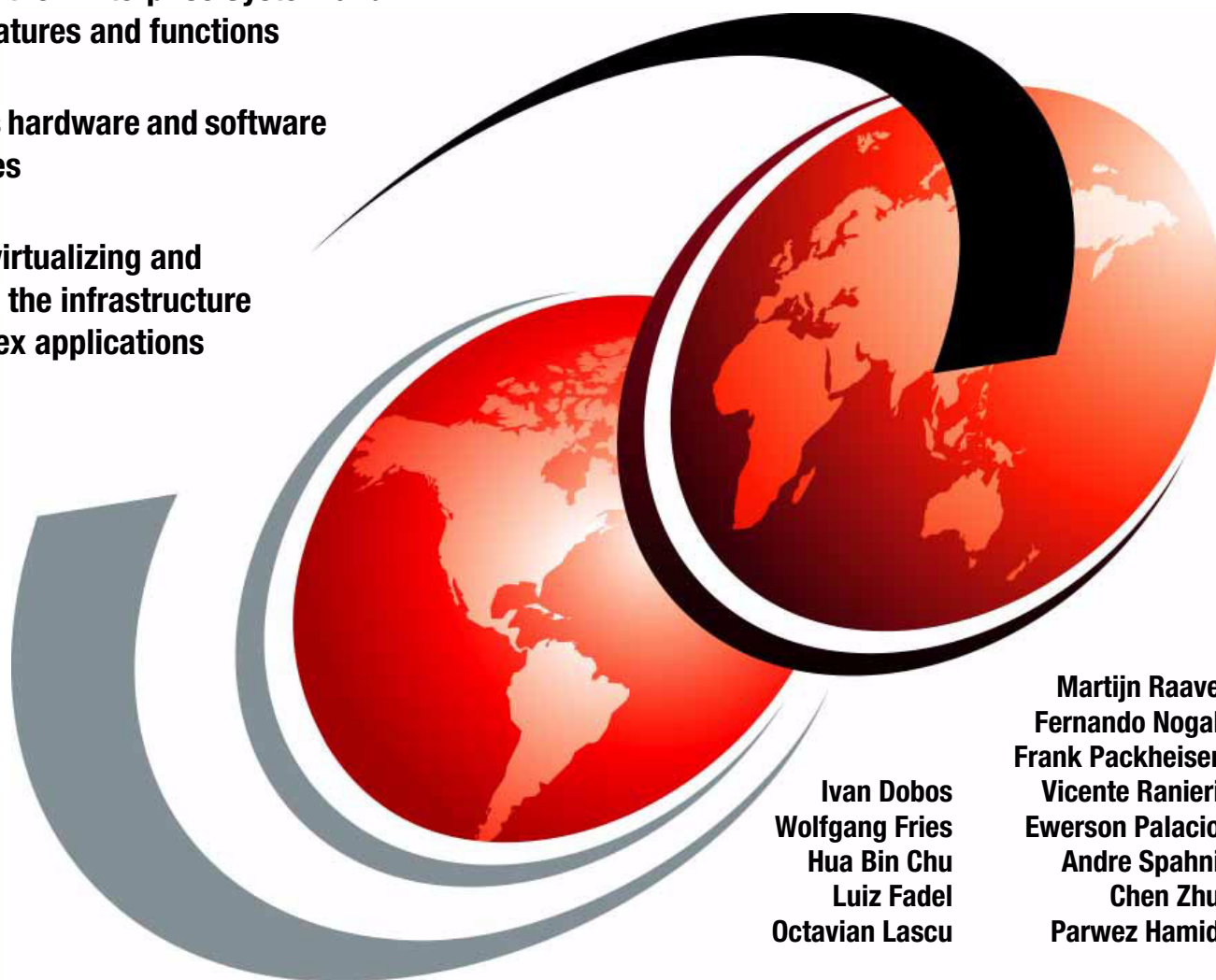


IBM zEnterprise EC12 Technical Guide

Describes the zEnterprise System and related features and functions

Discusses hardware and software capabilities

Explains virtualizing and managing the infrastructure for complex applications



Ivan Dobos
Wolfgang Fries
Hua Bin Chu
Luiz Fadel
Octavian Lascu

Martijn Raave
Fernando Nogal
Frank Packheiser
Vicente Ranieri
Ewerson Palacio
Andre Spahni
Chen Zhu
Parwez Hamid

Redbooks



International Technical Support Organization

IBM zEnterprise EC12 Technical Guide

December 2013

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

Second Edition (December 2013)

This edition applies to the IBM zEnterprise EC12 and the IBM zEnterprise BladeCenter Extension Model 003.

© Copyright International Business Machines Corporation 2013. 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

Notices	xv
Trademarks	xvi
Preface	xvii
Authors	xvii
Now you can become a published author, too!	xxi
Comments welcome	xxi
Stay connected to IBM Redbooks	xxi
Chapter 1. Introducing the IBM zEnterprise EC12	1
1.1 zEC12 highlights	4
1.1.1 Processor and memory	4
1.1.2 Capacity and performance	4
1.1.3 I/O subsystem and I/O features	5
1.1.4 Virtualization	7
1.1.5 Increased flexibility with z/VM mode logical partition	7
1.1.6 zAware mode logical partition	8
1.1.7 IBM System z Advanced Workload Analysis Reporter	8
1.1.8 Flash Express	8
1.1.9 10GbE RoCE Express	9
1.1.10 zEDC Express	9
1.1.11 IBM Mobile Systems Remote	9
1.1.12 Reliability, availability, and serviceability	9
1.2 zEC12 technical overview	10
1.2.1 Models	10
1.2.2 Model upgrade paths	11
1.2.3 Frames	12
1.2.4 Processor cage	13
1.2.5 I/O connectivity: PCIe and InfiniBand	15
1.2.6 I/O subsystems	15
1.2.7 Coupling and Server Time Protocol connectivity	19
1.2.8 Special purpose features	21
1.2.9 Reliability, availability, and serviceability	24
1.3 Hardware Management Consoles and Support Elements	24
1.4 IBM zEnterprise BladeCenter Extension (zBX) Model 003	25
1.4.1 Blades	25
1.4.2 IBM WebSphere DataPower Integration Appliance XI50 for zEnterprise	26
1.5 Unified Resource Manager	26
1.6 Operating systems and software	27
1.6.1 Supported operating systems	27
1.6.2 IBM compilers	28
Chapter 2. Central processor complex hardware components	29
2.1 Frames and cage	30
2.1.1 A frame	31
2.1.2 Z frame	32
2.1.3 I/O cage, I/O drawer, and PCIe I/O drawer features	32
2.1.4 Top exit I/O cabling	34
2.2 Book concept	34

2.2.1	Book interconnect topology	35
2.2.2	Dual external clock facility (ECF)	36
2.2.3	Oscillator	37
2.2.4	System control	38
2.2.5	Book power	38
2.3	Multiple chip module (MCM)	39
2.4	Processor unit (PU) and storage control (SC) chips	39
2.4.1	PU chip	40
2.4.2	Processor unit (core)	41
2.4.3	PU characterization	43
2.4.4	Storage control (SC) chip	44
2.4.5	Cache level structure	45
2.5	Memory	45
2.5.1	Memory subsystem topology	47
2.5.2	Redundant array of independent memory (RAIM)	47
2.5.3	Memory configurations	48
2.5.4	Memory upgrades	51
2.5.5	Book replacement and memory	51
2.5.6	Flexible Memory Option	51
2.5.7	Pre-planned memory	52
2.6	Reliability, availability, and serviceability (RAS)	53
2.6.1	RAS in the CPC memory subsystem	53
2.6.2	General zEC12 RAS features	54
2.7	Connectivity	54
2.7.1	Redundant I/O interconnect	56
2.7.2	Enhanced book availability	58
2.7.3	Book upgrade	58
2.8	Model configurations	59
2.8.1	Upgrades	60
2.8.2	Concurrent PU conversions	61
2.8.3	Model capacity identifier	62
2.8.4	Model capacity identifier and MSU value	63
2.8.5	Capacity Backup	63
2.8.6	On/Off Capacity on Demand and CPs	66
2.9	Power and cooling	66
2.9.1	Power consumption	67
2.9.2	High voltage DC power	67
2.9.3	Internal Battery Feature (IBF)	68
2.9.4	Power capping and power saving	68
2.9.5	Power estimation tool	68
2.9.6	Cooling	69
2.9.7	Radiator Unit	69
2.9.8	Water Cooling Unit (WCU)	71
2.9.9	Backup air cooling system	74
2.10	Summary of zEC12 structure	74
Chapter 3. Central processor complex system design		77
3.1	Overview	78
3.2	Design highlights	78
3.3	Book design	79
3.3.1	Cache levels and memory structure	80
3.3.2	Book interconnect topology	82
3.4	Processor unit design	82

3.4.1	Out-of-order (OOO) execution	83
3.4.2	Superscalar processor	86
3.4.3	Compression and cryptography accelerators on a chip	86
3.4.4	Decimal floating point (DFP) accelerator	87
3.4.5	IEEE floating point	88
3.4.6	Processor error detection and recovery	88
3.4.7	Branch prediction	89
3.4.8	Wild branch	89
3.4.9	Translation lookaside buffer (TLB)	90
3.4.10	Instruction fetching, decoding, and grouping	90
3.4.11	Extended Translation Facility	90
3.4.12	Instruction set extensions	91
3.4.13	Transactional execution (TX)	91
3.4.14	Runtime instrumentation (RI)	91
3.5	Processor unit functions	91
3.5.1	Overview	91
3.5.2	Central processors	93
3.5.3	Integrated Facility for Linux	94
3.5.4	Internal Coupling Facility	94
3.5.5	System z Application Assist Processors (zAAPs)	96
3.5.6	System z Integrated Information Processor (zIIP)	99
3.5.7	zAAP on zIIP capability	101
3.5.8	System assist processors (SAPs)	102
3.5.9	Reserved processors	102
3.5.10	Integrated firmware processor	103
3.5.11	Processor unit assignment	103
3.5.12	Sparing rules	104
3.5.13	Increased flexibility with z/VM mode partitions	104
3.6	Memory design	105
3.6.1	Overview	105
3.6.2	Central storage (CS)	106
3.6.3	Expanded storage	107
3.6.4	Hardware system area (HSA)	107
3.7	Logical partitioning	108
3.7.1	Overview	108
3.7.2	Storage operations	114
3.7.3	Reserved storage	117
3.7.4	Logical partition storage granularity	118
3.7.5	LPAR dynamic storage reconfiguration (DSR)	118
3.8	Intelligent Resource Director (IRD)	118
3.9	Clustering technology	120
3.9.1	Coupling facility control code (CFCC)	121
3.9.2	Dynamic CF dispatching	124
Chapter 4	Central processor complex I/O system structure	127
4.1	Introduction to InfiniBand and PCIe	128
4.1.1	InfiniBand specification	129
4.1.2	Data, signaling, and link rates	129
4.1.3	PCIe	129
4.2	I/O system overview	130
4.2.1	Characteristics	130
4.2.2	Summary of supported I/O features	131
4.3	I/O cages	131

4.4 I/O drawers	134
4.5 PCIe I/O drawers.	136
4.6 I/O cage, I/O drawer, and PCIe I/O drawer offerings	140
4.7 Fanouts	141
4.7.1 HCA2-C fanout (FC 0162).	143
4.7.2 PCIe copper fanout (FC 0169)	143
4.7.3 HCA2-O (12x IFB) fanout (FC 0163).	143
4.7.4 HCA2-O LR (1x IFB) fanout (FC 0168).	144
4.7.5 HCA3-O (12x IFB) fanout (FC 0171).	145
4.7.6 HCA3-O LR (1x IFB) fanout (FC 0170).	146
4.7.7 Fanout considerations.	147
4.7.8 Fanout summary	150
4.8 I/O feature cards	150
4.8.1 I/O feature card ordering information	151
4.8.2 PCHID report	152
4.9 Connectivity.	154
4.9.1 I/O feature support and configuration rules.	154
4.9.2 IBM ESCON channels	158
4.9.3 FICON channels	158
4.9.4 OSA-Express5S	162
4.9.5 OSA-Express4S features	164
4.9.6 OSA-Express3 features	167
4.9.7 OSA-Express for ensemble connectivity.	170
4.9.8 HiperSockets.	171
4.10 Parallel Sysplex connectivity.	173
4.10.1 Coupling links	173
4.10.2 External clock facility.	179
4.11 Cryptographic functions	179
4.11.1 CPACF functions (FC 3863)	179
4.11.2 Crypto Express4S feature (FC 0865)	179
4.11.3 Crypto Express3 feature (FC 0864)	179
4.12 Integrated firmware processor	180
4.13 Flash Express	180
4.14 10 GbE RoCE Express	181
4.15 zEDC Express.	182
Chapter 5. Central processor complex channel subsystem	185
5.1 Channel subsystem.	186
5.1.1 Multiple channel subsystems concept.	186
5.1.2 CSS elements	187
5.1.3 Multiple subchannel sets.	188
5.1.4 Parallel access volumes and extended address volumes.	190
5.1.5 Logical partition name and identification.	191
5.1.6 Physical channel ID	192
5.1.7 Channel spanning	192
5.1.8 Multiple CSS construct	194
5.1.9 Adapter ID (AID)	194
5.1.10 Channel subsystem enhancement for I/O resilience.	195
5.2 I/O configuration management	195
5.3 Channel subsystem summary.	196
5.4 System-initiated CHPID reconfiguration	197
5.5 Multipath initial program load (IPL)	198

Chapter 6. Cryptography	199
6.1 Cryptographic synchronous functions	200
6.2 Cryptographic asynchronous functions	200
6.2.1 Secure key functions	201
6.2.2 Additional functions	202
6.3 CPACF protected key	203
6.4 PKCS #11 overview	205
6.4.1 PKCS #11 model	205
6.4.2 z/OS PKCS #11 implementation	206
6.4.3 Secure IBM Enterprise PKCS #11 (EP11) Coprocessor	208
6.5 Cryptographic feature codes	209
6.6 CP Assist for Cryptographic Function (CPACF)	210
6.7 Crypto Express4S	210
6.8 Crypto Express3	212
6.9 Tasks that are run by PCIe Crypto Express	215
6.9.1 PCIe Crypto Express as a CCA coprocessor	216
6.9.2 PCIe Crypto Express as an EP11 coprocessor	217
6.9.3 PCIe Crypto Express as an accelerator	217
6.9.4 IBM Common Cryptographic Architecture (CCA) Enhancements	218
6.10 TKE workstation feature	219
6.10.1 TKE 7.0 Licensed Internal Code (LIC)	219
6.10.2 TKE 7.1 Licensed Internal Code (LIC)	220
6.10.3 TKE 7.2 Licensed Internal Code (LIC)	222
6.10.4 Logical partition, TKE host, and TKE target	223
6.10.5 Optional smart card reader	223
6.11 Cryptographic functions comparison	223
6.12 Software support	225
Chapter 7. zEnterprise BladeCenter Extension (zBX) Model 003	227
7.1 zBX concepts	228
7.2 zBX hardware description	228
7.2.1 zBX racks	229
7.2.2 Top of Rack (ToR) switches	231
7.2.3 zBX BladeCenter chassis	233
7.2.4 zBX blades	236
7.2.5 Power distribution unit (PDU)	241
7.3 zBX entitlements, firmware, and upgrades	241
7.3.1 zBX management	243
7.3.2 zBX firmware	244
7.4 zBX connectivity	244
7.4.1 Intranode management network (INMN)	245
7.4.2 Primary and alternate HMCs	247
7.4.3 Intraensemble data network (IEDN)	249
7.4.4 Network connectivity rules with zBX	252
7.4.5 Network security considerations with zBX	252
7.4.6 zBX storage connectivity	254
7.5 zBX connectivity examples	257
7.5.1 Single-node ensemble with a zBX	257
7.5.2 Dual-node ensemble with a single zBX	259
7.5.3 Dual-node ensemble with two zBXs	260
7.6 References	260
Chapter 8. Software support	263
8.1 Operating systems summary	264

8.2 Support by operating system	264
8.2.1 z/OS	265
8.2.2 z/VM	265
8.2.3 z/VSE	265
8.2.4 z/TPF	265
8.2.5 Linux on System z	266
8.2.6 zEC12 function support summary	266
8.3 Support by function	278
8.3.1 Single system image	278
8.3.2 zAAP support	279
8.3.3 zIIP support	280
8.3.4 zAAP on zIIP capability	281
8.3.5 Transactional Execution (TX)	281
8.3.6 Maximum main storage size	282
8.3.7 Flash Express	282
8.3.8 zEnterprise Data Compression (zEDC) Express	284
8.3.9 10GbE RoCE Express	284
8.3.10 Large page support	285
8.3.11 Guest support for execute-extensions facility	285
8.3.12 Hardware decimal floating point	285
8.3.13 Up to 60 logical partitions	286
8.3.14 Separate LPAR management of PUs	286
8.3.15 Dynamic LPAR memory upgrade	287
8.3.16 LPAR physical capacity limit enforcement	287
8.3.17 Capacity Provisioning Manager	287
8.3.18 Dynamic PU add	288
8.3.19 HiperDispatch	288
8.3.20 The 63.75-K subchannels	289
8.3.21 Multiple subchannel sets	289
8.3.22 Third subchannel set	290
8.3.23 IPL from an alternate subchannel set	290
8.3.24 MIDAW facility	290
8.3.25 HiperSockets Completion Queue	290
8.3.26 HiperSockets integration with the intraensemble data network (IEDN)	291
8.3.27 HiperSockets Virtual Switch Bridge	291
8.3.28 HiperSockets Multiple Write Facility	292
8.3.29 HiperSockets IPv6	292
8.3.30 HiperSockets Layer 2 support	292
8.3.31 HiperSockets network traffic analyzer for Linux on System z	293
8.3.32 FICON Express8S	293
8.3.33 FICON Express8	294
8.3.34 z/OS discovery and autoconfiguration (zDAC)	294
8.3.35 High performance FICON (zHPF)	295
8.3.36 Request node identification data	297
8.3.37 24k subchannels for the FICON Express	297
8.3.38 Extended distance FICON	297
8.3.39 Platform and name server registration in FICON channel	298
8.3.40 FICON link incident reporting	298
8.3.41 FCP provides increased performance	298
8.3.42 N-Port ID virtualization (NPIV)	298
8.3.43 OSA-Express5S 10-Gigabit Ethernet LR and SR	299
8.3.44 OSA-Express5S Gigabit Ethernet LX and SX	299
8.3.45 OSA-Express5S 1000BASE-T Ethernet	300

8.3.46	OSA-Express4S 10-Gigabit Ethernet LR and SR	301
8.3.47	OSA-Express4S Gigabit Ethernet LX and SX	302
8.3.48	OSA-Express4S 1000BASE-T Ethernet	303
8.3.49	OSA-Express3 10-Gigabit Ethernet LR and SR	303
8.3.50	OSA-Express3 Gigabit Ethernet LX and SX	304
8.3.51	OSA-Express3 1000BASE-T Ethernet	305
8.3.52	Open Systems Adapter for IBM zAware	307
8.3.53	Open Systems Adapter for Ensemble	307
8.3.54	Intranode management network (INMN)	307
8.3.55	Intraensemble data network (IEDN)	308
8.3.56	OSA-Express5S and OSA-Express4S NCP support (OSN)	308
8.3.57	Integrated Console Controller	309
8.3.58	VLAN management enhancements	309
8.3.59	GARP VLAN Registration Protocol	309
8.3.60	Inbound workload queuing (IWQ) for OSA-Express5S, OSA-Express4S, and OSA-Express3	310
8.3.61	Inbound workload queuing (IWQ) for Enterprise Extender	310
8.3.62	Query and display OSA configuration	311
8.3.63	Link aggregation support for z/VM	311
8.3.64	QDIO data connection isolation for z/VM	311
8.3.65	QDIO interface isolation for z/OS	312
8.3.66	QDIO optimized latency mode (OLM)	312
8.3.67	Large send for IPv6 packets	312
8.3.68	OSA-Express5S and OSA-Express4S checksum offload	313
8.3.69	Checksum offload for IPv4 packets when in QDIO mode	313
8.3.70	Adapter interruptions for QDIO	313
8.3.71	OSA Dynamic LAN idle	314
8.3.72	OSA Layer 3 virtual MAC for z/OS environments	314
8.3.73	QDIO Diagnostic Synchronization	314
8.3.74	Network Traffic Analyzer	315
8.3.75	Program directed re-IPL	315
8.3.76	Coupling over InfiniBand	315
8.3.77	Dynamic I/O support for InfiniBand CHPIDs	316
8.4	Cryptographic support	316
8.4.1	CP Assist for Cryptographic Function (CPACF)	316
8.4.2	Crypto Express4S	317
8.4.3	Crypto Express3	318
8.4.4	Web deliverables	318
8.4.5	z/OS ICSF FMIDs	318
8.4.6	ICSF migration considerations	320
8.5	z/OS migration considerations	320
8.5.1	General guidelines	321
8.5.2	Hardware Configuration Definition (HCD)	321
8.5.3	InfiniBand coupling links	321
8.5.4	Large page support	321
8.5.5	HiperDispatch	321
8.5.6	Capacity Provisioning Manager	322
8.5.7	Decimal floating point and z/OS XL C/C++ considerations	323
8.5.8	IBM System z Advanced Workload Analysis Reporter (IBM zAware)	323
8.6	Coupling facility and CFCC considerations	324
8.7	MIDAW facility	326
8.7.1	MIDAW technical description	326
8.7.2	Extended format data sets	328

8.7.3 Performance benefits	329
8.8 IOCP	330
8.9 Worldwide port name (WWPN) tool	330
8.10 ICKDSF	331
8.11 zEnterprise BladeCenter Extension (zBX) Model 003 software support	331
8.11.1 IBM blades	331
8.11.2 IBM WebSphere DataPower Integration Appliance XI50 for zEnterprise	332
8.12 Software licensing considerations	332
8.12.1 MLC pricing metrics	333
8.12.2 Advanced Workload License Charges (AWLC)	334
8.12.3 System z new application license charges (zNALC)	335
8.12.4 Select application license charges (SALC)	335
8.12.5 Midrange Workload License Charges (MWLC)	335
8.12.6 Parallel Sysplex License Charges (PSLC)	336
8.12.7 System z International Program License Agreement (IPLA)	336
8.13 References	337
Chapter 9. System upgrades	339
9.1 Upgrade types	340
9.1.1 Overview of upgrade types	340
9.1.2 Terminology related to CoD for zEC12 systems	341
9.1.3 Permanent upgrades	343
9.1.4 Temporary upgrades	344
9.2 Concurrent upgrades	344
9.2.1 Model upgrades	345
9.2.2 Customer Initiated Upgrade (CIU) facility	347
9.2.3 Summary of concurrent upgrade functions	350
9.3 Miscellaneous equipment specification (MES) upgrades	351
9.3.1 MES upgrade for processors	352
9.3.2 MES upgrades for memory	354
9.3.3 MES upgrades for I/O	355
9.3.4 MES upgrades for the zBX	356
9.3.5 Summary of plan-ahead features	358
9.4 Permanent upgrade through the CIU facility	358
9.4.1 Ordering	360
9.4.2 Retrieval and activation	361
9.5 On/Off Capacity on Demand	362
9.5.1 Overview	362
9.5.2 Ordering	363
9.5.3 On/Off CoD testing	366
9.5.4 Activation and deactivation	367
9.5.5 Termination	368
9.5.6 z/OS capacity provisioning	369
9.6 Capacity for Planned Event (CPE)	373
9.7 Capacity Backup (CBU)	374
9.7.1 Ordering	374
9.7.2 CBU activation and deactivation	376
9.7.3 Automatic CBU enablement for GDPS	377
9.8 Nondisruptive upgrades	378
9.8.1 Components	378
9.8.2 Concurrent upgrade considerations	379
9.9 Summary of Capacity on Demand offerings	383
9.10 Reference	383

Chapter 10. Reliability, availability, and serviceability	385
10.1 zEC12 availability characteristics	386
10.2 zEC12 RAS functions	388
10.2.1 Scheduled outages	389
10.2.2 Unscheduled outages	390
10.3 zEC12 enhanced book availability (EBA)	391
10.3.1 EBA planning considerations	391
10.3.2 Enhanced book availability processing	393
10.4 zEC12 enhanced driver maintenance (EDM)	399
10.5 RAS capability for the HMC and SE	400
10.6 RAS capability for zBX	401
10.7 Considerations for PowerHA in zBX environment.	404
10.8 IBM System z Advanced Workload Analysis Reporter (IBM zAware).	405
10.9 RAS capability for Flash Express	406
Chapter 11. Environmental requirements	409
11.1 zEC12 power and cooling	410
11.1.1 Power consumption	410
11.1.2 Internal Battery Feature	411
11.1.3 Emergency power-off switch.	412
11.1.4 Cooling requirements	412
11.2 IBM zEnterprise EC12 physical specifications	415
11.3 IBM zEnterprise EC12 physical planning	416
11.3.1 Raised floor or non-raised floor.	416
11.3.2 Top Exit Power feature	418
11.3.3 Top Exit I/O Cabling feature	418
11.3.4 Weight distribution plate	419
11.3.5 Three-in-one bolt-down kit for raised floor	419
11.4 zBX environmental requirements	420
11.4.1 zBX configurations	420
11.4.2 zBX power components	420
11.4.3 zBX cooling	421
11.4.4 zBX physical specifications.	423
11.5 Energy management.	424
11.5.1 Power estimation tool	426
11.5.2 Query maximum potential power	426
11.5.3 System Activity Display and Monitors Dashboard.	427
11.5.4 IBM Systems Director Active Energy Manager.	427
11.5.5 Unified Resource Manager: Energy management	428
Chapter 12. Hardware Management Console and Support Element	431
12.1 Introduction to the HMC and SE	432
12.2 SE driver support with the new HMC	432
12.2.1 HMC Feature Code 0092 changes	433
12.3 HMC and SE enhancements and changes.	433
12.3.1 HMC media support	436
12.3.2 Tree Style User Interface and Classic Style User Interface	436
12.4 HMC and SE connectivity	436
12.4.1 Hardware prerequisite changes	438
12.4.2 TCP/IP Version 6 on the HMC and SE	439
12.4.3 Assigning addresses to the HMC and SE.	439
12.5 Remote Support Facility (RSF)	440
12.5.1 Security characteristics.	440
12.5.2 RSF connections to IBM and the Enhanced IBM Service Support System	441

12.5.3 HMC and SE remote operations	442
12.6 HMC and SE key capabilities	442
12.6.1 Central processor complex (CPC) management	443
12.6.2 Logical partition management	443
12.6.3 Operating system communication	444
12.6.4 HMC and SE microcode	445
12.6.5 Monitoring	447
12.6.6 IBM Mobile Systems Remote	450
12.6.7 Capacity on Demand (CoD) support	451
12.6.8 Feature on Demand (FoD) support	452
12.6.9 Server Time Protocol support	453
12.6.10 NTP client and server support on the HMC	454
12.6.11 Security and user ID management	456
12.6.12 System Input/Output Configuration Analyzer on the SE and HMC	458
12.6.13 Automated operations	458
12.6.14 Cryptographic support	459
12.6.15 z/VM virtual machine management	460
12.6.16 Installation support for z/VM using the HMC	461
12.7 HMC in an ensemble	461
12.7.1 Unified Resource Manager	461
12.7.2 Ensemble definition and management	465
12.7.3 HMC availability	466
12.7.4 Considerations for multiple HMCs	467
12.7.5 HMC browser session to a primary HMC	467
12.7.6 HMC ensemble topology	467
Chapter 13. Performance	469
13.1 LSPR workload suite	470
13.2 Fundamental components of workload capacity performance	470
13.3 Relative nest intensity	472
13.4 LSPR workload categories based on relative nest intensity	474
13.5 Relating production workloads to LSPR workloads	474
13.6 Workload performance variation	476
Appendix A. IBM System z Advanced Workload Analysis Reporter (IBM zAware)	479
A.1 Troubleshooting in complex IT environments	480
A.2 Introducing IBM zAware	481
A.2.1 Value of IBM zAware	482
A.2.2 IBM z/OS Solutions to improve problem diagnostic procedures	483
A.3 Understanding IBM zAware technology	484
A.3.1 Training period	489
A.3.2 Priming IBM zAware	489
A.3.3 IBM zAware ignore message support	489
A.3.4 IBM zAware graphical user interface	490
A.3.5 IBM zAware is complementary to your existing tools	490
A.4 Learning IBM zAware prerequisites	490
A.4.1 IBM zAware features and ordering	491
A.4.2 IBM zAware operating requirements	493
A.5 Configuring and using IBM zAware virtual appliance	494
Appendix B. Channel options	497
Appendix C. Flash Express	501
C.1 Flash Express overview	502

C.2 Using Flash Express	504
C.3 Security on Flash Express	508
C.3.1 Integrated Key Controller	508
C.3.2 Key serving topology	509
C.3.3 Error recovery scenarios	510
Appendix D. Remote Direct Memory Access over Converged Ethernet (RoCE)	513
D.1 Overview	514
D.1.1 Remote Direct Memory Access (RDMA) technology overview	514
D.1.2 Shared Memory Communications – RDMA (SMC-R)	515
D.2 Hardware	516
D.2.1 10GbE RoCE Express feature	516
D.2.2 10GbE RoCE Express configuration sample	517
D.2.3 Hardware configuration definitions	519
D.3 Software exploitation	520
D.3.1 SMC-R support overview	520
D.3.2 SMC-R use cases for z/OS to z/OS	521
D.3.3 Enabling SMC-R support in z/OS Communications Server	523
Appendix E. zEnterprise Data Compression (zEDC) Express	525
E.1 Overview	526
E.2 zEDC Express	526
E.3 Software support	527
Appendix F. Native Peripheral Component Interconnect Express (PCIe)	529
F.1 Design of native PCIe I/O adapter management	530
F.2 Native PCIe adapter	530
F.3 Integrated firmware processor (IFP)	530
F.4 Resource Groups (RGs)	531
F.5 Native PCIe feature plugging rules	532
F.6 Management tasks	533
F.6.1 Firmware update	533
F.6.2 Error recovery	533
F.6.3 Maintenance tasks	533
F.7 zEDC Express	533
F.8 10GbE RoCE Express	534
F.9 Native PCIe feature definitions	534
Related publications	537
IBM Redbooks	537
Other publications	537
Online resources	537
Help from IBM	539

Notices

This information was developed for products and services offered in the U.S.A.

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, Armonk, NY 10504-1785 U.S.A.

IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM's sole discretion. Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision. The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code, or functionality.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: 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 states 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 supply in any way it believes appropriate without incurring any obligation to you.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

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.

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 the names and addresses used by an actual business enterprise 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.

Trademarks

IBM, the IBM logo, and [ibm.com](http://www.ibm.com) are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. These and other IBM trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at <http://www.ibm.com/legal/copytrade.shtml>

The following terms are trademarks of the International Business Machines Corporation in the United States, other countries, or both:

AIX®	IBM®	Resource Measurement Facility™
BladeCenter®	IBM Systems Director Active Energy Manager™	RETAIN®
CICS®	IMS™	RMF™
Cognos®	Language Environment®	Sysplex Timer®
DataPower®	MQSeries®	System p®
DB2®	OMEGAMON®	System Storage®
DB2 Connect™	Parallel Sysplex®	System x®
developerWorks®	Passport Advantage®	System z®
Distributed Relational Database Architecture™	POWER®	System z10®
Domino®	Power Systems™	System z9®
DRDA®	POWER6®	SystemMirror™
DS8000®	POWER7®	Tivoli®
ECKD™	PowerHA®	VTAM®
ESCON®	PowerPC®	WebSphere®
FICON®	PowerVM®	z/Architecture®
FlashCopy®	PR/SM™	z/OS®
GDPS®	Processor Resource/Systems Manager™	z/VM®
Geographically Dispersed Parallel Sysplex™	RACF®	z/VSE®
Global Technology Services®	Redbooks®	z10™
HACMP™	Redpapers™	z9®
HyperSockets™	Redbooks (logo)  ®	zEnterprise®
HyperSwap®	Resource Link®	zSeries®

The following terms are trademarks of other companies:

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

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.

Preface

The popularity of the Internet and the affordability of IT hardware and software have resulted in an explosion of applications, architectures, and platforms. Workloads have changed. Many applications, including mission-critical ones, are deployed on various platforms, and the IBM® System z® design has adapted to this change. It takes into account a wide range of factors, including compatibility and investment protection, to match the IT requirements of an enterprise.

This IBM Redbooks® publication addresses the new IBM zEnterprise® System. This system consists of the **IBM zEnterprise EC12** (zEC12), an updated IBM zEnterprise Unified Resource Manager, and the IBM zEnterprise BladeCenter® Extension (zBX) Model 003.

The **zEC12** is designed with improved scalability, performance, security, resiliency, availability, and virtualization. The superscalar design allows the zEC12 to deliver a record level of capacity over the prior System z servers. It is powered by 120 of the world's most powerful microprocessors. These microprocessors run at 5.5 GHz and are capable of running more than 75,000 millions of instructions per second (MIPS). The **zEC12** Model HA1 is estimated to provide up to 50% more total system capacity than the IBM zEnterprise 196 (z196) Model M80.

The zBX Model 003 infrastructure works with the **zEC12** to enhance System z virtualization and management. It does so through an integrated hardware platform that spans mainframe, IBM POWER7®, and IBM System x® technologies. Through the Unified Resource Manager, the zEnterprise System is managed as a single pool of resources, integrating system and workload management across the environment.

This book provides information about the zEnterprise System 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 zEnterprise System 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 System z® technology and terminology.

Authors

This book was produced by a team of specialists from around the world working at the International Technical Support Organization, Poughkeepsie Center.

Ivan Doboš is an IBM Certified Consulting IT Specialist working as a mainframe consultant at STG Lab Services Central & Eastern Europe. He has 15 years of experience with IBM System z. He joined IBM in 2003 and worked in different sales and technical roles supporting mainframe clients, as the Technical Leader for Linux on System z projects in the System z Benchmark Center, IT Optimization Consultant in the System z New Technology Center, and Mainframe Technical Sales Manager in Central & Eastern Europe. During the past ten years, he has worked with many clients and spent most of his time supporting new workloads on System z projects. Ivan has authored several IBM Redbooks and Redpapers™ publications.

Wolfgang Fries is a Senior Consultant in the System z Hardware Support Center in Germany. He spent several years at the European Support Center in Montpellier, France, providing international support for System z servers. Wolfgang has 35 years of experience in

supporting large System z clients. His areas of expertise include System z servers and connectivity.

Hua Bin Chu is an Advisory IT Specialist in China. He has seven years of experience with IBM Global Technology Services® and in supporting clients of large System z products. His areas of expertise include IBM z/OS®, Parallel Sysplex®, System z high availability solutions, and IBM Geographically Dispersed Parallel Sysplex (GDPS®).

Luiz Fadel is an IBM Distinguished Engineer responsible for supporting System z for the Latin America region, part of the Growth Markets Unit. He joined IBM in 1969 and has supported large systems ever since, including working on two assignments with the International Technical Support Organization (ITSO). Luiz is a member of the Latin America Advanced Technical Support team, which is responsible for handling Client Critical Situations and client claims within System z, Early Support Programs, new product installations, internal product announcements, and second-level client support, as well as managing complex proofs of concepts (POCs). He is a member of the zChampions team and the co-author of several IBM Redbooks publications.

Octavian Lascu is a Senior IT Consultant for IBM Romania with over 20 years of experience. He specializes in designing and supporting complex IT infrastructure environments (systems, storage, and networking), including high availability and disaster recovery solutions and high-performance computing deployments. He has developed and taught over 50 workshops for technical audiences around the world. He has authored several Redbooks and Redpapers.

Martijn Raave is an IBM Certified System z Client Technical Specialist for IBM Systems and Technology Group (STG) in the Netherlands. Over a period of 15 years, his professional career has revolved around the mainframe platform. Before joining IBM through a strategic outsourcing deal in 2005, he worked for a large Dutch client as a systems programmer with expertise in the areas of z/OS, (Globally Dispersed) Parallel Sysplex, and hardware. Four years ago, he decided to explore the other aspects of the mainframe ecosystem within IBM and joined STG in his current role. As a Client Technical Specialist, he supports several Dutch System z clients, IBM Business Partners, and IBM Sales Representatives on technical topics and in sales engagements. He is also a board member of Guide Share Europe (GSE) Netherlands.

Fernando Nogal is an IBM Certified Consulting IT Specialist working as an STG Technical Consultant for the Spain, Portugal, Greece, and Israel IMT. He specializes in advanced infrastructures and architectures. In his 30+ years with IBM, he has held various technical positions, mainly providing support for mainframe clients. Previously, he was on assignment to the Europe Middle East and Africa (EMEA) System z Technical Support group, working full-time on complex solutions for e-business. His job includes presenting and consulting in architectures and infrastructures, and providing strategic guidance to System z clients about the establishment and enablement of advanced technologies on System z, including the z/OS, IBM z/VM®, and Linux environments. He is a zChampion and a member of the System z Business Leaders Council. An accomplished writer, he has authored and co-authored over 28 IBM Redbooks publications and several technical papers. Other activities include serving as a University Ambassador. He travels extensively on direct client engagements, and as a speaker at IBM and client events and trade shows.

Frank Packheiser is a Senior zIT Specialist at the Field Technical Sales Support office in Germany. He has 21 years of experience in zEnterprise, System z, IBM zSeries®, and predecessor mainframe servers. He has worked for 10 years for the IBM education center in Germany, developing and providing professional training. He also provides professional services to System z and mainframe clients. In 2008 and 2009, he supported clients in Middle East/North Africa (MENA) as a zIT Architect. Besides co-authoring several Redbooks

publications since 1999, he has been an ITSO guest speaker at ITSO workshops for the last two years.

Vicente Ranieri is an Executive IT Specialist and the Lead Architect at the High End Design Center in Latin America. He has more than 30 years of experience working for IBM. Vicente used to be a member of the Advanced Technical Skills (ATS) team in Latin America. His areas of expertise include System z security, IBM Parallel Sysplex, System z hardware, and z/OS. Vicente has co-authored several Redbooks publications and has been an ITSO guest speaker since 2001, teaching System z security update workshops worldwide. Vicente is certified as a Distinguished IT Specialist by the Open group and he is a member of the zChampions team, of the Technology Leadership Council - Brazil, and of the IBM Academy of Technology.

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 System z Server technical and client support, mainframe architecture, infrastructure implementation, and design. He is an ITSO System z 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 System z 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.

André Spahni is a Senior System Service Representative working for IBM Global Technology Services in Switzerland. He has 10 years of experience working with and supporting System z clients. André has worked for the Technical Support Competence Center (TSCC) Hardware FE System z for Switzerland, Germany, and Austria since 2008. His areas of expertise include System z hardware, Parallel Sysplex, and connectivity.

Chen Zhu is a Consulting System Service Representative at IBM Global Technology Services in Shanghai, China. He joined IBM in 1998 to support and maintain System z products for clients throughout China. Chen has worked in the Technical Support Group (TSG) providing second-level support to System z clients since 2005. His areas of expertise include System z hardware, IBM Parallel Sysplex, Tape Library, and IBM Fibre Channel connection (FICON®).

Special thanks to **Parwez Hamid** for guiding the team through the massive amount of information that was required to put this book together.

Parwez Hamid has been an Executive IT Consultant with the IBM Server and Technology Group and a Technical Staff member of the IBM UK Technical Council. During the past 39 years, he has worked in various IT roles within IBM. Since 1988, he has worked with many IBM mainframe clients, mainly introducing new technology. Currently, he works as a Consultant for System z in Poughkeepsie and provides technical support for the IBM System z hardware product portfolio. Parwez continues to co-author ITSO Redbooks publications and he prepares technical material for the worldwide announcement of System z servers. Parwez works closely with System z product development in Poughkeepsie, and provides input and feedback for future product plans. Parwez teaches and presents at numerous IBM user group and internal conferences and teaches at ITSO workshops.

Thanks to the following people for their contributions to this project (in no particular order):

Debbie Beatrice
James Caffrey
Ellen Carbarnes

Edward Chencinski
Doris Conti
Kathleen Fadden
Darelle Gent
Susan Greenlee
Christian Jacobi
Paul Jones
Amy Lander
Kenneth Kerr
Georgette Kurdt
Rob Overton
Patrick Rausch
Charles Reimers
Peter Relson
Robert (Bob) Rogers
Donald M. Smith
James Sweeny
Barbara Weiler
Martin Ziskind
Doug Zobre
IBM Poughkeepsie

Dean St Pier
IBM Fishkill

Gregory Hutchison
IBM Gaithersburg

Leslie Geer
Romney White
IBM Endicott

Jim Coon
IBM Rochester

Marcel Mitran
IBM Canada

Jia Lei Ma
IBM China

Juergen Kuhn
Horst Sinram
IBM Germany

Ella Buslovich
ITSO Poughkeepsie Center

The authors of the first edition of this book, *IBM zEnterprise EC12 Technical Guide*, SG24-8049-00, are Ivan Dobos, Octavian Lascu, Parwez Hamid, Hand Wijngaard, Vicente Ranieri, Jr., Swee Send NG, Zhaoxu Zhang, Esra Ufacik, Andre Spahni, Frank Packheiser, Wolfgang Fries, Karan Singh, Gerard Laumay, and Fernando Nogal.

This project was supervised by Bill White, ITSO Global Content Services System z Portfolio Manager, Poughkeepsie Center.

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 the IBM zEnterprise EC12

The IBM zEnterprise EC12 (zEC12) builds on the strengths of its predecessor, the IBM zEnterprise 196. It is designed to help overcome problems in today's IT infrastructures and provide a foundation for the future. True to its zEnterprise heritage, the zEC12 continues the evolution of integrated hybrid systems, introducing the zEnterprise BladeCenter Extension (zBX) Model 003, and an updated zEnterprise Unified Resource Manager.

The IBM zEnterprise BladeCenter Extension (zBX) Model 003, similarly to its predecessor, the zBX Model 002, houses up to 112 general-purpose IBM POWER7 and IBM System x blades, and specialized solutions, such as the IBM WebSphere® DataPower® X150 for zEnterprise.

The zEC12, when managed by the Unified Resource Manager, with or without a zBX attached, constitutes a *node* in an *ensemble*. An ensemble is a collection of up to eight highly virtualized heterogeneous zEnterprise nodes. It has dedicated networks for system management and data transfer across the virtualized system images. The ensemble is managed as a single logical entity by the Unified Resource Manager functions. Multiple, diverse workloads can be deployed across its resources.

Figure 1-1 on page 2 shows the elements of an ensemble node with the IBM zEnterprise EC12.

The zEC12 has a redesigned zEnterprise chip. It is the first six-core chip in mainframe history, and operates at an industry leading, high frequency, 5.5 GHz. The zEC12 is a scalable symmetric multiprocessor (SMP) that can be configured with up to 101 processors that run concurrent production tasks, and with up to 3 TB of memory.

The zEC12 introduces several Peripheral Component Interconnect Express (PCIe) I/O features, such as exploitation of Storage Class Memory through the Flash Express feature, and Remote Direct Memory Access over Converged Ethernet. It also introduces technologies, such as data compression and decompression acceleration, and the IBM System z Advanced Workload Analysis Reporter (IBM zAware). This appliance has leading-edge pattern recognition analytics that use *heuristic techniques*, and represents the next generation of system health monitoring.

The zEC12 goes beyond previous designs while continuing to enhance the traditional mainframe qualities, delivering unprecedented performance and capacity growth. The zEC12 has a well-balanced general-purpose design that allows it to be equally at ease with compute-intensive and I/O-intensive workloads.

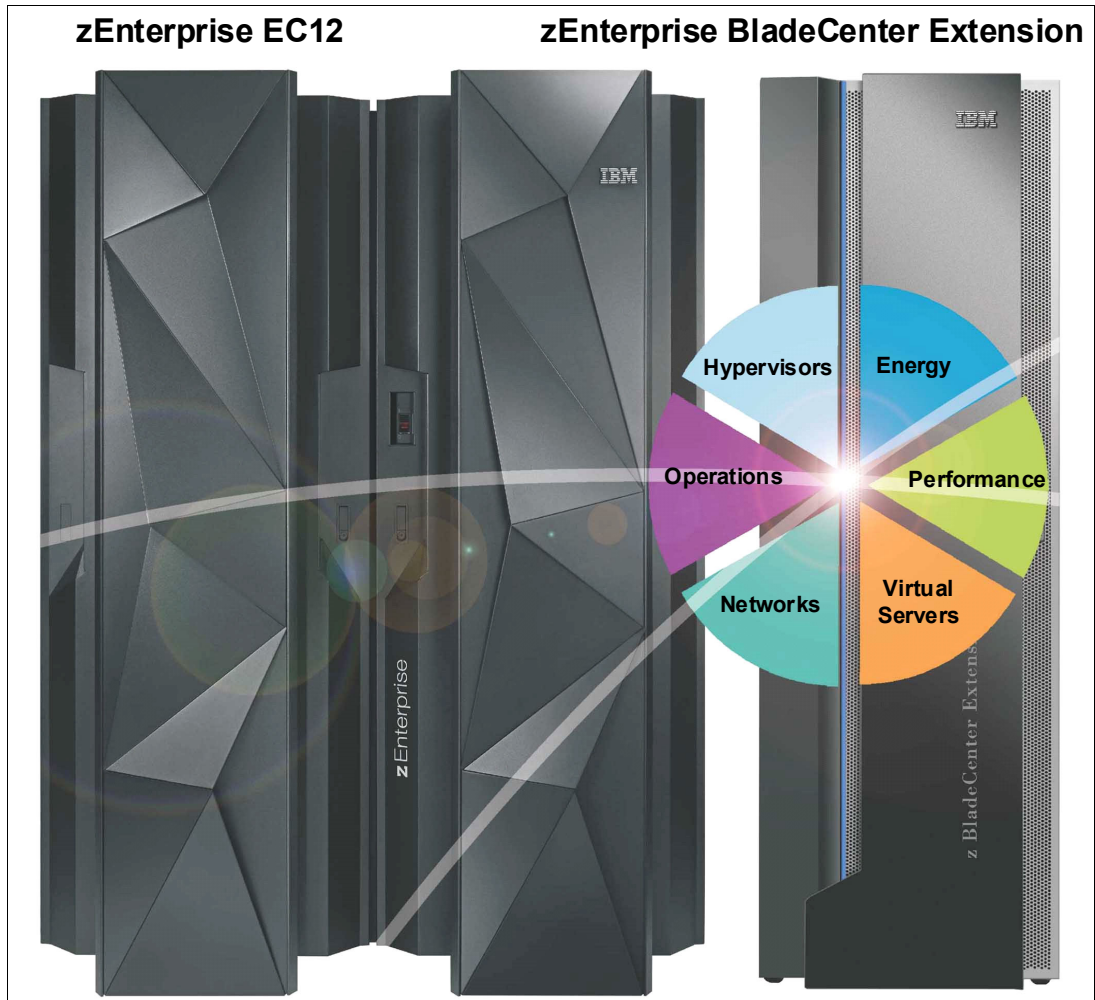


Figure 1-1 Elements of an ensemble node with the zEC12

Workloads continue to change. Multi-tier application architectures and their deployment on heterogeneous infrastructures are common today. But what is uncommon is the infrastructure setup that is needed to provide the high qualities of service that are required by mission-critical applications.

Creating and maintaining these high-level qualities of service while using a large collection of distributed components takes a great amount of knowledge and effort. It implies acquiring and installing extra equipment and software to ensure availability, security, monitoring, and management. Additional staff is required to configure, administer, troubleshoot, and tune such a complex set of separate and diverse environments. Because of platform functional differences, the resulting infrastructure is not uniform regarding those qualities of service and serviceability.

While undeniably being a key piece of the IT infrastructure, the zEC12 is also the place of choice for a large and diversified stack of software, which, when complemented with services, places the zEC12 at the heart of leading-edge solution offerings, including mobility-based and cloud-enabled applications, and big data. Its traditional strengths and characteristics, such as security, are increasingly recognized as indispensable for public acceptability of these new IT services.

The IBM holistic approach to System z design includes hardware, software, and procedures. It takes into account a wide range of factors, including compatibility and investment protection, therefore ensuring a tighter fit with the IT requirements of the entire enterprise.

IBM zEnterprise EC12 elements

The remainder of this chapter provides an overview of IBM zEnterprise EC12 features and functions.

1.1 zEC12 highlights

This section reviews some of the most important features and functions of zEC12:

- ▶ Processor and memory
- ▶ Capacity and performance
- ▶ I/O subsystem and I/O features
- ▶ Virtualization
- ▶ Increased flexibility with z/VM mode logical partition
- ▶ zAware mode logical partition
- ▶ IBM System z Advanced Workload Analysis Reporter
- ▶ Flash Express
- ▶ 10GbE RoCE Express
- ▶ zEDC Express
- ▶ Flash Express
- ▶ IBM Mobile Systems Remote
- ▶ Reliability, availability, and serviceability

1.1.1 Processor and memory

IBM continues its technology leadership with the zEC12. The zEC12 is built using the IBM modular multi-book design that supports one to four books per central processor complex (CPC). Each book contains a multiple chip module (MCM), which hosts the redesigned complementary metal-oxide semiconductor (CMOS) 13S¹ processor units, storage control chips, and connectors for I/O. The superscalar processor has enhanced out-of-order instruction execution, redesigned caches, and an expanded instruction set that includes a Transactional Execution facility, for better performance.

Depending on the model, the zEC12 can support from a minimum of 32 GB to a maximum of 3040 GB of usable memory, with up to 768 GB per book. In addition, a fixed amount of 32 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). To use the RAIM function, up to 960 GB can be physically installed per book, for a system total of 3840 GB.

1.1.2 Capacity and performance

The zEC12 provides increased processing and I/O capacity over its predecessor, the z196 system. This capacity is achieved both by increasing the performance of the individual processor units and by increasing the number of processor units (PUs) per system. 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 features helps to ensure that the zEC12 is an innovative, security-rich platform. It is designed to maximize resource exploitation and utilization, and provides the ability to integrate applications and data across the enterprise IT infrastructure.

¹ CMOS 13S is a 32-nanometer CMOS logic fabrication process.

zEC12 has five model offerings that range from one to 101 configurable PUs. The first four models (H20, H43, H66, and H89) have 27 PUs per book, and the high-capacity model (the HA1) has four 30 PU books. Model HA1 is estimated to provide up to 50% more total system capacity than the z196 Model M80, which has the same memory and power requirements. Uniprocessor performance has also increased significantly. A zEC12 Model 701 offers, on average, performance improvements of about 1.25 times over the z196 Model 701. However, variations on the observed performance increase are dependent on the workload type.

The zEC12 expands the subcapacity settings, offering three subcapacity levels for up to 20 processors that are characterized as central processors (CPs). This configuration gives a total of 161 distinct capacity settings in the system, and provides a range of over 1:320 in processing power. The zEC12 delivers 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. The zEC12 continues to offer all the specialty engines available on previous System z systems.

This comparison is based on the Large Systems Performance Reference (LSPR) mixed workload analysis. For a discussion about performance and workload variation on zEC12, see Chapter 13, “Performance” on page 469.

Workload variability

Consult the Large Systems Performance Reference (LSPR) when considering performance on the zEC12. The range of performance ratings across the individual LSPR workloads is 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 more PUs available. For more information, see Chapter 13, “Performance” on page 469.

For detailed performance information, see the LSPR website:

<https://www-304.ibm.com/servers/resourceLink/lib03060.nsf/pages/lsprindex>

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

<http://www-03.ibm.com/systems/z/resources/swprice/reference/exhibits/>

Capacity on demand (CoD)

On-demand enhancements enable clients to have more flexibility in managing and administering their temporary capacity requirements. The zEC12 supports the same architectural approach for temporary offerings as the z196. Within the zEC12, one or more flexible configuration definitions can be available to solve multiple temporary situations and multiple capacity configurations can be active simultaneously.

Up to 200 staged records can be created for many scenarios. Up to eight of these records can be installed on the server at any given time. After the records are installed, the activation of the records can be done manually, or the z/OS Capacity Provisioning Manager can automatically invoke the activation when Workload Manager (WLM) policy thresholds are reached. Tokens are available that can be purchased for On/Off Capacity on Demand (CoD) either before or after execution.

1.1.3 I/O subsystem and I/O features

The zEC12 supports both a PCIe and an InfiniBand I/O infrastructure. PCIe features are installed in PCIe I/O drawers. Up to five PCIe I/O drawers are supported, providing space for up to 160 I/O features. When upgrading a z196 or IBM z10™ EC to a zEC12, up to two I/O

drawers, that were introduced with the IBM z10 BC, and one traditional I/O cage are also supported.

There are up to 48 high-performance fanout connections for data communications between the books and the I/O infrastructure. The multiple channel subsystem (CSS) architecture allows up to four CSSs, each with 256 channels.

For I/O constraint relief, three subchannel sets are available per CSS, allowing access to a larger number of logical volumes. For improved device connectivity for parallel access volumes (PAVs), Peer-to-Peer Remote Copy (PPRC) secondaries, and IBM FlashCopy® devices, this third subchannel set allows extending the amount of addressable external storage. The zEC12 allows you to IPL from subchannel set 1 (SS1) or subchannel set 2 (SS2), in addition to subchannel set 0.

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

zEC12 connectivity supports the following I/O or special purpose features:

- ▶ Storage connectivity:
 - Fibre Channel connection (FICON):
 - FICON Express8S 10 KM long wavelength (LX) and short wavelength (SX)
 - FICON Express8 10 KM LX and SX
 - FICON Express4 10 KM LX and SX (four port cards 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
 - OSA-Express4S GbE LX and SX
 - OSA-Express4S 1000BASE-T Ethernet
 - OSA-Express3 10 GbE LR and SR
 - OSA-Express3 GbE LX and SX
 - OSA-Express3 1000BASE-T Ethernet
 - IBM HiperSockets™
 - 10 GbE Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE)
- ▶ Coupling and Server Time Protocol (STP) connectivity:
 - Parallel Sysplex InfiniBand coupling links (IFB)
 - Internal Coupling links (IC)
 - InterSystem Channel-3 (ISC-3), peer mode only

In addition, zEC12 supports the following special function features, which are installed on the PCIe I/O drawers or I/O drawers:

- Cryptography:
 - Crypto Express4S
 - Crypto Express3

- Flash Express
- zEnterprise Data Compression (zEDC) Express

1.1.4 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 SMP system. This virtualization enables full sharing of the installed resources with high security and efficiency, by configuring up to 60 LPARs, each of which has logical processors, memory, and I/O resources, assigned from the installed books and features.

LPAR configurations can be dynamically adjusted in order to optimize the virtual servers' workloads. For details, see “Modes of operation” on page 111.

On zEC12, PR/SM is enhanced to support an option to limit the amount of physical processor capacity consumed by an individual LPAR when a PU defined as a CP or an Integrated Facility for Linux (IFL) is shared across a set of LPARs. For a definition of these PU types, see 1.2.1, “Models” on page 10.

This enhancement 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 in units of CPs or IFLs.

zEC12 provides 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-book communication, and optimizes z/OS work dispatching, with the result of increasing throughput.

zEC12 supports 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 ensemble internode data network.

1.1.5 Increased flexibility with z/VM mode logical partition

The zEC12 provides for the definition of a z/VM mode logical partition (LPAR) containing a mix of processor types. These types include CPs and specialty processors, such as IFLs, System z Application Assist Processors (zAAPs), System z Integrated Information Processors (zIIPs), and Internal Coupling Facilities (ICFs). For a definition of these types, see 1.2.1, “Models” on page 10.

z/VM V5R4 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 System z 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
- ▶ Exploit zAAPs to provide an economical Java execution environment under z/OS

1.1.6 zAware mode logical partition

The zEC12 introduces the zAware mode logical partition. Either CPs or IFLs can be configured to the partition. This special partition is defined for the exclusive use of the IBM System z Advanced Workload Analysis Reporter (IBM zAware) offering. IBM zAware requires a special license.

1.1.7 IBM System z Advanced Workload Analysis Reporter

IBM System z Advanced Workload Analysis Reporter (IBM zAware) is a feature introduced with the zEC12 that embodies the next generation of system monitoring. IBM zAware is designed to offer a near real-time, continuous learning, diagnostics, 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 grow to large problems. Systems often experience "soft failures" (sick but not dead), which are much more difficult or unusual to detect.

IBM zAware is designed to help in those circumstances. For more information, see Appendix A, "IBM System z Advanced Workload Analysis Reporter (IBM zAware)" on page 479.

1.1.8 Flash Express

Flash Express is an innovative optional feature introduced with the zEC12. It is intended to provide performance improvements and better availability for critical business workloads that cannot afford any hit to service levels. Flash Express is easy to configure, requires no special skills, 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 exploited by these products:

- ▶ z/OS V1R13 (or later), for handling z/OS paging activity and supervisor call (SVC) memory dumps. Although Flash Express is not used for the actual writing of an SVC dump, there are indirect benefits for SVC dumps when paged-out data must be paged-in to memory for inclusion in the dump. Such data can be paged-in faster from Flash Express than from DASD paging data sets. This can result in reduced SVC dump times thus improving system availability.
- ▶ Coupling facility control code (CFCC) Level 19, to use Flash Express as an overflow device for shared queue data. This provides emergency capacity to handle WebSphere MQ shared queue buildups during abnormal situations, such as when "putters" are putting to the shared queue, but "getters" are transiently not getting from the shared queue.
- ▶ Linux (Red Hat Enterprise Linux (RHEL)), for use as temporary storage.

Additional functions of Flash Express are expected to be introduced later, including 2 GB page support and dynamic reconfiguration.

For more information, see Appendix C, “Flash Express” on page 501.

1.1.9 10GbE RoCE Express

The 10 Gigabit Ethernet (10GbE) RoCE Express feature exploits Remote Direct Access Memory over Converged Ethernet and is designed to provide very fast memory-to-memory communications between two CPCs. It is transparent to applications.

Use of the 10GbE RoCE Express feature helps reduce consumption of CPU resources for applications utilizing the TCP/IP stack (such as WebSphere accessing a DB2 database), and might also help to reduce network latency with memory-to-memory transfers utilizing Shared Memory Communications - Remote Direct Memory Access (SMC-R) in z/OS V2R1.

This feature is exclusive to the zEC12 and IBM zEnterprise BC12 (zBC12) and is installed in the PCIe I/O drawer. A maximum of 16 features can be installed. One port per feature is supported by z/OS.

1.1.10 zEDC Express

The growth of data that needs to be captured, transferred, and stored for large periods of time is not relenting. On the contrary, software-implemented compression algorithms are costly in terms of processor resources, and storage costs are not negligible either.

zEDC Express, an optional feature exclusive to zEC12 and zBC12, addresses those requirements by providing hardware-based acceleration for data compression and decompression. zEDC provides data compression with lower CPU consumption than previously existing compression technology on System z.

For more information, see Appendix E, “zEnterprise Data Compression (zEDC) Express” on page 525.

1.1.11 IBM Mobile Systems Remote

IBM Mobile Systems Remote is a mobile application developed by IBM, which is intended to help clients monitor and manage their zEnterprise environment from a personal mobile communication device (such as smartphones or tablets).

By interfacing with the zEnterprise Hardware Management Console (HMC), the application allows authorized personnel to hold on the palm of their hands almost all the information normally viewed on the HMC. Clients will be able to monitor their zEnterprise CP and, in the case of an ensemble, also the IBM BladeCenters and installed blades in the zBX.

For more information about this freely downloadable application and links to the different application stores, check the IBM Mobile Systems Remote website:

<http://www.ibmremote.com/>

1.1.12 Reliability, availability, and serviceability

System reliability, availability, and serviceability (RAS) are areas of continuous IBM focus. The objective is to reduce, or eliminate if possible, all sources of planned and unplanned outages, with the objective of keeping the system running. It is a design objective to provide higher availability with a focus on reducing outages. With a properly configured zEC12, further reduction of outages can be attained through improved nondisruptive replace, repair, and

upgrade functions for memory, books, and I/O adapters. In addition, zEC12 has extended nondisruptive capability to download and install LIC updates.

Enhancements include removing pre-planning requirements with the fixed 32 GB hardware system area (HSA). Client-purchased memory is *not* used for I/O configurations, and it is no longer required to reserve capacity to avoid disruption when adding new features. With a fixed amount of 32 GB for the HSA, maximums are configured and IPLed so that later insertion can be dynamic, which eliminates the need for a power-on reset of the server.

This approach provides many high-availability and nondisruptive operational capabilities that differentiate it 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, competitive price/performance, scalable growth, and continuous availability.

1.2 zEC12 technical overview

This section briefly reviews the major elements of zEC12:

- ▶ Models
- ▶ Model upgrade paths
- ▶ Frames
- ▶ Processor cage
- ▶ I/O connectivity: PCIe and InfiniBand
- ▶ I/O subsystems
- ▶ Coupling and Server Time Protocol connectivity
- ▶ Special purpose features:
 - Cryptography
 - Flash Express
 - zEDC Express
- ▶ Reliability, availability, and serviceability

1.2.1 Models

The zEC12 has a machine type of 2827. Five models are offered: H20, H43, H66, H89, and HA1 (“A1” stands for 101). The model name indicates the maximum number of processor units (PUs) available for purchase. A *PU* is the generic term for the IBM z/Architecture® processor on the MCM.

On the zEC12, 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:

- ▶ System Assist Processor (SAP) used by the channel subsystem. The number of predefined SAPs is dependent on the zEC12 model.
- ▶ One integrated firmware processor (IFP). The IFP is used in support of select features, such as zEDC and 10GbE RoCE.
- ▶ Two spare PUs that can transparently assume any characterization in 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 exploitation of Linux on System z.
- ▶ System z Application Assist Processor (zAAP). One CP must be installed with or before the installation of any zAAPs.
- ▶ System z Integrated Information Processor (zIIP). One CP must be installed with or before the installation of any zIIPs.

zIIP and zAAP: At least one central processor (CP) must be purchased with, or before, a System z Application Assist Processor (zAAP) or System z Integrated Information Processor (zIIP) can be purchased. Clients can purchase up to two zAAPs and up to two zIIPs for each purchased CP (assigned or unassigned) on the system.

However, an LPAR definition can go behind the 1:2 ratio. For example, on a system with two CPs, a maximum of four zAAPs and four zIIPs can be installed. An LPAR definition for that system can contain up to two logical CPs, four logical zAAPs, and four logical zIIPs. Another possible configuration is one logical CP, three logical zAAPs, and four logical zIIPs.

- ▶ Internal Coupling Facility (ICF) to be used by the Coupling Facility Control Code (CFCC).
- ▶ An additional System Assist Processor (SAP) to be used by the channel subsystem.

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

- ▶ In the five-model structure, at least one CP, ICF, or IFL must be purchased and activated for any 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 available for that model.
- ▶ The number of installed zAAPs cannot exceed twice the number of installed CPs.
- ▶ The number of installed zIIPs cannot exceed twice the number of installed CPs.

The multi-book system design provides the capability to concurrently increase the capacity of the system in these ways:

- ▶ Add capacity by concurrently activating more CPs, IFLs, ICFs, zAAPs, or zIIPs on an existing book.
- ▶ Add a book concurrently and activate more CPs, IFLs, ICFs, zAAPs, or zIIPs.
- ▶ Add a book to provide more memory, or one or more adapters to support a greater number of I/O features.

1.2.2 Model upgrade paths

Any zEC12 can be upgraded to another zEC12 hardware model. All upgrades to Models H20, H43, H66, and H89 are concurrent and all upgrades to HA1 are disruptive (that is, the system is unavailable during the upgrade). Any z196 or z10 EC model can be upgraded to any zEC12 model, which is also disruptive. Figure 1-2 on page 12 presents a diagram of the upgrade path.

Restriction: An air-cooled zEC12 cannot be converted to a water-cooled zEC12, and vice versa.

z196 upgrade to zEC12

When a z196 is upgraded to a zEC12, the z196 driver level must be at least 93. If a zBX is involved, the driver 93 must be at Bundle 27 or higher. When upgrading a z196 that controls a zBX Model 002 to a zEC12, the zBX is upgraded to a Model 003. That upgrade is disruptive.

Not offered

The following processes are not supported:

- ▶ Downgrades within the zEC12 models
- ▶ Upgrade from a z114 with zBX to zEC12
- ▶ Removal of a zBX without a controlling system
- ▶ Upgrades from IBM System z9® or earlier systems
- ▶ Attachment of a zBX Model 002 to zEC12

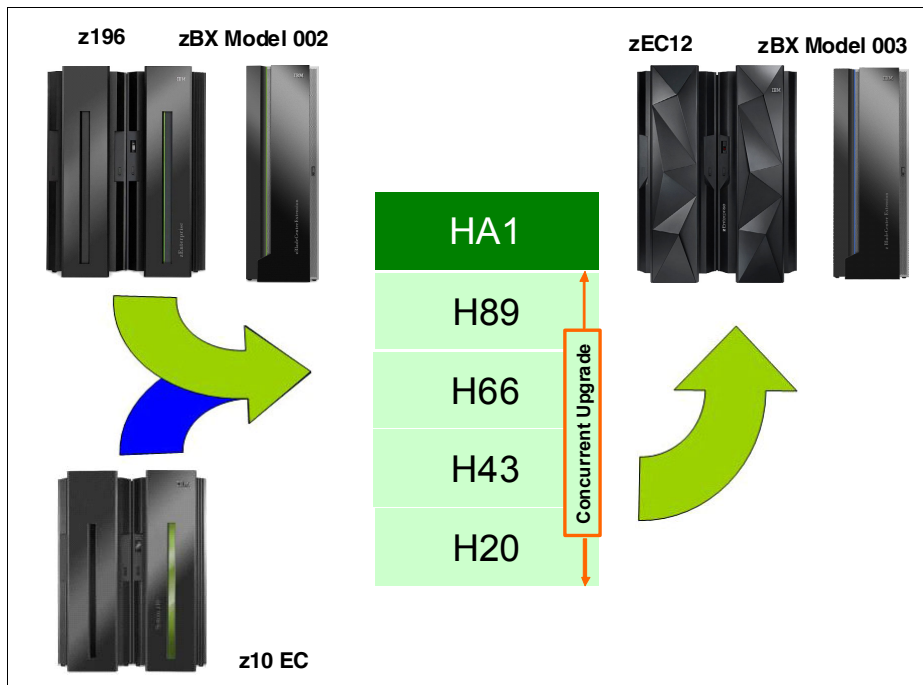


Figure 1-2 zEC12 upgrades

1.2.3 Frames

The zEC12 has two frames, which are bolted together and are known as the A frame and the Z frame. The frames contain the following CPC components:

- ▶ The processor cage, with up to four books
- ▶ PCIe I/O drawers, I/O drawers, and I/O cage, which hold I/O features and special purpose features
- ▶ Power supplies
- ▶ An optional Internal Battery Feature (IBF)
- ▶ Cooling units for either air or water cooling
- ▶ Support elements

1.2.4 Processor cage

The *processor cage* houses up to four processor books. Each book houses an MCM, memory, and I/O interconnects.

MCM technology

The zEC12 is built on the proven superscalar microprocessor architecture of its predecessor, and provides several enhancements over the z196. Each book has one MCM. The MCM has six PU chips and two storage control (SC) chips. The PU chip has six cores, with four, five, or six active cores, which can be characterized as CPs, IFLs, ICFs, zIIPs, zAAPs, or System Assist Processors (SAPs). Two MCM sizes are offered: 27 and 30 cores.

The MCM provides a significant increase in system scalability and an additional opportunity for server consolidation. All books are interconnected with high-speed internal communication links, in a full star topology, through the L4 cache. This configuration allows the system to be operated and controlled by the PR/SM facility as a memory-coherent and cache-coherent SMP.

The PU configuration includes two designated spare PUs per CPC and a variable number of SAPs. The SAPs scale with the number of books that are installed in the server. For example, there are four SAPs with one book installed, and up to 16 when four books are installed. 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, zAAPs, zIIPs, ICF processors, or additional SAPs.

The zEC12 offers a water cooling option for increased system and data center energy efficiency. In a water-cooled system, the MCM is cooled by a cold plate that is connected to the internal water cooling loop. In an air-cooled system, radiator units (RUs) exchange the heat from the internal water loop with air, and air backup. Both cooling options are fully redundant.

Processor features

The processor chip has a six-core design, with either four, five, or six active cores, and operates at 5.5 GHz. Depending on the MCM version (27 PU or 30 PU), from 27 to 120 PUs are available on one to four books.

Each core on the PU chip includes a dedicated coprocessor for data compression and cryptographic functions, such as the Central Processor Assist for Cryptographic Function (CPACF). This configuration is an improvement over z196, where two cores shared a coprocessor.

Hardware data compression can play a significant role in improving performance and saving costs over performing compression in software. The zEDC Express feature offers additional performance and savings over the coprocessor. Their functions are not interchangeable.

Having standard clear key cryptographic coprocessors that are integrated with the processor provides high-speed cryptography for protecting data.

Each core has its own hardware decimal floating point unit that is designed according to a standardized, open algorithm. Much of today's commercial computing is decimal floating point, so on-core hardware decimal floating point meets the requirements of business and user applications, and provides improved performance, precision, and function.

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.

Transactional Execution facility

The z/Architecture is expanded with the Transactional Execution facility. This set of instructions 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

The zEC12 has a superscalar microprocessor with out-of-order (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 instance, OOO allows a few instructions to complete while another instruction is waiting. Up to three instructions can be decoded per system cycle, and up to seven instructions can be in execution.

Concurrent processor unit conversions

The zEC12 supports concurrent conversion between various PU types, providing flexibility to meet changing business environments. CPs, IFLs, zAAPs, zIIPs, ICFs, and optional SAPs can be converted to CPs, IFLs, zAAPs, zIIPs, ICFs, and optional SAPs.

Memory subsystem and topology

zEC12 uses the buffered dual inline memory modules (DIMM) design that was developed for the z196. For this purpose, IBM has developed a chip that controls communication with the PU, and drives address and control from DIMM to DIMM. The DIMM capacities are 4 GB, 16 GB, and 32 GB.

Memory topology provides the following benefits:

- ▶ Redundant array of independent memory (RAIM) for protection at the dynamic random access memory (DRAM), DIMM, and memory channel levels
- ▶ A maximum of 3.0 TB of user configurable memory with a maximum of 3,840 GB of physical memory (with a maximum of 1 TB configurable to a single LPAR)
- ▶ One memory port for each PU chip, and up to three independent memory ports per book
- ▶ Increased bandwidth between memory and I/O
- ▶ Asymmetrical memory size and DRAM technology across books
- ▶ Large memory pages (1 MB and 2 GB)
- ▶ Key storage
- ▶ Storage protection key array that is kept in physical memory
- ▶ Storage protection (memory) key is also kept in every L2 and L3 cache directory entry
- ▶ The large (32 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 I/O cards through the PCIe 8 GBps bus. The PCIe fanout is hot-pluggable. In 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 eight PCIe fanouts are available per book.

Host channel adapter fanout hot-plug

The host channel adapter (HCA) fanout provides the path for data between memory and the I/O cards through InfiniBand (IFB) cables. The HCA fanout is hot-pluggable. In an outage, an HCA fanout can be concurrently repaired without the loss of access to its associated I/O features, using redundant I/O interconnect. Up to eight HCA fanouts are available per book.

1.2.5 I/O connectivity: PCIe and InfiniBand

The zEC12 offers various improved features and exploits technologies, such as PCIe, InfiniBand, and Ethernet. This section briefly reviews the most relevant I/O capabilities.

The zEC12 takes advantage of PCIe Generation 2 to implement the following features:

- ▶ An I/O bus that implements the PCIe infrastructure. This is the preferred infrastructure and can be used with InfiniBand.
- ▶ PCIe fanouts that provide 8 GBps connections to the PCIe I/O features.

The zEC12 takes advantage of InfiniBand to implement the following features:

- ▶ A 6 GBps I/O bus that includes the InfiniBand infrastructure.
This configuration replaces the self-timed interconnect bus in System z systems before z9.
- ▶ Parallel Sysplex coupling links using IFB: 12x InfiniBand coupling links for local connections and 1x InfiniBand coupling links for extended distance connections between any two zEnterprise CPCs and z10 CPCs. The 12x IB link has a bandwidth of 6 GBps.
- ▶ Host Channel Adapters for InfiniBand (HCA3), which can deliver up to 40% faster coupling link service times than HCA2.

1.2.6 I/O subsystems

The zEC12 I/O subsystem is similar to the one on z196 and includes a PCIe infrastructure. The I/O subsystem is supported by both a PCIe bus and an I/O bus similar to that of z196. It includes the InfiniBand Double Data Rate (IB-DDR) infrastructure, which replaces the self-timed interconnect that was in previous System z systems. This infrastructure is designed to reduce processor usage and latency, and provide increased throughput. The I/O expansion network uses the InfiniBand Link Layer (IB-2, Double Data Rate).

zEC12 also offers three I/O infrastructure elements for holding the I/O features: PCIe I/O drawers, for PCIe features; and I/O drawers and I/O cages, for non-PCIe features.

PCIe I/O drawer

The *PCIe I/O drawer*, together with the PCIe I/O features, offers improved granularity and capacity over previous I/O infrastructures. It can be concurrently added and removed in the field, easing planning. Two PCIe I/O drawers occupy the same space as an I/O cage, yet each offers 32 I/O card slots, a 14% increase in capacity. Only PCIe cards (features) are supported, in any combination. Up to five PCIe I/O drawers are supported.

I/O drawer

On the zEC12, I/O drawers are supported only when carried forward on upgrades from z196, z10 EC, or zBC12 model H13 to zEC12 model H20 (air-cooled only). For a new zEC12 installation, it is not possible to have an I/O drawer.

The zEC12 can have up to two I/O drawers. I/O drawers can accommodate up to eight I/O features in any combination. Based on the number of I/O features that are carried forward, the configurator determines the number of required I/O drawers.

I/O cage

On the zEC12, a maximum of one I/O cage is supported and is only available on upgrades from z196 or z10 to zEC12. The I/O cage is housed in the A frame. The I/O cage can accommodate up to 28 I/O features in any combination. Based on the number of I/O features that are carried forward, the configurator determines the required number of I/O drawers and I/O cages.

Native PCIe and integrated firmware processor (IFP)

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 diagnostics for the adapter layer functions of the native PCIe features are managed by LIC that is designated as a resource group, 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

The zEC12 supports the following PCIe features, which can only be installed in the PCIe I/O drawers:

- ▶ FICON Express8S Short Wave (SX) and 10 KM (6.2 miles) Long Wave (LX) (Fibre Channel connection)
- ▶ OSA-Express5S 10 GbE Long Reach (LR) and Short Reach (SR), GbE LX and SX, and 1000BASE-T
- ▶ OSA-Express4S 10 GbE LR and SR, GbE LX and SX, and 1000BASE-T
- ▶ 10GbE RoCE Express
- ▶ Crypto Express4S
- ▶ Flash Express
- ▶ zEDC Express

When carried forward on an upgrade, the zEC12 also supports up to one I/O cage and up to two I/O drawers on which the following features can be installed:

- ▶ FICON Express8 10 KM LX and SX
- ▶ FICON Express4 10 KM LX and SX (four port cards only)
- ▶ OSA-Express3 10 GbE LR and SR
- ▶ OSA-Express3 GbE LX and SX
- ▶ OSA-Express3 1000BASE-T
- ▶ Crypto Express3
- ▶ InterSystem Channel (ISC-3) coupling links (peer-mode only)

In addition, InfiniBand coupling links, which attach directly to the processor books, are supported.

FICON channels

Up to 160 features with up to 320 FICON Express8S channels are supported. The FICON Express8S features support a link data rate of 2 Gbps, 4 Gbps, or 8 Gbps.

Up to 44 features with up to 176 FICON Express8 or FICON Express4 channels are supported:

- ▶ The FICON Express8 features support a link data rate of 2, 4, or 8 Gbps.
- ▶ The FICON Express4 features support a link data rate of 1, 2, or 4 Gbps.

The zEC12 FICON features support the following protocols:

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

FICON also offers the following capabilities:

- ▶ Modified Indirect Data Address Word (MIDAW) facility: 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 Fibre Channel Protocol (FCP).

Open Systems Adapter

The zEC12 allows 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, and up to 24 OSA Express3 features, with a maximum of 96 ports, are supported. OSA-Express5S and OSA-Express4S features are plugged into the PCIe I/O drawer. OSA-Express3 features are plugged into an I/O drawer or I/O cage.

The maximum number of combined OSA-Express5S, OSA-Express4S, and OSA-Express3 features cannot exceed 48.

OSM and OSX channel path identifier (CHPID) types

The zEC12 provides OSA-Express5S, OSA-Express4S, and OSA-Express3 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). Connects the zEC12 to the zBX through the bulk power hubs (BPHs) for use of the Unified Resource Manager functions in the Hardware Management Console (HMC). Exclusively uses OSA-Express5S 1000BASE-T Ethernet, OSA-Express4S 1000BASE-T Ethernet, or OSA-Express3 1000BASE-T Ethernet.
- ▶ OSA-Express for zBX (OSX)
Connectivity to the intraensemble data network (IEDN). Provides a data connection from the zEC12 to the zBX. Uses OSA-Express5S 10 GbE, preferably, and also OSA-Express4S 10 GbE, or OSA-Express3 10 GbE features.

OSA-Express5S, OSA-Express4S, and OSA-Express3 feature highlights

The zEC12 supports five different types of OSA-Express5S, OSA-Express4S, and OSA-Express3 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 Ethernet 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 Ethernet 1000BASE-T Ethernet
- ▶ OSA-Express3 10 GbE Long Reach
- ▶ OSA-Express3 10 GbE Short Reach
- ▶ OSA-Express3 GbE Long Wave
- ▶ OSA-Express3 GbE Short Wave
- ▶ OSA-Express3 1000BASE-T Ethernet

OSA-Express features provide the important benefits for TCP/IP traffic, namely reduced latency and improved throughput for standard and jumbo frames. Performance enhancements are the result of the data router function present in all OSA-Express features. For functions that were previously performed in firmware, the OSA Express5S, OSA-Express4S, and OSA-Express3 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, allowing 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 to increase throughput for standard frames (1492 byte) and jumbo frames (8992 byte).

For more information about the OSA features, see 4.9, “Connectivity” on page 154.

HiperSockets

The HiperSockets function is also known as *Internal Queued Direct Input/Output (internal QDIO or iQDIO)*. It is an integrated function of the zEC12 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 up system and network resources, eliminating attachment costs while improving availability and performance.

For communications between LPARs in the same zEC12 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 10 GbE RoCE Express feature is exclusive to the zEC12 and zBC12 and is for use exclusively 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 use of a multimode fiber optic cable terminated with an LC Duplex connector. Both point-to-point connection and switched connection with an enterprise-class 10 GbE switch are supported.

Support is provided by z/OS, which supports one port per feature.

For more information, see Appendix D, “Remote Direct Memory Access over Converged Ethernet (RoCE)” on page 513.

1.2.7 Coupling and Server Time Protocol connectivity

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

Coupling links support

Coupling connectivity in support of Parallel Sysplex environments is provided on the zEC12 by the following features:

- ▶ 12x InfiniBand coupling links offering up to 6 GBps of bandwidth between zEC12, zBC12, z196, z114, and z10 systems, for a distance of up to 150 m (492 feet). With the introduction of InfiniBand coupling links (HCA3-O 12xIFB), improved service times can be obtained.
- ▶ 1x InfiniBand up to 5 Gbps connection bandwidth between zEC12, zBC12, z196, z114, and z10 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).
- ▶ Internal Coupling Channels (ICs), operating at memory speed.
- ▶ InterSystem Channel-3² (ISC-3) supports an unrepeated link data rate of 2 Gbps over 9 µm single mode fiber optic cabling with an LC Duplex connector.

All coupling link types can be used to carry Server Time Protocol (STP) messages. The zEC12 does not support ICB4 connectivity.

Removal of ISC-3 support on System z: The IBM zEnterprise EC12 and IBM zEnterprise BC12 are planned to be the last high-end System z servers to offer support of the InterSystem Channel-3 (ISC-3) for Parallel Sysplex environments at extended distances. ISC-3 will not be supported on future high-end System z servers to carry forward on an upgrade. Previously, the IBM zEnterprise 196 (z196) and IBM zEnterprise 114 (z114) servers were announced to be the last to offer ordering of ISC-3. Enterprises need to continue upgrading from ISC-3 features to 12x InfiniBand or 1x InfiniBand coupling links.

Coupling Facility Control Code Level 19

CFCC Level 19 is delivered on the zEC12 with driver level 15. CFCC Level 19 introduces the following enhancements:

- ▶ Performance improvements:
 - Coupling Thin Interrupts:
 - Improve the performance in shared CF engine environments.
 - Improve the response time of asynchronous CF requests.
- ▶ Resiliency enhancements:

² Only available on zEC12 when carried forward during an upgrade.

- Flash Express supports and provides cost-effective standby capacity to help manage the potential overflow of WebSphere MQ shared queues.

CF structure sizing changes are expected when upgrading from CFCC Level 17 (or earlier) to CFCC Level 18 or CFCC Level 19 as well as from CFCC Level 18 to CFCC Level 19. Review the CF LPAR size by using the CFSizer tool:

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

Coupling Facility Control Code Level 18

The initial support of the CFCC on the zEC12 with driver level 12 is level 18. CFCC Level 18 offers the following enhancements:

- ▶ Performance enhancements:
 - Dynamic structure size alter improvement
 - DB2 group buffer pool (GBP) cache bypass
 - Cache structure management
- ▶ Coupling channel reporting improvement, enabling IBM Resource Measurement Facility (RMF™) to differentiate between various IFB link types, and detect whether a condition information block (CIB) link is running in a degraded state.
- ▶ Serviceability enhancements:
 - Additional structure control information in CF dumps
 - Enhanced CFCC tracing support
 - Enhanced triggers for CF nondisruptive dumping

Server Time Protocol facility

Server Time Protocol (STP) is a server-wide facility that is implemented in the LIC of System z servers (including servers 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 System z server can be enabled for STP by installing the STP feature. Each server that needs to be configured in a Coordinated Timing Network (CTN) must be STP-enabled.

The STP feature is designed to be the supported method for maintaining time synchronization between System z systems 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*.

Network Time Protocol (NTP) client support is available to the STP code on the zEC12, zBC12, z196, z114, and z10. With this function, the zEC12, zBC12, z196, z114, and z10 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 zEC12, zBC12, IBM zEnterprise 196 (z196), z114, and z10, as well as other servers that have NTP clients, such as UNIX and Microsoft Windows systems. NTP can only be used as ETS for an STP-only CTN where no server can have an active connection to an IBM Sysplex Timer®.

The time accuracy of an STP-only CTN is improved by adding as the ETS device an NTP server with the pulse per second output signal (PPS). 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 Support Element (SE). The HMC is normally attached to the private dedicated LAN for

System z maintenance and support. For zEC12, authentication support is added to the HMC's NTP communication with NTP time servers.

A zEC12 cannot be connected to a Sysplex Timer. Generally, change to an STP-only CTN for existing environments. A zEC12 can be a Stratum 2 or Stratum 3 server in a Mixed CTN if at least one IBM System z10® is attached to the Sysplex Timer operating as Stratum 1 server, but two System z10s acting as Stratum 1 need to be used whenever possible.

Statement of direction: The IBM zEnterprise EC12 will be the last server to support connections to a Mixed CTN (external time reference provided by the Sysplex Timer 9037). After zEC12, if time synchronization is needed (such as to support a base or Parallel Sysplex), STP is required. In addition, all servers participating in the Coordinated Timing Network must be configured in STP-only mode.

1.2.8 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 actually performing I/O operations. That is, no data is moved between the CPC and externally attached devices.

Cryptography

Integrated cryptographic features provide leading cryptographic performance and functionality. Reliability, availability, and serviceability (RAS) support is unmatched in the industry, and the cryptographic solution has received the highest standardized security certification (FIPS 140-2 Level 4³). The crypto cards permit you to add or move crypto coprocessors to LPARs without pre-planning.

The zEC12 implements the PKCS#11, one of the industry-accepted standards called public key cryptographic standards (PKCS) provided by RSA Laboratories from 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 System z also invoke 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 shipped enabled with each server.

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

Configurable Crypto Express4S feature

The Crypto Express4S represents the newest generation of cryptographic feature that is designed to complement the cryptographic capabilities of the CPACF. It is an optional feature of zEC12 server generation. The Crypto Express4S feature is designed to provide port

³ Federal Information Processing Standards (FIPS) 140-2 Security Requirements for Cryptographic Modules

granularity for increased flexibility with one PCIe adapter per feature. For availability reasons, a minimum of two features are required.

The Crypto Express4S 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:

- ▶ Secure IBM CCA coprocessor is for secure key encrypted transactions that use CCA callable services (default).
- ▶ 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.

- ▶ Accelerator for public key and private key cryptographic operations is used with Secure Sockets Layer/Transport Layer Security (SSL/TLS) acceleration.

Federal Information Processing Standards (FIPS) 140-2 certification is supported only when Crypto Express4S is configured as a CCA or an EP11 coprocessor.

Configurable Crypto Express3 feature

The Crypto Express3 is an optional feature available only on a carry-forward basis in zEC12. Each feature has two PCIe adapters. Each adapter can be configured as a secure coprocessor or as an accelerator:

- ▶ Crypto Express3 Coprocessor is for secure key encrypted transactions (default).
- ▶ Crypto Express3 Accelerator is for Secure Sockets Layer/Transport Layer Security (SSL/TLS) acceleration.

TKE workstation and support for smart card readers

The Trusted Key Entry (TKE) workstation and the TKE 7.3 LIC are optional features on the zEC12. 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, backup, and a secure hardware-based key loading for operational and master keys. TKE also provides a secure management of host cryptographic module and host capabilities.

Support for an optional smart card reader 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).

When Crypto Express4S is configured as a Secure IBM Enterprise PKCS #11 (EP11) coprocessor, the TKE workstation is required to manage the Crypto Express4S feature. If the smart card reader feature is installed in the TKE workstation, the new smart card part 74Y0551 is required for EP11 mode.

For more information about the Cryptographic features, see Chapter 6, “Cryptography” on page 199.

Flash Express

The *Flash Express* optional feature is intended to provide performance improvements and better availability for critical business workloads that cannot afford any hits to service levels.

Flash Express is easy to configure, requires no special skills, and provides rapid time to value.

Flash Express implements storage-class memory (SCM) in a PCIe card form factor. Each Flash Express card implements an internal NAND Flash solid-state drive (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 (eight features) can be installed on a zEC12, for a maximum capacity of 5.6 TB of storage.

The Flash Express feature is designed to allow each LPAR to be configured with its own SCM address space. It is used for paging. Flash Express can be used, for instance, to hold pageable 1 MB pages.

Encryption is included to improve data security. Data security is ensured through a unique key that is stored on the Support Element (SE) hard disk drive. It is mirrored for redundancy. Data on the Flash Express feature is protected with this key, and is only usable on the system with the key that encrypted it. The Secure Key Store is implemented by using a smart card that is installed in the Support Element. The smart card (one pair, so you have 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 (at minimum) for handling z/OS paging activity, support for 1 MB pageable pages, and SVC memory dumps. Support was added to the coupling facility control code (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” are transiently not getting from the shared queue.

Flash memory is assigned to a CF image via HMC panels. Coupling facility resource management (CFRM) policy definition permits the desired amount of storage class memory (SCM) to be used by a particular structure, on a structure-by-structure basis. Additionally, Linux (RHEL) can now use Flash Express for temporary storage. Additional functions of Flash Express are expected to be introduced later, including 2-GB page support and dynamic reconfiguration for Flash Express.

For more information, see Appendix C, “Flash Express” on page 501.

zEDC Express

zEDC Express, an optional feature exclusive to zEC12 and zBC12, provides hardware-based acceleration for data compression and decompression with lower CPU consumption than the previous compression technology on System z.

Exploitation of the zEDC Express feature by the z/OS V2R1 zEnterprise Data Compression acceleration capability is designed to deliver an integrated solution to help reduce CPU consumption, optimize performance of compression-related tasks, and enable more efficient use of storage resources, while providing a lower cost of computing and also helping to optimize the cross-platform exchange of data.

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 LPARs.

For more information, see Appendix E, “zEnterprise Data Compression (zEDC) Express” on page 525.

1.2.9 Reliability, availability, and serviceability

The zEC12 reliability, availability, and serviceability (RAS) strategy is a building-block approach developed to meet the client’s stringent requirements for continuous operation requirements. Those building blocks are error prevention, error detection, recovery, problem determination, service structure, change management, and measurement and analysis.

The initial focus is on preventing failures from occurring in the first place. This is accomplished by using *Hi-Rel* (highest reliability) components; using screening, sorting, burn-in, and run-in; and by taking advantage of technology integration. For LIC and hardware design, failures are eliminated 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 that is necessary 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 total transparency cannot be achieved, you can restart the server with the maximum capacity possible.

The zEC12 has the following RAS improvements, among others:

- ▶ Improved error detection for the L3/L4 memory cache
- ▶ IBM System z Advanced Workload Analysis Reporter to detect abnormal behavior of z/OS
- ▶ OSA firmware changes to increase concurrent maintenance change level (MCL) capability
- ▶ Digital Temperature Sensor (DTS) and On Chip Temperature Sensor on the PU chips

The following examples show the reduced effect of planned and unplanned system outages:

- ▶ Enhanced book availability
- ▶ Hot pluggable PCIe I/O drawers and I/O drawers
- ▶ Redundant I/O interconnect
- ▶ Concurrent PCIe fanout and Host Channel Adapter (HCA-O, HCA-C) fanout card hot-plug
- ▶ Enhanced driver maintenance

For more information, see Chapter 10, “Reliability, availability, and serviceability” on page 385.

1.3 Hardware Management Consoles and Support Elements

The Hardware Management Consoles (HMCs) and Support Elements (SEs) are appliances that together provide hardware platform management for a System z server and for the ensemble nodes, when the zEnterprise CPC is a member of an ensemble.

In an ensemble, the HMC is used to manage, monitor, and operate one or more zEnterprise CPCs and their associated LPARs, and zBXs. Also, when the zEnterprise is a member of an ensemble, the HMC⁴ has a global (ensemble) management scope, compared to the SE, which has local (node) management responsibility. When tasks are performed on the HMC, the commands are sent to one or more SEs, which then issue commands to their zEnterprise

⁴ From Version 2.11. For more information, see 12.7, “HMC in an ensemble” on page 461.

CPCs and zBXs. To promote high availability, an ensemble configuration requires a pair of HMCs in primary and alternate roles.

1.4 IBM zEnterprise BladeCenter Extension (zBX) Model 003

The IBM zEnterprise BladeCenter Extension (zBX) Model 003 is improving infrastructure reliability by extending the mainframe systems management and service across a set of heterogeneous compute elements in an ensemble.

The zBX Model 003 is available as an optional system to work with the zEC12 server and consists of the following components:

- ▶ 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.
- ▶ Up to 112⁵ blades.
- ▶ Intranode management network (INMN) Top of Rack (ToR) switches. The INMN provides connectivity between the zEC12 Support Elements and the zBX, for management purposes.
- ▶ Intraensemble data network (IEDN) ToR switches. The IEDN is used for data paths between the zEC12 and the zBX, and the other ensemble members, and also for customer data access. The IEDN point-to-point connections use MAC addresses, not IP addresses (Layer 2 connection).
- ▶ 8-Gbps Fibre Channel switch modules for connectivity to a 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 System z, such as the capability for concurrent upgrades and repairs.

GDPS/PPRC and GDPS/GM (Geographically Dispersed Parallel Sysplex/ Global Mirror) support zBX hardware components, providing workload failover for automated multi-site recovery. These capabilities will help facilitate the management of planned and unplanned outages across IBM zEnterprise EC12.

1.4.1 Blades

There are two types of blades that can be installed and operated in the IBM zEnterprise BladeCenter Extension (zBX):

- ▶ Optimizer Blades:
 - IBM WebSphere DataPower Integration Appliance XI50 for zEnterprise blades
- ▶ IBM Blades:
 - A selected subset of IBM POWER7 blades
 - A selected subset of IBM BladeCenter HX5 blades

⁵ The maximum number of blades varies according to the blade type and blade function.

These blades have been thoroughly tested to ensure compatibility and manageability in the IBM zEnterprise System environment:

- ▶ IBM POWER7 blades are virtualized by PowerVM® Enterprise Edition, and the virtual servers run the IBM AIX® operating system.
- ▶ IBM BladeCenter HX5 blades are virtualized by using an integrated hypervisor for System x and the virtual servers run Linux on System x (Red Hat Enterprise Linux (RHEL) and SUSE Linux Enterprise Server (SLES) operating systems) and select Microsoft Windows Server operating systems.

zEnterprise enablement for the blades is specified with an entitlement feature code to be configured on zEnterprise CPCs.

1.4.2 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) function, including routing, bridging, transformation, and event handling. It can help to simplify, govern, and enhance the network security for XML and web services.

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 System z service support representative (SSR), as needed.

1.5 Unified Resource Manager

The zEnterprise Unified Resource Manager is the integrated management fabric that executes on the Hardware Management Console (HMC) and Support Element (SE). The Unified Resource Manager consists of six management areas, as shown in Figure 1-1 on page 2.

- ▶ 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 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.

1.6 Operating systems and software

The zEC12 is supported by a large set of software, including independent software vendor (ISV) applications. This section lists only the supported operating systems. Exploitation of various features might require the latest releases. For more information, see Chapter 8, “Software support” on page 263.

1.6.1 Supported operating systems

Using some features might require the latest releases. The following operating systems are supported for zEC12:

- ▶ z/OS Version 2 Release 1
- ▶ z/OS Version 1 Release 13 with program temporary fixes (PTFs)
- ▶ z/OS Version 1 Release 12 with PTFs
- ▶ z/OS Version 1 Release 11 with the IBM Lifecycle Extension with PTFs⁶
- ▶ z/VM Version 6 Release 3 with PTFs
- ▶ z/VM Version 6 Release 2 with PTFs
- ▶ z/VM Version 5 Release 4 with PTFs
- ▶ z/VSE Version 4 Release 3 or later with PTFs
- ▶ z/TPF Version 1 Release 1
- ▶ Linux on System z distributions:
 - SUSE Linux: SLES⁷ 10 and SLES 11
 - Red Hat: RHEL⁸ 5 and RHEL 6

Statement of Direction: Systems zEC12 and zBC12 are the last servers planned to be supported by z/VM 5.4.

The following operating systems support IBM blades on the zBX Model 003:

- ▶ For the POWER7 blades, AIX Version 5 Release 3 or later, with PowerVM Enterprise Edition
- ▶ For the System x blades:
 - Linux on System x (64-bit only):
 - Red Hat RHEL 5.5, 5.6, 5.7, 6.0, and 6.1
 - SUSE SLES 10 (Service Pack (SP4)), SLES 11 SP1
 - Microsoft Windows Server 2012, Windows Server 2008 R2, and Windows Server 2008 SP2 (Datacenter Edition is recommended), 64-bit only

⁶ The z/OS V1 Release 11 requires IBM Lifecycle Extension.

⁷ SUSE Linux Enterprise Server

⁸ Red Hat Enterprise Linux

Together with support for IBM WebSphere software, full support for SOA, web services, Java Platform, Enterprise Edition, Linux, and Open Standards, the zEnterprise BC12 is intended to be a platform of choice for the integration of the newest generations of applications with existing applications and data.

1.6.2 IBM compilers

The following IBM compilers for z/OS can exploit zEC12:

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

The compilers increase the return on your investment in zEC12 hardware by maximizing application performance on System z by leveraging the compilers' advanced optimization technology to exploit the z/Architecture. Through their support of web services, XML, and Java, they allow for the modernization of existing assets in web-based applications. They support the latest IBM middleware products (CICS, DB2, and IMS), allowing applications to leverage their latest capabilities.

