any TKE 8.0 workstation can be upgraded to TKE 8.1 tower workstation (FC 0847), which uses the same crypto adapter. The TKE 7.3 workstation (FC 0842) can be upgraded to the TKE 8.1 tower workstation by purchasing a 4767 crypto adapter.

The Omnikey Cardman 3821 smart card readers can be carried forward to any TKE 8.1 workstation. Smart cards 45D3398, 74Y0551, and 00JA710 can be used with TKE 8.1

When doing a MES upgrade to a TKE 8.1 installation, the following steps must be performed.

Upgrade from TKE 8.0 to TKE 8.1:

1. Save Upgrade Data on TKE 8.0 to USB memory to save client data.
2. Upgrade firmware to TKE 8.1.
3. Perform a Frame Roll install to apply Save Upgrade Data (client data) to the TKE 8.1 system.
4. Run the TKE Workstation Setup wizard.

Upgrade from TKE 7.3 to TKE 8.1:

1. Save Upgrade Data on TKE 7.3 to USB memory to save client data.
2. Replace the 4765 crypto adapter with the 4767 crypto adapter.
3. Upgrade the firmware to TKE 8.1.
4. Perform a Frame Roll install to apply Save Upgrade Data (client data) to the TKE 8.1 system
5. Run the TKE Workstation Setup wizard.

For more information about TKE hardware support, see Table 6-3.

Table 6-3 TKE Compatibility Matrix

TKE Workstation	TKE Release LIC	7.2	7.3 ^a	8.0	8.1
	HW Feature Code	0814	0842	0847	0847 or 0097
	LICC	0850	0872	0877	0878
	Smart Card Reader	0885	0885	0891	0891
	Smart Card	0884	0884 ^b	0892	0892

Server supported	z196	Yes	Yes	Yes	Yes
	z114	Yes	Yes	Yes	Yes
	zEC12	Yes	Yes	Yes	Yes
	zBC12		Yes	Yes	Yes
	z13			Yes	Yes
	z13s			Yes	Yes
Manage Host Crypto Module	CEC3C (CCA)	Yes	Yes	Yes	Yes
	CEX4C (CCA)	Yes	Yes	Yes	Yes
	CEX4P (EP11)	Yes ^c	Yes ^c	Yes ^c	Yes ^d
	CEX5C (CCA)			Yes	Yes
	CEX5P (EP11)			Yes	Yes

- a. TKE workstation (FC 0842) with LIC 7.3 can be upgraded to TKE tower workstation (FC 0847) with LIC 8.0 or LIC 8.1. The MES generates FC 0894 to add the IBM 4767 adapter.
- b. Older smart cards 45D3398 (FC 0884) and 74Y0551 (FC 0884) can be used on TKE workstation with LIC 8.0 (available from System z10)
- c. A Crypto Express4S running in EP11 mode requires smart cards to hold administrator certificates and master key material. The smart card must be P/N 74Y0551.
- d. A Crypto Express4S running in EP11 mode requires smart cards to hold administrator certificates and master key material. The smart card must be P/N 74Y0551.

Attention: The TKE is unaware of the CPC type where the host crypto module is installed. That is, the TKE does not care whether a Crypto Express is running on a z196, z114, zEC12, zBC12, z13, or z13s servers. Therefore, the LIC can support any CPC where the coprocessor is supported, but the TKE LIC must support the specific crypto module.

6.7 Cryptographic functions comparison

Table 6-4 lists the functions or attributes on z13s servers for the two cryptographic hardware features. In the table, “X” indicates that the function or attribute is supported.

Table 6-4 Cryptographic functions on z13s servers

Functions or attributes	CPACF	CEX5C	CEX5P	CEX5A
Supports z/OS applications using ICSF	X	X	X	X
Supports Linux on z Systems CCA applications	X	X	-	X
Encryption and decryption using secret-key algorithm	-	X	X	-
Provides the highest SSL/TLS handshake performance	-	-	-	X
Supports SSL/TLS functions	X	X	-	X
Provides the highest symmetric (clear key) encryption performance	X	-	-	-
Provides the highest asymmetric (clear key) encryption performance	-	-	-	X

Functions or attributes	CPACF	CEX5C	CEX5P	CEX5A
Provides the highest asymmetric (encrypted key) encryption performance	-	X	X	-
Disruptive process to enable	-	Note ^a	Note ^a	Note ^a
Requires IOCDs definition	-	-	-	-
Uses CHPID numbers	-	-	-	-
Uses PCHIDs (one PCHID)	-	X	X	X
Requires CPACF enablement (FC 3863)	X ^b	X ^b	X ^b	X ^b
Requires ICSF to be active	-	X	X	X
Offers UDX	-	X	-	-
Usable for data privacy: Encryption and decryption processing	X	X	X	-
Usable for data integrity: Hashing and message authentication	X	X	X	-
Usable for financial processes and key management operations	-	X	X	-
Crypto performance IBM RMF™ monitoring	-	X	X	X
Requires system master keys to be loaded	-	X	X	-
System (master) key storage	-	X	X	-
Retained key storage	-	X	-	-
Tamper-resistant hardware packaging	-	X	X	X ^c
Designed for FIPS 140-2 Level 4 certification	-	X	X	X
Supports Linux applications that perform SSL handshakes	-	-	-	X
RSA functions	-	X	X	X
High-performance SHA-1 and SHA2	X	X	X	-
Clear key DES or triple DES	X	-	-	-
AES for 128-bit, 192-bit, and 256-bit keys	X	X	X	-
Pseudorandom number generator (PRNG)	X	X	X	-
Clear key RSA	-	-	-	X
Europay, MasterCard and Visa (EMV) support	-	X	-	-
Public Key Decrypt (PKD) support for Zero-Pad option for clear RSA private keys	-	X	-	-
Public Key Encrypt (PKE) support for Mod_Raised_to Power (MRP) function	-	X	X	-
Remote loading of initial keys in ATM	-	X	-	-
Improved key exchange with non-CCA systems	-	X	-	-

Functions or attributes	CPACF	CEX5C	CEX5P	CEX5A
ISO 16609 CBC mode triple DES message authentication code (MAC) support	-	X	-	-
AES GMAC, AES GCM, AES XTS mode, CMAC	-	X	-	-
SHA-2 (384,512), HMAC	-	X	-	-
Visa Format Preserving Encryption	-	X	-	-
AES PIN support for the German banking industry				
ECDSA (192, 224, 256, 384, 521 Prime/NIST)	-	X	-	-
ECDSA (160, 192, 224, 256, 320, 384, 512 BrainPool)	-	X	-	-
ECDH (192, 224, 256, 384, 521 Prime/NIST)	-	X	-	-
ECDH (160, 192, 224, 256, 320, 384, 512 BrainPool)	-	X	-	-
PNG (Prime Number Generator)	-	X	-	-

- a. To make the addition of the Crypto Express features nondisruptive, the logical partition must be predefined with the appropriate PCI Express cryptographic adapter number. This number must be selected in its candidate list in the partition image profile.
- b. This feature is not required for Linux if only RSA clear key operations are used. DES or triple DES encryption requires CPACF to be enabled.
- c. This feature is physically present but is not used when configured as an accelerator (clear key only).

6.8 Cryptographic software support

The software support levels for cryptographic functions are listed in Chapter 7.4, “Cryptographic support” on page 285.



Software support

This chapter lists the minimum operating system requirements and support considerations for the IBM z13s (z13s) and its features. It addresses z/OS, z/VM, z/VSE, z/TPF, Linux on z Systems, and KVM for IBM z Systems. Because this information is subject to change, always see the Preventive Service Planning (PSP) bucket for 2965DEVICE for the most current information. Also included is generic software support for IBM z BladeCenter Extension (zBX) Model 004.

Support of z13s functions depends on the operating system, its version, and release.

This chapter covers the following topics:

- ▶ Operating systems summary
- ▶ Support by operating system
- ▶ Support by function
- ▶ Cryptographic support
- ▶ GDPS Virtual Appliance
- ▶ z/OS migration considerations
- ▶ IBM z Advanced Workload Analysis Reporter (zAware)
- ▶ Coupling facility and CFCC considerations
- ▶ Simultaneous multithreading (SMT)
- ▶ Single-instruction multiple-data (SIMD)
- ▶ The MIDAW facility
- ▶ IOCP
- ▶ Worldwide port name (WWPN) tool
- ▶ ICKDSF
- ▶ IBM z BladeCenter Extension (zBX) Model 004 software support
- ▶ Software licensing
- ▶ References

7.1 Operating systems summary

Table 7-1 lists the minimum operating system levels that are required on the z13s. For similar information about the IBM zBX Model 004, see 7.15, “IBM z BladeCenter Extension (zBX) Model 004 software support” on page 304.

End-of-service operating systems: The operating system levels earlier than those in Table 7-1 are no longer in service, and are not covered in this publication. These older levels might provide support for some features.

Table 7-1 z13s minimum operating systems requirements

Operating Systems	ESA/390 (31-bit mode)	z/Architecture (64-bit mode)	Notes
z/OS V1R12 ^a	No	Yes	Service is required. See the following box, which is titled “Features”.
z/VM V6R2 ^b	No	Yes ^c	
z/VSE V5R1	No	Yes	
z/TPF V1R1	Yes	Yes	
Linux on z Systems	No ^d	See Table 7-2 on page 232.	
KVM for IBM z Systems	No	Yes	

a. Regular service support for z/OS V1R12 ended September 2014. However, by ordering the IBM Software Support Services - Service Extension for z/OS V1.12, fee-based corrective service can be obtained up to September 2017.

b. z/VM V6R2 with PTF provides compatibility support (CEX5S with enhanced crypto domain support)

c. z/VM supports both 31-bit and 64-bit mode guests.

d. 64-bit distributions included the 31-bit emulation layer to run 31-bit software products.

Features: Usage of certain features depends on the operating system. In all cases, PTFs might be required with the operating system level that is indicated. Check the z/OS, z/VM, z/VSE, and z/TPF subsets of the 2965DEVICE PSP buckets. The PSP buckets are continuously updated, and contain the latest information about maintenance:

- ▶ Hardware and software buckets contain installation information, hardware and software service levels, service guidelines, and cross-product dependencies.
- ▶ For Linux on z Systems distributions, consult the distributor’s support information.
- ▶ For KVM for IBM z Systems, see *Getting Started with KVM for IBM z Systems*, SG24-8332.

7.2 Support by operating system

The IBM z13s introduces several new functions. This section addresses support of those functions by the current operating systems. Also included are some of the functions that were introduced in previous z Systems and carried forward or enhanced in z13s. Features and

functions that are available on previous servers but no longer supported by z13s have been removed.

For a list of supported functions and the z/OS and z/VM minimum required support levels, see Table 7-3 on page 234. For z/VSE, z/TPF, and Linux on z Systems, see Table 7-4 on page 239. The tabular format is intended to help you determine, by a quick scan, which functions are supported and the minimum operating system level that is required.

7.2.1 z/OS

z/OS Version 1 Release 13 is the earliest in-service release that supports z13s. After September 2016, a fee-based service extension for defect support (for up to three years) can be obtained for z/OS V1R13. Although service support for z/OS Version 1 Release 12 ended in September of 2014, a fee-based extension for defect support (for up to three years) can be obtained by ordering IBM Software Support Services - Service Extension for z/OS 1.12. Also, z/OS.e is not supported on z13s, and z/OS.e Version 1 Release 8 was the last release of z/OS.e.

z13s capabilities differ depending on the z/OS release. Toleration support is provided on z/OS V1R12. Exploitation support is provided only on z/OS V2R1 and later. For a list of supported functions and their minimum required levels, see Table 7-3 on page 234.

7.2.2 z/VM

At general availability, z/VM V6R2 and V6R3 provide compatibility support with limited use of new z13s functions. For a list of supported functions and their minimum required support levels, see Table 7-3 on page 234.

For the capacity of any z/VM logical partition (LPAR), and any z/VM guest, in terms of the number of Integrated Facility for Linux (IFL) processors and central processors (CPs), real or virtual, you might want to adjust the number to accommodate the processor unit (PU) capacity of the z13s.

In light of the IBM cloud strategy and adoption of OpenStack, the management of z/VM environments in zManager is now stabilized and will not be further enhanced. zManager therefore does not provide systems management support for z/VM V6R2 on IBM z13s or for z/VM V6.3 and later releases. However, zManager continues to play a distinct and strategic role in the management of virtualized environments that are created by integrated firmware hypervisors of the IBM z Systems. These hypervisors include IBM Processor Resource/Systems Manager (PR/SM), PowerVM, and System x hypervisor based on a kernel-based virtual machine (KVM).

Statements of Direction:

- ▶ *Removal of support for Expanded Storage (XSTORE):* z/VM V6.3 is the last z/VM release that supports XSTORE for either host or guest use. The IBM z13 and z13s servers will be the last z Systems servers to support XSTORE.
- ▶ *Stabilization of z/VM V6.2 support:* The IBM z13 and z13s servers are planned to be the last z Systems servers supported by z/VM V6.2 and the last z Systems server that will be supported where z/VM V6.2 is running as a guest (second level). This is in conjunction with the statement of direction that the IBM z13 and z13s servers will be the last to support ESA/390 architecture mode, which z/VM V6.2 requires. z/VM V6.2 will continue to be supported until December 31, 2016, as announced in announcement letter # 914-012.
- ▶ *Product Delivery of z/VM on DVD/Electronic only:* z/VM V6.3 is the last release of z/VM that is available on tape. Subsequent releases will be available on DVD or electronically.

7.2.3 z/VSE

Support is provided by z/VSE V5R1 and later. Note the following considerations:

- ▶ z/VSE runs in z/Architecture mode only.
- ▶ z/VSE supports 64-bit real and virtual addressing.
- ▶ z/VSE V5 has an architecture level set (ALS) that requires IBM System z9 or later.
- ▶ z/VSE V6 has an ALS that requires IBM System z10 or later.

For a list of supported functions and their required support levels, see Table 7-4 on page 239.

7.2.4 z/TPF

For a list of supported functions and their required support levels, see Table 7-4 on page 239.

7.2.5 Linux on z Systems

Linux on z Systems distributions are built separately for the 31-bit and 64-bit addressing modes of the z/Architecture. The newer distribution versions are built for 64-bit only. Using the 31-bit emulation layer on a 64-bit Linux on z Systems distribution provides support for running 31-bit applications. None of the current versions of Linux on z Systems distributions (SUSE Linux Enterprise Server 12 and 11, and Red Hat Enterprise Linux (RHEL) 7 and 6) require z13s toleration support. Table 7-2 shows the service levels of SUSE and Red Hat releases supported at the time of writing.

Table 7-2 Current Linux on z Systems distributions

Linux on z Systems distribution	z/Architecture (64-bit mode)
SUSE Linux Enterprise Server 12	Yes
SUSE Linux Enterprise Server 11	Yes
Red Hat RHEL 7	Yes
Red Hat RHEL 6	Yes

For the latest information about supported Linux distributions on z Systems, see this website:

<http://www.ibm.com/systems/z/os/linux/resources/testedplatforms.html>

IBM is working with its Linux distribution partners to provide further use of selected z13s functions in future Linux on z Systems distribution releases.

Consider the following guidelines:

- ▶ Use SUSE Linux Enterprise Server 12 or Red Hat RHEL 7 in any new projects for the z13s.
- ▶ Update any Linux distributions to their latest service level before the migration to z13s.
- ▶ Adjust the capacity of any z/VM and Linux on z Systems LPAR guests, and z/VM guests, in terms of the number of IFLs and CPs, real or virtual, according to the PU capacity of the z13s.

7.2.6 KVM for IBM z Systems

KVM for IBM z Systems (KVM for IBM z) is an open virtualization alternative for z Systems built on Linux and KVM. KVM for IBM z delivers a Linux-familiar administrator experience that can enable simplified virtualization management and operation. See Table F-1 on page 513 for a list of supported features.

7.2.7 z13s function support summary

The following tables summarize the z13s functions and their minimum required operating system support levels:

- ▶ Table 7-3 on page 234 is for z/OS and z/VM.
- ▶ Table 7-4 on page 239 is for z/VSE, z/TPF, and Linux on z Systems.

Information about Linux on z Systems refers exclusively to the appropriate distributions of SUSE and Red Hat.

Both tables use the following conventions:

- Y** The function is supported.
- N** The function is not supported.
- The function is not applicable to that specific operating system.

Although the following tables list all functions that require support, the PTF numbers are not given. Therefore, for the current information, see the PSP bucket for 2965DEVICE.

Table 7-3 shows the minimum support levels for z/OS and z/VM.

Table 7-3 z13s function minimum support requirements summary (part 1 of 2)

Function	z/OS V2 R2	z/OS V2 R1	z/OS V1R13	z/OS V1R12	z/VM V6R3	z/VM V6R2
z13s ^a	Y	Y	Y	Y	Y	Y
Maximum processor unit (PUs) per system image	18 ^b	18 ^b	18 ^b	18 ^b	20 ^c	20
Support of IBM zAware	Y	Y	Y	N	Y ^d	Y ^d
z Systems Integrated Information Processors (zIIPs)	Y	Y	Y	Y	Y	Y
Java Exploitation of Transactional Execution	Y	Y	Y	N	N	N
Large memory support (TB)	4TB	4TB ^e	1 TB	1 TB	1 TB	256 GB
Large page support of 1 MB pageable large pages	Y	Y	Y	Y	N	N
2 GB large page support	Y	Y	Y ^f	N	N	N
Out-of-order execution	Y	Y	Y	Y	Y	Y
Hardware decimal floating point ^g	Y	Y	Y	Y	Y	Y
40 LPARs	Y	Y	Y	N	Y	Y
CPU measurement facility	Y	Y	Y	Y	Y	Y
Enhanced flexibility for Capacity on Demand (CoD)	Y	Y	Y	Y	Y	Y
HiperDispatch	Y	Y	Y	Y	Y	N
Three logical channel subsystems (LCSSs)	Y	Y	Y	N	N	N
Three subchannel sets per LCSS	Y	Y	Y	N	Y	Y
Simultaneous MultiThreading (SMT)	Y	Y	N	N	Y ^h	N
Single-instruction multiple-data (SIMD)	Y	Y	N	N	Y ^h	N
Multi-vSwitch Link Aggregation	Y	N	N	N	Y ^h	N
Cryptography						
CP Assist for Cryptographic Function (CPACF) greater than 16 Domain Support	Y	Y	Y	Y	Y	Y
CPACF AES-128, AES-192, and AES-256	Y	Y	Y	Y	Y	Y
CPACF SHA-1, SHA-224, SHA-256, SHA-384, and SHA-512	Y	Y	Y	Y	Y	Y
CPACF protected key	Y	Y	Y	Y	Y	Y
Secure IBM Enterprise PKCS #11 (EP11) coprocessor mode	Y	Y	Y	Y	Y	Y

Function	z/OS V2 R2	z/OS V2 R1	z/OS V1R13	z/OS V1R12	z/VM V6R3	z/VM V6R2
Crypto Express5S	Y	Y	Y	Y	Y	Y
Elliptic Curve Cryptography (ECC)	Y	Y	Y	Y	Y	Y
HiperSocketsⁱ						
32 HiperSockets	Y	Y	Y	Y	Y	Y
HiperSockets Completion Queue	Y	Y	Y	N	Y	Y
HiperSockets integration with intraensemble data network (IEDN)	Y	Y	Y	N	N	N
HiperSockets Virtual Switch Bridge	-	-	-	-	Y	Y
HiperSockets Network Traffic Analyzer	N	N	N	N	Y	Y
HiperSockets Multiple Write Facility	Y	Y	Y	Y	N	N
HiperSockets support of IPV6	Y	Y	Y	Y	Y	Y
HiperSockets Layer 2 support	Y	Y	Y	Y	Y	Y
HiperSockets	Y	Y	Y	Y	Y	Y
Flash Express Storage						
Flash Express	Y	Y	Y	N	N	N
zEnterprise Data Compression (zEDC)						
zEDC Express	Y	Y	N	N	Y	N
Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE)						
10GbE RoCE Express	Y	Y	Y	Y	Y	N
Shared RoCE environment	Y	Y	N	N	Y	N
Shared Memory Communications - Direct Memory Access (SMC-D)						
SMC-D over ISM (Internal Shared Memory)	Y	N	N	N	Y ^{jh}	N
FICON (Fibre Connection) and FCP (Fibre Channel Protocol)						
FICON Express 8S (channel-path identifier (CHPID) type FC) when using z13s FICON or channel-to-channel (CTC)	Y	Y	Y	Y	Y	Y
FICON Express 8S (CHPID type FC) for support of High Performance FICON for z Systems (zHPF) single-track operations	Y	Y	Y	Y	Y	Y
FICON Express 8S (CHPID type FC) for support of zHPF multitrack operations	Y	Y	Y	Y	Y	Y
FICON Express 8S (CHPID type FCP) for support of SCSI devices	N	N	N	N	Y	Y
FICON Express 8S (CHPID type FCP) support of hardware data router	N	N	N	N	Y	N

Function	z/OS V2 R2	z/OS V2 R1	z/OS V1R13	z/OS V1R12	z/VM V6R3	z/VM V6R2
T10-DIF support by the FICON Express8S and FICON Express8 features when defined as CHPID type FCP	N	N	N	N	Y	Y
Global resource serialization (GRS) FICON CTC toleration	Y	Y	Y	Y	N	N
FICON Express8 CHPID 10KM Long Wave (LX) and Short Wave (SX) type FC	Y	Y	Y	Y	Y	Y
FICON Express 16S (CHPID type FC) when using FICON or CTC	Y	Y	Y	Y	Y	Y
FICON Express 16S (CHPID type FC) for support of zHPF single-track operations	Y	Y	Y	Y	Y	Y
FICON Express 16S (CHPID type FC) for support of zHPF multitrack operations	Y	Y	Y	Y	Y	Y
FICON Express 16S (CHPID type FCP) for support of SCSI devices	Y	N	N	N	Y	Y
FICON Express 16S (CHPID type FCP) support of hardware data router	Y	N	N	N	Y	N
T10-DIF support by the FICON Express16S features when defined as CHPID type FCP	N	N	N	N	Y	Y
Health Check for FICON Dynamic routing	Y	Y	Y	Y	N	N
OSA (Open Systems Adapter)						
OSA-Express5S 10 Gigabit Ethernet Long Reach (LR) and Short Reach (SR) CHPID type OSD	Y	Y	Y	Y	Y	Y
OSA-Express5S 10 Gigabit Ethernet LR and SR CHPID type OSX	Y	Y	Y	Y	N ^k	N ^k
OSA-Express5S Gigabit Ethernet LX and SX CHPID type OSD (two ports per CHPID)	Y	Y	Y	Y	Y	Y
OSA-Express5S Gigabit Ethernet LX and SX CHPID type OSD (one port per CHPID)	Y	Y	Y	Y	Y	Y
OSA-Express5S 1000BASE-T Ethernet CHPID type OSC	Y	Y	Y	Y	Y	Y
OSA-Express5S 1000BASE-T Ethernet CHPID type OSD (two ports per CHPID)	Y	Y	Y	Y	Y	Y

Function	z/OS V2 R2	z/OS V2 R1	z/OS V1R13	z/OS V1R12	z/VM V6R3	z/VM V6R2
OSA-Express5S 1000BASE-T Ethernet CHPID type OSD (one port per CHPID)	Y	Y	Y	Y	Y	Y
OSA-Express5S 1000BASE-T Ethernet CHPID type OSE	Y	Y	Y	Y	Y	Y
OSA-Express5S 1000BASE-T Ethernet CHPID type OSM	Y	Y	Y	Y	N ^k	N ^k
OSA-Express5S 1000BASE-T Ethernet CHPID type OSN	Y	Y	Y	Y	Y	Y
OSA-Express4S 10-Gigabit Ethernet LR and SR CHPID type OSD	Y	Y	Y	Y	Y	Y
OSA-Express4S 10-Gigabit Ethernet LR and SR CHPID type OSX	Y	Y	Y	Y	N ^k	N ^k
OSA-Express4S Gigabit Ethernet LX and SX CHPID type OSD (two ports per CHPID)	Y	Y	Y	Y	Y	Y
OSA-Express4S Gigabit Ethernet LX and SX CHPID type OSD (one port per CHPID)	Y	Y	Y	Y	Y	Y
OSA-Express4S 1000BASE-T CHPID type OSC (one or two ports per CHPID)	Y	Y	Y	Y	Y	Y
OSA-Express4S 1000BASE-T CHPID type OSD (two ports per CHPID)	Y	Y	Y	Y	Y	Y
OSA-Express4S 1000BASE-T CHPID type OSD (one port per CHPID)	Y	Y	Y	Y	Y	Y
OSA-Express4S 1000BASE-T CHPID type OSE (one or two ports per CHPID)	Y	Y	Y	Y	Y	Y
OSA-Express4S 1000BASE-T CHPID type OSM (one port per CHPID)	Y	Y	Y	Y	N ^k	N ^k
OSA-Express4S 1000BASE-T CHPID type OSN	Y	Y	Y	Y	Y	Y
Inbound workload queuing enterprise extender	Y	Y	Y	Y	Y	Y
Checksum offload IPV6 packets	Y	Y	Y	Y	Y	Y
Checksum offload for LPAR-to-LPAR traffic	Y	Y	Y	Y	Y	Y
Large send for IPV6 packets	Y	Y	Y	Y	Y	Y

Function	z/OS V2 R2	z/OS V2 R1	z/OS V1R13	z/OS V1R12	z/VM V6R3	z/VM V6R2
Parallel Sysplex and other						
Server Time Protocol (STP)	Y	Y	Y	Y	-	-
Coupling over InfiniBand CHPID type CIB	Y	Y	Y	Y	N	N
InfiniBand coupling links 12x at a distance of 150 m (492 ft.)	Y	Y	Y	Y	N	N
InfiniBand coupling links 1x at an unrepeated distance of 10 km (6.2 miles)	Y	Y	Y	Y	N	N
CFCC Level 21	Y	Y	Y	Y	Y	Y
CFCC Level 21 Flash Express usage	Y	Y	Y	N	N	N
CFCC Level 21 Coupling Thin Interrupts	Y	Y	Y	Y	N	N
CFCC Level 21 Coupling Large Memory support	Y	Y	Y	Y	N	N
CFCC 21 Support for 256 Coupling CHPIDs	Y	Y	Y	Y	N	N
IBM Integrated Coupling Adapter (ICA)	Y	Y	Y	Y	N	N
Dynamic I/O support for InfiniBand and ICA CHPIDs	-	-	-	-	Y	Y
IBM z/OS Resource Measurement Facility (RMF) coupling channel reporting	Y	Y	Y	Y	N	N

- a. PTFs might be required for toleration support or use of z13s functions and features.
- b. Hardware limit of 6 CP and 2zIIP:1CP ratio.
- c. With SMT, 40 threads.
- d. zAware V2.0, Linux Guest support.
- e. 4 TB of real storage is supported with PTF for APAR OA47439.
- f. PTF support required and with RSM enabled web delivery.
- g. Packed decimal conversion support.
- h. PTF support required.
- i. On z13s servers, the CHPID statement of HiperSockets devices requires the keyword VCHID, so the z13s IOCP definitions need to be migrated to support the HiperSockets definitions (CHPID type IQD). VCHID specifies the virtual channel identification number associated with the channel path. Valid range is 7E0 - 7FF. VCHID is not valid on z Systems before z13s servers.
- j. For guest exploitation only.
- k. Dynamic I/O support only.

Table 7-4 shows the minimum support levels for z/VSE, z/TPF, and Linux on z Systems.

Table 7-4 z13 functions minimum support requirements summary (part 2 of 2)

Function	z/VSE V6R1	z/VSE V5R2	z/VSE V5R1	z/TPF V1R1	Linux on z Systems
z13s	Y	Y ^a	Y ^a	Y	Y
Maximum PUs per system image	6	6	6	20	20 ^b
Support of IBM zAware	-	-	-	-	Y
IBM z Integrated Information Processors (zIIPs)	-	-	-	-	-
Java Exploitation of Transactional Execution	N	N	N	N	Y
Large memory support ^c	32 GB	32 GB	32 GB	4 TB	4 TB ^d
Large page (1 MB) support	Y	Y	Y	N	Y
Pageable 1 MB page support	N	N	N	N	Y
2 GB Large Page Support	-	-	-	-	-
Out-of-order execution	Y	Y	Y	Y	Y
40 logical partitions	Y	Y	Y	Y	Y
HiperDispatch	N	N	N	N	N
Three logical channel subsystems (LCSSs)	N	N	N	N	Y
Three subchannel set per LCSS	Y	Y	Y	N	Y
Simultaneous multithreading (SMT)	N	N	N	N	Y
Single instruction, multiple data (SIMD)	N	N	N	N	N
Multi-vSwitch Link Aggregation	N	N	N	N	N
Cryptography					
CP Assist for Cryptographic Function (CPACF)	Y	Y	Y	Y	Y
CPACF AES-128, AES-192, and AES-256	Y	Y	Y	Y ^e	Y
CPACF SHA-1/SHA-2, SHA-224, SHA-256, SHA-384, and SHA-512	Y	Y	Y	Y ^f	Y
CPACF protected key	N	N	N	N	N
Secure IBM Enterprise PKCS #11 (EP11) coprocessor mode	N	N	N	N	N
Crypto Express5S	Y	Y	Y	Y ^{gh}	Y
Elliptic Curve Cryptography (ECC)	N	N	N	N	N ^k
HiperSocketsⁱ					
32 HiperSockets	Y	Y	Y	Y	Y

Function	z/VSE V6R1	z/VSE V5R2	z/VSE V5R1	z/TPF V1R1	Linux on z Systems
HiperSockets Completion Queue	Y	Y	Y ^g	N	Y
HiperSockets integration with IEDN	N	N	N	N	N
HiperSockets Virtual Switch Bridge	-	-	-	-	Y ^j
HiperSockets Network Traffic Analyzer	N	N	N	N	Y ^k
HiperSockets Multiple Write Facility	N	N	N	N	N
HiperSockets support of IPV6	Y	Y	Y	N	Y
HiperSockets Layer 2 support	N	N	N	N	Y
HiperSockets	Y	Y	Y	N	Y
Flash Express Storage					
Flash Express	N	N	N	N	Y
zEnterprise Data Compression (zEDC)					
zEDC Express	N	N	N	N	N
Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE)					
10GbE RoCE Express	N	N	N	N	N ^k
Shared Memory Communications - Direct (SMC-D)					
SMC-D over Internal Shared Memory (ISM)	N	N	N	N	N
FICON and FCP					
FICON Express8S support of zHPF single track CHPID type FC	N	N	N	N	Y
FICON Express8 support of zHPF multitrack CHPID type FC	N	N	N	N	Y
High Performance FICON for z Systems (zHPF)	N	N	N	N	Y ^l
GRS FICON CTC toleration	-	-	-	-	-
N_Port ID Virtualization (NPIV) for FICON CHPID type FCP	Y	Y	Y	N	Y
FICON Express8S support of hardware data router CHPID type FCP	N	N	N	N	Y ^m
FICON Express8S and FICON Express8 and FICON Express8S support of T10-DIF CHPID type FCP	N	N	N	N	Y ^l

Function	z/VSE V6R1	z/VSE V5R2	z/VSE V5R1	z/TPF V1R1	Linux on z Systems
FICON Express8S, FICON Express8, FICON Express16S 10KM LX, and FICON Express4 SX support of SCSI disks CHPID type FCP	Y	Y	Y	N	Y
FICON Express8S CHPID type FC	Y	Y	Y	Y	Y
FICON Express8 CHPID type FC	Y	Y	Y	Y ⁿ	Y ⁿ
FICON Express 16S (CHPID type FC) when using FICON or CTC	Y	Y	Y	Y	Y
FICON Express 16S (CHPID type FC) for support of zHPF single-track operations	N	N	N	N	Y
FICON Express 16S (CHPID type FC) for support of zHPF multitrack operations	N	N	N	N	Y
FICON Express 16S (CHPID type FCP) for support of SCSI devices	Y	Y	Y	N	Y
FICON Express 16S (CHPID type FCP) support of hardware data router	N	N	N	N	Y
T10-DIF support by the FICON Express16S features when defined as CHPID type FCP	N	N	N	N	Y
OSA					
Large send for IPv6 packets	-	-	-	-	-
Inbound workload queuing for Enterprise Extender	N	N	N	N	N
Checksum offload for IPV6 packets	N	N	N	N	N
Checksum offload for LPAR-to-LPAR traffic	N	N	N	N	N
OSA-Express5S 10 Gigabit Ethernet LR and SR CHPID type OSD	Y	Y	Y	Y ^o	Y
OSA-Express5S 10 Gigabit Ethernet LR and SR CHPID type OSX	Y	Y	Y	Y ^p	Y ^q
OSA-Express5S Gigabit Ethernet LX and SX CHPID type OSD (two port per CHPID)	Y	Y	Y	Y ^o	Y ^r
OSA-Express5S Gigabit Ethernet LX and SX CHPID type OSD (one port per CHPID)	Y	Y	Y	Y ^o	Y
OSA-Express5S 1000BASE-T Ethernet CHPID type OSC	Y	Y	Y	Y	-

Function	z/VSE V6R1	z/VSE V5R2	z/VSE V5R1	z/TPF V1R1	Linux on z Systems
OSA-Express5S 1000BASE-T Ethernet CHPID type OSD (two ports per CHPID)	Y	Y	Y	Y ^o	Y ^r
OSA-Express5S 1000BASE-T Ethernet CHPID type OSD (one port per CHPID)	Y	Y	Y	Y ^o	Y
OSA-Express5S 1000BASE-T Ethernet CHPID type OSE	Y	Y	Y	N	N
OSA-Express5S 1000BASE-T Ethernet CHPID type OSM	N	N	N	N	Y ^s
OSA-Express5S 1000BASE-T Ethernet CHPID type OSN	Y	Y	Y	Y	Y
OSA-Express4S 10-Gigabit Ethernet LR and SR CHPID type OSD	Y	Y	Y	Y	Y
OSA-Express4S 10-Gigabit Ethernet LR and SR CHPID type OSX	Y	Y	Y	Y ^t	Y
OSA-Express4S Gigabit Ethernet LX and SX CHPID type OSD (two ports per CHPID)	Y	Y	Y	Y ^t	Y
OSA-Express4S Gigabit Ethernet LX and SX CHPID type OSD (one port per CHPID)	Y	Y	Y	Y	Y
OSA-Express4S 1000BASE-T CHPID type OSC (one or two ports per CHPID)	Y	Y	Y	N	-
OSA-Express4S 1000BASE-T CHPID type OSD (two ports per CHPID)	Y	Y	Y	Y ^g	Y
OSA-Express4S 1000BASE-T CHPID type OSD (one port per CHPID)	Y	Y	Y	Y	Y
OSA-Express4S 1000BASE-T CHPID type OSE (one or two ports per CHPID)	Y	Y	Y	N	N
OSA-Express4S 1000BASE-T CHPID type OSM	N	N	N	N	Y
OSA-Express4S 1000BASE-T CHPID type OSN (one or two ports per CHPID)	Y	Y	Y	Y ^g	Y
Parallel Sysplex and other					
STP enhancements	-	-	-	-	Y
Server Time Protocol (STP)	-	-	-	Y ^u	Y
Coupling over InfiniBand CHPID type CIB	-	-	-	Y	-

Function	z/VSE V6R1	z/VSE V5R2	z/VSE V5R1	z/TPF V1R1	Linux on z Systems
InfiniBand coupling links 12x at a distance of 150 m (492 ft.)	-	-	-	-	-
InfiniBand coupling links 1x at unrepeatd distance of 10 km (6.2 miles)	-	-	-	-	-
Dynamic I/O support for InfiniBand CHPIDs	-	-	-	-	-
CFCC Level 21	-	-	-	Y	-

- a. Toleration support PTFs are required.
- b. SUSE Linux Enterprise Server 12 and RHEL can support up to 256 PUs (IFLs or CPs).
- c. 4 TB of real storage is supported per server.
- d. Red Hat (RHEL) supports a maximum of 3 TB.
- e. z/TPF supports only AES-128 and AES-256.
- f. z/TPF supports only SHA-1 and SHA-256.
- g. Service is required.
- h. Supported only when running in accelerator mode.
- i. On z13s servers, the CHPID statement of HiperSockets devices requires the keyword VCHID. The z13s IOCP definitions therefore must be migrated to support the HiperSockets definitions (CHPID type IQD). VCHID specifies the virtual channel identification number associated with the channel path. Valid range is 7E0 - 7FF. VCHID is not valid on z Systems before z13s servers.
- j. Applicable to guest operating systems.
- k. IBM is working with its Linux distribution partners to include support in future Linux on z Systems distribution releases. Check the Linux distribution partners for the latest information.
- l. Supported by SUSE Linux Enterprise Server 11.
- m. Supported by SUSE Linux Enterprise Server 11 SP3 and RHEL 6.4.
- n. For more information, see 7.3.39, "FCP provides increased performance" on page 267.
- o. Requires PUT 5 with program temporary fixes (PTFs).
- p. Requires PUT 8 with PTFs.
- q. Supported by SUSE Linux Enterprise Server 11 SP1, SUSE Linux Enterprise Server 10 SP4, and RHEL 6, RHEL 5.6.
- r. Supported by SUSE Linux Enterprise Server 11, SUSE Linux Enterprise Server 10 SP2, RHEL 6, and RHEL 5.2.
- s. Supported by SUSE Linux Enterprise Server 11 SP2, SUSE Linux Enterprise Server 10 SP4, RHEL 6, and RHEL 5.2.
- t. Requires PUT 4 with PTFs.
- u. SSTP is supported in z/TPF with APAR PJ36831 in PUT 07.

7.3 Support by function

This section addresses operating system support by function. Only the currently in-support releases are covered.

Tables in this section use the following convention:

N/A Not applicable
NA Not available

7.3.1 Single system image

A single system image can control several processor units, such as CPs, zIIPs, or IFLs.

Maximum number of PUs per system image

Table 7-5 lists the maximum number of PUs supported by each OS image and by special purpose LPARs. On the z13s, the image size is restricted by the number of PUs available:

- ▶ Maximum of 6 CPs or 6 CPs with 12 zIIPs
- ▶ Maximum of 20 IFLs or Internal Coupling Facilities (ICFs)

Table 7-5 Single system image size software support

Operating system	Maximum number of PUs per system image ^a
z/OS V2R2	141 ^{bc}
z/OS V2R1	141 ^{bc}
z/OS V1R13	100 ^c .
z/OS V1R12	100 ^b .
z/VM V6R3	64 ^d .
z/VM V6R2	32.
z/VSE V5R1 and later	z/VSE Turbo Dispatcher can exploit up to 4 CPs, and tolerates up to 10-way LPARs.
z/TPF V1R1	86 CPs.
CFCC Level 21	16 CPs or ICFs: CPs and ICFs cannot be mixed.
IBM zAware	80
Linux on z Systems ^e	SUSE Linux Enterprise Server 12: 256 CPs or IFLs. SUSE Linux Enterprise Server 11: 64 CPs or IFLs. Red Hat RHEL 7: 256 CPs or IFLs. Red Hat RHEL 6 64 CPs or IFLs.
KVM for IBM z Systems	141 ^c

a. z13s servers have 20 characterizable PUs.

b. 128 PUs in multithreading mode and 141 PUs supported without multithreading.

c. Total characterizable PUs including zIIPs and CPs on z13s servers.

d. 64 PUs without SMT mode and 32 PUs with SMT.

e. IBM is working with its Linux distribution partners to provide the use of this function in future Linux on z Systems distribution releases.

The z Appliance Container Infrastructure- mode LPAR

zEC12 introduced an LPAR mode, called zAware-mode, that is exclusively for running the IBM zAware virtual appliance.

IBM z Appliance Container Infrastructure (zACI) is a new partition type that, along with an appliance installer, enables the secure deployment of software and firmware appliances. zACI will shorten the deployment and implementation of firmware solutions or software solutions delivered as virtual software appliances.

The zACI framework enforces a common set of standards and behaviors, and a new zACI partition mode for a virtual appliance that requires a new zACI LPAR type. The IBM zAware partition mode has been renamed to zACI, and the IBM zAware firmware will now run in this partition. There are no hardware dependencies. zACI will be delivered as part of the base code on each z13s and z13 (driver level 27) server.

The IBM zAware virtual appliance can pinpoint deviations in z/OS normal system behavior. It also improves real-time event diagnostic tests by monitoring the z/OS operations log

(OPERLOG). It looks for unusual messages, unusual message patterns that typical monitoring systems miss, and unique messages that might indicate system health issues. The IBM zAware virtual appliance requires the monitored clients to run z/OS V1R13 with PTFs or later. The newer version of IBM zAware is enhanced to work with messages without message IDs. This includes support for Linux running natively or as a guest under z/VM on z Systems.

For more information about zAware, see Appendix B, “IBM z Systems Advanced Workload Analysis Reporter (IBM zAware)” on page 453.

The z/VM-mode LPAR

z13s supports an LPAR mode, called *z/VM-mode*, that is exclusively for running z/VM as the first-level operating system. The z/VM-mode requires z/VM V6R2 or later, and allows z/VM to use a wider variety of specialty processors in a single LPAR. For example, in a z/VM-mode LPAR, z/VM can manage Linux on z Systems guests running on IFL processors while also managing z/VSE and z/OS guests on CPs. It also allows z/OS to fully use zIIPs.

IBM Dynamic Partition Manager

A new administrative mode is being introduced for Linux only CPCs for z13 and z13s servers with SCSI storage attached through FCP channels. IBM Dynamic Partition Manager (DPM) provides simplified z Systems hardware and virtual infrastructure management including integrated dynamic I/O management for users who intend to run KVM for IBM z Systems as hypervisor or Linux on z Systems running in a partition (LPAR). The new mode, DPM, provides simplified, consumable, and enhanced partition lifecycle and dynamic I/O management capabilities by using the Hardware Management Console (HMC).

For more information, see Appendix E, “IBM Dynamic Partition Manager” on page 501.

7.3.2 zIIP support

zIIPs do not change the model capacity identifier of the z13s. IBM software product license charges based on the model capacity identifier are not affected by the addition of zIIPs. On a z13s, z/OS Version 1 Release 12 is the minimum level for supporting zIIPs.

No changes to applications are required to use zIIPs. It can be used by these applications:

- ▶ DB2 V8 and later for z/OS data serving, for applications that use data Distributed Relational Database Architecture (DRDA) over TCP/IP, such as data serving, data warehousing, and selected utilities.
- ▶ z/OS XML services.
- ▶ z/OS CIM Server.
- ▶ z/OS Communications Server for network encryption (Internet Protocol Security (IPSec)) and for large messages that are sent by HiperSockets.
- ▶ IBM GBS Scalable Architecture for Financial Reporting.
- ▶ IBM z/OS Global Mirror (formerly XRC) and System Data Mover.
- ▶ IBM OMEGAMON® XE on z/OS, OMEGAMON XE on DB2 Performance Expert, and DB2 Performance Monitor.
- ▶ Any Java application that is using the current IBM SDK.
- ▶ WebSphere Application Server V5R1 and later, and products that are based on it, such as WebSphere Portal, WebSphere Enterprise Service Bus (WebSphere ESB), and WebSphere Business Integration for z/OS.

- ▶ CICS/TS V2R3 and later.
- ▶ DB2 UDB for z/OS Version 8 and later.
- ▶ IMS Version 8 and later.
- ▶ zIIP Assisted HiperSockets for large messages.

The functioning of a zIIP is transparent to application programs.

In z13s, the zIIP processor is designed to run in SMT mode, with up to two threads per processor. This new function is designed to help improve throughput for zIIP workloads and provide appropriate performance measurement, capacity planning, and SMF accounting data. This support is available for z/OS V2.1 with PTFs or later at z13s general availability.

Use the **PROJECTCPU** option of the IEAOPTxx parmlib member to help determine whether zIIPs can be beneficial to the installation. Setting PROJECTCPU=YES directs z/OS to record the amount of eligible work for zIIPs in SMF record type 72 subtype 3. The field APPL% IIPCP of the Workload Activity Report listing by WLM service class indicates the percentage of a processor that is zIIP eligible. Because of the zIIP lower price as compared to a CP, a utilization as low as 10% might provide benefits.

7.3.3 Transactional Execution

The IBM zEnterprise EC12 introduced an architectural feature called *Transactional Execution* (TX). This capability is known in academia and industry as *hardware transactional memory*. Transactional execution is also implemented in z13, z13s, and zBC12 servers.

This feature enables software to indicate to the hardware the beginning and end of a group of instructions that need to be treated in an atomic way. Either all of their results happen or none happens, in true transactional style. The execution is optimistic. The hardware provides a memory area to record the original contents of affected registers and memory as the instruction's execution takes place. If the transactional execution group is canceled or must be rolled back, the hardware transactional memory is used to reset the values. Software can implement a fallback capability.

This capability enables more efficient software by providing a way to avoid locks (*lock elision*). This advantage is of special importance for speculative code generation and highly parallelized applications.

TX is used by IBM JVM, but potentially can be used by other software. z/OS V1R13 with PTFs or later is required. The feature also is enabled for the Linux distributions SUSE Linux Enterprise Server 11 SP3, RHEL6.4, SUSE Linux Enterprise Server 12, and RHEL7.

7.3.4 Maximum main storage size

Table 7-6 lists the maximum amount of main storage that is supported by current operating systems. A maximum of 4 TB of main storage can be defined for an LPAR on a z13s. If an I/O drawer is present (as carry-forward), the LPAR maximum memory is limited to 1 TB. Expanded storage, although part of the z/Architecture, is used only by z/VM.

Table 7-6 Maximum memory that is supported by the operating system

Operating system	Maximum supported main storage ^a
z/OS	z/OS V2R1 and later support 4 TB ^a .
z/VM	z/VM V6R3 supports 1 TB. z/VM V6R2 supports 256 GB
z/VSE	z/VSE V5R1 and later support 32 GB.
z/TPF	z/TPF supports 4 TB. ^a
CFCC	Level 21 supports up to 3 TB per server. ^a
zACI	Supports up to 3 TB per server. ^a
Linux on z Systems (64-bit)	SUSE Linux Enterprise Server 12 supports 4 TB. ^a SUSE Linux Enterprise Server 11 supports 4 TB. ^a SUSE Linux Enterprise Server 10 supports 4 TB. ^a Red Hat RHEL 7 supports 3 TB. ^a Red Hat RHEL 6 supports 3 TB. ^a Red Hat RHEL 5 supports 3 TB. ^a

a. z13s supports 4 TB user configurable memory per server.

7.3.5 Flash Express

IBM z13s continues support for *Flash Express*, which can help improve the resilience and performance of the z/OS system. Flash Express is designed to assist with the handling of workload spikes or increased workload demand that might occur at the opening of the business day, or in a workload shift from one system to another.

z/OS is the first OS to use Flash Express storage as storage-class memory (SCM) for paging store and supervisor call (SAN Volume Controller) memory dumps. Flash memory is a faster paging device compared to a hard disk drive (HDD). SAN Volume Controller memory dump data capture time is expected to be substantially reduced. As a paging store, Flash Express storage is suitable for workloads that can tolerate paging. It does not benefit workloads that cannot afford to page. The z/OS design for Flash Express storage does not completely remove the virtual storage constraints that are created by a paging spike in the system. However, some scalability relief is expected because of faster paging I/O with Flash Express storage.

Flash Express storage is allocated to an LPAR similarly to main memory. The initial and maximum amount of Flash Express Storage that is available to a particular LPAR is specified at the Support Element (SE) or Hardware Management Console (HMC) by using a new Flash Storage Allocation window.

The Flash Express storage granularity is 16 GB. The amount of Flash Express storage in the partition can be changed dynamically between the initial and the maximum amount at the SE or HMC. For z/OS, this change also can be made by using an operator command. Each

partition's Flash Express storage is isolated like the main storage, and sees only its own space in the flash storage space.

Flash Express provides 1.4 TB of storage per feature pair. Up to four pairs can be installed, for a total of 5.6 TB. All paging data can easily be on Flash Express storage, but not all types of page data can be on it. For example, virtual I/O (VIO) data always is placed on an external disk. Local page data sets are still required to support peak paging demands that require more capacity than provided by the amount of configured SCM.

The z/OS paging subsystem works with a mix of internal Flash Express storage and external disk. The placement of data on Flash Express storage and external disk is self-tuning, based on measured performance. At IPL time, z/OS detects whether Flash Express storage is assigned to the partition. z/OS automatically uses Flash Express storage for paging unless specified otherwise by using PAGESCM=NONE in IEASYSxx. No definition is required for placement of data on Flash Express storage.

The support is delivered in the z/OS V1R13 real storage manager (RSM) Enablement Offering Web Deliverable (FMID JBB778H) for z/OS V1R13. Dynamic reconfiguration support for SCM is available as a web deliverable. The installation of this web deliverable requires careful planning because the size of the Nucleus, extended system queue area (ESQA) per CPU, and RSM stack is increased. Also, a new memory pool for pageable large pages is added. For web-deliverable code on z/OS, see the z/OS downloads website:

<http://www.ibm.com/systems/z/os/zos/downloads/>

The support also is delivered in z/OS V2R1 (included with the base product) or later.

Linux on z Systems also offers the support for Flash Express as an SCM device. This support is useful for workloads with large write operations with a block size of 256 KB or more. The SCM increments are accessed through extended asynchronous data mover (EADM) subchannels.

Table 7-7 lists the minimum support requirements for Flash Express.

Table 7-7 Minimum support requirements for Flash Express

Operating system	Support requirements
z/OS	z/OS V1R13 ^a
Linux on z Systems	SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 SP3 RHEL6 RHEL7

a. Web deliverable and PTFs are required.

Flash Express usage by CFCC

Coupling facility control code (CFCC) Level 21 supports Flash Express. Initial CF Flash usage is targeted for WebSphere MQ shared queues application structures. Structures can now be allocated with a combination of real memory and SCM that is provided by the Flash Express feature. For more information, see “Flash Express exploitation by CFCC” on page 294.

Flash Express usage by Java

z/OS Java SDK 7 SR3, CICS TS V5.1, WebSphere Liberty Profile 8.5, DB2 V11, and IMS V12 are targeted for Flash Express usage. There is a statement of direction to support traditional WebSphere V8. The support is for just-in-time (JIT) Code Cache and Java Heap to improve performance for pageable large pages.

7.3.6 z Enterprise Data Compression Express

The growth of data that must be captured, transferred, and stored for a long time is unrelenting. Software-implemented compression algorithms are costly in terms of processor resources, and storage costs are not negligible either.

zEDC is an optional feature that is available to z13, z13s, zEC12, and zBC12 servers that addresses those requirements by providing hardware-based acceleration for data compression and decompression. zEDC provides data compression with lower CPU consumption than the compression technology that previously was available on z Systems.

Exploitation support of zEDC Express functions is provided by z/OS V2R1zEnterprise Data Compression or higher for both data compression and decompression.

Support for data recovery (decompression) when the zEDC is not installed, or installed but not available on the system, is provided through software on z/OS V2R2, z/OS V2R1, V1R13, and V1R12 with the correct PTFs. Software decompression is slow and uses considerable processor resources. It is therefore not suggested for production environments.

zEDC is enhanced to support QSAM/BSAM (non-VSAM) data set compression. This support can be achieved by any of the following ways

- ▶ Data class level: Two new values, zEDC Required (ZR) and zEDC Preferred (ZP), can be set with the **COMPACTION** option in the data class.
- ▶ System Level: Two new values, zEDC Required (ZEDC_R) and zEDC Preferred (ZEDC_P), can be specified with the **COMPRESS** parameter found in the IGDSMSXX member of the SYS1.PARMLIB data sets.

Data class takes precedence over system level.

Table 7-8 shows the minimum support requirements for zEDC Express.

Table 7-8 Minimum support requirements for zEDC Express

Operating system	Support requirements
z/OS	z/OS V2R1 ^a z/OS V1R13 ^a (Software decompression support only) z/OS V1R12 ^a (Software decompression support only)
z/VM	z/VM 6.3 ^b

a. PTFs are required.

b. For guest use. PTF support required.

For more information about zEDC Express, see Appendix J, “IBM zEnterprise Data Compression Express” on page 551.

7.3.7 10GbE RoCE Express

The IBM z13s supports the RoCE Express feature. It extends this support by providing support to the second port on the adapter and by sharing the ports to up to 31 partitions, per adapter, using both ports.

The 10 Gigabit Ethernet (10GbE) RoCE Express feature is designed to help reduce consumption of CPU resources for applications that use the TCP/IP stack (such as WebSphere accessing a DB2 database). Use of the 10GbE RoCE Express feature also can help reduce network latency with memory-to-memory transfers using Shared Memory

Communications over Remote Direct Memory Access (SMC-R) in z/OS V2R1 or later. It is transparent to applications and can be used for LPAR-to-LPAR communication on a single z13s or for server-to-server communication in a multiple CPC environment.

z/OS V2R1 or later with PTFs provides support for the SMC-R protocol. It does not roll back to previous z/OS releases. z/OS V1R12 and z/OS V1R13 with PTFs provide only compatibility support.

IBM is working with its Linux distribution partners to include support in future Linux on z Systems distribution releases.

Table 7-9 lists the minimum support requirements for 10GbE RoCE Express.

Table 7-9 Minimum support requirements for RoCE Express

Operating system	Support requirements
z/OS	z/OS V2R1 with supporting PTFs (SPE for IBM VTAM®, TCP/IP, and IOS). The IOS PTF is a minor change to allow greater than 256 (xFF) PFIDs for RoCE.
z/VM	z/VM V6.3 with supporting PTFs. The z/VM V6.3 SPE for base PCIe support is required. When running z/OS as a guest z/OS, APAR OA43256 is required for RoCE, and APARs OA43256 and OA44482 are required for zEDC. A z/VM web page that details the prerequisites for using RoCE and zEDC in a guest can be found at following location: http://www.vm.ibm.com/zvm630/apars.html
Linux on z Systems	Currently limited to experimental support in these operating systems: <ul style="list-style-type: none"> ▶ SUSE Linux Enterprise Server 12 SP1^a ▶ SUSE Linux Enterprise Server 11 SP3 with the latest maintenance. ▶ RHEL 7.0.

a. See SUSE website: https://www.suse.com/releasenotes/x86_64/SUSE-SLES/12-SP1/

For more information, see Appendix D, “Shared Memory Communications” on page 475.

7.3.8 Large page support

In addition to the existing 1-MB large pages, 4-KB pages, and page frames, z13s supports pageable 1-MB large pages, large pages that are 2 GB, and large page frames. For more information, see “Large page support” on page 114.

Table 7-10 lists the support requirements for 1-MB large pages.

Table 7-10 Minimum support requirements for 1-MB large page

Operating system	Support requirements
z/OS	z/OS V1R11 z/OS V1R13 ^a for <i>pageable</i> 1-MB large pages
z/VM	Not supported, and not available to guests
z/VSE	z/VSE V5R1: Supported for data spaces
Linux on z Systems	SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 Red Hat RHEL 7 Red Hat RHEL 6

a. Web deliverable and PTFs are required, plus the Flash Express hardware feature.

Table 7-11 lists the support requirements for 2-GB large pages.

Table 7-11 Minimum support requirements for 2-GB large pages

Operating system	Support requirements
z/OS	z/OS V1R13

7.3.9 Hardware decimal floating point

Industry support for decimal floating point is growing, with IBM leading the open standard definition. Examples of support for the draft standard IEEE 754r include Java BigDecimal, C#, XML, C/C++, GCC, COBOL, and other key software vendors, such as Microsoft and SAP.

Decimal floating point support was introduced with z9 EC. z13s inherited the decimal floating point accelerator feature that was introduced with z10 EC. For more information, see 3.4.6, “Decimal floating point accelerator” on page 96.

Table 7-12 lists the operating system support for decimal floating point. For more information, see 7.6.6, “Decimal floating point and z/OS XL C/C++ considerations” on page 291.

Table 7-12 Minimum support requirements for decimal floating point

Operating system	Support requirements
z/OS	z/OS V1R12.
z/VM	z/VM V6R2: Support is for guest use.
Linux on z Systems	SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 Red Hat RHEL 7 Red Hat RHEL 6

7.3.10 Up to 40 LPARs

This feature, first made available in z13s, allows the system to be configured with up to 40 LPARs. Because channel subsystems can be shared by up to 15 LPARs, you must configure three channel subsystems to reach the 40 LPARs limit. Table 7-13 lists the minimum operating system levels for supporting 40 LPARs.

Table 7-13 Minimum support requirements for 40 LPARs

Operating system	Support requirements
z/OS	z/OS V1R13
z/VM	z/VM V6R2
z/VSE	z/VSE V5R1
z/TPF	z/TPF V1R1
Linux on z Systems	SUSE Linux Enterprise Server 11 SUSE Linux Enterprise Server 12 Red Hat RHEL 7 Red Hat RHEL 6

Remember: The IBM zAware virtual appliance runs in a dedicated LPAR, so when it is activated, it reduces the maximum number of available LPARs by one.

7.3.11 Separate LPAR management of PUs

The z13s uses separate PU pools for each optional PU type. The separate management of PU types enhances and simplifies capacity planning and management of the configured LPARs and their associated processor resources. Table 7-14 lists the support requirements for the separate LPAR management of PU pools.

Table 7-14 Minimum support requirements for separate LPAR management of PUs

Operating system	Support requirements
z/OS	z/OS V1R12
z/VM	z/VM V6R2
z/VSE	z/VSE V5R1
z/TPF	z/TPF V1R1
Linux on z Systems	SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 Red Hat RHEL 7 Red Hat RHEL 6

7.3.12 Dynamic LPAR memory upgrade

An LPAR can be defined with both an initial and a reserved amount of memory. At activation time, the initial amount is made available to the partition and the reserved amount can be added later, partially or totally. Those two memory zones do not have to be contiguous in real memory, but appear as logically contiguous to the operating system that runs in the LPAR.

z/OS can take advantage of this support and nondisruptively acquire and release memory from the reserved area. z/VM V6R2 and later can acquire memory nondisruptively, and immediately make it available to guests. z/VM virtualizes this support to its guests, which can also increase their memory nondisruptively if supported by the guest operating system. Releasing memory from z/VM is not supported. Releasing memory from the z/VM guest depends on the guest's operating system support.

Linux on z Systems also supports both acquiring and releasing memory nondisruptively. This feature is enabled for SUSE Linux Enterprise Server 11 and RHEL 6.

Dynamic LPAR memory upgrade is not supported for zACI-mode LPARs.

7.3.13 LPAR physical capacity limit enforcement

On the IBM z13s, PR/SM is enhanced to support an option to limit the amount of physical processor capacity that is consumed by an individual LPAR when a PU that is defined as a central processor (CP) or an IFL is shared across a set of LPARs. This enhancement is designed to provide a physical capacity limit that is enforced as an absolute (versus a relative) limit. It is not affected by changes to the logical or physical configuration of the system. This physical capacity limit can be specified in units of CPs or IFLs.

Table 7-15 lists the minimum operating system level that is required on z13s.

Table 7-15 Minimum support requirements for LPAR physical capacity limit enforcement

Operating system	Support requirements
z/OS	z/OS V1R12 ^a
z/VM	z/VM V6R3
z/VSE	z/VSE V5R1 ^a

a. PTFs are required.

7.3.14 Capacity Provisioning Manager

The provisioning architecture enables clients to better control the configuration and activation of the On/Off CoD. The new process is inherently more flexible, and can be automated. This capability can result in easier, faster, and more reliable management of the processing capacity. For more information, see 8.8, “Nondisruptive upgrades” on page 349.

The Capacity Provisioning Manager, a function that is first available with z/OS V1R9, interfaces with z/OS Workload Manager (WLM) and implements capacity provisioning policies. Several implementation options are available, from an analysis mode that issues only guidelines, to an autonomic mode that provides fully automated operations.

Replacing manual monitoring with autonomic management or supporting manual operation with guidelines can help ensure that sufficient processing power is available with the least possible delay. Support requirements are listed in Table 7-16.

Table 7-16 Minimum support requirements for capacity provisioning

Operating system	Support requirements
z/OS	z/OS V1R12
z/VM	Not supported: Not available to guests

7.3.15 Dynamic PU add

Planning of an LPAR configuration allows defining reserved PUs that can be brought online when extra capacity is needed. Operating system support is required to use this capability without an IPL, that is, nondisruptively. This support has been in z/OS for a long time.

The dynamic PU add function enhances this support by allowing you to define and dynamically change the number and type of reserved PUs in an LPAR profile, removing any planning requirements. Table 7-17 lists the minimum required operating system levels to support this function.

Table 7-17 Minimum support requirements for dynamic PU add

Operating system	Support requirements
z/OS	z/OS V1R12
z/VM	z/VM V6R2
z/VSE	z/VSE V5R1

Operating system	Support requirements
Linux on z Systems	SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 Red Hat RHEL 7 Red Hat RHEL 6

The new resources are immediately made available to the operating system and, in the case of z/VM, to its guests. The dynamic PU add function is not supported for zAware-mode LPARs.

7.3.16 HiperDispatch

The **HIPERDISPATCH=YES/NO** parameter in the IEAOPTxx member of SYS1.PARMLIB and on the **SET OPT=xx** command controls whether HiperDispatch is enabled or disabled for a z/OS image. It can be changed dynamically, without an IPL or any outage.

The default is that HiperDispatch is disabled on all releases, from z/OS V1R10 (requires PTFs for zIIP support) through z/OS V1R12.

Beginning with z/OS V1R13, when running on a z13, z13s, zEC12, zBC12, z196, or z114 server, the IEAOPTxx keyword **HIPERDISPATCH** defaults to YES. If **HIPERDISPATCH=NO** is specified, the specification is accepted as it was on previous z/OS releases.

The use of SMT in the z13s requires that HiperDispatch is enabled on the operating system (z/OS V2R1/z/OS V2R2 or z/VM V6R3).

Additionally, with z/OS V1R12 or later, any LPAR running with more than 64 logical processors is required to operate in HiperDispatch Management Mode.

The following rules control this environment:

- ▶ If an LPAR is defined at IPL with more than 64 logical processors, the LPAR automatically operates in HiperDispatch Management Mode, regardless of the **HIPERDISPATCH=** specification.
- ▶ If more logical processors are added to an LPAR that has 64 or fewer logical processors and the additional logical processors raise the number of logical processors to more than 64, the LPAR automatically operates in HiperDispatch Management Mode regardless of the **HIPERDISPATCH=YES/NO** specification. That is, even if the LPAR has the **HIPERDISPATCH=NO** specification, that LPAR is converted to operate in HiperDispatch Management Mode.
- ▶ An LPAR with more than 64 logical processors running in HiperDispatch Management Mode cannot be reverted to run in non-HiperDispatch Management Mode.

HiperDispatch in the z13s uses a new chip and CPC drawer configuration to improve the access cache performance. Beginning with z/OS V1R13, HiperDispatch changed to use the new node cache structure of z13s. The base support is provided by PTFs identified by `IBM.device.server.z13s-2965.requireservice`.

The PR/SM in the System z9 EC to zEC12 servers stripes the memory across all books in the system to take advantage of the fast book interconnection and spread memory controller work. The PR/SM in the z13s seeks to assign all memory in one drawer striped across the two nodes to take advantage of the lower latency memory access in a drawer and smooth performance variability across nodes in the drawer.

The PR/SM in the System z9 EC to zEC12 servers attempts to assign all logical processors to one book, packed into PU chips of that book in cooperation with operating system HiperDispatch optimize shared cache usage. The PR/SM in the z13s seeks to assign all logical processors of a partition to one CPC drawer, packed into PU chips of that CPC drawer in cooperation with the operating system HiperDispatch optimize shared cache usage.

The PR/SM automatically keeps partition's memory and logical processors on the same CPC drawer. This arrangement looks simple for a partition, but it is a complex optimization for multiple logical partitions because some must be split among processors drawers.

To use HiperDispatch effectively, WLM goal adjustment might be required. Review the WLM policies and goals, and update them as necessary. You might want to run with the new policies and HiperDispatch on for a period, turn it off, and then run with the older WLM policies. Compare the results of using HiperDispatch, readjust the new policies, and repeat the cycle, as needed. WLM policies can be changed without turning off HiperDispatch.

A health check is provided to verify whether HiperDispatch is enabled on a system image that is running on z13s.

z/VM V6R3

z/VM V6R3 also uses the HiperDispatch facility for improved processor efficiency by better use of the processor cache to take advantage of the cache-rich processor, node, and drawer design of the z13s system. The supported processor limit has been increased to 64, whereas with SMT, it remains at 32, supporting up to 64 threads running simultaneously.

The operating system support requirements for HiperDispatch are listed in Table 7-18.

Table 7-18 Minimum support requirements for HiperDispatch

Operating system	Support requirements
z/OS	z/OS V1R11 with PTFs
z/VM	z/VM V6R3
Linux on z Systems	SUSE Linux Enterprise Server 12 ^a SUSE Linux Enterprise Server 11 ^a Red Hat RHEL 7 ^a Red Hat RHEL 6 ^a

a. For more information about CPU polarization support, see http://www.ibm.com/support/knowledgecenter/linuxonibm/com.ibm.linux.z.lgdd/lgdd_t_cpu_pol.html

7.3.17 The 63.75-K subchannels

Servers before z9 EC reserved 1024 subchannels for internal system use, out of a maximum of 64 K subchannels. Starting with z9 EC, the number of reserved subchannels was reduced to 256, increasing the number of subchannels that are available. Reserved subchannels exist only in subchannel set 0. One subchannel is reserved in each of subchannel sets 1, 2, and 3.

The informal name, *63.75-K subchannels*, represents 65280 subchannels, as shown in the following equation:

$$63 \times 1024 + 0.75 \times 1024 = 65280$$

This equation is applicable for subchannel set 0. For subchannel sets 1, 2 and 3, the available subchannels are derived by using the following equation:

(64 X 1024) -1=65535

Table 7-19 lists the minimum operating system level that is required on the z13s.

Table 7-19 Minimum support requirements for 63.75-K subchannels

Operating system	Support requirements
z/OS	z/OS V1R12
z/VM	z/VM V6R2
Linux on z Systems	SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 Red Hat RHEL 7 Red Hat RHEL 6

7.3.18 Multiple Subchannel Sets

Multiple Subchannel Sets (MSS), first introduced in z9 EC, provide a mechanism for addressing more than 63.75-K I/O devices and aliases for ESCON¹ (CHPID type CNC) and FICON (CHPID types FC) on the z13, z13s, zEC12, zBC12, z196, z114, z10 EC, and z9 EC. z196 introduced the third subchannel set (SS2), which is not available in zBC12. With z13s servers, three subchannel sets are now available.

Table 7-20 lists the minimum operating system levels that are required on the z13s.

Table 7-20 Minimum software requirement for MSS

Operating system	Support requirements
z/OS	z/OS V1R12
z/VM	z/VM V6R3 ^a
Linux on z Systems	SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 Red Hat RHEL 7 Red Hat RHEL 6

a. For specific Geographically Dispersed Parallel Sysplex (GDPS) usage only

z/VM V6R3 MSS support for mirrored direct access storage device (DASD) provides a subset of host support for the MSS facility to allow using an alternative subchannel set for a Peer-to-Peer Remote Copy (PPRC) secondary volumes.

7.3.19 Three subchannel sets

With z13s, a *third subchannel set (SS2)* is introduced. Together with the second subchannel set (SS1), SS2 can be used for disk alias devices of both primary and secondary devices, and as Metro Mirror secondary devices. These subchannel sets help facilitate storage growth and complement other functions, such as extended address volume (EAV) and Hyper Parallel Access Volumes (HyperPAV).

¹ ESCON features are not supported on z13s servers.

Table 7-21 lists the minimum operating systems level that is required on the z13s.

Table 7-21 Minimum software requirement for four subchannel sets

Operating system	Support requirements
z/OS	z/OS V1R13 with PTFs
z/VM	z/VM V6R2 with PTF
Linux on z Systems	SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 Red Hat RHEL 7 Red Hat RHEL 6

7.3.20 IPL from an alternative subchannel set

z13s supports IPL from subchannel set 1 (SS1), or subchannel set 2 (SS2) in addition to subchannel set 0. For more information, see “IPL from an alternate subchannel set” on page 192.

7.3.21 Modified Indirect Data Address Word facility

The Modified Indirect Data Address Word (MIDAW) facility improves FICON performance. It provides a more efficient channel command word (CCW)/indirect data address word (IDAW) structure for certain categories of data-chaining I/O operations.

Support for the MIDAW facility when running z/OS as a guest of z/VM requires z/VM V6R2 or later. For more information, see 7.9, “Simultaneous multithreading” on page 295.

Table 7-22 lists the minimum support requirements for MIDAW.

Table 7-22 Minimum support requirements for MIDAW

Operating system	Support requirements
z/OS	z/OS V1R12
z/VM	z/VM V6R2 for guest use

7.3.22 HiperSockets Completion Queue

The HiperSockets Completion Queue function is implemented on z13, z13s, zEC12, zBC12, z196, and z114 servers. The HiperSockets Completion Queue function is designed to allow HiperSockets to transfer data synchronously if possible, and asynchronously if necessary. Therefore, it combines ultra-low latency with more tolerance for traffic peaks. HiperSockets Completion Queue can be especially helpful in burst situations.

Table 7-23 lists the minimum support requirements for HiperSockets Completion Queue.

Table 7-23 Minimum support requirements for HiperSockets Completion Queue

Operating system	Support requirements
z/OS	z/OS V1R13 ^a
z/VSE	z/VSE V5R1 ^a
z/VM	z/VM V6R2 ^a

Operating system	Support requirements
Linux on z Systems	SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 SP2 Red Hat RHEL 7 Red Hat RHEL 6.2

a. PTFs are required.

7.3.23 HiperSockets integration with the intraensemble data network

The HiperSockets integration with the intraensemble data network (IEDN) is available on z13, z13s, zEC12, zBC12, z196, and z114 servers. HiperSockets integration with the IEDN combines the HiperSockets network and the physical IEDN to be displayed as a single Layer 2 network. This configuration extends the reach of the HiperSockets network outside the CPC to the entire ensemble, displaying as a single Layer 2 network.

Table 7-24 lists the minimum support requirements for HiperSockets integration with the IEDN.

Table 7-24 Minimum support requirements for HiperSockets integration with IEDN

Operating system	Support requirements
z/OS	z/OS V1R13 ^a

a. PTFs are required.

7.3.24 HiperSockets Virtual Switch Bridge

The HiperSockets Virtual Switch Bridge is implemented on z13, z13s, zEC12, zBC12, z196, and z114 servers. HiperSockets Virtual Switch Bridge can integrate with the IEDN through OSA-Express for zBX (OSX) adapters. It can then bridge to another central processor complex (CPC) through OSD adapters. This configuration extends the reach of the HiperSockets network outside of the CPC to the entire ensemble and hosts that are external to the CPC. The system is displayed as a single Layer 2 network.

Table 7-25 lists the minimum support requirements for HiperSockets Virtual Switch Bridge.

Table 7-25 Minimum support requirements for HiperSockets Virtual Switch Bridge

Operating system	Support requirements
z/VM	z/VM V6R2 ^a , z/VM V6R3
Linux on z Systems ^b	SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 SUSE Linux Enterprise Server 10 SP4 update (kernel 2.6.16.60-0.95.1) Red Hat RHEL 7 Red Hat RHEL 6 Red Hat RHEL 5.8 (GA-level)

a. PTFs are required.

b. Applicable to guest operating systems.

7.3.25 HiperSockets Multiple Write Facility

The HiperSockets Multiple Write Facility allows the streaming of bulk data over a HiperSockets link between two LPARs. Multiple output buffers are supported on a single Signal Adapter (SIGA) write instruction. The key advantage of this enhancement is that it allows the receiving LPAR to process a much larger amount of data per I/O interrupt. This process is transparent to the operating system in the receiving partition. HiperSockets Multiple Write Facility with fewer I/O interrupts is designed to reduce processor utilization of the sending and receiving partitions.

Support for this function is required by the sending operating system. For more information, see 4.8.6, “HiperSockets” on page 172. Table 7-26 lists the minimum support requirements for HiperSockets Virtual Multiple Write Facility.

Table 7-26 Minimum support requirements for HiperSockets multiple write

Operating system	Support requirements
z/OS	z/OS V1R12

7.3.26 HiperSockets IPv6

IPv6 is expected to be a key element in future networking. The IPv6 support for HiperSockets allows compatible implementations between external networks and internal HiperSockets networks.

Table 7-27 lists the minimum support requirements for HiperSockets IPv6 (CHPID type IQD).

Table 7-27 Minimum support requirements for HiperSockets IPv6 (CHPID type IQD)

Operating system	Support requirements
z/OS	z/OS V1R12
z/VM	z/VM V6R2
z/VSE	z/VSE V5R1
Linux on z Systems	SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 Red Hat RHEL 7 Red Hat RHEL 6

7.3.27 HiperSockets Layer 2 support

For flexible and efficient data transfer for IP and non-IP workloads, the HiperSockets internal networks on z13s can support two transport modes. These modes are Layer 2 (Link Layer) and the current Layer 3 (Network or IP Layer). Traffic can be Internet Protocol (IP) Version 4 or Version 6 (IPv4, IPv6) or non-IP (AppleTalk, DECnet, IPX, NetBIOS, or SNA).

HiperSockets devices are protocol-independent and Layer 3-independent. Each HiperSockets device has its own Layer 2 Media Access Control (MAC) address. This MAC address allows the use of applications that depend on the existence of Layer 2 addresses, such as Dynamic Host Configuration Protocol (DHCP) servers and firewalls.

Layer 2 support can help facilitate server consolidation. Complexity can be reduced, network configuration is simplified and intuitive, and LAN administrators can configure and maintain the mainframe environment the same way as they do a non-mainframe environment.

Table 7-28 lists the minimum support requirements for HiperSockets Layer 2.

Table 7-28 Minimum support requirements for HiperSockets Layer 2

Operating system	Support requirements
z/OS	z/OS V1R12
z/VM	z/VM V6R2 for guest use
Linux on z Systems	SUSE Linux Enterprise Server 11 SUSE Linux Enterprise Server 12 Red Hat RHEL 7 Red Hat RHEL 6

7.3.28 HiperSockets network traffic analyzer for Linux on z Systems

HiperSockets network traffic analyzer (HS NTA), introduced with IBM System z10, provides support for tracing Layer2 and Layer3 HiperSockets network traffic in Linux on z Systems. This support allows Linux on z Systems to control the trace for the internal virtual LAN to capture the records into host memory and storage (file systems).

Linux on z Systems tools can be used to format, edit, and process the trace records for analysis by system programmers and network administrators.

7.3.29 FICON Express16S

FICON Express16S supports a link data rate of 16 gigabits per second (Gbps) and autonegotiation to 4 or 8 Gbps for synergy with existing switches, directors, and storage devices. With support for native FICON, zHPF, and FCP, the z13s server enables SAN for even higher performance, helping to prepare for an end-to-end 16 Gbps infrastructure to meet the increased bandwidth demands of your applications.

The new features for the multimode and single mode fiber optic cabling environments have reduced latency for large read/write operations and increased bandwidth compared to the FICON Express8S features.

Table 7-29 lists the minimum support requirements for FICON Express16S.

Table 7-29 Minimum support requirements for FICON Express16S

Operating system	z/OS	z/VM	z/VSE	z/TPF	Linux on z Systems
Native FICON and CTC CHPID type FC	V1R12 ^a	V6R2	V5R1	V1R1	SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 Red Hat RHEL 7 Red Hat RHEL 6
zHPF single-track operations CHPID type FC	V1R12	V6R2 ^b	N/A	N/A	SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 Red Hat RHEL 7 Red Hat RHEL 6
zHPF multitrack operations CHPID type FC	V1R12 ^b	V6R2 ^b	N/A	N/A	SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 Red Hat RHEL 7 Red Hat RHEL 6

Operating system	z/OS	z/VM	z/VSE	z/TPF	Linux on z Systems
Support of SCSI devices CHPID type FCP	N/A	V6R2 ^b	V5R1	N/A	SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 Red Hat RHEL 7 Red Hat RHEL 6
Support of hardware data router CHPID type FCP	N/A	V6R3	N/A	N/A	SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 SP3 Red Hat RHEL 7 Red Hat RHEL 6.4
Support of T10-DIF CHPID type FCP	N/A	V6R2 ^b	N/A	N/A	SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 SP3 Red Hat RHEL 7 Red Hat RHEL 6.4

a. PTFs are required to support GRS FICON CTC toleration.

b. PTFs are required.

7.3.30 FICON Express8S

The *FICON Express8S* feature is exclusively installed in the Peripheral Component Interconnect Express (PCIe) I/O drawer. It provides a link rate of 8 Gbps, with auto negotiation to 4 or 2 Gbps for compatibility with previous devices and investment protection. Both 10 km (6.2 miles) LX and SX connections are offered (in a feature, all connections must have the same type).

With FICON Express 8S, clients might be able to consolidate existing FICON, FICON Express², and FICON Express4² channels, while maintaining and enhancing performance.

FICON Express8S introduced a hardware data router for more efficient zHPF data transfers. It is the first channel with hardware that is designed to support zHPF, as contrasted to FICON Express8, FICON Express4², and FICON Express2², which have a firmware-only zHPF implementation.

Table 7-30 lists the minimum support requirements for FICON Express8S.

Table 7-30 Minimum support requirements for FICON Express8S

Operating system	z/OS	z/VM	z/VSE	z/TPF	Linux on z Systems
Native FICON and CTC CHPID type FC	V1R12 ^a	V6R2	V5R1	V1R1	SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 Red Hat RHEL 7 Red Hat RHEL 6
zHPF single-track operations CHPID type FC	V1R12	V6R2 ^b	N/A	N/A	SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 Red Hat RHEL 7 Red Hat RHEL 6

² All FICON Express4, FICON Express2, and FICON features have been withdrawn from marketing.

Operating system	z/OS	z/VM	z/VSE	z/TPF	Linux on z Systems
zHPF multitrack operations CHPID type FC	V1R12 ^b	V6R2 ^b	N/A	N/A	SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 Red Hat RHEL 7 Red Hat RHEL 6
Support of SCSI devices CHPID type FCP	N/A	V6R2 ^b	V5R1	N/A	SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 Red Hat RHEL 7 Red Hat RHEL 6
Support of hardware data router CHPID type FCP	N/A	V6R3	N/A	N/A	SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 SP3 Red Hat RHEL 7 Red Hat RHEL 6.4
Support of T10-DIF CHPID type FCP	N/A	V6R2 ^b	N/A	N/A	SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 SP3 Red Hat RHEL 7 Red Hat RHEL 6.4

a. PTFs are required to support GRS FICON CTC toleration.

b. PTFs are required.

7.3.31 FICON Express8

The FICON Express8 features provide a link rate of 8 Gbps, with auto-negotiation to 4 Gbps or 2 Gbps for compatibility with previous devices and investment protection. Both 10 km (6.2 miles) LX and SX connections are offered (in a feature, all connections must have the same type).

Important: The z13 and z13s servers are the last z Systems servers to support FICON Express 8 channels. Enterprises should begin upgrading from FICON Express8 channel features (FC 3325 and FC 3326) to FICON Express16S channel features (FC 0418 and FC 0419). FICON Express 8 will not be supported on future z Systems servers as carry-forward on an upgrade.

All statements regarding IBM plans, directions, and intent are subject to change or withdrawal without notice. Any reliance on these statements of general direction is at the relying party's sole risk and will not create liability or obligation for IBM.

Table 7-31 lists the minimum support requirements for FICON Express8.

Table 7-31 Minimum support requirements for FICON Express8

Operating system	z/OS	z/VM	z/VSE	z/TPF	Linux on z Systems
Native FICON and CTC CHPID type FC	V1R12	V6R2	V5R1	V1R1	SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 Red Hat RHEL 7 Red Hat RHEL 6

Operating system	z/OS	z/VM	z/VSE	z/TPF	Linux on z Systems
zHPF single-track operations CHPID type FC	V1R12	V6R2 ^a	N/A	N/A	SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 Red Hat RHEL 7 Red Hat RHEL 6
zHPF multitrack operations CHPID type FC	V1R12 ^a	V6R2 ^a	N/A	N/A	SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 Red Hat RHEL 7 Red Hat RHEL 6
Support of SCSI devices CHPID type FCP	N/A	V6R2 ^a	V5R1	N/A	SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 Red Hat RHEL 7 Red Hat RHEL 6
Support of T10-DIF CHPID type FCP	N/A	V6R2 ^a	N/A	N/A	SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 Red Hat RHEL 7 Red Hat RHEL 6

a. PTFs are required.

7.3.32 z/OS Discovery and Auto-Configuration

z/OS Discovery and Auto-Configuration (zDAC) is designed to automatically run a number of I/O configuration definition tasks for new and changed disk and tape controllers that are connected to a switch or director, when attached to a FICON channel.

The zDAC function is integrated into the existing hardware configuration definition (HCD). Clients can define a policy that can include preferences for availability and bandwidth that include parallel access volume (PAV) definitions, control unit numbers, and device number ranges. When new controllers are added to an I/O configuration or changes are made to existing controllers, the system discovers them and proposes configuration changes that are based on that policy.

zDAC provides real-time discovery for the FICON fabric, subsystem, and I/O device resource changes from z/OS. By exploring the discovered control units for defined logical control units (LCUs) and devices, zDAC compares the discovered controller information with the current system configuration. It then determines delta changes to the configuration for a proposed configuration.

All added or changed logical control units and devices are added into the proposed configuration. They are assigned proposed control unit and device numbers, and channel paths that are based on the defined policy. zDAC uses channel path chosen algorithms to minimize single points of failure. The zDAC proposed configurations are created as work I/O definition files (IODF) that can be converted to production IODFs and activated.

zDAC is designed to run discovery for all systems in a sysplex that support the function. Therefore, zDAC helps to simplify I/O configuration on z13s systems that run z/OS, and reduces complexity and setup time.

zDAC applies to all FICON features that are supported on z13s when configured as CHPID type FC. Table 7-32 lists the minimum support requirements for zDAC.

Table 7-32 Minimum support requirements for zDAC

Operating system	Support requirements
z/OS	z/OS V1R12 ^a

a. PTFs are required.

7.3.33 High-performance FICON

zHPF, first provided on System z10, is a FICON architecture for protocol simplification and efficiency. It reduces the number of information units (IUs) processed. Enhancements have been made to the z/Architecture and the FICON interface architecture to provide optimizations for online transaction processing (OLTP) workloads.

When used by the FICON channel, the z/OS operating system, and the DS8000 control unit or other subsystems, the FICON channel processor usage can be reduced and performance improved. Appropriate levels of Licensed Internal Code (LIC) are required. Additionally, the changes to the architectures provide end-to-end system enhancements to improve reliability, availability, and serviceability (RAS).

zHPF is compatible with these standards:

- ▶ Fibre Channel Framing and Signaling standard (FC-FS)
- ▶ Fibre Channel Switch Fabric and Switch Control Requirements (FC-SW)
- ▶ Fibre Channel Single-Byte-4 (FC-SB-4) standards

The zHPF channel programs can be used, for example, by the z/OS OLTP I/O workloads, DB2, VSAM, the partitioned data set extended (PDSE), and the z/OS file system (zFS).

At the zHPF announcement, zHPF supported the transfer of small blocks of fixed size data (4 K) from a single track. This capability has been extended, first to 64 KB, and then to multitrack operations. The 64 KB data transfer limit on multitrack operations was removed by z196. This improvement allows the channel to fully use the bandwidth of FICON channels, resulting in higher throughputs and lower response times.

The multitrack operations extension applies exclusively to the FICON Express8S, FICON Express8, and FICON Express16S, on the z13, z13s, zEC12, zBC12, z196, and z114 servers, when configured as CHPID type FC and connecting to z/OS. zHPF requires matching support by the DS8000 series. Otherwise, the extended multitrack support is transparent to the control unit.

From the z/OS point of view, the existing FICON architecture is called *command mode* and the zHPF architecture is called *transport mode*. During link initialization, the channel node and the control unit node indicate whether they support zHPF.

Requirement: All FICON CHPIDs that are defined to the same LCU must support zHPF. The inclusion of any non-compliant zHPF features in the path group causes the entire path group to support command mode only.

The mode that is used for an I/O operation depends on the control unit that supports zHPF and its settings in the z/OS operating system. For z/OS use, a parameter in the IECIOSxx member of SYS1.PARMLIB (ZHPF=YES or NO) and in the **SETI0S** system command controls whether zHPF is enabled or disabled. The default is ZHPF=NO.

Support is also added for the **D IOS,ZHPF** system command to indicate whether zHPF is enabled, disabled, or not supported on the server.

Similar to the existing FICON channel architecture, the application or access method provides the channel program (CCWs). The way that zHPF (transport mode) manages channel program operations is different from the CCW operation for the existing FICON architecture (command mode). While in command mode, each CCW is sent to the control unit for execution. In transport mode, multiple channel commands are packaged together and sent over the link to the control unit in a single control block. Fewer processors are used compared to the existing FICON architecture. Certain complex CCW chains are not supported by zHPF.

zHPF is available to z13, z13s, zEC12, zBC12, z196, z114, and System z10 servers. The FICON Express8S, FICON Express8, and FICON Express16S (CHPID type FC) concurrently support both the existing FICON protocol and the zHPF protocol in the server LIC.

zHPF is enhanced to allow all large write operations (> 64 KB) at distances up to 100 km to be run in a single round trip to the control unit. This enhancement avoids elongating the I/O service time for these write operations at extended distances. This enhancement to zHPF removes a key inhibitor for clients adopting zHPF over extended distances, especially when using the IBM HyperSwap® capability of z/OS.

Table 7-33 lists the minimum support requirements for zHPF.

Table 7-33 Minimum support requirements for zHPF

Operating system	Support requirements
z/OS	Single-track operations: z/OS V1R12. Multitrack operations: z/OS V1R12 with PTFs. 64 K enhancement: z/OS V1R12 with PTFs.
z/VM	z/VM V6.2 for guest use only.
Linux on z Systems	SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 SP1 Red Hat RHEL 7 Red Hat RHEL 6 IBM continues to work with its Linux distribution partners on use of appropriate z Systems (z13s, zBC12) functions to be provided in future Linux on z Systems distribution releases.

For more information about FICON channel performance, see the performance technical papers on the z Systems I/O connectivity website at:

http://www.ibm.com/systems/z/hardware/connectivity/ficon_performance.html

7.3.34 Request node identification data

First offered on z9 EC, the request node identification data (RNID) function for native FICON CHPID type FC allows isolation of cabling-detected errors.

Table 7-34 lists the minimum support requirements for RNID.

Table 7-34 Minimum support requirements for RNID

Operating system	Support requirements
z/OS	z/OS V1R12

7.3.35 32 K subchannels for the FICON Express16S

To help facilitate growth and continue to enable server consolidation, the z13s supports up to 32 K subchannels per FICON Express16S channel (CHPID). More devices can be defined per FICON channel, which includes primary, secondary, and alias devices. The maximum number of subchannels across all device types that are addressable within an LPAR remains at 63.75 K for subchannel set 0 and 64 K (64 X 1024)-1 for subchannel sets 1 and 2.

This support is exclusive to z13 and z13s servers and applies to the FICON Express16S feature (defined as CHPID type FC). FICON Express8S and FICON Express8 remain at 24 subchannel support when defined as CHPID type FC.

Table 7-35 lists the minimum support requirements of 32 K subchannel support for FICON Express.

Table 7-35 Minimum support requirements for 32K subchannel

Operating system	Support requirements
z/OS	z/OS V1R12 ^a
z/VM	z/VM V6R2
Linux on z Systems	SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 RHEL 7 RHEL 6

a. PTFs are required.

7.3.36 Extended distance FICON

An enhancement to the industry-standard FICON architecture (FC-SB-3) helps avoid degradation of performance at extended distances by implementing a new protocol for persistent IU pacing. Extended distance FICON is transparent to operating systems and applies to all FICON Express16S, FICON Express8S, and FICON Express8 features that carry native FICON traffic (CHPID type FC).

To use this enhancement, the control unit must support the new IU pacing protocol. IBM System Storage DS8000 series supports extended distance FICON for IBM z Systems environments. The channel defaults to current pacing values when it operates with control units that cannot use extended distance FICON.

7.3.37 Platform and name server registration in FICON channel

The FICON Express16S, FICON Express8S, and FICON Express8 features (on the zBC12 servers) support platform and name server registration to the fabric for CHPID types FC and FCP.

Information about the channels that are connected to a fabric, if registered, allows other nodes or storage area network (SAN) managers to query the name server to determine what is connected to the fabric.

The following attributes are registered for the z13s servers:

- ▶ Platform information
- ▶ Channel information
- ▶ Worldwide port name (WWPN)

- ▶ Port type (N_Port_ID)
- ▶ FC-4 types that are supported
- ▶ Classes of service that are supported by the channel

The platform and name server registration service are defined in the FC-GS-4 standard.

7.3.38 FICON link incident reporting

FICON link incident reporting allows an operating system image (without operator intervention) to register for link incident reports. Table 7-36 lists the minimum support requirements for this function.

Table 7-36 Minimum support requirements for link incident reporting

Operating system	Support requirements
z/OS	z/OS V1R12

7.3.39 FCP provides increased performance

The FCP LIC is modified to help provide increased I/O operations per second for both small and large block sizes, and to support 16-Gbps link speeds.

For more information about FCP channel performance, see the performance technical papers on the z Systems I/O connectivity website at:

http://www.ibm.com/systems/z/hardware/connectivity/fcp_performance.html

7.3.40 N_Port ID Virtualization

N_Port ID Virtualization (NPIV) allows multiple system images (in LPARs or z/VM guests) to use a single FCP channel as though each were the sole user of the channel. This feature, first introduced with z9 EC, can be used with earlier FICON features that have been carried forward from earlier servers.

Table 7-37 lists the minimum support requirements for NPIV.

Table 7-37 Minimum support requirements for NPIV

Operating system	Support requirements
z/VM	z/VM V6R2 provides support for guest operating systems and VM users to obtain virtual port numbers. Installation from DVD to SCSI disks is supported when NPIV is enabled.
z/VSE	z/VSE V5R1.
Linux on z Systems	SUSE Linux Enterprise Server 12. SUSE Linux Enterprise Server 11. Red Hat RHEL 7. Red Hat RHEL 6.

7.3.41 OSA-Express5S 10-Gigabit Ethernet LR and SR

The *OSA-Express5S 10-Gigabit Ethernet* feature, introduced with the zEC12 and zBC12, is installed exclusively in the PCIe I/O drawer. Each feature has one port, which is defined as either CHPID type OSD or OSX. CHPID type OSD supports the queued direct input/output (QDIO) architecture for high-speed TCP/IP communication. The z196 introduced the CHPID type OSX. For more information, see 7.3.50, “Intraensemble data network” on page 274.

Table 7-38 lists the minimum support requirements for OSA-Express5S 10-Gigabit Ethernet LR and SR features.

Table 7-38 Minimum support requirements for OSA-Express5S 10-Gigabit Ethernet LR and SR

Operating system	Support requirements
z/OS	OSD: z/OS V1R12 ^a OSX: z/OS V1R12 ^a
z/VM	OSD: z/VM V6R2
z/VSE	OSX: z/VSE V5R1 OSD: z/VSE V5R1
z/TPF	OSD: z/TPF V1R1 PUT 5 ^a OSX: z/TPF V1R1 PUT 8 ^a
IBM zAware	OSD OSX
Linux on z Systems	OSD: SUSE Linux Enterprise Server 11 SUSE Linux Enterprise Server 12 Red Hat RHEL 7 Red Hat RHEL6 OSX: SUSE Linux Enterprise Server 11 SP1 ^b SUSE Linux Enterprise Server 12 Red Hat RHEL 6 Red Hat RHEL 7

a. IBM Software Support Services is required for support.

b. Maintenance update is required.

7.3.42 OSA-Express5S Gigabit Ethernet LX and SX

The *OSA-Express5S Gigabit Ethernet feature* is installed exclusively in the PCIe I/O drawer. Each feature has one PCIe adapter and two ports. The two ports share a channel path identifier (CHPID type OSD exclusively). Each port supports attachment to a 1 Gbps Ethernet local area network (LAN). The ports can be defined as a spanned channel, and can be shared among LPARs and across logical channel subsystems.

Operating system support is required to recognize and use the second port on the OSA-Express5S Gigabit Ethernet feature. Table 7-39 lists the minimum support requirements for OSA-Express5S Gigabit Ethernet LX and SX.

Table 7-39 Minimum support requirements for OSA-Express5S Gigabit Ethernet LX and SX

Operating system	Support requirements using two ports per CHPID	Support requirements using one port per CHPID
z/OS	OSD: z/OS V1R12	OSD: z/OS V1R12
z/VM	OSD: z/VM V6R2	OSD: z/VM V6R2
z/VSE	OSD: z/VSE V5R1	OSD: z/VSE V5R1
z/TPF	OSD: z/TPF V1R1	OSD: z/TPF V1R1
IBM zAware	OSD	OSD
Linux on z Systems	OSD: SUSE Linux Enterprise Server 12, SUSE Linux Enterprise Server 11, Red Hat RHEL 7, and Red Hat RHEL 6	OSD: SUSE Linux Enterprise Server 12, SUSE Linux Enterprise Server 11, Red Hat RHEL 7, and Red Hat RHEL 6

7.3.43 OSA-Express5S 1000BASE-T Ethernet

The *OSA-Express5S 1000BASE-T Ethernet* feature is installed exclusively in the PCIe I/O drawer. Each feature has one PCIe adapter and two ports. The two ports share a CHPID, which can be defined as OSC, OSD, OSE, OSM, or OSN. The ports can be defined as a spanned channel, and can be shared among LPARs and across logical channel subsystems. The OSM CHPID type was introduced with z196. For more information, see 7.3.49, “Intranode management network” on page 274.

Each adapter can be configured in the following modes:

- ▶ QDIO mode, with CHPID types OSD and OSN.
- ▶ Non-QDIO mode, with CHPID type OSE.
- ▶ Local 3270 emulation mode, including OSA-ICC, with CHPID type OSC.
 - OSA-ICC (OSC Channel) supports Secure Sockets Layer on z13s and z13 Driver 27 servers.
 - Designed to improve security of console operations.
 - Up to 48 secure sessions per CHPID (the overall maximum of 120 connections unchanged).
- ▶ Ensemble management, with CHPID type OSM for dynamic I/O only.

Table 7-40 lists the minimum support requirements for OSA-Express5S 1000BASE-T.

Table 7-40 Minimum support requirements for OSA-Express5S 1000BASE-T Ethernet

Operating system	Support requirements using two ports per CHPID	Support requirements using one port per CHPID
z/OS	OSC, OSD, OSE, and OSN ^b : z/OS V1R12 ^a	OSC, OSD, OSE, OSM, and OSN ^b : z/OS V1R12 ^a
z/VM	OSC, OSD ^a , OSE, and OSN ^b : z/VM V6R2	OSC, OSD, OSE, OSM ^{ac} , and OSN ^b : z/VM V6R2
z/VSE	OSC, OSD, OSE, and OSN ^b : z/VSE V5R1	OSC, OSD, OSE, and OSN ^b : z/VSE V5R1
z/TPF	OSD, OSC, OSN ^b : z/TPF V1R1	OSD, OSC, OSN ^b : z/TPF
IBM zAware	OSD	OSD
Linux on z Systems	OSD: SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 Red Hat RHEL 7 Red Hat RHEL 6 OSN ^b : SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 Red Hat RHEL 7 Red Hat RHEL 6	OSD: SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 Red Hat RHEL 7 Red Hat RHEL 6 OSM: SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 SP2 Red Hat RHEL 7 Red Hat RHEL 6 OSN ^b : SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 Red Hat RHEL 7 Red Hat RHEL 6

a. PTFs are required.

b. Although CHPID type OSN does not use any ports (because all communication is LPAR to LPAR), it is listed here for completeness.

c. OSM support in V6R2 and V6R3 for dynamic I/O only.

7.3.44 OSA-Express4S 10-Gigabit Ethernet LR and SR

The *OSA-Express4S 10-Gigabit Ethernet* feature, introduced with the zEC12, is installed exclusively in the PCIe I/O drawer. Each feature has one port, which is defined as either CHPID type OSD or OSX. CHPID type OSD supports the QDIO architecture for high-speed TCP/IP communication. The z196 introduced the CHPID type OSX. For more information, see 7.3.50, “Intraensemble data network” on page 274.

The OSA-Express4S features have half the number of ports per feature when compared to OSA-Express3, and half the size as well. This configuration results in an increased number of installable features. It also facilitates the purchase of the correct number of ports to help satisfy your application requirements and to better avoid redundancy.

Table 7-41 lists the minimum support requirements for OSA-Express4S 10-Gigabit Ethernet LR and SR features.

Table 7-41 Minimum support requirements for OSA-Express4S 10-Gigabit Ethernet LR and SR

Operating system	Support requirements
z/OS	OSD: z/OS V1R12 ^a OSX: z/OS V1R12 ^a
z/VM	OSD: z/VM V6R2 OSX: z/VM V6R2 ^a and V6R3 for dynamic I/O only
z/VSE	OSD: z/VSE V5R1 OSX: z/VSE V5R1
z/TPF	OSD: z/TPF V1R1 OSX: z/TPF V1R1
IBM zAware	OSD OSX
Linux on z Systems	OSD: SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 Red Hat RHEL 7 Red Hat RHEL 6 OSX: SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 SP1 ^b Red Hat RHEL 7 Red Hat RHEL 6

a. PTFs are required.

b. Maintenance update is required.

7.3.45 OSA-Express4S Gigabit Ethernet LX and SX

The *OSA-Express4S Gigabit Ethernet* feature is installed exclusively in the PCIe I/O drawer. Each feature has one PCIe adapter and two ports. The two ports share a channel path identifier (CHPID type OSD exclusively). Each port supports attachment to a 1 Gbps Ethernet LAN. The ports can be defined as a spanned channel, and can be shared among LPARs and across logical channel subsystems.

Operating system support is required to recognize and use the second port on the OSA-Express4S Gigabit Ethernet feature. Table 7-42 lists the minimum support requirements for OSA-Express4S Gigabit Ethernet LX and SX.