In order to fully exploit the capabilities of the zEC12, you must compile using the minimum level of each compiler that is specified in Table 1-1.

Table 1-1 Supported compiler levels

Compiler	Level
C/C++	z/OS 1.13 XL C/C++ with PTFs: UK80670, UK80671, UK80039, UK79899 or z/OS 2.1 XL C/C++
COBOL	Enterprise COBOL for z/OS 5.1
PL/I	Enterprise PL/I for z/OS 4.4

In order to obtain the best performance, you must specify the ARCH(10) option, which grants the compiler permission to use machine instructions that are only available in the zEC12. Because the ARCH(10) option results in the generated application using instructions that are only available in the zEC12, the application will not run on earlier versions of hardware. If the application needs to run on the zEC12 as well as on older hardware, specify the ARCH option corresponding to the oldest hardware on which the application needs to run. For more information, see the documentation for the ARCH option in the guide for the corresponding compiler product.



Central processor complex hardware components

This chapter introduces IBM zEnterprise EC12 (zEC12) hardware components. It also describes the significant features and functions with their characteristics and options. The objective is to explain the zEC12 hardware building blocks and how these components interconnect from a physical point of view. This information is useful for planning purposes and can help in defining configurations that fit your requirements.

This chapter includes the following sections:

- ▶ Frames and cage
- ▶ Book concept
- ▶ Multiple chip module (MCM)
- ▶ Processor unit (PU) and storage control (SC) chips
- ▶ Memory
- ▶ Reliability, availability, and serviceability (RAS)
- ▶ Connectivity
- ▶ Model configurations
- ▶ Power and cooling
- ▶ Summary of zEC12 structure

2.1 Frames and cage

System z frames are enclosures that are built to Electronic Industries Association (EIA) standards. The zEC12 has two 42U EIA frames, which are shown in Figure 2-1. The two frames, A and Z, are bolted together and have positions for one processor cage and a combination of Peripheral Component Interconnect Express (PCIe) I/O drawers, I/O drawers, and one I/O cage.

All books, including the distributed converter assemblies (DCAs) on the books and the cooling components, are in the processor cage in the A frame. Figure 2-1 shows the front view of the A frame (with four books installed) and the Z frame of an air cooled zEC12.

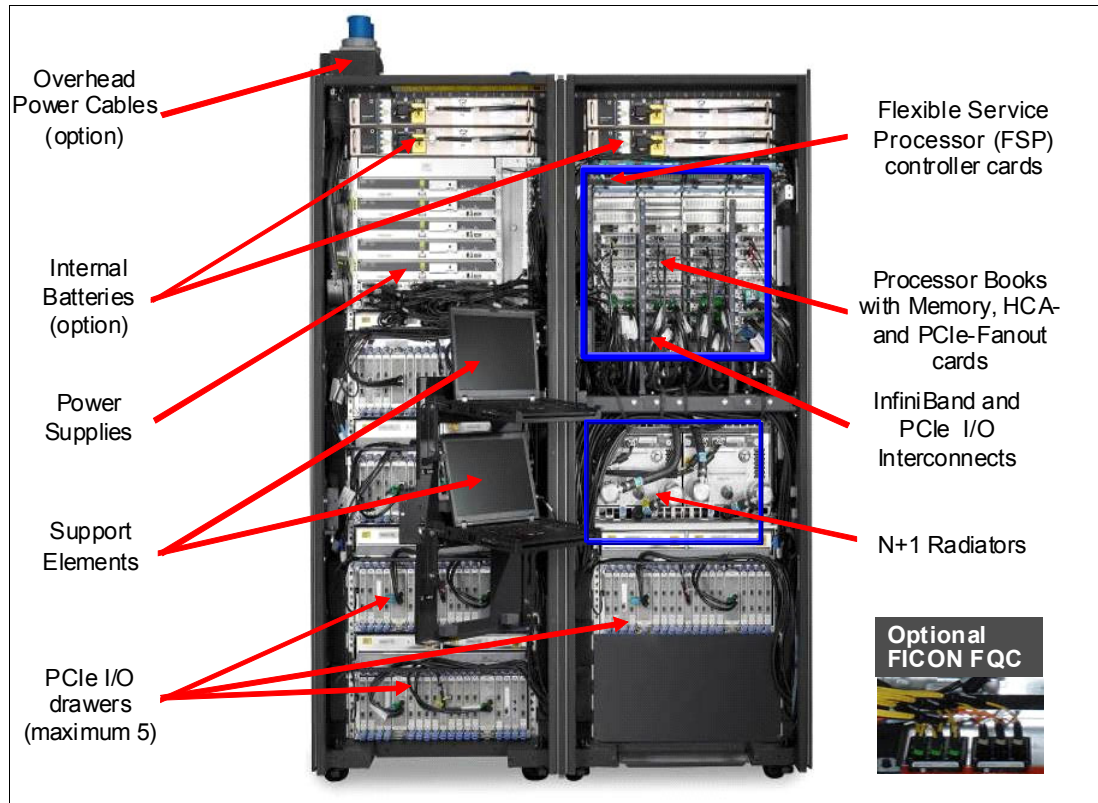


Figure 2-1 CPC cage, I/O drawers, and I/O cage locations for air-cooled system, front view

Figure 2-2 shows the front view of a water-cooled zEC12.

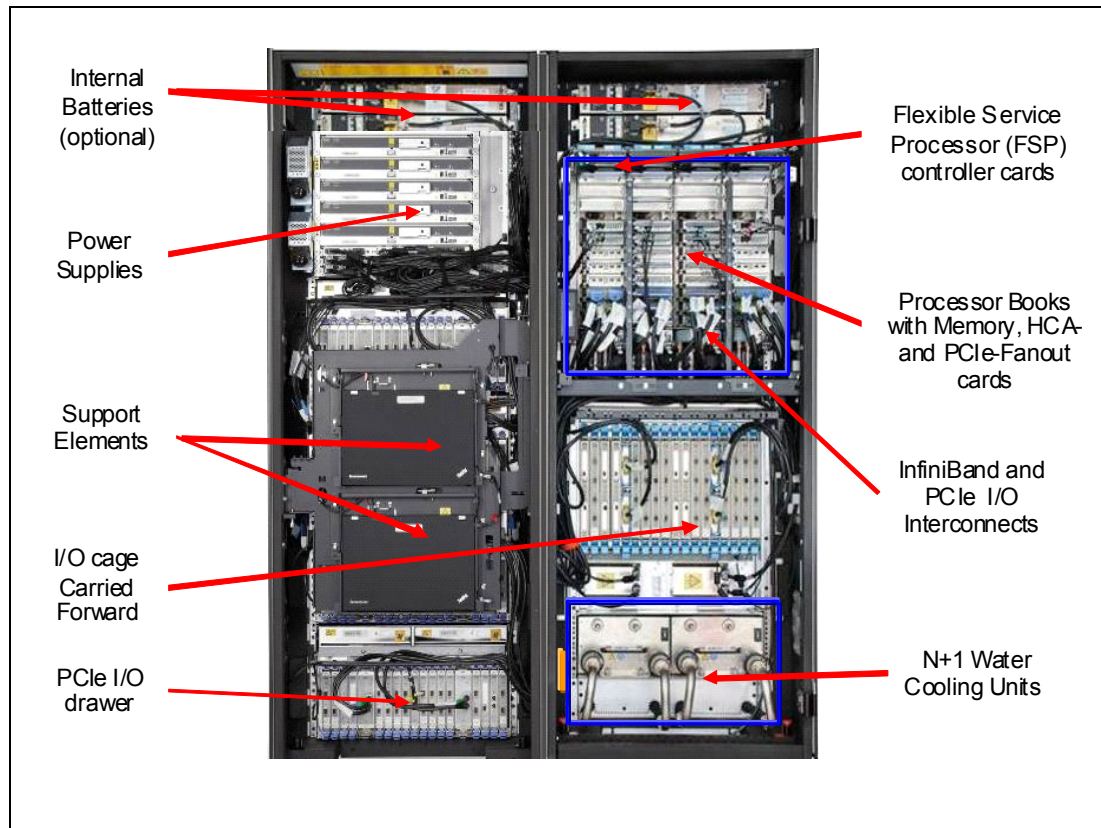


Figure 2-2 CPC cage, I/O drawers, and I/O cage locations for water-cooled system, front view

2.1.1 A frame

As shown in Figure 2-1 on page 30 and Figure 2-2, the A frame has the following major components (from top to bottom of the frame):

- ▶ Two optional **Internal Battery Features (IBFs)**, which 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 all connected AC or DC feeds from the utility provider. The IBF provides battery power to preserve full system function despite the loss of power at all system power cords. It allows continuous operation through intermittent losses, brownouts, and power source switching. It can also provide time for an orderly shutdown during a longer outage. The IBF provides up to 10 minutes of full power, depending on the I/O configuration. The batteries are installed in pairs. Two to six battery units can be installed. The number is based on the zEC12 model and configuration.
- ▶ Two, fully redundant **radiator units** that contain water pumps to feed the internal closed water loops for the MCM cooling, heat exchanger, manifold assembly, and blowers. This configuration provides cooling for the MCMs.
- ▶ Instead of the radiator units, the customer can specify two **Water Conditioning Units (WCUs)** that are connected to a chilled water supply. The WCUs maintain the same internal water loops for MCM cooling as the radiator units. When the WCUs option is used for cooling the books, an additional exhaust air heat exchanger is installed in the rear of the frame.

- ▶ **Processor cage**, which contains up to four books, which are connected to the internal water cooling system.
- ▶ Depending on the configuration, the following **I/O assemblies** can be used. Any combination of up to two drawers of any kind or only the I/O cage is possible:
 - Up to two **PCIe I/O drawers** for installation of PCIe I/O features. The PCIe I/O drawer is used for new installations or can be carried forward via an MES from zEnterprise 196 (System z196) or IBM zEnterprise BC12 (zBC12) Model H13 to zEC12 Model H20 (air-cooled only). It is equipped with a maximum of 32 features.
 - One **I/O drawer** that contains up to eight existing I/O features. The I/O drawer can be carried forward only with an MES from z10, z196, or zBC12 Model H13 to zEC12 Model H20 (air-cooled only). For a new zEC12 installation, it is not possible to have an I/O drawer.
 - One **I/O cage**, which can house 28 existing I/O card slots. One I/O cage is supported. The I/O cage can be carried forward only with an MES from z10 or z196. For a new zEC12 installation, it is not possible to have an I/O cage.
- ▶ **Air moving devices** (AMDs), which provide N+1 redundant cooling for the fanouts, memory, and DCAs.

2.1.2 Z frame

As shown in Figure 2-1 on page 30, the Z frame has these major components (from top to bottom of the frame):

- ▶ Two optional **IBFs**.
- ▶ **Bulk Power Assembly (BPA)**. The number of BPAs varies depending on the configuration of the zEC12. For more information about the required number of BPAs, see 2.9.1, “Power consumption” on page 67.
- ▶ The **Support Element (SE) tray**, which is in front of the I/O drawer slots, contains the two SEs.
- ▶ Up to **four drawers**, which can be any combination of up to two **I/O drawers** and up to four **PCIe I/O drawers**:
 - The PCIe I/O drawer is used for all new installations or can be carried forward via MES from z196 or z12BC Model H13 to zEC12 Model H20 (air-cooled only).
 - The I/O drawer itself can be carried forward only with an MES¹ from z10, z196, or zBC12 Model H13 to zEC12 Model H20 (air-cooled only).
 - The I/O cage is not supported in the Z frame.
- ▶ When the WCUs option is used for cooling the books, an additional exhaust air heat exchanger is installed in the rear of the frame.
- ▶ An optional overhead power cable feature is shown in Figure 2-1 on page 30. When this feature is ordered, it is present on the Z frame.

2.1.3 I/O cage, I/O drawer, and PCIe I/O drawer features

Each book has up to eight dual-port fanouts to support two types of I/O infrastructures for data transfer:

- ▶ PCIe I/O infrastructure with bandwidth of 8 GBps
- ▶ InfiniBand I/O infrastructure with bandwidth of 6 GBps

¹ Miscellaneous equipment specification

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

- ▶ FICON Express8S (two port card, long wavelength (LX) or short wavelength (SX), and two channel-path identifiers (CHPIDs)).
- ▶ Open System Adapter (OSA)-Express5S features (available with driver 15):
 - OSA-Express5S 10 Gb Ethernet (one port card, Long Reach (LR) or Short Reach (SR), and one CHPID)
 - OSA-Express5S Gb Ethernet (two port card, LX or SX, and one CHPID)
 - OSA-Express5S 1000BASE-T Ethernet (two port card, RJ-45, and one CHPID)
- ▶ OSA-Express4S features (only for a carry-forward MES):
 - OSA-Express4S 10 Gb Ethernet (one port card, LR or SR, and one CHPID)
 - OSA-Express4S Gb Ethernet (two port card, LX or SX, and one CHPID)
 - OSA-Express4S 1000BASE-T Ethernet (two port card, RJ-45, and one CHPID)
- ▶ Crypto Express4S. Each Crypto Express4S 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.
- ▶ Flash Express. Each Flash Express feature occupies two I/O slots, 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. Two port card, but the ports cannot be used simultaneously. It depends on the configuration. A port is activated via function ID (FID) to a dedicated LPAR. A CHPID is not required.
- ▶ zEnterprise Data Compression (zEDC) Express. The zEnterprise Data Compression Express feature occupies one I/O slot, but it does not have a CHPID type. Up to 15 partitions can share the feature concurrently.

InfiniBand I/O infrastructure uses the HCA2-C fanout to connect to I/O drawers or I/O cages. The drawers and cages can contain various Fibre Channel connection (FICON), coupling link, OSA-Express, and Cryptographic features:

- ▶ FICON features (FICON or Fibre Channel Protocol (FCP) modes):
 - FICON Express4 (four port card, LX or SX, and four CHPIDs)
 - FICON Express8 (four port card, LX or SX, and four CHPIDs)
- ▶ ISC-3 links (up to four coupling links, two links per daughter card). Two daughter cards (ISC-D) plug into one mother card (ISC-M).
- ▶ OSA-Express features:
 - OSA-Express3 10 Gb Ethernet (two port card, LR or SR, and two CHPIDs)
 - OSA-Express3 Gb Ethernet (four port card, LX or SX, and two CHPIDs)
 - OSA-Express3 1000BASE-T Ethernet (four port card, RJ-45, and two CHPIDs)
- ▶ Crypto Express3 is an optional feature, and it is available only on a carry-forward basis when you are upgrading from earlier generations to zEC12. The Crypto Express3 feature has two cryptographic coprocessors per feature. Each feature can be configured as a cryptographic coprocessor for secure key operations or as an accelerator for clear key operations.

2.1.4 Top exit I/O cabling

On zEC12, you can order the infrastructure to support top exit for fiber optic cables (FICON, OSA, 12x InfiniBand, 1x InfiniBand, ISC-3, and RoCE), and copper cables for the 1000BASE-T Ethernet features.

Top exit I/O cabling is designed to provide you with an additional option. Instead of all your cables exiting under the CPC or under the raised floor, you can select the option that best meets the requirements of your data center.

Top exit I/O cabling can also help to increase the air flow. This option is offered on new build and MES orders.

2.2 Book concept

The central processor complex (CPC) uses a packaging design for its processors that is based on books. A *book* contains a multiple chip module (MCM), memory, and connectors to I/O drawers, or an I/O cage and to other CPCs. Books are in the processor cage in the A frame. The zEC12 has one to four books installed. A book and its components are shown in Figure 2-3.

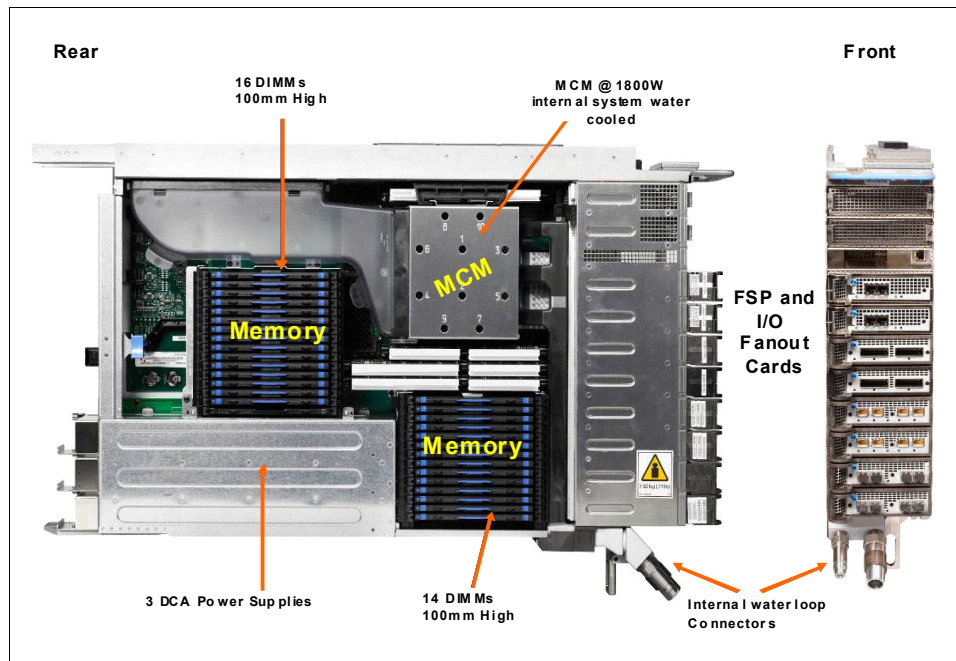


Figure 2-3 Book structure and components

Each book contains the following components:

- ▶ One MCM with six hex core microprocessor chips, having either 27 or 30 processor units (PUs), depending on the model, and two storage control chips with 384 MB of Level 4 cache.
- ▶ Memory dual inline memory modules (DIMMs) plugged into 30 available slots, providing 60 GB - 960 GB of physical memory.

- ▶ A combination of up to eight (host channel adapter (HCA) or PCIe) fanout cards. HCA2-Copper connections are for 6 GBps links to the I/O cage or I/O drawers in the CPC. PCIe fanouts are used for 8 GBps links to the PCIe I/O drawers, and the HCA-Optical fanouts connect to external CPCs (coupling links).
- ▶ Three DCAs that provide power to the book. Loss of a DCA leaves enough book power to satisfy the book's power requirements ($n+1$ redundancy). The DCAs can be concurrently maintained.
- ▶ Two flexible service processor (FSP) cards for system control.

Figure 2-4 shows the book logical structure, showing its component connections, including the PUs on MCM.

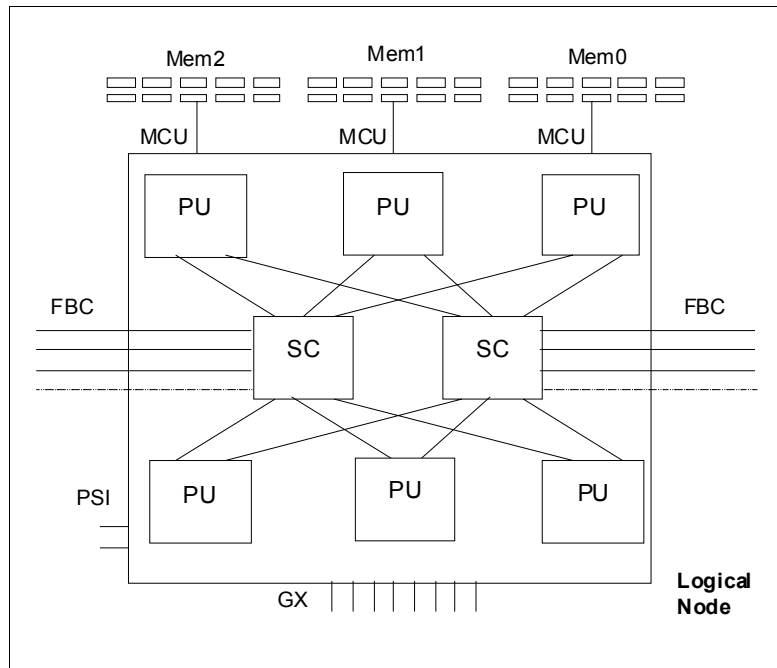


Figure 2-4 Book logical structure

Memory is connected to MCM through three memory control units (MCUs). GX0 to GX7 are the I/O bus interfaces to HCAs. They have full store buffering, a maximum of 10 GBps per bus direction, and support for InfiniBand and PCIe.

Processor support interfaces (PSIs) are used to communicate with FSP cards for system control.

Fabric book connectivity (FBC) provides the point-to-point connectivity between books.

2.2.1 Book interconnect topology

Figure 2-5 on page 36 shows the point-to-point topology for book communication. Each book communicates directly to all other books in the CPC.

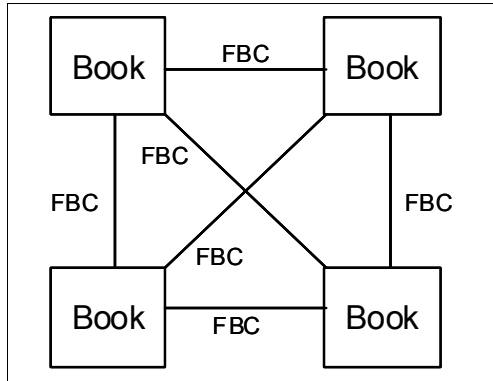


Figure 2-5 Book-to-book communication

Up to four books can be in the processor cage. Books slide into a mid-plane card that supports up to four books and is in the top of the A frame. The mid-plane card is also the location of two oscillator cards.

Position the books in the following manner:

- ▶ In a one-book model, the first book slides in the second slot from the left (processor cage slot location LG06).
- ▶ In a two-book model, the second book slides in the rightmost slot (processor cage slot location LG15).
- ▶ In a three-book model, the third book slides in the third slot from the left (processor cage slot location LG10).
- ▶ In a four-book model, the fourth book slides into the leftmost slot (processor cage slot location LG01).

Table 2-1 indicates the order of book installation and position in the processor cage.

Table 2-1 Book installation order and position in the processor cage

Book	Book0	Book1	Book2	Book3
Installation order	Fourth	First	Third	Second
Position in cage (LG)	01	06	10	15

Book installation is concurrent, except for the upgrade to the model HA1. Concurrent book repair requires a minimum of two books.

Consideration: The processor cage slot locations are important in the physical channel ID (PCHID) report, resulting from the IBM configurator tool. Locations 01, 06, 10, and 15 are used to indicate whether book features, such as fanouts and assigned adapter IDs (AIDs), relate to the first, second, third, or fourth book in the processor cage.

2.2.2 Dual external clock facility (ECF)

Two external clock facility (ECF) cards are already installed and shipped with the CPC. They provide a dual-path interface for the pulse per second (PPS) signal. This redundancy allows continued operation if an ECF card fails.

This redundant design also allows concurrent maintenance. The two connectors that connect to the PPS output of a Network Time Protocol (NTP) server are located above the books. They are connected on the mid-plane to which the books are connected.

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 with the time provided by a Network Time Protocol (NTP) server. This configuration allows a heterogeneous platform environment to synchronize to the same time source.

The accuracy of an STP-only CTN is improved by adding an NTP server with the PPS output signal as the external time source (ETS) device. 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 ECFcard to the PPS output of the NTP server is required when the zEC12 is using STP and is configured in an STP-only CTN using NTP with pulse per second (PPS) as the external time source.

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 the IBM zEnterprise EC12.

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

Figure 2-6 shows the location of the two ECF cards on the CPC, which is above the Book 0 and Book 3.

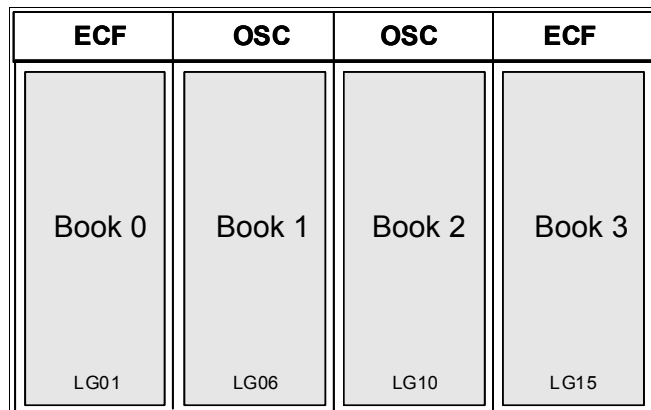


Figure 2-6 ECF and OSC cards

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*, SG24-7280-03
- ▶ *Server Time Protocol Implementation Guide*, SG24-7281-03
- ▶ *Server Time Protocol Recovery Guide*, SG24-7380-01

2.2.3 Oscillator

The zEC12 has two oscillator cards (OSC): a primary and a backup. Although not part of the book design, they are above the books, and are connected to the same mid-plane to which the books are connected. If the primary fails, the secondary detects the failure, takes over transparently, and continues to provide the clock signal to the CPC.

Figure 2-6 on page 37 shows the location of the two OSC cards on the CPC, which is above the Book 1 and Book 2 locations.

2.2.4 System control

Various system elements use FSPs². An FSP is based on the IBM Power PC microprocessor technology. It connects to an internal Ethernet LAN to communicate with the SEs and provides a subsystem interface (SSI) for controlling components. Figure 2-7 is a conceptual overview of the system control design.

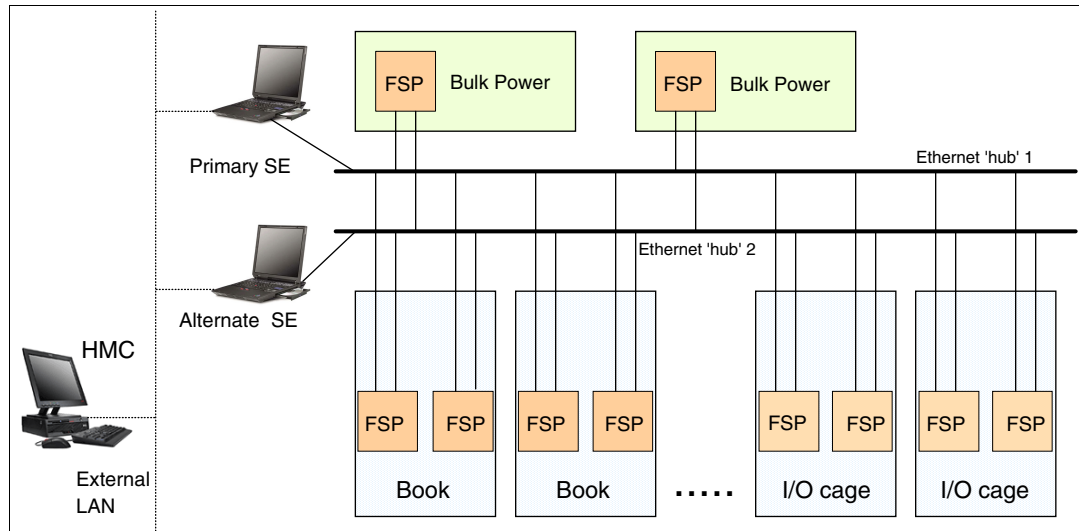


Figure 2-7 Conceptual overview of system control elements

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 (using 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 an FSP. Two internal Ethernet LANs and two SEs, for redundancy, and crossover capability between the LANs are available so that both SEs can operate on both LANs.

The SEs, in turn, are connected to one or two (external) LANs (Ethernet only), and the Hardware Management Consoles (HMCs) are connected to these external LANs. One or more HMCs can be used, but two (a primary and an alternate) are mandatory with an ensemble. Additional HMCs can operate a zEC12 when it is not a member of an ensemble. For more information, see 12.7, “HMC in an ensemble” on page 461.

2.2.5 Book power

Each book gets its power from three DCAs in the book. The DCAs provide the required power for the book in an $n+1$ design. Loss of one DCA leaves enough book power to satisfy its power requirements. The DCAs can be concurrently maintained, and are accessed from the rear of the frame.

² Flexible Service Processor

2.3 Multiple chip module (MCM)

The MCM is a 103-layer glass ceramic substrate (size is 96 x 96 mm) containing eight chip sites and 7,356 land grid array (LGA) connections. There are six PU chips and two storage control (SC) chips. Figure 2-8 illustrates the chip locations. The total number of transistors on all chips on the MCM is more than 23 billion.

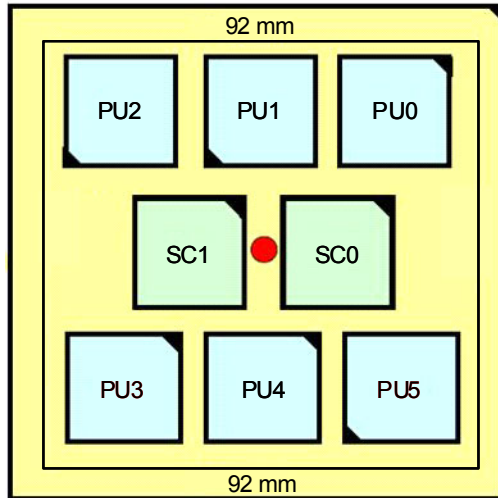


Figure 2-8 zEC12 multi-chip module

The MCM plugs into a card that is part of the book packaging. The book itself is plugged into the mid-plane system board to provide interconnectivity between the books. This configuration allows a multibook system to be displayed as a symmetric multiprocessor (SMP) system.

2.4 Processor unit (PU) and storage control (SC) chips

Both PU and SC chips on the MCM use CMOS 13S chip technology. CMOS 13S is state-of-the-art microprocessor technology that is based on 15-layer copper interconnections and Silicon-On-Insulator (SOI) technologies. The chip lithography line width is 0.032 μm (32 nm). On the MCM, four serial electrically erasable programmable ROM (SEEPROM) chips are rewritable memory chips with these characteristics:

- ▶ Hold data without power
- ▶ Use the same technology
- ▶ Used for retaining product data for the MCM and engineering information

Two of them are active, and the other two are used for redundancy.

Figure 2-9 is the MCM structure diagram, showing the PUs and SCs, and their connections.

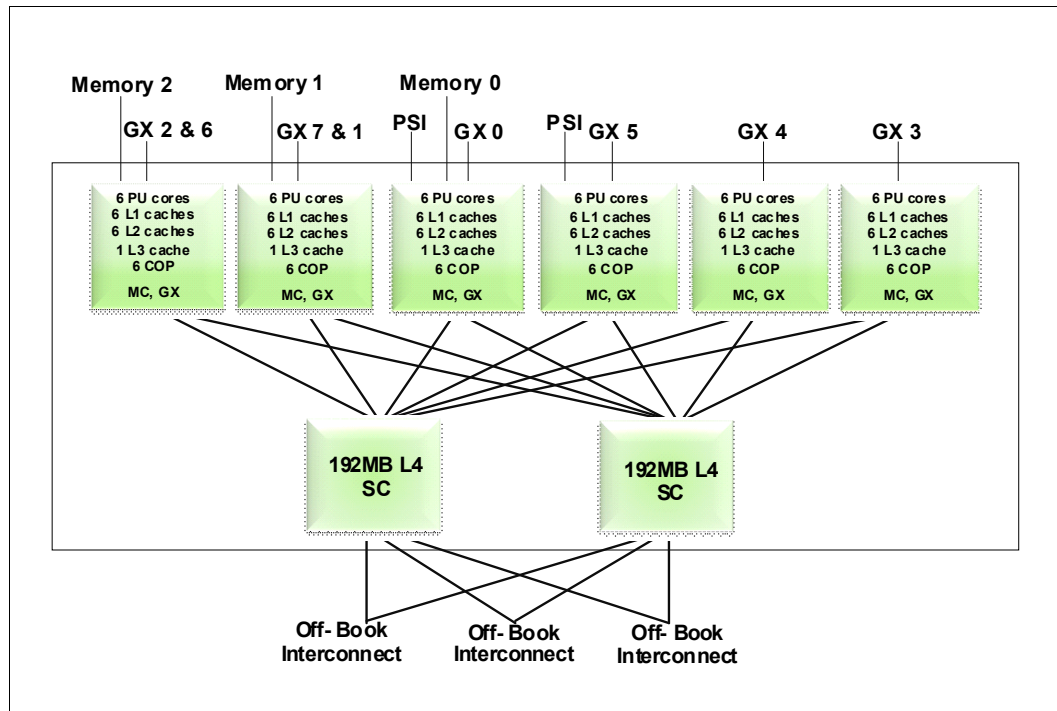


Figure 2-9 PU MCM structure

2.4.1 PU chip

The zEC12 PU chip is an evolution of the z196 core design. It uses CMOS 13S technology, out-of-order instruction processing, higher clock frequency, and redesigned and larger caches. Compute-intensive workloads can achieve more performance improvements through compiler enhancements, and larger caches can improve system performance on many production workloads.

Each PU chip has up to six active cores that run at 5.5 GHz, which means that the cycle time is slightly shorter than 0.18 ns. There are six PU chips on each MCM. The PU chips come in three versions: four, five, and six active cores. For models H20, H43, H66, and H89, the processor units in the MCM in each book are implemented with 27 active cores per MCM. This configuration means that model H20 has 27, model H43 has 54, model H66 has 81, and model H89 has 108 active cores.

Model HA1 has 30 active cores per MCM. This configuration means that there are 120 active cores on model HA1.

A schematic representation of the PU chip is shown in Figure 2-10.

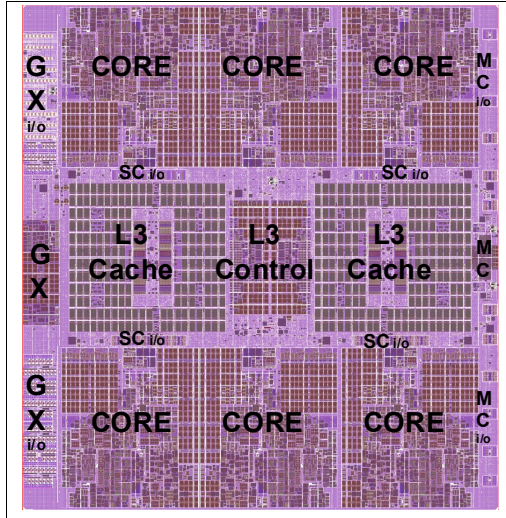


Figure 2-10 PU chip diagram

Each PU chip has 2.75 billion transistors. Each one of the six cores has its own L1 cache with 64 KB for instructions and 96 KB for data. Next to each core is its private L2 cache, with 1 MB for instructions and 1 MB for data.

Each one of the six cores has one L3 cache, with 48 MB. This 48 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 160 GBps bandwidth to each core simultaneously. The L3 cache interconnects the six cores, GX I/O buses, and memory controllers (MCs) with storage control (SC) chips.

The memory controller (MC) function controls access to memory. The GX I/O bus controls the interface to the HCAs accessing the I/O. The chip controls traffic between the cores, memory, I/O, and the L4 cache on the SC chips.

One coprocessor (CoP) 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 3.4.3, “Compression and cryptography accelerators on a chip” on page 86.

2.4.2 Processor unit (core)

Each processor unit, or core, is a superscalar, out-of-order processor that has six execution units:

- ▶ Two fixed point (integer)
- ▶ Two load/store
- ▶ One binary floating point
- ▶ One decimal floating point

Up to three instructions can be decoded per cycle, and up to seven instructions/operations can be initiated to run per clock cycle (<0.18 ns). The execution 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.

The following functional areas are on each core, as shown in Figure 2-11 on page 43:

- ▶ 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 fetching unit (IFU) (prediction): These units contain the instruction cache, branch prediction logic, instruction fetching controls, and buffers. Its relative size is the result of the elaborate branch prediction design. For more information, see 3.4.2, “Superscalar processor” on page 86.
- ▶ Instruction decode unit (IDU): The IDU is fed from the IFU buffers, and is responsible for 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 as defined in the z/Architecture.
- ▶ Translation unit (XU): The XU has a large translation lookaside buffer (TLB) and the Dynamic Address Translation (DAT) function that handles the dynamic translation of logical to physical addresses.
- ▶ 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 DU runs both floating-point and decimal fixed-point operations and fixed-point division operations.
- ▶ 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 Co-Processor (COP): The dedicated coprocessor is responsible for data compression and encryption functions for each core.

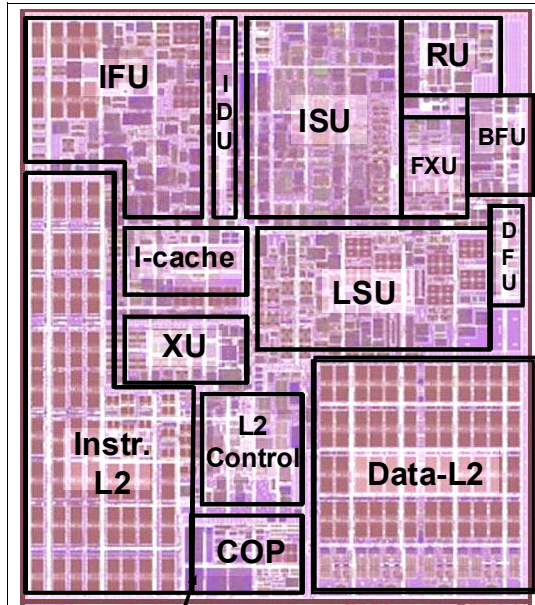


Figure 2-11 Core layout

2.4.3 PU characterization

In each MCM, some PUs can be characterized for client use. The characterized PUs can be used for general purposes to run supported operating systems, such as z/OS, z/VM, and Linux on System z. They can also be specialized to run specific workloads, such as Java, XML services, IPSec, and some DB2 workloads, or functions such as Coupling Facility Control Code. For more information about PU characterization, see 3.5, “Processor unit functions” on page 91.

The maximum number of characterized PUs depends on the zEC12 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 system that are used to assume the function of a failed PU. The remaining installed PUs can be characterized for client use. A zEC12 model nomenclature includes a number that represents this maximum number of PUs that can be characterized for client use, as shown in Table 2-2.

Table 2-2 Number of PUs per zEC12 model

Model	Books	Installed PUs	Standard SAPs	Minimum spare PUs	Maximum characterized PUs	Integrated firmware processor (IFP)
H20	1	27 (1 x 27)	4	2	20	1
H43	2	54 (2 x 27)	8	2	43	1
H66	3	81 (3 x 27)	12	2	66	1
H89	4	108 (4 x 27)	16	2	89	1
HA1	4	120 (4 x 30)	16	2	101	1

2.4.4 Storage control (SC) chip

The SC chip uses the same CMOS 13S 32nm SOI technology, with 15 layers of metal. It measures 28.4 x 23.9 mm, has 3.3 billion transistors, and has 2.1 billion cells for eDRAM. Each MCM has two SC chips. The L4 cache on each SC chip has 192 MB, resulting in 384 MB of L4 cache that is shared per book.

Figure 2-12 shows a schematic representation of the SC chip.

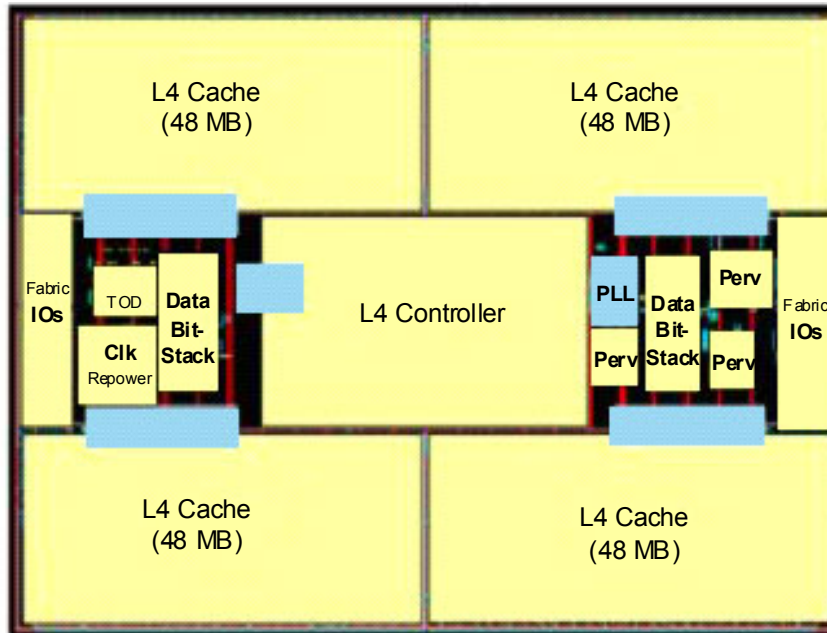


Figure 2-12 SC chip diagram

Most of the SC chip space is taken by the L4 controller and the 192 MB L4 cache. The cache consists of four 48 MB quadrants with 256 x 1.5 MB eDRAM macros per quadrant. The L4 cache is logically organized as 16 address-sliced banks, with 24-way set associativity. The L4 cache controller is a single pipeline with multiple individual controllers, 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 uni-directional buses. L3 is divided into two logical slices. Each slice is 24 MB, and consists of two 12 MB banks. L3 is 12-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 books by three bidirectional data buses. The integrated SMP fabric transport and system coherency manager use the L4 directory to filter snoop traffic from remote books. This process uses an enhanced synchronous fabric protocol for improved latency and cache management. There are two clock domains, and the clock function is distributed between both SC chips.

2.4.5 Cache level structure

The zEC12 implements a four level cache structure, as shown in Figure 2-13.

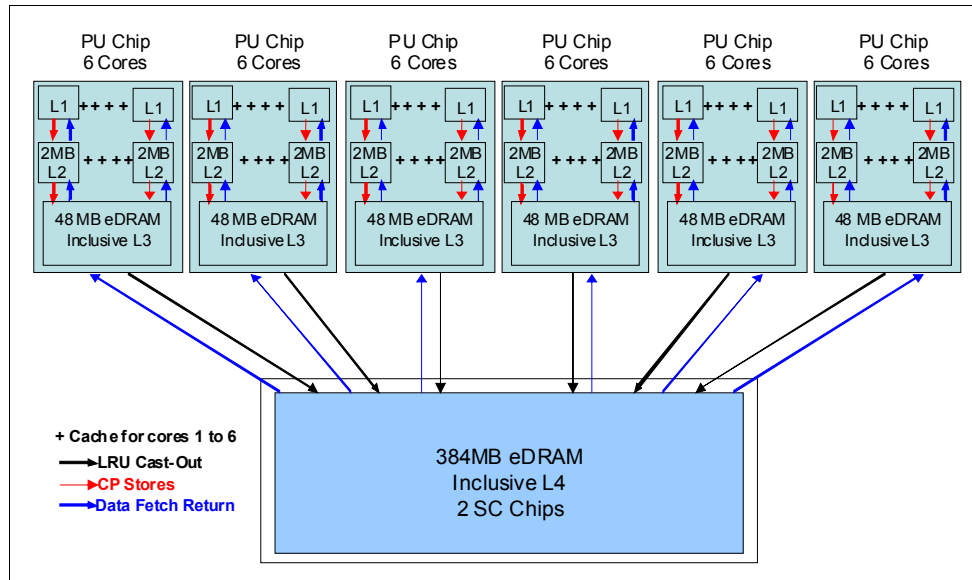


Figure 2-13 Cache level structure

Each core has its own 160-KB cache Level 1 (L1), split into 96 KB for data (D-cache) and 64 KB for instructions (I-cache). The L1 cache is designed as a store-through cache, meaning that altered data is also stored to the next level of memory.

The next level is the private cache Level 2 (L2) on each core. This cache has 2 MB, split into 1 MB D-cache and 1 MB I-cache, and also designed as a store-through cache.

The cache Level 3 (L3) is also on the PU chip. It is shared by the six cores, has 48 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 large shared cache L4, which is on the two SC chips. Each SC chip has 192 MB, resulting in 384 MB of L4 cache, which is shared by all PUs on the MCM. The L4 cache uses a store-in design.

2.5 Memory

Maximum physical memory size is directly related to the number of books in the system. Each book can contain up to 960 GB of physical memory, for a total of 3,840 GB (3.75 TB) of installed memory per system.

A zEC12 has more memory installed than ordered. Part of the physical installed memory is used to implement the redundant array of independent memory (RAIM) design. This configuration results in up to 768 GB of available memory per book and up to 3,072 GB (3 TB) per system.

Table 2-3 shows the maximum and minimum memory sizes you can order for each zEC12 model.

Table 2-3 zEC12 memory sizes

Model	Number of books	Customer memory (GB)
H20	1	32 - 704
H43	2	32 - 1392
H66	3	32 - 2272
H89	4	32 - 3040
HA1	4	32 - 3040

The minimum physical installed memory is 80 GB per book. The minimum initial amount of memory that can be ordered is 32 GB for all zEC12 models. The maximum customer memory size is based on the physical installed memory minus RAIM, and minus HSA memory, which has a fixed amount of 32 GB.

Table 2-4 shows the memory granularity that is based on the installed customer memory.

Table 2-4 Memory granularity

Granularity (GB)	Customer memory (GB)
32	32 - 256
64	320 - 512
96	608 - 896
112	1008
128	1136 - 1520
240	1760
256	2016 - 3040

With the zEC12, the memory granularity varies from 32 GB (for customer memory sizes from 32 to 256 GB) up to 256 GB (for CPCs having from 1760 GB to 3040 GB of customer memory).

2.5.1 Memory subsystem topology

The zEC12 memory subsystem uses high speed, differential-ended communications memory channels to link a host memory to the main memory storage devices.

Figure 2-14 shows an overview of the book memory topology of a zEC12.

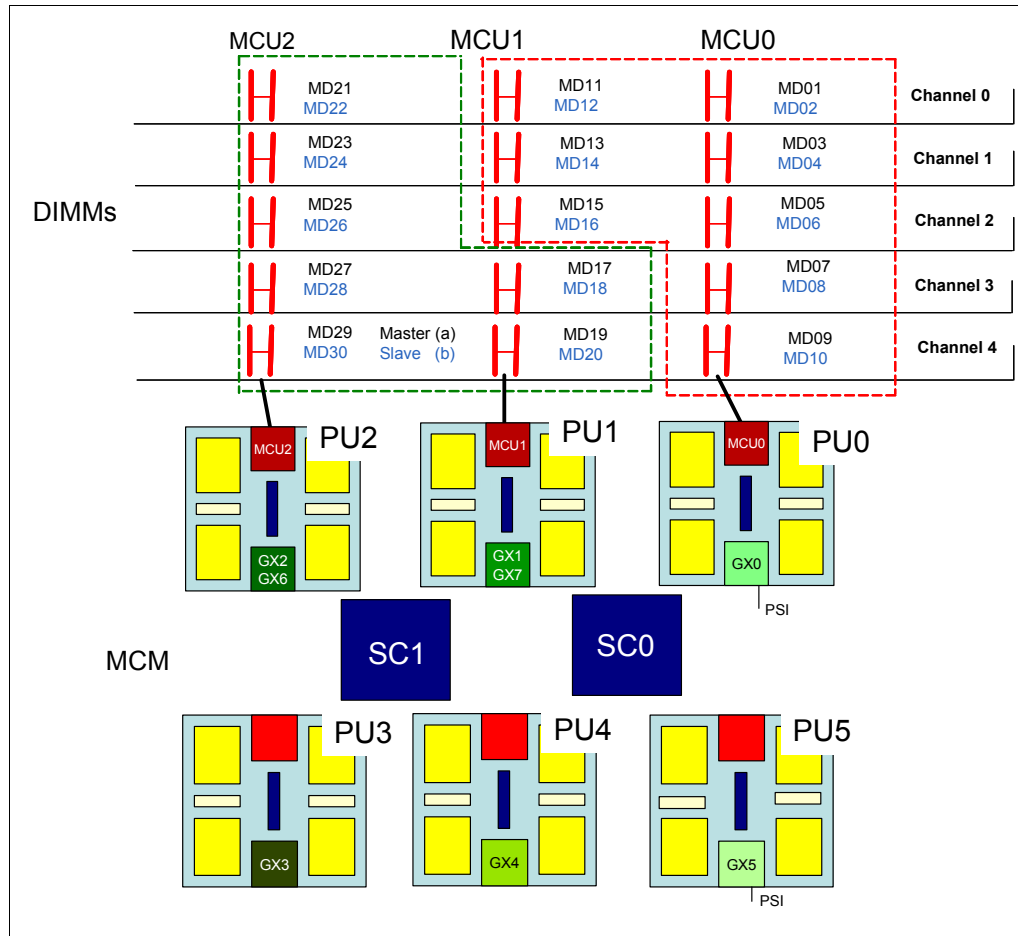


Figure 2-14 Book memory topology

Each book has from 10 to 30 dual inline memory modules (DIMMs). DIMMs are connected to the MCM through three MCUs on PU0, PU1, and PU2. Each MCU uses five channels, one of them for RAIM implementation, in a 4 + 1 (parity) design. Each channel has one or two chained DIMMs, so a single MCU can have five or ten DIMMs. Each DIMM has 4 GB, 16 GB, or 32 GB. You cannot mix DIMM sizes in a book.

2.5.2 Redundant array of independent memory (RAIM)

For a fully fault-tolerant $N+1$ design, the zEC12 use the RAIM technology. The RAIM design detects and recovers from DRAM, socket, memory channel, or DIMM failures.

The RAIM design requires the addition of one memory channel that is dedicated for RAS, as shown in Figure 2-15.

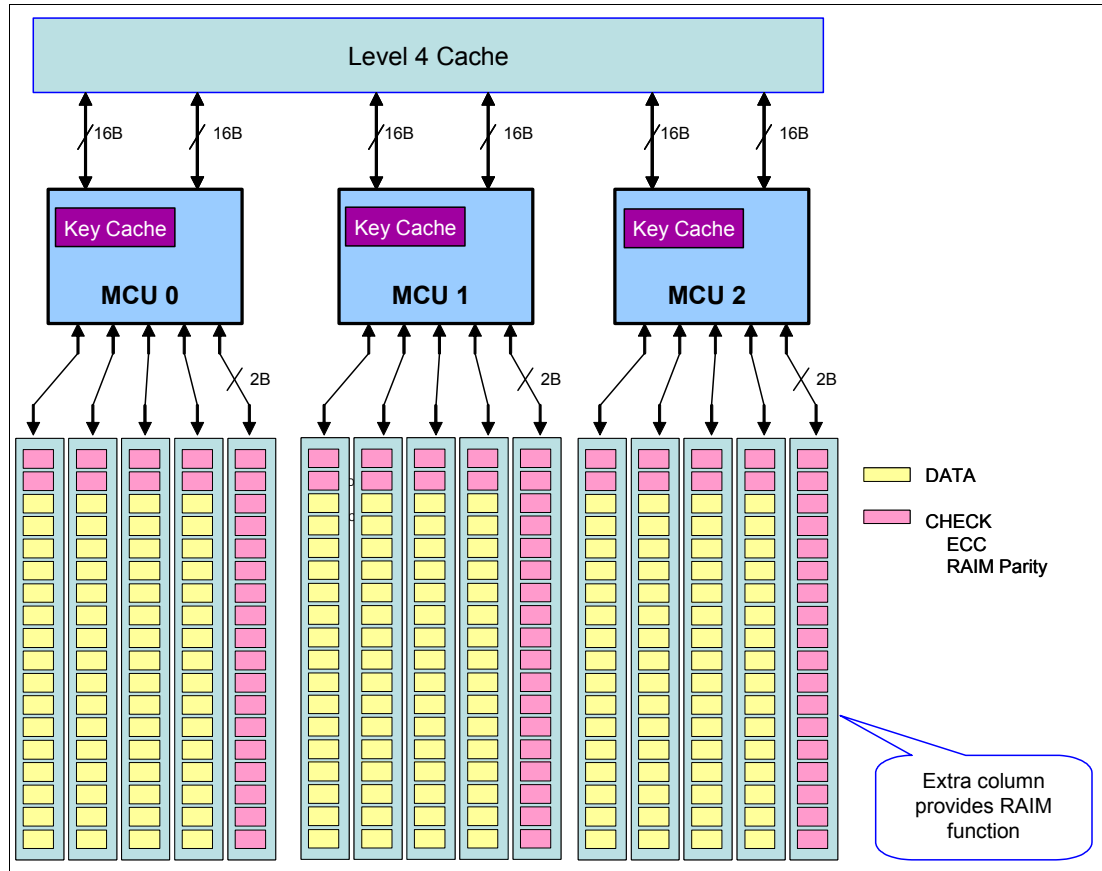


Figure 2-15 RAIM DIMMs

The parity of the four “data” DIMMs are stored in the DIMMs attached to the 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.

2.5.3 Memory configurations

Memory sizes in each book do not have to be similar. Different books can contain different amounts of memory. Table 2-5 shows the physically installed memory on each book for all zEC12 models.

Table 2-5 Physically installed memory

Customer Memory	Model H20	Model H43		Model H66			Model H89 and Model HA1			
(GB)	Book 1	Book 1	Book 3	Book 1	Book 2	Book 3	Book 0	Book 1	Book 2	Book 3
32	80 ^a	80 ^a	80 ^a	40	40	40	40	40	40	40
64	120	80	80	40	40	40	40	40	40	40
96	160	80	80	60	40	40	40	40	40	40
128	240	100	100	60	60	60	60	40	40	40

Customer Memory	Model H20	Model H43		Model H66			Model H89 and Model HA1			
(GB)	Book 1	Book 1	Book 3	Book 1	Book 2	Book 3	Book 0	Book 1	Book 2	Book 3
160	240	120	120	80	80	60	60	60	60	40
192	320	160	120	100	80	80	80	60	60	60
224	320	160	160	100	100	100	80	80	80	60
256	400	240	160	120	120	100	100	80	80	80
320	480	240	240	160	160	100	120	100	100	100
384	640	320	240	240	160	160	160	120	120	100
448	640	320	320	240	240	160	160	160	160	100
512	800	400	320	240	240	240	240	160	160	160
608	800	400	400	320	240	240	240	240	160	160
704	960	480	480	320	320	320	240	240	240	240
800	N/A	640	480	400	320	320	320	240	240	240
896	N/A	640	640	400	400	400	320	320	320	240
1008	N/A	640	640	480	400	400	320	320	320	320
1136	N/A	800	640	480	480	480	400	400	320	320
1264	N/A	800	800	640	480	480	400	400	400	400
1392	N/A	960	800	640	640	480	480	480	400	400
1520	N/A	960	960	640	640	640	480	480	480	480
1760	N/A	N/A	N/A	800	800	640	640	640	480	480
2016	N/A	N/A	N/A	960	800	800	640	640	640	640
2272	N/A	N/A	N/A	960	960	960	800	800	640	640
2528	N/A	N/A	N/A	N/A	N/A	N/A	800	800	800	800
2784	N/A	N/A	N/A	N/A	N/A	N/A	960	960	800	800
3040	N/A	N/A	N/A	N/A	N/A	N/A	960	960	960	960

a. 80 GB for a one or two book system. However, if the system is ordered with more than two books, 40 GB is installed.

Physically, memory is organized in the following manner:

- ▶ A book always contains a minimum of 10 DIMMs of 4 GB each (40 GB).
- ▶ A book has more memory that is installed than enabled. The amount of memory that can be enabled by the client is the total physically installed memory minus the RAIM amount and minus the 32 GB of HSA memory.
- ▶ A book can have available unused memory, which can be ordered as a memory upgrade.

Figure 2-16 illustrates how the physical installed memory is allocated on a zEC12, showing HSA memory, RAIM, customer memory, and the remaining available unused memory that can be enabled by LIC when required.

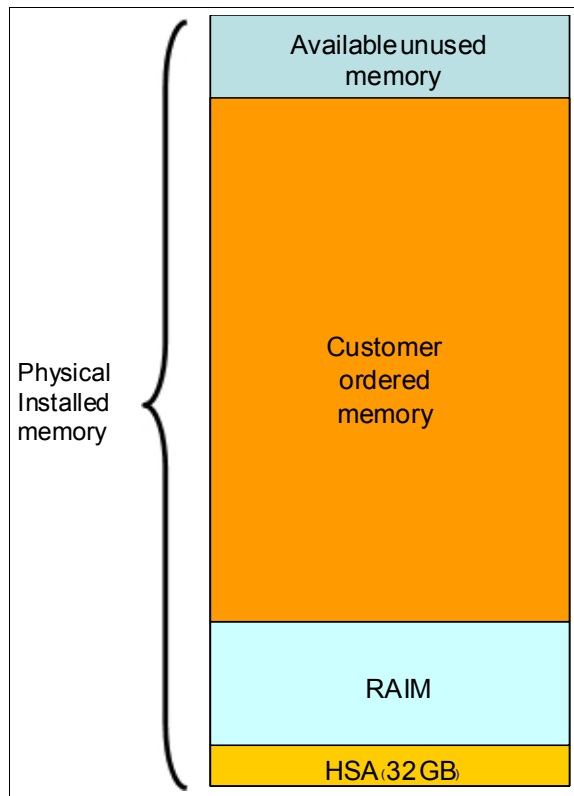


Figure 2-16 Memory allocation diagram

As an example, a zEC12 Model H43 (two books) ordered with 192 GB of memory has the following memory sizes:

- ▶ Physical installed memory is 280 GB: 160 GB on book 1 and 120 GB on book 3.
- ▶ Book 1 has 32 GB of HSA memory and up to 96 GB for customer memory. Book 2 has up to 96 GB for customer memory, resulting in 192 GB of available memory for the customer.
- ▶ Because the customer ordered 176 GB, provided the granularity rules are met, there are still 16 GB (192 - 176 GB) available to be used for future upgrades by 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 book with more memory must be added.
- ▶ Memory cards (DIMMs) must be added.

A memory upgrade is concurrent when it requires no change of the physical memory cards. A memory card change is disruptive when no use is made of enhanced book availability. For more information, see 2.7.2, “Enhanced book availability” on page 58.

When activated, an LPAR can use memory resources that are in any book. No matter where the memory is, an LPAR has access to that memory up to a maximum of 1 TB. This is possible because despite the book structure, the zEC12 is still an SMP system. For more information, see 3.7, “Logical partitioning” on page 108.

2.5.4 Memory upgrades

For a model upgrade that results in the addition of a book, the minimum memory increment is added to the system. The following minimum physical memory sizes are available in book configuration:

- ▶ A one-book system has 80 GB of physical memory.
- ▶ A two-book system has 80 GB of physical memory for each book.
- ▶ Three-book and four-book systems have 80 GB of physical memory in the first book and second book, and 40 GB of physical memory for the third book and fourth book.

During a model upgrade, the addition of a book is a concurrent operation. The addition of the physical memory in the added book is also concurrent. If all or part of the added memory is enabled for use, it might become available to an active LPAR if the partition has reserved storage defined. For more information, see 3.7.3, “Reserved storage” on page 117.

Alternately, the added memory can be used by an already-defined LPAR that is activated after the memory addition.

2.5.5 Book replacement and memory

With enhanced book availability as supported for zEC12, sufficient resources must be available to accommodate resources that are lost when a book is removed for upgrade or repair. For more information, see 2.7.2, “Enhanced book availability” on page 58. Most of the time, removal of a book results in removal of active memory. With the flexible memory option, evacuating the affected memory and reallocating its use elsewhere in the system are possible. For more information, see 2.5.6, “Flexible Memory Option” on page 51. This process requires more available memory to compensate for the memory that is lost with the removal of the book.

2.5.6 Flexible Memory Option

With the Flexible Memory Option, more physical memory is supplied to support activation of the actual purchased memory entitlement in a single book failure. This memory is also available during an enhanced book availability action.

When you order memory, you can request additional flexible memory. The additional physical memory, if required, is calculated by the configurator and priced accordingly.

Flexible memory is available only on the H43, H66, H89, and HA1 models. Table 2-6 shows the flexible memory sizes that are available for the zEC12.

Table 2-6 zEC12 memory sizes

Model	Standard memory (GB)	Flexible memory (GB)
H20	32 - 704	N/A
H43	32 - 1392	32 - 704
H66	32 - 2272	32 - 1392
H89	32 - 3040	32 - 2272
HA1	32 - 3040	32 - 2272

Table 2-7 shows the memory granularity for the Flexible Memory Option.

Table 2-7 Flexible memory granularity

Granularity (GB)	Flexible memory (GB)
32	32 - 256
64	320 - 512
96	608 - 896 ^a
112	1008
128	1136 - 1520 ^b
240	1760
256	1776 - 2288

a. The Model H43 limit is 704 GB.

b. The Model H66 limit is 1392 GB.

Flexible memory can be purchased, but cannot be used for normal everyday use. For that reason, a different purchase price for the flexible memory is offered to increase the overall availability of the system.

2.5.7 Pre-planned memory

Pre-planned memory helps you plan for nondisruptive permanent memory upgrades. It differs from the flexible memory option. The flexible memory option is meant to anticipate nondisruptive book replacement. The usage of flexible memory is therefore temporary, in contrast with plan-ahead memory.

When you are preparing for a future memory upgrade, memory can be pre-plugged, based on a target capacity. The pre-plugged memory can be made available through a LIC configuration code (LICCC) update. You can order this LICCC through these channels:

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

The installation and activation of any pre-planned memory requires the purchase of the required feature codes (FC), as shown in Table 2-8.

Table 2-8 Feature codes for plan-ahead memory

Memory	zEC12 feature code
Pre-planned memory Charged when physical memory is installed. Used for tracking the quantity of physical increments of plan-ahead memory capacity.	FC 1996
Pre-planned memory activation Charged when plan-ahead memory is enabled. Used for tracking the quantity of increments of plan-ahead memory that are being activated.	FC 1901

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 usage. For the exact terms and conditions, contact your IBM representative.

Installation of pre-planned memory is done by ordering FC 1996. The ordered amount of plan-ahead memory is charged at a reduced price compared to the normal price for memory. One FC 1996 is needed for each 16 GB of usable memory (20 GB RAIM).

Activation of installed pre-planned memory is achieved by ordering FC 1901, which causes the other portion of the previously contracted charge price to be invoiced. One FC 1901 is needed for each additional 16 GB to be activated.

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

2.6 Reliability, availability, and serviceability (RAS)

IBM System z continues to deliver enterprise class RAS with the IBM zEnterprise EC12. The main philosophy behind RAS is about preventing or masking outages. 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 System z 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

These examples are highly unlikely on a zEC12 server. The System z hardware has decades of intense engineering, which have resulted in a robust and reliable platform. The hardware has many RAS features built in, as does the software. The IBM Parallel Sysplex architecture is a good example.

2.6.1 RAS in the CPC memory subsystem

Patented error correction technology in the memory subsystem provides the most robust error correction from IBM to date. Two full DRAM failures per rank can be spared and a third full DRAM failure 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 MCM, 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 by the zEC12:

- ▶ Improved error detection for L3/L4 eDRAM configuration latches
- ▶ Improved error detection for L3/L4 wordline failures
- ▶ Improved recovery for unresponsive memory DIMM ASIC
- ▶ Improved recovery of memory command words by changing code points
- ▶ Improved chip marking technology (marking a chip as defective)

Taken together, these improvements provide the most robust memory subsystem of all generations of System z servers.

2.6.2 General zEC12 RAS features

The zEC12 has the following RAS features:

- ▶ The zEC12 provides true $N+1$ (fully redundant) cooling functionality for both radiator-cooled and water-cooled models. In the unlikely case of a failure of both components (both water pumps or both radiators), there is backup in the form of air cooling.
- ▶ The power supplies for the zEC12 are also based on the $N+1$ design. A defective power supply will therefore not cause an unplanned outage of the system.
- ▶ The IBM zEnterprise System CPCs have improved chip packaging (encapsulated chip connectors) and use soft error rate (SER) hardened latches throughout the design.
- ▶ Fully fault protected $N+2$ voltage transformation module (VTM) power conversion. This redundancy protects the processor from the loss of voltage because of VTM failures. System z uses triple redundancy on the environmental sensors (ambient temperature, relative humidity, air density³, and corrosion) for reliability.
- ▶ Improved zBX Fibre Channel link/path testing and diagnostic procedures.
- ▶ Coupling Facility Control Code (CFCC) service enhancements:
 - Structure control information is included in CF dumps
 - Enhanced tracing support
 - Method to measure burst activity
 - Trigger nondisruptive dumping for other soft-failure cases beyond break-duplexing

IBM zEnterprise EC12 continues to deliver robust server designs through exciting new technologies, hardening both new and classic redundancy.

For more information, see Chapter 10, “Reliability, availability, and serviceability” on page 385.

2.7 Connectivity

Connections to I/O cages, I/O drawers, and Parallel Sysplex InfiniBand (PSIFB) coupling are driven from the HCA fanouts. These fanouts are located on the front of the book. Connections to PCIe I/O drawers are driven from the PCIe fanouts that are also located on the front of the books.

³ EPO air density sensor measures air pressure and it is used to indicate altitude.

Figure 2-17 shows the location of the fanouts and connectors for a two-book system. ECF is the External Clock Facility card for the PPS connectivity. OSC is the oscillator card. FSP is the flexible service processor, and LG is the location code for the logic card.

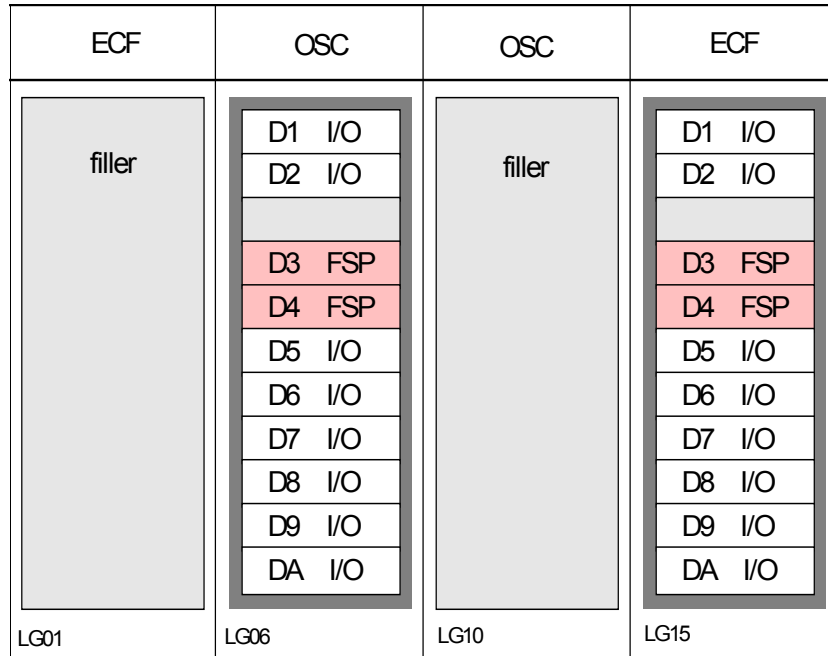


Figure 2-17 Location of the host channel adapter fanouts

Up to eight fanouts (numbered D1, D2, and D5 through DA) can be installed in each book.

A fanout can be repaired concurrently with the use of redundant I/O interconnect. For more information, see 2.7.1, “Redundant I/O interconnect” on page 56.

Six types of fanouts are available:

- ▶ Host Channel Adapter2-C (HCA2-C): Copper connections for InfiniBand I/O interconnect to all FICON, OSA, and crypto cards in I/O cages and I/O drawers.
- ▶ PCIe fanout: Copper connections for PCIe I/O interconnect to all FICON, OSA, Crypto, Flash Express, 10 GbE RoCE Express, and zEDC Express features in PCIe I/O drawers.
- ▶ Host Channel Adapter2-O (HCA2-O): Optical connections for 12x InfiniBand (IFB) for coupling links. The HCA2-O (12xIFB) provides a point-to-point connection over a distance of up to 150 m (492 ft) by using OM3 fiber optic cables (50/125 μm). zEC12 to zBC12, z196, z114, or IBM System z10 connections use 12-lane InfiniBand link at 6 Gbps.
- ▶ Host Channel Adapter2-O Long Reach (HCA2-O LR): Optical connections for 1x IFB that support IFB Long Reach (IFB LR) coupling links for distances of up to 10 km (6.2 miles). Also support up to 100 km (62 miles) when repeated through a System z qualified dense wavelength division multiplexing (DWDM). IFB LR coupling links operate at up to 5.0 Gbps between two CPCs, or automatically scale down to 2.5 Gbps depending on the capability of the attached equipment.
- ▶ Host Channel Adapter3-O (HCA3-O): Optical connections for 12x IFB or 12x IFB3 for coupling links. For more information, see “12x IFB and 12x IFB3 protocols” on page 146. The HCA3-O (12xIFB) provides a point-to-point connection over a distance of up to 150 m (492 ft) by using OM3 fiber optic cables (50/125 μm).

- ▶ Host Channel Adapter3-O Long Reach (HCA3-O LR): Optical connections for 1x InfiniBand and supports IFB Long Reach (IFB LR) coupling links for distances of up to 10 km (6.2 miles). It can support up to 100 km (62 miles) when repeated through a System z qualified DWDM.

IFB LR coupling links operate at up to 5.0 Gbps between two CPCs, or automatically scale down to 2.5 Gbps depending on the capability of the attached equipment.

Fanout positions:

- ▶ On a model H20 and a model H43, all fanout positions can be populated.
- ▶ On a model H66, all fanout positions can be populated only on the first book. Positions D1 and D2 must remain free of fanouts on both the second and third books.
- ▶ On models H89 and HA1, all D1 and D2 positions must remain free of any fanout.

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

Enhanced book availability (EBA) allows a single book in a multibook CPC to be concurrently removed and reinstalled for an upgrade or a repair. Removing a book means that the connectivity to the I/O devices that are connected to that book is lost. To prevent connectivity loss, the redundant I/O interconnect feature allows you to maintain connection to critical devices, except for Parallel Sysplex InfiniBand coupling, when a book is removed.

2.7.1 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 book. IFB-MP cards are half-high cards and are interconnected with cards called STI-A8 and STI-A4. This configuration allows redundant I/O interconnect if the connection coming from a book ceases to function. This situation can happen when, for example, a book is removed for maintenance.

A conceptual view of how redundant I/O interconnect is accomplished is shown in Figure 2-18.

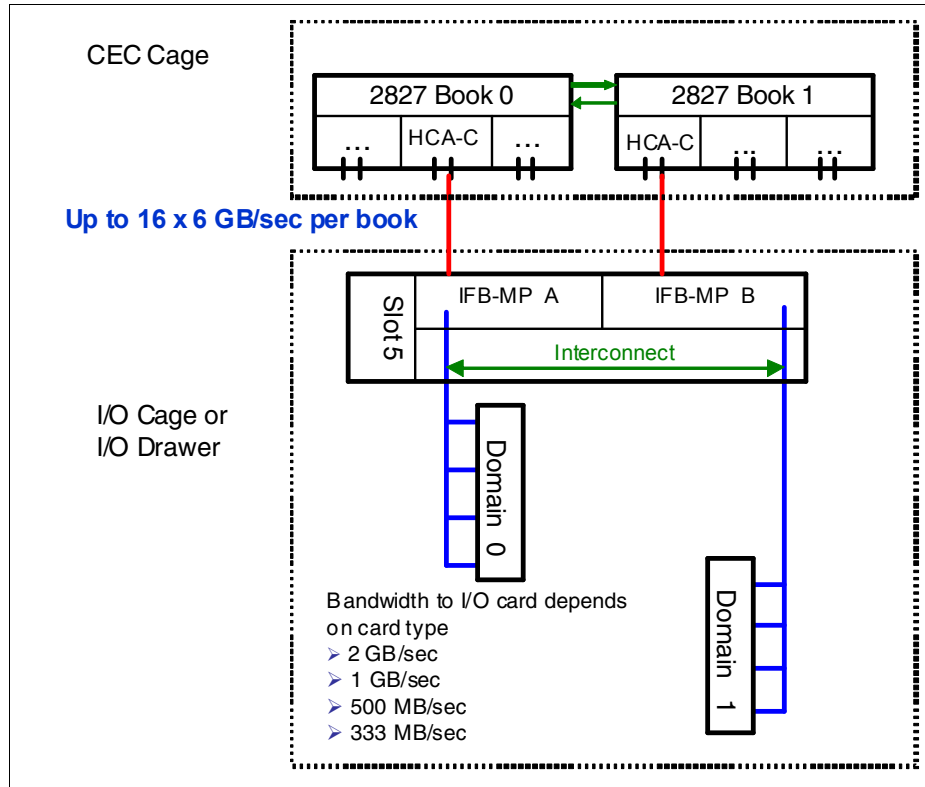


Figure 2-18 Redundant I/O interconnect

Normally, the HCA2-C fanout in the first book connects to the IFB-MP (A) card and services domain 0 in an I/O cage or I/O drawer. In the same fashion, the HCA2-C fanout of the second book connects to the IFB-MP (B) card and services domain 1 in an I/O cage or I/O drawer. If the second book is removed, or the connections from the second book to the cage or drawer are removed, connectivity to domain 1 is maintained. The I/O is guided to domain 1 through the interconnect between IFB-MP (A) and IFB-MP (B).

PCIe I/O connection

The PCIe I/O Drawer supports up to 32 I/O cards. They are organized in four hardware domains per drawer, as shown in Figure 2-19 on page 58.

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 I/O cards 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 I/O cards in its domain. In backup operation mode, one PCIe interconnect supports all 16 I/O cards in the domain pair.

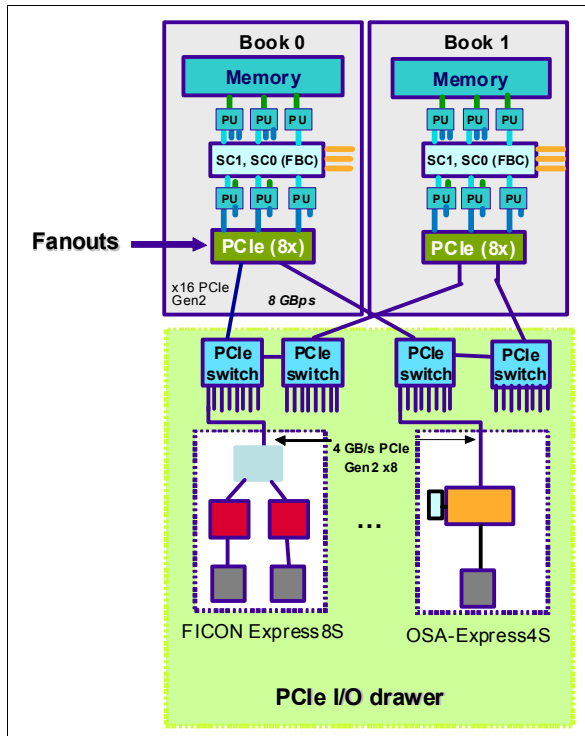


Figure 2-19 Redundant I/O interconnect for PCIe I/O drawer

2.7.2 Enhanced book availability

With enhanced book availability, the effect of book replacement is minimized. In a multiple book system, a single book can be concurrently removed and reinstalled for an upgrade or repair. Removing a book without affecting the workload requires sufficient resources in the remaining books.

Before removing the book, the contents of the PUs and memory from the book to be removed must be relocated. More PUs must be available on the remaining books to replace the deactivated book. Also, sufficient redundant memory must be available if no degradation of applications is allowed. To ensure that the CPC configuration supports removal of a book with minimal effect to the workload, consider the flexible memory option. Any book can be replaced, including the first book that initially contains the HSA.

Removal of a book also removes the book connectivity to the I/O cages and PCIe I/O drawers. The effect of the removal of the book on the system is limited by the use of redundant I/O interconnect. For more information, see 2.7.1, “Redundant I/O interconnect” on page 56. However, all PSIFB links on the removed book must be configured offline.

If the enhanced book availability and flexible memory options are *not* used when a book needs to be replaced, the memory in the failing book is also removed. This might be the case in an upgrade or a repair action. Until the removed book is replaced, a power-on reset of the system with the remaining books is supported.

2.7.3 Book upgrade

All fanouts that are used for I/O and HCA fanouts that are used for IFB are concurrently rebalanced as part of a book addition.

2.8 Model configurations

When a zEC12 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 System z 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 System z 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.
zAAP	An IBM System z Application Assist Processor (zAAP) purchased and activated to run eligible workloads, such as Java code under control of z/OS Java virtual machine (JVM) or z/OS XML System Services.
zIIP	An IBM System z Integrated Information Processor (zIIP) purchased and activated to run eligible workloads, such as DB2 Distributed Relational Database Architecture (DRDA) and 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 characterized as a CP, IFL, or ICF is required per system. The maximum number of CPs, IFLs, and ICFs is 101. The maximum number of zAAPs is 50, but requires an equal or greater number of characterized CPs. The maximum number of zIIPs is also 50, and requires an equal or greater number of characterized CPs. The sum of all zAAPs plus zIIPs cannot be larger than two times the number of characterized CPs.

Not all PUs on a model are required to be characterized.

Also, the following items are present in the zEC12, but they are not part of the PUs that clients purchase and require no characterization:

- ▶ System Assist Processor (SAP) to be used by the channel subsystem. The number of predefined SAPs is dependent on the zEC12 model.
- ▶ One integrated firmware processor (IFP). The IFP is used in the support of designated features, such as zEDC and 10GbE RoCE.
- ▶ Two spare PUs, which can transparently assume any characterization, in the permanent failure of another PU.

The zEC12 model nomenclature is based on the number of PUs available for client use in each configuration. The models are summarized in Table 2-9.

Table 2-9 zEC12 configurations

Model	Books	PUs per MCM	CPs	IFLs/uIFL	ICFs	zAAPs	zIIPs	Add. SAPs	Std. SAPs	Spares	IFP
H20	1	27	0 - 20	0 - 20 0 - 19	0 - 20	0 - 10	0 - 10	0 - 34	4	2	1
H43	2	54	0 - 43	0 - 43 0 - 42	0 - 43	0 - 21	0 - 21	0 - 8	8	2	1
H66	3	81	0 - 66	0 - 66 0 - 65	0 - 66	0 - 33	0 - 33	0 - 12	12	2	1
H89	4	108	0 - 89	0 - 89 0 - 88	0 - 89	0 - 44	0 - 44	0 - 16	16	2	1
HA1	4	120	0 - 101	0 - 101 0 - 100	0 - 101	0 - 44	0 - 44	0 - 16	16	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 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.

2.8.1 Upgrades

Concurrent upgrades of CPs, IFLs, ICFs, zAAPs, zIIPs, or SAPs are available for the zEC12. However, concurrent PU upgrades require that more PUs are installed, but not activated.

Spare PUs are used to replace defective PUs. There are always two spare PUs on a zEC12. In the rare event of a PU failure, a spare PU is concurrently and transparently activated, and assigned the characteristics of the failing PU.

If an upgrade request cannot be accomplished within the configuration, a hardware upgrade is required. The upgrade enables the addition of one or more books to accommodate the wanted capacity. Additional books can be installed concurrently.

Although upgrades from one zEC12 model to another zEC12 model are concurrent, meaning that one or more books can be added, there is one exception. Upgrades from any zEC12 (model H20, H43, H66, or H89) to a model HA1 are disruptive, because the upgrade requires the replacement of all four books.

Table 2-10 shows the possible upgrades within the zEC12 configuration range.

Table 2-10 zEC12 to zEC12 upgrade paths

To 2827 From 2827	Model H20	Model H43	Model H66	Model H89	Model HA1 ^a
Model H20	-	Yes	Yes	Yes	Yes
Model H43	-	-	Yes	Yes	Yes
Model H66	-	-	-	Yes	Yes
Model H89	-	-	-	-	Yes

a. Disruptive upgrade

You can also upgrade a z10 EC or a z196 to a zEC12 and preserve the CPC serial number (S/N). The I/O cards can also be carried forward (with certain restrictions) to the zEC12.

Important: Upgrades from System z10 and zEnterprise 196 (System z196) are always disruptive.

Upgrade paths from any z10 Enterprise Class (EC) to any zEC12 are supported as listed in Table 2-11.

Table 2-11 z10 EC to zEC12 upgrade paths

To 2827 From 2097	Model H20	Model H43	Model H66	Model H89	Model HA1
Model E12	Yes	Yes	Yes	Yes	Yes
Model E26	Yes	Yes	Yes	Yes	Yes
Model E40	Yes	Yes	Yes	Yes	Yes
Model E56	Yes	Yes	Yes	Yes	Yes
Model E64	Yes	Yes	Yes	Yes	Yes

Upgrades from any z196 to any zEC12 are supported as listed in Table 2-12.

Table 2-12 z196 to zEC12 upgrade paths

To 2827 From 2817	Model H20	Model H43	Model H66	Model H89	Model HA1
Model M15	Yes	Yes	Yes	Yes	Yes
Model M32	Yes	Yes	Yes	Yes	Yes
Model M49	Yes	Yes	Yes	Yes	Yes
Model M66	Yes	Yes	Yes	Yes	Yes
Model M80	Yes	Yes	Yes	Yes	Yes

2.8.2 Concurrent PU conversions

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

Most conversions are nondisruptive. In exceptional cases, the conversion can be disruptive, for example, when a model H20 with 20 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 2-13.

Table 2-13 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	-

2.8.3 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, a zAAP, or a zIIP is not reflected in the output of the STSI instruction. This is because characterization has no effect on software charging. For more information about STSI output, see “Processor identification” on page 379.

Four distinct model capacity identifier ranges are recognized (one for full capacity and three for granular capacity):

- ▶ For full-capacity engines, model capacity identifiers 701 to 7A1 are used. They express the 101 possible capacity settings from one to 101 characterized CPs.
- ▶ Three model capacity identifier ranges offer a unique level of granular capacity at the low end. They are available when no more than 20 CPs are characterized. These three subcapacity settings are applied to up to 20 CPs, which offer 60 more capacity settings. For more information, see “Granular capacity” on page 62.

Granular capacity

The zEC12 offers 60 capacity settings at the low end of the processor. Only 20 CPs can have granular capacity. When subcapacity settings are used, other PUs beyond 20 can be characterized only as specialty engines.

The three defined ranges of subcapacity settings have model capacity identifiers numbered from 401- 420, 501 - 520, and 601 - 620.

Consideration: Within a zEC12, all CPs have the same capacity identifier. Specialty engines (IFLs, zAAPs, zIIPs, and ICFs) operate at full speed.

List of model capacity identifiers

Table 2-14 shows that regardless of the number of books, a configuration with one characterized CP is possible. For example, model HA1 might have only one PU characterized as a CP.

Table 2-14 Model capacity identifiers

zEC12	Model capacity identifier
Model H20	701 - 720, 601 - 620, 501 - 520, and 401 - 420
Model H43	701 - 743, 601 - 620, 501 - 520, and 401 - 420
Model H66	701 - 766, 601 - 620, 501 - 520, and 401 - 420
Model H89	701 - 789, 601 - 620, 501 - 520, and 401 - 420
Model HA1	701 - 7A1, 601 - 620, 501 - 520, and 401 - 420

Important: On zEC12, model capacity identifier 400 is used for only IFL or ICF configurations.

2.8.4 Model capacity identifier and MSU value

All model capacity identifiers have a related millions of service units (MSU) value. The MSU values are used to determine the software license charge for MLC software. Tables with MSU values are available on the Mainframe Exhibits for IBM Servers website:

<http://www-03.ibm.com/systems/z/resources/swprice/reference/exhibits/hardware.html>

2.8.5 Capacity Backup

Capacity Backup (CBU) delivers temporary backup capacity in addition to the capacity that an installation might have already available in numbers of assigned CPs, IFLs, ICFs, zAAPs, zIIPs, and optional SAPs. CBU has the following types:

- ▶ CBU for CP
- ▶ CBU for IFL
- ▶ CBU for ICF
- ▶ CBU for zAAP
- ▶ 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, zAAPs, 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 CBUs 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, zAAPs, and zIIPs) always run at full capacity, also when running as CBU to replace lost capacity for disaster recovery.

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

You can test the CBU. The number of CBU test activations that you can run for no additional fee in each CBU record is now determined by the number of years that are purchased with the CBU record. For example, a three-year CBU record has three test activations, as compared to a one-year CBU record that 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 that are related to the CBU:

- ▶ Number of CP CBUs allowed to be activated
- ▶ Number of IFL CBUs allowed to be activated
- ▶ Number of ICF CBUs allowed to be activated
- ▶ Number of zAAP CBUs allowed to be activated
- ▶ Number of zIIP CBUs allowed to be activated
- ▶ Number of SAP CBUs allowed to be activated
- ▶ Number of additional CBU tests that are allowed 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 IBM configurator output, which is 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(5)

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 configuration for two CP CBUs, two IFL CBUs, two zAAP CBUs, and two zIIP CBUs, with five more tests and a 5-year CBU contract. The result is now a total number of 10 years of CBU ordered: Five years in the first record and five years in the second record. The two CBU records are independent, and can be activated individually. Five more CBU tests have been requested. Because a total of five years are contracted for a total of three CP CBUs (two IFL CBUs, two zAAPs, and two zIIP CBUs), they are shown as 15, 10, 10, and 10 CBU years for their respective types.

Example 2-2 Second CBU record and resulting configuration features

NEW00001 - CBU - Replenishment is required to reactivate
Expiration(06/21/2017)
NEW00002 - CBU - CP(2) - IFL(2) - zAAP(2) - zIIP(2)

Total Tests(10) - Years(5)

Resulting cumulative feature numbers in configuration:

6805	5 Additional CBU Tests	5
6817	Total CBU Years Ordered	10
6818	CBU Records Ordered	2
6820	Single CBU CP-Year	15
6822	Single CBU IFL-Year	10
6826	Single CBU zAAP-Year	10
6828	Single CBU zIIP-Year	10

CBU for CP rules

Consider the following guidelines when you are 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 that change the capacity level. For example, if 2 CBU CPs are added to the current model 503, and the capacity level does not change, the 503 becomes 505:

$$(503 + 2 = 505)$$

If the capacity level changes from a 503 to a 606, the number of additional CPs (3) is added to the 3 CPs of the 503, resulting in a total number of CBU CP capacity features of 6:

$$(3 + 3 = 6)$$

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

Remember: Activation of CBU for CPs, IFLs, ICFs, zAAPs, zIIPs, and SAPs can be activated together with On/Off Capacity on Demand temporary upgrades. Both facilities can be on a single system, and can be activated simultaneously.

CBU for specialty engines

Specialty engines (ICFs, IFLs, zAAPs, and zIIPs) run at full capacity for all capacity settings. This also applies to CBU for specialty engines. Table 2-15 shows the minimum and maximum (min-max) numbers of all types of CBUs that might be activated on each of the models. Remember that the CBU record can contain larger numbers of CBUs than can fit in the current model.

Table 2-15 Capacity backup matrix

Model	Total PUs available	CBU CPs min - max	CBU IFLs min - max	CBU ICFs min - max	CBU zAAPs min - max	CBU zIIPs min - max	CBU SAPs min - max
Model H20	20	0 - 20	0 - 20	0 - 20	0 - 10	0 - 10	0 - 4
Model H43	43	0 - 43	0 - 43	0 - 43	0 - 21	0 - 21	0 - 8
Model H66	66	0 - 66	0 - 66	0 - 66	0 - 33	0 - 33	0 - 12
Model H89	89	0 - 89	0 - 89	0 - 89	0 - 44	0 - 44	0 - 16
Model HA1	101	0 - 101	0 - 101	0 - 101	0 - 50	0 - 50	0 - 16

Unassigned IFLs are ignored because they are considered spares, and are available for use as CBU. When an unassigned IFL is converted to an assigned IFL, or when more PUs are characterized as IFLs, the number of CBUs of any type that can be activated is decreased.

2.8.6 On/Off Capacity on Demand and CPs

On/Off Capacity on Demand (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 CPs equals the sum of the number of permanent CPs plus the number of temporary CPs ordered. For example, when a model capacity identifier 504 has two CP5s added temporarily, it becomes a model capacity identifier 506.

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 CPC with model capacity identifier 504 specifies six CP6 temporary CPs through On/Off CoD, the result is a CPC with model capacity identifier 606. A cross-over does not necessarily mean that the CP count for the additional temporary capacity will increase. The same 504 might temporarily be upgraded to a CPC with model capacity identifier 704. 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 CPC.
- ▶ Adding more engines than are currently installed is not possible.

For more information about temporary capacity increases, see Chapter 9, “System upgrades” on page 339.

2.9 Power and cooling

The zEC12 power and cooling system is a continuation of z196 with the addition of some significant newly developed technologies. The power service specifications of the zEC12 are the same as the z196. In the zEC12, a new radiator unit was designed to replace the modular refrigeration unit (MRU) in the z196. The water cooling system is still an option of the zEC12. The new Top Exit Power feature is available for the zEC12. Combined with the Top Exit I/O Cabling feature, this feature gives you more options when you are planning your computer room cabling. For more information about the zEC12 Top Exit features, see 11.3, “IBM zEnterprise EC12 physical planning” on page 416.

2.9.1 Power consumption

The system operates with two redundant power feeds. Each power feed is either 1 or 2 power cords. The number of required power cords depends on the system configuration (the number of Bulk Power Regulator (BPR) pairs installed). Power cords attach to either a 3 phase, 50/60 Hz, 200 - 480 V AC power source, or a 380 - 520 V DC power source. There is no effect on system operation with the total loss of one power feed.

A Balanced Power Plan Ahead feature is available for future growth, helping to ensure adequate and balanced power for all possible configurations. With this feature, system downtime for upgrading a server is 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 Hardware Management Console, its display, and its switch, more single phase outlets are required.

The power requirements depend on the cooling facility that is installed, and on the number of books and I/O units that are installed. For more information about the requirements related to the number of installed I/O units, see 11.1.1, "Power consumption" on page 410.

2.9.2 High voltage DC power

The High Voltage Direct Current power feature is an option for IBM zEnterprise EC12. It allows zEC12 to directly use the high voltage (HV) DC distribution as shown in Figure 2-20 on page 68. 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 zEC12 bulk power supplies have been modified to support HV DC, so the only difference in the shipped hardware to implement this option is the DC power cords. Because HV DC is a new technology, there are multiple proposed standards. The zEC12 supports both ground referenced and dual polarity HV DC supplies, such as +/-190 V, +/-260 V, and +380 V. Beyond the data center uninterruptible power supply and power distribution energy savings, a zEC12 that runs on HV DC power draws 1 - 3% less input power. HV DC does not change the number of power cords that a system requires.

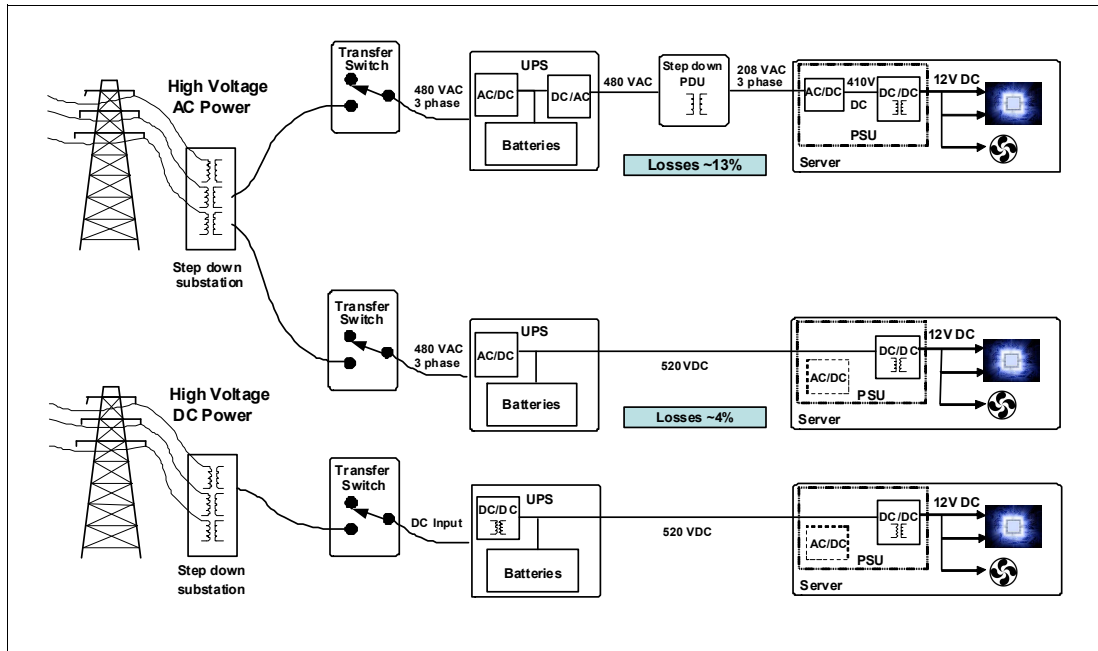


Figure 2-20 AC versus DC distribution

2.9.3 Internal Battery Feature (IBF)

The Internal Battery Feature (IBF) is an optional feature on the zEC12 server. It is shown in Figure 2-1 on page 30 for air-cooled (radiator) models and in Figure 2-2 on page 31 for water-cooled models. The IBF provides 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 all power feeds from the computer room. The IBF can hold power briefly during a brownout, or for orderly shutdown in a longer outage. The IBF provides up to 10 minutes of full power. For more information about the hold times, which depend on the I/O configuration, see 11.1.2, “Internal Battery Feature” on page 411.

2.9.4 Power capping and power saving

The zEC12 supports power capping, which limits the maximum power consumption and reduces the cooling requirements (especially with the zBX). To use power capping, FC 0020, Automate Firmware Suite, is required. This feature is used to enable the Automate suite of functions associated with the zEnterprise Unified Resource Manager. The Automate suite includes resources in a workload context, goal-oriented monitoring, the management of resources, and energy management.

2.9.5 Power estimation tool

The power estimation tool for the zEC12 allows you to enter your precise server configuration to produce an *estimate* of power consumption. Log in to the Resource link with your user ID. Click **Planning Tools Power Estimation Tools**. 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.

Tip: The exact power consumption for your system will vary. The object of the tool is to estimate the power requirements to aid you in planning for your system installation. Actual power consumption after installation can be confirmed by using the HMC Monitors Dashboard task.

2.9.6 Cooling

Closed water loop cooling technology is used in zEC12 MCM cooling. The zEC12 still has both air-cooled and water-cooled models.

Air-cooled models

In the zEC12, books, I/O drawers, and power enclosures are all cooled by forced air with blowers that are controlled by Move Device Assembly (MDA).

The MCMs in books are cooled by water that comes from a radiator unit. The radiator unit is a new cooling component in the zEC12 that replaces the modular refrigeration unit (MRU) that is used in the z196. In addition, the evaporator was redesigned and replaced by an MCM cold plate. The internal closed water loop takes heat away from the MCM by circulating between the radiator heat exchanger and the cold plate that is mounted on the MCM. For more information, see 2.9.7, “Radiator Unit” on page 69.

Although the MCM is cooled by water, the heat is exhausted into the room from the radiator heat exchanger by forced air with blowers. At the system level, it is still an air-cooled system.

Water-cooled models

The zEC12 has an available water-cooled system. With water cooling unit (WCU) technology, the zEC12 can transfer most of the heat that is generated into the building’s chilled water, effectively reducing the heat output to the computer room.

Unlike the radiator in air-cooled models, a WCU has two water loops: an internal closed water loop and an external (chilled) water loop. The external water loop connects to the client-supplied building’s chilled water. The internal water loop circulates between the WCU heat exchanger and MCM cold plate. The loop takes heat away from the MCM and transfers it to the external water loop in the WCU heat exchanger. For more information, see 2.9.8, “Water Cooling Unit (WCU)” on page 71.

In addition to the MCMs, the internal water loop also circulates through two heat exchangers that are in the path of the exhaust air in the rear of the frames. These heat exchangers remove approximately 60% - 65% of the residual heat from the I/O drawers, the air-cooled logic in the books, and the power enclosures. Almost two-thirds of the total heat that is generated can be removed from the room by the chilled water.

The selection of air-cooled models or water-cooled models is done when ordering, and the appropriate equipment is factory installed. An MES from an air-cooled model to a water-cooled model is not allowed, and vice versa.

2.9.7 Radiator Unit

The Radiator Unit provides cooling to MCM with closed loop water. No connection to an external chilled water supply is required. For the zEC12, the internal circulating water is conditioned water that is added to the radiator unit during system installation with a specific Fill and Drain Tool (FC 3378). The Water Fill and Drain Tool is shipped with the new zEC12.

As Figure 2-21 shows, the water pumps, manifold assembly, radiator assembly (includes heat exchanger), and blowers are the main components of the zEC12 radiator unit.

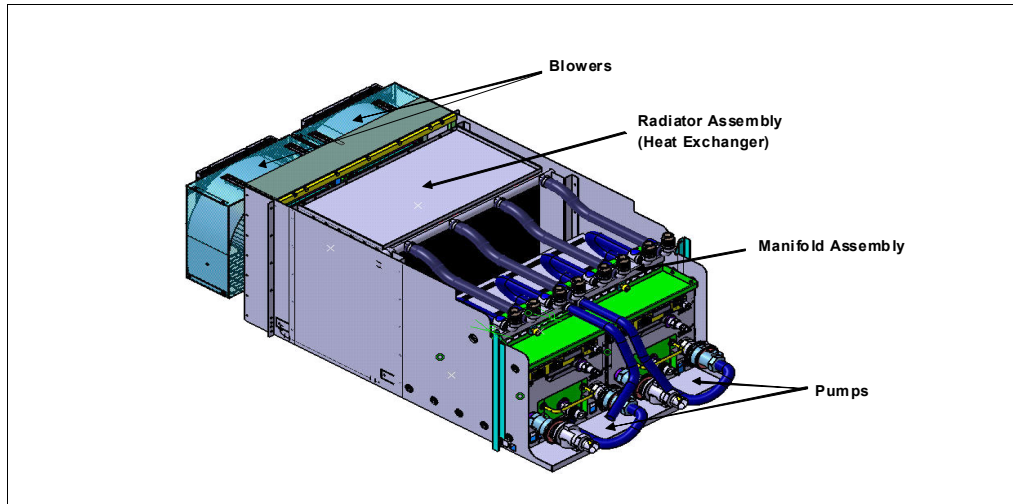


Figure 2-21 Radiator unit

The radiator unit can connect to all four books and cool all MCMs simultaneously. The cooling capability is a redundant design, so a single working pump and blower can support the entire load. The replacement of one pump or blower can be done concurrently and has no performance effect.

Figure 2-22 shows the closed water loop in the radiator unit. The warm water exiting from the MCM cold plates enters pumps through a common manifold. It is pumped through a heat exchanger where heat is extracted by the air flowing across the heat exchanger fins. The cooled water is then recirculated back into the MCM cold plates.

The radiator air-cooling system is open loop, so there is no specific target temperature. There is an estimated range that is based upon MCM power and ambient temperature. However, the radiator blowers are throttled up in increments of air flow when the MCM temperature increases.

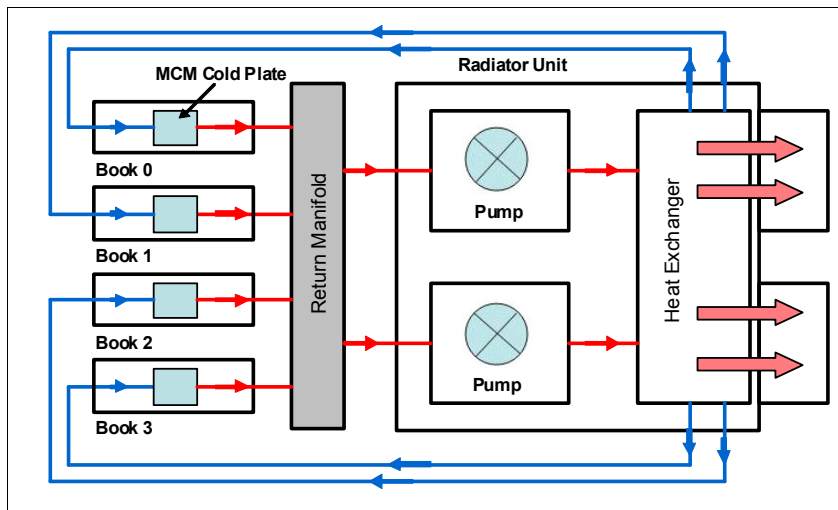


Figure 2-22 Radiator cooling system

Like the MRU design in the IBM zEnterprise 196 (z196), the backup blowers are the redundant radiator unit in the zEC12. If MCM temperature increases to the threshold limit because the radiator unit fails or the environment temperature is too high, the backup blowers are turned on to strengthen the MCM cooling. In this situation, the *cycle steering* mode is required. For more information about backup blower and cycle steering mode, see 2.9.9, “Backup air cooling system” on page 74.

In the zEC12, backup blowers are also provided primarily to allow the concurrent replacement of the radiator (heat exchanger) or the return manifold. During the replacement of these parts, cycle steering mode might be required.

2.9.8 Water Cooling Unit (WCU)

The zEC12 continues to provide the ability to cool systems with building chilled water by employing the WCU technology. The MCM in the book is cooled by internal closed loop water. The internal closed loop water exchanges heat with building chilled water in the WCU heat exchanger. The source of the building’s chilled water is provided by the client.

A simplified high-level diagram that illustrates the principle is shown in Figure 2-23.

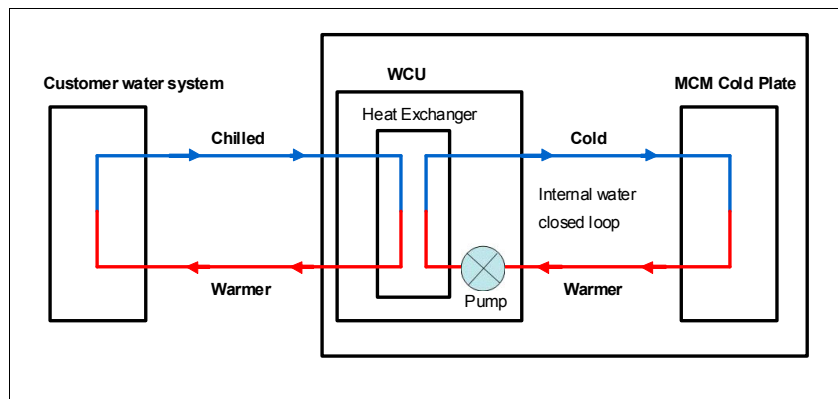


Figure 2-23 WCU water loop

The water in the closed loop within the system exchanges heat with the continuous supply of building chilled water. The internal water loop contains approximately 23 liters (6 US gallons) of conditioned water. This water circulates between the MCM cold plates and a heat exchanger within the WCU. Heat from the MCMs is transferred to the cold plates where it is in turn transferred to the circulating system water (closed loop). The system water then loses its heat to the building chilled water within the WCU’s heat exchanger. The MCMs are efficiently cooled in this manner.

The zEC12 operates with two fully redundant WCUs. These water cooling units each have their own facility feed and return water connections. If water is interrupted to one of the units, the other unit picks up the entire load, and the server continues to operate without interruption. You must provide independent redundant water loops to the water cooling units to obtain full redundancy.

In a total loss of building chilled water or if both water cooling units fail, the backup blowers turn on to keep the server running, just like in air-cooled models. Currently, cycle time degradation is required by this process. For more information, see 2.9.9, “Backup air cooling system” on page 74.

The internal circulating water in the water cooling unit is conditioned water. It is added to the two WCUs during system installation with a specific Fill and Drain Tool (FC 3378), just like the radiator unit. A Water Fill and Drain Tool is shipped with the new zEC12.

Exhaust Air Heat Exchanger

In the zEC12, all water-cooled models have two Exhaust Air Heat Exchanger units installed on the rear of the A and Z frames, as shown in Figure 2-24. These units remove heat from the internal system water loop and internal air exiting the server into the hot air exhaust aisle.



Figure 2-24 Water-cooled model - rear view

In addition to the MCM cold plates, the internal water loop also circulates through these two heat exchangers. These exchangers are in the path of the exhaust air in the rear of the frames. These heat exchangers remove approximately 65% of the residual heat from the I/O drawers, the air-cooled logic in the book, and the power enclosures. The goal is for two-thirds of the total heat that is generated to be removed from the room by the chilled water.

A diagram of the entire circulation system is shown in Figure 2-25.

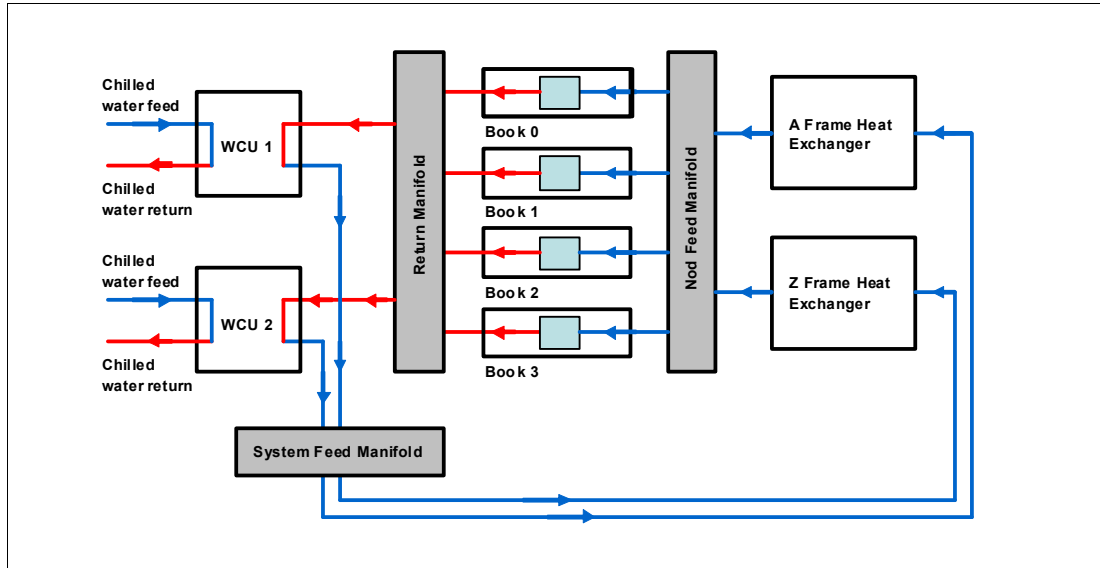


Figure 2-25 WCU complete water loop

This figure shows the complete water loop of WCU. Two client water feeds connect to the two redundant WCUs. The WCUs feed water to a common feed manifold that supplies water to the frame heat exchangers. Two frame heat exchangers are fed in parallel from this manifold. The node feed manifold feeds the four book positions in parallel. The heat is exchanged with the client-supplied chilled water in the WCUs, which also pump the system water around the loop.

If one client water supply or one WCU fails, the remaining feed maintains MCM cooling. The WCUs and the associated drive card are concurrently replaceable. Also, the heat exchangers can be disconnected and removed from the system concurrently.

Considerations before you order

The water cooling option is preferable because it can substantially lower the total power consumption of the zEC12 and, therefore, lower the total cost of ownership for the CPC. This savings is greater for the bigger models of the zEC12, as shown in Table 2-16.

The water cooling option cannot be installed in the field. Therefore, you must carefully consider the present and future computer room and CPC configuration options before you decide what cooling option to order. For more information, see 11.1.4, “Cooling requirements” on page 412.

Table 2-16 Power consumption based on temperature

Temperature	Three book typical configuration	Four book typical configuration	Four book maximum power configuration
Water-cooled system power in normal room/hot room -est.			
	12.9 kW/14.1 kW	17.4 kW/19.0 kW	24.7 kW/26.3 kW
Inlet air temperature Heat to water and as % of total system heat load			
18°C (64°F)	7.3 kW (57%)	9.8 kW (56%)	12.6 kW (51%)
23°C (73°F)	9.5 kW (74%)	12.6 kW (72%)	15.6 kW (63%)

Temperature	Three book typical configuration	Four book typical configuration	Four book maximum power configuration
27°C (80.6°F)	11.5 kW (89%)	14.8 kW (85%)	18.0 kW (73%)
32°C (89.6°F) (hot room)	14.8 kW (105%)	18.2 kW (96%)	21.6 kW (82%)

2.9.9 Backup air cooling system

The zEC12 has a backup air cooling system that is designed to lower the power consumption of MCM. It works on both air-cooled models and water-cooled models.

In the zEC12, the radiator water cooling system of the air-cooled models or the WCU water cooling system of the water-cooled models is the primary cooling source of MCM. If the water cooling system cannot provide enough cooling capacity to maintain MCM temperature within a normal range, two backup blowers are switched on to provide additional air cooling. At the same time, the oscillator card is set to a slower cycle time. This process slows the system down to allow the degraded cooling capacity to maintain the correct temperature range of the MCM. Running at a slower clock speed, the MCM produces less heat. The slowdown process is done in stages that are based on the temperature of the MCM.

2.10 Summary of zEC12 structure

Table 2-17 summarizes all aspects of the zEC12 structure.

Table 2-17 System structure summary

Description	Model H20	Model H43	Model H66	Model H89	Model HA1
Number of MCMs	1	2	3	4	4
Total number of PUs	27	54	81	108	120
Maximum number of characterized PUs	20	43	66	89	101
Number of CPs	0 - 20	0 - 43	0 - 66	0 - 89	0 - 101
Number of IFLs	0 - 20	0 - 43	0 - 66	0 - 89	0 - 101
Number of ICFs	0 - 20	0 - 43	0 - 66	0 - 89	0 - 101
Number of zAAPs	0 - 10	0 - 21	0 - 33	0 - 44	0 - 50
Number of zIIPs	0 - 10	0 - 21	0 - 33	0 - 44	0 - 50
Standard SAPs	4	8	12	16	16
Additional SAPs	0 - 4	0 - 8	0 - 12	0 - 16	0 - 16
Standard spare PUs	2	2	2	2	2
Enabled memory sizes	32 - 704 GB	32 - 11392 GB	32 - 2272 GB	32 - 3040 GB	32 - 3040 GB
L1 cache per PU	64-I/96-D KB	64-I/96-D KB	64-I/96-D KB	64-I/96-D KB	64-I/96-D KB
L2 cache per PU	1-I/1-D MB	1-I/1-D MB	1-I/1-D MB	1-I/1-D MB	1-I/1-D MB

Description	Model H20	Model H43	Model H66	Model H89	Model HA1
L3 shared cache per PU chip	48 MB	48 MB	48 MB	48 MB	48 MB
L4 shared cache	384 MB	384 MB	384 MB	384 MB	384 MB
Cycle time (ns)	0.178	0.178	0.178	0.178	0.178
Clock frequency	5.5 GHz	5.5 GHz	5.5 GHz	5.5 GHz	5.5 GHz
Maximum number of fanouts	8	16	20	24	24
I/O interface per IFB cable	6 GBps	6 GBps	6 GBps	6 GBps	6 GBps
I/O interface per PCIe cable	8 GBps	8 GBps	8 GBps	8 GBps	8 GBps
Number of support elements	2	2	2	2	2
External AC power	3 phase	3 phase	3 phase	3 phase	3 phase
Optional external DC	520 V/380 V	520 V/380 V	520 V/380 V	520 V/380 V	520 V/380 V
Internal Battery Feature	Optional	Optional	Optional	Optional	Optional



Central processor complex system design

This chapter explains how the **IBM zEnterprise EC12** (zEC12) processor unit is designed. This information can be used to understand the functions that make the zEC12 a system that suits a broad mix of workloads for large enterprises.

This chapter includes the following sections:

- ▶ Overview
- ▶ Design highlights
- ▶ Book design
- ▶ Processor unit design
- ▶ Processor unit functions
- ▶ Memory design
- ▶ Logical partitioning
- ▶ Intelligent Resource Director (IRD)
- ▶ Clustering technology

3.1 Overview

The design of the zEC12 symmetric multiprocessor (SMP) is the next step in an evolutionary trajectory that began with the introduction of CMOS technology back in 1994. Over time, the design was adapted to the changing requirements dictated by the shift toward new types of applications on which clients depend.

The zEC12 offers high levels of serviceability, availability, reliability, resilience, and security. It fits in the IBM strategy in which mainframes play a central role in creating an intelligent, energy efficient, and integrated infrastructure. The zEC12 is designed so that everything around it, such as operating systems, middleware, storage, security, and network technologies that support open standards, help you achieve your business goals.

The modular book design aims to reduce, or in some cases even eliminate, planned and unplanned outages. The design does so by offering concurrent repair, replace, and upgrade functions for processors, memory, and I/O. For more information about the zEC12 serviceability, availability, and reliability (RAS) features, see Chapter 10, “Reliability, availability, and serviceability” on page 385.

The zEC12 has ultra-high frequency, large high speed buffers (caches) and memory, superscalar processor design, out-of-order core execution, and flexible configuration options. It is the next implementation to address the ever-changing IT environment.

3.2 Design highlights

The physical packaging of the zEC12 is comparable to the packaging that is used for zEnterprise 196 (System z196) systems. Its modular book design addresses the increasing costs that are related to building systems with ever-increasing capacities. The modular book design is flexible and expandable, offering unprecedented capacity to meet consolidation needs, and might contain even larger capacities in the future.

zEC12 continues 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 zEC12 system design has the following main objectives:

- ▶ Offer a *flexible infrastructure* to concurrently accommodate a wide range of operating systems and applications. These range from the traditional systems (for example, z/OS and z/VM) to the world of Linux and cloud computing.
- ▶ Offer state-of-the-art *integration* capability for server consolidation by using virtualization capabilities:
 - Logical partitioning, which allows 60 independent logical servers
 - z/VM, which can virtualize hundreds to thousands of servers as independently running virtual machines
 - HiperSockets, which implement virtual LANs between logical partitions (LPARs) within the system

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

- ▶ Offer *high performance* to achieve the outstanding response times required by new workload-type applications. This performance is achieved by high frequency, superscalar processor technology; improved out-of-order core execution; large high-speed buffers (cache) and memory; architecture; and high-bandwidth channels. This configuration offers second-to-none data rate connectivity.
- ▶ Offer the *high capacity* and *scalability* that are required by the most demanding applications, both from 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 broad internal and external *connectivity* offerings, supporting open standards, such as Gigabit Ethernet (GbE), and Fibre Channel Protocol (FCP).
- ▶ Provide leading cryptographic performance. Every processor unit (PU) has a dedicated CP Assist for Cryptographic Function (CPACF). Optional Crypto Express features with Cryptographic Coprocessors provide the highest standardized security certification¹. 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 zEC12 system structure, showing a logical representation of the data flow from PUs, caches, memory cards, and various interconnect capabilities.

3.3 Book design

A zEC12 system can have up to four books in a fully connected topology, up to 101 PUs that can be characterized, and up to 3 TB of memory capacity. The topology is shown in Figure 3-3 on page 82. Memory has up to 12 memory controllers, using 5-channel redundant array of independent memory (RAIM) protection, with dual inline memory modules (DIMM) bus cyclic redundancy check (CRC) error try again. 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 made by IBM has negated that. In addition, eDRAM offers higher density, less power utilization, fewer soft errors, and better performance. Concurrent maintenance allows dynamic book add and repair.

The zEC12 uses 32 nm chip technology, with advanced low latency pipeline design, creating high-speed yet power-efficient circuit designs. The multiple chip module (MCM) has a dense packaging, allowing closed water loop cooling. The heat exchange from the closed loop is either air cooled by a radiator unit or, optionally, water cooled (WCU). The water cooling option can lower the total power consumption of the system. This benefit is particularly true for the larger configurations. For more information, see 2.9.6, “Cooling” on page 69.

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

3.3.1 Cache levels and memory structure

The zEC12 memory subsystem focuses on keeping data “closer” to the PU. With the current processor configuration, all cache levels beginning from L2 have increased, and chip-level shared cache (L3) and book-level shared cache (L4) have doubled in size.

Figure 3-1 shows the zEC12 cache levels and memory hierarchy.

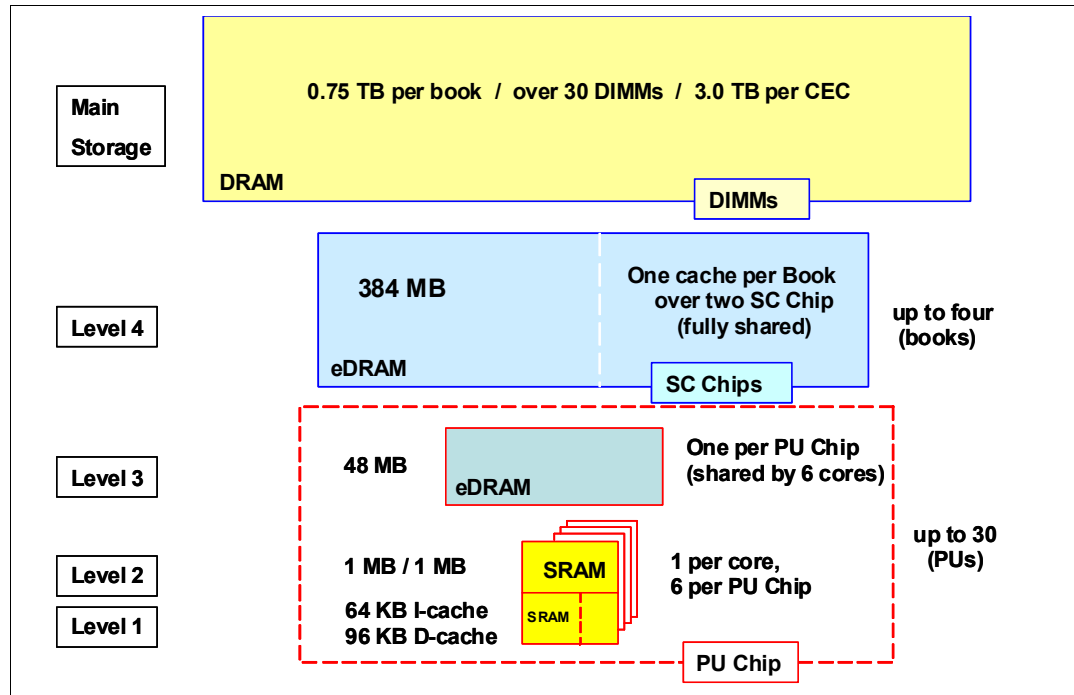


Figure 3-1 zEC12 cache levels and memory hierarchy

The 4-level cache structure is implemented within the MCM. The first three levels (L1, L2, and L3) are on each PU chip, and the last level (L4) is on the SC chips:

- ▶ L1 and L2 caches use static random access memory (SRAM), and are private for each core.
- ▶ L3 cache uses embedded dynamic static random access memory (eDRAM) and is shared by all six cores within the PU chip. Each book has six L3 caches. A four-book system therefore has 24 of them, resulting in 1152 MB (48 x 24 MB) of this shared PU chip-level cache.
- ▶ L4 cache also uses eDRAM, and is shared by all PU chips on the MCM. A four-book system has 1536 MB (4 x 384 MB) of shared L4 cache.
- ▶ Central storage has up to 0.75 TB per book, using up to 30 DIMMs. A four-book system can have up to 3 TB of central 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, avoiding increased latency.

Also, the distance to remote caches as seen from the microprocessor becomes a significant factor. An example is an L4 cache that is not on the microprocessor (and might not even be in the same book). Although the L4 cache is rather large, the reduced cycle time means that more cycles are needed to travel the same distance.

To avoid this potential latency, zEC12 uses two more cache levels (L2 and L3) within the PU chip, with denser packaging. This design reduces traffic to and from the shared L4 cache, which is on another chip (SC chip). Only when there is a cache miss in L1, L2, or L3 is the request sent to L4. L4 is the coherence manager, meaning that all memory fetches must be in the L4 cache before that data can be used by the processor.

Another approach is available for avoiding L4 cache access delays (latency). The L4 cache straddles up to four books. This configuration means that relatively large distances exist between the higher-level caches in the processors and the L4 cache content. To overcome the delays inherent to the book 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 given LPAR workload to the processors in the same book as the L4 cache. This configuration is achieved by having the IBM Processor Resource/Systems Manager (PR/SM) scheduler and the z/OS 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 book boundary) without affecting throughput and response times.

Figure 3-2 compares the cache structures of the zEC12 with the System z previous generation system, the z196.

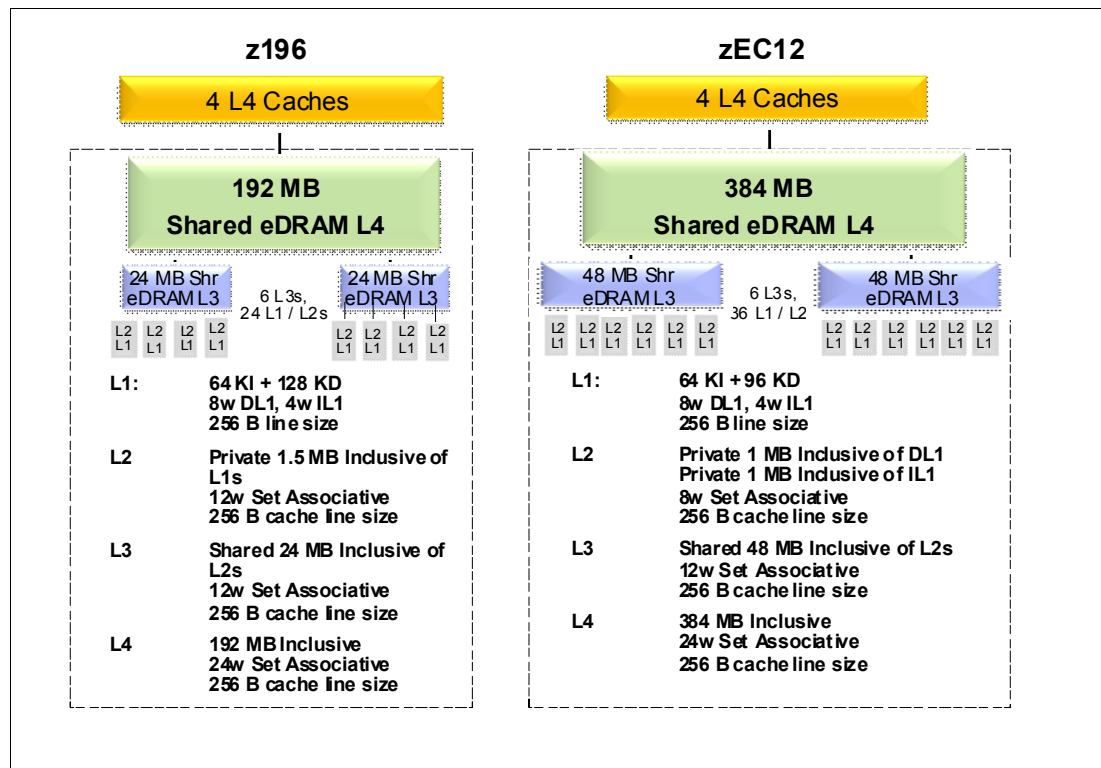


Figure 3-2 zEC12 and z196 cache level comparison

Compared to z196, the zEC12 cache design has much larger cache level sizes, except for the L1 private cache on each core. The access time of the private cache usually occurs in one cycle. The zEC12 cache level structure is focused on keeping more data closer to the PU. This design can improve system performance on many production workloads.

HiperDispatch

Preventing PR/SM and the dispatcher from scheduling and dispatching a workload on any processor available, and keeping the workload in as small a portion of the system as possible, contributes to overcoming latency in a high-frequency processor design, such as the zEC12. The cooperation between z/OS and PR/SM is bundled in a function called *HiperDispatch*. HiperDispatch exploits the zEC12 cache topology, with reduced cross-book “help”, and better locality for multi-task address spaces. For more information about HiperDispatch, see 3.7, “Logical partitioning” on page 108.

3.3.2 Book interconnect topology

Books are interconnected in a point-to-point topology, allowing every book to communicate with every other book. Data transfer never has to go through another book (cache) to address the requested data or control information.

Figure 3-3 shows a simplified topology of a four-book system.

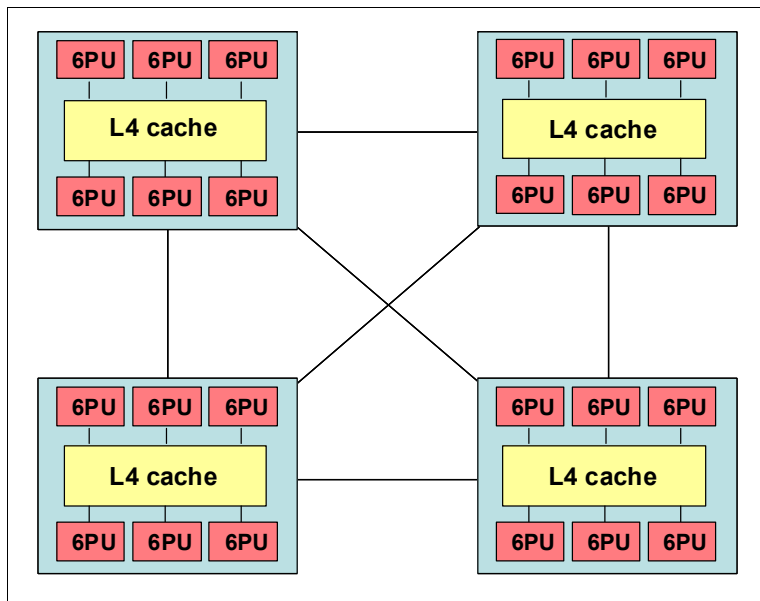


Figure 3-3 Point-to-point topology for book-to-book communication

Inter-book communication takes place at the L4 cache level, which is implemented on Storage Controller (SC) cache chips in each MCM. The SC function regulates coherent book-to-book traffic.

3.4 Processor unit design

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. Processor cycle time is especially important for CPU-intensive applications. The System z10 EC introduced a dramatic PU cycle time improvement. Its succeeding generation, the z196, lowers it even further, reaching 0.192 nanoseconds (5.2 GHz). The zEC12 improves even this industry leading number to 0.178 ns (5.5 GHz).

In addition to the cycle time, other processor design aspects, such as pipeline, execution order, branch prediction, and high-speed buffers (caches), also greatly contribute to the performance of the system. The zEC12 processor unit core is a superscalar, out-of-order (OOO) processor. It has six execution units where, for instructions that are not directly run by the hardware, some instructions are run by millicode, and others are split into multiple operations.

zEC12 introduces architectural extensions with instructions designed to allow reduced processor quiesce effects, reduced cache misses, and reduced pipeline disruption. The zEC12 new architecture includes the following features:

- ▶ Improvements in branch prediction and handling
- ▶ Performance per watt improvements when compared to the z196 system
- ▶ Numerous improvements in the out-of-order (OOO) design
- ▶ Enhanced instruction dispatch and grouping efficiency
- ▶ Enhanced branch prediction structure and sequential instruction fetching
- ▶ Millicode improvements
- ▶ Transactional execution (TX) facility
- ▶ Runtime instrumentation (RI) facility
- ▶ Enhanced DAT²-2 for 2-GB page support
- ▶ Decimal floating point (DFP) improvements

The zEC12 enhanced instruction set architecture (ISA) includes a set of instructions added to improve compiled code efficiency. These instructions optimize PUs to meet the demands of a wide variety of business workload types without compromising the performance characteristics of traditional workloads.

3.4.1 Out-of-order (OOO) execution

The z196 was the first System z to implement an OOO core. The zEC12 improves this technology by increasing the OOO resources, increasing the execution and completion throughput, and improving the instruction dispatch and grouping efficiency. 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 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.

² Dynamic Address Translation

The zEC12 is the second generation of OOO System z processor design with advanced micro-architectural innovations that provide these benefits:

- ▶ Maximized instruction level parallelism (ILP) for a better cycle per instruction (CPI) design by reviewing every part of the z196 design.
- ▶ Maximized performance per watt. Two cores are added, as compared to the z196 chip, at slightly higher chip power (~300 watts).
- ▶ Enhanced instruction dispatch and grouping efficiency.
- ▶ Increased OOO resources (Global Completion Table entries, physical GPR entries, and physical FPR entries).
- ▶ Improved completion rate.
- ▶ Reduced cache/translation lookaside buffer (TLB) miss penalty.
- ▶ Improved execution of D-Cache store and reload and new Fixed-point divide.
- ▶ New 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 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 zEC12 core is shown in Figure 3-4.

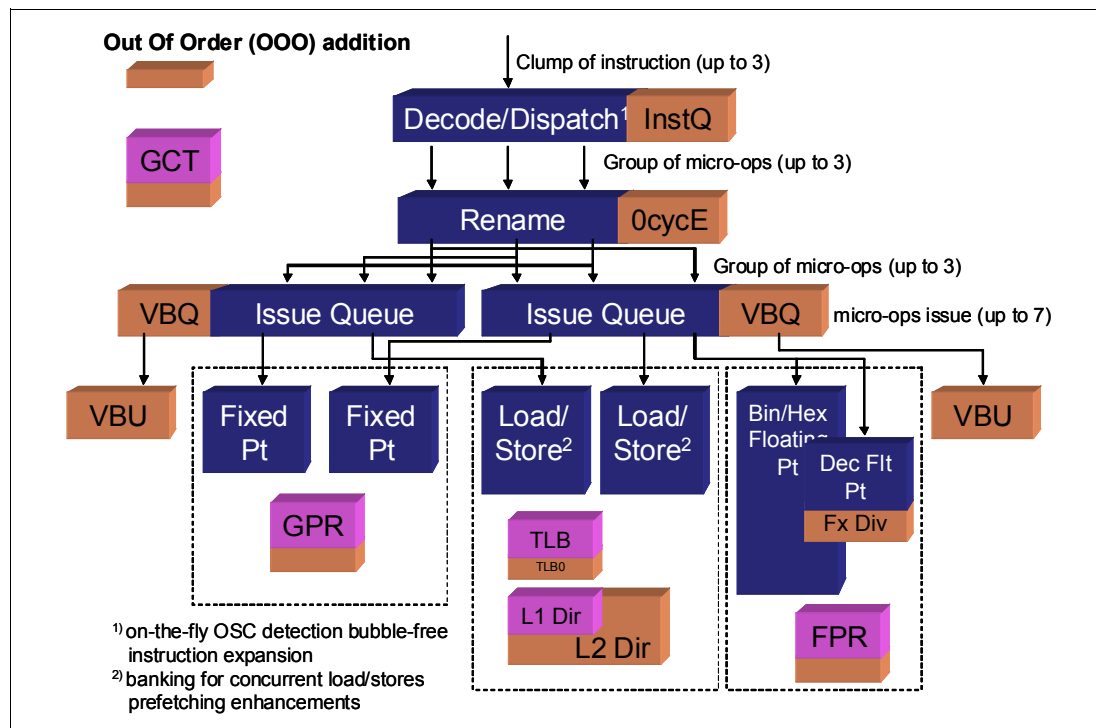


Figure 3-4 zEC12 PU core logical diagram

Memory address generation and memory accesses can occur out of (program) order. This capability can provide a greater exploitation of the zEC12 superscalar core, and can improve system performance. Figure 3-5 shows how OOO core execution can reduce the execution time of a program.

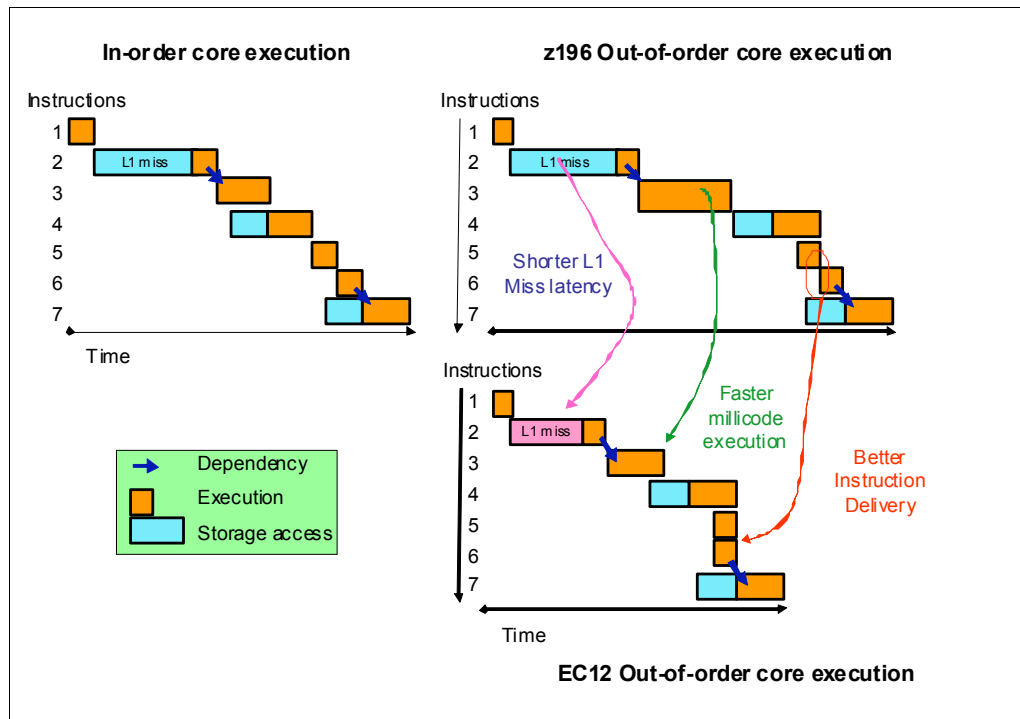


Figure 3-5 In-order and zEC12 out-of-order core execution improvements

The left side of the example 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 the example, instruction 4 can start its storage access and execution while instruction 2 is waiting for data. This situation occurs only if no dependencies exist between both instructions. When the L1 cache miss is solved, instruction 2 can also start its execution 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 zEC12 superscalar PU core can have up to seven instructions/operations in execution per cycle. Compared to the z196, the first IBM System z[®] CPC that used the OOO technology, further enhancements to the execution cycle are integrated in the cores. These improvements result in a shorter execution time.

Example of 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 more costly 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 zEC12 PU core logical diagram in Figure 3-4 on page 84. 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.

zEC12 microprocessor improves the branch prediction structure by increasing the size of the branch buffer (BTB2), which has a faster prediction throughput than BTB1 by using a fast reindexing table (FIT), and improving the sequential instruction stream delivery.

3.4.2 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 execution of instructions by adding more resources onto 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 the zEC12, up to three instructions can be decoded per cycle and up to seven instructions or operations can be in execution per cycle. Execution can occur out of (program) order.

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 to get 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.

3.4.3 Compression and cryptography accelerators on a chip

This section describes the compression and cryptography features.

Coprocessor units

There is one coprocessor (CoP) unit for compression and cryptography on each core in the chip. The compression engine uses static dictionary compression and expansion. The dictionary size is up to 64 KB, with 8 K entries, and has a local 16 KB cache for dictionary data. The cryptography engine is used for CP assist for cryptographic function (CPACF), which offers a set of symmetric cryptographic functions for high encrypting and decrypting performance of clear key operations.

Figure 3-6 shows the location of the coprocessor on the chip.

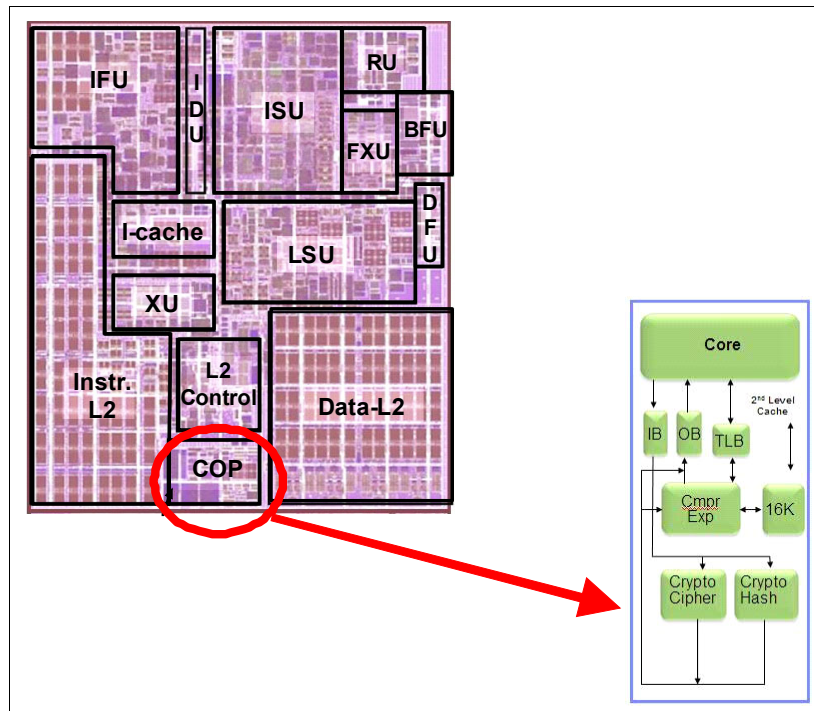


Figure 3-6 Compression and cryptography accelerators on a core in the chip

CP assist for cryptographic function (CPACF)

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 FIPS³ 140-2 level 4 security. The assist function uses a special instruction set for symmetrical clear key cryptographic encryption and decryption, as well as 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 zEC12, see Chapter 6, “Cryptography” on page 199.

3.4.4 Decimal floating point (DFP) accelerator

The decimal floating point (DFP) accelerator function is present on each of the microprocessors (cores) on the hex 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 necessary data conversions and approximation to represent decimal numbers. This 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

³ Federal Information Processing Standards (FIPS)140-2 Security Requirements for Cryptographic Modules

- ▶ Instructions for running decimal floating point computations
- ▶ An instruction that runs data conversions to and from the decimal floating point representation

The DFP architecture on zEC12 is improved to facilitate better performance on traditional zoned-decimal operations for COBOL programs. Additional instructions are provided to convert zoned-decimal data into DFP format in Floating Point Registers (FPRs).

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 standardization of the dominant decimal data and decimal operations in commercial computing supporting industry standardization (IEEE 754R) of decimal floating point operations. Instructions are added in support of the Draft Standard for Floating-Point Arithmetic, which is intended to supersede the ANSI/IEEE Standard 754-1985.
- ▶ Allows COBOL programs that use zoned-decimal operations to take advantage of the zEC12 introduced architecture.

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

3.4.5 IEEE floating point

Binary and hexadecimal floating-point instructions are implemented in zEC12. They incorporate IEEE standards into the system.

The key point is that Java and C/C++ applications tend to use IEEE Binary Floating Point operations more frequently than earlier applications. Therefore, the better the hardware implementation of this set of instructions, the better the performance of applications.

3.4.6 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 architected state in its buffer. The principle is shown in Figure 3-7 on page 89.

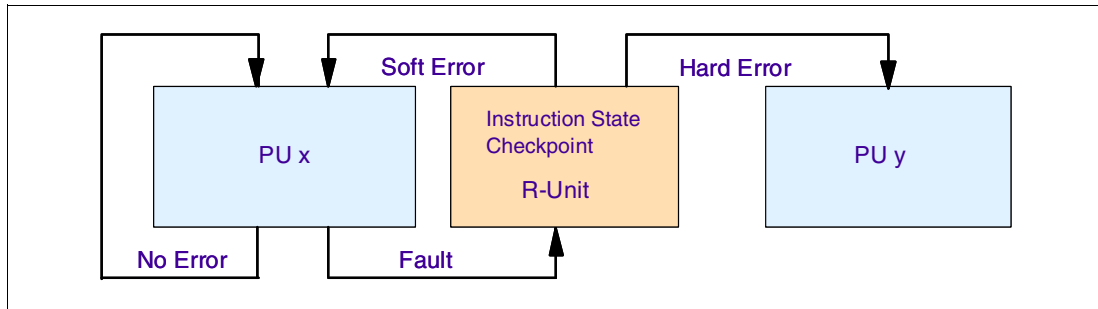


Figure 3-7 PU error detection and recovery

3.4.7 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 combined with other prediction mechanisms, is implemented on each microprocessor.

The branch history table (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, the zEC12 uses various techniques to improve the prediction of the correct branch to be run. The following techniques are included:

- ▶ 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 the zEC12. This is because the architecture rules prescribe that, for successful parallel execution of an instruction stream, the correctly predicted result of the branch is essential.

zEC12 architecture introduces the new instructions BPP/BPRL to allow software to preinstall a future branch and its target into the BTB.

3.4.8 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 from where a wild branch came. It might also collect data from that storage location. This approach decreases the many debugging steps that are necessary when you want to know from where the branch came.

3.4.9 Translation lookaside buffer (TLB)

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 low 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 page support is introduced and can be used when appropriate. For more information, see “Large page support” on page 105.

With the new enhanced DAT-2 (EDAT-2) improvements, zEC12 introduces architecture enhancements to allow support for 2 GB page frames.

3.4.10 Instruction fetching, decoding, and grouping

The superscalar design of the microprocessor allows for the decoding of up to three instructions per cycle and the execution of up to seven 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 three 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 five 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 System z compilers and the JVMs are constantly updated to benefit from new instructions and advances in microprocessor designs.

3.4.11 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 the Extended Translation Facility instructions.

3.4.12 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 Message-Security Assist (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.

3.4.13 Transactional execution (TX)

This capability, 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.

3.4.14 Runtime instrumentation (RI)

Runtime instrumentation is a hardware facility that was introduced with the zEC12 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 right decision by providing real-time feedback on the execution.

3.5 Processor unit functions

This section describes the processor unit (PU) functions.

3.5.1 Overview

All PUs on a zEC12 are physically identical. When the system is initialized, one integrated firmware processor (IFP) is allocated from the pool of PUs available for the whole system, the other PUs can be characterized to specific functions: CP, IFL, ICF, zAAP, 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 zEC12.

This design brings outstanding flexibility to the zEC12 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 8, “Software support” on page 263.

Concurrent upgrades

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

- ▶ Customer Initiated Upgrade (CIU) facility 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 MCMs in the installed books have no available remaining PUs, an upgrade results in a model upgrade and the installation of an additional book. However, there is a limit of four books. Book installation is nondisruptive, but takes more time than a simple LIC upgrade.

For more information about Capacity on Demand, see Chapter 9, “System upgrades” on page 339.

PU sparing

In the rare event of a PU failure, the failed PU’s characterization is dynamically and transparently reassigned to a spare PU. The zEC12 has two spare PUs. PUs that are not characterized on a CPC configuration can also be used as additional spare PUs. For more information about PU sparing, see 3.5.12, “Sparing rules” on page 104.

PU pools

PUs that are defined as CPs, IFLs, ICFs, zIIPs, and zAAPs 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, zAAP, and zIIP weights can be managed separately. For more information, see “PU weighting” on page 93.

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

As an example, consider a zEC12 with 10 CPs, three zAAPs, two IFLs, two zIIPs, and one ICF. This system has a PU pool of 18 PUs, called the *pool width*. Subdivision defines these pools:

- ▶ A CP pool of 10 CPs
- ▶ An ICF pool of one ICF
- ▶ An IFL pool of two IFLs
- ▶ A zAAP pool of three zAAPs

- ▶ A zIIP pool of two zIIPs

PU are placed in the pools in the following circumstances:

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

PUs 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 and the conversion of a PU. When a dedicated LPAR is activated, its PUs are taken from the correct pools. This is also the case when an LPAR logically configures a PU on, if the width of the pool allows.

By having different pools, a weight distinction can be made between CPs, zAAPs, and zIIPs. On earlier systems, specialty engines, such as zAAPs, automatically received the weight of the initial CP.

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

PU weighting

Because zAAPs, zIIPs, IFLs, and ICFs have their own pools from where they are dispatched, they can be given their own weights. For more information about PU pools and processing weights, see the *zEnterprise System Processor Resource/Systems Manager Planning Guide*, SB10-7156.

3.5.2 Central processors

A central processor (CP) is a PU 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 101 PUs can be characterized as CPs, depending on the configuration.

The zEC12 can be initialized only in LPAR mode. 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 101 logical CPs that are defined (the sum of active and reserved logical CPs). In practice, define no more CPs than the operating system supports. For example, the z/OS LPAR supports a maximum of 99 logical CPs.

All PUs 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.

The zEC12 recognizes four distinct capacity settings for CPs. Full-capacity CPs are identified as CP7. In addition to full-capacity CPs, three subcapacity settings (CP6, CP5, and CP4), each for up to 20 CPs, are offered.

The four capacity settings appear in hardware descriptions:

- ▶ CP7 feature code 1908
- ▶ CP6 feature code 1907
- ▶ CP5 feature code 1906
- ▶ CP4 feature code 1905

Granular capacity adds 60 subcapacity settings to the 101 capacity settings that are available with full capacity CPs (CP7). Each of the 60 subcapacity settings applies only to up to 20 CPs, independently of the model installed.

Information about CPs in the remainder of this chapter applies to all CP capacity settings, unless indicated otherwise. For more information about granular capacity, see 2.8, “Model configurations” on page 59.

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, and IBM zAware. Up to 101 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 can be shared by multiple Linux guests, z/VM LPARs, or IBM zAware running on the same zEC12. Only z/VM, Linux on System z Operating Systems, IBM zAware, and designated software products can run on IFLs. IFLs are orderable by using FC 1909.

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 the zEC12. Software product license charges 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 1914). When the system is later upgraded with an additional IFL, the system recognizes that an IFL was already 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 up to four books, up to 101 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 1910.

ICFs exclusively run CFCC. ICFs do not change the model capacity identifier of the zEC12. Software product license charges 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 HMC workplace.

The ICFs can be used only 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)), System z 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 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, therefore allowing 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 simply expedites the re-dispatching of the partition.

LPAR presents these coupling thin interrupts to the guest partition, so CFCC and z/OS both require interrupt handler support that is capable of dealing with them. CFCC also changed to give up control of the processor as soon as all available pending work is exhausted, or when the LPAR un-dispatches it off the shared processor, whichever comes first.

Coupling facility combinations

A coupling facility image can have one of the following combinations 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 desirable production configuration. It is preferable for a production CF to operate by using dedicated ICFs. With CFCC Level 19 and Coupling Thin Interrupts, you can experience CF response time improvements or more consistent CF response time when using coupling facilities with shared engines, while dedicated engines continue to be recommended to obtain the best coupling facility performance.

In Figure 3-8 on page 96, the CPC on the left has two environments defined (production and test), each has one z/OS and one coupling facility image. The coupling facility images share an ICF.

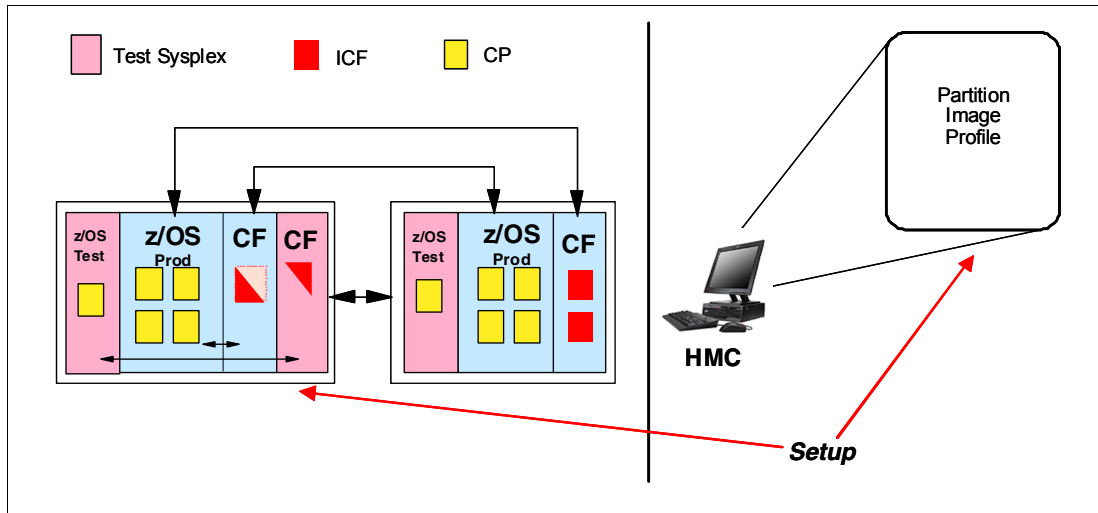


Figure 3-8 ICF options: Shared ICFs

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 image to protect the production environment.

Connections between these z/OS and coupling facility images can use internal coupling links (ICs) 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 introduces 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 3.9.2, “Dynamic CF dispatching” on page 124.

3.5.5 System z Application Assist Processors (zAAPs)

A *zAAP* reduces the standard processor (CP) capacity requirements for z/OS Java or XML system services applications, freeing up capacity for other workload requirements. zAAPs do not increase the millions of service units (MSU) value of the processor, and therefore do not affect the IBM software license charges.

The zAAP is a PU that is used for running IBM designated z/OS workloads, such as Java or z/OS XML System Services. IBM SDK for z/OS Java 2 Technology Edition, in cooperation with z/OS dispatcher, directs Java virtual machine (JVM) processing from CPs to zAAPs. Also, z/OS XML parsing that is performed in the TCB mode is eligible to be run on the zAAP processors.

zAAPs 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, 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.

One CP must be installed with or before you install a zAAP. The number of zAAPs in a CPC cannot exceed the number of purchased CPs. Within the capacity of the sum of all unassigned PUs in up to four books, up to 50 zAAPs on a model HA1 can be characterized. Table 3-1 shows the allowed number of zAAPs for each model.

Table 3-1 Number of zAAPs per model

Model	H20	H43	H66	H89	HA1
zAAPs	0 - 10	0 - 21	0 - 33	0 - 44	0 - 50

The number of permanent zAAPs plus temporary zAAPs cannot exceed the number of purchased (permanent plus unassigned) CPs plus temporary CPs. Also, the number of temporary zAAPs cannot exceed the number of permanent zAAPs.

PUs characterized as zAAPs within a configuration are grouped into the zAAP pool. This configuration allows zAAPs to have their own processing weights, independent of the weight of parent CPs. The zAAP pool can be seen on the hardware console.

zAAPs are orderable by using FC 1912. Up to one zAAP can be ordered for each CP or marked CP-configured in the CPC.

zAAPs and logical partition definitions

zAAPs are either dedicated or shared, depending on whether they are part of a logical partition (LPAR) with dedicated or shared CPs. In an LPAR, you must have at least one CP to be able to define zAAPs for that partition. You can define as many zAAPs for an LPAR as are available in the system.

Logical partition: A zEC12 cannot have more zAAPs than CPs. However, in an LPAR, as many zAAPs as are available can be defined together with at least one CP.

How zAAPs work

zAAPs are designed for supporting designated z/OS workloads. The initial user was Java code execution. When Java code must be run (for example, under control of WebSphere), the z/OS JVM calls the function of the zAAP. The z/OS dispatcher then suspends the JVM task on the CP that it is running on and dispatches it on an available zAAP. After the Java application code execution is finished, z/OS redispaches the JVM task on 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 3-9 shows the logical flow of Java code running on a zEC12 that has a zAAP available. When JVM starts the execution of a Java program, it passes control to the z/OS dispatcher that verifies the availability of a zAAP.

The availability is treated in the following manner:

- ▶ If a zAAP is available (not busy), the dispatcher suspends the JVM task on the CP, and assigns the Java task to the zAAP. 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 zAAP 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.

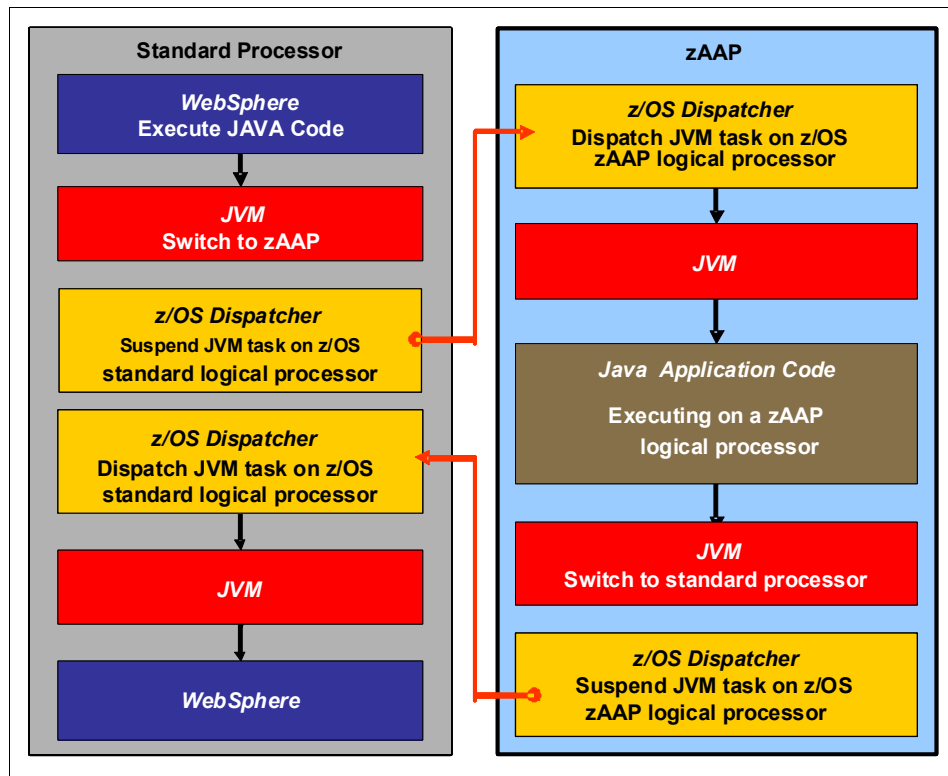


Figure 3-9 Logical flow of Java code execution on a zAAP

A zAAP 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 zAAP is not able to 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 zAAP. The installation can manage the use of CPs so that Java application code runs only on CPs, only on zAAPs, or on both.

Three execution options for Java 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 (IFAHONORPRIORITY=YES): This is the default option, and specifies that CPs must not automatically consider zAAP-eligible work for dispatching on them. The zAAP-eligible work is dispatched on the zAAP engines until Workload Manager (WLM) considers that the zAAPs are overcommitted. WLM then requests help from the CPs. When help is requested, the CPs consider dispatching zAAP-eligible work on the CPs themselves based on the dispatching priority relative to other workloads. When the zAAP engines are no longer overcommitted, the CPs stop considering zAAP-eligible work for dispatch.

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

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

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

- ▶ Option 3: Java discretionary crossover (IFACROSSOVER=YES or NO): As of z/OS V1R8 (and the IBM zIIP support for the z/OS V1R7 web deliverable), the IFACROSSOVER parameter is no longer honored.

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

3.5.6 System z Integrated Information Processor (zIIP)

A zIIP enables eligible z/OS workloads to have a portion of the workload's enclave service request block (SRB) work directed to the zIIP. The zIIPs do not increase the MSU value of the processor, and therefore do not affect the IBM software license charges.

z/OS Communications Server and DB2 UDB for z/OS Version 8 or later use the zIIP by indicating to z/OS which portions of the work are eligible to be routed to a zIIP.

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, therefore reducing the need for more gateway products that can affect performance and availability. The application uses the DRDA requestor 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 up 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 zEC12, 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 is eligible to run on the zIIP. This configuration requires z/OS V1R10 or later releases.
- ▶ The first IBM exploiter 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, as well as 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-03.ibm.com/systems/z/hardware/features/ziip/about.html>

zIIP installation information

One CP must be installed with or before any zIIP is installed. The number of zIIPs in a system cannot exceed the number of CPs and unassigned CPs in that system. Within the capacity of the sum of all unassigned PUs in up to four books, up to 66 zIIPs on a model HA1 can be characterized. Table 3-2 shows the allowed number of zIIPs for each model.

Table 3-2 Number of zIIPs per model

Model	H20	H43	H66	H89	HA1
Maximum zIIPs	0 - 13	0 - 28	0 - 44	0 - 59	0 - 66

zIIPs are orderable by using FC 1913. Up to one zIIP can be ordered for each CP or marked CP configured in the system. If the installed books have no remaining unassigned PUs, the assignment of the next zIIP might require the installation of an additional book.

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 a logical partition (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 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.

3.5.7 zAAP on zIIP capability

zAAPs and zIIPs support different types of workloads. However, there are installations that do not have enough eligible workloads to justify buying a zAAP or a zIIP. IBM now offers the capability to combine zAAP and zIIP workloads on zIIP processors, if no zAAPs are installed on the system. This combination can provide the following benefits:

- ▶ The combined eligible workloads can make the zIIP acquisition more cost-effective.
- ▶ When zIIPs are already present, the investment is maximized by running the Java and z/OS XML System Services-based workloads on existing zIIPs.

This capability does not eliminate the need to have one CP for every two zIIP processors in the system. Support is provided by z/OS. For more information, see 8.3.2, “zAAP support” on page 279.

When zAAPs are present, this capability is not available, because it is not intended as a replacement for zAAPs or as an overflow possibility for zAAPs. The zAAP on zIIP capability is available to z/OS when running as a guest of z/VM on systems with zAAPs installed, if no zAAPs are defined to the z/VM LPAR. This configuration allows, for instance, testing this capability to estimate usage before you commit to production. Do not convert zAAPs to zIIPs to take advantage of the zAAP to zIIP capability for the following reasons:

- ▶ Having both zAAPs and zIIPs maximizes the system potential for new workloads.
- ▶ zAAPs were announced on April 7, 2004. Having been available for so many years, there might be applications or middleware with zAAP-specific code dependencies. For example, the code might use the number of installed zAAP engines to optimize multithreading performance.

It is a good idea to plan and test before eliminating all zAAPs because there can be application code dependencies that might affect performance.

Statement of Direction: IBM zEnterprise EC12 and IBM zEnterprise BC12 are planned to be the last high-end System z servers to offer support for zAAP specialty engine processors. IBM intends to continue support for running zAAP workloads on zIIP processors (“zAAP on zIIP”). This change is intended to help simplify capacity planning and performance management, while still supporting all the currently eligible workloads. In addition, IBM provides a program temporary fix (PTF) for authorized program analysis report (APAR) OA38829 on V1.13. This PTF removes the restriction that prevents zAAP-eligible workloads from running on zIIP processors when a zAAP is installed on the server. This change is intended only to help facilitate migration and testing of zAAP workloads on zIIP processors.

3.5.8 System assist processors (SAPs)

An 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 zEC12 model, as shown in Table 3-3.

Table 3-3 SAPs per model

Model	H20	H43	H66	H89	HA1
Standard SAPs	4	8	12	16	16

SAP configuration

A standard SAP configuration provides a well-balanced system for most environments. However, there are application environments with high I/O rates (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 zEC12 systems, the number of SAPs can be greater than the number of CPs.

Optional additional orderable SAPs

The ability to order more SAPs is an option available on all models (FC 1911). These additional SAPs increase the capacity of the channel subsystem to run I/O operations, usually 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 3-4.

Table 3-4 Optional SAPs per model

Model	H20	H43	H66	H89	HA1
Optional SAPs	0 - 4	0 - 8	0 - 12	0 - 16	0 - 16

3.5.9 Reserved processors

Reserved processors are defined by the Processor Resource/Systems Manager (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 to an LPAR more logical processors than the number of available CPs, IFLs, ICFs, zAAPs, and zIIPs in the configuration. This process makes it possible to configure online, nondisruptively, more logical processors after additional CPs, IFLs, ICFs, zAAPs, 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 101. 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 definition, see 3.7, “Logical partitioning” on page 108.

3.5.10 Integrated firmware processor

An *integrated firmware processor* (IFP) is allocated from the pool of PUs 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 for the purpose of 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 (RG) Licensed Internal Code (LIC) to provide native PCIe I/O feature management and virtualization functions. For more information, see Appendix F, “Native Peripheral Component Interconnect Express (PCIe)” on page 529.

3.5.11 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 book boundaries to optimize shared cache usage.

The PU assignment is based on book plug ordering. This process defines the low-order and the high-order books:

- ▶ Book 0: Plug order 4 (when plugged, this is the low-order book)
- ▶ Book 1: Plug order 1 (when Book 0 is not plugged, this is the low-order book)
- ▶ Book 2: Plug order 3
- ▶ Book 3: Plug order 2

Assignment rules isolate the PUs as much as possible on different books and even on different PU chips. This configuration ensures that the operating systems do not use the same shared caches. For example, CPs, zAAPs, 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 LPAR to improve system performance. For more information, see “LPAR dynamic PU reassignment” on page 113.

When an additional book is added concurrently after POR and new LPARs are activated, or processor capacity for active partitions is dynamically expanded, the additional PU capacity can be assigned from the new book. The processor unit assignment rules consider the newly installed book only after the next POR.

3.5.12 Sparing rules

On a zEC12 system, two PUs are reserved as spares. The reserved spares are available to replace any two characterized PUs, whether they are CP, IFL, ICF, zAAP, 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 MCM to be replaced (referred to as a *pending repair*) can still be upgraded when sufficient PUs are available.

Transparent CP, IFL, ICF, zAAP, zIIP, SAP, and IFP sparing

Depending on the model, sparing of CP, IFL, ICF, zAAP, zIIP, SAP, and IFP is transparent and does not require an 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, zAAP, 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 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 in the 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.

3.5.13 Increased flexibility with z/VM mode partitions

zEC12 provides 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, zAAPs, 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 System z on IFLs, z/VSE, and z/OS on CPs
- ▶ Run designated z/OS workloads, such as parts of DB2 DRDA processing and XML, on zIIPs
- ▶ Provide an economical Java execution environment under z/OS on zAAPs

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 about the zEC12 memory design.

3.6.1 Overview

The zEC12 memory design also provides flexibility, high availability, and upgrades:

- ▶ Concurrent memory upgrades (if the physically installed capacity is not yet reached): The zEC12 can have more physically installed memory 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.
- ▶ Concurrent memory upgrades (if the physically installed capacity is reached): Physical memory upgrades require a book to be removed and reinstalled after replacing the memory cards in the book. Except for a model H20, the combination of enhanced book availability and the flexible memory option allows you to concurrently add memory to the system. For more information, see 2.5.5, “Book replacement and memory” on page 51, and 2.5.6, “Flexible Memory Option” on page 51.

When the total capacity installed has more usable memory than required for a configuration, the LIC configuration control (LICCC) determines how much memory is used from each book. The sum of the LICCC provided memory from each book is the amount available for use in the system.

Memory allocation

Memory assignment or allocation is done at POR when the system is initialized. PR/SM is responsible for the memory assignments.

PR/SM knows the amount of purchased memory and how it relates to the available physical memory in each of the installed books. PR/SM has control over all physical memory, and therefore is able to make physical memory available to the configuration when a book is nondisruptively added.

PR/SM also controls the reassignment of the content of a specific physical memory array in one book to a memory array in another book. This is the *memory copy/reassign* function, and is used to reallocate the memory content from the memory in a book to another memory location. It is used when *enhanced book availability* (EBA) is applied to concurrently remove and reinstall a book in an upgrade or repair action.

Because of the memory allocation algorithm, systems that undergo a number of miscellaneous equipment specification (MES) upgrades for memory can have various memory mixes in all books of the system. If, however unlikely, memory fails, it is technically feasible to POR the system with the remaining memory resources. After POR, the memory distribution across the books is now different, and so is the amount of available memory.

Large page support

By default, page frames are allocated with a 4 KB size. The zEC12 also supports 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 System z support for 1 MB pages is available in SUSE Linux Enterprise Server (SLES) 10 SP2 and Red Hat 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 when there is a TLB miss and both the segment table (needed to find the page table) and the page table (needed to find the entry for the particular page in question) are not in cache. In this case, there are 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 desirable 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 1 TLB line. Therefore, large page size exploiters 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.

Exploiters of large pages are better represented in the TLB and are expected to see performance improvement 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.

Very 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 needs to use large pages as a result of the system-level costs of tying up real storage. There is a balance between the performance of a process using large pages, and the performance of the remaining work on the system.

On zEC12, 1 MB large pages become pageable if Flash Express is enabled. They are only available for 64-bit virtual private storage, such as virtual memory located 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's contents. Correct sizing of the TLB is subject to complex statistical modeling to find the optimal trade-off between size and performance.

3.6.2 Central storage (CS)

CS consists of main storage, which is addressable by programs, and storage that is not directly addressable by programs. Non-addressable storage includes the hardware system area (HSA).

Central 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

Central storage can be accessed by all processors, but cannot be shared between LPARs. Any system image (LPAR) must have a central storage size defined. This defined central storage is allocated exclusively to the LPAR during partition activation.

3.6.3 Expanded storage

Expanded storage can optionally be defined on zEC12. Expanded storage is physically a section of processor storage. It is controlled by the operating system and transfers 4-KB pages to and from central 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 central 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: Expanded storage is not recommended for use with z/VM 6.3.

Defining expanded storage to a coupling facility image is not possible. However, any other image type can have expanded storage defined, even if that image runs a 64-bit operating system and does not use expanded storage.

The zEC12 runs only 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 central storage as needed when partitions are activated or deactivated.

LPAR single storage pool

In LPAR mode, storage is not split into central storage and expanded storage at POR. Rather, the storage is placed into a single central storage pool that is dynamically assigned to expanded storage and back to central 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 minus the 32 GB of hardware system area. LPARs are still defined to have central storage and, optionally, expanded storage.

Activation of LPARs and dynamic storage reconfiguration cause the storage to be assigned to the type 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 the zEC12, the HSA has a fixed size of 32 GB and is not part of the purchased memory that you order and install.

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:

- ▶ Four channel subsystems (CSSs)
- ▶ Fifteen LPARs in each CSS for a total of 60 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 (LPARs) is a function that is implemented by the Processor Resource/Systems Manager (PR/SM) on the zEC12. The zEC12 runs only in LPAR mode. Therefore, all system aspects are controlled by PR/SM functions.

PR/SM is aware of the book structure on the zEC12. LPARs, however, 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 the responsibility of PR/SM.

As shown in 3.5.11, “Processor unit assignment” on page 103, the initial PU assignment is done during POR by using rules to optimize cache usage. This is the “physical” step, where CPs, zIIPs, zAAPs, IFLs, ICFs, and SAPs are allocated on books.

When an LPAR is activated, PR/SM builds logical processors and allocates memory for the LPAR.

Memory allocation is spread across all books. This optimization is done by using a round-robin algorithm with three increments per book, to match the number of memory controllers (MCs) per book. This memory allocation design is driven by performance results, also minimizing variability for most workloads.

Logical processors are dispatched by PR/SM on physical processors. The assignment topology that is used by PR/SM to dispatch logical on physical PUs is also based on cache usage optimization.

Book-level assignments are more important because they optimize L4 cache usage. So, logical processors from a given LPAR are packed into a book (or books) as much as possible.

Then, PR/SM optimizes chip assignments within the assigned book (or books) 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 (six) matches the number of z/OS processor affinity queues 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.

HiperDispatch

PR/SM and z/OS work in tandem to more efficiently use processor resources. 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 redispershing units of work to same processor group, keeping processes running near their cached instructions and data, and minimizing transfers of data ownership among processors/books.

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 book level), and dynamically optimizes assignment of logical processors and units of work.

z/OS dispatcher manages multiple queues, called *affinity queues*, with a target number of six 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 a given LPAR workload. So, even if the LPAR is defined with many logical processors, HiperDispatch optimizes this number of processors nearest to the required capacity. The optimal number of processors to be used is kept within a book boundary where possible.

Tip: HiperDispatch is now supported with z/VM 6.3 also.

Logical partitions

PR/SM enables the zEC12 to be initialized for a logically partitioned operation, supporting up to 60 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, zAAPs, 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, zAAP, and zIIP logical processors. All three logical processor types can be defined as either all dedicated or all shared. The zAAP and 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 can be used to achieve the defined goals of this specific partition and of the overall system. The provisioning architecture of the zEC12, described in Chapter 9, “System upgrades” on page 339, 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 consumed by an individual logical partition (LPAR) when a processor unit (PU) is defined as a general-purpose processor (CP) or an Integrated Facility for Linux (IFL) shared across a set of LPARs.

This enhancement is designed to provide a physical capacity limit 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 in support of 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 millions of service units (MSUs) 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 8.12, “Software licensing considerations” on page 332.

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 case, and try to establish a defined capacity that is equal or close to the relative weight.

- **Memory:** Memory, either central storage or expanded storage, must be dedicated to an LPAR. The defined storage must be available during the LPAR activation. Otherwise, the 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 (DSR)” on page 118.

- 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 input/output configuration program (IOCP) or the hardware configuration definition (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.

IOCP 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.

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 3-5 shows the nodes 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 3-5 zEC12 modes of operation

Image mode	PU type	Operating system	Addressing mode
ESA/390	CP and zAAP/zIIP	z/OS z/VM	64-bit
	CP	z/VSE and Linux on System z (64-bit)	64-bit
	CP	Linux on System z (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 System z (64-bit)	64-bit
		z/VM	
		Linux on System z (31-bit)	31-bit
z/VM	CP, IFL, zIIP, zAAP, or ICF	z/VM	64-bit
zAware	IFL or CP	zAware	64-bit

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 IPLed in 31-bit mode 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 information about operating system support, see Chapter 8, “Software support” on page 263.

Logically partitioned mode

The zEC12 runs only in LPAR mode. Each of the 60 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 System z operating system, on dedicated or shared CPs
 - z/OS, on any of the following processor units:
 - Dedicated or shared CPs
 - Dedicated CPs *and* dedicated zAAPs *or* zIIPs
 - Shared CPs *and* shared zAAPs *or* zIIPs

zAAP and zIIP usage: zAAPs and zIIPs can be defined to an ESA/390 mode or z/VM mode image as shown in Table 3-5 on page 111. However, zAAPs and zIIPs are used only by z/OS. Other operating systems cannot use zAAPs or zIIPs even if they are defined to the LPAR. z/VM V5R4 and later provide support for real and virtual zAAPs and zIIPs to guest z/OS systems.

- ▶ 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 System z 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 zAAPs, zIIPs, and ICFs
- ▶ zAware mode to run by loading the zAware code into the LPAR that is defined as either of these types:
 - Dedicated or shared CPs
 - Dedicated or shared IFLs

Table 3-6 on page 113 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 can also have reserved processors defined, allowing nondisruptive LPAR upgrades.

Table 3-6 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 System z	CPs DED or CPs SHR
	CPs <i>and</i> zAAPs <i>or</i> zIIPs	z/OS z/VM (V5R4 and later for guest exploitation)	CPs DED <i>and</i> zAAPs DED, <i>and (or)</i> zIIPs DED <i>or</i> CPs SHR <i>and</i> zAAPs SHR <i>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 System z z/VM	IFLs DED <i>or</i> IFLs SHR, <i>or</i> CPs DED <i>or</i> CPs SHR
z/VM	CPs, IFLs, zAAPs, zIIPs, <i>or</i> ICFs	z/VM (V5R4 and later)	All PUs must be SHR or DED
zAware	IFLs, <i>or</i> CPs	zAware	IFLs DED <i>or</i> IFLs SHR, <i>or</i> CPs DED <i>or</i> CPs SHR

Dynamically adding or deleting a logical partition name

Dynamic add or delete of 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 later be 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 crypto feature to a logical partition

You can preplan the addition of Crypto Express3 or Crypto Express4S features to an LPAR on the crypto page in the image profile by defining the Cryptographic Candidate List, Cryptographic Online List, and 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 System z.

LPAR dynamic PU reassignment

System configuration is enhanced to optimize the PU-to-book assignment of physical processors dynamically. The initial assignment of client usable physical processors to physical books can change dynamically to better suit the actual LPAR configurations that are in use. For more information, see 3.5.11, “Processor unit assignment” on page 103.

Swapping of specialty engines and general processors with each other, with spare PUs, or with both, can occur as the system attempts to compact LPAR configurations into physical configurations that span the least number of books.

LPAR dynamic PU reassignment can swap client processors of different types between books. For example, reassignment can swap an IFL on book 1 with a CP on book 2. Swaps can also occur between PU chips within a book, and can include spare PUs. The goals are to further pack the LPAR on fewer books and also on fewer PU chips, which are based on the zEC12 book 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 zEC12 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.

3.7.2 Storage operations

In the zEC12, memory can be assigned as a combination of central storage and expanded storage, supporting up to 60 LPARs. Expanded storage is only used by the z/VM operating system.

Before you activate an LPAR, central storage (and, optionally, expanded storage) must be defined to the LPAR. All installed storage can be configured as central storage. Each individual LPAR can be defined with a maximum of 1 TB of central storage.

Central storage can be dynamically assigned to expanded storage and back to central storage as needed without a POR. For more information, see “LPAR single storage pool” on page 107.

Memory *cannot* be shared between system images. It is possible to dynamically reallocate 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 (DSR)” on page 118.

Operating systems that run under 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 3-7 shows the zEC12 storage allocation and usage possibilities, depending on the image mode.

Table 3-7 Central storage definition and usage possibilities

Image mode	Architecture mode (addressability)	Maximum central storage		Expanded storage	
		Architecture	zEC12 definition	zEC12 definable	Operating system usage ^a
ESA/390	z/Architecture (64-bit)	16 EB	1 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	256 GB	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
zAware	zAware (64-bit)	16 EB	1 TB	Yes	No

a. z/VM supports the use of expanded storage; however, expanded storage is not recommended for z/VM 6.3.

b. z/VM mode is supported by z/VM V5R4 and later.

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 central storage. However, the current central storage limit for LPARs is 1 TB of central storage. The operating system that runs in z/Architecture mode must be able to support the real storage. Currently, z/OS for example, supports up to 4 TB of real storage (z/OS V1R10 and later releases).

Expanded storage can also be configured to an image running an operating system in z/Architecture mode. However, only z/VM is able to use expanded storage. Any other operating system that runs in z/Architecture mode (such as a z/OS or a Linux on System z image) *does not* address the configured expanded storage. This expanded storage remains configured to this image and is *unused*.

- ▶ **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 central 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 central storage, the central 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 zEC12. Any ESA/390 image can be defined with more than 2 GB of central 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 zEC12 definition limit for LPARs is 1 TB of storage.

CFCC Level 18, which is available for the zEC12 with driver level 12K, 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 exploitation.

For more information, see 3.9.1, “Coupling facility control code (CFCC)” on page 121. Expanded storage cannot be defined for a coupling facility image. Only IBM CFCC can run in the coupling facility mode.

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 System z 64-bit distributions (SUSE SLES 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 System z 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 zAAPs or on zIIPs

zAware mode

In IBM zAware mode, storage addressing is 64-bit for a IBM zAware image that runs IBM System z Advanced Workload Analysis Reporter firmware. This configuration allows for an addressing range up to 16 EB. However, the current zEC12 definition limit for LPARs is 1 TB of storage.

The IBM zAware feature, which is exclusive to zEC12 and zBC12, allows the following capabilities:

- ▶ Helps detect and diagnose unusual behavior of z/OS images in near real time
- ▶ Reduces problem determination time and improves service availability beyond standard z/OS features
- ▶ Provides an easy-to-use graphical user interface with quick drill-down capabilities to view analytical data about z/OS behavior

For more information, see Appendix A, “IBM System z Advanced Workload Analysis Reporter (IBM zAware)” on page 479. Only IBM zAware firmware can run in IBM zAware mode.

3.7.3 Reserved storage

Reserved storage can optionally be 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 central storage and, optionally (if not a coupling facility image), an amount of expanded storage. Both central 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 simply 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 to 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 V5R4 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 (SLES) 11 supports both concurrent add and remove.

3.7.4 Logical partition storage granularity

Granularity of central storage for an LPAR depends on the largest central storage amount that is defined for either initial or reserved central storage, as shown in Table 3-8.

Table 3-8 Logical partition main storage granularity

Logical partition: Largest main storage amount	Logical partition: Central storage granularity
Central storage amount <= 128 GB	256 MB
128 GB < central storage amount <= 256 GB	512 MB
256 GB < central storage amount <= 512 GB	1 GB
512 GB < central storage amount <= 1 TB	2 GB

The granularity applies across all central 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 central storage granularity of both initial and reserved central storage is 256 MB.

Expanded storage granularity is fixed at 256 MB.

LPAR storage granularity information is required for LPAR image setup and for z/OS RSU definition. LPARs are limited to a maximum size of 1 TB of central storage. For z/VM V5R4 and later, the limitation is 256 GB.

3.7.5 LPAR dynamic storage reconfiguration (DSR)

Dynamic storage reconfiguration on the zEC12 allows an operating system running on an LPAR to add (nondisruptively) its reserved storage amount to its configuration. This process can occur only if any 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)

IRD is a zEC12 and System z capability that is used only by z/OS. IRD is a function that optimizes processor and channel resource utilization across LPARs within a single System z system.

This feature extends the concept of goal-oriented resource management. It does so by allowing grouping system images that are resident on the same zEC12 or System z 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 3-10 shows an LPAR cluster. It contains three z/OS images, and one Linux image that is 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.

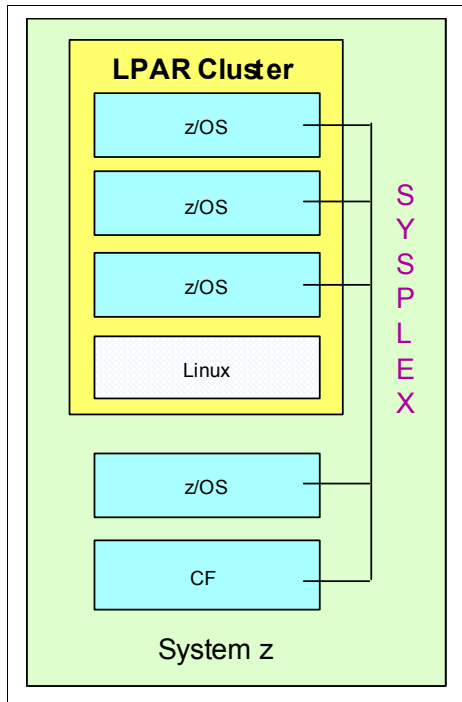


Figure 3-10 IRD LPAR cluster example

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 108.

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. The zEC12 supports DCM within a channel subsystem.

- ▶ Channel subsystem priority queuing: This function on the zEC12 and System z allows 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 continues to be the clustering technology that is used with zEC12. Figure 3-11 illustrates the components of a Parallel Sysplex as implemented within the System z architecture. The figure is intended only as an example. It shows one of many possible Parallel Sysplex configurations.

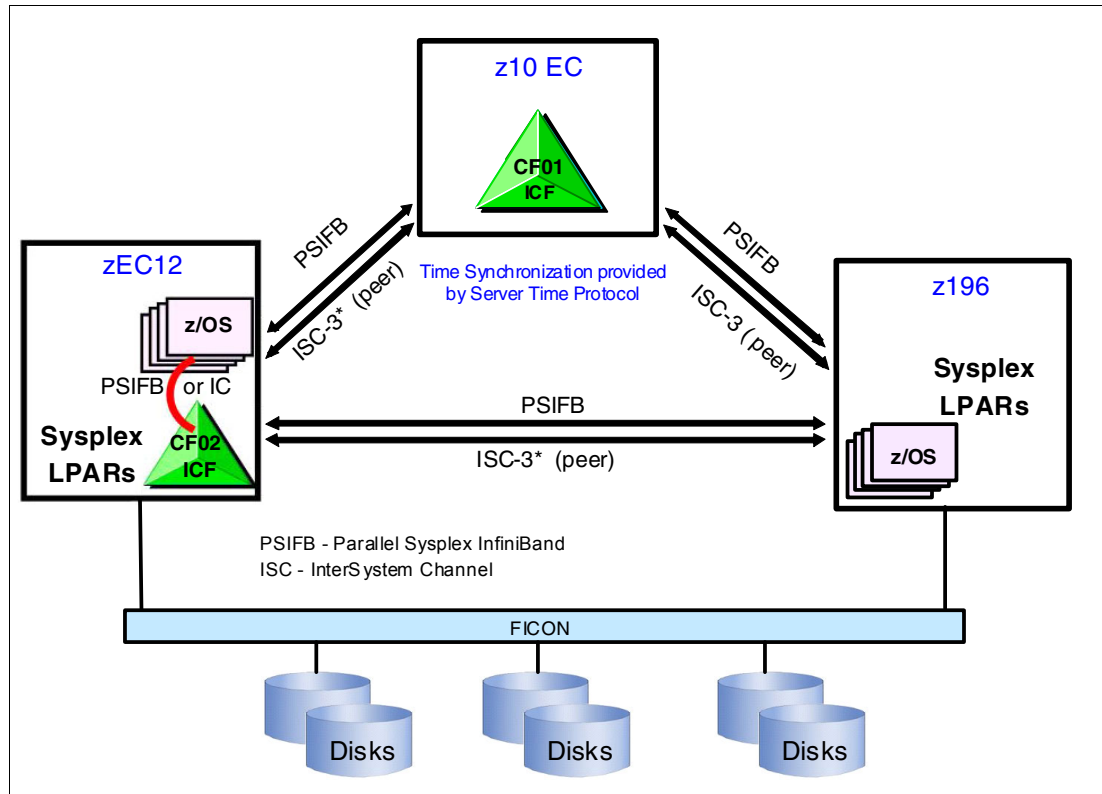


Figure 3-11 Sysplex hardware overview

Figure 3-11 shows a zEC12 system that contains multiple z/OS sysplex partitions. It contains an internal coupling facility (CF02), a z10 EC system that contains a stand-alone CF (CF01), and a z196 that contains multiple z/OS sysplex partitions. STP over coupling links provides time synchronization to all systems. Appropriate CF link technology (1x IFB or 12x IFB) selection depends on the system configuration and how distant they are physically located. ISC-3 links can be carried forward to zEC12 only when they are upgraded from either IBM zEnterprise 196 (z196) or z10 EC. The ICB-4 coupling link is not supported on both zEC12 and zEnterprise CPCs. For more information about link technologies, see 4.10.1, “Coupling links” on page 173.

Parallel Sysplex technology is an enabling technology, allowing highly reliable, redundant, and robust System z technology to achieve near-continuous availability. A Parallel Sysplex comprises 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 from 2 to 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: 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 System z Parallel Sysplex cluster takes the commercial strengths of the platform to improved levels of system management, competitive price for performance, scalable growth, and continuous availability.

3.9.1 Coupling facility control code (CFCC)

CFCC Level 18 and 19 are available on the [zEC12](#).

CFCC Level 18

CFCC Level 18 delivered with Driver level 12 introduces several enhancements in the performance, reporting, and serviceability areas.

Performance improvements

CFCC Level 18 introduces these improvements in cache structure management:

- ▶ “Dynamic structure size alter” is enhanced to improve the performance of changing cache structure size.
- ▶ “DB2 global buffer pool (GBP) write-around (cache bypass)” supports a new conditional write to GBP command. DB2 can use this enhancement during batch update/insert processing to intelligently decide which entries to write to the GBP cache, and which to write around the cache to disk. Before this enhancement, overrunning cache structures with useless directory entries and changed data during batch update/insert jobs (for example, reorganizations) caused several issues. These issues included CF processor usage, thrashing the cache through LRU processing, and cast out processing backlogs and delays.
- ▶ “CF castout class contention avoidance” reduces latch contention with more granular class assignments.
- ▶ “CF storage class contention avoidance” improves response time by changing the latching from a suspend lock to a spin lock.

CF Engine performance is improved by more efficient use of shared-processor CF images with good service times, and latency reduction for asynchronous CF operations and asynchronous CF notifications.

Coupling channel reporting

CFCC Level 18 provides more coupling channel characteristics reporting to z/OS by allowing it to know about the underlying InfiniBand hardware. This change enables RMF to distinguish between Coupling over InfiniBand (CIB) CHPID types (12x IFB, 12x IFB3, and 1x IFB), and to detect whether there is any degradation in performance on CIB channels. RMF uses the changed cross-system extended services (XES) interface and obtains new channel path characteristics. The channel path has these new characteristics:

- ▶ Stored in a new channel path data section of SMF record 74 subtype 4
- ▶ Added to the Subchannel Activity and CF To CF Activity sections of the Resource Measurement Facility (RMF) Postprocessor Coupling Facility Activity report
- ▶ Provided on the Subchannels Details panel of the RMF Monitor III Coupling Facility Systems report

Serviceability enhancements

Serviceability enhancements provide help for debugging in these areas:

- ▶ Additional structure control information in CF memory dumps: Before CFCC Level 18, only CF control structures were dumped and no structure-related controls were included. With CFCC Level 18, new structure control information is included in CF memory dumps, although data elements (client data) are still not dumped.
- ▶ Enhanced CFCC tracing support: CFCC Level 18 has enhanced trace points, especially in areas, such as latching, suspend queue management and dispatching, duplexing protocols, and sublist notification.
- ▶ Enhanced triggers for CF nondisruptive dumping for soft-failure cases beyond break-duplexing

CFCC Level 19

CFCC Level 19 delivered with Driver level 15 introduces several improvements in the performance and resiliency areas.

Performance improvements

CFCC Level 19 introduces Coupling Thin Interrupts to improve performance in environments which are sharing Coupling Facility engines. While dedicated engines continue to be recommended to obtain the best Coupling Facility performance, Coupling thin interrupts can help to 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 to be dispatched by PR/SM, if it is not already dispatched, allowing the request or signal to be processed in a more timely manner. The coupling facility will give up control when work is exhausted or when PR/SM takes the physical processor away from the logical processor.
- ▶ Use controlled by a new DYNDISP specification.

You can now experience CF response time improvements or more consistent CF response time when using Coupling Facilities with shared engines. This 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.

Resiliency enhancements

CFCC Level 19 now supports Flash Express. It improves resiliency while providing cost effective standby capacity to help manage the potential overflow of WebSphere MQ shared queues. Structures can now be allocated with a combination of real memory and Storage Class Memory (SCM) provided by the Flash Express feature.

For more information about Flash Express and CF Flash exploitation, refer to Appendix C, “Flash Express” on page 501.

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. This also means that the CF Control Code uses all the processor capacity (cycles) 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, you can see the details in 3.9.2, “Dynamic CF dispatching” on page 124.

Performance consideration: Dedicated-processor CF still provides best CF image performance for production.

CF structure sizing changes are expected when going from CFCC Level 17 (or earlier) to CFCC Level 18 or CFCC Level 19 as well as from CFCC Level 18 to CFCC Level 19. Ensure the CF LPAR has at least 512 MB storage and also review the CF structure size by using the CFSizer tool, which is available at this website:

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

3.9.2 Dynamic CF dispatching

Dynamic CF dispatching provides the following function 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 (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 there is no more work. 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 is used, the CFCC code is changed to handle these interrupts correctly. CFCC was also changed to voluntarily give up control of the processor whenever it runs out of work to do, relying 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 can also provide reasonably close to dedicated-engine CF performance, as long as the CF engines are not coupling facility control code thin interrupts. The introduction of thin interrupts allows a CF to run using a shared processor and to still have good performance. The shared engine is allowed to be undispached when there is no more work, just as in the past. The new thin interrupt now gets the shared processor dispatched as soon as a command or duplexing signal gets 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 3-12 on page 125 shows dynamic CF dispatching.

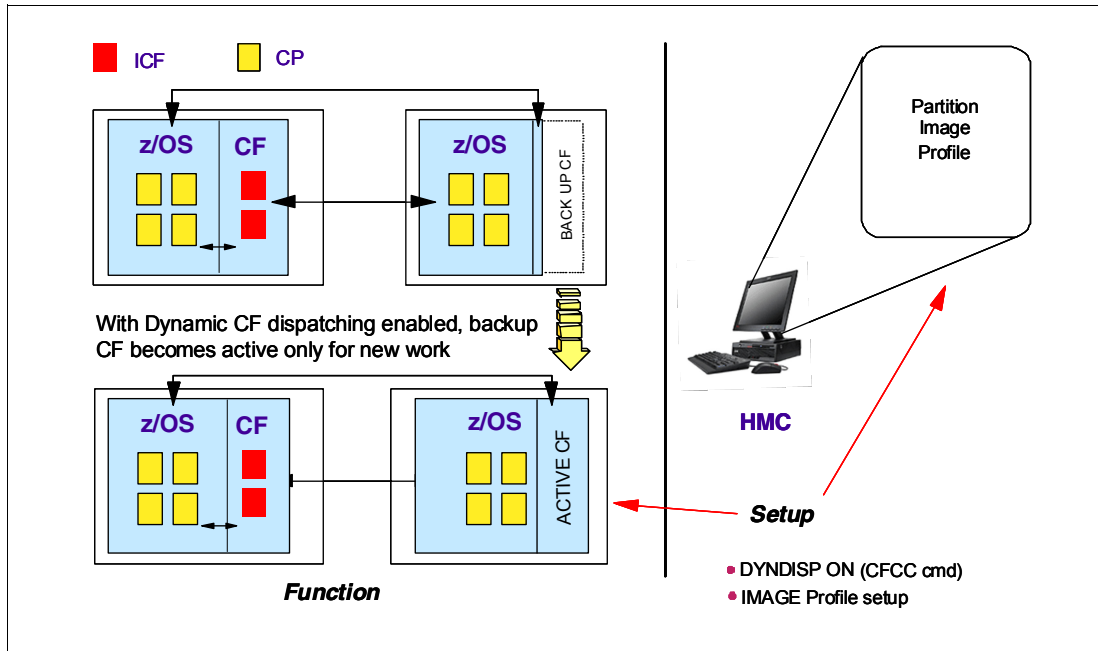


Figure 3-12 Dynamic CF dispatching (shared CPs or shared ICF PUs)

For more information about CF configurations, see *Coupling Facility Configuration Options*, GF22-5042, which is also available on the Parallel Sysplex website:

<http://www.ibm.com/systems/z/advantages/psa/index.html>



Central processor complex I/O system structure

This chapter describes the I/O system structure and connectivity options available on the IBM zEnterprise EC12 (zEC12).

This chapter includes the following topics:

- ▶ Introduction to InfiniBand and PCIe
- ▶ I/O system overview
- ▶ I/O cages
- ▶ I/O drawers
- ▶ PCIe I/O drawers
- ▶ I/O cage, I/O drawer, and PCIe I/O drawer offerings
- ▶ Fanouts
- ▶ I/O feature cards
- ▶ Connectivity
- ▶ Parallel Sysplex connectivity
- ▶ Cryptographic functions
- ▶ Integrated firmware processor
- ▶ Flash Express
- ▶ 10 GbE RoCE Express
- ▶ zEDC Express

4.1 Introduction to InfiniBand and PCIe

The zEC12 supports two types of internal I/O infrastructure:

- ▶ InfiniBand infrastructure for I/O cages and I/O drawers
- ▶ Peripheral Component Interconnect Express (PCIe)-based infrastructure for PCIe I/O drawers (new form factor drawer and I/O features)

InfiniBand I/O infrastructure

The InfiniBand I/O infrastructure was first made available on System z10 and is supported on the zEC12. It consists of these components:

- ▶ InfiniBand fanouts that support the current 6 GBps InfiniBand I/O interconnect
- ▶ InfiniBand I/O card domain multiplexers with redundant I/O interconnect in these form factors:
 - The 14U, 28-slot, 7-domain I/O cage
 - The 5U, 8-slot, 2-domain I/O drawer

PCIe I/O infrastructure

IBM extends the use of industry standards on the System z platform by offering a Peripheral Component Interconnect Express Generation 2 (PCIe Gen2) I/O infrastructure. The PCIe I/O infrastructure that is provided by the zEnterprise central processor complexes (CPCs) improves I/O capability and flexibility, while allowing for the future integration of PCIe adapters and accelerators.

The zEC12 PCIe I/O infrastructure consists of the following components:

- ▶ PCIe fanouts that support 8 GBps I/O bus interconnection for processor book connectivity to the PCIe I/O drawers
- ▶ The 7U, 32-slot, 4-domain PCIe I/O drawer for PCIe I/O features

The zEnterprise PCIe I/O infrastructure provides these benefits:

- ▶ **Bandwidth:** Increased bandwidth from the processor book or drawer to the I/O domain in the PCIe I/O drawer through an 8 GBps bus.
- ▶ **14% more capacity:** Two PCIe I/O drawers occupy the same space as one I/O cage. Up to 128 channels (64 PCIe I/O features) are supported versus the 112 channels (28 I/O features) offered with the I/O cage.
- ▶ **Better granularity for the storage area network (SAN) and the local area network (LAN):** For the Fibre Channel connection (FICON), High Performance FICON on System z (zHPF), and Fibre Channel Protocol (FCP) storage area networks, the FICON Express8S 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.
- ▶ **New PCIe features can be plugged into the PCIe I/O drawer, such as Flash Express, zEnterprise Data Compression (zEDC) Express, and 10GbE Remote Direct Memory Access over Converged Ethernet (RoCE) Express.**

4.1.1 InfiniBand specification

The InfiniBand specification defines the raw bandwidth of one lane (referred to as 1x) connection at 2.5 Gbps. Two more lane widths are specified, referred to as 4x and 12x, as multipliers of the base link width.