Table 7-42 Minimum support requirements for OSA-Express4S Gigabit Ethernet LX and SX

Operating system	Support requirements using two ports per CHPID	Support requirements using one port per CHPID
z/OS	OSD: z/OS V1R12	OSD: z/OS V1R12
z/VM	OSD: z/VM V6R2	OSD: z/VM V6R2
z/VSE	OSD: z/VSE V5R1	OSD: z/VSE V5R1
z/TPF	OSD: z/TPF V1R1	OSD: z/TPF V1R1
IBM zAware	OSD	OSD

Operating system	Support requirements using two ports per CHPID	Support requirements using one port per CHPID
Linux on z Systems	OSD: SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 Red Hat RHEL 7 Red Hat RHEL 6	OSD: SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 Red Hat RHEL 7 Red Hat RHEL 6

7.3.46 OSA-Express4S 1000BASE-T Ethernet

The *OSA-Express4S 1000BASE-T Ethernet* feature is installed exclusively in the PCIe I/O drawer. Each feature has one PCIe adapter and two ports. The two ports share a CHPID, which is defined as OSC, OSD, OSE, OSM, or OSN. The ports can be defined as a spanned channel, and can be shared among LPARs and across logical channel subsystems. The OSM CHPID type was introduced with z196. For more information, see 7.3.49, “Intranode management network” on page 274.

Each adapter can be configured in the following modes:

- ▶ QDIO mode, with CHPID types OSD and OSN
- ▶ Non-QDIO mode, with CHPID type OSE
- ▶ Local 3270 emulation mode, including OSA-ICC, with CHPID type OSC:
 - OSA-ICC (OSC Channel) supports Secure Sockets Layer on z13s and z13 Driver 27 servers
 - Up to 48 secure sessions per CHPID (the overall maximum of 120 connections is unchanged)
- ▶ Ensemble management, with CHPID type OSM

Operating system support is required to recognize and use the second port on the OSA-Express4S 1000BASE-T feature. Table 7-43 lists the minimum support requirements for OSA-Express4S 1000BASE-T.

Table 7-43 Minimum support requirements for OSA-Express4S 1000BASE-T Ethernet

Operating system	Support requirements using two ports per CHPID	Support requirements using one port per CHPID
z/OS	OSC, OSD, OSE, and OSN ^b : z/OS V1R12 ^a	OSC, OSD, OSE, OSM, and OSN ^b : z/OS V1R12 ^a
z/VM	OSC, OSD ^a , OSE, and OSN ^b : z/VM V6R2	OSC, OSD, OSE, OSM ^{a,c} , and OSN ^b : z/VM V6R2
z/VSE	OSC, OSD, OSE, and OSN ^b : z/VSE V5R1	OSC, OSD, OSE, and OSN ^b : z/VSE V5R1
z/TPF	OSD, OSC, OSN ^b : z/TPF V1R1	OSD, OSC, OSN ^b : z/TPF V1R1
IBM zAware	OSD	OSD

Operating system	Support requirements using two ports per CHPID	Support requirements using one port per CHPID
Linux on z Systems	OSD: SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 Red Hat RHEL 7 Red Hat RHEL 6 OSN ^b : SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 Red Hat RHEL 7 Red Hat RHEL 6	OSD: SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 Red Hat RHEL 7 Red Hat RHEL 6 OSM: SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 SP2 Red Hat RHEL 7 Red Hat RHEL 6 OSN ^b : SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 Red Hat RHEL 7 Red Hat RHEL 6

- a. PTFs are required.
- b. Although CHPID type OSN does not use any ports (because all communication is LPAR to LPAR), it is listed here for completeness.
- c. OSM Support in z/VM V6R2 and V6R3 for dynamic I/O only.

7.3.47 Open Systems Adapter for IBM zAware

The IBM zAware server requires connections to the graphical user interface (GUI) browser and z/OS monitored clients. An OSA channel is the most logical choice for allowing GUI browser connections to the server. By using this channel, users can view the analytical data for the monitored clients through the IBM zAware GUI. For z/OS monitored clients that connect an IBM zAware server, one of the following network options is supported:

- ▶ A client-provided data network that is provided through an OSA Ethernet channel.
- ▶ A HiperSockets subnetwork within the z13s.
- ▶ IEDN on the z13s to other CPC nodes in the ensemble. The z13s server also supports the use of HiperSockets over the IEDN.

7.3.48 Open Systems Adapter for Ensemble

Five different OSA-Express5S and OSA-Express4S features are used to connect the z13s to Unified Resource Manager, and other ensemble nodes. These connections are part of the ensemble's two private and secure internal networks.

For the intranode management network (INMN), use these features:

- ▶ OSA Express5S 1000BASE-T Ethernet, FC 0417
- ▶ OSA Express4S 1000BASE-T Ethernet, FC 0408

For the IEDN, use these features (CHPID type OSX):

- ▶ OSA-Express5S 10 Gigabit Ethernet (GbE) Long Reach (LR), FC 0415
- ▶ OSA-Express5S 10 Gigabit Ethernet (GbE) Short Reach (SR), FC 0416
- ▶ OSA-Express4S 10 Gigabit Ethernet (GbE) Long Reach (LR), FC 0406
- ▶ OSA-Express4S 10 Gigabit Ethernet (GbE) Short Reach (SR), FC 0407

7.3.49 Intranode management network

The INMN is one of the ensemble's two private and secure internal networks. The INMN is used by the Unified Resource Manager functions.

The INMN is a private and physically isolated 1000Base-T Ethernet internal platform management network. It operates at 1 Gbps, and connects all resources (CPC components) of an ensemble node for management purposes. It is pre-wired, internally switched, configured, and managed with full redundancy for high availability.

The z196 introduced the OSA-Express for Unified Resource Manager (OSM) CHPID type. INMN requires two OSA-Express5S 1000BASE-T or OSA-Express4S 1000BASE-T ports from two different OSA-Express5S 1000BASE-T or OSA-Express4S 1000BASE-T features, which are configured as CHPID type OSM. One port per CHPID is available with CHPID type OSM.

The OSA connection is through the system control hub (SCH) on the z13s to the HMC network interface.

7.3.50 Intraensemble data network

The IEDN is one of the ensemble's two private and secure internal networks. The IEDN provides applications with a fast data exchange path between ensemble nodes. Specifically, it is used for communications across the virtualized images (LPARs, z/VM virtual machines, and blade LPARs).

The IEDN is a private and secure 10 Gbps Ethernet network that connects all elements of an ensemble. It is access-controlled by using integrated virtual LAN (VLAN) provisioning. No client-managed switches or routers are required. The IEDN is managed by the primary HMC that controls the ensemble. This configuration helps reduce the need for firewalls and encryption, and simplifies network configuration and management. It also provides full redundancy for high availability.

The z196 introduced the OSX CHPID type. The OSA connection is from the z13s to the top-of-rack (ToR) switches on zBX.

The IEDN requires two OSA-Express5S or OSA-Express4S 10 GbE ports that are configured as CHPID type OSX.

7.3.51 OSA-Express5S and OSA-Express4S NCP support

The OSA-Express5S 1000BASE-T Ethernet and OSA-Express4S 1000BASE-T Ethernet features can provide channel connectivity from an operating system in a z13s to IBM Communication Controller for Linux on z Systems (CCL). This configuration uses the Open Systems Adapter for Network Control Program (NCP) (CHPID type OSN) in support of the Channel Data Link Control (CDLC) protocol. OSN eliminates the requirement for an external communication medium for communications between the operating system and the CCL image.

The data flow of the LPAR to the LPAR is accomplished by the OSA-Express5S or OSA-Express4S feature without ever exiting the card. The OSN support allows multiple connections between the CCL image and the operating system, such as z/OS or z/TPF. The operating system must be in the same physical server as the CCL image.

For CCL planning information, see *IBM Communication Controller for Linux on System z V1.2.1 Implementation Guide*, SG24-7223. For the most recent CCL information, see this website:

<http://www.ibm.com/software/network/cc1/>

CDLC, when used with CCL, emulates selected functions of IBM 3745/NCP operations. The port that is used with the OSN support is displayed as an ESCON channel to the operating system. This support can be used with OSA-Express5S 1000BASE-T and OSA-Express4S 1000BASE-T features.

Table 7-44 lists the minimum support requirements for OSN.

Table 7-44 Minimum support requirements for OSA-Express5S and OSA-Express4S OSN

Operating system	Support requirements
z/OS	z/OS V1R12 ^a
z/VM	z/VM V6R2
z/VSE	z/VSE V5R1
z/TPF	z/TPF V1R1 ^a
Linux on z Systems	SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 Red Hat RHEL 7 Red Hat RHEL 6

a. PTFs are required.

7.3.52 Integrated Console Controller

The *1000BASE-T Ethernet* features provide the Integrated Console Controller (OSA-ICC) function, which supports TN3270E (RFC 2355) and non-SNA DFT 3270 emulation. The OSA-ICC function is defined as CHPID type OSC and console controller, and has multiple LPAR support, both as shared or spanned channels.

With the OSA-ICC function, 3270 emulation for console session connections is integrated in the z13s through a port on the OSA-Express5S 1000BASE-T or OSA-Express4S 1000BASE-T features. This function eliminates the requirement for external console controllers, such as 2074 or 3174, helping to reduce cost and complexity. Each port can support up to 120 console session connections.

OSA-ICC can be configured on a PCHID-by-PCHID basis, and is supported at any of the feature settings (10, 100, or 1000 Mbps, half-duplex or full-duplex).

Starting with z13s (z13 Driver 27) TLS/SSL with Certificate Authentication will be added to the OSC CHPID to provide a secure and validated method for connecting clients to the z Systems host.

Note: OSA-ICC supports up to 48 secure sessions per CHPID (the overall maximum of 120 connections is unchanged).

7.3.53 VLAN management enhancements

Table 7-45 lists the minimum support requirements for VLAN management enhancements for the OSA-Express5S and OSA-Express4S features (CHPID type OSD).

Table 7-45 Minimum support requirements for VLAN management enhancements

Operating system	Support requirements
z/OS	z/OS V1R12.
z/VM	z/VM V6R2. Support of guests is transparent to z/VM if the device is directly connected to the guest (pass through).

7.3.54 GARP VLAN Registration Protocol

All OSA-Express5S and OSA-Express4S features support VLAN prioritization, a component of the IEEE 802.1 standard. GARP VLAN Registration Protocol (GVRP) support allows an OSA-Express port to register or unregister its VLAN IDs with a GVRP-capable switch and dynamically update its table as the VLANs change. This process simplifies the network administration and management of VLANs because manually entering VLAN IDs at the switch is no longer necessary.

The minimum support requirements are listed in Table 7-46.

Table 7-46 Minimum support requirements for GVRP

Operating system	Support requirements
z/OS	z/OS V1R12
z/VM	z/VM V6R2
Linux on z Systems ^a	SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 Red Hat RHEL 7 Red Hat RHEL 6

a. By using VLANs

7.3.55 Inbound workload queuing for OSA-Express5S and OSA-Express4S

OSA-Express3 introduced inbound workload queuing (IWQ), which creates multiple input queues and allows OSA to differentiate workloads “off the wire.” It then assigns work to a specific input queue (per device) to z/OS. The support is also available with OSA-Express5S and OSA-Express4S. CHPID types OSD and OSX are supported.

Each input queue is a unique type of workload, and has unique service and processing requirements. The IWQ function allows z/OS to preassign the appropriate processing resources for each input queue. This approach allows multiple concurrent z/OS processing threads to process each unique input queue (workload), avoiding traditional resource contention. In a heavily mixed workload environment, this “off the wire” network traffic separation is provided by OSA-Express5S and OSA-Express4S. IWQ reduces the conventional z/OS processing that is required to identify and separate unique workloads. This advantage results in improved overall system performance and scalability.

A primary objective of IWQ is to provide improved performance for business-critical interactive workloads by reducing contention that is created by other types of workloads. The following types of z/OS workloads are identified and assigned to unique input queues:

- ▶ z/OS Sysplex Distributor traffic: Network traffic that is associated with a distributed virtual Internet Protocol address (VIPA) is assigned to a unique input queue. This configuration allows the Sysplex Distributor traffic to be immediately distributed to the target host.
- ▶ z/OS bulk data traffic: Network traffic that is dynamically associated with a streaming (bulk data) TCP connection is assigned to a unique input queue. This configuration allows the bulk data processing to be assigned the appropriate resources and isolated from critical interactive workloads.

IWQ is exclusive to OSA-Express5S and OSA-Express4S CHPID types OSD and OSX, and the z/OS operating system. This limitation applies to z13, z13s, zEC12, zBC12, z196, z114, and System z10 servers. The minimum support requirements are listed in Table 7-47.

Table 7-47 Minimum support requirements for IWQ

Operating system	Support requirements
z/OS	z/OS V1R12
z/VM	z/VM V6R2 for guest use only, service required

7.3.56 Inbound workload queuing for Enterprise Extender

IWQ for the OSA-Express features is enhanced to differentiate and separate inbound Enterprise Extender traffic to a dedicated input queue.

IWQ for Enterprise Extender is exclusive to OSA-Express5S and OSA-Express4S, CHPID types OSD and OSX, and the z/OS operating system. This limitation applies to z13, z13s, zEC12, zBC12, z196, and z114 servers. The minimum support requirements are listed in Table 7-48.

Table 7-48 Minimum support requirements for IWQ

Operating system	Support requirements
z/OS	z/OS V1R13
z/VM	z/VM V6R2 for guest use only, service required

7.3.57 Querying and displaying an OSA configuration

OSA-Express3 introduced the capability for the operating system to query and directly display the current OSA configuration information (similar to OSA/SF). z/OS uses this OSA capability by introducing a TCP/IP operator command called **display OSAINFO**.

Using **display OSAINFO** allows the operator to monitor and verify the current OSA configuration. Doing so helps improve the overall management, serviceability, and usability of OSA-Express5S and OSA-Express4S.

The **display OSAINFO** command is exclusive to z/OS, and applies to CHPID types OSD, OSM, and OSX.

7.3.58 Link aggregation support for z/VM

Link aggregation (IEEE 802.3ad) that is controlled by the z/VM Virtual Switch (VSWITCH) allows the dedication of an OSA-Express5S or OSA-Express4S port to the z/VM operating system. The port must be part of an aggregated group that is configured in Layer 2 mode. Link aggregation (trunking) combines multiple physical OSA-Express5S or OSA-Express4S ports into a single logical link. This configuration increases throughput, and provides nondisruptive failover if a port becomes unavailable. The target links for aggregation must be of the same type.

Link aggregation is applicable to CHPID type OSD (QDIO). Link aggregation is supported by z/VM V6R2 and later.

7.3.59 Multi-VSwitch Link Aggregation

Multi-VSwitch Link Aggregation support allows a port group of OSA-Express features to span multiple virtual switches within a single z/VM system or between multiple z/VM systems. Sharing a Link Aggregation Port Group (LAG) with multiple virtual switches increases optimization and utilization of the OSA Express when handling larger traffic loads. Higher adapter utilization protects customer investments, which is increasingly important as 10 Gb deployments become more prevalent.

The minimum support requirements are listed in Table 7-49.

Table 7-49 Minimum support requirements for Multi-VSwitch Link Aggregation

Operating system	Support requirements
z/VM	z/VM V6R3 ^a

a. PTF support is required.

7.3.60 QDIO data connection isolation for z/VM

The QDIO data connection isolation function provides a higher level of security when sharing an OSA connection in z/VM environments that use VSWITCH. The VSWITCH is a virtual network device that provides switching between OSA connections and the connected guest systems.

QDIO data connection isolation allows disabling internal routing for each QDIO connected. It also allows creating security zones and preventing network traffic between the zones. It is supported by all OSA-Express5S and OSA-Express4S features on z13s and zBC12.

7.3.61 QDIO interface isolation for z/OS

Some environments require strict controls for routing data traffic between servers or nodes. In certain cases, the LPAR-to-LPAR capability of a shared OSA connection can prevent such controls from being enforced. With interface isolation, internal routing can be controlled on an LPAR basis. When interface isolation is enabled, the OSA discards any packets that are destined for a z/OS LPAR that is registered in the OSA Address Table (OAT) as isolated.

QDIO interface isolation is supported by Communications Server for z/OS V1R12 or later, and all OSA-Express5S and OSA-Express4S features on z13s.

7.3.62 QDIO optimized latency mode

QDIO optimized latency mode (OLM) can help improve performance for applications that have a critical requirement to minimize response times for inbound and outbound data.

OLM optimizes the interrupt processing in the following manner:

- ▶ For inbound processing, the TCP/IP stack looks more frequently for available data to process. This process ensures that any new data is read from the OSA-Express5S or OSA-Express4S without needing more program controlled interrupts (PCIs).
- ▶ For outbound processing, the OSA-Express5S or OSA-Express4S also look more frequently for available data to process from the TCP/IP stack. Therefore, the process does not require a Signal Adapter (SIGA) instruction to determine whether more data is available.

7.3.63 Large send for IPv6 packets

Large send for IPv6 packets improves performance by offloading outbound TCP segmentation processing from the host to an OSA-Express5S and OSA-Express4S feature by employing a more efficient memory transfer into OSA-Express5S and OSA-Express4S. Large send support for IPv6 packets applies to the OSA-Express5S and OSA-Express4S features (CHPID type OSD and OSX), and is exclusive to z13, z13s, zEC12, zBC12, z196, and z114 servers. With z13s, large send for IPv6 packets (segmentation offloading) for LPAR-to-LPAR traffic is supported. The minimum support requirements are listed in Table 7-50.

Table 7-50 Minimum support requirements for large send for IPv6 packets

Operating system	Support requirements
z/OS	z/OS V1R13 ^a
z/VM	z/VM V6R2 for guest use only

a. PTFs are required.

7.3.64 OSA-Express5S and OSA-Express4S checksum offload

OSA-Express5S and OSA-Express4S features, when configured as CHPID type OSD, provide checksum offload for several types of traffic, as indicated in Table 7-51.

Table 7-51 Minimum support requirements for OSA-Express5S and OSA-Express4S checksum offload

Traffic	Support requirements
LPAR to LPAR	z/OS V1R12 ^a z/VM V6R2 for guest use ^b
IPv6	z/OS V1R13 z/VM V6R2 for guest use ^b
LPAR-to-LPAR traffic for IPv4 and IPv6	z/OS V1R13 z/VM V6R2 for guest use ^b

a. PTFs are required.

b. Device is directly attached to guest, and PTFs are required.

7.3.65 Checksum offload for IPv4 and IPv6 packets when in QDIO mode

The *checksum offload function* supports z/OS and Linux on z Systems environments. It is offered on the OSA-Express5S GbE, OSA-Express5S 1000BASE-T Ethernet, OSA-Express4S GbE, and OSA-Express4S 1000BASE-T Ethernet features. Checksum offload provides the capability of calculating the Transmission Control Protocol (TCP), User Datagram Protocol (UDP), and IP header checksum. Checksum verifies the accuracy of files. By moving the checksum calculations to a Gigabit or 1000BASE-T Ethernet feature, host processor cycles are reduced and performance is improved.

When checksum is offloaded, the OSA-Express feature runs the checksum calculations for Internet Protocol version 4 (IPv4) and Internet Protocol version 6 (IPv6) packets. The checksum offload function applies to packets that go to or come from the LAN. When multiple IP stacks share an OSA-Express, and an IP stack sends a packet to a next hop address that is owned by another IP stack that is sharing the OSA-Express, OSA-Express sends the IP packet directly to the other IP stack. The packet does not have to be placed out on the LAN, which is termed LPAR-to-LPAR traffic. Checksum offload is enhanced to support the LPAR-to-LPAR traffic, which was not originally available.

Checksum offload is supported by the GbE features, which include FC 0404, FC 0405, FC 0413, and FC 0414. It also is supported by the 1000BASE-T Ethernet features, including FC 0408 and FC 0417, when it is operating at 1000 Mbps (1 Gbps). Checksum offload is applicable to the QDIO mode only (channel type OSD).

z/OS support for checksum offload is available in all in-service z/OS releases, and in all supported Linux on z Systems distributions.

7.3.66 Adapter interruptions for QDIO

Linux on z Systems and z/VM work together to provide performance improvements by using extensions to the QDIO architecture. Adapter interruptions, first added to z/Architecture with HiperSockets, provide an efficient, high-performance technique for I/O interruptions to reduce path lengths and processor usage. These reductions are in both the host operating system and the adapter (OSA-Express5S and OSA-Express4S when using CHPID type OSD).

In extending the use of adapter interruptions to OSD (QDIO) channels, the processor utilization to handle a traditional I/O interruption is reduced. This benefits OSA-Express TCP/IP support in z/VM, z/VSE, and Linux on z Systems.

Adapter interruptions apply to all of the OSA-Express5S and OSA-Express4S features on z13s when in QDIO mode (CHPID type OSD).

7.3.67 OSA Dynamic LAN idle

The OSA Dynamic LAN idle parameter change helps reduce latency and improve performance by dynamically adjusting the inbound blocking algorithm. System administrators can authorize the TCP/IP stack to enable a dynamic setting that previously was static.

For latency-sensitive applications, the blocking algorithm is modified to be latency-sensitive. For streaming (throughput-sensitive) applications, the blocking algorithm is adjusted to maximize throughput. In all cases, the TCP/IP stack determines the best setting based on the current system and environmental conditions, such as inbound workload volume, processor utilization, and traffic patterns. It can then dynamically update the settings. OSA-Express5S and OSA-Express4S features adapt to the changes, avoiding thrashing and frequent updates to the OAT. Based on the TCP/IP settings, OSA holds the packets before presenting them to the host. A dynamic setting is designed to avoid or minimize host interrupts.

OSA Dynamic LAN idle is supported by the OSA-Express5S, and OSA-Express4S features on z13s when in QDIO mode (CHPID type OSD). It is used by z/OS V1R12 and higher releases.

7.3.68 OSA Layer 3 virtual MAC for z/OS environments

To help simplify the infrastructure and facilitate load balancing when an LPAR is sharing an OSA MAC address with another LPAR, each operating system instance can have its own unique logical or virtual MAC (VMAC) address. All IP addresses that are associated with a TCP/IP stack are accessible by using their own VMAC address instead of sharing the MAC address of an OSA port. This situation also applies to Layer 3 mode and to an OSA port spanned among channel subsystems.

OSA Layer 3 VMAC is supported by the OSA-Express5S and OSA-Express4S features on z13s when in QDIO mode (CHPID type OSD). It is used by z/OS V1R12 and later.

7.3.69 QDIO Diagnostic Synchronization

QDIO Diagnostic Synchronization enables system programmers and network administrators to coordinate and simultaneously capture both software and hardware traces. It allows z/OS to signal OSA-Express5S and OSA-Express4S features (by using a diagnostic assist function) to stop traces and capture the current trace records.

QDIO Diagnostic Synchronization is supported by the OSA-Express5S and OSA-Express4S features on z13s when in QDIO mode (CHPID type OSD). It is used by z/OS V1R12 and later.

7.3.70 Network Traffic Analyzer

The z13s offers systems programmers and network administrators the ability to more easily solve network problems despite high traffic. With the OSA-Express Network Traffic Analyzer and QDIO Diagnostic Synchronization on the server, you can capture trace and trap data.

This data can then be forwarded to z/OS tools for easier problem determination and resolution.

The Network Traffic Analyzer is supported by the OSA-Express5S and OSA-Express4S features on z13s when in QDIO mode (CHPID type OSD). It is used by z/OS V1R12 and later.

7.3.71 Program-directed re-IPL

First available on System z9, program directed re-IPL allows an operating system on a z13s to re-IPL without operator intervention. This function is supported for both SCSI and IBM extended count key data (IBM ECKD™) devices. Table 7-52 lists the minimum support requirements for program directed re-IPL.

Table 7-52 Minimum support requirements for program directed re-IPL

Operating system	Support requirements
z/VM	z/VM V6R2
Linux on z Systems	SUSE Linux Enterprise Server 12 SUSE Linux Enterprise Server 11 Red Hat RHEL 7 Red Hat RHEL 6
z/VSE	V5R1 on SCSI disks

7.3.72 Coupling over InfiniBand and Integrated Coupling Adapter

InfiniBand (IFB) and ICA using PCIe Gen3 technology can potentially provide high-speed interconnection at short distances, longer distance fiber optic interconnection, and interconnection between partitions on the same system without external cabling. For more information, see 4.9, “Parallel Sysplex connectivity” on page 174.

Integrated Coupling Adapter

Support for PCIe Gen3 fanout (also known as Integrated Coupling Adapter Short Range (ICA SR)) that supports a maximum distance of 150 meters (492 feet) is listed in Table 7-53.

Table 7-53 Minimum support requirements for coupling links over InfiniBand

Operating system	Support requirements
z/OS	z/OS V1R12
z/VM	z/VM V6R2 (dynamic I/O support for ICA CHPIDs only; coupling over ICA is not supported for guest use.)
z/TPF	z/TPF V1R1

InfiniBand coupling links

Support for HCA3-O³ (12xIFB) fanout that supports InfiniBand coupling links 12x at a maximum distance of 150 meters (492 feet) is listed in Table 7-54.

Table 7-54 Minimum support requirements for coupling links over InfiniBand

Operating system	Support requirements
z/OS	z/OS V1R12
z/VM	z/VM V6R2 (dynamic I/O support for InfiniBand CHPIDs only; coupling over InfiniBand is not supported for guest use.)
z/TPF	z/TPF V1R1

InfiniBand coupling links at an unrepeated distance of 10 km (6.2 miles)

Support for HCA3-O LR⁴ (1xIFB) fanout that supports InfiniBand coupling links 1x at an unrepeated distance of 10 KM (6.2 miles) is listed in Table 7-55.

Table 7-55 Minimum support requirements for coupling links over InfiniBand at 10 km (6.2 miles)

Operating system	Support requirements
z/OS	z/OS V1R12
z/VM	z/VM V6R2 (dynamic I/O support for InfiniBand CHPIDs only; coupling over InfiniBand is not supported for guest use.)

7.3.73 Dynamic I/O support for InfiniBand and ICA CHPIDs

This function refers exclusively to the z/VM dynamic I/O support of InfiniBand and ICA coupling links. Support is available for the CIB and CS5 CHPID type in the z/VM dynamic commands, including the **change channel path** dynamic I/O command. Specifying and changing the system name when entering and leaving configuration mode are also supported. z/VM does not use InfiniBand or ICA, and does not support the use of InfiniBand or ICA coupling links by guests.

Table 7-56 lists the minimum support requirements for dynamic I/O support for InfiniBand CHPIDs.

Table 7-56 Minimum support requirements for dynamic I/O support for InfiniBand CHPIDs

Operating system	Support requirements
z/VM	z/VM V6R2

7.3.74 Simultaneous multithreading

SMT is the hardware capability to process up to two simultaneous threads in a single core, sharing the resources of the superscalar core. This capability improves the system capacity and the efficiency in the usage of the processor, increasing the overall throughput of the system.

SMT is supported only by zIIP and IFL speciality engines on z13s, and must be used by the operating system. An operating system with SMT support can be configured to dispatch work

³ HCA2-O is not supported on z13s.

⁴ HCA2-O is not supported on z13s.

to a thread on a zIIP (for eligible workloads in z/OS) or an IFL (for z/VM) core in single-thread or SMT mode. For more information, see 7.9, “Simultaneous multithreading” on page 295.

Table 7-57 lists the minimum support requirements for SMT.

Table 7-57 Minimum support requirements for SMT

Operating system	Support requirements
z/OS	z/OS V2R1 with APARs
z/VM	z/VM V6R3

Statement of Direction: IBM is working with its Linux distribution partners to include SMT support in future distribution releases.

All statements regarding IBM plans, directions, and intent are subject to change or withdrawal without notice. Any reliance on these statements of general direction is at the relying party’s sole risk and will not create liability or obligation for IBM.

7.3.75 Single Instruction Multiple Data

The SIMD feature introduces a new set of instructions with z13s to enable parallel computing that can accelerate code with string, character, integer, and floating point data types. The SIMD instructions allow a larger number of operands to be processed with a single complex instruction. For more information, see 3.4.2, “Single-instruction multiple-data” on page 90.

Table 7-58 lists the minimum support requirements for SIMD.

Table 7-58 Minimum support requirements for SIMD

Operating system	Support requirements
z/OS	z/OS V2R1 with small programming enhancement (SPE)
z/VM	z/VM V6.3 with PTFs for APAR VM65733

Statement of Direction: IBM is working with its Linux distribution partners to include SIMD support in future distribution releases.

All statements regarding IBM plans, directions, and intent are subject to change or withdrawal without notice. Any reliance on these statements of general direction is at the relying party’s sole risk and will not create liability or obligation for IBM.

7.3.76 Shared Memory Communication - Direct Memory Access

The SMC-D feature is introduced in the z13 and z13s servers. SMC-D maintains the socket-API transparency aspect of SMC-R so that applications using TCPI/IP communications can benefit immediately without requiring application software of IP topology changes. Similar to SMC-R, this protocol utilizes shared memory architectural concepts eliminating TCP/IP processing in the data path, yet preserving TCP/IP Qualities of Service for connection management purposes.

Table 7-59 lists the minimum support requirements for SMC-D.

Table 7-59 Minimum support requirements for SMC-D.

Operating system	Support requirements
z/OS	z/OS V2R2
z/VM	z/VM V6R3 with APAR VM65716 if running z/OS as guest

For more information, see Appendix D, “Shared Memory Communications” on page 475.

7.4 Cryptographic support

z13s provides two major groups of cryptographic functions:

- ▶ Synchronous cryptographic functions, which are provided by CPACF
- ▶ Asynchronous cryptographic functions, which are provided by the Crypto Express5S feature

The minimum software support levels are listed in the following sections. Obtain and review the current PSP buckets to ensure that the latest support levels are known and included as part of the implementation plan.

7.4.1 CP Assist for Cryptographic Function

In z13s, CPACF supports the following encryption types:

- ▶ The Advanced Encryption Standard (AES, symmetric encryption)
- ▶ The Data Encryption Standard (DES, symmetric encryption)
- ▶ The Secure Hash Algorithm (SHA, hashing)

For more information, see 6.4, “CP Assist for Cryptographic Functions” on page 207.

Table 7-60 lists the support requirements for CPACF on z13s.

Table 7-60 Support requirements for CPACF

Operating system	Support requirements
z/OS ^a	z/OS V1R12 and later with the Cryptographic Support for web deliverable that is specific for the operating system level
z/VM	z/VM V6R2 with PTFs and later: Supported for guest use
z/VSE	z/VSE V5R1 and later with PTFs
z/TPF	z/TPF V1R1
Linux on z Systems	SUSE Linux Enterprise Server 12 Red Hat RHEL 7

- a. CPACF also is used by several IBM software product offerings for z/OS, such as IBM WebSphere Application Server for z/OS.

7.4.2 Crypto Express5S

Support of *Crypto Express5S* functions varies by operating system and release. Table 7-61 lists the minimum software requirements for the Crypto Express5S features when configured as a coprocessor or an accelerator. For more information, see 6.5, “Crypto Express5S” on page 211.

Table 7-61 *Crypto Express5S support on z13s servers*

Operating system	Crypto Express5S
z/OS	z/OS V2R2, z/OS V2R1, z/OS V1R13, or z/OS V1R12 with the specific web deliverable
z/VM	For guest use: z/VM V6R2
z/VSE	z/VSE V5R1 with PTFs and later
z/TPF V1R1	Service required (accelerator mode only)
Linux on z Systems	IBM is working with its Linux distribution partners to include support in future Linux on z Systems distribution releases

7.4.3 Web deliverables

For web-deliverable code on z/OS, see the z/OS downloads website:

<http://www.ibm.com/systems/z/os/zos/downloads/>

For Linux on z Systems, support is delivered through IBM and the distribution partners. For more information, see Linux on z Systems on the IBM developerWorks® website:

<http://www.ibm.com/developerworks/linux/linux390/>

7.4.4 z/OS Integrated Cryptographic Service Facility FMIDs

Integrated Cryptographic Service Facility (ICSF) is a base component of z/OS. It is designed to transparently use the available cryptographic functions, whether CPACF or Crypto Express, to balance the workload and help address the bandwidth requirements of the applications.

Despite being a z/OS base component, ICSF functions are generally made available through web deliverable support a few months after a new z/OS release. Therefore, new functions are related to an ICSF function modification identifier (FMID) instead of a z/OS version.

For a list of ICSF versions and FMID cross-references, see the Technical Documents website:

<http://www.ibm.com/support/techdocs/atmastr.nsf/WebIndex/TD103782>

Table 7-62 lists the ICSF FMIDs and web-deliverable codes for z/OS V1R12 through V2R2. Later FMIDs include the functions of previous ones.

Table 7-62 z/OS ICSF FMIDs

ICSF FMID	z/OS	Web deliverable name	Supported function
HCR7770	V1R12 V1R11 V1R10	Cryptographic Support for z/OS V1R9-V1R11 Included as a base element of z/OS V1R12	<ul style="list-style-type: none"> ▶ Crypto Express3 and Crypto Express3-1P support ▶ PKA Key Management Extensions ▶ CPACF Protected Key ▶ Extended PKCS #11 ▶ ICSF Restructure (Performance, RAS, and ICSF-CICS Attach Facility)
HCR7780	V1R13 V1R12 V1R11 V1R10	Cryptographic Support for z/OS V1R10-V1R12 Included as a base element of z/OS V1R13	<ul style="list-style-type: none"> ▶ IBM zEnterprise 196 support ▶ Elliptic Curve Cryptography ▶ Message-Security-Assist-4 ▶ HMAC Support ▶ ANSI X9.8 Pin ▶ ANSI X9.24 (CBC Key Wrapping) ▶ CKDS constraint relief ▶ PCI Audit ▶ All callable services AMODE (64) ▶ PKA RSA OAEP with SHA-256 algorithm^a
HCR7790	V1R13 V1R12 V1R11	Cryptographic Support for z/OS V1R11-V1R13	<ul style="list-style-type: none"> ▶ Expanded key support for AES algorithm ▶ Enhanced ANSI TR-31 ▶ PIN block decimalization table protection ▶ Elliptic Curve Diffie-Hellman (ECDH) algorithm ▶ RSA in the Modulus-Exponent (ME) and Chinese Remainder Theorem (CRT) formats
HCR77A0	V2R1 V1R13 V1R12	Cryptographic Support for z/OS V1R12-V1R13 Included as a base element of z/OS V2R1	<ul style="list-style-type: none"> ▶ zEC12 & CEX4S Support, including Enterprise PKCS #11 (EP11) ▶ KDS Administration support for the PKDS (RSA-MK/ECC-MK) and TKDS (P11-MK) including improved I/O performance on these key datasets ▶ 24-byte DES Master Key support ▶ New controls for weak key wrapping ▶ DUKPT for MAC and Encryption Keys ▶ FIPS compliant RNG and Random Number cache ▶ Secure Cipher Text Translate ▶ EMV Enhancements for Amex

ICSF FMID	z/OS	Web deliverable name	Supported function
HCR77A1	V2R1 V1R13 V1R12	Cryptographic Support for z/OS V1R13-V2R1	<ul style="list-style-type: none"> ▶ AP Configuration Simplification ▶ KDS Key Utilization Statistics ▶ Dynamic SSM ▶ UDX Reduction and Simplification ▶ Europay-Mastercard-Visa (EMV) Enhancements ▶ Key wrapping and other security enhancements ▶ OWH/RNG Authorization Access ▶ SAF ACEE Selection ▶ Non-SAF Protected IQF ▶ RKX Key Export Wrapping ▶ AES MAC Enhancements ▶ PKCS #11 Enhancements ▶ Improved CTRACE Support
HCR77B0	V2R2 V2R1 V1R13	Cryptographic Support for z/OS V1R13-z/OS V2R1 Included as a base element of z/OS V2R2	<ul style="list-style-type: none"> ▶ z13 / z13s & CEX5 support, including support for sharing cryptographic coprocessors across a maximum of 85 (for z13) / 40 (for z13s servers) domains ▶ VISA Format Preserving Encryption (VFPE) services ▶ DK AES PIN and AES MAC Generate and Verify Services ▶ Support for exploitation of counter mode (CTR) for AES-based encryption on z196 and later processors ▶ Enhanced random number generation exploiting CPACF Deterministic Random Number Generate (DRNG) instruction with ability to disable RNG Cache ▶ Services and support for key archiving and key material validity; ▶ Enhancement to the ICSF Multi-Purpose service, CSFMPS, for change master key operation dry run
HCR77B1	V2R2 V2R1 V1R13	Cryptographic Support for z/OS V1R13 - z/OS V2R2	<ul style="list-style-type: none"> ▶ ICSF Console command support ▶ Support for EMV Simplification services ▶ Support for RSAES-OAEP formatting in PKA Decrypt and Encrypt services, Support in Key Generate for CIPHER, DATAC, and DATAM keys in OP, IM, or EX form ▶ Operational Key Load support for HMAC keys loaded from the TKE ▶ Additional DK AES PIN support

a. Service is required.

7.4.5 ICSF migration considerations

Consider the following if you have installed the Cryptographic Support for z/OS V1R13 – z/OS V2R1 web deliverable (FMID HCR77A1) to accommodate the change in the way master keys are processed to determine which coprocessors become active:

- ▶ The FMID HCR77A1 ICSF level is not integrated in z/OS V2R1 and needs to be downloaded and installed even after ordering a z/OS V2R1 ServerPac. The Cryptographic web deliverable is available at the following website:

<http://www.ibm.com/systems/z/os/zos/downloads/>

- ▶ Starting with FMID HCR77A1, the activation procedure now uses the master key verification patterns (MKVP) in the header record of the Cryptographic Key Data Set (CKDS) and PKDS to determine which coprocessors become active.
- ▶ You can use the IBM Health Checker check ICSFMIG77A1_UNSUPPORTED_HW to determine whether your current server can support HCR77A1. The migration check is available for HCR7770, HCR7780, HCR7790, and HCR77A0 through APAR OA42011.
- ▶ All systems in a sysplex that share a PKDS/TKDS must be at HCR77A1 to use the PKDS/TKDS Coordinated Administration support.

For more information, see *Migration from z/OS V1R13 and z/OS V1R12 to z/OS V2R1*, GA32-0889.

7.5 GDPS Virtual Appliance

The GDPS Virtual Appliance solution implements GDPS/PPRC Multiplatform Resilience for z Systems, also known as xDR. xDR coordinates near-continuous availability and a DR solution by using the following features:

- ▶ Disk error detection
- ▶ Heartbeat for smoke tests
- ▶ Re-IPL in place
- ▶ Coordinated site takeover
- ▶ Coordinated HyperSwap
- ▶ Single point of control

The GDPS Virtual Appliance software requirements are z/VM Version 5 Release 4 or later, or z/VM 6.2 or later. z/VM 5.4 is not supported on z13s.

For more information, see Appendix I, “GDPS Virtual Appliance” on page 541.

7.6 z/OS migration considerations

Except for base processor support, z/OS software changes do not require any of the functions that are introduced with the z13s. Also, the functions do not require functional software. The approach, where applicable, allows z/OS to automatically enable a function based on the presence or absence of the required hardware and software.

7.6.1 General guidelines

The IBM z13s introduces the latest z Systems technology. Although support is provided by z/OS starting with z/OS V1R12, use of z13s depends on the z/OS release. z/OS.e is *not* supported on z13s.

In general, consider the following guidelines:

- ▶ Do not change software releases and hardware at the same time.
- ▶ Keep members of the sysplex at the same software level, except during brief migration periods.
- ▶ Migrate to an STP-only network before introducing a z13s into a sysplex.
- ▶ Review z13s restrictions and migration considerations before creating an upgrade plan.

7.6.2 Hardware configuration definition

On z/OS V1R12 and later, the HCD or Hardware Configuration Manager (HCM) helps define a configuration for z13s.

7.6.3 Coupling links

Each system can use internal coupling links, InfiniBand coupling links, or ICA coupling links independently of what other systems are using. z13s does not support participating in a Parallel Sysplex with System z10 and earlier systems.

Coupling connectivity is available only when other systems also support the same type of coupling. When you plan to use the InfiniBand coupling or ICA coupling links technology, consult the *Coupling Facility Configuration Options* white paper, which is available at the following website:

<http://www.ibm.com/systems/z/advantages/ps0/whitepaper.html>

7.6.4 Large page support

The large page support function must not be enabled without the respective software support. If large page is not specified, page frames are allocated at the current size of 4 K.

In z/OS V1R9 and later, the amount of memory to be reserved for large page support is defined by using the **LFAREA** parameter in the IEASYSxx member of SYS1.PARMLIB:

```
LFAREA=xx%|xxxxxxM|xxxxxxG
```

The parameter indicates the amount of storage in percentage, megabytes, or gigabytes. The value cannot be changed dynamically.

7.6.5 Capacity Provisioning Manager

The installation of the capacity provision function on z/OS requires the following prerequisites:

- ▶ Setting up and customizing z/OS RMF, including the Distributed Data Server (DDS)
- ▶ Setting up the z/OS CIM Server (included in z/OS base)
- ▶ Performing capacity provisioning customization, as described in *z/OS MVS Capacity Provisioning User's Guide*, SA33-8299

Using the capacity provisioning function requires these prerequisites:

- ▶ TCP/IP connectivity to observed systems
- ▶ RMF Distributed Data Server must be active
- ▶ CIM server must be active
- ▶ Security and CIM customization
- ▶ Capacity Provisioning Manager customization

In addition, the Capacity Provisioning Control Center must be downloaded from the host and installed on a PC server. This application is used only to define policies. It is not required for regular operation.

Customization of the capacity provisioning function is required on the following systems:

- ▶ Observed z/OS systems: Systems in one or multiple sysplexes that are to be monitored. For more information about the capacity provisioning domain, see 8.8, “Nondisruptive upgrades” on page 349.
- ▶ Runtime systems: Systems where the Capacity Provisioning Manager is running, or to which the server can fail over after server or system failures.

7.6.6 Decimal floating point and z/OS XL C/C++ considerations

z/OS V1R13 with PTFs or higher is required to use the latest level (11) of the following C/C++ compiler options:

- ▶ **ARCHITECTURE**: This option selects the minimum level of system architecture on which the program can run. Certain features that are provided by the compiler require a minimum architecture level. ARCH(11) uses instructions that are available on the z13s.
- ▶ **TUNE**: This option allows optimization of the application for a specific system architecture, within the constraints that are imposed by the **ARCHITECTURE** option. The **TUNE** level must not be lower than the setting in the **ARCHITECTURE** option.

For more information about the **ARCHITECTURE** and **TUNE** compiler options, see the *z/OS V1R13.0 XL C/C++ User's Guide*, SC09-4767.

Important: Use the previous z Systems **ARCHITECTURE** or **TUNE** options for C/C++ programs if the same applications run on both the z13s and on previous z Systems servers. However, if C/C++ applications run only on z13s servers, use the latest **ARCHITECTURE** and **TUNE** options to ensure that the best performance possible is delivered through the latest instruction set additions.

For more information, see *Migration from z/OS V1R13 and z/OS V1R12 to z/OS V2R1*, GA32-0889.

7.7 IBM z Advanced Workload Analysis Reporter (zAware)

IBM zAware is designed to offer a real-time, continuous learning, diagnostic, and monitoring capability. This capability is intended to help you pinpoint and resolve potential problems quickly enough to minimize their impact on your business. IBM zAware runs analytics in firmware and intelligently examines the message logs for potential deviations, inconsistencies, or variations from the norm. Many z/OS environments produce such a large volume of OPERLOG messages that it is difficult for operations personnel to analyze them

easily. IBM zAware provides a simple GUI for easy identification of message anomalies, which can facilitate faster problem resolution.

IBM zAware is enhanced as part of z13s general availability to process messages without message IDs. This feature allows you to analyze the health of Linux operating systems by analyzing Linux SYSLOG. Linux on z Systems running natively and as a guest on z/VM is also supported. It must be a Linux distribution that has been tested for the z Systems server on which it runs.

7.7.1 z Appliance Container Infrastructure mode LPAR

IBM zAware is ordered through specific features of z13s, and requires z/OS V1R13 or later with IBM zAware exploitation support to collect specific log stream data. It requires a correctly configured LPAR. For more information, see “The z Appliance Container Infrastructure- mode LPAR” on page 244.

During a driver upgrade for z13 servers from Driver 22 to Driver 27, an existing zAware mode LPAR will be automatically converted to zACI LPAR type. With Driver 27 and later, any new zAware host will be deployed in an LPAR of type zACI.

To use the IBM zAware feature, complete the following tasks in z/OS:

1. For each z/OS that is to be monitored through the IBM zAware client, configure a network connection in the TCP/IP profile. If necessary, update the firewall settings.
2. Verify that each z/OS system meets the sysplex configuration and OPERLOG requirements for monitored clients of the IBM zAware virtual appliance.
3. Configure the z/OS system logger to send data to the IBM zAware virtual appliance server.
4. Prime the IBM zAware server with prior data from monitored clients.

For each Linux system to be monitored, complete these steps:

1. Configure a network connection. If necessary, update the firewall settings to ensure secure communications between IBM zAware and Linux clients.
2. Configure the syslog daemon of the Linux system to send messages to the IBM zAware server.

For more information about zAware, see Appendix B, “IBM z Systems Advanced Workload Analysis Reporter (IBM zAware)” on page 453 and the latest guide on Resource Link at the following website:

<https://www.ibm.com/servers/resource link/>

7.8 Coupling facility and CFCC considerations

Coupling facility connectivity to a z13s is supported on the z13, zEC12, zBC12, z196, z114, or another z13s server. The LPAR running the CFCC can be on any of the previously listed supported systems. For more information about CFCC requirements for supported systems, see Table 7-63 on page 294.

Consideration: Because coupling link connectivity to System z10 and previous systems is not supported, introduction of z13s into existing installations might require more planning. Also, consider the level of CFCC. For more information, see “Coupling link considerations” on page 181.

7.8.1 CFCC Level 21

CFCC level 21 is delivered on the z13s with driver level 27. CFCC Level 21 introduces the following enhancements:

- ▶ Usability enhancement:
 - Enable systems management applications to collect valid coupling facility (CF) LPAR information by using z/OS BCPii.
 - System type (CFCC), System Level (CFCC level).
 - Dynamic dispatch settings to indicate CF state (dedicated, shared, thin interrupt).
- ▶ Availability enhancement:
 - Asynchronous CF duplexing for lock structures. Secondary structure updates are performed asynchronously with respect to primary updates
 - Designed to drive out cross-site latencies that exist today when replicating CF data across distance
 - Designed to avoid the need for synchronous speed-of-light communication delays during the processing of every duplexed update operation
 - Improves performance with cross-site duplexing of lock structures at distance
 - Maintains robust failure recovery capability through the redundancy of duplexing
 - Reduces z/OS, CF, and link utilization overhead costs associated with synchronous duplexing of lock structures
 - Requires CFCC level 21 and service level 02.16 and z/OS V2.2 SPE with PTFs for APAR OA47796.
- ▶ Large memory support:
 - Improve availability/scalability for larger CF cache structures and data sharing performance with larger DB2 group buffer pools (GBPs).
 - This support removes inhibitors to using large CF structures, enabling use of Large Memory to scale to larger DB2 local buffer pools (LBPs) and GBPs in data sharing environments.
 - CF structure size remains at a maximum of 1 TB.
- ▶ Support for the new IBM ICA for short distance coupling.

z13s systems with CFCC Level 21 require z/OS V1R12 or later, and z/VM V6R2 or later for guest virtual coupling.

To support an upgrade from one CFCC level to the next, different levels of CFCC can be run concurrently while the coupling facility LPARs are running on different servers. CF LPARs that run on the same server share the CFCC level. The CFCC level for z13s servers is CFCC Level 21, as shown in Table 7-63.

z13s (CFCC level 21) can coexist in a sysplex with CFCC levels 17 and 19.

Table 7-63 z Systems CFCC code-level considerations

z Systems	Code level
z13	CFCC Level 20 or CFCC Level 21
z13s	CFCC Level 21
zEC12	CFCC Level 18 ^a or CFCC Level 19
zBC12	CFCC Level 19
z196 and z114	CFCC Level 17
z10 EC or z10 BC	CFCC Level 15 ^a or CFCC Level 16 ^a

a. This CFCC level cannot coexist in the same sysplex with CFCC level 20 or 21.

For more information about CFCC code levels, see the Parallel Sysplex website at:

<http://www.ibm.com/systems/z/psocftable.html>

For the latest CFCC code levels, see the current exception letter that is published on Resource Link at the following website:

<https://www.ibm.com/servers/resourceLink/lib03020.nsf/pages/exceptionLetters>

CF structure sizing changes are expected when upgrading from a previous CFCC Level to CFCC Level 21. Review the CF LPAR size by using the CFSizer tool with is available at the following website:

<http://www.ibm.com/systems/z/cfsizer>

Sizer Utility, an authorized z/OS program download, is useful when you are upgrading a CF. It is available at the following web page:

<http://www.ibm.com/systems/support/z/cfsizer/altsizer.html>

Before the migration, install the compatibility/coexistence PTFs. A planned outage is required when you upgrade the CF or CF LPAR to CFCC Level 21.

7.8.2 Flash Express exploitation by CFCC

CFCC Level 19 is the minimum level that supports Flash Express. Initial CF Flash Express exploitation is targeted for WebSphere MQ shared queue application structures. It is designed to help improve resilience and provide cost-effective standby capacity to help manage the potential overflow of WebSphere MQ shared queues. Structures now can be allocated with a combination of real memory and SCM that is provided by the Flash Express feature.

Flash memory in the CPC is assigned to a CF partition through hardware definition windows, which is similar to how it is assigned to the z/OS partitions. The CFRM policy definition allows the maximum amount of flash memory that you want to be used by a particular structure, on a structure-by-structure basis.

Important: Flash memory is *not* pre-assigned to structures at allocation time.

Structure size requirements for real memory get larger at initial allocation time to accommodate more control objects that are needed to use flash memory.

The CFSIZER structure recommendations consider these additional requirements, both for sizing the structure's flash usage and for the related real memory considerations.

The following are the minimum CFCC Flash Express use requirements:

- ▶ CFCC Level 19 support or later
- ▶ z/OS support for z/OS V1R13 with PTFs and z/OS V2R1 or later with PTFs

WebSphere MQ Version 7 is required.

7.8.3 CFCC Coupling Thin Interrupts

The Coupling Thin Interrupts enhancement, which is delivered with CFCC 19, improves the performance of a Coupling Facility partition and the dispatching of z/OS LPARs awaiting the arrival of returned asynchronous CF requests, when used in a shared engine environment.

7.9 Simultaneous multithreading

The z13s can run up two threads simultaneously in the same processor, dynamically sharing resources of the core such as cache, TLB, and execution resources. It provides better utilization of the cores and more processing capacity. This function is known as SMT, and is available only in zIIP and IFL cores. For more information about SMT, see 3.4.1, “Simultaneous multithreading” on page 89.

The z/OS and the z/VM have SMT support if the support is enabled by PTFs. z/OS 2.2 supports the operation of zIIP processors in SMT mode.

The following APARs must be applied to z/OS V2R1 to use SMT⁵:

- ▶ OA43366 (BCP)
- ▶ OA43622 (WLM)
- ▶ OA44439 (XCF)

The use of SMT on z/OS V2R1 and later requires enabling HiperDispatch, and defining the processor view (**PROCVIEW**) control statement in the LOADxx parmlib member and the **MT_ZIIP_MODE** parameter in the IEAOPTxx parmlib member.

The **PROCVIEW** statement is defined for the life of IPL, and can have the following values:

- ▶ **CORE**: This value specifies that z/OS should configure a processor view of core, where a core can have one or more threads. The number of threads is limited by z13s to two. If the underlying hardware does not support SMT, a core is limited to one thread.
- ▶ **CPU**: This value is the default. It specifies that z/OS should configure a traditional processor view of CPU and not use SMT.
- ▶ **CORE,CPU_OK**: This value specifies that z/OS should configure a processor view of core, as with the **CORE** value, but the CPU parameter is accepted as an alias for applicable commands.

⁵ SMT is only available for zIIP workload.

When **PROCVIEW CORE** or **CORE,CPU_OK** are specified in z/OS running in z13s, HiperDispatch is forced to run as enabled. You cannot disable it. The **PROCVIEW** statement cannot be changed dynamically, so you must run an IPL after changing it to make the new setting effective.

The **MT_ZIIP_MODE** parameter in the IEAOPTxx controls zIIP SMT mode. This parameter can be 1 (the default), where only one thread can be running in a core, or 2, where up two threads can be running in a core. If **PROCVIEW CPU** is specified, the **MT_ZIIP_MODE** is always 1. Otherwise, the use of SMT to dispatch two threads in a single zIIP logical processor (**MT_ZIIP_MODE=2**) can be changed dynamically by using the **SET OPT=xx** setting in the IEAOPTxx parmlib. Changing the MT mode for all cores can take some time to complete.

The activation of SMT mode also requires that the HMC Customize/Delete Activation Profiles task, "Do not end the time slice if a partition enters a wait state", must not be selected. This is the recommended default setting.

PROCVIEW CORE requires **DISPLAY M=CORE** and **CONFIG CORE** to display the core states and configure an entire core.

Figure 7-1 shows the result of the display core command with processor view core and SMT enabled.

```

-D M=CPU
IEE174I 21.26.22 DISPLAY M 557
CORE STATUS: HD=Y MT=2 MT_MODE CP=1 zIIP=2
ID ST ID RANGE VP ISCM CRU THREAD STATUS
0000 + 0000-0001 H 0000 +N
0001 + 0002-0003 H 0000 +N
0002 + 0004-0005 H 0000 +N
0003 + 0006-0007 H FC00 +N
0004 - 0008-0009
0005 - 000A-000B
0006 +I 000C-000D H 0200 ++
0007 - 000E-000F
0008 - 0010-0011
0009 -I 0012-0013

CPCND = 002965.N20IBM.02.00000008DA87
CPCSI = 2965.N20.IBM.02.000000000008DA87
Model: N20
CPCID = 00
CPCNAME = SCZP501
LP NAME = A01 LP ID = 1
CSS ID = 0
MIF ID = 1

+ ONLINE - OFFLINE N NOT AVAILABLE / MIXED STATE
W WLM-MANAGED

I INTEGRATED INFORMATION PROCESSOR (zIIP)
CPCND CENTRAL PROCESSING COMPLEX NODE DESCRIPTOR
CPCSI SYSTEM INFORMATION FROM STSI INSTRUCTION
CPCID CENTRAL PROCESSING COMPLEX IDENTIFIER
CPCNAME CENTRAL PROCESSING COMPLEX NAME
LP NAME LOGICAL PARTITION NAME
LP ID LOGICAL PARTITION IDENTIFIER
CSS ID CHANNEL SUBSYSTEM IDENTIFIER
MIF ID MULTIPLE IMAGE FACILITY IMAGE IDENTIFIER
***** BOTTOM OF DATA *****

```

Figure 7-1 Result of the display core command

The use of SMT in z/VM V6R3⁶ requires the application of a small programming enhancement (SPE) that has a new **MULTITHREADING** statement in the system configuration file, has HiperDispatch enabled, and has the dispatcher work distribution mode set to *reshuffle*.

⁶ The z/VM 6.3 SMT enablement APAR is VM65586.

The default in z/VM is multithreading disabled by default. There is no dynamic switching of the SMT mode. It can be enabled or disabled only by setting the **MULTITHREADING** statement in the configuration file. After the statement is set, you must run an IPL of the partition to enable or disable SMT.

z/VM supports up to 32 multithreaded cores (64 threads) for IFLs, and each thread is treated as an independent processor. z/VM dispatches virtual IFLs on the IFL logical processor so that the same or different guests can share a core. There is a single dispatch vector per core, and z/VM attempts to place virtual sibling IFLs on the same dispatch vector to maximize cache reuses. The guests have no awareness of SMT, and they cannot use it.

z/VM SMT exploitation does not include guest support for multithreading. The value of this support for guests is that the first-level z/VM hosts under the guests can achieve higher throughput from the multi-threaded IFL cores.

IBM is working with its Linux distribution partners to support SMT. KVM for IBM z Systems SMT support is planned for 1Q2016.

An operating system that uses SMT controls each core and is responsible for maximizing their throughput and meeting workload goals with the smallest number of cores. In z/OS, HiperDispatch cache optimization should be considered when you must choose the two threads to be dispatched in the same processor. HiperDispatch attempts to dispatch guest virtual CPU on the same logical processor on which they have run previously. PR/SM attempts to dispatch a vertical low logical processor in the same physical processor. If that is not possible, PR/SM attempt to dispatch in the same node, or if that is not possible, in the same CPC drawer where it was dispatched before to maximize cache reuse.

From the point of view of an application, SMT is transparent and no changes are required in the application for it to run in an SMT environment, as shown in Figure 7-2.

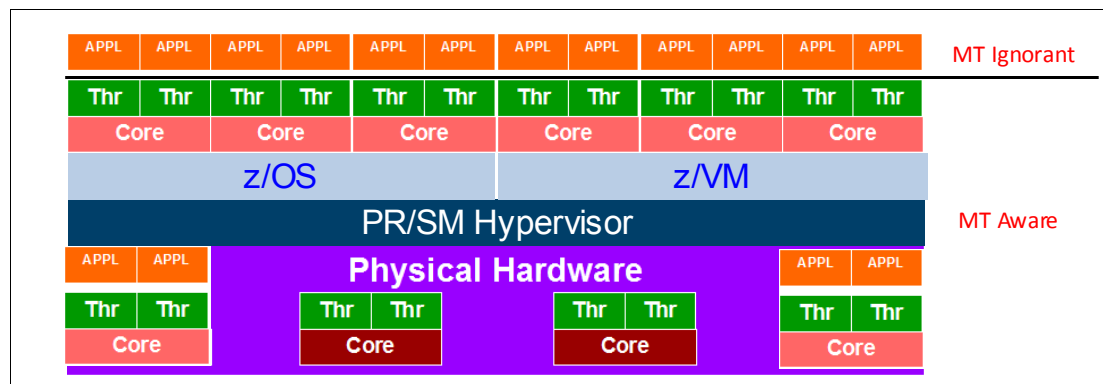


Figure 7-2 Simultaneous Multithreading

7.10 Single-instruction multiple-data

z13s is equipped with new set of instructions to improve the performance of complex mathematical models and analytic workloads through vector processing and new complex instructions, which can process a great deal of data with only a single instruction.

This new set of instructions, which is known as SIMD, enables more consolidation of analytic workloads and business transactions on z Systems.

z/OS V2R1 has support for SIMD through an SPE. The z/VM support for SIMD support will be delivered for z13/z13s with PTF for APAR WM65733 for guest exploitation.

OS support includes the following items:

- ▶ Enablement of vector registers.
- ▶ Use of vector registers using XL C/C++ ARCH(11) and TUNE(11).
- ▶ A math library with an optimized and tuned math function (Mathematical Acceleration Subsystem (MASS)) that can be used in place of some of the C standard math functions. It has a SIMD vectorized and non-vectorized version.
- ▶ A specialized math library, which is known as Automatically Tuned Linear Algebra Software (ATLAS), which is optimized for the hardware.
- ▶ IBM Language Environment® for C runtime function enablement for ATLAS.
- ▶ DBX to support the disassembly of the new vector instructions, and to display and set vector registers.
- ▶ XML SS exploitation to use new vector processing instructions to improve performance.

MASS and ATLAS can reduce the time and effort for middleware and application developers. IBM provides compiler built-in functions for SIMD that software applications can use as needed, such as for using string instructions.

The use of new hardware instructions through XL C/C++ ARCH(11) and TUNE(11) or SIMD usage by MASS and ATLAS libraries requires the z13s support for z/OS V2R1 XL C/C++ web deliverable.

The followings compilers have built-in functions for SIMD:

- ▶ IBM Java
- ▶ XL C/C++
- ▶ Enterprise COBOL
- ▶ Enterprise PL/I

Code must be developed to take advantage of the SIMD functions, and applications with SIMD instructionsabend if they run on a lower hardware level system. Some mathematical function replacement can be done without code changes by including the scalar MASS library before the standard math library.

The MASS and standard math library have different accuracies. Therefore, you need to assess the accuracy of the functions in the context of the user application when you decide whether to use the MASS and ATLAS libraries.

The SIMD functions can be disabled in z/OS partitions at IPL time by using the **MACHMIG** parameter in the LOADxx member. To disable SIMD code, use the MACHMIG VEF hardware-based vector facility. If you do not specify a **MACHMIG** statement, which is the default, the system will be unlimited in its use of the Vector Facility for z/Architecture (SIMD).

7.11 The MIDAW facility

The Modified Indirect Data Address Word (MIDAW) facility is a system architecture and software exploitation that is designed to improve FICON performance. This facility was first made available on System z9 servers, and is used by the Media Manager in z/OS.

The MIDAW facility provides a more efficient CCW/IDAW structure for certain categories of data-chaining I/O operations:

MIDAW can improve FICON performance for extended format data sets. Non-extended data sets can also benefit from MIDAW.

MIDAW can improve channel utilization, and can improve I/O response time. It reduces FICON channel connect time, director ports, and control unit processor usage.

IBM laboratory tests indicate that applications that use EF data sets, such as DB2, or long chains of small blocks can gain significant performance benefits by using the MIDAW facility.

MIDAW is supported on FICON channels that are configured as CHPID type FC.

7.11.1 MIDAW technical description

An IDAW is used to specify data addresses for I/O operations in a virtual environment.⁷ The existing IDAW design allows the first IDAW in a list to point to any address within a page. Subsequent IDAWs in the same list must point to the first byte in a page. Also, IDAWs (except the first and last IDAW) in a list must deal with complete 2 K or 4 K units of data.

Figure 7-3 shows a single CCW that controls the transfer of data that spans non-contiguous 4 K frames in main storage. When the IDAW flag is set, the data address in the CCW points to a list of words (IDAWs). Each IDAW contains an address that designates a data area within real storage.

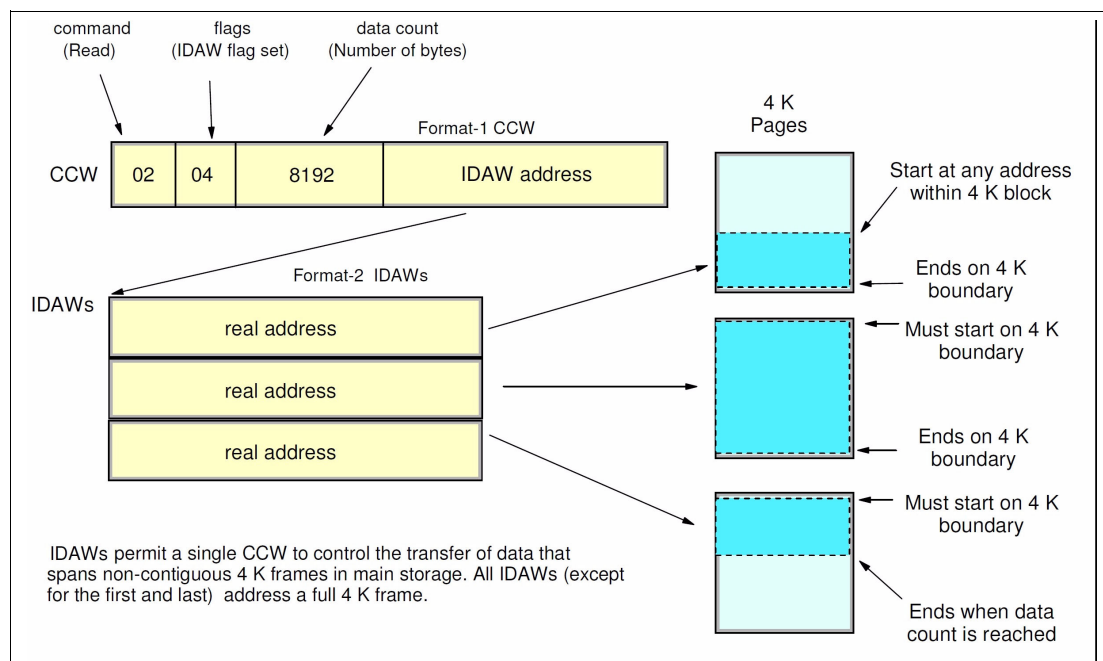


Figure 7-3 IDAW usage

The number of required IDAWs for a CCW is determined by these factors:

- ▶ The IDAW format as specified in the operation request block (ORB)
- ▶ The count field of the CCW

⁷ There are exceptions to this statement, and many details are skipped in this description. This section assumes that you can merge this brief description with an existing understanding of I/O operations in a virtual memory environment.

- ▶ The data address in the initial IDAW

For example, three IDAWS are required when these events occur:

- ▶ The ORB specifies format-2 IDAWs with 4 KB blocks.
- ▶ The CCW count field specifies 8 KB.
- ▶ The first IDAW designates a location in the middle of a 4 KB block.

CCWs with data chaining can be used to process I/O data blocks that have a more complex internal structure, in which portions of the data block are directed into separate buffer areas. This process is sometimes known as scatter-read or scatter-write. However, as technology evolves and link speed increases, data chaining techniques become less efficient because of switch fabrics, control unit processing and exchanges, and other issues.

The MIDAW facility is a method of gathering and scattering data from and into discontinuous storage locations during an I/O operation. The modified IDAW (MIDAW) format is shown in Figure 7-4. It is 16 bytes long and is aligned on a quadword.

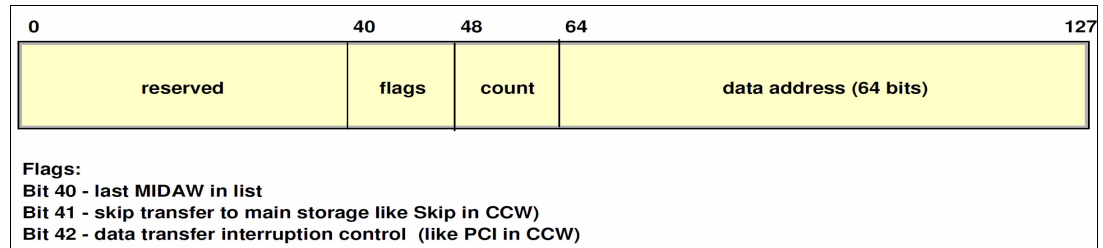


Figure 7-4 MIDAW format

An example of MIDAW usage is shown in Figure 7-5.

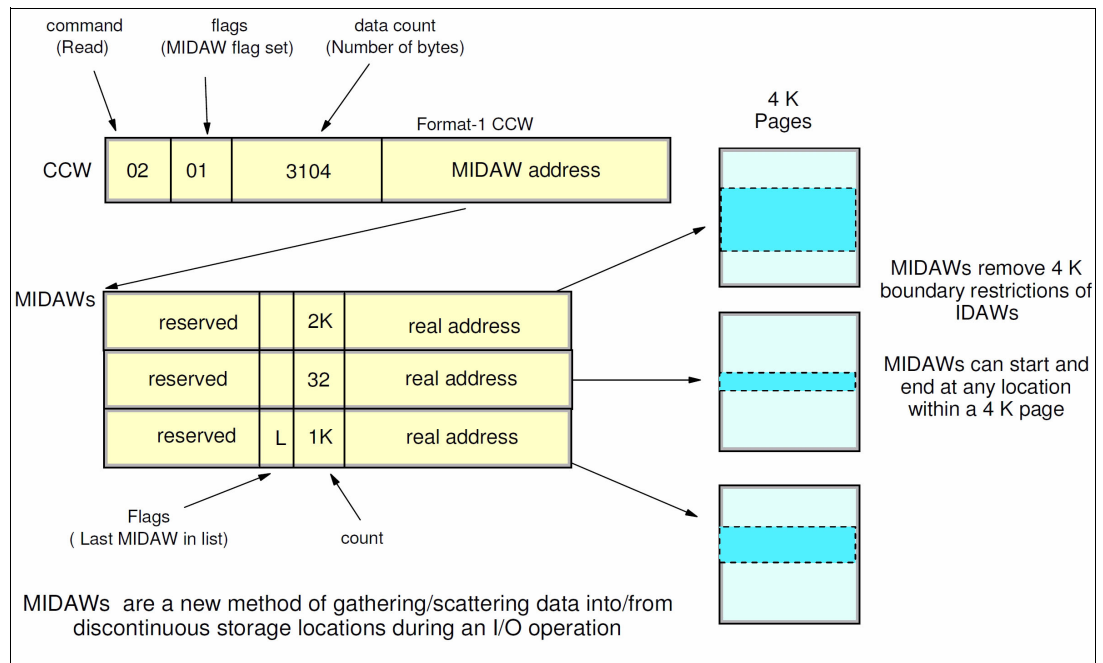


Figure 7-5 MIDAW usage

The use of MIDAWs is indicated by the MIDAW bit in the CCW. If this bit is set, the skip flag cannot be set in the CCW. The skip flag in the MIDAW can be used instead. The data count in the CCW must equal the sum of the data counts in the MIDAWs. The CCW operation ends

when the CCW count goes to zero or the last MIDAW (with the last flag) ends. The combination of the address and count in a MIDAW cannot cross a page boundary. This configuration means that the largest possible count is 4 K. The maximum data count of all the MIDAWs in a list cannot exceed 64 K, which is the maximum count of the associated CCW.

The scatter-read or scatter-write effect of the MIDAWs makes it possible to efficiently send small control blocks that are embedded in a disk record to separate buffers from those used for larger data areas within the record. MIDAW operations are on a single I/O block, in the manner of data chaining. Do not confuse this operation with CCW command chaining.

7.11.2 Extended format data sets

z/OS extended format (EF) data sets use internal structures (usually not visible to the application program) that require a scatter-read (or scatter-write) operation. Therefore, CCW data chaining is required, which produces less than optimal I/O performance. Because the most significant performance benefit of MIDAWs is achieved with EF data sets, a brief review of the EF data sets is included here.

Both VSAM and non-VSAM (DSORG=PS) sets can be defined as extended format data sets. For non-VSAM data sets, a 32-byte suffix is appended to the end of every physical record (that is, block) on disk. VSAM appends the suffix to the end of every control interval (CI), which normally corresponds to a physical record. A 32 K CI is split into two records to span tracks. This suffix is used to improve data reliability and facilitates other functions that are described in the following paragraphs. Therefore, for example, if the DCB BLKSIZE or VSAM CI size is equal to 8192, the actual block on storage consists of 8224 bytes. The control unit itself does not distinguish between suffixes and user data. The suffix is transparent to the access method and database.

In addition to reliability, EF data sets enable three other functions:

- ▶ DFSMS striping
- ▶ Access method compression
- ▶ Extended addressability (EA)

EA is useful for creating large DB2 partitions (larger than 4 GB). Striping can be used to increase sequential throughput, or to spread random I/Os across multiple logical volumes. DFSMS striping is useful for using multiple channels in parallel for one data set. The DB2 logs are often striped to optimize the performance of DB2 sequential inserts.

Processing an I/O operation to an EF data set normally requires at least two CCWs with data chaining. One CCW is used for the 32-byte suffix of the EF data set. With MIDAW, the additional CCW for the EF data set suffix is eliminated.

MIDAWs benefit both EF and non-EF data sets. For example, to read twelve 4 K records from a non-EF data set on a 3390 track, Media Manager chains 12 CCWs together by using data chaining. To read twelve 4 K records from an EF data set, 24 CCWs are chained (two CCWs per 4 K record). Using Media Manager track-level command operations and MIDAWs, an entire track can be transferred by using a single CCW.

7.11.3 Performance benefits

z/OS Media Manager has I/O channel program support for implementing EF data sets, and automatically uses MIDAWs when appropriate. Most disk I/Os in the system are generated by using Media Manager.

Users of the Executing Fixed Channel Programs in Real Storage (EXCPVR) instruction can construct channel programs that contain MIDAWs. However, doing so requires that they construct an IOBE with the IOBEMIDA bit set. Users of the EXCP instruction cannot construct channel programs that contain MIDAWs.

The MIDAW facility removes the 4 K boundary restrictions of IDAWs and reduces the number of CCWs for EF data sets. Decreasing the number of CCWs helps to reduce the FICON channel processor utilization. Media Manager and MIDAWs do not cause the bits to move any faster across the FICON link. However, they reduce the number of frames and sequences that flow across the link, so they use the channel resources more efficiently.

The MIDAW facility with FICON Express8S, operating at 8 Gbps, shows an improvement in throughput for all reads on DB2 table scan tests with EF data sets compared to the use of IDAWs with FICON Express2, operating at 2 Gbps.

The performance of a specific workload can vary according to the conditions and hardware configuration of the environment. IBM laboratory tests found that DB2 gains significant performance benefits by using the MIDAW facility in the following areas:

- ▶ Table scans
- ▶ Logging
- ▶ Utilities
- ▶ Use of DFSMS striping for DB2 data sets

Media Manager with the MIDAW facility can provide significant performance benefits when used with applications that use EF data sets (such as DB2) or long chains of small blocks.

For more information about FICON and MIDAW, see the following resources:

- ▶ The I/O Connectivity website contains material about FICON channel performance:
<http://www.ibm.com/systems/z/connectivity/>
- ▶ *DS8000 Performance Monitoring and Tuning*, SG24-7146

7.12 IOCP

All z Systems servers require a description of their I/O configuration. This description is stored in input/output configuration data set (IOCDs) files. The input/output configuration program (IOCP) allows the creation of the IOCDs file from a source file that is known as the *input/output configuration source (IOCS)*.

z13s IOCP definitions: A new parameter for HiperSockets IOCP definitions on z13s servers means that the z13s IOCP definitions need to be migrated to support the HiperSockets definitions (CHPID type IQD).

On z13s servers, the CHPID statement of HiperSockets devices requires the keyword VCHID. VCHID specifies the virtual channel identification number associated with the channel path. Valid range is 7E0 - 7FF. VCHID is not valid on z Systems before z13s servers.

IOCP required level for z13s servers: The required level of IOCP for z13s servers is V5 R1 L0 (IOCP 5.1.0) or later with PTFs. For more information, see the following manuals:

- ▶ *z Systems Stand-Alone Input/Output Configuration Program User's Guide*, SB10-7166.
- ▶ *z Systems Input/Output Configuration Program User's Guide for ICP IOCP*, SB10-7163.

The IOCS file contains detailed information for each channel and path assignment, each control unit, and each device in the configuration.

7.13 Worldwide port name tool

Part of the installation of your z13s system is the pre-planning of the SAN environment. IBM has a stand-alone tool to assist with this planning before the installation.

The capabilities of the WWPN are extended to calculate and show WWPNs for both virtual and physical ports ahead of system installation.

The tool assigns WWPNs to each virtual FCP channel/port by using the same WWPN assignment algorithms that a system uses when assigning WWPNs for channels using NPIV. Therefore, the SAN can be set up in advance, allowing operations to proceed much faster after the server is installed. In addition, the SAN configuration can be retained instead of altered by assigning the WWPN to physical FCP ports when a FICON feature is replaced.

The WWPN tool takes a .csv file that contains the FCP-specific I/O device definitions and creates the WWPN assignments that are required to set up the SAN. A binary configuration file that can be imported later by the system is also created. The .csv file can either be created manually, or exported from the HCD/HCM.

The WWPN tool on z13s (CHPID type FCP) requires the following levels:

- ▶ z/OS V1R12 and later
- ▶ z/VM V6R2 and later

The WWPN tool is applicable to all FICON channels that are defined as CHPID type FCP (for communication with SCSI devices) on z13s. It is available for download at the Resource Link at the following website:

<http://www.ibm.com/servers/resourceLink/>

Note: An optional feature can be ordered for WWPN persistency before shipment to keep the same I/O serial number on the new CEC. Current information must be provided during the ordering process.

7.14 ICKDSF

Device Support Facilities (ICKDSF) Release 17 is required on all systems that share disk subsystems with a z13s processor.

ICKDSF supports a modified format of the CPU information field that contains a two-digit LPAR identifier. ICKDSF uses the CPU information field instead of CCW reserve/release for concurrent media maintenance. It prevents multiple systems from running ICKDSF on the same volume, and at the same time allows user applications to run while ICKDSF is processing. To prevent data corruption, ICKDSF must be able to determine all sharing systems that can potentially run ICKDSF. Therefore, this support is required for z13s.

Remember: The need for ICKDSF Release 17 applies even to systems that are not part of the same sysplex, or are running an operating system other than z/OS, such as z/VM.

7.15 IBM z BladeCenter Extension (zBX) Model 004 software support

zBX Model 004 houses two types of blades:

- ▶ Power platform specific
- ▶ IBM Blades for Linux and Windows operating systems

For information about zBX upgrades, see 8.3.4, “MES upgrades for the zBX” on page 327.

7.15.1 IBM Blades

IBM offers a selected subset of IBM POWER7 blades that can be installed and operated on the zBX Model 004.

The blades are virtualized by PowerVM Enterprise Edition. Their LPARs run either AIX Version 5 Release 3 technology level (TL) 12 (IBM POWER6® mode), AIX Version 6 Release 1 TL5 (POWER7 mode), or AIX Version 7 Release 1 and subsequent releases. Applications that are supported on AIX can be deployed to blades.

Also offered are selected IBM System x HX5 blades. Virtualization is provided by an integrated kernel-based hypervisor (xHyp), supporting Linux on System x and Microsoft Windows operating systems.

Table 7-64 lists the operating systems that are supported by HX5 blades.

Table 7-64 Operating system support for zBX Model 003 HX5 blades

Operating system	Support requirements
Linux on System x	Red Hat RHEL 5.5 and later, 6.0 and later, and 7.0 and later SUSE Linux Enterprise Server 10 (SP4) and later, SUSE Linux Enterprise Server 11 (SP1) ^a and later, and SUSE Linux Enterprise Server 12.0 and later
Microsoft Windows	Microsoft Windows Server 2008 R2 ^b Microsoft Windows Server 2008 (SP2) ^b (Datacenter Edition preferred) Microsoft Windows Server 2012 ^b (Datacenter Edition preferred) Microsoft Windows Server 2012 ^b R2

a. Latest patch level required

b. 64-bit only

7.15.2 IBM WebSphere DataPower Integration Appliance XI50 for zEnterprise

The IBM WebSphere DataPower Integration Appliance XI50 for zEnterprise (DataPower XI50z) is a special-purpose, double-wide blade.

The DataPower XI50z is a multifunctional appliance that provides these features:

- ▶ Offers multiple levels of XML optimization
- ▶ Streamlines and secures valuable service-oriented architecture (SOA) applications
- ▶ Provides drop-in integration for heterogeneous environments by enabling core enterprise service bus (ESB) functions, including routing, bridging, transformation, and event handling
- ▶ Simplifies, governs, and enhances the network security for XML and web services

Table 7-65 lists the minimum support requirements for DataPower Sysplex Distributor support.

Table 7-65 Minimum support requirements for DataPower Sysplex Distributor support

Operating system	Support requirements
z/OS	z/OS V1R11 for IPv4 z/OS V1R12 for IPv4 and IPv6

7.16 Software licensing

This section describes the software licensing options that are available for the z13s. Basic information about software licensing for the IBM z BladeCenter Extension (zBX) Model 004 environments is also covered.

7.16.1 Software licensing considerations

The IBM z13s software portfolio includes operating system software (that is, z/OS, z/VM, z/VSE, and z/TPF) and middleware that runs on these operating systems. The portfolio also includes middleware for Linux on z Systems environments.

zBX software products are covered by the International Program License Agreement and more agreements, such as the IBM International Passport Advantage® Agreement, similar to other AIX, Linux on System x, and Windows environments. PowerVM Enterprise Edition licenses must be ordered for IBM POWER7 blades.

For the z13s, two metric groups for software licensing are available from IBM, depending on the software product:

- ▶ Monthly license charge (MLC)
- ▶ International Program License Agreement (IPLA)

MLC pricing metrics have a recurring charge that applies each month. In addition to the permission to use the product, the charge includes access to IBM product support during the support period. MLC metrics, in turn, include various offerings.

IPLA metrics have a single, up-front charge for an entitlement to use the product. An optional and separate annual charge, called subscription and support, entitles clients to access IBM product support during the support period. With this option, you can also receive future releases and versions at no additional charge.

For more information about software licensing, see the following websites:

- ▶ Learn about Software licensing:
http://www.ibm.com/software/lotus/passportadvantage/about_software_licensing.html
- ▶ Base license agreements:
<http://www.ibm.com/software/sla/slabd.nsf/sla/bla/>
- ▶ IBM z Systems Software Pricing Reference Guide:
<http://www.ibm.com/systems/z/resources/swprice/reference/index.html>
- ▶ IBM z Systems Software Pricing:
<http://www.ibm.com/systems/z/resources/swprice/index.html>

- ▶ The IBM International Passport Advantage Agreement can be downloaded from the “Learn about Software licensing” website:

ftp://ftp.software.ibm.com/software/passportadvantage/PA_Agreements/PA_Agreement_International_English.pdf

The remainder of this section describes the software licensing options that are available for the z13s.

7.16.2 Monthly license charge pricing metrics

MLC pricing applies to z/OS, z/VSE, and z/TPF operating systems. Any mix of z/OS, z/VM, Linux, z/VSE, and z/TPF images is allowed. Charges are based on processor capacity, which is measured in millions of service units (MSU) per hour.

Charge models

Various Workload License Charges (WLC) pricing structures support two charge models:

- ▶ Variable charges (several pricing metrics):

Variable charges apply to products such as z/OS, z/VSE, z/TPF, DB2, IMS, CICS, and WebSphere MQ. Several pricing metrics employ the following charge types:

- Full-capacity license charges:

The total number of MSUs of the CPC is used for charging. Full-capacity licensing is applicable when the CPC of the client is not eligible for subcapacity.

- Subcapacity license charges:

Software charges that are based on the utilization of the logical partitions where the product is running.

- ▶ Flat charges:

Software products that are licensed under flat charges are not eligible for subcapacity pricing. There is a single charge for each CPC on the z13s.

Subcapacity license charges

For eligible programs, subcapacity licensing allows software charges that are based on the measured utilization by logical partitions instead of the total number of MSUs of the CPC. Subcapacity licensing removes the dependency between the software charges and CPC (hardware) installed capacity.

Subcapacity licensed products are charged on a monthly basis based on the highest observed 4-hour rolling average utilization of the logical partitions in which the product runs (except for products that are licensed by using the select application license charge (SALC) pricing metric). This type of charge requires measuring the utilization and reporting it to IBM.

The 4-hour rolling average utilization of the logical partition can be limited by a defined capacity value on the image profile of the partition. This value activates the soft capping function of the PR/SM, limiting the 4-hour rolling average partition utilization to the defined capacity value. Soft capping controls the maximum 4-hour rolling average usage (the last 4-hour average value at every 5-minute interval), but does not control the maximum instantaneous partition use.

Also available is an LPAR group capacity limit, which sets soft capping by PR/SM for a group of logical partitions running on z/OS.

Even by using the soft capping option, the use of the partition can reach its maximum share based on the number of logical processors and weights in the image profile. Only the 4-hour rolling average utilization is tracked, allowing utilization peaks above the defined capacity value.

Some pricing metrics apply to stand-alone z Systems servers. Others apply to the aggregation of multiple z Systems server workloads within the same Parallel Sysplex.

For more information about WLC and details about how to combine logical partition utilization, see *z/OS Planning for Workload License Charges*, SA22-7506.

IBM z13s

Metrics that are applicable to a stand-alone z13s include the following charges:

- ▶ Advanced entry workload license charges (AEWLC)
- ▶ z Systems new application license charge (zNALC)
- ▶ Parallel Sysplex license charge (PSLC)

Metrics that are applicable to a z13s in an actively coupled Parallel Sysplex include the following charges:

- ▶ AWLC, when all nodes are z13, z13s, zEC12, zBC12, z196, or z114 servers.
- ▶ Variable workload license charge (VWLC), allowed only under the AWLC Transition Charges for Sysplexes when not all of the nodes are z13, z13s, zEC12, zBC12, z196, or z114 servers.
- ▶ zNALC
- ▶ PSLC

7.16.3 Advanced Workload License Charges

Advanced Workload License Charges (AWLC) was introduced with the IBM zEnterprise 196. They use the measuring and reporting mechanisms, and the existing MSU tiers, from VWLCs, although the prices for each tier were lowered.

AWLC can be implemented in full-capacity or subcapacity mode. The AWLC applies to z/OS and z/TPF and their associated middleware products, such as DB2, IMS, CICS, and WebSphere MQ, as well as IBM Lotus® and IBM Domino®.

With z13s, Technology Transition Offerings are available that extend the software price and performance of the AWLC pricing metric:

- ▶ Technology Update Pricing for the IBM z13s is applicable for clients that run on a stand-alone z13s or in an aggregated Parallel Sysplex consisting exclusively of z13s servers.
- ▶ New Transition Charges for Sysplexes (TC3) are applicable when z13, z13s, zEC12, and zBC12 are the only servers in an actively coupled Parallel Sysplex.
- ▶ Transition Charges for Sysplexes (TC2) are applicable when two or more servers exist in an actively coupled Parallel Sysplex consisting of one or more z13, z13s, zEC12, zBC12, z196, or z114 servers.

For more information, see the AWLC website:

<http://www.ibm.com/systems/z/resources/swprice/mlc/awlc.html>

7.16.4 Advanced Entry Workload License Charge

AEWLC was introduced with the z114. They use the measuring and reporting mechanisms, and the existing MSU tiers, from the entry workload license charges (EWLC) pricing metric and the midrange workload license charges (MWLC) pricing metric.

As compared to EWLC and MWLC, the software price performance has been extended. AEWLC also offers improved price performance as compared to PSLC.

Similarly to WLC, AEWLC can be implemented in full-capacity or subcapacity mode. AEWLC applies to z/OS and z/TPF and their associated middleware products, such as DB2, IMS, CICS, WebSphere MQ, and IBM Lotus Domino, when running on a z114.

AEWLC applies to z/VSE and several middleware products when running on z114, zBC12, and z13s servers. For more information, see the AEWLC web page:

<http://www.ibm.com/systems/z/resources/swprice/mlc/aewlc.html>

7.16.5 System z New Application License Charges

zNALCs offer a reduced price for the z/OS operating system on logical partitions that run a qualified new workload application. An example includes Java language business applications that run under the WebSphere Application Server for z/OS or SAP.

z/OS with zNALC provides a strategic pricing model that is available on the full range of z Systems servers for simplified application planning and deployment. zNALC allows for aggregation across a qualified Parallel Sysplex, which can provide a lower cost for incremental growth across new workloads that span a Parallel Sysplex.

For more information, see the zNALC website:

<http://www.ibm.com/systems/z/resources/swprice/mlc/znalc.html>

7.16.6 Midrange workload license charges

MWLCs apply to z/VSE V4 and later when running on z13, zEC12, z196, System z10, and z9 servers. The exceptions are the z10 BC and z9 BC servers at the capacity setting A01, to which zELC applies, and z114 and zBC12, where MWLC is not available.

Similar to workload license charges, MWLC can be implemented in full-capacity or subcapacity mode. An MWLC applies to z/VSE V4 and later, and several IBM middleware products for z/VSE. All other z/VSE programs continue to be priced as before.

The z/VSE pricing metric is independent of the pricing metric for other systems (for example, z/OS) that might be running on the same server. When z/VSE is running as a guest of z/VM, z/VM V5R4 or later is required.

To report usage, the subcapacity report tool is used. One Sub-Capacity Reporting Tool (SCRT) report per server is required.

For more information, see the MWLC website:

<http://www.ibm.com/systems/z/resources/swprice/mlc/mwlc.html>

7.16.7 Parallel Sysplex License Charges

PSLCs apply to a large range of mainframe servers. The list can be obtained from this website:

<http://www.ibm.com/systems/z/resources/swprice/reference/exhibits/hardware.html>

Although it can be applied to stand-alone CPCs, the metric provides aggregation benefits only when applied to a group of CPCs in an actively coupled Parallel Sysplex cluster according to IBM terms and conditions.

Aggregation allows charging a product that is based on the total MSU value of the systems where the product runs (as opposed to all the systems in the cluster). In an uncoupled environment, software charges are based on the MSU capacity of the system.

For more information, see the PSLC website:

<http://www.ibm.com/systems/z/resources/swprice/mlc/pslc.html>

7.16.8 z Systems International Program License Agreement

For z Systems systems, the following types of products are generally in the IPLA category:

- ▶ Data management tools
- ▶ DB2 for z/OS VUE
- ▶ CICS TS VUE V5 and CICS Tools
- ▶ IMS DB VUE V12 and IMS Tools
- ▶ Application development tools
- ▶ Certain WebSphere for z/OS products
- ▶ Linux middleware products
- ▶ z/VM V5 and V6

Generally, three pricing metrics apply to IPLA products for z13s and z Systems:

- ▶ Value unit (VU)

VU pricing applies to the IPLA products that run on z/OS. Value Unit pricing is typically based on the number of MSUs and allows for a lower cost of incremental growth. Examples of eligible products are IMS Tools, CICS Tools, DB2 Tools, application development tools, and WebSphere products for z/OS.

- ▶ Engine-based value unit (EBVU)

EBVU pricing enables a lower cost of incremental growth with more engine-based licenses that are purchased. Examples of eligible products include z/VM V5 and V6, and certain z/VM middleware, which are priced based on the number of engines.

- ▶ Processor value unit (PVU)

PVUs are determined from the number of engines, under the Passport Advantage terms and conditions. Most Linux middleware is also priced based on the number of engines. In z/VM environments, CPU pooling can be used to limit the number of engines that are used to determine the PVUs for a particular software product.

For more information, see the z Systems IPLA website:

<http://www.ibm.com/systems/z/resources/swprice/zipla/index.html>

7.16.9 zBX licensed software

The software licensing for the zBX select System x and POWER7 blades and DataPower XI50z follows the same rules as licensing for blades that are installed outside of zBX.

PowerVM Enterprise Edition *must* be licensed for POWER7 blades at the time of ordering the blades.

The hypervisor for the select System x blades for zBX is provided as part of the zEnterprise Unified Resource Manager.

IBM z Unified Resource Manager

The IBM z Unified Resource Manager is available through z13, z13s, zEC12, and zBC12 *hardware* features, either ordered with the system or ordered later. No separate software licensing is required.

7.17 References

For current planning information, see the support website for each of the following operating systems:

- ▶ z/OS:
<http://www.ibm.com/systems/support/z/zos/>
- ▶ z/VM:
<http://www.ibm.com/systems/support/z/zvm/>
- ▶ z/VSE:
<http://www.ibm.com/systems/z/os/zvse/support/preventive.html>
- ▶ z/TPF:
<http://www.ibm.com/software/http/tpf/pages/maint.htm>
- ▶ Linux on z Systems:
<http://www.ibm.com/systems/z/os/linux/>
- ▶ KVM for IBM z Systems:
<http://www.ibm.com/systems/z/solutions/virtualization/kvm/>



System upgrades

This chapter provides an overview of IBM z Systems z13s upgrade capabilities and procedures, with an emphasis on capacity on demand (CoD) offerings.

The upgrade offerings to the z13s central processor complex (CPC) have been developed from previous IBM z Systems servers. In response to customer demands and changes in market requirements, a number of features have been added. The changes and additions are designed to provide increased customer control over the capacity upgrade offerings with decreased administrative work and enhanced flexibility. The provisioning environment gives the customer an unprecedented flexibility, and a finer control over cost and value.

Given today's business environment, the growth capabilities of z13s servers provide these benefits, among others:

- ▶ Enabling exploitation of new business opportunities
- ▶ Supporting the growth of dynamic, smart environments
- ▶ Managing the risk of volatile, high-growth, and high-volume applications
- ▶ Supporting 24x365 application availability
- ▶ Enabling capacity growth during lockdown periods
- ▶ Enabling planned downtime changes without availability effects

This chapter provides information about the following topics:

- ▶ Upgrade types
- ▶ Concurrent upgrades
- ▶ Miscellaneous equipment specification upgrades
- ▶ Permanent upgrade through the CIU facility
- ▶ On/Off Capacity on Demand
- ▶ Capacity for Planned Event
- ▶ Capacity BackUp
- ▶ Nondisruptive upgrades
- ▶ Summary of capacity on-demand offerings

8.1 Upgrade types

This section summarizes the types of upgrades available for IBM z13s servers.

8.1.1 Overview of upgrade types

Upgrades can be categorized as described in the following scenario.

Permanent and temporary upgrades

In various situations, separate types of upgrades are needed. After a certain amount of time, depending on your growing workload, you might require more memory, extra I/O cards, or more processor capacity. However, in certain situations, only a short-term upgrade is necessary to handle a peak workload, or to temporarily replace lost capacity on a server that is down during a disaster or data center maintenance. z13s servers offer the following solutions for such situations:

► Permanent upgrades:

- Miscellaneous equipment specification (MES)

The MES upgrade order is always performed by IBM personnel. The result can be either real hardware added to the server, or the installation of Licensed Internal Code Configuration Control (LICCC) to the server. In both cases, installation is performed by IBM personnel.

- Customer Initiated Upgrade (CIU)

Using the CIU facility for a server requires that the online CoD buying feature (FC 9900) is installed on the server. The CIU facility supports LICCC upgrades only.

► Temporary upgrades:

All temporary upgrades are LICCC-based. The billable capacity offering is On/Off CoD. The two replacement capacity offerings that are available are Capacity BackUp (CBU) and Capacity for Planned Events (CPE).

For more information, see 8.1.2, “Terminology related to CoD for z13s systems” on page 313.

MES: The MES provides a system upgrade that can result in more enabled processors and a separate central processor (CP) capacity level, but also in a second CPC drawer, memory, I/O drawers, and I/O features (physical upgrade). An MES can also upgrade the IBM z BladeCenter Extension.

Additional planning tasks are required for nondisruptive logical upgrades. Order an MES through your IBM representative, and the MES is delivered by IBM service personnel.

Concurrent and nondisruptive upgrades

Depending on the effect on system and application availability, upgrades can be classified in one of the following ways:

Concurrent

In general, concurrency addresses the continuity of operations of the hardware part of an upgrade, for instance, whether a server (as a box) is required to be turned off during the upgrade. For more information, see 8.2, “Concurrent upgrades” on page 317.

Non-concurrent

This type of upgrade requires stopping the hardware system. Examples of these upgrades include a model upgrade from an N10 model to the N20 model, and physical memory capacity upgrades.

- Disruptive** An upgrade is disruptive when the resources that are added to an operating system (OS) image require that the OS is recycled to configure the newly added resources.
- Nondisruptive** Nondisruptive upgrades do not require that you restart the software that is running or the OS for the upgrade to take effect. Therefore, even concurrent upgrades can be disruptive to those OSs or programs that do not support the upgrades while being nondisruptive to others. For more information, see 8.8, “Nondisruptive upgrades” on page 349.

8.1.2 Terminology related to CoD for z13s systems

Table 8-1 briefly describes the most frequently used terms that relate to CoD for z13s systems.

Table 8-1 CoD terminology

Term	Description
Activated capacity	Capacity that is purchased and activated. Purchased capacity can be greater than activated capacity.
Billable capacity	Capacity that helps handle workload peaks, either expected or unexpected. The one billable offering available is On/Off CoD.
Capacity	Hardware resources (processor and memory) that are able to process the workload that can be added to the system through various capacity offerings.
CBU	A function that enables the use of spare capacity in a CPC to replace capacity from another CPC within an enterprise, for a limited time. Typically, CBU is used when another CPC of the enterprise has failed or is unavailable because of a disaster event. The CPC using CBU replaces the missing CPC's capacity.
CPE	Used when temporary replacement capacity is needed for a short-term event. CPE activates processor capacity temporarily to facilitate moving machines between data centers, upgrades, and other routine management tasks. CPE is an offering of CoD.
Capacity levels	Can be full capacity or subcapacity. For the z13s CPC, capacity levels for the CP engines are from A to Z: <ul style="list-style-type: none"> ▶ A full-capacity CP engine is indicated by Z. ▶ Subcapacity CP engines are indicated by A to Y.
Capacity setting	Derived from the capacity level and the number of processors. For the z13s CPC, the capacity levels are from A01 to Z06, where the last digit indicates the number of active CPs, and the letter from A to Z indicates the processor capacity.
CPC	A physical collection of hardware that consists of main storage, one or more CPs, timers, and channels.
CIU	A web-based facility where you can request processor and memory upgrades by using the IBM Resource Link and the system's Remote Support Facility (RSF) connection.
CoD	The ability of a computing system to increase or decrease its performance capacity as needed to meet fluctuations in demand.

Term	Description
Capacity Provisioning Manager (CPM)	As a component of z/OS Capacity Provisioning, CPM monitors business-critical workloads that are running on z/OS systems on z13s CPCs.
Customer profile	This information is on IBM Resource Link, and contains customer and machine information. A customer profile can contain information about more than one machine.
Full capacity CP feature	For z13s servers, capacity settings Zxx are full capacity settings.
High water mark	Capacity purchased and owned by the customer.
Installed record	The LICCC record was downloaded, staged to the Support Element (SE), and installed on the CPC. A maximum of eight records can be concurrently installed and active.
Licensed Internal Code (LIC)	LIC is microcode, basic I/O system code, utility programs, device drivers, diagnostics, and any other code that is delivered with an IBM machine to enable the machine's specified functions.
LICCC	Configuration control by the LIC provides for a server upgrade without hardware changes by enabling the activation of extra, previously installed capacity.
MES	An upgrade process initiated through an IBM representative and installed by IBM personnel.
Model capacity identifier (MCI)	Shows the current active capacity on the server, including all replacement and billable capacity. For z13s servers, the model capacity identifier is in the form of Axx to Zxx, where xx indicates the number of active CPs. Note that xx can have a range of 01-06.
Model permanent capacity identifier (MPCI)	Keeps information about capacity settings that were active before any temporary capacity was activated.
Model temporary capacity identifier (MTCI)	Reflects the permanent capacity with billable capacity only, without replacement capacity. If no billable temporary capacity is active, MTCI equals MPCI.
On/Off CoD	Represents a function that enables a spare capacity in a CPC to be made available to increase the total capacity of a CPC. For example, On/Off CoD can be used to acquire more capacity to handle a workload peak.
Feature on demand (FoD)	FoD is a new centralized way to flexibly entitle features and functions on the system. FoD contains, for example, the zBX High Water Marks (HWM). HWMs refer to highest quantity of blade entitlements by blade type that the customer has purchased. On z196/z114, the HWMs are stored in the processor and memory LICCC record. On zEC12, zBC12, z13, and z13s servers, the HWMs are found in the FoD record.
Permanent capacity	The capacity that a customer purchases and activates. This amount might be less capacity than the total capacity purchased.
Permanent upgrade	LIC licensed by IBM to enable the activation of applicable computing resources, such as processors or memory, for a specific CIU-eligible machine on a permanent basis.
CPC drawer	Packaging technology that contains the single chip modules (SCMs) for the processor units (PUs) and shared cache (SC) chips, and memory and connections to I/O and coupling links.