Similar to Fibre Channel, PCI Express, Serial Advanced Technology Attachment (SATA), and many other contemporary interconnects, InfiniBand is a point-to-point, bidirectional serial link. It is intended for the connection of processors with high-speed peripheral devices, such as disks. InfiniBand supports various signaling rates and, as with PCI Express, links can be bonded together for more bandwidth.

The serial connection's signaling rate is 2.5 Gbps on one lane in each direction, per physical connection. Currently, InfiniBand also supports 5 Gbps and 10 Gbps signaling rates.

4.1.2 Data, signaling, and link rates

Links use 8b/10b encoding (every 10 bits sent carry 8 bits of data). Therefore, the useful data transmission rate is four-fifths of the signaling rate (signaling rate equals raw bit rate). Therefore, links carry 2 Gbps, 4 Gbps, or 8 Gbps of useful data for a 1x link.

Links can be aggregated in units of 4 or 12, indicated as 4x¹ and 12x. A 12x link therefore carries 120 Gbps raw or 96 Gbps of payload (useful) data. Larger systems with 12x links are typically used for cluster and supercomputer interconnects, as implemented on the zEC12, and for inter-switch connections.

For details and the standard for InfiniBand, see the InfiniBand website:

<http://www.infinibandta.org>

InfiniBand functions on zEC12: Not all properties and functions that are offered by InfiniBand are implemented on the zEC12. Only a subset is used to fulfill the interconnect requirements that are defined for the zEC12.

4.1.3 PCIe

PCIe is a serial bus with an embedded clock. It uses 8b/10b encoding, where every 8 bits are encoded into a 10-bit symbol that is then decoded at the receiver. Therefore, the bus must transfer 10 bits to send 8 bits of actual usable data. A PCIe bus generation 2 (gen2) single lane can transfer 5 Gbps of raw data (duplex connection), which is 10 Gbps of raw data. From these 10 Gbps, only 8 Gbps are actual data (payload). Therefore, an x16 (16 lanes) PCIe gen2 bus transfers 160 Gbps encoded, which is 128 Gbps of unencoded data (payload). This result is 20 GBps raw data and 16 GBps of encoded data.

The new measuring unit for transfer rates for PCIe is gigatransfers per second (GT/s). This measurement refers to the raw data, even though only 80% of this transfer is actual payload data. The translation between GT/s to GBps is 5 GT/s equals 20 GBps, or 1 GT/s equals 4 GBps.

The 16 lanes of the PCIe bus are virtual lanes that consist of one transmit and one receive lane. Each of these lanes consists of two physical copper wires. The physical method that is used to transmit signals is a differential bus. A *differential bus* means that the signal is encoded into the different voltage levels between two wires. This is as opposed to one voltage

¹ zEC12 does not support this data rate.

level on one wire in comparison to the ground signal. Therefore, each of the 16 PCIe lanes actually uses four copper wires for the signal transmissions.

4.2 I/O system overview

This section lists the zEC12 I/O characteristics and a summary of supported features.

4.2.1 Characteristics

The zEC12 I/O subsystem design provides great flexibility, high availability, and excellent performance characteristics:

- ▶ High bandwidth

The zEC12 uses PCIe as an internal interconnect protocol to drive PCIe I/O drawers. The I/O bus infrastructure data rate increases up to 8 GBps.

The zEC12 uses InfiniBand as the internal interconnect protocol to drive I/O drawers and CPC to CPC connection. InfiniBand supports I/O bus infrastructure data rates up to 6 GBps.

- ▶ Connectivity options:

- The zEC12 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 connection, zEC12 uses Parallel Sysplex InfiniBand (IFB) or InterSystem Channel (ISC)-3² coupling links. The 10GbE RoCE Express feature provides high speed access to a remote CPC's memory by using Shared Memory Communications - Remote Direct Memory Access (SMC-R) protocol.

- ▶ Concurrent I/O upgrade

You can concurrently add I/O features to the zEC12 if unused I/O slot positions are available.

- ▶ Concurrent PCIe I/O drawer upgrade

Additional PCIe I/O drawers can be installed concurrently if there is free space in one of the I/O drawer slots.

- ▶ 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 Express8S, FICON Express8, and FICON Express4, OSA Express5S, and 10GbE RoCE Express features have Small Form-Factor Pluggable (SFP) optics. These optics allow each channel to be individually serviced in 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 plugged in an I/O cage, I/O drawer, or PCIe I/O drawer supports concurrent card replacement during a repair action.

² Available on carry-forward upgrades only

4.2.2 Summary of supported I/O features

The following I/O features are supported:

- ▶ Up to 320 FICON Express8S channels
- ▶ Up to 176 FICON Express8 channels
- ▶ Up to 176 FICON Express4 channels
- ▶ Up to 96 OSA-Express5S ports
- ▶ Up to 96 OSA-Express4S ports
- ▶ Up to 96 OSA-Express3 ports
- ▶ Up to 48 ISC-3 coupling links
- ▶ Up to 16 InfiniBand fanouts with one of these options:
 - Up to 32 12x InfiniBand coupling links with HCA2-O fanout
 - Up to 32 1x InfiniBand coupling links with HCA2-O long range (LR) (1xIFB) fanout
 - Up to 32 12x InfiniBand coupling links with HCA3-O fanout
 - Up to 64 1x InfiniBand coupling links with HCA3-O LR (1xIFB) fanout

Coupling links: The maximum number of external coupling links combined (ISC-3 and IFB coupling links) cannot exceed 112 for each zEC12.

In addition to I/O features, new PCIe features are supported exclusively in the PCIe I/O drawer:

- ▶ Up to 8 zFlash Express features
- ▶ Up to 8 zEDC Express features
- ▶ Up to sixteen 10 GbE RoCE Express features

4.3 I/O cages

The I/O cage is 14 Electronic Industries Alliance (EIA) units high. Each cage supports up to seven I/O domains for a total of 28 I/O card slots. Each I/O domain supports four I/O card slots. Each uses an IFB-MP card in the I/O cage and a copper cable that is connected to a host channel adapter (HCA-C) fanout in the CPC books. An eight IFB-MP card is installed to provide an alternate path to I/O cards in slots 29, 30, 31, and 32 during a repair action. Figure 4-1 on page 132 shows an I/O cage diagram.

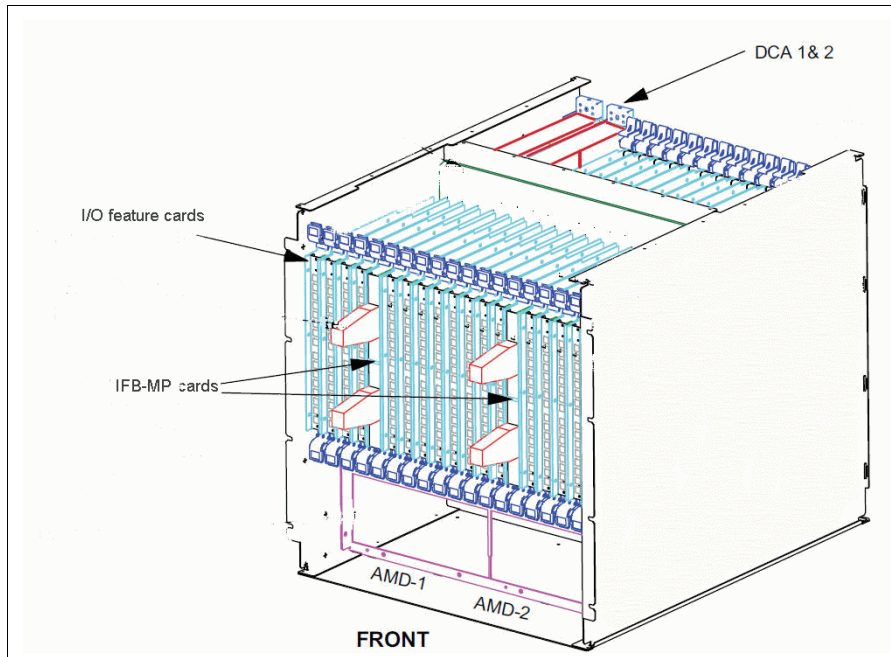


Figure 4-1 I/O cage

Figure 4-2 illustrates the I/O structure of an I/O cage. An InfiniBand (IFB) cable connects the HCA2-C fanout to an IFB-MP card in the I/O cage. The passive connection between two IFB-MP cards allows for redundant I/O interconnection. The IFB cable between an HCA2-C fanout in a book and each IFB-MP card in the I/O cage supports a 6 Gbps bandwidth.

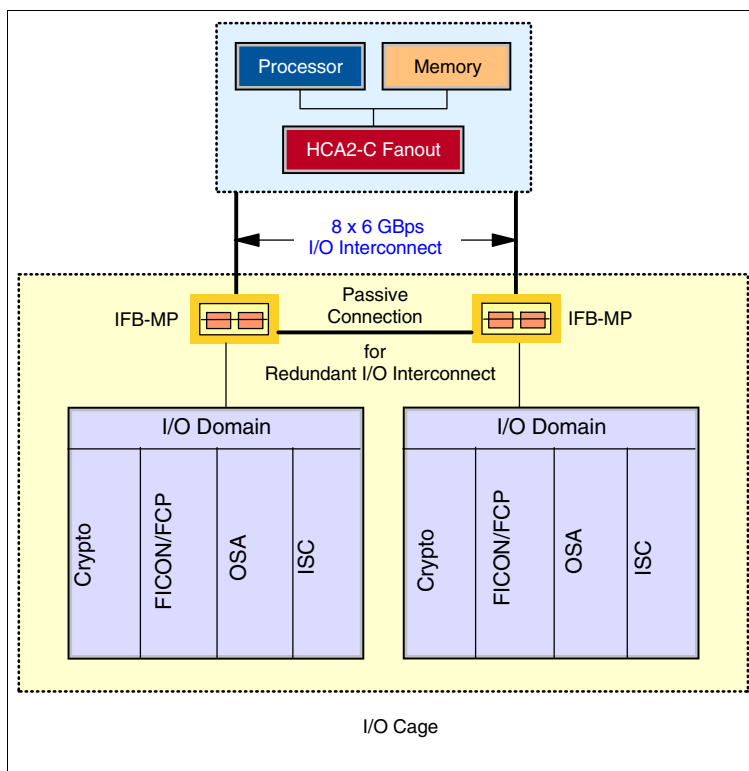


Figure 4-2 EC12 I/O structure when using I/O cage

Restriction: Only one I/O cage is supported in zEC12 on a carry-forward basis only.

The I/O cage domains and their related I/O slots are shown in Figure 4-3.

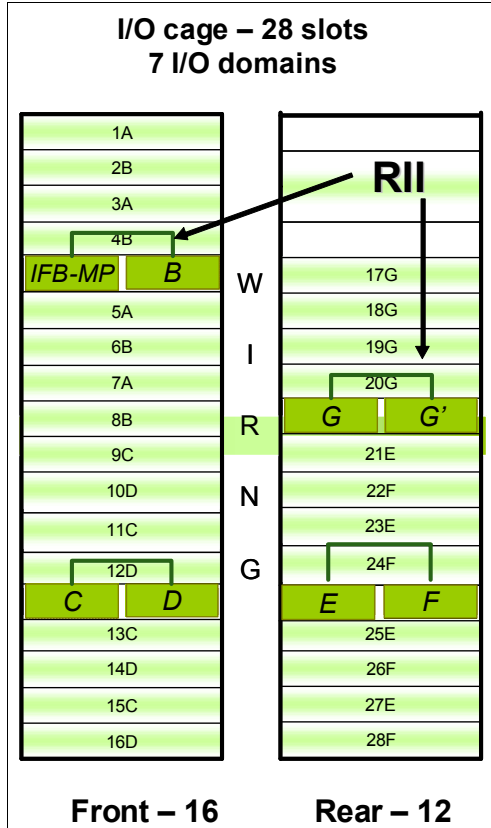


Figure 4-3 I/O domains of I/O cage

Each I/O domain supports up to four I/O cards (FICON, OSA, Crypto, or ISC). All I/O cards are connected to the IFB-MP cards through the backplane board.

Table 4-1 lists the I/O domains and their related I/O slots.

Table 4-1 I/O domains of I/O cage

Domain number (name)	I/O slot in domain
0 (A)	01, 03, 06, 08
1 (B)	02, 04, 07, 09
2 (C)	10, 12, 15, 17
3 (D)	11, 13, 16, 18
4 (E)	19, 21, 24, 26
5 (F)	20, 22, 25, 27
6 (G)	29, 30, 31, 32

Restriction: The Power Sequence Controller (PSC) feature is not supported on zEC12.

4.4 I/O drawers

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 IFB-MP card in the I/O drawer and a copper cable to connect to a Host Channel Adapter (HCA) fanout in the CPC books.

The link between the HCA in the CPC books 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. The 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 4-4.

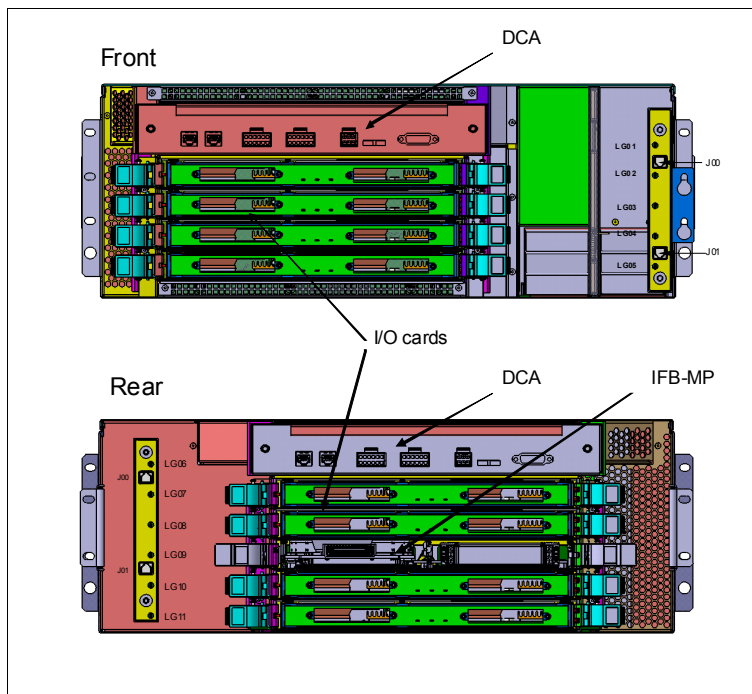


Figure 4-4 I/O drawer

The I/O structure in a zEC12 CPC is illustrated in Figure 4-5. An IFB cable connects the HCA fanout card in the book to an IFB-MP card in the I/O drawer. The passive connection between two IFB-MP cards allows redundant I/O interconnect (RII). RII provides connectivity between an HCA fanout card, and I/O cards in a 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.

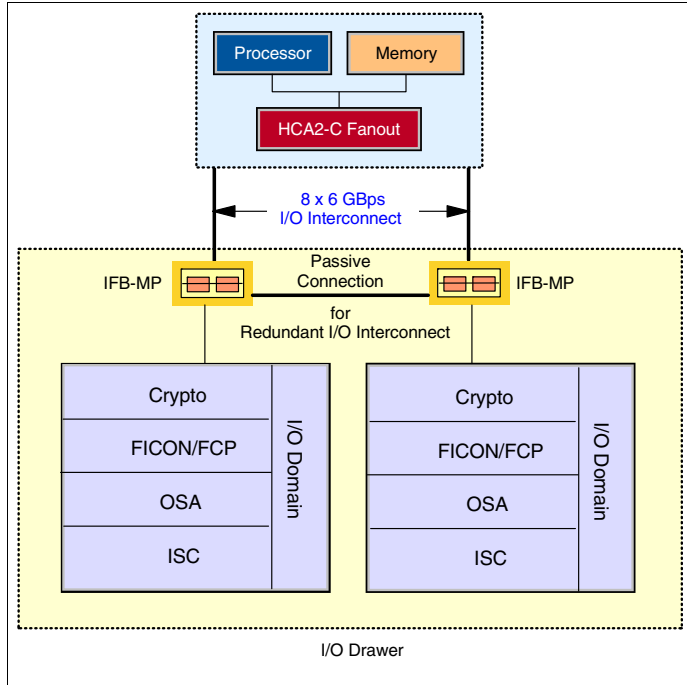


Figure 4-5 zEC12 I/O structure when you are using an I/O drawer

Restriction: Only two I/O drawers are supported in zEC12 on a carry-forward basis only.

The I/O drawer domains and their related I/O slots are shown in Figure 4-6. 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 (FICON, OSA, Crypto, or ISC). The I/O cards are connected to the IFB-MP card through the backplane board.

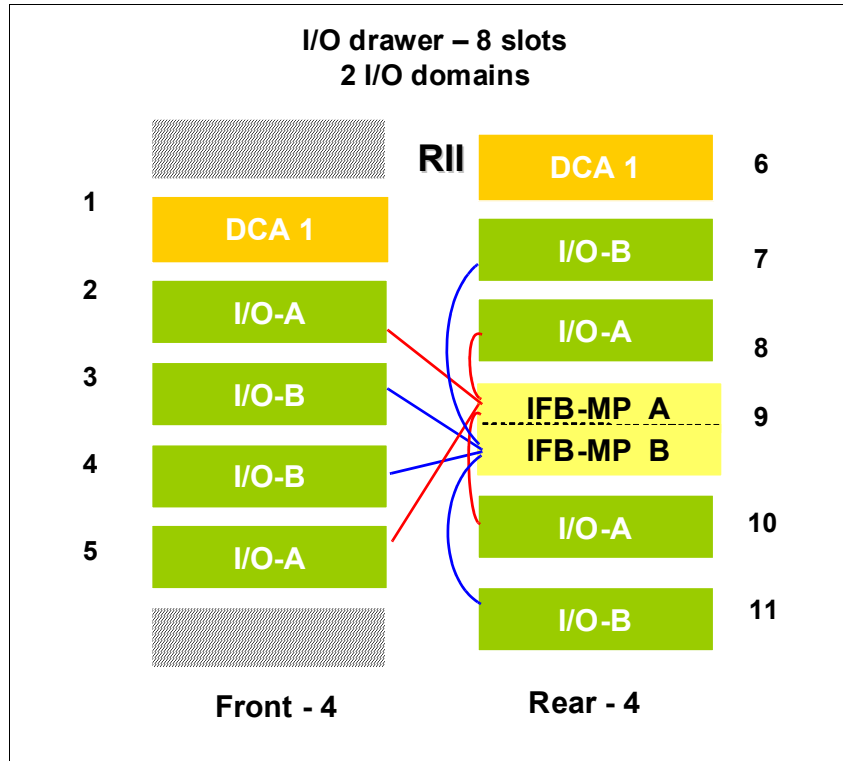


Figure 4-6 I/O domains of an I/O drawer

Each I/O domain supports four I/O card slots. Balancing I/O cards across both I/O domains on new build servers, or on upgrades, is automatically done when the order is placed. Table 4-2 lists the I/O domains and their related I/O slots.

Table 4-2 I/O domains of I/O drawer

Domain	I/O slot in domain
A	02, 05, 08, 10
B	03, 04, 07, 11

Restriction: The Power Sequence Controller (PSC) feature is not supported on zEC12.

4.5 PCIe I/O drawers

The PCIe I/O drawers are attached to the CPC books 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 8 GBps. PCIe switch application-specific integrated circuits (ASICs) are used to fan out the host bus from the CPC books to the individual I/O features. Up to 128 channels (32 PCIe I/O features) are supported versus the 112 channels (28 I/O features) offered with the I/O cage.

The PCIe drawer is a two-sided drawer (I/O cards on both sides) that is 7U high (one half of I/O cage). The drawer contains 32 I/O slots for feature cards, four switch cards (two in front, two in rear), two DCAs to provide the 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 4-7.

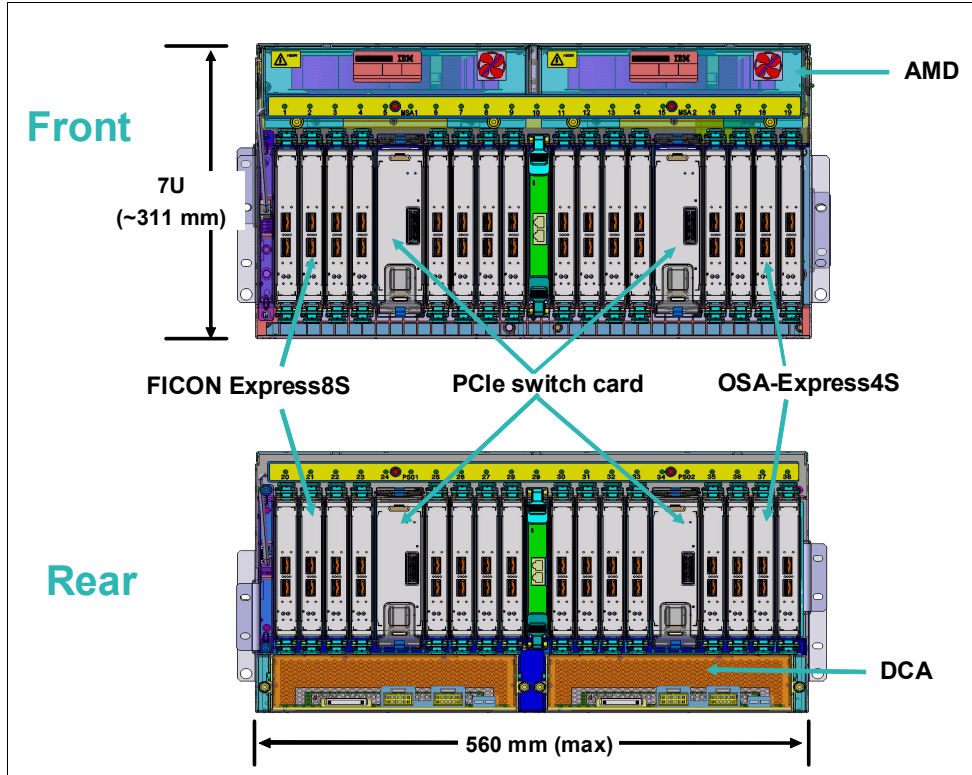


Figure 4-7 PCIe I/O drawer

The I/O structure in a zEC12 CPC is illustrated in Figure 4-8 on page 138. 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 books through a single x16 PCIe Gen 2 bus from a PCIe fanout card. The fanout card converts the book internal bus into two PCIe buses.

A switch card in the front is connected to a switch card in the rear through the PCIe I/O drawer board. This configuration provides a failover capability during a PCIe fanout card failure or book upgrade. 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.

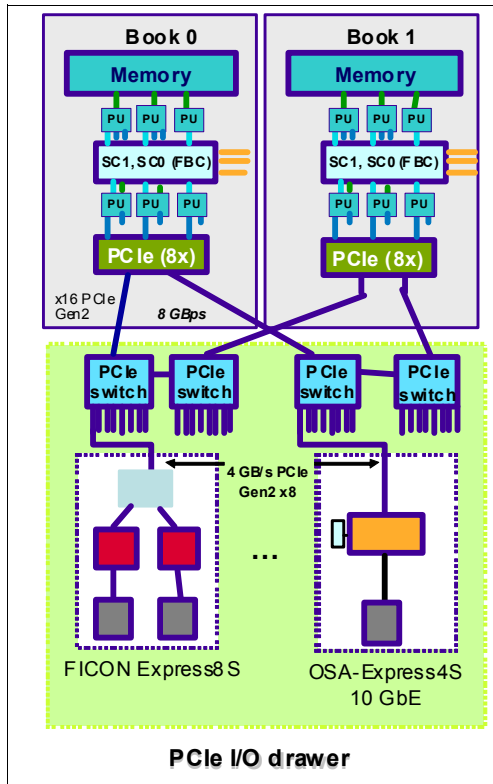


Figure 4-8 zEC12 structure when you are using PCIe I/O drawers

The PCIe I/O Drawer supports up to 32 I/O features. They are organized in 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. In the case of 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.

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 from two different PCIe fanouts in the books. All four domains in one of these drawers can be activated with two 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 4-9.

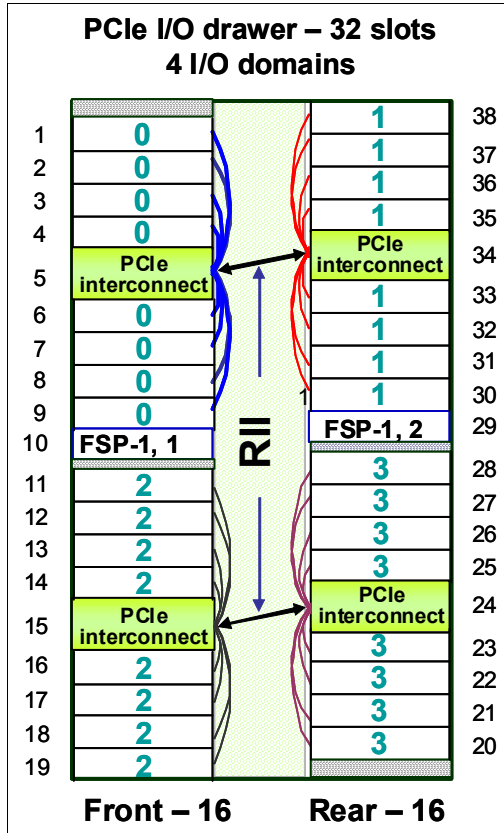


Figure 4-9 I/O domains of PCIe I/O drawer

Each I/O domain supports up to eight features (FICON, OSA, and Crypto), and up to two native PCIe features (Flash Express, zEDC Express, and 10GbE RoCE Express). All I/O cards connect to the PCIe switch card through the backplane board.

Table 4-3 lists the I/O domains and slots.

Table 4-3 I/O domains of PCIe I/O drawer

Domain	I/O slot in domain
0	01, 02, 03, 04, 06, 07, 08, 09
1	30, 31, 32, 33, 35, 36, 37, 38
2	11, 12, 13, 14, 16, 17, 18, 19
3	20, 21, 22, 23, 25, 26, 27, 28

Power Sequence Controller: The PCIe I/O drawer does not support the Power Sequence Controller (PSC) feature.

4.6 I/O cage, I/O drawer, and PCIe I/O drawer offerings

A maximum of five PCIe I/O drawers can be installed, which support up to 160 PCIe I/O features. Clients no longer order I/O cages or I/O drawers. They now order I/O features, and the configurator determines the correct mix of I/O cages, I/O drawers, and PCIe I/O drawers.

Restriction: On a *new build* zEC12, only PCIe I/O drawers are supported. A mixture of I/O cages, I/O drawers, and PCIe I/O drawers are only available on upgrades to a zEC12.

Some I/O and speciality features are supported only by I/O cages and I/O drawers:

- ▶ FICON Express8
- ▶ FICON Express4
- ▶ OSA-Express3
- ▶ ISC-3
- ▶ Crypto Express3

The PCIe I/O drawers support the following PCIe features:

- ▶ FICON Express8S
- ▶ OSA-Express5S
- ▶ OSA-Express4S
- ▶ 10 GbE RoCE Express
- ▶ Crypto Express4S
- ▶ Flash Express
- ▶ zEDC Express

Table 4-4 is an overview of the number of I/O cages, I/O drawers, and PCIe drawers that can be in a zEC12.

Table 4-4 I/O cage and drawer summary

Description	New build	Carry-forward basis	MES addition
I/O cage	0	0 - 1	0
I/O drawer	0	0 - 2	0
PCIe I/O drawer	0 - 5	0 - 5	0 - 5

A maximum of 44 I/O features can be carried forward. Table 4-5 lists the number and mix of I/O cages and I/O drawers, depending on the number of original I/O features.

Table 4-5 Number and mix of I/O cages and I/O drawers

Number of I/O features carried forward on upgrades	Number of I/O cages	Number of I/O drawers
0	0	0
1 - 8	0	1
9 - 16	0	2
17 - 28	1	0
29 - 36	1	1
37 - 44	1	2

4.7 Fanouts

The zEC12 server uses fanout cards to connect the I/O hardware subsystem to the CPC books. They also provide the InfiniBand coupling links for Parallel Sysplex. All fanout cards support concurrent add, delete, and move.

zEC12 supports two different internal I/O infrastructures for the internal connection. zEC12 uses InfiniBand based infrastructure for the internal connection to I/O cages and I/O drawers, and uses PCIe-based infrastructure for PCIe I/O drawers in which the different supported features are located.

The InfiniBand and PCIe fanouts are on the front of each book. Each book has eight fanout slots. They are named D1 to DA, top to bottom. Slots D3 and D4 are not used for fanouts. Six types of fanout cards are supported by zEC12. Each slot holds one of the following six fanouts:

- ▶ Host Channel Adapter (HCA2-C): This copper fanout provides connectivity to the IFB-MP card in the I/O cage and I/O drawer.
- ▶ PCIe fanout: This copper fanout provides connectivity to the PCIe switch card in the PCIe I/O drawer.
- ▶ Host Channel Adapter (HCA2-O (12xIFB)): This optical fanout provides 12x InfiniBand coupling link connectivity up to a 150-meter (492-ft) distance to a zEC12, IBM zEnterprise BC12 (zBC12), z196, z114, and System z10.
- ▶ Host Channel Adapter (HCA2-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 System z qualified dense wavelength division multiplexing (DWDM) equipment) distance to zEC12, zBC12, z196, z114, and System z10 servers.
- ▶ Host Channel Adapter (HCA3-O (12xIFB)): This optical fanout provides 12x InfiniBand coupling link connectivity up to 150-meter (492-ft) distance to a zEC12, zBC12, z196, z114, and System z10.
- ▶ Host Channel Adapter (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 System z qualified DWDM equipment) distance to zEC12, zBC12, z196, z114, and System z10.

The HCA3-O LR (1xIFB) fanout comes with four ports, and other fanouts come with two ports.

Figure 4-10 illustrates the IFB connection from the CPC cage to an I/O cage and an I/O drawer, and the PCIe connection from the CPC books to a PCIe I/O drawer.

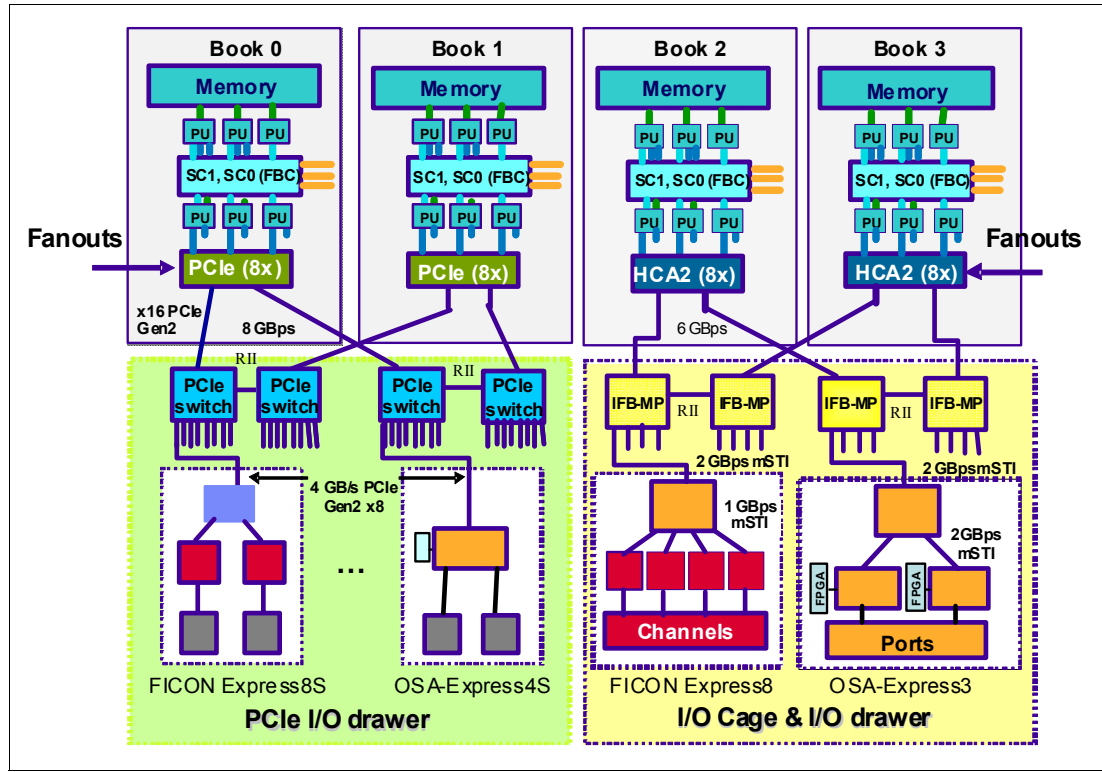


Figure 4-10 PCIe and InfiniBand I/O infrastructure

Figure 4-11 illustrates the zEC12 coupling link options.

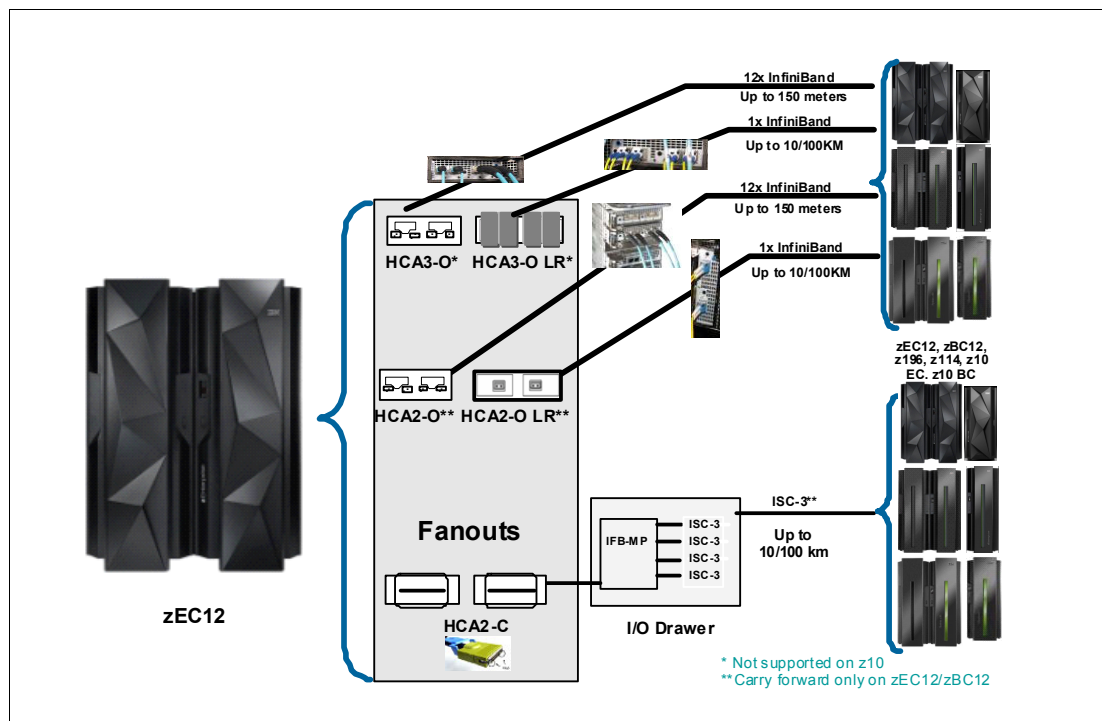


Figure 4-11 zEC12 coupling links

4.7.1 HCA2-C fanout (FC 0162)

The HCA2-C fanout is used to connect to an I/O cage or an I/O drawer by using a copper cable. The two ports on the fanout are dedicated to I/O. The bandwidth of 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 cage or the I/O drawer. An HCA2-C fanout is supported only if carried forward with an MES from z10, z196, z114, or zBC12 model H13 to zEC12 model H20 (air-cooled only). For a new zEC12 installation, it is not possible to have HCA2-C.

HCA2-C fanout: The HCA2-C fanout is used exclusively for connection to the I/O cage and I/O drawer. It cannot be shared for any other purpose.

4.7.2 PCIe copper fanout (FC 0169)

The PCIe fanout card supports the PCIe Gen2 bus and is used to connect to the PCIe I/O drawer. PCIe fanout cards are always plugged in pairs. The bandwidth of each port on the PCIe fanout supports a link rate of up to 8 GBps.

The PCIe fanout supports FICON Express8S, OSA Express5S, OSA Express4S, 10GbE RoCE Express, Crypto Express 4S, Flash Express, and zEDC Express in PCIe I/O drawers.

PCIe fanout: The PCIe fanout is used exclusively for I/O and cannot be shared for any other purpose.

4.7.3 HCA2-O (12x IFB) fanout (FC 0163)

The HCA2-O fanout for 12x InfiniBand (12x IFB) provides an optical interface that is used for coupling links. The two ports on the fanout are dedicated to coupling links to connect to zEC12, zBC12, z196, z114, and System z10 servers. Up to 16 HCA2-O (12x IFB) fanouts are supported by zEC12, and provide up to 32 ports for coupling links. An HCA2-O fanout is supported only if carried forward with a miscellaneous equipment specification (MES) from the z196, z114, or z10.

The HCA2-O fanout supports InfiniBand 12x optical links that offer configuration flexibility, and high bandwidth for enhanced performance of coupling links. There are 12 lanes (two fibers per lane) in the cable, which means that 24 fibers are used in parallel for data transfer. Each port provides one connector for transmit data and one connector for receive data.

The fiber optic cables are industry standard OM3 (2000 MHz-km) 50- μ m multimode optical cables with Multi-Fiber Push-On (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.

Each connection supports a link rate of 6 GBps when connected to a zEC12, zBC12, z196, z114, or System z10 server.

Important: The HCA2-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 4-12 on page 144.

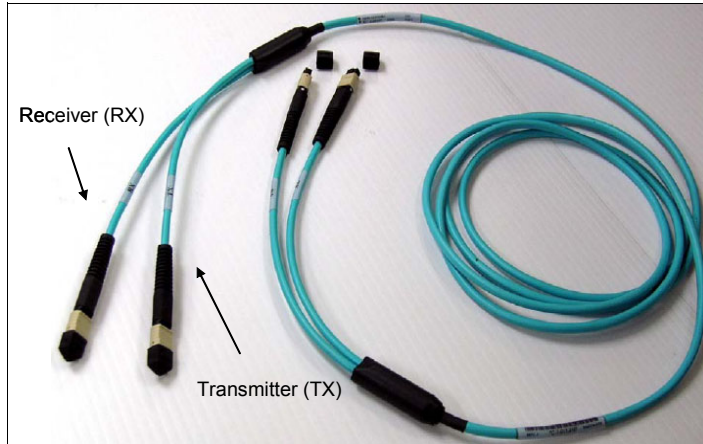


Figure 4-12 OM3 50/125 μm multimode fiber cable with MPO connectors

A fanout has two ports for optical link connections and supports up to 16 channel-path identifiers (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 channel subsystem (CSS). They can also be spanned across multiple CSSs in a CPC.

Each HCA2-O (12xIFB) fanout that is used for coupling links 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 CHPID number. For more information about AID numbering, see “Adapter ID number assignment” on page 148.

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 System z SG24-7539*.

When Server Time Protocol (STP) is enabled, IFB coupling links can be defined as timing-only links to other zEC12, zBC12, z196, z114, and System z10 CPCs.

4.7.4 HCA2-O LR (1x IFB) fanout (FC 0168)

The HCA2-O LR (1xIFB) fanout for 1x InfiniBand provides an optical interface that is used for coupling links. The two ports on the fanout are dedicated to coupling links to connect to zEC12, zBC12, z196, z114, and System z10 servers. Up to 16 HCA2-O LR (1xIFB) fanouts are supported by zEC12 and provide 32 ports for coupling links. An HCA2-O LR fanout is supported only if carried forward with an MES from z196, z114, or z10.

The HCA2-O LR (1x IFB) fanout has 1x optical links that offer a longer distance of coupling links. The cable has one lane that contains two fibers. One fiber is used for transmitting and one fiber is used for receiving data.

Each connection supports a link rate of up to 5 Gbps if connected to a zEC12, zBC12, z196, z114, System z10 CPC, or to a System z qualified dense wavelength division multiplexer (DWDM). It supports a data link rate of 2.5 Gbps when connected to a System z qualified DWDM. The link rate is auto-negotiated to the highest common rate.

The fiber optic cables are 9- μm single mode (SM) optical cables that are terminated with an LC Duplex connector. The maximum unrepeated distance is 10 km (6.2 miles), and up to 100 km (62 miles) with System z qualified DWDM. Request for price quotation (RPQ) 8P2263 or

RPQ 8P2340 is required for 20 km (12.4 miles) support. Going over 100 km (62 miles) requires RPQ 8P2263 or RPQ 8P2340.

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 IOCDS. The coupling links can be defined as shared between images within a channel subsystem. They can also be spanned across multiple CSSs in a server.

Each HCA2-O LR (1xIFB) fanout can be used for link definitions to another server, or a link from one port to a port in another fanout on the same server.

The source and target operating system image, coupling facility (CF) image, and the CHPIDs used on both ports in both servers are defined in IOCDS.

Each HCA2-O LR (1xIFB) fanout that is used for coupling links has an AID number that must be used for definitions in IOCDS. This process creates a relationship between the physical fanout location and the CHPID number. For more information about AID numbering, see “Adapter ID number assignment” on page 148.

For more information about how the AID is used and referenced in HCD, see *Implementing and Managing InfiniBand Coupling Links on System z SG24-7539*.

When STP is enabled, IFB LR coupling links can be defined as timing-only links to other zEC12, zBC12, z196, z114, and System z10 CPCs.

4.7.5 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 to connect to zEC12, zBC12, z196, z114, and System z10 CPCs. Up to 16 HCA3-O (12x IFB) fanouts are supported, and provide up to 32 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 4-13 on page 146.

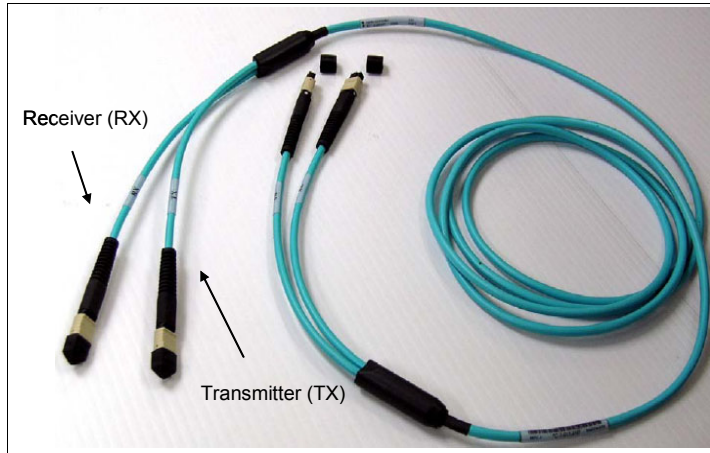


Figure 4-13 OM3 50/125 μm multimode fiber cable with MPO connectors

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 IOCDS. The coupling links can be defined as shared between images within a CSS. They can also be spanned across multiple CSSs in a CPC.

Each HCA3-O (12x IFB) fanout that is used for coupling links 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 148.

For more information about how the AID is used and referenced in the HCD, see *Implementing and Managing InfiniBand Coupling Links on System z SG24-7539*.

When STP is enabled, IFB LR coupling links can be defined as timing-only links to other zEC12, zBC12, z196, z114, and System z10 CPCs.

12x IFB and 12x IFB3 protocols

There are two protocols that 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.

4.7.6 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 zEC12, zBC12, z196, z114, and System z10 servers. Up to 16 HCA3-O LR (1xIFB) fanouts are supported by zEC12, and provide up to 64 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 zEC12, zBC12, z196, z114, or z10. It supports a link rate of 2.5 Gbps when connected to a System z 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 are terminated with an LC Duplex connector. The maximum unrepeat distance is 10 km (6.2 miles), and up to 100 km (62 miles) with System z qualified DWDM. RPQ 8P2263 or RPQ 8P2340 is required for 20 km (12.4 miles) support. Going beyond 100 km (62 miles) requires RPQ 8P2263 or RPQ 8P2340.

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 can also be spanned across multiple CSSs in a server. This configuration is compatible with the HCA2-O LR (1xIFB) fanout, which has two ports.

Each HCA3-O LR (1xIFB) fanout can be used for link definitions to another server, or a link from one port to a port in another fanout on the same server.

The source and target operating system image, CF image, and the CHPIDs used on both ports in both servers are defined in IOCDs.

Each HCA3-O LR (1xIFB) fanout that is used for coupling links 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 148.

For more information about how the AID is used and referenced in HCD, see *Implementing and Managing InfiniBand Coupling Links on System z* SG24-7539.

When STP is enabled, IFB LR coupling links can be defined as timing-only links to other zEC12, zBC12, z196, z114, and System z10 CPCs.

4.7.7 Fanout considerations

Fanout slots in each book can be used to plug different fanouts, where each fanout is designed for a special purpose. As a result, certain restrictions might apply to the number of available channels in the I/O cage, I/O drawer, and PCIe I/O drawer. Depending on the model, the number of fanouts varies.

The plugging rules for fanouts for each model are illustrated in Figure 4-14.

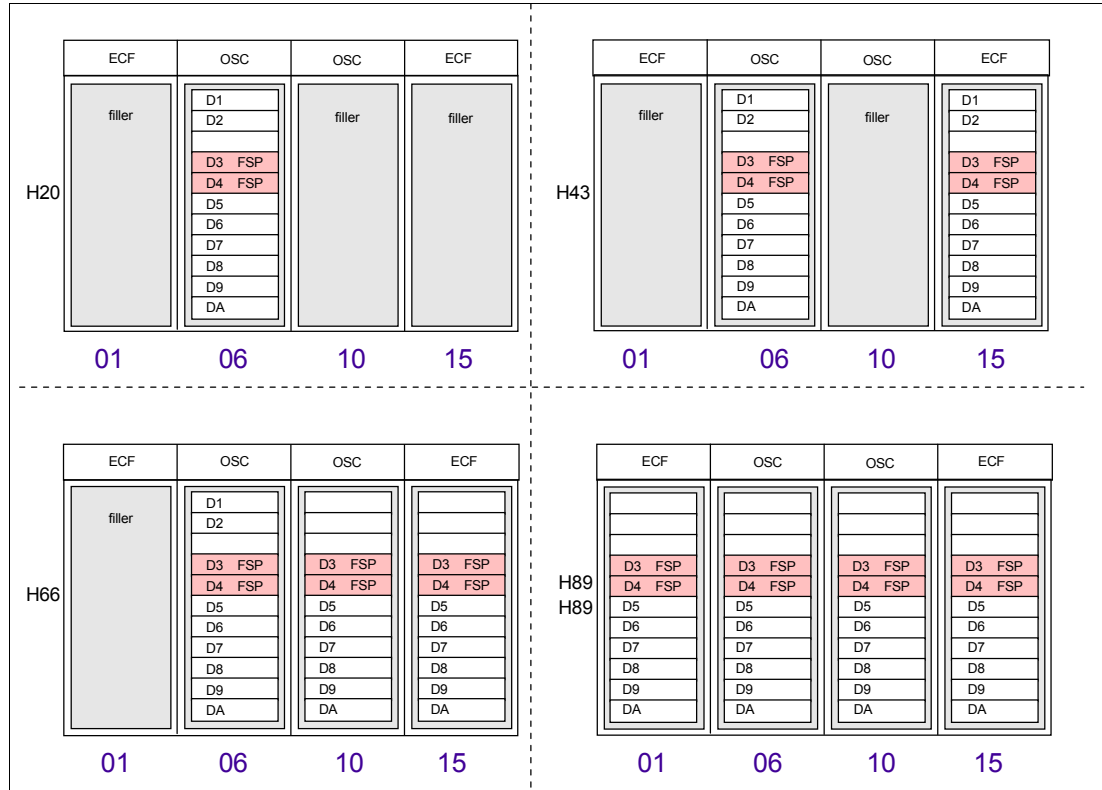


Figure 4-14 Fanout plugging rules

Adapter ID number assignment

IFB fanouts and ports are identified by an AID that is initially dependent on their physical locations, which is unlike channels that are installed in an I/O cage. 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 4-6 illustrates the AID assignment for each fanout slot relative to the book location on a new build system.

Table 4-6 AID number assignment

Book	Slot	Fanout slot	AIDs
First	06	D1, D2, D5-DA	08, 09, 0A - 0F
Second	15	D1, D2, D5-DA	18, 19, 1A - 1F
Third	10	D1, D2, D5-DA	10, 11, 12 - 17
Fourth	01	D1, D2, D5-DA	00, 01, 02 - 07

Fanout slots

The fanout slots are numbered D1 to DA top to bottom, as shown in Table 4-7. All fanout locations and their AIDs for all four books are shown in the table for reference only. Fanouts in locations D1 and D2 are not available on all models. Slots D3 and D4 never have a fanout installed because they are dedicated for FSPs.

Important: Slots D1 and D2 are not used in a 4-book system, and only partially in a 3-book system.

Table 4-7 Fanout AID numbers

Fanout location	Fourth book	First book	Third book	Second book
D1	00	08	10	18
D2	01	09	11	19
D3	-	-	-	-
D4	-	-	-	-
D5	02	0A	12	1A
D6	03	0B	13	1B
D7	04	0C	14	1C
D8	05	0D	15	1D
D9	06	0E	16	1E
DA	07	0F	17	1F

Important: The AID numbers in Table 4-7 are valid only for a new build system or if new books are added. If a fanout is moved, the AID follows the fanout to its new physical location.

The AID assigned to a fanout is 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 report, named PCHID REPORT, for a model M32. In this example, one fanout is installed in the first book (location 06) and one fanout is installed in the second book (location 15), and both are in location D5. The assigned AID for the fanout in the first book is 0A. The AID assigned to the fanout in the second book is 1A.

Example 4-1 AID assignment in PCHID REPORT

```

CHPIDSTART
12345675                PCHID REPORT                Jun xx,2012
Machine: xxxx-H43 SNXXXXXXX
-----
Source      Cage Slot F/C   PCHID/Ports or AID      Comment
06/D5      A25B D506 0163  AID=0A
15/D5      A25B D515 0163  AID=1A

```

4.7.8 Fanout summary

Fanout features that are supported by the zEC12 server are shown in Table 4-8. The table provides the feature type, feature code, and information about the link that is supported by the fanout feature.

Table 4-8 Fanout summary

Fanout feature	Feature code	Use	Cable type	Connector type	Maximum distance	Link data rate
HCA2-C	0162	Connect to I/O cage or I/O drawer	Copper	N/A	3.5 m (11.48 ft)	6 Gbps
HCA2-O (12xIFB)	0163	Coupling link	50- μ m MM OM3 (2000 MHz-km)	MPO	150 m (492 ft)	6 Gbps
HCA2-O LR (1xIFB)	0168	Coupling link	9- μ m SM	LC Duplex	10 km ^a (6.2 miles)	5.0 Gbps 2.5 Gbps ^b
PCIe fanout	0169	Connect to PCIe I/O drawer	Copper	N/A	3 m (9.8 ft)	8 Gbps
HCA3-O (12xIFB)	0171	Coupling link	50- μ m MM OM3 (2000 MHz-km)	MPO	150 m (492 ft)	6 Gbps ^c
HCA3-O LR (1xIFB)	0170	Coupling link	9- μ m SM	LC Duplex	10 km ^a (6.2 miles)	5.0 Gbps 2.5 Gbps ^b

a. Up to 100 km (62 miles) with repeaters (System z qualified DWDM).

b. Auto-negotiated, depending on DWDM equipment.

c. When using the 12x IFB3 protocol, synchronous service times are 40% faster than when you use the 12x IFB protocol.

4.8 I/O feature cards

I/O cards have ports to connect the zEC12 to external devices, networks, or other servers. I/O cards are plugged into the I/O cage, I/O drawer, and 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 I/O cards, Crypto Express features can be installed in I/O drawers, I/O cages, or PCIe drawers. Flash Express, 10GbE RoCE Express, and zEDC Express features can be installed in a PCIe drawer only. These feature types occupy one or more I/O slots.

4.8.1 I/O feature card ordering information

Table 4-9 lists the I/O features supported by zEC12 and the ordering information for them.

Table 4-9 I/O features and ordering information

Channel feature	Feature code	New build	Carry forward
FICON Express8S 10KM LX	0409	Y	N/A
FICON Express8S SX	0410	Y	N/A
FICON Express8 10KM LX	3325	N	Y
FICON Express8 SX	3326	N	Y
FICON Express4 10KM LX	3321	N	Y
FICON Express4 SX	3322	N	Y
OSA-Express5S 10 GbE LR	0415	Y	N/A
OSA-Express5S 10 GbE SR	0416	Y	N/A
OSA-Express5S GbE LX	0413	Y	N/A
OSA-Express5S GbE SX	0414	Y	N/A
OSA-Express5S 1000BASE-T Ethernet	0417	Y	N/A
OSA-Express4S 10 GbE LR	0406	Y	Y
OSA-Express4S 10 GbE SR	0407	Y	Y
OSA-Express4S GbE LX	0404	Y	Y
OSA-Express4S GbE SX	0405	Y	Y
OSA-Express4S 1000BASE-T Ethernet	0408	Y	N/A
OSA-Express3 10 GbE LR	3370	N	Y
OSA-Express3 10 GbE SR	3371	N	Y

Channel feature	Feature code	New build	Carry forward
OSA-Express3 GbE LX	3362	N	Y
OSA-Express3 GbE SX	3363	N	Y
OSA-Express3 1000BASE-T Ethernet	3367	N	Y
ISC-3	0217 (ISC-M) 0218 (ISC-D)	N	Y
ISC-3 up to 20 km ^a (12.4 miles)	RPQ 8P2197 (ISC-D)	N	Y
HCA2-O (12xIFB)	0163	N	Y
HCA2-O LR (1xIFB)	0168	N	Y
HCA3-O (12xIFB)	0171	Y	Y
HCA3-O LR (1xIFB)	0170	Y	Y
Crypto Express4S	0864	Y	N/A
Crypto Express3	0865	N	Y
Flash Express	0402	Y	N/A
10GbE RoCE Express	0411	Y	N/A
zEDC Express	0420	Y	N/A

a. RPQ 8P2197 enables the ordering of a daughter card that supports 20 km (12.4 miles) unrepeated distance for 1 Gbps peer mode. RPQ 8P2262 is a requirement for that option. Other than the normal mode, the channel increment is two, which means that both ports (FC 0219) at the card must be activated.

4.8.2 PCHID report

A Physical Channel ID (PCHID) reflects the physical location of a channel-type interface.

A PCHID number is based on these factors:

- ▶ The I/O cage, 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 hardware configuration definition (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 installed, the physical slot location, and the assigned PCHID. Example 4-2 on page 153 shows a portion of a sample PCHID REPORT. For more information about the AID numbering rules for InfiniBand coupling links, see “Adapter ID number assignment” on page 148.

Example 4-2 PCHID REPORT

```

CHPIDSTART
12345675          PCHID REPORT          Jun xx,2012
Machine: xxxx-H43 SNXXXXXXX
-----
Source           Cage Slot F/C   PCHID/Ports or AID          Comment
-----
06/D7            A25B D706 0170  AID=0C
15/D5            A25B D515 0171  AID=1A
06/DA/J01       Z15B 04   0405  130/J00J01 131/J02J03
06/D8/J01       Z15B D120 0218  210/J00 211/J01
06/D8/J01       Z15B D220 0218  218/J00 219/J01
06/D9/J02       Z15B 04   0409  520/D1 521/D2
06/DA/J02       Z22B 02   0865  580/P00
15/DA/J02       Z22B 03   0407  590/J00
15/DA/J02       Z22B 04   0408  5A0/J00J01 5A1/J02J03
06/DA/J02       Z22B 05   0408  5B0/J00J01 5B1/J02J03

```

The following list explains the content of the sample PCHID REPORT:

- ▶ Feature code 0170 (HCA3-O LR (1xIFB)) is installed in the first book (cage A25B, slot 06) location D7 and has AID 0C assigned.
- ▶ Feature code 0171 (HCA3-O (12xIFB)) is installed in the second book (cage A25B, slot 15) location D5 and has AID 1A assigned.
- ▶ Feature code 0405 (OSA-Express4S GbE short wavelength (SX)) is installed in cage Z15B slot 4 and has PCHIDs 130 and 131 assigned. PCHID 130 is shared by port 00 and 01; PCHID 131 is shared by ports 02 and 03.
- ▶ Feature code 0218 (ISC-3) is installed in cage Z15B slot 20. It has PCHIDs 210 and 211 assigned to the two ports on the upper daughter card, and PCHIDs 218 and 219 assigned to the two ports on the lower daughter card.
- ▶ Feature code 0409 (FICON Express8S long wavelength (LX) 10 km (6.2 miles)) is installed in drawer Z15B slot 4 and has PCHIDs 520 and 521 assigned.
- ▶ Feature code 0865 (Crypto Express4S) is installed in drawer Z22B slot 2 and has PCHIDs 580 assigned.
- ▶ Feature code 0407 (OSA-Express4S 10 GbE SR) is installed in drawer Z22B slot 3 and has PCHIDs 590 assigned.
- ▶ Feature code 0408 (OSA-Express4S 1000BASE-T) is installed in drawer Z22B slot 4 and has PCHIDs 5A0 and 5A1 assigned. PCHID 5A0 is shared by ports 00 and 01, PCHID 5A1 is shared by ports 02 and 03.

A resource group (RG) parameter is shown in the PCHID REPORT for native PCIe features (see Example 4-3). There is a balanced plugging of native PCIe features between two resource groups (RG1 and RG2).

Example 4-3 Resource group assignment

Source	Cage	Slot	F/C*	PCHID/Ports or AID	Comment
A21/D8/J01	A02B	01	0420	100/	RG1
A21/D8/J01	A02B	09	0411	11C/D1D2	RG1

A21/D1/J02	A02B	11	0411	120/D1D2	RG2
A21/D1/J02	A02B	14	0420	12C/	RG2
A21/D8/J02	A02B	20	0420	140/	RG2
A21/D8/J02	A02B	21	0411	144/D1D2	RG2
A21/D1/J01	A02B	37	0411	178/D1D2	RG1
A21/D1/J01	A02B	38	0420	17C/	RG1

For details about resource groups, see Appendix F, “Native Peripheral Component Interconnect Express (PCIe)” on page 529.

The preassigned PCHID number of each I/O port relates directly to its physical location (jack location in a specific slot).

4.9 Connectivity

I/O channels are part of the channel subsystem (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 10GbE RoCE Express, InterSystem Channel-3 (ISC-3), coupling through InfiniBand, or channel-to-channel (CTC) connections.

Communication to LANs is provided by the OSA-Express5S, OSA-Express4S, and OSA-Express 3 features.

Connectivity to I/O subsystems to exchange data is provided by FICON channels.

4.9.1 I/O feature support and configuration rules

Table 4-10 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 4-10 zEC12 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 Express8S LX/SX	2	2	320	160	Yes	FC, FCP
FICON Express8 LX/SX	4	4	176	44	Yes	FC, FCP

I/O feature	Number of		Maximum number of		PCHID	CHPID definition
	Ports per card	Port increments	Ports	I/O slots		
FICON Express4 ^a LX/SX	4	4	176	44	Yes	FC, FCP
OSA-Express5S 10 GbE LR/SR	2	2	96	48	Yes	OSD
OSA-Express5S GbE LX/SX	1	1	48	48	Yes	OSD, OSX
OSA-Express5S 1000BASE-T	2	2	96	48	Yes	OSE, OSD, OSC, OSN, OSM
OSA- Express4S 10 GbE LR/SR	1	1	48	48 ^b	Yes	OSD, OSX
OSA-Express4S GbE LX/SX	2	2	96	48 ^b	Yes	OSD
OSA-Express4S 1000BASE-T	2	2	96	48 ^b	Yes	OSE, OSD, OSC, OSN, OSM
OSA- Express3 10 GbE LR/SR	2	2	48	24 ^b	Yes	OSD, OSX
OSA-Express3 GbE LX/SX	4	4	96	24 ^b	Yes	OSD, OSN
OSA-Express3 1000BASE-T	4	4	96	24 ^b	Yes	OSE, OSD, OSC, OSN, OSM
10 GbE RoCE Express	16	1	16	16	Yes	N/A ^c
HCA3-O for 12x IFB and 12x IFB3	2	2	32	16	No	CIB
HCA2-O for 12x IFB	2	2	32	16	No	CIB
HCA3-O LR for 1x IFB	4	4	64	16	No	CIB
HCA2-O LR for 1x IFB	2	2	32	16	No	CIB
ISC-3 2 Gbps (10 km (6.2 miles))	2/ISC-D	1	48	12	Yes	CFP
ISC-3 1 Gbps (20 km (12.4 miles))	2/ISC-D	2	48	12	Yes	CFP

a. FICON Express4 4 km LX feature (FC 3324) is not supported on zEC12.

b. Each OSA-Express3 feature installed in an I/O cage/drawer reduces by two the number of OSA-Express4S/5S features allowed.

c. Defined by Function ID (FID).

At least one I/O feature (FICON) or one coupling link feature (IFB or ISC-3) must be present in the minimum configuration. A maximum of 256 channels is configurable per channel subsystem and per operating system image.

The following features can be shared and spanned:

- ▶ FICON channels that are defined as FC or FCP
- ▶ OSA-Express5S that are defined as OSC, OSD, OSE, OSM, OSN, or OSX
- ▶ OSA-Express4S that are defined as OSC, OSD, OSE, OSM, OSN, or OSX
- ▶ OSA-Express3 that are defined as OSC, OSD, OSE, OSM, OSN, or OSX
- ▶ Coupling links that are defined as CFP, ICP, or CIB
- ▶ HiperSockets that are defined as IQD

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 16 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 60 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 only one LPAR.
- ▶ 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 zEC12 installations and upgrades. Fiber optic conversion kits and mode conditioning patch cables are not orderable as features on zEC12 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 zEC12 for quick connection. The FQC is offered as a feature on zEC12 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 4-11 on page 157 lists the required connector and cable type for each I/O feature on the zEC12.

Table 4-11 I/O feature connector and cable types

Feature code	Feature name	Connector type	Cable type
0409	FICON Express8S LX 10 km	LC Duplex	9 μ m SM ^a
0410	FICON Express8S SX	LC Duplex	50, 62.5 μ m MM ^b
3325	FICON Express8 LX 10 km	LC Duplex	9 μ m SM
3326	FICON Express8 SX	LC Duplex	50, 62.5 μ m MM
3321	FICON Express4 LX 10 km	LC Duplex	9 μ m SM
3322	FICON Express4 SX	LC Duplex	50, 62.5 μ m MM
0415	OSA-Express5S 10 GbE LR	LC Duplex	9 μ m SM
0416	OSA-Express5S 10 GbE SR	LC Duplex	50, 62.5 μ m MM
0413	OSA-Express5S GbE LX	LC Duplex	9 μ m SM
0414	OSA-Express5S GbE SX	LC Duplex	50, 62.5 μ m MM
0417	OSA-Express5S 1000BASE-T	RJ-45	Category 5 UTP ^c
0404	OSA-Express4S GbE LX	LC Duplex	9 μ m SM
0405	OSA-Express4S GbE SX	LC Duplex	50, 62.5 μ m MM
0406	OSA-Express4S 10 GbE LR	LC Duplex	9 μ m SM
0407	OSA-Express4S 10 GbE SR	LC Duplex	50, 62.5 μ m MM
0408	OSA-Express4S 1000BASE-T	RJ-45	Category 5 UTP
3370	OSA-Express3 10 GbE LR	LC Duplex	9 μ m SM
3371	OSA-Express3 10 GbE SR	LC Duplex	50, 62.5 μ m MM
3362	OSA-Express3 GbE LX	LC Duplex	9 μ m SM
3363	OSA-Express3 GbE SX	LC Duplex	50, 62.5 μ m MM
3367	OSA-Express3 1000BASE-T	RJ-45	Category 5 UTP
0411	RoCE	LC Duplex	50, 62.5 μ m MM
0171	InfiniBand coupling (IFB)	MPO	50 μ m MM OM3 (2000 MHz-km)
0170	InfiniBand coupling (IFB LR)	LC Duplex	9 μ m SM
0163	InfiniBand coupling (IFB)	MPO	50 μ m MM OM3 (2000 MHz-km)
0168	InfiniBand coupling (IFB LR)	LC Duplex	9 μ m SM
0219	ISC-3	LC Duplex	9 μ m 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.

4.9.2 IBM ESCON channels

IBM ESCON® channels support the ESCON architecture and directly attach to ESCON supported I/O devices. The zEC12 server does not support the ESCON feature to attach to ESCON devices directly.

ESCON to FICON: zEC12 does not support ESCON features and does not offer ordering of ESCON channels on new builds, migration offerings, upgrades, and System z exchange programs. Enterprises need to migrate from ESCON to FICON. Alternate solutions are available for connectivity to ESCON devices.

IBM Global Technology Services (through IBM Facilities Cabling Services) offers ESCON to FICON migration services. For more information, see this website:

<http://www-935.ibm.com/services/us/index.wss/offering/its/c337386u66547p02>

The PRIZM Protocol Converter Appliance from Optica Technologies Incorporated provides a FICON-to-ESCON conversion function that is System z qualified. For more information, see this website:

<http://www.opticatech.com>

Vendor inquiries: IBM cannot confirm the accuracy of compatibility, performance, or any other claims by vendors for products that have not been System z qualified. Address any questions regarding these capabilities and device support to the suppliers of these products.

4.9.3 FICON channels

The FICON Express8S, FICON Express8, and FICON Express4³ features conform to the following architectures:

- ▶ Fibre Connection (FICON)
- ▶ High Performance FICON on System z (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.

Important: FICON Express and FICON Express2 features installed in previous systems are not supported on a zEC12 and cannot be carried forward on an upgrade.

Each FICON Express8 or FICON Express4 feature occupies one I/O slot in the I/O cage or I/O drawer. Each feature has four ports, each supporting an LC Duplex connector, with one PCHID and one CHPID associated with each port.

Each 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.

³ FICON Express4 4 km LX (FC 3324) is not supported on zEC12.

All FICON Express8S, FICON Express8, and FICON Express4 features use SFP optics that allow for concurrent repair 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 Express8S, FICON Express8, and FICON Express4 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.

All FICON Express8S, FICON Express8, and FICON Express4 features support 24k I/O devices (subchannels) for base and alias devices.

Each FICON Express8S, FICON Express8, and FICON Express4 channel can be defined independently for connectivity to servers, switches, directors, disks, tapes, and printers:

- ▶ CHPID type 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 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.

FICON Express8S

The FICON Express8S feature is 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 two types of FICON Express8S optical transceivers that are supported are the long wavelength (LX) and the short wavelength (SX):

- ▶ 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 a 1300 nanometer (nm) 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 Express8S SX feature uses an 850 nanometer (nm) optical transceiver, which supports various distances depending on the fiber used (50 or 62.5-μm multimode fiber).

Auto-negotiation: FICON Express8S features do not support auto-negotiation to a data link rate of 1 Gbps. The FICON Express8S feature will be the last FICON feature that supports auto-negotiation to a data link rate of 2 Gbps.

FICON Express8

The FICON Express8 feature can be in an I/O cage or I/O drawer. 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 two types of FICON Express8 optical transceivers that are supported are the long wavelength (LX) and the short wavelength (SX):

- ▶ 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 a 1300 nanometer (nm) fiber bandwidth transceiver. This transceiver 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 850 nanometer (nm) optical transceiver. This transceiver supports various distances, depending on the fiber used (50 or 62.5- μ m multimode fiber).

Limitation: FICON Express8 features do not support auto-negotiation to a data link rate of 1 Gbps.

FICON Express4

The FICON Express4 feature can be in an I/O cage or I/O drawer. Each of the four independent ports is capable of 1 Gbps, 2 Gbps, or 4 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 IBM zEnterprise EC12 is planned to be the last high-end System z server to offer support of the FICON Express4 features (FC 3321 and FC 3322). FICON Express4 will not be supported on future high-end System z servers as a carry-forward option on an upgrade. Enterprises need to continue upgrading from the FICON Express4 features to the FICON Express8S features (FC 0409 and FC 0410).

The two types of FICON Express4 optical transceivers that are supported are two long wavelength (LX) and one short wavelength (SX):

- ▶ FICON Express4 10 km LX feature FC 3321, with four ports per feature, supporting LC Duplex connectors
- ▶ FICON Express4 SX feature FC 3322, with four ports per feature, supporting LC Duplex connectors

FICON Express4: It is intended that the zEC12 and zBC12 are the last servers to support FICON Express4 features. Clients need to review the usage of their installed FICON Express4 channels and, where possible, migrate to FICON Express8S channels.

The FICON Express4 LX features use 1300 nanometer (nm) optical transceivers. One supports an unrepeated distance of 10 km (6.2 miles), and the other an unrepeated distance of 4 km (2.4 miles), by using 9 μ m single-mode fiber. Use of mode conditioning patch cables limits the link speed to 1 Gbps and the unrepeated distance to 550 meters (1804 ft).

The FICON Express4 SX feature uses 850 nanometer (nm) optical transceivers. These transceivers support various distances, depending on the fiber used (50 or 62.5- μ m multimode fiber).

Link speed: FICON Express4 is the last FICON family able to negotiate link speed down to 1 Gbps.

FICON feature summary

Table 4-12 shows the FICON card feature codes, cable type, maximum unrepeated distance, and the link data rate on a zEC12. All FICON features use LC Duplex connectors. For longwave FICON features that can use a data rate of 1 Gbps, mode conditioning patch cables (50 or 62.5 MM) can be used. The maximum distance for these connections is reduced to 550 m (1804 ft) at a link data rate of 1 Gbps.

Table 4-12 EC12 channel feature support

Channel feature	Feature codes	Bit rate	Cable type	Maximum unrepeated distance ^a
FICON Express8S 10KM LX	0409	2, 4, or 8 Gbps	SM 9 μ m	10 km
FICON Express8S SX	0410	8 Gbps	MM 62.5 μ m MM 50 μ m	21 m (200) 50 m (500) 150 m (2000)
		4 Gbps	MM 62.5 μ m MM 50 μ m	70 m (200) 150 m (500) 380 m (2000)
		2 Gbps	MM 62.5 μ m MM 50 μ m	150 m (200) 300 m (500) 500 m (2000)
FICON Express8 10KM LX	3325	2, 4, or 8 Gbps	SM 9 μ m	10 km
FICON Express8 SX	3326	8 Gbps	MM 62.5 μ m MM 50 μ m	21 m (200) 50 m (500) 150 m (2000)
		4 Gbps	MM 62.5 μ m MM 50 μ m	70 m (200) 150 m (500) 380 m (2000)
		2 Gbps	MM 62.5 μ m MM 50 μ m	150 m (200) 300 m (500) 500 m (2000)
FICON Express4 10KM LX	3321	1, 2, or 4 Gbps	SM 9 μ m	10 km/20 km

Channel feature	Feature codes	Bit rate	Cable type	Maximum unrepeat distance ^a
FICON Express4 SX	3322	4 Gbps	MM 62.5 µm MM 50 µm	70 m (200) 150 m (500) 380 m (2000)
		2 Gbps	MM 62.5 µm MM 50 µm	150 m (200) 300 m (500) 500 m (2000)
		1 Gbps	MM 62.5 µm MM 50 µm	300 m (200) 500 m (500) 860 m (2000)

a. Minimum fiber bandwidths in MHz/km for multimode fiber optic links are included in parentheses were applicable.

4.9.4 OSA-Express5S

The OSA-Express5S feature resides exclusively in the PCIe I/O drawer. The following OSA-Express5S features can be installed on zBC12 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 4-13 lists the OSA-Express5S features.

Table 4-13 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	0413	1	1	48	48	OSD, OSX
OSA-Express5S 10 GbE SR	0414	1	1	48	48	OSD, OSX
OSA-Express5S GbE LX	0415	2	2	96	48	OSD
OSA-Express5S GbE SX	0416	2	2	96	48	OSD
OSA-Express5S 1000BASE-T	0417	2	2	96	48	OSD, OSC, 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 channel path identifier (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 zEC12 to zEnterprise 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 a long-reach (LR) transceiver. The sending and receiving transceivers must be the same (LR to LR, which might also be referred to as LW or LX).

A 9 μm single-mode fiber optic cable terminated 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 channel path identifier (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 zEC12 to zEnterprise BladeCenter Extension (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 a Short Reach (SR) transceiver. The sending and receiving transceivers must be the same (SR to SR).

A 50 or a 62.5 μm multimode fiber optic cable terminated 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 1Gbps 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 a long-wavelength (LX) transceiver. The sending and receiving transceivers must be the same (LX to LX).

A 9 μm single-mode fiber optic cable terminated 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 cable for each end of the link.

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 a small form factor pluggable (SFP) with an RJ-45 receptacle for cabling to an Ethernet switch. The RJ-45 receptacle is required to be attached by using EIA/TIA Category 5 or Category 6 unshielded twisted pair (UTP) cable with a maximum length of 100 m (328 ft). The SFP allows 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 and 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, OSN, or OSM. Non-QDIO operation mode requires CHPID type OSE. 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 will attempt 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 will not connect.

4.9.5 OSA-Express4S features

This section addresses the characteristics of all OSA-Express4S features that are supported on zEC12 servers.

The OSA-Express4S feature is exclusively in the PCIe I/O drawer. The following OSA-Express4S features can be installed on zEC12 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 4-14 lists the characteristics of the OSA-Express4S features.

Table 4-14 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 ^a	OSD, OSX
OSA-Express4S 10 GbE SR	0407	1	1	48	48 ^b	OSD, OSX
OSA-Express4S GbE LX	0404	2	2	96 ^b	48 ^b	OSD
OSA-Express4S GbE SX	0405	2	2	96 ^b	48 ^b	OSD
OSA-Express4S 1000BASE-T	0408	2	2	96 ^b	48 ^b	OSC, OSD, OSE, OSM, OSN

a. Each OSA-Express3 feature installed in an I/O cage/drawer reduces by two the number of OSA-Express4S features allowed.

b. Both ports on each feature share one PCHID/CHPID.

OSA-Express4S Gigabit Ethernet LX (FC 0404)

The OSA-Express4S 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-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 is terminated 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) 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-Express4S GbE SX feature supports 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 is terminated with an LC Duplex connector is required for connecting each port on this feature to the selected device.

OSA-Express4S 10 Gigabit Ethernet LR (FC 0406)

The OSA-Express4S 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 zEC12 to zEnterprise BladeCenter Extension (zBX). The 10 GbE feature is designed to support attachment to a single mode fiber 10-Gbps 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 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, which might also be referred to as LW or LX).

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 is terminated 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) 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 zEC12 to zBX. The 10 GbE feature is designed to support attachment to a multimode fiber 10-Gbps 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 is terminated 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)

Feature code 0408 occupies one slot in the PCIe 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 EIA/TIA Category 5 or Category 6 unshielded twisted pair (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 intranode management network (INMN).

The following settings are supported on the OSA-Express3 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 will not connect.

Statement of Direction: The OSA-Express4S 1000BASE-T Ethernet feature is planned to be the last copper Ethernet feature to support half-duplex operation and a 10-Mbps link data rate. The zEnterprise EC12 servers are planned to be the last IBM System z servers to support half-duplex operation and a 10-Mbps link data rate for copper Ethernet environments. Any future 1000BASE-T Ethernet feature will support full-duplex operation and auto-negotiation to 100 or 1000 Mbps exclusively.

4.9.6 OSA-Express3 features

This section addresses the connectivity options that are offered by the OSA-Express3 features.

Statement of Direction: The IBM zEnterprise EC12 is planned to be the last high-end System z server to offer support of the Open System Adapter-Express3 (OSA-Express3 FC 3362, FC 3363, FC 3367, FC 3370, and FC 3371) family of features. OSA-Express3 will not be supported on future high-end System z servers as a carry-forward option on an upgrade. Continue upgrading from the OSA-Express3 features to the OSA-Express5S features (#0415, #0416, #0413, #0414, #0417).

The OSA-Express3 feature is exclusively in an I/O cage or I/O drawer. The following OSA-Express3 features can be installed on zEC12 servers:

- ▶ OSA-Express3 10 Gigabit Ethernet (GbE) Long Range (LR), FC 3370
- ▶ OSA-Express3 10 Gigabit Ethernet (GbE) Short Reach (SR), FC 3371
- ▶ OSA-Express3 Gigabit Ethernet (GbE) Long wavelength (LX), FC 3362
- ▶ OSA-Express3 Gigabit Ethernet (GbE) Short wavelength (SX), FC 3363
- ▶ OSA-Express3 1000BASE-T Ethernet, FC 3367

Table 4-15 on page 168 lists the characteristics of OSA-Express3 features.

Table 4-15 OSA-Express3 features

I/O feature	Feature code	Number of ports per feature	Port increment	Maximum number of ports	Maximum number of features	CHPID type
OSA-Express3 10 GbE LR	3370	2	2	48	24	OSD, OSX
OSA-Express3 10 GbE SR	3371	2	2	48	24	OSD, OSX
OSA-Express3 GbE LX	3362	4	4	96	24	OSD, OSN
OSA-Express3 GbE SX	3363	4	4	96	24	OSD, OSN
OSA-Express3 1000BASE-T	3367	4	4	96	24	OSC, OSD, OSE, OSN, OSM

OSA-Express3 10 GbE LR (FC 3370)

The OSA-Express3 10 GbE LR feature occupies one slot in the I/O cage or I/O drawer. It has two ports that connect to a 10-Gbps Ethernet LAN through a 9- μ m single mode fiber optic cable that is terminated with an LC Duplex connector. Each port on the card has a CHPID assigned. The feature supports an unrepeated maximum distance of 10 km (6.2 miles).

Tip: Each OSA-Express3 feature that is installed in an I/O cage/drawer reduces by two the number of OSA-Express4S features allowed.

Compared to the OSA-Express2 10 GbE LR feature, the OSA-Express3 10 GbE LR feature has double port density (two ports for each feature). This configuration improves performance for standard and jumbo frames.

The OSA-Express3 10 GbE LR feature does not support auto-negotiation to any other speed, and runs in full-duplex mode only. It supports only 64B/66B encoding as opposed to GbE, which supports 8B/10B encoding.

The OSA-Express3 10 GbE LR feature has two CHPIDs, with each CHPID having one port, and supports CHPID types OSD (QDIO mode) and OSX. CHPID type OSD is supported by z/OS, z/VM, z/VSE, Transaction Processing Facility (TPF), and Linux on System z to provide client-managed external network connections. CHPID type OSX is dedicated for connecting the zEC12 to an IEDN, providing a private data exchange path across ensemble nodes.

OSA-Express3 10 GbE SR (FC 3371)

The OSA-Express3 10 GbE SR feature (FC 3371) occupies one slot in the I/O cage or I/O drawer. It has two CHPIDs, with each CHPID having one port.

External connection to a 10-Gbps Ethernet LAN is done through a 62.5- μ m or 50- μ m multimode fiber optic cable that is terminated with an LC Duplex connector. The following distances are the maximum supported unrepeated distances:

- ▶ 33 meters (108 ft) on a 62.5- μ m multimode (200 MHz) fiber optic cable
- ▶ 82 meters (269 ft) on a 50- μ m multimode (500 MHz) fiber optic cable
- ▶ 300 meters (984 ft) on a 50- μ m multimode (2000 MHz) fiber optic cable

The OSA-Express3 10 GbE SR feature does not support auto-negotiation to any other speed, and runs in full-duplex mode only. OSA-Express3 10 GbE SR supports 64B/66B encoding, as opposed to GbE, which supports 8B/10 encoding, making auto-negotiation to any other speed impossible.

The OSA-Express3 10 GbE SR feature supports CHPID types OSD (QDIO mode) and OSX. CHPID type OSD is supported by z/OS, z/VM, z/VSE, TPF, and Linux on System z to provide client-managed external network connections. CHPID type OSX is dedicated for connecting the zEC12 to an IEDN, providing a private data exchange path across ensemble nodes.

OSA-Express3 GbE LX (FC 3362)

Feature code 3362 occupies one slot in the I/O cage or I/O drawer. It has four ports that connect to a 1-Gbps Ethernet LAN through a 9- μ m single mode fiber optic cable. This cable is terminated with an LC Duplex connector, and supports an unrepeated maximum distance of 5 km (3.2 miles). Multimode (62.5 or 50 μ m) fiber optic cable can be used with this feature.

Requirement: The use of these multimode cable types requires a mode conditioning patch cable at each end of the fiber optic link. The use of the single mode to multimode mode conditioning patch cables reduces the supported distance of the link to a maximum of 550 meters (1804 ft).

The OSA-Express3 GbE LX feature does not support auto-negotiation to any other speed, and runs in full-duplex mode only.

The OSA-Express3 GbE LX feature has two CHPIDs, with each CHPID (OSD or OSN) having two ports for a total of four ports per feature. Exploitation of all four ports requires operating system support.

OSA-Express3 GbE SX (FC 3363)

Feature code 3363 occupies one slot in the I/O cage or I/O drawer. It has four ports that connect to a 1-Gbps Ethernet LAN through a 50- μ m or 62.5- μ m multimode fiber optic cable. This cable is terminated with an LC Duplex connector over an unrepeated distance of 550 meters (1804 ft) (for 50- μ m fiber) or 220 meters (721.7 ft) (for 62.5- μ m fiber).

The OSA-Express3 GbE SX feature does not support auto-negotiation to any other speed, and runs in full-duplex mode only.

The OSA-Express3 GbE SX feature has two CHPIDs (OSD or OSN), with each CHPID having two ports for a total of four ports per feature. Exploitation of all four ports requires operating system support.

OSA-Express3 1000BASE-T Ethernet feature (FC 3367)

Feature code 3367 occupies one slot in the I/O cage or I/O drawer. It has four 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 is required to be attached by using EIA/TIA Category 5 or Category 6 UTP cable with a maximum length of 100 meters (328 ft).

The OSA-Express3 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. It then connects at the speed and duplex mode of the device at the other end of the cable.

The OSA-Express3 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-Express3 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 these settings do not match the speed and duplex mode of the signal on the cable, the OSA-Express port will not connect.

4.9.7 OSA-Express for ensemble connectivity

The following OSA-Express features are used to connect the zEC12 to its attached IBM zEnterprise BladeCenter Extension (zBX) Model 003 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, feature code 0408
- ▶ OSA-Express3 10 Gigabit Ethernet (GbE) Long Reach (LR), FC 3370
- ▶ OSA-Express3 10 Gigabit Ethernet (GbE) Short Reach (SR), FC 3371
- ▶ OSA-Express3 1000BASE-T Ethernet, FC 3367

Intraensemble data network (IEDN)

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
- ▶ OSA-Express3 10 GbE

For redundancy, one port each from two OSA-Express 10 GbE features must be configured.

The connection is from the zEC12 to the IEDN Top of Rack (ToR) switches on the zBX Model 003. With a stand-alone zEC12 node (no-zBX), the connection is interconnect pairs of OSX ports through LC Duplex directly connected cables, not wrap cables as has previously been recommended.

For more information about OSA-Express5S, OSA-Express4S, and OSA-Express3 in an ensemble network, see 7.4, “zBX connectivity” on page 244.

Intranode management network (INMN)

The INMN is a private and physically isolated 1000BASE-T Ethernet internal management network that operates at 1 Gbps. It connects all resources (zEC12 and zBX Model 003 components) of an ensemble node for management purposes. It is prewired, 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, OSA-Express4S 1000BASE-T features, or OSA-Express3 1000BASE-T features; CHPID port 1 is not used at all in this case) configured as CHPID type OSM. The connection is through port J07 of the bulk power hubs (BPHs) in the zEC12. The INMN ToR switches on zBX Model 003 also connect to the BPHs.

For more information about OSA-Express3 and OSA Express4S in an ensemble network, see 7.4, “zBX connectivity” on page 244.

Ensemble HMC management functions

An HMC can manage multiple System z servers and can be at a local or a remote site. If the zEC12 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 12.7, “HMC in an ensemble” on page 461 and 10.5, “RAS capability for the HMC and SE” on page 400.

4.9.8 HiperSockets

The HiperSockets function of IBM zEnterprise EC12 provides up to 32 high-speed virtual LAN attachments, just like the IBM zEnterprise BC12, IBM zEnterprise 196, and IBM zEnterprise 114 servers. Previous servers provided 16 attachments.

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 zEC12 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.

Datagrams are 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 can also be used to connect to the HiperSockets Layer 2 networks of different servers.

HiperSockets Layer 2 in the zEC12 is supported by Linux on System z, and by z/VM for Linux guest exploitation.

zEC12, and the other zEnterprise CPCs (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 available in its inbound queue. HiperSockets Completion Queue function requires the following applications at a minimum:

- ▶ z/OS V1.13
- ▶ Linux on System z distributions:
 - Red Hat Enterprise Linux (RHEL) 6.2
 - SUSE Linux Enterprise Server (SLES) 11 SP2
- ▶ z/VSE 5.1.1
- ▶ z/VM 6.2 with maintenance

The zEC12 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 System z networks, the combination allows System z virtual servers to use the optimal path for communications.

In z/VM 6.2, 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 also directly communicate with these systems:

- ▶ Other guest virtual machines on the virtual switch
- ▶ External network hosts through the virtual switch OSA UPLINK port

Statement of Direction fulfillment: In z/VM 6.2 with a short programming enhancement (SPE), IBM provided support for both synchronous and asynchronous transport of HiperSockets messages on the zEnterprise systems, therefore fulfilling this Statement of Direction.

4.10 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 System z technology solution to achieve near-continuous availability. A Parallel Sysplex comprises one or more z/OS operating system images that are coupled through one or more coupling facilities.

4.10.1 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:

- ▶ **IFB:** Parallel Sysplex using InfiniBand (IFB) connects to zEC12, zBC12, z196, z114, or System z10 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.

zEC12 supports two types of 12x InfiniBand coupling links:

- FC 0171 HCA3-O (12xIFB) fanout
- FC 0163 HCA2-O (12xIFB) fanout

- ▶ **IFB LR:** IFB LR (Long Reach) connects to zEC12, zBC12, z196, z114, or System z10 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 System z 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.

zEC12 supports two types of 1x InfiniBand coupling links:

- FC 0170 HCA3-O LR (1xIFB) fanout
- FC 0168 HCA2-O LR (1xIFB) fanout

- ▶ **Internal coupling (IC):** CHPIDs (type ICP) defined for internal coupling can connect a CF to a z/OS LPAR in the same zEC12. IC connections require two CHPIDs to be defined, which can be defined only in peer mode. The bandwidth is greater than 2 GBps. A maximum of 32 IC CHPIDs (16 connections) can be defined.

- **ISC-3:** The InterSystem Channel-3 (ISC-3) type is available in peer mode only. ISC-3 links can be used to connect to zEC12, zBC12, z196, z114, or System z10. These optic fiber links support a maximum distance of 10 km (6.2 miles), 20 km (12.4 miles) with RPQ 8P2197, and 100 km (62 miles) with a System z qualified DWDM. ISC-3 supports 9- μ m single mode fiber optic cabling. The link data rate is 2 Gbps at distances up to 10 km (6.2 miles), and 1 Gbps when RPQ 8P2197 is installed. Each port operates at 2 Gbps. Ports are ordered in increments of one. The maximum number of ISC-3 links per zEC12 is 48. ISC-3 supports transmission of Server Time Protocol (STP) messages.

ISC-3: The zEC12 and zBC12 are intended to be the last servers that will support ISC-3 coupling links. Review the usage of your installed ISC-3 coupling links, and where possible upgrade to IFB (FC 0163 and FC 0171) or IFB LR (FC 0168 and FC 0170) coupling links.

Table 4-16 shows the coupling link options.

Table 4-16 Coupling link options

Type	Description	Use	Link rate	Distance	zEC12-H20 maximum	zEC12-H43 to HA1 maximum
IFB	12x InfiniBand (HCA3-O) ^a	zEC12 to zEC12, zBC12, z196, z114, z10	6 Gbps	150 meters (492 feet)	16 ^b	32
	12x InfiniBand (HCA2-O)	zEC12 to zEC12, zBC12, z196, z114, z10	6 Gbps	150 meters (492 feet)	16 ^b	32
IFB LR	1x IFB (HCA3-O LR)	zEC12 to zEC12, zBC12, z196, z114, z10	2.5 Gbps 5.0 Gbps	10 km unrepeated (6.2 miles) 100 km repeated (62 miles)	32 ^b	64
	1x IFB (HCA2-O LR)	zEC12 to zEC12, zBC12, z196, z114, z10	2.5 Gbps 5.0 Gbps	10 km unrepeated (6.2 miles) 100 km repeated (62 miles)	16 ^b	32
IC	Internal coupling channel	Internal communication	Internal speeds	N/A	32	32
ISC-3	InterSystem Channel-3	zEC12 to zEC12, zBC12, z196, z114, z10	2 Gbps	10 km unrepeated (6.2 miles) 100 km repeated (62 miles)	48	48

- a. 12x IFB3 protocol supports a maximum of 4 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 4.7.5, "HCA3-O (12x IFB) fanout (FC 0171)" on page 145.
- b. Uses all available fanout slots. Allows no other I/O or coupling.

The maximum for IFB links is 64. The maximum number of combined external coupling links (active ISC-3 links and IFB LR) cannot exceed 112 per server. There is a maximum of 128 coupling CHPIDs limitation, including ICP for IC, CIB for IFB and IFB LR, and CFP for ISC-3.

The zEC12 supports various connectivity options, depending on the connected zEC12, zBC12, z196, z114, or System z10 server. Figure 4-15 shows zEC12 coupling link support for zEC12, zBC12, z196, z114, and System z10 servers.

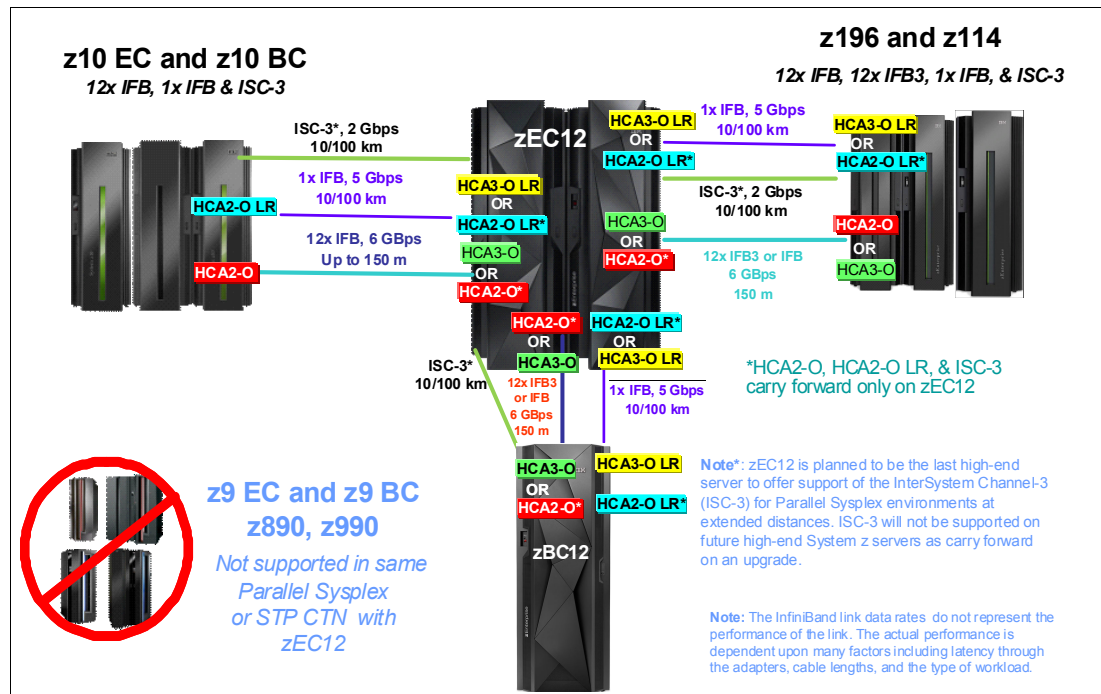


Figure 4-15 zEC12 Parallel Sysplex coupling connectivity

When defining IFB coupling links (CHPID type CIB), HCD now defaults to seven subchannels. Thirty-two subchannels are only supported on HCA2-O LR (1xIFB) and HCA3-O LR (1xIFB) on zEC12 when both sides of the connection use 32 subchannels. Otherwise, change the default value from 7 to 32 subchannels on each CIB definition.

z/OS and coupling facility images can run on the same or on separate servers. There must be at least one CF 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 zEC12 supports five types of coupling link options:

- ▶ HCA3-O fanout for 12x InfiniBand, FC 0171
- ▶ HCA2-O fanout for 12x InfiniBand, FC 0163
- ▶ HCA3-O LR fanout for 1x InfiniBand, FC 0170
- ▶ HCA2-O LR fanout for 1x InfiniBand, FC 0168
- ▶ InterSystem Channel-3 (ISC-3), FC 0217, FC 0218, and FC 0219

The coupling link features available on the zEC12 connect the zEC12 servers to the identified System z servers by various link options:

- ▶ 12x InfiniBand using HCA3-O and HCA2-O (12xIFB) fanout card at 6 Gbps to zEC12, zBC12, z196, z114, and System z10
- ▶ 1x InfiniBand using both HCA3-O LR and HCA2-O LR at 5.0 or 2.5 Gbps to zEC12, zBC12, z196, z114, and System z10 servers
- ▶ ISC-3 at 2 Gbps to zEC12, zBC12, IBM zEnterprise 196 (z196), z114, and System z10

HCA2-O fanout for 12x InfiniBand (FC 0163)

For more information, see 4.7.3, “HCA2-O (12x IFB) fanout (FC 0163)” on page 143.

HCA2-O LR fanout for 1x InfiniBand (FC 0168)

For more information, see 4.7.4, “HCA2-O LR (1x IFB) fanout (FC 0168)” on page 144.

HCA3-O fanout for 12x InfiniBand (FC 0171)

For more information, see 4.7.5, “HCA3-O (12x IFB) fanout (FC 0171)” on page 145.

HCA3-O LR fanout for 1x InfiniBand (FC 0170)

For more information, see 4.7.6, “HCA3-O LR (1x IFB) fanout (FC 0170)” on page 146.

Internal coupling links

IC links are LIC-defined links to connect a CF to a z/OS LPAR in the same server. These links are available on all System z servers. The IC link is a System z server coupling connectivity option. It 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 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 System z servers.
- ▶ Provide the fastest connectivity, significantly 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 System z server.

ISC-3 coupling links

Three feature codes are available to implement ISC-3 coupling links:

- ▶ FC 0217, ISC-3 mother card
- ▶ FC 0218, ISC-3 daughter card
- ▶ FC 0219, ISC-3 port

The ISC mother card (FC 0217) occupies one slot in the I/O cage or I/O drawer, and supports up to two daughter cards. The ISC daughter card (FC 0218) provides two independent ports with one CHPID associated with each enabled port. The ISC-3 ports are enabled and activated individually (one port at a time) by Licensed Internal Code (LIC).

When the quantity of ISC links (FC 0219) is selected, the quantity of ISC-3 port features selected determines the appropriate number of ISC-3 mother and daughter cards. There is a maximum of 12 ISC-M cards.

Each active ISC-3 port in peer mode supports a 2 Gbps (200 MBps) connection through 9- μ m single mode fiber optic cables that are terminated with an LC Duplex connector. The maximum unrepeated distance for an ISC-3 link is 10 km (6.2 miles). With repeaters, the maximum distance extends to 100 km (62 miles). ISC-3 links can be defined as *timing-only links* when STP is enabled. Timing-only links are coupling links that allow two servers to be synchronized by using STP messages when a CF does not exist at either end of the link.

Statement of Direction: The IBM zEnterprise EC12 and the zBC12 are planned to be the last System z servers to offer support of the InterSystem Channel-3 (ISC-3) for Parallel Sysplex environments at extended distances. ISC-3 will not be supported on future System z servers as a carry-forward option on an upgrade. Enterprises need to continue migrating from ISC-3 features (FC 0217, FC 0218, and FC 0219) to 12x InfiniBand (FC 0171 - HCA3-O fanout) or 1x InfiniBand (FC 0170 - HCA3-O LR fanout) coupling links.

RPQ 8P2197 extended distance option

The RPQ 8P2197 daughter card provides two ports that are active and enabled when installed, and do not require activation by LIC.

This RPQ allows the ISC-3 link to operate at 1 Gbps (100 MBps) instead of 2 Gbps (200 MBps). This lower speed allows an extended unrepeated distance of 20 km (12.4 miles). One RPQ daughter is required on both ends of the link to establish connectivity to other servers. This RPQ supports STP if defined as either a coupling link or timing-only.

Coupling link considerations

For more information about changing to InfiniBand coupling links, see the *Coupling Facility Configuration Options* white paper:

<http://www.ibm.com/systems/z/advantages/ps0/whitepaper.html>

Coupling links and Server Time Protocol (STP)

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 links, can meet more stringent synchronization requirements than servers that exchange messages over long ISC-3 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 necessary in a Parallel Sysplex. Therefore, there is a potential benefit of minimizing the number of cross-site links 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 zEC12 server does not support attachment to the IBM Sysplex Timer. A zEC12 can be added into a Mixed Coordinated Timing Network (CTN) only when there is a System z10 attached to the Sysplex Timer operating as Stratum 1 server. Connections to two Stratum 1 servers are preferable to provide redundancy and avoid a single point of failure.

Important: A Parallel Sysplex in an external time reference (ETR) network *must* change to Mixed CTN or STP-only CTN *before* introducing a zEC12.

STP recovery enhancement

The new generation of host channel adapters (HCA3-O (12xIFB) and HCA3-O LR (1xIFB)), introduced for coupling, is designed to send a reliable unambiguous “going away signal.” This signal indicates that the server on which the HCA3 is running is about to enter a failed (check stopped) state. 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 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 zEnterprise CPCs. It is available only if you have an HCA3-O (12x IFB) or HCA3-O LR (1x IFB) on the CTS communicating with an 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.

Important: For more information about configuring an STP CTN with three or more servers, see the white paper:

<http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP101833>

If the guidelines are not followed, this 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 resources:

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

4.10.2 External clock facility

The external clock facility (ECF) card in the CPC cage provides a Pulse Per Second (PPS) input. This PPS signal can be received from a Network Time Protocol (NTP) server that acts as an external time source (ETS). Two ECF cards are installed in the card slots above the books to provide redundancy for continued operation and concurrent maintenance when a single ECF card fails. Each ECF card has a Bayonet Neill-Concelman (BNC) connector for PPS connection support, attaching to two different ETSs. Two PPS connections from two different NTP servers are preferable for redundancy.

The time accuracy of an STP-only CTN is improved by adding an NTP server with the PPS output signal as the ETS device. STP tracks the highly stable accurate PPS signal from the NTP server. It maintains accuracy of 10 μ s as measured at the PPS input of the zEC12 server. If STP uses an NTP server without PPS, a time accuracy of 100 ms to the ETS is maintained. NTP servers with PPS output are available from various vendors that offer network timing solutions.

4.11 Cryptographic functions

Cryptographic functions are provided by the CP Assist for Cryptographic Function (CPACF) and the PCI Express cryptographic adapters. zEC12 supports the Crypto Express4S feature, and, on a carry-forward only basis, the Crypto Express3 cards when upgrading from earlier generations.

4.11.1 CPACF functions (FC 3863)

Feature code (FC) 3863 is required to enable CPACF functions.

4.11.2 Crypto Express4S feature (FC 0865)

Crypto Express4S is an optional zEC12 feature that was not available in previous generations. On the initial order, a minimum of two features are installed. Then, the number of features increases one at a time up to a maximum of 16 features. Each Crypto Express4S 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 Express4S feature occupies one I/O slot in the PCIe I/O drawer, and it has no CHPID assigned. However, it uses one PCHID.

4.11.3 Crypto Express3 feature (FC 0864)

Crypto Express3 is an optional feature, and it is available only on a carry-forward basis when you are upgrading from earlier generations to zEC12. The minimum number of carry-forward features is two, and the maximum number that is supported is eight features. Each Crypto Express3 feature holds two PCI Express cryptographic adapters. Either of the adapters can be configured by the installation as a Secure IBM CCA coprocessor or as an accelerator.

Statement of Direction: The IBM zEnterprise EC12 generation is planned to be the last high-end System z server to offer support of the Crypto Express3 feature (FC 0864). You need to upgrade from the Crypto Express3 feature to the Crypto Express4S feature (FC 0865).

Each Crypto Express3 feature occupies one I/O slot in the I/O cage or in the I/O drawer. It has no CHPIDs assigned, but it uses two PCHIDS.

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

4.12 Integrated firmware processor

The integrated firmware processor (IFP) is introduced by zEC12 and zBC12 servers. The IFP is reserved for managing a new generation of PCIe features. These new features are installed exclusively into the PCIe I/O drawer:

- ▶ zEDC Express
- ▶ 10 GbE RoCE Express

All native PCIe features need to 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 (RGs) managed by the IFP. If two features of same type are installed, one is managed by resource group 1 (RG 1) and the other feature is managed by resource group 2 (RG 2). This provides redundancy in case of maintenance or failure in one of the features or resource groups.

The IFP provides support for the following infrastructure management functions:

- ▶ Handle adapter layer functions
- ▶ Firmware update
- ▶ Error recovery
- ▶ Adapter test
- ▶ Maintenance functions

For detailed information about the IFP and RGs, see Appendix F, “Native Peripheral Component Interconnect Express (PCIe)” on page 529.

4.13 Flash Express

The Flash Express cards are supported in PCIe I/O drawer with other PCIe I/O cards. They are plugged into PCIe I/O drawers in pairs for availability. Like the Crypto Express4S cards, each card takes up a CHPID, and no HCD/IOCP definition is required. Flash Express subchannels are predefined, and are allocated from the .25K reserved in subchannel set 0.

Flash Express cards are internal to the CPC, and are accessible by using the new System z architected 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.

zEC12 supports a maximum of four pairs of Flash Express cards. Only one Flash Express card is allowed per domain. The PCIe 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 slot of the PCIe I/O drawer before being plugged into the rear of drawer. These four slots are reserved for Flash Express and must not be filled by other types of I/O cards until there is no spare slot.

Figure 4-16 shows a PCIe I/O drawer that is fully populated with Flash Express cards.

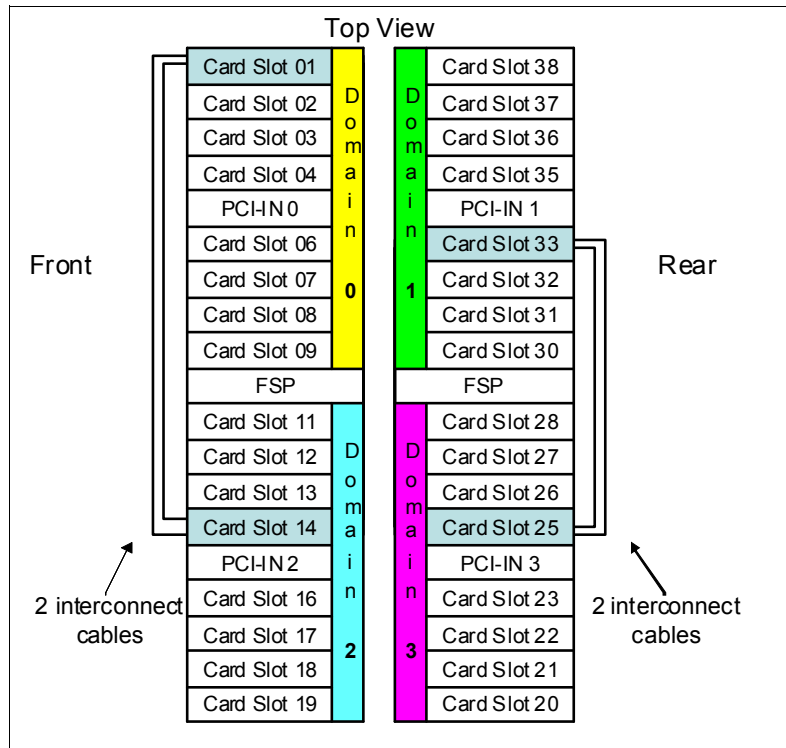


Figure 4-16 PCIe I/O drawer that is fully populated with Flash Express cards

4.14 10 GbE RoCE Express

Feature code 0411 RoCE Express resides exclusively in the PCIe I/O drawer (FC 4009) and is exclusive to the zEC12 and zBC12. 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 Hardware Configuration Definition (HCD). Each feature must be dedicated to an LPAR. Only one of the two ports is supported by z/OS.

The 10GbE RoCE Express feature utilizes a short reach (SR) laser as the optical transceiver, and supports use of a multimode fiber optic cable terminated with an LC Duplex connector.

Both point-to-point connection and switched connection 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 need to support the following requirements:

- ▶ Global Pause function enabled
- ▶ Priority Flow Control (PFC) disabled
- ▶ No firewalls, no routing, and no IEDN

The maximum supported unrepeated 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 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 meters (984 ft))
- ▶ OM2 50 micron multimode fiber optic cable rated at 500 MHz-km terminated with an LC Duplex connector (supports 82 meters (269 ft))
- ▶ OM1 62.5 micron multimode fiber optic cable rated at 200 MHz-km terminated with an LC Duplex connector (supports 33 meters (108 ft))

4.15 zEDC Express

The zEDC Express is an optional feature (FC 0420), exclusive to the zEC12 and zBC12. It is designed to provide hardware-based acceleration for data compression and decompression.

The feature installs exclusively on the PCIe I/O drawer. Up to two zEDC Express features can be installed per PCIe I/O drawer domain. However, if the domain contains a Flash Express or 10GbE RoCE feature, only one zEDC feature can be installed on that domain.

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 LPARs.

Adapter support for zEDC is provided by Resource Group (RG) code running on the system IFP. For resilience, there are always two independent RGs on the system, sharing the IFP. It is, therefore, suggested that a minimum of two zEDC features be installed, one per RG. For the best data throughput and availability, two features per RG, for a total of four features, need to be installed.

Exploitation support of zEDC Express functionality is provided by z/OS V2R1 for both data compression and decompression. Support for data recovery (decompression), if zEDC is not available on the system, is provided via software on z/OS V2R1, V1R13, and V1R12 with the appropriate PTFs. Software decompression is slow and uses considerable processor resources; therefore, it is not suggested for production environments.

Statement of Direction: IBM plans to add exploitation support of zEDC Express in z/VM 6.3.

For detailed information about zEDC, see Appendix E, “zEnterprise Data Compression (zEDC) Express” on page 525.

IBM System z Batch Network Analyzer

The IBM System z 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, provides graphical and text reports, including Gantt charts, and support for alternate 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 can be candidates for zEDC
- ▶ Those jobs that might be zEDC candidates but are not in extended format

Therefore, zBNA can help estimate utilization of zEDC features and help size the number of features needed:

- ▶ IBM Employees can obtain zBNA and other Customer Support Plans (CSP) tools via the IBM intranet:

<http://w3-03.ibm.com/support/techdocs/atmsastr.nsf/WebIndex/PRS5126>

- ▶ IBM Business Partners can obtain zBNA and other CSP tools via the Internet:

https://www.ibm.com/partnerworld/wps/servlet/mem/ContentHandler/tech_PRS5133

- ▶ IBM clients can obtain zBNA and other CSP tools via the Internet:

<http://www-03.ibm.com/support/techdocs/atmsastr.nsf/WebIndex/PRS5132>



Central processor complex channel subsystem

This chapter addresses the concepts of the IBM zEnterprise EC12 (zEC12) channel subsystem, including multiple channel subsystems. Also described are 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
- ▶ System-initiated CHPID reconfiguration
- ▶ Multipath initial program load (IPL)

5.1 Channel subsystem

The role of the channel subsystem (CSS) is to control communication of internal and external channels to control units and devices. The CSS configuration defines the operating environment for the correct execution of all system I/O operations.

The CSS provides the server communications to external devices through channel connections. The channels transfer data between main storage and I/O devices or other servers under the control of a channel program. The CSS allows channel I/O operations to continue independently of other operations within the central processors (CPs) and Integrated Facility for Linux processors (IFLs).

The building blocks that make up a channel subsystem are shown in Figure 5-1.

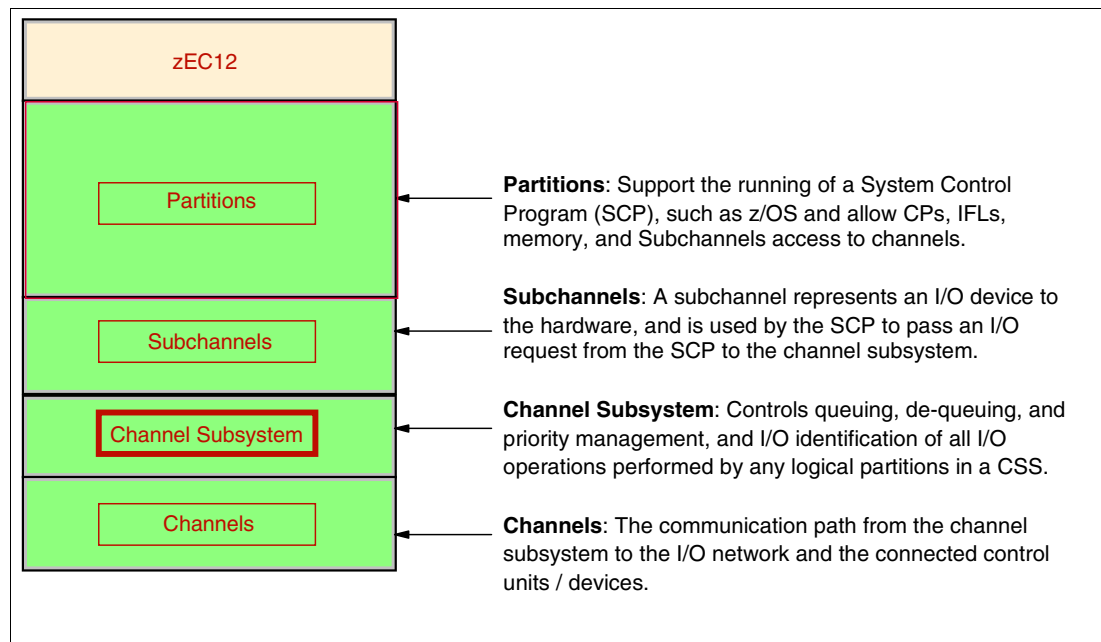


Figure 5-1 Channel subsystem overview

5.1.1 Multiple channel subsystems concept

The design of System z systems offers considerable processing power, memory size, and I/O connectivity. In support of the larger I/O capability, the CSS concept is scaled up correspondingly. The increased size provides relief for the number of supported logical partitions (LPARs), channels, and devices available to the system.

A single channel subsystem allows the definition of up to 256 channel paths. To overcome this limit, the multiple channel subsystems concept was introduced. The architecture provides for up to four channel subsystems. The structure of the multiple CSSs provides channel connectivity to the defined LPARs in a manner that is transparent to subsystems and application programs. This configuration enables the definition of a balanced configuration for the processor and I/O capabilities.

Each CSS can have from 1 to 256 channels, and can be configured to 1 to 15 LPARs. Therefore, four CSSs support a maximum of 60 LPARs. CSSs are numbered from 0 to 3, which is referred to as the *CSS image ID* (CSSID 0, 1, 2, or 3). These CSSs are also referred to as *logical channel subsystems* (LCSSs).

5.1.2 CSS elements

The CSS is composed of the following elements:

- ▶ Subchannels
- ▶ Channel paths
- ▶ Channel path identifier
- ▶ Control units
- ▶ I/O devices

Subchannels

A *subchannel* provides the logical representation of a device to a program. It contains the information that is required for sustaining a single I/O operation. A subchannel is assigned for each device that is defined to the LPAR.

Multiple subchannel sets, described in 5.1.3, “Multiple subchannel sets” on page 188, are available to increase addressability. Three subchannel sets per CSS are supported on a zEC12. Subchannel set 0 can have up to 63.75 K subchannels, and subchannel sets 1 and 2 can have up to 64 K minus 1 subchannel each.

Channel paths

Each CSS can have up to 256 channel paths. A *channel path* is a single interface between a server and one or more control units. Commands and data are sent across a channel path to run I/O requests.

Channel path identifier

Each channel path in the system is assigned a unique identifier value that is known as a *channel path identifier* (CHPID). A total of 256 CHPIDs are supported by the CSS, and a maximum of 1024 are supported per system (CPC).

The channel subsystem communicates with I/O devices through channel paths between the channel subsystem and control units. On System z, a CHPID number is assigned to a physical location (slot/port) by the client, by using the hardware configuration definition (HCD) tool or input/output configuration program (IOCP).

Control units

A *control unit* provides the logical capabilities 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 system 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.

5.1.3 Multiple subchannel sets

Do not confuse the multiple subchannel sets (MSS) functionality with multiple channel subsystems. In most cases, a *subchannel* represents an addressable device. For example, a disk control unit with 30 drives uses 30 subchannels for base addresses. An addressable device is associated with a device number, which is commonly (but incorrectly) known as the device address.

Subchannel numbers

Subchannel numbers (including their implied path information to a device) are limited to four hexadecimal digits by the architecture (0x0000 to 0xFFFF). Four hexadecimal digits provide 64 K addresses, which are known as a *set*.

IBM has reserved 256 subchannels, leaving over 63 K subchannels for general use¹. Again, addresses, device numbers, and subchannels are often used as synonyms, although this is not technically accurate. You might hear that there is a *maximum of 63.75 K addresses* or a *maximum of 63.75 K device numbers*.

The processor architecture allows for sets of subchannels (addresses), with a current implementation of three sets. Each set provides 64 K addresses. Subchannel set 0, the first set, reserves 256 subchannels for IBM use. Each of subchannel sets 1 and 2 provides 64 K minus one subchannel. In principle, subchannels in either set can be used for any device-addressing purpose. These subchannels are referred to as *special devices*.

Figure 5-2 summarizes the multiple channel subsystems and multiple subchannel sets.

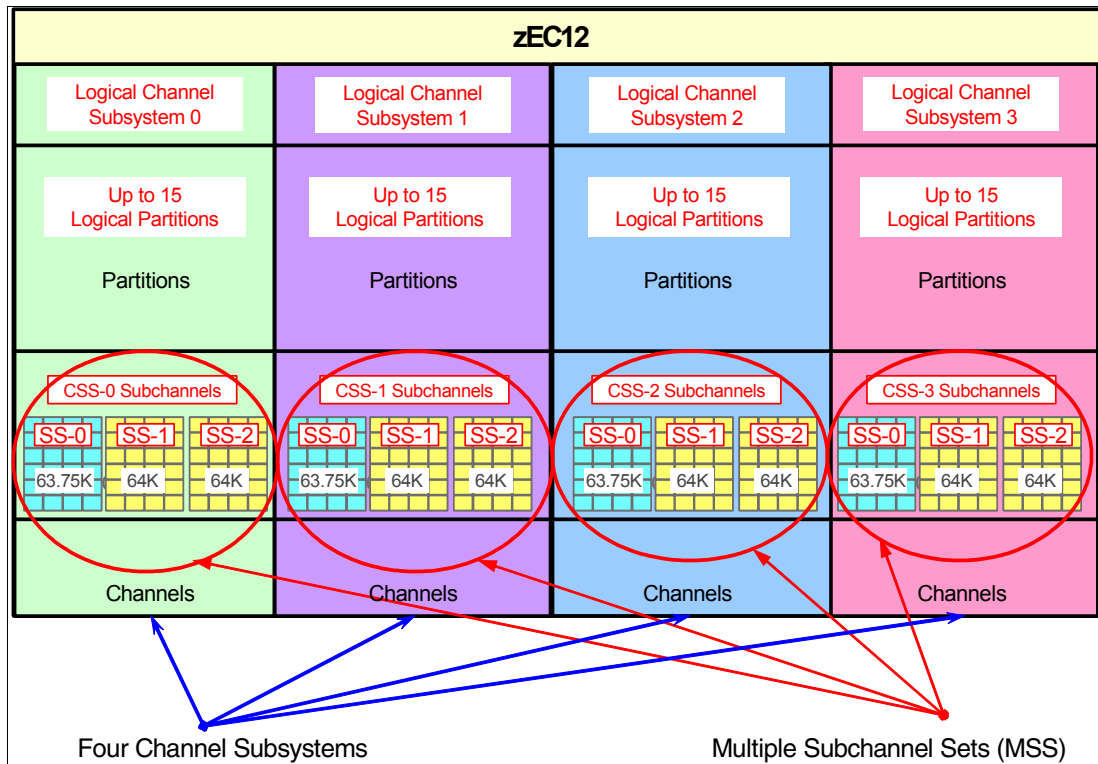


Figure 5-2 Multiple channel subsystems and multiple subchannel sets

¹ The number of reserved subchannels is 256. This is abbreviated to 63.75 K in this discussion to easily differentiate it from the 64-K minus one subchannel available in subchannel sets 1 and 2. The informal name, 63.75 K subchannel, represents the following equation: $(63 \times 1024) + (0.75 \times 1024) = 65280$.

The additional subchannel sets, in effect, add an extra high-order digit (either 0, 1, or 2) to existing device numbers. For example, you might think of an address as 08000 (subchannel set 0), 18000 (subchannel set 1), or 28000 (subchannel set 2). Adding a digit is not done in system code or in messages because of the architectural requirement for four-digit addresses (device numbers or subchannels). However, certain messages contain the subchannel set number. You can mentally use that as a high-order digit for device numbers. Only a few requirements refer to the subchannel sets 1 and 2 because they are only used for these special devices. JCL, messages, and programs rarely refer directly to these special devices.

Moving these special devices into an alternate subchannel set creates more space for device number growth. The appropriate subchannel set number must be included in the input/output configuration program (IOCP) definitions or in the HCD definitions that produce the input/output configuration data set (IOCDS). The subchannel set number defaults to zero.

IPL from an alternate subchannel set

zEC12 supports IPL from subchannel set 1 (SS1) or subchannel set 2 (SS2), in addition to subchannel set 0. Devices that are used early during IPL processing can now be accessed by using subchannel set 1 or subchannel set 2. This configuration allows the users of Metro Mirror (PPRC) secondary devices defined using the same device number and a new device type in an alternate subchannel set to be used for IPL, I/O definition file (IODF), and stand-alone dump volumes when needed.

IPL from an alternate subchannel set is supported by z/OS V1.13 or later, and V1.12 and V1.11 with program temporary fixes (PTFs). IPL applies to the Fibre Channel connection (FICON) and High Performance FICON for System z (zHPF) protocols.

The display ios,config command

The z/OS **display ios,config(all)** command that is shown in Figure 5-3 on page 190 includes information about the MSSs.

```

D IOS,CONFIG(ALL)
IOS506I 18.21.37 I/O CONFIG DATA 610
ACTIVE IODF DATA SET = SYS6.IODF45
CONFIGURATION ID = TESTxxxx EDT ID = 01
TOKEN:  PROCESSOR DATE      TIME      DESCRIPTION
SOURCE: SCZP201 12-03-04 09:20:58 SYS6      IODF45
ACTIVE CSS: 0 SUBCHANNEL SETS CONFIGURED: 0, 1, 2
CHANNEL MEASUREMENT BLOCK FACILITY IS ACTIVE
HARDWARE SYSTEM AREA AVAILABLE FOR CONFIGURATION CHANGES
PHYSICAL CONTROL UNITS          8131
CSS 0 - LOGICAL CONTROL UNITS   4037
  SS 0 SUBCHANNELS              62790
  SS 1 SUBCHANNELS              61117
  SS 2 SUBCHANNELS              60244
CSS 1 - LOGICAL CONTROL UNITS   4033
  SS 0 SUBCHANNELS              62774
  SS 1 SUBCHANNELS              61117
  SS 2 SUBCHANNELS              60244
CSS 2 - LOGICAL CONTROL UNITS   4088
  SS 0 SUBCHANNELS              65280
  SS 1 SUBCHANNELS              65535
  SS 2 SUBCHANNELS              62422
CSS 3 - LOGICAL CONTROL UNITS   4088
  SS 0 SUBCHANNELS              65280
  SS 1 SUBCHANNELS              65535
  SS 2 SUBCHANNELS              62422
ELIGIBLE DEVICE TABLE LATCH COUNTS
      0 OUTSTANDING BINDS ON PRIMARY EDT

```

Figure 5-3 Output for display ios,config(all) command with MSS

5.1.4 Parallel access volumes and extended address volumes

Parallel access volume (PAV) support enables a single System z system to simultaneously process multiple I/O operations to the same logical volume. This feature can help to significantly reduce device queue delays. Dynamic PAV allows the dynamic assignment of aliases to volumes to be under Workload Manager (WLM) control.

With the availability of HyperPAV, the requirement for PAV devices is greatly reduced. HyperPAV allows an alias address to be used to access any base on the same control unit image per I/O base. It also allows different HyperPAV hosts to use one alias to access different bases, which reduces the number of alias addresses required. HyperPAV is designed to enable applications to achieve equal or better performance than the original PAV feature alone, while also using the same or fewer z/OS resources. HyperPAV is an optional feature on the IBM DS8000® series.

To further reduce the complexity of managing large I/O configurations, System z introduced extended address volume (EAV). EAV is designed to build large disk volumes by using virtualization technology. By being able to extend the disk volume size, a client might potentially need fewer volumes to hold the data. This configuration makes systems management and data management less complex.

5.1.5 Logical partition name and identification

No logical partitions (LPARs) can exist without at least one defined CSS. LPARs are defined to a CSS, not to a system. An LPAR is associated with one CSS only.

An LPAR is identified through its name, its identifier, and its multiple image facility (MIF) image ID (MIF ID). The LPAR name is defined through HCD or the IOCP, and is the partition name in the RESOURCE statement in the configuration definitions. Each name must be unique across the CPC.

The LPAR identifier is a number in the range of 00 - 3F. It is assigned by the user on the image profile through the Support Element (SE) or the Hardware Management Console (HMC). It is unique across the CPC, and might also be referred to as the *user LPAR ID (UPID)*.

The MIF ID is a number that is defined through the HCD tool or directly through the IOCP. It is specified in the RESOURCE statement in the configuration definitions. It is in the range of 1 - F and is unique within a CSS. However, because of the multiple CSSs, the MIF ID is not unique within the CPC.

The multiple image facility enables resource sharing across LPARs within a single CSS, or across the multiple CSSs. When a channel resource is shared across LPARs in multiple CSSs, the configuration is called *spanning*. Multiple CSSs can specify the same MIF image ID. However, the combination CSSID.MIFID is unique across the CPC.

Dynamic addition or deletion of a logical partition name

All undefined LPARs are reserved partitions. They are automatically predefined in the hardware system area (HSA) with a name placeholder and a MIF ID.

Summary of identifiers

It is a preferred practice to establish a naming convention for the LPAR identifiers. Figure 5-4 summarizes the identifiers and how they are defined. You can use the CSS number concatenated to the MIF ID, which means that LPAR ID 3A is in CSS 3 with MIF ID A. This fits within the allowed range of LPAR IDs, and conveys helpful information to the user.

CSS0			CSS1			CSS2	CSS3		Specified in HCD / IOCP
Logical TST1	Partition PROD1	Name PROD2	Logical TST2	Partition PROD3	Name PROD4	Log Part TST3	Logical TST4	Partition PROD5	Specified in HCD / IOCP
Logical Partition ID 02 04 0A			Logical Partition ID 14 16 1D			Log Part ID 22	Logical Partition ID 35 3A		Specified in HMC Image Profile
MIF ID 2	MIF ID 4	MIF ID A	MIF ID 4	MIF ID 6	MIF ID D	MIF ID 2	MIF ID 5	MIF ID A	Specified in HCD / IOCP

Figure 5-4 CSS, LPAR, and identifier example

5.1.6 Physical channel ID

A physical channel ID (PCHID) reflects the physical identifier of a channel-type interface. A PCHID number is based on the I/O drawer or I/O cage location, the channel feature slot number, and the port number of the channel feature. A hardware channel is identified by a PCHID. The physical channel, which uniquely identifies a connector jack on a channel feature, is known by its PCHID number.

Do not confuse PCHIDs with CHPIDs. A CHPID does not directly correspond to a hardware channel port, and can be arbitrarily assigned. Within a single channel subsystem, 256 CHPIDs can be addressed. That gives a maximum of 1,024 CHPIDs when four CSSs are defined. Each CHPID number is associated with a single channel.

CHPIDs are not pre-assigned. Assign the CHPID numbers during installation by using the CHPID Mapping Tool (CMT) or HCD/IOCP. Assigning CHPIDs means that a CHPID number is associated with a physical channel/port location and a CSS. The CHPID number range is still from 00 - FF, and must be unique within a CSS. Any non-internal CHPID that is not defined with a PCHID can fail validation when you are attempting to build a production IODF or an IOCDs.

5.1.7 Channel spanning

Channel spanning extends the MIF concept of sharing channels across LPARs to sharing physical channels across LPARs and channel subsystems.

Spanning is the ability for a physical channel (PCHID) to be mapped to CHPIDs defined in multiple channel subsystems. When so defined, channels can be transparently shared by any or all of the configured LPARs, regardless of the channel subsystem to which the LPAR is configured.

A channel is considered a spanned channel if the same CHPID number in different CSSs is assigned to the same PCHID in IOCP, or is defined as spanned in HCD.

For internal channels, such as Internal Coupling (IC) links and HiperSockets, the same applies, but with no PCHID association. They are defined with the same CHPID number in multiple CSSs.

In Figure 5-5 on page 193, CHPID 04 is spanned to CSS0 and CSS1. Because it is an internal channel link, no PCHID is assigned. CHPID 06 is an external spanned channel and has a PCHID assigned.

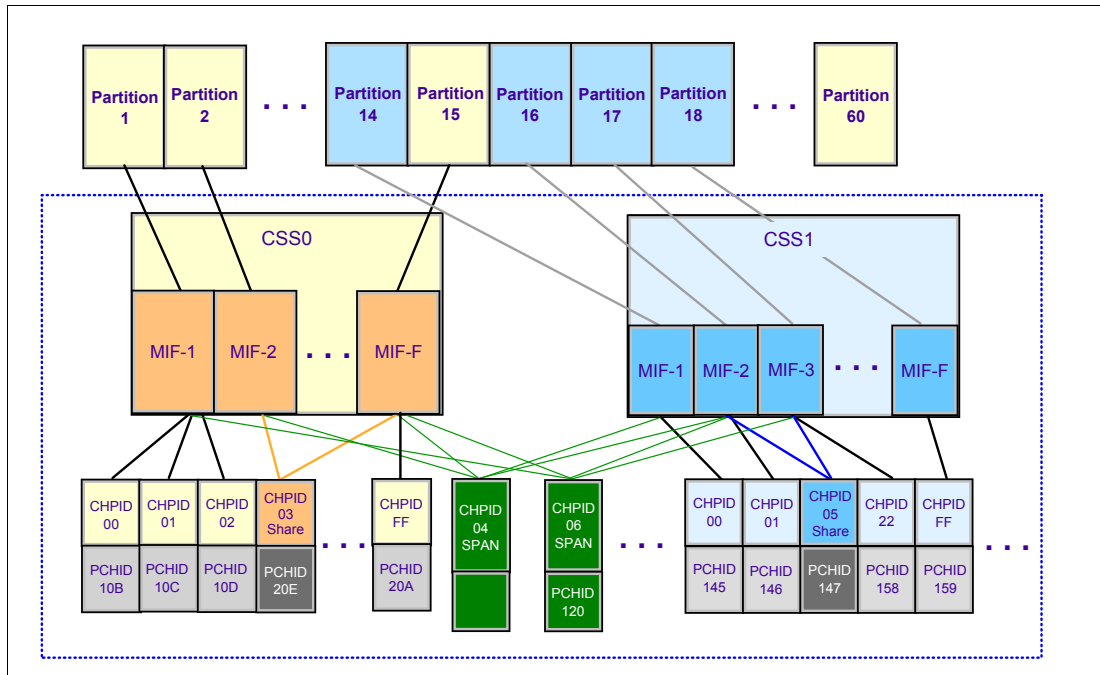


Figure 5-5 zEC12 CSS: Channel subsystems with channel spanning

CHPIDs that span CSSs reduce the total number of channels available. The total is reduced because no CSS can have more than 256 CHPIDs. For a zEC12 with two CSSs defined, a total of 512 CHPIDs is supported. If all CHPIDs are spanned across the two CSSs, only 256 channels are supported. For a zEC12 with four CSSs defined, a total of 1024 CHPIDs are supported. If all CHPIDs are spanned across the four CSSs, only 256 channels are supported.

Channel spanning is supported for internal links (HiperSockets and Internal Coupling (IC) links) and for certain external links. External links that are supported include FICON Express8S, FICON Express8 channels, OSA-Express4S, OSA-Express3, and Coupling Links.

5.1.8 Multiple CSS construct

A zEC12 with multiple CSSs defined is shown in Figure 5-6. In this example, two channel subsystems are defined (CSS0 and CSS1). Each CSS has three LPARs with their associated MIF image identifiers.

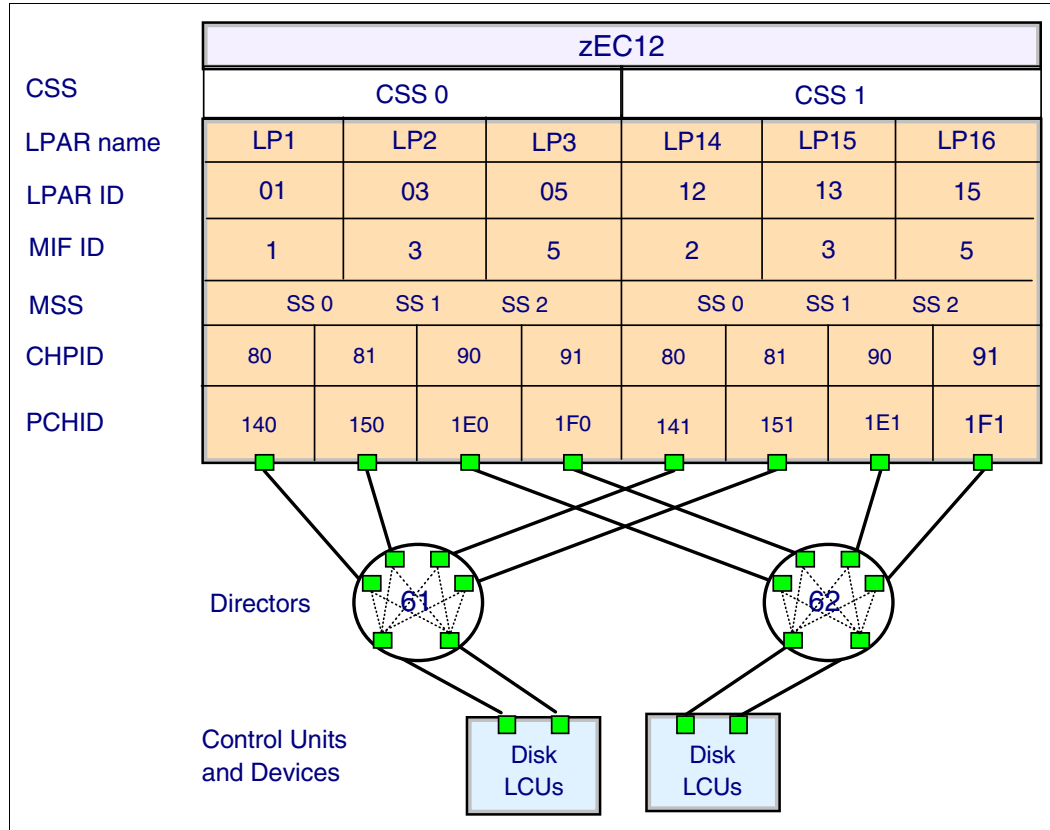


Figure 5-6 zEC12 CSS connectivity

In each CSS, the CHPIDs are shared across all LPARs. The CHPIDs in each CSS can be mapped to their designated PCHIDs by using the CMT or manually by using the HCD or IOCP. The output of the CHPID mapping tool (CMT) is used as input to HCD or the IOCP to establish the CHPID to PCHID assignments.

5.1.9 Adapter ID (AID)

When you use HCD or IOCP to assign a CHPID to a Parallel Sysplex over an InfiniBand (IFB) coupling link port, an AID number is required.

The AID is bound to the serial number of the fanout. If the fanout is moved, the AID moves with it. No IOCDs update is required if adapters are moved to a new physical location.

For detailed information, see “Adapter ID number assignment” on page 148.

5.1.10 Channel subsystem enhancement for I/O resilience

The zEC12 channel subsystem is enhanced to provide improved throughput and I/O service times when the following abnormal conditions occur:

- ▶ Multiple system workload spikes
- ▶ Multiple system resource contention in the storage area network (SAN) or at the control unit ports
- ▶ SAN congestion
- ▶ Improperly defined SAN configurations
- ▶ Dynamic changes in fabric routing
- ▶ Destination port congestion

It might also apply to Licensed Internal Code (LIC) failures in the SAN, channel extenders, Wavelength Division Multiplexers, and control units. When abnormal conditions occur that can cause an imbalance in I/O performance characteristics across a set of channel paths to the control unit, the channel subsystem intelligently uses the channels that provide optimal performance. This enhancement is accomplished by using the in-band I/O instrumentation and metrics of the System z FICON and zHPF protocols.

This channel subsystem enhancement is exclusive to zEC12, and is supported on all FICON channels when configured as CHPID type FC. This enhancement is transparent to operating systems.

5.2 I/O configuration management

For ease of management, use HCD to build and control the I/O configuration definitions. 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.

The following tools are provided 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 new configurations or upgrades of an existing configuration, and it maintains the installed features of those configurations. Reports that are produced by eConfig are helpful in understanding the changes being made for a system upgrade, and what the final configuration will look like.
- ▶ Hardware configuration definition (HCD): HCD supplies an interactive dialog to generate the I/O definition file (IODF), and later the input/output configuration data set (IOCDs). Generally, use HCD or HCM to generate the I/O configuration rather than writing IOCP statements. The validation checking that HCD runs as data is entered helps minimize the risk of errors before the I/O configuration is implemented.
- ▶ Hardware Configuration Management (HCM): HCM is a priced optional feature that supplies a graphical interface to HCD. It is installed on a PC and allows you to manage both the physical and the logical aspects of a mainframe hardware configuration.
- ▶ CHPID mapping tool (CMT): The CMT provides a mechanism to map CHPIDs onto PCHIDs as required. Additional enhancements have been built into the CMT to meet the requirements of the zEC12. It provides the best availability choices for the installed features and defined configuration. The CMT is available for download from the IBM Resource Link website:

<http://www.ibm.com/servers/resourceLink>

The health checker function in z/OS V1.10 introduces a health check in the I/O Supervisor. This application can help system administrators identify single points of failure in the I/O configuration.

5.3 Channel subsystem summary

The zEC12 provides support for the full architecture. Table 5-1 shows CSS-related information in terms of maximum values for devices, subchannels, LPARs, and CHPIDs.

Table 5-1 zEC12 CSS overview

Setting	zEC12
Maximum number of CSSs	4
Maximum number of CHPIDs	1024
Maximum number of LPARs supported per CSS	15
Maximum number of LPARs supported per system	60
Maximum number of HSA subchannels	11505 K (191.75 K per partition x 60 partitions)
Maximum number of devices	255 K (4 CSSs x 63.75 K devices)
Maximum number of CHPIDs per CSS	256
Maximum number of CHPIDs per LPAR	256
Maximum number of subchannels per LPAR	191.75 K (63.75 K + 2 x(64 K - 1))

All channel subsystem images (CSS images) are defined within a single IOCDS. The IOCDS is loaded and initialized into the hardware system area (HSA) during system power-on reset (POR). The HSA is pre-allocated in memory with a fixed size of 32 GB. This configuration eliminates planning for HSA and pre-planning for HSA expansion because HCD/IOCP always reserves the following items during the IOCDS process:

- ▶ Four CSSs
- ▶ Fifteen LPARs in each CSS
- ▶ Subchannel set 0 with 63.75 K devices in each CSS
- ▶ Subchannel set 1 with 64 K minus one devices in each CSS
- ▶ Subchannel set 2 with 64 K minus one devices in each CSS

All these items are activated and used with dynamic I/O changes.

Figure 5-7 shows a logical view of the relationships. Each CSS supports up to 15 LPARs. System-wide, a total of up to 60 LPARs are supported.

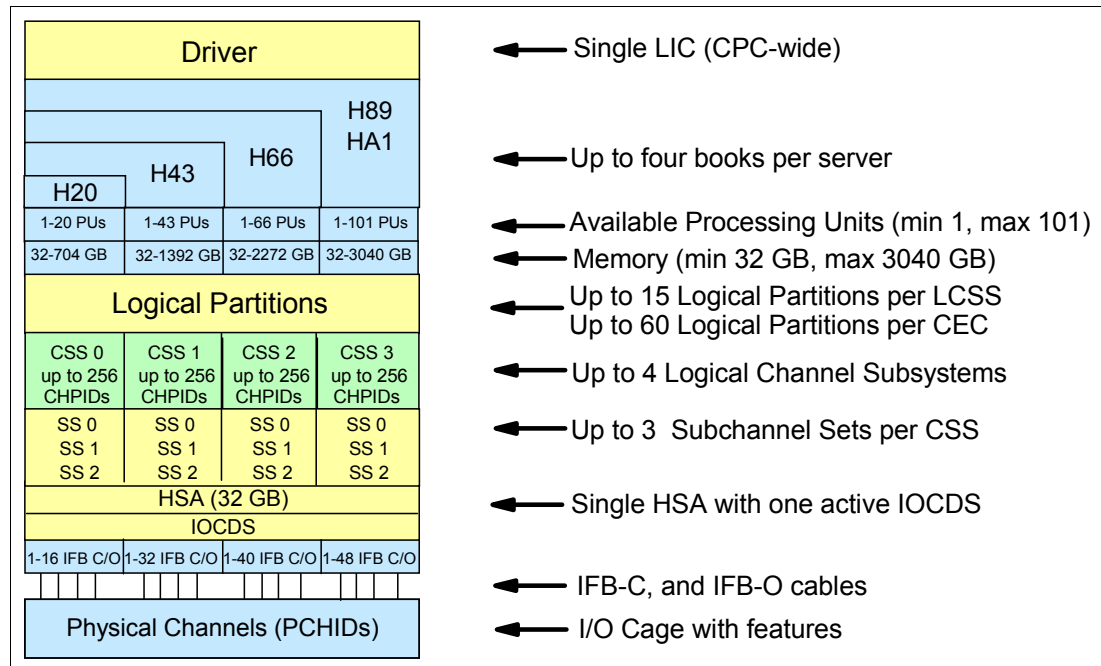


Figure 5-7 Logical view of zEC12 models, CSSs, IOCDS, and HSA

Book repair: The HSA can be moved from one book to a different book in an enhanced availability configuration as part of a concurrent book repair action.

The channel definitions of a CSS are not bound to a single book. A CSS can define resources that are physically connected to any InfiniBand cable of any book in a multibook CPC.

5.4 System-initiated CHPID reconfiguration

The system-initiated CHPID reconfiguration function is designed to reduce the duration of a repair action and minimize operator interaction. It is used when a FICON channel, an Open Systems Adapter (OSA) port, or an InterSystem Channel-3 (ISC-3) link is shared across LPARs on a zEC12 server. When an I/O card is to be replaced (for a repair), it usually has a few failed channels and others that are still functioning.

To remove the card, all channels must be configured offline from all LPARs that share those channels. Without system-initiated CHPID reconfiguration, the IBM service support representative (SSR) must contact the operators of each affected LPAR and have them set the channels offline. After the repair, the SSR must contact them again to configure the channels back online.

With system-initiated CHPID reconfiguration support, the Support Element sends a signal to the channel subsystem that a channel needs to be configured offline. The channel subsystem determines all the LPARs that share that channel and sends an alert to the operating systems in those LPARs. The operating system then configures the channel offline without any operator intervention. This cycle is repeated for each channel on the card.

When the card is replaced, the Support Element sends another signal to the channel subsystem for each channel. This time, the channel subsystem alerts the operating system that the channel must be configured back online. This process minimizes operator interaction when you are configuring channels offline and online.

System-initiated CHPID reconfiguration is supported by z/OS.

5.5 Multipath initial program load (IPL)

Multipath IPL helps increase availability and eliminate manual problem determination during IPL execution. This support allows IPL to complete, if possible, using alternate paths when you are running an IPL from a device that is connected through FICON channels. If an error occurs, an alternate path is selected.

Multipath IPL is applicable to FICON channels (CHPID type FC). z/OS supports multipath IPL.



Cryptography

This chapter describes the hardware cryptographic functions available on the IBM zEnterprise EC12 (zEC12). The CP Assist for Cryptographic Function (CPACF) with the Peripheral Component Interconnect Express (PCIe) Cryptographic Coprocessors offers a balanced use of resources and unmatched scalability.

The zEC12 includes 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 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, FIPS 140-2 Level 4¹.

The cryptographic functions include the full range of cryptographic operations 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 the zEC12 offers.

Secure Sockets Layer/Transport Layer Security (SSL/TLS) is a key technology for conducting secure e-commerce using web servers. It has been adopted by a rapidly increasing number of applications, demanding new levels of security, performance, and scalability.

This chapter includes the following sections:

- ▶ Cryptographic synchronous functions
- ▶ Cryptographic asynchronous functions
- ▶ CPACF protected key
- ▶ PKCS #11 overview
- ▶ Cryptographic feature codes
- ▶ CP Assist for Cryptographic Function (CPACF)
- ▶ Crypto Express4S
- ▶ Crypto Express3
- ▶ Tasks that are run by PCIe Crypto Express
- ▶ TKE workstation feature
- ▶ Cryptographic functions comparison

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

6.1 Cryptographic synchronous functions

Cryptographic synchronous functions are provided by the CPACF. The CPACF must be explicitly enabled by using an enablement feature (feature code (FC) 3863) that is available for no extra fee. For IBM and client-written programs, CPACF functions can be started by instructions that are described in *z/Architecture Principles of Operation*, SA22-7832. As a group, these instructions are known as the Message-Security Assist (MSA). z/OS Integrated Cryptographic Service Facility (ICSF) callable services on z/OS and in-kernel crypto APIs and libica cryptographic functions library running on Linux on System z can also start CPACF synchronous functions.

The zEC12 hardware includes the implementation of algorithms as hardware synchronous operations. This configuration holds the processor unit (PU) processing of the instruction flow until the operation completes. The zEC12 offers 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) for cryptographic key generation

Requirement: The keys must be provided in clear form only.

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. The CPACF functions are supported by z/OS, z/VM, z/VSE, z/TPF, and Linux on System z.

6.2 Cryptographic asynchronous functions

Cryptographic asynchronous functions are provided by the optional Peripheral Component Interconnect Express (PCIe) cryptographic coprocessors Crypto Express4S or Crypto Express3.

6.2.1 Secure key functions

The following secure key functions are provided as cryptographic asynchronous functions. System internal messages are passed to the cryptographic coprocessors to initiate the operation. The messages then are passed back from the coprocessors to signal completion of the operation:

- ▶ Data encryption and decryption algorithms for data protection:

Data Encryption Standard (DES):

- Single-length key DES
- Double-length key DES
- Triple-length key DES (Triple-DES)

- ▶ DES key generation and distribution
- ▶ Personal identification number (PIN) generation, verification, and translation functions
- ▶ Random number generator
- ▶ Public Key Cryptography Standards (PKCS) #11 functions²:

z/OS Integrated Cryptographic Service Facility (ICSF) implements callable services in support of the PKCS #11 standard. Secure IBM Enterprise PKCS #11 (EP11) coprocessor mode implements secure keys for PKCS #11 functions.

- ▶ Public key algorithm (PKA) functions:

The following supported callable services are intended for application programs that use PKA:

- Importing Rivest-Shamir-Adleman algorithm (RSA) public-private key pairs in clear and encrypted forms
- Rivest-Shamir-Adleman algorithm (RSA):
 - Key generation, up to 4096-bit
 - Signature generation and verification, up to 4096-bit
 - Import and export of DES keys under an RSA key, up to 4096-bit

- Public key encryption (PKE)/Public key decryption (PKD):

The PKE and PKD callable services are provided for assisting the SSL/TLS handshake. They are used to offload compute-intensive portions of the protocol onto the cryptographic adapters.

- Europay MasterCard and Visa (EMV) standard:

Applications can be written to comply with the EMV standard for financial transactions between heterogeneous hardware and software. EMV standards have been updated to use improved security properties of EMV contact and contactless cards. ICSF HRC770A improved support of EMV card applications that support American Express cards.

² Requires Crypto Express4S and Trusted Key Entry (TKE) workstation

6.2.2 Additional functions

Other key functions of the Crypto Express features serve to enhance the security of the cryptographic processing:

- ▶ Remote loading of initial Automated Teller Machine (ATM) keys:

This function remotely loads the initial keys for capable Automated Teller Machines (ATMs) and Point of Sale (POS) systems. Remote key loading is the process of loading DES keys to the ATM from a central administrative site without requiring manually loading keys on each machine. The standard ANSI X9.24-2 defines the acceptable methods of performing this using public key cryptographic techniques. The process uses ICSF callable services with the Crypto Express4S or Crypto Express3 features to perform the remote load.

Trusted Block Create (CSNDTBC) is a callable service that is used to create a trusted block that contains a public key and certain processing rules. The rules define the ways and formats in which keys are generated and exported. Remote Key Export (CSNDRKX) is a callable service that uses the trusted block to generate or export DES keys for local use, and for distribution to an ATM or other remote device. The PKA Key Import (CSNDPKI), PKA Key Token Change (CSNDKTC), and Digital Signature Verify (CSFNDFV) callable services support the remote key loading process.

- ▶ Key exchange with non-CCA cryptographic systems:

This function allows the exchange of operational keys between the Crypto Express4S or Crypto Express3 coprocessors and non-CCA systems, such as ATMs. IBM Common Cryptographic Architecture (CCA) employs control vectors to control the usage of cryptographic keys. Non-CCA systems use other mechanisms, or can use keys that have no associated control information. Enhancements to key exchange functions allow the CCA to exchange keys between CCA systems and systems that do not use control vectors. It allows the CCA system owner to define permitted types of keys to be imported and exported while preventing uncontrolled key exchange that can open the system to an attack.

- ▶ Elliptic Curve Cryptography (ECC) Digital Signature Algorithm support:

Elliptic Curve Cryptography is an emerging public-key algorithm that is intended to replace RSA cryptography in many applications. ECC provides digital signature functions and key agreement functions. The CCA functions provide ECC key generation and key management. They also provide digital signature generation and verification functions compliant with the Elliptic Curve Digital Signature Algorithm (ECDSA) method. For more information, see ANSI X9.62 “Public Key Cryptography for the Financial Services Industry: The Elliptic Curve Digital Signature Algorithm (ECDSA)”. ECC uses keys that are shorter than RSA keys for equivalent strength-per-key-bit. So, the strength-per-key-bit is substantially greater in an algorithm that uses elliptic curves.

This cryptographic function is supported by z/OS, z/VM, and Linux on System z.

- ▶ Elliptic Curve Diffie-Hellman (ECDH) algorithm support:

The Common Cryptographic Architecture is extended to include the ECDH algorithm.

ECDH is a key agreement protocol that allows two parties, each having an elliptic curve public-private key pair, to establish a shared secret over an insecure channel. This shared secret can be used directly as a key. It can also be used to derive another key that can then be used to encrypt subsequent communications that use a symmetric key cipher, such as AES KEK.

ECDH includes these enhancements:

- Key management function to support AES key-encrypting key (KEK)
- Generating an ECC private key that is wrapped with an AES KEK
- Importing and exporting an ECC private key that is wrapped with an AES KEK

- Support for ECDH with a new service
- ▶ User-Defined Extensions (UDX) 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 CryptoCards 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 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 System z is tied to specific versions of the CCA card code and the related host code. UDX is available for the Crypto Express4S (Secure IBM CCA coprocessor mode only) and Crypto Express3 feature. An UDX migration is no more disruptive than a normal maintenance change level (MCL) or ICSF release migration.

In zEC12, you can import up to four UDX files. These files can be imported only from a DVD. The UDX configuration panel was updated to include a *Reset to IBM Default* bottom.

Consideration: CCA will have a new code level at zEC12, and the UDX clients will require a new UDX.

For more information, see the IBM CryptoCards website:

<http://www.ibm.com/security/cryptocards>

6.3 CPACF protected key

The zEC12 supports the protected key implementation. Since 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 Express4s or Crypto Express3 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 Express4S and Crypto Express3 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 were implemented in CPACF code with 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. There is no customizable option available at SE or HMC that permits or avoids the wrapping key generation. Figure 6-1 shows this function flow.

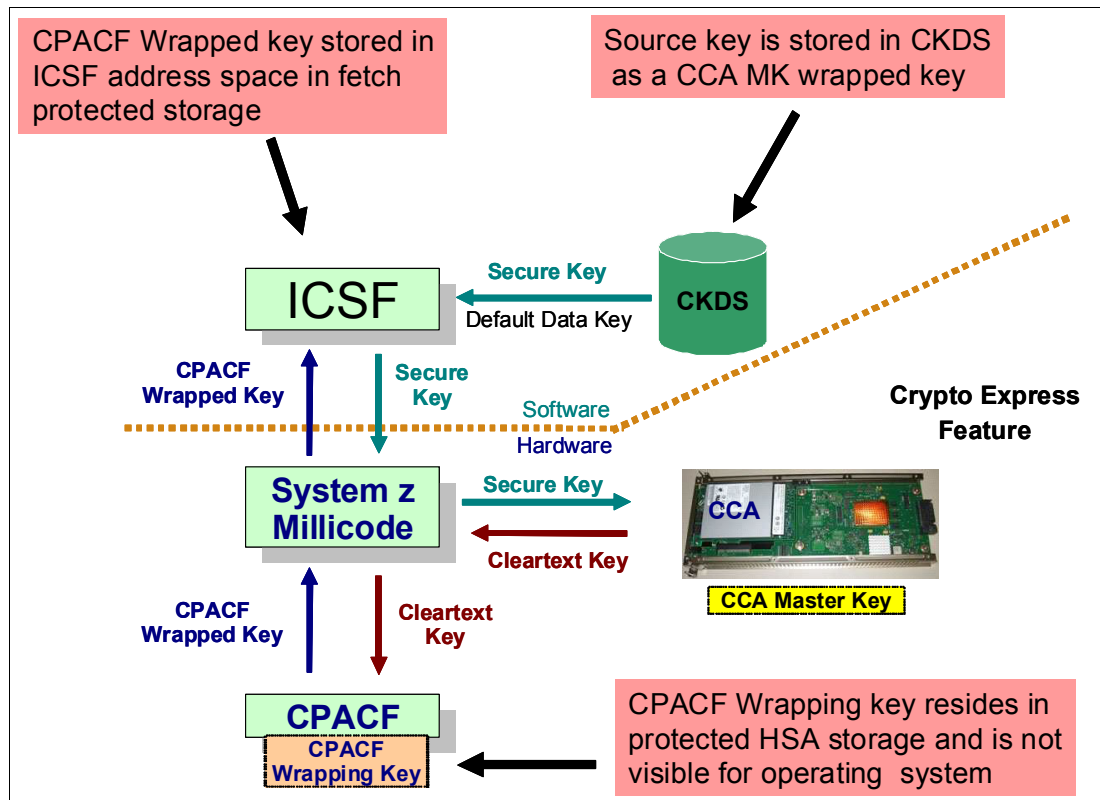


Figure 6-1 CPACF key wrapping

If a Crypto Express4 coprocessor (CEX4C) or Crypto Express3 coprocessor (CEX3C) 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 the Crypto Express4S or the Crypto Express3 is not available.

A new segment in the profiles at 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 6-1 considers a secure key as being the source of a protected key.

The source key in this case was already stored in the ICSF Cryptographic Key Data Set (CKDS) as a secure key (encrypted under the master key). This secure key is sent to Crypto Express4S or to the Crypto Express3 to be deciphered, and sent to Central Processor Assist for Cryptographic Function (CPACF) in clear text. At 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 will be 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 HCR7770 as a minimum release.

6.4 PKCS #11 overview

Public Key Cryptography Standards (PKCS) #11 is one of the industry-accepted standards. It is provided by RSA Laboratories of RSA Security Inc. PKCS #11 specifies an application programming interface (API) to devices, referred to as *tokens*, that hold cryptographic information and run cryptographic functions. PKCS #11 provides an alternative to IBM CCA.

PKCS #11 describes the cryptographic token interface standard and its API, which is also known as the Cryptographic Token Interface (*Cryptoki*). It is a de facto industry standard on many computing platforms today. It is a higher-level API when compared to CCA, and it is easier to use by C language-based applications. The persistent storage and retrieval of objects are part of the standard. The objects are certificates, keys, and application-specific data objects.

6.4.1 PKCS #11 model

On most single-user systems, a *token* is a smart card or other plug-installed cryptographic device that is accessed through a card reader or *slot*. Cryptoki provides a logical view of slots and tokens. This view makes each cryptographic device look logically like every other device regardless of the technology that is used. The PKCS #11 specification assigns numbers to slots, which are known as *slot IDs*. An application identifies the token that it wants to access by specifying the appropriate slot ID. On systems that have multiple slots, the application determines which slot to access.

The PKCS #11 logical view of a token is a device that stores objects and can run cryptographic functions. PKCS #11 defines three types of objects:

- ▶ A data object that is defined by an application.
- ▶ A certificate object that stores a digital certificate.
- ▶ A key object that stores a cryptographic key. The key can be a public key, a private key, or a secret key.

Objects are also classified according to their lifetime and visibility:

- ▶ *Token objects* are visible to all applications connected to the token that have sufficient permission. They remain on the token even after the sessions are closed, and the token is removed from its slot. *Sessions* are connections between an application and the token.
- ▶ *Session objects* are more temporary. When a session is closed by any means, all session objects that were created by that session are automatically deleted. Furthermore, session objects are visible only to the application that created them.

Attributes are characteristics that distinguish an instance of an object. General attributes in PKCS #11 distinguish, for example, whether the object is public or private. Other attributes are specific to a particular type of object, such as a Modulus or exponent for RSA keys.

The PKCS #11 standard was designed for systems that grant access to token information based on a PIN. The standard recognizes two types of token user:

- ▶ Security officer (SO)
- ▶ Standard user (USER)

The role of the SO is to initialize a token (zeroize the content) and set the user's PIN. The SO can also access public objects on the token but not private ones. The USER can access private objects on the token. Access is granted only after the user is authenticated. Users can also change their own PINs. Users cannot, however, reinitialize a token.

The PKCS #11 general model components are represented in Figure 6-2:

- ▶ Token: Logical view of a cryptographic device, such as a smart card or Hardware Security Module (HSM)
- ▶ Slot: Logical view of a smart card reader
- ▶ Objects: Items that are stored in a token, such as digital certificates and cryptographic keys
- ▶ User: The owner of the private data on the token, who is identified by the access Personal Identification Number (PIN)
- ▶ Security Officer: Person who initializes the token and the USER PIN

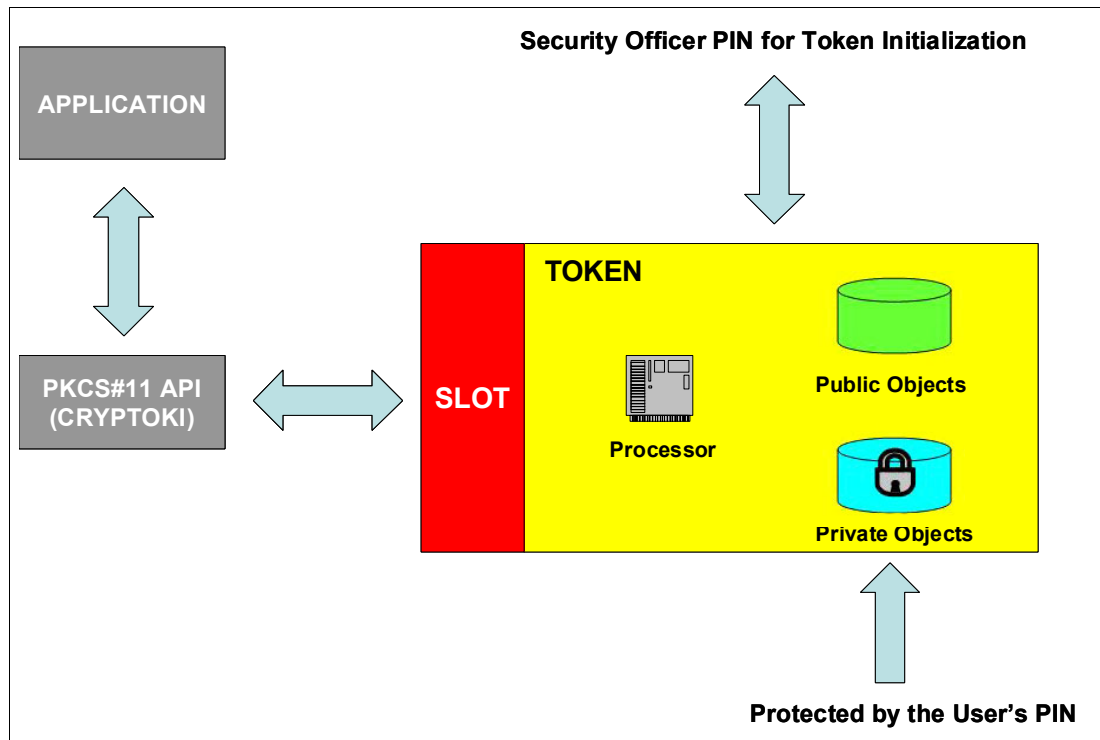


Figure 6-2 The PKCS #11 general model

6.4.2 z/OS PKCS #11 implementation

ICSF supports the PKCS #11 standard, increasing the number of cryptographic applications that use System z cryptography. PKCS #11 support was introduced in ICSF FMID HCR7740 within z/OS V1R9. In zEC12, with Crypto Express4S and FMID HCR77A0, ICSF expanded the support and introduced PKCS #11 secure keys.

On z/OS, PKCS #11 tokens are not physical cryptographic devices, but rather virtual smart cards. New tokens can be created at any time. The tokens can be application-specific or system-wide, depending on the access control definitions that are used instead of PINs. The tokens and their contents are stored in a new ICSF VSAM data set, the Token Key Data Set (TKDS). The TKDS serves as the repository for cryptographic keys and certificates that are used by PKCS #11 applications.

z/OS provides several facilities to manage tokens:

- ▶ A C language API that implements a subset of the PKCS #11 specification.
- ▶ Token management ICSF callable services, which are also used by the C API.

- ▶ The ICSF ISPF panel, called Token Browser, that allows you to see a formatted view of TKDS objects and make minor, limited updates to them.
- ▶ The **RACF RACDCERT** command supports the certificate, public key, and private key objects, and provides subfunctions to manage these objects together with tokens.
- ▶ The **gskkyman** command supports management of certificates and keys similar to the way that RACFDCERT does.

ICSF supports PKCS #11 session objects and token objects. ICSF supports PKCS #11 session objects and token objects. Session objects exist in memory only. They are not maintained on the direct access storage device (DASD). An application has only one session objects database, even if the application creates multiple PKCS #11 sessions.

Token objects are stored in the TKDS with one record per object. They are visible to all applications that have sufficient permission to the token. The objects are persistent and remain associated with the token even after a session is closed.

The PKCS #11 standard was designed for systems that grant access to token information based on a PIN. z/OS does not use PINs. Instead, profiles in the SAF CRYPTOZ class control access to tokens. Each token has two resources in the CRYPTOZ class:

- ▶ The resource `USER.token-name` controls the access of the user role to the token.
- ▶ The resource `SO.token-name` controls the access of the SO role to the token.

A user's access level to each of these resources (read, update, and control) determines the user's access level to the token. Figure 6-3 shows the concepts that were introduced by the PKCS #11 model to the z/OS PKCS #11 implementation.

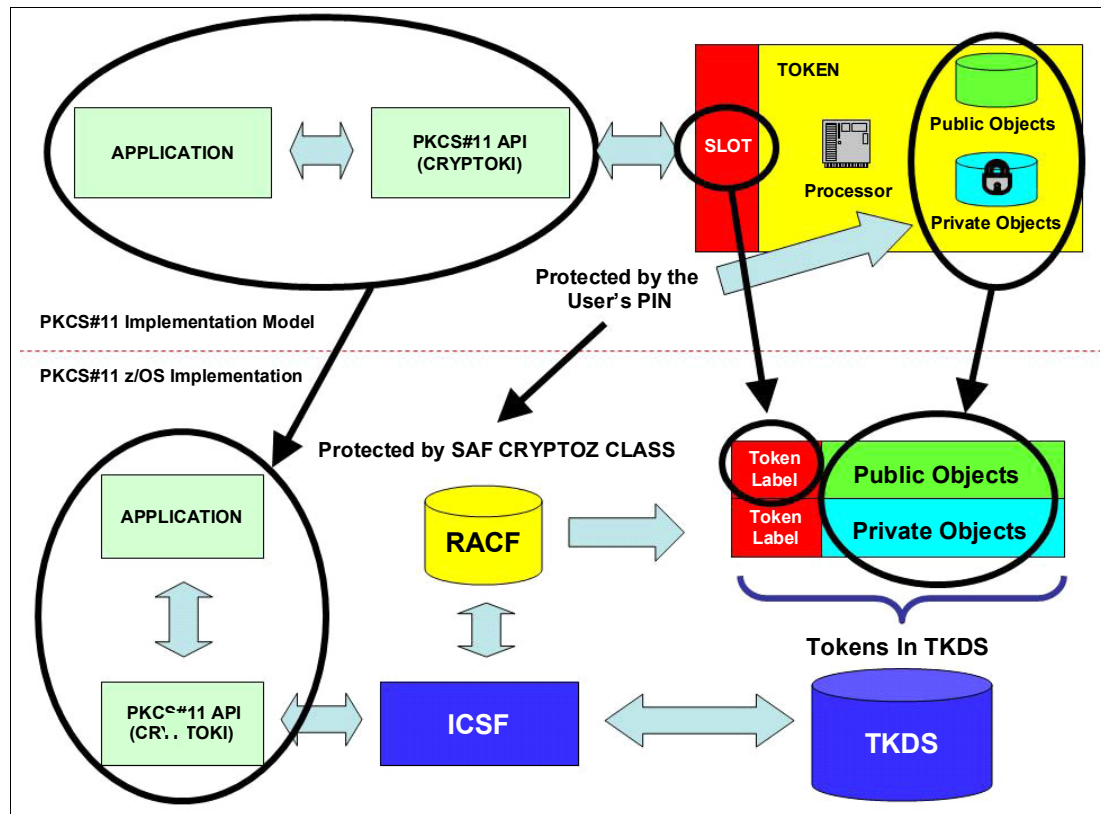


Figure 6-3 Mapping the PKCS #11 model to the z/OS PKCS #11 implementation

Tokens

The PKCS #11 tokens on z/OS are virtual, and are similar to RACF (SAF) key rings. An application can have one or more z/OS PKCS #11 tokens, depending on its requirements. z/OS PKCS #11 tokens are created by using system software, such as RACF, the **gskkyman** utility, or applications using the C API. Also, ICSF panels can be used for token management with limited usability.

Each token has a unique token name or label that is specified by the user or application when the token is created. Like z/OS PKCS #11 token creation, the PKCS #11 tokens can be deleted by using the same system software tools that are used to create them.

Token Key Data Set (TKDS)

The TKDS is a VSAM data set that serves as the repository for cryptographic keys and certificates that are used by z/OS PKCS #11 applications. Before an installation can run PKCS #11 applications, the TKDS must be created. The ICSF installation options data set must then be updated to identify the name of the TKDS data set. To optimize performance, ICSF creates a data space that contains an in-storage copy of the TKDS.

Important: Until ICSF FMID HCR7790, keys in the TKDS were not encrypted. Therefore, the RACF profile must be created to protect the TKDS from unauthorized access.

6.4.3 Secure IBM Enterprise PKCS #11 (EP11) Coprocessor

The IBM Enterprise PKCS #11 Licensed Internal Code (LIC) implements an industry standardized set of services. These services adhere to the PKCS #11 specification V2.20 and more recent amendments. It is designed to meet the Common Criteria (EAL 4+) and FIPS 140-2 Level 4 certifications. It conforms to the Qualified Digital Signature (QDS) Technical Standards that are mandated by the European Union.

The PKCS #11 secure key support is provided by the Crypto Express4S card that is configured in Secure IBM Enterprise PKCS #11 (EP11) coprocessor mode. Before EP11, ICSF PKCS #11 implementation only supported clear keys, and the key protection was provided only by RACF CRYPTOZ class protection. In EP11, keys now can be generated and securely wrapped under the EP11 Master Key. The secure keys never leave the secure coprocessor boundary unencrypted.

The Crypto Express4S firmware has a unique code for EP11 separated from the CCA code. Crypto Express4S with EP11 configuration is known as CEX4P. There is no change in the domain configuration in the LPAR activation profiles. The configuration selection is run in the Cryptographic Configuration panel on the Support Element. A coprocessor in EP11 mode is configured off after being zeroized.

Important: The Trusted Key Entry (TKE) workstation is required for management of the Crypto Express4S when defined as an EP11 coprocessor.

6.5 Cryptographic feature codes

Table 6-1 lists the cryptographic features that are available.

Table 6-1 Cryptographic features for zEnterprise CPC

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 Crypto Express features.
0864	Crypto Express3 feature: A maximum of eight features can be carried forwarded. This is an optional feature, and each feature contains two PCI Express cryptographic adapters (adjunct processors). This feature is not supported as a new build. It is available only on a carry-forward basis when you are upgrading from earlier generations to zEC12.
0865	Crypto Express4S feature: A maximum of 16 features can be ordered. This is an optional feature, and each feature contains one PCI Express cryptographic adapter (adjunct processor).
0841	Trusted Key Entry (TKE) workstation: This feature is optional. TKE provides a basic key management (key identification, exchange, separation, update, backup), and security administration. 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 zEC12 can be installed.
0850	TKE 7.2 Licensed Internal Code (TKE 7.2 LIC): The 7.2 LIC requires Trusted Key Entry workstation feature code 0841. It is required to support CEX4P. The 7.2 LIC can also be used to control z196, z114, z10 EC, z10 BC, z9 EC, z9 BC, z990, and z890 servers.
0885	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 to the TKE workstation, and 20 smart cards. Smart card part 74Y0551 is required to support CEX4P.
0884	TKE additional smart cards: When one feature code is ordered, 10 smart cards are shipped. Order increment is 1 - 99 (990 blank smart cards). Smart card part 74Y0551 is required to support CEX4P.

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 Express features in a zEC12, a TKE workstation with the TKE 7.2 LIC or later is required. For more information, see 6.10, “TKE workstation feature” on page 219.

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.

6.6 CP Assist for Cryptographic Function (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 Secure Sockets Layer (SSL), Virtual Private Network (VPN), and data-storing applications that do not require Federal Information Processing Standard (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 6.3, “CPACF protected key” on page 203.

The CPACF feature provides hardware acceleration for DES, Triple-DES, MAC, AES-128, AES-192, AES-256, SHA-1, SHA-224, SHA-256, SHA-384, and SHA-512 cryptographic services. It provides high-performance hardware encryption, decryption, and hashing 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, zIIP, and zAAP. For more information about MSA instructions, see *z/Architecture Principles of Operation*, SA22-7832.

The functions of the CPACF must be explicitly enabled by using FC 3863 during the manufacturing process or at the client site as an MES installation. The exceptions are SHA-1, and SHA-2 support for SHA-224, SHA-256, SHA-384, and SHA-512, which are always enabled.

6.7 Crypto Express4S

The Crypto Express4S feature (FC 0865) is an optional zEC12 exclusive feature. Each feature has one PCIe cryptographic adapter. The Crypto Express4S feature occupies one I/O slot in a zEC12 PCIe I/O drawer. This feature 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 Express4S feature provides tamper-sensing and tamper-responding, high-performance cryptographic operations.

Each Crypto Express4S PCI Express adapter can be in one of these configurations:

- ▶ Secure IBM CCA coprocessor (CEX4C) for Federal Information Processing Standard (FIPS) 140-2 Level 4 certification. This configuration includes secure key functions. It is

optionally programmable to deploy more functions and algorithms by using user programming function (UDX).

- ▶ Secure IBM Enterprise PKCS #11 (EP11) coprocessor (CEX4P) 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.

Trusted Key Entry (TKE) workstation is required to support the administration of the Crypto Express4S when configured as EP11 mode.

- ▶ Accelerator (CEX4A) for acceleration of public key and private key cryptographic operations that are used with SSL/Transport Layer Security (TLS) processing.

These modes can be configured by using the Support Element, and the PCIe adapter must be configured offline to change the mode.

Remember: 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 Express4S uses the IBM 4765 PCIe Coprocessor³. The Crypto Express4S feature does not have external ports and does not use fiber optic 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. The zEC12 supports a maximum of 16 Crypto Express4S features. 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. There are up to 16 LPARs per adapter. Having access to the adapter requires setup in the image profile for each partition. The adapter must be in the candidate list.

Each zEC12 supports up to 16 Crypto Express4S features. Table 6-2 shows configuration information for Crypto Express4S.

Table 6-2 *Crypto Express4S 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	16

a. The minimum initial order of Crypto Express4S features is two. After the initial order, more Crypto Express4S can be ordered one feature at a time, up to a maximum of 16.

b. More than one partition, defined to the same CSS or to different CSSs, can use the same domain number when assigned to different PCIe cryptographic adapters.

³ For more information, see <http://www-03.ibm.com/security/cryptocards/pciicc/overview.shtml>

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 Express4S features are assigned PCIe cryptographic adapter numbers sequentially during the power-on reset that follows the installation. When a Crypto Express4S feature is installed concurrently, the installation can select an out-of-sequence number from the unused range. When a Crypto Express4S 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 ahead to allow for nondisruptive changes:

- ▶ Operational changes can be made by using the Change LPAR Cryptographic Controls task from the Support Element, which reflects the cryptographic definitions in the image profile for the partition. With this function, adding and removing the cryptographic feature without stopping 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 coupled with the usage domain index 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. For example, you might define a configuration for backup situations. However, only one of the LPARs can be active at a time.

The zEC12 allows up to 60 LPARs to be active concurrently. Each PCI Express supports 16 domains, whether it is configured as a Crypto Express4S accelerator or a Crypto Express4S coprocessor. The server configuration must include at least four Crypto Express4S features (four PCIe adapters and 16 domains per PCIe adapter) when all 60 LPARs require concurrent access to cryptographic functions. More Crypto Express4S features might be needed to satisfy application performance and availability requirements.

6.8 Crypto Express3

The Crypto Express3 feature (FC 0864) is an optional feature, and is available only on a carry-forward basis when you upgrade from earlier generations to zEC12. Each feature has two PCIe cryptographic adapters. The Crypto Express3 feature occupies one I/O slot in an I/O cage or an I/O drawer.

Statement of Direction: The IBM zEnterprise EC12 is planned to be the last high-end System z server to offer support of the Crypto Express3 feature (FC 0864). Plan to upgrade from the Crypto Express3 feature to the Crypto Express4S feature (FC 0865).

Each Crypto Express3 PCI Express adapter can have one of these configurations:

- ▶ Secure coprocessor (CEX3C) for Federal Information Processing Standard (FIPS) 140-2 Level 4 certification. This configuration includes secure key functions, and is optionally programmable to deploy more functions and algorithms by using UDX.
- ▶ Accelerator (CEX3A) for acceleration of public key and private key cryptographic operations that are used with SSL/TLS processing.

These modes can be configured by using the Support Element. The PCIe adapter must be configured offline to change the mode.

The Crypto Express3 feature is designed to complement the functions of CPACF. This feature is tamper-sensing and tamper-responding. Unauthorized removal of the adapter or feature *zeroizes* its content. It provides dual processors that operate in parallel, supporting cryptographic operations with high reliability.

The CEX3 uses the 4765 PCIe Coprocessor. It holds a secured subsystem module, batteries for backup power, and a full-speed USB 2.0 host port available through a mini-A connector. On System z, these USB ports are not used. The securely encapsulated subsystem contains two 32-bit IBM PowerPC® 405D5 RISC processors that run in lockstep with cross-checking to detect malfunctions. The subsystem also includes a separate service processor that is used to manage these items:

- ▶ Self-test and firmware updates
- ▶ RAM, flash memory, and battery-powered memory
- ▶ Cryptographic-quality random number generator
- ▶ AES, DES, TDES, SHA-1, SHA-224, SHA-256, SHA-384, SHA-512, and modular-exponentiation (for example, RSA, DSA) hardware
- ▶ Full-duplex direct memory access (DMA) communications

Figure 6-4 shows the Crypto Express3 feature physical layout.

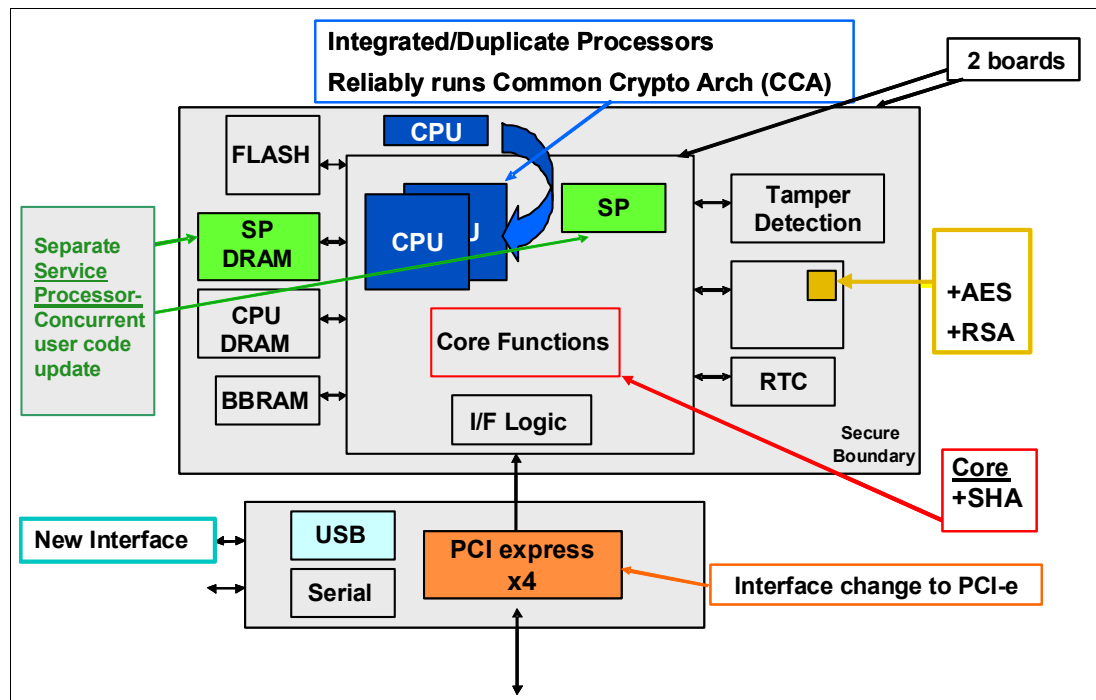


Figure 6-4 Crypto Express3 feature layout

The Crypto Express3 feature does not have external ports, and does not use fiber optic or other cables. It does not use CHPIDs, but requires one slot in the I/O cage and one PCHID for each PCIe cryptographic adapter. Removal of the feature or card *zeroizes* the content.

The zEC12 supports a maximum of eight Crypto Express3 features on a carry-forward basis, offering a combination of up to 16 coprocessors and accelerators. Access to the PCIe

cryptographic adapter is controlled through the setup in the image profiles on the Support Element (SE).

Adapter: Although PCIe cryptographic adapters have no CHPID type and are not identified as external channels, all LPARs in all channel subsystems have access to the adapter. There can be up to 16 LPARs per adapter. Having access to the adapter requires setup in the image profile for each partition. The adapter must be in the candidate list.

Each zEC12 supports up to eight Crypto Express3 features, which means a maximum of 16 PCIe cryptographic adapters. Table 6-3 shows configuration information for Crypto Express3.

Table 6-3 Crypto Express3 features

Feature	Quantity
Minimum number of carry-forward features for each server	2
Maximum number of features for each server	8
Number of PCIe cryptographic adapters for each feature (coprocessor or accelerator)	2
Maximum number of PCIe adapters for each server	16
Number of cryptographic domains for each PCIe adapter ^a	16

a. More than one partition, 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 coprocessor or accelerator, the PCIe cryptographic adapter is made available to an LPAR. The availability is directed by the domain assignment and the candidate list in the LPAR image profile. This availability is not affected by the shared or dedicated status that is given to the CPs in the partition.

When installed non-concurrently, Crypto Express3 features are assigned PCIe cryptographic adapter numbers sequentially during the power-on reset that follows the installation. When a Crypto Express3 feature is installed concurrently, the installation can select an out-of-sequence number from the unused range. When a Crypto Express3 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 ahead to allow for nondisruptive changes:

- ▶ Operational changes can be made by using the Change LPAR Cryptographic Controls task from the Support Element. The Support Element reflects the cryptographic definitions in the image profile for the partition. With this function, adding and removing the cryptographic feature without stopping 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 coupled with the usage domain index 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. For example, you might define a configuration for backup situations. However, only one of the LPARs can be active at any one time.

The zEC12 allows up to 60 LPARs to be active concurrently. Each PCI Express supports 16 domains, whether it is configured as a Crypto Express3 accelerator or a Crypto Express3 coprocessor. The server configuration must include at least two Crypto Express3 features (four PCIe adapters and 16 domains per PCIe adapter) when all 60 LPARs require concurrent

access to cryptographic functions. More Crypto Express3 features might be needed to satisfy application performance and availability requirements.

6.9 Tasks that are run by PCIe Crypto Express

The Crypto Express features running at zEC12 support all cryptographic functions that were introduced on zEnterprise CPC:

- ▶ Expanded key support for the AES algorithm

CCA supports the AES algorithm to allow the use of AES keys to encrypt data. Expanded key support for AES adds a framework to support a much broader range of application areas. It also lays the groundwork for future use of AES in areas where standards and client applications are expected to change.

As stronger algorithms and longer keys become increasingly common, security requirements dictate that these keys must be wrapped by using key-encrypting keys (KEKs) of sufficient strength. This feature adds support for AES key-encrypting keys. These AES wrapping keys have adequate strength to protect other AES keys for transport or storage. This support introduced AES key types that use the variable length key token. The supported key types are EXPORTER, IMPORTER, and for use in the encryption and decryption services, CIPHER.

- ▶ Enhanced ANSI TR-31 interoperable secure key exchange

ANSI TR-31 defines a method of cryptographically protecting Triple Data Encryption Standard (TDES) cryptographic keys and their associated usage attributes. The TR-31 method complies with the security requirements of the ANSI X9.24 Part 1 standard. However, use of TR-31 is not required to comply with that standard. CCA has added functions that can be used to import and export CCA TDES keys in TR-31 formats. These functions are designed primarily as a secure method of wrapping TDES keys for improved and more secure key interchange between CCA and non-CCA devices and systems.

- ▶ PIN block decimalization table protection

To help avoid a decimalization table attack to learn a PIN, a solution is now available in the CCA to thwart this attack by protecting the decimalization table from manipulation. PINs are most often used for ATMs, but are increasingly used at point-of sale (POS) devices for debit and credit cards.

- ▶ ANSI X9.8 PIN security

This function facilitates compliance with the processing requirements defined in the new version of the ANSI X9.8 and ISO 9564 PIN Security Standards. It provides added security for transactions that require PINs.

- ▶ Enhanced CCA key wrapping to comply with ANSI X9.24-1 key bundling requirements

This support allows that CCA key token wrapping method to use cipher block chaining (CBC) mode in combination with other techniques to satisfy the key bundle compliance requirements. The standards include ANSI X9.24-1 and the recently published Payment Card Industry Hardware Security Module (PCI HSM) standard.

- ▶ Secure key hashed message authentication code (HMAC)

HMAC is a method for computing a message authentication code by using a secret key and a secure hash function. It is defined in the standard FIPS 198, "The Keyed-Hash Message Authentication Code (HMAC)". The CCA function supports HMAC by using SHA-1, SHA-224, SHA-256, SHA-384, and SHA-512 hash algorithms. The HMAC keys are variable length and are securely encrypted so that their values are protected. This Crypto function is supported by z/OS, z/VM, and Linux on System z.

6.9.1 PCIe Crypto Express as a CCA coprocessor

The PCIe Crypto Express coprocessors enable the user to perform the following tasks:

- ▶ Encrypt and decrypt data by using secret-key algorithms. Triple-length key DES, double-length key DES, and AES algorithms are supported.
- ▶ Generate, install, and distribute cryptographic keys securely by using both public and secret-key cryptographic methods that generate, verify, and translate PINs.
- ▶ Crypto Express coprocessors support 13-digit through 19-digit personal account numbers (PANs)
- ▶ Ensure the integrity of data by using message authentication codes (MACs), hashing algorithms, and RSA PKA digital signatures, as well as ECC digital signatures

The Crypto Express coprocessors also provide the functions that are listed for the Crypto Express accelerator. However, they provide a lower performance than the Crypto Express accelerator can provide.

Three methods of master key entry are provided by Integrated Cryptographic Service Facility (ICSF) for the Crypto Express feature coprocessors:

- ▶ A passphrase initialization method, which generates and enters all master keys that are necessary to fully enable the cryptographic system in a minimal number of steps
- ▶ A simplified master key entry procedure that is provided through a series of Clear Master Key Entry panels from a TSO terminal
- ▶ A TKE workstation, which is available as an optional feature in enterprises that require enhanced key-entry security

Linux on System z also permits the master key entry through panels or through the TKE workstation.

The security-relevant portion of the cryptographic functions is run inside the secure physical boundary of a tamper-resistant card. Master keys and other security-relevant information are also maintained inside this secure boundary.

The Processor Resource/Systems Manager (PR/SM) fully supports the Crypto Express coprocessor features to establish a logically partitioned environment on which multiple LPARs can use the cryptographic functions. The following keys are provided for each of 16 cryptographic domains that a cryptographic adapter can serve:

- ▶ A 128-bit data-protection symmetric master key
- ▶ A 256-bit AES master key
- ▶ A 256-bit ECC master key
- ▶ One 192-bit PKA master key

Use the dynamic addition or deletion of an LPAR name to rename an LPAR. Its name can be changed from NAME1 to * (single asterisk) and then changed again from * to NAME2. The LPAR number and Multiple Image Facility (MIF) ID are retained across the LPAR name change. The master keys in the Crypto Express feature coprocessor that were associated with the old LPAR NAME1 are retained. No explicit action is taken against a cryptographic component for this dynamic change.

Coprocessors: Cryptographic coprocessors are not tied to LPAR numbers or MIF IDs. They are set up with PCIe adapter numbers and domain indexes that are defined in the partition image profile. You can dynamically configure them to a partition, and change or clear them when needed.

6.9.2 PCIe Crypto Express as an EP11 coprocessor

The Crypto Express4S card configured in Secure IBM Enterprise PKCS #11 (EP11) coprocessor mode provides PKCS #11 secure key support. Before EP11, ICSF PKCS #11 implementation only supported 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

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 Express4S feature.

6.9.3 PCIe Crypto Express as an accelerator

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. This reconfiguration has the following characteristics:

- ▶ It is done through the Support Element.
- ▶ It is done at the PCIe cryptographic adapter level. A Crypto Express3 feature can host a coprocessor and an accelerator, two coprocessors, or two accelerators.
- ▶ It works both ways, from coprocessor to accelerator and from accelerator to coprocessor. Master keys in the coprocessor domain can be optionally preserved when it is reconfigured to be an accelerator.
- ▶ 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 Express 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 bit to 4,096 bit, in the Modulus Exponent (ME) and Chinese Remainder Theorem (CRT) formats.

6.9.4 IBM Common Cryptographic Architecture (CCA) Enhancements

A new set of cryptographic functions and callable services are provided by IBM CCA LIC to enhance the functions that secure financial transactions and keys. These functions require ICSF FMID HCR77A0 and Secure IBM CCA coprocessor mode.

- ▶ Improved wrapping key strength: To comply with cryptographic standards, including ANSI X9.24 Part 1 and PCI-HSM, a key must not be wrapped with a key weaker than itself. Many CCA verbs allow the client to select the key wrapping key. With this release, CCA allows the client to configure the coprocessor to ensure that the client's system meets these key wrapping requirements. It can be configured to respond in one of three ways when a key is wrapped with a weaker key:

- Ignore weak wrapping
- Complete the requested operation but return a warning message
- Prohibit weak wrapping altogether

- ▶ Derived Unique Key Per Transaction (DUKPT) for MAC and encryption keys: DUKPT is defined in the ANSI X9.24 Part 1 standard. It provides a method in which a separate key is used for each transaction or other message that is sent from a device. Therefore, an attacker who is able to discover the value of a key is able to gain information only about a single transaction. The other transactions remain secure. The keys are derived from a base key that is initially loaded into the device, but erased as soon as the first keys are derived from it. Those keys, in turn, are erased as subsequent keys are derived.

The original definition of DUKPT only allowed derivation of keys to be used in encryption of PIN blocks. The purpose was to protect PINs that were entered at a POS device and then sent to a host system for verification. Recent versions of X9.24 Part 1 expanded this process so that DUKPT can also be used to derive keys for MAC generation and verification, and for data encryption and decryption. Three separate variations of the DUKPT key derivation process are used so that there is key separation between the keys that are derived for PIN, MAC, and encryption purposes.

- ▶ Secure Cipher Text Translate2 (CTT2): CTT2 is a new data encryption service that takes input data that is encrypted with one key and returns the same data encrypted under a different key. This verb can securely change the encryption key for cipher text without exposing the intermediate plain text. The decryption of data and reencryption of data happen entirely inside the secure module on the Crypto Express feature.
- ▶ Compliance with new random number generation standards: The standards that define acceptable methods for generating random numbers have been enhanced to include improved security properties. The Crypto Express coprocessor function is updated to support methods compliant with these new standards.

In this release, the random number generation in the Crypto Express feature when defined as a coprocessor conforms to the Deterministic Random Bit Generator (DRBG) requirements by using the SHA-256-based DRBG mechanism. These requirements are defined in NIST Special Publication 800-90/90A. The methods in these NIST standards supersede those previously defined in NIST FIPS 186-2, ANSI X9.31, and ANSI X9.62. These improvements help meet the timeline that is outlined in Chapter 4 of NIST SP800-131 for switching to the new methods and discontinuing the old methods.

- ▶ EMV enhancements for applications that support American Express cards: Two changes have been made to the CCA APIs to help improve support of EMV card applications that support American Express cards. The Transaction_Validation verb is used to generate and verify American Express card security codes (CSCs).

This release also adds support for the American Express CSC version 2.0 algorithm that is used by contact and contactless cards. The PIN_Change/Unblock verb is used for PIN maintenance. It prepares an encrypted message portion for communicating an original or

replacement PIN for an EMV smart card. The verb embeds the PINs in an encrypted PIN block from information that is supplied. With this CCA enhancement, PIN_Change/Unblock adds support for the message format that is used to change or unblock the PIN on American Express EMV cards.

6.10 TKE workstation feature

The TKE workstation is an optional feature that offers key management functions. The TKE workstation, FC 0841, 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 (FC 0841) was the first to have the 4765 crypto card installed. TKE LIC V7.2 requires CEX4 or CEX3, and TKE workstation FC 0841.

Adapters: The TKE workstation supports Ethernet adapters only to connect to a LAN.

A TKE workstation is part of a customized solution for using Integrated Cryptographic Service Facility for z/OS (ICSF for z/OS) or Linux for System z. This program provides a basic key management system for the cryptographic keys of a zEC12 that has Crypto Express features installed. It is configured for using DES, AES, ECC, and PKA cryptographic keys.

The TKE provides a secure, remote, and flexible method of providing Master Key Part Entry, and to remotely manage PCIe Cryptographic Coprocessors. The cryptographic functions on the TKE are run by one PCIe Cryptographic Coprocessor. The TKE workstation communicates with the System z 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 FC 0841 can be used to control the zEC12. It can also be used to control the z196, z114, z10 EC, z10 BC, z9 EC, z9 BC, z990, and z890 servers.

6.10.1 TKE 7.0 Licensed Internal Code (LIC)

The TKE workstation FC 0841 with LIC 7.0 offers a significant number of enhancements:

- ▶ ECC master key support
ECC keys are protected by using a new ECC master key (256-bit AES key). From the TKE, administrators can generate key material, load or clear the new ECC master key register, and clear the old ECC master key register. The ECC key material can be stored on the TKE or on a smart card.
- ▶ CBC default settings support
The TKE provides function that allows the TKE user to set the default key wrapping method that is used by the host crypto module.
- ▶ TKE Audit Record Upload Configuration Utility support
The TKE Audit Record Upload Configuration Utility allows TKE workstation audit records to be sent to a System z host. They are then saved on the host as z/OS System Management Facilities (SMF) records. The SMF records have a record type of 82 (ICSF) and a subtype of 29. TKE workstation audit records are sent to the same TKE Host Transaction Program that is used for TKE operations.
- ▶ USB flash memory drive support
The TKE workstation now supports a USB flash memory drive as a removable media device. When a TKE application displays media choices, you can choose a USB flash

memory drive if the IBM supported drive is plugged into a USB port on the TKE. The drive must have been formatted for the specified operation.

- ▶ Stronger pin strength support

TKE smart cards that are created on TKE 7.0 require a 6-digit pin rather than a 4-digit pin. TKE smart cards that were created before TKE 7.0 continue to use 4-digit pins, and work on TKE 7.0 without changes. Take advantage of the stronger pin strength by initializing new TKE smart cards and copying the data from the old TKE smart cards to the new smart cards.

- ▶ Stronger password requirements for TKE passphrase user profile support

New rules are required for the passphrase that is used for the passphrase logon to the TKE workstation crypto adapter. The passphrase must meet the following requirements:

- Be 8 - 64 characters long
- Contain at least two numeric and two non-numeric characters
- Not contain the user ID

These rules are enforced when you define a new user profile for passphrase logon, or when you change the passphrase for an existing profile. Your current passphrases will continue to work.

- ▶ Simplified TKE usability with Crypto Express migration wizard

A wizard is now available to allow users to collect data, including key material, from a Crypto Express coprocessor and migrate the data to a different Crypto Express coprocessor. The target Crypto Express coprocessor must have the same or greater capabilities. This wizard helps migrate from Crypto Express2 to Crypto Express3. Crypto Express2 is not supported on the zEC12, IBM zEnterprise 196 (z196), and z114. The following benefits are obtained when you use this wizard:

- Reduces migration steps, minimizing user errors
- Minimizes the number of user clicks
- Reduces migration task duration significantly

6.10.2 TKE 7.1 Licensed Internal Code (LIC)

The TKE workstation FC 0841 with LIC 7.1 offers these enhancements:

- ▶ New access control support for all TKE applications

Every TKE application and the ability to create and manage the crypto module and domain groups now require the TKE local cryptographic adapter profile to have explicit access to the TKE application or function that you want to run. This change was made to provide more control of the functions that TKE users are allowed to perform.