Term	Description
Purchased capacity	Capacity delivered to and owned by the customer. It can be higher than the permanent capacity.
Permanent/Temporary entitlement record	The internal representation of a temporary (TER) or permanent (PER) capacity upgrade processed by the CIU facility. An entitlement record contains the encrypted representation of the upgrade configuration with the associated time limit conditions.
Replacement capacity	A temporary capacity that is used for situations in which processing capacity in other parts of the enterprise is lost during a planned event or an unexpected disaster. The two replacement offerings available are CPE and CBU.
Resource Link	IBM Resource Link is a technical support website that is included in the comprehensive set of tools and resources available from the IBM Systems technical support site: http://www.ibm.com/servers/resourceLink/
Secondary approval	An option, selected by the customer, that a second approver control each CoD order. When a secondary approval is required, the request is sent for approval or cancellation to the Resource Link secondary user identifier (ID).
SCM	The packaging technology that is used to hold the PUs and SC chips.
Staged record	The point when a record representing a capacity upgrade, either temporary or permanent, has been retrieved and loaded on the SE disk.
Subcapacity	For z13s servers, CP features A01 to Y06 represent subcapacity configurations, and CP features Z01 to Z06 represent full capacity configurations.
Temporary capacity	An optional capacity that is added to the current server capacity for a limited amount of time. It can be capacity that is owned or not owned by the customer.
Vital product data (VPD)	Information that uniquely defines the system, hardware, software, and microcode elements of a processing system.

8.1.3 Permanent upgrades

Permanent upgrades can be ordered through an IBM marketing representative, or started by the customer with the CIU on IBM Resource Link.

CIU: The use of the CIU facility for a server requires that the online CoD buying feature code (FC 9900) is installed on the server. The CIU facility itself is enabled through the permanent upgrade authorization feature code (FC 9898).

Permanent upgrades ordered through an IBM representative

Through a permanent upgrade, you can perform these tasks:

- ▶ Add a CPC drawer.
- ▶ Add I/O drawers and features.
- ▶ Add model capacity.
- ▶ Add specialty engines.
- ▶ Add memory.
- ▶ Activate unassigned model capacity or Integrated Facility for Linux (IFL) processors.
- ▶ Deactivate activated model capacity or IFLs.

- ▶ Activate channels.
- ▶ Activate cryptographic engines.
- ▶ Change specialty engine (recharacterization).
- ▶ Add zBX features. Blades can only be added if Blade entitlement is already present.

Important: Most of the MESs can be concurrently applied without disrupting the existing workload (see 8.2, “Concurrent upgrades” on page 317 for details). However, certain MES changes are disruptive (for example, an upgrade of Model N10 to N20, or adding memory).

Permanent upgrades initiated through CIU on IBM Resource Link

Ordering a permanent upgrade by using the CIU application through the IBM Resource Link enables you to add capacity to fit within your existing hardware:

- ▶ Add model capacity.
- ▶ Add specialty engines.
- ▶ Add memory.
- ▶ Activate unassigned model capacity or IFLs.
- ▶ Deactivate activated model capacity or IFLs.

8.1.4 Temporary upgrades

IBM z13s servers offer the following types of temporary upgrades:

▶ On/Off CoD

This offering enables you to temporarily add more capacity or specialty engines due to seasonal activities, period-end requirements, peaks in workload, or application testing.

This temporary upgrade can only be ordered by using the CIU application through the Resource Link.

▶ CBU

This offering enables you to replace model capacity or specialty engines to a backup server during an unforeseen loss of server capacity because of an emergency.

▶ CPE

This offering enables you to replace model capacity or specialty engines due to a relocation of workload during system migrations or a data center move.

CBU or CPE temporary upgrades can be ordered by using the CIU application through Resource Link, or by calling your IBM marketing representative. Temporary upgrades capacity changes can be billable or replacement capacity.

Billable capacity

To handle a peak workload, processors can be activated temporarily on a daily basis. You can activate up to twice the purchased millions of service units (MSU) capacity of any PU type. The one billable capacity offering is On/Off CoD.

Replacement capacity

When processing capacity is lost in another part of an enterprise, replacement capacity can be activated. It enables you to activate any PU type up to the authorized limit.

Two replacement capacity offerings exist:

- ▶ CBU
- ▶ CPE

8.2 Concurrent upgrades

Concurrent upgrades on z13s servers can provide extra capacity with no server outage. In most cases, with prior planning and OS support, a concurrent upgrade is nondisruptive to the OS.

Given today's business environment, the benefits of the concurrent capacity growth capabilities that are provided by z13s servers are plentiful. They include, but are not limited to, the following benefits:

- ▶ Enabling the exploitation of new business opportunities
- ▶ Supporting the growth of smart environments
- ▶ Managing the risk of volatile, high-growth, and high-volume applications
- ▶ Supporting 24x365 application availability
- ▶ Enabling capacity growth during *lockdown* periods
- ▶ Enabling planned-downtime changes without affecting availability

These capabilities are based on the flexibility of the design and structure, which enables concurrent hardware installation and LIC control over the configuration.

Subcapacity models provide for a CP capacity increase in two dimensions that can be used together to deliver configuration granularity. The first dimension is by adding CPs to the configuration, and the second dimension is by changing the capacity setting of the CPs currently installed to a higher MCI. In addition, a capacity increase can be delivered by increasing the CP capacity setting, and at the same time decreasing the number of active CPs.

z13s servers support the concurrent addition of processors to a running logical partition (LPAR). As a result, you can have a flexible infrastructure in which you can add capacity without pre-planning. The OS running on the CPC also needs to support the dynamic addition of processors resources.

Another function concerns the system assist processor (SAP). When extra SAPs are concurrently added to the configuration, the SAP-to-channel affinity is dynamically remapped on all SAPs on the z13s server to rebalance the I/O configuration.

Additionally, zBX Blade features can be installed concurrently if their entitlement exists.

8.2.1 Model upgrades

z13s servers have the following machine types and models, and MCIs:

- ▶ Machine type and model are 2965-N vv . The vv can be 10 or 20. The model number indicates how many PUs (vv) are available for customer characterization. Model 10 has one CPC drawer, and Model N20 can have one or two CPC drawers.
- ▶ MCIs are A01 to Z06. The MCI describes how many CPs are characterized (01 - 06) and the capacity setting (A to Z) of the CPs.

A hardware configuration upgrade might require more physical hardware (processor or I/O drawers, or both). A z13s upgrade can change either, or both, the server model and the MCI.

Note the following model upgrade information:

- ▶ LICCC upgrade:
 - Does not change the server model 2965 N10 because an additional node (3 SCMs and memory) or CPC drawer is not added.
 - Can change the MCI, the capacity setting, or both.
- ▶ Hardware installation upgrade:
 - Can change the server model 2965 N10 to N20, if an additional node or an additional CPC drawer is included. This upgrade is non-concurrent.
 - Can change the model capacity identifier, the capacity setting, or both.

The MCI can be concurrently changed. Concurrent upgrades can be accomplished for both *permanent* and *temporary* upgrades.

Model upgrades: A model upgrade from the N10 to the N20 is disruptive. Upgrades from a N20 single CPC drawer to a N20 with two CPC drawers is also disruptive.

Licensed Internal Code upgrades (MES-ordered)

The LICCC provides for server upgrade without hardware changes by activation of additional (previously installed) unused capacity. Concurrent upgrades through LICCC can be done for the following components and situations:

- ▶ Processors (logical CPs, IFL processors, Internal Coupling Facility (ICF) processors, IBM z Systems Integrated Information Processors (zIIPs), and SAPs.
- ▶ Unused PUs, if they are available on the installed CPC drawers, or if the MCI for the CPs can be increased.
- ▶ Memory, when unused capacity is available on the installed memory cards. The plan-ahead memory option is available for customers to gain better control over future memory upgrades. See 2.4.4, “Memory upgrades” on page 63 for more details.

Concurrent hardware installation upgrades (MES-ordered)

Configuration upgrades can be concurrent when installing extra cards, PCIe I/O drawers, I/O drawers, and features:

- ▶ HCA2-C, HCA3-O, or PCIe fanout cards
- ▶ I/O cards (features), when slots are available in the installed I/O drawers and PCIe I/O drawers
- ▶ I/O drawers and PCIe I/O drawers
- ▶ zBX Blade features

The concurrent I/O upgrade capability can be better used if a future target configuration is considered during the initial configuration.

Concurrent PU conversions (MES-ordered)

z13s servers support concurrent conversion between all PU types, such as any-to-any PUs, including SAPs, to provide flexibility to meet changing business requirements.

LICCC-based PU conversions: The LICCC-based PU conversions require that at least one PU, either CP, ICF, or IFL, remains unchanged. Otherwise, the conversion is disruptive. The PU conversion generates a new LICCC that can be installed concurrently in two steps:

1. The assigned PU is removed from the configuration.
2. The newly available PU is activated as the new PU type.

LPARs might also have to free the PUs to be converted, and the OSs must have support to configure processors offline or online for the PU conversion to be done non-disruptively.

PU conversion: Customer planning and operator action are required to use concurrent PU conversion. Consider the following information about PU conversion:

- ▶ It is disruptive if *all* current PUs are converted to separate types.
- ▶ It might require individual LPAR outages if dedicated PUs are converted.

Unassigned CP capacity is recorded by an MCI. CP feature conversions change (increase or decrease) the MCI.

8.2.2 Customer Initiated Upgrade facility

The CIU facility is an IBM online system through which a customer can order, download, and install permanent and temporary upgrades for z Systems servers. Access to and use of the CIU facility requires a contract between the customer and IBM, through which the terms and conditions for use of the CIU facility are accepted.

The use of the CIU facility for a server requires that the online CoD buying feature code (FC 9900) is installed on the server. The CIU facility itself is controlled through the permanent upgrade authorization feature code (FC 9898).

After you place an order through the CIU facility, you will receive a notice that the order is ready for download. You can then download and apply the upgrade by using functions that are available through the Hardware Management Console (HMC), along with the RSF. After all the prerequisites are met, you perform the entire process, from ordering to activation of the upgrade.

After the download, the actual upgrade process is fully automated, and does not require any onsite presence of IBM service personnel.

CIU prerequisites

The CIU facility supports LICCC upgrades only. It does not support I/O upgrades. All additional capacity that is required for an upgrade must be previously installed. An additional CPC drawer or I/O cards cannot be installed as part of an order placed through the CIU facility. The sum of CPs, unassigned CPs, ICFs, zIIPs, IFLs, and unassigned IFLs cannot exceed the PU count of the installed drawers. The number of zIIPs (each) cannot exceed twice the number of purchased CPs.

CIU registration and agreed contract for CIU

To use the CIU facility, a customer must be registered and the system must be set up. After completing the CIU registration, access the CIU application through the IBM Resource Link website:

<http://www.ibm.com/servers/resourceLink/>

As part of the setup, the customer provides one resource link ID for configuring and placing CIU orders and, if required, a second ID as an approver. The IDs are then set up for access to the CIU support. The CIU facility is beneficial by enabling upgrades to be ordered and delivered much faster than through the regular MES process.

To order and activate the upgrade, log on to the IBM Resource Link website and start the CIU application to upgrade a server for processors or memory. Requesting a customer order approval to conform to your operation policies is possible. You can enable the definition of additional IDs to be authorized to access the CIU. Extra IDs can be authorized to enter or approve CIU orders, or only view existing orders.

Permanent upgrades

Permanent upgrades can be ordered by using the CIU facility. Through the CIU facility, you can generate online permanent upgrade orders to concurrently add processors (CPs, ICFs, zIIPs, IFLs, and SAPs) and memory, or change the MCI, up to the limits of the installed CPC drawers on an existing z13s CPC.

Temporary upgrades

The z13s base model describes permanent and dormant capacity (Figure 8-1) using the capacity marker and the number of PU features installed on the CPC. Up to eight temporary offerings (or records) can be present.

Each offering has its own policies and controls, and each offering can be activated or deactivated independently in any sequence and combination. Although multiple offerings can be active at any time if enough resources are available to fulfill the offering specifications, only one On/Off CoD offering can be active at any time.

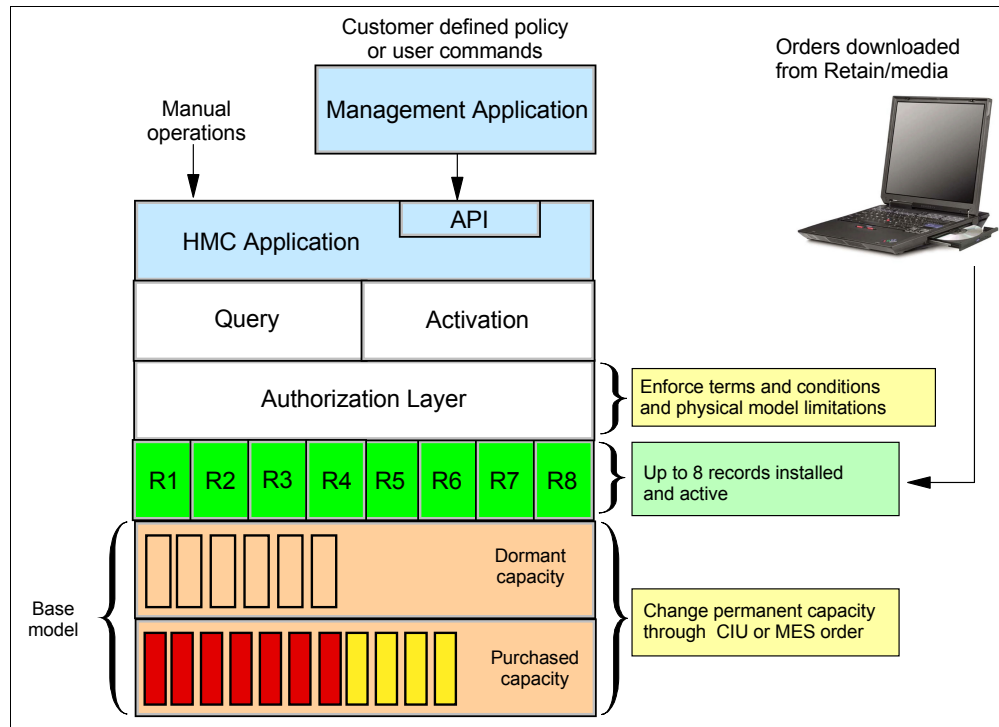


Figure 8-1 The provisioning architecture

Temporary upgrades are represented in the z13s server by a *record*. All temporary upgrade records, downloaded from the RSF or installed from portable media, are resident on the SE hard disk drive (HDD). At the time of activation, the customer can control everything locally. Figure 8-1 shows a representation of the provisioning architecture.

The authorization layer enables administrative control over the temporary offerings. The activation and deactivation can be driven either manually or under the control of an application through a documented application programming interface (API).

By using the API approach, you can customize, at activation time, the resources necessary to respond to the current situation, up to the maximum specified in the order record. If the situation changes, you can add more, or remove resources, without having to go back to the base configuration. This capability eliminates the need for temporary upgrade specifications for all possible scenarios. However, for CPE, the specific ordered configuration is the only possible activation.

In addition, this approach enables you to update and replenish temporary upgrades, even in situations where the upgrades are already active. Likewise, depending on the configuration, permanent upgrades can be performed while temporary upgrades are active. Figure 8-2 shows examples of activation sequences of multiple temporary upgrades.

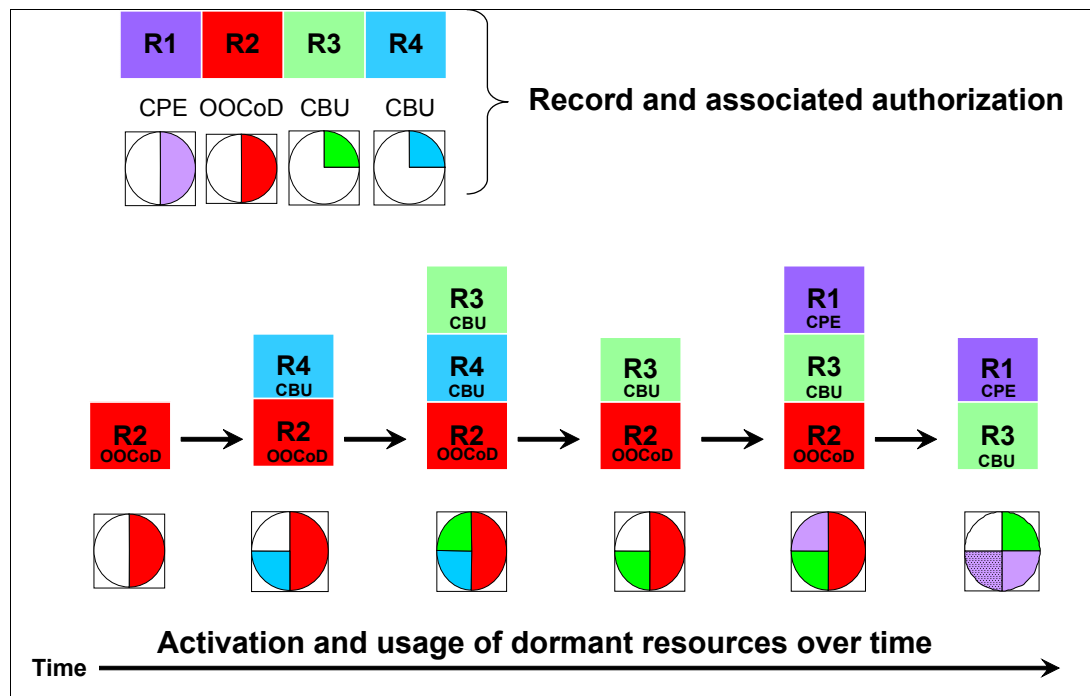


Figure 8-2 Example of temporary upgrade activation sequence

If R2, R3, and R1 are active at the same time, only parts of R1 can be activated because not enough resources are available to fulfill all of R1. When R2 is then deactivated, the remaining parts of R1 can be activated as shown.

Temporary capacity can be billable as On/Off CoD, or replacement as CBU or CPE:

- On/Off CoD is a function that enables *concurrent* and *temporary* capacity growth of the server.

On/Off CoD *can* be used for customer peak workload requirements, for any length of time, and has a daily hardware and maintenance charge.

The software charges can vary according to the license agreement for the individual products. See your IBM Software Group (SWG) representative for exact details.

On/Off CoD can concurrently add processors (CPs, ICFs, zIIPs, IFLs, and SAPs), increase the MCI, or both, up to the limit of the installed CPC drawers of an existing server. It is restricted to twice the currently installed capacity. On/Off CoD requires a contractual agreement between the customer and IBM.

The customer decides whether to pre-pay or post-pay On/Off CoD. Capacity tokens inside the records are used to control activation time and resources.

- CBU is a concurrent and temporary activation of extra CPs, ICFs, zIIPs, IFLs, and SAPs, an increase of the MCI, or both.

CBU *cannot* be used for peak load management of customer workload, or for CPE. A CBU activation can last up to 90 days when a disaster or recovery situation occurs.

CBU features are optional, and require unused capacity to be available on the installed CPC drawers of the backup server, either as unused PUs or as a possibility to increase the MCI, or both. A CBU contract must be in place before the special code that enables this capability can be loaded on the server. The standard CBU contract provides for five 10-day tests and one 90-day disaster activation over a five-year period. Contact your IBM representative for details.

Note: Customers can order more 10-day test periods. Buying options vary from 1 to 10 extra test periods of 10 days each.

- CPE is a concurrent and temporary activation of extra CPs, ICFs, zIIPs, IFLs, and SAPs, an increase of the MCI, or both.

The CPE offering is used to replace temporary lost capacity within a customer's enterprise for planned downtime events, such as during data center changes. CPE cannot be used for peak load management of the customer workload, or for a disaster situation.

The CPE feature requires unused capacity to be available on installed CPC drawers of the backup server, either as unused PUs or as a possibility to increase the MCI, or both. A CPE contract must be in place before the special code that enables this capability can be loaded on the server. The standard CPE contract provides for one 3-day planned activation at a specific date. Contact your IBM representative for details.

8.2.3 Summary of concurrent upgrade functions

Table 8-2 summarizes the possible concurrent upgrade combinations.

Table 8-2 Concurrent upgrade summary

Type	Name	Upgrade	Process
Permanent	MES	CPs, ICFs, zIIPs, IFLs, SAPs, CPC drawer, memory, and I/O	Installed by IBM service personnel
	Online permanent upgrade	CPs, ICFs, zIIPs, IFLs, SAPs, and memory	Performed through the CIU facility

Type	Name	Upgrade	Process
Temporary	On/Off CoD	CPs, ICFs, zIIPs, IFLs, and SAPs	Performed through the On/Off CoD facility
	CBU	CPs, ICFs, zIIPs, IFLs, and SAPs	Performed through the CBU facility
	CPE	CPs, ICFs, zIIPs, IFLs, and SAPs	Performed through the CPE facility

8.3 Miscellaneous equipment specification upgrades

MES upgrades enable concurrent and permanent capacity growth. MES upgrades enable the concurrent addition of processors (CPs, ICFs, zIIPs, IFLs, and SAPs), memory capacity, I/O ports, hardware, and entitlements to the zBX. MES upgrades enable the concurrent adjustment of both the number of processors and the capacity level. The MES upgrade can be done by using LICCC only, by installing an additional CPC drawer, adding I/O cards, or a combination:

- ▶ MES upgrades for processors are done by any of the following methods:
 - LICCC assigning and activating unassigned PUs up to the limit of the installed CPC drawers.
 - LICCC to adjust the number and types of PUs, to change the capacity setting, or both.
 - Installing an additional CPC drawer, and LICCC assigning and activating unassigned PUs in the installed drawer.
- ▶ MES upgrades for memory are done by either of the following methods:
 - Using LICCC to activate extra memory capacity up to the limit of the memory cards on the currently installed CPC drawers. Plan-ahead memory features enable you to have better control over future memory upgrades. For more information about the memory features, see 2.4.4, “Memory upgrades” on page 63.
 - Installing an additional CPC drawer, and using LICCC to activate extra memory capacity on the installed CPC drawer.
- ▶ MES upgrades for I/O are done by the following method:
 - Installing additional I/O cards and supporting infrastructure, if required, in the I/O drawers or PCIe I/O drawers that are already installed, or installing extra PCIe I/O drawers to hold the new cards.
- ▶ MES upgrades for the zBX can only be performed through your IBM service support representative (SSR) and are restricted to add Blades to existing entitlements.

An MES upgrade requires IBM service personnel for the installation. In most cases, the time that is required for installing the LICCC and completing the upgrade is short. To better use the MES upgrade function, carefully plan the initial configuration to enable a concurrent upgrade to a target configuration.

By planning ahead, it is possible to enable nondisruptive capacity and I/O growth with no system power down and no associated power-on resets (PORs) or initial program loads (IPLs). The availability of I/O drawer and PCIe I/O drawers has improved the flexibility to perform unplanned I/O configuration changes concurrently.

The store system information (STSI) instruction gives more useful and detailed information about the base configuration, and about temporary upgrades. STSI enables you to more easily resolve billing situations where independent software vendor (ISV) products are in use.

The model and MCI returned by the STSI instruction are updated to coincide with the upgrade. See “Store system information instruction” on page 351 for more details.

Upgrades: The MES provides the physical upgrade, resulting in more enabled processors, separate capacity settings for the CPs, additional memory, and more I/O ports. Extra planning tasks are required for nondisruptive logical upgrades (see “Suggestions to avoid disruptive upgrades” on page 353).

8.3.1 MES upgrade for processors

An MES upgrade for processors can concurrently add CPs, ICFs, zIIPs, IFLs, and SAPs to a z13s server by assigning available PUs that are on the CPC drawers, through LICCC. Depending on the quantity of the additional processors in the upgrade, an additional CPC drawer might be required before the LICCC is enabled. Addition of a second CPC drawer is disruptive. More capacity can be provided by adding CPs, by changing the MCI on the current CPs, or by doing both.

Maximums: The sum of CPs, inactive CPs, ICFs, zIIPs, IFLs, unassigned IFLs, and SAPs cannot exceed the maximum limit of PUs available for customer use. The number of zIIPs cannot exceed twice the number of purchased CPs.

8.3.2 MES upgrade for memory

An MES upgrade for memory can concurrently add more memory by enabling, through LICCC, extra capacity up to the limit of the currently installed memory cards. Installing an additional CPC drawer, and enabling the memory capacity on the new drawer with LICCC, are disruptive upgrades.

The Preplanned Memory feature (FC 1996 - 16 GB increment or FC 1993 - 8 GB increment) is available to enable better control over future memory upgrades. See 2.4.5, “Preplanned memory” on page 63, for details about plan-ahead memory features.

The N10 model has, as a minimum, ten 16 GB dual inline memory modules (DIMMs), resulting in 160 GB of installed memory. The minimum customer addressable storage is 64 GB. If you require more than that amount, a *non-concurrent* upgrade can install up to 984 GB of memory for customer use by changing the existing DIMMs.

The N20 single CPC drawer model has, as a minimum, ten 16 GB DIMMs (five DIMMs per node), resulting in 160 GB of installed memory. The minimum customer addressable storage is 64 GB. If you require more than that amount, a *non-concurrent* upgrade can install up to 2008 GB of memory for customer use in a single drawer N20.

The N20 two CPC drawer model has, as a minimum, twenty 16 GB DIMMs (five DIMMs per node in both drawers), resulting in 320 GB of installed memory. The minimum customer addressable storage is 64 GB. If you require more than that amount, a nondisruptive upgrade can install up to 256 GB or, by using a disruptive upgrade, up to 4056 GB, by changing the existing DIMMs.

An LPAR can dynamically take advantage of a memory upgrade if reserved storage was defined to that LPAR. The reserved storage is defined to the LPAR as part of the image profile. Reserved memory can be configured online to the LPAR by using the LPAR dynamic storage reconfiguration (DSR) function.

DSR enables a z/OS OS image and z/VM partitions to add reserved storage to their configuration if any unused storage exists. The nondisruptive addition of storage to a z/OS and z/VM partition necessitates that pertinent OS parameters have been prepared. If reserved storage has not been defined to the LPAR, the LPAR must be deactivated, the image profile changed, and the LPAR reactivated to enable the additional storage resources to be available to the OS image.

8.3.3 Preplanned Memory feature

The Preplanned Memory feature (FC 1996 or FC 1993) enables you to install memory for future use:

- ▶ FC 1996 specifies memory to be installed but not used. Order one feature for each 16 GB that will be usable by the customer.
- ▶ FC 1899 is used to activate previously installed preplanned memory, and can activate all the preinstalled memory or subsets of it. For each additional 16 GB of memory to be activated, one FC 1899 must be added, and one FC 1996 must be removed.
- ▶ FC 1993 specifies memory to be installed but not used. Order one feature for each 8 GB that will be usable by the customer.
- ▶ FC 1903 is used to activate previously installed preplanned memory, and can activate all the preinstalled memory or subsets of it. For each additional 8 GB of memory to be activated, one FC 1903 must be added, and one FC 1993 must be removed.

See Figure 8-3 for details of memory configurations and upgrades for a z13s Model N10.

N10 Physical Memory RAIM GB Addressable Memory GB	Client GB	Increment GB
32 RAIM (2 x 16GB)- 128 for HSA (40) + Client (88)	64 72 80 88	8
64 RAIM (2 x 32GB)- 256 for HSA (40) + Client (216)	120 152 184 216	32
128 RAIM (2 x 64GB)- 384 for HSA (40) + Client (344)	248 280 312 344	32
128 RAIM (2 x 64GB)- 512 for HSA (40) + Client (472)	408 472	64
256 RAIM (2 x 128GB)- 640 for HSA (40) + Client (600)	536 600	64
256 RAIM (2 x 128GB) 1024 for HSA (40) + Client (984)	728 856 984	128

Figure 8-3 Memory sizes and upgrades for the N10

Figure 8-4 shows the memory configurations and upgrades for a z13s Model N20 single CPC drawer.

N20 one drawer Physical Memory RAIM GB Addressable Memory GB	Client GB	Increment GB
32 RAIM (2 x 16GB) + 128 for HSA (40) + Client (88)	64 72 80 88	8
64 RAIM (4 x 16GB) + 256 for HSA (40) + Client (216)	120 152 184 216	32
128 RAIM (4 x 32GB) + 384 for HSA (40) + Client (344)	248 280 312 344	32
128 RAIM (4 x 32GB) + 512 for HSA (40) + Client (472)	408 472	64
256 RAIM (4 x 64GB) + 640 for HSA (40) + Client (600)	536 600	64
256 RAIM (4 x 64GB) + 1024 for HSA (40) + Client (984)	728 856 984	128
512 RAIM (4 x 128GB) + 2048 for HSA (40) + Client (2008)	1112 1240 1368 1496 1624 1752 1880 2008	128

Figure 8-4 Memory sizes and upgrades for the single drawer N20

Figure 8-5 shows the memory configurations and upgrades for a z13s Model N20 with two CPC drawers.

N20 two drawers Physical Memory RAIM GB Addressable Memory GB	Client GB	Increment GB
576 RAIM (4 x 128GB + 4 x 16GB)+ 2304 for HSA (40) + Client (2264)	2264	256
640 RAIM (4 x 128 GB + 4 x 32GB)+ 2560 for HSA (40) + Client (2520)	2520	256
768 RAIM (4 x 128GB + 4 x 64GB)+ 3072 for HSA (40) + Client (3032)	2776 3032	256
1024 RAIM (8 x 128GB)+ 4096 for HSA (40) + Client (4056)	3288 3544 3800 4056	256

Figure 8-5 Memory sizes and upgrades for the two drawer N20

The accurate planning and definition of the target configuration are vital to maximize the value of these features.

8.3.4 MES upgrades for the zBX

The MES upgrades for zBX can concurrently add only the following components:

- ▶ Blades only if there are any slots available in the existing blade chassis and entitlements to support the new blades

Feature on Demand

FoD contains the zBX HWMs. HWMs refer to the highest quantities of blade entitlements by blade type that the customer has purchased. On IBM zEnterprise 196 (z196) and IBM zEnterprise 114 (z114), the HWMs are stored in the processor and memory LICCC record. On zBC12, zEC12, z13, and z13s servers, the HWMs are found in the FoD LICCC record.

The current zBX installed and staged feature values can be obtained by using the Perform a Model Conversion function on the SE, or from the HMC by using a Single Object Operation (SOO) to the servers' SE.

Figure 8-6 shows the window for the FoD Blades feature values shown under the **Perform a Model Conversion** → **Feature on Demand** → **Manage** function.

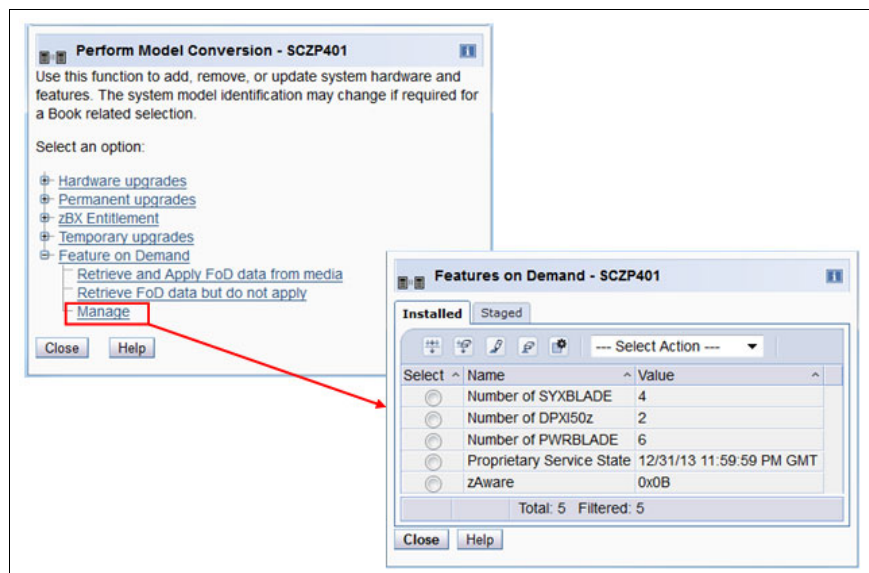


Figure 8-6 Feature on Demand window for zBX Blades features HWMs

Only one FoD LICCC record is installed or staged at any time in the system, and its contents can be viewed under the Manage window, as shown in Figure 8-6. A staged record can be removed without installing it. An FoD record can only be installed completely. There are no selective-feature or partial-record installations, and the features that are installed will be merged with the central electronic complex (CEC) LICCC after activation.

An FoD record can only be installed once. If it is removed, a new FoD record is needed to install it again. A remove action cannot be undone.

If upgrading from an existing z114 with zBX model 002 or a zBC12 with zBX model 003 attached, the zBX must be upgraded to a zBX model 004. The zBX model 004 becomes a stand-alone ensemble member. Two 1U SEs are installed inside the frame B01 to monitor

and control the zBX model 004. With the zBX model 004, the concept of owning the CPC does not exist. Because the system upgrade is always disruptive, the zBX upgrade will also be a disruptive task.

If installing a new build z13s server and planning to take over an existing zBX attached to another existing CPC, and the zBX is not a model 004, the zBX conversion to a zBX model 004 can be done during the installation phase of the z13s server.

The following are some of the features and functions added by the zBX Model 004. Model 004 includes all of the enhancements implemented in zBX Model 003.

- ▶ Stand-Alone node based zBX defined as an *ensemble member* to the ensemble HMCs.
- ▶ Direct management connection to server eliminated with node-based zBX Model 004.
- ▶ Two SEs installed in the zBX Model 004.
- ▶ Firmware updates and service activities are no longer linked to the z Systems CPC.
- ▶ No need for a direct intranode management network (INMN) link between CPC and zBX. Configuration and connectivity complexity is reduced and resiliency improved.
- ▶ Synchronized time from HMC by using Network Time Protocol (NTP).
- ▶ Workload Optimization mapped to SLA requirements available with zManager and zBX Model 004.
- ▶ Available only as an upgrade from zBX Model 002 and from a zBX Model 003.
- ▶ No new blade types will be announced or delivered. As of December 31, 2014, all supported Blades have been withdrawn from marketing (WDfM).
- ▶ Except for DataPower XI50z Blades, supported x and p blades might still be available from distributors or on the used market.
- ▶ Empty slot entitlements for x and p blades, if already existing, can be populated.
- ▶ New entitlements for blades were WDfM as of December 2015.
- ▶ Consider upgrading zBX to a Model 004 until withdrawn (WDfM date not announced).

8.4 Permanent upgrade through the CIU facility

By using the CIU facility (through the IBM Resource Link on the web), you can start a permanent upgrade for CPs, ICFs, zIIPs, IFLs, SAPs, or memory. When performed an upgrade through the CIU facility, you add the resources yourself. IBM personnel do not have to be present at your location. You can also unassign previously purchased CPs and IFL processors through the CIU facility.

The capability to add permanent upgrades to a z13s server through the CIU facility requires that the permanent upgrade enablement feature (FC 9898) be installed on the z13s server. A permanent upgrade might change the MCI if additional CPs are requested or the change capacity identifier is part of the permanent upgrade. However, it cannot change the z13s model from an N10 to an N20. If necessary, more LPARs can be created concurrently to use the newly added processors.

Planning: A permanent upgrade of processors can provide a physical concurrent upgrade, resulting in more enabled processors available to a z13s configuration. Therefore, additional planning and tasks are required for *nondisruptive* logical upgrades. See “Suggestions to avoid disruptive upgrades” on page 353 for more information.

Maintenance charges are automatically adjusted as a result of a permanent upgrade.

Software charges, based on the total capacity of the server on which the software is installed, are adjusted to the new capacity in place after the permanent upgrade is installed. Software products that use the Workload License Charges (WLC) might not be affected by the server upgrade because the charges are based on LPAR usage rather than on the server total capacity. See 7.16.3, “Advanced Workload License Charges” on page 307 for more information about the WLC.

Figure 8-7 illustrates the CIU facility process on IBM Resource Link.

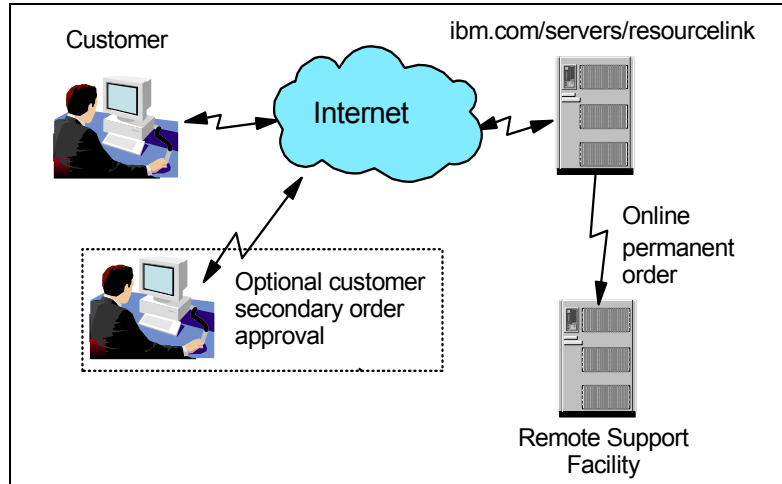


Figure 8-7 Permanent upgrade order example

The following sample sequence on IBM Resource Link initiates an order:

1. Sign on to Resource Link.
2. Select **Customer Initiated Upgrade** from the main Resource Link page. The customer and server details that are associated with the user ID are listed.
3. Select the server that will receive the upgrade. The current configuration (PU allocation and memory) is shown for the selected server.
4. Select **Order Permanent Upgrade**. Resource Link limits the options to those that are valid or possible for this configuration.
5. After the target configuration is verified by the system, accept or cancel the order.
An order is created and verified against the pre-established agreement.
6. Accept or reject the price that is quoted. A secondary order approval is optional.
Upon confirmation, the order is processed. The LICCC for the upgrade will be available within hours.

Figure 8-8 illustrates the process for a permanent upgrade. When the LICCC is passed to the remote support facility, you are notified through email that the upgrade is ready to be downloaded.

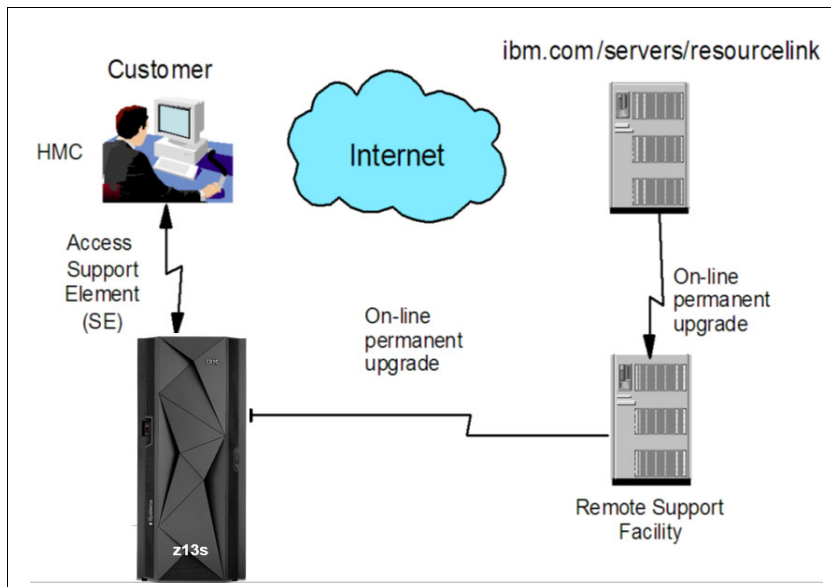


Figure 8-8 CIU-eligible order activation example

The major components in the process are *ordering* and *retrieval* (along with *activation*).

8.4.1 Ordering

Resource Link provides the interface that enables you to order a concurrent upgrade for a server. You can create, cancel, view the order, and view the history of orders that were placed through this interface. Configuration rules enforce that only valid configurations are generated within the limits of the individual server. Warning messages are issued if you select invalid upgrade options. The process enables only one permanent CIU-eligible order for each server to be placed at a time. For a tutorial, see the following website:

<https://www.ibm.com/servers/resourceLink/hom03010.nsf/pages/CIUInformation>

Figure 8-9 shows the initial view of the machine profile on Resource Link.

IBM Systems > System z > Resource Link > Customer Initiated Upgrade >

Machine profile

2818 - CEC04 - 5555556

Current configuration	
Model Capacity:	W02 (2 CPs)
ICF:	0
zAAP:	0
zIIP:	0
IFL:	1
SAP:	2
Memory:	24
Unassigned IFLs:	0
Management enablement level: 1. Manage	
Current configuration as of 27 May 2011 12:16:48	

Machine summary	
Type, model, serial: 2818 - M05 - CEC04	

Customer summary	
Company name: IBM	
Customer number: 5555556	
GEO, country: Americas - zDutchy of Merwyn	

Ordering options	
→ Order permanent upgrade	
→ Order On/Off CoD record	
→ Order On/Off CoD test record	
→ Order On/Off CoD record with prepaid upgrades	
→ Order On/Off CoD record with spending limits	
→ Order administrative On/Off CoD test record	
→ Order Capacity Backup (CBU) record	
→ Order Capacity for Planned Events (CPE) record	
Display upgrade matrix	

About ordering	
Authorization to create orders	
User ID: ciutestuser@us.ibm.com	Ordering options
Name: ciu testuser	CIU Permanent: Enabled
Authorization to approve orders	On/Off CoD: Enabled
Not required	CBU: Enabled
	CPE: Enabled
Notes:	
<ul style="list-style-type: none"> A pre-negotiated price agreement exists for this machine. On/Off CoD Test: 0 staged out of 1 remaining 	

To update profile	
Upload VPD	
Upload upgrade billing XML data	

For more information	
→ View machine's On/Off CoD order billing history	
Download upgrade history CSV (2KB)	
Users authorized to order upgrades	
Users authorized to view orders	
Order status definitions	
→ Customer Initiated Upgrade information	

Permanent upgrades	
<p>Open orders Complete orders All orders</p> <p>There are no open orders for this machine.</p>	

Figure 8-9 Machine profile

The number of CPs, ICFs, zIIPs, IFLs, and SAPs, the memory size, CBU features, unassigned CPs, and unassigned IFLs on the current configuration are displayed on the left side of the web page.

Resource Link retrieves and stores relevant data that is associated with the processor configuration, such as the number of CPs and installed memory cards. It enables you to select only those upgrade options that are deemed valid by the order process. It supports upgrades only within the bounds of the currently installed hardware.

8.4.2 Retrieval and activation

After an order is placed and processed, the appropriate upgrade record is passed to the IBM support system for download.

When the order is available for download, you receive an email that contains an activation number. You can then retrieve the order by using the Perform Model Conversion task from the SE, or through SOO to the SE from an HMC. To retrieve the order, follow these steps:

1. In the Perform Model Conversion window, select **Permanent upgrades** to start the process, as shown in Figure 8-10.

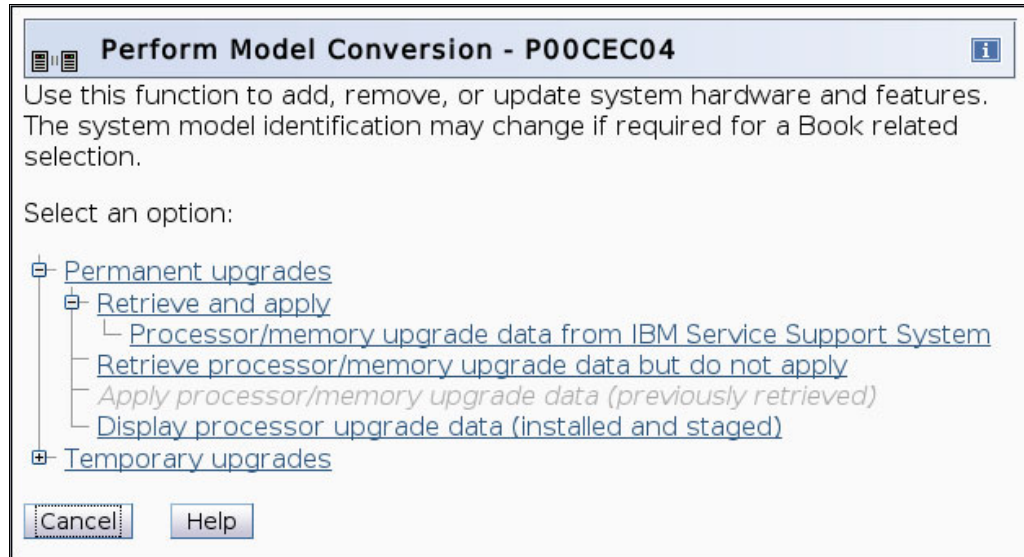


Figure 8-10 IBM z13s Perform Model Conversion window

2. The window provides several possible options. If you select the **Retrieve and apply** option, you are prompted to enter the order activation number to start the permanent upgrade, as shown in Figure 8-11.

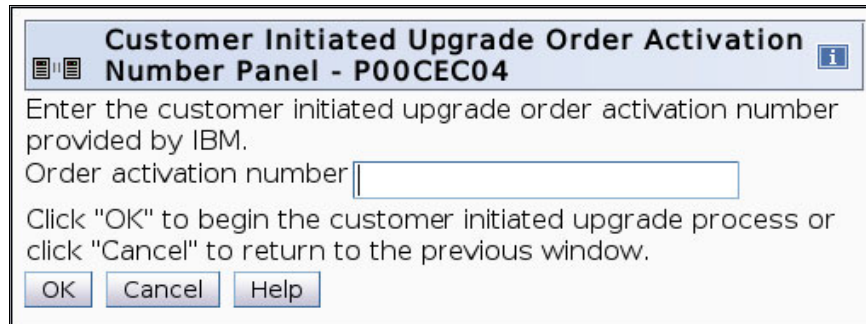


Figure 8-11 Customer Initiated Upgrade Order Activation Number window

8.5 On/Off Capacity on Demand

On/Off CoD enables you to temporarily enable PUs and unassigned IFLs that are available within the current model, or to change capacity settings for CPs to help meet your peak workload requirements.

8.5.1 Overview

The capacity for CPs is expressed in MSUs. The capacity for speciality engines is expressed in the number of speciality engines. *Capacity tokens* are used to limit the resource consumption for all types of processor capacity.

Capacity tokens are introduced to provide better control over resource consumption when On/Off CoD offerings are activated. Tokens are represented in the following manner:

- ▶ For CP capacity, each token represents the amount of CP capacity that will result in one MSU of software cost for one day (an *MSU-day token*).
- ▶ For speciality engines, each token is equivalent to one speciality engine capacity for one day (an *engine-day token*).

Tokens are by capacity type, MSUs for CP capacity, and number of engines for speciality engines. Each speciality engine type has its own tokens, and each On/Off CoD record has separate token pools for each capacity type. During the ordering sessions on Resource Link, you decide how many tokens of each type must be created in an offering record. Each engine type must have tokens for that engine type to be activated. Capacity that has no tokens cannot be activated.

When the resources from an On/Off CoD offering record that contains capacity tokens are activated, a *billing window* is started. A billing window is always 24 hours in length. Billing takes place at the end of each billing window. The resources that are billed are the highest resource usage inside each billing window for each capacity type.

An *activation period* is one or more complete billing windows. It represents the time from the first activation of resources in a record until the end of the billing window in which the last resource in a record is deactivated.

At the end of each billing window, the tokens are decremented by the highest usage of each resource during the billing window. If any resource in a record does not have enough tokens to cover usage for the next billing window, the entire record is deactivated.

On/Off CoD requires that the Online CoD Buying feature (FC 9900) be installed on the server that is to be upgraded.

The On/Off CoD to Permanent Upgrade Option is a new offering, which is an offshoot of On/Off CoD and takes advantage of aspects of the architecture. The customer is given a window of opportunity to assess the capacity additions to the customer's permanent configurations using On/Off CoD. If a purchase is made, the hardware On/Off CoD charges during this window, three days or less, are waived. If no purchase is made, the customer is charged for the temporary use.

The resources that are eligible for temporary use are CPs, ICFs, zIIPs, IFLs, and SAPs. The temporary addition of memory is not supported. Unassigned PUs that are on the installed CPC drawers can be temporarily and concurrently activated as CPs, ICFs, zIIPs, IFLs, and SAPs through LICCC, up to twice the currently installed CP capacity and up to twice the number of ICFs, zIIPs, or IFLs.

An On/Off CoD upgrade cannot change the model from an N10 to an N20. The addition of a new CPC drawer is not supported. However, the activation of an On/Off CoD upgrade can increase the MCI.

8.5.2 Ordering

Concurrently installing temporary capacity by ordering On/Off CoD is possible in the following quantities:

- ▶ CP features equal to the MSU capacity of installed CPs
- ▶ IFL features up to the number of installed IFLs
- ▶ ICF features up to the number of installed ICFs
- ▶ zIIP features up to the number of installed zIIPs
- ▶ SAPs up to two for both models

On/Off CoD can provide CP temporary capacity by increasing the number of CPs, by changing the capacity setting of the CPs, or both. The capacity setting for all CPs must be the same. If the On/Off CoD is adding CP resources that have a capacity setting that differs from the installed CPs, the base capacity settings are changed to match.

On/Off CoD has the following limits associated with its use:

- ▶ The number of CPs cannot be reduced.
- ▶ The target configuration capacity is limited in these ways:
 - Twice the currently installed capacity, expressed in MSUs for CPs.
 - Twice the number of installed IFLs, ICFs, and zIIPs. The number of additional SAPs that can be activated is two.

See Appendix D, “Shared Memory Communications” on page 475 for the valid On/Off CoD configurations for CPs.

On/Off CoD can be ordered as prepaid or postpaid:

- ▶ A prepaid On/Off CoD offering record contains the resource descriptions, MSUs, number of speciality engines, and tokens that describe the total capacity that can be consumed. For CP capacity, the token contains MSU-days. For speciality engines, the token contains speciality engine-days.
- ▶ When resources on a prepaid offering are activated, they must have enough capacity tokens to support the activation for an entire billing window, which is 24 hours. The resources remain active until you deactivate them or until one resource has used all of its capacity tokens. When that happens, all activated resources from the record are deactivated.
- ▶ A postpaid On/Off CoD offering record contains resource descriptions, MSUs, and speciality engines, and can contain capacity tokens for MSU-days and speciality engine-days.
- ▶ When resources in a postpaid offering record without capacity tokens are activated, those resources remain active until they are deactivated, or until the offering record expires, which is usually 180 days after its installation.
- ▶ When resources in a postpaid offering record with capacity tokens are activated, those resources must have enough capacity tokens to support the activation for an entire billing window (24 hours). The resources remain active until they are deactivated or until one of the resource tokens is consumed, or until the record expires, usually 180 days after its installation. If one capacity token type is consumed, resources from the entire record are deactivated.

As an example, for a z13s server with capacity identifier D02, you can use the following methods to deliver a capacity upgrade through On/Off CoD:

- ▶ One option is to add CPs of the same capacity setting. With this option, the MCI can be changed to a D03, which adds one extra CP (making a three-way configuration) or to a D04, which adds two extra CPs (making a four-way configuration).
- ▶ Another option is to change to a separate capacity level of the current CPs, and change the MCI to an E02 or to an F02. The capacity level of the CPs is increased, but no additional CPs are added. The D02 can also be temporarily upgraded to an E03 as indicated in the appendix pointer, increasing the capacity level and adding another processor.

Use the Large Systems Performance Reference (LSPR) information to evaluate the capacity requirements according to your workload type. LSPR data for current IBM processors is available at the following website:

<https://www.ibm.com/servers/resourceLink/lib03060.nsf/pages/lsprindex>

The On/Off CoD hardware capacity is charged on a 24-hour basis. A grace period at the end of the On/Off CoD day provides up to an hour after the 24-hour billing period to either change the On/Off CoD configuration for the next 24-hour billing period, or deactivate the current On/Off CoD configuration. The times when the capacity is activated and deactivated are maintained in the z13s server and sent back to the support systems.

If On/Off capacity is already active, more On/Off capacity can be added without having to return the z13s server to its original capacity. If the capacity is increased multiple times within a 24-hour period, the charges apply to the highest amount of capacity active in the period.

If additional capacity is added from an already active record that contains capacity tokens, a check is made to ensure that the resource in question has enough capacity to be active for an entire billing window (24 hours). If that criteria is not met, no additional resources are activated from the record.

If necessary, extra LPARs can be activated concurrently to use the newly added processor resources.

Planning: On/Off CoD provides a concurrent *hardware* upgrade, resulting in more enabled processors available to a z13s configuration. Additional planning tasks are required for *nondisruptive* upgrades. For more information, see “Suggestions to avoid disruptive upgrades” on page 353.

To participate in this offering, you must have accepted contractual terms for purchasing capacity through the Resource Link, established a profile, and installed an On/Off CoD enablement feature on the z13s server. You can then concurrently install temporary capacity up to the limits in On/Off CoD, and use it for up to 180 days.

Monitoring occurs through the z13s call-home facility, and an invoice is generated if the capacity was enabled during the calendar month. The customer continues to be billed for use of temporary capacity until the z13s server is returned to the original configuration. If the On/Off CoD support is no longer needed, the enablement code should be removed.

On/Off CoD orders can be pre-staged in Resource Link to enable multiple optional configurations. The pricing of the orders is done at the time of the order, and the pricing can vary from quarter to quarter.

Staged orders can have separate pricing. When the order is downloaded and activated, the daily costs are based on the pricing at the time of the order. The staged orders do not have to be installed in order sequence. If a staged order is installed out of sequence, and later an order that had a higher price is downloaded, the daily cost is based on the lower price.

Another possibility is to store unlimited On/Off CoD LICCC records on the SE with the same or separate capacities at any time, giving greater flexibility to quickly enable needed temporary capacity. Each record can be easily identified with descriptive names, and you can select from a list of records that can be activated.

Resource Link provides the interface that enables you to order a dynamic upgrade for a specific z13s server. You are able to create, cancel, and view the order. Configuration rules are enforced, and only valid configurations are generated based on the configuration of the individual z13s server. After completing the prerequisites, orders for the On/Off CoD can be placed. The order process is to use the CIU facility on Resource Link.

You can order temporary capacity for CPs, ICFs, zIIPs, IFLs, or SAPs. Memory and channels are not supported on On/Off CoD. The amount of capacity is based on the amount of owned capacity for the various types of resources. An LICCC record is established and staged to Resource Link for this order. After the record is activated, it has no expiration date. However, an individual record can only be activated once. Subsequent sessions require a new order to be generated, producing a new LICCC record for that specific order.

Alternatively, the customer can use an automatic renewal feature to eliminate the need for a manual replenishment of the On/Off CoD order. This feature is implemented in Resource Link, and the customer must enable the feature in the machine profile. The default for the feature is disabled. See Figure 8-12 for more details.

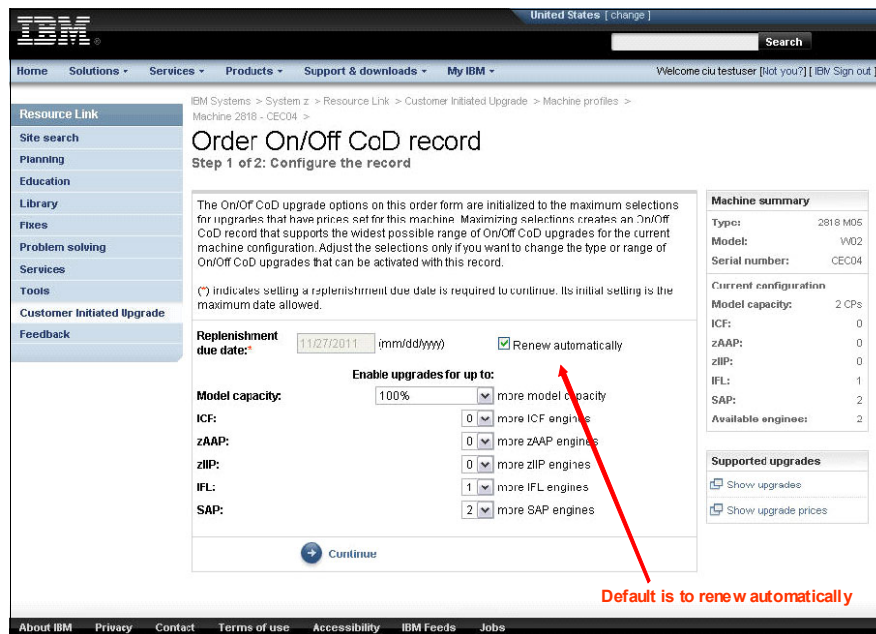


Figure 8-12 Order On/Off CoD record window

8.5.3 On/Off CoD testing

Each On/Off CoD-enabled z13s server is entitled to *one* no-charge 24-hour test. The test has no IBM charges that are associated with temporary hardware capacity, IBM software, or IBM maintenance. The test can be used to validate the processes to download, stage, install, activate, and deactivate On/Off CoD capacity.

The test can last up to of 24 hours, commencing on the activation of any capacity resource that is contained in the On/Off CoD record. Activation levels of capacity can change during the 24-hour test period. The On/Off CoD test automatically stops at the end of the 24-hour period.

In addition, you can perform administrative testing. During this test, no additional capacity is added to the z13s server, but you can test all of the procedures and automation for the management of the On/Off CoD facility.

Figure 8-13 is an example of an On/Off CoD order on the Resource Link web page.

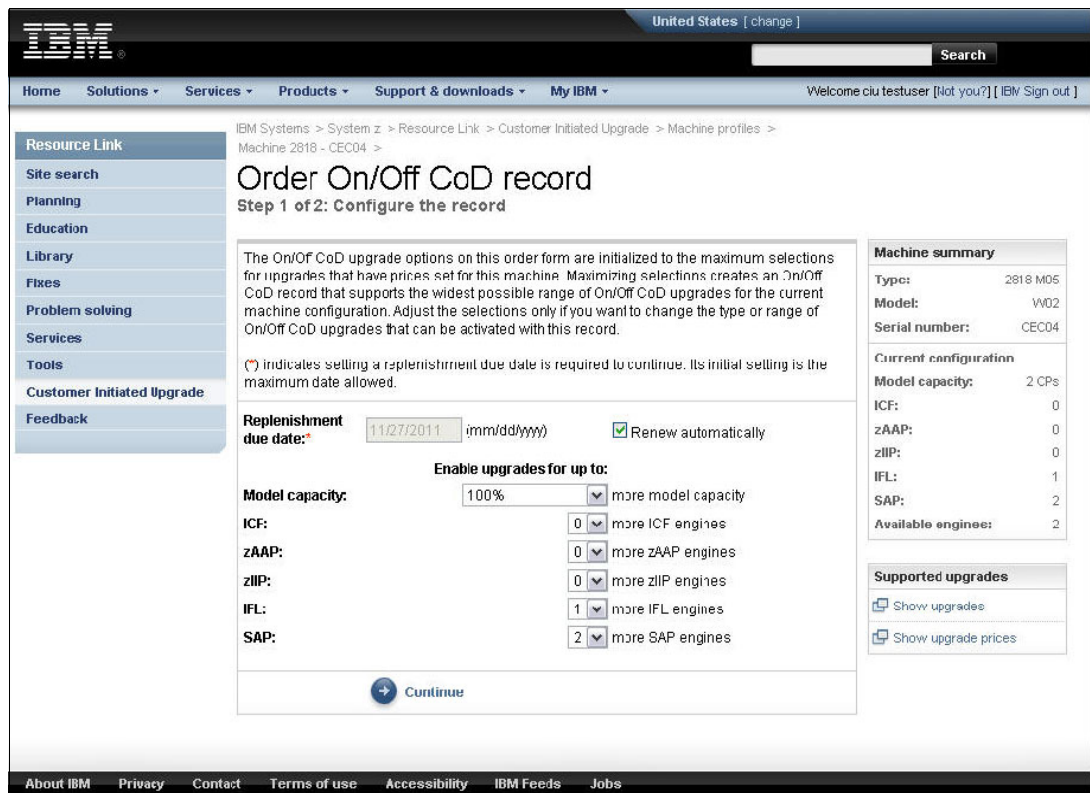


Figure 8-13 On/Off CoD order example

The example order in Figure 8-13 is an On/Off CoD order for 100% more CP capacity, and for one IFL, and two SAPs. The maximum number of CPs, ICFs, zIIPs, and IFLs is limited by the current number of available unused PUs in the installed CPC drawers. The maximum number of SAPs is determined by the model number and the number of available PUs on the already installed CPC drawers.

8.5.4 Activation and deactivation

When a previously ordered On/Off CoD is retrieved from Resource Link, it is downloaded and stored on the SE hard disk. You can activate the order when the capacity is needed, either manually or through automation.

If the On/Off CoD offering record does not contain resource tokens, you must take action to deactivate the temporary capacity. Deactivation is accomplished from the SE, and is nondisruptive. Depending on how the additional capacity was added to the LPARs, you might be required to perform tasks at the LPAR level to remove the temporary capacity. For example, you might have to configure offline CPs that had been added to the partition, or deactivate additional LPARs created to use the temporary capacity, or both.

On/Off CoD orders can be staged in Resource Link so that multiple orders are available. An order can only be downloaded and activated one time. If a separate On/Off CoD order is required, or a permanent upgrade is needed, it can be downloaded and activated without having to restore the system to its original purchased capacity.

In support of automation, an API is provided that enables the activation of the On/Off CoD records. The activation is performed from the HMC, and requires specifying the order number. With this API, the automation code can be used to send an activation command along with the order number to the HMC to enable the order.

8.5.5 Termination

A customer is contractually obliged to terminate the On/Off CoD right-to-use feature when a transfer in asset ownership occurs. A customer can also choose to terminate the On/Off CoD right-to-use feature without transferring ownership.

Application of FC 9898 terminates the right to use the On/Off CoD. This feature cannot be ordered if a temporary session is already active. Similarly, the CIU enablement feature cannot be removed if a temporary session is active. Anytime that the CIU enablement feature is removed, the On/Off CoD right-to-use is simultaneously removed. Reactivating the right-to-use feature subjects the customer to the terms and fees that apply then.

Upgrade capability during On/Off CoD

Upgrades involving physical hardware are supported while an On/Off CoD upgrade is active on a particular z13s server. LICCC-only upgrades can be ordered and retrieved from Resource Link and applied while an On/Off CoD upgrade is active. LICCC-only memory upgrades can also be retrieved and applied while an On/Off CoD upgrade is active.

Repair capability during On/Off CoD

If the z13s server requires service while an On/Off CoD upgrade is active, the repair can take place without affecting the temporary capacity.

Monitoring

When you activate an On/Off CoD upgrade, an indicator is set in the vital product data. This indicator is part of the call-home data transmission, which is sent on a scheduled basis. A time stamp is placed into call-home data when the facility is deactivated. At the end of each calendar month, the data is used to generate an invoice for the On/Off CoD that was used during that month.

Maintenance

The maintenance price is adjusted as a result of an On/Off CoD activation.

Software

Software Parallel Sysplex license charge (PSLC) customers are billed at the MSU level that is represented by the combined permanent and temporary capacity. All PSLC products are billed at the peak MSUs enabled during the month, regardless of usage. Customers with WLC licenses are billed by product at the highest four-hour rolling average for the month. In this instance, temporary capacity does not necessarily increase the software bill until that capacity is allocated to LPARs and actually consumed.

Results from the STSI instruction reflect the current permanent and temporary CPs. See “Store system information instruction” on page 351 for more details.

8.5.6 IBM z/OS capacity provisioning

The z13s provisioning capability, combined with CPM functions in z/OS, provides a flexible, automated process to control the activation of On/Off Capacity on Demand. The z/OS provisioning environment is shown in Figure 8-14.

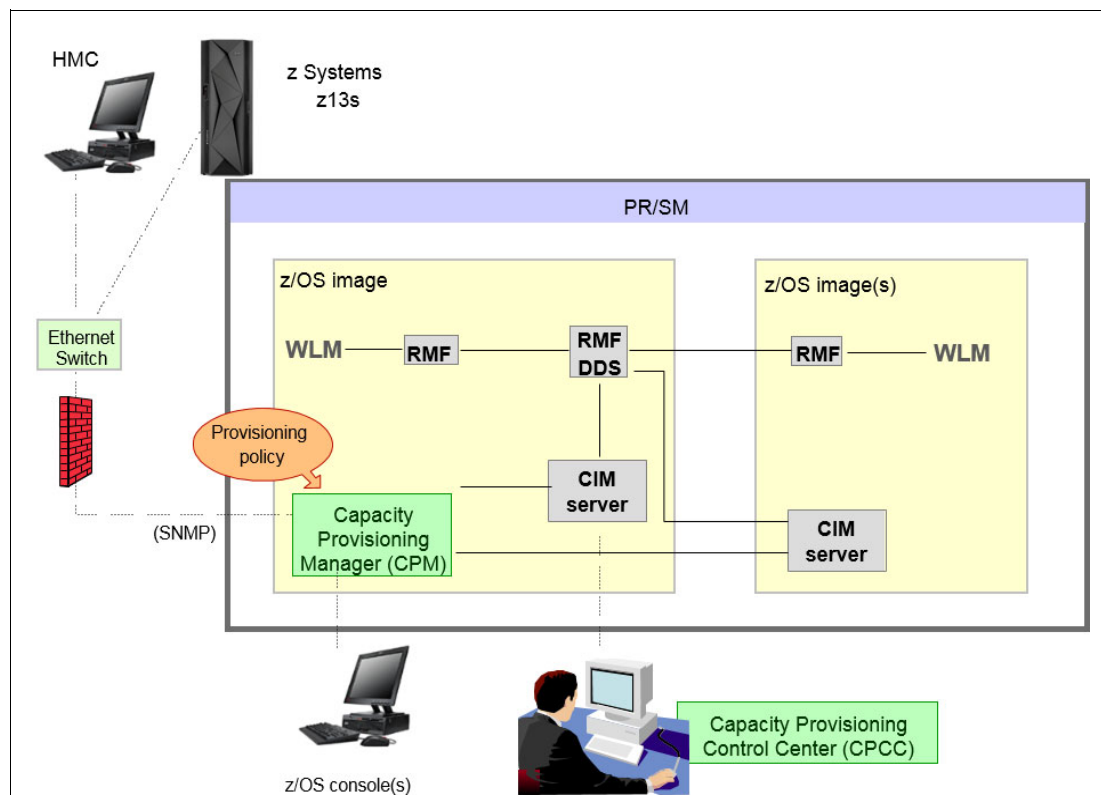


Figure 8-14 The capacity provisioning infrastructure

The z/OS WLM manages the workload by the goals and business importance on each z/OS system. WLM metrics are available through existing interfaces, and are reported through IBM Resource Measurement Facility™ (RMF) Monitor III, with one RMF gatherer for each z/OS system.

Sysplex-wide data aggregation and propagation occur in the RMF Distributed Data Server (DDS). The RMF Common Information Model (CIM) providers and associated CIM models publish the RMF Monitor III data.

The CPM, a function inside z/OS, retrieves critical metrics from one or more z/OS systems through the CIM structures and protocol. CPM communicates to (local or remote) SEs and HMCs through Simple Network Management Protocol (SNMP).

CPM has visibility of the resources in the individual offering records, and the capacity tokens. When CPM decides to activate resources, a check is performed to determine whether enough capacity tokens remain for the specified resource to be activated for at least 24 hours. If insufficient tokens remain, no resource from the On/Off CoD record is activated.

If a capacity token is completely consumed during an activation driven by the CPM, the corresponding On/Off CoD record is deactivated prematurely by the system. This deactivation occurs even if the CPM has activated this record, or parts of it. Warning messages are issued when capacity tokens are getting close to being fully consumed. The messages start being generated five days before a capacity token is fully consumed. The five days are based on the assumption that the consumption will be constant for the five days.

Put operational procedures in place to handle these situations. You can either deactivate the record manually, let it happen automatically, or replenish the specified capacity token by using the Resource Link application.

The Capacity Provisioning Control Center (CPCC), which runs on a workstation, provides an interface to administer capacity provisioning policies. The CPCC is not required for regular CPM operation. The CPCC will over time be moved into the z/OS Management Facility (z/OSMF). Parts of the CPCC have been included in z/OSMF V1R13.

The control over the provisioning infrastructure is run by the CPM through the Capacity Provisioning Domain (CPD), which is controlled by the Capacity Provisioning Policy (CPP). An example of a CPD is shown in Figure 8-15.

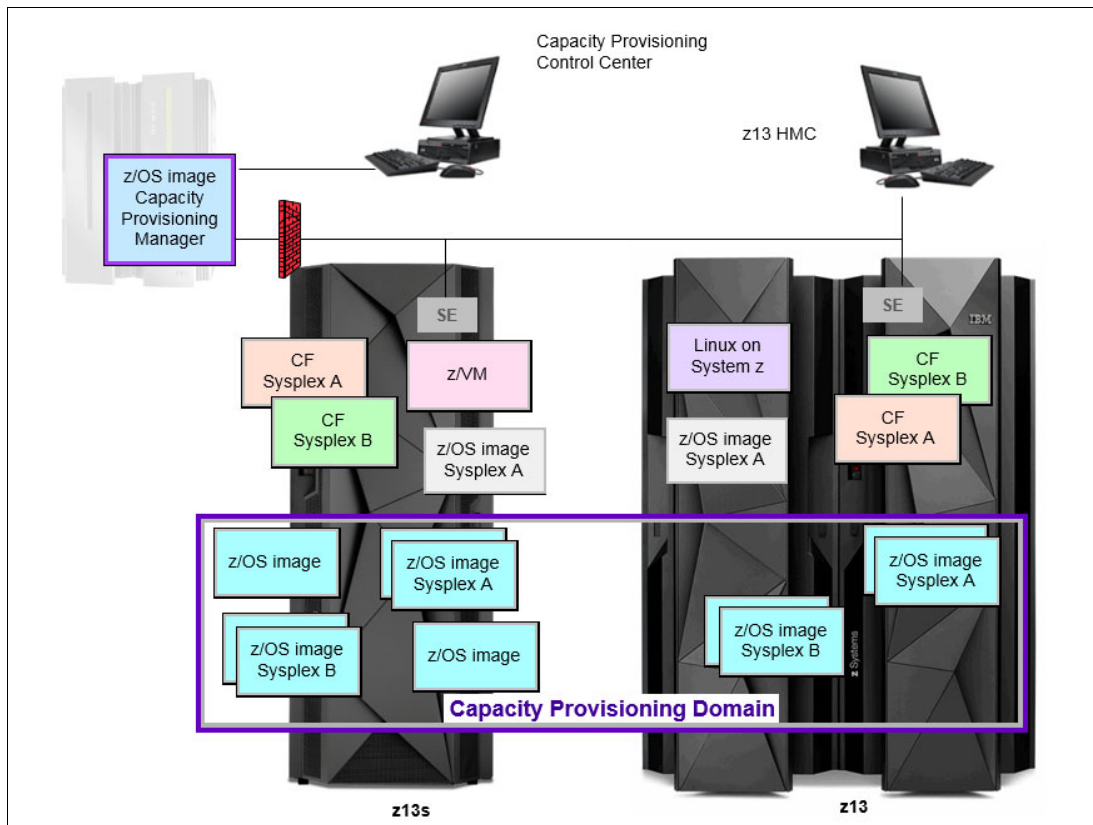


Figure 8-15 A Capacity Provisioning Domain

The CPD represents the central processor complexes (CPCs) that are controlled by the CPM. The HMCs of the CPCs within a CPD must be connected to the same processor LAN. Parallel Sysplex members can be part of a CPD. There is no requirement that all members of a Parallel Sysplex must be part of the CPD, but participating members must all be part of the same CPD.

The CPCC is the user interface component. Administrators work through this interface to define domain configurations and provisioning policies, but it is not needed during production. The CPCC is installed on a Microsoft Windows workstation.

CPM operates in four modes, enabling various levels of automation:

- ▶ Manual mode: Use this command-driven mode when no CPM policy is active.
- ▶ Analysis mode: In analysis mode:
 - CPM processes capacity-provisioning policies and informs the operator when a provisioning or deprovisioning action is required according to the policy criteria.
 - The operator determines whether to ignore the information, or to manually upgrade or revert the system by using the HMC, SE, or available CPM commands.

- ▶ Confirmation mode: In this mode, CPM processes capacity-provisioning policies and interrogates the installed temporary offering records. Every action that is proposed by the CPM needs to be confirmed by the operator.
- ▶ Autonomic mode: This mode is similar to the confirmation mode, but no operator confirmation is required.

A number of reports are available in all modes, containing information about the workload, provisioning status, and the rationale for provisioning recommendations. User interfaces are through the z/OS console and the CPCC application.

The provisioning policy defines the circumstances under which additional capacity can be provisioned (when, which, and how). The criteria has three elements:

- ▶ A time condition is when provisioning is enabled:
 - Start time indicates when provisioning can begin.
 - Deadline indicates that the provisioning of the additional capacity is no longer enabled.
 - End time indicates when the deactivation of the additional capacity must begin.
- ▶ A workload condition is a description of which workload qualifies for provisioning. The parameters include this information:
 - The z/OS systems that might run the eligible work.
 - The importance filter indicates eligible service class periods, which are identified by Workload Manager (WLM) importance.
 - Performance index (PI) criteria:
 - Activation threshold. The PI of the service class periods must exceed the activation threshold for a specified duration before the work is considered to be suffering.
 - Deactivation threshold. The PI of the service class periods must fall below the deactivation threshold for a specified duration before the work is considered to no longer be suffering.
 - Included service classes are eligible service class periods.
 - Excluded service classes are service class periods that must not be considered.

If no workload condition is specified: The full capacity that is described in the policy is activated and deactivated at the start and end times that are specified in the policy.

- ▶ Provisioning scope is how much extra capacity can be activated, expressed in MSUs. CPD can contain multiple CPCs. There can be only one additional capacity specification per CPC, when this is expressed in MSUs.

The maximum provisioning scope is the maximum additional capacity that can be activated for all of the rules in the CPD.

The provisioning rule

The provisioning rule is *in the specified time interval, if the specified workload is behind its objective, up to the defined additional capacity can be activated*. The rules and conditions are named and stored in the CPP. For more information about z/OS Capacity Provisioning functions, see *z/OS MVS Capacity Provisioning User's Guide*, SA33-8299.

Planning considerations for using automatic provisioning

Although only one On/Off CoD offering can be active at any one time, several On/Off CoD offerings can be present on the zBC12. Changing from one to another requires that the active

one be stopped before the inactive one can be activated. This operation decreases the current capacity during the change.

The provisioning management routines can interrogate the installed offerings, their content, and the status of the content of the offering. To avoid the decrease in capacity, only create one On/Off CoD offering on the zBC12 by specifying the maximum possible capacity. The CPM can then, at the time that an activation is needed, activate a subset of the contents of the offering sufficient to satisfy the demand. Later, if more capacity is needed, the CPM can activate more capacity up to the maximum supported in the offering.

Having an unlimited number of offering records pre-staged on the SE hard disk is possible. Changing the content of the offerings if necessary is also possible.

Important: The CPM has control over capacity tokens for the On/Off CoD records. In a situation where a capacity token is completely consumed, the z13s server deactivates the corresponding offering record.

Therefore, prepare routines for detecting the warning messages about capacity tokens being nearly consumed, and have administrative routines in place for these situations. The messages from the system begin five days before a capacity token is fully used. To avoid capacity records from being deactivated in this situation, replenish the necessary capacity tokens before they are completely consumed.

In a situation where a CBU offering is active on a z13s server, and that CBU offering is 100% or more of the base capacity, activating any On/Off CoD is not possible. The On/Off CoD offering is limited to 100% of the base configuration.

The CPM operates based on WLM indications, and the construct used is the PI of a service class period. It is important to select service class periods that are appropriate for the business application that needs more capacity.

For example, the application in question might be running through several service class periods, where the first period might be the important one. The application might be defined as importance level 2 or 3, but it might depend on other work running with importance level 1. Therefore, determining which workloads to control and which service class periods to specify is important.

8.6 Capacity for Planned Event

CPE is offered with z13s servers to provide replacement backup capacity for planned downtime events. For example, if a server room requires an extension or repair work, replacement capacity can be installed temporarily on another z13s server in your environment.

Important: CPE is for planned replacement capacity only, and *cannot* be used for peak workload management.

CPE has these feature codes:

- ▶ FC 6833: Capacity for Planned Event enablement
- ▶ FC 0116: 1 CPE Capacity Unit
- ▶ FC 0117: 100 CPE Capacity Unit
- ▶ FC 0118: 10000 CPE Capacity Unit

- ▶ FC 0119: 1 CPE Capacity Unit-IFL
- ▶ FC 0120: 100 CPE Capacity Unit-IFL
- ▶ FC 0121: 1 CPE Capacity Unit-ICF
- ▶ FC 0122: 100 CPE Capacity Unit-ICF
- ▶ FC 0123: 1 CPE Capacity Unit-zAAP
- ▶ FC 0124: 100 CPE Capacity Unit-zAAP
- ▶ FC 0125: 1 CPE Capacity Unit-zIIP
- ▶ FC 0126: 100 CPE Capacity Unit-zIIP
- ▶ FC 0127: 1 CPE Capacity Unit-SAP
- ▶ FC 0128: 100 CPE Capacity Unit-SAP

The feature codes are calculated automatically when the CPE offering is configured. Whether using the eConfig tool or the Resource Link, a target configuration must be ordered consisting of a model identifier, a number of speciality engines, or both. Based on the target configuration, a number of feature codes from the previous list are calculated automatically, and a CPE offering record is constructed.

CPE is intended to replace capacity lost within the enterprise because of a planned event, such as a facility upgrade or system relocation. CPE is intended for short duration events that last up to a maximum of three days. Each CPE record, after it is activated, gives you access to dormant PUs on the server for which you have a contract (as described by the feature codes listed).

PUs can be configured in any combination of CP or specialty engine types (zIIP, SAP, IFL, and ICF). At the time of CPE activation, the contracted configuration is activated. The general rule of two zIIPs for each configured CP is maintained for the contracted configuration.

The processors that can be activated by CPE come from the available unassigned PUs on any installed CPC drawer. CPE features can be added to an existing z13s server non-disruptively. A one-time fee is applied for each individual CPE event, depending on the contracted configuration and its resulting feature codes. Only one CPE contract can be ordered at a time.

The base z13s configuration must have sufficient memory and channels to accommodate the potential requirements of the large CPE configuration. It is important to ensure that all required functions and resources are available on the z13s server where CPE is activated, including coupling facility (CF) LEVELs for CF partitions, memory, cryptographic functions, and connectivity capabilities.

The CPE configuration is activated temporarily, and provides extra PUs in addition to the amount the z13 server had installed as its original, permanent configuration. The number of additional PUs is predetermined by the number and type of feature codes configured, as described by the list of the feature codes. The number of PUs that can be activated is limited by the unused capacity that is available on the server.

When the planned event is over, the server must be returned to its original configuration. You can deactivate the CPE features at any time before the expiration date.

A CPE contract must be in place before the special code that enables this capability can be installed. CPE features can be added to an existing z13s server non-disruptively.

8.7 Capacity BackUp

CBU provides reserved emergency backup processor capacity for unplanned situations in which capacity is lost in another part of your enterprise, and you want to recover by adding the reserved capacity on a designated z13s server.

CBU is the quick, temporary activation of PUs, and is available in the following durations:

- ▶ For up to 90 consecutive days, in a loss of processing capacity as a result of an emergency or disaster recovery situation
- ▶ For 10 days for testing your disaster recovery procedures

Important: CBU is for disaster and recovery purposes only, and *cannot* be used for peak workload management or for a planned event.

8.7.1 Ordering

The CBU process enables CBU to activate CPs, ICFs, zAAPs, zIIPs, IFLs, and SAPs. To be able to use the CBU process, a CBU enablement feature (FC 9910) must be ordered and installed. You must order the quantity and type of PUs that you require. CBU has these feature codes:

- ▶ FC 9910: CBU enablement
- ▶ FC 6805: Additional test activations
- ▶ FC 6817: Total CBU years ordered
- ▶ FC 6818: CBU records ordered
- ▶ FC 6820: Single CBU CP-year
- ▶ FC 6821: 25 CBU CP-year
- ▶ FC 6822: Single CBU IFL-year
- ▶ FC 6823: 25 CBU IFL-year
- ▶ FC 6824: Single CBU ICF-year
- ▶ FC 6825: 25 CBU ICF-year
- ▶ FC 6826: Single CBU zAAP-year
- ▶ FC 6827: 25 CBU zAAP-year
- ▶ FC 6828: Single CBU zIIP-year
- ▶ FC 6829: 25 CBU zIIP-year
- ▶ FC 6830: Single CBU SAP-year
- ▶ FC 6831: 25 CBU SAP-year
- ▶ FC 6832: CBU replenishment

The CBU entitlement record (FC 6818) contains an expiration date that is established at the time of order, and depends upon the quantity of CBU years (FC 6817). You can extend your CBU entitlements through the purchase of more CBU years. The number of FC 6817 per instance of FC 6818 remains limited to five, and fractional years are rounded up to the nearest whole integer when calculating this limit.

For instance, if you have two years and eight months to the expiration date at the time of order, the expiration date can be extended by no more than two additional years. One test activation is provided for each additional CBU year added to the CBU entitlement record.

Feature code 6805 enables ordering additional tests in increments of one. The total number of tests that are supported is 15 for each feature code 6818.

The processors that can be activated by CBU come from the available unassigned PUs on any CPC drawer. The maximum number of CBU features that can be *ordered* is 20. The number of features that can be *activated* is limited by the number of unused PUs on the server.

However, the ordering system permits over-configuration in the order itself. You can *order* up to 20 CBU features regardless of the current configuration. However, at *activation*, only the capacity that is already installed can be *activated*. Note that at activation, you can decide to activate only a subset of the CBU features that are ordered for the system.

Subcapacity makes a difference in the way that the CBU features are done. On the full-capacity models, the CBU features indicate the amount of extra capacity needed. If the amount of necessary CBU capacity is equal to four CPs, the CBU configuration is four CBU CPs.

The number of CBU CPs must be equal to or greater than the number of CPs in the base configuration, and all of the CPs in the CBU configuration must have the same capacity setting. For example, if the base configuration is a two-way D02, providing a CBU configuration of a four-way of the same capacity setting requires two CBU feature codes.

If the required CBU capacity changes the capacity setting of the CPs, going from model capacity identifier D02 to a CBU configuration of a four-way E04 requires four CBU feature codes with a capacity setting of Exx.

If the capacity setting of the CPs is changed, more CBU features are required, not more physical PUs. Therefore, your CBU contract requires more CBU features if the capacity setting of the CPs is changed.

CBU can add CPs through LICCC only, and the z13s server must have the correct number of CPC drawers installed to support the required upgrade. CBU can change the MCI to a *higher* value than the base setting, but it does not change the *z13s* model. The CBU feature cannot *decrease* the capacity setting.

A CBU contract must be in place before the special code that enables this capability can be installed on the z13s server. CBU features can be added to an existing z13s server non-disruptively. For each machine enabled for CBU, the authorization to use CBU is available for a definite number of years (one - five).

The installation of the CBU code provides an alternative configuration that can be activated during an actual emergency. Five CBU tests, lasting up to 10 days each, and one CBU activation, lasting up to 90 days for a real disaster and recovery, are typically available in a CBU contract.

The alternative configuration is activated *temporarily*, and provides extra capacity greater than the server's original, *permanent* configuration. At activation time, you determine the capacity that is required for a situation, and you can decide to activate only a subset of the capacity that is specified in the CBU contract.

The base server configuration must have sufficient memory and channels to accommodate the potential requirements of the large CBU target configuration. Ensure that all required functions and resources are available on the backup z13s server, including CF LEVELS for CF partitions, memory, cryptographic functions, and connectivity capabilities.

When the emergency is over (or the CBU test is complete), the z13s server must be taken back to its original configuration. The CBU features can be deactivated by the customer at any time before the expiration date.

Failure to deactivate the CBU feature before the expiration date can cause the system to degrade gracefully back to its original configuration. The system does *not* deactivate dedicated engines, or the last of in-use shared engines.

Planning: CBU for processors provides a concurrent upgrade, resulting in more enabled processors, or changed capacity settings available to a z13s configuration, or both. You decide, at activation time, to activate a subset of the CBU features ordered for the system. Therefore, more planning and tasks are required for *nondisruptive* logical upgrades. See “Suggestions to avoid disruptive upgrades” on page 353.

For detailed instructions, see the *z Systems Capacity on Demand User's Guide*, SC28-6943-01.

8.7.2 CBU activation and deactivation

The activation and deactivation of the CBU function is a customer responsibility, and does not require the onsite presence of IBM service personnel. The CBU function is activated and deactivated concurrently from the HMC by using the API. On the SE, CBU is activated either by using the Perform Model Conversion task or through the API (the API enables task automation).

CBU activation

CBU is activated from the SE by using the Perform Model Conversion task, or through automation by using the API on the SE or the HMC. During a real disaster, use the Activate CBU option to activate the 90-day period.

Image upgrades

After the CBU activation, the z13s server can have more capacity, more active PUs, or both. The additional resources go into the resource pools, and are available to the LPARs. If the LPARs must increase their share of the resources, the LPARs' weight can be changed, or the number of logical processors can be concurrently increased by configuring reserved processors online.

The OS must have the capability to concurrently configure more processors online. If necessary, additional LPARs can be created to use the newly added capacity.

CBU deactivation

To deactivate the CBU, the additional resources must be released from the LPARs by the OSs. In certain cases, releasing resources is a matter of varying the resources offline. In other cases, it can mean shutting down OSs or deactivating LPARs. After the resources have been released, the same facility on the SE is used to turn off CBU. To deactivate CBU, click **Undo temporary upgrade** from the Perform Model Conversion task on the SE.

CBU testing

Test CBUs are provided as part of the CBU contract. CBU is activated from the SE by using the Perform Model Conversion task. Select the test option to start a 10-day test period. A standard contract can support five tests of this type. However, you can order more tests in increments of one up to a maximum of 15 for each CBU order.

Tip: The CBU test activation is done the same way as the real activation, by using the same SE Perform a Model Conversion window and selecting the **Temporary upgrades** option. The HMC windows have been changed to avoid real CBU activations by setting the test activation as the default option.

The test CBU has a 10-day limit, and must be deactivated in the same way as the real CBU, using the same facility through the SE. Failure to deactivate the CBU feature before the expiration date can cause the system to degrade gracefully back to its original configuration. The system does *not* deactivate dedicated engines, or the last of the in-use shared engine.

CBU example

Figure 8-16 shows an example of a CBU operation. The permanent configuration is C02, and a record contains three CP CBU features. During an activation, you can choose among many target configurations. With three CP CBU features, you can add one to three CPs, which enable you to activate C03, C04, or C05. Alternatively, two CP CBU features can be used to change the capacity level of permanent CPs, which means that you can activate D02, E02, and F02 through Z02.

Z01	Z02	Z03	Z04	Z05	Z06
Y01	Y02	Y03	Y04	Y05	Y06
X01	X02	X03	X04	X05	X06
W01	W02	W03	W04	W05	W06
V01	V02	V03	V04	V05	V06
U01	U02	U03	U04	U05	U06
T01	T02	T03	T04	T05	T06
S01	S02	S03	S04	S05	S06
R01	R02	R03	R04	R05	R06
Q01	Q02	Q03	Q04	Q05	Q06
P01	P02	P03	P04	P05	P06
O01	O02	O03	O04	O05	O06
N01	N02	N03	N04	N05	N06
M01	M02	M03	M04	M05	M06
L01	L02	L03	L04	L05	L06
K01	K02	K03	K04	K05	K06
J01	J02	J03	J04	J05	J06
I01	I02	I03	I04	I05	I06
H01	H02	H03	H04	H05	H06
G01	G02	G03	G04	G05	G06
F01	F02	F03	F04	F05	F06
E01	E02	E03	E04	E05	E06
D01	D02	D03	D04	D05	D06
C01	C02	C03	C04	C05	C06
B01	B02	B03	B04	B05	B06
A01	A02	A03	A04	A05	A06
1-way	2-way	3-way	4-way	5-way	6-way
Specialty Engine	Specialty Engine	Specialty Engine	Specialty Engine	Specialty Engine	Specialty Engine

Figure 8-16 Example of C02 with three CBU features

In addition, two CP CBU features can be used to change the capacity level of permanent CPs, and the third CP CBU feature can be used to add a CP, which enables the activation of D03, E03, and F03 through Z03. In this example, you are offered 49 possible configurations at activation time. While CBU is active, you can change the target configuration at any time.

8.7.3 Automatic CBU for Geographically Dispersed Parallel Sysplex

The intent of the IBM Geographically Dispersed Parallel Sysplex (GDPS) CBU is to enable automatic management of the PUs provided by the CBU feature during a server or site failure. Upon detection of a site failure or planned disaster test, GDPS concurrently adds CPs to the servers in the takeover site to restore processing power for mission-critical production workloads. GDPS automation does the following tasks:

- ▶ Performs the analysis that is required to determine the scope of the failure. This function minimizes operator intervention and the potential for errors.
- ▶ Automates the authentication and activation of the reserved CPs.
- ▶ Automatically restarts the critical applications after reserved CP activation.
- ▶ Reduces the outage time to restart critical workloads from several hours to minutes.

The GDPS service is for z/OS only, or for z/OS in combination with Linux on z Systems.

8.8 Nondisruptive upgrades

Continuous availability is an increasingly important requirement for most customers, and even planned outages are no longer acceptable. Although Parallel Sysplex clustering technology is the best continuous availability solution for z/OS environments, nondisruptive upgrades within a single server can avoid system outages, and are suitable to additional OS environments.

z13s servers support *concurrent* upgrades, meaning that dynamically adding more capacity to the CPC is possible. If OS images running on the upgraded CPC do not require disruptive tasks to use the new capacity, the upgrade is also *nondisruptive*. Therefore, POR, LPAR deactivation, and IPL do not have to take place.

If the concurrent upgrade is intended to satisfy an *image upgrade* to an LPAR, the OS running in this partition must also have the capability to concurrently configure more capacity online. IBM z/OS OSs have this capability. IBM z/VM can concurrently configure new processors and I/O devices online, and memory can be dynamically added to z/VM partitions.

If the concurrent upgrade is intended to satisfy the need for more OS images, more LPARs can be created *concurrently* on the z13s CPC, including all resources needed by such LPARs. These additional LPARs can be activated concurrently.

Linux OSs in general do *not* have the capability to add more resources concurrently. However, Linux, and other types of virtual machines running under z/VM, can benefit from the z/VM capability to non-disruptively configure more resources online (processors and I/O).

With z/VM, Linux guests can manipulate their logical processors by using the Linux CPU hotplug daemon. The daemon can start and stop logical processors based on the Linux *load average* value. The daemon is available in Linux SLES 10 SP2 and up, and in Red Hat Enterprise Linux (RHEL) V5R4 and up.

8.8.1 Processors

CPs, ICFs, zIIPs, IFLs, and SAPs can be concurrently added to a z13s server if unassigned PUs are available in any installed CPC drawer. The number of zIIPs cannot exceed twice the number of CPs, plus unassigned CPs. If necessary, more LPARs can be created concurrently to use the newly added processors.

The coupling facility control code (CFCC) can also configure more processors online to CF LPARs by using the CFCC image operations window.

8.8.2 Memory

Memory can be concurrently added up to the physically installed memory limit.

Using the previously defined reserved memory, z/OS OS images and z/VM partitions can dynamically configure more memory online, enabling nondisruptive memory upgrades. Linux on z Systems supports Dynamic Storage Reconfiguration.

8.8.3 I/O

I/O adapters can be added concurrently if all of the required infrastructure (I/O slots and fanouts) is present on the configuration.

Dynamic I/O configurations are supported by certain OSs (z/OS and z/VM), enabling nondisruptive I/O upgrades. However, having dynamic I/O reconfiguration on a stand-alone CF server is not possible because there is no OS with this capability running on this server.

8.8.4 Cryptographic adapters

Crypto Express5S features can be added concurrently if all of the required infrastructure is in the configuration.

8.8.5 Special features

Special features, such as Flash Express, IBM zEnterprise Data Compression (zEDC), and Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE), can also be added concurrently if all of the infrastructure is available in the configuration.

8.8.6 Concurrent upgrade considerations

By using MES upgrade, On/Off CoD, CBU, or CPE, a z13s server can be concurrently upgraded from one capacity identifier to another, either temporarily or permanently.

Enabling and using the additional processor capacity is transparent to most applications. However, certain programs depend on processor model-related information (for example, ISV products). You need to consider the effect on the software that is running on a z13s server when you perform any of these configuration upgrades.

Processor identification

The following instructions are used to obtain processor information:

- ▶ The STSI instruction: STSI reports the processor model and model capacity identifier for the base configuration and for any additional configuration changes through temporary upgrade actions. It fully supports the concurrent upgrade functions and is the preferred way to request processor information.
- ▶ The store CPU ID (STIDP) instruction: STIDP is provided for purposes of compatibility with a previous version.

Store system information instruction

The STSI instruction returns the MCI for the permanent configuration, and the MCI for any temporary capacity. This process is key to the functioning of CoD offerings.

Figure 8-17 shows the relevant output from the STSI instruction.

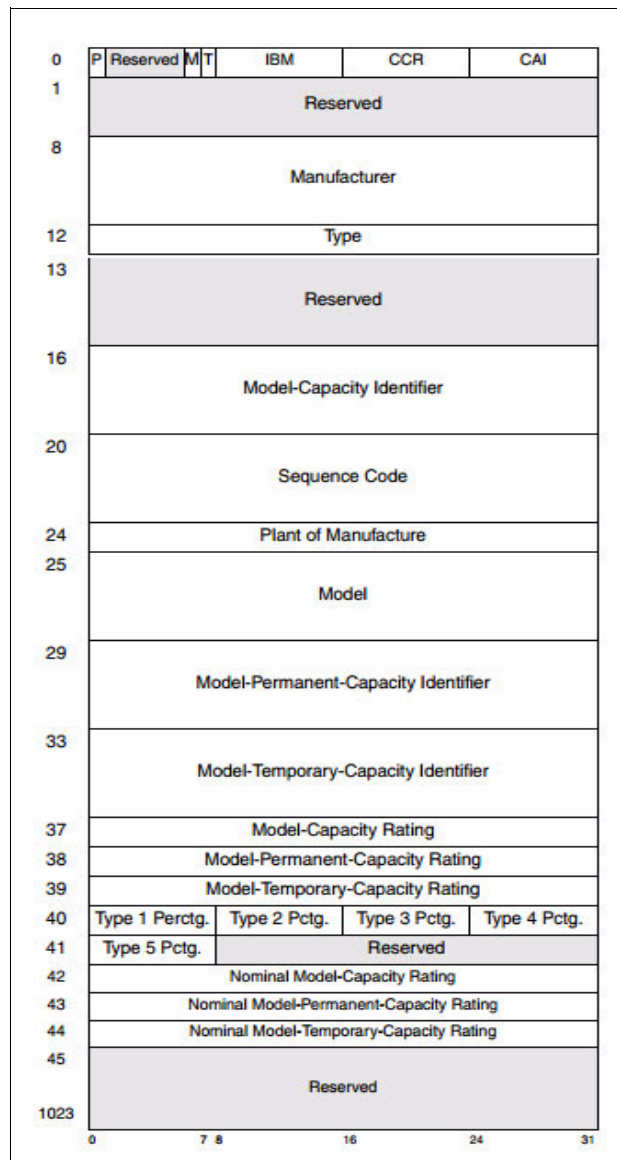


Figure 8-17 STSI output on z13s server

The MCI contains the base capacity, the On/Off CoD, and the CBU. The MPCl and the model permanent capacity rating (MPCR) contain the base capacity of the system, and the MTCI and model temporary capacity rating (MTCR) contain the base capacity and the On/Off CoD.

Store CPU ID instruction

The STIDP instruction provides information about the processor type, serial number, and LPAR identifier, as shown in Table 8-3. The LPAR identifier field is a full byte to support greater than 15 LPARs.

Table 8-3 STIDP output for z13s servers

Description	Version code	CPU ID number		Machine type number	LPAR two-digit indicator
Bit position	0 - 7	8 - 15	16 - 31	32 - 48	48 - 63
Value	x00 ^a	LPAR ID ^b	Six-digit number derived from the CPC serial number	x2965	x8000 ^c

- a. The version code for z13s servers is x00.
- b. The LPAR ID is a two-digit number in the range of 00 - 3F. It is assigned by the user on the image profile through the Support Element or HMC.
- c. High order bit on indicates that the LPAR ID value returned in bits 8 - 15 is a two-digit value.

When issued from an OS running as a guest under z/VM, the result depends on whether the **SET CPUID** command was used:

- ▶ Without the use of the **SET CPUID** command, bits 0 - 7 are set to FF by z/VM. However, the remaining bits are unchanged, which means that they are the same as they are without running as a z/VM guest.
- ▶ If the **SET CPUID** command was issued, bits 0 - 7 are set to FF by z/VM, and bits 8 - 31 are set to the value entered in the **SET CPUID** command. Bits 32 - 63 are the same as they are without running as a z/VM guest.

Table 8-4 lists the possible output returned to the issuing program for an OS running as a guest under z/VM.

Table 8-4 IBM z/VM guest STIDP output for z13s servers

Description	Version code	CPU identification number		Machine type number	LPAR two-digit indicator
Bit position	0 - 7	8 - 15	16 - 31	32 - 48	48 - 63
Without SET CPUID command	xFF	LPAR ID	4-digit number derived from the CPC serial number	x2965	x8000
With SET CPUID command	xFF	6-digit number as entered by the SET CPUID = nnnnnn command		x2965	x8000

Planning for nondisruptive upgrades

Online permanent upgrades, On/Off CoD, CBU, and CPE can be used to concurrently upgrade a z13s server. However, certain situations require a disruptive task to enable the new capacity that was recently added to the CPC. Several of these situations can be avoided if

planning is done in advance. Planning ahead is a key factor for enabling nondisruptive upgrades.