- ▶ New migration utility

During a migration from a lower release of TKE to TKE 7.1 LIC, you must add access control points to the existing roles. The new access control points can be added through the new Migrate Roles Utility or by manually updating each role through the Cryptographic Node Management Utility. The IBM supplied roles created for TKE 7.1 LIC have all of the access control points needed to run the functions permitted in TKE releases before TKE 7.1 LIC.

- ▶ Single process for loading an entire key

The TKE now has a wizard-like feature that takes users through the entire key loading procedure for a master or operational key. The feature preserves all of the existing separation of duties and authority requirements for clearing, loading key parts, and completing a key. The procedure saves time by guiding users through the key loading procedure. However, this feature does not reduce the number of people required to perform the key load procedure.
- ▶ Single process for generating multiple key parts of the same type

The TKE now has a wizard-like feature that allows a user to generate more than one key part at a time. The procedure saves time because the user must start the process only one time, and the TKE efficiently generates the wanted number of key parts.
- ▶ AES operational key support

CCA V4.2 for the Crypto Express feature includes three new AES operational key types. From the TKE, users can load and manage the new AES EXPORTER, IMPORTER, and CIPHER operational keys from the TKE workstation crypto module notebook.
- ▶ Decimalization table support

CCA V4.2 for the Crypto Express feature includes support for 100 decimalization tables for each domain on a Crypto Express feature. From the TKE, users can manage the decimalization tables on the Crypto Express feature from the TKE workstation crypto module notebook. Users can manage the tables for a specific domain, or manage the tables of a set of domains if they are using the TKE workstation Domain Grouping function.
- ▶ Host cryptographic module status support

From the TKE workstation crypto module notebook, users are able to display the status of the host cryptographic module that is being managed. If they view the Crypto Express feature module information from a crypto module group or domain group, they see only the status of the group's master module.
- ▶ Display of active IDs on the TKE console

A user can be logged on to the TKE workstation in privileged access mode. In addition, the user can be signed onto the TKE workstation's local cryptographic adapter. If a user is signed on in privileged access mode, that ID is shown on the TKE console. With this new support, both the privileged access mode ID and the TKE local cryptographic adapter ID are displayed on the TKE console.
- ▶ Increased number of key parts on smart card

If a TKE smart card is initialized on a TKE workstation with a 7.1 level of LIC, it is able to hold up to 50 key parts. Previously, TKE smart cards held only 10 key parts.
- ▶ Use of ECDH to derive shared secret

When the TKE workstation with a 7.1 level of LIC exchanges encrypted material with a Crypto Express card at CCA level V4.2, ECDH is used to derive the shared secret. This process increases the strength of the transport key that is used to encrypt the material.

6.10.3 TKE 7.2 Licensed Internal Code (LIC)

The TKE workstation FC 0841 with LIC 7.2 offers even more enhancements:

- ▶ Support for the Crypto Express4S feature when configured as an EP11 coprocessor

The TKE workstation is required to manage a Crypto Express4S feature that is configured as an EP11 coprocessor. Allow domain grouping between Crypto Express4S features that are defined only as EP11. The TKE smart card reader (FC 0885) is mandatory.

EP11 requires the use of the new smart card part 74Y0551 (FC 0884 and FC 0885). The new smart card can be used for any of the six types of smart cards that are used on TKE. Two items must be placed on the new smart cards:

 - Master key material: The Crypto Express4S feature has master keys for each domain. The key material must be placed on a smart card before the key material can be loaded.
 - Administrator signature keys: When commands are sent to the Crypto Express4S feature, they must be signed by administrators. Administrator signature keys must be on smart cards.
- ▶ Support for the Crypto Express4S feature when configured as a CCA coprocessor

Crypto Express4S (defined as a CCA coprocessor) is managed the same way as any other CCA-configured coprocessor. A Crypto Express4S can be in the same crypto module group or domain group as a Crypto Express4S, Crypto Express3, and Crypto Express2 feature.
- ▶ Support for 24-byte DES master keys

CCA supports both 16-byte and 24-byte DES master keys. The DES master key length for a domain is determined by a new domain control bit that can be managed by using the TKE.

Two Access Control Points (ACPs) allow the user to choose between warning about or prohibiting the loading of a weak master key. The latest CCA version is required.
- ▶ Protection of generated RSA keys with AES importer keys

TKE-generated RSA keys are encrypted by AES keys before they are sent to System z. It allows the generation of 2046-bit and 4096-bit RSA keys for target crypto card use.
- ▶ New DES operational keys

Four new DES operational keys can be managed from the TKE workstation (FC 0841). The DES keys can be any of the following types:

 - CIPHERXI
 - CIPHERXL
 - CIPHERXO
 - DUKPT-KEYGENKY

The new keys are managed the same way as any other DES operational key.
- ▶ New AES CIPHER key attribute

A new attribute, “key can be used for data translate only,” can be specified now when you create an AES CIPHER operational key part.
- ▶ Creation of corresponding keys allowed

In certain cases, operational keys must be loaded to different host systems to serve an opposite purpose. For example, one host system needs an exporter key encrypting key, and another system needs a corresponding importer key encrypting key with the same value. The TKE workstation now allows nine types of key material to be used for creating a corresponding key.

- ▶ Support for four smart card readers

The TKE workstation supports two, three, or four smart card readers when smart cards are used. The additional readers are added to help reduce the number of smart card swaps needed while you manage EP11-configured coprocessors. EP11 can be managed with only two smart card readers. CCA-configured coprocessors can be managed with three or four smart card readers.

6.10.4 Logical partition, TKE host, and TKE target

If one or more logical partitions (LPARs) are configured to use Crypto Express 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 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 as set for an LPAR through the Support Element determine whether the workstation is a TKE host or TKE target.

6.10.5 Optional smart card reader

You can add an optional smart card reader (FC 0885) to the TKE workstation. One FC 0885 includes two Smart Card Readers, two cables to connect 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 additional 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).

6.11 Cryptographic functions comparison

Table 6-4 lists functions or attributes on zEC12 of the three cryptographic hardware features. In the table, X indicates that the function or attribute is supported.

Table 6-4 Cryptographic functions on zEC12

Functions or attributes	CPACF ^a	CEX4C ^a	CEX4P ^a	CEX4A ^a	CEX3C ^{ab}	CEX3A ^{ab}
Supports z/OS applications using ICSF	X	X	X	X	X	X
Supports Linux on System z CCA applications	X	X	-	X	X	X
Encryption and decryption using secret-key algorithm	-	X	X	-	X	-
Provides highest SSL/TLS handshake performance	-	-	-	X	-	X

Functions or attributes	CPACF ^a	CEX4C ^a	CEX4P ^a	CEX4A ^a	CEX3C ^{ab}	CEX3A ^{ab}
Supports SSL/TLS functions	X	X	-	X	X	X
Provides highest symmetric (clear key) encryption performance	X	-	-	-	-	-
Provides highest asymmetric (clear key) encryption performance	-	-	-	X	-	X
Provides highest asymmetric (encrypted key) encryption performance	-	X	X	-	X	-
Disruptive process to enable	-	Note ^c	Note ^c	Note ^c	Note ^c	Note ^c
Requires IOCDs definition	-	-	-	-	-	-
Uses CHPID numbers	-	-	-	-	-	-
Uses PCHIDs		X ^d	X ^d	X ^d	X ^d	X ^d
Requires CPACF enablement (FC 3863)	X ^e	X ^e	X ^e	X ^e	X ^e	X ^e
Requires ICSF to be active	-	X	X	X	X	X
Offers user programming function (UDX)	-	X	-	-	X	-
Usable for data privacy: Encryption and decryption processing	X	X	X	-	X	-
Usable for data integrity: Hashing and message authentication	X	X	X	-	X	-
Usable for financial processes and key management operations	-	X	X	-	X	-
Crypto performance RMF monitoring	-	X	X	X	X	X
Requires system master keys to be loaded	-	X	X	-	X	-
System (master) key storage	-	X	X	-	X	-
Retained key storage	-	X	-	-	X	-
Tamper-resistant hardware packaging	-	X	X	X ^f	X	X ^f
Designed for FIPS 140-2 Level 4 certification	-	X	X	-	X	-
Supports Linux applications performing SSL handshakes	-	-	-	-	-	X
RSA functions	-	X	X	X	X	X
High performance SHA-1 and SHA2	X	X	X	-	X	-
Clear key DES or triple DES	X	-	-	-	-	-

Functions or attributes	CPACF ^a	CEX4C ^a	CEX4P ^a	CEX4A ^a	CEX3C ^{ab}	CEX3A ^{ab}
Advanced Encryption Standard (AES) for 128-bit, 192-bit, and 256-bit keys	X	X	X	-	X	-
Pseudorandom number generator (PRNG)	X	X	X	-	X	-
Clear key RSA	-	-	-	X	-	X
Europay MasterCard VISA (EMV) support	-	X	-	-	X	-
Public Key Decrypt (PKD) support for Zero-Pad option for clear RSA private keys	-	X	-	X	X	X
Public Key Encrypt (PKE) support for Mod_Raised_to Power (MRP) function	-	X	-	X	X	X
Remote loading of initial keys in ATM	-	X	-	-	X	-
Improved key exchange with non-CCA systems	-	X	-	-	X	-
ISO 16609 CBC mode triple DES message authentication code (MAC) support	-	X	-	-	X	-

- a. This configuration requires CPACF enablement feature code 3863.
- b. Available only in a carry-forward basis when you are upgrading from earlier generations to zEC12.
- c. 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.
- d. One PCHID is required for each PCIe cryptographic adapter.
- e. 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.
- f. This feature is physically present but is not used when configured as an accelerator (clear key only).

6.12 Software support

The software support levels are listed in 8.4, “Cryptographic support” on page 316.



zEnterprise BladeCenter Extension (zBX) Model 003

IBM has extended the mainframe system by bringing select IBM BladeCenter product lines under the same management umbrella. This is called the zEnterprise BladeCenter Extension (zBX) Model 003, and the common management umbrella is the IBM zEnterprise Unified Resource Manager.

The zBX brings the computing capacity of systems in blade form factor to the IBM zEnterprise System. It is designed to provide a redundant hardware infrastructure that supports the multi-platform environment of the zEC12 in a seamless, integrated way.

Also key to this hybrid environment is the Unified Resource Manager. The Unified Resource Manager helps deliver end-to-end infrastructure virtualization and management, as well as the ability to optimize multi-platform technology deployment according to complex workload requirements. For more information about the Unified Resource Manager, see Chapter 12, “Hardware Management Console and Support Element” on page 431 and *IBM zEnterprise Unified Resource Manager*, SG24-7921.

This chapter introduces the zEnterprise BladeCenter Extension (zBX) Model 003 and describes its hardware components. It also explains the basic concepts and building blocks for zBX connectivity.

You can use this information for planning purposes and to help define the configurations that best fit your requirements.

This chapter includes the following sections:

- ▶ zBX concepts
- ▶ zBX hardware description
- ▶ zBX entitlements, firmware, and upgrades
- ▶ zBX connectivity
- ▶ zBX connectivity examples
- ▶ References

7.1 zBX concepts

The integration of System z in a hybrid computing infrastructure represents a new height for mainframe functionality and qualities of service. It is a cornerstone for the IT infrastructure, especially when flexibility for rapidly changing environments is needed.

The IBM zEC12 system characteristics make it valuable for mission-critical workloads. Today, most of these workloads employ a multi-tiered architecture that span various hardware and software platforms. However, there are differences in the quality of service offered by the various platforms. There are also distinct configuration procedures for hardware and software, operational management, software servicing, and failure detection and correction. These procedures in turn require personnel with distinct skill sets, various sets of operational procedures, and an integration effort that is not trivial and, therefore, not often achieved. Failure to achieve integration translates into lack of flexibility and agility, which can affect the bottom line.

IBM mainframe systems have been providing specialized hardware and fit-for-purpose (tuned to the task) computing capabilities for a long time. In addition to the machine instruction assists, another early example is the vector facility of the IBM 3090¹. Other specialty hardware includes the System Assist Processor for I/O handling (implemented in the 370-XA architecture), the Coupling Facility, and the Cryptographic processors. Furthermore, all the I/O cards are specialized, dedicated hardware components with sophisticated firmware that offload processing from the System z processor units (PUs).

The common theme with all of these specialized hardware components is their seamless integration within the mainframe. The zBX components are configured, managed, and serviced in the same way as the other components of the System z central processor complex (CPC). Although the zBX processors are not z/Architecture PUs, the zBX is handled by System z management firmware that is called the IBM zEnterprise Unified Resource Manager. The zBX hardware features are integrated into the mainframe system (are not add-ons).

System z has long been an integrated, heterogeneous system. With zBX, that integration reaches a new level. zEnterprise with its zBX infrastructure allows you to run an application that spans z/OS, z/VM, Linux on System z, AIX, Linux on System x, and Microsoft Windows, yet have it under a single management umbrella. Also, zBX can host and integrate special-purpose workload optimizers, such as the WebSphere DataPower Integration Appliance XI50 for zEnterprise (DataPower XI50z).

7.2 zBX hardware description

The zBX has a machine type of 2458-003 and attaches to the zEC12. It can host integrated multi-platform systems and heterogeneous workloads, with integrated advanced virtualization management. The zBX Model 003 is configured with the following key components:

- ▶ One to four standard 19-inch IBM 42U zEnterprise racks with required network and power infrastructure
- ▶ One to eight BladeCenter chassis with a combination of up to 112² different blades

¹ <http://domino.watson.ibm.com/tchjr/journalindex.nsf/9fe6a820aae67ad785256547004d8af0/212e35ae6fecea7185256bfa00685bff!OpenDocument>

² The number of chassis and blades varies depending on the type of the blades that are configured within zBX. For more information, see 7.2.4, “zBX blades” on page 236.

- ▶ Redundant power and I/O infrastructure for fault tolerance and higher availability
- ▶ Management support through the zEC12 Hardware Management Console (HMC) and Support Element (SE)

For more information about zBX reliability, availability, and serviceability (RAS), see 10.6, “RAS capability for zBX” on page 401.

The zBX can be ordered with a new zEC12, or as a miscellaneous equipment specification (MES) to an existing zEC12. If an IBM zEnterprise 196 (z196) controlling a zBX is upgraded to a zEC12, the controlled zBX Model 002 will be also upgraded to a Model 003. Either way, the zBX is treated as an extension to zEC12 and cannot be ordered as a stand-alone feature.

Figure 7-1 shows a zEC12 with a maximum zBX configuration. The first rack (Rack B) in the zBX is the primary rack, which has one or two BladeCenter chassis and four Top of Rack (ToR) switches. The other three racks (C, D, and E) are expansion racks hosting up to two BladeCenter chassis each.



Figure 7-1 zEC12 with a maximum zBX configuration

7.2.1 zBX racks

The zBX Model 003 (2458-003) hardware is housed in up to four IBM zEnterprise racks. Each rack is an industry-standard 19-inch, 42U high rack with four sidewall compartments to support the installation of power distribution units (PDUs) and switches, with additional space for cable management.

Figure 7-2 on page 230 shows the rear view of a two-rack zBX configuration, including the following components:

- ▶ Two 1000BASE-T Top of Rack (ToR) switches (Rack B only) for the intranode management network (INMN).
- ▶ Two 10 GbE ToR switches (Rack B only) for the intraensemble data network (IEDN).

- ▶ Up to two BladeCenter chassis in each rack with the following components:
 - Up to 14 blade server slots per chassis
 - 1 Gbps Ethernet Switch Modules (ESMs)
 - 10 Gbps High speed switch Ethernet (HSS) modules
 - 8 Gbps Fibre Channel switches for connectivity to the SAN environment³
 - Blower modules
- ▶ PDUs

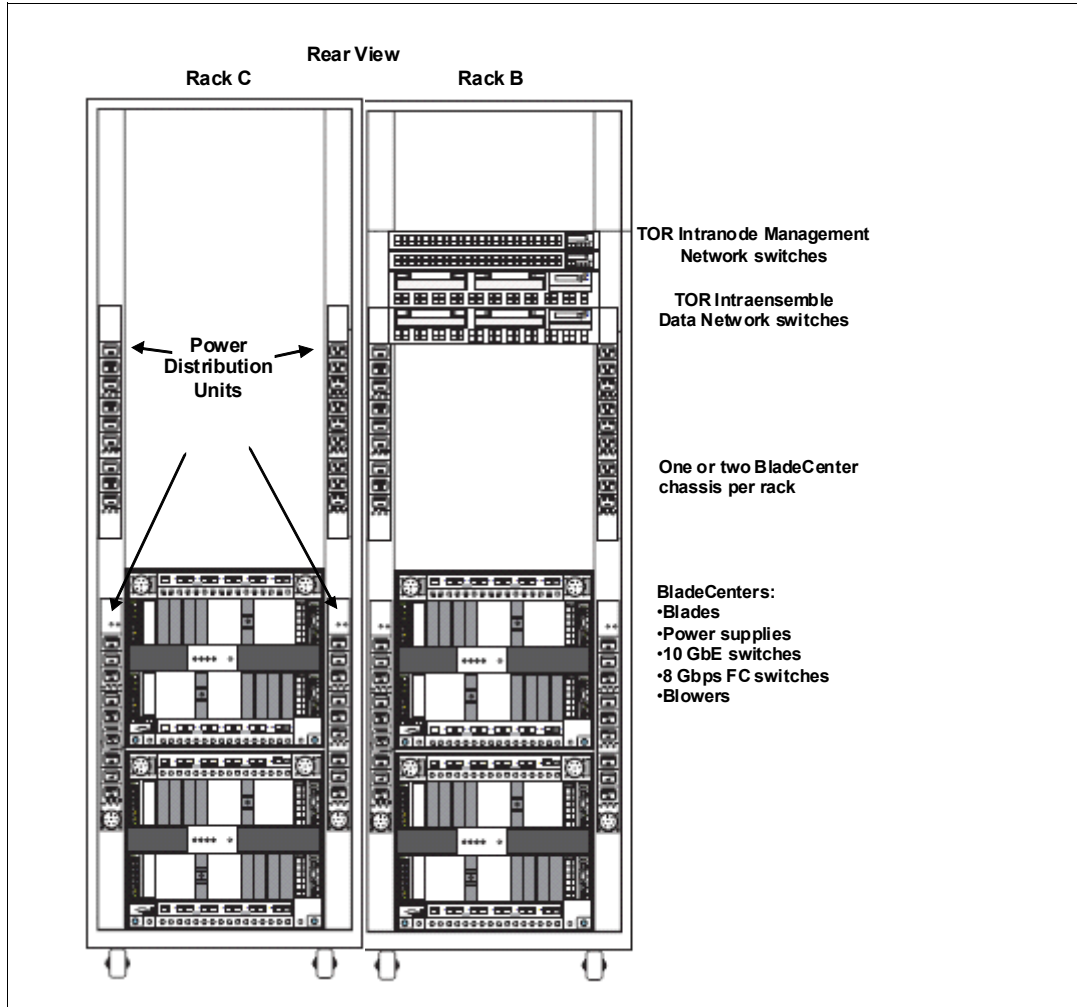


Figure 7-2 zBX racks rear view with BladeCenter chassis

A zBX rack supports a maximum of two BladeCenter chassis. Each rack is designed for enhanced air flow and ships loaded with the initial configuration. It can be upgraded onsite.

The zBX racks are shipped with lockable standard non-acoustic doors and side panels. The following optional features are also available:

- ▶ IBM rear door heat eXchanger (FC 0540) reduces the heat load of the zBX that is emitted into ambient air. The rear door heat eXchanger is an air-to-water heat exchanger that diverts the heat of the zBX to chilled water (client-supplied data center infrastructure). The rear door heat eXchanger requires external conditioning units for its use.

³ Client-supplied FC switches are required that must support N-Port ID Virtualization (NPIV). Some FC switch vendors also require “interop” mode. Check the interoperability matrix for the latest details:
<http://www-03.ibm.com/systems/support/storage/ssic/interoperability.wss>

- ▶ IBM acoustic door (FC 0543) can be used to reduce the noise from the zBX.
- ▶ Height reduction (FC 0570) reduces the rack height to 36U high and accommodates doorway openings as low as 1832 mm (72.1 inches). Order this choice if you have doorways with openings less than 1941 mm (76.4 inches) high.
- ▶ Top exit I/O and power exit: On a zBX, you now have the option of ordering the infrastructure to support the top exit of your fiber optic cables and your copper cables for the 1000BASE-T Ethernet features, as well as the power cords. Top exit I/O cabling is designed to provide you with an additional option. Instead of all of your cables exiting under the server and under the raised floor, you now have the flexibility to choose the option that best meets the requirements of your data center (see Figure 7-3). Top exit is FC 0545 and is available for zBX Model 003 only. There is one feature per rack. Therefore, if you order two racks, you need to order two features.

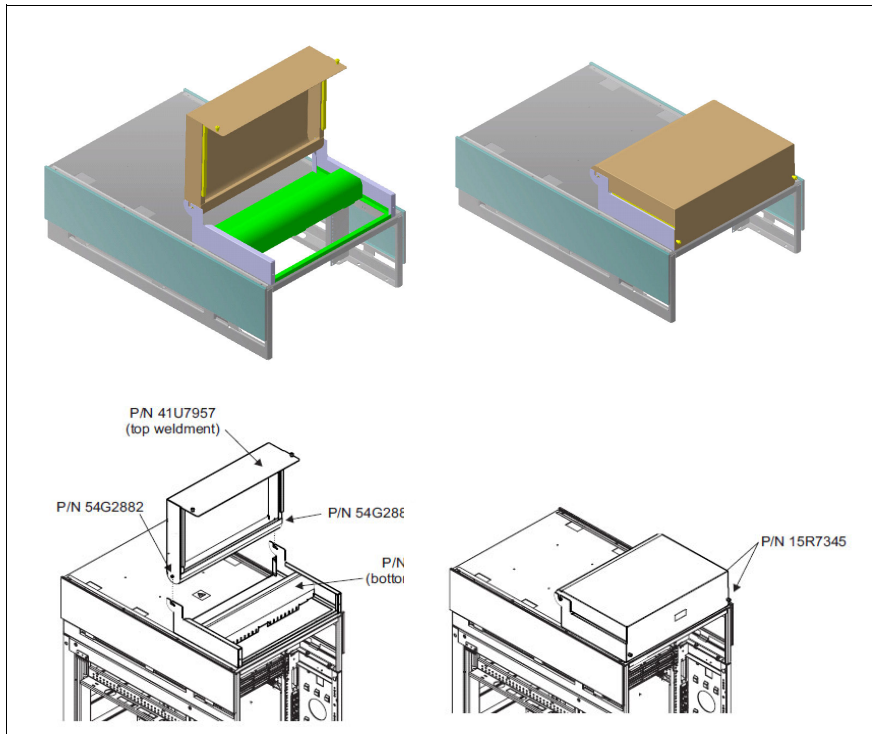


Figure 7-3 zBX top I/O exit

7.2.2 Top of Rack (ToR) switches

The four ToR switches are installed in the first rack (Rack B). Expansion racks (Rack C, D, and E) do not require additional ToR switches.

The ToR switches are located near the top of the rack, and are mounted from the rear of the rack. From the top down, there are two 1000BASE-T switches for the INMN and two 10 GbE switches for the IEDN.

INMN switches

A zBX Model 003 can be managed only by one zEC12 through the INMN connections. Each virtual LAN (VLAN)-capable 1000BASE-T switch has 48 ports. The switch ports are reserved in the following manner:

- ▶ One port for each of the two bulk power hubs (BPHs) on the controlling zEC12

- ▶ One port for each of the advanced management modules (AMMs) and ESMs in each zBX BladeCenter chassis
- ▶ One port for each of the two IEDN 10 GbE ToR switches
- ▶ Two ports each for interconnecting the two switches

Both switches have the same connections to the corresponding redundant components (BPH, AMM, ESM, and IEDN ToR switches) to avoid any single point of failure. Table 7-5 on page 247 shows port assignments for the 1000BASE-T ToR switches.

Tip: Although IBM provides a 26-m (85.3 ft) cable for the INMN connection, zBX needs to be installed next to, or near the *controlling zEC12*. This configuration provides easy access to the zBX for service-related operations.

IEDN switches

Each (VLAN-capable) 10 GbE ToR switch has 40 ports dedicated to the IEDN. The switch ports have the following connections:

- ▶ Up to 16 ports are used for connections to an HSS module (SM07 and SM09) of each BladeCenter chassis in the same zBX (as part of IEDN). These connections provide data paths to blades.
- ▶ Up to eight ports are used for OSA-Express5S 10GBE, OSA-Express4S 10 GbE, or OSA-Express3 10 GbE (long reach (LR) or short reach (SR)) connections to the ensemble CPCs (as part of IEDN). These connections provide data paths between the ensemble CPCs and the blades in a zBX.
- ▶ Up to seven ports are used for zBX to zBX connections within the same ensemble (as part of the IEDN).
- ▶ Up to seven ports are used for the client-managed data network. Client-network connections are not part of IEDN, and they cannot be managed or provisioned by the Unified Resource Manager. The Unified Resource Manager recognizes them as migration connections, and provides access control from the client network to the 10 GbE ToR switches.
- ▶ The management port is connected to the INMN 1000BASE-T ToR switch.
- ▶ Two ports are used for interconnections between the two switches (as a failover path), using two Direct Attach Cables (DAC).

Figure 7-4 shows the connections of ToR switches and the first BladeCenter chassis in frame B. For more information about the connectivity options for the INMN and the IEDN, as well as the connectivity rules, see 7.4, “zBX connectivity” on page 244.

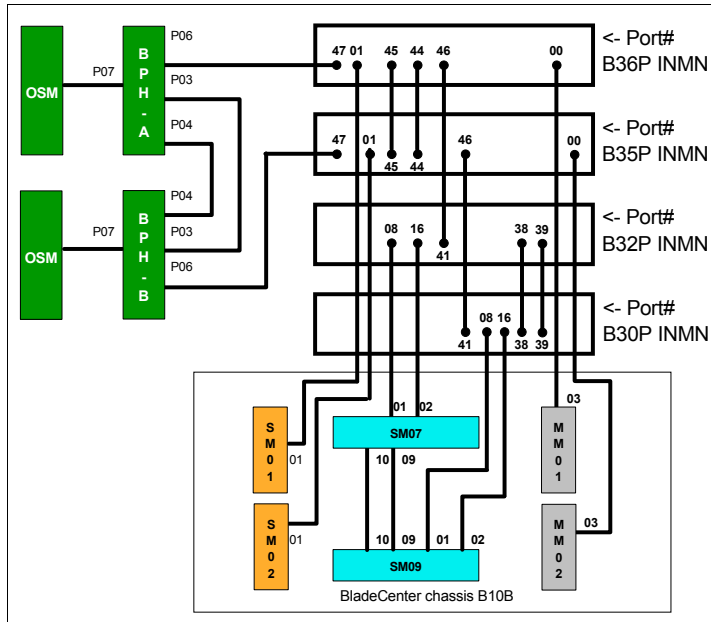


Figure 7-4 Graphical illustration of zEnterprise network connections

7.2.3 zBX BladeCenter chassis

Each zBX BladeCenter chassis is designed with additional components installed for high levels of resiliency.

The front of a zBX BladeCenter chassis has the following components:

- ▶ Blade server slots

There are 14 blade server slots (BS01 - BS14) available in a zBX BladeCenter chassis. Each slot can house any zBX supported blades, with the following restrictions:

- Slot 14 cannot hold a double-wide blade.
- The DataPower XI50z blades are double-wide. Each feature takes two adjacent BladeCenter slots, so the maximum number of DataPower blades per BladeCenter is seven. The maximum number of DataPower blades per zBX is 28.

- ▶ Power module

The power module includes a power supply and a three-pack of fans. Two of three fans are needed for power module operation. Power modules 1 and 2 (PM01 and PM02) are installed as a pair to provide a power supply for the seven blade server slots from BS01 to BS07. Power modules 3 and 4 (PM03 and PM04) support the BS08 to BS14.

The two different power connectors (marked with “1” and “2” in Figure 7-5 on page 234) provide power connectivity for the power modules (PMs) and blade slots. PM01 and PM04 are connected to power connector 1. PM02 and PM03 are connected to power connector 2. Therefore, each slot has fully redundant power from a different power module that is connected to a different power connector.

Figure 7-5 shows the rear view of a zBX BladeCenter chassis.

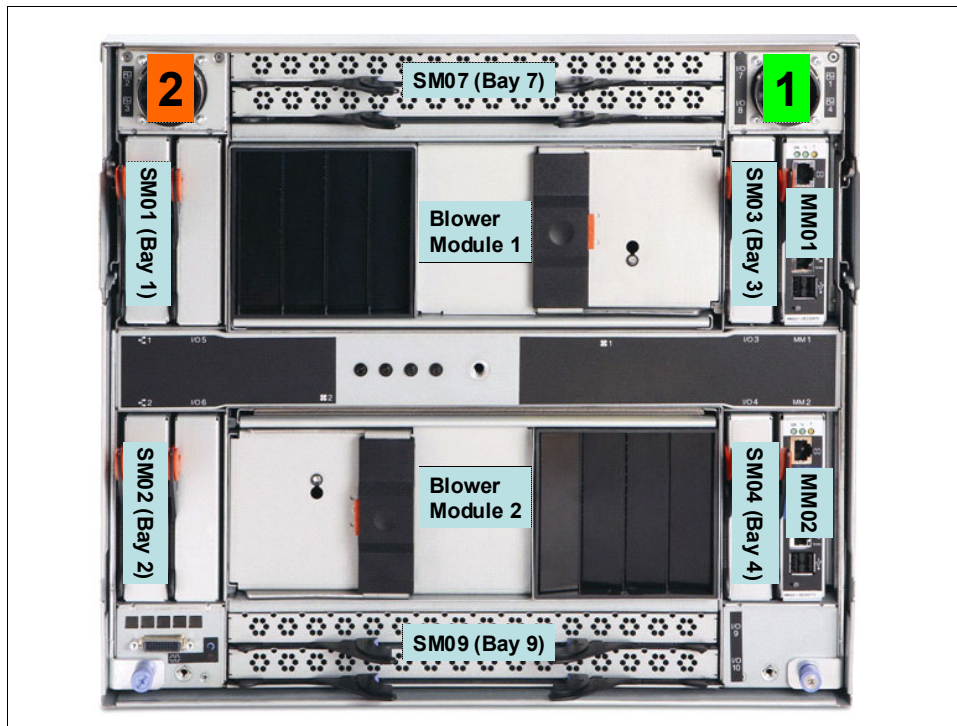


Figure 7-5 zBX BladeCenter chassis rear view

The rear of a zBX BladeCenter chassis has following components:

- ▶ Advanced management module (AMM)

The AMM provides systems management functions and kernel-based virtual machine (KVM) multiplexing for all blade servers that support KVM. It controls the external keyboard, mouse, and video connections, for use by a local console, and a 10/100 Mbps Ethernet remote management connection.

Blade console support: Use of KVM is not supported on zBX. All required management functions are available on the controlling zEC12 SE or the HMC.

The management module communicates with all components in the BladeCenter unit, detecting their presence, reporting their status, and sending alerts for error conditions when required.

The service processor in the management module communicates with the service processor (integrated management module (IMM)) in each blade server. This process supports features, such as blade server power-on requests, error and event reporting, and requests to use the BladeCenter shared media tray.

The AMMs are connected to the INMN through the 1000BASE-T ToR switches. Therefore, firmware and configuration for the AMM are controlled by the SE of the controlling zEC12, together with all service management and reporting functions of AMMs.

Two AMMs (MM01 and MM02) are installed in the zBX BladeCenter chassis. Only one AMM has primary control of the chassis (it is active). The second module is in passive (standby) mode. If the active module fails, the second module is automatically enabled with all of the configuration settings of the primary module.

► Ethernet switch module (ESM)

Two 1000BASE-T (1 Gbps) Ethernet switch modules (SM01 and SM02) are installed in switch bays 1 and 2 in the chassis. Each ESM has 14 internal full-duplex Gigabit ports. One is connected to each of the blade servers in the BladeCenter chassis, and two internal full-duplex 10/100 Mbps ports are connected to the AMM modules. Six 1000BASE-T copper RJ-45 connections are used for INMN connections to the ToR 1000BASE-T switches.

The ESM port 01 is connected to one of the 1000BASE-T ToR switches. As part of the INMN, configuration and firmware of ESM are controlled by the controlling zEC12 SE.

► High-speed switch module (HSS)

Two high-speed switch modules (SM07 and SM09) are installed to switch bays 7 and 9. The HSS modules provide 10 GbE uplinks to the 10 GbE ToR switches, and 10 GbE downlinks to the blades in the chassis.

Ports 01 and 02 are connected to one of the 10 GbE ToR switches. Ports 09 and 10 are used to interconnect HSS in bays 7 and 9 as a redundant failover path.

► 8-Gbps Fibre Channel switch module

Two 8 Gbps Fibre Channel (FC) switches (SM03 and SM04) are installed in switch bays 3 and 4. Each switch has 14 internal ports that are reserved for the blade servers in the chassis, and six external Fibre Channel ports to provide connectivity to the SAN environment.

► Blower module

There are two hot swap blower modules installed. The blower speeds vary depending on the ambient air temperature at the front of the BladeCenter unit and the temperature of internal BladeCenter components. If a blower fails, the remaining blowers run full speed.

► BladeCenter mid-plane fabric connections

The BladeCenter mid-plane provides redundant power, control, and data connections to a blade server. It does so by internally routed chassis components (power modules, AMMs, switch modules, and media tray) to connectors in a blade server slot.

There are six connectors in a blade server slot on the mid-plane (from top to bottom):

- Top 1X fabric connects blade to MM01, SM01, and SM03
- Power connector from power module 1 (blade server slots 1 - 7) or power module 3 (blade server slots 8 - 14)
- Top 4X fabric connects blade to SM07
- Bottom 4X fabric connects blade to SM09
- Bottom 1X fabric connects blade to MM02, SM02, and SM04
- Power connector from power module 2 (blade server slot 1 - 7) or power module 4 (blade server slot 8 - 14)

Each blade server therefore has redundant power, data, and control links from separate components.

7.2.4 zBX blades

The zBX Model 003 supports the following blade types:

- ▶ IBM BladeCenter PS701 Express blades

Three configurations of IBM POWER® blades are supported, depending on their memory sizes (see Table 7-1 on page 237).

Up to 112 PS701 Express blades are supported.

- ▶ IBM WebSphere DataPower Integration Appliance XI50 for zEnterprise (DataPower XI50z)

Up to 28 IBM WebSphere DataPower XI50 for zEnterprise (DataPower XI50z) blades are supported. These blades are double-wide (each one occupies two blade server slots).

- ▶ IBM BladeCenter HX5 (7873) blades

Up to 56 IBM System x HX5 blades are supported.

All zBX blades are connected to AMMs and ESMs through the chassis mid-plane. The AMMs are connected to the INMN.

zBX blade expansion cards

Each zBX blade has two PCI Express connectors: Combination input/output vertical (CIOv), and combination form factor horizontal (CFFh). I/O expansion cards are attached to these connectors and connected to the mid-plane fabric connectors. Therefore, a zBX blade can expand its I/O connectivity through the mid-plane to the high-speed switches and switch modules in the chassis.

Depending on the blade type, 10-GbE CFFh expansion cards and 8-Gbps Fibre Channel CIOv expansion cards provide I/O connectivity to the IEDN, SAN, or client-supplied FC-attached storage.

INMN connectivity is established by using an onboard 1GbE adapter.

POWER7 blade

The POWER7 blade (Table 7-1 on page 237) is a single width blade that includes a POWER7 processor, up to 16 dual inline memory modules (DIMMs), and a hard disk drive (HDD). The POWER7 blade supports 10 GbE connections to IEDN. It also supports 8-Gbps FC connections to client-provided Fibre Channel storage through the FC switches (SM03 and SM04) in the chassis.

The POWER7 blade is loosely integrated into the zBX so that you can acquire supported blades through traditional channels from IBM. The primary HMC and SE of the controlling [zEC12](#) run entitlement management for installed POWER7 blades on a one-blade basis.

PowerVM Enterprise Edition *must* be ordered *with each* POWER7 blade. AIX 5.3, AIX 6.1, AIX 7.1, and subsequent releases are supported⁴.

⁴ As per PS701 Operating System support specification

Table 7-1 Supported configuration of POWER7 blades

Feature	FC	Config 1 quantity	Config 2 quantity	Config 3 quantity
Processor (3.0 GHz@150 W)		1	1	1
Processor Activations (quantity must be equal to eight total)	8411 8412	4 4	4 4	4 4
Memory kits 8-GB memory (2 x 4 GB) 16-GB memory (2 x 8 GB)	8208 8209	4 0	8 0	0 8
Internal HDD (300 GB)	8274	1	1	1
CFFh 10-GbE expansion	8275	1	1	1
CIOv 8-Gb FC expansion	8242	1	1	1
PowerVM Enterprise Edition	5228	8	8	8

DataPower XI50z blades

The IBM WebSphere DataPower Integration Appliance XI50 for zEnterprise (DataPower XI50z) is integrated into the zBX. It is a high-performance hardware appliance that offers these benefits:

- ▶ Provides fast and flexible integration with any-to-any transformation between disparate message formats with integrated message-level security and superior performance.
- ▶ Provides web services enablement for core System z applications to enable web-based workloads. As a multifunctional appliance, DataPower XI50z can help provide multiple levels of XML optimization. This optimization streamlines and secures valuable service-oriented architecture (SOA) applications.
- ▶ Enables SOA and XML applications with System z web services for seamless integration of distributed and System z platforms. It can help simplify, govern, and enhance the network security for XML and web services.
- ▶ Provides drop-in integration for heterogeneous environments by enabling core enterprise service bus (ESB) functions, including routing, bridging, transformation, and event handling.
- ▶ Offers standards-based, centralized System z governance, and extreme reliability through integrated operational controls, “call home,” and integration with RACF security through a secured private network.

The zBX provides more benefits to the DataPower appliance environment in these areas:

- ▶ Blade hardware management:
 - Improved cooling and power management controls, including cooling of the frame and energy monitoring and management of the DataPower blades
 - Virtual network provisioning
 - Call home for current and expected problems
- ▶ Hardware Management Console integration:
 - Single view that shows the System z environment together with the DataPower blades in an overall hardware operational perspective
 - Group GUI operations for functions that are supported on HMC, such as activate or deactivate blades

- ▶ Improved availability:
 - Guided placement of blades to optimize built-in redundancy in all components at the rack, BladeCenter, and HMC levels. These components include the ToR switch, ESM switches, and physical network.
 - Detection and reporting by the HMC/SE on appliance failures. The HMC/SE can also be used to recycle the DataPower appliance.
- ▶ Networking:
 - Virtual network provisioning
 - Enforced isolation of network traffic by VLAN support
 - 10-Gbps end-to-end network infrastructure
 - Built-in network redundancy
 - Network protection via IEDN, possibly meeting your need for encryption of data flowing between DataPower and the target System z server
- ▶ Monitoring and reporting:
 - Monitoring and reporting of DataPower hardware health and degraded operation by HMC
 - Monitoring of all hardware, call-home, and service actions
 - Consolidation and integration of DataPower hardware problem reporting with other problems reported in zBX
- ▶ System z value
 - Simplified ordering process of the DataPower appliance by System z allows the correct blade infrastructure to be transparently ordered.
 - Simplified upgrades keep MES history so the upgrades flow based on what is installed.
 - System z service on the zBX and DataPower blade using a common maintenance package and a single point of control. The DataPower appliance becomes part of the data center and comes under data center control.

In addition, although not specific to the zBX environment, dynamic load balancing for DataPower appliances is available by using the z/OS Communications Server Sysplex Distributor.

Configuration

The DataPower XI50z is a double-wide blade based on IBM HS22. Each one takes two BladeCenter slots. Therefore, the maximum number of DataPower blades per BladeCenter is seven, and the maximum number of DataPower blades per zBX is 28. It can coexist with POWER7 blades and with IBM BladeCenter HX5 blades in the same zBX BladeCenter. Although DataPower XI50z blades are configured and ordered as zBX (machine type 2458-003) features, they have their own machine type (2462-4BX).

The DataPower XI50z with the DataPower expansion unit has the following specifications:

- ▶ 2.13 GHz processor speed.
- ▶ Two quad core processors.
- ▶ 8 MB cache.
- ▶ 3 x 4 GB or 6 x 2 GB DIMMs (12 GB of RAM).
- ▶ 4 GB USB Flash Key that contains the DataPower XI50z firmware load.
- ▶ Two 300 GB HDDs used for logging, storing style sheets, and XML files. The hard disk array consists of two hard disk drives in a RAID-1 (mirrored) configuration.

- ▶ Two Broadcom BCM5709SEthernet adapters with TOE⁵ (integrated on system board).
- ▶ BPE4⁶ Expansion Unit, which is a sealed FRU with one-way tamper-proof screws that contains the crypto for secure SOA⁷ applications.
- ▶ XG5 accelerator Peripheral Component Interconnect Express (PCIe) card.
- ▶ CN1620 Cavium crypto PCIe card.
- ▶ Dual 10-Gb Ethernet cards.

2462 Model 4BX (DataPower XI50z)

The 2462 Model 4BX is designed to work together with the 2458 Model 003 (zBX). It is functionally equivalent to an IBM 4195-4BX with similar feature codes. The IBM 2462 Model 4BX is ordered through certain feature codes for the 2458-003.

When configuring the IBM 2458 Model 003 with FC 0611 (DataPower XI50z), order a machine type IBM 2462 Model 4BX for each configured feature code. It requires Software PID 5765-G84.

A Software Maintenance Agreement (SWMA) must be active for the IBM software that runs on the DataPower XI50z before you can obtain service or other support. Failure to maintain an SWMA results in your not being able to obtain service for the IBM software. This is true even if the DataPower XI50z is under a warranty or post-warranty IBM hardware maintenance service contract.

The DataPower XI50z includes the following license entitlements:

- ▶ DataPower Basic Enablement (FC 0650)
- ▶ IBM Tivoli® Access Manager (FC 0651)
- ▶ TIBCO (FC 0652)
- ▶ Database Connectivity (DTB) (FC 0653)
- ▶ Application Optimization (AO) (FC 0654)
- ▶ Month Indicator (FC 0660)
- ▶ Day Indicator (FC 0661)
- ▶ Hour Indicator (FC 0662)
- ▶ Minute Indicator (FC 0663)

The 5765-G84 IBM WebSphere DataPower Integration Blade XI50B feature code descriptions follow:

- ▶ 0001 License with 1-year SWMA
- ▶ 0002 Option for TIBCO
- ▶ 0003 Option for Application Optimization
- ▶ 0004 Option for Database Connectivity
- ▶ 0005 Option for Tivoli Access Manager

Every IBM 2462 Model 4BX includes feature codes 0001, 0003, and 0005 (they are optional on DataPower XI50B). Optional Software feature codes 0002 and 0004 are required if FC 0652 TIBCO or FC 0653 Database Connectivity is ordered.

The TIBCO option (FC 0002) extends the DataPower XI50z so you can send and receive messages from TIBCO Enterprise Message Service (EMS).

The option for Database Connectivity (FC 0004) extends the DataPower XI50z to read and write data from relational databases, such as IBM DB2, Oracle, Sybase, and Microsoft SQL Server.

⁵ TOE - TCP/IP Offload Engine

⁶ BPE4- BladeCenter PCI Express Gen 2 Expansion Blade

⁷ Service-oriented architecture

For software PID number 5765-G85 (registration and renewal), every IBM 2462 Model 4BX includes feature code 0001. Feature code 0003 is available at the end of the first year to renew software maintenance for one more year.

For software PID number 5765-G86 (maintenance reinstatement 12 months), feature code 0001 is available if software PID 5765-G85 feature code 0003 was not ordered before the year expired.

For software PID number 5765-G87 (3-year registration), feature code 0001 can be ordered instead of software PID 5765-G85 feature code 0003. This code makes the initial period three years rather than one year.

For software PID number 5765-G88 (3-year renewal), feature code 0001 can be used as an alternative to software PID 5765-G85 feature code 0003 if you want a three-year renewal. The maximum duration is five years.

For software PID number 5765-G89 (three years after license), feature code 0001 is available if software PID 5765-G85 feature code 0003 was not ordered before the year expired. Use this option if you want a 3-year renewal.

IBM BladeCenter HX5 (7873) blades

The IBM BladeCenter HX5 is a scalable blade server that provides new levels of utilization, performance, and reliability. It is suitable for compute and memory intensive workloads, such as database, virtualization, business intelligence, modeling and simulation, and other enterprise applications.

Select System x blades running Linux on System x and Microsoft Windows on IBM System x servers are supported in the zBX. They use the zBX integrated hypervisor for IBM System x blades (Kernel Virtual Machine-based), providing logical device integration between System z and System x blades for multi-tiered applications. System x blades are licensed separately, and are enabled and managed as part of the ensemble by Unified Resource Manager.

The following operating systems are supported:

- ▶ Linux on System x (64-bit only):
 - Red Hat Enterprise Linux (RHEL) 5.5, 5.6, 5.7, 6.0, and 6.1
 - SUSE (SLES) 10 (Service Pack (SP) P4) and up, SLES 11 SP1 and up
- ▶ Microsoft Windows Server 2008 R2, Microsoft Windows Server 2008 SP2, and Microsoft Windows Server 2012 (Datacenter Edition is preferred), 64-bit only

Support of select IBM System x blades in the zBX allows the zEnterprise to access a whole new application portfolio. Front-end applications that need access to centralized data serving are a good fit for running on the blades, as well as applications that are a front end to core CICS or IMS transaction processing, such as IBM WebSphere Application Server. You can acquire BladeCenter HX5 blades through existing channels or through IBM. POWER7, DataPower XI50z, and System x blades can be mixed in the same BladeCenter chassis. Supported configuration options are listed in Table 7-2 on page 241.

IBM BladeCenter HX5 7873 is a dual-socket 16-core blade with the following features:

- ▶ Intel 8-core processor
- ▶ Two processor sockets
- ▶ 2.13 GHz 105-W processor
- ▶ Up to 14 A16Ms features per BC-H

- ▶ Up to 16 DIMM dynamic device reconfiguration (DDR)-3 with 64, 128, 192, or 256 GB of memory
- ▶ 200 GB SSD Internal Disk

Table 7-2 Supported configurations of System x blades

System x blades	Part number	Feature code	Config 0 7873-AAx	Config 1 7873-ABx	Config 2 7873-ACx	Config 3 7873-ADx
Blades base - HX5	MT 7873	A16M	1	1	1	1
Processor 2.13 GHz 105 W	69Y3071 69Y3072	A16S A179	1 1	1 1	1 1	1 1
Intel processors			2	2	2	2
Blade width			Single	Single	Single	Single
Total cores			16	16	16	16
Memory kits: 8 GB 1333 MHz 16 GB 1333 MHz	46C0558 49Y1527	A17Q 2422	8 0	16 0	8 8	0 16
GB/core			4	8	12	16
Speed Burst	46M6843	1741	1	1	1	1
SSD Exp Card 100-GB SSD MLC No Internal RAID	46M6906 00W1122	5765 A3HQ 9012	1 2 1	1 2 1	1 2 1	1 2 1
Updated Broadcom 10 GB virtual fabric CFFh	81Y3134	A1QR	1	1	1	1
CIOv 8-Gb FC	44X1946	1462	1	1	1	1

7.2.5 Power distribution unit (PDU)

The PDUs provide connection to the main power source for intranode management network and intraensemble data network ToR switches, and the BladeCenter. The number of necessary power connections is based on the zBX configuration. A rack contains two PDUs if one BladeCenter is installed, and four PDUs if two BladeCenters are installed.

7.3 zBX entitlements, firmware, and upgrades

When you order a zBX, the controlling **zEC12** node has the entitlements features for the configured blades. The entitlements are similar to a high water mark or maximum purchased flag. Only a blade quantity equal to or less than that installed in the zBX can communicate with the CPC.

In addition, Unified Resource Manager has two management suites: Manage suite (FC 0019) and Automate/Advanced Management Suite (FC 0020):

- If the controlling **zEC12** has Manage suite (FC 0019), the same quantity that is entered for any blade enablement feature code (FC 0611, FC 0612, or FC 0613) is used for Manage Firmware (FC 0047, FC 0048, or FC 0049) of the corresponding blades.

- If the controlling zEC12 has Automate/Advanced Management Suite (FC 0020), the same quantity that is entered for Blade Enablement feature codes (FC 0611, FC 0612, or FC 0613) is used for the Manage Firmware (FC 0047, FC 0048, or FC 0049) and Automate/Advanced Management Firmware (FC 0050, FC 0051, FC 0069, or FC0071) of the corresponding blades.

Table 7-3 lists these features. The minimum quantity to order depends on the number of corresponding blades that are configured in the zBX Model 003.

Table 7-3 Feature codes for blade enablements and Unified Resource Manager suites

	Blade enablement	Manage (per connection)	Automate/Advanced Management (per connection)
z/OS only	N/A	FC 0019	FC 0020
IFL	N/A	N/C	FC 0054
DataPower XI50z	FC 0611	FC 0047	FC 0050
POWER7 Blade	FC 0612	FC 0048	FC 0051
IBM System x HX5 Blade	FC 0613	FC 0049	FC 0069/FC 0071

Important: If any attempt is made to install more blades that exceed the FC 0611, FC 0612, or FC 0613 count, those blades are not powered on by the system. The blades are also checked for minimum hardware requirements.

Table 7-4 shows the maximum quantities for Unified Resource Manager feature codes.

Table 7-4 Maximum quantities for Unified Resource Manager feature codes

Feature code	Maximum quantity	Feature description
FC 0047	28	Manage firmware DataPower
FC 0050	28	Automate/Advanced Management firmware DataPower
FC 0048	112	Manage firmware POWER blade
FC 0051	112	Automate/Advanced Management firmware POWER blade
FC 0049	56	Manage firmware System x blade
FC 0069	56	Advanced Management firmware System x blade (zBC12)
FC 0071	56	Advanced Management firmware System x blade (zEC12)
FC 0054	101	Automate/Advanced Management firmware IFL

FC 0047, FC 0048, FC 0049, FC 0050, FC 0051, FC 0069, FC 0071, and FC 0054 are priced features. To obtain ensemble member management and cables for zEC12 nodes, FC 0025 must also be ordered.

Feature codes are available to “detach” a zBX from an existing CPC and to “attach” a zBX to another CPC. FC 0030 indicates that the zBX will be detached. FC 0031 indicates that the detached zBX is going to be attached to another CPC.

Only zBX Model 003 (2458-003) is supported with zEC12. If you are upgrading a z196 with zBX Model 002 attachment to zEC12, the zBX Model 002 (2458-002) will be upgraded to a zBX Model 003, as well. Upgrades from zBX Model 002 to zBX Model 003 are disruptive.

A zBX Model 003 has the following improvements over Model 002:

- ▶ New AMM⁸ firmware in BladeCenter chassis with enhanced functions.
- ▶ Additional Ethernet connectivity to IEDN network for redundancy and increased bandwidth between ToR switches and BladeCenter switch modules.
- ▶ New firmware levels with improved functionality.
- ▶ New Unified Resource Manager support for ensembles with zEC12, zBC12, z196, and z114 with zBX Models 002 and 003.
- ▶ zBX Fibre Channel link/path testing and diagnostics.
- ▶ Layer 2 (L2) integration support between the customer data network and zBX.
- ▶ CPU management for System x blades enables Unified Resource Manager to dynamically manage processors for x86 blades in the zBX. It allows clients to monitor availability of workload resources to satisfy defined service policy goals by using KVM Control Groups (*cgroups*⁹). The cgroup exploitation by Unified Resource Manager platform performance manager (PPM) is based on assigning virtual servers to cgroups, allowing dynamic CPU sharing management based on policy goals.
- ▶ Ensemble availability manager (EAM) provides basic availability services for the ensemble by the Unified Resource Manager. It allows clients to monitor for zBX blade errors, including conditions affecting the availability of resources, and performs complete error analysis.

7.3.1 zBX management

One key feature of the zBX is its integration under the System z management umbrella. Therefore, the initial firmware installation, updates, and patches follow the already familiar pattern of System z. The same integration applies to the configuration and definitions.

Similar to channels and processors, the SE has a view for the zBX blades. This view shows icons for each of the zBX component's objects, including an overall status (power, operational, and so on).

The following functions and actions are managed and controlled from the zEC12 HMC/SE:

- ▶ View firmware information for the BladeCenter and blades.
- ▶ Retrieve firmware changes.
- ▶ Change firmware level.
- ▶ Back up and restore critical data: zBX configuration data is backed up as part of the System zEC12 SE backup. It is restored on the replacement of a blade.

For more information, see *IBM zEnterprise Unified Resource Manager*, SG24-7921.

⁸ Advanced Management Module (BladeCenter feature)

⁹ With cgroups, you can restrict a set of tasks to a set of resources, prevent denial-of-service situations in KVM environments, and monitor resource use.

7.3.2 zBX firmware

The firmware for the zBX is managed, controlled, and delivered in the same way as for the zEC12. It is packaged and tested with System z microcode, and changes are supplied and applied with MCL bundle releases.

The zBX firmware that is packaged with System z microcode has these benefits:

- ▶ Tested together with System z driver code and MCL bundle releases.
- ▶ Retrieved code uses the same integrated process of System z (IBM RETAIN® or media).
- ▶ No need to use separate tools or connect to websites to obtain code.
- ▶ Uses new upcoming System z firmware features, such as Digitally Signed Firmware.
- ▶ Infrastructure incorporates System z concurrency controls where possible.
- ▶ zBX firmware update is fully concurrent; blades are similar to Config Off/On controls.
- ▶ Audit trail of all code changes in security log.
- ▶ Automatic backout of changes to previous working level on code apply failures.
- ▶ Optimizer firmware.
- ▶ zBX requires the use of broadband RSF capability of HMC.

7.4 zBX connectivity

There are three types of LANs (each with redundant connections) that attach to the zBX:

- ▶ INMN
- ▶ IEDN
- ▶ Customer-managed data network

INMN

The INMN is fully isolated, and only established between the controlling zEC12 and the zBX. The zBX is managed by the HMC through the physically isolated INMN, which interconnects all resources of the zEC12 and zBX components.

IEDN

The IEDN connects the zBX to a maximum of eight zEC12s; you can have any combination of up to eight zEnterprise Systems (z196, z114, zEC12, and zBC12).

The IEDN provides private and secure 10 GbE high-speed data paths between all elements of a zEnterprise ensemble (up to eight zEC12, zBC12, z196, or z114 with optional zBXs).

Note: IEDN can expand to all members of an ensemble.

Each zEC12 must have a minimum of two connections to the zBX. The IEDN is also used to connect a zBX to a maximum of seven other zBXs. The IEDN is a VLAN-capable network that allows enhanced security by isolating data traffic between virtual servers.

Figure 7-6 on page 245 shows the connectivity that is required for the zBX environment. The zEC12 connects through two OSA-Express5S 1000BASE-T, OSA-Express4S 1000BASE-T, or OSA-Express3 1000BASE-T¹⁰ features (CHPID type OSM) to the Bulk Power Hubs (BPH) and to the INMN ToR switches. The OSA-EXPRESS5S 1000BASE-T, OSA-Express4S 10 GbE, or OSA-Express3 10 GbE¹⁰ features (CHPID type OSX) connect directly to the two IEDN ToR switches.

¹⁰ Carry forward only for zEC12 when you are upgrading from earlier generations.

Depending on the requirements, any OSA-Express5S, OSA-Express4S, or OSA-Express3 features (CHPID type OSD) can connect to the customer-managed data network.

Terminology: If not specifically stated otherwise, the term “OSA1000BASE-T” applies to the OSA-Express5S 1000BASE-T, OSA-Express4S 1000BASE-T, and OSA-Express3 1000BASE-T features throughout this chapter.

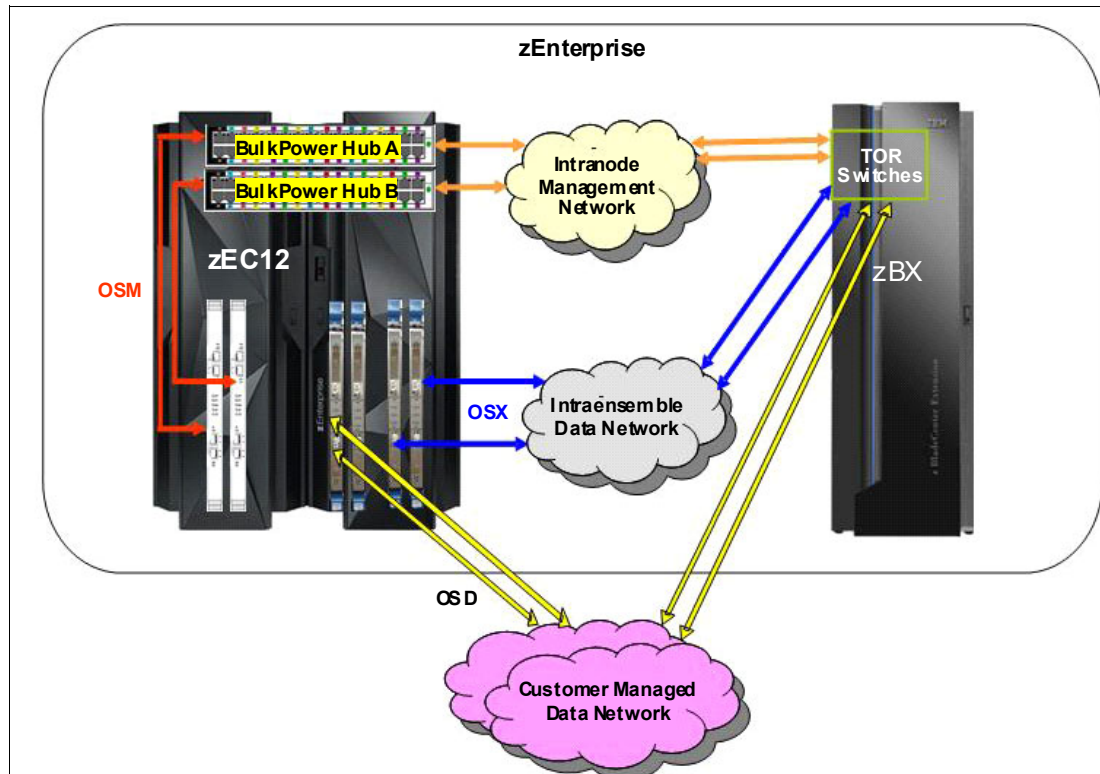


Figure 7-6 INMN, IEDN, and customer-managed local area networks

7.4.1 Intranode management network (INMN)

The scope of the INMN is within an ensemble *node*. A node consists of a z114, z196, zBC12, or a zEC12 and its optional zBX. INMNs of different nodes are not connected to each other. The INMN connects the SE of the z114, z196, zBC12, or zEC12 to the hypervisor, optimizer, and guest management agents within the node.

INMN communication

Communication across the INMN is exclusively for enabling the Unified Resource Manager of the HMC to run its various management disciplines for the node. These disciplines include virtual server, performance, network virtualization, energy, storage management, and so on. The zEC12 connection to the INMN is achieved through the definition of a CHPID type OSM, which can be defined over an OSA 1000BASE-T Ethernet feature. There is also a 1 GbE infrastructure within the zBX.

INMN configuration

Consider the following key points for an INMN:

- ▶ Each zEC12 must have two OSA 1000BASE-T ports that are connected to the Bulk Power Hub in the same zEC12:
 - The two ports provide a redundant configuration for failover purposes in case one link fails.
 - For availability, each connection must be from a different OSA 1000BASE-T feature within the same zEC12.

Figure 7-7 shows both the OSA 1000BASE-T features and required cable type.

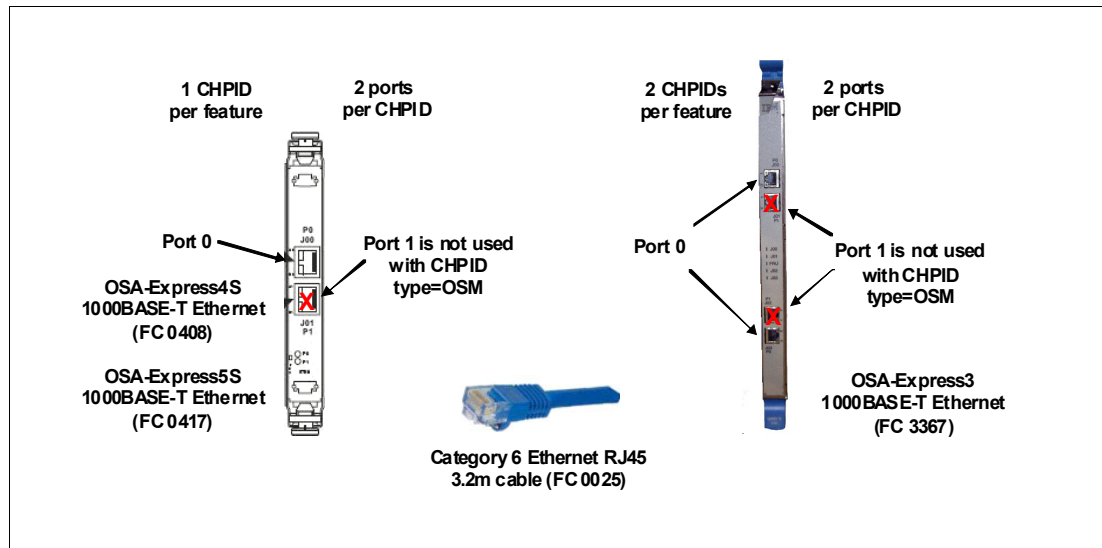


Figure 7-7 OSA-Express5S/4S 1000BASE-T and OSA-Express3 1000BASE-T features and cable type

- ▶ OSA 1000BASE-T ports (CHPID type OSM) can be defined in the input/output configuration data set (IOCDS) as SPANNED, SHARED, or DEDICATED:
 - DEDICATED restricts the OSA 1000BASE-T port to a single LPAR.
 - SHARED allows the OSA 1000BASE-T port to be used by all or selected LPARs in the same channel subsystem of a zEC12.
 - SPANNED allows the OSA 1000BASE-T port to be used by all or selected LPARs across multiple channel subsystems in the same zEC12.
 - SPANNED and SHARED ports can be restricted by the PARTITION keyword in the CHPID statement to allow only a subset of LPARs in the zEC12 to use the Open Systems Adapter (OSA) 1000BASE-T port.
 - SPANNED, SHARED, and DEDICATED link pairs can be defined within the maximum of 16 links that are supported by the zBX.
- ▶ The z/OS Communication server TCP/IP stack must be enabled for IPv6. The CHPID type OSM-related definitions are dynamically created. No IPv4 address is needed. A IPv6 link local address is dynamically applied.
- ▶ z/VM (before z/VM 6.3) virtual switch types provide INMN access:
 - The uplink can be a virtual machine network interface card (NIC).
 - Ensemble membership conveys the Universally Unique Identifier (UUID) and Media Access Control (MAC) prefix.

- ▶ Two 1000BASE-T ToR switches in the zBX (Rack B) are used for the INMN. No additional 1000BASE-T Ethernet switches are required. Figure 7-8 shows the 1000BASE-T ToR switches.



Figure 7-8 Two 1000BASE-T ToR switches (INMN)

The port assignments for both 1000BASE-T ToR switches are listed in Table 7-5.

Table 7-5 Port assignments for the 1000BASE-T ToR switches

Ports	Description
J00 - J03	Management for BladeCenters in zBX Rack-B
J04 - J07	Management for BladeCenters in zBX Rack-C
J08 - J11	Management for BladeCenters in zBX Rack-D
J12 - J15	Management for BladeCenters in zBX Rack-E
J16 - J43	Not used
J44 - J45	INMN switch B36P(Top) to INMN switch B35P(Bottom)
J46	INMN-A to IEDN-A port J41 / INMN-B to IEDN-B port J41
J47	INMN-A to zEC12 BPH-A port J06 / INMN-B to zEC12 BPH-B port J06

- ▶ 1000BASE-T supported cable:
 - 3.2-meter (10.49-ft) Category 6 Ethernet cables are shipped with the zEC12 ensemble management flag feature (FC 0025). These cables connect the OSA 1000BASE-T (OSM) ports to the Bulk Power Hubs (port 7).
 - 26-meter (85.3-ft) Category 5 Ethernet cables are shipped with the zBX. These cables are used to connect the zEC12 Bulk Power Hubs (port 6) and the zBX ToR switches (port J47).

7.4.2 Primary and alternate HMCs

The zEnterprise System HMC that has management responsibility for a particular zEnterprise ensemble is called a *primary HMC*. Only one primary HMC is active for a single ensemble. This HMC requires an alternate HMC to provide redundancy. The alternate HMC is not available for use until it becomes the primary HMC in a failover situation. To manage ensemble resources, the primary HMC for that ensemble must be used. A primary HMC can run all HMC functions. For more information about the HMC network configuration, see Chapter 12, “Hardware Management Console and Support Element” on page 431.

Figure 7-9 shows the primary and alternate HMC configuration that connects into the two BPHs in the zEC12.

Important: All ports on the zEC12 BPHs are reserved for specific connections. Any deviations or miscabling will affect the operation of the zEC12 system.

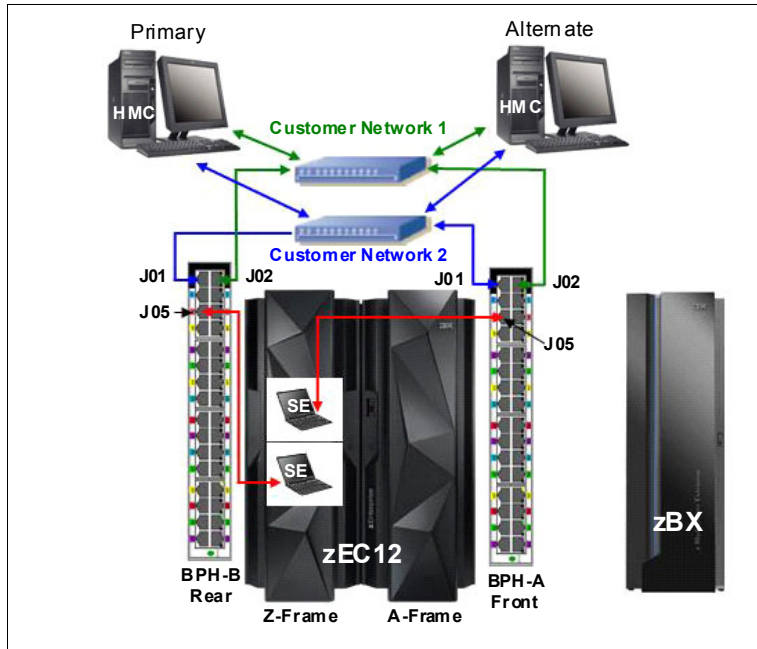


Figure 7-9 HMC configuration in an ensemble node

Table 7-6 shows the port assignments for both BPHs.

Table 7-6 Port assignments for the BPHs

BPH A		BPH B	
Port number	Connects to	Port number	Connects to
J01	HMC to SE Customer Network2 (VLAN 0.40)	J01	HMC to SE Customer Network2 (VLAN 0.40)
J02	HMC to SE Customer Network1 (VLAN 0.30)	J02	HMC to SE Customer Network1 (VLAN 0.30)
J03	BPH B J03	J03	BPH A J03
J04	BPH B J04	J04	BPH A J04
J05	SE A-Side (Top SE)	J05	SE B-Side (Bottom SE)
J06	zBX ToR Switch B36P, Port 47 (INMN-A)	J06	zBX ToR Switch B35P, Port 47 (INMN-B)
J07	OSA 1000BASE-T (CHPID type OSM)	J07	OSA 1000BASE-T (CHPID type OSM)
J08	Not used	J08	Not used
J09 - J32	Used for internal zEC12 components	J09 - J32	Used for internal zEC12 components

For more information, see Chapter 12, “Hardware Management Console and Support Element” on page 431.

7.4.3 Intraensemble data network (IEDN)

The IEDN is the main application data path that is provisioned and managed by the Unified Resource Manager of the controlling zEC12. Data communications for ensemble-defined workloads flow over the IEDN between nodes of an ensemble.

All of the physical and logical resources of the IEDN are configured and managed by the Unified Resource Manager. The IEDN extends from the zEC12 through the OSA-Express5S 10GbE, OSA-Express4S 10 GbE, or OSA-Express3 10 GbE ports when defined as CHPID type OSX. The minimum number of OSA 10 GbE features is two per zEC12. Similarly, a 10 GbE networking infrastructure within the zBX is used for IEDN access.

Terminology: If not specifically stated otherwise, the term “OSA10 GbE” applies to the OSA-Express5S 10 GbE, OSA-Express4S 10 GbE, and OSA-Express3 10GbE features throughout this chapter.

IEDN configuration

The IEDN connections can be configured in a number of ways. Consider the following key points for IEDN:

- ▶ Each zEC12 must have a minimum of two OSA 10 GbE ports that are connected to the zBX through the IEDN:
 - The two ports provide a redundant configuration for failover purposes in case one link fails.
 - For availability, each port connection must be from a different OSA 10 GbE feature within the same zEC12.
 - The zBX can have a maximum of 16 IEDN connections (eight pairs of OSA 10 GbE ports).
 - Four connections are between IEDN ToR switches and high-speed switch modules in each BladeCenter chassis (two pairs of 10 GbE ports).
 - For redundancy, two connections are between both high-speed switch modules in each BladeCenter.

Figure 7-10 shows the OSA 10 GbE feature (long reach or short reach) and the required fiber optic cable types.

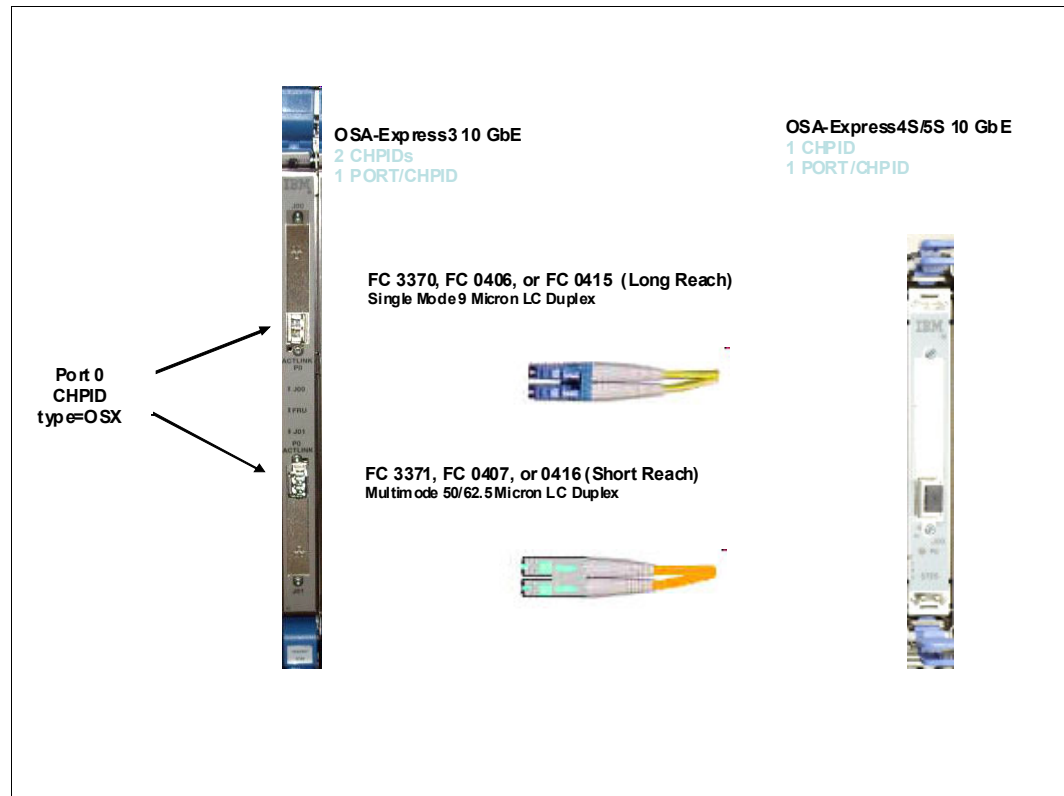


Figure 7-10 OSA-Express4S/5S 10 GbE and OSA-Express3 10 GbE features and cables

- ▶ OSA 10 GbE ports (CHPID type OSX) can be defined in the IOCDS as SPANNED, SHARED, or DEDICATED:
 - DEDICATED restricts the OSA 10 GbE port to a single LPAR.
 - SHARED allows the OSA 10 GbE port to be used by all or selected LPARs in the same channel subsystem of a zEC12.
 - SPANNED allows the OSA 10 GbE port to be used by all or selected LPARs across multiple channel subsystems (CSSs) in the same zEC12.
 - SHARED and SPANNED ports can be restricted by the PARTITION keyword in the CHPID statement to allow only a subset of LPARs on the zEC12 to use the OSA 10 GbE port.
 - SPANNED, SHARED, and DEDICATED link pairs can be defined within the maximum of 16 links that are supported by the zBX.
- ▶ z/OS Communication Server requires minimal configuration:
 - IPv4 or IPv6 addresses are used.
 - VLAN must be configured to match the HMC (Unified Resource Manager) configuration.
- ▶ z/VM (before 6.3) virtual switch types provide IEDN access:
 - The uplink can be a virtual machine NIC.
 - Ensemble membership conveys the ensemble UUID and MAC prefix.
- ▶ IEDN network definitions are completed from the primary HMC “Manage Virtual Network” task.

- ▶ Two 10 GbE ToR switches in the zBX (Rack B) are used for the IEDN. No additional Ethernet switches are required. Figure 7-11 shows the 10 GbE ToR switches.

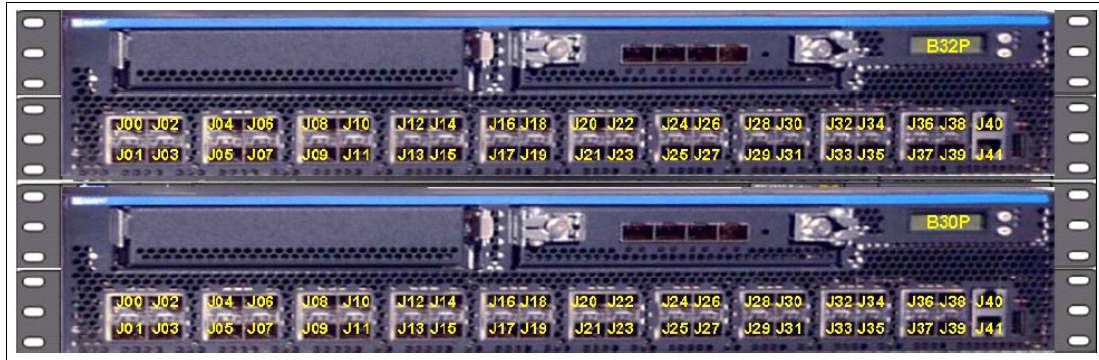


Figure 7-11 Two 10 GbE ToR switches

The IBM zEnterprise EC12 and BC12 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. It is displayed as a single Layer 2. Because HiperSockets and IEDN are both internal System z networks, the combination allows System z virtual servers to use the optimal path for communications.

The support of HiperSockets integration with the IEDN function is available on z/OS Communication Server V1R13 and up as well as on z/VM V6R2 and up.

Port assignments

The port assignments for both 10 GbE ToR switches are listed in Table 7-7.

Table 7-7 Port assignments for the 10 GbE ToR switches

Ports	Description
J00 - J07	SFP and reserved for zEC12 (OSX) IEDN connections
J08 - J23	DAC reserved for BladeCenter SM07/SM09 IEDN connections
J24 - J30	SFP reserved for zBX-to-zBX IEDN connections
J31 - J37	SFP reserved for client IEDN connections
J38 - J39	DAC for ToR switch-to-ToR switch IEDN communication
J40	RJ-45 (not used)
J41	RJ-45 IEDN Switch Management Port to INMN ToR switch port 46

- ▶ All IEDN connections must be point-to-point to the 10 GbE switch:
 - Through J31-J37 IEDN supports Layer 2 and Layer 3 connections to the customer network.
 - No additional switches or routers are needed.
 - This limits the distances between CPCs and the 10 GbE ToR switches in an ensemble.

- ▶ The 10 GbE ToR switches use small form-factor pluggable (SFP) optics for the external connections and DACs for connections in the following manner:

Ports J00 - J07 are reserved for the **zEC12** OSX IEDN connections. These ports use SFPs plugged according to the zBX order:

- FC 0632 LR SFP to FC 0415 OSA-Express5S 10 GbE LR
- FC 0633 SR SFP to FC 0416 OSA-Express5S 10 GbE SR
- FC 0632 LR SFP to FC 0406 OSA-Express4S 10 GbE LR
- FC 0633 SR SFP to FC 0407 OSA-Express4S 10 GbE SR
- FC 0632 LR SFP to FC 3370 OSA-Express3 10 GbE LR
- FC 0633 SR SFP to FC 3371 OSA-Express3 10 GbE SR

Ports J08 - J23 are reserved for IEDN to BladeCenter attachment. The cables that are used are Direct Attach Cables (DACs), and are included with the zBX. These are hard-wired 10 GbE SFP cables. The feature codes indicate the length of the cable:

- FC 0626: 1 meter (3.28 ft) for Rack B BladeCenters and IEDN to IEDN
- FC 0627: 5 meters (16.4 ft) for Rack C BladeCenter
- FC 0628: 7 meters (22.97 ft) for Racks D and E BladeCenters

- ▶ Ports J31 - J37 are reserved for the customer external network to IEDN connections. These ports use SFP modules plugged according to the zBX order. You must provide all IEDN cables except for zBX internal connections. The following 10 GbE fiber optic cable types are available, and their maximum distances are listed:

- Multimode fiber:
 - 50-micron fiber at 2000 MHz-km: 300 meters (984 ft)
 - 50-micron fiber at 500 MHz-km: 82 meters (269 ft)
 - 62.5-micron fiber at 200 MHz-km: 33 meters (108 ft)
- Single mode fiber:
 - 10 km (6.2 miles)

7.4.4 Network connectivity rules with zBX

Interconnecting a zBX must follow these network connectivity rules:

- ▶ Only one zBX is allowed per controlling zEC12.
- ▶ The zBX can be installed next to the controlling zEC12, within the limitation of the 26-meter (85-ft) cable.
- ▶ Customer-managed data networks are outside the ensemble. A customer-managed data network is connected with these components:
 - CHPID type OSD from zEC12
 - IEDN ToR switch ports J31 - J37 from zBX

7.4.5 Network security considerations with zBX

The private networks that are involved in connecting the zEC12 to the zBX are constructed with extreme security in mind:

- ▶ The INMN is entirely private and can be accessed only by the SE (standard HMC security applies). There are also additions to Unified Resource Manager “role-based” security. Therefore, not just any user can reach the Unified Resource Manager panels even if that user can perform other functions of the HMC. There are strict authorizations for users and programs to control who is allowed to take advantage of the INMN.

- ▶ The INMN network uses “link-local” IP addresses. “Link-local” addresses are not advertised and are accessible only within a single LAN segment. There is no routing in this network because it is a “flat network” with all virtual servers on the same IPv6 network. The Unified Resource Manager communicates with the virtual servers through the SE over the INMN. The virtual servers cannot communicate with each other directly through INMN. They can communicate only with the SE.
- ▶ Only authorized programs or agents can take advantage of the INMN. Currently, the Performance Agent can do so. However, there can be other platform management applications in the future that will need to be authorized to access the INMN.
- ▶ The IEDN is built on a flat network design (same IPv4 or IPv6 network). Each server that accesses the IEDN must be an authorized virtual server, and must belong to an authorized virtual LAN (VLAN) within the physical IEDN. VLAN enforcement is part of the hypervisor functions of the ensemble. The controls are in the OSA (CHPID type OSX), in the z/VM VSWITCH, and in the VSWITCH hypervisor function of the blades on the zBX.

The VLAN IDs and the virtual MACs that are assigned to the connections from the virtual servers are tightly controlled through the Unified Resource Manager. Therefore, there is no chance of either MAC or VLAN spoofing for any of the servers on the IEDN. If you decide to attach your network to the ToR switches of the zBX to communicate with the virtual servers on the zBX blades, access must be authorized in the ToR switches (MAC-based or VLAN-based).

Although the ToR switches enforce the VMACs and VLAN IDs, you must take the usual network security measures to ensure that the devices in the customer-managed data network are not subject to MAC or VLAN spoofing. The Unified Resource Manager functions cannot control the assignment of VLAN IDs and VMACs in those devices. Whenever you decide to interconnect the external network to the secured IEDN, the security of that external network must involve all the usual layers of the IBM Security Framework. These layers include physical security, platform security, application and process security, and data and information security.

- ▶ The INMN and the IEDN are both subject to network access controls as implemented in z/OS and z/VM. Therefore, not just any virtual server on the zEC12 can use these networks. INMN is not accessible at all from within the virtual servers.
- ▶ It is unnecessary to implement firewalls, IP filtering, or encryption for data flowing over the IEDN. However, if company policy or security requires these measures, they are supported. You can implement any of the available security technologies, such as Secure Sockets Layer (SSL)/Transport Layer Security (TLS) or IP filtering.
- ▶ The centralized and internal network design of both the INMN and the IEDN limit the vulnerability to security breaches. Both networks reduce the amount of network equipment and the number of administration tasks. They also reduce routing hops that are under the control of multiple individuals and subject to security threats. Both use only IBM equipment (switches and blades) that have been tested, and in certain cases, preinstalled.

In summary, many more technologies than in the past are integrated in a more robust, secure fashion into the client network. This configuration is achieved with the help of either the Unified Resource Manager, or more System Authorization Facility (SAF) controls specific to zEnterprise System and the ensemble:

- ▶ MAC filtering
- ▶ VLAN enforcement
- ▶ ACCESS control
- ▶ Role-based security

- ▶ The following standard security implementations are still available for use in the IEDN:
 - Authentication.
 - Authorization and access control that include multilevel security (MLS) and firewall IP filtering. Only stateless firewalls or IP filtering implementations can be installed in a virtual server in the ensemble.
 - Confidentiality.
 - Data integrity.
 - Non-repudiation.

7.4.6 zBX storage connectivity

The FC connections can be established between the zBX and a SAN environment. Client-supplied FC switches are required, and must support NPIV¹¹. Some FC switch vendors also require “interop” mode. Check the interoperability matrix for the latest details:

<http://www-03.ibm.com/systems/support/storage/ssic/interoperability.wss>

Remember: It is the client’s responsibility to supply the cables for IEDN, the customer-managed network, and the connection between the zBX and the SAN environment.

Each BladeCenter chassis in the zBX has two 20-port 8-Gbps FC switch modules. Each switch has 14 internal ports and six shortwave (SX) external ports. The internal ports are reserved for the blades in the chassis. The six external ports (J00, and J15 - J19) are used to connect to the SAN. Figure 7-12 on page 255 shows an image of the external ports.

¹¹ N-Port ID Virtualization

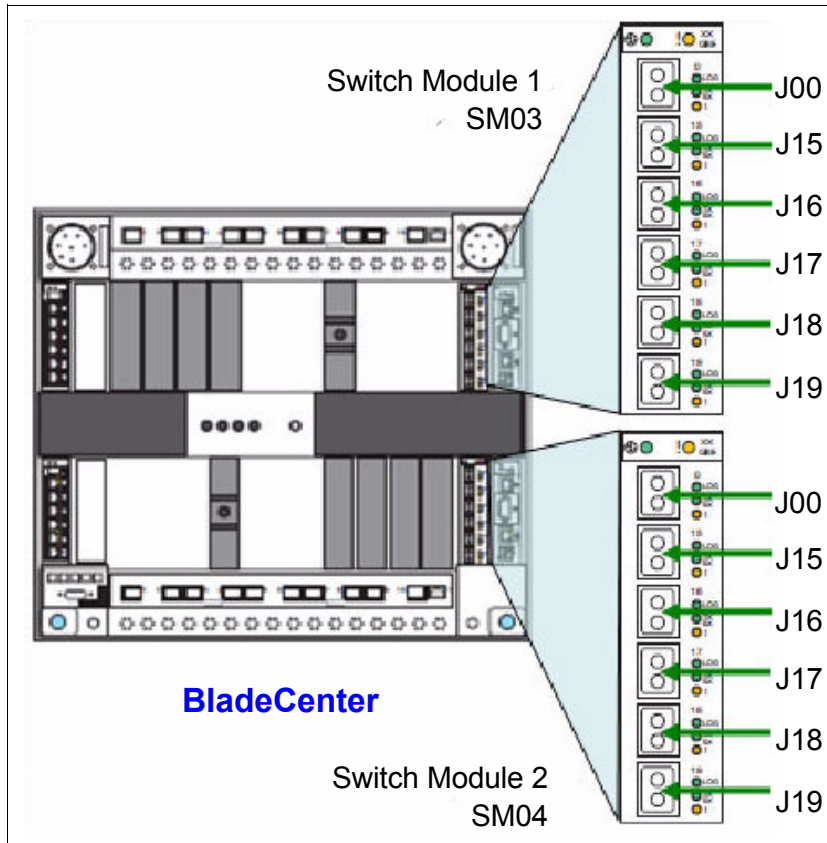


Figure 7-12 8-Gb FC switch external ports

You must provide multi-mode LC duplex cables to use for the FC switch connections to support speeds of 8 Gbps, 4 Gbps, or 2 Gbps (1 Gbps is not supported). The maximum distance depends on the speed and fiber type.

Cabling specifications are defined by the Fibre Channel - Physical Interface - 4 (FC-PI-4) standard.

Table 7-8 on page 256 identifies cabling types and link data rates that are supported in the zBX SAN environment, including their allowable maximum distances and link loss budget. The *link loss budget* is derived from the channel insertion loss budget that is defined by the FC-PI-4 standard (Revision 8.00).

Table 7-8 Fiber optic cabling for zBX FC switch: Maximum distances and link loss budget

FC-PI-4	2 Gbps		4 Gbps		8 Gbps	
	Distance in meters	Link loss budget (dB)	Distance in meters	Link loss budget (dB)	Distance in meters	Link loss budget (dB)
50 µm MM ^a (SX laser)	500	3.31	380	2.88	150	2.04
50 µm MM ^b (SX laser)	300	2.62	150	2.06	50	1.68
62.5 µm MM ^c (SX laser)	150	2.1	70	1.78	21	1.58

- a. OM3: 50/125 µm laser optimized multimode fiber with a minimum overfilled launch bandwidth of 1500 MHz-km at 850nm and an effective laser launch bandwidth of 2000 MHz-km at 850 nm in accordance with IEC 60793-2-10 Type A1a.2 fiber.
- b. OM2: 50/125 µm multimode fiber with a bandwidth of 500 MHz-km at 850 nm and 500 MHz-km at 1300 nm in accordance with IEC 60793-2-10 Type A1a.1 fiber.
- c. OM1: 62.5/125 µm multimode fiber with a minimum overfilled launch bandwidth of 200 MHz-km at 850 nm and 500 MHz-km at 1300 nm in accordance with IEC 60793-2-10 Type A1b fiber.

Cabling: IBM does not support a mix of 50 µm and 62.5-µm fiber optic cabling in a physical link.

IBM blade storage connectivity

IBM blades use six ports in both FC switch modules (SM03 and SM04) of the BladeCenter chassis, and must connect through an FC switch to FC disk storage. Figure 7-13 illustrates the FC connectivity with two FC switches for redundancy and high availability.

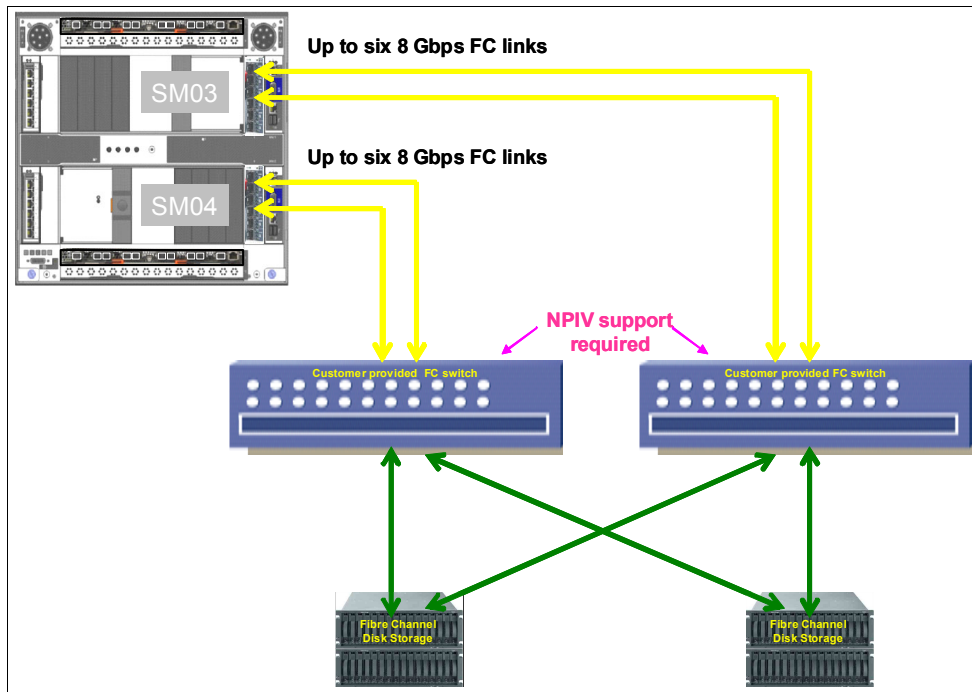


Figure 7-13 BladeCenter chassis storage connectivity

Up to six external ports of each BladeCenter switch module can be used to connect to the SAN. All fiber links of a switch module must be attached to the same SAN switch as shown in Figure 7-13 on page 256. SAN switches must support NPIV to allow virtualization.

The client has to provide all cables, FC disk storage, and FC switches, and must also configure and cable the FC switches that connect to the zBX.

Supported FC disk storage

Supported FC disk types and vendors with IBM blades are listed on the IBM System Storage® Interoperation Center (SSIC) website:

http://www-03.ibm.com/systems/support/storage/config/ssic/displayessearchwithoutjs.wss?start_over=yes

7.5 zBX connectivity examples

This section illustrates various ensemble configuration examples containing a zBX and the necessary connectivity. For simplicity, redundant connections are not shown in the configuration examples.

Subsequent configuration diagrams build on the previous configuration, and only additional connections are noted.

7.5.1 Single-node ensemble with a zBX

Figure 7-14 shows a single-node ensemble with a zBX. The necessary components include the controlling zEC12 (CPC1), and the attached zBX, FC switches, and FC disk storage.

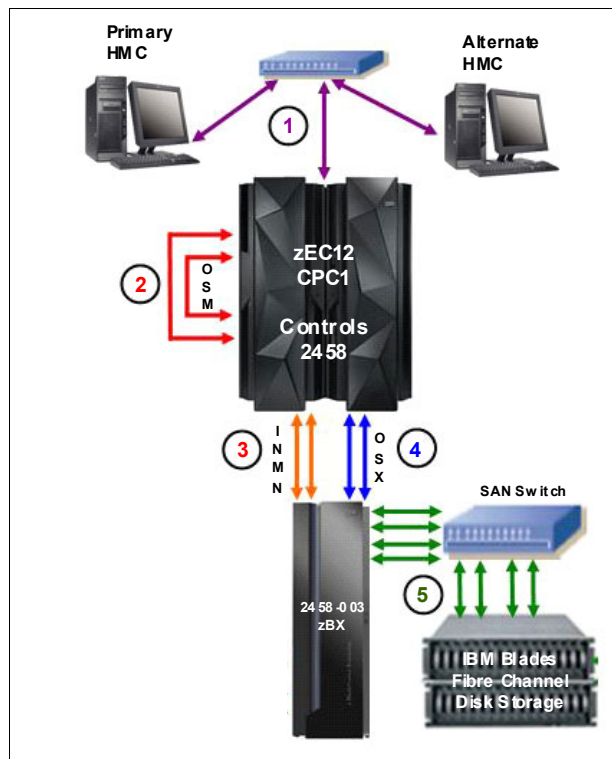


Figure 7-14 Single-node ensemble with zBX

The diagram shows the following components:

1. Client-provided management network:

- IBM supplies a 15-meter (49-ft) Ethernet RJ-45 cable with the 1000BASE-T (1 GbE) switch (FC 0070).

Important: The Ethernet switch (FC 0070) can only be carried forward on zEC12. If FC0070 is not carried forward, the client has to provide their own Ethernet switch.

- The 1000BASE-T switch (FC 0070 or client-provided) connects to the reserved client network ports of the Bulk Power Hubs in zEC12 - Z29BPS11 (on A side) and Z29BPS31 (on B side) - port J02. A second switch, if present, connects the reserved client network to the BPHs in zEC12 Z29BPS11 and Z29BPS31 on port J01.

2. Intranode management network:

- Two CHPIDs from two different OSA1000BASE-T features are configured as CHPID type OSM.
- IBM supplies two 3.2-meter (10.49-ft) Ethernet Category 6 cables from the OSM CHPIDs (ports) to both Z29BPS11 and Z29BPS31 on port J07. This is a zEC12 internal connection that is supplied with FC 0025 - Intranode management network extension.
- IBM supplies two 26-meter (85-ft) Category 5 Ethernet cables (chrome gray plenum rated cables) from zBX Rack B INMN-A/B switches port J47 to Z29BPS11 and Z29BPS31 on port J06.

3. Intraensemble data network:

- Two ports from two different OSA10 GbE ((SR) Short Reach or (LR) Long Reach) features are configured as CHPID type OSX.
- The client supplies the fiber optic cables (single mode or multimode).

4. 8-Gbps Fibre Channel switch:

- You supply all Fibre Channel cables (multimode) from the zBX to the attached FC switch.
- You are responsible for the configuration and management of the FC switch.

7.5.2 Dual-node ensemble with a single zBX

A second zEC12 (CPC2) is introduced in Figure 7-15, showing the additional hardware. Up to eight more nodes (CPCs) can be added in the same fashion.

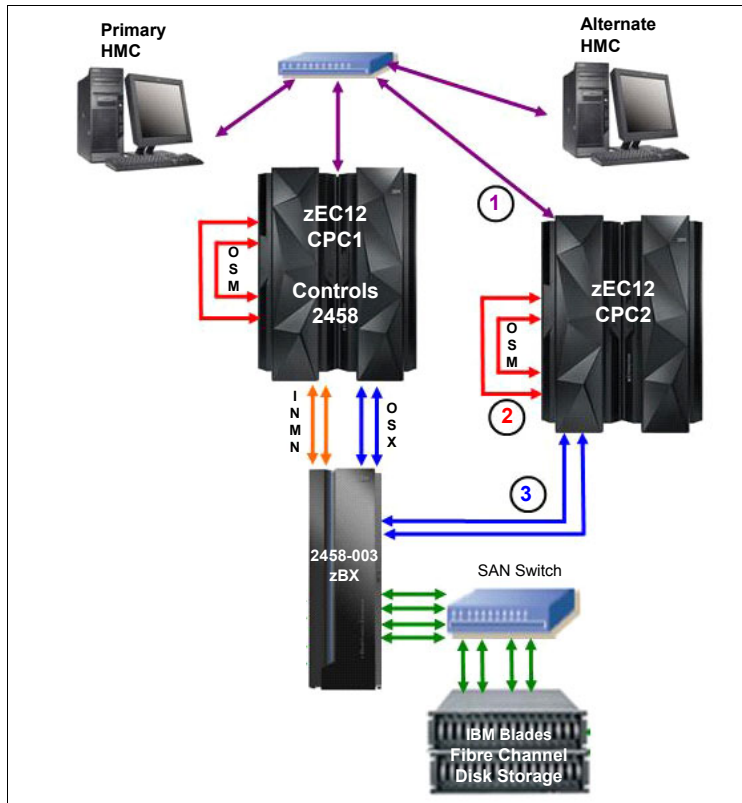


Figure 7-15 Dual-node ensemble with a single zBX

The diagram shows the following components:

1. Client-provided management network:
 - You supply an Ethernet RJ-45 cable.
 - The 1000BASE-T switch (FC 0070 or client provided) connects to the reserved client network ports of Z29BPS11 and Z29BPS31 - J02. A second switch connects to Z29BPS11 and Z29BPS31 on port J01.
2. Intranode management network:
 - Two ports from two different OSA1000BASE-T features are configured as CHPID type OSM.
 - IBM supplies two 3.2-meter (10.49-ft) Ethernet Category 6 cables from the OSM CHPIDs (ports) to both Z29BPS11 and Z29BPS31 on port J07. This is a zEC12 internal connection that is supplied with feature code 0025.
3. Intraensemble data network:
 - Two ports from two different OSA10 GbE ((SR) Short Reach or (LR) Long Reach) features are configured as CHPID type OSX.
 - The client must supply the fiber optic cables (single mode or multi-mode).