The following list describes the major reasons for disruptive upgrades. However, by carefully planning, and by reviewing “Suggestions to avoid disruptive upgrades”, you can minimize the need for these outages:

- ▶ When reserved storage was not previously defined, LPAR memory upgrades are disruptive to image upgrades. IBM z/OS and z/VM support this function.
- ▶ Installation of an extra CPC drawer to upgrade a Model N10 to a Model N20 (one or two drawers) is non-concurrent.
- ▶ When the OS cannot use the dynamic I/O configuration function, an I/O upgrade is disruptive to that LPAR. Linux, IBM z/Virtual Storage Extended (z/VSE), Transaction Processing Facility (TPF), IBM z/Transaction Processing Facility (z/TPF), and CFCC do not support the dynamic I/O configuration.

Suggestions to avoid disruptive upgrades

Based on the previous list of reasons for disruptive upgrades (“Planning for nondisruptive upgrades” on page 352), here are several suggestions for avoiding or at least minimizing these situations, increasing the potential for nondisruptive upgrades:

- ▶ Using an SE function under Operational Customization tasks called *Logical Processor add*, CPs and zIIPs can be added concurrently to a running partition. The initial or reserved number of processors can be dynamically changed.
- ▶ The OS that runs in the targeted LPAR must support the dynamic addition of resources, and be able to configure processors online. The total number of defined and reserved CPs cannot exceed the number of CPs supported by the OS. IBM z/OSV1R11, z/OS V1R12, and z/OS V1R13 with program temporary fixes (PTFs) supports up to 100 processors. IBM z/OS V2R1 also supports up to 141-way in non-SMT mode and up to 128-way in SMT mode. IBM z/VM supports up to 64 processors.
- ▶ Configure reserved storage to LPARs. Configuring reserved storage for all LPARs *before* their activation enables them to be non-disruptively upgraded. The OS running in the LPAR must be able to configure memory online. The amount of reserved storage can be higher than the CPC drawer threshold limit, even if the second CPC drawer is not installed. The current partition storage limit is 4 terabytes (TB). IBM z/OS and z/VM support this function.
- ▶ Consider the plan-ahead memory options. Use a convenient entry point for memory capacity, and consider the memory options to enable future upgrades within the memory cards that are already installed on the CPC drawers. For details about the offerings, see 2.4.3, “Memory configurations” on page 56.

Considerations when installing an additional CEC drawer

During an upgrade, an additional CEC drawer can be installed non-concurrently. Depending on your I/O configuration, a fanout rebalancing might be desirable for availability reasons.

8.9 Summary of capacity on-demand offerings

The CoD infrastructure and its offerings are major features for z13s servers. The introduction of these features was based on numerous customer requirements for more flexibility, granularity, and better business control over the z Systems infrastructure, operationally and financially.

One major customer requirement is to eliminate the necessity for a customer authorization connection to the IBM Resource Link system at the time of activation of any offering. This requirement was first met by the zBC12 and is also available with z13s servers.

After the offerings have been installed on z13s servers, they can be activated at any time, completely at the customer's discretion. No intervention by IBM or IBM personnel is necessary. In addition, the activation of the CBU does not require a password.

z13s servers can have up to eight offerings installed at the same time, with the limitation that only one of them can be an On/Off CoD offering. The others can be any combination of types. The installed offerings can be activated fully or partially, and in any sequence and any combination. The offerings can be controlled manually through command interfaces on the HMC, or programmatically through a number of APIs, so that IBM applications, ISV programs, or customer-written applications, can control the use of the offerings.

Resource consumption (and therefore financial exposure) can be controlled by using capacity tokens in On/Off CoD offering records.

The CPM is an example of an application that uses the CoD APIs to provision On/Off CoD capacity based on the requirements of the running workload. The CPM cannot control other offerings.

For detailed tutorials on all aspects of system upgrades, go to IBM Resource Link and select **Customer Initiated Upgrade Information** → **Education**. A list of available servers will help you select your particular product:

<https://www.ibm.com/servers/resourceLink/hom03010.nsf/pages/CIUInformation>

Registration is required to access IBM Resource Link.

For more information, see *z Systems Capacity on Demand User's Guide*, SC28-6943.



Reliability, availability, and serviceability

This chapter describes the reliability, availability, and serviceability (RAS) features of z13s servers.

The z13s design is focused on providing higher availability by reducing planned and unplanned outages. RAS can be accomplished with improved concurrent replace, repair, and upgrade functions for processors, memory, drawers, and I/O. RAS also extends to the nondisruptive capability for installing Licensed Internal Code (LIC) updates. In most cases, a capacity upgrade can be concurrent without a system outage. As an extension to the RAS capabilities, environmental controls are implemented in the system to help reduce power consumption and meet cooling requirements.

The design of the memory on z13s servers is based on the fully redundant memory infrastructure, redundant array of independent memory (RAIM). RAIM was first introduced with the z196. The z Systems servers are the only systems in the industry that offer this memory design.

RAS also provides digitally signed delivery and transmission of LIC, fixes, and restoration/backup files. Any data that is transmitted to IBM Support is encrypted. The design goal for z13s servers is to remove all sources of planned outages.

This chapter includes the following sections:

- ▶ The RAS strategy
- ▶ Availability characteristics
- ▶ RAS functions
- ▶ Enhanced Driver Maintenance
- ▶ RAS capability for the HMC and SE
- ▶ RAS capability for zBX Model 004
- ▶ Considerations for PowerHA in zBX environment
- ▶ IBM z Advanced Workload Analysis Reporter
- ▶ RAS capability for Flash Express

9.1 The RAS strategy

The RAS strategy is to manage change by learning from previous generations and investing in new RAS function to eliminate or minimize all sources of outages. Enhancements to z Systems RAS designs are implemented on the z13s system through the introduction of new technology, structure, and requirements. Continuous improvements in RAS are associated with new features and functions to ensure that z Systems servers deliver exceptional value to clients.

As described throughout this book, the z13s server introduced several changes from prior z Systems generations. Although the RAS design objective has not changed, many new RAS functions were introduced to mitigate changes in the server design. The RAS design on z13s servers is based on continuous improvements to address changes in technology, structure, complexity, and touches.

9.2 Availability characteristics

The following functions include the availability characteristics on z13s servers:

- ▶ Concurrent LIC memory upgrade

Memory can be upgraded concurrently by using LIC configuration code (LICCC) update if physical memory is available in the CPC drawers. If the physical memory cards must be changed, the z13s server needs to be powered down. To help ensure that the appropriate level of memory is available in a configuration, consider the plan-ahead memory feature.

The plan-ahead memory feature that is available with z13s servers provides the ability to plan for nondisruptive memory upgrades by having the system pre-plugged with dual inline memory modules (DIMMs) based on a target configuration. Pre-plugged memory is enabled when you place an order through LICCC.

- ▶ Enhanced Driver Maintenance (EDM)

One of the greatest contributors to downtime during planned outages is LIC driver updates performed in support of new features and functions. z13s servers are designed to support activating a selected new driver level concurrently.

- ▶ Concurrent fanout addition or replacement

A Peripheral Component Interconnect Express (PCIe) fanout, or host channel adapter (HCA2) fanout card provides the path for data between memory and I/O by using PCIe or InfiniBand cables. There is also a PCIe Integrated Coupling Adapter (ICA SR) fanout for external coupling connections. With z13s servers, a hot-pluggable and concurrently upgradeable fanout card is available.

Up to eight fanout cards are available per CPC drawer for a total of 16 fanout cards when both CPC drawers are installed. During an outage, a fanout card that is used for internal I/O can be concurrently repaired while redundant I/O interconnect ensures that no I/O connectivity is lost.

- ▶ Dynamic oscillator switchover

z13s servers have two oscillator cards: A primary and a backup. During a primary card failure, the backup card is designed to transparently detect the failure, switch over, and provide the clock signal to the system.

- ▶ IBM zAware

IBM z Systems Advanced Workload Analysis Reporter (IBM zAware) is an availability feature that is designed to use near real-time continuous learning algorithms, providing a

diagnostics capability that is intended to help you quickly pinpoint problems. This process in turn can help you to more rapidly address service disruptions.

IBM zAware uses analytics to examine z/OS and Linux on z Systems messages to find unusual patterns, inconsistencies, and variations. For more information about IBM zAware, see 9.8, “IBM z Advanced Workload Analysis Reporter” on page 367.

Deploying zAware Version 2: The zAware LPAR type is not supported on z13s or z13 at Driver level 27 servers. The IBM z Appliance Container Infrastructure (zACI) LPAR type is used instead. The new type of LPAR called *zACI* will be used to deploy the zAware firmware. During driver maintenance (see 9.4, “Enhanced Driver Maintenance” on page 360), the zAware LPAR is converted automatically to zACI.

► Flash Express

Internal flash storage is spread over two PCIe adapters, which mirror each other. If either card fails, the data is available on the other card. Data is stored over multiple flash devices in pairs, in a RAID configuration. If the flash device fails, the data is reconstructed dynamically. For more information about Flash Express, see 9.9, “RAS capability for Flash Express” on page 368.

► Redundant IBM z BladeCenter Extension configurations

Redundant hardware configurations in the IBM z BladeCenter Extension (zBX) provide the capacity to concurrently repair the BladeCenter components. Top-of-rack (TOR) switches present on the first zBX rack (frame B) are also redundant, which enables firmware application and repair actions to be fully concurrent.

Power Distribution Units (PDUs) provide redundant ($N+1$) connections to the main power source, improving zBX availability. The internal and external network connections are redundant throughout all of the zBX racks, TORs, and Blade Centers.

► Balanced Power Plan Ahead (FC 3003)

The Balanced Power Plan Ahead feature enables you to order the maximum number of Bulk Power Regulators (BPRs) on any server configuration. This feature helps to ensure that your configuration will be in a balanced power environment if you intend to add books and I/O drawers to your server in the future. Regardless of your configuration, all three BPR pairs will be shipped, installed, and activated. When this feature is ordered, a corequisite feature, the Line Cord Plan Ahead (FC 1901) is automatically selected.

► Processor unit (PU) sparing

The z13s Model N20 has two dedicated spare PUs to maintain performance levels should an active central processor (CP), Internal Coupling Facility (ICF), Integrated Facility for Linux (IFL), IBM z Integrated Information Processor (zIIP), integrated firmware processor (IFP), or system assist processor (SAP) fail.

Transparent sparing for failed processors is supported. A z13s Model N10 has no dedicated spares, but if not all processors are characterized, those unassigned PUs can be used as spares.

9.3 RAS functions

Unscheduled, scheduled, and planned outages have been addressed for the mainframe family of servers for many years:

- Unscheduled** This outage occurs because of an unrecoverable malfunction in a hardware component of the server.
- Scheduled** This outage is caused by changes or updates that must be done to the server in a timely fashion. A scheduled outage can be caused by a disruptive patch that must be installed, or other changes that must be made to the system.
- Planned** This outage is caused by changes or updates that must be done to the server. A planned outage can be caused by a capacity upgrade or a driver upgrade. A planned outage is usually requested by the customer, and often requires pre-planning. The z13s design phase focused on this pre-planning effort, and was able to simplify or eliminate it.

A fixed-size hardware system area (HSA) of 40 GB helps eliminate pre-planning requirements for HSA, and helps provide flexibility to dynamically update the configuration.

It is possible to perform the following tasks dynamically:

- ▶ Add a logical partition (LPAR) (up to 40)
- ▶ Add a logical channel subsystem (LCSS) (up to 3)
- ▶ Add a subchannel set (up to 3)
- ▶ Add a logical CP to an LPAR
- ▶ Add a cryptographic coprocessor
- ▶ Remove a cryptographic coprocessor
- ▶ Enable I/O connections
- ▶ Swap processor types
- ▶ Add memory by using LICCC
- ▶ Add a physical processor

In addition, by addressing the elimination of planned outages, the following tasks are also possible:

- ▶ Concurrent driver upgrades
- ▶ Periodic concurrent microcode control level (MCL) installation
- ▶ Concurrent and flexible Customer Initiated Upgrades (CIUs)

For a description of the flexible CIUs, see Chapter 8, “System upgrades” on page 311.

9.3.1 Unscheduled outages

An unscheduled outage occurs because of an unrecoverable malfunction in a hardware component of the z13s server.

The following improvements are designed to minimize unscheduled outages:

- ▶ Continued focus on firmware quality

For LIC and hardware design, failures are eliminated through rigorous design rules, design walk-throughs, and peer reviews. Failures are also eliminated through element, subsystem, and system simulation, and extensive engineering and manufacturing testing.

- ▶ Memory subsystem improvements

z13s servers use RAIM, which is a concept that is known in the disk industry as RAID. RAIM design detects and recovers from DRAM, socket, memory channel, or DIMM failures. The RAIM design includes the addition of one memory channel that is dedicated for RAS. The parity of the four *data* DIMMs is stored in the DIMMs that are attached to a fifth memory channel. Any failure in a memory component can be detected and corrected dynamically.

This design takes the RAS of the memory subsystem to another level, making it essentially a fully fault-tolerant $N+1$ design. The memory system on z13s servers is implemented with an enhanced version of the Reed-Solomon error correction code (ECC) that is known as 90B/64B, and includes protection against memory channel and DIMM failures.

A precise marking of faulty chips helps assure timely DRAM replacements. The key cache on the z13s memory is completely mirrored. For a full description of the memory system on z13s servers, see 2.4, “Memory” on page 53.

- ▶ Improved thermal and condensation management

- ▶ Soft-switch firmware

The capabilities of soft-switching firmware have been enhanced. Enhanced logic in this function ensures that every affected circuit is powered off during the soft switching of firmware components. For example, if you must upgrade the microcode of a Fibre Channel connection (FICON) feature, enhancements have been implemented to avoid any unwanted side effects that have been detected on previous servers.

- ▶ Server Time Protocol (STP) recovery enhancement

When HCA3-O (12xIFB) or HCA3-O Long Reach (LR) (1xIFB) or PCIe based ICA SR coupling links are used, an unambiguous “going away signal” is sent when the server on which the HCA3 is running is about to enter a failed (check stopped) state.

When the “going away signal” sent by the Current Time Server (CTS) in an STP-only Coordinated Timing Network (CTN) is received by the Backup Time Server (BTS), the BTS can safely take over as the CTS without relying on the previous Offline Signal (OLS) in a two-server CTN, or as the Arbiter in a CTN with three or more servers.

Enhanced Console Assisted Recovery (ECAR) is new with z13s and z13 GA2, and provides better recovery algorithms during a failing PTS scenario. It uses communication over the HMC/SE network to speed up the process of BTS takeover. See “Enhanced Console Assisted Recovery” on page 419 for more information.

9.3.2 Scheduled outages

Concurrent hardware upgrades, concurrent parts replacement, concurrent driver upgrade, and concurrent firmware fixes are available with z13s servers, and address the elimination of scheduled outages. Furthermore, the following indicators and functions that address scheduled outages are included:

- ▶ Double memory data bus lane sparing

This sparing reduces the number of repair actions for memory

- ▶ Single memory clock sparing

- ▶ Double-dynamic random access memory (DRAM) IBM Chipkill tolerance

- ▶ Field repair of the cache fabric bus

- ▶ Power distribution $N+2$ design

- ▶ Redundant humidity sensors
- ▶ Redundant altimeter sensors
- ▶ Corrosion sensors¹
- ▶ Unified support for the zBX

The zBX is supported the same as any other feature on z13s servers.

- ▶ Single processor core checkstop and sparing

This indicator implies that a processor core has malfunctioned and has been *spared*. IBM support personnel must consider what actions to perform, and also consider the history of z13s servers by asking the question, “Has this type of incident happened previously on this server?”

- ▶ Hot swap PCIe / InfiniBand hub cards

When properly configured for redundancy, hot swapping (replacing) these InfiniBand and PCIe fanout cards is possible:

- HCA2-optical, or HCA2-O (12xIFB)
- HCA3-optical, or HCA3-O (12xIFB)
- ICA-SR
- PCIe Gen3

This feature avoids any kind of interruption when the need for replacing these types of cards occurs.

- ▶ Redundant 1 gigabits per second (Gbps) Ethernet (GbE) support network with virtual local area network (VLAN)

The support network in the machine gives the machine code the capability to monitor each internal function in the machine. This capability helps to identify problems, maintain redundancy, and provide assistance while concurrently replacing a part. Through the implementation of the VLAN to the redundant internal Ethernet support network, the VLAN makes the support network itself easier to handle and more flexible.

- ▶ PCIe I/O drawer

The PCIe I/O drawer is available for z13s servers. It can be installed concurrently, as can all PCIe I/O drawer-supported features.

9.4 Enhanced Driver Maintenance

EDM is one more step toward reducing both the necessity for and the duration of a scheduled outage. One of the components to planned outages is LIC Driver updates that are run in support of new features and functions.

When correctly configured, z13s servers support concurrently activating a selected new LIC Driver level. Concurrent activation of the selected new LIC Driver level is supported only at specific released sync points. Concurrently activating a selected new LIC Driver level anywhere in the maintenance stream is not possible. A concurrent update/upgrade might not be possible for certain LIC updates.

Consider the following key points of EDM:

- ▶ The HMC can query whether a system is ready for a concurrent driver upgrade.

¹ The current implementation is just for collecting field data for analysis. System operation will not be affected by the availability or functions of this sensor.

- ▶ Previous firmware updates, which require an initial machine load (IML) of the z13s server to be activated, can block the ability to run a concurrent driver upgrade.
- ▶ An icon on the Support Element (SE) allows you or your IBM service support representative (SSR) to define the concurrent driver upgrade sync point to be used for an EDM.
- ▶ The ability to concurrently install and activate a driver can eliminate or reduce a planned outage.
- ▶ z13s servers introduce concurrent driver upgrade (CDU) cloning support to other CPCs for CDU preinstallation and activation.
- ▶ Concurrent crossover from Driver level N to Driver level $N+1$, and to Driver level $N+2$, must be done serially. No composite moves are allowed.
- ▶ Disruptive upgrades are permitted at any time, and allow for a composite upgrade (Driver N to Driver $N+2$).
- ▶ Concurrently backing up to the previous driver level is not possible. The driver level must move forward to driver level $N+1$ after EDM is initiated. Unrecoverable errors during an update can require a scheduled outage to recover.

The EDM function does not eliminate the need for planned outages for driver-level upgrades. Upgrades might require a system level or a functional element scheduled outage to activate the new LIC. The following circumstances require a scheduled outage:

- ▶ Specific complex code changes might dictate a disruptive driver upgrade. You are alerted in advance so that you can plan for the following changes:
 - Design data or hardware initialization data fixes
 - CFCC release level change
- ▶ OSA channel-path identifier (CHPID) code changes might require physical channel ID (PCHID) Vary OFF/ON to activate new code.
- ▶ Crypto CHPID code changes might require PCHID Vary OFF/ON to activate new code.

Note: zUDX clients should contact their UDX provider before installation of MCLs. Any changes to Segment 2 or 3 from a previous MCL level might require a change to the client's UDX. Attempting to install an incompatible UDX at this level will result in a Crypto checkstop

- ▶ Changes to the code of native PCIe features might require extra action from the client if the specific feature must be offline to the connecting LPARs before the new code can be applied and brought back online later.
- ▶ In changes to the Resource Group (RG) code, all native PCIe features within that RG might need to be varied offline to all connection LPARs by the client and back online after the code is applied.

z13s servers continue to support to activate MCLs concurrently on various channel types

9.5 RAS capability for the HMC and SE

The Hardware Management Console (HMC) and the SE have the following RAS capabilities:

- ▶ Back up from HMC and SE

On a scheduled basis, the HMC hard disk drive (HDD) is backed up to the USB flash memory drive, a customer provided FTP server, or both.

SE HDDs are backed up on to the primary SE HDD and alternate SE HDD. In addition, you can save the backup to a customer-provided FTP server.

For more information, see 11.2.4, “New backup options for HMCs and primary SEs” on page 393.

- ▶ Remote Support Facility (RSF)

The HMC RSF provides the important communication to a centralized IBM support network for hardware problem reporting and service. For more information, see 11.4, “Remote Support Facility” on page 403.

- ▶ Microcode Change Level (MCL)

Regular installation of MCLs is key for RAS, optimal performance, and new functions. Generally, plan to install MCLs quarterly at a minimum. Review hiper MCLs continuously. You must decide whether to wait for the next scheduled apply session, or schedule one earlier if your risk assessment of the hiper MCLs warrants.

For more information, see 11.5.4, “HMC and SE microcode” on page 409.

- ▶ Support Element (SE)

z13s servers are provided with two 1U System x servers inside the z Systems frame. One is always the primary SE and the other is the alternate SE. The primary SE is the active one. The alternate acts as the backup. Once per day, information is mirrored. The SE servers have N+1 redundant power supplies.

For more information, see 11.2.3, “New Support Elements” on page 393.

- ▶ Hardware Management Console (HMC) in an ensemble

The serviceability function for the components of an ensemble is delivered through the traditional HMC/SE constructs, as for earlier z Systems servers. From a serviceability point of view, all the components of the ensemble, including the zBX, are treated as z13s features. The zBX receives all of its serviceability and problem management through the HMC and SE infrastructure. All service reporting, including RSF functions, is delivered in a similar fashion to z13s servers.

The primary HMC for the ensemble is where portions of the Unified Resource Manager routines run. The Unified Resource Manager is an active part of the ensemble and z13s infrastructure. Therefore, the HMC is in a stateful state that needs high availability features to ensure the survival of the system during a failure. Therefore, each ensemble must be equipped with two HMCs: A primary and an alternate. The primary HMC performs all HMC activities (including Unified Resource Manager activities). The alternate is the backup, and cannot be used for tasks or activities.

Failover: The primary HMC and its alternate must be connected to the same LAN segment. This configuration allows the alternate HMC to take over the IP address of the primary HMC during failover processing.

For more information, see 11.6, “HMC in an ensemble” on page 428.

- ▶ Alternate HMC preload function

The Manage Alternate HMC task allows you to reload internal code onto the alternate HMC to minimize HMC downtime during an upgrade to a new driver level. After the new driver is installed on the alternate HMC, it can be made active by running an HMC switchover.

9.6 RAS capability for zBX Model 004

The zBX Model 004 exists only as a result of an MES from a zBX Model 002 or zBX Model 003. The zBX Model 004 is a stand-alone node that contains its own monitoring and controlling SEs. The two new 1U rack-mounted server SEs, along with their displays and keyboards, are added to the stand-alone zBX Model 004 frame B, below the TOR switches. zBX Model 004 includes all the RAS capabilities that are available in the previous models, and it was built with the traditional z Systems quality of service (QoS) to include RAS capabilities. The zBX Mod 004 offering provides extended service capability independent of the z13s hardware management structure.

zBX Model 004 is a stand-alone machine that can be added to an existing ensemble HMC as an individual ensemble member. The zBX Model 004 is required to be defined to the HMC. The ensemble HMC is used to run the zBX configuration and monitoring functions by using its connectivity to its internal SEs.

Apart from a zBX configuration with one chassis that are installed, the zBX is configured to provide $N + 1$ components. All the components are designed to be replaced concurrently. In addition, zBX configuration upgrades can be performed concurrently.

The zBX has two ToR switches. These switches provide $N + 1$ connectivity for the private networks between the zBX and its internal support elements for monitoring, controlling, and managing the zBX components.

BladeCenter components

Each BladeCenter has the following components:

- ▶ Up to 14 blade server slots. Blades can be removed, repaired, and replaced concurrently.
- ▶ ($N + 1$) Power Distribution Units (PDUs). If the PDUs have power inputs from two separate sources, during a single source failure, the second PDU takes over the total load of its BladeCenter.
- ▶ ($N + 1$) hot-swap power module with fan. A pair of power modules provides power for seven blades. A fully configured BladeCenter with 14 blades has a total of four power modules.
- ▶ ($N + 1$) 1 GbE switch modules for the power system control network (PSCN).
- ▶ ($N + 1$) 10 GbE High Speed switches for the intraensemble data network (IEDN).
- ▶ ($N + 1$) 1000BaseT switches for the intranode management network (INMN).
- ▶ ($N + 1$) 8 Gb FC switches for the external disk.
- ▶ ($N+1$) Internal 1U rack-mounted SEs with fully redundant ($N+1$) components.
- ▶ Two hot-swap advanced management modules (AMMs).
- ▶ Two hot-swap fans/blowers.

Maximums: Certain BladeCenter configurations do not physically fill up the rack with their components, but do reach other maximums, such as power usage.

zBX firmware

The testing, delivery, installation, and management of the zBX firmware is handled in the same way as z13s servers. The same processes and controls are used. All fixes to the zBX are downloaded to the zBX Model 004 internal SEs and applied to the zBX.

The MCLs for the zBX are designed to be concurrent and their status can be viewed at the z13s HMC.

9.6.1 zBX RAS and the IBM z Unified Resource Manager

The Hypervisor Management function of the Unified Resource Manager provides tasks for managing the hypervisor lifecycle, managing storage resources, performing RAS, and using the first-failure data capture (FFDC) features, and monitoring the supported hypervisors.

For blades that are deployed in a solution configuration, such as DataPower, the solution handles the complete end-to-end management for them and for their operating systems, middleware, and applications.

For blades that are deployed by the client, the Unified Resource Manager controls the blades:

- ▶ The client must have an entitlement for each blade in the configuration.
- ▶ When the blade is deployed in the BladeCenter chassis, the Unified Resource Manager powers up the blade, verifies that there is an entitlement for the blade, and verifies that the blade can participate in an ensemble. If these conditions are not met, the Unified Resource Manager powers down the blade.
- ▶ The blade is populated with the necessary microcode and firmware.
- ▶ The appropriate hypervisor is loaded onto the blade.
- ▶ The management scope is deployed according to which management enablement level is present in the configuration.
- ▶ The administrator can define the blade profile, and the profiles for virtual servers to run on the blade, through the HMC.

Based on the profile for individual virtual servers inside the deployed hypervisor, the virtual servers can be activated and an operating system can be loaded following the activation. For client-deployed blades, all of the application, database, operating system, and network management is handled by the client's usual system management disciplines.

9.6.2 zBX Model 004: 2458-004

z13s servers can only be in an ensemble that supports a zBX Model 004. When upgrading a z114 or a zBC12 with zBX to a z13s server, the zBX must also be upgraded from a Model 002 or a Model 003 to a zBX Model 004.

The zBX Model 004 is based on the BladeCenter and blade hardware offerings that contain IBM certified components. zBX Model 004 BladeCenter and blade RAS features are extended considerably for IBM z Systems servers:

- ▶ Hardware redundancy at various levels:
 - Redundant power infrastructure
 - Redundant power and switch units in the BladeCenter chassis
 - Redundant cabling for management of zBX and data connections
 - Redundant 1U rack-mounted support elements
- ▶ Concurrent to system operations:
 - Install more blades
 - Hardware repair
 - Firmware fixes and driver upgrades
 - Automated call home for hardware/firmware problems

Important: Depending on the type of hardware repair being performed and the firmware fixes being installed or activated, a deactivation of a target blade might be required.

The zBX has two pairs of ToR switches, the INMN *N*+1 pair and the IEDN *N*+1 switch pair, which are installed in Frame B. The management switch pair (INMN) provides *N*+1 connectivity for the private networks between the internal zBX support elements SEs and the zBX hardware. The connection is used for monitoring, controlling, and managing the zBX components. The data switch pair (IEDN) provides *N*+1 connectivity for the data traffic between the defined virtual servers and client’s networks.

Because of the MES conversion from previous zBX Models, a zBX Model 004 became a stand-alone machine, independent of the z Systems servers for management and control functions. The zBX Model 004 internal SEs are used for firmware updates and other related service activities. Therefore, the distance restriction of 23 meters from the CPC to the zBX that existed with the previous models was removed.

Not only hardware and firmware provide RAS capabilities. The operating system can also contribute to improving RAS. IBM PowerHA® SystemMirror® for AIX (PowerHA) supports the zBX PS701 8406-71Y blades. PowerHA enables setting up a PowerHA environment on the stand-alone zBX.

Table 9-1 provides more information about PowerHA and the required AIX levels that are needed for a PowerHA environment on zBX. AIX 6.1 Technology Level (TL)06 Service Pack (SP) 3 with RSCT 3.1.0.4 (packaged in Cluster Systems Management (CSM) PTF 1.7.1.10 installed with AIX 6.1.6.3) is the preferred baseline for zBX Virtual Servers running AIX.

Table 9-1 PowerHA and required AIX levels

IBM zBX Model 004	AIX V6.1	AIX V7.1
PowerHA V6.1	AIX V6.1 TL05 RSCT 2.5.5.0	PowerHA V6.1 SP3 AIX V7.1 RSCT V3.1.0.3
PowerHA V7.1	AIX V6.1 TL06 RSCT V3.1.0.3	AIX V7.1 RSCT V3.1.0.3

zBX Model 004 includes major firmware changes in almost all areas compared to the zBX Model 003. zBX Model 004 takes the RAS concept of the zBX to higher levels.

9.7 Considerations for PowerHA in zBX environment

An application that runs on AIX can be provided with high availability by using the PowerHA SystemMirror for AIX (formerly known as IBM HACMP™). PowerHA is easy to configure because it is menu-driven, and provides high availability for applications that run on AIX.

PowerHA helps define and manage resources that are required by applications that run on AIX. It provides service/application continuity through system resources and application monitoring, and automated actions (start/manage/monitor/restart/move/stop).

Tip: Resource movement and application restart on the second server are known as *failover*.

Automating the failover process speeds up recovery and allows for unattended operations, improving application availability. In an ideal situation, an application must be available 24 x 7. Application availability can be measured as the amount of time that the service is available, divided by the amount of time in a year, as a percentage.

A PowerHA configuration (also known as a *cluster*) consists of two or more servers² (up to 32) that have their resources managed by PowerHA cluster services. The configuration provides automated service recovery for the managed applications. Servers can have physical or virtual I/O resources, or a combination of both.

PowerHA performs the following functions at the cluster level:

- ▶ Manage and monitor operating system and hardware resources
- ▶ Manage and monitor application processes
- ▶ Manage and monitor network resources (service IP addresses)
- ▶ Automate application control (start/stop/restart/move)

The virtual servers that are defined and managed in zBX use only virtual I/O resources. PowerHA can manage both physical and virtual I/O resources (virtual storage and virtual network interface cards).

PowerHA can be configured to perform automated service recovery for the applications that run in virtual servers that are deployed in zBX. PowerHA automates application failover from one virtual server in an IBM System p blade to another virtual server in a different System p blade with a similar configuration.

Failover protects service (masks service interruption) in an unplanned or planned (scheduled) service interruption. During failover, you might experience a short service unavailability while resources are configured by PowerHA on the new virtual server.

The PowerHA configuration for the zBX environment is similar to standard Power environments, except that it uses only virtual I/O resources. Currently, PowerHA for zBX support is limited to fail over inside the same ensemble. All zBXs participating in the PowerHA cluster must have access to the same storage.

The PowerHA configuration includes the following tasks:

- ▶ Network planning (VLAN and IP configuration definition and for server connectivity)
- ▶ Storage planning (shared storage must be accessible to all blades that provide resources for a PowerHA cluster)

² Servers can be also virtual servers. One server is one instance of the AIX operating system.

- ▶ Application planning (start/stop/monitoring scripts and operating system, processor, and memory resources)
- ▶ PowerHA software installation and cluster configuration
- ▶ Application integration (integrating storage, networking, and application scripts)
- ▶ PowerHA cluster testing and documentation

A typical PowerHA cluster is shown in Figure 9-1.

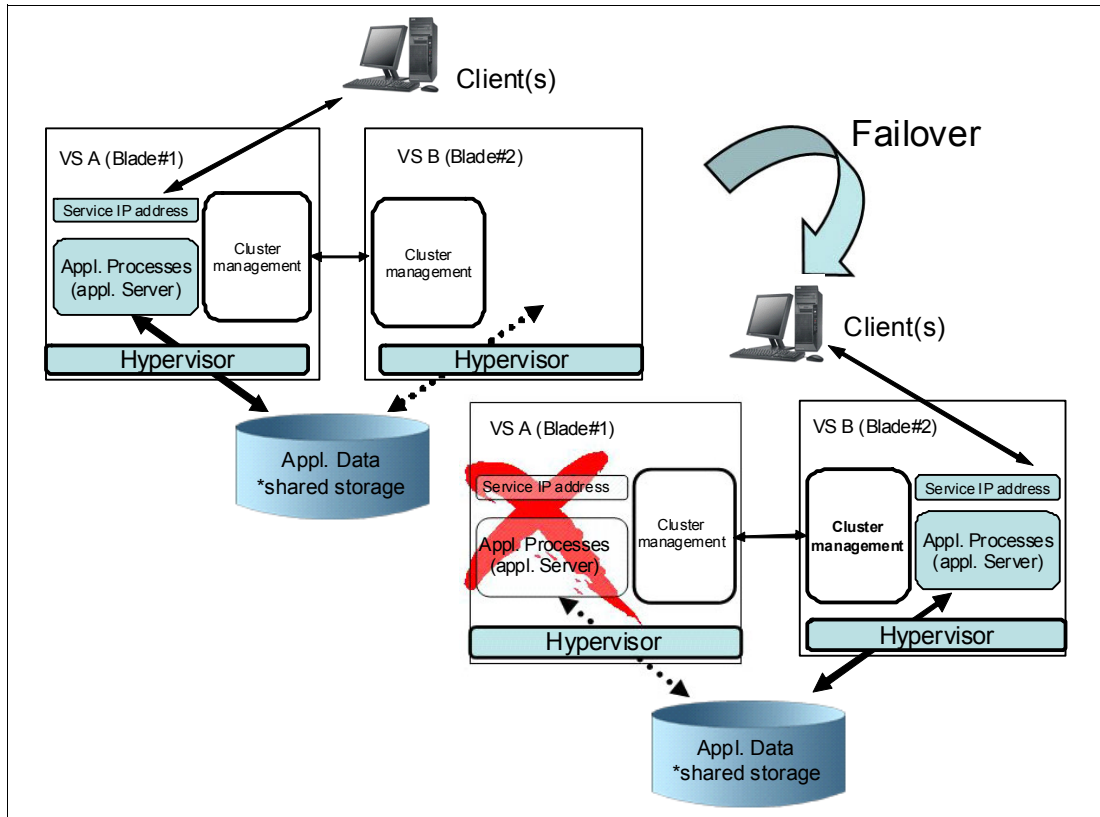


Figure 9-1 Typical PowerHA cluster diagram

For more information about IBM PowerHA SystemMirror for AIX, see this website:

<http://www.ibm.com/systems/power/software/availability/aix/index.html>

9.8 IBM z Advanced Workload Analysis Reporter

IBM z Advanced Workload Analysis Reporter (IBM zAware) provides a solution for detecting and diagnosing anomalies in z/OS and Linux on z Systems by analyzing software logs and highlighting abnormal events. It represents a first in a new generation of “smart monitoring” products with pattern-based message analysis.

IBM zAware runs as a firmware virtual appliance in a z13s LPAR. It is an integrated set of analytic applications that creates a model of normal system behavior that is based on prior system data. It uses pattern recognition techniques to identify unexpected messages in current data from the z/OS or Linux on z Systems servers that it is monitoring. This analysis of events provides nearly real-time detection of anomalies. These anomalies can then be easily viewed through a graphical user interface (GUI).

IBM zAware improves the overall RAS capability of z13s servers by providing these advantages:

- ▶ Identify when and where to look for a problem
- ▶ Drill down to identify the cause of the problem
- ▶ Improve problem determination in near real time
- ▶ Reduce problem determination efforts significantly

For more information about IBM zAware, see Appendix B, “IBM z Systems Advanced Workload Analysis Reporter (IBM zAware)” on page 453.

9.9 RAS capability for Flash Express

Flash Express cards come in pairs for availability, and are exclusively in PCIe I/O drawers. Similar to other PCIe I/O cards, redundant PCIe paths to Flash Express cards are provided by redundant IO interconnect. Unlike other PCIe I/O cards, they can be accessed only by the host by using a unique protocol.

In each Flash Express card, data is stored in four solid-state drives (SSDs) in a RAID configuration. If an SSD fails, the data is reconstructed dynamically. The cards in a pair mirror each other over a pair of SAS cables, in a RAID 10 configuration. If either card fails, the data is available on the other card. Card replacement is concurrent, and does not cause disruption to your operations.

The data is always stored encrypted with a volatile key, and the card is usable only on the system with the key that encrypted it. For key management, both the Primary and Alternate SEs have a smart card reader installed.

Flash Express cards support concurrent firmware upgrades.

Figure 9-2 shows the various components that support Flash Express RAS functions.

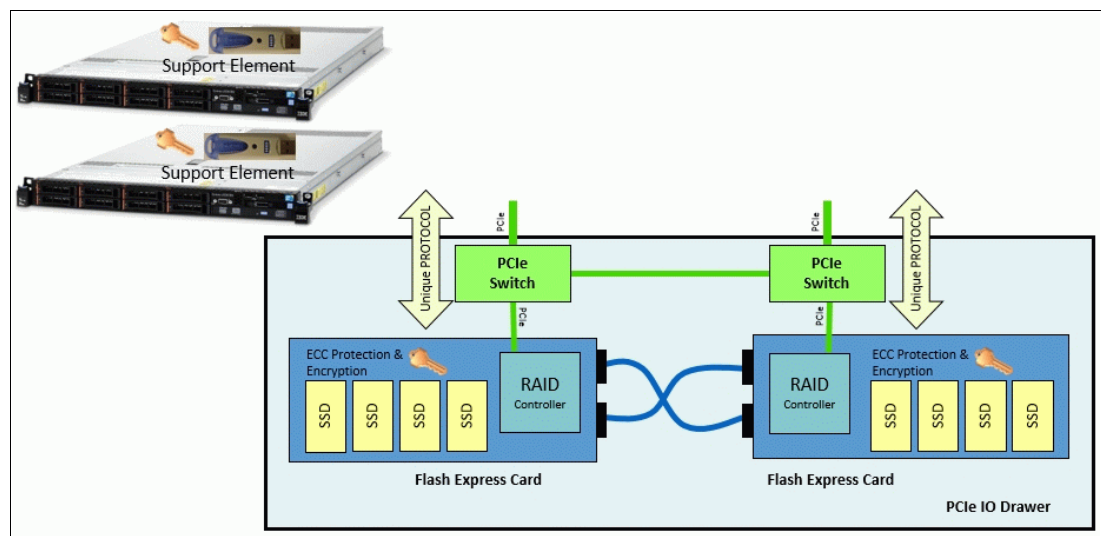


Figure 9-2 Flash Express RAS components

For more information about Flash Express, see Appendix H, “Flash Express” on page 529



Environmental requirements

“You can’t make a product greener, whether it’s a car, a refrigerator, or a city, without making it smarter: Smarter materials, smarter software, or smarter design.”

— “The Green Road Less Traveled” by Thomas L. Friedman, *The New York Times*, July 15, 2007

This chapter describes the environmental requirements for IBM z Systems servers. It lists the dimensions, weights, power, and cooling requirements as an overview of what is needed to plan for the installation of an IBM z Systems z13s server and IBM z BladeCenter Extension (zBX).

The IBM z13s server is an air cooled system with front to back airflow. Air temperature, altitude, and humidity requirements meet the ASHRAE¹ class A3 specifications.

The physical installation of z13s servers has a number of options, including raised floor and non-raised floor options, cabling from the bottom of the frame or off the top of the frame, and the option to have a high-voltage DC power supply directly into the z13s server, instead of the usual AC power supply.

For comprehensive physical planning information, see *z13s Installation Manual for Physical Planning*, GC28-6953.

This chapter includes the following sections:

- ▶ IBM z13s power and cooling
- ▶ IBM z13s physical specifications
- ▶ IBM zBX environmental components
- ▶ Energy management

¹ Refer to ASHRAE’s document “Thermal Guidelines for Data Processing Environments, Third Edition” at: <https://www.ashrae.org/resources--publications/bookstore/datacom-series#thermalguidelines>.

10.1 IBM z13s power and cooling

z13s servers have a number of options for installation on a raised floor, or on a non-raised floor. Furthermore, the cabling can come into the bottom of the machine, or into the top of the machine.

10.1.1 Power and I/O cabling

The z13s server is an air-cooled, one-frame system. Installation can be on a raised floor or non-raised floor, with numerous options for top exit or bottom exit for all cabling (power cords and I/O cables) as shown in Figure 10-1.

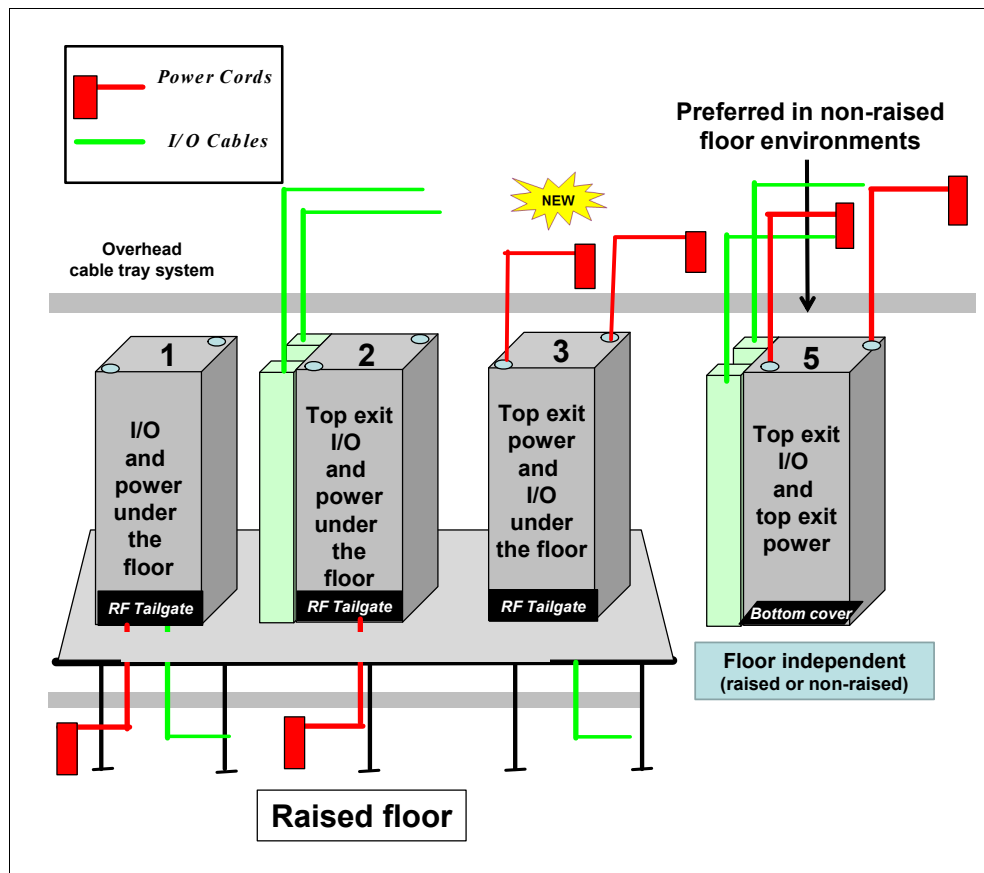


Figure 10-1 IBM z13s cabling options

Top Exit I/O

Like the zBC12, z13s servers support the Top Exit I/O Cabling feature (FC 7920). This feature routes all coupling links and all I/O cables, including a 1000BASE-T Ethernet cable from I/O drawers or PCIe I/O drawers through two frame extensions, out the top of the frame.

Figure 10-2 shows the frame extensions, also called *chimneys*, that are installed to each corner on the left side of the A frame when the Top Exit I/O Cabling feature (FC 7920) is ordered. The bottom of the chimney is closed with welded sheet metal.

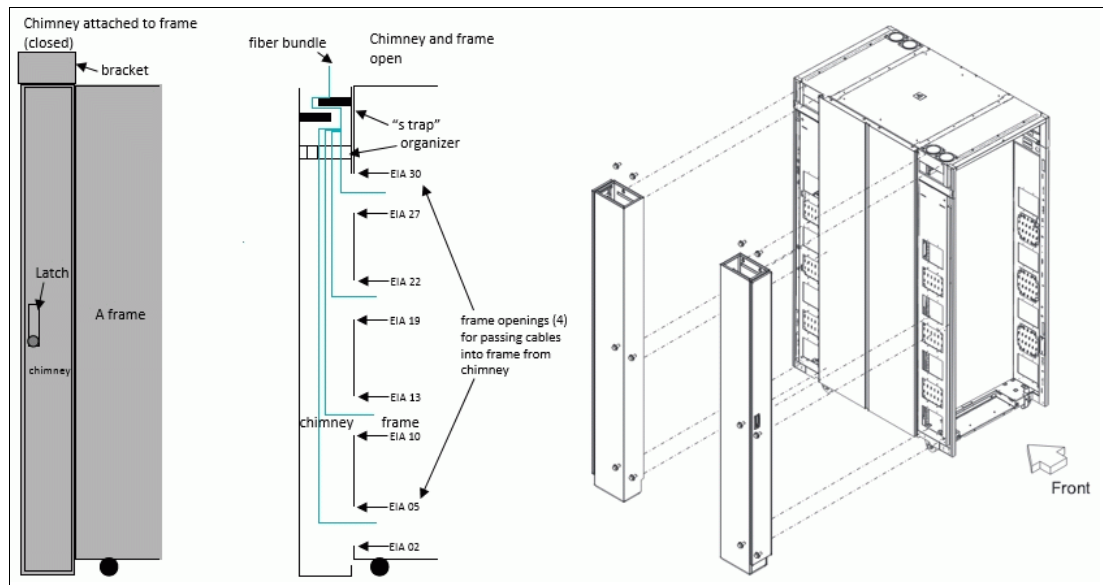


Figure 10-2 Top Exit I/O cabling feature

The Top Exit I/O Cabling feature adds 15 cm (6 in.) to the width of the frame and about 60 lbs (27.3 kg) to the weight.

10.1.2 Power consumption

The system operates with two redundant power supplies. Each of the power supplies has its individual power cords. For redundancy, the server must have two power feeds. Each power feed is one power cord. Power cords attach to either 3-phase, 50/60 hertz (Hz), 250 or 450 volt (V) alternating current (AC) power or 380 to 520 V direct current (DC) power.

Depending on the configuration, single-phase power cords can be used (200 V 30 amps (A)). See the shaded area in Table 10-2 on page 372. The total loss of one power feed has no effect on the system operation.

For ancillary equipment, such as the Hardware Management Console (HMC) and its display, extra single-phase outlets are required.

The power requirements depend on the number of central processor complex (CPC) drawers, and the number of I/O drawers, that are installed. Table 10-1 lists the maximum power requirements for z13s servers.

These numbers assume the maximum memory configurations, all drawers are fully populated, and all fanout cards installed. Use the Power Estimation tool to obtain a precise indication for a particular configuration. For more information, see 10.4.1, “Power estimation tool” on page 381.

Table 10-1 Utility power consumption

I/O drawer / PCIe I/O drawer	N10 (FC 1043)	N20 - one CPC drawer (FC 1044)	N20 - two CPC drawers (FC 1045)
	(kW) ^a		
0/0	2.8 ^b	4.0 ^b	6.4
1/0	3.6 ^b	4.9 ^b	7.4
0/1	4.4 ^b	5.6 ^b	8.0
1/1	5.3	6.7	8.9
0/2	N/A	7.4	9.6
1/2	N/A	8.2	N/A
Notes:			
<ul style="list-style-type: none"> ▶ Three-phase input power is available for all configurations. ▶ A balanced three-phase option is offered which BPRs per side depending on the configuration. This option is available on all configurations that are not already balanced. ▶ DC input power is available on all configurations. There is no balanced power option for DC input. A mix of AC and DC power on any system is not permitted. ▶ Power is shown as maximum values for worst case I/O configurations. Your actual power usage will be something less than shown here. 			

a. Environmental conditions: Warm room (40oC) and altitude of less than 1500 feet with system blowers speed increased for the appropriate environmental zone.

b. Indicates that single phase input power is available. There is no balanced power option for single phase input power.

10.1.3 Internal Battery Feature

The optional Internal Battery Feature (IBF) provides sustained system operations for a relatively short time, enabling an orderly shutdown. In addition, an external uninterruptible power supply system can be connected, enabling longer periods of sustained operation.

If the batteries are not older than three years and have been discharged regularly, the IBF is capable of providing emergency power for the periods of time that are listed in Table 10-2.

Table 10-2 IBM hold-up times

I/O drawer / PCIe I/O drawer	N10 (FC 1043)	N20 - one CPC drawer (FC 1044)	N20 - two CPC drawers (FC 1045)
	minutes		
0/0	22.8	13.3	7.2
1/0	15.5	9.9	6.2
0/1	11.7	8.3	4.9

I/O drawer / PCIe I/O drawer	N10 (FC 1043)	N20 - one CPC drawer (FC 1044)	N20 - two CPC drawers (FC 1045)
	minutes		
1/1	9.0	6.9	4.2
0/2	N/A	6.2	3.7
1/2	N/A	4.8	N/A
Notes:			
<ul style="list-style-type: none"> ▶ The numbers shown assume both BPAs feeding power with the batteries charged at full capacity. ▶ Hold-up times are influenced by temperature, battery age, and fault conditions within the system. 			

10.1.4 Balanced Power Plan Ahead

A Balanced Power Plan Ahead feature is available for future growth, also assuring adequate and balanced power for all possible configurations. With this feature, downtimes for upgrading a system are eliminated by including the maximum power requirements in terms of Bulk Power Regulators (BPRs) and power cords with the initial installation. The Balanced Power Plan Ahead feature is not available with DC and 1-phase power cords.

10.1.5 Emergency power-off

On the front of the frame is an emergency power-off switch that, when activated, immediately disconnects utility and battery power from the server. This method causes all volatile data in the z13s servers to be lost.

If the z13s server is connected to a machine room's emergency power-off switch, and the IBF is installed, the batteries take over if the switch is engaged. To avoid this switchover, connect the machine room emergency power-off switch to the z13s power-off switch. Then, when the machine room emergency power-off switch is engaged, all power is disconnected from the power cords and the IBF. However, all volatile data in the z13s server will be lost.

Unit Emergency Power Off switch cover

z13s servers provide Unit Emergency Power Off (UEPO) switch protection from accidental tripping. The option consists of a clear cover that can be installed into the UEPO switch opening. The cover itself and the additional hardware that is required to install the option are included in the z13s Shipping Group material. Determine whether you want this optional UEPO cover installed.

10.1.6 Cooling requirements

z13s servers are air cooled. They require chilled air, ideally coming from under a raised floor, to fulfill the air-cooling requirements. However, a non-raised floor option is available. The requirements for cooling are indicated in *z13s Installation Manual for Physical Planning*, GC28-6953.

The front and the rear of z13s frame dissipate separate amounts of heat. Most of the heat comes from the rear of the system. To calculate the heat output expressed in kilo British Thermal Units (kBTU) per hour for z13s configurations, multiply the table entries from Table 10-2 on page 372 by 3.4. The planning phase must consider correct placement of z13s servers in relation to the cooling capabilities of the data center.

10.1.7 New environmental class for IBM z13s servers: ASHRAE Class A3

ASHRAE² (American Society of Heating, Refrigerating, and Air-Conditioning Engineers) is an organization devoted to the advancement of indoor-environment-control technology in the heating, ventilation, and air conditioning industry.

The IBM z13s server is the first z Systems server to meet the ASHRAE Class A3 specifications. Environmental specifications are presented in two categories: *Required* and *Recommended*. Obviously, meeting the required specifications is prerequisite to using the z13s servers.

IBM strongly suggests you strive for more than the minimum requirements. The powerful computing that z13s servers provide generates a great deal of heat. That heat must be removed from the equipment to keep it operating at peak efficiency. Cooling the servers can result in condensation on critical internal parts, leading to equipment failure, unless the computer room environment is adequately maintained to prevent it. That is where operating your data center with the goal of reaching recommended specifications instead of just the required numbers will pay off for you.

Figure 10-3 depicts the environmental limits for previous and current generation of z Systems servers.

zEC12	A1	Temp range 15C (59F) to 32C (89.6F) RH range 20% to 80% Max dew point 17C (62.6F) (note 1) Alt 10,000 ft.	
zBC12	A2	Temp range 10C (50F) to 35C (95F) RH range 20% to 80% Max dew point 21C (69.8F) (note 1) Alt 10,000 ft.	z13
	A3	Temp range 5C (41F) to 40C (104F) RH range -12C (10.4F) min dewpoint; 8% to 85% Max dew point 24C (75.2F) (note 1) Alt 10,000 ft.	z13s
Recommended (all)		Temp range 18C (64.4F) to 27C (80.6F) RH range 5.5C (41.9F) min dewpoint; up to 60% Max dew point 15C (59F) (note 1)	
Note 1: Actual inlet air moisture content range (grams moisture/Kg dry air)			
A1: 2 - 12			
A2: 1.5 - 16			
A3: 1.5 - 19			
Recommended: 6.2 - <10			

Figure 10-3 Recommended environmental conditions

10.2 IBM z13s physical specifications

This section describes the weights and dimensions of z13s servers.

² <http://www.ashrae.org>

10.2.1 Weights and dimensions

Installation can be on a raised floor or a non-raised floor. You can find the most up-to-date details about the installation requirements for z13s servers in *z13s Installation Manual for Physical Planning*, GC28-6953.

Table 10-3 indicates the maximum system dimensions and weights for the z13s models.

Table 10-3 IBM z13s physical dimensions and weights

Maximum	z13s single frame								
	Model N10			Model N20- 1 drawer			Model N20- 2 drawers		
	W / O IBFs	With IBF	With IBF and overhead cabling	W / O IBF	With IBF	With IBF and overhead cabling	W / O IBFs	With IBF	With IBF and overhead cabling
Weight in kilograms (kg)	810	896	924	944	1030	1058	944	1030	1058
Weight in pounds (lbs.)	1785	1975	2037	2081	2270	2232	2081	2270	2232
Height with covers ^a Width with covers ^a Depth with covers	201.5 centimeters (cm) or 79.3 inches (in.) (42 Electronic Industries Alliance (EIA)) 78.5 cm (30.9 in.) 159.5 cm (62.8 in.)								
Height reduction Width reduction	180.9 cm (71.2 in.) None								
Machine area Service clearance	1.24 square meters (13.3 square feet) 3.22 square meters (34.7 square feet) (IBF contained within the frame)								
Note:									
▶ The Balanced Power Plan Ahead feature adds approximately 51 kg (112 lbs.)									

a. The width increases by 15.2 cm (6 in.), and the height increases by 14.0 cm (5.5 in.) if overhead I/O cabling is configured.

10.2.2 Four-in-one (4-in-1) bolt-down kit

A bolt-down kit can be ordered for the z13s frame. The kit provides hardware to enhance the ruggedness of the frame, and to tie down the frame to a concrete floor. The kit is offered in the following configurations:

- ▶ The Bolt-Down Kit for a raised floor installation (FC 8021) provides frame stabilization and bolt-down hardware for securing the frame to a concrete floor beneath the raised floor. The kit covers raised floor heights from 15.2 cm (6 in.) to 91.4 cm (36 in.).

The kit helps to secure the frame and its content from damage caused by shocks and vibrations, such as those generated by an earthquake.

10.3 IBM zBX environmental components

The following sections provide information about the environmental components in summary for zBX. For a full description of the environmental components for the zBX, see *zBX Model 004 Installation Manual for Physical Planning*, GC27-2630.

Note: Only zBX Model 004 is supported by z13s servers. zBX Model 004 only exists as a result of an miscellaneous equipment specification (MES) from a previous zBX Model 002 or Model 003.

Also, the only additional features that can be ordered to a zBX Model 004 are the PS501 and HX5 Blades to fill up empty slot entitlements.

10.3.1 IBM zBX configurations

The zBX can have from one to four racks. The racks are shipped separately, and are bolted together at installation time. Each rack can contain up to two BladeCenter chassis, and each chassis can contain up to 14 single-wide blades. The number of required blades determines the actual components that are required for each configuration. The number of BladeCenters and racks depends on the quantity of blades (see Table 10-4).

Table 10-4 IBM zBX configurations

Number of blades	Number of BladeCenters	Number of racks
7	1	1
14	1	1
28	2	1
42	3	2
56	4	2
70	5	3
84	6	3
98	7	4
112	8	4

A zBX can be populated by up to 112 Power 701 blades. A maximum of 56 IBM BladeCenter HX5 blades can be installed in a zBX. For DataPower blades, the maximum number is 28. Note that the DataPower blade is a double-wide blade.

10.3.2 IBM zBX power components

The zBX has its own power supplies and cords, which are independent of the z13s server power. Depending on the configuration of the zBX, up to 16 customer-supplied power feeds can be required. A fully configured four-rack zBX has 16 Power Distribution Units (PDUs). The zBX has these power specifications:

- ▶ 50/60 Hz AC power
- ▶ Voltage (240 V)
- ▶ Both single-phase and three-phase wiring

PDUs and power cords

The following PDU options are available for the zBX:

- ▶ FC 0520 - 7176: Model 3NU with attached power cord (US)
- ▶ FC 0521 - 7176: Model 2NX (worldwide (WW))

The following power cord options are available for the zBX:

- ▶ FC 0531: 4.3 m (14.1 ft.), 60A/208 V, US power cord, Single Phase.
- ▶ FC 0532: 4.3 m (14.1 ft.), 63A/230 V, non-US power cord, Single Phase.
- ▶ FC 0533: 4.3 m (14.1 ft.), 32A/380 V-415 V, non-US power cord, Three Phase. Note that 32 A WYE 380 V or 415 V gives you 220 V or 240 V line to neutral. This voltage ensures that the BladeCenter maximum of 240 V is not exceeded.

Power installation considerations

Each zBX BladeCenter operates from two fully redundant PDUs that are installed in the rack with the BladeCenter. Each PDU has its own power cords as shown in Table 10-5, enabling the system to survive the loss of customer power to either power cord. If power is interrupted to one of the PDUs, the other PDU picks up the entire load, and the BladeCenter continues to operate without interruption.

Table 10-5 Number of BladeCenter power cords

Number of BladeCenters	Number of power cords
1	2
2	4
3	6
4	8
5	10
6	12
7	14
8	16

For maximum availability, the power cords on each side of the racks need to be powered from separate building PDUs.

Actual power consumption depends on the zBX configuration in terms of the number of BladeCenters and blades installed.

Input power in kilovolt-ampere (kVA) is equal to the outgoing power in kilowatt (kW). Heat output expressed in kBTU per hour is derived by multiplying the table entries by a factor of 3.4. For 3-phase installations, phase balancing is accomplished with the power cable connectors between the BladeCenters and the PDUs.

10.3.3 IBM zBX cooling

The individual BladeCenter configuration is air cooled with two hot-swap blower modules. The blower speeds vary depending on the ambient air temperature at the front of the BladeCenter unit and the temperature of the internal BladeCenter components:

- ▶ If the ambient temperature is 25°C (77°F) or below, the BladeCenter unit blowers run at their minimum rotational speed, increasing their speed as required to control internal BladeCenter temperature.
- ▶ If the ambient temperature is above 25°C (77°F), the blowers run faster, increasing their speed as required to control the internal BladeCenter unit temperature.
- ▶ If a blower fails, the remaining blower runs at full speed and continues to cool the BladeCenter unit and blade servers.

Heat released by configurations

Table 10-6 shows the typical heat that is released by the various zBX solution configurations.

Table 10-6 IBM zBX power consumption and heat output

Number of blades	Maximum utility power (kW)	Heat output (kBTU/hour)
7	7.3	24.82
14	12.1	41.14
28	21.7	73.78
42	31.3	106.42
56	40.9	139.06
70	50.5	171.70
84	60.1	204.34
98	69.7	236.98
112	79.3	269.62

Optional Rear Door Heat eXchanger (FC 0540)

For data centers that have limited cooling capacity, using the Rear Door Heat eXchanger (Figure 10-4 on page 379) is a more cost-effective solution than adding another air conditioning unit.

Tip: The Rear Door Heat eXchanger is not a requirement for BladeCenter cooling. It is a solution for customers who cannot upgrade a data center's air conditioning units due to space, budget, or other constraints.

The Rear Door Heat eXchanger has the following features:

- ▶ A water-cooled heat exchanger door is designed to dissipate heat that is generated from the back of the computer systems before it enters the room.
- ▶ An easy-to-mount rear door design attaches to customer-supplied water by using industry-standard fittings and couplings.
- ▶ Up to 50,000 BTUs (or approximately 15 kW) of heat can be removed from the air exiting the back of a zBX rack.

Note: The Rear Door Heat eXchanger feature has been withdrawn from marketing and cannot be ordered if not already present on the zBX Model 004.

Figure 10-4 shows the IBM Rear Door Heat eXchanger details.

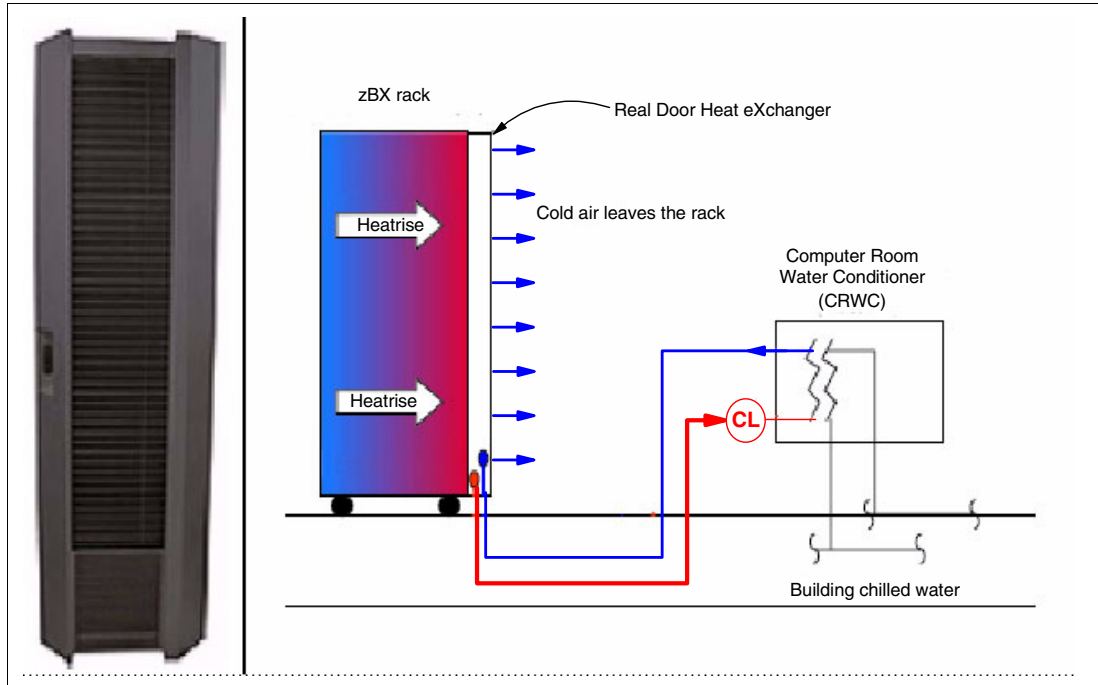


Figure 10-4 Rear Door Heat eXchanger (left) and functional diagram

The IBM Rear Door Heat eXchanger also offers a convenient way to handle hazardous “hot spots”, which might help you lower the total energy cost of your data center.

10.3.4 IBM zBX physical specifications

The zBX solution is delivered either with one rack (Rack B) or four racks (Rack B, C, D, and E). Table 10-7 shows the physical dimensions of the zBX minimum and maximum solutions.

Table 10-7 Dimensions of zBX racks

Racks with covers	Width mm (in.)	Depth mm (in.)	Height mm (in.)
B	648 (25.5)	1105 (43.5)	2020 (79.5)
B+C	1296 (51.0)	1105 (43.5)	2020 (79.5)
B+C+D	1994 (76.5)	1105 (43.5)	2020 (79.5)
B+C+D+E	2592 (102)	1105 (43.5)	2020 (79.5)

Top Exit Support FC 0545

This feature enables you to route I/O and power cabling through the top of the zBX rack. The feature as shown in Figure 10-5 on page 380 will add 177 mm (7 in.) to the height and 9.75 kg (21.5 lbs) to the weight of the zBX rack after it is installed. It can be ordered with a new zBX, but can also added later. You need one feature per installed rack.

Note: The Top Exit Support feature (FC 0545) has been withdrawn from marketing and cannot be ordered if not already present on the zBX Model 004.

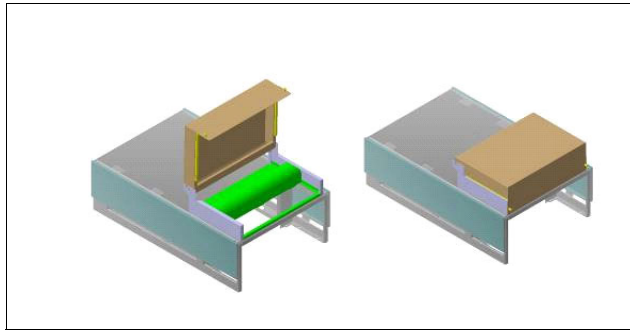


Figure 10-5 Top Exit Support for the zBX

Height Reduction FC 0570

This feature is required if it is necessary to reduce the shipping height for the zBX. Select this feature when it is deemed necessary for delivery clearance purposes. Order it if you have doorways with openings less than 1941 mm (76.4 in.) high. It accommodates doorway openings as low as 1832 mm (72.1 in.).

Note: The Height Reduction feature (FC 0570) has been withdrawn from marketing and cannot be ordered if not already present on the zBX Model 004.

IBM zBX weight

Table 10-8 shows the maximum weights of fully populated zBX racks and BladeCenters.

Table 10-8 Weights of zBX racks

Rack description	Weight kg (lbs.)
B with 28 blades	740 (1630)
B + C full	1234 (2720)
B + C + D full	1728 (3810)
B + C + D + E full	2222 (4900)

Rack weight: A fully configured Rack B is heavier than a fully configured Rack C, D, or E, because Rack B has the two TOR switches installed.

For a complete view of the physical requirements, see *zBX Model 004 Installation Manual for Physical Planning*, GC27-2630.

10.4 Energy management

This section provides information about the elements of energy management in areas of tooling to help you understand the requirement for power and cooling, monitoring and trending, and reducing power consumption.

The hardware components in the z Systems CPC and the optional zBX are monitored and managed by the Energy Management component in the SE and HMC. The GUI of the SE and the HMC provide views, for instance, with the System Activity Display or Monitors Dashboard.

Through a Simple Network Management Protocol (SNMP) application programming interface (API), energy information is available to, for example, Active Energy Manager, which is a plug-in of IBM Systems Director. See 10.4.4, “IBM Systems Director Active Energy Manager” on page 383 for more information.

When z Unified Resource Manager features are installed (see 11.6.1, “Unified Resource Manager” on page 428), several monitoring and control functions can be used to perform Energy Management. For more details, see 10.4.5, “Unified Resource Manager: Energy management” on page 384.

A few aids are available to plan and monitor the power consumption and heat dissipation of z13s servers. This section summarizes the tools that are available to plan and monitor the energy consumption of z13s servers:

- ▶ Power estimation tool
- ▶ Query maximum potential power
- ▶ System Activity Display and Monitors Dashboard
- ▶ IBM Systems Director Active Energy Manager™

10.4.1 Power estimation tool

The power estimation tool for z Systems servers is available through the IBM Resource Link website:

<http://www.ibm.com/servers/resourceLink>

The tool provides an estimate of the anticipated power consumption of a machine model, given its configuration. You enter the machine model, memory size, number of I/O drawers, and quantity of each type of I/O feature card. The tool outputs an estimate of the power requirements for that configuration.

If you have a registered machine in Resource Link, you can access the power estimation tool on the machine information page of that particular machine. In the Tools section of Resource Link, you also can enter the power estimator and enter any system configuration for which you want to calculate power requirements. This tool helps with power and cooling planning for installed and planned z Systems servers.

10.4.2 Query maximum potential power

The maximum potential power that is used by the system is less than the *label power*, as depicted in a typical power usage report that is shown in Figure 10-6 on page 382. The Query maximum potential power function shows what *your* systems maximum power usage and heat dissipation can be so that you are able to allocate the correct power and cooling resources.

The output values of this function for *maximum potential power* and *maximum potential heat load* are displayed on the Energy Management tab of the central processor complex (CPC) Details view of the HMC.

This function enables operations personnel with no z Systems knowledge to query the maximum power draw of the system, as shown in Figure 10-6. The implementation helps to avoid capping enforcement through dynamic capacity reduction. The customer controls are implemented in the HMC, the SE, and the Active Energy Manager.

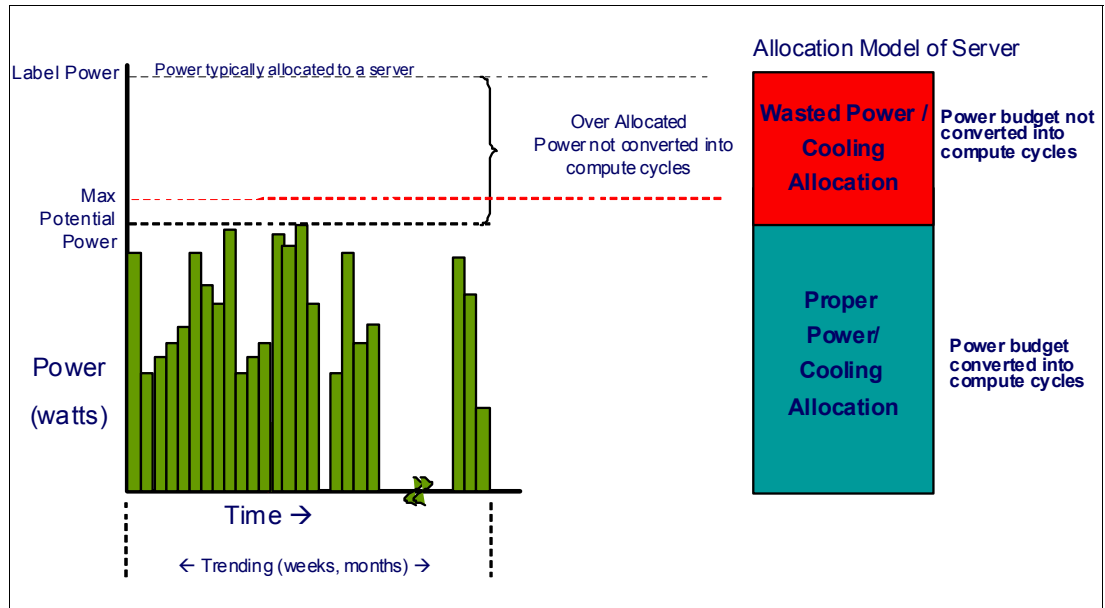


Figure 10-6 Maximum potential power

Use this function with the power estimation tool to plan for power and cooling requirements. See 10.4.1, “Power estimation tool” on page 381.

10.4.3 System Activity Display and Monitors Dashboard

The System Activity Display presents you with the current power usage, among other information, as shown in Figure 10-7.

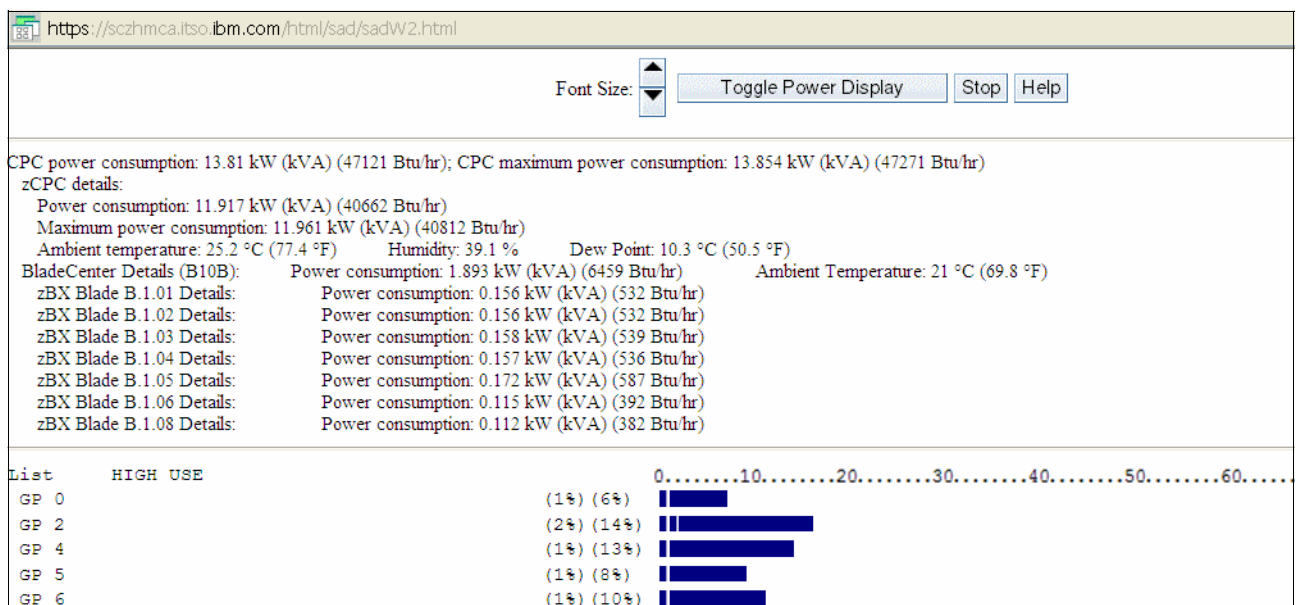


Figure 10-7 Power usage on the System Activity Display

The Monitors Dashboard of the HMC enables you to display power and other environmental data. It also enables you to start a Dashboard Histogram Display, where you can trend a particular value of interest, such as the power consumption of a blade or the ambient temperature of a CPC.

10.4.4 IBM Systems Director Active Energy Manager

IBM Systems Director Active Energy Manager is an energy management solution building block that returns true control of energy costs to the customer. Active Energy Manager is an industry-leading cornerstone of the IBM energy management framework.

Active Energy Manager Version 4.4 is a plug-in to IBM Systems Director Version 6.2.1 and is available for installation on Linux on z Systems. It can also run on Windows, Linux on IBM System x, and AIX and Linux on IBM Power Systems™. For more information, see *Implementing IBM Systems Director Active Energy Manager 4.1.1*, SG24-7780. Version 4.4 supports IBM z13s servers and their optional attached zBX.

Use Active Energy Manager to monitor the power and environmental values of resources, for z Systems servers and other IBM products, such as IBM Power Systems, IBM System x, or devices and hardware that are acquired from another vendor. You can view historical trend data for resources, calculate energy costs and savings, view properties and settings for resources, and view active energy-related events.

Active Energy Manager does not directly connect to the z Systems servers, but it attaches through a LAN connection to the HMC. See 11.3, “HMC and SE connectivity” on page 398. Active Energy Manager discovers the HMC managing the server by using a discovery profile that specifies the HMC’s IP address and the SNMP credentials for that z Systems HMC. As the system is discovered, the z Systems servers that are managed by the HMC are also discovered.

Active Energy Manager is a management software tool that can provide a single view of the actual power usage across multiple platforms, as opposed to the benchmarked or rated power consumption. It can effectively monitor and control power in the data center at the system, chassis, or rack level. By enabling these power management technologies, data center managers can more effectively manage the power of their systems while lowering the cost of computing.

The following data is available through Active Energy Manager:

- ▶ System name, machine type, model, serial number, and firmware level of the z Systems servers and optional zBX.
- ▶ Ambient temperature.
- ▶ Exhaust temperature.
- ▶ Average power usage.
- ▶ Peak power usage.
- ▶ Limited status and configuration information. This information helps to explain the changes to the power consumption, which are called *events*:
 - Changes in fan speed
 - Changes between power-off, power-on, and IML-complete states
 - CBU records expirations

IBM Systems Director Active Energy Manager provides customers with the intelligence necessary to effectively manage power consumption in the data center. Active Energy Manager, which is an extension to IBM Director Systems Management software, enables you to *meter* actual power usage and trend data for any single physical system or group of systems. Active Energy Manager uses monitoring circuitry, which was developed by IBM, to help identify how much actual power is being used, and the temperature of the system.

10.4.5 Unified Resource Manager: Energy management

The energy management capabilities for z Unified Resource Manager can be used in an ensemble depending on which suite is installed in the ensemble:

- ▶ Manage Suite (feature code 0019)
- ▶ Automate Suite (feature code 0020)

Manage Suite

For energy management, the Manage Suite focuses on the monitoring capabilities. Energy monitoring can help you better understand the power and cooling demand of the z Systems servers. It provides complete monitoring and trending capabilities for z13s servers and the zBX by using one or more of the following options:

- ▶ Monitors dashboard
- ▶ Environmental Efficiency Statistics
- ▶ Details view

Automate Suite

The z Unified Resource Manager offers multiple energy management tasks as part of the automate suite. These tasks enable you to change system behaviors for optimized energy usage and energy saving:

- ▶ Power cap
- ▶ Group power cap

Various options are presented, depending on the scope that is selected in the z Unified Resource Manager graphical user interface (GUI).

Set Power Cap

The power cap function can be used to limit the maximum amount of energy that is used by the ensemble. If enabled, it enforces power caps for the hardware by throttling the processors in the system.

The z Unified Resource Manager shows all of the components of an ensemble in the Set Power Cap window. Because not all components that are used in a specific environment can support power capping, only those components that are marked as **Enabled** can perform power capping functions.

z13s servers support power capping, which gives you the ability to limit the maximum power consumption and reduce cooling requirements. To use power capping, Automate Firmware Suite (FC 0020) must be ordered. This feature is used to enable the Automate suite of functions associated with the IBM z Unified Resource Manager. The Automate suite includes representation of resources in a workload context, goal-oriented monitoring and management of resources, and energy management. The Automate suite is included in the base z Systems CPC at no charge for CPs and zIIPs.

When capping is enabled for a z13s server, this capping level is used as a threshold for a warning message that informs you that the z13s server went above the set cap level. Being under the limit of the cap level is equal to the maximum potential power value (see 10.4.2, “Query maximum potential power” on page 381).

More information about energy management with zManager is available in *IBM zEnterprise Unified Resource Manager*, SG24-7921.



Hardware Management Console and Support Elements

The Hardware Management Console (HMC) supports many functions and tasks to extend the management capabilities of z13s servers. When tasks are performed on the HMC, the commands are sent to one or more Support Elements (SEs), which then issue commands to their central processor complexes (CPCs) or IBM z BladeCenter Extension (zBX).

This chapter addresses the HMC and SE in general, and adds appropriate information for HMCs that manage ensembles with the IBM z Unified Resource Manager.

This chapter includes the following sections:

- ▶ Introduction to the HMC and SE
- ▶ HMC and SE enhancements and changes
- ▶ HMC and SE connectivity
- ▶ Remote Support Facility
- ▶ HMC and SE key capabilities
- ▶ HMC in an ensemble

11.1 Introduction to the HMC and SE

The HMC is a stand-alone computer that runs a set of management applications. The HMC is a closed system, which means that no other applications can be installed on it.

The HMC is used to set up, manage, monitor, and operate one or more z Systems CPCs. It manages z Systems hardware, its logical partitions (LPARs), and provides support applications. At least one HMC is required to operate an IBM z Systems server. An HMC can manage multiple z Systems CPCs and can be at a local or a remote site.

If the z13s server is defined as a member of an ensemble, a pair of HMCs is required (a primary and an alternate). When a z13s server is defined as a member of an ensemble, certain restrictions apply. For more information, see 11.6, “HMC in an ensemble” on page 428.

The SEs are two integrated servers in the CPC frame that are supplied together with the z13s server. One is the primary SE and the other is the alternate SE. The primary SE is the active one. The alternate SE acts as the backup. Like the HMCs, the SEs are closed systems, and no other applications can be installed on them.

When tasks are performed at the HMC, the commands are routed to the active SE of the z Systems CPC. The SE then issues those commands to their CPC and or zBX (if any). One HMC can control up to 100 SEs and one SE can be controlled by up to 32 HMCs.

Some functions are available only on the SE. With Single Object Operation (SOO), these functions can be used from the HMC. For more information, see “Single object operating” on page 405.

With Driver 27 (Version 2.13.1), the IBM Dynamic Partition Manager (DPM) is introduced for Linux only CPCs with Fibre Channel Protocol (FCP)-attached storage. DPM is a new mode of operation that enables customers with little or no knowledge of z Systems servers to set up the system efficiently and with ease. For more information, see Appendix E, “IBM Dynamic Partition Manager” on page 501

The HMC Remote Support Facility (RSF) provides the important communication to a centralized IBM support network for hardware problem reporting and service. For more information, see 11.4, “Remote Support Facility” on page 403.

11.2 HMC and SE enhancements and changes

z13s servers come with the new HMC application Version 2.13.1. Use the What’s New task to explore the new features that are available for each release. For a complete list of HMC and SE functions, use the HMC and SE (Version 2.13.1) console help system or go to the IBM Knowledge Center at the following website:

<http://www.ibm.com/support/knowledgecenter>

After you get to the IBM Knowledge Center, click **z Systems**, and then click **z13s**.

11.2.1 Driver Level 27 HMC and SE enhancements and changes

The HMC and SE with Driver 27 has several enhancements and changes for z13s servers:

- ▶ Rack-mounted HMC

A new rack-mounted HMC (FC 0094) is available. For more information, see 11.2.2, “Rack-mounted HMC” on page 392.

- ▶ New SE server

The SEs are no longer two notebooks in one z13s server. They are now two 1U servers that are installed in the top of the CPC frame. For more information, see 11.2.3, “New Support Elements” on page 393.

- ▶ New backup options

The backup options of the HMC and SE have changed. You can now have an FTP destination for the backup. For more information, see 11.2.4, “New backup options for HMCs and primary SEs” on page 393.

- ▶ New UFD

A new optional 32 GB USB flash memory drive (UFD) is available for previous servers to create a backup. For more information, see 11.2.4, “New backup options for HMCs and primary SEs” on page 393.

- ▶ Monitor activity enhancements

You can now display the activity for an LPAR by processor type, and the Monitors Dashboard is enhanced to show simultaneous multithreading (SMT) usage. For more information, see “The Monitors Dashboard task” on page 412.

- ▶ HMC Data Replication

If you are using HMC Data Replication and you have an HMC at Driver 27, this HMC can only do Data Replication with another HMC at Driver 27 or Driver 22. Data Replication with driver levels older than Driver 22 are not supported on an HMC at Driver 27. If IBM z Unified Resource Manager is used, data replication is not available, Mirroring between the primary and alternate Ensemble HMCs is used instead.

- ▶ STP enhancements

Enhancements and changes have been added for the Server Time Protocol (STP), which are described in 11.5.8, “Server Time Protocol support” on page 416.

- ▶ Help infrastructure updates

The content from the following publications is incorporated into the HMC and SE help system:

- *z Systems Hardware Management Console Operations Guide Version 2.13.1*
- *z Systems Hardware Management Console Operations Guide for Ensembles Version 2.13.1*
- *z Systems Support Element Operations Guide Version 2.13.1*

Alternatively, see the HMC and SE (Version 2.13.1) console help system or go to the IBM Knowledge Center at the following website:

<http://www.ibm.com/support/knowledgecenter>

After you get to the IBM Knowledge Center, click **z Systems**, and then click **z13s**.

- ▶ Disable boot from removable media

By default, starting at level 2.13.0, booting from removable media has been disabled. An admin password for the UEFI (the successor of BIOS) can be set. This action protects the

SE and HMC against unauthorized booting from removable media. An IBM Service Support Representative (IBM SSR) might perform the following tasks, which require booting from removable media:

- Engineering change (EC) upgrade
- Save or restore of Save/Restore data
- Hard disk drive (HDD) restore

Note: When a UEFI admin password is set, it must be available for the SSR to perform these tasks. A lost password requires a system board replacement.

- ▶ Support multi-partitions for LPAR weights. Ability to change LPAR weight related information at the same time
- ▶ New option for CPUMF Diagnostic Sampling Collection. Can only be selected if Basic sampling is selected as shown in Figure 11-1.

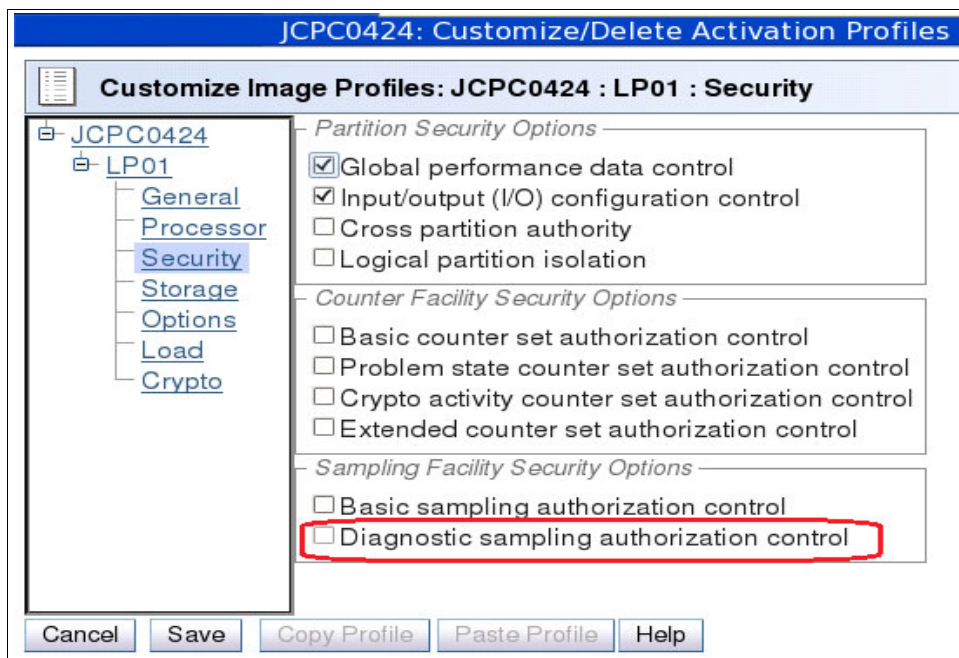


Figure 11-1 Diagnostic sampling authorization control

Enabling **Diagnostic Sampling** for a specific LPAR is shown in Figure 11-2.

Select	Logical Partition	Active	Performance Data Control	I/O Config Control	Cross Partition Authority	Partition Isolation	Basic Counter	Problem State Counter	Crypto Activity Counter	Extended Counter	Basic Sampling	Diagnostic Sampling
<input type="checkbox"/>	CF01	No	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	CF02	No	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	LP01	Yes	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	LP02	Yes	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	LP04	No	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	LP05	No	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	LP06	No	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	LP07	No	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	LP08	No	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	LP09	No	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	LP11	No	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	LP14	No	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	LP15	No	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	ZAWARE	No	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	ZAWARE2	No	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Figure 11-2 Change LPAR security

► zBX firmware management.

As a result of a MES upgrade from a zBX Model 002 or zBX Model 003, the zBX Model 004 becomes a stand-alone machine. It must be added to an existing ensemble HMC as an ensemble node member. As part of the upgrade, the zBX Model 004 is detached from the owning CPC and receives two redundant internal SEs that are installed on its B Frame. The zBX Model 004 object can then be added to the ensemble HMC by using the Add Object Definition task. The ensemble HMC and the zBX Model 004 SE along with the zEnterprise Unified Resource Manager can perform the following monitoring and management functions:

- zBX firmware upgrades are downloaded from IBM RETAIN by using the HMC broadband RSF connection. Firmware updates are saved locally and installed during a scheduled MCL apply session.
- Firmware updates are installed from the HMC and zBX SEs by using the same process and controls as used for z Systems servers.
- zBX hardware-related and firmware-related failures are reported to IBM, and the IBM support structure is engaged, by using the HMC RSF. This is the same process that is used for reporting z Systems problems.

► zBX lifecycle management: zBX Model 004 supports the same System x, POWER7, and DataPower XI50z blade types that are supported in zBX Model 003 and zBX Model 002.

► Help infrastructure updates

The content from the following publications is incorporated into the HMC and SE help system:

- *z Systems Hardware Management Console Operations Guide Version 2.13.1*
- *z Systems Hardware Management Console Operations Guide for Ensembles Version 2.13.1*
- *z Systems Support Element Operations Guide Version 2.13.0 (zBX 004)*

Alternatively, see the HMC and SE (Version 2.13.1) console help system or go to the IBM Knowledge Center at the following website:

<http://www.ibm.com/support/knowledgecenter>

After you get to the IBM Knowledge Center, click **z Systems**, and then click **z13s**.

11.2.2 Rack-mounted HMC

Feature codes (FC) 0094 and FC 0096 provide a rack-mounted HMC. They cannot be ordered as a feature code for zEC12, zBC12, or previous z Systems servers.

The HMC is a 1U IBM server and comes with an IBM 1U standard tray containing a monitor and a keyboard. The system unit and tray must be mounted in the rack in two adjacent 1U locations in the “ergonomic zone” between 21U and 26U in a standard 19” rack.

The customer must provide the rack. Three C13 power receptacles are required: Two for the system unit and one for the display and keyboard, as shown in Figure 11-3.

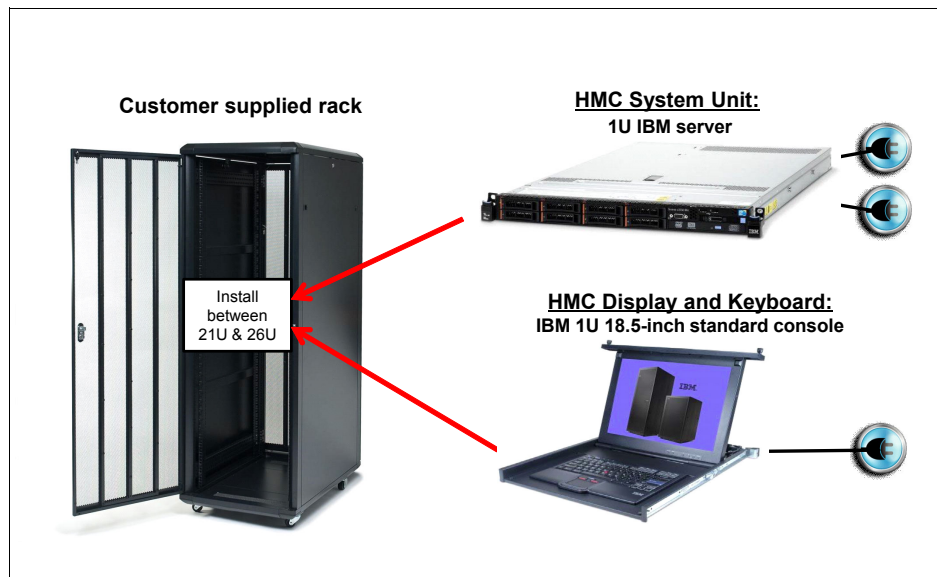


Figure 11-3 Rack-mounted HMC

11.2.3 New Support Elements

The SEs are no longer two notebooks in one z13s server. They are now two servers that are installed at the top of the CPC frame. They are managed by the keyboard, pointing device, and display that are mounted in the front of the tray of the CPC frame (where the SE notebooks were in previous z Systems servers), as shown in Figure 11-4. The SEs have an internal USB attached smart card reader to support Flash Express and Features on Demand (FoD).

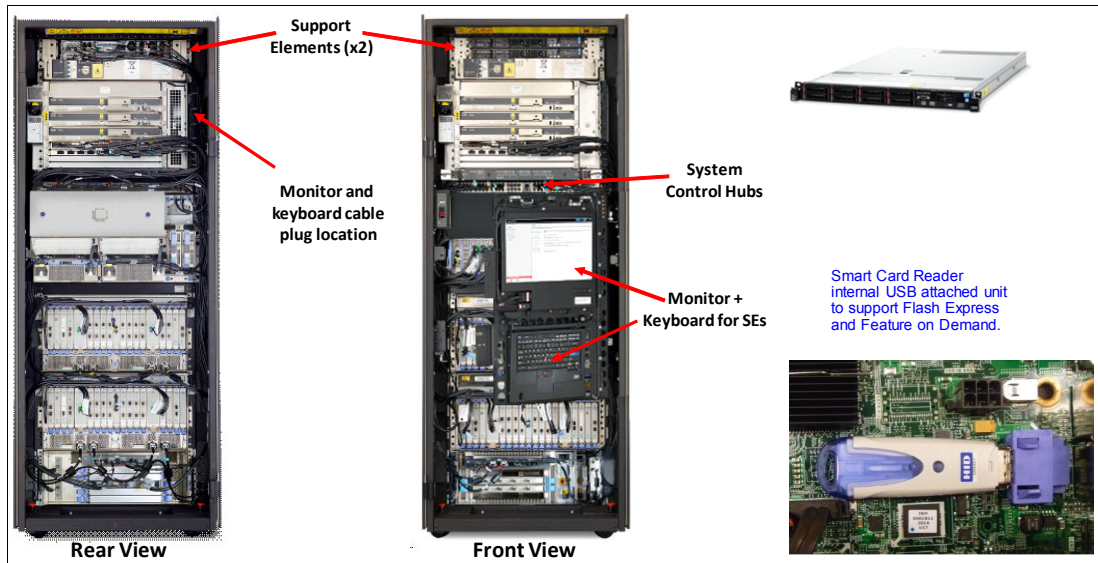


Figure 11-4 SEs location

11.2.4 New backup options for HMCs and primary SEs

This section provides a short description of the new backup options that are available for HMC Version 2.13.1.

Backup of primary SEs or HMCs to an FTP server

With Driver 22 or later, you can perform a backup of primary SEs or HMCs to an FTP server.

Note: If you do a backup to an FTP server for a z13s or zBX Model 004 server, ensure that you have set up a connection to the FTP server by using the Configure Backup Setting task. If you have not set up a connection to the FTP server, a message appears that prompts you to configure your FTP server.

The FTP server must be supplied by the customer. You can enable a secure FTP connection to your server.

Figure 11-5 shows the information that is required to configure your backup FTP server.

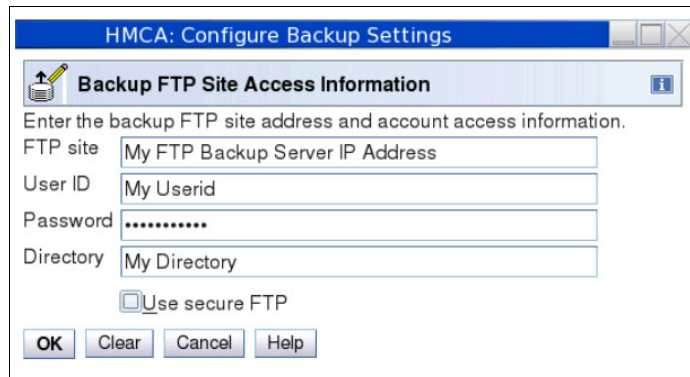


Figure 11-5 Configure backup FTP server

Note: Backup FTP site is a static setting for an HMC. If an alternate FTP site is needed to perform a backup, then this should be done from another HMC.

Backup of HMCs

A backup of the HMC can be performed to the following media

- ▶ A UFD
- ▶ An FTP server
- ▶ A UFD and FTP server

Figure 11-6 shows the destination options of the Backup Critical Console Data task.

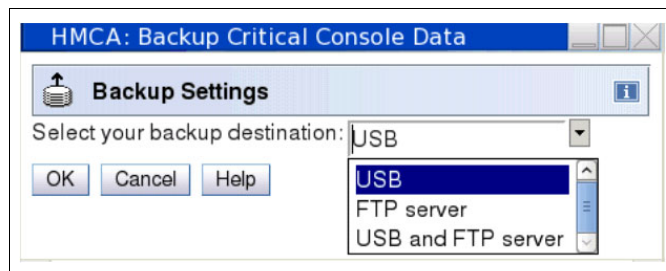


Figure 11-6 Backup Critical Console Data destinations

Optional 32 GB UFD FC 0848

A new, optional 32 GB UFD is available for backups. The default that is shipped with the system is an 8 GB UFD. SE and HMC backup files have been getting larger with newer z Systems. The following features require more storage space for a backup:

- IBM zEnterprise Unified Resource Manager
- zBX
- IBM zAware
- Several previous z Systems SE backups

Backup of primary SEs

The backup for the primary SE of a z13s or zBX Model 004 server can be made to the following media:

- ▶ The primary SE HDD and alternate SE HDD
- ▶ The primary SE HDD and alternate SE HDD and FTP server

It is no longer possible to do the primary SE backup to a UFD of a z13s or zBX Model 004 SE. Table 11-1 shows the SE Backup options for external media.

Table 11-1 SE Backup options

System Type	UFD Media	FTP Server
z13 / z13s	No	Yes
zBX 004	No	Yes
zEC12 / zBC12	Yes	No
z196 / z114	Yes	No
z10EC / z10BC	Yes	No
z9EC / /z9BC	Yes	No

Figure 11-7 shows examples of the different destination options of the SE Backup Critical Data for different CPC machine types.

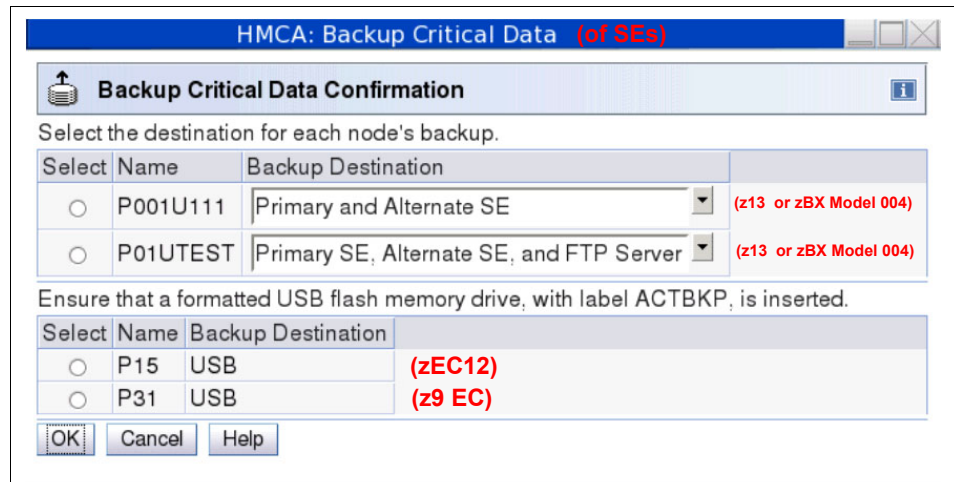


Figure 11-7 Backup Critical Data destinations of SEs

For more information, see the HMC and SE (Version 2.13.1) console help system or go to the IBM Knowledge Center at the following link:

<http://www.ibm.com/support/knowledgecenter>

After you get to the IBM Knowledge Center, click **z Systems**, and then click **z13s**.

Scheduled operations for the backup of HMCs and SEs

The Scheduled Operation task with the new backup options for HMC is changed, as shown in Figure 11-8.

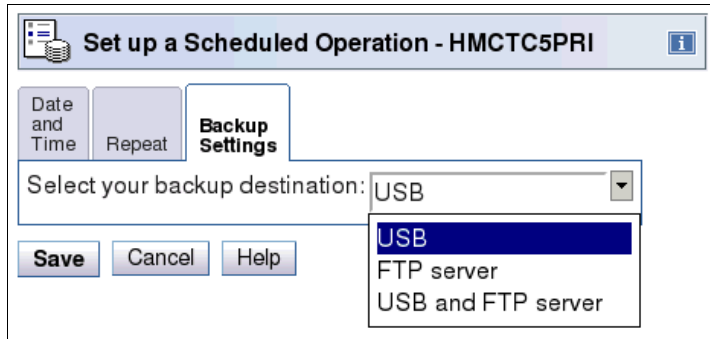


Figure 11-8 Scheduled Operation for HMC backup

The Scheduled Operation task with the new backup options for the SEs is changed, as shown in Figure 11-9.

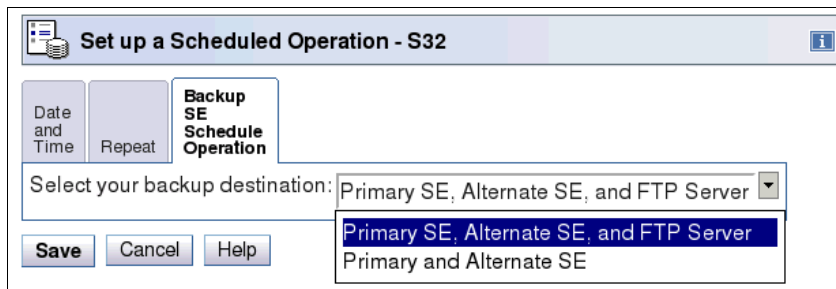


Figure 11-9 Scheduled Operation for SEs backup

11.2.5 SE driver support with the HMC driver

The driver of the HMC and SE is equivalent to a specific HMC and SE version, as shown in these examples:

- ▶ Driver 27 is equivalent to Version 2.13.1
- ▶ Driver 22 is equivalent to Version 2.13.0
- ▶ Driver 86 is equivalent to Version 2.11.0
- ▶ Driver 79 is equivalent to Version 2.10.2

An HMC with Version 2.13.1 can support different z Systems types. Some functions that are available on Version 2.13.1 and later are supported only when the HMC is connected to a z13s server with Driver 27.

Table 11-2 shows a summary of the SE drivers and versions that are supported by the new HMC Version 2.13.1 (Driver 27).

Table 11-2 z13s HMC at Driver 27: z Systems support summary

z Systems family name	Machine type	SE driver	SE version	Ensemble node potential
z13s	2965	27	2.13.1	Yes ^a
z13	2964	22,27	2.13.0, 2.13.1	Yes ^a
zBX Node	2458 Model 004	22	2.13.0	Required
zBC12	2828	15	2.12.1	Yes
zEC12	2827	15	2.12.1	Yes
z114	2818	93	2.11.1	Yes
z196	2817	93	2.11.1	Yes
z10 BC	2098	79	2.10.2	No
z10 EC	2097	79	2.10.2	No
z9 BC	2096	67	2.9.2	No
z9 EC	2094	67	2.9.2	No

a. A CPC in DPM mode cannot be a member of a ensemble, but that CPC can be managed by the ensemble HMC.

Note: z900/z800 (Driver 3G, SE Version 1.7.3) and z990/z890 (Driver 55, SE Version 1.8.2) systems are no longer supported. If you have these older systems, consider managing these systems by using separate HMCs running older drivers.

11.2.6 HMC feature codes

HMCs older than FC 0091 are not supported for z13s servers at Driver 27.

FC 0091 M/T 7327

FC 0091 can be carried forward. These HMCs need an upgrade to driver 27. ECA398 is required and can be ordered by the local IBM support representative. The upgrade process includes upgrading memory if necessary and installing the necessary firmware

FC 0092 M/T 7382

FC 0092 is an HMC that contains 16 GB of memory. It is a tower model. The physical dimensions from FC 0092 compared to FC 0090 and FC 0091 are similar, except the depth for FC 0092 is in round numbers, that is, 95 mm (3.75 in.) more.

FC 0094 M/T 7914

FC 0094 is a rack-mounted HMC. It contains 16 GB of memory.

FC 0096 M/T 2461

FC 0096 is the new rack-mounted HMC. It contains also 16 GB of memory.

For more information, see 11.2.2, “Rack-mounted HMC” on page 392.

11.2.7 Tree Style User Interface and Classic Style User Interface

Two user interface styles are provided with an HMC. The *Tree Style User Interface* (default) uses a hierarchical model that is common in newer operating systems, and features context-based task launching. The *Classic Style User Interface* uses the drag-and-drop interface style.

Statements of Direction:

Removal of support for Classic Style User Interface on the Hardware Management Console and Support Element: The IBM z13 and z13s servers will be the last z Systems servers to support Classic Style User Interface. In the future, user interface enhancements will be focused on the Tree Style User Interface.

Removal of support for the Hardware Management Console Common Infrastructure Model (CIM) Management Interface: IBM z13 and z13s servers will be the last z Systems servers to support the Hardware Console Common Information Model (CIM) Management Interface. The Hardware Management Console Simple Network Management Protocol (SNMP), and Web Services application programming interfaces (APIs) will continue to be supported.

All statements regarding IBM plans, directions, and intent are subject to change or withdrawal without notice. Any reliance on these statements of general direction is at the relying party's sole risk and will not create liability or obligation for IBM.

The Classic Style User Interface is not available on zBX Model 004 SEs.

If at least one CPC defined on the HMC is running in DPM mode, the welcome window on the HMC changes. Setting up a machine in DPM mode is wizard-based. For more information, see 11.5.15, “Dynamic Partition Manager” and Appendix E, “IBM Dynamic Partition Manager” on page 501.

Tutorials: IBM Resource Link^a provides tutorials that demonstrate how to change from the Classic Style User Interface to the Tree Style Interface, and introduce the function of the Tree Style Interface on the HMC. You can find Resource Link at the following website:

<http://www.ibm.com/servers/resourceLink>

After you go to the Resource Link website, click **Education** → **IBM z13** → **Course Z121-0255-00**.

a. Registration is required to access IBM Resource Link.

11.3 HMC and SE connectivity

The HMC has two Ethernet adapters, which are supported by HMC Driver 27 for connectivity to up to two different Ethernet LANs.

The SEs on z13s servers are connected to the System Control Hubs (SCHs) to control the internal network. In previous z Systems servers, the customer network was connected to the BPH. Now the SEs are directly connected to the customer network. The HMC to CPC communication is now only possible through an Ethernet switch connected to the J03 or

J04 port on the SEs. Other z Systems servers and HMCs also can be connected to the switch. To provide redundancy, install two Ethernet switches.

Only the switch (and not the HMC directly) can be connected to the SEs.

Figure 11-10 shows the connectivity between HMCs and the SEs.

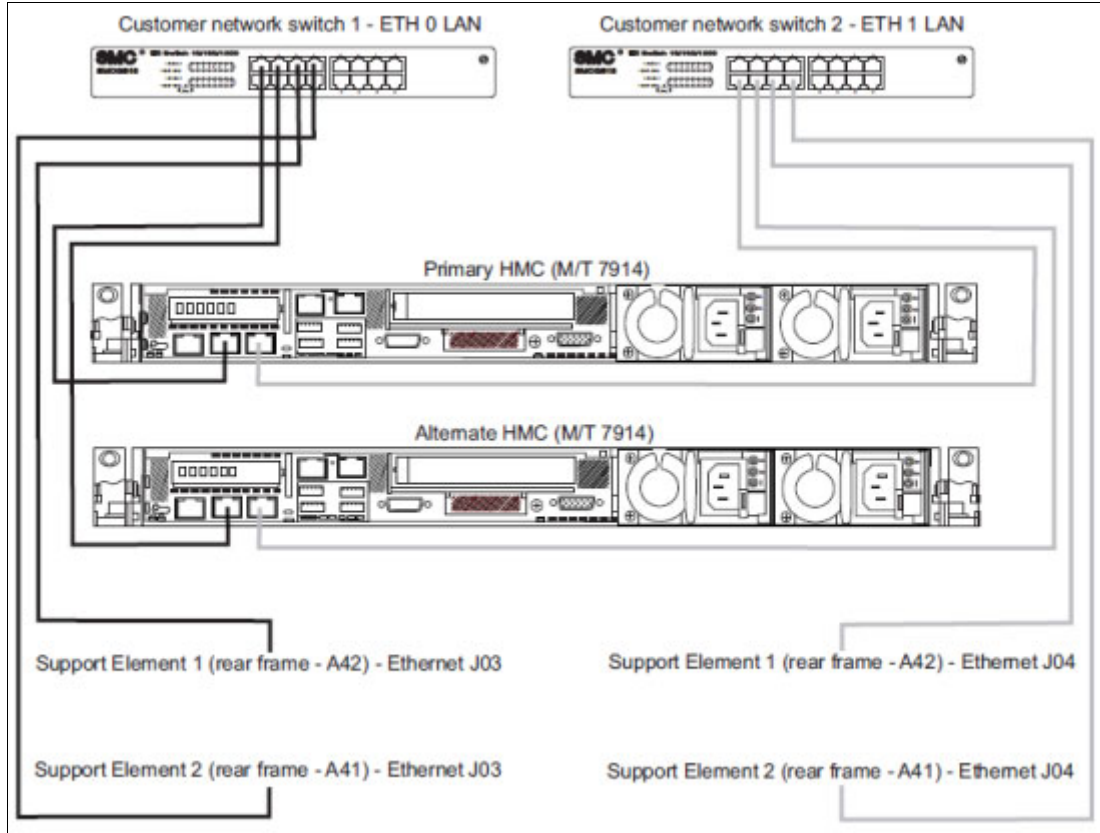


Figure 11-10 HMC and SE connectivity

The LAN ports for the SEs installed in the CPC are shown in Figure 11-11.



Figure 11-11 SE Physical connection

Various methods are available for setting up the network. It is your responsibility to plan and design the HMC and SE connectivity. Select the method based on your connectivity and security requirements.

Security: Configuration of network components, such as routers or firewall rules, is beyond the scope of this book. Whenever the networks are interconnected, security exposures can exist. For more information about HMC security, see *z Systems Integrating the Hardware Management Console's Broadband Remote Support Facility into your Enterprise*, SC28-6951. It is available on IBM Resource Link.^a

For more information about the HMC settings that are related to access and security, see the HMC and SE (Version 2.13.1) console help system or go to the IBM Knowledge Center at the following link:

<http://www.ibm.com/support/knowledgecenter>

After you get to the IBM Knowledge Center, click **z Systems**, and then click **z13s**.

a. Registration is required to access IBM Resource Link.

11.3.1 Network planning for the HMC and SE

Plan the HMC and SE network connectivity carefully to allow for current and future use. Many of the z Systems capabilities benefit from the network connectivity options that are available.

For example, these functions, which depend on the HMC connectivity, are available to the HMC:

- ▶ Lightweight Directory Access Protocol (LDAP) support, which can be used for HMC user authentication
- ▶ Network Time Protocol (NTP) client/server support
- ▶ RSF through broadband
- ▶ HMC access through a remote web browser
- ▶ Enablement of the SNMP and CIM APIs to support automation or management applications, such as IBM System Director Active Energy Manager (AEM)

These examples are shown in Figure 11-12.

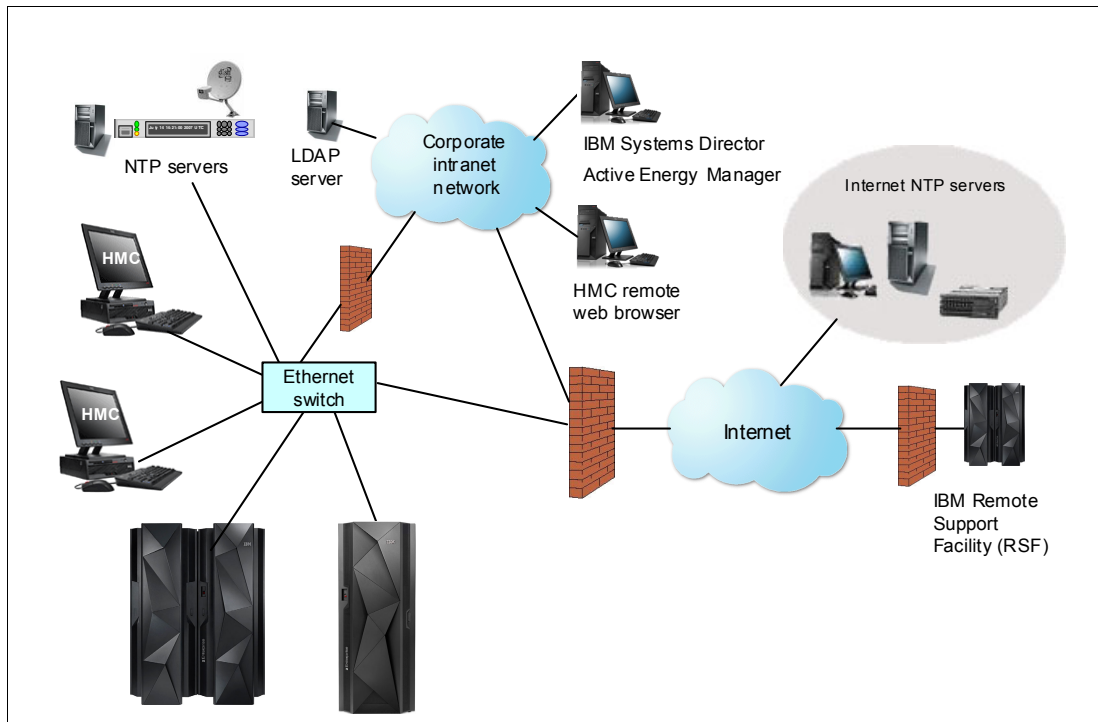


Figure 11-12 HMC connectivity examples

For more information, see the following resources:

- ▶ The HMC and SE (Version 2.13.1) console help system, or go to the IBM Knowledge Center at the following link:

<http://www.ibm.com/support/knowledgecenter>

After you get to the IBM Knowledge Center, click **z Systems**, and then click **z13**.

- ▶ Section 10.4.4, “IBM Systems Director Active Energy Manager” on page 383.
- ▶ *z13s Installation Manual for Physical Planning*, GC28-6953.

11.3.2 Hardware prerequisite changes

The following HMC changes are important for z13s servers:

- ▶ IBM does not provide Ethernet switches with the system.
- ▶ RSF is broadband-only.

Ethernet switches

Ethernet switches for HMC and SE connectivity are customer provided. Existing supported switches can still be used.

Ethernet switches/hubs typically have these characteristics:

- ▶ Sixteen auto-negotiation ports
- ▶ 100/1000 Mbps data rate
- ▶ Full or half duplex operation
- ▶ Auto-medium dependent interface crossover (MDI-X) on all ports
- ▶ Port status LEDs

Note: Generally, use 1000 Mbps/Full duplex.

11.3.3 RSF is broadband-only

RSF through a modem *is not supported* on the z13s HMC. Broadband is needed for hardware problem reporting and service. For more information, see 11.4, “Remote Support Facility” on page 403.

11.3.4 TCP/IP Version 6 on the HMC and SE

The HMC and SE can communicate by using IPv4, IPv6, or both. Assigning a static IP address to a SE is unnecessary if the SE communicates only with HMCs on the same subnet. The HMC and SE can use IPv6 link-local addresses to communicate with each other.

IPv6 link-local addresses have the following characteristics:

- ▶ Every IPv6 network interface is assigned a link-local IP address.
- ▶ A link-local address is used only on a single link (subnet) and is never routed.
- ▶ Two IPv6-capable hosts on a subnet can communicate by using link-local addresses, without having any other IP addresses assigned.

11.3.5 Assigning addresses to the HMC and SE

An HMC can have the following IP configurations:

- ▶ Statically assigned IPv4 or statically assigned IPv6 addresses
- ▶ Dynamic Host Configuration Protocol (DHCP)-assigned IPv4 or DHCP-assigned IPv6 addresses
- ▶ Auto-configured IPv6:
 - Link-local is assigned to every network interface.
 - Router-advertised, which is broadcast from the router, can be combined with a Media Access Control (MAC) address to create a unique address.
 - Privacy extensions can be enabled for these addresses as a way to avoid using the MAC address as part of the address to ensure uniqueness.

An SE can have the following IP addresses:

- ▶ Statically assigned IPv4 or statically assigned IPv6
- ▶ Auto-configured IPv6 as link-local or router-advertised

IP addresses on the SE cannot be dynamically assigned through DHCP to ensure repeatable address assignments. Privacy extensions are not used.

The HMC uses IPv4 and IPv6 multicasting¹ to automatically discover SEs. The HMC Network Diagnostic Information task can be used to identify the IP addresses (IPv4 and IPv6) that are being used by the HMC to communicate to the CPC SEs.

¹ For a customer-supplied switch, multicast must be enabled at the switch level.

IPv6 addresses are easily identified. A fully qualified IPV6 address has 16 bytes. It is written as eight 16-bit hexadecimal blocks that are separated by colons, as shown in the following example:

```
2001:0db8:0000:0000:0202:b3ff:fe1e:8329
```

Because many IPv6 addresses are not fully qualified, shorthand notation can be used. In shorthand notation, the leading zeros can be omitted, and a series of consecutive zeros can be replaced with a double colon. The address in the previous example also can be written in the following manner:

```
2001:db8::202:b3ff:fe1e:8329
```

For remote operations that use a web browser, if an IPv6 address is assigned to the HMC, navigate to it by specifying that address. The address must be surrounded with square brackets in the browser's address field:

```
https://[fdab:1b89:fc07:1:201:6cff:fe72:ba7c]
```

Using link-local addresses must be supported by your browser.

11.4 Remote Support Facility

The HMC RSF provides important communication to a centralized IBM support network for hardware problem reporting and service. The following types of communication are provided:

- ▶ Problem reporting and repair data
- ▶ Microcode Change Level (MCL) delivery
- ▶ Hardware inventory data, which is also known as vital product data (VPD)
- ▶ On-demand enablement

Consideration: RSF through a modem *is not supported* on the z13s HMC. Broadband connectivity is needed for hardware problem reporting and service. Modems on installed HMC FC 0091 hardware do not work with HMC Version 2.13.1, which is required to support the z13s server.

11.4.1 Security characteristics

The following security characteristics are in effect:

- ▶ RSF requests always are initiated from the HMC to IBM. An inbound connection is never initiated from the IBM Service Support System.
- ▶ All data that is transferred between the HMC and the IBM Service Support System is encrypted with high-grade Secure Sockets Layer (SSL)/Transport Layer Security (TLS) encryption.
- ▶ When starting the SSL/TLS-encrypted connection, the HMC validates the trusted host with the digital signature that is issued for the IBM Service Support System.
- ▶ Data that is sent to the IBM Service Support System consists of hardware problems and configuration data.

For more information about the benefits of Broadband RSF and the SSL/TLS-secured protocol, and a sample configuration for the Broadband RSF connection, see *Integrating the HMC Broadband Remote Support Facility into Your Enterprise*, SC28-6951.

11.4.2 RSF connections to IBM and Enhanced IBM Service Support System

If the HMC and SE are at Driver 27, the driver uses a new remote infrastructure at IBM when the HMC connects through RSF for certain tasks. Check your network infrastructure settings to ensure that this new infrastructure will work.

At the time of writing, RSF still uses the “traditional” RETAIN connection. You must add access to the new Enhanced IBM Service Support System to your current RSF infrastructure (proxy, firewall, and so on).

To have the best availability and redundancy and to be prepared for the future, the HMC must have access to the Internet to IBM through RSF in the following manner. Transmission to the enhanced IBM Support System requires a domain name server (DNS). The DNS must be configured on the HMC if you are not using a proxy for RSF. If you are using a proxy for RSF, the proxy must provide the DNS.

The following host names and IP addresses are used and your network infrastructure must allow the HMC to have access to them:

- ▶ Host names:
 - www-945.ibm.com on port 443
 - esupport.ibm.com on port 443
- ▶ IP addresses. IPv4, IPv6, or both can be used:
 - IPv4:
 - 129.42.26.224:443
 - 129.42.34.224:443
 - 129.42.42.224:443
 - 129.42.50.224:443
 - 129.42.56.189:443 (enhanced)
 - 129.42.58.189:443 (enhanced)
 - 129.42.60.189:443 (enhanced)
 - 129.42.54.189:443 (enhanced)
 - IPv6:
 - 2620:0:6C0:1::1000:443
 - 2630:0:6C1:1::1000:443
 - 2630:0:6C2:1::1000:443
 - 2620:0:6C4:1::1000:443
 - 2620:0:6C4:200:129:42:54:189:443 (enhanced)
 - 2620:0:6C0:200:129:42:56:189:443 (enhanced)
 - 2630:0:6C1:200:129:42:58:189:443 (enhanced)
 - 2630:0:6C2:200:129:42:60:189:443 (enhanced)

Note: All other previous existing IP addresses are no longer supported.

11.4.3 HMC and SE remote operations

These methods are available to perform remote manual operations on the HMC:

- ▶ Using a remote HMC

A remote HMC is a physical HMC that is on a different subnet from the SE. This configuration prevents the SE from being automatically discovered with IP multicast. A remote HMC requires TCP/IP connectivity to each SE to be managed. Therefore, any

existing customer-installed firewalls between the remote HMC and its managed objects must permit communications between the HMC and the SE. For service and support, the remote HMC also requires connectivity to IBM, or to another HMC with connectivity to IBM through RSF. For more information, see 11.4, “Remote Support Facility” on page 403.

- ▶ Using a web browser to connect to an HMC

The z13s HMC application simultaneously supports one local user and any number of remote users. The user interface in the web browser is the same as the local HMC and has the same functions. Some functions are not available. Access by the UFD requires physical access to the HMC. Logon security for a web browser is provided by the local HMC user logon procedures. Certificates for secure communications are provided, and can be changed by the user. A remote browser session to the primary HMC that is managing an ensemble allows a user to perform ensemble-related actions.

Microsoft Internet Explorer, Mozilla Firefox, and Google Chrome were tested as remote browsers. For detailed web browser requirements, see the HMC and SE (Version 2.13.1) console help system or go to the IBM Knowledge Center at the following link:

<http://www.ibm.com/support/knowledgecenter>

After you get to the IBM Knowledge Center, click **z Systems**, and then click **z13s**.

Single object operating

It is not necessary to be physically close to a SE to use it. The HMC can be used to access the SE remotely by using the SOO task. The interface that is displayed is the same as the one on the SE. For more information, see the HMC and SE (Version 2.13.1) console help system or go to the IBM Knowledge Center at the following link:

<http://www.ibm.com/support/knowledgecenter>

After you get to the IBM Knowledge Center, click **z Systems**, and then click **z13s**.

11.5 HMC and SE key capabilities

The HMC and SE have many capabilities. This section covers the key areas. For a complete list of capabilities, see the HMC and SE (Version 2.13.1) console help system or go to the IBM Knowledge Center at the following link:

<http://www.ibm.com/support/knowledgecenter>

After you get to the IBM Knowledge Center, click **z Systems**, and then click **z13s**.

11.5.1 Central processor complex management

The HMC is the primary place for CPC control. For example, the input/output configuration data set (IOCDS) contains definitions of LPARs, channel subsystems, control units, and devices, and their accessibility from LPARs. IOCDS can be created and put into production from the HMC.

The HMC is used to start the power-on reset (POR) of the server. During the POR, processor units (PUs) are characterized and placed into their respective pools, memory is put into a single storage pool, and the IOCDS is loaded and initialized into the hardware system area (HSA).

The Hardware messages task displays hardware-related messages at the CPC level, LPAR level, or SE level. It also displays hardware messages that relate to the HMC itself.

11.5.2 Logical partition management

Use the HMC to define LPAR properties, such as the number of processors of each type, how many are reserved, or how much memory is assigned to it. These parameters are defined in LPAR profiles, and are stored on the SE.

Because Processor Resource/Systems Manager (PR/SM) must manage LPAR access to processors and the initial weights of each partition, weights are used to prioritize partition access to processors.

You can use the Load task on the HMC to perform an IPL of an operating system. This task causes a program to be read from a designated device, and starts that program. You can perform the IPL of the operating system from storage, the HMC DVD-RAM drive, the UFD, or an FTP server.

When an LPAR is active and an operating system is running in it, you can use the HMC to dynamically change certain LPAR parameters. The HMC provides an interface to change partition weights, add logical processors to partitions, and add memory.

LPAR weights can also be changed through a scheduled operation. Use the Customize Scheduled Operations task to define the weights that are set to LPARs at the scheduled time.

Channel paths can be dynamically configured on and off as needed for each partition from an HMC.

The Change LPAR Controls task for z13s servers can export the Change LPAR Controls table data to a comma-separated value (.csv)-formatted file. This support is available to a user when connected to the HMC remotely by a web browser.

Partition capping values can be scheduled and are specified on the Change LPAR Controls scheduled operation support. Viewing details about an existing Change LPAR Controls scheduled operation is available on the SE.

Absolute physical HW LPAR capacity setting

Driver 15 introduced the capability to define, in the image profile for shared processors, the absolute processor capacity that the image is allowed to use (independent of the image weight or other cappings).

To indicate that the LPAR can use the undedicated processors absolute capping, select **Absolute capping** on the Image Profile Processor settings to specify an absolute number of processors to cap the LPAR's activity to. The absolute capping value can either be **None** or a value for the number of processors (0.01 - 255.0).

LPAR group absolute capping

This is the next step in partition capping options available on z13s and z13 at Driver level 27 servers. Follow on to LPAR absolute capping, using a similar methodology to enforce:

- ▶ Customer licensing
- ▶ Non-z/OS partitions where group soft capping is not an option
- ▶ z/OS partitions where ISV does not support software capping

A group name, processor capping value, and partition membership are specified at HW console:

- ▶ Set an absolute capacity cap by CPU type on a group of LPARs.
- ▶ Allows each of the partitions to consume capacity up to their individual limits as long as the group's aggregate consumption does not exceed the group absolute capacity limit.
- ▶ Includes updated SysEvent Query Virtual Server (QVS) support (used by vendors who implement software pricing).
- ▶ Only shared partitions managed in these groups.
- ▶ Can specify caps for one or more processor types in the group.
- ▶ Specified in absolute processor capacity (for example, 2.5 processors).
- ▶ Use Change LPAR Group Controls (same windows that are used for software group defined capacity. Figure 11-13 shows a window on a z13.

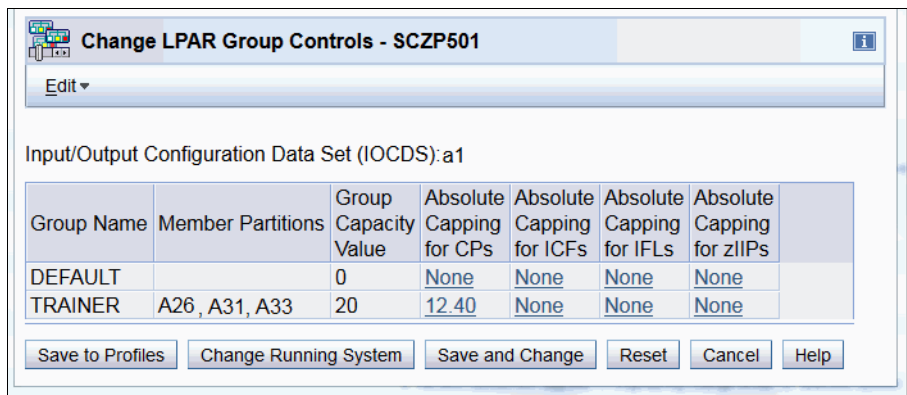


Figure 11-13 Change LPAR Group Controls - Group absolute capping

Absolute capping is specified as an absolute number of processors to cap the group's activity to. The value is specified to hundredths of a processor (for example, 4.56 processors) worth of capacity.

The value is not tied to the LICCC. Any value from 0.01 to 255.00 can be specified. This makes the profiles more portable and means you will not have issues in the future when profiles are migrated to new machines.

Although the absolute cap can be specified to hundredths of a processor, the exact amount might not be that precise. The same factors that influence the “machine capacity” influence the precision to which the absolute capping actually works.

11.5.3 Operating system communication

The Operating System Messages task displays messages from an LPAR. You also can enter operating system commands and interact with the system. This task is especially valuable for entering Coupling Facility Control Code (CFCC) commands.

The HMC also provides integrated 3270 and ASCII consoles. These consoles allow an operating system to be accessed without requiring other network or network devices, such as TCP/IP or control units.

Updates to x3270 support

The Configure 3270 Emulators task on the HMC and TKE consoles was enhanced with Driver 15 to verify the authenticity of the certificate that is returned by the 3270 server when a secure and encrypted SSL connection is established to an IBM host. This process is also known as *Secure 3270*.

Use the Certificate Management task if the certificates that are returned by the 3270 server are not signed by a well-known trusted certificate authority (CA) certificate, such as VeriSign or Geotrust. An advanced action within the Certificate Management task, Manage Trusted Signing Certificates, is used to add trusted signing certificates.

For example, if the certificate that is associated with the 3270 server on the IBM host is signed and issued by a corporate certificate, it must be imported, as shown in Figure 11-14.

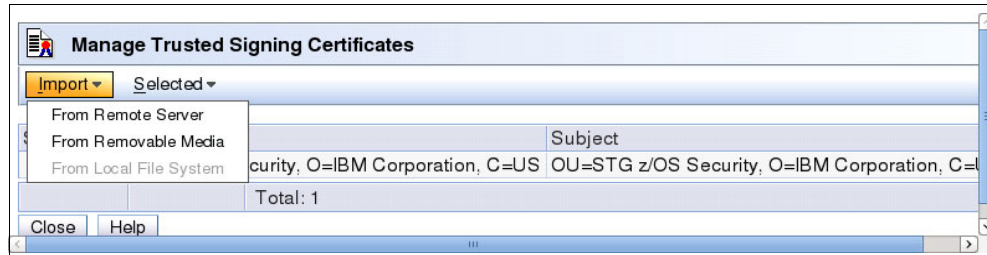


Figure 11-14 Manage Trusted Signing Certificates

If the connection between the console and the IBM host can be trusted when importing the certificate, the import from the remote server option can be used, as shown in Figure 11-15. Otherwise, import the certificate from removable media.

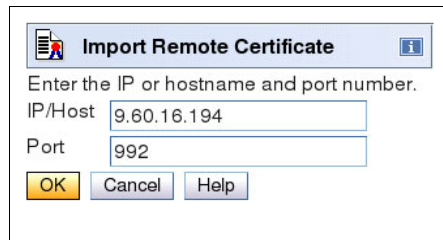


Figure 11-15 Import Remote Certificate example

A secure Telnet connection is established by adding the prefix L: to the IP address port of the IBM host, as shown in Figure 11-16.

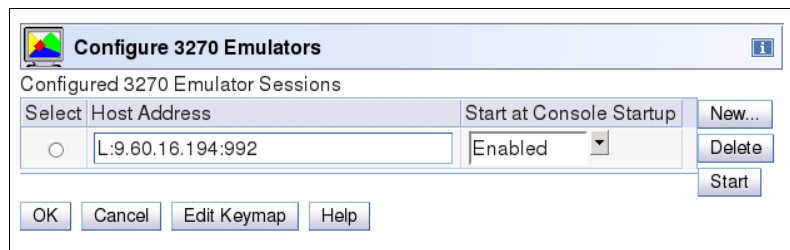


Figure 11-16 Configure 3270 Emulators

11.5.4 HMC and SE microcode

The microcode for the HMC, SE, CPC, and zBX is included in the driver/version. The HMC provides the management of the driver upgrade through Enhanced Driver Maintenance (EDM). EDM also provides the installation of the latest functions and patches (Microcode Change Levels (MCLs)) of the new driver.

When you perform a driver upgrade, always check the Driver xx Customer Exception Letter option in the Fixes section at the IBM Resource Link.

For more information, see Appendix 9.4, “Enhanced Driver Maintenance” on page 360.

Microcode Change Level

Regular installation of MCLs is key for reliability, availability, and serviceability (RAS), optimal performance, and new functions:

- ▶ Install MCLs on a quarterly basis at a minimum.
- ▶ Review hiper MCLs continuously to decide whether to wait for the next scheduled fix application session or to schedule one earlier if the risk assessment warrants.

Tip: The IBM Resource Link^a provides access to the system information for your z Systems according to the system availability data that is sent on a scheduled basis. It provides more information about the MCL status of your z13s server. To access the Resource Link, go to the following website:

<http://www.ibm.com/servers/resourcelink>

After you reach the Resource Link, click **Tools** → **Machine Information**, select your z Systems server, and click **EC/MCL**.

a. Registration is required to access the IBM Resource Link.

Microcode terms

The microcode has these characteristics:

- ▶ The driver contains EC streams.
- ▶ Each EC stream covers the code for a specific component of z13s servers. It has a specific name and an ascending number.
- ▶ The EC stream name and a specific number are one MCL.
- ▶ MCLs from the same EC stream must be installed in sequence.
- ▶ MCLs can have installation dependencies on other MCLs.
- ▶ Combined MCLs from one or more EC streams are in one bundle.
- ▶ An MCL contains one or more Microcode Fixes (MCFs).

Figure 11-17 shows how the driver, bundle, EC stream, MCL, and MCFs interact with each other.

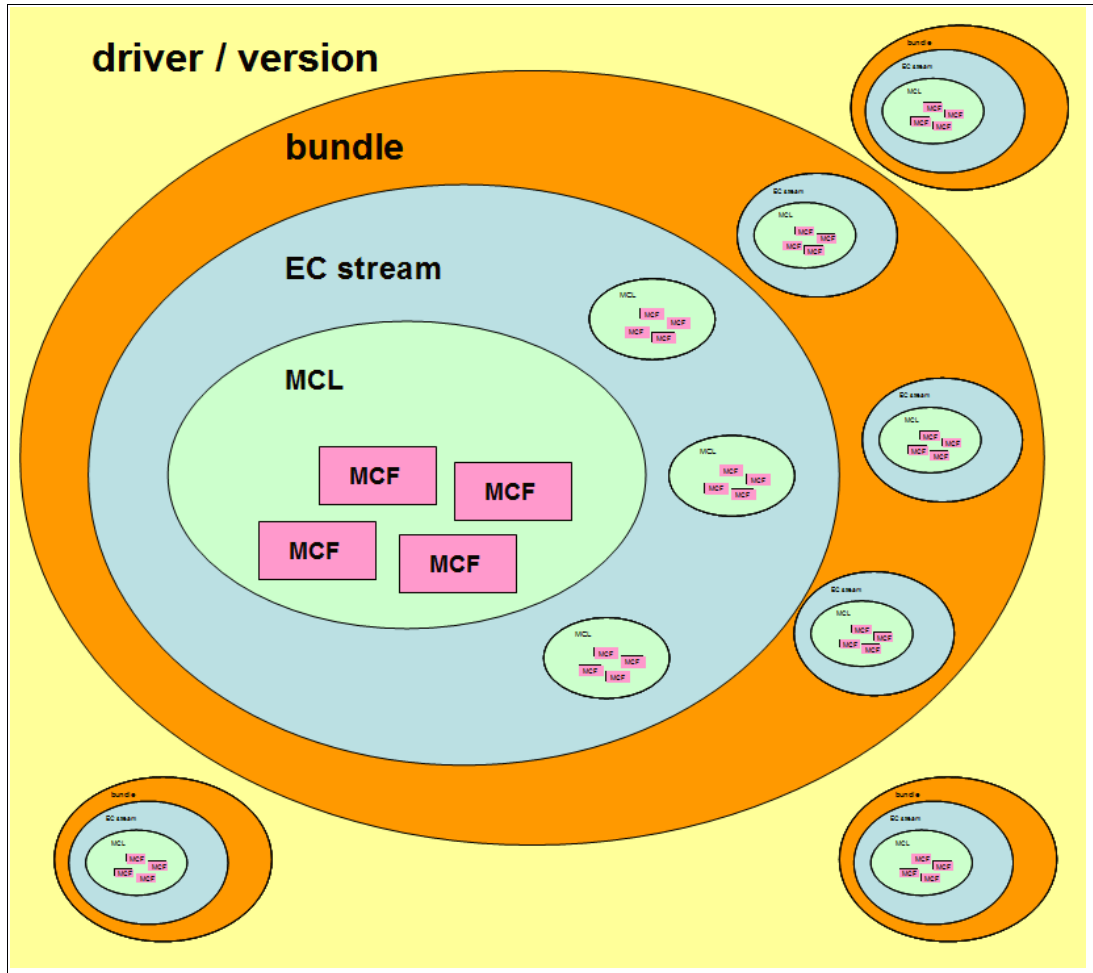


Figure 11-17 Microcode terms and interaction

Microcode installation by MCL bundle target

A *bundle* is a set of MCLs that are grouped during testing and released as a group on the same date. You can install an MCL to a specific target bundle level. The System Information window is enhanced to show a summary bundle level for the activated level, as shown in Figure 11-18.

The screenshot displays the System Information window with the following sections:

- Machine Information:** EC number: P00339, LIC control level: 0001, Engineering Changes AROM, Type: 2965, Model number: N20, Serial number: 00002008DA87, Version: 2.13.1, Driver level: 27, Bundle level: S01a (circled in red).
- Internal Code Change Information:** A table with columns: Select, EC Number, Retrieved Level, Installable Concurrent, Activated Level, Accepted Level, and Description.
- Pending Actions:** A message stating "Some actions might be pending. Click **Query Additional Actions...** for more information." with a button labeled "Query Additional Actions..."

Select	EC Number	Retrieved Level	Installable Concurrent	Activated Level	Accepted Level	Description
<input type="radio"/>	P00339	001	001	001		SE Framework
<input type="radio"/>	P08471					Concurrent Upgrade Sync Point
<input type="radio"/>	P08413	002	002	002		SE Licensed Internal Code Alerts
<input type="radio"/>	P08414	002	002	002		I390/PU-ML LIC
<input type="radio"/>	P08415					LPAR HV LIC
<input type="radio"/>	P08416	001	001	001		CFCC LIC
<input type="radio"/>	P08417					FCS Ficon Express8 LIC
<input type="radio"/>	P08422					Ficon Express 16S
<input type="radio"/>	P08454					z/VSE IP Assist
<input type="radio"/>	P08418					PCX LIC
<input type="radio"/>	P08419					CHANNEL DIAGS
<input type="radio"/>	P08420					SCSI IPL Machine Loader

Figure 11-18 System Information: Bundle level

OSC Concurrent Patch

Concurrent patch for OSC channels is now supported

11.5.5 Monitoring

This section addresses monitoring considerations.

Monitor task group

The Monitor task group on the HMC and SE includes monitoring-related tasks for the z13s server, as shown in Figure 11-19.

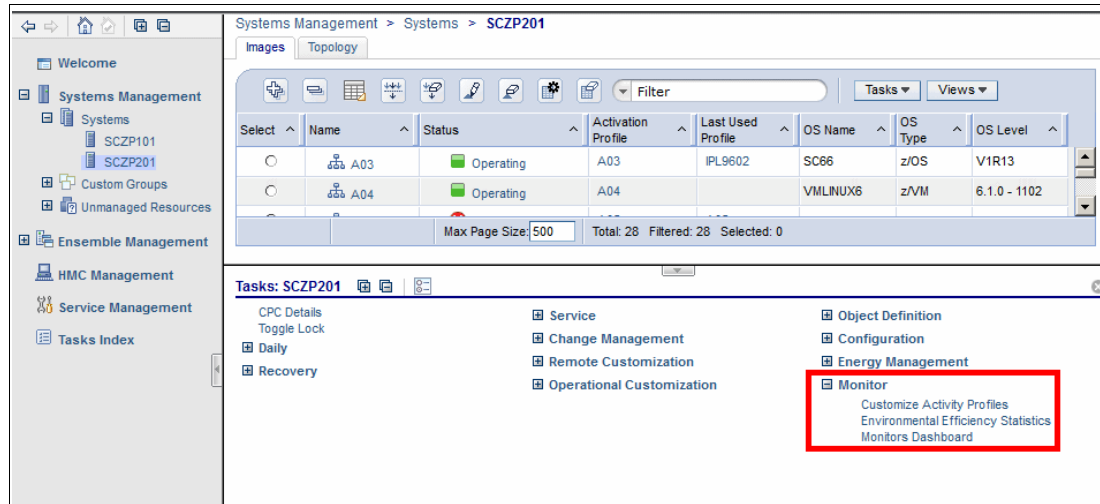


Figure 11-19 HMC Monitor task group

The Monitors Dashboard task

The Monitors Dashboard task supersedes the System Activity Display (SAD). In the z13s server, the Monitors Dashboard task in the Monitor task group provides a tree-based view of resources. Multiple graphical views exist for displaying data, including history charts. The Open Activity task, which is known as SAD, monitors processor and channel usage. It produces data that includes power monitoring information, power consumption, and the air input temperature for the server.

Figure 11-20 shows an example of the Monitors Dashboard task.

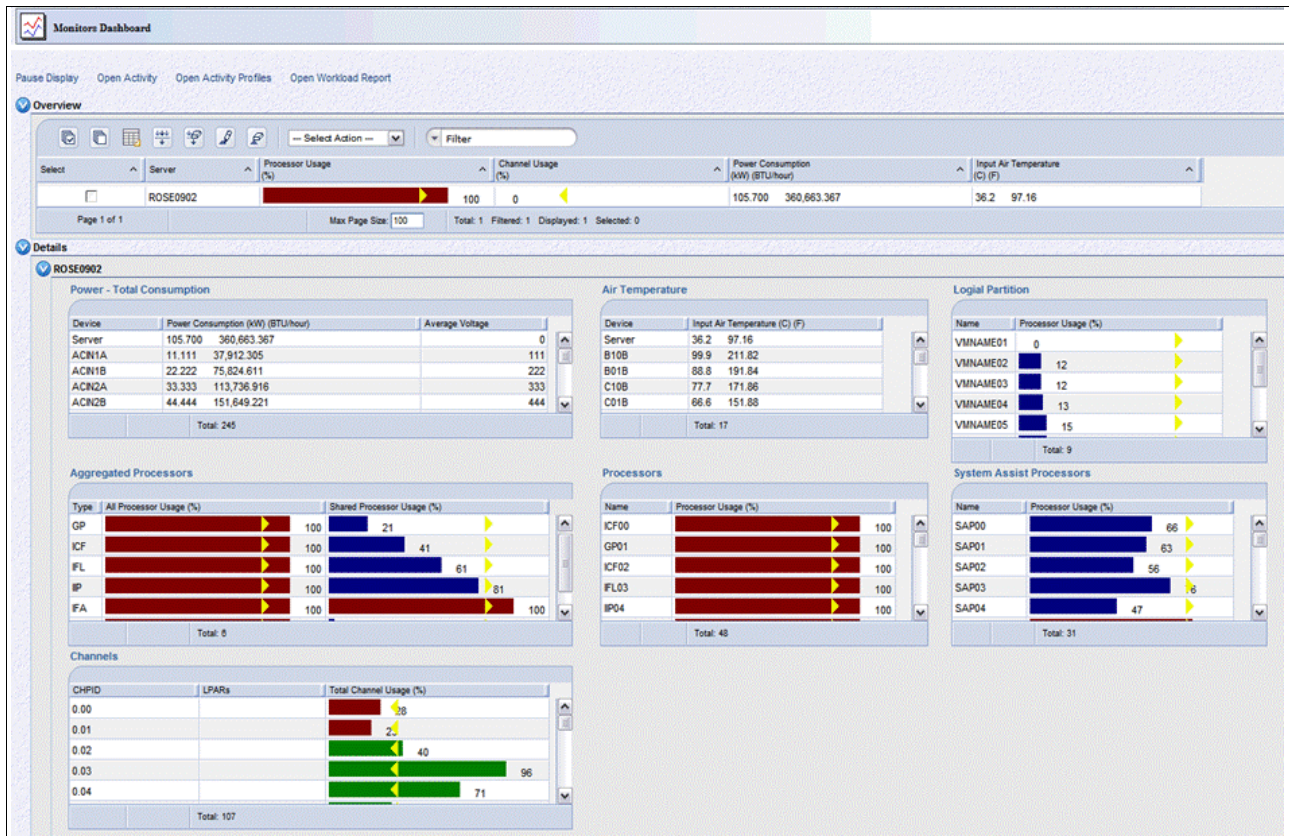


Figure 11-20 Monitors Dashboard task

Starting with Driver 27, you can display the activity for an LPAR by processor type, as shown in Figure 11-21.

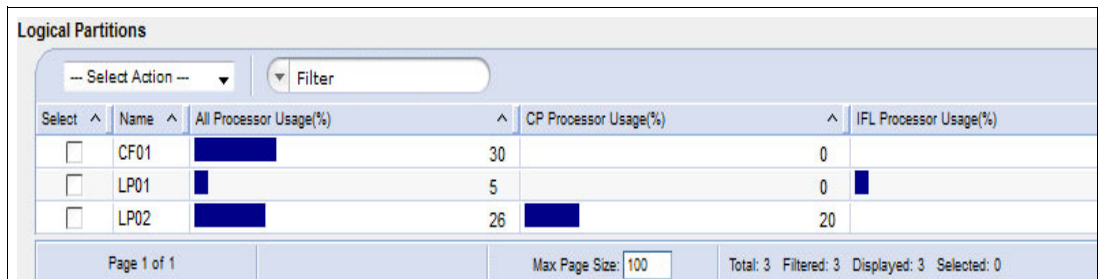


Figure 11-21 Display the activity for an LPAR by processor type

The Monitors Dashboard is enhanced to show SMT usage, as shown in Figure 11-22.

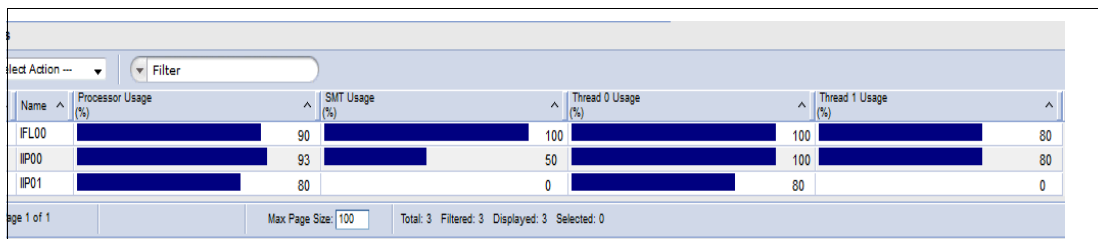


Figure 11-22 Display the SMT usage

The crypto utilization percentage is displayed on the Monitors Dashboard according to the physical channel ID (PCHID) number. The associated crypto number (Adjunct Processor Number) for this PCHID is also shown in the table. It provides information about the usage rate on a system-wide basis, not per LPAR, as shown in Figure 11-23.

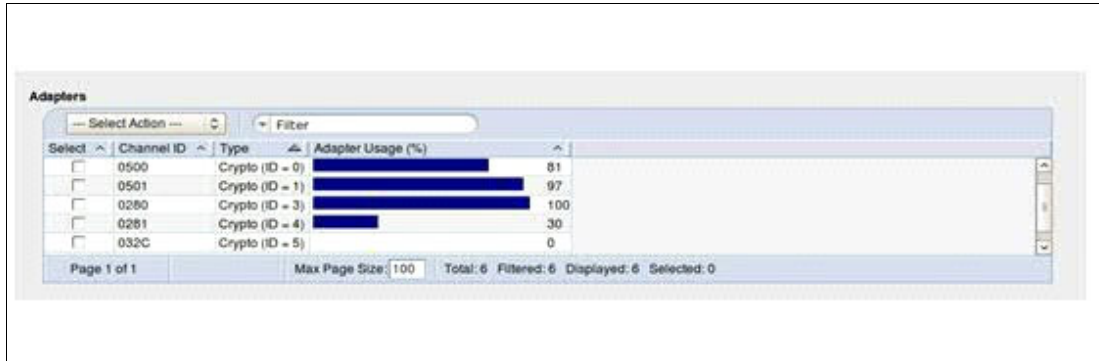


Figure 11-23 Monitors Dashboard: Crypto function integration

For Flash Express, a new window is added, as shown in Figure 11-24.

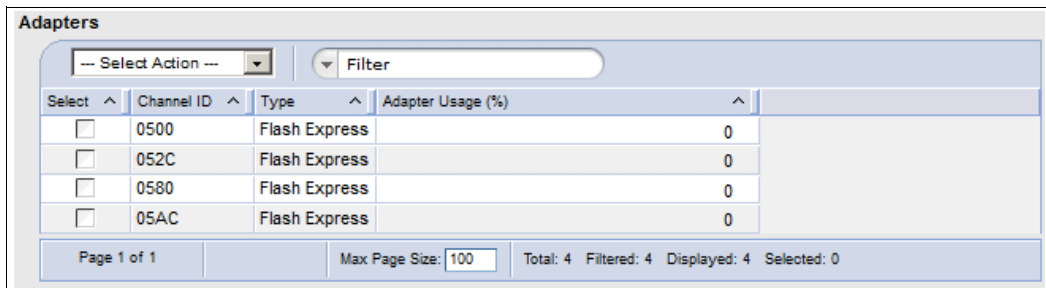


Figure 11-24 Monitors Dashboard - Flash Express function integration

Environmental Efficiency Statistics task

The Environmental Efficiency Statistics task (Figure 11-25) is part of the Monitor task group. It provides historical power consumption and thermal information for the zEnterprise CPC, and is available on the HMC.

The data is presented in table format and graphical “histogram” format. The data can also be exported to a .csv-formatted file so that the data can be imported into a spreadsheet. For this task, you must use a web browser to connect to an HMC.

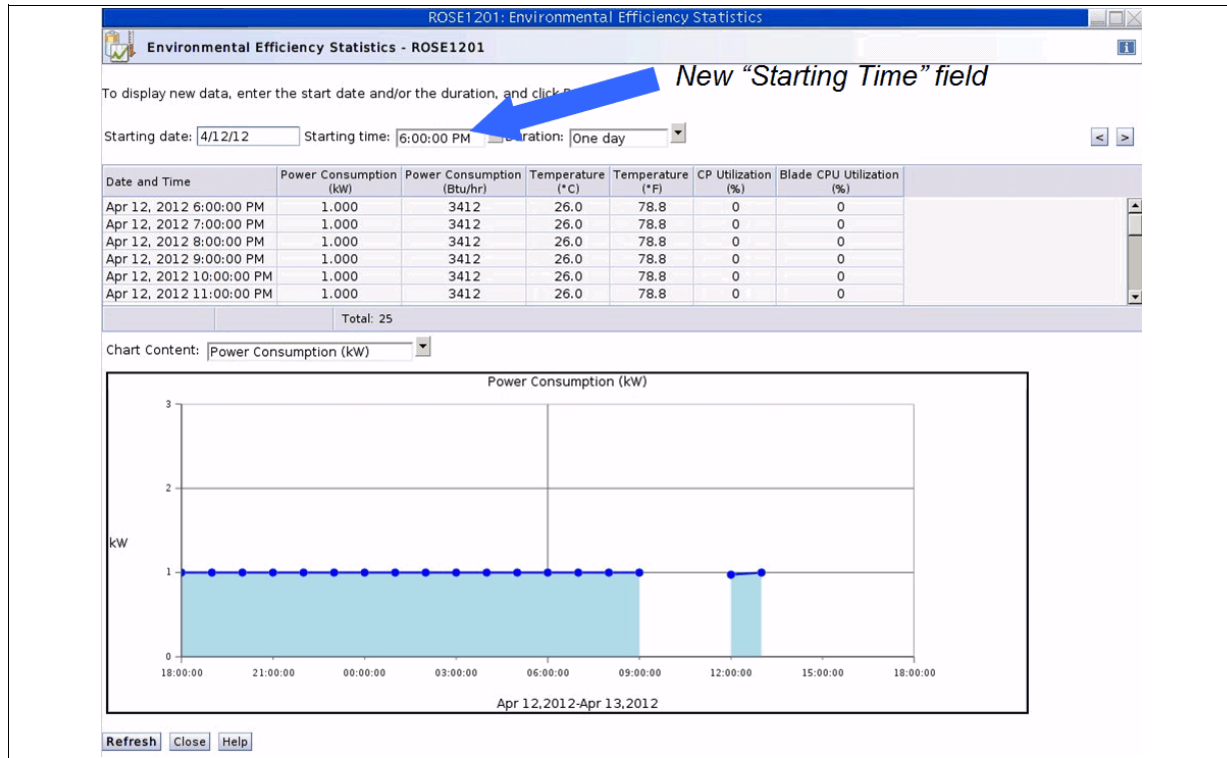


Figure 11-25 Environmental Efficiency Statistics

11.5.6 Capacity on demand support

All capacity on demand (CoD) upgrades are performed by using the SE Perform a Model Conversion task. Use the task to retrieve and activate a permanent upgrade, and to retrieve, install, activate, and deactivate a temporary upgrade. The task shows a list of all installed or staged LIC configuration code (LICCC) records to help you manage them. It also shows a history of recorded activities.

The HMC for z13s servers has these CoD capabilities:

- ▶ SNMP API support:
 - API interfaces for granular activation and deactivation
 - API interfaces for enhanced CoD query information
 - API event notification for any CoD change activity on the system
 - CoD API interfaces, such as On/Off CoD and Capacity BackUp (CBU)
- ▶ SE window features (accessed through HMC SOO):
 - Window controls for granular activation and deactivation
 - History window for all CoD actions
 - Description editing of CoD records

- ▶ HMC/SE Version 2.13.1 provides the following CoD information:
 - Millions of service units (MSU) and processor tokens
 - Last activation time
 - Pending resources that are shown by processor type instead of just a total count
 - Option to show the details of installed and staged permanent records
 - More details for the Attention state by providing seven more flags

HMC and SE are a part of the z/OS Capacity Provisioning environment. The Capacity Provisioning Manager (CPM) communicates with the HMC through z Systems APIs, and enters CoD requests. For this reason, SNMP must be configured and enabled by using the Customize API Settings task on the HMC.

For more information about using and setting up CPM, see these publications:

- ▶ *z/OS MVS™ Capacity Provisioning User's Guide, SC33-8299*
- ▶ *z Systems Capacity on Demand User's Guide, SC28-6943*

11.5.7 Features on Demand support

FoD is a new centralized way to entitle flexibly features and functions on the system. FoD contains, for example, the zBX High Water Marks (HWMs). HWMs refer to highest quantity of blade entitlements by blade type that the customer has purchased. On the z196/z114, the zBX HWMs are stored in the processor and memory LICCC record. On z13s servers, they are in the FoD record.

FoD allows separate LICCC controls for z Systems processors (central processors (CPs), Integrated Facility for Linux (IFL) and IBM z Integrated Information Processors (zIIPs)), and zBX HWMs, providing entitlement controls for each blade type. It is also used as LICCC support for the following features:

- ▶ IBM zAware: Enablement/max connections
- ▶ Base/proprietary service: Expiration date
- ▶ New features: Yet to be announced or developed

11.5.8 Server Time Protocol support

With the STP functions, the role of the HMC is extended to provide the user interface for managing the Coordinated Timing Network (CTN):

- ▶ The z13s server relies solely on STP for time synchronization, and continues to provide support of a pulse per second (PPS) port. It maintains accuracy of 10 microseconds as measured at the PPS input of the z13s server. If STP uses a Network Time Protocol (NTP) server without PPS, a time accuracy of 100 milliseconds to the ETS is maintained.
- ▶ z13s servers cannot be in the same CTN with a System z10 (n-2) or earlier systems. As a consequence, z13s servers cannot become member of an STP mixed CTN.
- ▶ An STP-only CTN can be managed by using different HMCs. However, the HMC must be at the same driver level (or later) than any SE that is to be managed. Furthermore, all SEs to be managed must be known (defined) to that HMC.

In a STP-only CTN, the HMC can be used to perform the following tasks:

- ▶ Initialize or modify the CTN ID.
- ▶ Initialize the time, manually or by contacting an NTP server.
- ▶ Initialize the time zone offset, Daylight Saving Time offset, and leap second offset.
- ▶ Assign the roles of preferred, backup, and current time servers, and arbiter.

- ▶ Adjust time by up to plus or minus 60 seconds.
- ▶ Schedule changes to the offsets listed. STP can automatically schedule Daylight Saving Time, based on the selected time zone.
- ▶ Monitor the status of the CTN.
- ▶ Monitor the status of the coupling links that are initialized for STP message exchanges.
- ▶ For diagnostic purposes, the PPS port state on a z13s server can be displayed and fenced ports can be reset individually.

STP window enhancements

The z13s STP windows are enhanced to show the following additional or enhanced information:

- ▶ The new Integrated Coupling Adapters (ICAs) are added to the list of possible STP-supporting Coupling Links.
- ▶ A new task role is added within the Customize User Controls window that allows the enablement of the View System (Sysplex) Time option. This option provides view-only access to the STP windows for each individual user role.
- ▶ The Initialize Time window shows the current values of leap second offset, time zone, and 'date and time' if the CTN time was set before.
- ▶ Set Date and Time is modified so that Use External Time Source (ETS) is the first option. This configuration should encourage the user to select the NTP option (if their system is configured for NTP) because a NTP-set time is more precise compared to a user-set time.
- ▶ With Driver 22 and later, the HMC Enable for time synchronization task was moved from the "Add (or Modify) object definition" window to the Customize Console Date and Time window. This function synchronizes the HMCs date and time to either an NTP server (if defined) or to a SE Date and Time, with the NTP server option provided as the first choice.

As shown in Figure 11-26, the NTP option is the recommended option, if an NTP server is available. If an NTP server is not available for this HMC, any defined CPC SE can be selected after you select **Selected CPCs**.

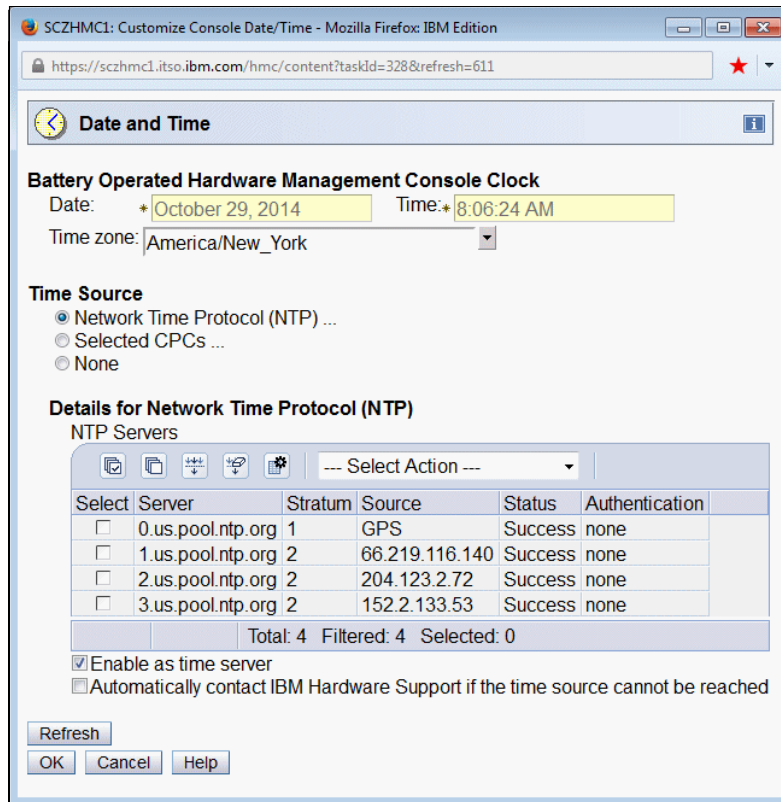


Figure 11-26 Customize Console Date and Time

- ▶ The Timing Network window now includes the next scheduled Daylight Saving Time change and the next leap second adjustment, as shown in Figure 11-27. The schedules that are shown are the ones for the next DLS change (either given per automatic or scheduled adjustment) and for the next Leap Second change (given per scheduled adjustment).

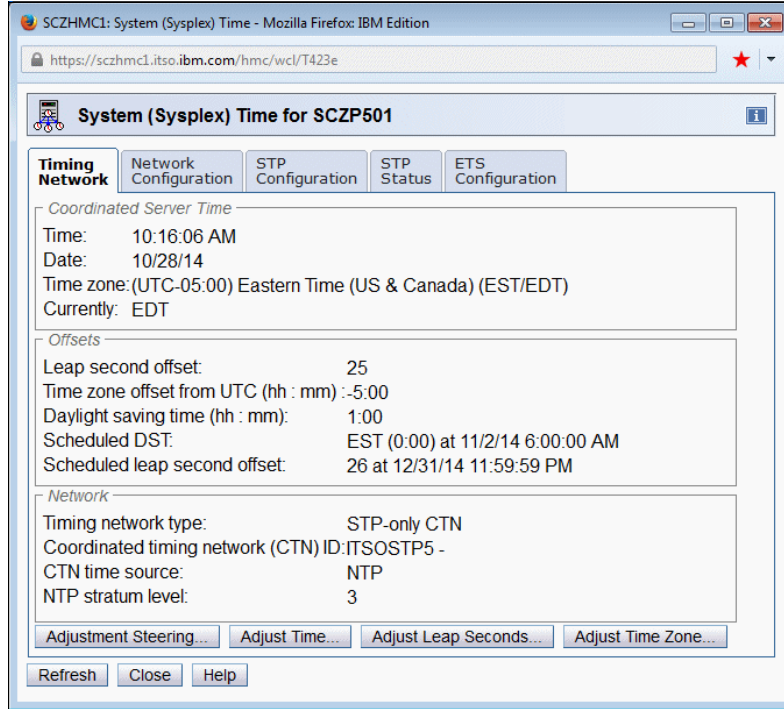


Figure 11-27 Timing Network window with Scheduled DST and Scheduled leap second offset

Attention: Figure 11-27 shows a ‘Schedule leap second offset’ change to 26 seconds that is scheduled for 12/31/2014. This is *not* a real leap second released by the International Earth Rotation and Reference System Services. It was temporarily set *only* to show the window appearance.

Enhanced Console Assisted Recovery

Enhanced Console Assisted Recovery (ECAR) goes through these steps:

1. When the Primary Time Server (PTS/CTS) encounters a check-stop condition, the CEC informs its SE and HMC.
2. The PTS SE recognizes the check-stop pending condition, and the PTS SE STP code is called.
3. The PTS SE sends an ECAR request through the HMC to the BTS SE.
4. The Backup Time Server (BTS) SE communicates with the BTS to start the takeover.

ECAR support is faster than the original CAR support for these reasons:

- ▶ The console path goes from a 2-way path to a one-way path.
- ▶ There is almost no lag time between the system check-stop and the start of ECAR processing.
- ▶ Because the request is generated from the PTS before system logging, it avoids the potential of recovery being held up.

Requirements

ECAR is only available on z13 Driver 27 and z13s servers. In a mixed environment with previous generation machines, you should define a z13 or z13s server as the PTS and CTS.

For more planning and setup information, see the following publications:

- ▶ *Server Time Protocol Planning Guide*, SG24-7280
- ▶ *Server Time Protocol Implementation Guide*, SG24-7281
- ▶ *Server Time Protocol Recovery Guide*, SG24-7380

11.5.9 NTP client and server support on the HMC

The NTP client support allows a STP-only CTN to use an NTP server as an external time source (ETS).

This capability addresses the following requirements:

- ▶ Clients who want time accuracy for the STP-only CTN
- ▶ Clients who use a common time reference across heterogeneous systems

The NTP server becomes the single time source, the ETS, for STP and other servers that are not z Systems (such as AIX, Microsoft Windows, and others) that have NTP clients.

The HMC can act as an NTP server. With this support, the z13s server can get time from the HMC without accessing a LAN other than the HMC/SE network. When the HMC is used as an NTP server, it can be configured to get the NTP source from the Internet. For this type of configuration, a LAN that is separate from the HMC/SE LAN can be used.

HMC NTP broadband authentication support

HMC NTP authentication can be used since HMC Driver 15. The SE NTP support is unchanged. To use this option on the SE, configure the HMC with this option as an NTP server for the SE.

Authentication support with a proxy

Some client configurations use a proxy for external access outside the corporate data center. NTP requests are User Datagram Protocol (UDP) socket packets and cannot pass through the proxy. The proxy must be configured as an NTP server to get to target servers on the web. Authentication can be set up on the client's proxy to communicate with the target time sources.

Authentication support with a firewall

If you use a firewall, HMC NTP requests can pass through it. Use HMC authentication to ensure untampered time stamps.

NTP symmetric key and autokey authentication

With symmetric key and autokey authentication, the highest level of NTP security is available. HMC Level 2.12.0 and later provide windows that accept and generate key information to be configured into the HMC NTP configuration. They can also issue NTP commands, as shown in Figure 11-28.

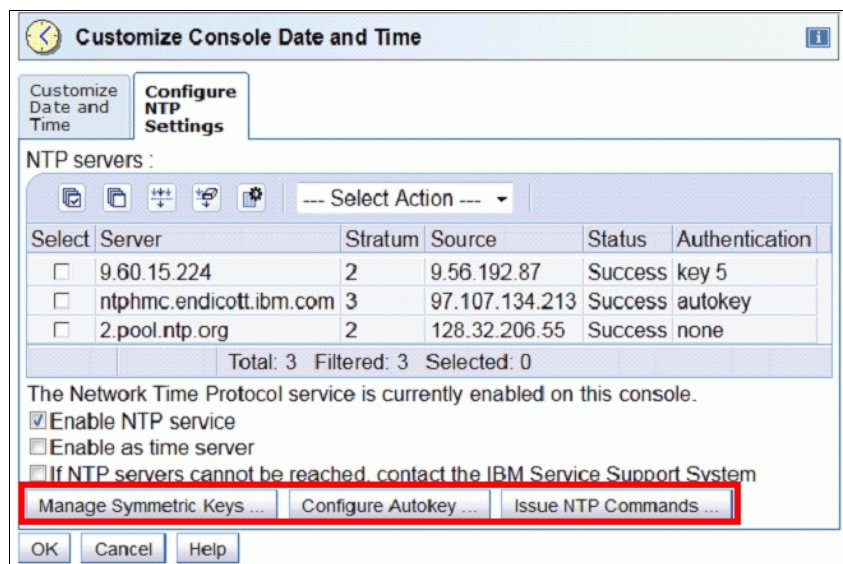


Figure 11-28 HMC NTP broadband authentication support

The HMC offers symmetric key and autokey authentication and NTP commands:

- ▶ Symmetric key (NTP V3-V4) authentication
Symmetric key authentication is described in RFC 1305, which was made available in NTP Version 3. Symmetric key encryption uses the same key for both encryption and decryption. Users who are exchanging data keep this key secret. Messages encrypted with a secret key can be decrypted only with the same secret key. Symmetric key authentication supports network address translation (NAT).
- ▶ Symmetric key autokey (NTP V4) authentication
This autokey uses public key cryptography, as described in RFC 5906, which was made available in NTP Version 4. You can generate keys for the HMC NTP by clicking **Generate Local Host Key** in the Autokey Configuration window. This option issues the **ntp-keygen** command to generate the specific key and certificate for this system. Autokey authentication is not available with the NAT firewall.
- ▶ Issue NTP commands
NTP command support is added to display the status of remote NTP servers and the current NTP server (HMC).

For more information about planning and setup for STP and NTP, see the following publications:

- ▶ *Server Time Protocol Planning Guide*, SG24-7280
- ▶ *Server Time Protocol Implementation Guide*, SG24-7281
- ▶ *Server Time Protocol Recovery Guide*, SG24-7380

Time coordination for zBX components

NTP clients that run on blades in the zBX can synchronize their time to the SE battery operated clock (BOC). The SE BOC is synchronized to the z13s time-of-day (TOD) clock every hour. This process allows the SE clock to maintain a time accuracy of 100 milliseconds to an NTP server that is configured as the ETS in a STP-only CTN. This configuration is shown in Figure 11-29. For more information, see the *Server Time Protocol Planning Guide*, SG24-7280.

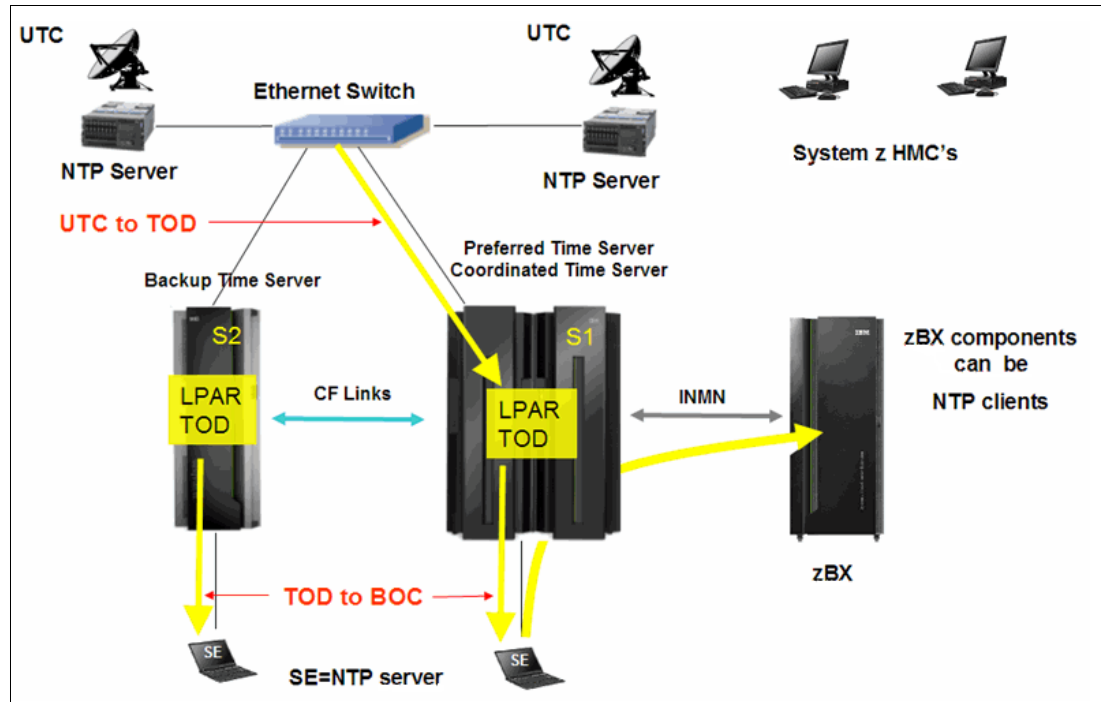


Figure 11-29 Time coordination for zBX components

11.5.10 Security and user ID management

This section addresses security and user ID management considerations.

HMC and SE security audit improvements

With the Audit and Log Management task, audit reports can be generated, viewed, saved, and offloaded. The Customize Scheduled Operations task allows you to schedule audit report generation, saving, and offloading. The Monitor System Events task allows Security Logs to send email notifications by using the same type of filters and rules that are used for both hardware and operating system messages.

With z13s servers, you can offload the following HMC and SE log files for customer audit:

- ▶ Console event log
- ▶ Console service history
- ▶ Tasks performed log
- ▶ Security logs
- ▶ System log

Full log offload and delta log offload (since the last offload request) are provided. Offloading to removable media and to remote locations by FTP is available. The offloading can be manually started by the new Audit and Log Management task or scheduled by the Customize Scheduled Operations task. The data can be offloaded in the HTML and XML formats.

HMC user ID templates and LDAP user authentication

LDAP user authentication and HMC user ID templates enable the addition and removal of HMC users according to your corporate security environment. These processes use an LDAP server as the central authority. Each HMC user ID template defines the specific authorization levels for the tasks and objects for the user who is mapped to that template. The HMC user is mapped to a specific user ID template by user ID pattern matching. The system then obtains the name of the user ID template from content in the LDAP server schema data.

Default HMC user IDs

It is no longer possible to change the Managed Resource or Task Roles of the default user IDs operator, advanced, sysprog, acsadmin, and service.

If you want the capability to change the roles for a default user ID, create your own version by copying an existing default user ID.

View-only user IDs and view-only access for HMC and SE

With HMC and SE user ID support, users can be created that have “view-only” access to selected tasks. Support for “view-only” user IDs is available for the following purposes:

- ▶ Hardware messages
- ▶ Operating system messages
- ▶ Customize or delete activation profiles
- ▶ Advanced facilities
- ▶ Configure on and off

HMC and SE secure FTP support

You can use a secure FTP connection from a HMC/SE FTP client to a customer FTP server location. This configuration is implemented by using the Secure Shell (SSH) File Transfer Protocol, which is an extension of SSH. You can use the Manage SSH Keys console action, which is available to both the HMC and SE, to import public keys that are associated with a host address.

The Secure FTP infrastructure allows HMC and SE applications to query whether a public key is associated with a host address and to use the Secure FTP interface with the appropriate public key for a host. Tasks that use FTP now provide a selection for the secure host connection.

When selected, the task verifies that a public key is associated with the specified host name. If none is provided, a message window is displayed that points to the Manage SSH Keys task to input a public key. The following tasks provide this support:

- ▶ Import/Export IOCDs
- ▶ Advanced Facilities FTP IBM Content Collector Load
- ▶ Audit and Log Management (Scheduled Operations only)
- ▶ FCP Configuration Import/Export
- ▶ OSA view Port Parameter Export
- ▶ OSA Integrated Console Configuration Import/Export

11.5.11 System Input/Output Configuration Analyzer on the SE and HMC

The System Input/Output Configuration Analyzer task supports the system I/O configuration function.

The information that is needed to manage a system's I/O configuration must be obtained from many separate sources. The System Input/Output Configuration Analyzer task enables the system hardware administrator to access, from one location, the information from those sources. Managing I/O configurations then becomes easier, particularly across multiple servers.

The System Input/Output Configuration Analyzer task runs the following functions:

- ▶ Analyzes the current active IOCDs on the SE.
- ▶ Extracts information about the defined channel, partitions, link addresses, and control units.
- ▶ Requests the channels' node ID information. The Fibre Channel connection (FICON) channels support remote node ID information, which is also collected.

The System Input/Output Configuration Analyzer is a view-only tool. It does not offer any options other than viewing. With the tool, data is formatted and displayed in five different views. The tool provides various sort options, and data can be exported to an UFD for later viewing.

The following five views are available:

- ▶ PCHID Control Unit View shows PCHIDs, channel subsystems (CSS), channel-path identifiers (CHPIDs), and their control units.
- ▶ PCHID Partition View shows PCHIDs, CSS, CHPIDs, and the partitions in which they exist.
- ▶ Control Unit View shows the control units, their PCHIDs, and their link addresses in each CSS.
- ▶ Link Load View shows the Link address and the PCHIDs that use it.
- ▶ Node ID View shows the Node ID data under the PCHIDs.

11.5.12 Automated operations

As an alternative to manual operations, an application can interact with the HMC and SE through an API. The interface allows a program to monitor and control the hardware components of the system in the same way you can. The HMC APIs provide monitoring and control functions through SNMP and the CIM. These APIs can get and set a managed object's attributes, issue commands, receive asynchronous notifications, and generate SNMP traps.

The HMC supports the CIM as an extra systems management API. The focus is on attribute query and operational management functions for z Systems servers, such as CPCs, images, and activation profiles. z13s servers contain a number of enhancements to the CIM systems management API. The function is similar to that provided by the SNMP API.

For more information about APIs, see *z Systems Application Programming Interfaces*, SB10-7164.

11.5.13 Cryptographic support

This section lists the cryptographic management and control functions that are available in the HMC and the SE.

Cryptographic hardware

z13s servers include both standard cryptographic hardware and optional cryptographic features for flexibility and growth capability.

The HMC/SE interface provides the following capabilities:

- ▶ Defining the cryptographic controls
- ▶ Dynamically adding a Crypto feature to a partition for the first time
- ▶ Dynamically adding a Crypto feature to a partition that already uses Crypto
- ▶ Dynamically removing a Crypto feature from a partition

The Crypto Express5S, a new Peripheral Component Interconnect Express (PCIe) cryptographic coprocessor, is an optional z13s exclusive feature. Crypto Express5S provides a secure programming and hardware environment on which crypto processes are run. Each Crypto Express5S adapter can be configured by the installation as a Secure IBM Common Cryptographic Architecture (CCA) coprocessor, a Secure IBM Enterprise Public Key Cryptography Standards (PKCS) #11 (EP11) coprocessor, or an accelerator.

When EP11 mode is selected, unique Enterprise PKCS #11 firmware is loaded into the cryptographic coprocessor. It is separate from the CCA firmware that is loaded when a CCA coprocessor is selected. CCA firmware and PKCS #11 firmware cannot coexist at the same time in a card.

The Trusted Key Entry (TKE) Workstation with smart card reader feature is required to support the administration of the Crypto Express5S when configured as an Enterprise PKCS #11 coprocessor.

To support the new Crypto Express5S card, the Cryptographic Configuration window was changed to support the following card modes:

- ▶ Accelerator mode (CEX5A)
- ▶ CCA Coprocessor mode (CEX5C)
- ▶ PKCS #11 Coprocessor mode (CEX5P)

The Cryptographic Configuration window also has the following updates:

- ▶ Support for a Client-Initiated Self-test (CIS) for Crypto running EP11 Coprocessor mode.
- ▶ TKE commands are always permitted for EP11 mode.
- ▶ The Test RN Generator function was modified and generalized to also support CIS, depending on the mode of the crypto card.
- ▶ The Crypto Details window was changed to display the crypto part number.
- ▶ Support is now provided for up to four User Defined Extensions (UDX) files. Only UDX CCA is supported for z13s servers.
- ▶ UDX import now supports importing from DVD only.

Figure 11-30 shows an example of the Cryptographic Configuration window.

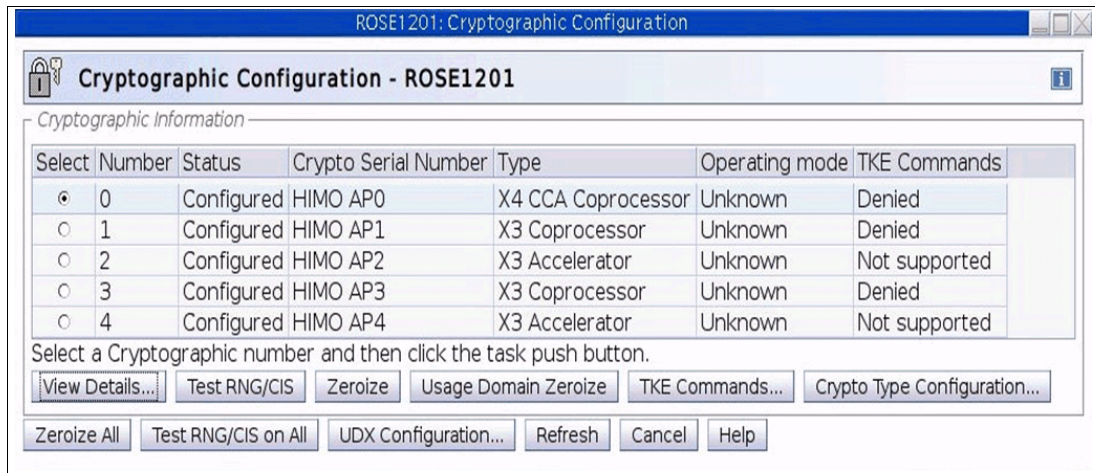


Figure 11-30 Cryptographic Configuration window

The Usage Domain Zeroize task is provided to clear the appropriate partition crypto keys for a usage domain when you remove a crypto card from a partition. Crypto Express5S in EP11 mode is configured to the standby state after zeroize.

For more information, see *IBM z13 Configuration Setup*, SG24-8260.

Digitally signed firmware

Critical issues with firmware upgrades are security and data integrity. Procedures are in place to use a process to digitally sign the firmware update files that are sent to the HMC, the SE, and the TKE. Using a hash algorithm, a message digest is generated that is then encrypted with a private key to produce a digital signature.

This operation ensures that any changes that are made to the data are detected during the upgrade process by verifying the digital signature. It helps ensure that no malware can be installed on z Systems products during firmware updates. It enables the z13s Central Processor Assist for Cryptographic Function (CPACF) functions to comply with Federal Information Processing Standard (FIPS) 140-2 Level 1 for Cryptographic Licensed Internal Code (LIC) changes. The enhancement follows the z Systems focus on security for the HMC and the SE.

11.5.14 Installation support for z/VM using the HMC

Starting with z/VM V5R4 and System z10, Linux on z Systems can be installed in a z/VM virtual machine from HMC workstation media. This Linux on z Systems installation can use the existing communication path between the HMC and the SE. No external network or extra network setup is necessary for the installation.

11.5.15 Dynamic Partition Manager

DPM is a z Systems mode of operation that provides a simplified approach to create and manage virtualized environments, reducing the barriers of its adoption for new and existing customers. For more information about Dynamic Partition Manager (DPM), see Appendix E., "IBM Dynamic Partition Manager" on page 501.

Setting up is a disruptive action. The selection of DPM mode of operation is done by using a function called **Enable Dynamic Partition Manager** under the SE CPC Configuration menu as shown in Figure 11-31.

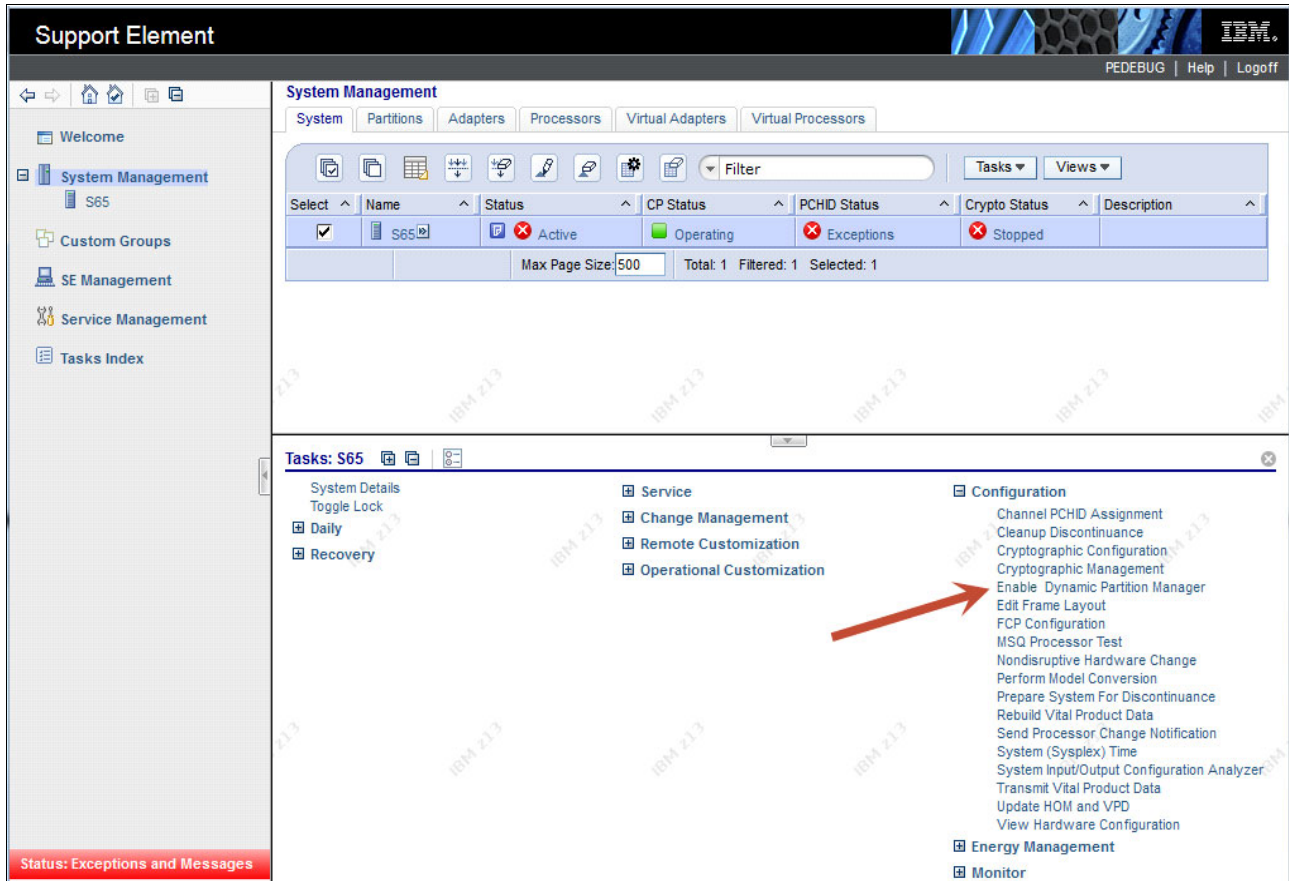


Figure 11-31 Enabling Dynamic Partition Manager

After the CPC is restarted and you log on to the HMC that has this CPC defined, the HMC shows another welcome window as shown in Figure 11-32.

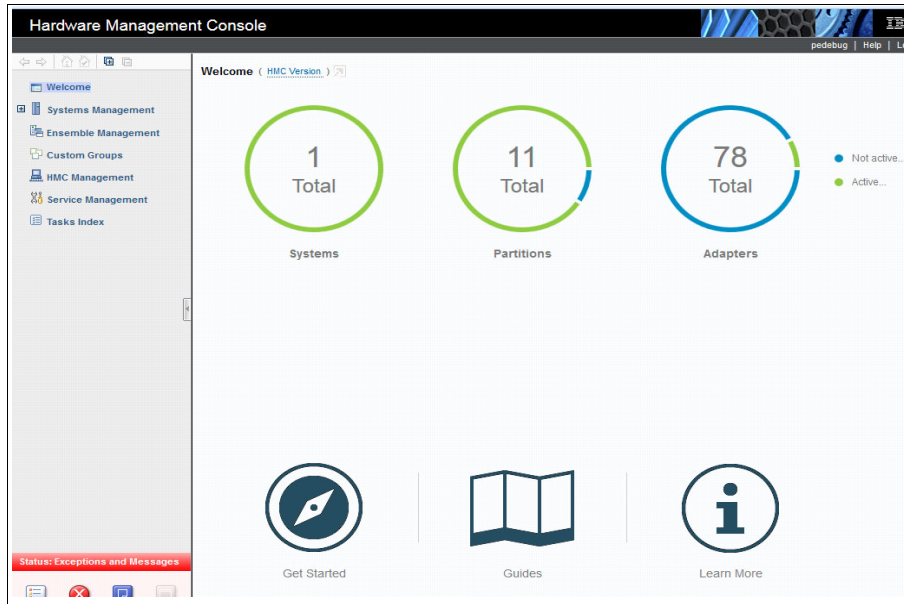


Figure 11-32 HMC welcome window (CPC in DPM Mode)

New LPARs can be added by selecting **Get Started**. More information can be found in Appendix E., “IBM Dynamic Partition Manager” on page 501

11.6 HMC in an ensemble

An *ensemble* is a platform systems management domain that consists of up to eight z13s or IBM zEnterprise System nodes, and up to eight zBX Model 004 systems. Each node is a zEnterprise CPC or a zBX Model 004. The ensemble provides an integrated way to manage virtual server resources and the workloads that can be deployed on those resources. The zEnterprise is a workload-optimized technology system that delivers a multiple platform, integrated hardware system. This system spans z Systems, System p, and System x blade server technologies.

Management of the ensemble is provided by the IBM zEnterprise Unified Resource Manager.

Consideration: The ensemble HMC mode is available only for managing IBM z Systems servers (z13s, z13, zEC12, zBC12, z196, and z114).

11.6.1 Unified Resource Manager

The ensemble is provisioned and managed through the Unified Resource Manager, which is in the HMC. The Unified Resource Manager provides a large set of functions for system management.

Figure 11-33 shows the Unified Resource Manager functions and suites.

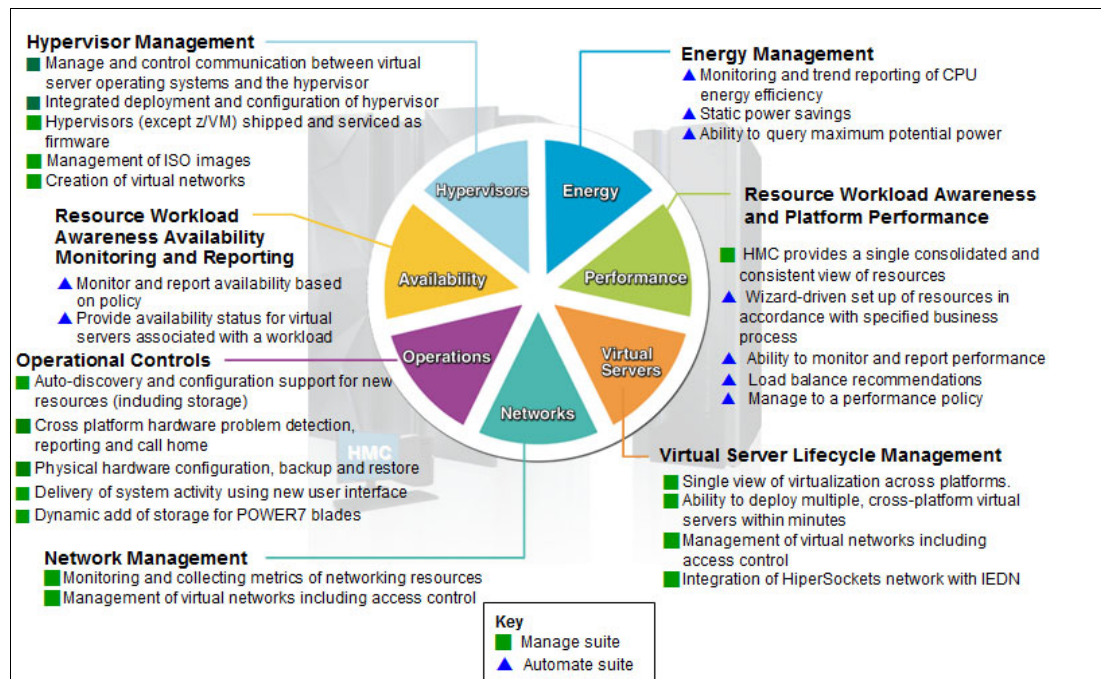


Figure 11-33 Unified Resource Manager functions and suites

Overview

Unified Resource Manager provides the following functions:

- ▶ Hypervisor management
 - Provides tasks for managing the hypervisor lifecycle, managing storage resources, providing RAS and first-failure data capture (FFDC) features, and monitoring the supported hypervisors.
- ▶ Ensemble membership management
 - Provides tasks for creating an ensemble and controlling membership of the ensemble.
- ▶ Storage management
 - Provides a common user interface for allocation and deallocation of physical and virtual storage resources for an ensemble.
- ▶ Virtual server management
 - Provides lifecycle management to create, delete, activate, deactivate, and modify the definitions of virtual servers.
- ▶ Virtual network management
 - Allows management of networking resources for an ensemble.
- ▶ Availability management
 - The resource workload “Awareness availability” function monitors and reports virtual servers’ availability status, based on the workloads of which they are a part and their associated workload policies.

► Performance management

Provides a global performance view of all the virtual servers that support workloads that are deployed in an ensemble. The virtual server workload performance goal is like a simplified z/OS Workload Manager (WLM) policy:

- You can define, monitor, report, and manage the performance of virtual servers based on workload performance policies.
- Policies are associated to the workload:
 - From the overall Workload performance health report, you can review contributions of individual virtual servers.
 - You can manage resources across virtual servers within a hypervisor instance.

► Ensemble Availability Management (EAM)

EAM implements basic availability services for the ensemble as part of the Unified Resource Manager. It provides consistent high availability management across virtual servers running on the zEnterprise and zBX in an ensemble, allowing error monitoring and identifying conditions that affect the availability of resources.

The EAM availability assessment is based on user-defined policies for the following objects:

- PR/SM LPARs running on zEnterprise
- Kernel-based virtual machine (KVM) virtual servers running on zBX
- PowerVM virtual servers running on zBX

► EAM enhancements

EAM availability enhancements are based on Workload Resource Group (WRG) definitions. A WRG is a grouping mechanism and management view of the virtual servers that support a business application. The availability definitions are created at the HMC and include these functions:

- Creation of element groups (an *element* is a virtual server that is associated to a specific workload. Elements are grouped to form a *Resource Group*. Resource Groups are associated, based on a defined workload, to form a WRG.)
- Addition of virtual servers and element groups to a workload.
- Definition of new availability policies.
- Definition of workload status: Performance and availability compliance.
- Providing workload details summary and reports.

► Energy management:

- Monitors energy usage and controls power-saving settings, which are accessed through the new Monitors Dashboard task.
- Monitoring virtual server resources for processor use and delays, with the capability to create a graphical trend report.

Unified Resource Manager supports different levels of system management. These features determine the management functions and operational controls that are available for a zEnterprise mainframe and any stand-alone zBX:

► Manage suite

Provides the Unified Resource Manager function for core operational controls, installation, and energy monitoring. It is configured by default and activated when an ensemble is created.

► Automate/Advanced Management suite

Advanced Management functions for IBM System x blades delivers a workload definition and performance policy monitoring and reporting. The Automate function adds goal-oriented resource monitoring management and energy management for CPC components such as System x blades, POWER7 Blades, and the IBM DataPower XI50z. This function is in addition to the Advanced Management function.

Table 11-3 lists the feature codes that must exist to enable Unified Resource Manager. To get ensemble membership, ensure that you also have FC 0025 for the zBC12.

Restriction: No new features can be ordered for IBM z Unified Resource Manager with IBM z13s servers.

Table 11-3 Unified Resource Manager feature codes and charge indicators

Unified Resource Manager managed component	Manage ^a (per connection)	Automate ^a (per connection)
Base features	FC 0019 ^b - N/C	FC 0020 ^c - N/C
POWER7 blade	FC 0178 ^d - Yes	FC 0179 ^d - Yes
DataPower blade	FC 0184 ^d - Yes	FC 0185 ^d - N/C
IBM System x blades	FC 0182 ^d - Yes	FC 0183 ^d - Yes (also covers the previous Advanced Management function)

- a. Yes = charged feature, N/C = no charge, N/A = not applicable. All components are either managed through the Manage suite or the Automate/Advanced Management suite. The Automate/Advanced Management suite contains the functions of the Managed suite.
- b. FC 0019 is a prerequisite for FC 0020, FC 0184, and FC 0178.
- c. FC 0020 is a prerequisite for FC 0185 and FC 0183.
- d. All these feature codes are now associated to the zBX Model 004.

APIs for the Unified Resource Manager

The API is a web-oriented programming interface that makes the underlying Unified Resource Manager capabilities available for use by higher-level management applications, system automation functions, and custom scripting. The functions that are available through the API support several important usage scenarios. These scenarios are in virtualization management, resource inventory, provisioning, monitoring, automation, workload-based optimization, and so on.

The Web Services API consists of two major components that are accessed by customer applications through Internet Protocol network connections with the HMC.

For more information about the API and the Unified Resource Manager, see *z Systems Hardware Management Console Web Services API (Version 2.13.1)*, SC27-2634, and *IBM zEnterprise Unified Resource Manager*, SG24-7921.

z/VM V6R3 and Unified Resource Manager: Because of the IBM cloud strategy and adoption of OpenStack, the management of z/VM environments in zManager is now stabilized and will not be further enhanced. *zManager will not provide systems management support for z/VM 6.3 and later releases.* However, zManager continues to play a distinct and strategic role in the management of virtualized environments that are created by the integrated firmware hypervisors (PR/SM, PowerVM, and x hypervisor, which is based on KVM) of zEnterprise.

11.6.2 Ensemble definition and management

The ensemble starts with a pair of HMCs that are designated as the primary and alternate HMCs, and are assigned an ensemble identity. The zEnterprise CPCs and zBXs are then added to the ensemble through an explicit action at the primary HMC.

Ensemble Membership Flag

The Ensemble Membership Flag feature, FC 0025, is associated with a HMC when a z13s server is ordered.

The new Create Ensemble task allows the Ensemble Administrator user to create an ensemble that contains CPCs and zBXs (Model 004) as members along with images, workloads, virtual networks, and storage pools.

If a z13s server is entered into an ensemble, the CPC Details task on the SE and the HMC reflects the ensemble name.

The Unified Resource Manager actions for the ensemble are conducted from a single primary HMC. All other HMCs that are connected to the ensemble can run system management tasks (but not ensemble management tasks) for any CPC or zBX Model 004 within the ensemble. The primary HMC also can be used to run system management tasks on CPCs that are not part of the ensemble. These tasks include Load, Activate, and so on.

The ensemble-specific managed objects include the following objects:

- ▶ Ensemble
- ▶ Members
- ▶ Blades
- ▶ BladeCenters
- ▶ Hypervisors
- ▶ Storage resources
- ▶ Virtual servers
- ▶ Workloads

When another HMC accesses an ensemble node's CPC, the HMC can perform the same tasks as though the CPC were not a part of an ensemble. A few of those tasks are extended so that you can configure certain ensemble-specific properties. You can, for example, set the virtual network that is associated with Open Systems Adapters (OSAs) for an LPAR. Showing ensemble-related data in certain tasks is allowed. Generally, if the data affects the operation of the ensemble, the data is read-only on another HMC.