7.5.3 Dual-node ensemble with two zBXs

Figure 7-16 introduces a second zBX added to the original configuration. The two zBXs are interconnected through fiber optic cables to SFPs in the IEDN switches for isolated communication (SR or LR) over the IEDN network.

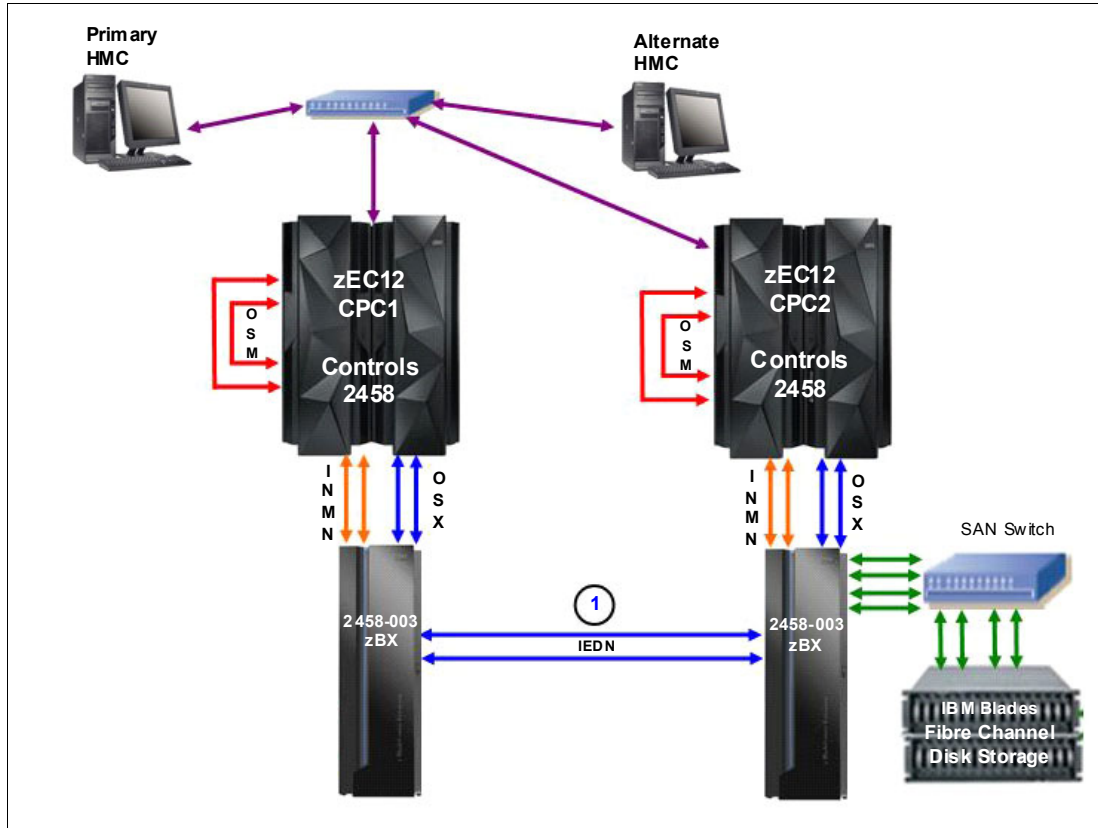


Figure 7-16 Dual-node ensemble

The diagram shows the following components:

1. Intraensemble data network:

- Two 10 GbE ports in the ToRs are used to connect the two zBXs (10 GbE ToR switch to 10 GbE ToR switch). This connection is represented by the number “1”.

Up to eight CPCs can be connected to a zBX by using the IEDN. More CPCs are added and connected to the zBX through the OSA 10 GbE ((SR) Short Reach or (LR) Long Reach) features configured as CHPID type OSX.

7.6 References

For more information about installation details, see *IBM zEnterprise BladeCenter Extension Model 003 Installation Manual for Physical Planning*, GC27-2611, and *IBM zEnterprise BladeCenter Extension Model 003 Installation Manual*, GC27-2610.

For more information about the BladeCenter components, see *IBM BladeCenter Products and Technology*, SG24-7523.

For more information about DataPower XI50z blades, visit the website:

<http://www-01.ibm.com/software/integration/datapower/xi50z>

Additional documentation is available on the IBM Resource Link:

<http://www.ibm.com/servers/resourceLink>



Software support

This chapter lists the minimum operating system requirements and support considerations for the **IBM zEnterprise EC12** (zEC12) and its features. It addresses z/OS, z/VM, z/VSE, z/TPF, and Linux on System z. Because this information is subject to change, see the Preventive Service Planning (PSP) bucket for 2827DEVICE for the most current information. Also included is generic software support for zEnterprise BladeCenter Extension (zBX) Model 003.

Support of zEC12 functions depends on the operating system, its version, and release.

This chapter includes the following sections:

- ▶ Operating systems summary
- ▶ Support by operating system
- ▶ Support by function
- ▶ Cryptographic support
- ▶ z/OS migration considerations
- ▶ Coupling facility and CFCC considerations
- ▶ MIDAW facility
- ▶ IOCP
- ▶ ICKDSF
- ▶ zEnterprise BladeCenter Extension (zBX) Model 003 software support
- ▶ Software licensing considerations
- ▶ References

8.1 Operating systems summary

Table 8-1 lists the minimum operating system levels that are required on the zEC12. For similar information about the IBM zEnterprise BladeCenter Extension (zBX) Model 003, see 8.11, “zEnterprise BladeCenter Extension (zBX) Model 003 software support” on page 331.

Note that operating system levels that are no longer in service are not covered in this publication. These older levels might provide support for some features.

Table 8-1 zEC12 minimum operating systems requirements

Operating systems	ESA/390 (31-bit mode)	z/Architecture (64-bit mode)	Notes
z/OS V1R11 ^a	No	Yes	Service is required. See the following box, which is titled “Features”.
z/VM V5R4 ^b	No	Yes ^c	
z/VSE V4R3	No	Yes	
z/TPF V1R1	Yes	Yes	
Linux on System z	No ^d	See Table 8-2 on page 266.	

- a. Regular service support for z/OS V1R11 ended September 2012. However, by ordering the IBM Lifecycle Extension for z/OS V1R11 product, fee-based corrective service can be obtained up to September 2014.
- b. z/VM V5R4, V6R2, and V6R3 provide compatibility support only. z/VM V6R2 and V6R3 require an architecture level set (ALS) at z10 or higher. Support for z/VM V6R1 ended April 2013.
- c. 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: Exploitation of certain features depends on the operating system. In all cases, PTFs might be required with the operating system level indicated. Check the z/OS, z/VM, z/VSE, and z/TPF subsets of the 2827DEVICE 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 System z distributions, consult the distributor’s support information.

8.2 Support by operating system

IBM zEnterprise EC12 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 System z servers and carried forward or enhanced in zEC12. Features and functions available on previous servers but no longer supported by zEC12 have been removed.

For a list of supported functions and the z/OS and z/VM minimum required support levels, see Table 8-3 on page 267. For z/VSE, z/TPF, and Linux on System z, see Table 8-4 on page 273. The tabular format is intended to help determine, by a quick scan, which functions are supported and the minimum operating system level required.

8.2.1 z/OS

z/OS Version 1 Release 12 is the earliest in-service release that supports zEC12. After September 2014, a fee-based Extended Service for defect support (for up to three years) can be obtained for z/OS V1R12. Although service support for z/OS Version 1 Release 11 ended in September of 2012, a fee-based extension for defect support (for up to two years) can be obtained by ordering the IBM Lifecycle Extension for z/OS V1R11. Also, z/OS.e is not supported on zEC12, and z/OS.e Version 1 Release 8 was the last release of z/OS.e.

zEC12 capabilities differ depending on z/OS release. Toleration support is provided on z/OS V1R11. Exploitation support is only provided on z/OS V1R12 and higher. For a list of supported functions and their minimum required support levels, see Table 8-3 on page 267.

8.2.2 z/VM

At general availability, z/VM V5R4, V6R2, and V6R3 provide compatibility support with limited exploitation of new zEC12 functions.

For a list of supported functions and their minimum required support levels, see Table 8-3 on page 267.

Capacity: 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 want to adjust the number to accommodate the processor unit (PU) capacity of the zEC12.

z/VM V6R3 and Unified Resource Manager: 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. Accordingly, zManager will not provide systems management support for z/VM 6.3 and later releases. However, zManager will continue to play a distinct and strategic role in the management of virtualized environments created by integrated firmware hypervisors (IBM Processor Resource/Systems Manager (PR/SM), PowerVM, and x hypervisor based on a kernel-based virtual machine (KVM)) of the zEnterprise.

8.2.3 z/VSE

Support is provided by z/VSE V4R3 and later. Note the following considerations:

- ▶ z/VSE runs in z/Architecture mode only.
- ▶ z/VSE uses 64-bit real memory addressing.
- ▶ Support for 64-bit virtual addressing is provided by z/VSE V5R1.
- ▶ z/VSE V5R1 requires an architectural level set that is specific to the IBM System z9.

For a list of supported functions and their minimum required support levels, see Table 8-4 on page 273.

8.2.4 z/TPF

For a list of supported functions and their minimum required support levels, see Table 8-4 on page 273.

8.2.5 Linux on System z

Linux on System z 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 System z distribution provides support for running 31-bit applications. None of the current versions of Linux on System z distributions (SUSE Linux Enterprise Server (SLES) 10 and SLES 11, and Red Hat (RHEL) 5 and RHEL 6)¹ require zEC12 toleration support. Table 8-2 shows the service levels of SUSE and Red Hat releases supported at the time of writing.

Table 8-2 Current Linux on System z distributions

Linux on System z distribution	z/Architecture (64-bit mode)
SUSE SLES 10 SP4	Yes
SUSE SLES 11 SP2	Yes
Red Hat RHEL 5.8	Yes
Red Hat RHEL 6.3	Yes

For the latest information about supported Linux distributions on System z, 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 exploitation of selected zEC12 functions in future Linux on System z distribution releases.

Consider the following guidelines:

- ▶ Use SUSE SLES 11 or Red Hat RHEL 6 in any new projects for the zEC12.
- ▶ Update any Linux distributions to their latest service level before the migration to zEC12.
- ▶ Adjust the capacity of any z/VM and Linux on System z 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 zEC12.

8.2.6 zEC12 function support summary

The following tables summarize the zEC12 functions and their minimum required operating system support levels:

- ▶ Table 8-3 on page 267 is for z/OS and z/VM.
- ▶ Table 8-4 on page 273 is for z/VSE, z/TPF, and Linux on System z.

Information about Linux on System z 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.

¹ SLES is SUSE Linux Enterprise Server
RHEL is Red Hat Enterprise Linux

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 2827DEVICE.

Table 8-3 zEC12 function minimum support requirements summary (part 1 of 2)

Function	z/OS V2 R1	z/OS V1 R13	z/OS V1 R12	z/OS V1 R11	z/OS V1 R10	z/VM V6 R3	z/VM V6 R2	z/VM V5 R4
zEC12	Y	Y ^q	Y ^q	Y ^q	Y ^q	Y ^q	Y ^q	Y ^q
Support of Unified Resource Manager	Y	Y	Y ^q	Y ^q	Y ^q	N	Y	N
Greater than 64 PUs single system image	Y	Y	Y	Y	Y ^q	N	N	N
Greater than 54 PUs single system image	Y	Y	Y	Y	Y	N ^a	N ^a	N ^a
Support of IBM zAware	Y	Y ^q	N	N	N	-	-	-
System z Integrated Information Processors (zIIPs)	Y	Y	Y	Y	Y	Y ^b	Y ^b	Y ^b
System z Application Assist Processors (zAAPs)	Y	Y	Y	Y	Y	Y ^b	Y ^b	Y ^b
zAAP on zIIP	Y	Y	Y	Y	Y ^q	Y ^c	Y ^c	Y ^c
Java Exploitation of Transactional Execution	Y	Y ^q	N	N	N	N	N	N
Large memory (> 128 GB)	Y	Y	Y	Y	Y	Y ^d	Y ^e	Y ^e
Large page support	Y	Y ^{f g}	Y	Y	Y	N ^h	N ^h	N ^h
Out-of-order execution	Y	Y	Y	Y	Y	Y	Y	Y
Guest support for execute-extensions facility	-	-	-	-	-	Y	Y	Y
Hardware decimal floating point	Y ⁱ	Y ⁱ	Y ⁱ	Y ⁱ	Y ⁱ	Y ^b	Y ^b	Y ^b
Zero address detection	Y	Y	Y	N	N	N	N	N
60 Logical partitions (LPARs)	Y	Y	Y	Y	Y	Y	Y	Y
Logical partition (LPAR) group capacity limit	Y	Y	Y	Y	Y	-	-	-
LPAR physical capacity limit	Y ^q	Y ^q	Y ^q	N	N	Y	N	N
CPU measurement facility	Y	Y ^q	Y ^q	Y ^q	Y ^q	Y ^b	Y ^{bq}	Y ^{bq}
Separate LPAR management of PUs	Y	Y	Y	Y	Y	Y	Y	Y
Dynamic add and delete LPAR name	Y	Y	Y	Y	Y	Y	Y	Y
Capacity provisioning	Y	Y	Y	Y	Y	N ^h	N ^h	N ^h
Enhanced flexibility for Capacity on Demand (CoD)	Y	Y	Y ⁱ	Y ⁱ	Y ⁱ	Y ⁱ	Y ⁱ	Y ⁱ
HiperDispatch	Y	Y	Y	Y	Y	Y	N ^h	N ^h
63.75 K subchannels	Y	Y	Y	Y	Y	Y	Y	Y
Four logical channel subsystems (LCSSs)	Y	Y	Y	Y	Y	Y	Y	Y
Dynamic I/O support for multiple LCSSs	Y	Y	Y	Y	Y	Y	Y	Y
Third subchannel set	Y	Y	Y ^q	Y ^q	Y ^q	N ^h	N ^h	N ^h
Multiple subchannel sets	Y	Y	Y	Y	Y	N ^h	N ^h	N ^h
IPL from alternate subchannel set	Y	Y ^q	Y ^q	Y ^q	N	N ^h	N ^h	N ^h

Function	z/OS V2 R1	z/OS V1 R13	z/OS V1 R12	z/OS V1 R11	z/OS V1 R10	z/VM V6 R3	z/VM V6 R2	z/VM V5 R4
MIDAW facility	Y	Y	Y	Y	Y	Y ^b	Y ^b	Y ^b
Cryptography								
CP Assist for Cryptographic Function (CPACF)		Y	Y	Y	Y	Y ^b	Y ^b	Y ^b
CPACF AES-128, AES-192, and AES-256		Y	Y	Y	Y	Y ^b	Y ^b	Y ^b
CPACF SHA-1, SHA-224, SHA-256, SHA-384, and SHA-512		Y	Y	Y	Y	Y ^b	Y ^b	Y ^b
CPACF protected key		Y	Y	Y ^j	Y ^j	Y ^b	Y ^b	Y ^{bq}
Crypto Express4S		Y ^{kl}	Y ^{kl}	Y ^k	Y ^k	Y ^b	Y ^{bq}	Y ^{bq}
Secure IBM Enterprise PKCS #11 (EP11) coprocessor mode		Y ^{kl}	Y ^{kl}	Y ^k	Y ^k	Y ^b	Y ^{bq}	Y ^{bq}
Crypto Express3		Y	Y	Y ^j	Y ^j	Y ^b	Y ^b	Y ^{bq}
Elliptic Curve Cryptography (ECC)		Y	Y ^j	Y ^j	Y ^j	Y ^b	Y ^b	Y ^{bq}
HiperSockets								
32 HiperSockets	Y	Y	Y ^q	Y ^q	Y ^q	Y	Y	Y ^q
HiperSockets Completion Queue	Y	Y ^q	N	N	N	Y	Y ^q	N
HiperSockets integration with IEDN	Y	Y ^q	N	N	N	N	Y ^q	N
HiperSockets Virtual Switch Bridge	-	-	-	-	-	Y	Y ^q	N
HiperSockets Network Traffic Analyzer	N	N	N	N	N	Y ^b	Y ^b	Y ^{bq}
HiperSockets Multiple Write Facility	Y	Y	Y	Y	Y	N ^h	N ^h	N ^h
HiperSockets support of IPV6	Y	Y	Y	Y	Y	Y	Y	Y
HiperSockets Layer 2 support	Y	Y	Y	N	N	Y ^b	Y ^b	Y ^b
HiperSockets	Y	Y	Y	Y	Y	Y	Y	Y
Flash Express Storage								
Flash Express	Y	Y ^{jm}	N	N	N	N	N	N
zEnterprise Data Compression (zEDC)								
zEDC Express	Y ^q	Y ^{qn}	Y ^{qn}	N	N	N ^o	N	N
Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE)								
10GbE RoCE Express	Y ^q	Y ^{qp}	Y ^{qp}	N	N	N ^o	N	N
FICON (Fibre Connection) and FCP (Fibre Channel Protocol)								
z/OS Discovery and Auto Configuration (zDAC)	Y	Y	Y ^q	N	N	N	N	N
24k subchannel support for FICON Express8S, FICON Express8, and the FICON Express4 CHPID type FC	Y ^q	Y ^q	Y ^q	Y ^q	Y ^q	Y	Y	Y

Function	z/OS V2 R1	z/OS V1 R13	z/OS V1 R12	z/OS V1 R11	z/OS V1 R10	z/VM V6 R3	z/VM V6 R2	z/VM V5 R4
FICON Express8S support of High Performance FICON for System z (zHPF) enhanced multitrack CHPID type FC	Y	Y	Y	Y ^q	Y ^q	Y ^b	Y ^q	N
FICON Express8 support of zHPF enhanced multitrack CHPID type FC	Y	Y	Y	Y ^q	Y ^q	Y ^b	Y ^b	N ^h
High Performance FICON for System z (zHPF)	Y	Y	Y	Y	Y ^q	Y ^b	Y ^b	N ^h
FCP increased performance for small block sizes	N	N	N	N	N	Y	Y	Y
Request node identification data	Y	Y	Y	Y	Y	N	N	N
FICON link for incident reporting	Y	Y	Y	Y	Y	N	N	N
GRS FICON CTC toleration	Y	Y ^q	Y ^q	Y ^q	Y ^q	N	N	N
N-Port ID Virtualization for FICON (NPIV) CHPID type FCP	N	N	N	N	N	Y	Y	Y
FCP point-to-point attachments	N	N	N	N	N	Y	Y	Y
FICON SAN platform and name server registration	Y	Y	Y	Y	Y	Y	Y	Y
FCP SAN management	N	N	N	N	N	N	N	N
SCSI IPL for FCP	N	N	N	N	N	Y	Y	Y
Cascaded FICON Directors CHPID type FC	Y	Y	Y	Y	Y	Y	Y	Y
Cascaded FICON Directors CHPID type FCP	N	N	N	N	N	Y	Y	Y
FICON Express8S support of hardware data router CHPID type FCP	N	N	N	N	N	Y ^b	N	N
FICON Express8S and FICON Express8 support of T10-DIF CHPID type FCP	N	N	N	N	N	Y ^b	Y ^b	Y ^{bq}
FICON Express8S, FICON Express8, FICON Express4 10KM LX, and FICON Express4 SX support of SCSI disks CHPID type FCP	N	N	N	N	N	Y	Y	Y ^q
FICON Express8S CHPID type FC	Y	Y	Y	Y ^q	Y ^q	Y	Y	Y
FICON Express8 CHPID type FC	Y	Y	Y	Y ^r	Y ^r	Y ^r	Y ^r	Y ^r
FICON Express4 10KM LX and SX ^s CHPID type FC	Y	Y	Y	Y	Y	Y	Y	Y
OSA (Open Systems Adapter)								
VLAN management	Y	Y	Y	Y	Y	Y	Y	Y
VLAN (IEE 802.1q) support	Y	Y	Y	Y	Y	Y	Y	Y
Queued direct I/O (QDIO) data connection isolation for z/VM virtualized environments	-	-	-	-	-	Y	Y	Y ^q
OSA Layer 3 Virtual MAC	Y	Y	Y	Y	Y	Y ^b	Y ^b	Y ^b

Function	z/OS V2 R1	z/OS V1 R13	z/OS V1 R12	z/OS V1 R11	z/OS V1 R10	z/VM V6 R3	z/VM V6 R2	z/VM V5 R4
OSA Dynamic LAN idle	Y	Y	Y	Y	Y	Y ^b	Y ^b	Y ^b
OSA/SF enhancements for IP, Media Access Control (MAC) addressing (CHPID type OSD)	Y	Y	Y	Y	Y	Y	Y	Y
QDIO diagnostic synchronization	Y	Y	Y	Y	Y	Y ^b	Y ^b	Y ^b
Network Traffic Analyzer	Y	Y	Y	Y	Y	Y ^b	Y ^b	Y ^b
Large send for IPv6 packet	Y	Y ^q	N	N	N	Y ^b	Y ^b	Y ^b
Broadcast for IPv4 packets	Y	Y	Y	Y	Y	Y	Y	Y
Checksum offload for IPv4 packets	Y	Y	Y	Y	Y	Y ^t	Y ^u	Y ^u
OSA-Express4S and OSA-Express3 inbound workload queuing for Enterprise Extender	Y	Y	N	N	N	Y ^b	Y ^{b q}	Y ^{b q}
OSA-Express5S 10 Gigabit Ethernet Long Reach (LR) and Short Reach (SR) CHPID type OSD	Y	Y	Y	Y ^q	Y ^q	Y	Y ^q	Y
OSA-Express5S 10 Gigabit Ethernet LR and SR CHPID type OSX	Y	Y	Y	Y ^q	Y ^q	N ^v	Y ^q	N ^v
OSA-Express5S Gigabit Ethernet Long Wave (LX) and Short Wave (SX) CHPID type OSD (two ports per CHPID)	Y	Y	Y	Y ^q	Y ^q	Y	Y ^q	Y ^q
OSA-Express5S Gigabit Ethernet LX and SX CHPID type OSD (one port per CHPID)	Y	Y	Y	Y ^q	Y ^q	Y	Y ^q	Y
OSA-Express5S 1000BASE-T Ethernet CHPID type OSC	Y	Y	Y	Y ^q	Y ^q	Y	Y	Y
OSA-Express5S 1000BASE-T Ethernet CHPID type OSD (two ports per CHPID)	Y	Y	Y	Y ^q	Y ^q	Y	Y ^q	Y ^q
OSA-Express5S 1000BASE-T Ethernet CHPID type OSD (one port per CHPID)	Y	Y	Y	Y ^q	Y ^q	Y	Y ^q	Y
OSA-Express5S 1000BASE-T Ethernet CHPID type OSE	Y	Y	Y	Y ^q	Y ^q	Y	Y	Y
OSA-Express5S 1000BASE-T Ethernet CHPID type OSM	Y	Y	Y	Y ^q	Y ^q	N ^v	Y	N ^v
OSA-Express5S 1000BASE-T Ethernet CHPID type OSN	Y	Y	Y	Y ^q	Y ^q	Y	Y	Y
OSA-Express4S 10-Gigabit Ethernet LR and SR CHPID type OSD	Y	Y	Y	Y ^q	Y ^q	Y	Y	Y
OSA-Express4S 10-Gigabit Ethernet LR and SR CHPID type OSX	Y	Y	Y ^q	Y ^q	Y ^q	N ^v	Y	N ^v
OSA-Express4S Gigabit Ethernet LX and SX CHPID type OSD (two ports per CHPID)	Y	Y	Y	Y ^q	Y ^q	Y	Y	Y ^q
OSA-Express4S Gigabit Ethernet LX and SX CHPID type OSD (one port per CHPID)	Y	Y	Y	Y ^q	Y ^q	Y	Y	Y

Function	z/OS V2 R1	z/OS V1 R13	z/OS V1 R12	z/OS V1 R11	z/OS V1 R10	z/VM V6 R3	z/VM V6 R2	z/VM V5 R4
OSA-Express4S 1000BASE-T CHPID type OSC (one or two ports per CHPID)	Y	Y	Y	Y ^q	Y ^q	Y	Y	Y
OSA-Express4S 1000BASE-T CHPID type OSD (two ports per CHPID)	Y	Y	Y	Y ^q	Y ^q	Y	Y	Y ^q
OSA-Express4S 1000BASE-T CHPID type OSD (one port per CHPID)	Y	Y	Y	Y ^q	Y ^q	Y	Y	Y
OSA-Express4S 1000BASE-T CHPID type OSE (one or two ports per CHPID)	Y	Y	Y	Y ^q	Y ^q	Y	Y	Y
OSA-Express4S 1000BASE-T CHPID type OSM ^x (one port per CHPID)	Y	Y	Y ^q	Y ^q	Y ^q	N ^v	Y	N ^v
OSA-Express4S 1000BASE-T CHPID type OSN ^w	Y	Y	Y	Y ^q	Y ^q	Y	Y	Y
OSA-Express3 10-Gigabit Ethernet LR and SR CHPID type OSD	Y	Y	Y	Y	Y	Y	Y	Y
OSA-Express3 10-Gigabit Ethernet LR and SR CHPID type OSX	Y	Y	Y ^q	Y ^q	Y ^q	N ^v	Y	N ^v
OSA-Express3 Gigabit Ethernet LX and SX CHPID types OSD, OSN ^w (two ports per CHPID)	Y	Y	Y	Y	Y	Y	Y	Y ^q
OSA-Express3 Gigabit Ethernet LX and SX CHPID types OSD, OSN ^w (one port per CHPID)	Y	Y	Y	Y	Y	Y	Y	Y
OSA-Express3 1000BASE-T CHPID type OSC (two ports per CHPID)	Y	Y	Y	Y	Y	Y	Y	Y
OSA-Express3 1000BASE-T CHPID type OSD (two ports per CHPID)	Y	Y	Y	Y	Y	Y	Y	Y ^q
OSA-Express3 1000BASE-T CHPID types OSC and OSD (one port per CHPID)	Y	Y	Y	Y	Y	Y	Y	Y
OSA-Express3 1000BASE-T CHPID type OSE (one or two ports per CHPID)	Y	Y	Y	Y	Y	Y	Y	Y
OSA-Express3 1000BASE-T CHPID type OSM ^x (one port per CHPID)	Y	Y	Y ^q	Y ^q	Y ^q	N ^v	Y	N ^v
OSA-Express3 1000BASE-T CHPID type OSN ^w	Y	Y	Y	Y	Y	Y	Y	Y
Parallel Sysplex and other								
z/VM integrated systems management	-	-	-	-	-	Y	Y	Y
System-initiated CHPID reconfiguration	Y	Y	Y	Y	Y	-	-	-

Function	z/OS V2 R1	z/OS V1 R13	z/OS V1 R12	z/OS V1 R11	z/OS V1 R10	z/VM V6 R3	z/VM V6 R2	z/VM V5 R4
Program-directed re-IPL	-	-	-	-	-	Y	Y	Y
Multipath IPL	Y	Y	Y	Y	Y	N	N	N
Server Time Protocol (STP) enhancements	Y	Y	Y	Y	Y	-	-	-
STP	Y	Y	Y	Y	Y	-	-	-
Coupling over InfiniBand CHPID type CIB	Y	Y	Y	Y	Y	Y ^y	Y ^y	Y ^y
InfiniBand coupling links 12x at a distance of 150 m (492 ft)	Y	Y	Y	Y ^q	Y ^q	Y ^y	Y ^y	Y ^y
InfiniBand coupling links 1x at an unrepeat distance of 10 km (6.2 miles)	Y	Y	Y	Y ^q	Y ^q	Y ^y	Y ^y	Y ^y
Dynamic I/O support for InfiniBand CHPIDs	-	-	-	-	-	Y ^y	Y ^y	Y ^y
Coupling facility control code (CFCC) Level 18	Y	Y ^q	Y ^q	N	N	Y ^b	Y ^b	Y ^b
CFCC Level 19	Y ^q	Y ^q	Y ^q	N	N	Y ^b	Y ^b	Y ^b
CFCC Level 19 Flash Express exploitation	Y ^q	Y ^q	N	N	N	N	N	N
CFCC Level 19 Coupling Thin Interrupts	Y ^q	Y ^q	Y ^q	N	N	N	N	N

- a. A maximum of 32 PUs per system image are supported. Guests can be defined with up to 64 virtual PUs. z/VM V5R4 and later support up to 32 real PUs.
- b. Support is for guest use only.
- c. Available for z/OS on virtual machines without virtual zAAPs defined when zAAPs are not defined on the z/VM LPAR.
- d. z/VM V6R3 supports 1 TB of real memory and up to 1 TB of central storage for individual virtual machines.
- e. 256 GB of central memory are supported by z/VM V5R4 and later. z/VM V5R4 and later are designed to support more than 1 TB of virtual memory in use for guests.
- f. A web deliverable is required for Pageable 1 M large page support.
- g. 2 GB large page support available as a web deliverable.
- h. Not available to guests.
- i. Support varies by operating system, and by version and release.
- j. FMIDs are shipped in a web deliverable.
- k. Crypto Express4S Toleration requires a web deliverable and PTFs.
- l. Crypto Express4S Exploitation requires a web deliverable.
- m. Dynamic Reconfiguration Support for Flash Express available as a web deliverable.
- n. Software decompression support only.
- o. Statement of Direction - IBM intends to add support for z/VM 6.3 in the future.
- p. Compatibility support only.
- q. Service is required.
- r. Support varies with operating system and level. For more information, see 8.3.41, "FCP provides increased performance" on page 298.
- s. FICON Express4 features are withdrawn from marketing.
- t. Supported for dedicated devices only.
- u. Supported for dedicated devices only.
- v. Only supported for dynamic I/O.
- w. CHPID type OSN does not use ports. All communication is LPAR to LPAR.
- x. One port is configured for OSM. The other port in the pair is unavailable.
- y. Support is for dynamic I/O configuration only.

Table 8-4 zEC12 functions minimum support requirements summary (part 2 of 2)

Function	z/VSE V5R1 ^a	z/VSE V4R3 ^b	z/TPF V1R1	Linux on System z
zEC12	Y ^g	Y ^g	Y	Y
Support of Unified Resource Manager	N	N	N	N
Greater than 64 PUs single system image	N	N	Y	N
Greater than 54 PUs single system image	N	N	Y	Y
Support of IBM zAware	-	-	-	-
System z Integrated Information Processors (zIIPs)	-	-	-	-
System z Application Assist Processors (zAAPs)	-	-	-	-
zAAP on zIIP	-	-	-	-
Java Exploitation of Transactional Execution	N	N	N	N
Large memory (> 128 GB)	N	N	Y	Y
Large page support	Y	Y	N	Y
Out-of-order execution	Y	Y	Y	Y
Guest support for Execute-extensions facility	-	-	-	-
Hardware decimal floating point ^c	N	N	N	Y
Zero address detection	N	N	N	N
Sixty logical partitions	Y	Y	Y	Y
CPU measurement facility	N	N	N	N
LPAR group capacity limit	-	-	-	-
LPAR physical capacity limit	Y ^g	N	N	N
Separate LPAR management of PUs	Y	Y	Y	Y
Dynamic add/delete logical partition name	N	N	N	Y
Capacity provisioning	-	-	N	-
Enhanced flexibility for CoD	-	-	N	-
HiperDispatch	N	N	N	N
63.75 K subchannels	N	N	N	Y
Four logical channel subsystems (LCSSs)	Y	Y	N	Y
Dynamic I/O support for multiple LCSSs	N	N	N	Y
Third subchannel set	N	N	N	N
Multiple subchannel sets	N	N	N	Y
IPL from alternate subchannel set	N	N	N	N
MIDAW facility	N	N	N	N

Function	z/VSE V5R1 ^a	z/VSE V4R3 ^b	z/TPF V1R1	Linux on System z
Cryptography				
CP Assist for Cryptographic Function (CPACF)	Y	Y	Y	Y
CPACF AES-128, AES-192, and AES-256	Y	Y	Y ^d	Y
CPACF SHA-1, SHA-224, SHA-256, SHA-384, and SHA-512	Y	Y	Y ^e	Y
CPACF protected key	N	N	N	N
Crypto Express4S Toleration	Y ^f	N	Y ^{gh}	Y ^j
Secure IBM Enterprise PKCS #11 (EP11) coprocessor mode	N	N	N	N
Crypto Express3	Y	Y	Y ^{gh}	Y
Elliptic Curve Cryptography (ECC)	N	N	N	N ^j
HiperSockets				
32 HiperSockets	Y	Y	Y	Y
HiperSockets Completion Queue	Y ^g	N	N	Y
HiperSockets integration with IEDN	N	N	N	N
HiperSockets Virtual Switch Bridge	-	-	-	Y ⁱ
HiperSockets Network Traffic Analyzer	N	N	N	Y ^j
HiperSockets Multiple Write Facility	N	N	N	N
HiperSockets support of IPV6	Y	Y	N	Y
HiperSockets Layer 2 support	N	N	N	Y
HiperSockets	Y	Y	N	Y
Flash Express Storage				
Flash Express	N	N	N	N ^j
zEnterprise Data Compression (zEDC)				
zEDC Express	N	N	N	N
Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE)				
10GbE RoCE Express	N	N	N	N ^j
FICON and FCP				
z/OS Discovery and autoconfiguration (zDAC)	N	N	N	N
24k subchannel support for FICON Express8S, FICON Express8, and the FICON Express4 CHPID type FC	N	N	N	Y
FICON Express8S support of zHPF enhanced multitrack CHPID type FC	N	N	N	Y
FICON Express8 support of zHPF enhanced multitrack CHPID type FC	N	N	N	N
High Performance FICON for System z (zHPF)	N	N	N	Y ^k

Function	z/VSE V5R1 ^a	z/VSE V4R3 ^b	z/TPF V1R1	Linux on System z
FCP increased performance for small block sizes	Y	Y	N	Y
Request node identification data	-	-	-	-
FICON link incident reporting	N	N	N	N
GRS FICON CTC toleration	-	-	-	-
N-Port ID Virtualization for FICON (NPIV) CHPID type FCP	Y	Y	N	Y
FCP point-to-point attachments	Y	Y	N	Y
FICON SAN platform and name registration	Y	Y	Y	Y
FCP SAN management	N	N	N	Y
SCSI IPL for FCP	Y	Y	N	Y
Cascaded FICON Directors CHPID type FC	Y	Y	Y	Y
Cascaded FICON Directors CHPID type FCP	Y	Y	N	Y
FICON Express8S support of hardware data router CHPID type FCP	N	N	N	Y ^l
FICON Express8S and FICON Express8 support of T10-DIF CHPID type FCP	N	N	N	Y ^k
FICON Express8S, FICON Express8, FICON Express4 10KM LX, and FICON Express4 SX support of SCSI disks CHPID type FCP	Y	Y	N	Y
FICON Express8S ^c CHPID type FC	Y	Y	Y	Y
FICON Express8 ^c CHPID type FC	Y	Y ^m	Y ^m	Y ^m
FICON Express4 10KM LX and SX ^{c n} CHPID type FC	Y	Y	Y	Y
OSA				
VLAN management	N	N	N	N
VLAN (IEE 802.1q) support	Y	N	N	Y
Queued direct I/O (QDIO) data connection isolation for z/VM virtualized environments	-	-	-	-
OSA Layer 3 Virtual MAC	N	N	N	N
OSA Dynamic LAN idle	N	N	N	N
OSA/SF enhancements for IP, MAC addressing (CHPID=OSD)	N	N	N	N
QDIO Diagnostic Synchronization	N	N	N	N
Network Traffic Analyzer	N	N	N	N
Large send for IPv6 packets	-	-	-	-
Broadcast for IPv4 packets	N	N	N	Y
Checksum offload for IPv4 packets	N	N	N	Y

Function	z/VSE V5R1 ^a	z/VSE V4R3 ^b	z/TPF V1R1	Linux on System z
OSA-Express4S and OSA-Express3 inbound workload queuing for Enterprise Extender	N	N	N	N
OSA-Express5S 10 Gigabit Ethernet LR and SR CHPID type OSD	Y	Y	Y ^o	Y
OSA-Express5S 10 Gigabit Ethernet LR and SR CHPID type OSX	Y	N	Y ^p	Y ^q
OSA-Express5S Gigabit Ethernet LX and SX CHPID type OSD (two port per CHPID)	Y	Y	Y ^o	Y ^r
OSA-Express5S Gigabit Ethernet LX and SX CHPID type OSD (one port per CHPID)	Y	Y	Y ^o	Y
OSA-Express5S 1000BASE-T Ethernet CHPID type OSC	Y	Y	N	-
OSA-Express5S 1000BASE-T Ethernet CHPID type OSD (two port per CHPID)	Y	Y	Y ^o	Y ^r
OSA-Express5S 1000BASE-T Ethernet CHPID type OSD (one port per CHPID)	Y	Y	Y ^o	Y
OSA-Express5S 1000BASE-T Ethernet CHPID type OSE	Y	Y	N	N
OSA-Express5S 1000BASE-T Ethernet CHPID type OSM	N	N	N	Y ^s
OSA-Express5S 1000BASE-T Ethernet CHPID type OSN	Y	Y	Y	Y
OSA-Express4S 10-Gigabit Ethernet LR and SR CHPID type OSD	Y	Y	Y	Y
OSA-Express4S 10-Gigabit Ethernet LR and SR CHPID type OSX	Y	N	Y ^t	Y
OSA-Express4S Gigabit Ethernet LX and SX CHPID type OSD (two ports per CHPID)	Y	Y	Y ^t	Y
OSA-Express4S Gigabit Ethernet LX and SX CHPID type OSD (one port per CHPID)	Y	Y	Y	Y
OSA-Express4S 1000BASE-T CHPID type OSC (one or two ports per CHPID)	Y	Y	N	-
OSA-Express4S 1000BASE-T CHPID type OSD (two ports per CHPID)	Y	Y	Y ^g	Y
OSA-Express4S 1000BASE-T CHPID type OSD (one port per CHPID)	Y	Y	Y	Y
OSA-Express4S 1000BASE-T CHPID type OSE (one or two ports per CHPID)	Y	Y	N	N
OSA-Express4S 1000BASE-T CHPID type OSM ^v	N	N	N	Y
OSA-Express4S 1000BASE-T CHPID type OSN ^u (one or two ports per CHPID)	Y	Y	Y ^g	Y

Function	z/VSE V5R1 ^a	z/VSE V4R3 ^b	z/TPF V1R1	Linux on System z
OSA-Express3 10-Gigabit Ethernet LR and SR CHPID type OSD	Y	Y	Y	Y
OSA-Express3 10-Gigabit Ethernet LR and SR CHPID type OSX	Y	N	N	Y ^j
OSA-Express3 Gigabit Ethernet LX and SX CHPID types OSD, OSN ^u (two ports per CHPID)	Y	Y	Y ^t	Y
OSA-Express3 Gigabit Ethernet LX and SX CHPID types OSD, OSN ^u (one port per CHPID)	Y	Y	Y	Y
OSA-Express3 1000BASE-T CHPID type OSC (four ports)	Y	Y	Y	-
OSA-Express3 1000BASE-T (two ports per CHPID) CHPID type OSD	Y	Y	Y ^t	Y
OSA-Express3 1000BASE-T (one port per CHPID) CHPID type OSD	Y	Y	Y	Y
OSA-Express3 1000BASE-T (one or two ports per CHPID) CHPID type OSE	Y	Y	N	N
OSA-Express3 1000BASE-T Ethernet CHPID type OSN ^u	Y	Y	Y	Y
OSA-Express3 1000BASE-T CHPID type OSM ^v (two ports)	N	N	N	N
Parallel Sysplex and other				
z/VM integrated systems management	-	-	-	-
System-initiated CHPID reconfiguration	-	-	-	Y
Program-directed re-IPL ^w	Y	Y	-	Y
Multipath IPL	-	-	-	-
Server Time Protocol (STP) enhancements	-	-	-	-
STP Server Time Protocol	-	-	Y ^x	-
Coupling over InfiniBand CHPID type CIB	-	-	Y	-
InfiniBand coupling links 12x at a distance of 150 m (492 ft)	-	-	-	-
InfiniBand coupling links 1x at unrepeated distance of 10 km (6.2 miles)	-	-	-	-
Dynamic I/O support for InfiniBand CHPIDs	-	-	-	-
CFCC Level 18	-	-	Y	-
CFCC Level 19	-	-	Y	-
CFCC Level 19 Flash Express exploitation	-	-	-	-
CFCC Level 19 Coupling Thin Interrupts	-	-	-	-

a. z/VSE V5R1 is designed to use z/Architecture, specifically 64-bit real and virtual-memory addressing. z/VSE V5R1 requires an architectural level set available with IBM System z9 or later.

b. z/VSE V4 is designed to use z/Architecture, specifically 64-bit real-memory addressing, but does not support 64-bit virtual-memory addressing.

- c. Support varies with operating system and level.
- d. z/TPF supports only AES-128 and AES-256.
- e. z/TPF supports only SHA-1 and SHA-256.
- f. Crypto Express4S Exploitation requires PTFs.
- g. Service is required.
- h. Supported only when running in accelerator mode.
- i. Applicable to Guest Operating Systems.
- j. IBM is working with its Linux distribution partners to include support in future Linux on System z distribution releases.
- k. Supported by SLES 11.
- l. Supported by SLES 11 SP3 and RHEL 6.4.
- m. For more information, see 8.3.41, "FCP provides increased performance" on page 298.
- n. FICON Express4 features are withdrawn from marketing.
- o. Requires PUT 5 with PTFs.
- p. Requires PUT 8 with PTFs.
- q. Supported by SLES 11 SP1, SLES 10 SP4, and RHEL 6, RHEL 5.6.
- r. Supported by SLES 11, SLES 10 SP2, and RHEL 6, RHEL 5.2.
- s. Supported by SLES 11 SP2, SLES 10 SP4, and RHEL 6, RHEL 5.2.
- t. Requires PUT 4 with PTFs.
- u. CHPID type OSN does not use ports. All communication is LPAR to LPAR.
- v. One port is configured for OSM. The other port is unavailable.
- w. For FCP-SCSI disks.
- x. Server Time Protocol (STP) is supported in z/TPF with APAR PJ36831 in PUT 07.

8.3 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

8.3.1 Single system image

A single system image can control several processor units, such as CPs, zIIPs, zAAPs, or IFLs, as appropriate.

Maximum number of PUs per system image

Table 8-5 lists the maximum number of PUs supported by each operating system image and by special-purpose LPARs.

Table 8-5 Single system image size software support

Operating system	Maximum number of PUs per system image
z/OS V2R1	100 ^a
z/OS V1R13	100 ^a
z/OS V1R12	100 ^a
z/OS V1R11	100 ^a
z/VM V6R3	32 ^b
z/VM V6R2	32 ^b

Operating system	Maximum number of PUs per system image
z/VM V5R4	32 ^b
z/VSE V4R3 and later	z/VSE Turbo Dispatcher can use up to 4 CPs, and tolerates up to 10-way LPARs
z/TPF V1R1	86 CPs
CFCC Level 19	16 CPs or ICFs: CPs and ICFs cannot be mixed
zAware	80
Linux on System z ^c	SUSE SLES 11: 64 CPs or IFLs SUSE SLES 10: 64 CPs or IFLs Red Hat RHEL 6: 80 CPs or IFLs Red Hat RHEL 5: 80 CPs or IFLs

- a. The number of purchased zAAPs and the number of purchased zIIPs each cannot exceed the number of purchased CPs. An LPAR can be defined with any number of the available zAAPs and zIIPs. The total refers to the sum of these PU characterizations.
- b. When running on a VM-mode LPAR, z/VM can manage CPs, IFLs, zAAPs, and zIIPs. Otherwise, only CPs or IFLs (but not both simultaneously) are supported.
- c. Values are for z196 support. IBM is working with its Linux distribution partners to provide use of this function in future Linux on System z distribution releases.

The zAware-mode logical partition (LPAR)

zEC12 introduces an LPAR mode, called *zAware-mode*, that is exclusively for running the IBM zAware virtual appliance. 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 z/VM-mode LPAR

zEC12 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 V5R4 or later, and allows z/VM to use a wider variety of specialty processors in a single LPAR. For instance, in a z/VM-mode LPAR, z/VM can manage Linux on System z guests running on IFL processors while also managing z/VSE and z/OS on central processors (CPs). It also allows z/OS to fully use IBM System z Integrated Information Processors (zIIPs) and IBM System z Application Assist Processors (zAAPs).

8.3.2 zAAP support

zAAPs do not change the model capacity identifier of the zEC12. IBM software product license charges based on the model capacity identifier are not affected by the addition of zAAPs. On a zEC12, z/OS V1R11 is the minimum level for supporting zAAPs, together with the current IBM software development kits (SDKs) for z/OS Java 2 Technology Edition.

zAAPs can be used by the following applications:

- ▶ Any Java application that is using the current IBM SDK.
- ▶ WebSphere Application Server V5R1 and later, and also, products that are based on it, such as WebSphere Portal, WebSphere Enterprise Service Bus (WebSphere ESB), and WebSphere Business Integration (WBI) for z/OS.

- ▶ CICS/TS V2R3 and later.
- ▶ DB2 UDB for z/OS Version 8 and later.
- ▶ IMS Version 8 and later.
- ▶ All z/OS XML System Services validation and parsing that run in TCB mode, which might be eligible for zAAP processing. This eligibility requires z/OS V1R9 and later. For z/OS 1R10 (with appropriate maintenance), middleware and applications that request z/OS XML System Services can have z/OS XML System Services processing running on the zAAP.

To use zAAPs, DB2 V9 has the following prerequisites:

- ▶ DB2 V9 for z/OS in new function mode
- ▶ The C application programming interface (API) for z/OS XML System Services, available with z/OS V1R9 with rollback APARs to z/OS V1R7, and z/OS V1R8
- ▶ One of the following items:
 - z/OS V1R9² has native support.
 - z/OS V1R8² requires an APAR for zAAP support.

The functioning of a zAAP is transparent to all Java programming on Java virtual machine (JVM) V1.4.1 and later.

Use the PROJECTCPU option of the IEAOPTxx parmlib member to help determine whether zAAPs can be beneficial to the installation. Setting PROJECTCPU=YES directs z/OS to record the amount of eligible work for zAAPs and zIIPs in SMF record type 72 subtype 3. The field APPL% AAPCP of the Workload Activity Report listing by the Workload Manager (WLM) service class indicates the percentage of a processor that is zAAP eligible. Because of the zAAP lower price as compared to a CP, a utilization as low as 10% might provide benefits.

8.3.3 zIIP support

zIIPs do not change the model capacity identifier of the zEC12. IBM software product license charges based on the model capacity identifier are not affected by the addition of zIIPs. On a zEC12, z/OS Version 1 Release 11 is the minimum level for supporting zIIPs.

No changes to applications are required to use zIIPs. zIIPs 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

The functioning of a zIIP is transparent to application programs.

² z/OS V1R11 is the minimum z/OS level to support zAAP on zEC12.

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 zAAPs and zIIPs in SMF record type 72 subtype 3. The field APPL% IPCP 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.

8.3.4 zAAP on zIIP capability

This capability, first made available on System z9 servers under defined circumstances, enables workloads eligible to run on zAAPs to run on zIIP. It is intended as a means to optimize the investment on existing zIIPs, not as a replacement for zAAPs. The rule of at least one CP installed per zAAP and zIIP installed still applies.

IBM has released PTF for APAR OA38829 on z/OS V1R12 and V1R13. This PTF allows zAAP-eligible workloads to be dispatched on zIIPs even when there are active zAAPs installed, provided that the number of installed zAAPs plus the number of installed zIIPs does not exceed the number of installed standard CPs.

This PTF is intended only to help facilitate migration and testing of zAAP workloads on zIIP processors.

Statement of Direction: IBM zEnterprise EC12 is planned to be the last high-end System z server to offer support for zAAP specialty engine processors. IBM intends to continue support for running zAAP workloads on zIIP processors (“zAAP on zIIP”). This configuration is intended to help simplify capacity planning and performance management, while still supporting all the currently eligible workloads.

Because z/VM can dispatch both virtual zAAPs and virtual zIIPs on real CPs³, the z/VM partition does not require any real zIIPs defined to it. However, in general, you need to use real zIIPs for software licensing reasons.

Support is available on z/OS V1R11 and later. This capability is enabled by default (ZAAPZIIP=YES). To disable it, specify NO for the ZAAPZIIP parameter in the IEASYSxx PARMLIB member.

On z/OS V1R10, support is provided by PTF for APAR OA27495 and the default setting in the IEASYSxx PARMLIB member is ZAAPZIIP=NO. Enabling or disabling this capability is disruptive. After you change the parameter, run IPL for z/OS again for the new setting to take effect.

8.3.5 Transactional Execution (TX)

The IBM zEnterprise EC12 introduces an architectural feature called *Transactional Execution* (TX). This capability is known in academia and industry as “*hardware transactional memory*”.

³ The z/VM system administrator can use the SET CPUAFFINITY command to influence the dispatching of virtual specialty engines on CPs or real specialty engines.

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 designed to be used by IBM JVM, but potentially can be used by other software. z/OS V1R13 with PTFs or later is required.

8.3.6 Maximum main storage size

Table 8-6 lists the maximum amount of main storage that is supported by current operating systems. A maximum of 1 TB of main storage can be defined for an LPAR on a zEC12.

Expanded storage, although part of the z/Architecture, is used only by z/VM.

Statement of direction: In a future z/VM deliverable, IBM plans to withdraw support for expanded storage (XSTOR). With the enhanced memory management support added in z/VM V6R3, expanded storage is no longer advised as part of the paging configuration. z/VM can now run efficiently in all central storage configurations.

Table 8-6 Maximum memory that is supported by the operating system

Operating system	Maximum supported main storage ^a
z/OS	z/OS V1R11 and later support 4 TB and up to 3 TB per server ^a
z/VM	z/VM V6R3 and later support 1 TB z/VM V5R4 and z/VM V6R1 support 256 GB
z/VSE	z/VSE V4R3 and later support 32 GB
z/TPF	z/TPF supports 4 TB ^a
CFCC	Level 19 supports up to 3 TB per server ^a
zAware	Supports up to 3 TB per server ^a
Linux on System z (64-bit)	SUSE SLES 11 supports 4 TB ^a SUSE SLES 10 supports 4 TB ^a Red Hat RHEL 5 supports 3 TB ^a Red Hat RHEL 6 supports 3 TB ^a

a. zEC12 restricts the maximum LPAR memory size to 1 TB.

8.3.7 Flash Express

IBM zEnterprise EC12 introduces *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 exploiter to use Flash Express storage as storage-class memory (SCM) for paging store and supervisor call (SVC) dumps. Flash memory is a faster paging device as compared to hard disk. SVC 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 will 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 available to a particular LPAR is specified at the Support Element (SE) or Hardware Management Console (HMC) by using a new Flash Storage Allocation panel. 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 can also 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 is always 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 management (RSM) Enablement Offering Web Deliverable (FMID JBB778H) for z/OS V1R13⁴. The installation of this web deliverable requires careful planning as the size of the Nucleus, extended system queue area (ESQA) per CPU, and RSM stack is increased. Also, there is a new memory pool for pageable large pages. For web-deliverable code on z/OS, see the z/OS downloads:

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

The support is also delivered in z/OS V2R1 (shipped in base product) or later. Table 8-7 list the minimum support requirements for Flash Express.

Table 8-7 Minimum support requirements for Flash Express

Operating system	Support requirements
z/OS	z/OS V1R13 ^a

a. Web deliverable and PTFs required.

Flash Express exploitation by CFCC

CFCC Level 19 supports Flash Express. Initial CF Flash exploitation is targeted for MQ shared queues application structures. Structures can now be allocated with a combination of real memory and Storage Class Memory (SCM) provided by the Flash Express feature. For more information, see “Flash Express exploitation by CFCC” on page 325.

⁴ Dynamic reconfiguration support for SCM available as a web deliverable.

8.3.8 zEnterprise Data Compression (zEDC) Express

The growth of data that needs to be captured, transferred, and stored for a long time is not relenting. Software-implemented compression algorithms are costly in terms of processor resources, and storage costs are not negligible either.

zEDC Express, an optional feature exclusive to zEC12 and zBC12, 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 System z.

Exploitation support of zEDC Express functionality is provided exclusively by z/OS V2R1 zEnterprise Data Compression 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 via software on z/OS V2R1, V1R13, and V1R12 with the correct PTFs. Software decompression is slow and uses considerable processor resources; therefore, it is not suggested for production environments. Table 8-8 shows the minimum support requirements for zEDC Express.

Table 8-8 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)

a. PTFs are required.

Statement of Direction: IBM plans to provide exploitation support of zEDC Express functionality in z/VM 6.3.

8.3.9 10GbE RoCE Express

The IBM zEnterprise EC12 delivers a significant paradigm shift in network communications by using existing System z and industry standard (mature and secure) communications technology along with emerging new network technology. It introduces the Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE).

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 can also help to reduce network latency with memory-to-memory transfers using Shared Memory Communications - Remote Direct Memory Access (SMC-R) in z/OS V2R1. It is transparent to applications and can be used for LPAR-to-LPAR communication on a single z/OS system or server-to-server communication in a multiple CPC environment.

z/OS V2R1 with PTFs is the only supporting OS for SMC-R protocol. It does not roll back to previous z/OS releases. z/OS V1R12 and z/OS V1R13 with PTFs only provide compatibility support.

Statement of direction: IBM plans to offer future z/VM support for guests to take advantage of the 10GbE RoCE Express feature on the IBM zEnterprise EC12 and IBM zEnterprise BC12 systems.

IBM is also working with its Linux distribution partners to include support in future Linux on System z distribution releases.

Table 8-9 list the minimum support requirements for 10GbE RoCE Express.

Table 8-9 Minimum support requirements for RoCE Express

Operating system	Support requirements
z/OS	z/OS V2R1 with PTFs

8.3.10 Large page support

In addition to the existing 1-MB large pages, 4-KB pages, and page frames, zEC12 supports *pageable 1-MB large pages, large pages that are 2 GB in size, and large page frames*. For more information, see “Large page support” on page 105.

Table 8-10 lists the support requirements for 1-MB large pages.

Table 8-10 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 V4R3: Supported for data spaces
Linux on System z	SUSE SLES 10 SP2 Red Hat RHEL 5.2

a. Web deliverable and PTFs required

Table 8-11 lists the support requirements for 2-GB large pages.

Table 8-11 Minimum support requirements for 2-GB large pages

Operating system	Support requirements
z/OS	z/OS V1R13

8.3.11 Guest support for execute-extensions facility

The *execute-extensions facility* contains several machine instructions. Support is required in z/VM so that guests can use this facility. Table 8-12 lists the minimum support requirements.

Table 8-12 Minimum support requirements for execute-extensions facility

Operating system	Support requirements
z/VM	z/VM V5R4

8.3.12 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. zEC12 inherited the decimal floating point accelerator feature that was introduced with z10 EC. For more information, see 3.4.4, “Decimal floating point (DFP) accelerator” on page 87.

Table 8-13 lists the operating system support for decimal floating point. For more information, see 8.5.7, “Decimal floating point and z/OS XL C/C++ considerations” on page 323.

Table 8-13 Minimum support requirements for decimal floating point

Operating system	Support requirements
z/OS	z/OS V1R11: Support includes XL, C/C++, HLASM, IBM Language Environment®, DBX, and CDA RTLE.
z/VM	z/VM V5R4: Support is for guest use.
Linux on System z	SUSE SLES 11 SP1 Red Hat RHEL 6

8.3.13 Up to 60 logical partitions

This feature, first made available in z9 EC, allows the system to be configured with up to 60 logical partitions (LPARs). Because channel subsystems can be shared by up to 15 LPARs, it is necessary to configure four channel subsystems to reach 60 LPARs. Table 8-14 lists the minimum operating system levels for supporting 60 LPARs.

Table 8-14 Minimum support requirements for 60 LPARs

Operating system	Support requirements
z/OS	z/OS V1R11
z/VM	z/VM V5R4
z/VSE	z/VSE V4R3
z/TPF	z/TPF V1R1
Linux on System z	SUSE SLES 10 Red Hat RHEL 5

Remember: The IBM zAware virtual appliance runs in a dedicated LPAR so, when activated, it reduces by one the maximum number of available LPARs.

8.3.14 Separate LPAR management of PUs

The zEC12 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 8-15 lists the support requirements for the separate LPAR management of PU pools.

Table 8-15 Minimum support requirements for separate LPAR management of PUs

Operating system	Support requirements
z/OS	z/OS V1R11
z/VM	z/VM V5R4

Operating system	Support requirements
z/VSE	z/VSE V4R3
z/TPF	z/TPF V1R1
Linux on System z	SUSE SLES 10 Red Hat RHEL 5

8.3.15 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 is able to take advantage of this support and nondisruptively acquire and release memory from the reserved area. z/VM V5R4 and higher are able to acquire memory nondisruptively, and immediately make it available to guests. z/VM virtualizes this support to its guests, which now can also increase their memory nondisruptively if supported by the guest operating system. Releasing memory from z/VM is a disruptive operation to z/VM. Releasing memory from the z/VM guest depends on the guest's operating system support.

Dynamic LPAR memory upgrade is not supported for IBM zAware-mode LPARs.

8.3.16 LPAR physical capacity limit enforcement

On the IBM zEnterprise EC12, PR/SM is enhanced to support an option to limit the amount of physical processor capacity consumed by an individual LPAR when a processor unit (PU) that is defined as a general-purpose processor (CP) or an Integrated Facility for Linux (IFL) is shared across a set of LPARs. This enhancement is designed to provide a physical capacity limit 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 8-16 lists the minimum operating system level that is required on zEC12.

Table 8-16 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

8.3.17 Capacity Provisioning Manager

The provisioning architecture enables clients to better control the configuration and activation of the On/Off Capacity on Demand. For more information, see 9.8, "Nondisruptive upgrades" on page 378. 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.

The Capacity Provisioning Manager, a function 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 only issues 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 on Table 8-17.

Table 8-17 Minimum support requirements for capacity provisioning

Operating system	Support requirements
z/OS	z/OS V1R11
z/VM	Not supported: Not available to guests

8.3.18 Dynamic PU add

Preplanning of an LPAR configuration allows defining reserved PUs that can be brought online when extra capacity is needed. Operating system support is required to exploit this capability without a re-IPL, that is, nondisruptively. This support has been in z/OS for a long time.

Dynamic PU add enhances this support by providing the ability to dynamically define and change the number and type of reserved PUs in an LPAR profile, removing any preplanning requirements. Table 8-18 lists the minimum required operating system levels to support this function.

The new resources are immediately made available to the operating system and, in the z/VM case, to its guests. Dynamic PU add is not supported for IBM zAware-mode LPARs.

Table 8-18 Minimum support requirements for dynamic PU add

Operating system	Support requirements
z/OS	z/OS V1R11
z/VM	z/VM V5R4
z/VSE	z/VSE V4R3

8.3.19 HiperDispatch

HiperDispatch, which is available for System z10 and later servers, represents a cooperative effort between the operating system and the zEC12 hardware. It improves efficiencies in both the hardware and the software in the following ways:

- ▶ Work can be dispatched across fewer logical processors, therefore reducing the multiprocessor (MP) effects and lowering the interference across multiple partitions.
- ▶ Operating system tasks can be dispatched to a small subset of logical processors. Processor Resource/Systems Manager (PR/SM) then ties these logical processors to the same physical processors, improving the hardware cache reuse and locality of reference characteristics. This configuration also reduces the rate of cross-book communications.

For more information, see 3.7, “Logical partitioning” on page 108. Table 8-19 on page 289 lists HiperDispatch support requirements.

Table 8-19 Minimum support requirements for HiperDispatch

Operating system	Support requirements
z/OS	z/OS V1R11 with PTFs
z/VM	z/VM V6R3

8.3.20 The 63.75-K subchannels

Servers before z9 EC reserved 1024 subchannels for internal system use out of the maximum of 64 K subchannels. Starting with z9 EC, the number of reserved subchannels was reduced to 256, increasing the number of subchannels available. Reserved subchannels exist only in subchannel set 0. One subchannel is reserved in each of subchannel sets 1 and 2.

The informal name, *63.75-K subchannels*, represents 65280 subchannels, as shown in the following equation:

$$63 \times 1024 + 0.75 \times 1024 = 65280$$

Table 8-20 lists the minimum operating system level that is required on the zEC12.

Table 8-20 Minimum support requirements for 63.75-K subchannels

Operating system	Support requirements
z/OS	z/OS V1R11
z/VM	z/VM V5R4
Linux on System z	SUSE SLES 10 Red Hat RHEL 5

8.3.21 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 ESCON⁵ (CHPID type CNC) and FICON (CHPID types FC) on the zEC12, zBC12, z196, z114, z10 EC, and z9 EC. z196 introduced the third subchannel set (SS2).

Table 8-21 lists the minimum operating system levels that are required on the zEC12.

Table 8-21 Minimum software requirement for MSS

Operating system	Support requirements
z/OS	z/OS V1R11
z/VM	z/VM V6R3 ^a
Linux on System z	SUSE SLES 10 Red Hat RHEL 5

a. For specific Geographically Dispersed Parallel Sysplex (GDPS) usage only

⁵ ESCON features are not supported on zEC12.

z/VM V6R3 Multiple Subchannel Set (MSS) support for mirrored DASD provides a subset of host support for the MSS facility to allow using an alternate subchannel set for Peer-to-Peer Remote Copy (PPRC) secondary volumes.

8.3.22 Third subchannel set

With z196, a *third subchannel set* (SS2) was introduced. It applies to ESCON⁵ (CHPID type CNC) and FICON (CHPID type FC for both FICON and zHPF paths) channels.

Together with the second set (SS1), SS2 can be used for disk alias devices of both primary and secondary devices, and as Metro Mirror secondary devices. This set helps facilitate storage growth and complements other functions such as extended address volume (EAV) and Hyper Parallel Access Volumes (HyperPAV).

Table 8-22 lists the minimum operating systems level that is required on the zEC12.

Table 8-22 Minimum software requirement for SS2

Operating system	Support requirements
z/OS	z/OS V1R11 with PTFs

8.3.23 IPL from an alternate subchannel set

zEC12 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 189.

8.3.24 MIDAW facility

The *Modified Indirect Data Address Word (MIDAW) facility* improves FICON performance. The MIDAW facility 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 V5R4 or higher. For more information, see 8.7, “MIDAW facility” on page 326.

Table 8-23 lists the minimum support requirements for MIDAW.

Table 8-23 Minimum support requirements for MIDAW

Operating system	Support requirements
z/OS	z/OS V1R11
z/VM	z/VM V5R4 for guest exploitation

8.3.25 HiperSockets Completion Queue

The *HiperSockets Completion Queue* is exclusive to zEC12, zBC12, z196, and z114. The Completion Queue function is designed to allow HiperSockets to transfer data synchronously if possible, and asynchronously if necessary. It therefore combines ultra-low latency with more tolerance for traffic peaks. This benefit can be especially helpful in burst situations.

Table 8-24 lists the minimum support requirements for HiperSockets Completion Queue.

Table 8-24 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
Linux on System z	SLES 11 SP2 Red Hat RHEL 6.2

a. PTFs are required.

8.3.26 HiperSockets integration with the intraensemble data network (IEDN)

The HiperSockets integration with IEDN is exclusive to zEC12, zBC12, z196, and z114. HiperSockets integration with 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 8-25 lists the minimum support requirements for HiperSockets integration with IEDN.

Table 8-25 Minimum support requirements for HiperSockets integration with IEDN

Operating system	Support requirements
z/OS	z/OS V1R13 ^a
z/VM	z/VM V6R2 ^{ab}

a. PTFs required

b. Not supported in z/VM 6.3

8.3.27 HiperSockets Virtual Switch Bridge

The HiperSockets Virtual Switch Bridge is exclusive to zEC12, zBC12, z196, and z114. *HiperSockets Virtual Switch Bridge* can integrate with the intraensemble data network (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 8-26 lists the minimum support requirements for HiperSockets Virtual Switch Bridge.

Table 8-26 Minimum support requirements for HiperSockets Virtual Switch Bridge

Operating system	Support requirements
z/VM	z/VM V6R2 ^{ab}
Linux on System z ^c	SLES 10 SP4 update (kernel 2.6.16.60-0.95.1) Red Hat RHEL 5.8 (GA-level)

a. PTFs are required.

b. Not supported in z/VM 6.3.

c. Applicable to Guest Operating Systems.

8.3.28 HiperSockets Multiple Write Facility

This capability 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 4.9.8, “HiperSockets” on page 171. Table 8-27 lists the minimum support requirements for HiperSockets Virtual Multiple Write Facility.

Table 8-27 Minimum support requirements for HiperSockets multiple write

Operating system	Support requirements
z/OS	z/OS V1R11

8.3.29 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 8-28 lists the minimum support requirements for HiperSockets IPv6 (CHPID type IQD).

Table 8-28 Minimum support requirements for HiperSockets IPv6 (CHPID type IQD)

Operating system	Support requirements
z/OS	z/OS V1R11
z/VM	z/VM V5R4
Linux on System z	SUSE SLES 10 SP2 Red Hat RHEL 5.2

8.3.30 HiperSockets Layer 2 support

For flexible and efficient data transfer for IP and non-IP workloads, the HiperSockets internal networks on zEC12 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 8-29 on page 293 lists the minimum support requirements for HiperSockets Layer 2.

Table 8-29 Minimum support requirements for HiperSockets Layer 2

Operating system	Support requirements
z/OS	z/OS V1R12
z/VM	z/VM V5R4 for guest exploitation
Linux on System z	SUSE SLES 10 SP2 Red Hat RHEL 5.2

8.3.31 HiperSockets network traffic analyzer for Linux on System z

HiperSockets network traffic analyzer (HS NTA), introduced with z196, is an enhancement to HiperSockets architecture. It provides support for tracing Layer2 and Layer3 HiperSockets network traffic in Linux on System z. This support allows Linux on System z to control the trace for the internal virtual LAN to capture the records into host memory and storage (file systems).

Linux on System z tools can be used to format, edit, and process the trace records for analysis by system programmers and network administrators.

8.3.32 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 Express2⁶, and FICON Express4⁶ 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 designed to support zHPF, as contrasted to FICON Express8, FICON Express4⁶, and FICON Express2⁶, which have a firmware-only zHPF implementation.

Table 8-30 lists the minimum support requirements for FICON Express8S.

Table 8-30 Minimum support requirements for FICON Express8S

Operating system	z/OS	z/VM	z/VSE	z/TPF	Linux on System z
Native FICON and Channel-to-Channel (CTC) CHPID type FC	V1R11 ^a	V5R4	V4R3	V1R1	SUSE SLES 10 Red Hat RHEL 5
zHPF single track operations CHPID type FC	V1R11	V6R2 ^b	N/A	N/A	SUSE SLES 11 SP1 Red Hat RHEL 6
zHPF multitrack operations CHPID type FC	V1R11 ^b	V6R2 ^b	N/A	N/A	SUSE SLES 11 SP2 Red Hat RHEL 6.1
Support of SCSI devices CHPID type FCP	N/A	V5R4 ^b	V4R3	N/A	SUSE SLES 10 Red Hat RHEL 5

⁶ All FICON Express4, FICON Express2, and FICON features are withdrawn from marketing.

Operating system	z/OS	z/VM	z/VSE	z/TPF	Linux on System z
Support of hardware data router CHPID type FCP	N/A	V6R3	N/A	N/A	N/A ^c
Support of T10-DIF CHPID type FCP	N/A	V5R4 ^b	N/A	N/A	SUSE SLES 11 SP2 ^c

- a. PTFs are required to support global resource serialization (GRS) FICON channel-to-channel (CTC) toleration.
- b. PTFs are required.
- c. IBM is working with its Linux distribution partners to provide exploitation of this function in future Linux on System z distribution releases.

8.3.33 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).

With FICON Express 8, clients might be able to consolidate existing FICON, FICON Express2⁶, and FICON Express4⁶ channels, while maintaining and enhancing performance.

Table 8-31 lists the minimum support requirements for FICON Express8.

Table 8-31 Minimum support requirements for FICON Express8

Operating system	z/OS	z/VM	z/VSE	z/TPF	Linux on System z
Native FICON and CTC CHPID type FC	V1R11	V5R4	V4R3	V1R1	SUSE SLES 10 Red Hat RHEL 5
zHPF single track operations CHPID type FC	V1R11	V6R2 ^b	N/A	N/A	N/A ^b
zHPF multitrack operations CHPID type FC	V1R11 ^a	V6R2 ^b	N/A	N/A	N/A
Support of SCSI devices CHPID type FCP	N/A	V5R4 ^a	V4R3	N/A	SUSE SLES 10 Red Hat RHEL 5
Support of T10-DIF CHPID type FCP	N/A	V5R4 ^a	N/A	N/A	SUSE SLES 11 SP2 ^b

- a. PTFs are required.
- b. IBM is working with its Linux distribution partners to provide exploitation of this function in future Linux on System z distribution releases.

8.3.34 z/OS discovery and autoconfiguration (zDAC)

zDAC is designed to automatically run a number of I/O configuration definition tasks for new and changed disk and tape controllers 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 zEC12 systems that run z/OS, and reduces complexity and setup time.

zDAC applies to all FICON features supported on zEC12 when configured as CHPID type FC. Table 8-32 lists the minimum support requirements for zDAC.

Table 8-32 Minimum support requirements for zDAC

Operating system	Support requirements
z/OS	z/OS V1R12 ^a

a. PTFs are required.

8.3.35 High performance FICON (zHPF)

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 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 instance, 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 announcement, zHPF supported the transfer of small blocks of fixed size data (4 K) from a single track. This capability is 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 Express4⁷, on the zEC12, zBC12, z196, and z114, 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 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 channel path identifiers (CHPIDs) defined to the same logical control unit (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 settings in the z/OS operating system. For z/OS exploitation, there is a parameter in the IECIOSxx member of SYS1.PARMLIB (ZHPF=YES or NO) and in the SETIOS system command to control 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 (channel command words (CCWs)). The way that zHPF (transport mode) manages channel program operations is significantly 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. Less processor is used compared to the existing FICON architecture. Certain complex CCW chains are not supported by zHPF.

zHPF is exclusive to zEC12, zBC12, z196, z114, and System z10. The FICON Express8S, FICON Express8, and FICON Express4^{7,8} (CHPID type FC) concurrently support both the existing FICON protocol and the zHPF protocol in the server Licensed Internal Code (LIC).

Table 8-33 lists the minimum support requirements for zHPF.

Table 8-33 Minimum support requirements for zHPF

Operating system	Support requirements
z/OS	Single track operations: z/OS V1R11 Multitrack operations: z/OS V1R11 with PTFs 64K enhancement: z/OS V1R11 with PTFs
z/VM	z/VM 6.2 for guest exploitation only
Linux on System z	SLES 11 SP1 supports zHPF. IBM continues to work with its Linux distribution partners on exploitation of appropriate zEC12 functions to be provided in future Linux on System z distribution releases.

For more information about FICON channel performance, see the performance technical papers on the System z I/O connectivity website:

http://www-03.ibm.com/systems/z/hardware/connectivity/ficon_performance.html

⁷ FICON Express4 LX 4 KM is not supported on zEC12.

⁸ All FICON Express4, FICON Express2, and FICON features are withdrawn from marketing.

8.3.36 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 8-34 lists the minimum support requirements for RNID.

Table 8-34 Minimum support requirements for RNID

Operating system	Support requirements
z/OS	z/OS V1R11

8.3.37 24k subchannels for the FICON Express

To help facilitate growth and continue to enable server consolidation, the IBM zEnterprise EC12 supports up to 24k subchannels per FICON Express channel (channel path identifier - 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 addressable within an LPAR remains at 63.75k for subchannel set 0 and 64k-1 for subchannel sets 1 and 2.

This support is exclusive to the zEC12 and the zBC12 and applies to the FICON Express8S, FICON Express8, and the FICON Express4 features when defined as CHPID type FC.

Table 8-35 lists the minimum support requirements of 24k subchannel support for FICON Express.

Table 8-35 Minimum support requirements for 24k subchannel

Operating system	Support requirements
z/OS	z/OS V1R11 ^a
z/VM	z/VM V5R4
Linux on System z	SLES 10 RHEL 5

a. PTFs are required.

8.3.38 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* information unit (IU) pacing. Extended distance FICON is transparent to operating systems and applies to all FICON Express 8S, FICON Express8, and FICON Express4⁸ features that carry native FICON traffic (CHPID type FC).

For exploitation, the control unit must support the new IU pacing protocol. IBM System Storage DS8000 series supports extended distance FICON for IBM System z environments. The channel defaults to current pacing values when it operates with control units that cannot use extended distance FICON.

8.3.39 Platform and name server registration in FICON channel

The FICON Express8S, FICON Express8, and FICON Express4⁹ features on the zEC12 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 zEC12 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 Fibre Channel - Generic Services 4 (FC-GS-4) standard.

8.3.40 FICON link incident reporting

FICON link incident reporting allows an operating system image (without operator intervention) to register for link incident reports. Table 8-36 lists the minimum support requirements for this function.

Table 8-36 Minimum support requirements for link incident reporting

Operating system	Support requirements
z/OS	z/OS V1R11

8.3.41 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 8-Gbps link speeds.

For more information about FCP channel performance, see the performance technical papers on the System z I/O connectivity website:

http://www-03.ibm.com/systems/z/hardware/connectivity/fcp_performance.html

8.3.42 N-Port ID virtualization (NPIV)

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 8-37 on page 299 lists the minimum support requirements for NPIV.

⁹ FICON Express4 LX 4KM is not supported on zEC12.

Table 8-37 Minimum support requirements for NPIV

Operating system	Support requirements
z/VM	z/VM V5R4 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 V4R3
Linux on System z	SUSE SLES 10 SP3 Red Hat RHEL 5.4

8.3.43 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 8.3.55, “Intraensemble data network (IEDN)” on page 308.

Table 8-38 lists the minimum support requirements for OSA-Express5S 10-Gigabit Ethernet LR and SR features.

Table 8-38 Minimum support requirements for OSA-Express5S 10-Gigabit Ethernet LR and SR

Operating system	Support requirements
z/OS	OSD: z/OS V1R11 ^a OSX: z/OS V1R11 ^a
z/VM	OSD: z/VM V5R4 OSX: z/VM V5R4 ^a and V6R3 for dynamic I/O only
z/VSE	OSD: z/VSE V4R3 OSX: 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 System z	OSD: SUSE SLES 10, Red Hat RHEL 5 OSX: SUSE SLES 10 SP4, Red Hat RHEL 5.6

a. PTFs are required.

8.3.44 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 Gigabit per second (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 8-39 lists the minimum support requirements for OSA-Express5S Gigabit Ethernet LX and SX.

Table 8-39 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 V1R11 ^a	OSD: z/OS V1R11 ^a
z/VM	OSD: z/VM V5R4 ^a	OSD: z/VM V5R4
z/VSE	OSD: z/VSE V4R3	OSD: z/VSE V4R3
z/TPF	OSD: z/TPF V1R1 PUT 5 ^a	OSD: z/TPF V1R1 PUT 5 ^a
IBM zAware	OSD	
Linux on System z	OSD: SUSE SLES 10 SP2 Red Hat RHEL 5.2	OSD: SUSE SLES 10 Red Hat RHEL 5

a. PTFs are required.

8.3.45 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 channel path identifier, 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 8.3.54, “Intranode management network (INMN)” on page 307.

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
- ▶ Ensemble management, with CHPID type OSM

Table 8-40 on page 301 lists the minimum support requirements for OSA-Express5S 1000BASE-T.

Table 8-40 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, OSN ^b : z/OS V1R11 ^a	OSC, OSD, OSE, OSM, OSN ^b : z/OS V1R11 ^a
z/VM	OSC, OSD ^a , OSE, OSN ^b : z/VM V5R4	OSC, OSD, OSE, OSM ^{ac} , OSN ^b : z/VM V5R4
z/VSE	OSC, OSD, OSE, OSN ^b : z/VSE V4R3	OSC, OSD, OSE, OSN ^b : z/VSE V4R3
z/TPF	OSD: z/TPF V1R1 PUT 5 ^a OSN ^b : z/TPF V1R1 PUT 5 ^a	OSD: z/TPF V1R1 PUT 5 ^a OSN ^b : z/TPF V1R1 ^a
IBM zAware	OSD	OSD
Linux on System z	OSD: SUSE SLES 10 SP2 Red Hat RHEL 5.2 OSN ^b : SUSE SLES 10 Red Hat RHEL 5	OSD: SUSE SLES 10 Red Hat RHEL 5 OSM: SUSE SLES 10 SP4 Red Hat RHEL 5.2 OSN ^b : SUSE SLES 10 Red Hat RHEL 5

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 V5R4 and V6R3 for dynamic I/O only.

8.3.46 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 queued direct input/output (QDIO) architecture for high-speed TCP/IP communication. The z196 introduced the CHPID type OSX. For more information, see 8.3.55, “Intraensemble data network (IEDN)” on page 308.

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 actually 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 optimize for redundancy. Table 8-41 on page 302 lists the minimum support requirements for OSA-Express4S 10-Gigabit Ethernet LR and SR features.

Table 8-41 Minimum support requirements for OSA-Express4S 10-Gigabit Ethernet LR and SR

Operating system	Support requirements
z/OS	OSD: z/OS V1R11 ^a OSX: z/OS V1R11 ^a
z/VM	OSD: z/VM V5R4 OSX: z/VM V5R4 ^a and V6R3 for dynamic I/O only
z/VSE	OSD: z/VSE V4R3 OSX: z/VSE V5R1
z/TPF	OSD: z/TPF V1R1 OSX: z/TPF V1R1 PUT4 ^a
IBM zAware	OSD OSX
Linux on System z	OSD: SUSE SLES 10, Red Hat RHEL 5 OSX: SUSE SLES 10 SP4, Red Hat RHEL 5.6

a. PTFs are required.

8.3.47 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 Gigabit per second (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-Express4S Gigabit Ethernet feature. Table 8-42 lists the minimum support requirements for OSA-Express4S Gigabit Ethernet LX and SX.

Table 8-42 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 V1R11 ^a	OSD: z/OS V1R11 ^a
z/VM	OSD: z/VM V5R4 ^a	OSD: z/VM V5R4
z/VSE	OSD: z/VSE V4R3	OSD: z/VSE V4R3
z/TPF	OSD: z/TPF V1R1 PUT 4 ^a	OSD: z/TPF V1R1
IBM zAware	OSD	
Linux on System z	OSD: SUSE SLES 10 SP2 Red Hat RHEL 5.2	OSD: SUSE SLES 10 Red Hat RHEL 5

a. PTFs are required.

8.3.48 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 channel path identifier, 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 8.3.54, “Intranode management network (INMN)” on page 307.

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
- ▶ 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 8-43 lists the minimum support requirements for OSA-Express4S 1000BASE-T.

Table 8-43 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, OSN ^b : z/OS V1R11 ^a	OSC, OSD, OSE, OSM, OSN ^b : z/OS V1R11 ^a
z/VM	OSC, OSD ^a , OSE, OSN ^b : z/VM V5R4	OSC, OSD, OSE, OSM ^{ac} , OSN ^b : z/VM V5R4
z/VSE	OSC, OSD, OSE, OSN ^b : z/VSE V4R3	OSC, OSD, OSE, OSN ^b : z/VSE V4R3
z/TPF	OSD: z/TPF V1R1 PUT 4 ^a OSN ^b : z/TPF V1R1 PUT 4 ^a	OSD: z/TPF V1R1 OSN ^b : z/TPF V1R1 PUT4 ^a
IBM zAware	OSD	OSD
Linux on System z	OSD: SUSE SLES 10 SP2 Red Hat RHEL 5.2 OSN ^b : SUSE SLES 10 Red Hat RHEL 5	OSD: SUSE SLES 10 Red Hat RHEL 5 OSM: SUSE SLES 10 SP4 Red Hat RHEL 5.6 OSN ^b : SUSE SLES 10 Red Hat RHEL 5

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 V5R4 and V6R3 for dynamic I/O only.

8.3.49 OSA-Express3 10-Gigabit Ethernet LR and SR

The *OSA-Express3 10-Gigabit Ethernet features* offer two ports, which are defined as 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 8.3.55, “Intraensemble data network (IEDN)” on page 308.

Table 8-44 lists the minimum support requirements for OSA-Express3 10-Gigabit Ethernet LR and SR features.

Table 8-44 Minimum support requirements for OSA-Express3 10-Gigabit Ethernet LR and SR

Operating system	Support requirements
z/OS	OSD: z/OS V1R11 OSX: z/OS V1R11, with service
z/VM	OSD: z/VM V5R4 OSX: z/VM V5R4 and V6R3 for dynamic I/O only
z/VSE	OSD: z/VSE V4R3
z/TPF	OSD: z/TPF V1R1
IBM zAware	OSD
Linux on System z	OSD: SUSE SLES 10 OSD: Red Hat RHEL 5

8.3.50 OSA-Express3 Gigabit Ethernet LX and SX

The *OSA-Express3 Gigabit Ethernet* features offer two cards with two PCI Express adapters each. Each PCI Express adapter controls two ports, giving a total of four ports per feature. Each adapter has its own CHPID, defined as either OSD or OSN, supporting the QDIO architecture for high-speed TCP/IP communication. Therefore, a single feature can support both CHPID types, with two ports for each type.

Operating system support is required to recognize and use the second port on each PCI Express adapter. The minimum support requirements for OSA-Express3 Gigabit Ethernet LX and SX features are listed in Table 8-45 (four ports).

Table 8-45 Minimum support requirements for OSA-Express3 Gigabit Ethernet LX and SX, four ports

Operating system	Support requirements
z/OS	z/OS V1R11
z/VM	z/VM V5R4, service required
z/VSE	z/VSE V4R3
z/TPF	z/TPF V1R1, service required
IBM zAware	
Linux on System z	SUSE SLES 10 SP2 Red Hat RHEL 5.2

The minimum support requirements for OSA-Express3 Gigabit Ethernet LX and SX features are listed in Table 8-46 on page 305 (two ports).

Table 8-46 Minimum support requirements for OSA-Express3 Gigabit Ethernet LX and SX, two ports

Operating system	Support requirements
z/OS	z/OS V1R11
z/VM	z/VM V5R4
z/VSE	z/VSE V4R3
z/TPF	z/TPF V1R1
IBM zAware	OSD
Linux on System z	SUSE SLES 10 Red Hat RHEL 5

8.3.51 OSA-Express3 1000BASE-T Ethernet

The *OSA-Express3 1000BASE-T Ethernet features* offer two cards with two PCI Express adapters each. Each PCI Express adapter controls two ports, giving a total of four ports for each feature. Each adapter has its own CHPID, defined as OSC, OSD, OSE, OSM, or OSN. A single feature can support two CHPID types, with two ports for each type. The OSM CHPID type is introduced with the z196. For more information, see 8.3.54, “Intranode management network (INMN)” on page 307.

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
- ▶ Ensemble management, with CHPID type OSM

Operating system support is required to recognize and use the second port on each PCI Express adapter. Minimum support requirements for the OSA-Express3 1000BASE-T Ethernet features are listed in Table 8-47 on page 306 (four ports) and Table 8-48 on page 306 (two ports).

Table 8-47 Minimum support requirements for OSA-Express3 1000BASE-T Ethernet, four ports

Operating system	Support requirements ^{a b}
z/OS	OSD: z/OS V1R11 OSE: z/OS V1R11 OSM: z/OS V1R11, with service OSN ^b : z/OS V1R11
z/VM	OSD: z/VM V5R4, service required OSE: z/VM V5R4 OSM: z/VM service required, V5R4 and V6R3 for dynamic I/O only OSN ^b : z/VM V5R4
z/VSE	OSD: z/VSE V4R3, service required OSE: z/VSE V4R3 OSN ^b : z/VSE V4R3
z/TPF	OSD and OSN ^b : z/TPF V1R1, service required
IBM zAware	OSD
Linux on System z	OSD: SUSE SLES 10 SP2 Red Hat RHEL 5.2 OSN ^b : SUSE SLES 10 SP2 Red Hat RHEL 5.2

a. Applies to CHPID types OSC, OSD, OSE, OSM, and OSN. For more information, see Table 8-48.

b. Although CHPID type OSN does not use any ports (because all communication is LPAR to LPAR), it is listed here for completeness.

Table 8-48 lists the minimum support requirements for OSA-Express3 1000BASE-T Ethernet (two ports).

Table 8-48 Minimum support requirements for OSA-Express3 1000BASE-T Ethernet, two ports

Operating system	Support requirements
z/OS	OSD, OSE, OSM ^a , and OSN: V1R11
z/VM	OSD, OSE, OSM, and OSN: V5R4
z/VSE	V4R3
z/TPF	OSD, OSN, and OSC: V1R1
IBM zAware	OSD
Linux on System z	OSD: SUSE SLES 10 Red Hat RHEL 5 OSN: SUSE SLES 10 SP3 Red Hat RHEL 5.4

a. OSM support in z/VM 5.4 and 6.3 for dynamic I/O only

8.3.52 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 zEC12.
- ▶ IEDN on the zEC12 to other CPC nodes in the ensemble. The zEC12 server also supports the use of HiperSockets over the IEDN.

8.3.53 Open Systems Adapter for Ensemble

Five different OSA-Express5S and OSA-Express4S features are used to connect the zEC12 to its attached zEnterprise BladeCenter Extension (zBX) Model 003, 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, feature code 0417
- ▶ OSA Express4S 1000BASE-T Ethernet, feature code 0408
- ▶ OSA-Express3 1000BASE-T Ethernet, feature code 3367

For the IEDN, use these features:

- ▶ OSA-Express5S 10 Gigabit Ethernet (GbE) Long Reach (LR), feature code 0415
- ▶ OSA-Express5S 10 Gigabit Ethernet (GbE) Short Reach (SR), feature code 0416
- ▶ OSA-Express4S 10 Gigabit Ethernet (GbE) Long Reach (LR), feature code 0406
- ▶ OSA-Express4S 10 Gigabit Ethernet (GbE) Short Reach (SR), feature code 0407
- ▶ OSA-Express3 10 Gigabit Ethernet (GbE) Long Reach (LR), feature code 3370
- ▶ OSA-Express3 10 Gigabit Ethernet (GbE) Short Reach (SR), feature code 3371

For more information about OSA-Express3, OSA-Express4S and OSA-Express5S in an ensemble network, see 7.4, “zBX connectivity” on page 244.

8.3.54 Intranode management network (INMN)

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 and zBX 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, OSA-Express4S 1000BASE-T, or OSA Express3 1000BASE-T ports, from two different OSA-Express5S 1000BASE-T, OSA-Express4S 1000BASE-T, or OSA-Express3 1000BASE-T features, that are configured as CHPID type OSM. One port per CHPID is available with CHPID type OSM.

The OSA connection is through the bulk power hub (BPH) port J07 on the zEC12 to the Top of Rack (ToR) switches on zBX.

8.3.55 Intraensemble data network (IEDN)

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 OSA-Express for zBX (OSX) CHPID type. The OSA connection is from the zEC12 to the ToR switches on zBX.

The IEDN requires two OSA-Express5S, OSA-Express4S, or OSA-Express3 10 GbE ports that are configured as CHPID type OSX.

8.3.56 OSA-Express5S and OSA-Express4S NCP support (OSN)

OSA-Express5S 1000BASE-T Ethernet and OSA-Express4S 1000BASE-T Ethernet features can provide channel connectivity from an operating system in a zEC12 to IBM Communication Controller for Linux on System z (CCL). This configuration uses the Open Systems Adapter for network control program (NCP) (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.

Because ESCON channels are not supported on zEC12, OSN is the only option. 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-01.ibm.com/software/network/ccl/>

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-Express5S 1000BASE-T features.

Table 8-49 lists the minimum support requirements for OSN.

Table 8-49 Minimum support requirements for OSA-Express5S and OSA-Express4S OSN

Operating system	Support requirements
z/OS	z/OS V1R11 ^a
z/VM	z/VM V5R4
z/VSE	z/VSE V4R3
z/TPF	z/TPF V1R1 PUT 4 ^a
Linux on System z	SUSE SLES 10 Red Hat RHEL 5

a. PTFs are required.

8.3.57 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 zEC12 through a port on the OSA-Express5S 1000BASE-T, OSA-Express4S 1000BASE-T, or OSA-Express3 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).

8.3.58 VLAN management enhancements

Table 8-50 lists the minimum support requirements for VLAN management enhancements for the OSA-Express5S, OSA-Express4S, and OSA-Express3 features (CHPID type OSD).

Table 8-50 Minimum support requirements for VLAN management enhancements

Operating system	Support requirements
z/OS	z/OS V1R11
z/VM	z/VM V5R4 . Support of guests is transparent to z/VM if the device is directly connected to the guest (pass through).

8.3.59 GARP VLAN Registration Protocol

All OSA-Express5S, OSA-Express4S, and OSA-Express3 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.

Minimum support requirements are listed in Table 8-51.

Table 8-51 Minimum support requirements for GVRP

Operating system	Support requirements
z/OS	z/OS V1R11
z/VM	z/VM V5R4

8.3.60 Inbound workload queuing (IWQ) for OSA-Express5S, OSA-Express4S, and OSA-Express3

OSA-Express3 introduced 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 provided by OSA-Express5S, OSA-Express4S, and OSA-Express3 IWQ reduces the conventional z/OS processing 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, OSA-Express4S, and OSA-Express3 CHPID types OSD and OSX, and the z/OS operating system. This limitation applies to zEC12, zBC12, z196, z114, and System z10. The minimum support requirements are listed in Table 8-52.

Table 8-52 Minimum support requirements for IWQ

Operating system	Support requirements
z/OS	z/OS V1R12
z/VM	z/VM V5R4 for guest exploitation only, service required

8.3.61 Inbound workload queuing (IWQ) 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, OSA-Express4S, and OSA-Express3 CHPID types OSD and OSX, and the z/OS operating system. This limitation applies to zEC12, zBC12, z196, and z114. The minimum support requirements are listed in Table 8-53.

Table 8-53 Minimum support requirements for IWQ

Operating system	Support requirements
z/OS	z/OS V1R13
z/VM	z/VM V5R4 for guest exploitation only, service required

8.3.62 Query and display OSA configuration

OSA-Express3 introduced the capability for the operating system to directly query and 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, OSA-Express4S, and OSA-Express3.

The **display OSAINFO** command is exclusive to z/OS, and applies to OSA-Express5S, OSA-Express4S, and OSA-Express3 features, and CHPID types OSD, OSM, and OSX.

8.3.63 Link aggregation support for z/VM

Link aggregation (IEEE 802.3ad) controlled by the z/VM Virtual Switch (VSWITCH) allows the dedication of an OSA-Express5S, OSA-Express4S, or OSA-Express3 port to the z/VM operating system. The port must be participating in an aggregated group that is configured in Layer 2 mode. Link aggregation (trunking) combines multiple physical OSA-Express5S, OSA-Express4S, or OSA-Express3 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 the OSA-Express5S, OSA-Express4S, and OSA-Express3 features when configured as CHPID type OSD (QDIO). Link aggregation is supported by z/VM V5R4 and later.

8.3.64 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 provides a means for creating security zones and preventing network traffic between the zones.

VSWITCH isolation support is provided by APAR VM64281. z/VM 5R4 and later support is provided by CP APAR VM64463 and TCP/IP APAR PK67610.

QDIO data connection isolation is supported by all OSA-Express5S, OSA-Express4S, and OSA-Express3 features on zEC12.

8.3.65 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 V1R11 or later and all OSA-Express5S, OSA-Express4S, and OSA-Express3 features on zEC12.

8.3.66 QDIO optimized latency mode (OLM)

QDIO 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, OSA-Express4S, or OSA-Express3 without needing more program controlled interrupts (PCIs).
- ▶ For outbound processing, the OSA-Express5S, OSA-Express4S, or OSA-Express3 also look more frequently for available data to process from the TCP/IP stack. The process therefore does not require a Signal Adapter (SIGA) instruction to determine whether more data is available.

8.3.67 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 zEC12, zBC12, z196, and z114. Large send is not supported for LPAR-to-LPAR packets. The minimum support requirements are listed in Table 8-54.

Table 8-54 Minimum support requirements for large send for IPv6 packets

Operating system	Support requirements
z/OS	z/OS V1R13 ^a
z/VM	z/VM V5R4 for guest exploitation only

a. PTFs are required.

8.3.68 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 8-55.

Table 8-55 Minimum support requirements for OSA-Express5S and OSA-Express4S checksum offload

Traffic	Support requirements
LPAR to LPAR	z/OS V1R12 ^a z/VM V5R4 for guest exploitation ^b
IPv6	z/OS V1R13 z/VM V5R4 for guest exploitation ^b
LPAR-to-LPAR traffic for IPv4 and IPv6	z/OS V1R13 z/VM V5R4 for guest exploitation ^b

a. PTFs are required.

b. Device is directly attached to guest, and PTFs are required.

8.3.69 Checksum offload for IPv4 packets when in QDIO mode

The *checksum offload function* supports z/OS and Linux on System z environments. It is offered on the OSA-Express5S GbE, OSA-Express5S 1000BASE-T Ethernet, OSA-Express4S GbE, OSA-Express4S 1000BASE-T Ethernet, OSA-Express3 GbE, and OSA-Express3 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) 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 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. Checksum offload does not apply to such IP packets.

Checksum offload is supported by the GbE features, which include FC 0404, FC 0405, FC 3362, and FC 3363. It is also supported by the 1000BASE-T Ethernet features, including FC 0408 and FC 3367, 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 System z distributions.

8.3.70 Adapter interruptions for QDIO

Linux on System z 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, OSA-Express4S, and OSA-Express3 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 System z.

Adapter interruptions apply to all of the OSA-Express5S, OSA-Express4S, and OSA-Express3 features on zEC12 when in QDIO mode (CHPID type OSD).

8.3.71 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, OSA-Express4S, and OSA-Express3 features adapt to the changes, avoiding thrashing and frequent updates to the OSA address table (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, OSA-Express4S, and OSA-Express3 features on zEC12 when in QDIO mode (CHPID type OSD). It is used by z/OS V1R8 and higher, with program temporary fixes (PTFs).

8.3.72 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 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 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, OSA-Express4S, and OSA-Express3 features on zEC12 when in QDIO mode (CHPID type OSD). It is used by z/OS V1R8 and later.

8.3.73 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, OSA-Express4S, and OSA-Express3 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, OSA-Express4S, and OSA-Express3 features on zEC12 when in QDIO mode (CHPID type OSD). It is used by z/OS V1R10 and later.

8.3.74 Network Traffic Analyzer

The zEC12 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, OSA-Express4S, and OSA-Express3 features on zEC12 when in QDIO mode (CHPID type OSD). It is used by z/OS V1R8 and later.