The following tasks show ensemble-related data on another HMC:

- ▶ **Scheduled operations:** Displays ensemble-introduced scheduled operations, but you can only view these scheduled operations.
- ▶ **User role:** Shows ensemble tasks. You can modify and delete those roles.
- ▶ **Event monitoring:** Displays ensemble-related events, but you cannot change or delete the event.

HMC considerations when you use IBM zEnterprise Unified Resource Manager to manage an ensemble

The following considerations are valid when you use Unified Resource Manager to manage an ensemble:

- ▶ All HMCs at the supported code level are eligible to create an ensemble. Only HMCs with FC 0094, FC 0092, or FC 0091 at Driver 27 or later can be primary or alternate HMCs for z13s servers.
- ▶ The primary HMC and the alternate HMC must be the same machine type and feature code.
- ▶ A single HMC pair manages the ensemble that consists of a primary HMC and an alternate HMC.
- ▶ Only one primary HMC manages an ensemble, which can consist of a maximum of eight CPCs and up to eight zBX Model 004 systems.
- ▶ The HMC that ran the Create Ensemble wizard becomes the primary HMC. An alternate HMC is elected and paired with the primary.
- ▶ The Primary HMC (Version 2.13.1 or later) and Alternate HMC (Version 2.13.1 or later) are displayed on the HMC banner. When the ensemble is deleted, the titles change back to the default.
- ▶ A primary HMC is the only HMC that can run ensemble-related management tasks. These tasks include create virtual server, manage virtual networks, and create workload.
- ▶ A zEnterprise ensemble can have a maximum of 16 nodes (eight CPCs plus eight zBX Model 004 systems), and is managed by one primary HMC and its alternate. Each node comprises a zEnterprise CPC or a zBX Model 004.
- ▶ Any HMC can manage up to 100 CPCs. The primary HMC can run all non-ensemble HMC functions on CPCs that are not members of the ensemble.
- ▶ The primary and alternate HMCs must be on the same LAN segment.
- ▶ The alternate HMC's role is to mirror the ensemble configuration and policy information from the primary HMC.
- ▶ When failover happens, the alternate HMC becomes the primary HMC. This behavior is the same as primary and alternate SEs.

11.6.3 HMC availability

The HMC is attached to the same LAN as the server's and zBX Model 004 SEs. This LAN is referred to as the *Customer Managed Management Network*. The HMC communicates with each CPC and with each zBX Model 004 SE.

If the z13s node is defined as a member of an ensemble, the primary HMC is the authoritative controlling (stateful) component for the Unified Resource Manager configuration. It is also the stateful component for policies that have a scope that spans all of the managed CPCs and SEs in the ensemble. The managing HMC has an active role in ongoing system monitoring and adjustment.

This configuration requires the HMC to be configured in a primary/alternate configuration. It also cannot be disconnected from the managed ensemble members.

Failover: The primary HMC and its alternate must be connected to the same LAN segment. This configuration allows the alternate HMC to take over the IP address of the primary HMC during failover processing.

11.6.4 Considerations for multiple HMCs

Customers often deploy multiple HMC instances to manage an overlapping collection of systems. Until the emergence of ensembles, all of the HMCs were peer consoles to the managed systems. Using this configuration, all management actions are possible to any of the reachable systems while logged in to a session on any of the HMCs (subject to access control). With the Unified Resource Manager, this paradigm has changed. One ensemble is managed by one primary and alternate HMC pair. Multiple ensembles require an equal number of primary and alternate HMC pairs to manage them. If a z13s server, a zEnterprise System, or a zBX Model 004 is added to an ensemble, management actions that target that object can be done only from the managing (primary) HMC for that ensemble.

11.6.5 HMC browser session to a primary HMC

A remote HMC browser session to the primary HMC that manages an ensemble allows an user who is logged on to another HMC or a workstation to perform ensemble-related actions.

11.6.6 HMC ensemble topology

The system management functions that pertain to an ensemble use the HMC and the z13s, zEnterprise System, or zBX Model 004 SEs through the internode management network (INMN) to provide the required connectivity.

Figure 11-34 depicts an ensemble with a zEC12, a z13s, and a stand-alone zBX servers that are managed by the Unified Resource Manager in the primary and alternate HMCs.

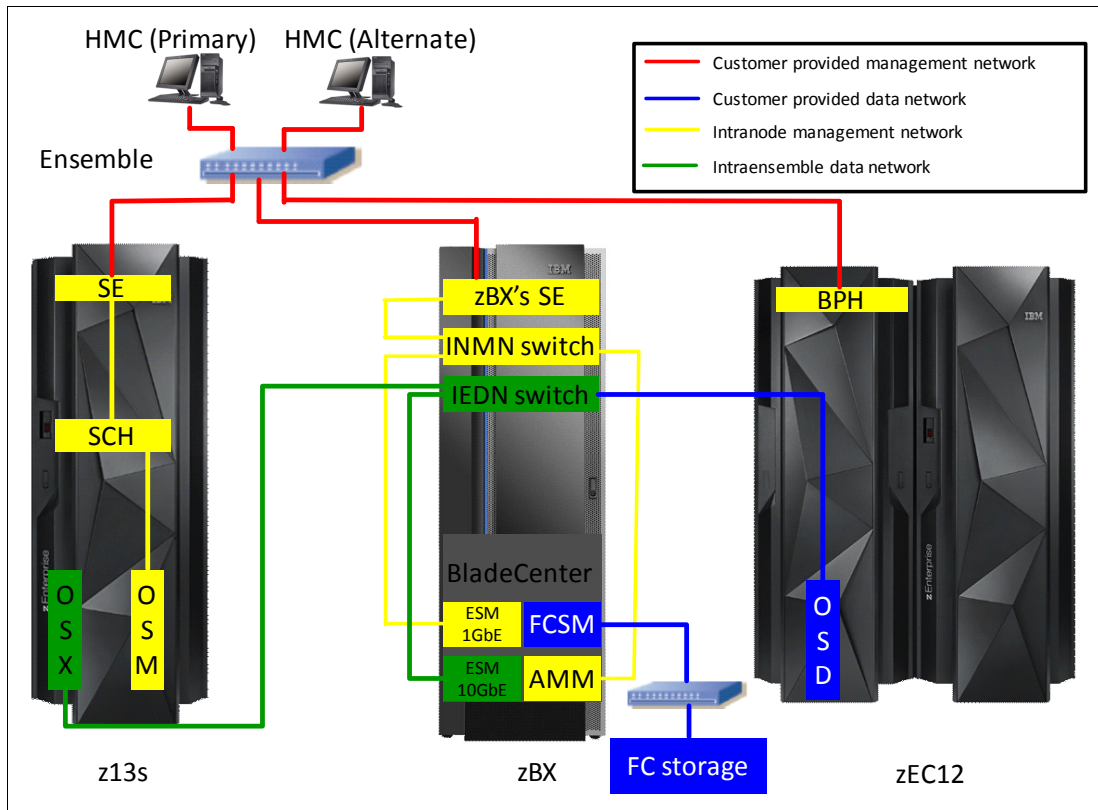


Figure 11-34 Ensemble example with primary and alternate HMCs

For the stand-alone CPC ensemble node (zEC12), an OSA-Express4S (CHPID type OSD) connects to the zBX intraensemble data network (IEDN) switch. The HMCs also communicate with all the components of the ensemble by using the System Control Hub (SCH) in the z13s server, the BPH in the zEC12, and the INMN switch in the zBX Model 004.

The OSA-Express5S (or OSA Express 4S) 10 GbE ports (CHPID type OSX) in the z13s server are plugged with customer-provided 10 GbE cables to the IEDN zBX switch. These cables are either short reach (SR) or long reach (LR), depending on the OSA feature.



Performance

This chapter describes the performance considerations for IBM z13s.

This chapter includes the following sections:

- ▶ Performance information
- ▶ LSPR workload suite
- ▶ Fundamental components of workload capacity performance
- ▶ Relative nest intensity
- ▶ LSPR workload categories based on relative nest intensity
- ▶ Relating production workloads to LSPR workloads
- ▶ Workload performance variation

12.1 Performance information

The IBM z13s Model Z06 is designed to offer approximately 51% more capacity and 8 times the amount of memory than the IBM zEnterprise BC12 (zBC12) Model Z06 system. Uniprocessor performance also has increased. A z13s Model Z01 offers, on average, performance improvements of 34% over the zBC12 Model Z01.

Figure 12-1 shows the estimated capacity ratios for z13s, zBC12, z114, z10 BC, and z9 BC.

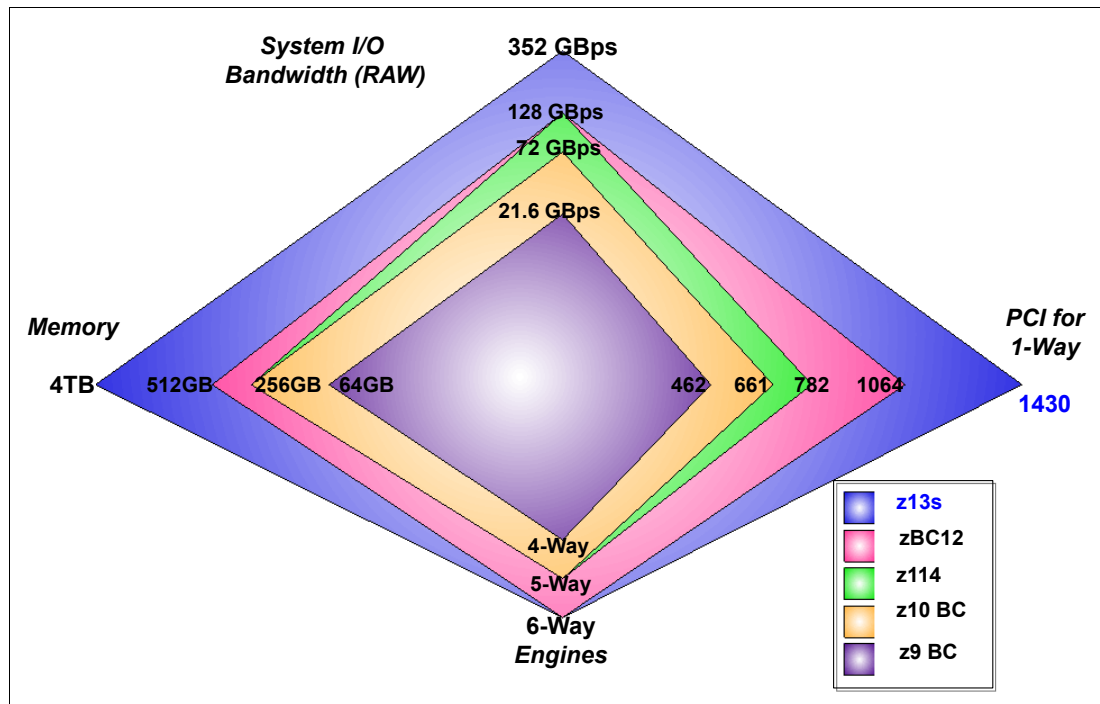


Figure 12-1 z13s to zBC12, z114, z10 BC, and z9 BC performance comparison

The Large System Performance Reference (LSPR) numbers that are given for z13s were obtained with z/OS V2R1. The numbers for zBC12 were obtained with z/OS V1R13. The numbers for the z114, z10 BC, and z9 BC systems were obtained with the z/OS V1R11 operating system.

On average, the z13s can deliver up to 51% more performance in a 6-way configuration than a zBC12 6-way. However, variations on the observed performance increase depend on the workload type.

Consult the LSPR when you consider performance on the zBC12. The performance ratings across the individual LSPR workloads are likely to have a large spread. More performance variation of individual logical partitions (LPARs) exists because the fluctuating resource requirements of other partitions can be more pronounced with the increased number of partitions and the availability of more processor units (PUs). For more information, see 12.7, "Workload performance variation" on page 445.

For detailed performance information, see the LSPR website at:

<https://www.ibm.com/servers/resourceLink/lib03060.nsf/pages/lsprindex>

The millions of service units (MSU) ratings are available from the following website:

<http://www.ibm.com/systems/z/resources/swprice/reference/exhibits/>

12.2 LSPR workload suite

Historically, LSPR capacity tables, including pure workloads and mixes, have been identified with application names or a *software* characteristic such as:

- ▶ CICS
- ▶ Information Management System (IMS)
- ▶ Traditional online transaction processing workload (formerly known as IMS) (OLTP-T)
- ▶ Commercial batch with long-running jobs (CB-L)
- ▶ Low I/O Content Mix Workload (LoIO-mix)
- ▶ Transaction Intensive Mix Workload (TI-mix)

However, capacity performance is more closely associated with how a workload uses and interacts with a particular processor *hardware* design. The CPU Measurement Facility (CPU MF) data that was introduced on the z10 provides insight into the interaction of workload and *hardware design* in production workloads. CPU MF data helps LSPR to adjust workload capacity curves based on the underlying hardware sensitivities, in particular, the processor access to caches and memory. This is known as *nest activity intensity*. Using this data, LSPR introduces three new workload capacity categories that replace all prior primitives and mixes.

LSPR contains the internal throughput rate ratios (ITRRs) for the zBC12 and the previous generation processor families. These ratios are based on measurements and projections that use standard IBM benchmarks in a controlled environment. The throughput that any user experiences can vary depending on the amount of multiprogramming in the user's job stream, the I/O configuration, and the workload processed. Therefore, no assurance can be given that an individual user can achieve throughput improvements equivalent to the stated performance ratios.

12.3 Fundamental components of workload capacity performance

Workload capacity performance is sensitive to three major factors:

- ▶ Instruction path length
- ▶ Instruction complexity
- ▶ Memory hierarchy and memory nest

This section examines each of these three factors.

12.3.1 Instruction path length

A transaction or job runs a set of instructions to complete its task. These instructions are composed of various paths through the operating system, subsystems, and application. The total count of instructions that are run across these software components is referred to as the *transaction or job path length*. The path length varies for each transaction or job, and depends on the complexity of the tasks that must be run. For a particular transaction or job, the application path length tends to stay the same, assuming that the transaction or job is asked to run the same task each time.

However, the path length that is associated with the operating system or subsystem might vary based on a number of factors:

- ▶ Competition with other tasks in the system for shared resources. As the total number of tasks grows, more instructions are needed to manage the resources.
- ▶ The *n*-way (number of logical processors) of the image or LPAR. As the number of logical processors grows, more instructions are needed to manage resources that are serialized by latches and locks.

12.3.2 Instruction complexity

The type of instructions and the sequence in which they are run interacts with the design of a microprocessor to affect a performance component. This factor is defined as *instruction complexity*. Many design alternatives affect this component:

- ▶ Cycle time (GHz)
- ▶ Instruction architecture
- ▶ Pipeline
- ▶ Superscalar
- ▶ Out-of-order execution
- ▶ Branch prediction

As workloads are moved between microprocessors with various designs, performance varies. However, when on a processor, this component tends to be similar across all models of that processor.

12.3.3 Memory hierarchy and memory nest

The *memory hierarchy* of a processor generally refers to the caches, data buses, and memory arrays that stage the instructions and data that must be run on the microprocessor to complete a transaction or job.

Many design choices affect this component:

- ▶ Cache size
- ▶ Latencies (sensitive to distance from the microprocessor)
- ▶ Number of levels, the Modified, Exclusive, Shared, Invalid (MESI) protocol, controllers, switches, the number and bandwidth of data buses, and others

Certain caches are *private* to the microprocessor core, which means that only that microprocessor core can access them. Other caches are shared by multiple microprocessor cores. The term *memory nest* for a z Systems processor refers to the shared caches and memory along with the data buses that interconnect them.

Figure 12-2 shows a memory nest in a z13s single CPC drawer system with two nodes.

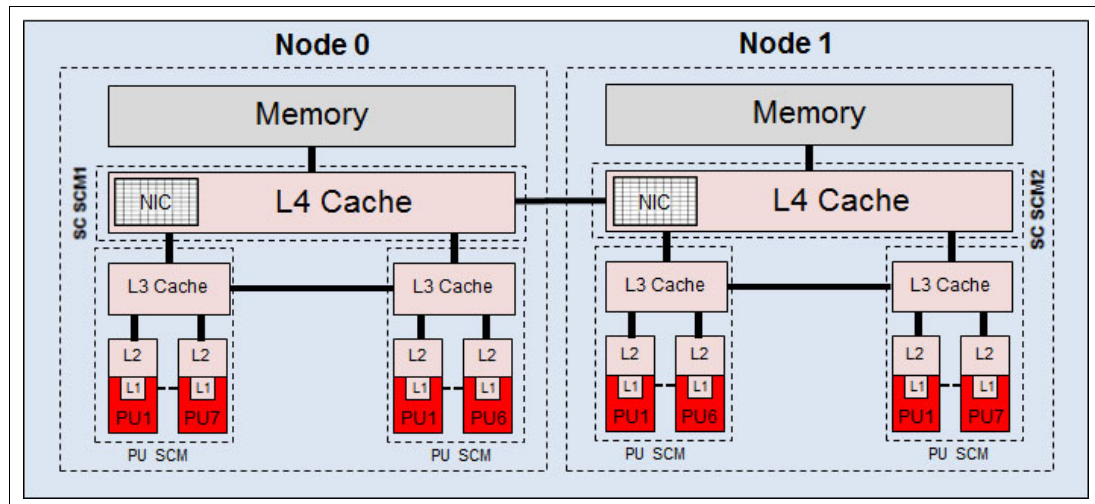


Figure 12-2 Memory hierarchy on the z13s one CPC drawer system (two nodes)

Workload capacity performance is sensitive to how deep into the memory hierarchy the processor must go to retrieve the workload instructions and data for running. The best performance occurs when the instructions and data are in the caches nearest the processor. In this configuration, little time is spent waiting before the task runs. If the instructions and data must be retrieved from farther out in the hierarchy, the processor spends more time waiting for their arrival.

As workloads are moved between processors with various memory hierarchy designs, performance varies because the average time to retrieve instructions and data from within the memory hierarchy varies. Additionally, when on a processor, this component continues to vary. This variation is because the location of a workload's instructions and data within the memory hierarchy is affected by many factors that include, but are not limited to, these factors:

- ▶ Locality of reference
- ▶ I/O rate
- ▶ Competition from other applications and LPARs

12.4 Relative nest intensity

The most performance-sensitive area of the memory hierarchy is the activity to the memory nest, which is the distribution of activity to the shared caches and memory. The term Relative Nest Intensity (RNI) indicates the level of activity to this part of the memory hierarchy. Using data from CPU MF, the RNI of the workload running in an LPAR can be calculated. The higher the RNI, the deeper into the memory hierarchy the processor must go to retrieve the instructions and data for that workload.

RNI reflects the distribution and latency of sourcing data from shared caches and memory, as shown in Figure 12-3.

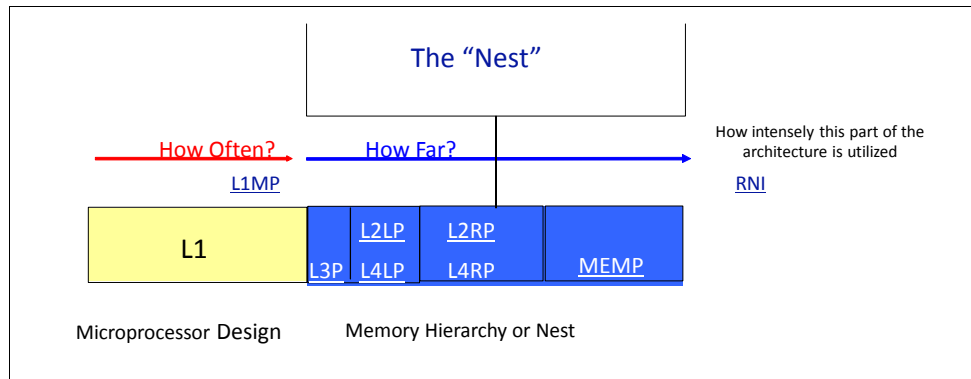


Figure 12-3 Relative Nest Intensity

Many factors influence the performance of a workload. However, usually what these factors are influencing is the RNI of the workload. The interaction of all these factors results in a net RNI for the workload, which in turn directly relates to the performance of the workload.

These factors are tendencies rather than absolutes. For example, a workload might have a low I/O rate, intensive processor use, and a high locality of reference, which all suggest a low RNI. But it might be competing with many other applications within the same LPAR and many other LPARs on the processor, which tend to create a higher RNI. It is the net effect of the interaction of all these factors that determines the RNI.

The traditional factors that were used to categorize workloads in the past are listed along with their RNI tendency in Figure 12-4.

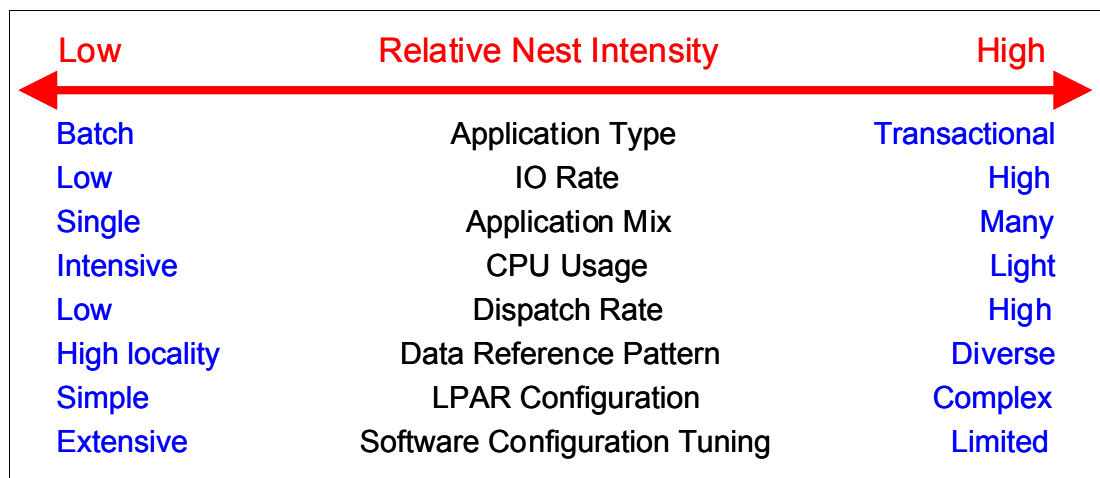


Figure 12-4 The traditional factors that were used to categorize workloads

Little can be done to affect most of these factors. An application type is whatever is necessary to do the job. The data reference pattern and processor usage tend to be inherent to the nature of the application. The LPAR configuration and application mix are mostly a function of what must be supported on a system. The I/O rate can be influenced somewhat through buffer pool tuning.

However, one factor, *software configuration tuning*, is often overlooked but can have a direct effect on RNI. This term refers to the number of address spaces (such as CICS

application-owning regions (AORs) or batch initiators) that are needed to support a workload. This factor always has existed, but its sensitivity is higher with the current high frequency microprocessors. Spreading the same workload over more address spaces than necessary can raise a workload's RNI. This increase occurs because the working set of instructions and data from each address space increases the competition for the processor caches.

Tuning to reduce the number of simultaneously active address spaces to the correct number that is needed to support a workload can reduce RNI and improve performance. In the LSPR, the number of address spaces for each processor type and *n*-way configuration is tuned to be consistent with what is needed to support the workload. Therefore, the LSPR workload capacity ratios reflect a presumed level of software configuration tuning. Retuning the software configuration of a production workload as it moves to a larger or faster processor might be needed to achieve the published LSPR ratios.

12.5 LSPR workload categories based on relative nest intensity

A workload's RNI is the most influential factor in determining workload performance. Other more traditional factors, such as application type or I/O rate, have RNI tendencies. However, it is the net RNI of the workload that is the underlying factor in determining the workload's capacity performance. The LSPR now runs various combinations of former workload primitives, such as CICS, DB2, IMS, OSAM, VSAM, WebSphere, COBOL, and utilities, to produce capacity curves that span the typical range of RNI.

Three new workload categories are represented in the LSPR tables:

▶ *LOW* (relative nest intensity)

A workload category that represents light use of the memory hierarchy. This category is similar to past high-scaling primitives.

▶ *AVERAGE* (relative nest intensity)

A workload category that represents average use of the memory hierarchy. This category is similar to the past LoIO-mix workload, and is expected to represent most production workloads.

▶ *HIGH* (relative nest intensity)

A workload category that represents a heavy use of the memory hierarchy. This category is similar to the past TI-mix workload.

These categories are based on the RNI. The RNI is influenced by many variables, such as application type, I/O rate, application mix, processor usage, data reference patterns, LPAR configuration, and the software configuration that is running. CPU MF data can be collected by z/OS System Measurement Facility on SMF 113 records.

12.6 Relating production workloads to LSPR workloads

Historically, a number of techniques were used to match production workloads to LSPR workloads:

- ▶ Application name (a client running CICS can use the CICS LSPR workload)
- ▶ Application type (create a mix of the LSPR online and batch workloads)
- ▶ I/O rate (the low I/O rates used a mix of low I/O rate LSPR workloads)

The previous LSPR workload suite was composed of the following workloads:

- ▶ OLTP-T (formerly known as IMS)
- ▶ Web-enabled online transaction processing workload (OLTP-W, also known as Web/CICS/DB2)
- ▶ A heavy Java based online stock trading application (WASDB, previously referred to as Trade2-EJB)
- ▶ Batch processing, represented by the CB-L (also called CBW2)
- ▶ A new ODE-B Java batch workload, replacing the CB-J workload

The traditional Commercial Batch Short Job Steps (CB-S) workload (formerly CB84) was dropped. Figure 12-4 on page 442 shows the traditional factors that have been used to categorize workloads.

The previous LSPR provided performance ratios for individual workloads and for the default mixed workload. This default workload was composed of equal amounts of four of the previous workloads (OLTP-T, OLTP-W, WASDB, and CB-L). Guidance in converting the previous LSPR categories to the new ones is given in Figure 12-5. The IBM Processor Capacity Reference for z Systems (zPCR) tool¹ has been changed to support the new z/OS workload categories.

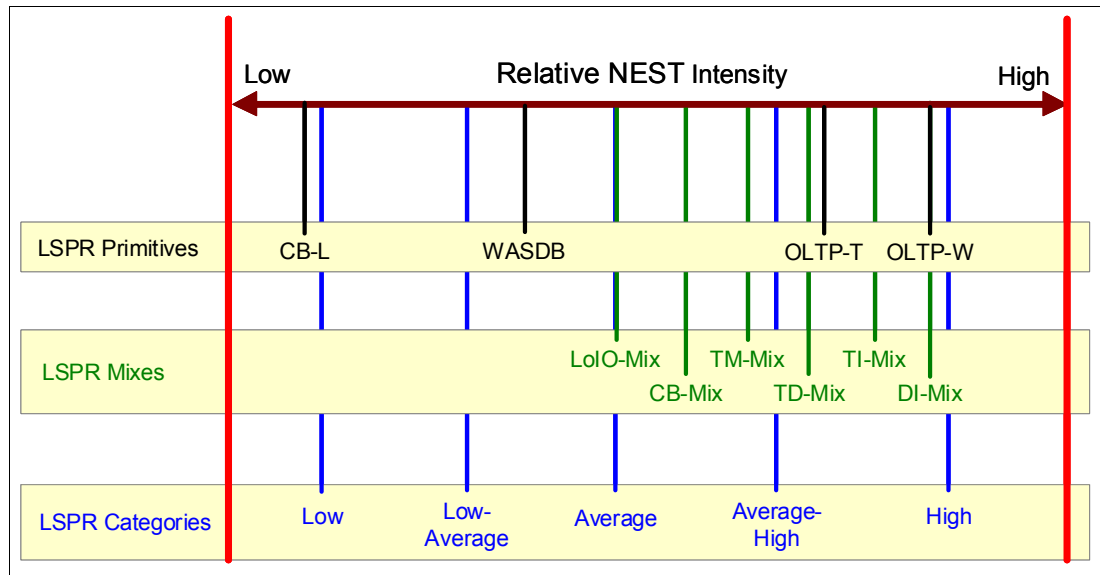


Figure 12-5 New z/OS workload categories defined

However, as addressed in 12.5, “LSPR workload categories based on relative nest intensity” on page 443, the underlying performance sensitive factor is how a workload interacts with the

¹ The IBM zPCR tool reflects the latest IBM LSPR measurements. It is available at no extra charge at <http://www.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/PRS1381>.

processor hardware. These past techniques approximated the hardware characteristics that were not available through software performance reporting tools.

Beginning with the z10 processor, the hardware characteristics can now be measured by using CPU MF (SMF 113) counters data. A production workload can now be matched to an LSPR workload category through these hardware characteristics. For more information about RNI, see 12.5, “LSPR workload categories based on relative nest intensity” on page 443.

The AVERAGE RNI LSPR workload is intended to match most client workloads. When no other data is available, use it for capacity analysis.

Direct access storage device (DASD) I/O rate was used for many years to separate workloads into two categories:

- ▶ Those whose DASD I/O per MSU (adjusted) is <30 (or DASD I/O per Peripheral Component Interconnect (PCI) is <5)
- ▶ Those higher than these values

Most production workloads fell into the “low I/O” category, and a LoIO-mix workload was used to represent them. Using the same I/O test, these workloads now use the AVERAGE RNI LSPR workload. Workloads with higher I/O rates can use the HIGH RNI workload or the AVG-HIGH RNI workload that is included with IBM zPCR. Low-Average and Average-High categories allow better granularity for workload characterization.

For z10 and newer processors, the CPU MF data can be used to provide an extra hint as to workload selection. When available, this data allows the RNI for a production workload to be calculated. By using the RNI and another factor from CPU MF, the L1MP (percentage of data and instruction references that miss the L1 cache), a workload can be classified as LOW, AVERAGE, or HIGH RNI. This classification and resulting hit are automated in the zPCR tool. It is preferable to use zPCR for capacity sizing.

12.7 Workload performance variation

Performance variability from application to application is expected. This variation is similar to that seen on the zBC12 and z114. This variability can be observed in certain ways. The range of performance ratings across the individual workloads is likely to have a spread, but not as large as with the z10 BC.

The memory and cache designs affect various workloads in many ways. All workloads are improved, with cache-intensive loads benefiting the most. When comparing moving from z9 BC to z10 BC with moving from z10 BC to z114 or from z114 to zBC12, it is likely that the relative benefits per workload will vary. Those workloads that benefited more than the average when moving from z9 BC to z10 BC will benefit less than the average when moving from z10 BC to z114. Nevertheless, the workload variability for moving from zBC12 to z13s is expected to be less than the last few upgrades.

The effect of this variability is increased deviations of workloads from single-number metric-based factors, such as millions of instructions per second (MIPS), MSUs, and CPU time charge-back algorithms.

Experience demonstrates that z Systems servers can be run at up to 100% utilization levels, sustained. However, most clients prefer to leave a bit of room and run at 90% or slightly under. For any capacity comparison exercise, using a single metric, such as MIPS or MSU, is not a valid method. When deciding the number of processors and the uniprocessor capacity,

remember both the workload characteristics and LPAR configuration. For these reasons, when you plan capacity, zPCR and involving IBM technical support are recommended.

12.7.1 Main performance improvement drivers with z13s

The z13s is designed to deliver new levels of performance and capacity for large-scale consolidation and growth. The following attributes and design points of the z13s contribute to overall performance and throughput improvements as compared to the zBC12.

The z/Architecture implementation has the following enhancements:

- ▶ Transactional Execution (TX) designed for z/OS, Java, DB2, and other users
- ▶ Runtime Instrumentation (RI) provides dynamic and self-tuning online recompilation capability for Java workloads
- ▶ Enhanced DAT-2 for supporting 2-GB pages for DB2 buffer pools, Java heap size, and other large structures
- ▶ Software directives implementation to improve hardware performance
- ▶ Decimal format conversions for COBOL programs

The z13s microprocessor design has the following enhancements:

- ▶ Six or Seven active processor cores per chip
- ▶ Improved out-of-order (OOO) execution design
- ▶ Improved pipeline balance, with up to six instructions that can be decoded per cycle, and up to 10 instructions/operations that can be initiated to run per clock cycle
- ▶ Simultaneous multithreading
- ▶ Single-instruction multiple-data (SIMD) unit and 139 new instructions for vector operations
- ▶ Enhanced branch prediction latency and instruction fetch throughput
- ▶ Improvements in execution bandwidth and throughput: 10 execution units and two load/store units, which are divided into two symmetric pipelines:
 - Four fixed-point units (FXUs) (integer)
 - Two load/store units (LSUs)
 - Two binary floating-point units (BFUs)
 - Two binary coded decimal floating-point units (DFUs)
 - Two vector execution units (VXUs)
- ▶ Redesigned cache structure:
 - Increased L1I and L1D caches (96 KB instruction and 128 KB data per core)
 - Increased 2 MB + 2 MB eDRAM split (instruction and data) private L2 cache per core
 - On chip 64 MB eDRAM L3 Cache, shared by all cores (six or seven) - 256 MB per CPC drawer (two nodes)
 - New Inclusive L4 Design: 480 MB L4 with 224 MB NIC Directory (960 MB L4 per CPC drawer) (two nodes)
- ▶ One cryptographic/compression co-processor per core, redesigned
- ▶ CPACF (hardware) runs additional UTF conversion operations: UTF8 to UTF32, UTF8 to UTF16, UTF32 to UTF8, and UTF32 to UTF16

- ▶ Clock frequency at 4.3 GHz
- ▶ IBM CMOS 14S0 22 nm SOI technology with IBM eDRAM technology

The z13s design has the following enhancements as compared with the zBC12:

- ▶ Increased total number of PUs that are available on the system, from 18 to 26, and number of characterizable cores, from 13 to 20. There is a maximum of 6 CPs.
- ▶ Hardware system area (HSA) increased from 16 GB to 40 GB
- ▶ Increased number of supported LPARs from 30 to 40
- ▶ 2 TB of addressable memory (configurable to LPARs) with up to 2 TB of memory per LPAR
- ▶ Increased default number of SAP processors up to 3 on a Model N20
- ▶ New Coupling Facility Control Code (CFCC) that is available for improved performance:
 - Elapsed time improvements when dynamically altering the size of a cache structure
 - DB2 conditional writes to a group buffer pool (GBP)
 - Performance improvements for coupling facility cache structures to avoid flooding the coupling facility cache with changed data, and avoid excessive delays and backlogs for cast-out processing
 - Performance throughput enhancements for parallel cache castout processing by extending the number of record code check (RCC) cursors beyond 512
 - Coupling facility (CF) storage class and castout class contention avoidance by breaking up individual storage class and castout class queues to reduce storage class and castout class latch contention

The following new features are available on the z13s:

- ▶ Integrated Coupling Adapter (ICA SR)
- ▶ FICON Express16S
- ▶ Shared Memory Communications over RDMA (SMC-R). The 10GbE Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE) Express (10GbE RoCE Express) feature now supports using both physical ports and can be shared between up to 31 LPARs.
- ▶ Shared Memory Communications-Direct Memory Access (SMC-D) over Internal Shared Memory (ISM)
- ▶ Crypto Express5S with up to 40 domains



A

IBM z Appliance Container Infrastructure

This appendix introduces the IBM z Appliance Container Infrastructure (zACI) framework available on z13s and z13 Driver Level 27 servers. It briefly describes the reason why IBM has created the framework and provides a description about how the zACI environment is intended to be used.

This appendix includes the following sections:

- ▶ What is zACI?
- ▶ Why use zACI?
- ▶ IBM z Systems servers and zACI

A.1 What is zACI?

An appliance is an integration of operating system, middleware, and software components that work autonomously and provide core services and infrastructure that focus on consumability and security.

Appliances can be implemented as firmware or software, depending on the environment in which the appliance runs and the function it must provide.

Introduced with z13 and z13s Driver Level 27 servers, the common framework called zACI is intended to provide a standard set of behaviors and operations that help simplify deploying infrastructure functions.

A.2 Why use zACI?

Frameworks like zACI are embedded technologies. They are not product deliverables themselves. They are only delivered as part of the solution that uses them.

The zACI framework reduces the amount of work a team must do to create an appliance, and enforces a common set of behaviors for operations that all appliances adhere to.

The zACI framework provides a set of utilities to implement common functions that all appliances need, such as first-failure data capture (FFDC), network setup, appliance configuration, and so on.

The framework enables you to release a product as software or firmware based on a business decision, not on a technical decision.

A.3 IBM z Systems servers and zACI

An appliance that is based on the zACI framework has the following features:

- ▶ Encapsulated Operating Systems
- ▶ Remote APIs (RESTful) and web interfaces
- ▶ Embedded monitoring and self-healing
- ▶ Tamper-protection
- ▶ Protected IP

An appliance that is built based on the zACI framework has these advantages:

- ▶ Tested and qualified by IBM for a specific use case
- ▶ Can be delivered either as firmware, platform, or software

Figure A-1 shows the basic outline of the zACI framework.

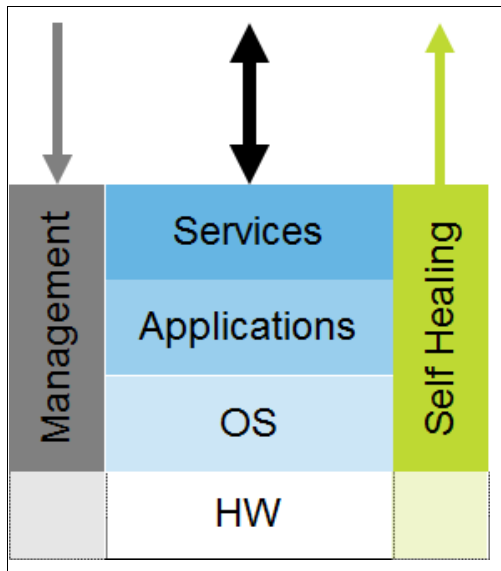


Figure A-1 zACI Framework basic outline

Statement of Direction: IBM plans to make available the z/VSE Network Appliance based on zACI in June 2016.

All statements regarding IBM plans, directions, and intent are subject to change or withdrawal without notice. Any reliance on these statements of general direction is at the relying party's sole risk and will not create liability or obligation for IBM.

12.7.2 Example: Deploying IBM zAware

The new type of LPAR called *zACI* will be used to deploy the IBM zAware server application. The IBM zAware logical partition (LPAR) type is not supported on z13s or z13 at Driver level 27 servers. The zACI LPAR type is used instead.

Figure A-2 shows the new LPAR type.

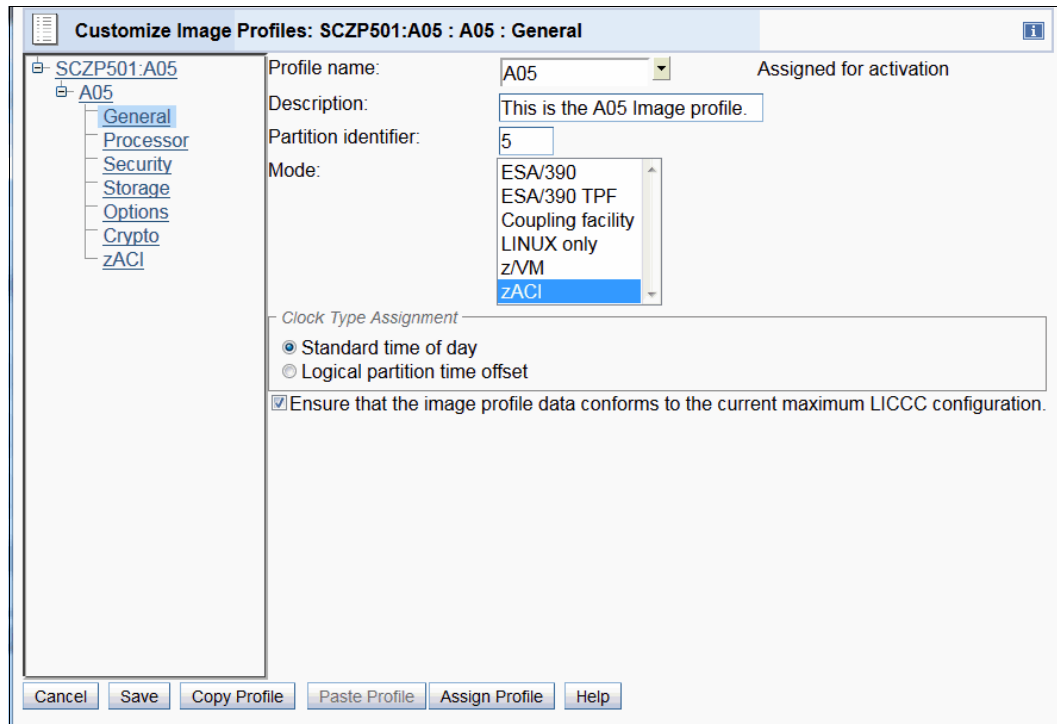


Figure A-2 IBM zAware Image Profile based on zACI

Existing IBM zAware LPARs will be automatically converted during Enhanced Driver Maintenance from Driver 22 to Driver 27. No reconfiguration of IBM zAware is required. A new icon will identify the zACI LPARs in the HMC user interface (UI), as shown in Figure A-3 (right - Classic UI, left - tree style UI).

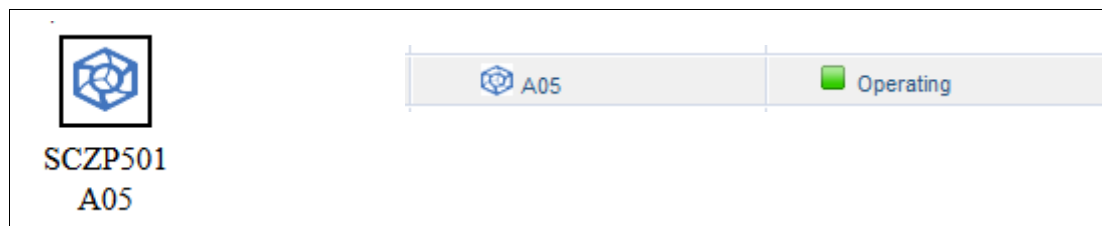


Figure A-3 zACI icon in the HMC interface



IBM z Systems Advanced Workload Analysis Reporter (IBM zAware)

This appendix introduces IBM z Systems Advanced Workload Analysis Reporter (IBM zAware), which was first available with IBM zEnterprise zEC12. This feature is designed to offer near real-time, continuous learning, diagnostics and monitoring capabilities. IBM zAware helps you pinpoint and resolve potential problems quickly enough to minimize impacts to your business.

This appendix includes the following sections:

- ▶ Troubleshooting in complex IT environments
- ▶ Introducing IBM zAware
- ▶ Understanding IBM zAware technology
- ▶ IBM zAware prerequisites
- ▶ Configuring and using IBM zAware virtual appliance

For more information about IBM zAware, see *Extending z/OS System Management Functions with IBM zAware*, SG24-8070, and *z Advanced Workload Analysis Reporter (IBM zAware) Guide Version 2.0*, SC27-2632.

B.1 Troubleshooting in complex IT environments

In a 24x7 operating environment, a system problem or incident can drive up operations costs and disrupt service to clients for hours or even days. Current IT environments cannot afford recurring problems or outages that take too long to repair. These outages can result in damage to a company's reputation and limit the company's ability to remain competitive in the marketplace.

However, as systems become more complex, errors can occur anywhere. Some problems begin with symptoms that can go undetected for long periods of time. Systems often experience "soft failures" (sick but not dead) that are much more difficult to detect. Moreover, problems can grow, cascade, and get out of control.

The following everyday activities can introduce system anomalies and trigger either hard or soft failures in complex, integrated data centers:

- ▶ Increased volume of business activity
- ▶ Application modifications to comply with changing regulatory requirements
- ▶ IT efficiency efforts, such as consolidating images
- ▶ Standard operational changes:
 - Adding or upgrading hardware
 - Adding or upgrading software, such as operating systems, middleware, and independent software vendor products
 - Modifying network configurations
 - Moving workloads (provisioning, balancing, deploying, disaster recovery (DR) testing, and so on)

Using a combination of existing system management tools helps to diagnose problems. However, they cannot quickly identify messages that precede system problems and cannot detect every possible combination of change and failure.

When using these tools, you might need to look through message logs to understand the underlying issue. But the number of messages makes this process a challenging and skills-intensive task, and also error-prone.

To meet IT service challenges and to effectively sustain high levels of availability, a proven way is needed to identify, isolate, and resolve system problems quickly. Information and insight are vital to understanding baseline system behavior along with possible deviations. Having this knowledge reduces the time that is needed to diagnose problems, and address them quickly and accurately.

The current complex, integrated data centers require a team of experts to monitor systems and perform the real-time diagnosis of events. However, it is not always possible to afford this level of skill for these reasons:

- ▶ A z/OS sysplex might produce more than 40 GB of message traffic per day for its images and components alone. Application messages can significantly increase that number.
- ▶ There are more than 40,000 unique message IDs defined in z/OS and the IBM software that runs on z/OS. Independent software vendor (ISV) or client messages can increase that number.

B.2 Introducing IBM zAware

IBM zAware Version 2 is an integrated expert solution that contains sophisticated analytics, IBM insight into the problem domain, and web-browser-based visualization. It is an adaptive analytics solution that learns your unique system characteristics and helps you to detect and diagnose unusual behavior of z/OS images in near real time, accurately and rapidly.

IBM zAware introduces a new generation of technology with improved analytics to provide better results. It can process message streams that do not have message IDs, which makes it possible to handle a broader variety of unstructured data.

IBM zAware Version 2 delivered on z13s and z13 at Driver 27 servers provides the following capabilities:

- ▶ Support for Linux on z Systems message log analysis
 - ▶ Support for native or guest Linux on z Systems images
 - ▶ The ability to group multiple systems that have similar operational characteristics for modeling and analysis
 - ▶ Recognition of dynamic activation and deactivation of a Linux image into a group and appropriate modeling and analysis:
 - Aggregated Sysplex view for z/OS and system views
 - User-defined grouping
- For Linux on z Systems, the user can group multiple systems' data into a combined model by workload (one for all web servers, one for all databases, and so on), by "solution" (for example, one model for your cloud), and by VM host.
- ▶ Heat map display that provides a consolidated, aggregated, and higher-level view, with the ability to drill down to detail views
 - ▶ Improved usability and GUI functional enhancements that address customer requirements
 - ▶ Enhanced filtering and visualization, with better use of GUI space
 - ▶ Improved UI navigation
 - ▶ Display of local time in addition to Coordinated Universal Time
 - ▶ Enhancements based on IBM One UI guidelines
 - ▶ Enhanced analytics
 - ▶ More robust data stores
 - ▶ Expanded browser support with Mozilla Firefox 31 and Internet Explorer 9, 10, and 11

IBM zAware is designed to use near real-time continuous learning algorithms, providing a diagnostics capability that is intended to help you pinpoint problems, which in turn can lead to better availability and a more efficient system. IBM zAware uses analytics to intelligently examine z/OS or Linux on z Systems messages to find unusual patterns, inconsistencies, and variations.

Large operating system environments can sometimes generate more than 25 million messages per day. This large volume of messages can make manual analysis time-consuming and error-prone when exceptional problems occur. IBM zAware provides a simple GUI and APIs to help you find message anomalies quickly, which can help speed problem resolution when seconds count.

IBM zAware and Tivoli® Service Management can be integrated by using the IBM zAware API to provide the following capabilities:

- ▶ Provide visibility into IBM zAware anomalies by using Event Management
- ▶ Improve mean time to repair (MTTR) through integration with existing problem determination and performance monitoring tools
- ▶ Identify system errors and eliminate subsequent occurrences through automation and more sophisticated analysis

B.2.1 Hardware requirements overview

IBM zAware runs on a client-visible logical partition (LPAR) with the following resources:

- ▶ LPAR type: zACI
- ▶ Shared or dedicated Open Systems Adapter (OSA) port
- ▶ Shared or dedicated Integrated Facilities for Linux (IFLs) or central processors (CPs)
- ▶ Storage and memory

For more information, see B.4.2, “IBM zAware operating requirements” on page 467.

Figure B-1 shows how IBM zAware complements an existing environment.

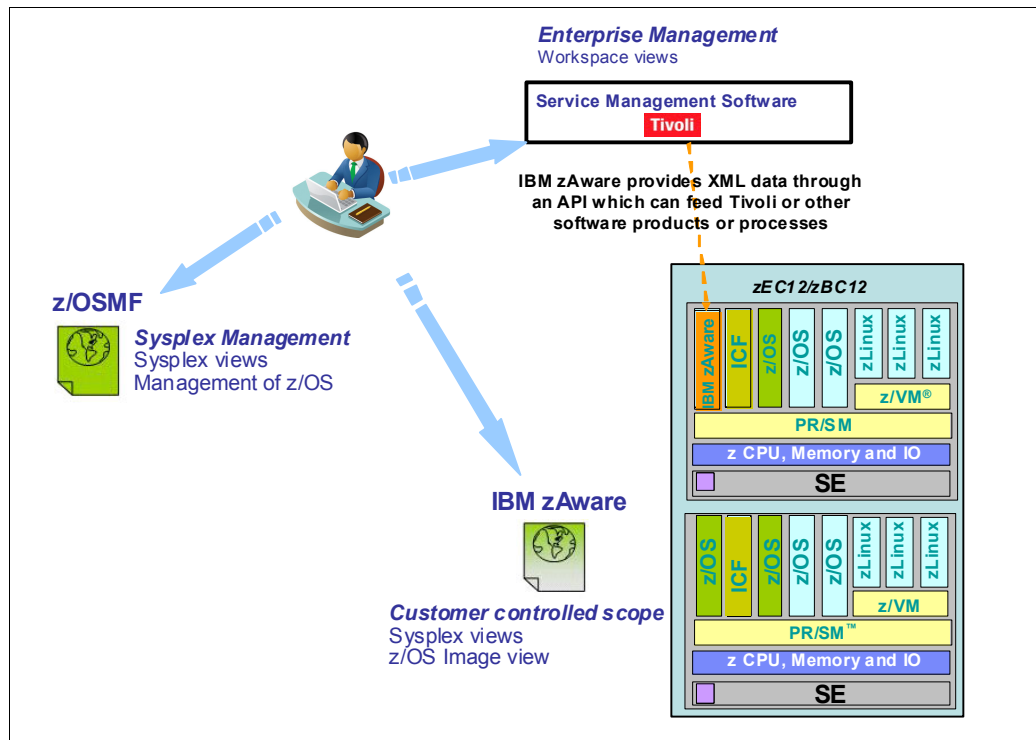


Figure B-1 IBM zAware complements an existing environment

B.2.2 Value of IBM zAware

Early detection and focused diagnosis can help improve the time that is needed to recover from complex z/OS problems. These problems can be cross sysplex, across a set of z Systems servers, and beyond central processing complex (CPC) boundaries. IBM zAware is enhanced to identify unusual system behavior of Linux on z Systems images running natively or as a guest on z/VM.

IBM zAware delivers sophisticated detection and diagnostic capabilities that identify when and where to look for a problem. The cause of the anomalies can be difficult to determine. High-speed analytics on large quantities of log data reduces the problem determination and isolation efforts, time to repair, and impact to service levels. They also provide system awareness for more effective monitoring.

Figure B-2 depicts how IBM zAware shortens the business impact of a problem.

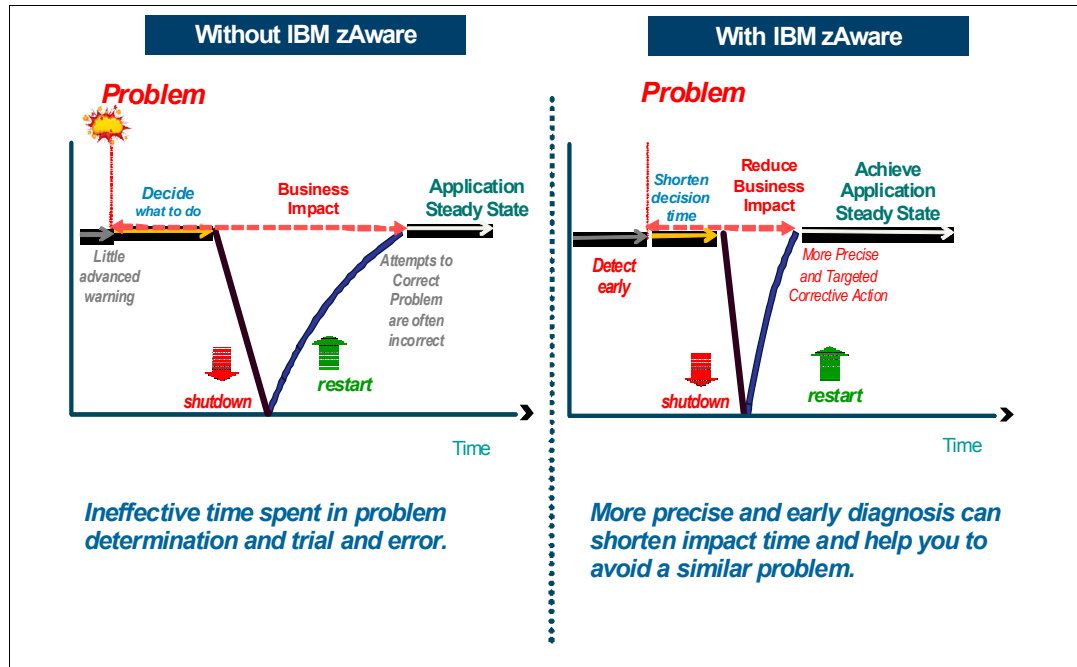


Figure B-2 IBM zAware shortens the business impact of a problem

The IBM zAware GUI also provides quick drill-down capabilities. You can view analytical data that indicates which system is experiencing deviations in behavior, when the anomaly occurred, and whether the message was issued out of context. The IBM zAware GUI fits into existing monitoring structure and can also feed other processes or tools so that they can take corrective action for faster problem resolution.

B.2.3 IBM z/OS Solutions to improve problem diagnostic procedures

Table B-1 shows how IBM zAware can be combined with other monitoring and problem determination tools.

Table B-1 Using IBM zAware with other monitoring and problem determination tools

Solution	Available functions	Rules based	Analytics/ Statistical model	Examines message traffic	Self-learning	Method
z/OS Health Checker ^a	<ul style="list-style-type: none"> ▶ Checks configurations ▶ Programmatic, applies to IBM and ISV tools ▶ Can escalate notifications 	Yes				Rules based to screen for conditions
z/OS PFA ^a	<ul style="list-style-type: none"> ▶ Trending analysis of z/OS system resources and performance ▶ Can start z/OS Runtime Diagnostics 		Yes		Yes	Early detection

Solution	Available functions	Rules based	Analytics/ Statistical model	Examines message traffic	Self-learning	Method
z/OS RTD ^a	<ul style="list-style-type: none"> ▶ Real-time diagnostics of specific z/OS system issues 	Yes		Yes		Rules based after an incident
Linux on z Systems health checker ^b	<ul style="list-style-type: none"> ▶ Checks configurations ▶ Programmatic, applies to IBM and ISV tools 	Yes				Rules based to screen for conditions
IBM zAware	<ul style="list-style-type: none"> ▶ Pattern-based message analysis ▶ Self-learning ▶ Aids in diagnosing complex z/OS problems, including cross sysplex and problems that might bring the system down 		Yes	Yes	Yes	Diagnosis before or after an incident

a. Included in z/OS.

b. Installable as Red Hat Package Manager (RPM).

You can use IBM zAware along with problem diagnosis solutions that are included in z/OS with any large and complex z/OS installation with mission-critical applications and middleware.

Notes:

- ▶ IBM zAware uniquely analyzes messages in context to determine unusual behaviors.
- ▶ IBM zAware uniquely understands and tunes its baseline to compare against your current activity.
- ▶ IBM zAware does not depend on other solutions or manual coding of rules, and is always enabled to watch your system.

B.3 Understanding IBM zAware technology

IBM zAware creates a base model of system behavior by accessing prior system log (SYSLOG) and applying mathematical modeling to these logs. The base model is then used to compare with current SYSLOG for the monitored z/OS, and Linux on z Systems images. This process helps in detecting unusual message pattern that might pinpoint deviations from normal system behavior, which improves real-time diagnostics. IBM zAware automatically manages the creation of the behavioral model that is used to compare current message log data from the connected z/OS and Linux on z Systems instances.

Figure B-3 depicts the basic components of an IBM zAware environment.

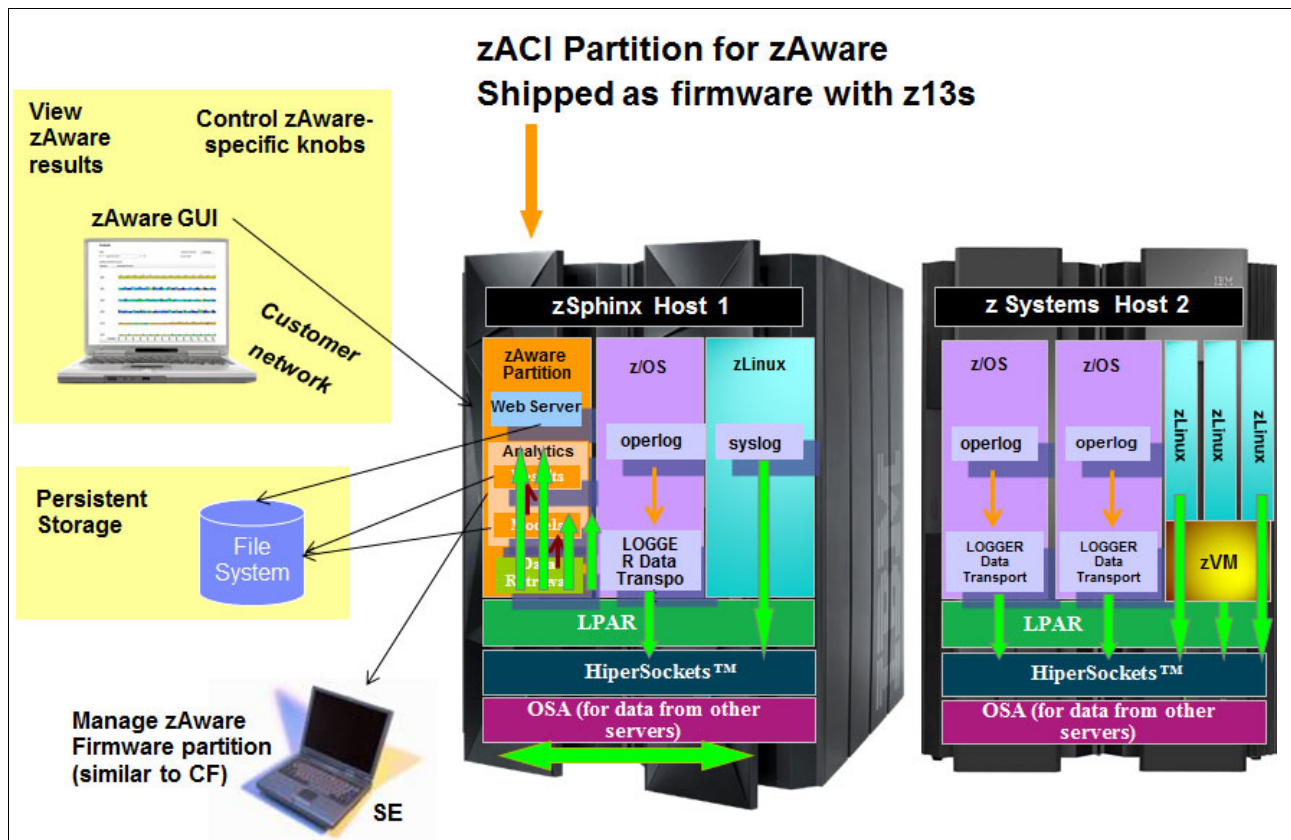


Figure B-3 Basic components of the IBM zAware environment

IBM zAware runs in an independent LPAR as firmware. IBM zAware has the following characteristics:

- ▶ IBM zAware V2 requires the z13s or z13 systems with a priced feature code.

Important: IBM zAware Version 2 server DR configuration requires z13 or z13s servers. IBM zAware Version 1 cannot be in a DR configuration with IBM zAware Version 2.

- ▶ IBM zAware V1 requires zEC12 or zBC12 systems with a priced feature code.
- ▶ Needs processor, memory, disk storage, and network resources to be assigned to the LPAR that it runs.
- ▶ Is updated like all other firmware, with a separate engineering change stream.
- ▶ Is loaded from the Support Element (SE) hard disk by activating an Image Profile configured with a mode of IBM zAware
- ▶ Employs out-of-band monitoring with minimal effect on product workloads.

Figure B-4 shows IBM zAware Image Profile on the Hardware Management Console (HMC). IBM zAware is displayed on the General tab of the image profile.

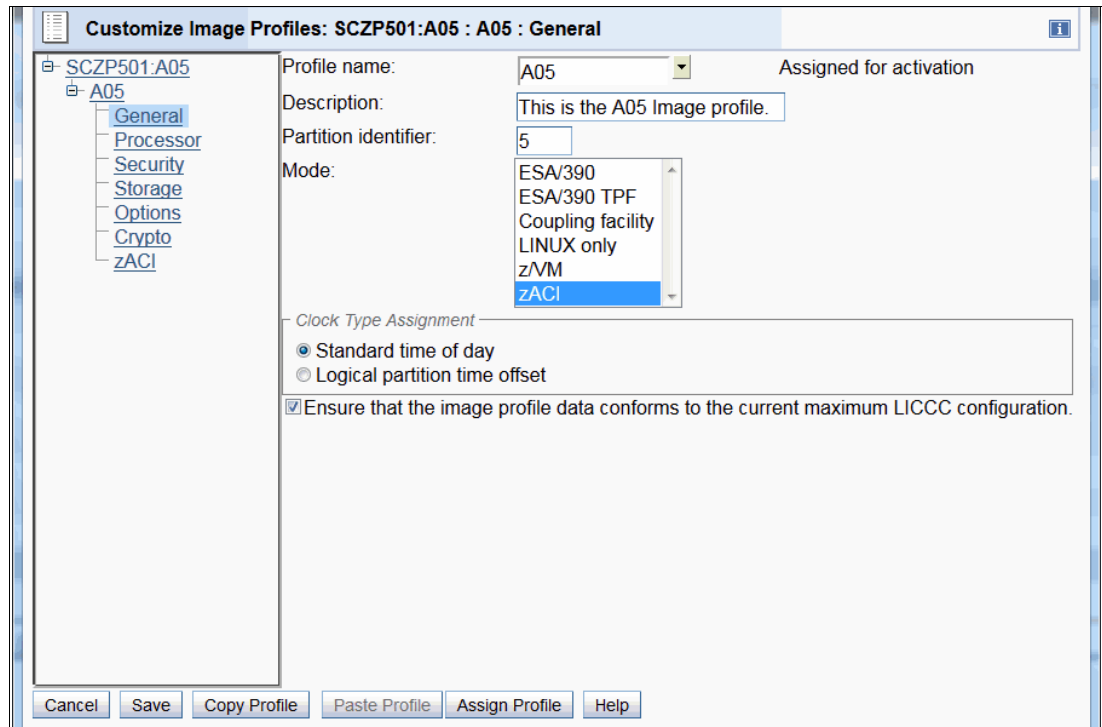


Figure B-4 HMC Image Profile for an IBM zAware LPAR

Figure B-5 shows the tab of the image profile on the HMC. The setup information is configured on the tab of the image profile. This setup must be performed for the initial activation of the image profile.

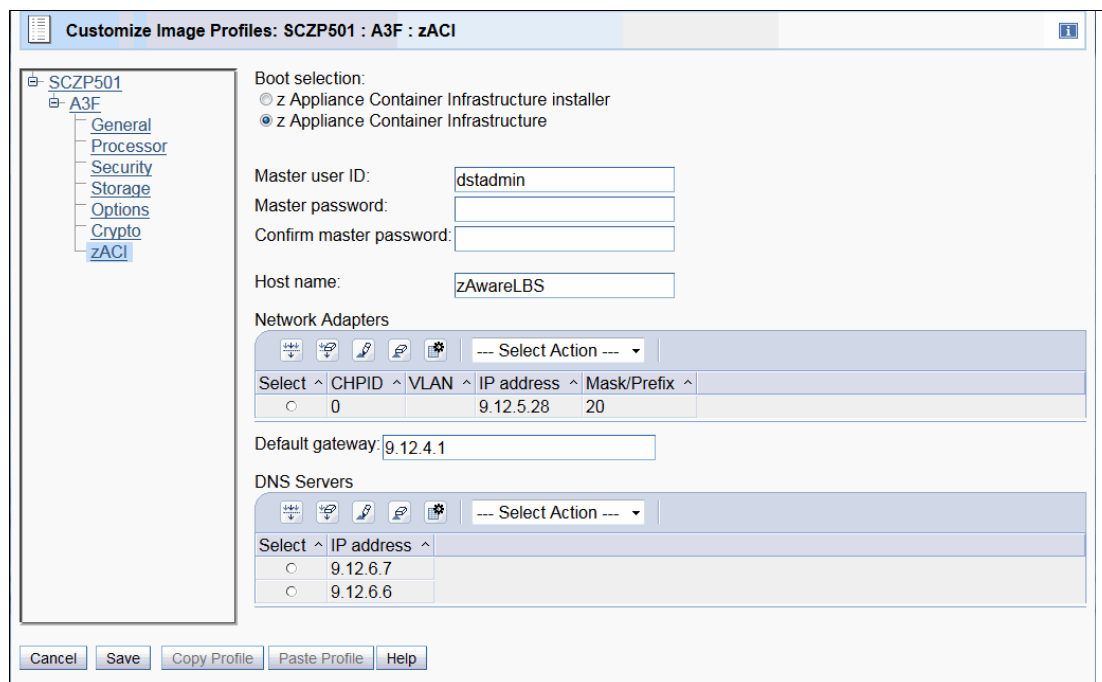


Figure B-5 HMC Image Profile for an IBM zAware LPAR

The following settings are configured in the IBM zAware image profile:

- ▶ Host name: A unique host name for the LPAR
- ▶ Master user ID and password: Used after the LPAR is initialized by the workstation through the intranet to configure or use the IBM zAware appliance.
- ▶ Default gateway / DNS Servers: Network information that defines the network interfaces of the IBM zAware appliance. These interfaces allow a workstation on the intranet to access the IBM zAware appliance, and an internal hipersocket channel to other LPAR clients on the same hosting system.

IBM zAware analyzes massive amounts of OPERLOG messages, including all z/OS console messages, and ISV and application-generated messages, to build sysplex and LPAR detailed views in the IBM zAware graphical user interface (GUI). Linux on z Systems images must be configured so that the syslog daemon sends data to IBM zAware. IBM zAware can create model groups based on similar operational characteristics for Linux images that run on z Systems.

Figure B-6 shows the IBM zAware Heat Map view.

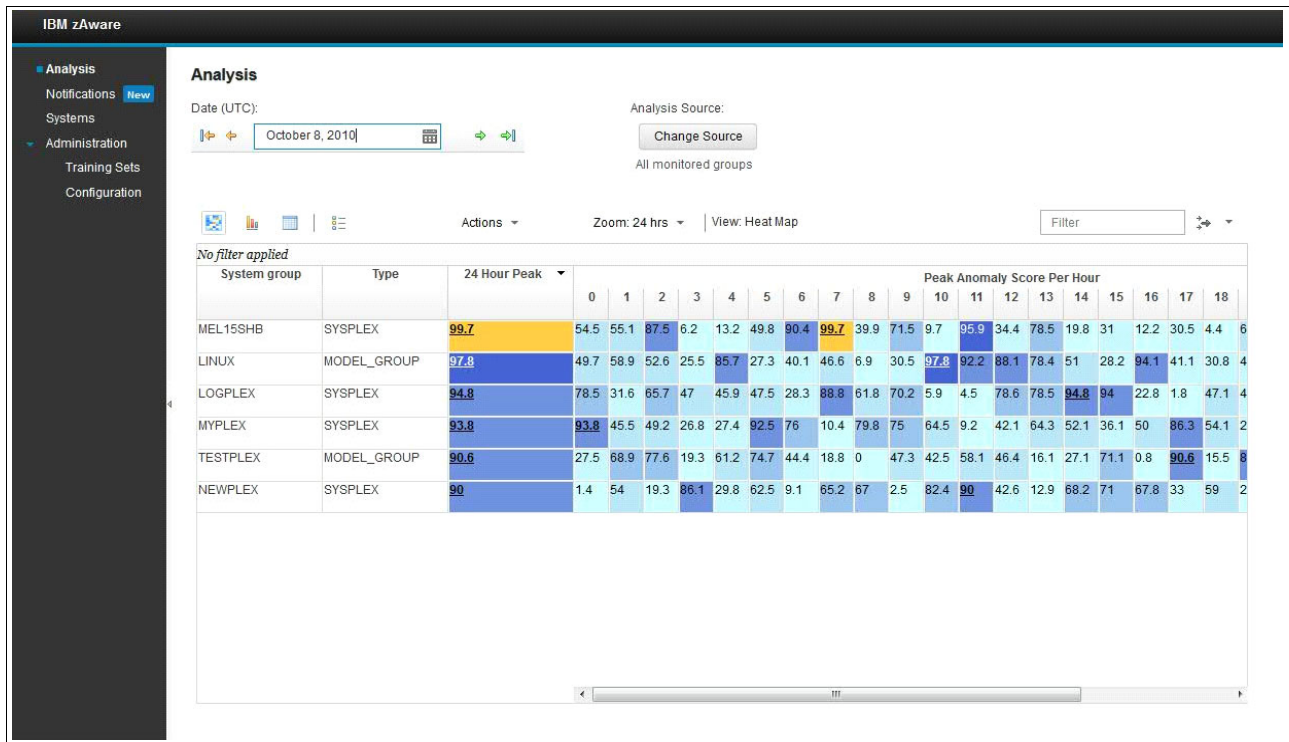


Figure B-6 IBM zAware Heat Map view analysis

The Analysis Heat Map Table displays peak anomaly scores per hour for the groups or systems indicated by the Analysis Source setting. Use this display to quickly find which system groups or which monitored systems have the highest anomaly scores within a specific day or hour.

Figure B-7 shows a sample bar score view.

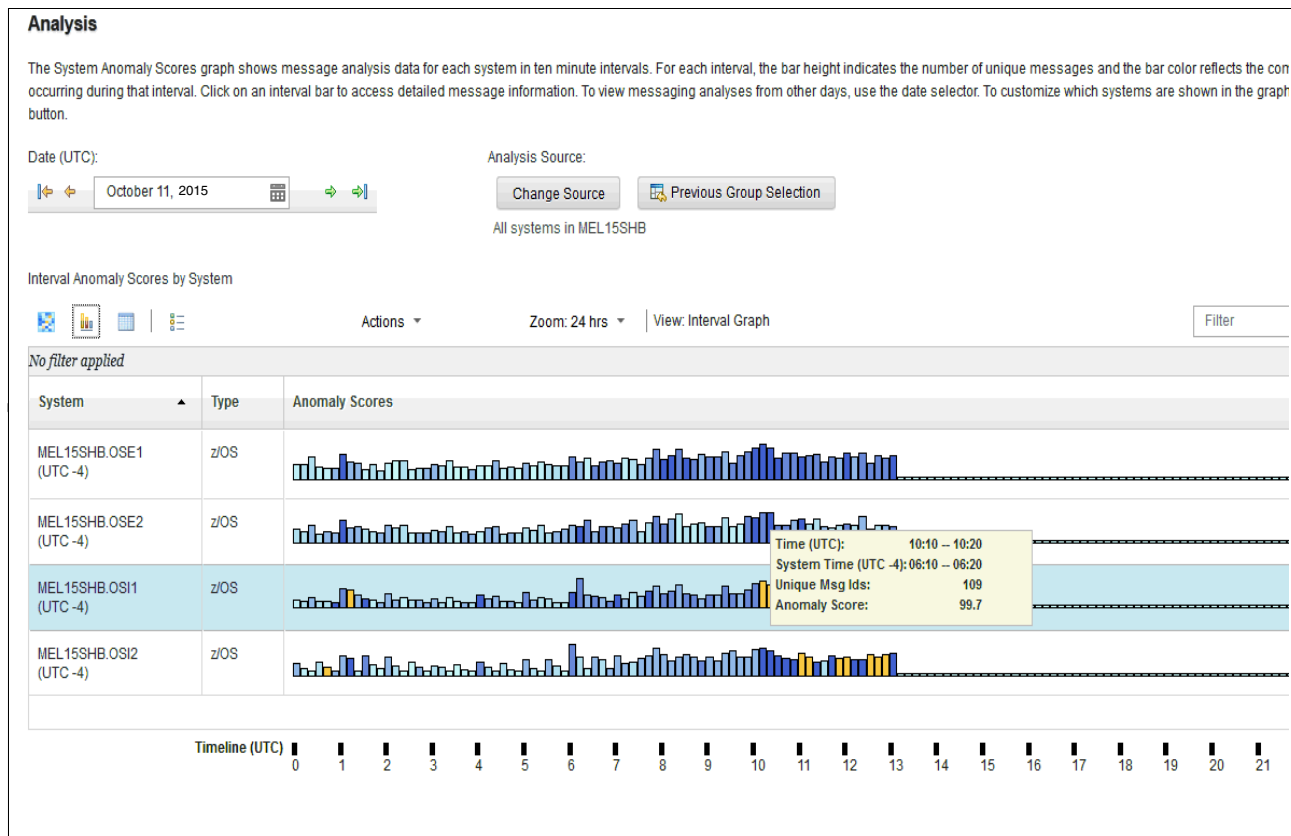


Figure B-7 IBM zAware bar score with intervals

The analytics create a statistical model of the normal message traffic that is generated by each monitored client (system or group of systems, such as z/OS or Linux on z Systems). This model is stored in a database and is used to identify out-of-the-ordinary messages and patterns of messages.

Using a sliding 10-minute interval that is updated every two minutes, a current score for the interval is created based on the uniqueness of the message traffic:

- ▶ A stable system requires a lower interval score to be marked as *interesting* or *rare*.
- ▶ An unstable system requires a larger interval score to be marked as *interesting* or *rare*.

For each interval, IBM zAware provides details of all of the unique and unusual message IDs within the interval. This data includes how many, how rare, and how much the messages contributed to the intervals score (anomaly score, interval contribution score, rarity score, and appearance count) when they first appeared. IBM zAware also helps in identifying messages with the following conditions:

- ▶ Whether the unusual message IDs are coming from a single component
- ▶ Whether the message is a critical z/OS or Linux kernel message
- ▶ Whether the messages are related to changes, such as new software levels (operating system, middleware, and applications) or updated system settings and configurations

IBM zAware detects conditions that typical monitoring systems miss because of these challenges:

- ▶ Message suppression (message too common): Common messages are useful for long-term health issues.
- ▶ Uniqueness (message not common enough): These messages are useful for real-time event diagnostic procedures.

IBM zAware assigns a color to an interval based on the distribution of interval score:

- ▶ Blue (Normal)
Interval score between 1- 99.5
- ▶ Gold (Interesting)
Interval score between 99.6 - 100
- ▶ Orange (Rare)
An interval score of 101

B.3.1 Training period

The IBM zAware server starts receiving current data from the z/OS system logger that runs on z/OS monitored clients and from Linux SYSLOG for Linux on z Systems monitored clients. However, the server cannot use this data for analysis until a model of normal system behavior exists.

The minimum amount of data for building the most accurate models is 90 days of data for each client. By default, training automatically runs every 30 days. You can modify the number of days that are required for this training period, based on your knowledge of the workloads that run on z/OS monitored clients. This training period applies for all monitored clients. Different training periods cannot be defined for each client.

B.3.2 Priming IBM zAware

Instead of waiting for the IBM zAware server to collect data over the course of the training period, you can *prime* the server. You do so by transferring prior data for monitored clients and requesting that the server build a model for each client from the transferred data. Currently, the bulk transfer of Linux historical data is not supported.

B.3.3 IBM zAware ignore message support

When a new workload is added to a system that is monitored by IBM zAware or is moved to a different system, it often generates messages that are not part of that system's model. Subsequently, these messages are flagged as anomalous and cause orange bars to appear on the IBM zAware analysis window.

Sometimes, the reporting of anomalous behavior is caused solely by the new workload, but sometimes a real problem is present as well. Therefore, it is not appropriate to automatically mark all the messages as "normal" when new workloads are introduced. IBM zAware provides the ignore message support to give you input into the IBM zAware rules. This function allows you to mark messages as "ignore." An ignored message is not part of the IBM zAware interval anomaly scoring, but it does appear in the output.

The first iteration of this work requires you to mark each message to be ignored on a per system basis. That is, for each message that you want to ignore, you must mark that particular message on each system for which IBM zAware is to ignore the message. You can choose from one of two types of ignore message: Until the next training period occurs (automatic or manual train) or Forever.

B.3.4 IBM zAware graphical user interface

IBM zAware creates XML data with the status of the z/OS, Linux image, and details about the message traffic. This data is rendered by the web server that runs as a part of IBM zAware. The web server is available by using a standard web browser (Internet Explorer or Mozilla Firefox).

IBM zAware provides an easy-to-use, browser-based GUI with relative weighting and color coding. For IBM messages, IBM zAware GUI has a link to the message description that often includes a corrective action for the issue that is highlighted by the message.

B.3.5 IBM zAware complements your existing tools

Compared to existing tools, IBM zAware works with relatively little customization. It does not depend on other solutions or manual coding of rules, and is always enabled to watch your system. The XML output that is created by IBM zAware is consumed by existing system monitoring tools, such as IBM Tivoli OMEGAMON XE for z/OS and IBM Tivoli NetView® for z/OS, by using published APIs.

B.4 IBM zAware prerequisites

This section describes the hardware and software requirements for IBM zAware.

B.4.1 IBM zAware features and ordering

IBM zAware is available with IBM z13s, IBM z13, IBM EC12 (zEC12), and BC12 (zBC12) models.

IBM zAware Version 2: IBM zAware Version 2 partition is not supported on hardware earlier than z13. DR configuration must use same generation HW (z13/z13s) for primary and DR IBM zAware host.

IBM zAware partition is not supported on a CPC running in Dynamic Partition Manager (DPM) Mode. For more information, see Appendix E, “IBM Dynamic Partition Manager” on page 499.

IBM zAware feature-related definitions are listed in Table B-2.

Table B-2 IBM zAware feature code definitions

Name	Related feature code	Description
IBM zAware host system	FC0011	Represents the z13s, z13, zEC12, or zBC12 server that hosts the IBM zAware partition. In most cases, the host server also has partitions on it that are being monitored. There can be multiple IBM zAware host partitions on one z13s, z13, zEC12, or zBC12 server, but only one IBM zAware FC0011 feature (no additional charge for multiple host partitions).
IBM zAware monitored client		Represents the z/OS partition that sends OPERLOG files for processing to an IBM zAware partition. Multiple z/OS partitions (monitored clients) can be on the server. The clients can also include Linux running natively or as a guest on a hypervisor (z/VM)
IBM zAware environment		Represents the collection of the IBM zAware host system and the IBM zAware monitored clients that are sending information to the IBM zAware host system.
IBM zAware connection	FC0101 and so on ^a	Represents a set of central processors that are associated with servers that are either the IBM zAware host system or IBM zAware monitored clients.
Disaster recovery (DR) IBM zAware server	FC0102 and so on ^b	Represents the z13, zEC12, or zBC12 with no-charge firmware to run IBM zAware in a disaster situation.

a. The following feature codes are supported:

FC0101: IBM zAware CP 10 pack (z13, zEC12)

FC0138: IBM zAware CP 2 pack (zBC12)

FC0140: IBM zAware CP 4 pack (zBC12)

FC0142: IBM zAware CP 6 pack (zBC12)

FC0150: IBM zAware CP 10 pack (zBC12)

b. The following feature codes are supported:

FC0102: IBM zAware DR CP 10 pack (z13, zEC12)

FC0139: IBM zAware DR CP 2 pack (zBC12)

FC0141: IBM zAware DR CP 4 pack (zBC12)

FC0143: IBM zAware DR CP 6 pack (zBC12)

FC0151: IBM zAware DR CP 10 pack (zBC12)

12.7.3 Feature on Demand (FoD)

Feature on Demand is a centralized way to flexibly entitle features and functions on the system. For example, Feature on Demand contains the IBM z BladeCenter Extension (zBX) Model 004 High Water Marks (HWMs). HWMs refer to the highest quantity of blade entitlements by blade type that the client has purchased. On the z196 and z114, the HWMs are stored in the processor and memory Licensed Internal Code (LIC) configuration code (LICCC) record. From zEC12 onwards, the HWMs are in the FoD record.

The IBM zAware feature availability and installed capacity are also controlled by the FoD LICCC record. The current IBM zAware installed and staged feature values can be obtained by using the Perform Model Conversion function on the SE or from the HMC by using a single object operation (SOO) to the server SE.

Figure B-8 shows the window for IBM zAware Feature on Demand status and value shown under the Perform Model Conversion, Feature on Demand Manage function.

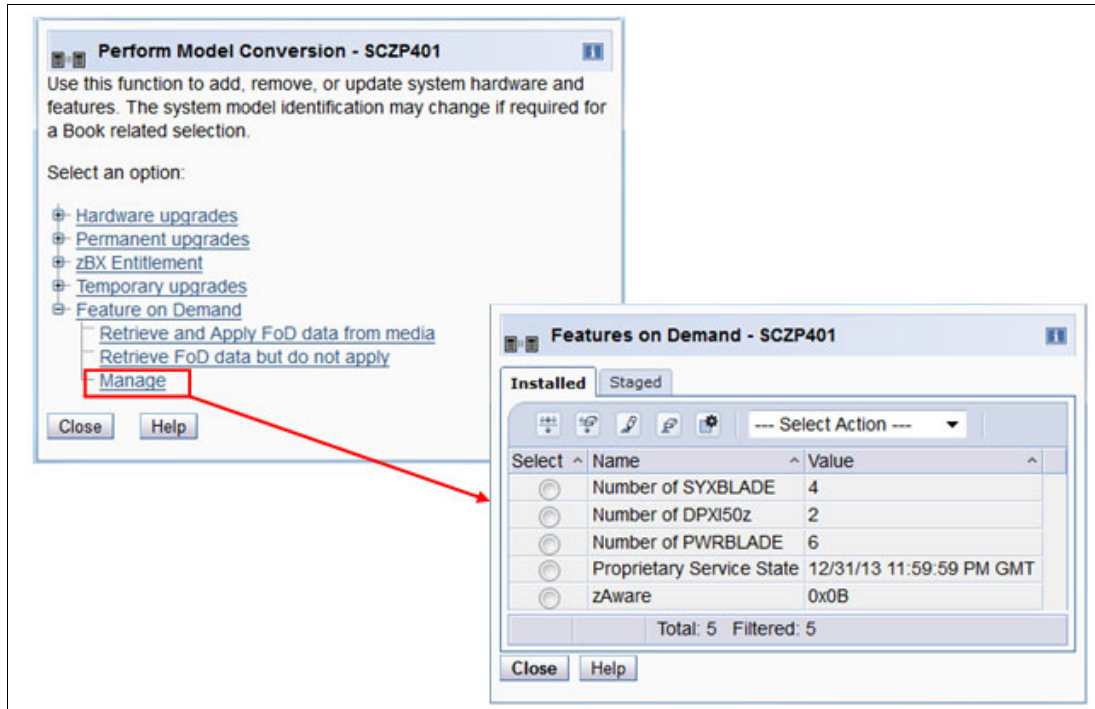


Figure B-8 Feature on Demand window for IBM zAware feature

Only one FoD LICCC record can be installed or staged at any time in the system. Its contents can be viewed under the Manage window as shown in Figure B-8. A staged record can be removed without installing it. A FoD record can be installed only as a complete installation. No selective feature or partial record installation is installed, and the features that are installed will be merged with the CPC LICCC after activation.

A FoD record can be installed only once. If it is removed, a new FoD record is needed to install it again. A remove action cannot be undone.

The IBM zAware host system feature code (FC 0011) must be ordered for the z13s, z13, zEC12, or zBC12 server that hosts the IBM zAware partition.

You do not need to order IBM zAware connections for client systems. The number of IBM zAware connections to be ordered can be calculated by completing the following steps:

1. Determine which systems have z/OS images to be monitored by IBM zAware, including the z13s, z13, zEC12, or zBC12 server where the IBM zAware LPAR resides.
2. Count the number of CPs on the systems that were identified in the previous step. Include banked CPs (HWM), and round up to the nearest factor of 10 (z13s or z13).

Example: z13s 3 CPs + z13 20 CPs + 5 IFLs + zEC12 16 CPs = 44

44 is rounded up to nearest factor of 10 = 50

A disaster recovery option (IBM zAware DR CP packs) is also available and indicates that IBM zAware is installed on a DR z13s, z13, zEC12, or zBC12 server. This feature is available at no additional fee, but is exclusive to the IBM zAware connection.

For example, FC 0102 represents the quantity of DR CPs. FC 0101 represents the quantity of CPs associated with servers that are either the IBM zAware host system or the IBM zAware

monitored clients. FC 0101 and FC 0102 are mutually exclusive. Therefore, if you have one, you cannot have the other. In addition, in most cases, the number of FC 0102 features on DR must match the number of FC 0101 features on the IBM zAware host server.

B.4.2 IBM zAware operating requirements

This section describes the components that are required for IBM zAware.

IBM zAware host system requirements

The z13s and z13 servers can host the IBM zAware Version 2 server. The zEC12, or zBC12 can host a IBM zAware Version 1 server. The IBM zAware server requires a dedicated LPAR and runs its own self-contained firmware stack.

Note: Host system resources (processors, memory, direct access storage devices (DASDs), and so on) depend on the number of monitored clients, amount of message traffic, and length of time that data is retained.

The following components are required:

- ▶ Processors
 - General-purpose CP or IFL that can be shared with other LPARs in the z13s, z13, zEC12, or zBC12 server
 - Usage estimates between a partial engine to two engines, depending on the size of the configuration
- ▶ Memory
 - Minimum 4 GB initial memory for the first six z/OS clients
 - 256 MB required for each additional z/OS client above the first six z/OS clients
 - 256 MB required for each additional LINUX client
 - Flash Express is not supported
- ▶ DASDs
 - 500 GB persistent DASD storage
 - Only extended count key data (ECKD) format is supported
 - Fibre Channel Protocol (FCP) devices are not supported
 - IBM zAware manages its own data store
- ▶ Network (for both instrumentation data gathering and outbound alerting/communications)
 - HiperSockets for the z/OS and Linux LPARs running on the same z13s, z13, zEC12, or zBC12 server as the IBM zAware LPAR
 - OSA ports for the z/OS LPARs running on a different CPC than where the IBM zAware LPAR runs and for browser access to the GUI
 - Dedicated IP address for IBM zAware LPAR

IBM zAware monitored client requirements

IBM zAware monitored clients can be in the same CPC as the IBM zAware host system or in different CPCs. They can be in the same site or multiple sites.

- ▶ Distance between the IBM zAware host systems and monitored clients can be up to a maximum of 3500 km (2174.79 miles).
- ▶ IBM zAware monitored clients can be on any z Systems servers (IBM z13s, z13, zEC12, zBC12, z196, z114, z10, and so on) if they fulfill the operating system requirements. Monitoring can be done by transmitting log files through an Internet Protocol network with IBM zAware servers.

Operating system requirements

IBM zAware Version 2 monitored clients have the following requirements (IBM zAware Version 1 only monitors z/OS clients):

- ▶ Linux on z Systems (SUSE or Red Hat)
- ▶ z/OS V2.1 or higher
- ▶ z/OS V1.13 with program temporary fixes (PTFs)
- ▶ 90 days historical SYSLOG or formatted OPERLOG data to initially prime IBM zAware

B.5 Configuring and using IBM zAware virtual appliance

The following checklist provides a task summary for configuring and using IBM zAware:

1. Phase 1: Planning
 - a. Plan the configuration of the IBM zAware environment.
 - b. Plan the LPAR characteristics of the zACI LPAR to host IBM zAware firmware.
 - c. Plan the network connections that are required for the partition and each monitored client.
 - d. Plan the security requirements for the IBM zAware server, its monitored clients, and users of the IBM zAware GUI.
 - e. Plan for using the IBM zAware GUI.
2. Phase 2: Configuring the zACI partition for IBM zAware
 - a. Verify that your installation meets the prerequisites for using the IBM zAware virtual appliance.
 - b. Configure network connections for the partition through the hardware configuration definition (HCD) or the input/output configuration program (IOCP).
 - c. Configure persistent storage for the partition through the HCD or IOCP.
 - d. Define the LPAR characteristics of the partition through the HMC.
 - e. Define network settings for the partition through the HMC.
 - f. Activate the partition through the HMC.
3. Phase 3: Configuring the IBM zAware server and its monitored clients
 - a. Assign storage devices for the IBM zAware server through the IBM zAware GUI.
 - b. Optional: Replace the self-signed certificate authority (CA) certificate that is configured in the IBM zAware server.
 - c. Optional: Configure a Lightweight Directory Access Protocol (LDAP) directory or local file-based repository for authenticating users of the IBM zAware GUI.

- d. Optional: Authorize users or groups to access the IBM zAware GUI.
- e. Optional: Modify the configuration values that control the IBM zAware analytics operation.
- f. Configure a network connection for each z/OS and Linux monitored client through the TCP/IP profile. If necessary, update firewall settings.
- g. Verify that each z/OS system meets the sysplex configuration and OPERLOG requirements for IBM zAware virtual appliance monitored clients.
- h. Configure the z/OS system logger to send data to the IBM zAware virtual appliance server.
- i. Configure the Linux SYSLOG to send data to the IBM zAware virtual appliance server.
- j. Prime the IBM zAware server with prior data from monitored clients.
- k. Build a model of normal system behavior for each monitored client.¹ The IBM zAware server uses these models for analysis.
- l. Optional: Use the IBM zAware ignore message support to provide input to the IBM zAware rules. It allows you to mark messages as “ignore.” An ignored message is not part of IBM zAware analysis and scoring.

¹ For Linux on z Systems, model groups are defined based on host names. Models are built with available data for systems matching the model group definition.



Channel options

This appendix describes all channel attributes, the required cable types, the maximum unrepeated distance, and the bit rate for z13s servers.

This appendix includes the following sections:

- ▶ C.1, “Channel options supported on z13s servers” on page 472
- ▶ C.2, “Maximum unrepeated distance for FICON SX features” on page 473

C.1 Channel options supported on z13s servers

For all optical links, the connector type is LC Duplex, except for the 12xInfiniBand (IFB) and the Integrated Coupling Adapter (ICA SR) connection, which are established with multifiber push-on (MPO) connectors. The MPO connector of the 12xIFB connection has one row of 12 fibers, and the MPO connector of the ICA connection has two rows of 12 fibers. The electrical Ethernet cable for the Open Systems Adapter (OSA) connectivity is connected through an RJ45 jack.

Table C-1 lists the attributes of the channel options that are supported on z13s servers.

Table C-1 z13s channel feature support

Channel feature	Feature codes	Bit rate ^a in Gbps (or stated)	Cable type	Maximum unrepeated distance ^b	Ordering information
Fiber Connection (FICON)					
FICON Express16S 10KM LX	0418	4, 8, or 16	SM 9 μm	10 km (6.2 miles)	New build
FICON Express16S SX	0419	4, 8, or 16	OM2, OM3	See Table C-2 on page 473.	New build
FICON Express8S 10KM LX	0409	2, 4, or 8	SM 9 μm	10 km (6.2 miles)	New build
FICON Express8 10KM LX	3325				Carry forward
FICON Express8S SX	0410	2, 4, or 8	OM1, OM2, OM3	See Table C-2 on page 473.	New build
FICON Express8 SX	3326				Carry forward
Open Systems Adapter (OSA)					
OSA-Express5S 10 GbE LR	0415	10	SM 9 μm	10 km (6.2 miles)	New build
OSA-Express4S 10 GbE LR	0406				Carry forward
OSA-Express5S 10 GbE SR	0416	10	MM 62.5 μm MM 50 μm	33 m (200) 82 m (500) 300 m (2000)	New build
OSA-Express4S 10 GbE SR	0407				Carry forward
OSA-Express5S GbE LX	0413	1.25	SM 9 μm	5 km (3.1 miles)	New build
OSA-Express4S GbE LX	0404				Carry forward
OSA-Express5S GbE SX	0414	1.25	MM 62.5 μm MM 50 μm	275 m (200) 550 m (500)	New build
OSA-Express4S GbE SX	0405				Carry forward
OSA-Express5S 1000BASE-T	0417	100 or 1000 Mbps	Cat 5, Cat 6 unshielded twisted pair (UTP)	100 m	New build
OSA-Express4S 1000BASE-T	0408	10, 100, or 1000 Mbps			Carry forward

Channel feature	Feature codes	Bit rate ^a in Gbps (or stated)	Cable type	Maximum unrepeated distance ^b	Ordering information
10GbE Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE) Express	0411	10	OM3	300 m	New build
Parallel Sysplex					
ICA (PCIe-O SR)	0172	8 Gbps	OM4	150 m	New build
			OM3	100 m	New build
HCA3-O (12x IFB)	0171	6 Gbps	OM3	150 m	New build
HCA3-O LR (1x IFB)	0170	2.5 or 5 Gbps	SM 9 μm	10 km (6.2 miles)	New build
IC	N/A		N/A	N/A	N/A
Cryptography					
Crypto Express5S	0890	N/A	N/A	N/A	New build
Flash Express	0403	N/A	N/A	N/A	New build
zEDC Express	0420	N/A	N/A	N/A	New build

a. The link data rate does not represent the actual performance of the link. The actual performance depends on many factors that include latency through the adapters, cable lengths, and the type of workload.

b. Where applicable, the minimum fiber bandwidth distance in MHz-km for multi-mode fiber optic links is included in parentheses.

C.2 Maximum unrepeated distance for FICON SX features

Table C-2 shows the maximum unrepeated distances for FICON SX features.

Table C-2 Maximum unrepeated distance for FICON SX features

Cable type/bit rate	1 Gbps	2 Gbps	4 Gbps	8 Gbps	16 Gbps
OM1 (62.5 μm at 200 MHz·km)	300 meters	150 meters	70 meters	21 meters	N/A
	984 feet	492 feet	230 feet	69 feet	N/A
OM2 (50 μm at 500 MHz·km)	500 meters	300 meters	150 meters	50 meters	35 meters
	1640 feet	984 feet	492 feet	164 feet	115 feet
OM3 (50 μm at 2000 MHz·km)	860 meters	500 meters	380 meters	150 meters	100 meters
	2822 feet	1640 feet	1247 feet	492 feet	328 feet



Shared Memory Communications

This appendix briefly describes the optional Shared Memory Communications (SMC) functionality implemented on IBM z Systems servers as Shared Memory Communications - Remote Direct Memory Access (SMC-R) and the new Shared Memory Communications - Direct Memory Access (SMC-D) of the IBM z13 and z13s servers.

This appendix includes the following sections:

- ▶ Shared Memory Communications overview
- ▶ Shared Memory Communication over RDMA
- ▶ Shared Memory Communications - Direct Memory Access

D.1 Shared Memory Communications overview

The volume of data being generated and transmitted by technologies driven by cloud, mobile, analytics, and social computing applications is growing. This increases the pressure on business IT organizations to be able to provide fast access to that data across the web, application, and database tiers that comprise most enterprise workloads. Shared Memory Communications helps to access data faster and with less latency, and helps reduce CPU resource consumption over traditional TCP/IP for communications.

D.2 Shared Memory Communication over RDMA

IBM z13s delivers improvements for the RoCE exploitation over the previous zEnterprise generation (zEC12, zBC12). On z13 and z13s servers, the IBM 10GbE RoCE Express feature can be shared between up to 31 partitions and the two ports are enabled to be used in z/OS. IBM z13 and z13s servers improve the usability of the RoCE feature by using existing z Systems servers and industry standard communications technology along with emerging new network technology:

- ▶ RDMA technology provides low latency, high bandwidth, high throughput, and low processor utilization attachment between hosts.
- ▶ SMC-R is a protocol that allows existing TCP applications to benefit transparently from RDMA for transferring data:
 - SMC-R uses a 10GbE RoCE Express adapter as the physical transport layer.
 - Initial deployment is limited to z/OS to z/OS communications with a goal to expand exploitation to more operating systems and possibly appliances and accelerators.
- ▶ Single root I/O virtualization (SR-IOV) technology provides the capability to share the 10GbE RoCE Express adapter between logical partitions (LPARs).

D.2.1 RDMA technology overview

RoCE is part of the InfiniBand Architecture Specification that provides InfiniBand transport over Ethernet fabrics. It encapsulates InfiniBand transport headers into Ethernet frames by using an IEEE-assigned Ethertype. One of the key InfiniBand transport mechanisms is RDMA, which is designed to allow transfer of data to or from memory on a remote system with low latency, high throughput, and low CPU utilization.

Traditional Ethernet transports, such as TCP/IP, typically use software-based mechanisms for error detection and recovery, and are based on the underlying Ethernet fabric using a “best-effort” policy. With the traditional policy, the switches typically discard packets in congestion and rely on the upper-level transport for packet retransmission.

RoCE, however, uses hardware-based error detection and recovery mechanisms that are defined by the InfiniBand specification. A RoCE transport performs best when the underlying Ethernet fabric provides a lossless capability, where packets are not routinely dropped. This goal can be accomplished by using Ethernet flow control where Global Pause frames are enabled for both transmission and reception on each of the Ethernet switches in the path between the 10GbE RoCE Express features. This capability is enabled by default in the 10GbE RoCE Express feature.

RDMA has two key requirements as shown in Figure D-1:

- ▶ A reliable “lossless” Ethernet network fabric (LAN for layer 2 data center network distance)
- ▶ An RDMA network interface card (RNIC)

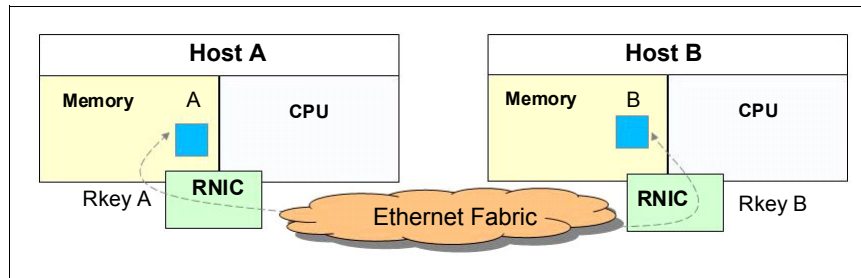


Figure D-1 RDMA technology overview

RDMA technology is now available on Ethernet. RoCE uses an existing Ethernet fabric (switches with Global Pause enabled) and requires advanced Ethernet hardware (RDMA-capable NICs on the host).

D.2.2 Shared Memory Communications over RDMA

SMC-R is a protocol that allows TCP socket applications to transparently use RDMA.

SMC-R is a “hybrid” solution as shown in Figure D-2 on page 478:

- ▶ It uses an existing TCP connection to establish the SMC-R connection.
- ▶ A TCP option (SMCR) controls switching from TCP to “out of band” SMC-R.
- ▶ The SMC-R information is exchanged within the TCP data stream.
- ▶ Socket application data is exchanged through RDMA (write operations).
- ▶ The TCP connection remains to control the SMC-R connection.
- ▶ This model preserves many critical operational and network management features of TCP/IP.

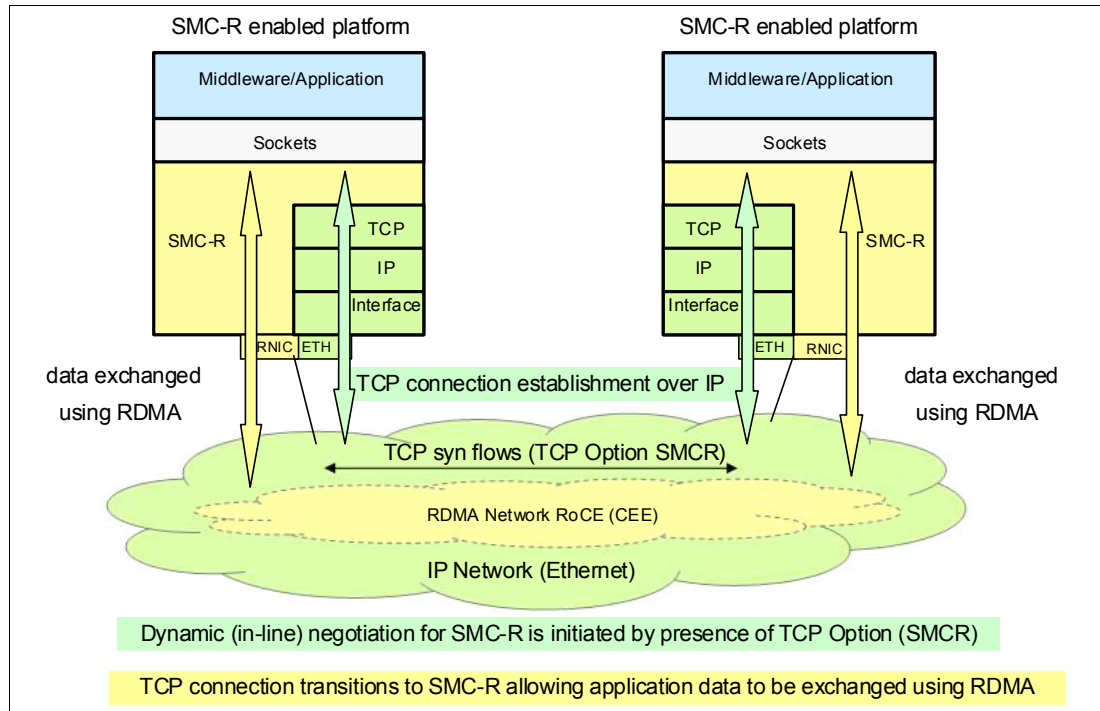


Figure D-2 Dynamic transition from TCP to SMC-R

The hybrid model of SMC-R uses these key existing attributes:

- ▶ It follows the standard TCP/IP connection setup.
- ▶ The hybrid model switches to RDMA (SMC-R) dynamically.
- ▶ The TCP connection remains active (idle) and is used to control the SMC-R connection.
- ▶ The hybrid model preserves the following critical operational and network management TCP/IP features:
 - Minimal (or zero) IP topology changes
 - Compatibility with TCP connection-level load balancers
 - Preservation of the existing IP security model, such as IP filters, policies, virtual LANs (VLANs), and Secure Sockets Layer (SSL)
 - Minimal network administration and management changes
- ▶ Host application software is not required to change, so all host application workloads can benefit immediately.

D.2.3 Single Root I/O Virtualization

SR-IOV is a technology that is designed to provide the capability to share the adapter between up to 31 LPARs. SR-IOV is also designed to provide isolation of virtual functions within the Peripheral Component Interconnect Express (PCIe) 10GbE RoCE Express adapter. For example, one LPAR cannot cause errors visible to other virtual functions or other LPARs. Each operating system LPAR has its own application queue in its own memory space.

Figure D-3 shows the concept of the Shared RoCE Mode.

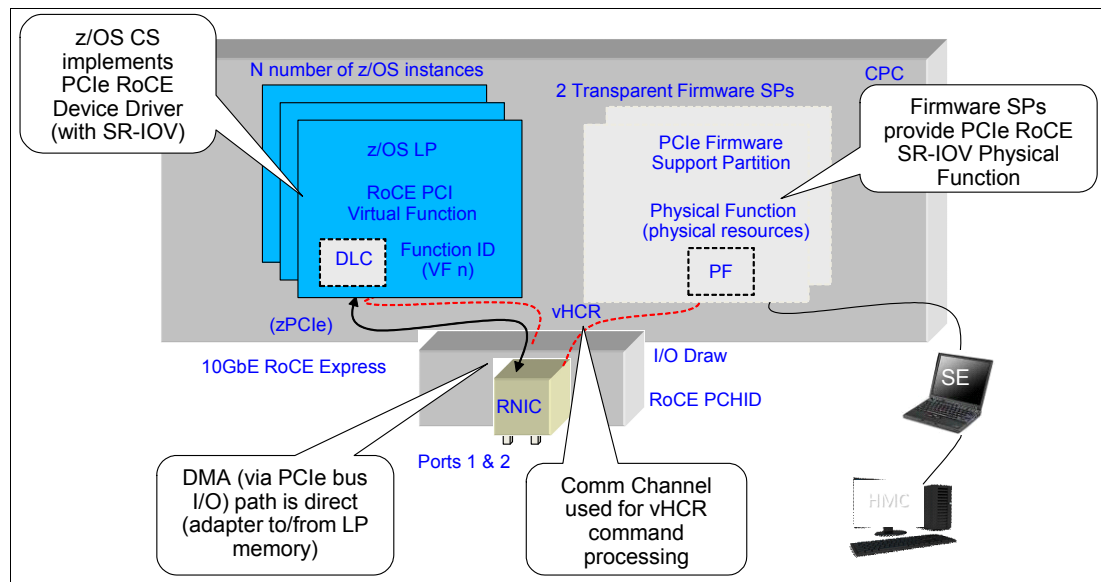


Figure D-3 Shared RoCE mode concepts

The Physical Function (PF) Driver communicates to the Physical Function in the PCIe adapter. The PF Driver has a relatively limited function:

- ▶ Manages resource allocation
- ▶ Perform hardware error handling
- ▶ Perform code updates
- ▶ Run diagnostics

The Device-specific z Systems Licensed Internal Code (LIC) connects PF Driver to Support Elements (SEs) and limited system level firmware required services.

D.2.4 Hardware

The 10 Gigabit Ethernet (10GbE) RoCE Express feature (FC 0411) is an RDMA-capable network interface card (NIC). The integrated firmware processor (IFP) has two Resource Groups (RGs) which have firmware for the 10GbE RoCE Express feature. See G.1.3, “Resource groups” on page 523 in this book for reference.

D.2.5 10GbE RoCE Express feature

The 10GbE RoCE Express feature is designed to help reduce the consumption of CPU resources for applications that use the TCP/IP stack, such as WebSphere accessing a DB2 database. Using the 10GbE RoCE Express feature also helps to reduce network latency with memory-to-memory transfers that use SMC-R in z/OS V2.1 or later. It is transparent to applications and can be used for LPAR-to-LPAR communications on a single z/OS system or server-to-server communications in a multiple CPC environment.

Table D-1 shows the differences in the number of ports and shared support for different systems.

Table D-1 RoCE number of enabled ports and shared/dedicated environment

System Name	z/OS Supported Ports	Shared mode	Dedicated mode
z13	2	YES	No
z13s	2	YES	No
zEC12	1	NO	YES
zBC12	1	NO	YES

The 10GbE RoCE Express feature shown in Figure D-4 is used exclusively in the PCIe I/O drawer. Each feature has one PCIe adapter and two ports. A maximum of 16 features can be installed. The 10GbE RoCE Express feature uses a short reach (SR) laser as the optical transceiver and supports the use of a multimode fiber optic cable terminated with an LC Duplex connector. Both point-to-point connection (with another 10GbE RoCE Express adapter) and switched connection with an enterprise-class 10 GbE switch are supported.



Figure D-4 10GbE RoCE Express

Although SMC-R can be used with direct RoCE Express to RoCE Express connectivity (without any switch), this type of direct physical connectivity forms a single physical point-to-point connection, disallowing any other connectivity with other LPARs (for example, any other SMC-R peers). Although this is a viable option for test scenarios, it is not practical (nor recommended) for production deployment.

If the IBM 10GbE RoCE Express features are connected to 10 GbE switches, the switches must support the following requirements:

- ▶ Global Pause function frame (as described in the IEEE 802.3x standard) should be enabled
- ▶ Priority Flow Control (PFC) disabled
- ▶ No firewalls, no routing, and no intraensemble data network (IEDN)

The maximum supported unrepeatd point-to-point distance is 300 meters (984.25 ft.).

A client-supplied cable is required. Three types of cables can be used for connecting the port to the selected 10 GbE switch or to the 10GbE RoCE Express feature on the attached server:

- ▶ OM3 50 micron multimode fiber optic cable rated at 2000 MHz-km terminated with an LC Duplex connector (supports 300 m (984.25 ft.))
- ▶ OM2 50 micron multimode fiber optic cable rated at 500 MHz-km terminated with an LC Duplex connector (support 82 m (269 ft.))
- ▶ OM1 62.5 micron multimode fiber optic cable rated at 200 MHz-km terminated with an LC Duplex connector (support 33 m (108.2 ft.))

D.2.6 10GbE RoCE Express configuration example

Figure D-5 illustrates a sample configuration that allows redundant SMC-R connectivity among LPAR A and C, and LPARs 1, 2, and 3. Each feature can be shared or dedicated to an LPAR. Like the sample configuration, two features per LPAR are advised for redundancy.

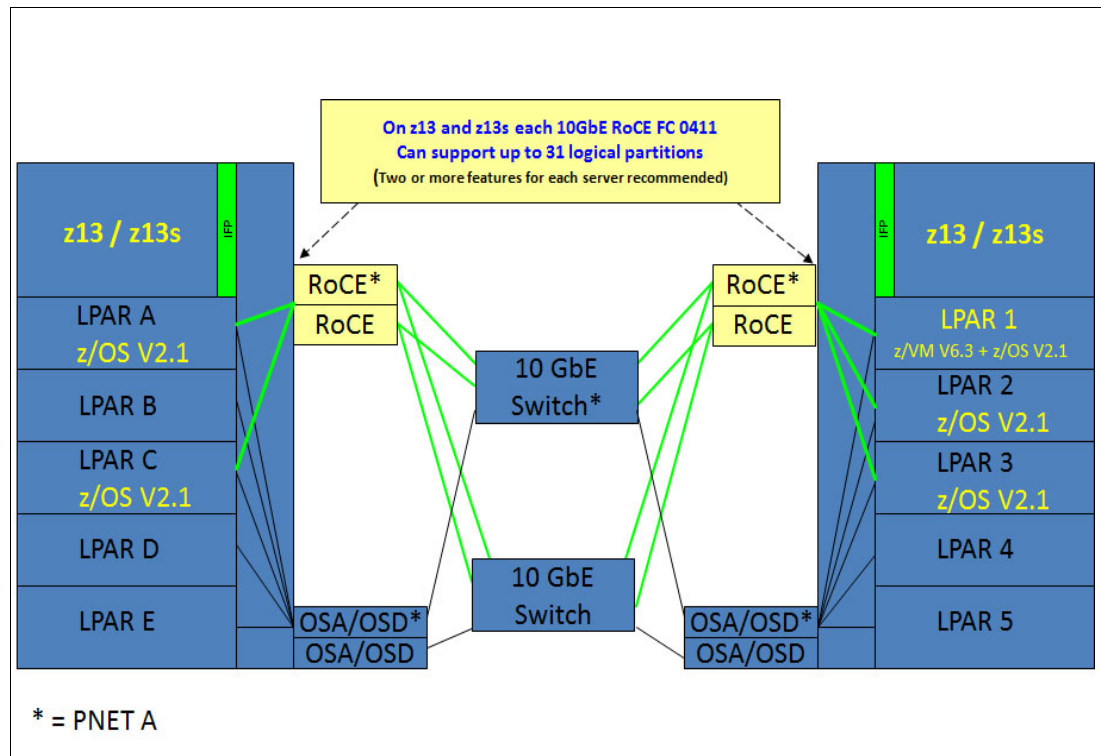


Figure D-5 10GbE RoCE Express sample configuration

The configuration that is shown in Figure D-5 allows redundant SMC-R connectivity among LPAR A, LPAR C, LPAR 1, LPAR 2, and LPAR 3. LPAR to LPAR OSD connections are required to establish the SMC-R communications. 1 GbE OSD connections can be used

instead of 10 GbE. OSD connections can flow through the same 10 GbE switches or different switches.

Note: The OSA-Express Adapter and the RoCE Express feature must be associated to each other by having equal PNET IDs (defined in HCD). Simultaneous use of both 10 GbE ports on a 10GbE RoCE Express feature and sharing by up to 31 LPARs on the same CPC is available on z13 and z13s servers.

An OSA-Express feature, defined as channel-path identifier (CHPID) type OSD, is required to establish SMC-R. Figure D-6 shows the interaction of OSD and the RNIC. The Open System Adapter (OSA) feature might be a single or pair of 10 GbE, 1 GbE, or 1000Base-T OSAs. The OSA needs to be connected to another OSA on the system with which the RoCE feature is communicating. In Figure D-5, 1 GbE OSD connections can still be used instead of 10 GbE and OSD connections can flow through the same 10 GbE switches.

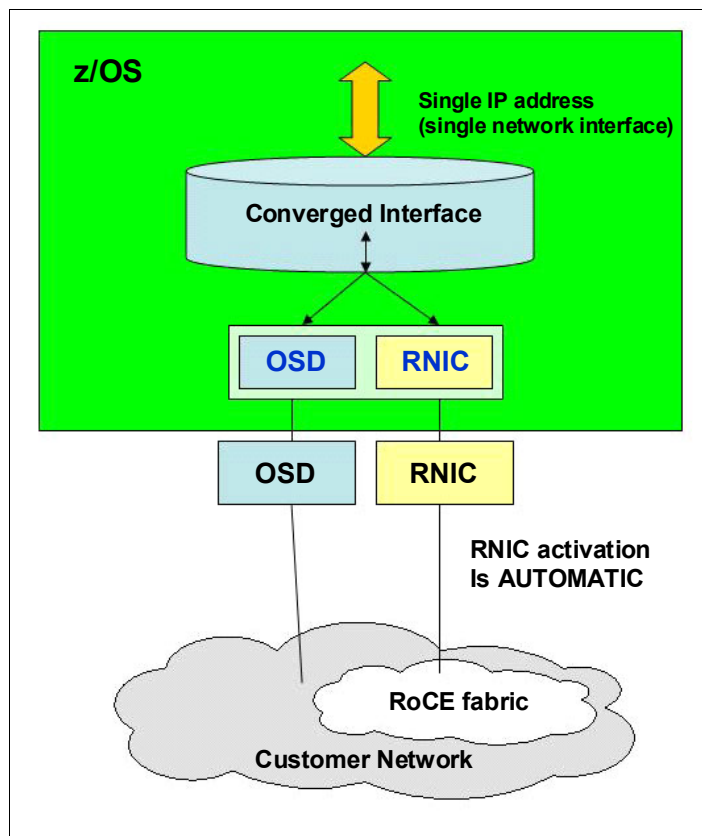


Figure D-6 RNIC and OSD interaction

The following notes refer to Figure D-6:

- ▶ The z/OS system administrator only must configure and manage the OSD interface.
- ▶ The Communications Server transparently splits and converges network traffic to and from the converged interface.
- ▶ Only OSD connectivity must be configured.

With SMC-R, the RNIC interface is dynamically and transparently added and configured.

D.2.7 Hardware configuration definitions

The following hardware configuration definitions (HCDs) are important.

Function ID

The RoCE feature is identified by a hexadecimal Function Identifier (FID). It has a dedicated limit in the range 00 - FF. The shared limit is 000 - 0FFF in the HCD or Hardware Management Console (HMC) to create the input/output configuration program (IOCP) input. A FID can only be configured to one LPAR, but it is reconfigurable. The RoCE feature, as installed in a specific PCIe I/O drawer and slot, is to be used for the defined function. The physical installation (drawer and slot) determines the physical channel identifier (PCHID). Only one FID can be defined for dedicated mode. Up to 31 FIDs can be defined for shared mode (on a z13 and a z13s server) for each physical cards (PCHID).

Virtual Function ID

Virtual Function ID is defined when PCIe hardware is shared between LPARs. Virtual Function ID has a decimal Virtual Function Identifier (VF=) in the range 1 – n, where n is the maximum number of partitions the PCIe feature supports. For example, the RoCE feature supports up to 31 partitions, and a zEDC Express feature supports up to 15.

Physical network ID

As one parameter for the FUNCTION statement, the physical network (PNet) ID is a client-defined value for logically grouping OSD interfaces and RNIC adapters based on physical connectivity. The PNet ID values are defined for both OSA and RNIC interfaces in the HCD. A PNet ID is also defined for each physical port. z/OS Communications Server gets the information during the activation of the interfaces and associates the OSD interfaces with the RNIC interfaces that have matching PNet ID values.

Attention: If you do not configure a PNet ID for the RNIC adapter, activation fails. If you do not configure a PNet ID for the OSA adapter, activation succeeds, but the interface is not eligible to use SMC-R.

Figure D-7 shows the three physically separate networks defined.

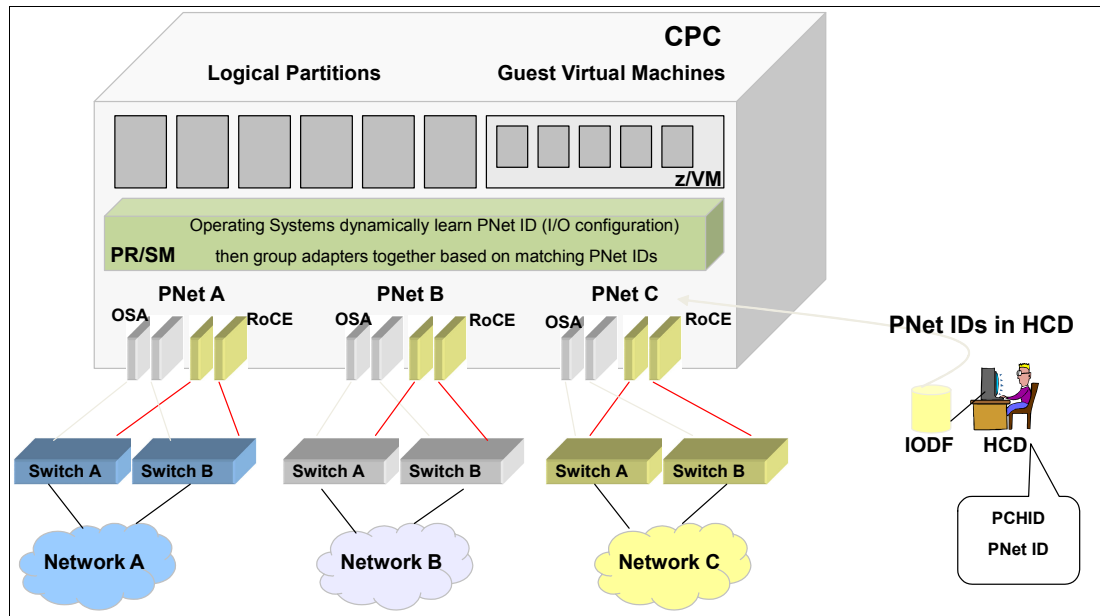


Figure D-7 Physical network ID example

Sample IOCP FUNCTION statement

Example D-1 shows one sample IOCP FUNCTION configuration to define an RoCE Express adapter shared between LPARs.

Example D-1 IOCP FUNCTION statements

```

FUNCTION FID=05,PCHID=100,PART=((LP08),(LP09)),VF=1,TYPE=ROCE,PNETID=(PNETA,PNETB)
FUNCTION FID=06,PCHID=12C,PART=((LP08),(LP09)),VF=1,TYPE=ROCE,PNETID=(PNETA,PNETB)
FUNCTION FID=07,PCHID=100,PART=((LP12),(LP06)),VF=2,TYPE=ROCE,PNETID=(PNETA,PNETB)
FUNCTION FID=08,PCHID=12C,PART=((LP12),(LP06)),VF=2,TYPE=ROCE,PNETID=(PNETA,PNETB)

```

This example has these characteristics:

- ▶ PNETID array identifies the network that the ports are associated with. Thus, all FIDs on a RoCE adapter that are associated with the same PCHID must have the same PNETID for each port.
- ▶ 10GbE RoCE Express Functions for LPAR 08 are reconfigurable to LP 09 with access to two networks.
- ▶ 10GbE RoCE Express Functions for LPAR 12 are reconfigurable to LP 06 with access to two networks.
- ▶ Physical RoCE Express adapters on PCHID 100 and 12C are shared between LPARs 08 and 12.

D.2.8 Software exploitation of SMC-R

SMC-R can be implemented on the RoCE and can communicate memory to memory, thus avoiding the CPU resources of TCP/IP by reducing network latency and improving wall clock time. It focuses on “time to value” and widespread performance benefits for all TCP socket-based middleware.

The following advantages are gained as shown in Figure D-8:

- ▶ No middleware or application changes (transparent)
- ▶ Ease of deployment (no IP topology changes)
- ▶ LPAR-to-LPAR communications on a single central processing complex (CPC)
- ▶ Server-to-server communications in a multi-CPC environment
- ▶ Retained key qualities of service that TCP/IP offers for enterprise class server deployments (high availability, load balancing, and an IP security-based framework)

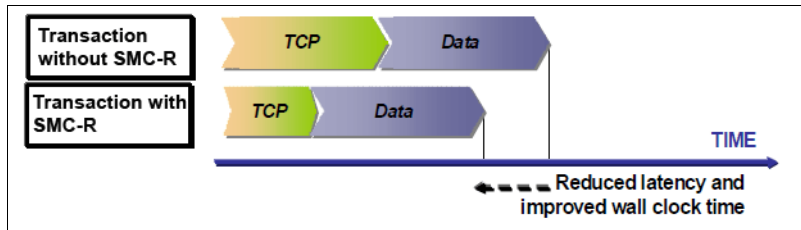


Figure D-8 Reduced latency and improved wall clock time with SMC-R

D.2.9 SMC-R support overview

SMC-R needs both hardware and software support.

Hardware

SMC-R requires the following hardware:

- ▶ PCIe-based RoCE Express
 - z13, z13s, zEC12, and zBC12 servers
 - Dual port 10 GbE adapter
 - Maximum of 16 RoCE Express features per CPC
- ▶ HCD and input/output configuration data set (IOCDs):
 - PCIe FID, VF (sharing), and RoCE configuration with PNet ID
- ▶ Optional: Standard 10 GbE switch (CEE-enabled switch is not required)
- ▶ Required queued direct input/output (QDIO) Mode OSA connectivity between z/OS LPARs as shown in Figure D-5 on page 481.
- ▶ Adapter *must* be dedicated to an LPAR on a zEC12 or zBC12. And it *must* be shared (or at least in shared mode) to one or more LPARs on a z13 or z13s server.
- ▶ SMC-R cannot be used in IEDN.

Software

IOCP required level for z13s: The required level of IOCP for the z13s server is V5 R2 L1 (IOCP 5.2.1) or later with program temporary fixes (PTFs). For more information, see the following manuals:

- ▶ *z Systems Stand-Alone Input/Output Configuration Program User's Guide*, SB10-7166.
- ▶ *z Systems Input/Output Configuration Program User's Guide for ICP IOCP*, SB10-7163.

SMC-R requires the following software:

- ▶ z/OS V2R1 (with PTFs) or higher are the only supported operating systems for the SMC-R protocol. You cannot roll back to previous z/OS releases.
- ▶ z/OS guests under z/VM 6.3 are supported to use 10GbE RoCE features
- ▶ IBM is working with its Linux distribution partners to include support in future Linux on z Systems distribution releases.

Other RoCE considerations

The following are other consideration when using RoCE:

- ▶ RoCE system limits:
 - 16 Physical cards per CPC (no change from zEC12)
 - 31 Virtual Functions per PCHID
 - 128 unique VLANs per PCHID physical port
 - Each VF guaranteed minimum of 2 VLANs with a maximum of 16 (31 VFs, and the maximum number of VLANs depends on the number of unique VLAN IDs)
- ▶ z/OS CS consumption of RoCE virtual resources:
 - One VF consumed per TCP stack (per PCIe function ID (PFID) / port).
 - One virtual message authentication code (VMAC) per VF (z/OS uses PF generated VMAC)
 - One VLAN ID (up to 16) per OSA VLAN (“inherited” as TCP connections occur)
- ▶ z/OS Communications Server Migration considerations:
 - RoCE HCD (IOCDS) configuration changes are required
 - Existing z/OS RoCE users might be required to make a TCP/IP configuration change (that is, existing TCP/IP profile (PFIDs) might be compatible with shared RoCE)
- ▶ Changes are required for existing RoCE users for the following cases:
 - z/OS users who use multiple TCP/IP stacks and both stacks currently use the same RoCE feature (single z/OS image sharing a physical card among multiple stacks).
 - z/OS users who need to use both physical RoCE ports from the same z/OS instance (not “best practices”, but is allowed).
 - z/OS users who could not continue using (coordinate) the same PFID values (continue using the existing PFID value used in the dedicated environment for a specific z/OS instance) when adding multiple PFIDs and VFs to the same card (for additional shared users).

D.2.10 SMC-R use cases for z/OS to z/OS

SMC-R with RoCE provides high-speed communications and “HiperSockets like” performance across physical processors. It can help all TCP-based communications across z/OS LPARs that are in different CPCs.

The following list shows several typical communications patterns:

- ▶ Optimized Sysplex Distributor intra-sysplex load balancing
- ▶ WebSphere Application Server type 4 connections to remote DB2, IMS, and CICS instances
- ▶ IBM Cognos® to DB2 connectivity

- ▶ CICS to CICS connectivity through Internet Protocol interconnectivity (IPIC)

Optimized Sysplex Distributor intra-sysplex load balancing

Dynamic virtual IP address (VIPA) and Sysplex Distributor support are often deployed for high availability (HA) and scalability in the sysplex environment.

When the clients and servers are all in the same sysplex, SMC-R offers a significant performance advantage. Traffic between client and server can flow directly between the two servers without having to traverse the Sysplex Distributor node for every inbound packet, which is the current model with TCP/IP. In the new model, only connection establishment flows must go through the Sysplex Distributor node.

Sysplex Distributor before RoCE

Figure D-9 shows a traditional Sysplex Distributor.

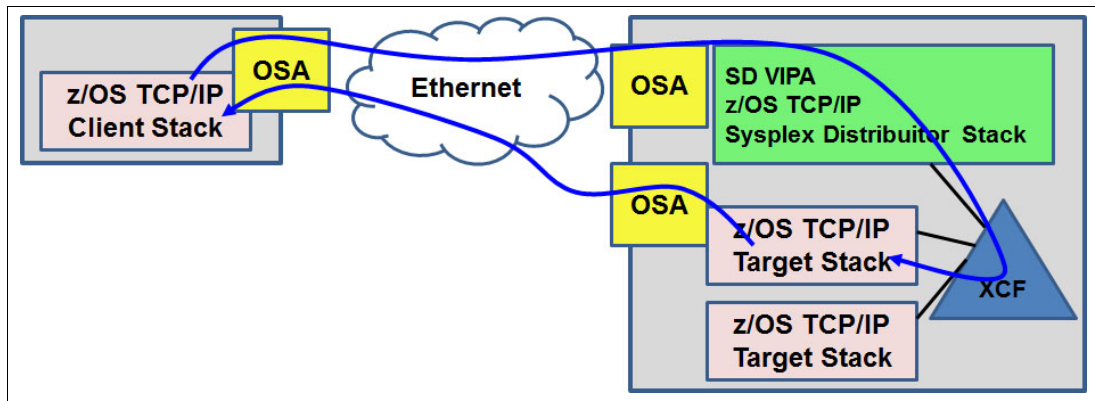


Figure D-9 Sysplex Distributor before RoCE

The traditional Sysplex Distributor has these characteristics:

- ▶ All traffic from the client to the target application goes through the Sysplex Distributor TCP/IP stack.
- ▶ All traffic from the target application goes directly back to the client using the TCP/IP routing table on the target TCP/IP stack.

Sysplex Distributor after RoCE

Figure D-10 shows a RoCE Sysplex Distributor:

- ▶ The initial connection request goes through the Sysplex Distributor stack.
- ▶ The session then flows directly between the client and the target over the RoCE cards.

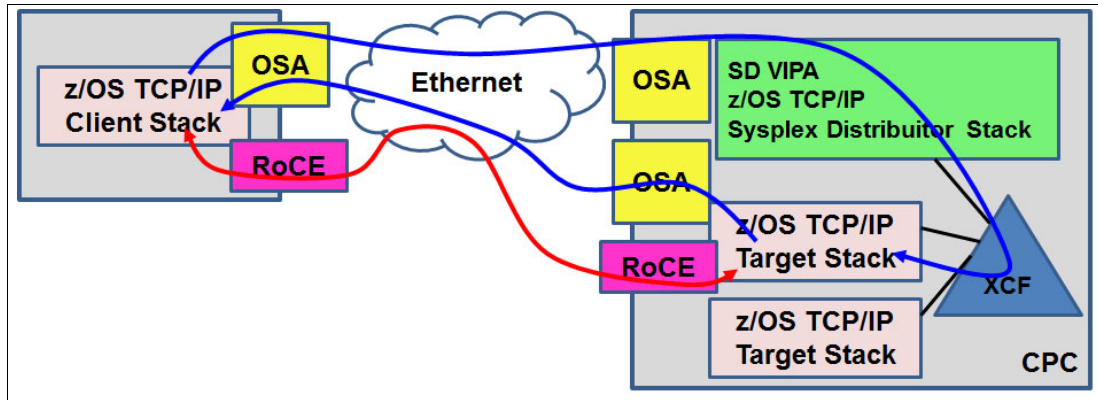


Figure D-10 Sysplex Distributor after RoCE

Note: As with all RoCE communications, the session end also flows over OSAs.

D.2.11 Enabling SMC-R support in z/OS Communications Server

The following checklist provides a task summary for enabling SMC-R support in z/OS Communications Server. This list assumes that you start with an existing IP configuration for LAN access using OSD:

- ▶ HCD definitions (install and configure RNICs in the HCD):
 - Add the PNet ID for the current OSD.
 - Define PFIDs for RoCE (with the same PNet ID).
- ▶ Specify the **GLOBALCONFIG SMCR** parameter (TCP/IP Profile):
 - Must specify at least one PCIe Function ID (PFID):
 - A PFID represents a specific RNIC adapter.
 - A maximum of 16 PFID values can be coded.
 - Up to eight TCP/IP stacks can share a RoCE PCHID (RoCE feature) in a specific LPAR (each stack must define a unique FID value).
- ▶ Start the IPAQENET or IPAQENET6 INTERFACE with CHPIDTYPE OSD:
 - SMC-R is enabled, by default, for these interface types.
 - SMC-R is not supported on any other interface types.

Note: The IPv4 INTERFACE statement (IPAQENET) must also specify an IP subnet mask

- ▶ Repeat in each host (at least two hosts).

Start the TCP/IP traffic and monitor it with Netstat and IBM VTAM displays.

D.3 Shared Memory Communications - Direct Memory Access

This section describes the new Shared Memory Communication - Direct Memory Access (SMC-D) functions implemented in IBM z13 and z13s Driver Level 27 servers.

Concepts

Co-location of multiple tiers of a workload onto a single z Systems physical server allows for the exploitation of HiperSockets, an internal LAN technology that provides low-latency communication between virtual machines within a physical z Systems CPC. HiperSockets is implemented fully within z Systems firmware, so it requires no physical cabling or external network connection to purchase, maintain, or replace. The lack of external components also provides for a secure and low latency network connection because data transfer occurs much like a cross-address-space memory move.

With the z13 (Driver 27) and z13s servers, IBM introduces SMC-D. SMC-D maintains the socket-API transparency aspect of SMC-R so that applications that use TCP/IP communications can benefit immediately without requiring any application software or IP topology changes. SMC-D completes the overall Shared Memory Communications solution, providing synergy with SMC-R. Both protocols use shared memory architectural concepts, eliminating TCP/IP processing in the data path, yet preserving TCP/IP Qualities of Service for connection management purposes.

From an operations standpoint, SMC-D is similar to SMC-R. The objective is to provide consistent operations and management tasks for both SMC-D and SMC-R. SMC-D uses a new virtual PCI adapter called Internal Shared Memory (ISM). The ISM Interfaces are associated with IP interfaces (for example HiperSockets or OSA, ISM interfaces do not exist without an IP interface). ISM interfaces are not defined in software. Instead, ISM interfaces are dynamically defined and created, and automatically started and stopped. You do not need to operate (Start or Stop) ISM interfaces. Unlike RoCE, ISM FIDs (PFIDs) are not defined in software. Instead, they are auto-discovered based on their PNet ID.

SMC-R uses RDMA (RoCE), which is based on Queue Pair (QP) technology:

- ▶ RC-QPs represent SMC Links (logical point-to-point connection).
- ▶ RC-QPs over unique RNICs are logically bound together to form Link Groups (used for HA and load balancing).
- ▶ Link Groups (LGs) and Links are provided in many Netstat displays (for operational and various network management tasks).

SMC-D over ISM does not use QPs:

- ▶ Links and LGs based on QPs (or other hardware constructs) are not applicable to ISM. So the SMC-D information in the Netstat command displays are related to ISM link information rather than LGs.
- ▶ SMC-D protocol (like SMC-R) has a design concept of a “logical point-to-point connection” and therefore preserves the concept of an SMC-D Link (for various reasons that include network administrative purposes).

Note: The SMC-D information in the Netstat command displays is related to ISM link information (not LGs).

D.3.1 Internal Shared Memory technology overview

ISM is a new function supported by the z13 and z13s machines. It is the firmware that provides the connectivity for shared memory access between multiple operating systems within the same CPC. It provides the same functionality as SMC-R but without physical adapters like the RoCE card, using instead virtual ISM devices as SMC-R. It is a Hipersocket like function that provides guest-to-guest communications within the same machine. Figure D-11 shows a possible solution using SMC-D only.

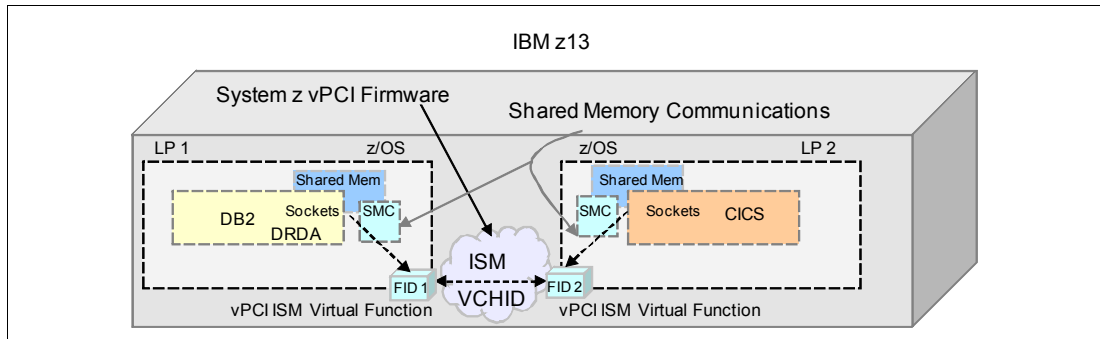


Figure D-11 Connecting two LPARs on the same CPC using SMC-D

Both SMC-D and SMC-R technologies can be used at the same time on the same CPCs.

Figure D-12 shows a fully configured three-tier solution using both SMC-D and SMC-R.

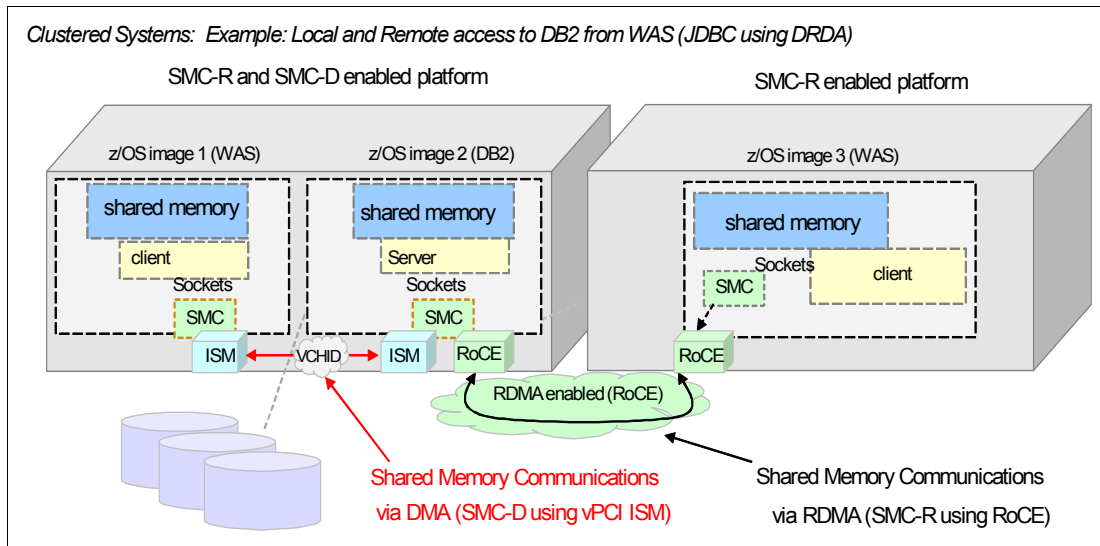


Figure D-12 Clustered systems: Multitier application solution. RDMA, and DMA

D.3.2 SMC-D over Internal Shared Memory

ISM is a virtual channel similar to IQD for Hipersockets. A virtual adapter is created in each OS. Using the SMC protocol, the memory is logically shared. The network is firmware provided. A new device is required to manage that virtual function. SMC is based on a TCP/IP connection and preserves the entire network infrastructure.

SMC-D is a protocol that allows TCP socket applications to transparently use ISM.

SMC-D is a “hybrid” solution as shown in Figure D-13:

- ▶ It uses a TCP connection to establish the SMC-D connection.
- ▶ The TCP connection can be either through the OSA adapter or IQD HiperSockets
- ▶ A TCP option (SMCD) controls switching from TCP to “out of band” SMC-D.
- ▶ The SMC-D information is exchanged within the TCP data stream.
- ▶ Socket application data is exchanged through ISM (write operations).
- ▶ The TCP connection remains to control the SMC-D connection.
- ▶ This model preserves many critical existing operational and network management features of TCP/IP.

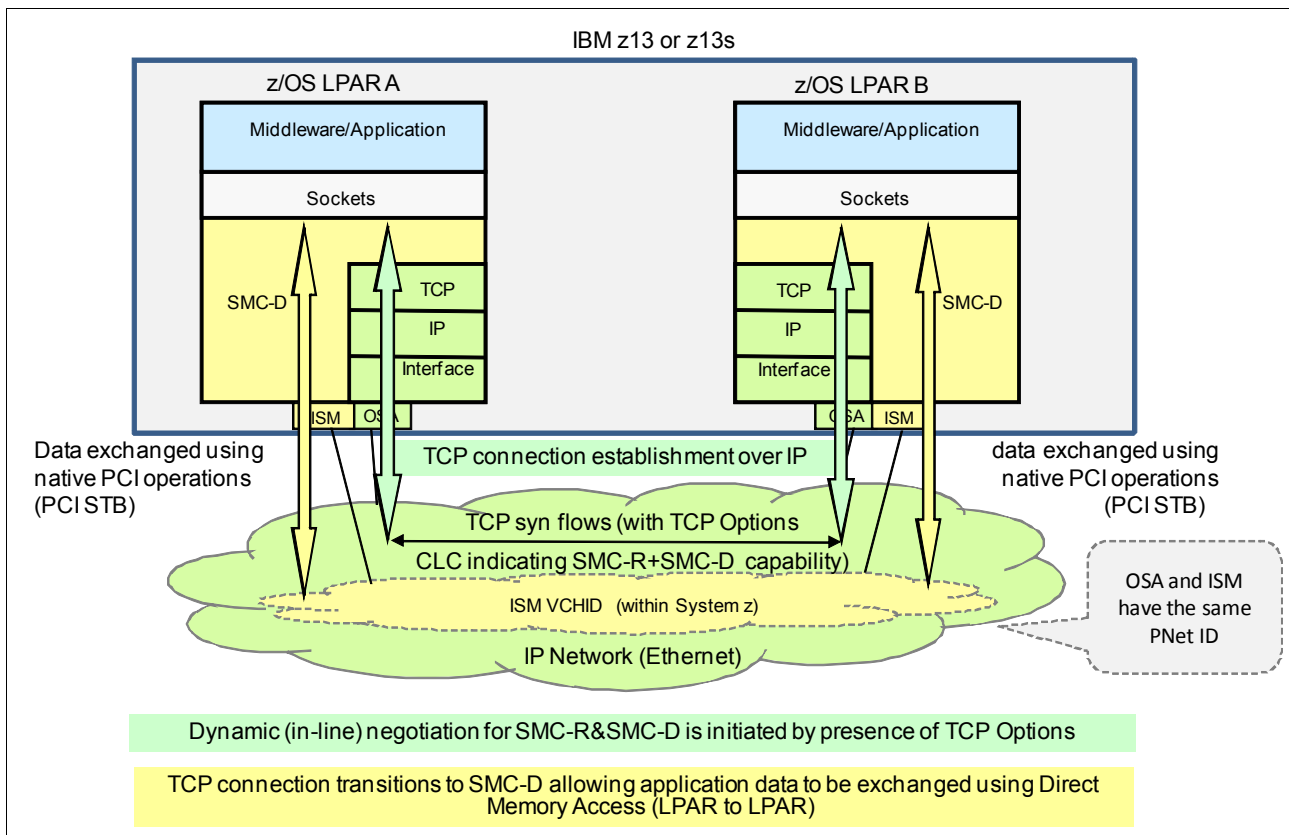


Figure D-13 Dynamic transition from TCP to SMC-D by using two OSA-Express adapters

The hybrid model of SMC-D uses these key existing attributes:

- ▶ It follows the standard TCP/IP connection setup.
- ▶ The hybrid model switches to ISM (SMC-D) dynamically.
- ▶ The TCP connection remains active (idle) and is used to control the SMC-D connection.
- ▶ The hybrid model preserves the following critical operational and network management TCP/IP features:
 - Minimal (or zero) IP topology changes
 - Compatibility with TCP connection-level load balancers

- Preservation of the existing IP security model, such as IP filters, policies, VLANs, and SSL
- Minimal network administration and management changes
- ▶ Host application software is not required to change, so all host application workloads can benefit immediately.

D.3.3 Internal Shared Memory - Introduction

The IBM z13 (Driver 27) and z13s servers introduce the ISM virtual PCI function. ISM is a virtual PCI network adapter that enables direct access to shared virtual memory providing a highly optimized network interconnect for z Systems intra-CPC communications. ISM introduces a new static virtual channel identifier (VCHID) Type. The VCHID is referenced in IOCDS / HCD. The ISM VCHID concepts are similar to the IQD (HiperSockets) type of virtual adapters. ISM is based on existing z Systems PCIe architecture (that is virtual PCI function / adapter). It introduces a new PCI Function Group and type (ISM virtual PCI). There will be a new virtual adapter.

The system admin, configuration, and operations tasks follow the same process (HCD/IOCDS) as existing PCI functions such as RoCE Express, zEDC Express, and so on. ISM supports dynamic I/O.

ISM Provides adapter virtualization (Virtual Functions) with high scalability:

- ▶ It supports up to 32 ISM VCHIDs per CPC (z13 or z13s servers, each VCHID represents a unique internal shared memory network each with a unique Physical Network ID)
- ▶ Each VCHID supports up to 255 VFs per VCHID (the maximum is 8k VFs per z13 or z13s CPC), which provide significant scalability.

Note: There is no concept of a PCI Physical Function to provide virtualization. There is no concept of MACs, or MTU or Frame size.

- Each ISM VCHID represents a unique and isolated internal network, each having a unique Physical Network ID (PNet IDs are configured in HCD/IOCDS).
- ▶ ISM VCHIDs support VLANs, so subdividing a VCHID using virtual LANs is supported.
- ▶ ISM provides a Global Identifier (GID) that is internally generated to correspond with each ISM FID.
- ▶ ISM is supported by z/VM in pass-through mode (PTF required).

D.3.4 Virtual PCI Function (vPCI Adapter)

Virtual Function ID is defined when PCIe hardware is shared between LPARs. Virtual Function ID has a decimal Virtual Function Identifier (VF=) in the range 1 – n, where n is the maximum number of partitions the PCIe feature supports. For example, the SMC-D ISM feature supports up to 32 partitions, and a zEDC Express feature supports up to 15.

The following basic infrastructure is available:

- ▶ zPCI architecture
- ▶ RoCE, zEDC, ISM
- ▶ zPCI layer in z/OS and Linux for z Systems
- ▶ vPCI for SD queues

Figure D-14 shows the basic concept of vPCI adapters:

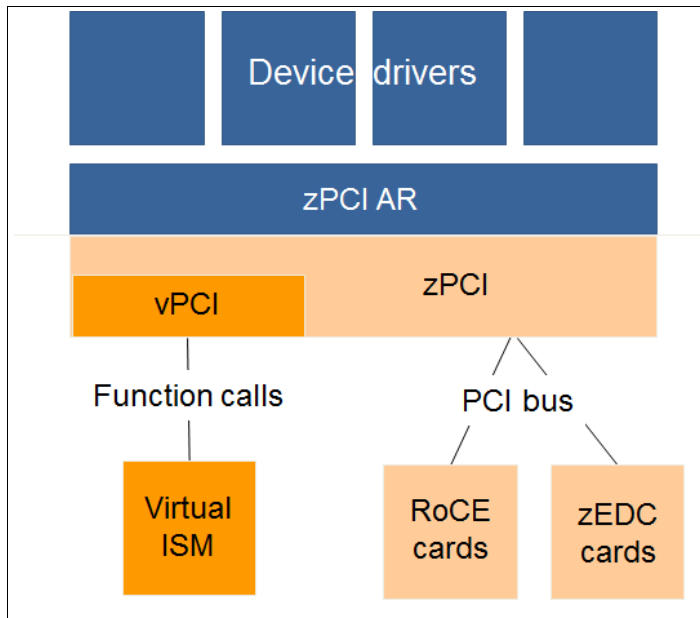


Figure D-14 Concept of vPCI adapter implementation

Note: Basic z/VM support is available:

- ▶ Generic zPCI pass-through support starting from z/VM 6.3
- ▶ The use of the zPCI architecture remains basically unchanged

Figure D-15 shows an SMC-D configuration in which Ethernet provides the connectivity.

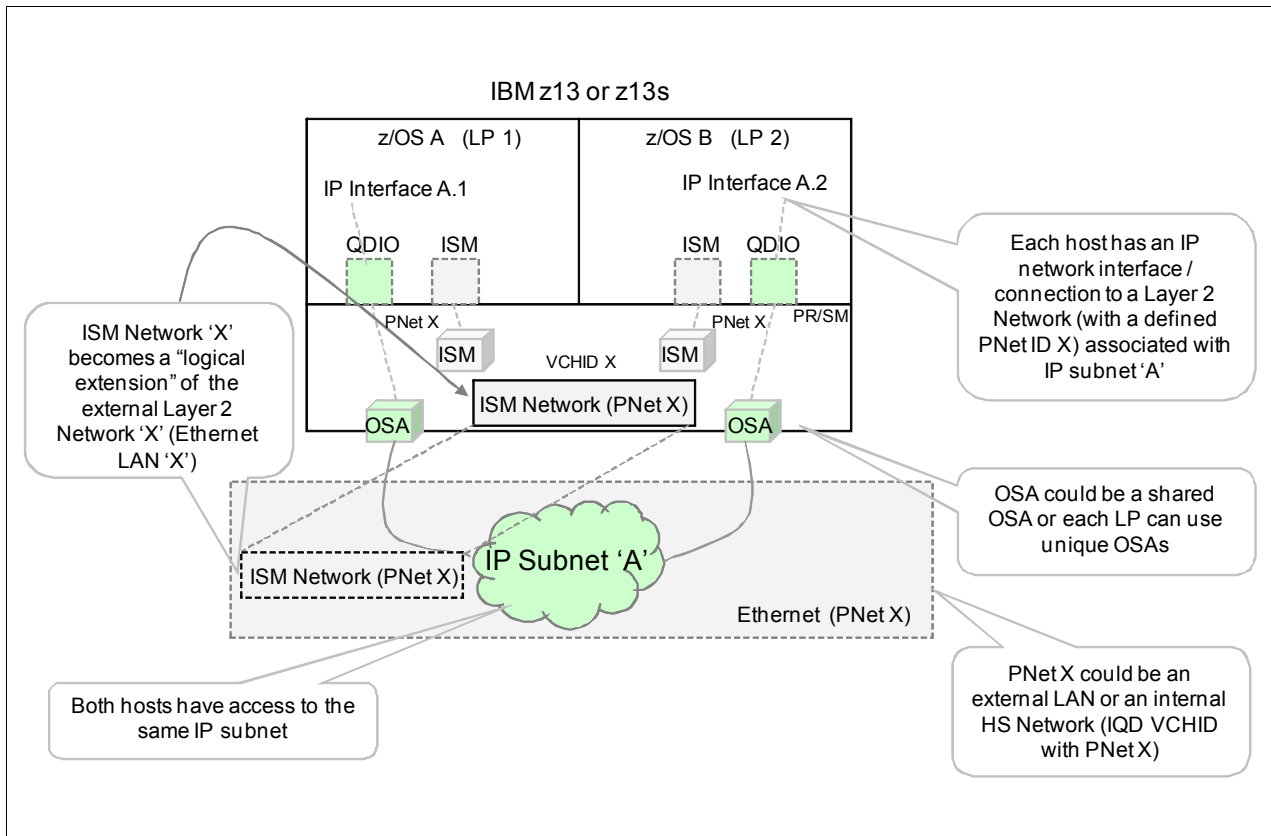


Figure D-15 SMC-D configuration that uses Ethernet to provide connectivity

Figure D-16 shows an SMC-D configuration in which HiperSockets provide the connectivity.

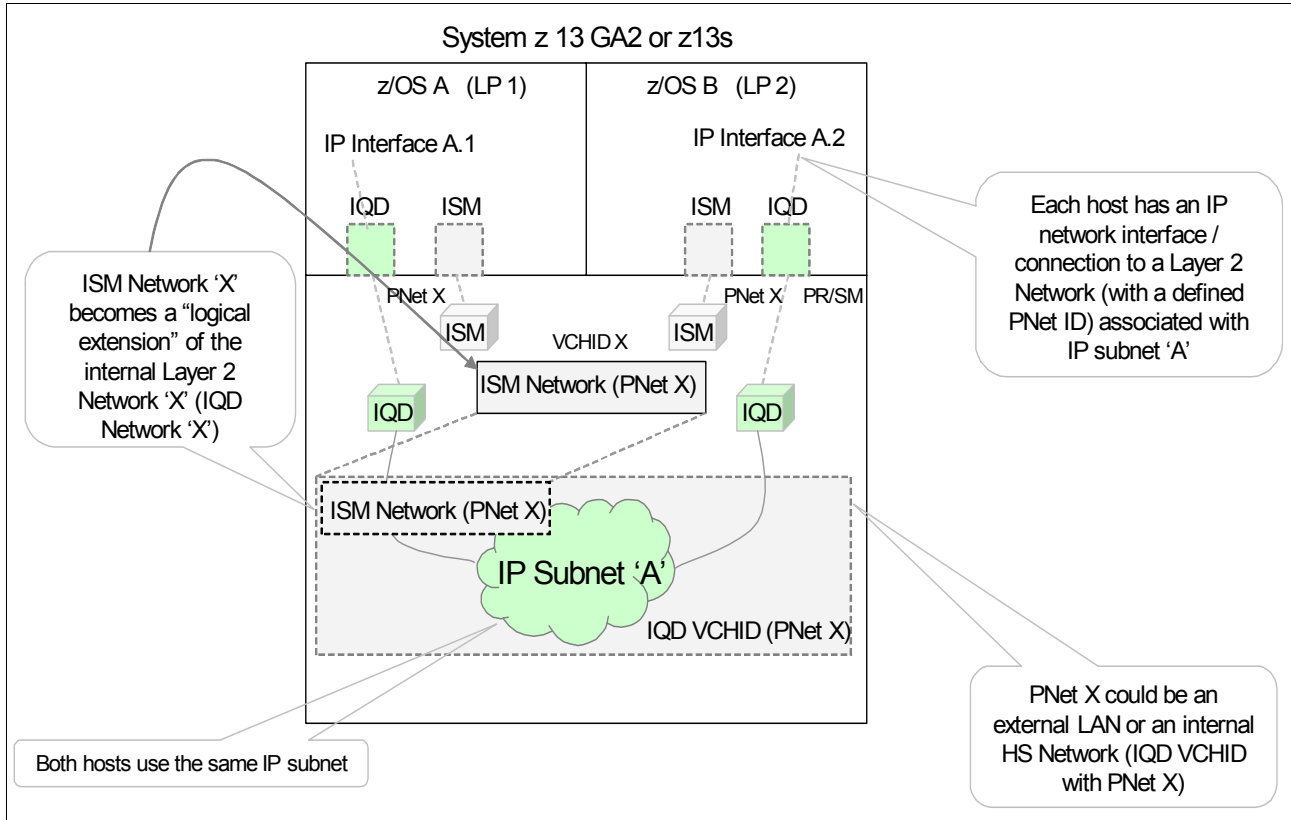


Figure D-16 SMC-D configuration that uses HiperSockets to provide connectivity

D.3.5 Planning considerations

In the z/OS SMC-D implementation, z/OS uses a single Virtual Function (VF) per ISM PNet. This is true for a single VLAN or for multiple VLANs per PNet (that is, the number of VLANs that are defined for a PNet does not affect the number of VFs required).

z/OS Communications Server requires one ISM FID per ISM PNet ID per TCP/IP stack. This requirement is not affected by the version of the IP (that is, it is true even if both IPv4 and IPv6 are used).

The following are reasons why z/OS might use extra ISM FIDs:

- ▶ IBM supports up to 8 TCP/IP stacks per z/OS LPAR. SMC-D can use up to 8 FIDs or VFs (one per TCP/IP stack).
- ▶ IBM supports up to 32 ISM PNet IDs per CEC. Each TCP/IP stack can have access to a PNet ID that consumes up to 32 FIDs (one VF per PNet ID).

D.3.6 Hardware configuration definitions

Complete these steps to create HCDs:

1. Configure ISM vPCI Functions (HCD/HCM).
2. Define PNet IDs (OSA, HiperSockets (IQD) and ISM) in HCD/HCM.
3. Activate the definition using HCD.

4. Enable SMC-D in at least two z/OS instances (single parameter in TCP/IP Global configuration). Both z/OS instances must execute on the same CPC.
5. Review and adjust as needed available real memory and fixed memory usage limits (z/OS and CS). SMC requires fixed memory. You might need to review the limits and provision additional real memory for z/OS.
6. Review IP topology, VLAN usage considerations, and IP security (SMC-R Security White paper).
7. Run the Shared Memory Communications Applicability Tool (SMC-AT) to evaluate applicability and potential value.
8. Review changes to messages, monitoring information and diagnostic tools. Similar to SMC-R there are numerous updates to these items:
 - Messages (VTAM and TCP stack)
 - Netstat (status, monitoring, and display information)
 - CS diagnostic tools (VIT, Packet trace, CTRACE, and IPCS formatted dumps)

Note: There are no application changes (transparent to Socket applications). There are no required optional operation changes (for example starting or stopping devices).

ISM Functions must be associated with another Channel:

- IQD (a single IQD HiperSockets) channel
- OSD channels

Note: A single ISM VCHID cannot be associated with both (IQD and OSD)

D.3.7 Sample IOCP FUNCTION statements

Example D-2 shows IOCP FUNCTION statements that describe the configuration that defines ISM adapters shared between LPARs on the same CPC as shown in Figure D-17.

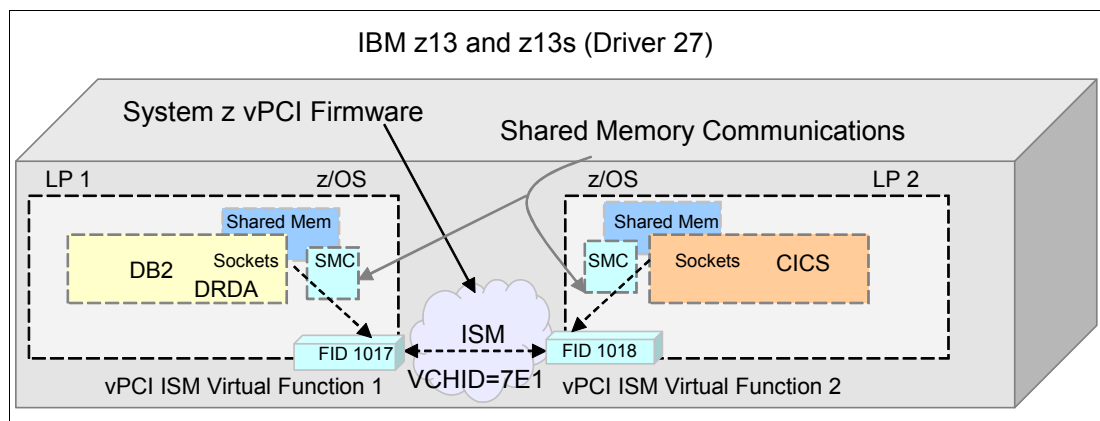


Figure D-17 ISM adapters shared between LPARs

Example D-2 IOCP FUNCTION statements

```
FUNCTION FID=1017,VCHID=7E1,VF=1,PART=((LP1),(LP1,LP2)),PNETID=(PNET1),TYPE=ISM
FUNCTION FID=1018,VCHID=7E1,VF=2,PART=((LP2),(LP1,LP2)),PNETID=(PNET1),TYPE=ISM
```

Note: On the IOCDs statement, the VCHID is defined as 7E1. In Figure D-17 on page 496, the ISM network “PNET 1” is referenced by the IOCDs VCHID statement. ISM (just like IQD) doesn’t use physical cards or card slots (PCHID), but only logical (firmware) instances that are defined as VCHIDs in IOCDs.

Example D-3 shows a sample IOCP FUNCTION configuration that defines ISM adapters shared between LPSRs and multiple VLANs on the same CPC as shown in Figure D-18.

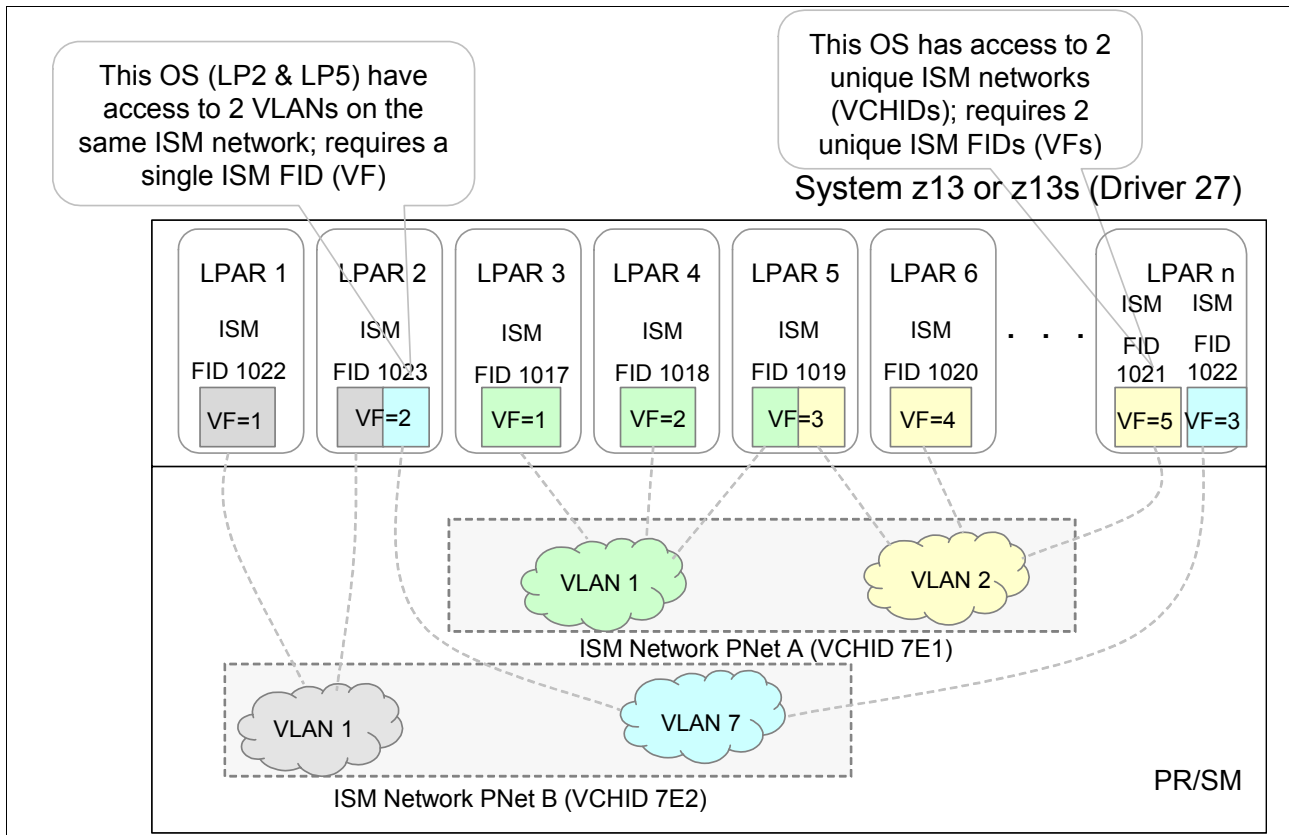


Figure D-18 Multiple LPARs connected through multiple VLANs

Workloads can be logically isolated on separate ISM VCHIDs. Alternatively, workloads can be isolated by exploiting VLANs. The ISM VLAN definitions are inherited from the associated IP network (OSA or HiperSockets).

Example D-3 Sample IOCP Function

```

FUNCTION FID=1017,VCHID=7E1,VF=1,PART=((LPAR3),(LPAR3,LPAR4)),PNETID=(PNETA),TYPE=ISM
FUNCTION FID=1018,VCHID=7E1,VF=2,PART=((LPAR4),(LPAR3,LPAR4)),PNETID=(PNETA),TYPE=ISM
FUNCTION FID=1019,VCHID=7E1,VF=3,PART=((LPAR5),(LPAR4,LPAR5)),PNETID=(PNETA),TYPE=ISM
FUNCTION FID=1020,VCHID=7E1,VF=4,PART=((LPAR6),(LPAR5,LPAR6)),PNETID=(PNETA),TYPE=ISM
FUNCTION FID=1021,VCHID=7E1,VF=5,PART=((LPARn),(LPAR6,LPARn)),PNETID=(PNETA),TYPE=ISM
FUNCTION FID=1022,VCHID=7E2,VF=1,PART=((LPAR1),(LPAR1,LPAR2)),PNETID=(PNETB),TYPE=ISM
FUNCTION FID=1023,VCHID=7E2,VF=2,PART=((LPAR2),(LPAR1,LPAR2)),PNETID=(PNETB),TYPE=ISM
FUNCTION FID=1024,VCHID=7E2,VF=3,PART=((LPARn),(LPAR1,LPARn)),PNETID=(PNETB),TYPE=ISM

```

Configuration considerations

The IOCDS (HCD) definitions for ISM PCI VFs are not directly related to the software (SMC-D) exploitation of ISM (that is, the z/OS TCP/IP and SMC-D implementation and usage are not directly related to the I/O definition).

The user defines a list of ISM FIDs (VFs) in IOCDS (HCD), and z/OS dynamically selects an eligible FID based on the required PNet ID. FIDs or VFs are NOT defined in Communications Server for z/OS TCP/IP. Instead, z/OS selects an available FID for a specific PNET. Access to additional VLANs does not require configuration of additional VFs.

Note: Consider over-provisioning the I/O definitions (e.g. consider defining eight FIDs instead of 5).

For native PCI devices, FIDs need to be defined. Each FID in turn will also define a corresponding VF. In terms of operating system administration tasks, the administrator will typically reference FIDs. Usually VFs (and VF numbers) are transparent.

D.3.8 Software exploitation of ISM

ISM enables SMC-D, which provides SMC capability within the CPC (SMC without requiring RoCE hardware/network equipment). Host virtual memory is managed by each OS (similar to SMC-R, logically shared memory) following existing z Systems PCI I/O translation architecture. Only minor changes required for z/VM guests. An OS can be enabled for both SMC-R and SMC-D. SMC-D is used when both peers are within the same CPC (and ISM PNet and VLAN). After the ISM HCD configuration is complete, SMC-D can be enabled in z/OS with a single TCP/IP parameter (GLOBALCONFIG SMCD). ISM FIDs must be associated with an IP network. The association is accomplished by matching PNet IDs (for example HiperSockets and ISM).

Note: ISM FIDs must be associated with HiperSockets or with an OSA adapter by using a PNet ID. It cannot be associated to both.

D.3.9 SMC-D over ISM prerequisites

SMC-D over ISM has these prerequisites:

- ▶ z13s or z13 server (Driver 27):
 - HMC/SE for ISM vPCI Functions.
- ▶ At least two z/OS V2.2 systems in two LPARs on the same CPC with required service installed:
 - SMC-D can only communicate with another z/OS V2.2 instance and peer hosts must be on the same CPC and ISM PNet.
 - SMC-D requires an IP Network with access through OSA or HiperSockets that has a defined PNet ID that matches the ISM PNet ID.
- ▶ If running as a z/OS guest under z/VM, z/VM 6.3 with APAR VM65716, including APARs is required for guest access to RoCE (Guest Exploitation only).
- ▶ Linux support is planned for a future deliverable.

D.3.11 SMC-D support overview

SMC-D requires IBM z13 and IBM z13s servers to be at driver level 27 or later for support of ISM.

IOCP required level for z13s: The required level of IOCP for the z13s server is V5 R2 L1 or later with PTFs. Defining ISM devices on machines other than the z13 or z13s servers will not be possible. For more information, see the following manuals:

- ▶ *z Systems Stand-Alone Input/Output Configuration Program User's Guide*, SB10-7166.
- ▶ *z Systems Input/Output Configuration Program User's Guide for ICP IOCP*, SB10-7163.

SMC-D requires the following software:

- ▶ z/OS V2R2 with PTFs (See Table D-2 on page 499) or later is the only supported operating systems for the SMC-D protocol:
 - HCD APAR (OA46010) is required.
 - You cannot roll back to previous z/OS releases.
- ▶ z/OS guests under z/VM 6.3 are supported to use SMC-D.
- ▶ IBM is working with its Linux distribution partners to include support in future Linux on z Systems distribution releases.

Other ISM considerations

ISM systems have the following limits:

- ▶ 32 ISM VCHIDs (in IOCDS / HCD) per CPC. Each IOCDS / HCD VCHID represents a unique internal shared memory network each with a unique Physical Network ID.
- ▶ 255 VFs per VCHID (8k VFs per CPC). For example, the maximum number of virtual servers that can communicate over the same ISM VCHID is 255.
- ▶ Each ISM VCHID in IOCDS / HCD represents a unique (isolated) internal network, each having a unique Physical Network ID (PNet IDs are configured in HCD/IOCDS).
- ▶ ISM VCHIDs support VLANs (can be subdivided into VLANs).
- ▶ ISM provides a GID (internally generated) to correspond with each ISM FID.
- ▶ MACs (VMACs), MTU, physical ports, and Frame size are all N/A.
- ▶ ISM is supported by z/VM (for pass-through guest access to support the new PCI function).

Additional documentation

A configuration example for SMC-D is presented in the Redbook *IBM z/OS V2R2 Communications Server TCP/IP Implementation - Volume 1*, SG24-8360.



IBM Dynamic Partition Manager

This appendix contains an introduction to IBM Dynamic Partition Manager (DPM) on z Systems. It provides a description about how Dynamic Partition Manager environment can be initialized and managed.

This appendix includes the following sections:

- ▶ What is IBM Dynamic Partition Manager?
- ▶ Why use DPM?
- ▶ IBM z Systems servers and DPM
- ▶ Setting up the DPM environment

E.1 What is IBM Dynamic Partition Manager?

DPM is a z Systems mode of operation that provides a simplified approach to creating and managing virtualized environments, reducing the barriers to its adoption for new and existing customers. The implementation provides built-in integrated capabilities that allow advanced virtualization management on z Systems servers. With DPM, customers can use their existing Linux and virtualization skills while getting the full value of z Systems hardware: Robustness and security in a workload optimized environment.

DPM provides facilities to define and run virtualized computing systems, using a firmware managed environment that coordinates the physical system resources shared by the partitions¹. The partitions' resources include processors, memory, network, storage, crypto, and accelerators.

DPM provides a new mode of operation for z Systems servers that provides these advantages:

- ▶ Facilitates defining, configuring, and operating partitions similar to the way that you perform these tasks on other platforms.
- ▶ Lays the foundation for a general z Systems new user experience.

DPM is not an additional hypervisor for z Systems servers. DPM uses the existing PR/SM hypervisor infrastructure and, on top of it, provides an intelligent interface that allows customers to define, use, and operate the platform virtualization with no z Systems experience or skills.

Note: When z Systems servers are set to run in DPM mode, only Linux virtual servers can be defined by using the provided user server definition interface. KVM for IBM z Systems is also supported in DPM mode.

E.2 Why use DPM?

DPM mode is targeted at customers (distributed market) with no specific z Systems skills or knowledge of z/VM, who want to implement and use the cloud infrastructure, consolidate and integrate their IT, and ease the management and administration of their Linux environment and workloads.

DPM is of special value for specific customer segments with the following characteristics:

- ▶ New z Systems, or Linux adopters, or distributed driven
 - Likely not z/VM users
 - Looking for integration into their distributed business models
 - Want ease of migration of distributed environments to z Systems servers and improve centralized management
- ▶ Currently not running on z Systems servers
 - No z Systems skills
 - Want to implement cloud
 - Have expectations acquired from another hypervisors' management suite (for example, VMware, KVM, or Citrix)

¹ DPM uses the term "partition", which is the same as logical partition (LPAR).

E.3 IBM z Systems servers and DPM

Traditional IBM z Systems servers are highly virtualized with the goal of maximizing the utilization of compute and I/O (storage and network) resources, and simultaneously lowering the total amount of resources needed for workloads. For decades, virtualization has been embedded in the z Systems architecture and built into the hardware and firmware.

Virtualization requires a hypervisor, which manages resources that are required for multiple independent virtual machines. The z Systems hardware hypervisor is known as IBM Processor Resource/Systems Manager (PR/SM). PR/SM is implemented in firmware as part of the base system. It fully virtualizes the system resources, and does not require additional software to run.

PR/SM allows the defining and managing of subsets of the z Systems resources in logical partitions (LPARs). The LPAR definitions include a number of logical processing units (LPUs), memory, and I/O resources. LPARs can be added, modified, activated, or deactivated in z Systems platforms by using the traditional Hardware Management Console (HMC) interface.

DPM uses all the capabilities that were previously mentioned as the foundation for the new user experience. On top of these capabilities, DPM provides an HMC user interface that allows customers to define, implement, and run Linux partitions without requiring deep knowledge of the underlying z Systems infrastructure management (for example, IOCP or HCD). The DPM infrastructure is depicted in Figure E-1.

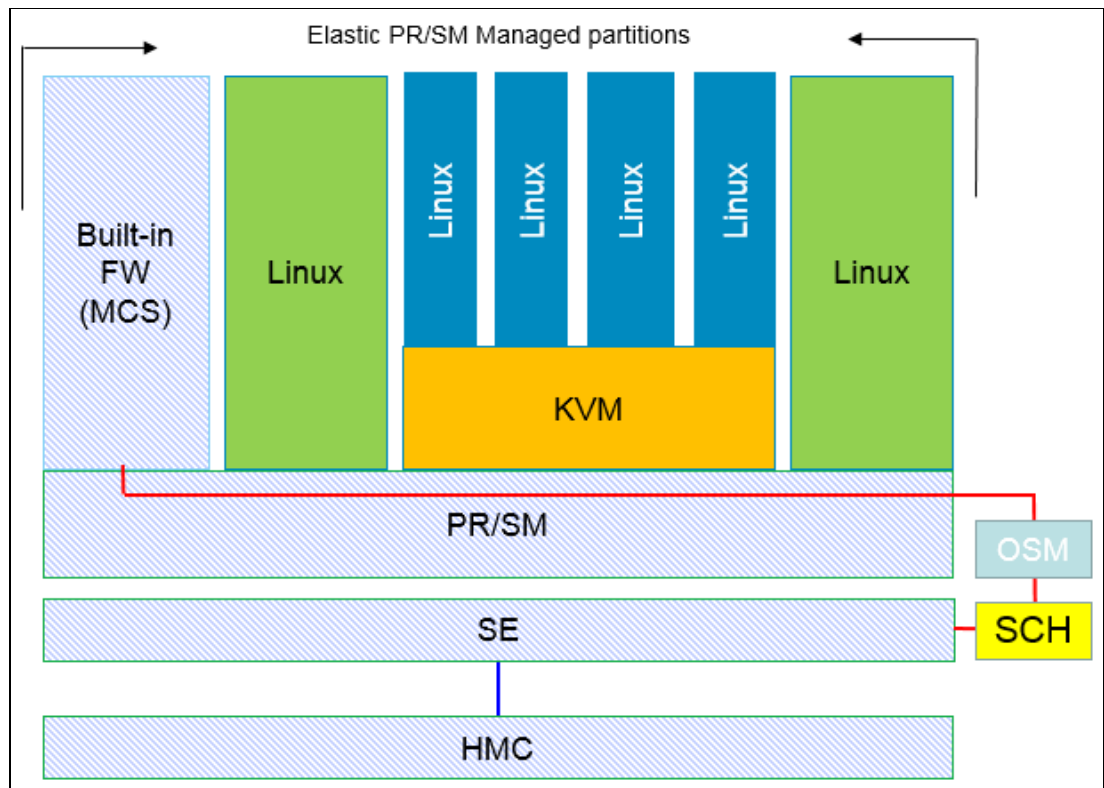


Figure E-1 High-level view of DPM implementation

The firmware partition master control services (MCS), along with the Support Element (SE), provides services to create and manage the Linux native partitions, or partitions that run KVM. The connectivity from the SE to the MCS is provided through the internal management network by two OSA Express5s 1000BASE-T acting as OSA Management adapters.

This implementation integrates platform I/O resource management and dynamic resource management.

E.4 Setting up the DPM environment

The DPM is in fact a z Systems mode of operation. Enabling DPM is a disruptive action. The selection of DPM mode of operation is done by using a function called **Enable Dynamic Partition Manager** under the **CPC Configuration** menu in the SE interface, as shown in Figure E-2. The DPM mode of operation is normally set at machine installation time by the service support representative (SSR).

Note: DPM is a feature code (FC 0016) that can be selected during the machine order process. After they are selected, a pair of OSA Express5s 1000BASE-T adapters must be included in the configuration.

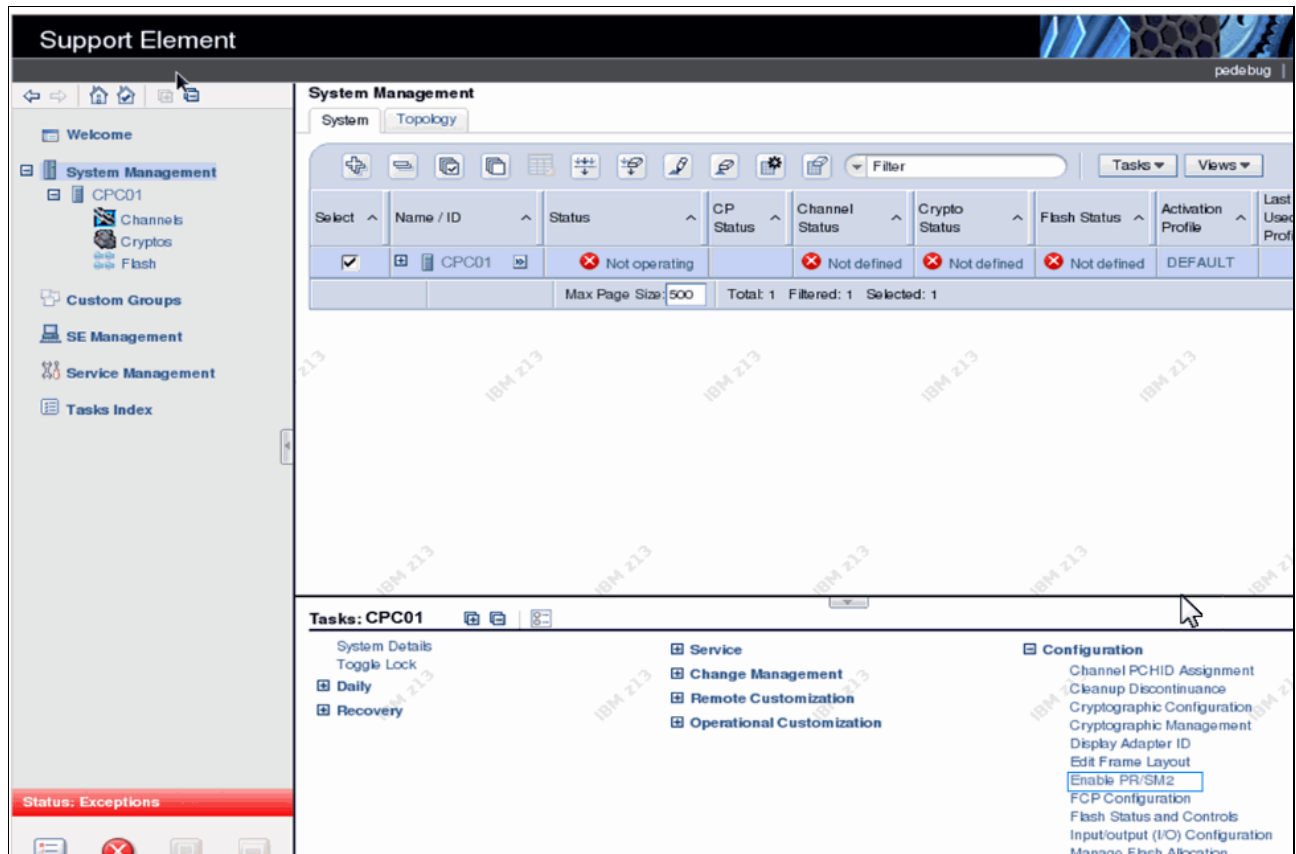


Figure E-2 Enabling DPM mode of operation from the SE CPC configuration options

After this option is selected, a new window is displayed where the user enters the two OSA Express5s 1000BASE-T ports selected and cabled to the System Control Hubs (SCHs) during the z Systems installation. This window is shown in the Figure E-3.

Note: During the installation process, the IBM SSR connects the two OSA Express5s 1000BASE-T cables to the SCHs provided ports.

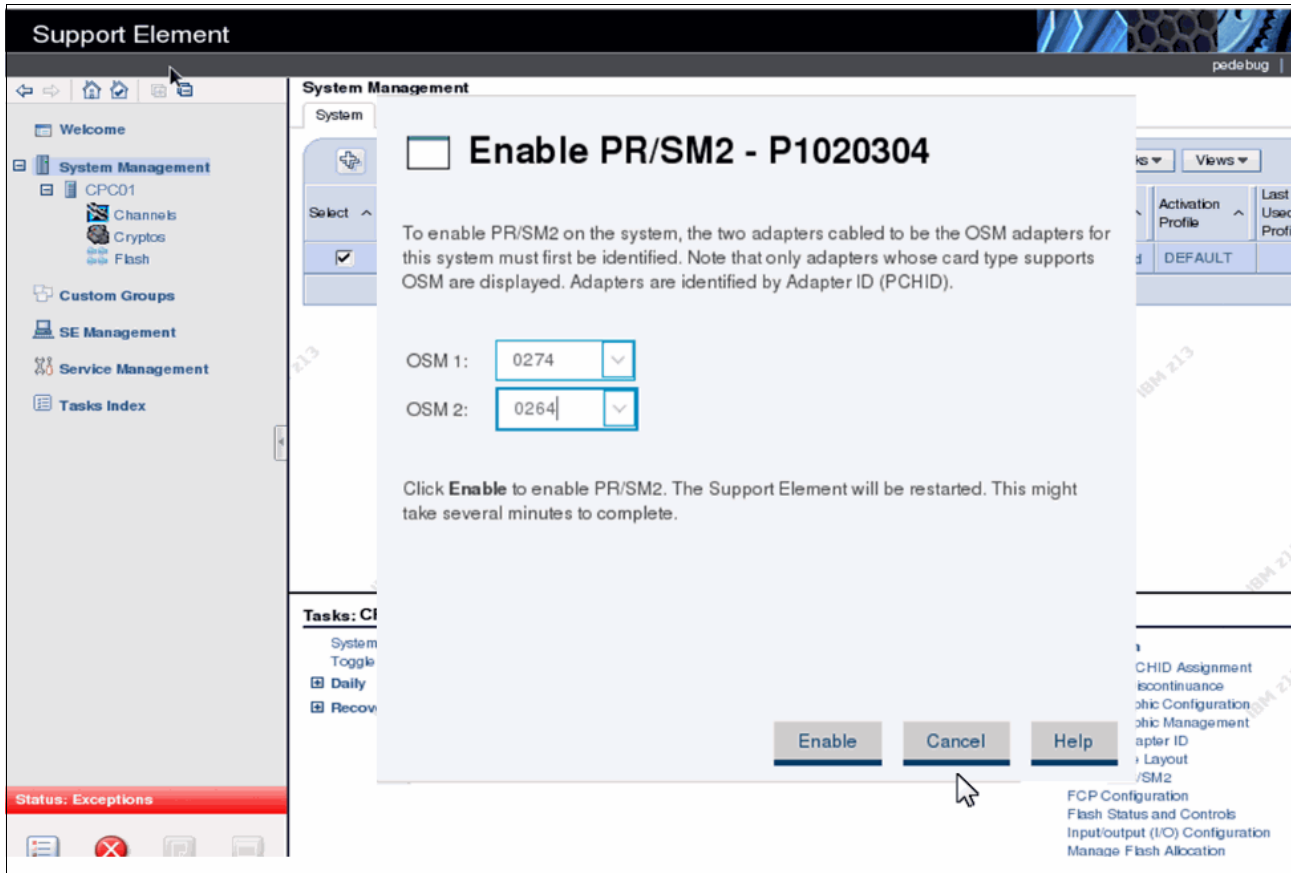


Figure E-3 Entering the OSA ports that will be used by the management network

After entering the OSA adapter port numbers that were previously cabled to the SCHs, click **Enable**. The SE then restarts, and, when finished, the DPM mode becomes active and operational.

Important:

- ▶ A CPC in DPM mode cannot be part of an Ensemble managed by Unified Resource Manager. The HMC used to enable the CPC in DPM mode must NOT be an Ensemble HMC (either Primary or Backup Ensemble HMC).
- ▶ All definitions that are made for the CPC, if any, before the DPM mode is activated are saved and can be brought back if customer chooses to revert to standard PR/SM mode. However, when switching the CPC into standard PR/SM mode, any definitions that were made with the CPC in DPM mode will be lost.

Figure E-4 shows the DPM mode welcome window. The three options at the bottom (**Getting Started**, **Guides**, and **Learn More**) have mouse-over functions that either briefly describe their meaning or provide more functions.

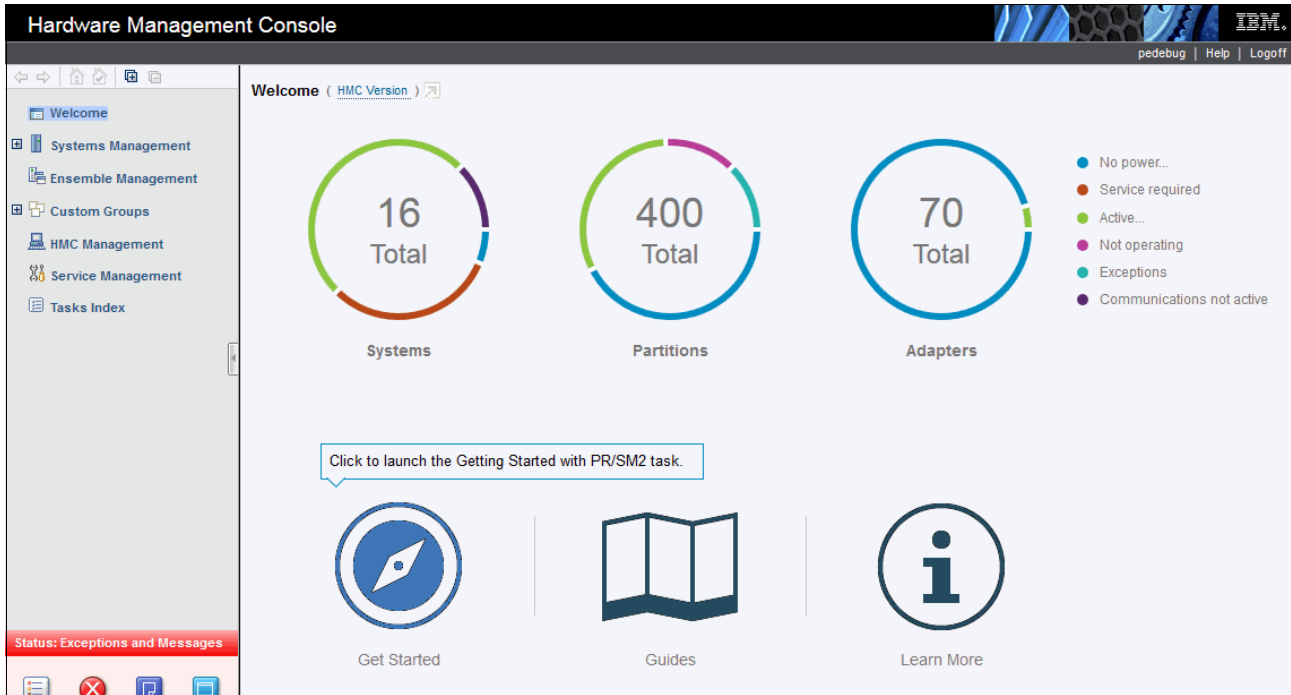


Figure E-4 DPM mode welcome window

The HMC can monitor and control up to 32 z Systems CPCs. The monitored and controlled CPCs need to be *defined* to the HMC by using the Object Definition task, and adding the CPC object.

The welcome window shown in Figure E-4 only appears when at least one HMC-defined CPC is active in DPM mode. Otherwise, the traditional HMC window is presented when you log on to the HMC.

Figure E-5 shows the welcome window from a traditional HMC when none of the defined CPC objects are running in DPM mode.

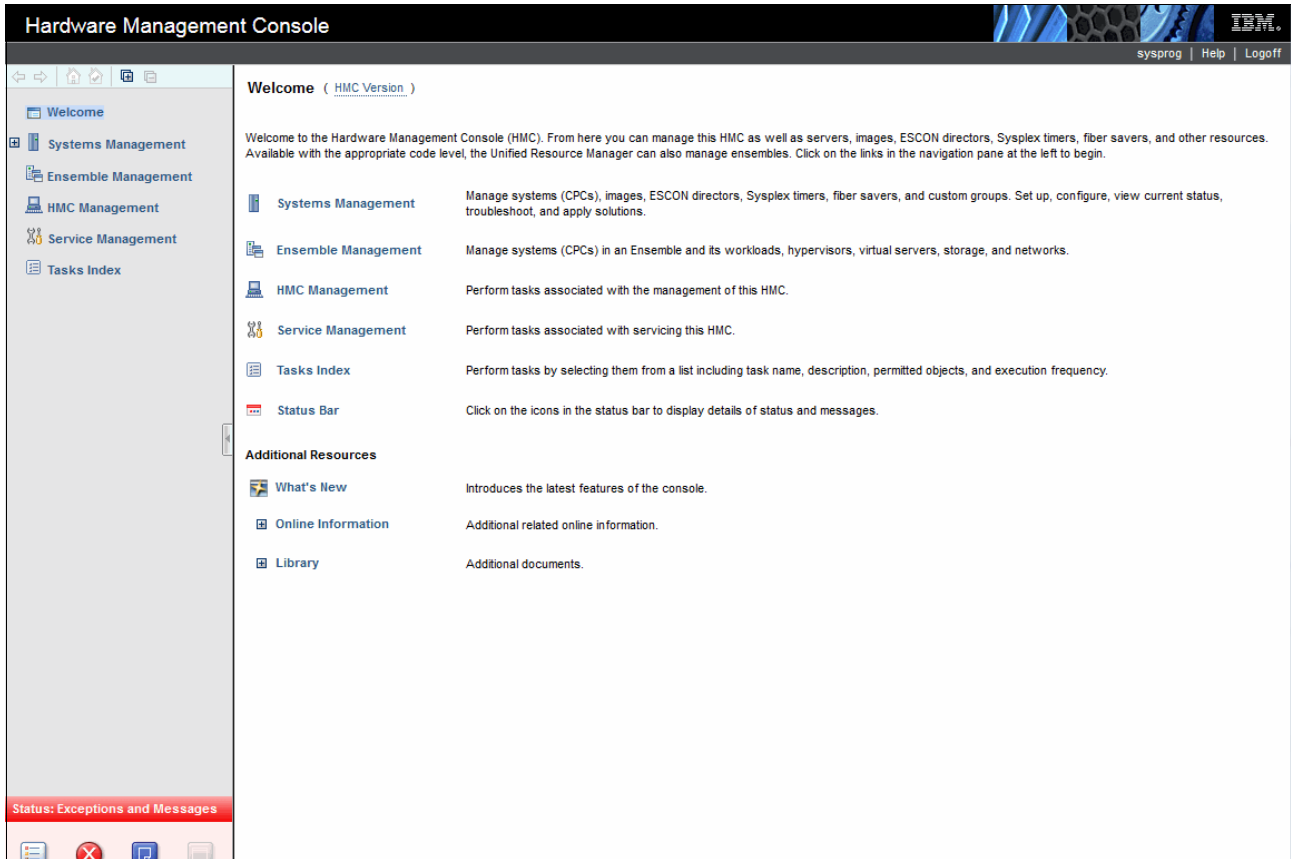


Figure E-5 Traditional HMC Welcome window when no CPCs are running in DPM mode

E.4.1 Defining partitions in DPM mode

After the CPC is in DPM mode, choose one of the options provided in the welcome window to learn more about the process of defining partitions, browse the tutorials, or start creating the environment by defining and activating the partitions.

Figure E-6 shows the available expanded views for two of the three options presented to the user in the HMC welcome page when there is a least one CPC running in DPM mode.



Figure E-6 User Options when the HMC presents the DPM welcome window

The **Guides** option provides Tutorials, videos, and information about What's New in DPM. The **Learn More** option covers the application programming interfaces (APIs), and the **Support** option takes the user to the IBM ResourceLink website.

The first option on the left of the window shown on Figure E-6 on page 507 is **Getting Started**. This option initiates the DPM wizard application on the HMC, which allows users to define their partitions, and then associate processor and memory resources, network and storage I/O, crypto adapters and accelerators to them.

From the Getting Started with DPM window, users can select the **Partition** option, which opens the Create Partition wizard. The Create Partition wizard can also be accessed clicking **Next** at the bottom of the window.

Figure E-7 on page 509 shows, on the left banner, the HMC create partition wizard steps available to define and activate a partition. It contains the following steps:

1. **Welcome:** Initial window that contains basic information about the process.
2. **Name:** This window is used to provide a name and description for partition that is being created.
3. **Processors:** The partition's processing resources are defined in this window.
4. **Memory:** This window is used to define partition's initial and maximum memory.
5. **Network:** This window is where users define the partition's network NICs resources.
6. **Storage:** Where storage connectivity Host Bus Adapters (HBA) are defined.
7. **Accelerators:** Partition resources such as zEDC can be added in this window.
8. **Cryptos:** A wizard window where users define their cryptographic resources
9. **Boot:** In this wizard window, you define the partition's operating system (OS) and that system's source. The following are the options for a source to load an OS:
 - FTP Server
 - Storage Device (SAN)
 - Network Server (PXE)
 - Hardware Management Console removable media
 - ISO image
10. **Summary:** This wizard window provides a view of all of the defined partition resources.

The final step after the partition creation process is to start it. After the partition is started (Status: Active), the user can start the messages or the Integrated ASCII console interface to operate it.

Figure E-7 also shows an option on the lower left called **Advanced**. The advanced option allows users to open a window that contains all definitions made for the partition. This window provides extra settings for some of the definitions, such as defining a processor as shared or dedicated, associating a weight with a partition, and so on.

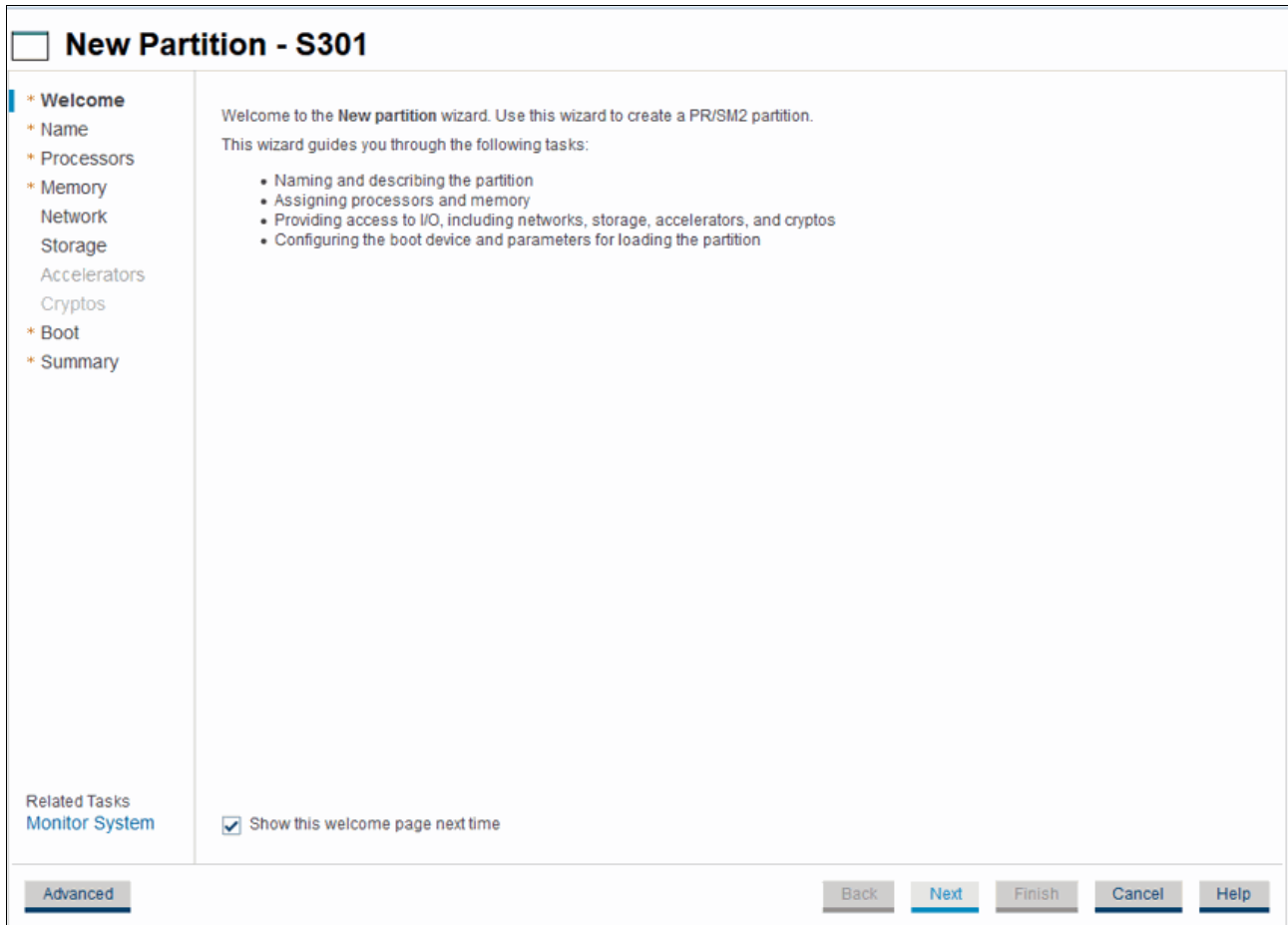


Figure E-7 DPM wizard welcome window options

An important additional facility that is provided by the DPM is **Monitor System**. This option allows users to monitor and manage their DPM environment in many ways and for various events. Some of the monitoring and management capabilities are:

- ▶ Partition overall performance, shown in usage percentages, including:
 - Processors
 - Storage utilization
 - Network adapters
 - Storage adapters
 - Cryptos
 - Accelerators
 - Power Consumption in KW
 - Environmentals - Ambient Temperature in Fahrenheit
- ▶ Adapters that exceed a user predefined threshold value
- ▶ Overall port utilization in the last 36 hours
- ▶ Utilization details (available by selecting one of the performance indicators)
- ▶ Manage Adapters Task

E.4.2 Summary

DPM provides simplified z Systems hardware and virtual infrastructure management that includes integrated dynamic I/O management for users who intend to run Linux on z Systems and KVM for IBM z as hypervisor, running in a partition.

The new mode, DPM, provides partition lifecycle and dynamic I/O management capabilities through the HMC to provide these capabilities:

- ▶ **Create and provision:** Creation of new partitions, assignment of processors and memory, configuration of I/O adapters (Network, FCP Storage, Crypto, and Accelerators).
- ▶ **Manage the environment:** Modification of system resources without disrupting running workloads.
- ▶ **Monitor and troubleshoot the environment:** Source identification of system failures, conditions, states, and events that might lead to workload degradation.

A CPC can be in either the DPM mode or standard PR/SM mode. The mode is enabled prior to the CPC power-on reset (POR).

DPM mode requires two OSA-Express 1000BASE-T Ethernet features for primary and backup connectivity (OSA-Express4S 1000BASE-T Ethernet #0408), along with associated cabling (HW for DPM FC0016).



KVM for IBM z Systems

This appendix contains an introduction to open fertilization with KVM for IBM z Systems and a description of how the environment can be managed.

This appendix includes the following sections:

- ▶ Why KVM for IBM z Systems
- ▶ IBM z Systems servers and KVM
- ▶ Managing the KVM for IBM z Systems environment
- ▶ Using IBM Cloud Manager with OpenStack

F.1 Why KVM for IBM z Systems

Organizations are challenged to drive new business opportunities while reducing budgets, managing increased IT complexity, and improving staff productivity. Although cost considerations are important, businesses are looking at other benefits of open source solutions, such as interoperability, flexibility, and access to the underlying code in their systems to address their challenges.

With KVM for IBM z Systems (KVM for IBM z), IT organizations can unleash the power of kernel-based virtual machine (KVM) open virtualization to improve productivity, and to simplify administration and management for a quick start on their journey to a highly virtualized environment on IBM z Systems servers.

KVM for IBM z Systems is an open source virtualization option for running Linux-centric workloads, using common Linux-based tools and interfaces, while taking advantage of the robust scalability, reliability, and security that is inherent to the IBM z Systems platform. The strengths of the z Systems platform have been developed and refined over several decades to provide more value to any type of IT-based services.

KVM-based virtualization on z Systems servers allows businesses to deploy fewer systems to run more workloads, sharing resources, and improving service levels to meet demand. KVM for IBM z has these advantages:

- ▶ Uses Linux administrative skills to allow for simplicity and familiarity for non-z Systems users, leading to greater operational efficiencies.
- ▶ Enables x86 workload consolidation and deployment on z Systems servers using KVM.
- ▶ Uses standard interfaces to enable single cross-platform virtualization and help simplify systems management.
- ▶ Integrates into existing cloud environments and enables cloud deployments by seamlessly working with OpenStack.
- ▶ Runs your Linux workloads on the z Systems platform, one of the most trusted, scalable, available, reliable, and highly secure platforms.
- ▶ Enhances performance, utilization, and delivery of service by allowing business applications to remain active while the workload is relocated for either load balancing or scheduled downtime.
- ▶ Provides flexibility and automatic provisioning of computing resources, as needed, to meet changing server demands, and to maintain service at a constant, high utilization rate.
- ▶ Provides flexibility and automatic provisioning of computing resources, as needed, to meet changing server demands, and to maintain service at a constant, high utilization rate.

Note: Both KVM for IBM z Systems and Linux on z Systems are the same KVM and Linux that run on other hardware platforms with the same look and feel.

F.1.1 Advantages of using KVM for IBM z Systems

KVM for IBM z Systems offers enterprises a cost-effective alternative to other hypervisors. It has simple and familiar standard user interfaces, offering easy integration of the z Systems platform into any IT infrastructure.

KVM for IBM z Systems can be managed to allow for over-commitment of system resources to optimize the virtualized environment. In addition, KVM for IBM z Systems can help make

platform mobility easier. Its live relocation capabilities enable you to move virtual machines and workloads without incurring downtime.

Table F-1 shows the features that are supported by KVM for IBM z Systems.

Table F-1 Supported features in KVM for IBM z Systems

Feature	Benefits
KVM hypervisor	Supports running multiple disparate Linux virtual machines on a single system
CPU sharing	Allows for the sharing of CPU resources by virtual machines
I/O sharing	Enables the sharing of I/O resources among virtual machines
Memory and CPU over-commitment	Supports the over-commitment of CPU, memory, and swapping of inactive memory
Live virtual machine relocation	Enables workload migration with minimal impact
Dynamic addition and deletion of virtual I/O devices	Reduces downtime to modify I/O device configurations for virtual machines
Thin provisioned virtual machines	Allows for copy-on-write virtual disks to save on storage
Hypervisor performance management	Supports policy based, goal-oriented management and monitoring of virtual CPU resources
Installation and configuration tools	Supplies tools to install and configure KVM for IBM z Systems
Transactional execution exploitation	Provides improved performance

F.2 IBM z Systems servers and KVM

The z Systems platform is highly virtualized, with the goal of maximizing the utilization of compute and I/O (storage and network) resources, and simultaneously lowering the total amount of resources needed for your workloads. For decades, virtualization has been embedded in the z Systems architecture and built into the hardware and firmware.

Virtualization requires a hypervisor, which manages resources that are required for multiple independent virtual machines. Hypervisors can be implemented in software or hardware, and the z Systems platform has both. The hardware hypervisor is known as IBM Processor Resource/Systems Manager (PR/SM). PR/SM is implemented in firmware as part of the base system. It fully virtualizes the system resources, and does not require extra software to run. KVM for IBM z is a software hypervisor that uses PR/SM functions to service its virtual machines.

PR/SM allows the defining and managing of subsets of the z Systems resources in logical partitions (LPARs). Each KVM for IBM z instance runs in a dedicated LPAR. The LPAR definition includes a number of logical processing units (LPUs), memory, and I/O resources. LPUs are defined and managed by PR/SM and are perceived by KVM for IBM z as real CPUs. PR/SM is responsible for accepting requests for work on LPUs and dispatching that

work on physical CPUs. LPUs can be dynamically added to and removed from an LPAR. LPARs can be added, modified, activated, or deactivated in z Systems platforms by using the Hardware Management Console (HMC).

KVM for IBM z Systems also uses PR/SM to access storage devices and the network for Linux on z Systems virtual machines as shown in Figure F-1.

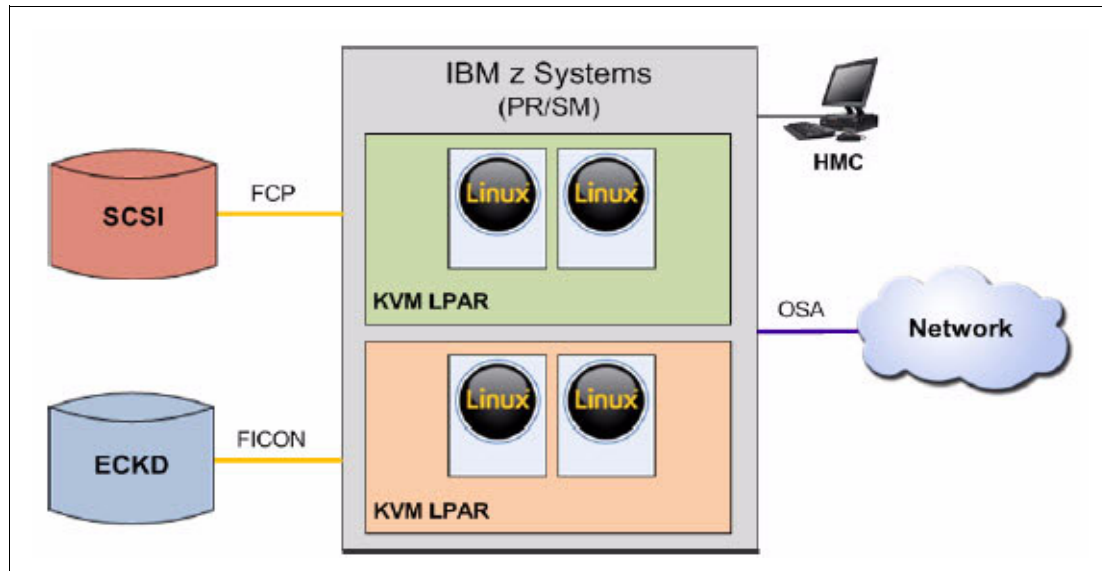


Figure F-1 KVM running in z Systems LPARs

F.2.1 Storage connectivity

Storage connectivity is provided on the z Systems platforms by host bus adapters (HBA) called Fibre Channel connection (IBM FICON) features. IBM FICON (FICON Express16S and FICON Express8S) features follow Fibre Channel (FC) standards. They support data storage and access requirements, and the latest FC technology in storage devices.

The FICON features support the following protocols:

- ▶ Native FICON: An enhanced protocol (over FC) providing for communication with FICON devices such as disks, tapes, and printers. Native FICON supports IBM extended count key data (ECKD) devices.
- ▶ Fibre Channel Protocol (FCP): A standard protocol for communicating with disk and tape devices. FCP supports Small Computer System Interface (SCSI) devices.

Linux on z Systems and KVM for IBM z Systems can make use of both protocols by using the FICON features.

F.2.2 Network connectivity

Network connectivity is provided on the z Systems platform by the network interface cards (NICs) called Open Systems Adapter (OSA) features. The OSA features (OSA-Express5S and OSA-Express4S), provide direct, industry-standard local area network (LAN) connectivity and communications in a networking infrastructure.

OSA features use the z Systems I/O architecture, called queued direct input/output (QDIO). QDIO is a highly efficient data transfer mechanism that uses system memory queues and a

signaling protocol to directly exchange data between the OSA microprocessor and network software.

KVM for IBM z Systems can use the OSA features by virtualizing them for Linux on IBM z Systems to use.

For more information about storage and network connectivity for Linux on z Systems, see *The Virtualization Cookbook for IBM z Systems Volume 3: SUSE Linux Enterprise Server 12*, SG24-8890.

F.2.3 Hardware Management Console

The HMC is a stand-alone computer that runs a set of management applications. The HMC is a closed system, which means that no other applications can be installed on it.

The HMC can set up, manage, monitor, and operate one or more z Systems platforms. It manages and provides support utilities for the hardware and its LPARs.

The HMC is used to install KVM for IBM z and to provide an interface to the IBM z Systems hardware for configuration management functions.

For details about the HMC, see *Introduction to the Hardware Management Console* at:

http://www.ibm.com/support/knowledgecenter/HW11P_2.13.1/com.ibm.hwmca.kc_hmc.doc/introductiontotheconsole/introduction.html

F.2.4 Open source virtualization

KVM technology is a cross-platform virtualization technology that turns the Linux kernel into an enterprise-class hypervisor by using the hardware virtualization support built into the z Systems platform. This configuration means that KVM for IBM z Systems can do things such as scheduling tasks, dispatching CPUs, managing memory, and interacting with I/O resources (storage and network) through PR/SM.

KVM for IBM z Systems creates virtual machines as Linux processes that run Linux on z Systems images using a modified version of another open source module, known as a quick emulator (QEMU). QEMU provides I/O device emulation inside the virtual machine.

The KVM for IBM z Systems kernel provides the core virtualized infrastructure. It can schedule virtual machines on real CPUs and manage their access to real memory. QEMU runs in a user space and implements virtual machines using KVM module functions. QEMU virtualizes real storage and network resources for a virtual machine, which in turn uses *virtio* drivers to access these virtualized resources as shown in Figure F-2.

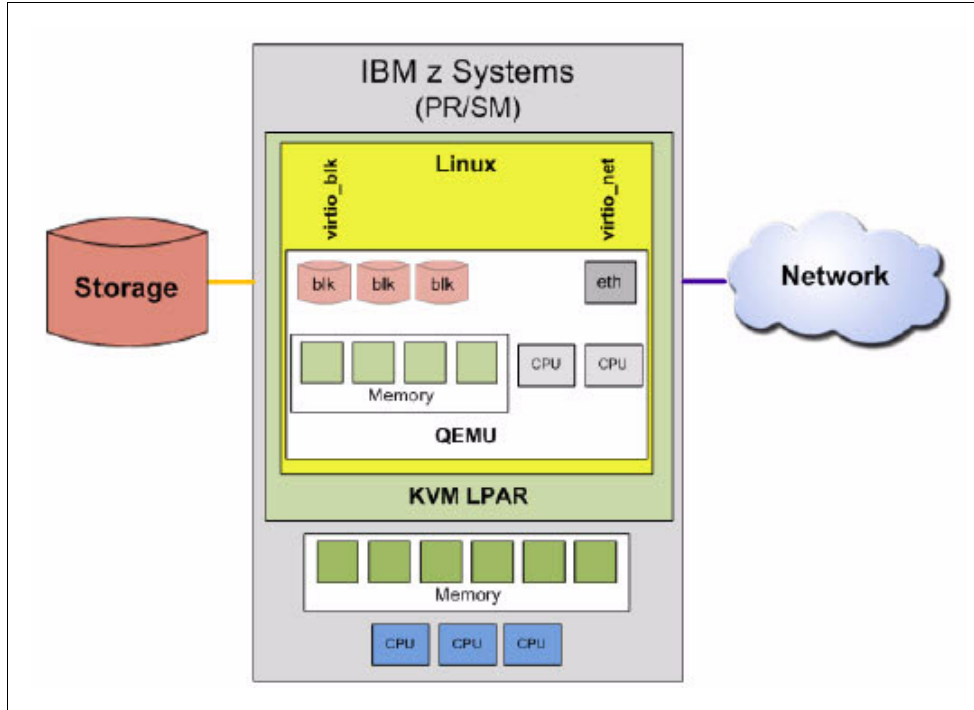


Figure F-2 Open source virtualization (KVM for IBM z Systems)

The network interface in Linux on z Systems is a virtual Ethernet interface. The interface name is *eth*. Multiple Ethernet interfaces can be defined to Linux and are handled by the *virtio_net* device driver module. This module must either be compiled into the Linux kernel or loaded automatically during the boot process.

In Linux, a generic virtual block device is used instead of specific devices, such as ECKD or SCSI devices. The virtual block devices are handled by the *virtio_blk* device driver module. This module must either be compiled into the Linux kernel or loaded automatically during the boot process.

For more information about KVM on IBM z Systems, see:

<http://www.ibm.com/systems/z/solutions/virtualization/kvm/>

F.2.5 What comes with KVM for IBM z Systems

KVM for IBM z Systems provides standard Linux and KVM interfaces for operational control of the environment, such as standard drivers and application programming interfaces (APIs), as well as system emulation support and virtualization management. Shipped as part of KVM for IBM z Systems are the following functions:

- ▶ The command-line interface (CLI) is a common, familiar Linux interface environment that is used to issue commands and interact with the KVM hypervisor. The users issues a series of successive lines of commands to change or control the environment.

- ▶ Libvirt is open source software that is on KVM and many other hypervisors to provide low-level virtualization capabilities that interface with KVM through a CLI called *virsh*. A list of virsh commands can be found in the Table F-2 on page 517

Table F-2 Basic virsh commands

virsh Command	Action
define	Creates a virtual server with the unique name specified in the domain configuration .xml file.
start	Starts a defined virtual server. Using the --console option grants initial access to the virtual server console and displays all messages that are issued to the console.onsole.
shutdown	Terminate a running virtual server, sending a shutdown signal to the VM. This allows proper shutdown of an operating system of a virtual machine (VM).
destroy	Immediately terminates a virtual server without any interaction with the operating system running on a VM.
undefine	Deletes the definition of a virtual server from libvirt.
list	Without an option, this lists the running virtual servers. With the --all option, this command lists all of the defined virtual servers.
edit	Opens the libvirt internal definition of a VM and allows it to be changed. These changes are not applied dynamically. Instead, they become effective after a restart of the VM.

Virsh is the main CLI for libvirt for managing virtual machines. A complete list of supported virsh commands in KVM for IBM z with detailed descriptions is listed in *KVM Virtual Server Management*, SC34-2752.

The Hypervisor Performance Manager (HPM) monitors virtual machines running on KVM

- ▶ The HPM monitors virtual machines running on KVM to achieve goal-oriented policy-based performance goals. For more information about HPM, see *Getting Started with KVM for IBM z Systems*, SG24-8332.
- ▶ Open vSwitch (OVS) is open source software that allows for network communication between virtual machines that are hosted by a KVM hypervisor.
<https://www.openswitch.org>
- ▶ MacVtap is a device driver that is used to virtualize bridge networking and is based on the mcvlan device driver.
<http://virt.kernelnewbies.org/MacVtap>
- ▶ QEMU is open source software that is a hardware emulator for virtual machines that run on KVM. It also provides management and monitoring function for the KVM virtual machines.
<http://wiki.qemu.org>
- ▶ The installer is a series of windows to assist and guide the user through the installation process. Each window has setting selections that can be made to customize the KVM installation. For more information about KVM installation process, see *Getting Started with KVM for IBM z Systems*, SG24-8332.
- ▶ Nagios Remote Plugin Executor (NRPE) can be used with KVM for IBM z. NRPE is an add-on that allows you to run plug-ins on KVM for IBM z. You can monitor resources, such as disk usage, CPU load, and memory usage. For more information about how to

configure the nagios monitoring, see *Getting Started with KVM for IBM z Systems*, SG24-8332.

F.3 Managing the KVM for IBM z Systems environment

KVM for IBM z Systems integrates with standard OpenStack virtualization managements, enabling enterprises to easily integrate Linux servers into their infrastructure and cloud offerings.

KVM for IBM z Systems supports libvirt APIs, enabling CLIs (and custom scripting) to be used to administer the hypervisor. Furthermore, KVM can be administered by using open source tools such as virt-manager or OpenStack. KVM for IBM z Systems can also be administered and managed by using IBM Cloud Manager with OpenStack (Figure F-3). IBM Cloud Manager is created and maintained by IBM and built on OpenStack.

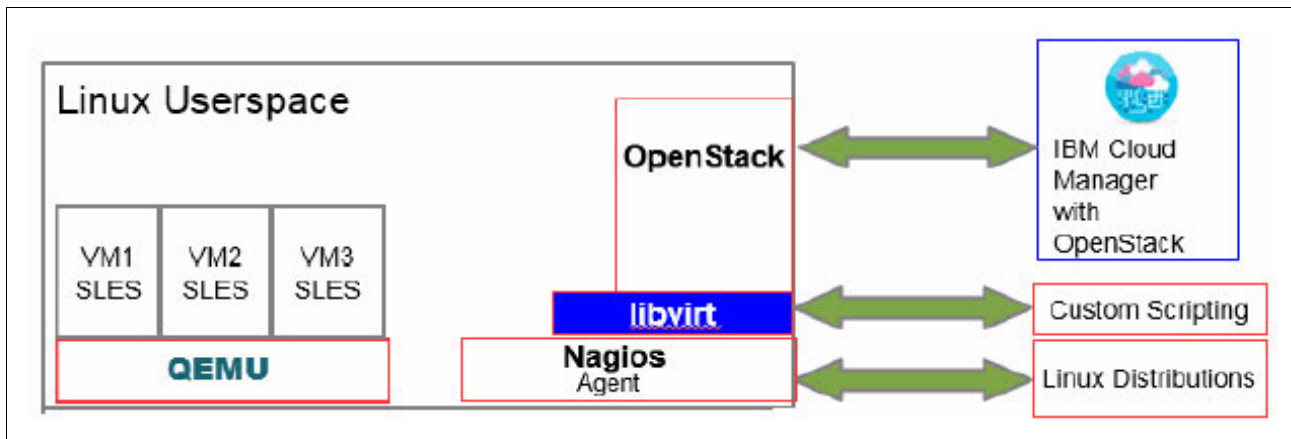


Figure F-3 KVM for IBM z Systems management interface

KVM for IBM z Systems can be managed just like any another KVM hypervisor by using the Linux CLI. The Linux CLI provides a familiar experience for platform management.

In addition, an open source tool that is called Nagios can be used to monitor the KVM for IBM z environment.

libvirt provides different methods of access, from a command line called **virsh** to a low-level API for many programming languages (see Figure F-4).

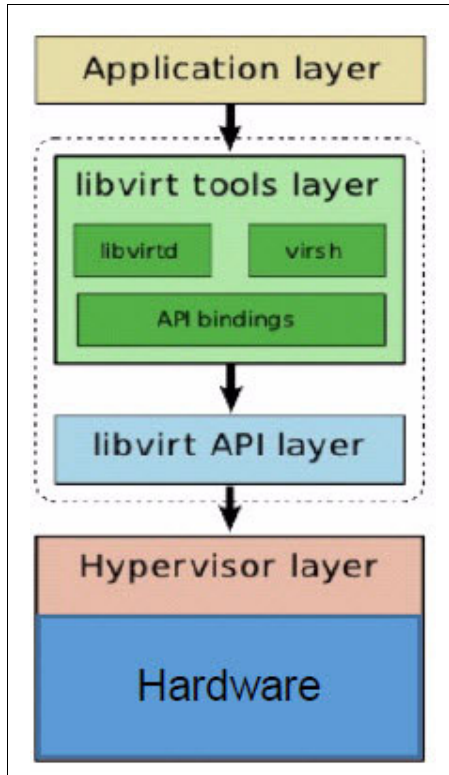


Figure F-4 KVM management by using the libvirt API layers

The main component of the libvirt software is the *libvirtd* daemon. This component interacts directly with QEMU and the KVM kernel.

QEMU manages and monitors the KVM virtual machines by performing the following tasks:

- ▶ Manage the I/O between virtual machines and KVM
- ▶ Create virtual disks
- ▶ Change the state of a virtual machine:
 - Start a virtual machine
 - Stop a virtual machine
 - Suspend a virtual machine
 - Resume a virtual machine
 - Delete a virtual machine
 - Take and restore snapshots

For more information about libvirt go to:

<http://libvirt.org>

F.3.1 Hypervisor Performance Manager

HPM monitors and manages workload performance of the virtual machines under KVM by performing the following operations:

- ▶ Detect when a virtual machine is not achieving its goals when it is a member of a Workload Resource Group.
- ▶ Determine whether the virtual machine performance can be improved with more resources.
- ▶ Project the impact on all virtual machines of the reallocation of resources.
- ▶ Redistribute processor resources if there is a good trade-off based policy.

F.4 Using IBM Cloud Manager with OpenStack

OpenStack is a cloud-based operating system that controls large pools of compute, storage, and networking resources throughout a data center and is based on the Open Stack project:

<http://www.openstack.org/>

IBM Cloud Manager with OpenStack is an advanced management solution that is created and maintained by IBM and built on OpenStack. IBM Cloud Manager with OpenStack can be used to get started with a cloud environment and continue to scale with users and workloads, providing advanced resource management with simplified cloud administration and full access to OpenStack APIs.

KVM for IBM z Systems compute nodes support these items:

- ▶ Nova, which is a libvirt driver
- ▶ Neutron agent for Open vSwitch
- ▶ Ceilometer support
- ▶ Cinder support

OpenStack compute node has an abstraction layer for compute drivers to support different hypervisors, including QEMU and KVM for IBM z through the libvirt API layer (Figure F-4 on page 519).



Native Peripheral Component Interconnect Express

This appendix introduces the design of native Peripheral Component Interconnect Express (PCIe) features management on z13s and includes the concepts of the integrated firmware processor (IFP) and resource groups (RGs).

The following topics are included:

- ▶ Design of native PCIe I/O adapter management
- ▶ Native PCIe feature plugging rules
- ▶ Native PCIe feature definitions

G.1 Design of native PCIe I/O adapter management

The native PCIe adapter is a new category of features introduced since zEC12. These features are 10GbE Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE) Express, zEnterprise Data Compression (zEDC) Express, and Flash Express. These adapters are exclusively installed into a PCIe I/O drawer, together with the existing I/O features and have a physical channel ID (PCHID) assigned according to its physical location.

For all the feature adapters installed in an I/O drawer, management functions in the form of device drivers and diagnostic tools are always implemented to support virtualization of the adapter, service, and maintenance.

Traditionally, these management functions are integrated on the adapter with specific hardware design. For the newly introduced native PCIe adapters, these functions are moved out of the adapter and are now handled by an IFP.

This section covers the following topics:

- ▶ Native PCIe adapter
- ▶ Integrated firmware processor
- ▶ Resource groups
- ▶ Management tasks

Note: Management of Flash Express feature is not covered in this section. All succeeding topics apply to RoCE Express and zEDC Express only.

G.1.1 Native PCIe adapter

For traditional I/O adapters, such as the Open Systems Adapter (OSA) and Fibre Channel connection (FICON) cards, the application-specific integrated circuit (ASIC) chip on the adapter always downloads the device drivers and diagnostic tools from the Support Element (SE) and runs the management functions on the adapter. In the new design, there is no ASIC chip for management function on the native PCIe feature adapters.

For the RoCE and zEDC, device drivers and diagnostic tools are now running on the IFP and use two RGs. Management functions, including virtualization, servicing and recovery, diagnostics, failover, firmware updates against an adapter, and other functions are still implemented.

G.1.2 Integrated firmware processor

The IFP is a processor unit (PU) exclusively used to manage native PCIe feature adapters that are installed in the PCIe I/O drawer. On previous systems, this processor was not used but known as a *reserved processor*. It is allocated from the system PU pool and is not counted in the PUs available for characterization.

If a native PCIe feature is installed in the system, the system allocates and initializes an IFP during its power-on reset (POR) phase. Although the IFP is allocated to one of the physical PUs, it is not visible to the users. In an error or failover scenario, PU sparing also happens for an IFP, with the same rules as other PUs.

G.1.3 Resource groups

The IFP allocates two RGs for running the management functions of native PCIe feature adapters. A native PCIe feature adapter is managed by one of the resource groups according to which I/O domain this adapter is in.

As shown in Figure G-1, each I/O domain in a PCIe I/O drawer of z13s is logically attached to one of the two resource groups: I/O domain 0 and 2 in the front of the drawer attached to RG1 and I/O domain 1 and 3 attached to RG2. The native PCIe I/O feature adapters in I/O domain 0 and 2 are managed by RG1 for device drivers and diagnostic tools functions, and adapters in I/O domain 1 and 3 are managed by RG2.

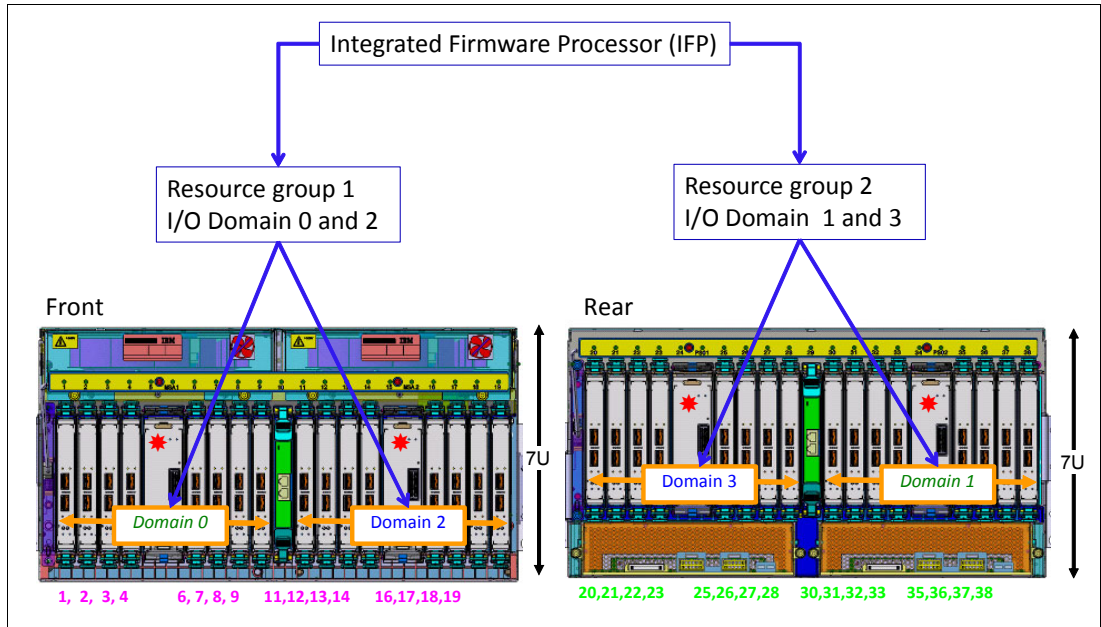


Figure G-1 I/O domains and resource groups managed by the IFP

Up to two PCIe I/O drawers are supported on the z13s. For multiple PCIe features of the same type, each feature is always assigned to a different I/O domain (of the two domains implemented), and, if available, installed in a different PCIe I/O drawer, to eliminate the possibility of a single point of failure.

Currently, an I/O domain of the PCIe I/O drawer can support a total of two native PCIe feature adapters, together with the existing PCIe feature cards (FICON, OSA, and Crypto).

The following native PCIe features are supported in the PCIe I/O drawer:

- ▶ Flash Express
- ▶ zEDC Express
- ▶ 10GbE RoCE Express

Though only zEDC Express and 10GbE RoCE Express features are managed by the IFP and resource groups, a Flash Express feature is always counted as a native PCIe feature installed in a PCIe I/O drawer.

G.1.4 Management tasks

The IFP and resource groups perform all management tasks on the native PCIe adapters:

- ▶ Firmware update of adapters and resource groups
- ▶ Error recovery and failure data collection
- ▶ Diagnostic and maintenance tasks

Firmware update of adapters and resource groups

Firmware of native PCIe adapters and resource groups are part of the system's microcode and can be updated by a Microcode Change Level (MCL) upgrade. MCL upgrades on adapters or on the code of the resource groups require the specific adapter or all native PCIe adapters managed by the specific resource group (depending on the type of u-code that it applies) to be offline during activation of the MCL. However, to maintain availability, MCLs can only be applied to one resource group at a time. While one resource group is offline, the second resource group and all adapters in it remain active. An MCL application for a native PCIe adapter or resource group is not possible if an error condition exists within the other resource group.

Error recovery and failure data collection

If an error in one of the resource groups or features assigned to one of the resource groups occurs, the IFP manages error recovery and collects error data. The error data is sent by the IFP to the SE, which then provides a message on the SE and the Hardware Management Console (HMC). If an error requires maintenance, a call home to the IBM Support system is started by the HMC.

Diagnostic and maintenance tasks

Any maintenance action on a native PCIe feature is managed by the IFP, including testing or replacing a feature card. Before configuring a feature offline, the IFP ensures that the same type of feature is available in the same or the other resource group (if applicable).

G.2 Native PCIe feature plugging rules

The following are maximum number of native PCIe adapters that can be installed in a z13s:

- ▶ Maximum of four Flash Express features, each of which requires two Flash Express adapters configured, and can only be installed into slot 1 and 14, or 25 and 33 of a PCIe I/O drawer.
- ▶ Maximum of 16 10GbE RoCE Express features, each of which has one adapter configured.
- ▶ Maximum of eight zEDC Express features, each of which has one adapter configured.

Table G-1 shows the combinations of maximum number of native PCIe features on a z13s.

Table G-1 Combinations of maximum number of native PCIe features on z13s

Number of Flash Express features (2 adapters/feature)	Maximum number of zEDC Express features	Maximum number of RoCE Express features	Minimum number of PCIe I/O drawers required
0	8	16	1
1	8	16	1

Number of Flash Express features (2 adapters/feature)	Maximum number of zEDC Express features	Maximum number of RoCE Express features	Minimum number of PCIe I/O drawers required
2	8	16	1
3 ^a	8	16	2
4 ^a	8	16	2

a. Maximum number of Flash Express per drawer is two.

Considering availability, install adapters of the same type in slots of different I/O domains, drawers, fanouts, and resource groups (for RoCE Express and zEDC Express). The next sections describe more information about achieving a highly available configuration.

G.3 Native PCIe feature definitions

During the ordering process of the native PCIe adapters, such as the zEDC Express and 10GbE RoCE Express features, features of the same type are evenly spread across two resource groups (resource group 1 and resource group 2) for availability and serviceability.

Notes:

- ▶ Although Flash Express features are counted as native PCIe cards for the total number of Native PCIe features, they are not part of any resource group.
- ▶ The Flash Express features are not defined by using the input/output configuration data set (IOCDs).

Figure G-2 shows a sample PCHID report of a z13s configuration with four zEDC Express features and four 10GbE RoCE Express features.

Source	Drwr	Slot	F/C	PCHID/Ports or AID	Comment
A21/LG12/J01	A13B	03	0420	108	RG1
A21/LG12/J01	A13B	04	0420	10C	RG1
A21/LG12/J01	A13B	06	0411	110/D1D2	RG1
A21/LG12/J01	A13B	07	0411	114/D1D1	RG1
A21/LG14/J01	A13B	35	0411	170/D1D2	RG2
A21/LG14/J01	A13B	36	0411	174/D1D2	RG2
A21/LG14/J01	A13B	37	0420	178	RG2
A21/LG14/J01	A13B	38	0420	17C	RG2

Figure G-2 Sample output of AO data or PCHID report

Figure G-2 on page 525 lists the following information for each adapter:

- ▶ PCHID and ports
- ▶ Resource Group the adapter is attached to (Comment column)
- ▶ Physical location (drawer, slot)

The report shows a balanced configuration where I/O domains and resource groups are distributed for both types of features:

- ▶ I/O domain 0 of drawer Z22B, attached to resource group RG1, has:
 - Two zEDC Express features, installed in slot 3 and 4
 - Two 10GbE RoCE Express features, installed in slot 6 and 7
- ▶ I/O domain 1 of drawer Z22B, attached to resource group RG2, has:
 - Two zEDC Express features, installed in slot 37 and 38
 - Two 10GbE RoCE Express features, installed in slot 35 and 36

Note: F/C 0420 = zEDC Express, F/C 0411 = 10GbE RoCE Express

The native PCIe features are not part of the traditional channel subsystem (CSS). They do not have a channel-path identifier (CHPID) assigned, but they have a PCHID assigned according to the physical location in the PCIe I/O drawer.

To define the native PCIe adapters in the HCD or HMC, a new I/O configuration program (IOCP) FUNCTION statement is introduced that includes several feature-specific parameters.

The IOCP example in Figure G-3 on page 527 define zEDC Express and 10GbE RoCE Express features to LPARs LP14 and LP15 as listed below:

- ▶ Both LPARs have access to two zEDC Express features with redundancy:
 - PCHID 108 in I/O domain 0 of drawer Z22B, in resource group RG1
 - PCHID 178 in I/O domain 1 of drawer Z22B, in resource group RG2
- ▶ Both LPARs have access to two 10GbE RoCE Express features with redundancy:
 - PCHID 110 in I/O domain 0 of drawer Z22B, in resource group RG1
 - PCHID 170 in I/O domain 1 of drawer Z22B, in resource group RG2
- ▶ Both LPARs have access to both networks, which PNETID as NET1 and NET2.

```

zEDC Express Functions for LPAR LP14, Reconfigurable to LP01:
FUNCTION FID=05,VF=1,PART=((LP14),(LP01)),TYPE=ZEDC,PCHID=108
FUNCTION FID=06,VF=1,PART=((LP14),(LP01)),TYPE=ZEDC,PCHID=178

zEDC Express Functions for LPAR LP15, Reconfigurable to LP02:
FUNCTION FID=07,VF=2,PART=((LP15),(LP02)),TYPE=ZEDC,PCHID=108
FUNCTION FID=08,VF=2,PART=((LP15),(LP02)),TYPE=ZEDC,PCHID=178

10GbE RoCE Express Functions for LPAR LP14, Reconfigurable to LP03 or LP04:
FUNCTION FID=9,VF=1,PART=((LP14),(LP03,LP04)),PNETID=(NET1,NET2), *
    TYPE=ROCE,PCHID=110
FUNCTION FID=A,VF=1,PART=((LP14),(LP03,LP04)),PNETID=(NET1,NET2), *
    TYPE=ROCE,PCHID=170

10GbE RoCE Express Functions for LPAR LP15, Reconfigurable to LP03 or LP04:
FUNCTION FID=B,VF=2,PART=((LP15),(LP03,LP04)),PNETID=(NET1,NET2), *
    TYPE=ROCE,PCHID=110
FUNCTION FID=C,VF=2,PART=((LP15),(LP03,LP04)),PNETID=(NET1,NET2), *
    TYPE=ROCE,PCHID=170

```

Figure G-3 Example of IOCP statements for zEDC Express and 10GbE RoCE Express

G.3.1 FUNCTION identifier

The FUNCTION identifier (FID) is a hexadecimal number between 000 and FFF that you use to assign a PCHID to the FUNCTION to identify the specific hardware feature in the PCIe I/O drawer. Because the FUNCTION is not related to a channel subsystem, all LPARs on a CPC can be defined to it. However, a FUNCTION cannot be shared between LPARs. It is only dedicated or reconfigurable by using the PART parameter. The TYPE parameter is new for z13s and is required.

G.3.2 Virtual function number

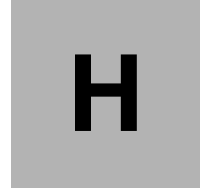
If you want several LPARs to be able to use a zEDC Express feature (the 10GbE RoCE Express feature cannot be shared between LPARs), you need to use a virtual function (VF) number. A *VF* number is a number between 1 and *n*, where *n* is the maximum number of LPARs that the feature supports, which is 15 for the zEDC Express feature and 31 for the RoCE Express feature.

G.3.3 Physical network identifier

The physical network identifier (PNETID) is required to set up the Shared Memory Communications over RDMA (SMC-R) communication between two 10GbE RoCE Express features. Each FUNCTION definition supports up to four PNETIDs.

Notes:

- ▶ For more information about the FUNCTION statement, see *z Systems Input/Output Configuration Program User's Guide for ICP IOCP*, SB10-7163.
- ▶ Definition of 10 GbE RoCE Express feature is required to pair up with an OSD CHPID definition, by the parameter of PNETID. The OSD CHPID definition statement is not listed in the example.
- ▶ The PNETID is limited to 2 for a 10GbE RoCE Express definition statement on z13s.



Flash Express

This appendix covers the IBM Flash Express feature introduced on the zEC12 server.

Flash memory is a non-volatile computer storage technology. It was introduced on the market decades ago. Flash memory is commonly used today in memory cards, USB flash drives, solid-state drives (SSDs), and similar products for general storage and data transfer. Until recently, the high cost per gigabyte and limited capacity of SSDs restricted deployment of these drives to specific applications. Recent advances in SSD technology and economies of scale have driven down the cost of SSDs, making them a viable storage option for I/O-intensive enterprise applications.

An SSD, sometimes called a *solid-state disk* or *electronic disk*, is a data storage device that uses integrated circuit assemblies as memory to store data persistently. SSD technology uses electronic interfaces compatible with traditional block I/O hard disk drives. SSDs do not employ any moving mechanical components. This characteristic distinguishes them from traditional magnetic disks, such as hard disk drives (HDDs), which are electromechanical devices that contain spinning disks and movable read/write heads. With no seek time or rotational delays, SSDs can deliver substantially better I/O performance than HDDs. Flash SSDs demonstrate latencies that are 10 - 50 times lower than the fastest HDDs, often enabling dramatically improved I/O response times.

This appendix includes the following sections:

- ▶ Flash Express overview
- ▶ Using Flash Express
- ▶ Security on Flash Express

H.1 Flash Express overview

Flash Express, which is implemented by using Flash SSDs mounted in Peripheral Component Interconnect Express (PCIe) Flash Express feature cards, is available as carry forward or as a new order for z13s servers.

Flash Express is an innovative solution that is designed to help improve availability and performance to provide a higher level of quality of service. It is designed to automatically improve availability for key workloads at critical processing times and to improve access time for critical business z/OS workloads. It can also reduce latency time during diagnostic collection (memory dump operations). In addition, CFCC Level 20 can take advantage of Flash Express as an overflow device for shared queue data in clustered WebSphere MQ deployments.

Figure H-1 illustrates how Flash Express introduces a new level in the z Systems storage hierarchy.

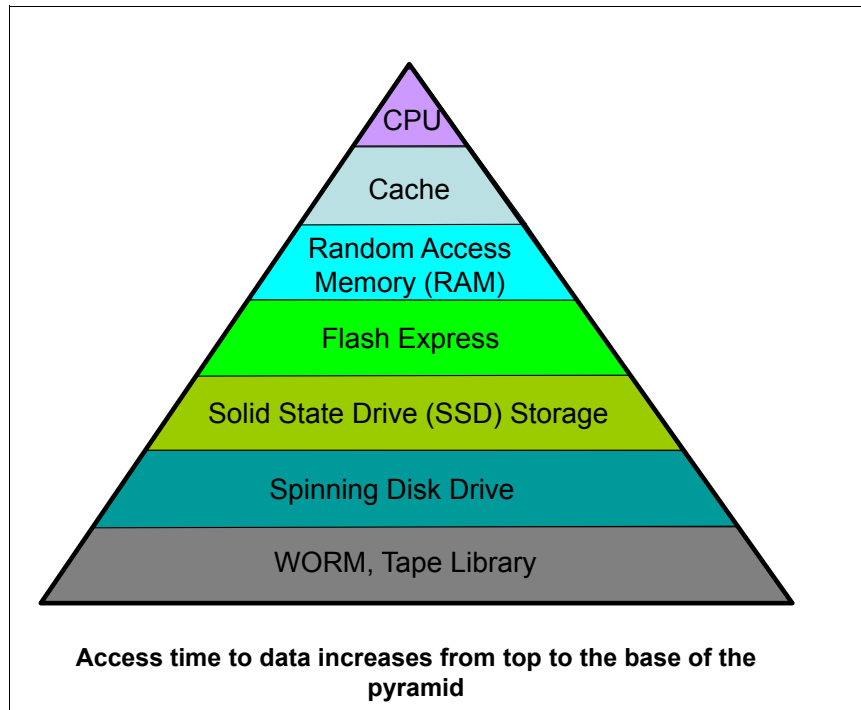


Figure H-1 z Systems storage hierarchy

Flash Express is an optional PCIe card feature that is available on zEC12, zBC12, z13, and z13s servers. Flash Express cards are supported in PCIe I/O drawers, and can be mixed with other PCIe I/O cards, such as Fibre Channel connection (FICON) Express16S, Crypto Express5S, and Open Systems Adapter (OSA) Express5S cards. You can order a minimum of two features (FC 0403) and a maximum of eight features.¹ The cards are ordered in increments of two.

Flash Express cards are assigned one physical channel ID (PCHID) although they have no ports. No hardware configuration definition (HCD) or input/output configuration program (IOCP) definition is required for Flash Express installation. Flash Express uses subchannels that are allocated from the 25K that is reserved in subchannel set 0. Similar to other PCIe I/O

¹ FC 0402 can also be carried forward. The total number of Flash Express features (carry forward + new order) cannot exceed eight per z Systems server.

cards, redundant PCIe paths to Flash Express cards are provided by redundant I/O interconnect. Unlike other PCIe I/O cards, they can be accessed from the host only by a unique protocol.

A Flash Express PCIe adapter integrates four SSD cards of 400 GB each for a total of 1.4 TB of usable data per card, as shown in Figure H-2.

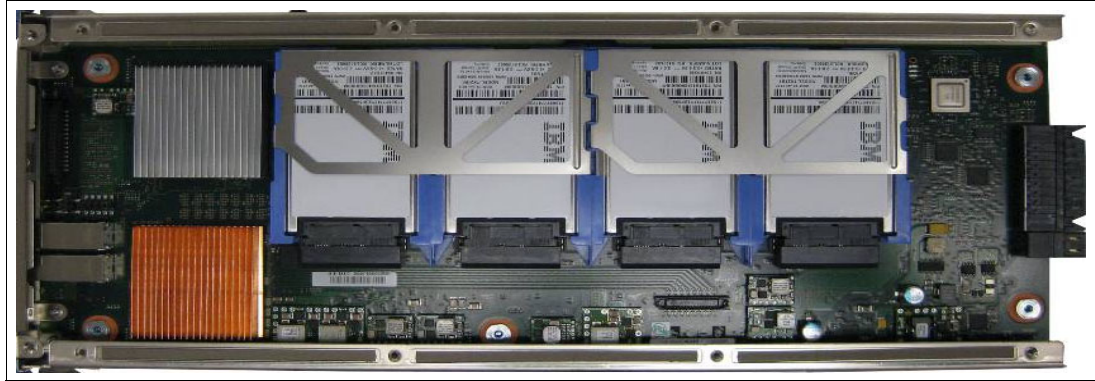


Figure H-2 Flash Express PCIe adapter

Each card (of a pair of cards) is installed in a PCIe I/O drawer in two different I/O domains. A maximum of two pairs are installed in a drawer with only one flash card per domain. Installing more than two pairs requires a second PCIe I/O drawer. Install the cards in the front of the installed drawers (slots 1 and 14) before you use the rear slots (25 and 33). Format each pair of cards before you use them.

Figure H-3 shows a PCIe I/O drawer that is fully populated with Flash Express cards.

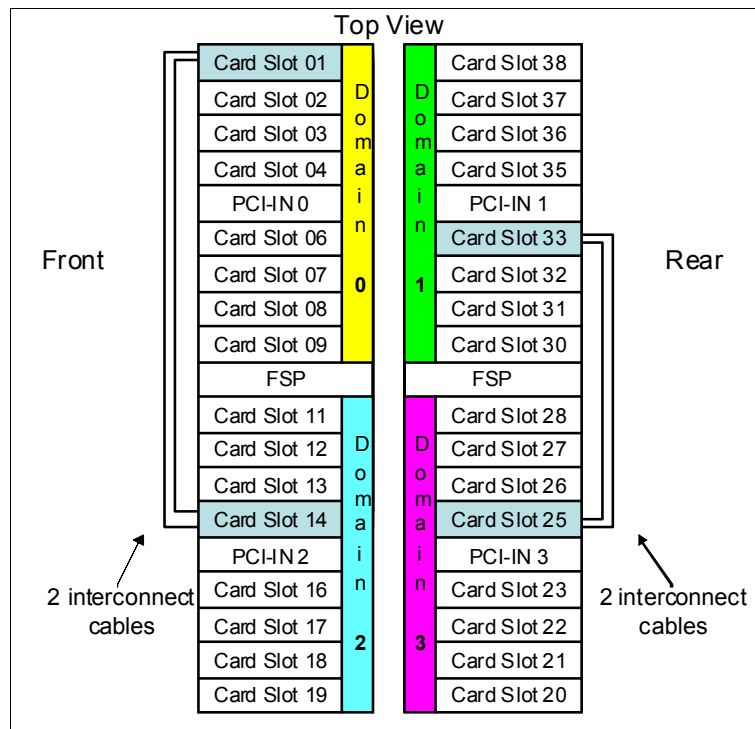


Figure H-3 PCIe I/O drawer that is fully populated with Flash Express cards

For higher resiliency and high availability, Flash Express cards are always installed in pairs. A maximum of four pairs are supported in a z13s system, providing a maximum of 5.6 TB of storage. In each Flash Express card, data is stored in a RAID configuration. If an SSD fails, data is reconstructed dynamically. The cards mirror each other over a pair of cables in a RAID 10 configuration that combines mirroring and striping RAID capabilities. If either card fails, the data is available on the other card. Card replacement is concurrent with the client's operations. In addition, Flash Express supports concurrent firmware upgrades.

The data that is written on the Flash Express cards is always stored encrypted with a volatile key. The card is only usable on the system with the key that encrypted it. For key management, both the primary and alternate Support Elements (SEs) have USB smart cards installed. The smart card contains both a unique key that is personalized for each system and a small Crypto engine that can run a set of security functions within the card.

H.2 Using Flash Express

Flash Express is designed to improve availability and latency from batch to interactive processing in z/OS environments, such as start of day. It helps accelerate start of day processing when there is heavy application activity. Flash Express also helps improve diagnostic procedures, such as supervisor call (SAN Volume Controller) memory dumps, and stand-alone memory dumps.

In z/OS, Flash Express memory is accessed by using the new z Systems Extended Asynchronous Data Mover (EADM) architecture. It is started with a Start subchannel instruction.

The Flash Express PCIe cards are shareable across logical partitions (LPARs). Flash Express memory can be assigned to z/OS LPARs, such as the main storage. It is dedicated to each LPAR. You can dynamically increase the amount of Flash Express memory that is allocated to an LPAR.

Flash Express is supported by z/OS 1.13 plus PTFs, and z/OS 2.1 for the z/OS paging activity and SAN Volume Controller memory dumps. Using Flash Express memory, 1 MB large pages become pageable. It is expected to provide applications with substantial improvement in SAN Volume Controller memory dump data capture time. Flash Express is expected to provide the applications with improved resiliency and speed, and make large pages pageable.

Flash Express memory in the CPC is assigned to a coupling facility (CF) partition by using hardware definition windows the same way that it is assigned to the z/OS partitions.

Flash Express use by the CF provides emergency capacity to handle WebSphere MQ shared queue buildups during abnormal situations. These situations include where "putters" are putting to the shared queue, but "getters" are transiently not getting from the shared queue, or other such transient producer or consumer mismatches on the queue. No new level of WebSphere MQ is required for this support.

Linux for z Systems (Red Hat Enterprise Linux and SUSE Linux enterprise) can use Flash Express as temporary storage. Other software subsystems might take advantage of Flash Express in the future.

Table H-1 gives the minimum support requirements for Flash Express.

Table H-1 Minimum support requirements for Flash Express

Operating system	Support requirements
z/OS	z/OS V1R13 ^a and V2R1
CFCC	CF Level 20

a. Web delivery and program temporary fixes (PTFs) are required.

You can use the Flash Express allocation windows on the SE or Hardware Management Console (HMC) to define the initial and maximum amount of Flash Express available to an LPAR. The maximum memory that is allocated to an LPAR can be dynamically changed. On z/OS, this process can also be done by using an operator command. Flash memory can also be configured offline to an LPAR.

Figure H-4 gives a sample SE/HMC interface that is used for Flash Express allocation.

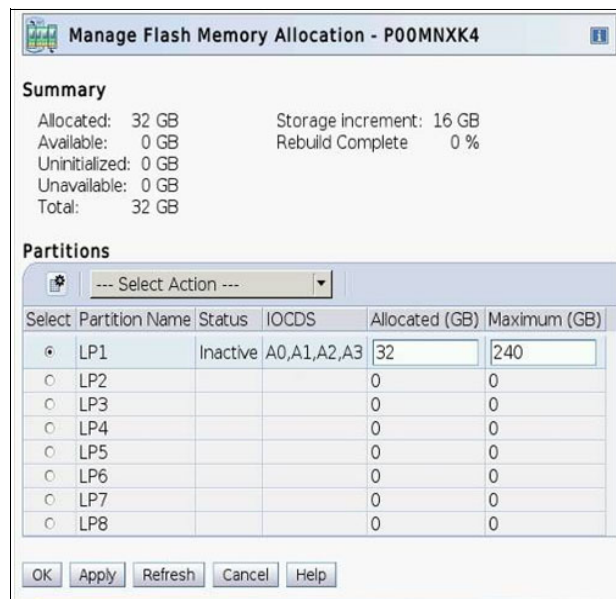


Figure H-4 Sample SE/HMC window for Flash Express allocation to LPAR

The SE user interface for Flash Express provides the following types of actions:

- ▶ Flash status and control: Displays the list of adapters that are installed in the system and their state.
- ▶ Manage Flash allocation: Displays the amount of flash memory on the system.
- ▶ View Flash allocations: Displays a table of flash information for one partition.
- ▶ View Flash: Displays information for one pair of flash adapters.

Physical Flash Express PCIe cards are fully virtualized across LPARs. Each LPAR can be configured with its own storage-class memory (SCM) address space. The size of Flash Express memory that is allocated to a partition is done by amount, not by card size. The hardware supports error isolation, transparent mirroring, centralized diagnostic procedures, hardware logging, and recovery, independently from the software.

At IPL, z/OS detects whether flash is assigned to the partition. z/OS automatically uses Flash Express for paging unless otherwise specified by using the new z/OS `PAGESCM=NONE` parameter. All paging data can be on Flash Express memory. The function is easy to use, and there is no need for capacity planning or placement of data on Flash Express cards.

Figure H-5 gives an example of Flash Express allocation between two z/OS LPARs.

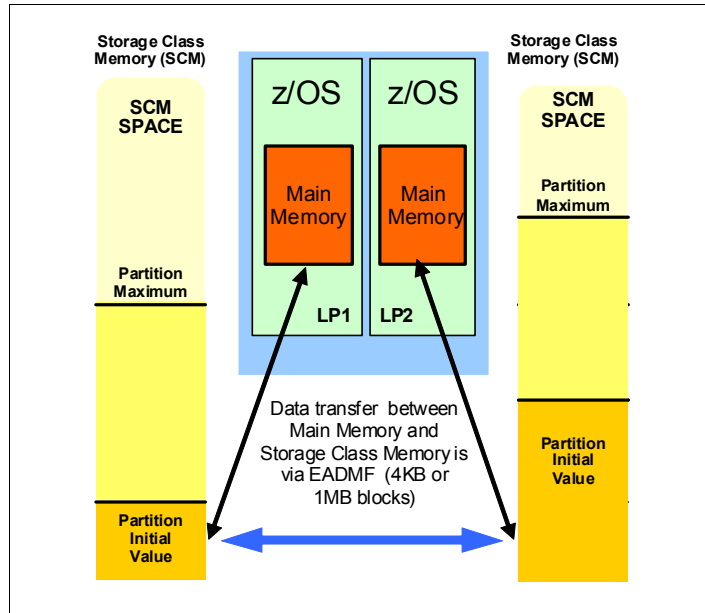


Figure H-5 Flash Express allocation in z/OS LPARs

Flash Express memory is a faster paging device than HDD. It replaces disks, not memory. It is suitable for workloads that can tolerate paging. It does not benefit workloads that cannot afford to page. The z/OS design for Flash Express memory does not completely remove the virtual constraints that are created by a paging spike in the system. The z/OS paging subsystem works with a mix of internal Flash Express and external disks.

Currently, 1 MB large pages are not pageable. With the introduction of Flash Express, 1 MB large pages can be on Flash Express and pageable.

Table H-2 introduces, for a few z/OS data types that are supported by Flash Express, the choice criteria for data placement on Flash Express or on disk.

Table H-2 Flash Express z/OS supported data types

Data type	Data page placement
Pageable link pack area (PLPA)	At IPL/NIP time, PLPA pages are placed both on flash and disk.
Virtual input/output (VIO)	VIO data is always placed on disk (first to VIO accepting data sets, with any spillover flowing to non-VIO data sets).
IBM HyperSwap Critical Address Space data	If flash space is available, all virtual pages that belong to a HyperSwap Critical Address Space are placed in flash memory. If flash space is not available, these pages are kept in memory and only paged to disk when the system is real storage constrained, and no other alternatives exist.

Data type	Data page placement
Pageable large pages	If contiguous flash space is available, pageable large pages are written to flash.
All other data	If space is available on both flash and disk, the system makes a selection that is based on response time.

Flash Express is used by the Auxiliary Storage Manager (ASM) with paging data sets to satisfy page-out and page-in requests received from the real storage manager (RSM). It supports 4 KB and 1 MB page sizes. ASM determines where to write a page based on space availability, data characteristics, and performance metrics. ASM still requires definition of a PLPA, Common, and at least one local paging data set. VIO pages are only written to DASD because persistence is needed for warm starts.

A new **PAGESCM** keyword in the IEASYSxx member defines the minimum amount of flash to be reserved for paging. The value can be specified in units of MB, GB, or TB. NONE indicates that the system does not use flash for paging. ALL (the default) indicates that all flash that is defined to the partition is available for paging.

The following new messages are issued during z/OS IPL and indicate the status of SCM:

```
IAR031I USE OF STORAGE-CLASS MEMORY FOR PAGING IS ENABLED - PAGESCM=ALL,
ONLINE=00001536M
```

```
IAR032I USE OF STORAGE-CLASS MEMORY FOR PAGING IS NOT ENABLED - PAGESCM=NONE
```

The **D ASM** and **D M** commands are enhanced to display flash-related information/status:

- ▶ **D ASM** lists the SCM status along with paging data set status.
- ▶ **D ASM, SCM** displays a summary of SCM usage.
- ▶ **D M=SCM** displays the SCM online/offline and increment information.
- ▶ **D M=SCM(DETAIL)** displays detailed increment-level information.

The **CONFIG ONLINE** command is enhanced to allow bringing more SCMs online:

```
CF SCM (amount), ONLINE
```

H.3 Security on Flash Express

Data that is stored on Flash Express is encrypted by a strong encryption symmetric key that is in a file on the SE hard disk. This key is also known as the *Flash encryption key/authentication key*. The firmware management of the Flash Express adapter can generate an asymmetric transport key in which the flash encryption key/authentication key is wrapped. This transport key is used while in transit from the SE to the firmware management of the Flash Express adapter.

The SE has an integrated card reader into which one smart card at a time can be inserted. When an SE is “locked down,” removing the smart card is not an option unless you have the key to the physical lock.

H.3.1 Integrated Key Controller

The SE initializes the environment by starting APIs within the Integrated Key Controller (IKC). The IKC loads an applet to a smart card inserted in the integrated card reader. The smart card applet, as part of its installation, creates a Rivest-Shamir-Adleman (RSA) algorithm key pair, the private component of which never leaves the smart card. However, the public key is

exportable. The applet also creates two Advanced Encryption Standard (AES) symmetric keys. One of these AES keys is known as the key-encrypting key (KEK), which is retained on the smart card. The KEK can also be exported. The other AES key becomes the *Flash encryption key/authentication key* and is encrypted by the KEK.

A buffer is allocated containing the KEK-encrypted flash encryption key/authentication key and the unique serial number of the SE. The buffer is padded per Public-Key Cryptography Standards #1 (PKCS #1) and then encrypted by the smart card RSA public key. The encrypted content is then written to a file on the SE hard disk.

This design defines a tight coupling of the file on the SE to the smart card. The coupling ensures that any other SE is not able to share the file or the smart card that is associated with an SE. It ensures that the encrypted files are unique and all such smart cards are uniquely tied to their SEs.

All key generation, encryption, and decryption occur on the smart card. Keys are never in the clear. The truly sensitive key, the flash encryption key/authentication key, is only in the file on the SE until it is served to the firmware management of the Flash Express adapter.

Figure H-6 shows the cryptographic keys that are involved in creating this tight-coupling design.

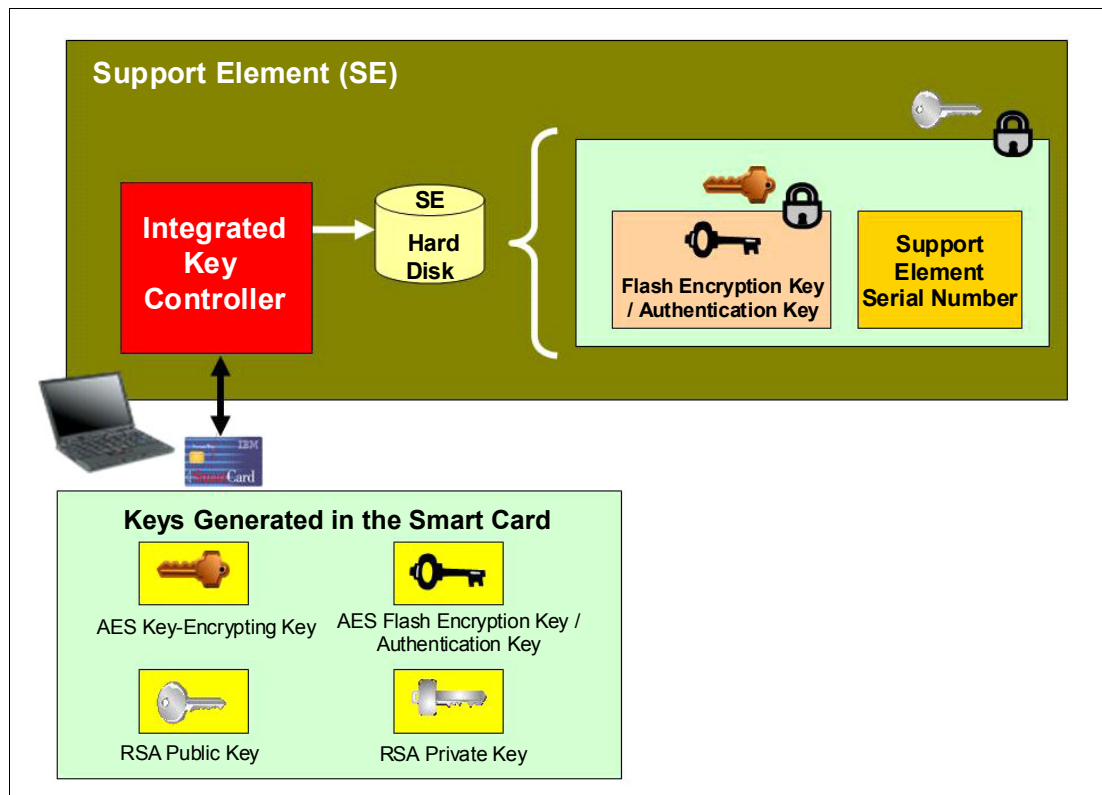


Figure H-6 Integrated Key Controller

The flash encryption key/authentication key can be served to the firmware management of the Flash Express adapter. This process can be either upon request from the firmware at initial microcode load (IML) time or from the SE as the result of a request to “change” or “roll” the key.

During the alternate SE initialization, application programming interfaces (APIs) are called to initialize the alternate smart card in it with the applet code and create the RSA public/private key pair. The API returns the public key of the smart card that is associated with the alternate SE. This public key is used to encrypt the KEK and the Flash encryption key/authentication key from the primary SE. The resulting encrypted file is sent to the alternate SE for redundancy.

H.3.2 Key serving topology

In a key serving topology, the SE is the key server and the IKC is the key manager. The SE is connected to the firmware management of the Flash Express adapter through a secure communications line. The firmware manages the transportation of the Flash encryption key/authentication key through internal system paths. Data in the adapter cache memory is backed up by a flash-backed DRAM module. This module can encrypt the data with the Flash encryption key/authentication key.

The firmware management of the Flash Express adapter generates its own transport RSA asymmetric key pair. This pair is used to wrap the Flash encryption key/authentication key while in transit between the SE and the firmware code.

Figure H-7 on page 538 shows the following key serving topology:

1. The firmware management of the Flash Express adapter requests the Flash encryption key/authentication key from the SE at IML time. When this request arrives, the firmware public key is passed to the SE to be used as the transport key.
2. The file that contains the KEK-encrypted Flash encryption key/authentication key and the firmware public key is passed to the IKC. The IKC sends the file contents and the public key to the smart card.
3. The applet on the smart card decrypts the file contents and the Flash encryption key/authentication key. It then re-encrypts the Flash encryption key/authentication key with the firmware public key.
4. This encrypted key is then passed back to the SE, which forwards it on to the firmware management of the Flash Express adapter code.

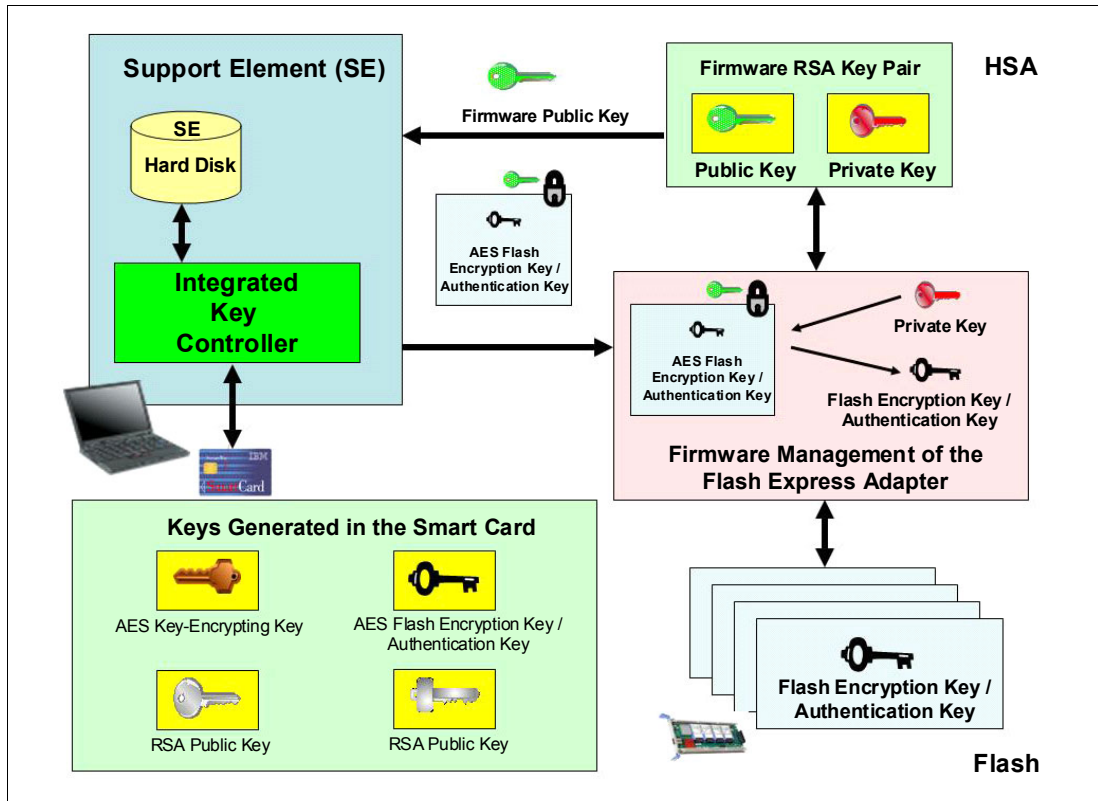


Figure H-7 Key serving topology

H.3.3 Error recovery scenarios

Possible error scenarios are described in this section.

Primary Support Element failure

When the primary SE fails, a switch is made to the alternate SE, which then becomes the new primary. When the former primary is brought back up, it becomes the alternate SE. The KEK and the Flash encryption key/authentication key from the primary SE were already sent to the alternate SE for redundancy at initialization time.

Removal of a smart card

If a smart card is removed from the card reader, the card reader signals the event to the IKC listening code. The IKC listener then calls the SE to take the appropriate action. The appropriate action can involve deleting the flash encryption key or authentication key file.

If the smart card is removed while the SE is powered off, the system has no knowledge of the event. However, when the SE is powered on, notification is sent to the system administrator.

Primary Support Element failure during IML serving of the flash key

If the primary SE fails during the serving of the key, the alternate SE takes over as the primary and restarts the key serving operation.

Alternate Support Element failure during switchover from the primary

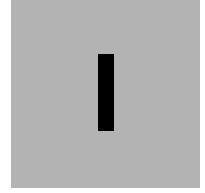
If the alternate SE fails during the switchover to become the primary SE, the key serving state is lost. When the primary comes back up, the key serving operation can be restarted.

Primary and alternate Support Elements fail

If the primary and the alternate SE both fail, the key cannot be served. If the devices are still up, the key is still valid. If either or both SEs are recovered, the files holding the Flash encryption key/authentication key can still be valid. This is true even in a key roll case. Both new and current (old) keys need to be available until the key serving operation is complete.

If both SEs are down, and the Flash Express goes down and comes back online before the SEs become available, all data on the Flash Express is lost. Reformatting is then necessary when the device is started.

If both Flash Express devices are still started, get the primary SE back online as fast as possible with the Flash encryption key/authentication key file and associated smart card still intact. After that happens, the alternate SE can be brought online with a new smart card and taken through the initialization procedure.



GDPS Virtual Appliance

This appendix discusses the Geographically Dispersed Parallel Sysplex (GDPS) Virtual Appliance.

This appendix includes the following sections:

- ▶ GDPS overview
- ▶ Overview of GDPS Virtual Appliance
- ▶ GDPS Virtual Appliance recovery scenarios

I.1 GDPS overview

GDPS is a collection of offerings, each addressing a different set of IT resiliency goals that can be tailored to meet the recovery point objective (RPO) and recovery time objective (RTO) for your business. Each offering uses a combination of server and storage hardware or software-based replication, automation, and clustering software technologies. In addition to the infrastructure that makes up a GDPS solution, IBM also includes services, particularly for the first installation of GDPS and optionally for subsequent installations, to ensure that the solution meets your business objectives.

Definitions:

- ▶ RPO defines the *amount of data* that you can afford to re-create during a recovery, by determining the most recent point in time for data recovery.
- ▶ RTO is the *time that is needed to recover* from a disaster or how long the business can survive without the systems.

Figure I-1 illustrates some of the GDPS offerings.

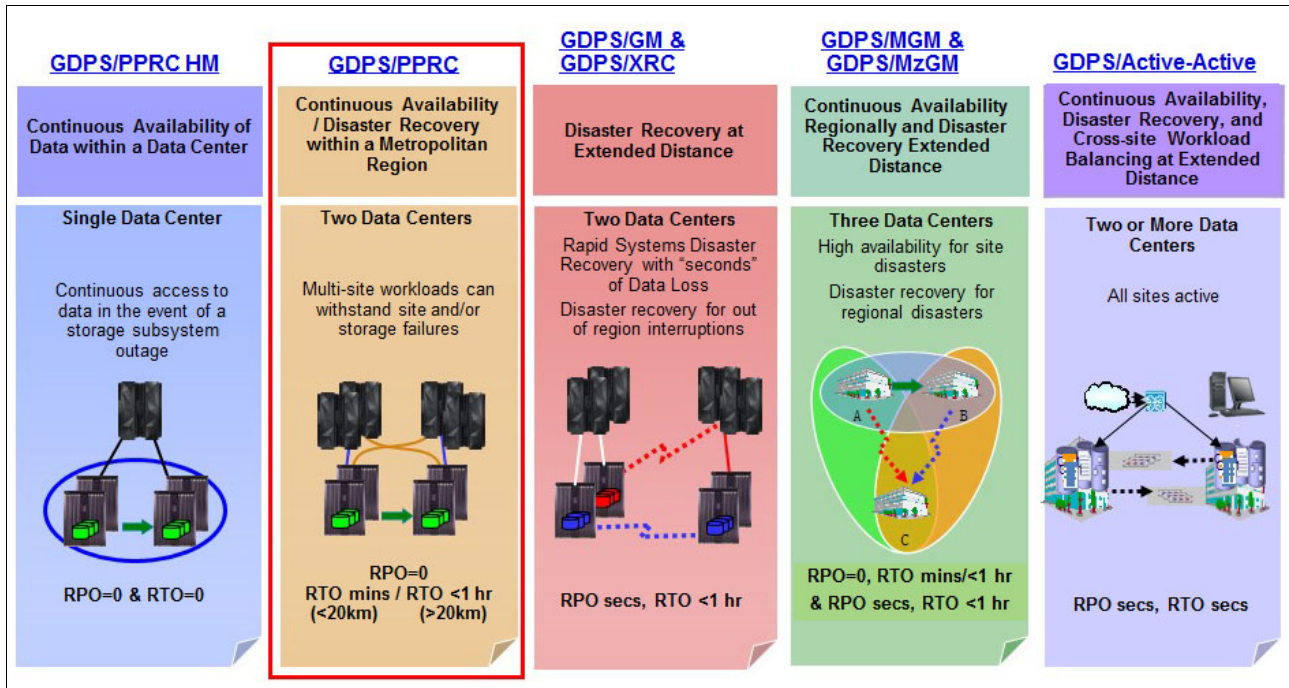


Figure I-1 GDPS offerings

For a complete description of GDPS solution and terminology, see *IBM GDPS Family of Products: An Introduction to Concepts and Capabilities*, SG24-6374.

It is typical to deploy IT environments that are only running Linux on z Systems. The GDPS Virtual Appliance is a building block of high availability (HA) and disaster recovery (DR) solutions for those environments that do not have or require z/OS skills.

The following are the major drivers behind implementing an HA/DR architecture:

- ▶ Regulatory compliance for business continuity (99.999% or higher availability)
- ▶ Avoid financial loss
- ▶ Maintaining reputation (which also translates into customer satisfaction, and thus money)

An HA/DR implementation has various levels:

- ▶ High availability (HA): The attribute of a system to provide service during defined periods at agreed upon levels by masking unplanned outages from users. HA employs component duplication (hardware and software), automated failure detection, retry, bypass, and reconfiguration.
- ▶ Continuous operations (CO): Attribute of a system to continuously operate and mask planned outages from users. CO provides the means to minimize planned downtime during maintenance windows. It employs nondisruptive hardware and software changes, nondisruptive configuration, and software coexistence.
- ▶ Continuous availability (CA): Attribute of a system to deliver nondisruptive service to the user 7 days a week, 24 hours a day (there are no planned or unplanned outages).

A system outage can occur for various reasons. Outages can be categorized as either *planned* or *unplanned*.

Planned outages can be caused by the following situations:

- ▶ Backups
- ▶ Operating system installation and maintenance
- ▶ Application software maintenance
- ▶ Hardware and software upgrades

Unplanned outages can be caused by the following situations:

- ▶ Non-disaster events such as:
 - Application failure
 - Operator errors (human error)
 - Power outages
 - Network failure
 - Hardware and software failures
- ▶ Disaster events such as:
 - Natural disasters or other catastrophes that damage the production facilities beyond usability (for example, fire, flood, earthquake, or bombing)
 - Failure of a regional power grid
 - Outages that require a recovery procedure at an off-site location

Automation is key when implementing a HA/DR solution. The major benefits of an automated solution are as follows:

- ▶ Provides reliable, consistent RTO
- ▶ Provides consistent and predictive recovery time as the environment scales
- ▶ Reduces infrastructure management cost and staff skills
- ▶ Reduces or eliminates human intervention, and therefore the probability of human error
- ▶ Facilitates regular testing for repeatable and reliable results of business continuity procedures
- ▶ Helps maintain recovery readiness by managing/monitoring servers, data replication, workload, and network with notification of events that occur within the environment

I.2 Overview of GDPS Virtual Appliance

To reduce IT costs and complexity, many enterprises are consolidating independent servers into Linux images (guests) running on z Systems servers. Linux on z Systems can be implemented either as guests running under z/VM or native Linux logical partitions (LPARs) on z Systems servers. Workloads with an application server that runs on Linux on z Systems and a database server that runs on z/OS are common. The following are some examples:

- ▶ WebSphere Application Server running on Linux and CICS, DB2 running under z/OS
- ▶ SAP application servers running on Linux and database servers running on z/OS

With a multi-tiered architecture, you need to provide a coordinated near-continuous availability and disaster recovery solution for both z/OS and Linux on z Systems.

GDPS Virtual Appliance is a fully integrated continuous availability and disaster recovery solution for Linux on z Systems customers that include these components:

- ▶ An operating system image
- ▶ The application components
- ▶ An appliance management layer, which makes the image self-contained
- ▶ Application programming interfaces (APIs) and user interfaces (UIs) for customization, administration, and operation tailored for the appliance function.

GDPS Virtual Appliance is designed to improve both consumability and time-to-value for customers.

Figure I-2 shows different solutions and how to position the GDPS Virtual Appliance.

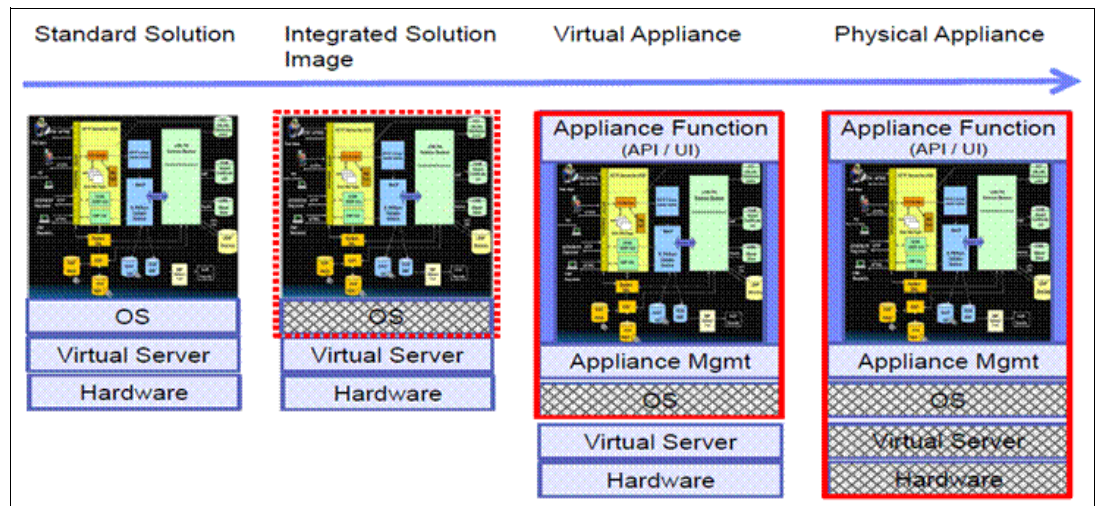


Figure I-2 Positioning a virtual appliance

A virtual appliance is a fully integrated software solution that has been targeted and optimized for a specific business problem:

- ▶ *Targeted* for a specific deployment platform to reduce potential configuration complexity while using any underlying capabilities of the platform
- ▶ *Purposed* for a specific, high-level business context or IT architecture by installing particular applications and hardening them before delivery
- ▶ *Optimized* by choosing the appropriate configuration, knowing all elements of the system and removing unnecessary attributes

The GDPS Virtual Appliance solution implements the GDPS/PPRC Multiplatform Resilience for z Systems (xDR), which coordinates near-continuous availability and DR solution by using these techniques:

- ▶ Disk error detection
- ▶ Heartbeat for smoke tests
- ▶ Initial program load (IPL) in place again
- ▶ Coordinated site takeover
- ▶ Coordinated HyperSwap
- ▶ Single point of control

The GDPS Virtual Appliance has these requirements:

- ▶ Hardware
 - Any supported hardware: zEC12, zBC12, z13, and z13s servers
 - 1 LPAR with one logical CP
 - 1 GB Memory
 - 3 extended count key data (ECKD) direct access storage devices (DASDs) (one for appliance image initial installation, and two for the upgrade scenario)
 - One or two Open Systems Adapter (OSA) attachments (dependent on the network setup and the wanted network resilience).
- ▶ System
 - z/VM Version 5 Release 4 or later, or z/VM 6.2 or later. Note that z/VM 5.4 is not supported on z13s servers.
 - The disks being used by z/VM and Linux to be mirrored must be ECKD disks
 - A supported distribution of Linux on z Systems with the latest recommended fix pack
 - IBM Tivoli System Automation for Multiplatforms with the latest recommended fix pack. The separately priced xDR for Linux feature is required (one Linux guest as xDR proxy).

Note: For the latest information about Tivoli System Automation for Multiplatforms, see:

<http://www.ibm.com/software/tivoli/products/sys-auto-multi>

Figure I-3 shows an overview of GDPS Virtual Appliance implementation.

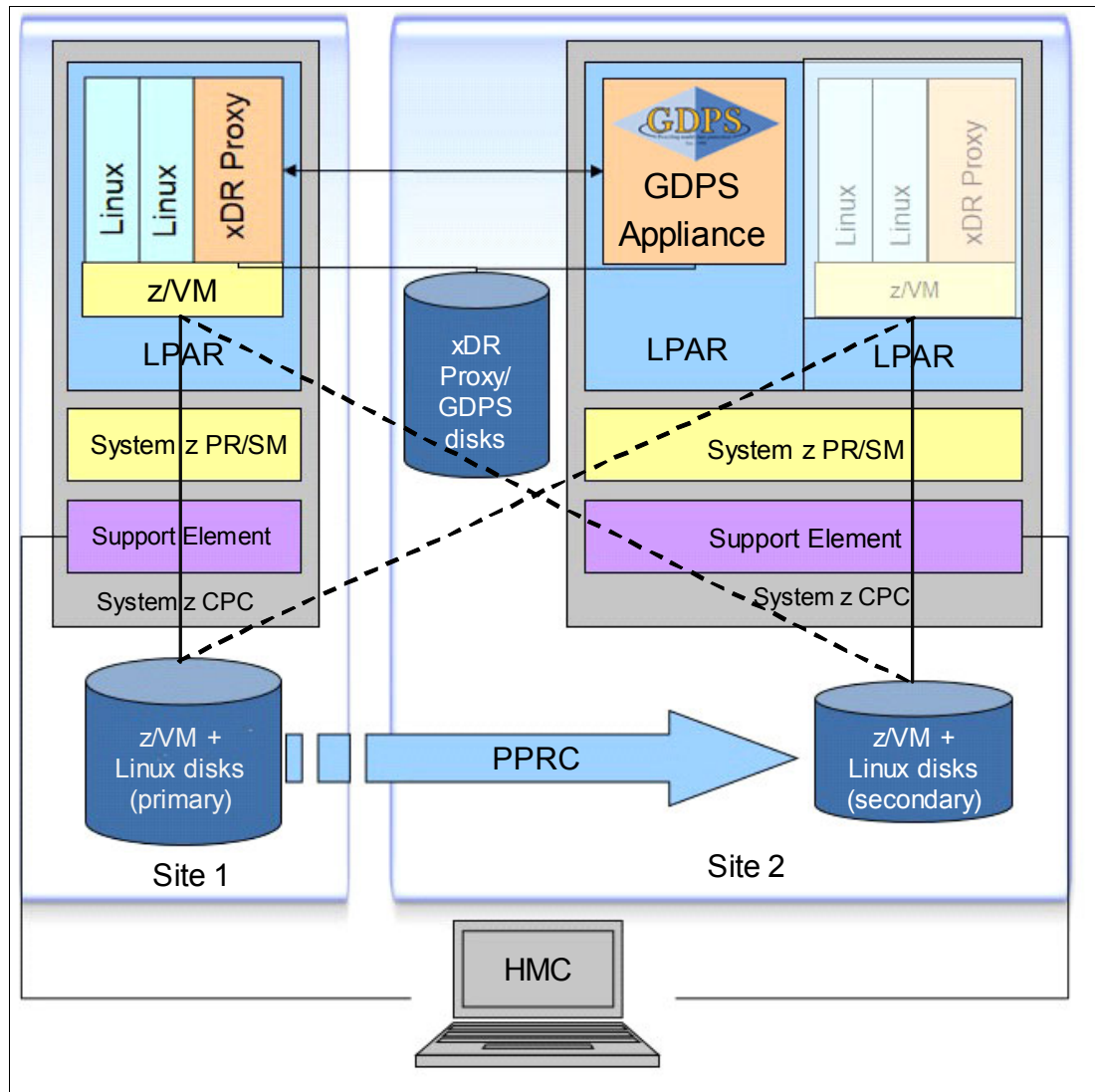


Figure I-3 GDPS virtual appliance architecture overview

Keep in mind the following considerations for GDPS Virtual Appliance architecture:

- ▶ PPRC ensures that the remote copy is identical to the primary data. The synchronization takes place at the time of I/O operation.
- ▶ One dedicated Linux guest is configured as the xDR Proxy for GDPS, which is used for tasks that have z/VM scope (HyperSwap, shut down z/VM, IPL z/VM guest).
- ▶ The remote copy environment is managed by using HyperSwap function, and data is kept available and consistent for operating systems and applications.
- ▶ Disaster detection ensures successful and faster recovery by using automated processes.
- ▶ A single point of control is implemented from the GDPS Virtual Appliance. There is no need for involvement of all experts (for example, storage team, hardware team, OS team, application team, and so on).

The GDPS Virtual Appliance implements the following functions:

- ▶ Awareness of a failure in a Linux on z Systems node or cluster by monitoring (heartbeats) all nodes and cluster master nodes. If a node or cluster fails, it can be set up to automatically IPL the node or all the nodes in the cluster again.
- ▶ Shutting down a Linux on z Systems node or cluster for service (planned maintenance).
- ▶ Initiation of z/VM Live Guest Relocation to move active guests from one member of a z/VM subsystem interface (SSI) cluster to another
- ▶ Graceful shutdown/startup of the Linux on z Systems cluster, nodes in the cluster, and the z/VM host. Graceful shutdown/startup of z/VM systems includes any z/VSE guests.
- ▶ Use of HyperSwap to non-disruptively swap z/VM and its guests from the primary to secondary Peer-to-Peer Remote Copy (PPRC) devices, for both planned disk subsystem maintenance and unplanned disk subsystem failure.

I.3 GDPS Virtual Appliance recovery scenarios

This section presents the following recovery scenarios using GDPS Virtual Appliance:

- ▶ Planned disk outage
- ▶ Unplanned disk outage
- ▶ Disaster recovery

I.3.1 Planned disk outage

In a planned disk outage, the HyperSwap provides the ability to non-disruptively swap from using the primary volume of a mirrored pair to using the secondary volume. A planned HyperSwap is started manually by operator action by using GDPS facilities. One example of a planned HyperSwap is where a HyperSwap operation is initiated in advance of a planned disruptive maintenance of a disk subsystem.

Figure I-4 shows the operation principle of a disk failover operation that uses HyperSwap.

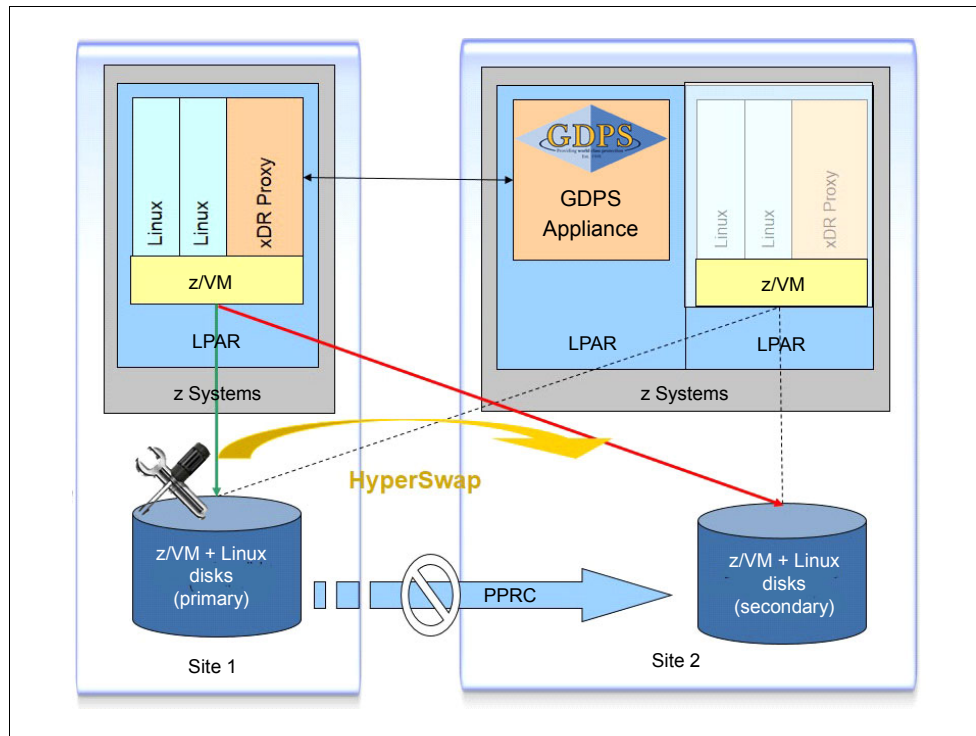


Figure I-4 GDPS Storage failover

Without HyperSwap, the procedure to change the primary disk to the secondary can take up to 2 hours, or even more, depending on the installation size. The procedures include shutting down the systems, removing systems from clusters, and when applicable, reversing PPRC (that is, suspending PPRC), and restarting the systems.

When using HyperSwap, disk swap takes seconds (for example, 6 seconds for 14 systems and 10,000 volume pairs), and the systems remain active.

I.3.2 Unplanned disk outage

An unplanned HyperSwap is started automatically by GDPS when triggered by events that indicate the failure of a primary disk device. HyperSwap events can include the following events:

- ▶ Hard failure triggers:
 - I/O errors
 - Boxed devices
 - Control unit failures
 - Loss of all channel paths
- ▶ Soft failures, such as I/O response time triggers

Again, without HyperSwap, this process can take more than an hour even when done properly. The systems are quiesced, removed from the cluster, and restarted on the other side. With HyperSwap, the same operation can take seconds.

I.3.3 Disaster recovery

In a site disaster, GDPS Appliance will immediately issue a *freeze* for all applicable primary devices. This action is taken to protect the integrity of the secondary data. The GDPS cluster then resets Site 1 and Site 2 systems and updates all the IPL information to point to the secondary devices, and IPLs all the production systems in the LPARs in Site 2 again. The GDPS Appliance scripting capability is key to recovering the systems in the shortest possible time following a disaster. All recovery operations are carried out without operator intervention.



IBM zEnterprise Data Compression Express

This appendix describes the optional IBM zEnterprise Data Compression (zEDC) Express feature of the z13 and z13s, zEC12, and IBM zBC12 servers.

The appendix includes the following sections:

- ▶ Overview
- ▶ zEDC Express
- ▶ Software support

J.1 Overview

The growth of data that needs to be captured, transferred, and stored for large periods of time is unrelenting. In addition, software-implemented compression algorithms can be costly in terms of processor resources and storage costs.

zEDC Express, an optional feature available on z13, z13s, zEC12, and zBC12 servers, addresses those requirements by providing hardware-based acceleration for data compression and decompression. zEDC provides data compression with lower CPU consumption than compression technology previously available on the IBM z Systems server.

Using the zEDC Express feature with the z/OS V2R1 zEnterprise Data Compression acceleration capability (or later releases) is designed to deliver an integrated solution to accomplish these goals:

- ▶ Reduce CPU consumption
- ▶ Optimize the performance of compression-related tasks
- ▶ Enable more efficient use of storage resources.

This solution provides a lower cost of computing and also helps to optimize the cross-platform exchange of data.

J.2 zEDC Express

zEDC Express is an optional feature (FC 0420). It is designed to provide hardware-based acceleration for data compression and decompression.

The feature installs exclusively on the Peripheral Component Interconnect Express (PCIe) I/O drawer. Between one and eight features can be installed on the system. There is one PCIe adapter/compression coprocessor per feature, which implements compression as defined by RFC1951 (DEFLATE).

A zEDC Express feature can be shared by up to 15 logical partitions (LPARs) on the same CPC.

Adapter support for zEDC is provided by Resource Group code that runs on the system Integrated Firmware Processor (IFP). The recommended high availability configuration per server is four features. This configuration provides high availability during concurrent update. For resilience, there are always two independent RGs on the system, sharing the IFP. Install a minimum of two zEDC features, one feature per RG, for resilience and throughput.

Figure J-1 illustrates the PCIe I/O drawer structure and the relationships among card slots, domains, and resource groups.

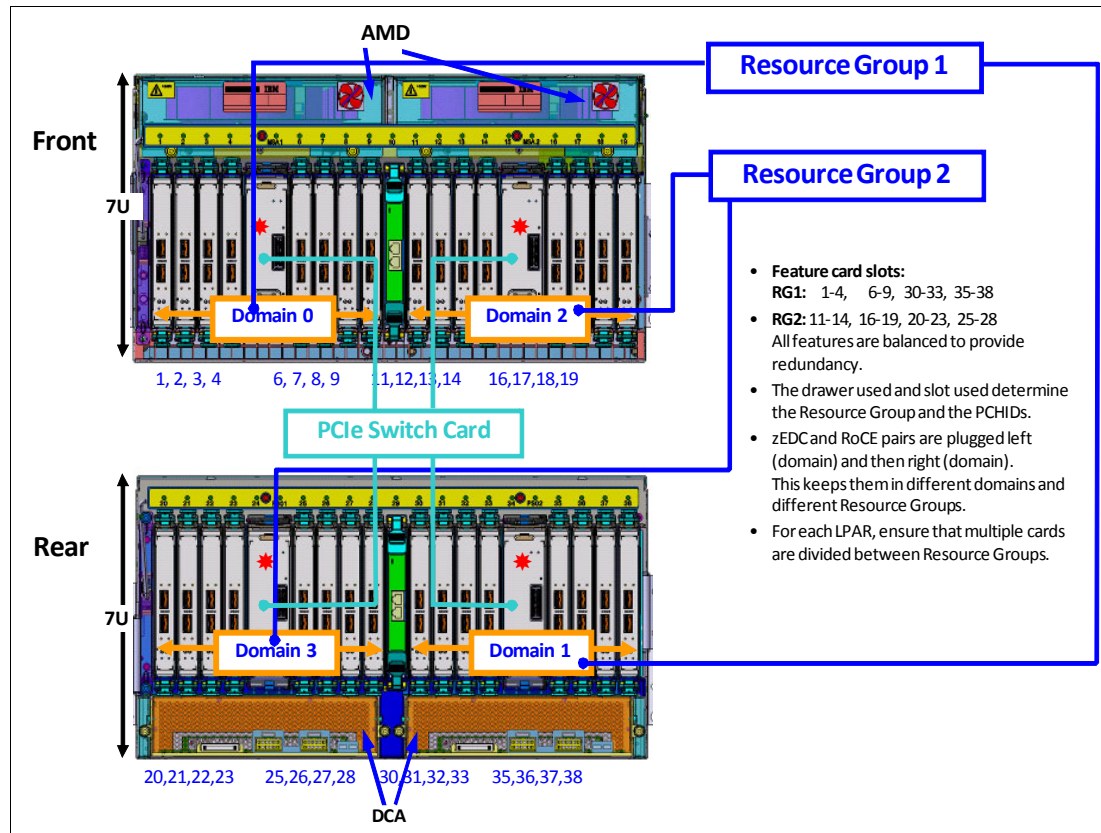


Figure J-1 Relationships between the PCIe I/O drawer card slots, I/O domains, and resource groups

J.3 Software support

Support of the zEDC Express function is provided by z/OS V2R1 zEnterprise Data Compression or later for both data compression and decompression. Support for data recovery (decompression) in the case that zEDC is not installed, or installed but not available, is provided through software in z/OS V2R2, V2R1, V1R13, and V1R12 with the appropriate program temporary fixes (PTFs). Software decompression is slow and can use considerable processor resources. Therefore, it is not suggested for production environments.

See the appropriate fixcat for SMP/E to install prerequisite PTFs. The fixcat function is explained here:

<http://www.ibm.com/systems/z/os/zos/features/smpe/fix-category.html>

The fix category named IBM.Function.zEDC that identifies the fixes that enable or use the zEDC function.

For more Information about Implementation and usage of the zEDC feature, see *Reduce Storage Occupancy and Increase Operations Efficiency with IBM zEnterprise Data Compression*, SG24-8259.

Reference: z/OS support for the zEDC can be found by using FIXCAT: IBM.Function.zEDC.

J.3.1 z/VM V6R3 support with PTFs

z/OS guests that run under z/VM V6.3 with PTFs can now use the zEDC Express feature. zEDC for z/OS V2.1 or later and the zEDC Express feature are designed to support a data compression function to help provide high-performance, low-latency compression without significant CPU processor usage. This feature can help to reduce disk usage, provide optimized cross-platform exchange of data, and provide higher write rates for SMF data. For more information, see:

<http://www.vm.ibm.com/zvm630/apars.html>

J.3.2 IBM SDK 7 for z/OS Java support

IBM 31-bit and 64-bit SDK for z/OS Java Technology Edition, Version 7 Release 1 (5655-W43 and 5655-W44) (IBM SDK 7 for z/OS Java) now provides use of the zEDC Express feature and Shared Memory Communications-Remote Direct Memory Access (SMC-R), which is used by the 10GbE RoCE Express feature.

J.3.3 IBM z Systems Batch Network Analyzer

z Systems Batch Network Analyzer (zBNA) is a no-charge, “as is” tool. It is available to clients, IBM Business Partners, and IBM employees. zBNA replaces the BWATOOL. It is based on Microsoft Windows. It provides graphical and text reports, including Gantt charts, and support for alternative processors.

zBNA can be used to analyze client-provided System Management Facilities (SMF) records to identify jobs and data sets that are candidates for zEDC compression, across a specified time window, typically a batch window. zBNA is able to generate lists of data sets by job:

- ▶ Those jobs that already perform hardware compression and might be candidates for zEDC
- ▶ Those jobs that might be zEDC candidates, but are not in extended format

Therefore, zBNA can help you estimate the use of zEDC features and help determine the number of features needed:

- ▶ IBM Employees can obtain zBNA and other CPS tools by using the IBM intranet:
<http://w3.ibm.com/support/techdocs/atmastr.nsf/WebIndex/PRS5126>
- ▶ IBM Business Partners can obtain zBNA and other CPS tools by using the Internet:
https://www.ibm.com/partnerworld/wps/servlet/mem/ContentHandler/tech_PRS5133
- ▶ IBM clients can obtain zBNA and other CPS tools by using the Internet:
<http://www.ibm.com/support/techdocs/atmastr.nsf/WebIndex/PRS5132>

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 z13 and IBM z13s Technical Introduction*, SG24-8250
- ▶ *IBM z13 Technical Guide*, SG24-8251
- ▶ *IBM z Systems Functional Matrix*, REDP-5157

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

Other publications

The IBM z13s product publications are also relevant as further information sources:

- ▶ *z13s Installation Manual for Physical Planning*, GC28-6953
- ▶ *z Systems Input/Output Configuration Program User's Guide for ICP IOCP*, SB10-7163
- ▶ *z Systems Processor Resource/Systems Manager Planning Guide*, SB10-7162

Online resources

These websites are also relevant as further information sources:

- ▶ IBM z Systems ResourceLink:
<http://www.ibm.com/servers/resourceLink/>

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services



Redbooks

IBM z13s Technical Guide

SG24-8294-00

ISBN 0738441678



(1.0" spine)

0.875" x 1.498"

460 x 788 pages



SG24-8294-00

ISBN 0738441678

Printed in U.S.A.

Get connected