8.3.75 Program directed re-IPL

First available on System z9, *program directed re-IPL* allows an operating system on a zEC12 to re-IPL without operator intervention. This function is supported for both SCSI and IBM extended count key data (IBM ECKD™) devices. Table 8-56 lists the minimum support requirements for program directed re-IPL.

Table 8-56 Minimum support requirements for program directed re-IPL

Operating system	Support requirements
z/VM	z/VM V5R4
Linux on System z	SUSE SLES 10 SP3 Red Hat RHEL 5.4
z/VSE	V4R3 on SCSI disks

8.3.76 Coupling over InfiniBand

InfiniBand 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. Several areas of this book address InfiniBand characteristics and support. For more information, see 4.10, “Parallel Sysplex connectivity” on page 173.

InfiniBand coupling links

Table 8-57 lists the minimum support requirements for coupling links over InfiniBand.

Table 8-57 Minimum support requirements for coupling links over InfiniBand

Operating system	Support requirements
z/OS	z/OS V1R11
z/VM	z/VM V5R4 (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 HCA2-O LR (1xIFB) fanout that supports InfiniBand coupling links 1x at an unrepeated distance of 10 KM (6.2 miles) is listed in Table 8-58.

Table 8-58 Minimum support requirements for coupling links over InfiniBand at 10 km (6.2 miles)

Operating system	Support requirements
z/OS	z/OS V1R11, service required
z/VM	z/VM V5R4 (dynamic I/O support for InfiniBand CHPIDs only, coupling over InfiniBand is not supported for guest use)

8.3.77 Dynamic I/O support for InfiniBand CHPIDs

This function refers exclusively to the z/VM dynamic I/O support of InfiniBand coupling links. Support is available for the CIB 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, and does not support the use of InfiniBand coupling links by guests.

Table 8-59 lists the minimum support requirements for dynamic I/O support for InfiniBand CHPIDs.

Table 8-59 Minimum support requirements for dynamic I/O support for InfiniBand CHPIDs

Operating system	Support requirements
z/VM	z/VM V5R4

8.4 Cryptographic support

IBM zEnterprise EC12 provides two major groups of cryptographic functions:

- ▶ Synchronous cryptographic functions, which are provided by the CP Assist for Cryptographic Function (CPACF)
- ▶ Asynchronous cryptographic functions, which are provided by the Crypto Express4S and Crypto Express3 features

The minimum software support levels are listed in the following sections. Obtain and review the current Preventive Service Planning (PSP) buckets to ensure that the latest support levels are known and included as part of the implementation plan.

8.4.1 CP Assist for Cryptographic Function (CPACF)

In zEC12, 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 6.6, “CP Assist for Cryptographic Function (CPACF)” on page 210.

Table 8-60 lists the support requirements for CPACF at zEC12.

Table 8-60 Support requirements for CPACF

Operating system	Support requirements
z/OS ^a	z/OS V1R10 and later with the Cryptographic Support for z/OS V1R10 - V1R12 web deliverable.
z/VM	z/VM V5R4 with PTFs and higher: Supported for guest use.
z/VSE	z/VSE V4R2 and later: Supports the CPACF features with the functionality supported on IBM System z10.
z/TPF	z/TPF V1R1
Linux on System z	SUSE SLES 11 SP1 Red Hat RHEL 6.1 For Message-Security-Assist-Extension 4 exploitation, IBM is working with its Linux distribution partners to include support in future Linux on System z distribution releases.

a. CPACF is also used by several IBM software product offerings for z/OS, such as IBM WebSphere Application Server for z/OS.

8.4.2 Crypto Express4S

Support of *Crypto Express4S* functions varies by operating system and release. Table 8-61 lists the minimum software requirements for the Crypto Express4S features when configured as a coprocessor or an accelerator. For more information, see 6.7, “Crypto Express4S” on page 210.

Table 8-61 Crypto Express4S support on zEC12

Operating system	Crypto Express4S
z/OS	<ul style="list-style-type: none"> ▶ z/OS V1R13, or z/OS V1R12 with the Cryptographic Support for z/OS V1R12-V1R13 web deliverable. ▶ z/OS V1R10, or z/OS V1R11 with toleration maintenance.
z/VM	For guest use: z/VM V6R3, z/VM V6R2, or z/VM V5R4 with maintenance.
z/VSE	z/VSE V5R1 with PTFs.
z/TPF V1R1	Service required (accelerator mode only).
Linux on System z	IBM is working with its Linux distribution partners to include support in future Linux on System z distribution releases.

8.4.3 Crypto Express3

Support of *Crypto Express3 functions* varies by operating system and release. Table 8-62 lists the minimum software requirements for the Crypto Express3 features when configured as a coprocessor or an accelerator. For more information, see 6.8, “Crypto Express3” on page 212.

Table 8-62 *Crypto Express3 support on zEC12*

Operating system	Crypto Express3
z/OS	<ul style="list-style-type: none"> ▶ z/OS V1R12 (ICSF FMID HCR7770) and later ▶ z/OS V1R10, or z/OS V1R11 with the Cryptographic Support for z/OS V1R9-V1R11 web deliverable. ▶ z/OS V1R8 with toleration maintenance: Crypto Express3 features handled as Crypto Express2 features.
z/VM	z/VM V5R4: Service required, supported for guest use only.
z/VSE	z/VSE V4R2.
z/TPF V1R1	Service required (accelerator mode only).
Linux on System z	<p>For toleration:</p> <ul style="list-style-type: none"> ▶ SUSE SLES10 SP3 and SLES 11. ▶ Red Hat RHEL 5.4 and RHEL 6.0. <p>For use:</p> <ul style="list-style-type: none"> ▶ SUSE SLES11 SP1. ▶ Red Hat RHEL 6.1.

8.4.4 Web deliverables

For web-deliverable code on z/OS, see the z/OS downloads:

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

For Linux on System z, support is delivered through IBM and the distribution partners. For more information, see Linux on System z on the IBM developerWorks® website:

<http://www.ibm.com/developerworks/linux/linux390/>

8.4.5 z/OS ICSF 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 must be related to an ICSF FMID instead of a z/OS version.

For a list of ICSF versions and FMID cross-references, see the Technical Documents page:

<http://www.ibm.com/support/techdocs/atmastr.nsf/WebIndex/TD103782>

Table 8-63 on page 319 lists the ICSF FMIDs and web-deliverable codes for z/OS V1R10 through V1R13. Later FMIDs include the functions of previous ones.

Table 8-63 z/OS ICSF FMIDs

ICSF FMID	z/OS	Web deliverable name	Supported function
HCR7750	V1R10	Included as a base element of z/OS V1R10	<ul style="list-style-type: none"> ▶ CPACF AES-192 and AES-256 ▶ CPACF SHA-224, SHA-384, and SHA-512 ▶ 4096-bit RSA keys ▶ ISO-3 PIN block format
HCR7751	V1R11 V1R10	<p>Cryptographic Support for z/OS V1R8-V1R10 and z/OS.e V1R8^a</p> <p>Included as a base element of z/OS V1R11</p>	<ul style="list-style-type: none"> ▶ IBM System z10 BC support ▶ Secure key AES ▶ Keystore policy ▶ PKDS Sysplex-wide consistency ▶ In-storage copy of the PKDS ▶ 13-digit through 19-digit PANs ▶ Crypto Query service ▶ Enhanced SAF checking
HCR7770	V1R12 V1R11 V1R10	<p>Cryptographic Support for z/OS V1R9-V1R11</p> <p>Included as a base element of z/OS V1R12</p>	<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, ICSF-CICS Attach Facility)
HCR7780	V1R13 V1R12 V1R11 V1R10	<p>Cryptographic Support for z/OS V1R10-V1R12</p> <p>Included as a base element of z/OS V1R13</p>	<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

ICSF FMID	z/OS	Web deliverable name	Supported function
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"> ▶ Support for the Crypto Express4S feature when configured as an EP11 coprocessor ▶ Support for the Crypto Express4S feature when configured as a CCA coprocessor ▶ Support for 24-byte DES master keys ▶ Improved wrapping key strength ▶ DUKPT for MAC and encryption keys ▶ Secure Cipher Text Translate2 ▶ Compliance with new random number generation standards ▶ EMV enhancements for applications that support American Express cards
HCR77A1	V2R1 V1R13 V1R12	Cryptographic Support for z/OS V1R13-V2R1	

a. Service is required.

8.4.6 ICSF migration considerations

Consider the following points about the Cryptographic Support for z/OS V1R12 - V1R13 web deliverable ICSF HCR77A0 code:

- ▶ It is not integrated in IBM ServerPac (even for new z/OS V1R13 orders).
- ▶ It is only required to use the functions available with zEC12.
- ▶ All systems in a sysplex that share a PKDS/TKDS must be at HCR77A0 to use the PKDS/TKDS Coordinated Administration support.
- ▶ The ICSF toleration PTFs are needed for these reasons:
 - Permit the use of a PKDS with RSA private key tokens encrypted under the ECC Master Key.
 - Support for installation options data sets that use the keyword BEGIN(FMID).
- ▶ A new SMP/E for z/OS Fix Category is created for ICSF coexistence: IBM.Coexistence.ICSF.z/OS_V1R12-V1R13-HCR77A0

8.5 z/OS migration considerations

Except for base processor support, z/OS software changes do not require any of the functions introduced with the zEC12. 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.

8.5.1 General guidelines

The IBM zEnterprise EC12 introduces the latest System z technology. Although support is provided by z/OS starting with z/OS V1R11, use of zEC12 is dependent on the z/OS release. The z/OS.e is *not* supported on zEC12.

In general, consider the following guidelines:

- ▶ Do not change software releases and hardware at the same time.
- ▶ Keep members of sysplex at same software level, except during brief migration periods.
- ▶ Migrate to an STP-only or Mixed-CTN network before introducing a zEC12 into a Sysplex.
- ▶ Review zEC12 restrictions and migration considerations before creating an upgrade plan.

8.5.2 Hardware Configuration Definition (HCD)

On z/OS V1R11 and later, the HCD or Hardware Configuration Manager (HCM) help to define a configuration for zEC12.

8.5.3 InfiniBand coupling links

Each system can use, or not use, InfiniBand coupling links independently of what other systems are doing, and do so with other link types.

InfiniBand coupling connectivity can be obtained only with other systems that also support InfiniBand coupling. zEC12 does not support InfiniBand connectivity with System z9 and earlier systems.

8.5.4 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 parameter **LFAREA** 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.

8.5.5 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 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.

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 will 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.

To effectively use HiperDispatch, 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 use the older WLM policies. Then, compare the results of using HiperDispatch, re-adjust 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 zEC12.

z/VM V6R3

z/VM V6R3 also uses the HiperDispatch facility for improved processor efficiency by better use of processor cache to take advantage of the cache-rich processor design of the zEC12 and zBC12 systems. The supported processor limit of 32 remains unchanged.

8.5.6 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 only used 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. These are the systems in one or multiple sysplexes that are to be monitored. For more information about the capacity provisioning domain, see 9.8, "Nondisruptive upgrades" on page 378.

- ▶ Runtime systems. These are the systems where the Capacity Provisioning Manager is running, or to which the server can fail over after server or system failures.

8.5.7 Decimal floating point and z/OS XL C/C++ considerations

z/OS V1R13 with PTFs or higher is required to use the latest level (10) of the following two C/C++ compiler options:

- ▶ ARCHITECTURE: This option selects the minimum level of system architecture on which the program can run. Note that certain features provided by the compiler require a minimum architecture level. ARCH(10) uses instructions that are available on the zEC12.
- ▶ 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 System z ARCHITECTURE or TUNE option for C/C++ programs if the same applications run on both the zEC12 and on previous System z servers. However, if C/C++ applications will run only on zEC12 servers, use the latest ARCHITECTURE and TUNE options to ensure that the best performance possible is delivered through the latest instruction set additions.

8.5.8 IBM System z Advanced Workload Analysis Reporter (IBM 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 impacts to 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 drill-down and identification of message anomalies, which can facilitate faster problem resolution.

IBM zAware is ordered through specific features of zEC12, and requires z/OS V1R13 or higher with IBM zAware exploitation support to collect specific log stream data. It requires a correctly configured LPAR. For more information, see “The zAware-mode logical partition (LPAR)” on page 279.

To use the IBM zAware feature, complete the following tasks in z/OS:

- ▶ 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 firewall settings.
- ▶ Verify that each z/OS system meets the sysplex configuration and OPERLOG requirements for monitored clients of the IBM zAware virtual appliance.
- ▶ Configure the z/OS system logger to send data to the IBM zAware virtual appliance server.
- ▶ Prime the IBM zAware server with prior data from monitored clients.

8.6 Coupling facility and CFCC considerations

Coupling facility connectivity to a zEC12 is supported on the zEC12, zBC12, z196, z114, and z10. The LPAR running the Coupling Facility Control Code (CFCC) can be on any of the previously listed supported systems. For more information about CFCC requirements for supported systems, see Table 8-64 on page 325.

Consideration: Because coupling link connectivity to System z9 and previous systems is not supported, introduction of zEC12 into existing installations might require more planning. Also, consider the level of CFCC. For more information, see “Coupling link considerations” on page 177.

CFCC Level 18

The initial support of the CFCC on the zEC12 with driver level 12 is Level 18. CFCC Level 18 offers the following enhancements:

- ▶ Coupling channel reporting enhancements:
 - Enables RMF to differentiate various IFB link types and detect whether the CIB link is running in a “degraded” state
- ▶ Serviceability enhancements:
 - Additional structure control information in CF dumps
 - Enhanced CFCC tracing support
 - Enhanced Triggers for CF nondisruptive dumping
- ▶ Performance enhancements:
 - Dynamic structure size alter improvement
 - DB2 GBP cache bypass
 - Cache structure management

Important: Having more than 1,024 structures requires a new version of the CFRM CDS. In addition, all systems in the sysplex must be at z/OS V1R12 (or later), or have the coexistence and preconditioning PTFs installed. Falling back to a previous level without the coexistence PTF installed is not supported at sysplex IPL.

CFCC Level 19

CFCC level 19 is delivered on the zEC12 with driver level 15. CFCC Level 19 introduces the following enhancements:

- ▶ Performance improvements:
 - Introduces Coupling Thin Interrupts:
 - Improves the performance in share CF engines environments
 - Improves the response time of asynchronous CF requests
- ▶ Resiliency enhancements:
 - Flash Express support
 - Provides cost-effective standby capacity to help manage the potential overflow of WebSphere MQ shared queues

z/OS V1R10 is no longer supported.

zEC12 systems with CFCC Level 19 require z/OS V1R12 or later, and z/VM V5R4 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 latest CFCC level for zEC12 servers is CFCC Level 19, as shown in Table 8-64.

Table 8-64 System z CFCC code level considerations

System z	Code level
zEC12	CFCC Level 18 or CFCC Level 19
zBC12	CFCC Level 19
z196 and z114	CFCC Level 17
z10 EC or z10 BC	CFCC Level 15 or CFCC Level 16
z9 EC or z9 BC	CFCC Level 14 or later
z990 or z890	CFCC Level 13 or later

For more information about CFCC code levels, see the Parallel Sysplex website:

<http://www.ibm.com/systems/z/psocftable.html>

CF structure sizing changes are expected when upgrading from CFCC Level 17 (or earlier) to CFCC Level 18 or CFCC Level 19, as well as from CFCC Level 18 to CFCC Level 19. Review the CF LPAR size by using the available CFSizer tool:

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

Before the migration, you want to install the compatibility/coexistence PTFs. A planned outage is required when you upgrade the CF or CF LPAR to CFCC Level 18 or to CFCC Level 19.

Flash Express exploitation by CFCC

CFCC Level 19 supports Flash Express. Initial CF Flash exploitation is targeted for WebSphere MQ shared queue application structures. It is designed to help improve resilience while providing cost-effective standby capacity to help manage the potential overflow of WebSphere MQ shared queues. Structures can now be allocated with a combination of real memory and Storage Class Memory (SCM) provided by the Flash Express feature.

Flash memory in the CPC is assigned to a CF partition via hardware definition panels, just as it is assigned to the z/OS partitions. The CFRM policy definition permits 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 somewhat larger at initial allocation time to accommodate additional control objects needed to make use of 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 current CFCC Flash Express exploitation requirements are listed:

- ▶ CFCC Level 19 support
- ▶ z/OS support for z/OS V1R13 with PTFs and z/OS V2R1 with PTFs

No new level of WebSphere MQ is required.

8.7 MIDAW facility

The *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 significantly improve FICON performance for extended format data sets. Non-extended data sets can also benefit from MIDAW.
- ▶ MIDAW can improve channel utilization, and can significantly 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.

8.7.1 MIDAW technical description

An indirect address word (IDAW) is used to specify data addresses for I/O operations in a virtual environment¹⁰. 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 8-1 shows a single channel command word (CCW) to control 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.

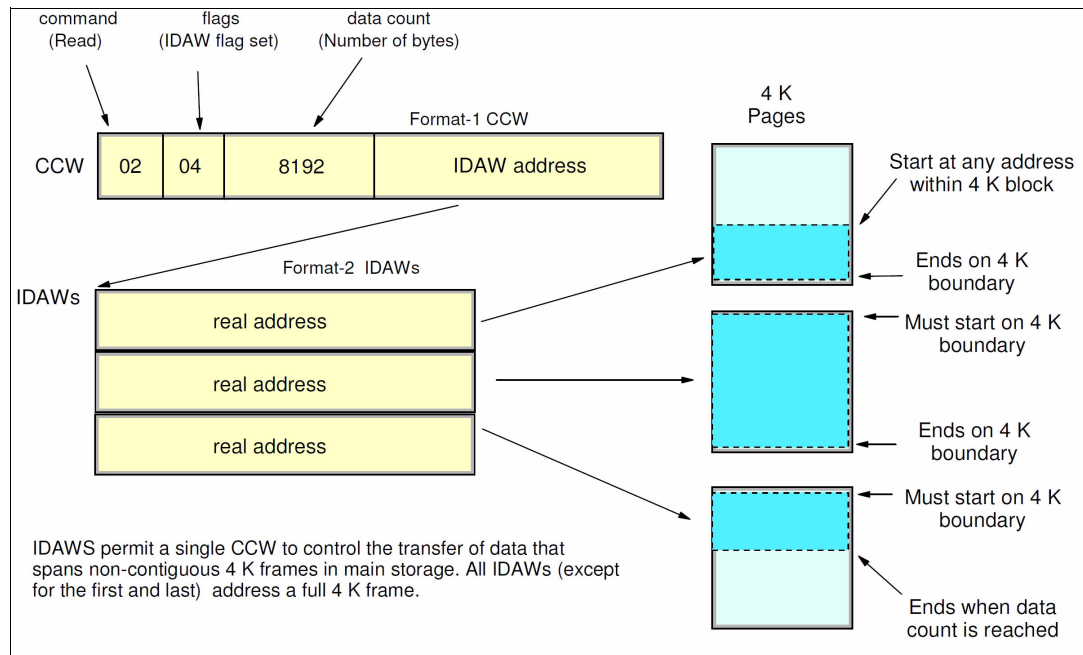


Figure 8-1 IDAW usage

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
- ▶ 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 are becoming 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 8-2. It is 16 bytes long and is aligned on a quadword.

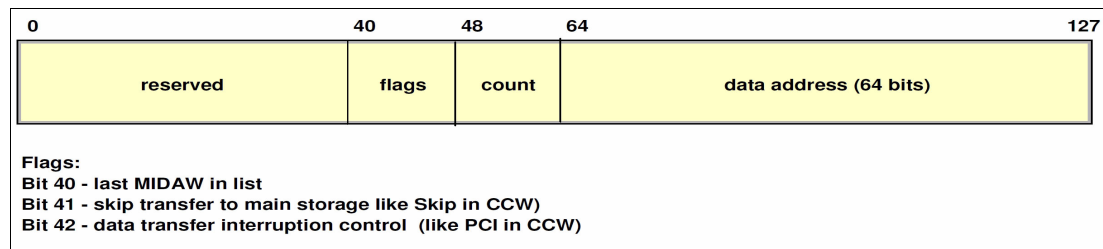


Figure 8-2 MIDAW format

An example of MIDAW usage is shown in Figure 8-3 on page 328.

¹⁰ There are exceptions to this statement, and a number of details are skipped in the description. This section assumes that you can merge this brief description with an existing understanding of I/O operations in a virtual memory environment.

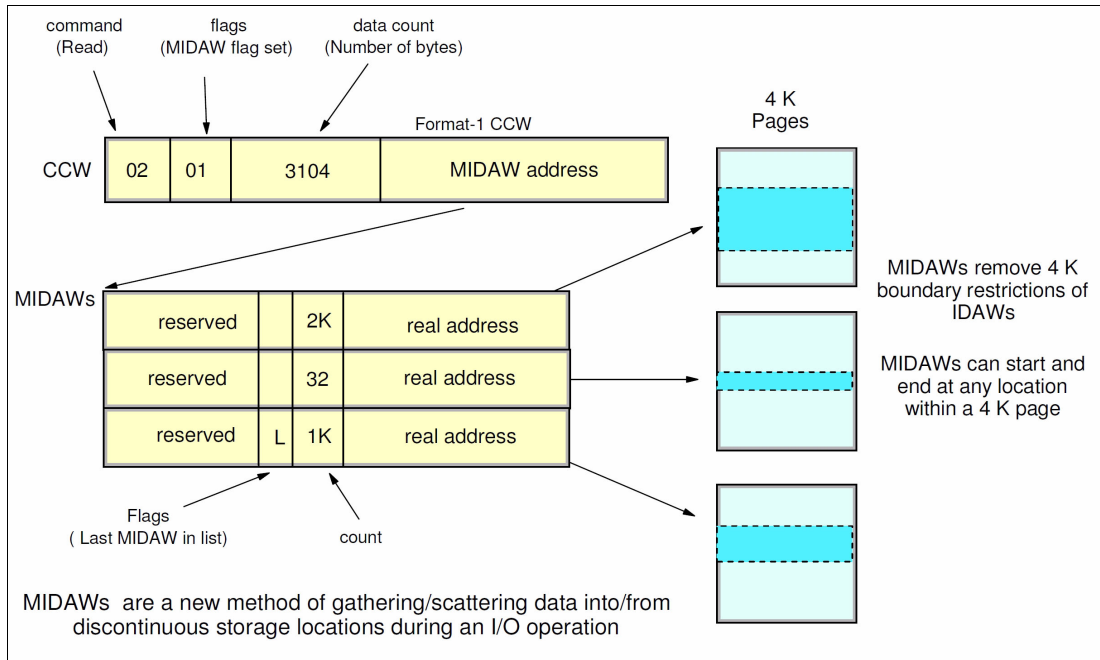


Figure 8-3 MIDAW usage

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 needs to 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 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.

8.7.2 Extended format data sets

z/OS extended format 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 extended format (EF) data sets, a brief review of the EF data sets is included here.

Both Virtual Storage Access Method (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 be able 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.

8.7.3 Performance benefits

z/OS Media Manager has I/O channel program support for implementing Extended Format 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, in the case of EF data sets, reduces the number of CCWs. 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, therefore using the channel resources more efficiently.

The MIDAW facility with FICON Express8S, operating at 8 Gbps, showed 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 in combination 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

8.8 IOCP

All System z 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)*.

The IOCS file contains detailed information for each channel and path assignment, each control unit, and each device in the configuration.

The required level of IOCP for the zEC12 is V3 R3 L0 (IOCP 3.3.0) or later with PTFs. For more information, see the *Input/Output Configuration Program User's Guide*, SB10-7037.

8.9 Worldwide port name (WWPN) tool

Part of the installation of your zEC12 system is the pre-planning of the SAN environment. IBM has a stand-alone tool to assist with this planning before the installation.

The capability of the *WWPN tool* is 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 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 Hardware Configuration Definition/Hardware Configuration Manager (HCD/HCM).

The WWPN tool on zEC12 (CHPID type FCP) requires the following levels:

- ▶ z/OS V1R11 with PTFs, or V1R12 and later
- ▶ z/VM V5R4 with PTFs, or V6R2 and later

The WWPN tool is applicable to all FICON channels defined as CHPID type FCP (for communication with SCSI devices) on zEC12. It is available for download at the Resource Link:

<http://www.ibm.com/servers/resourceLink/>

8.10 ICKDSF

Device Support Facilities, ICKDSF, Release 17 is required on all systems that share disk subsystems with a zEC12 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 zEC12.

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.

8.11 zEnterprise BladeCenter Extension (zBX) Model 003 software support

zBX Model 003 houses two types of blades: general purpose and solution specific.

8.11.1 IBM blades

IBM offers a selected subset of IBM POWER7 blades that can be installed and operated on the zBX Model 003. These blades have been thoroughly tested to ensure compatibility and manageability in the IBM zEnterprise EC12 environment.

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 hypervisor by using Kernel-based virtual machines, and supporting Linux on System x and Microsoft Windows operating systems.

Table 8-65 lists the operating systems that are supported by HX5 blades.

Table 8-65 Operating Support for zBX Model 003 HX5 Blades

Operating system	Support requirements
Linux on System x	Red Hat RHEL 5.5 and up, 6.0 and up SUSE SLES 10 (SP4) and up, SLES 11 (SP1) ^a and up
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)

a. Latest patch level required

b. 64 bit only

8.11.2 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 can help provide 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 8-66 lists the minimum support requirements for DataPower Sysplex Distributor support.

Table 8-66 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

8.12 Software licensing considerations

The IBM software portfolio for the zEC12 includes operating system software¹¹ (that is, z/OS, z/VM, z/VSE, and z/TPF) and middleware that runs on these operating systems. It also includes middleware for Linux on System z environments.

zBX software products are covered by the International Program License Agreement (IPLA) and other 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 POWER7 blades.

For the zEC12, 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 right 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 that is called *subscription and support* entitles you to access IBM product support during the support period. You also receive future releases and versions at no additional charge.

¹¹ Linux on System z distributions are not IBM products.

For more information, see the following references:

- ▶ The web page “Learn about Software licensing”, which has pointers to many documents:
http://www-01.ibm.com/software/lotus/passportadvantage/about_software_licensing.html
- ▶ The web page “Base license agreements” provides several documents:
<http://www-03.ibm.com/software/sla/sladb.nsf/viewbla>
- ▶ The IBM System z Software Pricing Reference Guide:
<http://www.ibm.com/systems/z/resources/swprice/reference/index.html>
- ▶ IBM System z Software Pricing web pages:
<http://www.ibm.com/systems/z/resources/swprice/mlc/index.html>
- ▶ The IBM International Passport Advantage Agreement can be downloaded from the “Learn about Software licensing” web page:
ftp://ftp.software.ibm.com/software/passportadvantage/PA_Agreements/PA_Agreement_International_English.pdf

The remainder of this section describes the software licensing options available on the zEC12.

8.12.1 MLC 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

There are various Workload License Charges (WLC) pricing structures that 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, IBM WebSphere MQ, and IBM Domino®. There are several pricing metrics that employ the following charge types:
 - Full-capacity:
The CPC’s total number of MSUs is used for charging. Full-capacity is applicable when your CPC is not eligible for subcapacity.
 - Subcapacity:
Software charges are based on the utilization of the LPARs 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 per CPC on the zEC12.

Subcapacity

For eligible programs, *subcapacity* allows software charges that are based on the measured utilization by LPARs instead of the CPC’s total number of MSUs. Subcapacity removes the dependency between software charges and CPC (hardware) installed capacity.

The subcapacity licensed products are charged monthly based on the highest observed 4-hour rolling average utilization of the LPARs in which the product runs. The exception is products that are licensed using the select application license charge (SALC) pricing metric. This pricing requires measuring the utilization and reporting it to IBM.

The LPAR's 4-hour rolling average utilization can be limited by a defined capacity value on the partition's image profile. This value activates the *soft capping* function of 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). However, it 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 LPARs running z/OS.

Even when using the soft capping option, the partition use 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 System z servers. Others apply to the aggregation of multiple zEC12 and System z servers' workloads within the same Parallel Sysplex.

For more information about WLC and how to combine the usage of LPARs, see *z/OS Planning for Workload License Charges*, SA22-7506:

http://www-03.ibm.com/systems/z/os/zos/bkserv/find_books.html

The following metrics are applicable to a stand-alone zEC12:

- ▶ Advanced Workload License Charges (AWLC)
- ▶ System z new application license charge (zNALC)
- ▶ Parallel Sysplex license charges (PSLC)

The following metrics are applicable to a zEC12 in an actively coupled Parallel Sysplex:

- ▶ AWLC when all CPCs are zEC12, zBC12, z196, or z114
 - Variable workload license charges (VWLCs) are only allowed under the AWLC Transition Charges for Sysplexes when not all CPCs are zEC12, zBC12, z196, or z114.
- ▶ zNALC
- ▶ PSLC

8.12.2 Advanced Workload License Charges (AWLC)

AWLCs were introduced with the IBM zEnterprise 196. They use the measuring and reporting mechanisms, as well as the existing MSU tiers, from VWLC.

Prices for tiers 4, 5, and 6 are different, allowing for lower costs for charges above 875 MSUs. AWLC offers improved price performance as compared to VWLC for all clients above 3 MSUs.

Similar to WLC, AWLC can be implemented in full-capacity or subcapacity mode. AWLC applies to z/OS and z/TPF, and their associated middleware products, such as DB2, IMS, CICS, WebSphere MQ, and IBM Domino, when running on a zEC12.

For more information, see the AWLC web page:

<http://www-03.ibm.com/systems/z/resources/swprice/mlc/awlc.html>

8.12.3 System z new application license charges (zNALC)

zNALC offers a reduced price for the z/OS operating system on LPARs that run a qualified new workload application, such as Java language business applications. These applications must run under WebSphere Application Server for z/OS, Domino, SAP, PeopleSoft, and Siebel.

z/OS with zNALC provides a strategic pricing model available on the full range of System z 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 web page:

<http://www-03.ibm.com/systems/z/resources/swprice/mlc/znalc.html>

8.12.4 Select application license charges (SALC)

SALC applies only to WebSphere MQ for System z. It allows a WLC client to license WebSphere MQ under product use rather than the subcapacity pricing provided under WLC.

WebSphere MQ is typically a low-usage product that runs pervasively throughout the environment. Clients that run WebSphere MQ at a low usage can benefit from SALC. Alternatively, you can still choose to license WebSphere MQ under the same metric as the z/OS software stack.

A reporting function, which IBM provides in the operating system IBM Software Usage Report Program, is used to calculate the daily MSU number. The rules to determine the billable SALC MSUs for WebSphere MQ use the following algorithm:

1. Determines the highest daily usage of a program family, which is the highest of 24 hourly measurements recorded each day. *Program* refers to all active versions of WebSphere MQ.
2. Determines the monthly usage of a program family, which is the fourth highest daily measurement that is recorded for a month.
3. Uses the highest monthly usage that is determined for the next billing period.

For more information about SALC, see the Other MLC Metrics web page:

<http://www.ibm.com/systems/z/resources/swprice/mlc/other.html>

8.12.5 Midrange Workload License Charges (MWLC)

MWLC applies to z/VSE V4 and later when running on zEC12, z196, and System z10 and z9 servers. The exceptions are the z10 BC and z9 BC servers at capacity setting A01, to which Entry Level License Charge (zELC) applies.

Similar to Workload License Charges, MWLC can be implemented in full-capacity or subcapacity mode. 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 instance, 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 SCRT report per server is required.

For more information, see the MWLC web page:

<http://www.ibm.com/systems/z/resources/swprice/mlc/mwlc.html>

8.12.6 Parallel Sysplex License Charges (PSLC)

Parallel Sysplex License Charges (PSLC) apply to a large range of mainframe servers. The list can be obtained at this website:

<http://www-03.ibm.com/systems/z/resources/swprice/reference/exhibits/hardware.html>

Although PSLC can be applied to stand-alone CPCs, the metric only provides aggregation benefits when applied to group of CPCs in an actively coupled Parallel Sysplex cluster according to IBM terms and conditions.

Aggregation allows charging a product 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 web page:

<http://www.ibm.com/systems/z/resources/swprice/mlc/pslc.html>

8.12.7 System z International Program License Agreement (IPLA)

For zEC12 and System z systems, the following types of products are generally in the IPLA category:

- ▶ Data management tools
- ▶ D2 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 Versions V5 and V6

Generally, three pricing metrics apply to IPLA products for zEC12 and System z:

- ▶ Value unit (VU):

Value unit pricing, which 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 purchased. Examples of eligible products include z/VM V5 and V6, and certain z/VM middleware. They are priced based on the number of engines.

- ▶ Processor value units (PVU):

In this metric, the number of engines is converted into processor value units under the Passport Advantage terms and conditions. Most Linux middleware is also priced based on the number of engines.

For more information, see the System z IPLA web page:

<http://www.ibm.com/systems/z/resources/swprice/zipla/index.html>

8.13 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/servers/eserver/zseries/zvse/support/preventive.html>

- ▶ z/TPF:

<http://www.ibm.com/software/htp/tpf/pages/maint.htm>

- ▶ Linux on System z:

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



System upgrades

This chapter provides an overview of IBM zEnterprise EC12 upgrade capabilities and procedures, with an emphasis on Capacity on Demand (CoD) offerings. The upgrade offerings to the zEC12 systems were developed from previous IBM System z systems. In response to client demands and changes in market requirements, many features were added. The provisioning environment gives you unprecedented flexibility and more control over cost and value.

For detailed tutorials about all aspects of system upgrades, use IBM Resource Link¹. Click **Resource Link** → **Client Initiated Upgrade Information**, and then select **Education**. Select your particular product from the list of available systems:

<https://www-304.ibm.com/servers/resourceLink/hom03010.nsf/pages/CIUInformation?OpenDocument>

The growth capabilities that are provided by the zEC12 include the following benefits:

- ▶ 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 impacts

This chapter includes the following sections:

- ▶ Upgrade types
- ▶ Concurrent upgrades
- ▶ Miscellaneous equipment specification (MES) upgrades
- ▶ Permanent upgrade through the CIU facility
- ▶ On/Off Capacity on Demand
- ▶ Capacity for Planned Event (CPE)
- ▶ Capacity Backup (CBU)
- ▶ Nondisruptive upgrades
- ▶ Summary of Capacity on Demand offerings

¹ Registration is required to access Resource Link.

9.1 Upgrade types

The types of upgrades for a zEC12 are summarized in this section.

9.1.1 Overview of upgrade types

Upgrades can be categorized as described in the following discussion.

Permanent and temporary upgrades

Permanent and temporary upgrades are different types of upgrades in different situations. For example, a growing workload might require more memory, more I/O cards, or more processor capacity. However, only a short-term upgrade might be necessary to handle a peak workload, or to temporarily replace a system that is down during a disaster or data center maintenance. The zEC12 offers the following solutions for such situations:

► Permanent:

- Miscellaneous equipment specification (MES):

The MES upgrade order is always performed by IBM personnel. The result can be either real hardware or installation of Licensed Internal Code Configuration Control (LICCC) to the system. In both cases, installation is performed by IBM personnel.

- Customer Initiated Upgrade (CIU):

Using the CIU facility for a system requires that the online CoD buying feature (FC 9900) is installed on the system. The CIU facility supports only LICCC upgrades.

► Temporary:

All temporary upgrades are LICCC-based. The one billable capacity offering is On/Off Capacity on Demand (On/Off CoD). The two replacement capacity offerings available are Capacity Backup (CBU) and Capacity for Planned Event (CPE).

For more information, see 9.1.2, “Terminology related to CoD for zEC12 systems” on page 341.

Tip: An MES provides system upgrades that can result in more enabled processors, a different central processor (CP) capacity level, and in additional books, memory, I/O drawers, and I/O features (physical upgrade). An MES can also upgrade the zEnterprise BladeCenter Extension. Additional planning tasks are required for nondisruptive logical upgrades. An MES is ordered through your IBM representative and installed by IBM service support representatives (SSRs).

Concurrent and nondisruptive upgrades

Depending on the impact on the system and application availability, upgrades can be classified in the following manner:

► Concurrent

In general, *concurrency* addresses the continuity of operations of the hardware part of an upgrade. For example, whether a system (hardware) is required to be switched off during the upgrade. For more information, see 9.2, “Concurrent upgrades” on page 344.

► Non-concurrent

This type of upgrade requires switching off the hardware that is being upgraded. Examples include model upgrades from any zEC12 model to the zEC12 HA1 model, and certain physical memory capacity upgrades.

► Disruptive

An upgrade is considered *disruptive* when resources modified or added to an operating system image require that the operating system be recycled to configure the newly added resources.

► Nondisruptive

Nondisruptive upgrades do not require the software or operating system to be restarted for the upgrade to take effect. Therefore, even concurrent upgrades can be disruptive to operating systems or programs that do not support the upgrades while being nondisruptive to others. For more information, see 9.8, “Nondisruptive upgrades” on page 378.

9.1.2 Terminology related to CoD for zEC12 systems

Table 9-1 lists the most frequently used terms that are related to Capacity on Demand for zEC12 systems.

Table 9-1 CoD terminology

Term	Description
Activated capacity	Capacity that is purchased and activated. Purchased capacity can be greater than the activated capacity.
Billable capacity	Capacity that helps handle workload peaks, either expected or unexpected. The one billable offering that is available is On/Off Capacity on Demand.
Capacity	Hardware resources (processor and memory) that are able to process the workload can be added to the system through various capacity offerings.
Capacity Backup (CBU)	Capacity Backup allows you to place model capacity or specialty engines in a backup system. CBU is used in an unforeseen loss of system capacity because of an emergency.
Capacity for planned event (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 Capacity on Demand.
Capacity levels	Can be full capacity or subcapacity. For the zEC12 system, capacity levels for the CP engine are 7, 6, 5, and 4: <ul style="list-style-type: none"> ► 1 - 99, A0, and A1 for capacity level 7nn. ► 1 - 20 for capacity levels 6yy and 5yy. ► 0 - 20 for capacity levels 4xx. An all Integrated Facility for Linux (IFL) or an all integrated catalog facility (ICF) system has a capacity level of 400.
Capacity setting	Derived from the capacity level and the number of processors. For the zEC12 system, the capacity levels are 7nn, 6yy, 5yy, and 4xx, where xx, yy, or nn indicates the number of active CPs. The number of processors can have the following ranges: <ul style="list-style-type: none"> ► 1 - 99, A0, and A1 for capacity level 7nn. ► 1 - 20 for capacity levels 6yy and 5yy. ► 0 - 20 for capacity levels 4xx. An all IFL or an all ICF system has a capacity level of 400.
Customer Initiated Upgrade (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.
Capacity on Demand (CoD)	The ability of a computing system to increase or decrease its performance capacity as needed to meet fluctuations in demand.
Capacity Provisioning Manager (CPM)	As a component of z/OS Capacity Provisioning, CPM monitors business-critical workloads that are running on z/OS on zEC12 systems.

Term	Description
Customer profile	This information is on Resource Link, and contains client and system information. A customer profile can contain information about more than one system.
Full capacity CP feature	For zEC12, feature (CP7) provides full capacity. Capacity settings 7nn are full capacity settings.
High water mark	Capacity that is purchased and owned by the client.
Installed record	The LICCC record is downloaded, staged to the Support Element (SE), and is installed on the CPC. A maximum of eight different records can be concurrently installed and active.
Model capacity identifier (MCI)	Shows the current active capacity on the system, including all replacement and billable capacity. For the zEC12, the model capacity identifier is in the form of 7nn, 6yy, 5yy, or 4xx, where xx, yy, or nn indicates the number of active CPs: <ul style="list-style-type: none"> ▶ nn can have a range of 01 - 99, A0, or A1. ▶ yy can have a range of 01 - 20. ▶ xx can have a range of 00 - 20. An all IFL or an all ICF system has a capacity level of 400.
Model Permanent Capacity Identifier (MPCI)	Keeps information about capacity settings active before any temporary capacity is activated.
Model Temporary Capacity Identifier (MTCI)	Reflects the permanent capacity with billable capacity only, without replacement capacity. If no billable temporary capacity is active, Model Temporary Capacity Identifier equals the MPCI.
On/Off Capacity on Demand (CoD)	Represents a function that allows 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 for handling 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 IBM zEnterprise BladeCenter Extension (zBX) Model 003 High Water Marks (HWM). <i>HWMs</i> refer to highest quantity of blade entitlements by blade type that the client has purchased. On z196 and z114, the HWMs are stored in the processor and memory LICCC record. On zEC12, the HWMs are stored in the Feature on Demand record.
Permanent capacity	The capacity that a client 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 system on a permanent basis.
Purchased capacity	Capacity that is delivered to and owned by the client. 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 <i>entitlement record</i> 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. This loss can be a planned event or an unexpected disaster. The two replacement offerings available are Capacity for Planned Events and Capacity Backup.
Resource Link	The IBM Resource Link is a technical support website that provides a comprehensive set of tools and resources. It is available from the IBM Systems technical support website: http://www.ibm.com/servers/resourceLink/
Secondary approval	An option, selected by the client, that requires second approver control for each Capacity on Demand order. When a secondary approval is required, the request is sent for approval or cancellation to the Resource Link secondary user ID.

Term	Description
Staged record	The point when a record that represents a capacity upgrade, either temporary or permanent, is retrieved and loaded on the Support Element (SE) disk.
Subcapacity	For the zEC12, CP features (CP4, CP5, and CP6) provide reduced capacity relative to the full capacity CP feature (CP7).
Temporary capacity	An optional capacity that is added to the current system capacity for a limited amount of time. It can be capacity that is owned or not owned by the client.
Vital product data (VPD)	Information that uniquely defines system, hardware, software, and microcode elements of a processing system.

9.1.3 Permanent upgrades

Permanent upgrades can be obtained by using these processes:

- ▶ Ordered through an IBM marketing representative
- ▶ Initiated by the client with the CIU on the IBM Resource Link

Tip: The use of the CIU facility for a system requires that the online CoD buying feature (FC 9900) is installed on the system. 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 accomplish these tasks:

- ▶ Add processor books
- ▶ Add Peripheral Component Interconnect Express (PCIe) drawers and features
- ▶ Add model capacity
- ▶ Add specialty engines
- ▶ Add memory
- ▶ Activate unassigned model capacity or IFLs
- ▶ Deactivate activated model capacity or IFLs
- ▶ Activate channels
- ▶ Activate cryptographic engines
- ▶ Change specialty engine (recharacterization)
- ▶ Add zBX and zBX features:
 - Chassis
 - Racks
 - DataPower blades
 - Entitlements

Considerations: Most of the MESs can be concurrently applied without disrupting the existing workload. For more information, see 9.2, “Concurrent upgrades” on page 344. However, certain MES changes are disruptive, such as model upgrades from any zEC12 model to the zEC12 HA1 model.

Memory upgrades that require dual in-line memory module (DIMM) changes can be made nondisruptively if there are multiple books and the flexible memory option is used.

Permanent upgrades initiated through CIU on the IBM Resource Link

Ordering a permanent upgrade by using the CIU application through Resource Link allows 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

9.1.4 Temporary upgrades

System zEC12 offers three types of temporary upgrades:

- ▶ On/Off Capacity on Demand (On/Off CoD):

This offering allows you to temporarily add more capacity or specialty engines to cover seasonal activities, period-end requirements, peaks in workload, or application testing. This temporary upgrade can be ordered only by using the CIU application through Resource Link.

- ▶ Capacity Backup (CBU):

This offering allows you to replace model capacity or specialty engines in a backup system used in an unforeseen loss of system capacity because of an emergency disaster.

- ▶ Capacity for Planned Event (CPE):

This offering allows you to replace model capacity or specialty engines because of 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 upgrade capacity changes can be billable or a replacement.

Billable capacity

To handle a peak workload, you can activate up to double the purchased capacity of any processor unit (PU) type temporarily. You are charged on a daily basis.

The one billable capacity offering is On/Off Capacity on Demand (On/Off CoD).

Replacement capacity

When a processing capacity is lost in another part of an enterprise, replacement capacity can be activated. It allows you to activate any PU type up to your authorized limit.

The following offerings are the two replacement capacity offerings:

- ▶ Capacity Backup
- ▶ Capacity for Planned Event

9.2 Concurrent upgrades

Concurrent upgrades on the zEC12 can provide more capacity with no system outage. In most cases, with prior planning and operating system support, a concurrent upgrade can also be nondisruptive to the operating system.

The benefits of the concurrent capacity growth capabilities that are provided by the zEC12 include, but are not limited to these benefits:

- ▶ Enabling 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* or *frozen* periods
- ▶ Enabling planned-downtime changes without affecting availability

This capability is based on the flexibility of the design and structure, which allows concurrent hardware installation and Licensed Internal Code (LIC) control over the configuration.

The subcapacity models allow more configuration granularity within the family. The added granularity is available for models that are configured with up to 20 CPs, and provides 60 additional capacity settings. Subcapacity models provide for CP capacity increase in two dimensions that can be used together to deliver configuration granularity. The first dimension is adding CPs to the configuration. The second is changing the capacity setting of the CPs currently installed to a higher model capacity identifier.

The zEC12 allows the concurrent and nondisruptive 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. This function is supported by z/OS, z/VM, and z/VSE. There are two ways to accomplish this addition:

- ▶ With planning ahead for the future need of extra processors. In the LPAR's profile, reserved processors can be specified. When the extra processors are installed, the number of active processors for that LPAR can be increased without the need for a partition reactivation and IPL.
- ▶ Another (easier) way is to enable the dynamic addition of processors through the z/OS LOADxx member. Set the parameter **DYNCPADD** in member LOADxx to ENABLE. The zEC12 supports dynamic processor addition in the same way that the z196 and z10 support it. The operating system must be z/OS V1R10 or higher.

Another function concerns the system assist processor (SAP). When more SAPs are concurrently added to the configuration, the SAP-to-channel affinity is dynamically remapped on all SAPs on the system to rebalance the I/O configuration.

9.2.1 Model upgrades

The zEC12 has a machine type and model, and model capacity identifiers:

- ▶ Machine type and model are 2827-Hvv

The vv can be 20, 43, 66, 89, or A1. The model number indicates how many PUs (vv) are available for client characterization (A1 stands for 101). Model H20 has one book installed; model H43 contains two books; model H66 contains three books; and models H89 and HA1 contain four books.

- ▶ Model capacity identifiers are 4xx, 5yy, 6yy, or 7nn

The xx is a range of 00 - 20²; yy is a range of 01 - 20; and nn is a range of 01 - 99, A0, and A1. A1 means 101. A zEC12 with 101 client usable processors is a zEC12 7A1. The model capacity identifier describes how many CPs are characterized (xx, yy, or nn) and the capacity setting (4, 5, 6, or 7) of the CPs.

² The zEC12 zero CP MCI is 400. This setting applies to an all IFL system or an all ICF system.

A hardware configuration upgrade always requires more physical hardware (books, cages, I/O drawers, or all of them³). A system upgrade can change either, or both, of the system model and the MCI.

Consider the following model upgrade information:

- ▶ LICCC upgrade:
 - Does not change the system model 2827-Hvv because more books are not added
 - Can change the model capacity identifier, the capacity setting, or both
- ▶ Hardware installation upgrade:
 - Can change the system model 2827-Hvv, if one or more books are added
 - Can change the model capacity identifier, the capacity setting, or both

The system model and the model capacity identifier can be concurrently changed. Concurrent upgrades can be performed for both permanent and temporary upgrades.

Tip: A model upgrade can be concurrent by using concurrent book add (CBA), except for upgrades to Model HA1.

Licensed Internal Code upgrades (MES ordered)

The LICCC provides for system upgrades without hardware changes by activation of additional (previously installed) unused capacity. Concurrent upgrades through LICCC can be performed for these resources:

- ▶ Processors (CPs, ICFs, System z Application Assist Processors (zAAPs), System z Integrated Information Processors (zIIPs), IFLs, and SAPs) if unused PUs are available on the installed books, or if the model capacity identifier for the CPs can be increased.
- ▶ Memory, when unused capacity is available on the installed memory cards. Plan-ahead memory and the flexible memory option are available to give you better control over future memory upgrades. For more information, see 2.5.6, “Flexible Memory Option” on page 51, and 2.5.7, “Pre-planned memory” on page 52.

Concurrent hardware installation upgrades (MES ordered)

Configuration upgrades can be concurrent when installing the following resources:

- ▶ Books (which contain processors, memory, and fanouts). Up to three books can be added concurrently on the model zEC12 H20.
- ▶ HCA and PCIe fanouts.
- ▶ I/O cards, when slots are still available on the installed PCIe I/O drawers.
- ▶ PCIe I/O drawers.
- ▶ All zBX and zBX features. However, the upgrade from a zBX Model 002 to a zBX Model 003 is disruptive.

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)

The zEC12 supports concurrent conversion between all PU types, which includes SAPs, to provide flexibility to meet changing business requirements.

³ I/O cage and the 8-slot I/O drawer cannot be ordered as an MES on zEC12. They are available as carry forward only.

Important: 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. Remove the assigned PU from the configuration.
2. Activate the newly available PU as the new PU type.

LPARs might also have to free the PUs to be converted. The operating systems must have support to configure processors offline or online so that the PU conversion can be done nondisruptively.

Considerations: Client 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 different types.
- ▶ It might require individual LPAR outages if dedicated PUs are converted.

Unassigned CP capacity is recorded by a model capacity identifier. CP feature conversions change (increase or decrease) the model capacity identifier.

9.2.2 Customer Initiated Upgrade (CIU) facility

The *CIU facility* is an IBM online system through which you can order, download, and install permanent and temporary upgrades for System z systems. Access to and use of the CIU facility requires a contract between the client and IBM, through which the terms and conditions for use of the CIU facility are accepted. The use of the CIU facility for a system requires that the online CoD buying feature code (FC 9900) is installed on the system. Although it can be installed on your zEC12 at any time, generally it is added when ordering a zEC12. 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 available through the Hardware Management Console (HMC), along with the Remote Support Facility (RSF). After all the prerequisites are met, the entire process, from ordering to activation of the upgrade, is performed by the client.

After download, the actual upgrade process is fully automated and does not require any onsite presence of IBM SSRs.

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. Additional books or I/O cards cannot be installed as part of an order that is placed through the CIU facility. The sum of CPs, unassigned CPs, ICFs, zAAPs, zIIPs, IFLs, and unassigned IFLs cannot exceed the client (characterized) PU count of the installed books. The total number of zAAPs or zIIPs cannot each exceed the number of purchased CPs.

CIU registration and contract for CIU

To use the CIU facility, a client must be registered and the system must be set up. After you complete the CIU registration, access to the CIU application is available through the IBM Resource Link website:

<http://www.ibm.com/servers/resourceLink/>

As part of the setup, you provide 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 allows 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 system for processors or memory. Requesting a client order approval to conform to your operational policies is possible. You can allow the definition of more IDs to be authorized to access the CIU. Additional 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, zAAPs, zIIPs, IFLs, and SAPs) and memory, or change the model capacity identifier. You can do so up to the limits of the installed books on an existing system.

Temporary upgrades

The base model zEC12 describes permanent and dormant capacity (Figure 9-1) using the capacity marker and the number of PU features installed on the system. Up to eight temporary offerings can be present. Each offering has its own policies and controls, and each 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.

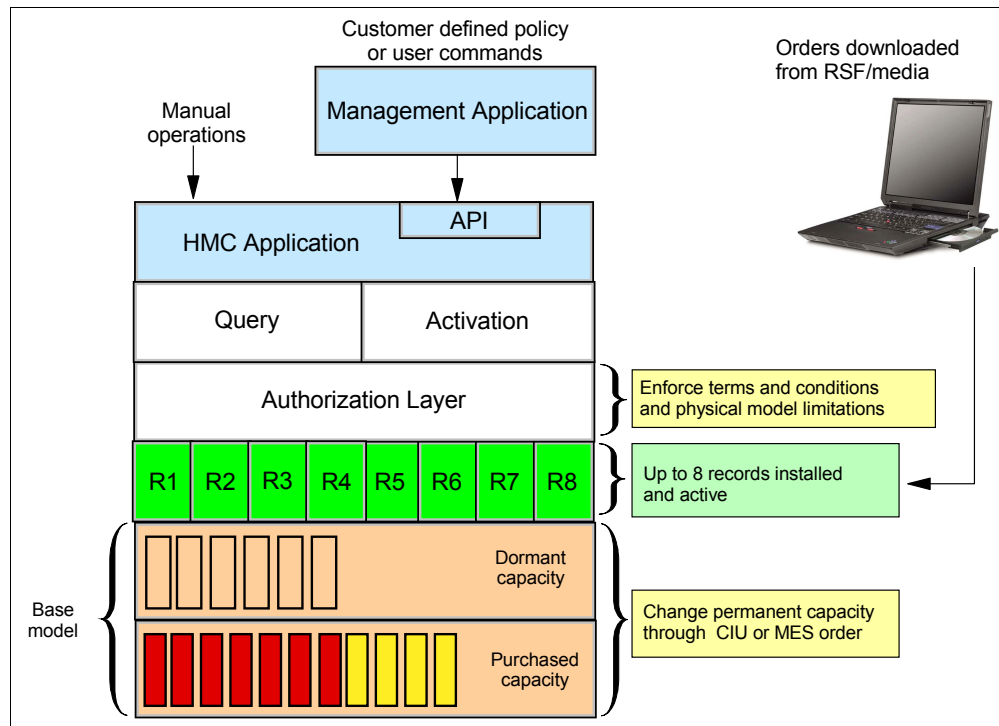


Figure 9-1 The provisioning architecture

Temporary upgrades are represented in the system by a *record*. All temporary upgrade records are resident on the Support Element (SE) hard disk drive. They can be downloaded from the RSF or installed from portable media. At the time of activation, you can control everything locally. Figure 9-1 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 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 or remove resources without having to go back to the base configuration. This process eliminates the need for temporary upgrade specification for all possible scenarios. However, for CPE, the 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 9-2 shows examples of the activation sequence of multiple temporary upgrades.

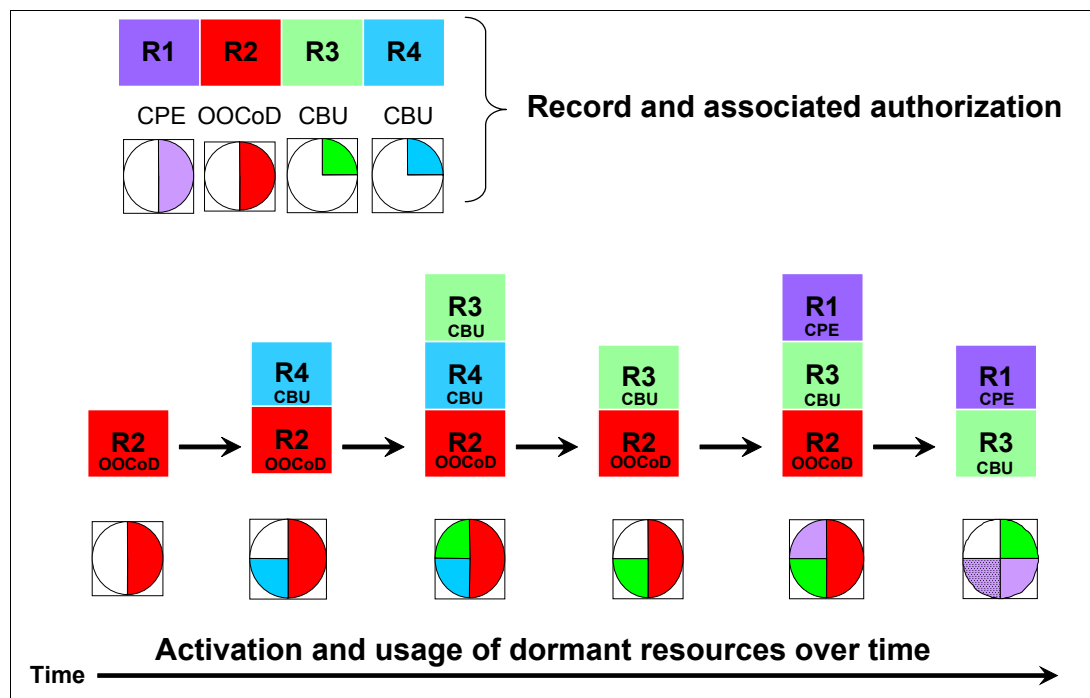


Figure 9-2 Example of temporary upgrade activation sequence

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 deactivated, the remaining parts of R1 can be activated as shown.

Temporary capacity can be billable as On/Off CoD, or replacement capacity as CBU or CPE:

- On/Off CoD is a function that enables *concurrent* and *temporary* capacity growth of the system.

On/Off CoD can be used for client 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 representative for exact details.

On/Off CoD can concurrently add processors (CPs, ICFs, zAAPs, zIIPs, IFLs, and SAPs), increase the model capacity identifier, or both. It can do so up to the limit of the installed books of an existing system, and is restricted to twice the currently installed capacity. On/Off CoD requires a contractual agreement between you and IBM.

You decide whether to either 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 more CPs, ICFs, zAAPs, zIIPs, IFLs, and SAPs, an increase of the model capacity identifier, or both.

CBU *cannot* be used for peak workload management in any form. As stated, On/Off CoD is the correct way to do that. 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 installed books of the backup system. They can be available as unused PUs, as a possibility to increase the model capacity identifier, or as both. A CBU contract must be in place before the special code that enables this capability can be loaded on the system. The standard CBU contract provides for five 10-day tests⁴ (the *CBU test activation*) and one 90-day activation over a five-year period. Contact your IBM representative for details.

You can run production workload on a CBU upgrade during a CBU test. At least an *equivalent amount* of production capacity must be shut down during the CBU test. If you already have existing CBU contracts, you also must sign an Amendment (US form #Z125-8145) with IBM to allow you to run production workload on a CBU upgrade during your CBU tests.

- ▶ CPE is a concurrent and temporary activation of additional CPs, ICFs, zAAPs, zIIPs, IFLs, and SAPs, an increase of the model capacity identifier, or both.

The CPE offering is used to replace temporary lost capacity within a client's enterprise for planned downtime events, for example, with data center changes. CPE cannot be used for peak load management of client workload or for a disaster situation.

The CPE feature requires unused capacity to be available on installed books of the backup system. The capacity must be available either as unused PUs, as a possibility to increase the model capacity identifier on a subcapacity system, or as both. A CPE contract must be in place before the special code that enables this capability can be loaded on the system. The standard CPE contract provides for one 3-day planned activation at a specific date. Contact your IBM representative for details.

9.2.3 Summary of concurrent upgrade functions

Table 9-2 on page 351 summarizes the possible concurrent upgrades combinations.

⁴ zEC12 provides more improvements in the CBU activation panels. These panels have been improved to prevent inadvertent CBU activation.

Table 9-2 Concurrent upgrade summary

Type	Name	Upgrade	Process
Permanent	MES	CPs, ICFs, zAAPs, zIIPs, IFLs, SAPs, book, memory, and I/Os	Installed by IBM SSRs
	Online permanent upgrade	CPs, ICFs, zAAPs, zIIPs, IFLs, SAPs, and memory	Performed through the CIU facility
Temporary	On/Off CoD	CPs, ICFs, zAAPs, zIIPs, IFLs, and SAPs	Performed through the OOCOD facility
	CBU	CPs, ICFs, zAAPs, zIIPs, IFLs, and SAPs	Performed through the CBU facility
	CPE	CPs, ICFs, zAAPs, zIIPs, IFLs, and SAPs	Performed through the CPE facility

9.3 Miscellaneous equipment specification (MES) upgrades

MES upgrades enable concurrent and permanent capacity growth. MES upgrades allow the concurrent adding of processors (CPs, ICFs, zAAPs, zIIPs, IFLs, and SAPs), memory capacity, and I/O ports, as well as hardware and entitlements to the zEnterprise BladeCenter Extension. For subcapacity models, MES upgrades allow the concurrent adjustment of both the number of processors and the capacity level. The MES upgrade can be performed by using LICCC only, by installing more books, by adding PCIe I/O drawers, by adding I/O⁵ features, or by 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 books
 - LICCC to adjust the number and types of PUs, to change the capacity setting, or both
 - Installing more books, and LICCC assigning and activating unassigned PUs on the installed books
- ▶ MES upgrades for memory are done by one of the following methods:
 - Using LICCC to activate more memory capacity up to the limit of the memory cards on the currently installed books. Plan-ahead and flexible memory features enable you to have better control over future memory upgrades. For more information about the memory features, see these descriptions:
 - 2.5.7, “Pre-planned memory” on page 52
 - 2.5.6, “Flexible Memory Option” on page 51
 - Installing more books and using LICCC to activate more memory capacity on installed books
 - Using the enhanced book availability (EBA), where possible, on multibook systems to add or change the memory cards
- ▶ MES upgrades for I/O⁵ are done by the following method:
 - Installing more I/O⁵ features and supporting infrastructure, if required, on PCIe drawers that are already installed, or installing more PCIe drawers to hold the new cards.
- ▶ MES upgrades for the zEnterprise BladeCenter Extension can be performed only through your IBM representative.

An MES upgrade requires IBM SSRs for the installation. In most cases, the time that is required for installing the LICCC and completing the upgrade is short.

⁵ Other adapter types, such as zFlash, zEOC, and Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE), also can be added to the PCIe I/O drawers via an MES.

To better use the MES upgrade function, carefully plan the initial configuration to allow a concurrent upgrade to a target configuration. The availability of PCIe I/O drawers improves 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 temporary upgrades. You can more easily resolve billing situations where independent software vendor (ISV) products are in use.

The model and model capacity identifiers that are returned by the STSI instruction are updated to coincide with the upgrade. For more information, see “Store System Information (STSI) instruction” on page 380.

Upgrades: The MES provides the physical upgrade, resulting in more enabled processors, different capacity settings for the CPs, and more memory, I/O ports, I/O adapters, and I/O drawers. Additional planning tasks are required for nondisruptive logical upgrades. For more information, see “Guidelines to avoid disruptive upgrades” on page 382.

9.3.1 MES upgrade for processors

An MES upgrade for processors can concurrently add CPs, ICFs, zAAPs, zIIPs, IFLs, and SAPs to a zEC12 by assigning available PUs on the books, through LICCC. Depending on the quantity of the additional processors in the upgrade, more books might be required and can be concurrently installed before the LICCC is enabled. With the subcapacity models, more capacity can be provided by adding CPs, by changing the capacity identifier on the current CPs, or by doing both.

Limits: The sum of CPs, inactive CPs, ICFs, zAAPs, zIIPs, IFLs, unassigned IFLs, and SAPs cannot exceed the maximum limit of PUs available for client use. The number of zAAPs and the number of zIIPs cannot each exceed the number of purchased CPs.

Figure 9-3 on page 353 is an example of an MES upgrade for processors, showing two upgrade steps.

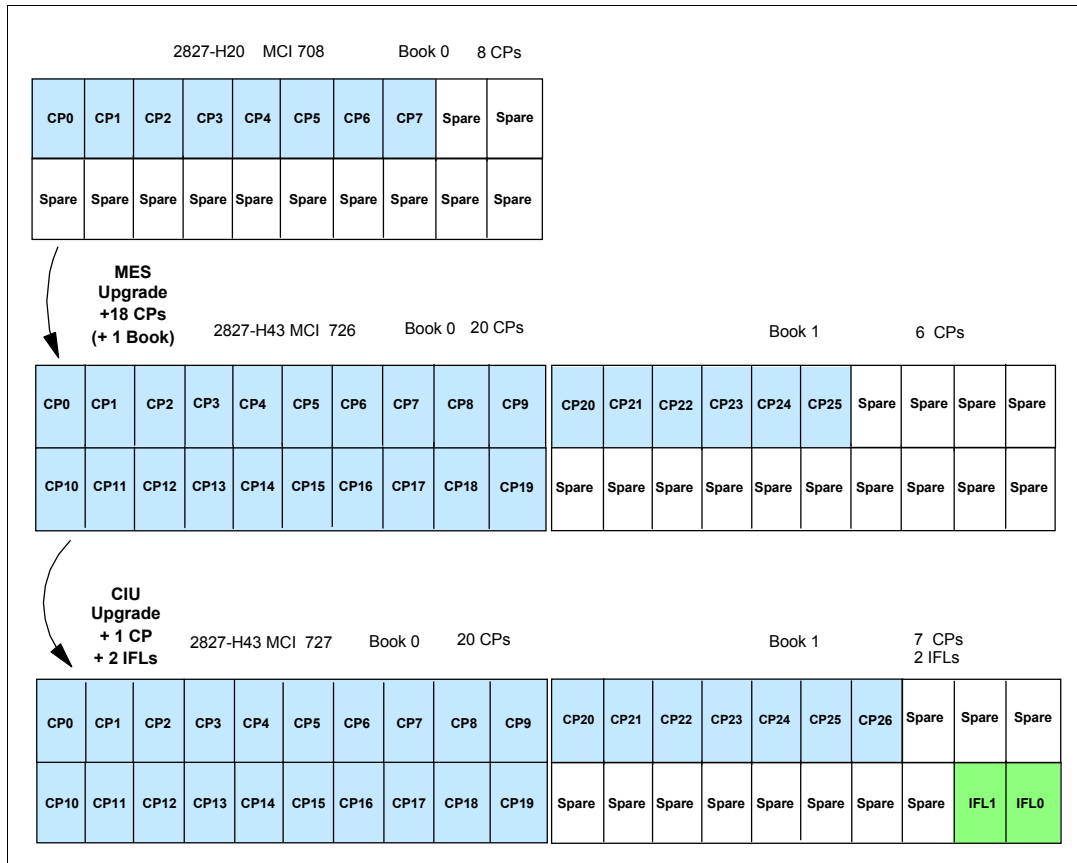


Figure 9-3 MES for processor example

A model H20 (one book), model capacity identifier 708 (eight CPs), is concurrently upgraded to a model H43 (two books), with model capacity identifier (MCI) 726 (26 CPs). The model upgrade requires adding a book and assigning and activating 18 PUs as CPs. Then, model H43, MCI 726, is concurrently upgraded to a capacity identifier 727 (27 CPs) with two IFLs. This process is done by assigning and activating three more unassigned PUs (one as CP and two as IFLs). If needed, more LPARs can be created concurrently to use the newly added processors.

The example in Figure 9-3 was used to show how the addition of PUs as CPs and IFLs, as well as the addition of a book, works. In reality, the addition of a book to a zEC12 Model H20 upgrades the machine model to H43. In addition, one of the two spare PUs on book 0 is moved over to book 1 to have one spare PU on each book. After the second book addition, book 0 has 21 configurable PUs and book 1 has 22 configurable PUs, which allow 43 PUs to be characterized on the new H43 model.

Restriction: Up to 101 logical processors, including reserved processors, can be defined to an LPAR. However, do not define more processors to an LPAR than the target operating system supports.

Table 9-3 on page 354 describes the number of processors that are supported by various z/OS and z/VM releases.

Table 9-3 Number of processors that are supported by the operating system

Operating System	Number of processors supported
z/OS V1R10 with PTFs	64
z/OS V1R11 with PTFs	100
z/OS V1R12 with PTFs	100
z/OS V1R13 with PTFs	100
z/VM V5R4 - z/VM V6R3	32
z/VSE	z/VSE Turbo Dispatcher can use up to 4 CPs, and tolerates up to 10-way LPARs
z/TPF	86 CPs
Linux on System z	SUSE SLES 10: 64 CPs or IFLs SUSE SLES 11: 64 CPs or IFLs Red Hat RHEL 5: 80 CPs or IFLs Red Hat RHEL 6: 80 CPs or IFLs

Software charges, which are based on the total capacity of the system on which the software is installed, are adjusted to the new capacity after the MES upgrade.

Software products that use Workload License Charges (WLC) might not be affected by the system upgrade. Their charges are based on partition utilization, not on the system total capacity. For more information about WLC, see 8.12, “Software licensing considerations” on page 332.

9.3.2 MES upgrades for memory

MES upgrades for memory can concurrently add more memory in the following ways:

- ▶ Enabling, through LICCC, more capacity up to the limit of the currently installed DIMM memory cards
- ▶ Concurrently installing more books and LICCC-enabling memory capacity on the new books

The Preplanned Memory Feature is available to allow better control over future memory upgrades. See 2.5.6, “Flexible Memory Option” on page 51 and 2.5.7, “Pre-planned memory” on page 52 for details about plan-ahead memory features.

If the zEC12 is a multiple-book configuration, you can use the EBA feature to remove a book and add DIMM memory cards. It can also be used to upgrade the already installed memory cards to a larger size, and you can then use LICCC to enable the additional memory. With correct planning, more memory can be added nondisruptively to z/OS partitions and z/VM partitions. If necessary, new LPARs can be created nondisruptively to use the newly added memory.

Concurrency: Upgrades requiring DIMM changes can be concurrent by using the EBA feature. Planning is required to see whether this is a viable option for your configuration. Using the flexible memory option and the Preplanned Memory Feature (FC 1996 for the 16-GB increment, or FC 1990 for the 32-GB increment) ensures that EBA can work with the least disruption.

The one-book model H20 has a minimum of 80 GB physical installed memory. The client addressable storage in this case is 32 GB. If you require more, an additional memory upgrade can install up to 704 GB of memory. It does so by changing the existing DIMM sizes and adding more DIMMs in all available slots in the book. You can also add memory by *concurrently* adding a second book with sufficient memory into the configuration and then using LICCC to enable that memory.

An LPAR can dynamically take advantage of a memory upgrade if reserved storage is 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 allows a z/OS operating system 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 requires that pertinent operating system parameters have been prepared. If reserved storage is not defined to the LPAR, the LPAR must be deactivated, the image profile changed, and the LPAR reactivated. This process allows the additional storage resources to be available to the operating system image.

9.3.3 MES upgrades for I/O

MES upgrades for I/O can concurrently add more I/O features by using one of the following methods:

- ▶ Installing more I/O features on an already installed PCIe I/O drawer.
- ▶ The installed PCIe I/O drawer must provide the number of I/O slots that are required by the target configuration.
- ▶ Adding a PCIe I/O drawer to hold the new I/O features.

Tip: Up to one I/O cage and up to two I/O drawers are supported if carried forward on an upgrade from a z196 or z10 EC.

For more information about I/O cages, I/O drawers, and PCIe I/O drawers, see 4.2, “I/O system overview” on page 130.

Table 9-4 gives an overview of the number of I/O cages, I/O drawers, and PCIe drawers that can be present in a zEC12.

Table 9-4 I/O cage and drawer summary

Description	New build	Carry forward	MES add
I/O cage	0	0 - 1	0
I/O drawer	0	0 - 2	0
PCIe I/O drawer	0 - 5	0 - 5	0 - 5

Table 9-5 on page 356 list the number of cards that can be on a carry forward.

Table 9-5 Number of I/O features and I/O cages and drawers on a carry forward

Number of features on a carry forward	Number of I/O drawers	Number of I/O cages
1 - 8	1	0
9 - 16	2	0
17 - 28	0	1
29 - 36	1	1
37 - 44	2	1

Restriction: The maximum number of original I/O features on a carry forward is 44.

Depending on the number of I/O features that are carried forward on an upgrade, the configurator determines the number and mix of I/O cages, I/O drawers, and PCIe I/O drawers.

To better use the MES for I/O capability, carefully plan the initial configuration to allow concurrent upgrades up to the target configuration. If original I/O features are removed from the I/O cage/drawer, the configurator does not physically remove the I/O cage/drawer unless the I/O frame slots are required to install a new PCIe I/O drawer.

If a PCIe I/O drawer is added to an existing zEC12 and original features must be physically moved to another I/O cage/drawer to empty the I/O cage/drawer for removal, original card moves are disruptive.

I/O cage removal is disruptive.

z/VSE, z/TPF, Linux on System z, and CFCC do *not* provide dynamic I/O configuration support. The installation of the new hardware is performed concurrently, but defining the new hardware to these operating systems requires an IPL.

Tip: The zEC12 has a hardware system area (HSA) of 32 GB. The z196 has a 16-GB HSA. It is *not* part of the client-purchased memory.

9.3.4 MES upgrades for the zBX

The MES upgrades for zBX can concurrently add blades if there are any slots available in the existing blade chassis, add chassis if there are any free spaces in existing racks, add racks up to a maximum of four, and add Feature on Demand (FoD) entitlements via LICCC, for connections to the zEC12.

Feature on Demand

FoD contains the zBX High Water Marks (HWM). HWMs refer to the highest quantities of blade entitlements by blade type that the client has purchased. On the z196/z114, the HWMs are stored in the processor and memory LICCC record. On the zEC12, the HWMs are in the Feature on Demand LICCC record.

The current zBX 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 servers' SE. Figure 9-4 on page 357 shows the panel for FoD Blades feature values shown under the Perform Model Conversion, Features on Demand Manage function.

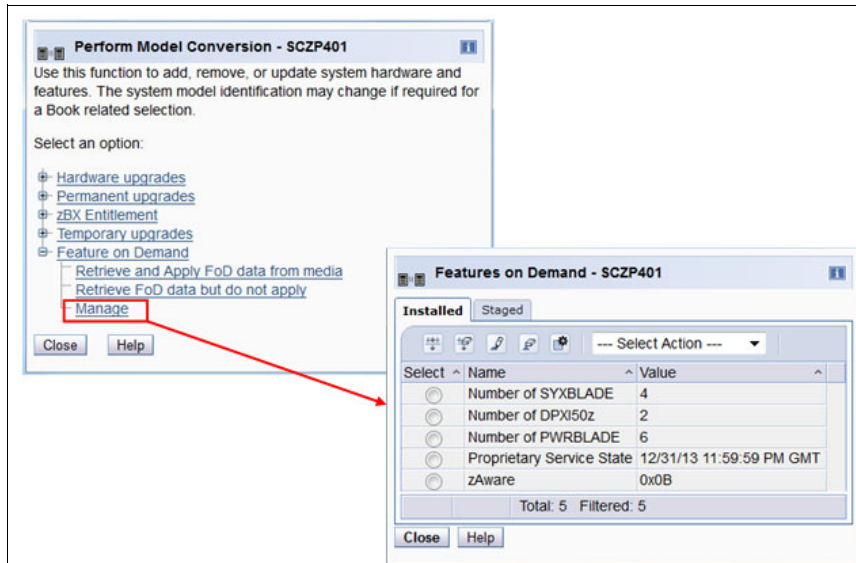


Figure 9-4 Features on Demand panel for zBX blade feature HWMs

There is only one FoD LICCC record installed or staged at any time in the system and its contents can be viewed under the Manage panel as shown in Figure 9-4. A staged record can be removed without installing it. An FoD record can only be installed completely. There is no selective feature or partial record installation. The features installed will be merged with the CEC LICCC after activation.

An FoD record can only be installed one time. 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 z196 with zBX-002 attached to a zEC12, the zBX Model 002 has to be upgraded to a zBX Model 003. The zBX has to be detached from z196 and attached to zEC12 during the system upgrade. Because the system upgrade is always disruptive, the zBX upgrade will be a disruptive task, as well.

If installing a new build zEC12 and planning to take over an existing zBX attached to a z196 or z114, the conversion of the zBX-002 to the zBX-003 can be done during the installation phase of the zBC12. Feature code 0030 must be ordered to detach the zBX from an existing z196 or z114. Feature code 0031 is required to re-attach the zBX to the zEC12.

If the model zBX Model 002 still has IBM Smart Analytics Optimizer blades installed, they need to be removed from the Model 002 prior to ordering the upgrade to a Model 003.

The zBX Model 003 adds the following features and functions:

- ▶ Broadband RSF support. The HMC application LIC for zEC12 and zBX Model 3 does not support dial modem use.
- ▶ Increased quantity of System x blade enablement to 56.
- ▶ Enables potential of 20-Gb Ethernet bandwidth by using link aggregation.
- ▶ Doubled 10 GbE cables between BladeCenter 10 GbE switch and 10 GbE Top of Rack (ToR) switch.
- ▶ Doubled 10 GbE cables between the BladeCenter 10 GbE switches.
- ▶ New version of the advanced management module (AMM) in the BladeCenter chassis.
- ▶ Upgraded hypervisors and other firmware changes.

9.3.5 Summary of plan-ahead features

A number of plan-ahead features exist for zEC12. The following list provides an overview of those features:

- ▶ Flexible memory

Flexible memory has no feature code (FC) associated with it. The purpose of flexible memory is to enable enhanced book availability. If a book is to be serviced, the flexible memory is activated to accommodate the storage of the book that is to be taken offline. After the repair action, the memory is taken offline again and is made unavailable for usage.

- ▶ Preplanned memory

Preplanned memory allows you to plan for nondisruptive memory upgrades. Any hardware required is pre-plugged, based on a target capacity that is specified in advance. Pre-plugged hardware is enabled by using an LICCC order when the additional memory capacity is needed. FC 1990 provides 32-GB preplanned memory, and FC 1996 provides 16-GB preplanned memory. FC 1901 is used to activate previously installed preplanned memory, and it can activate all the preinstalled memory or subsets of it.

- ▶ Balanced Power Plan Ahead

Balanced Power Plan Ahead is designed to anticipate future upgrade power needs on the zEC12. When more books are added to the system, the power consumption also rises. If necessary, one or more Bulk Power Regulators (BPRs) must be added. This process increases the time that is needed for the upgrade. When ordering this feature, regardless of the configuration, all six BPR pairs are installed and activated. Balanced Power Plan Ahead has FC 3003.

- ▶ Line Cord plan ahead

This option allows you to plan ahead for the second set of power cords. It is normally not configured until the addition of extra BPRs requires them. A plan-ahead option allows you to plan for a lengthy outage caused by installing circuit breakers or power feeds, or the routing of cables under the floor. The Line Cord plan-ahead option is FC 2000.

Tip: Accurate planning and the definition of the target configuration allow you to maximize the value of these plan-ahead features.

9.4 Permanent upgrade through the CIU facility

By using the CIU facility (through the IBM Resource Link on the web), you can initiate a permanent upgrade for CPs, ICFs, zAAPs, zIIPs, IFLs, SAPs, or memory. When performed through the CIU facility, you add the resources without having IBM personnel present at your location. You can also unassign previously purchased CPs and IFL processors through the CIU facility.

Adding permanent upgrades to a system through the CIU facility requires that the permanent upgrade enablement feature (FC 9898) is installed on the system. A permanent upgrade might change the system model capacity identifier (4xx, 5yy, 6yy, or 7nn) if more CPs are requested, or if the capacity identifier is changed as part of the permanent upgrade, but it cannot change the system model. If necessary, more LPARs can be created concurrently to use the newly added processors.

Consideration: A permanent upgrade of processors can provide a physical concurrent upgrade, resulting in more enabled processors available to a system configuration. Therefore, more planning and tasks are required for *nondisruptive* logical upgrades. For more information, see “Guidelines to avoid disruptive upgrades” on page 382.

Maintenance charges are automatically adjusted as a result of a permanent upgrade.

Software charges that are based on the total capacity of the system on which the software is installed are adjusted to the new capacity after the permanent upgrade is installed. Software products that use WLC might not be affected by the system upgrade because their charges are based on an LPAR utilization rather than system total capacity. For more information about WLC, see 8.12.2, “Advanced Workload License Charges (AWLC)” on page 334.

Figure 9-5 illustrates the CIU facility process on the IBM Resource Link.

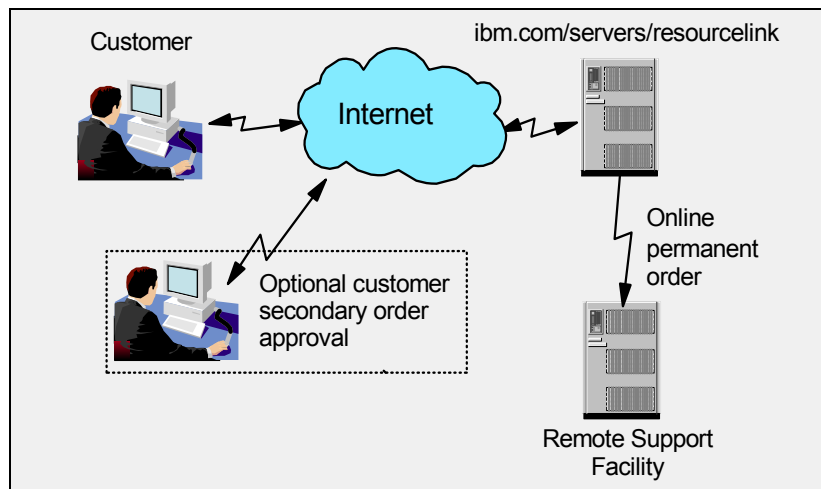


Figure 9-5 Permanent upgrade order example

The following sample sequence on the IBM Resource Link initiates an order:

1. Sign on to Resource Link.
2. Select **Customer Initiated Upgrade** from the main Resource Link page. Client and system details that are associated with the user ID are displayed.
3. Select the system to receive the upgrade. The current configuration (PU allocation and memory) is shown for the selected system.
4. Select **Order Permanent Upgrade**. The Resource Link limits the options to those that are valid or possible for the selected configuration (system).
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 9-6 illustrates the process for a permanent upgrade. When the LICCC is passed to the Remote Support Facility, you are notified through an email that the upgrade is ready to be downloaded.

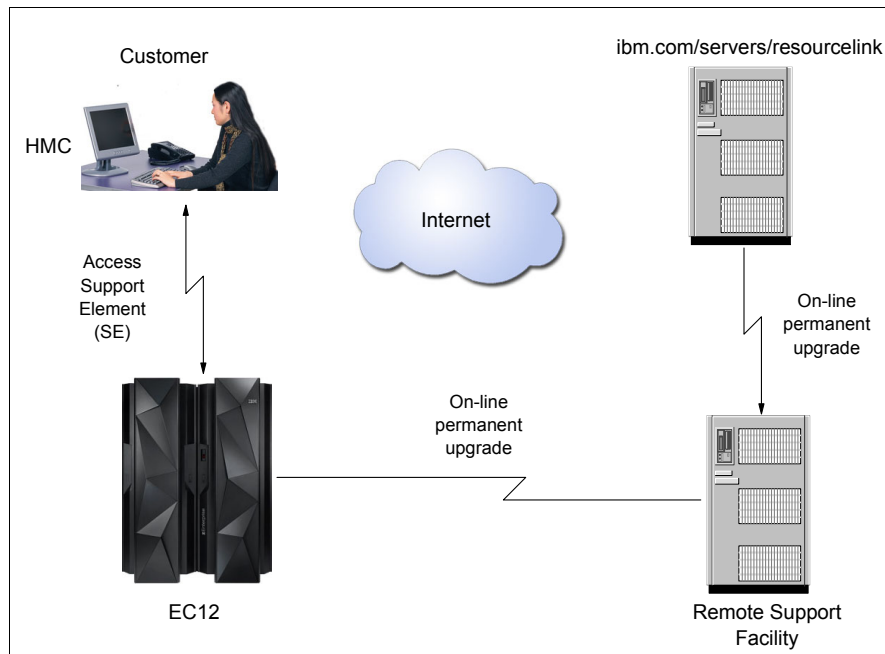


Figure 9-6 CIU-eligible order activation example

The two major components in the process are *ordering* and *retrieval* (along with activation).

9.4.1 Ordering

Resource Link provides the interface that enables you to order a concurrent upgrade for a system. 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 system. Warning messages are issued if you select invalid upgrade options. The process allows only one permanent CIU-eligible order for each system to be placed at a time. For a tutorial, see this website:

<https://www-304.ibm.com/servers/resourceLink/hom03010.nsf/pages/CIUInformation?OpenDocument>

Figure 9-7 shows the initial view of the Machine profile on Resource Link.

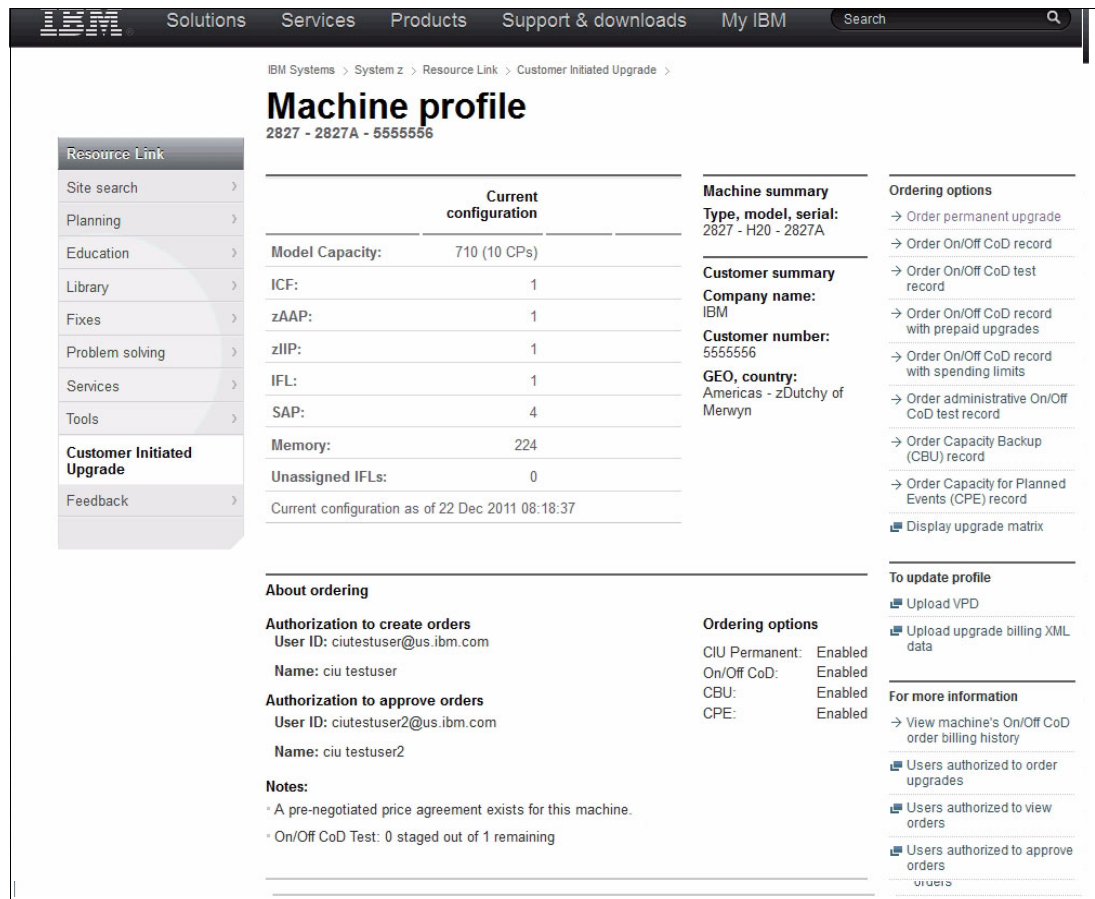


Figure 9-7 Machine profile window

The number of CPs, ICFs, zAAPs, zIIPs, IFLs, SAPs, memory size, 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 allows you to select only those upgrade options that are deemed valid by the order process. It allows upgrades only within the bounds of the currently installed hardware.

9.4.2 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 the Single Object Operation to the SE from a Hardware Management Console (HMC).

In the Perform Model Conversion window, select **Permanent upgrades** to start the process as shown in Figure 9-8.

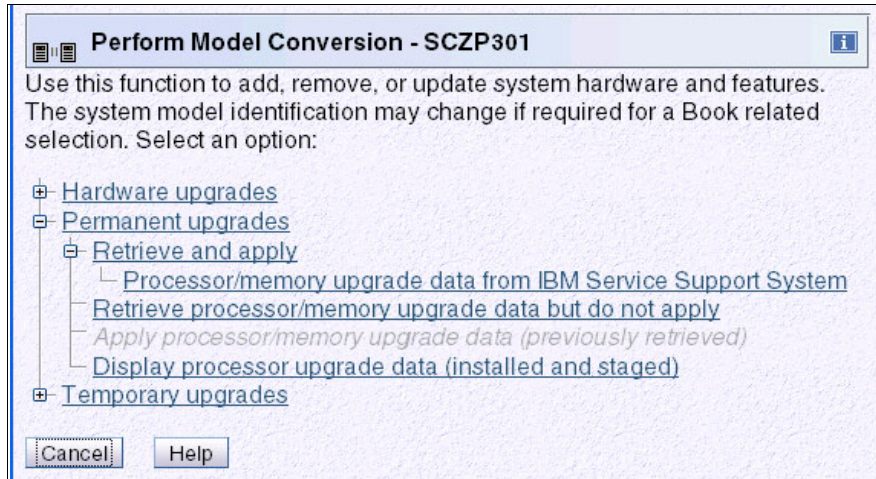


Figure 9-8 zEC12 Perform Model Conversion window

The window provides several possible options. If you select the **Retrieve and apply data** option, you are prompted to enter the order activation number to initiate the permanent upgrade. See Figure 9-9.

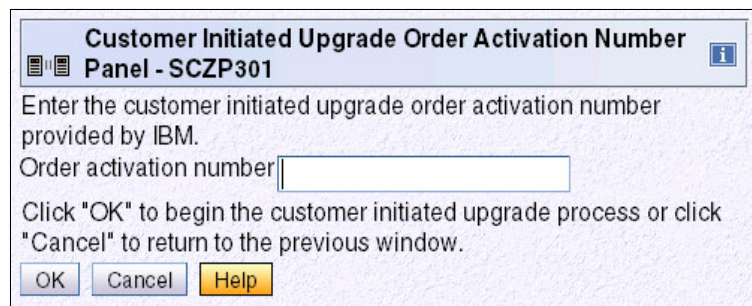


Figure 9-9 Customer Initiated Upgrade Order Activation Number Panel

9.5 On/Off Capacity on Demand

On/Off Capacity on Demand (On/Off CoD) allows you to temporarily enable PUs and unassigned IFLs available within the current hardware model. You can also use it to change capacity settings for CPs to help meet your peak workload requirements.

9.5.1 Overview

The capacity for CPs is expressed in millions of service units (MSUs). Capacity for speciality engines is expressed in 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 represent the following resource consumptions:

- ▶ For CP capacity, each token represents the amount of CP capacity that results 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*).

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, select how many tokens of each type to create for 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 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 billed are the highest resource usage inside each billing window for each capacity type. An activation period is one or more complete billing windows. The activation period is 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) is installed on the system that is to be upgraded.

The On/Off CoD to Permanent Upgrade Option is a new offering. It is an offshoot of On/Off CoD and takes advantage of the aspects of the architecture. You are given a window of opportunity to assess capacity additions to your permanent configurations by 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, you are charged for the temporary use.

The resources eligible for temporary use are CPs, ICFs, zAAPs, zIIPs, IFLs, and SAPs. The temporary addition of memory and I/O ports or adapters is not supported. Unassigned PUs that are on the installed books can be temporarily and concurrently activated as CPs, ICFs, zAAPs, zIIPs, IFLs, and SAPs through LICCC. You can assign PUs up to twice the currently installed CP capacity, and up to twice the number of ICFs, zAAPs, zIIPs, or IFLs. Therefore, an On/Off CoD upgrade cannot change the system model. The addition of new books is not supported. However, the activation of an On/Off CoD upgrade can increase the model capacity identifier (4xx, 5yy, 6yy, or 7nn).

9.5.2 Ordering

Concurrently installing temporary capacity by ordering On/Off CoD is possible in the following manner:

- ▶ 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
- ▶ zAAP features up to the number of installed zAAPs
- ▶ zIIP features up to the number of installed zIIPs
- ▶ SAPs up to four for model H20, eight for an H43, 12 for an H66, and 16 for an H89 and HA1

On/Off CoD can provide CP temporary capacity in two ways:

- ▶ By increasing the number of CPs.
- ▶ For subcapacity models, capacity can be added by increasing the number of CPs, 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 different from the installed CPs, the base capacity settings are changed to match.

On/Off CoD has the following limits that are associated with its use:

- The number of CPs cannot be reduced.
- The target configuration capacity is limited to these amounts:
 - Twice the currently installed capacity, expressed in MSUs for CPs.
 - Twice the number of installed IFLs, ICFs, zAAPs, and zIIPs. The number of SAPs that can be activated depends on the model. For more information, see 9.2.1, “Model upgrades” on page 345.

On/Off CoD can be ordered as prepaid or postpaid:

- ▶ A prepaid On/Off CoD offering record contains resource descriptions, MSUs, a number of speciality engines, and tokens that describe the total capacity that can be used. 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 allow the activation for an entire billing window, which is 24 hours. The resources remain active until you deactivate them or until one resource consumes 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, speciality engines, and can contain capacity tokens that denote 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. The record usually expires 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 allow the activation for an entire billing window (24 hours). The resources remain active until they are deactivated, until one of the resource tokens are consumed, or until the record expires. The record usually expires 180 days after its installation. If one capacity token type is consumed, resources from the entire record are deactivated.

As an example, for a zEC12 with capacity identifier 502, there are two ways to deliver a capacity upgrade through On/Off CoD:

- ▶ The first option is to add CPs of the same capacity setting. With this option, the model capacity identifier can be changed to a 503, adding one more CP to make it a 3-way. It can also be changed to a 504, which adds two CPs, making it a 4-way.
- ▶ The second option is to change to a different capacity level of the current CPs and change the model capacity identifier to a 602 or to a 702. The capacity level of the CPs is increased, but no additional CPs are added. The 502 can also be temporarily upgraded to a 603 as indicated in the table, increasing the capacity level and adding another processor. The capacity setting 420 does not have an upgrade path through On/Off CoD.

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 this website:

<https://www-304.ibm.com/servers/resourceLink/lib03060.nsf/pages/lsprindex>

The On/Off CoD hardware capacity is charged on a 24-hour basis. There is a grace period at the end of the On/Off CoD day. This grace period allows 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 zEC12 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 system 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 that period. If more capacity is added from an already active record that contains capacity tokens, the systems checks whether the resource 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, more LPARs can be activated concurrently to use the newly added processor resources.

Consideration: On/Off CoD provides a concurrent hardware upgrade, resulting in more enabled processors available to a system configuration. Additional planning tasks are required for nondisruptive upgrades. For more information, see “Guidelines to avoid disruptive upgrades” on page 382.

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 system. Later, you can concurrently install temporary capacity up to the limits in On/Off CoD and use it for up to 180 days. Monitoring occurs through the system call-home facility, and an invoice is generated if the capacity is enabled during the calendar month. You will be billed for use of temporary capacity until the system is returned to the original configuration. If the On/Off CoD support is no longer needed, the enablement code needs to be removed.

On/Off CoD orders can be pre-staged in Resource Link to allow multiple optional configurations. The pricing of the orders is done at the time you order them, and the pricing can vary from quarter to quarter. Staged orders can have different 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 was staged 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 Support Element with the same or different capacities, giving you greater flexibility to quickly enable needed temporary capacity. Each record is easily identified with descriptive names, and you can select from a list of records that can be activated.

Resource Link provides the interface to order a dynamic upgrade for a specific system. 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 system. After you complete the prerequisites, orders for the On/Off CoD can be placed. The order process uses the CIU facility on Resource Link.

You can order temporary capacity for CPs, ICFs, zAAPs, 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 different 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, you can use an *auto renewal* feature to eliminate the need for a manual replenishment of the On/Off CoD order. This feature is implemented in Resource Link, and you must also select this feature in the machine profile. See Figure 9-10 on page 366 for more details.

Order On/Off CoD record

Step 1 of 2: Configure the record

The On/Off CoD upgrade options on this order form are initialized to the maximum selections for upgrades that have prices set for this machine. Maximizing selections creates an On/Off CoD record that supports the widest possible range of On/Off CoD upgrades for the current machine configuration. Adjust the selections only if you want to change the type or range of On/Off CoD upgrades that can be activated with this record.

(*) indicates setting a replenishment due date is required to continue. Its initial setting is the maximum date allowed.

Replenishment due date: (mm/dd/yyyy) Renew automatically

Enable upgrades for up to:

Model capacity: more model capacity

ICF: more ICF engines

zAAP: more zAAP engines

zIIP: more zIIP engines

IFL: more IFL engines

SAP: more SAP engines

Machine summary

Type: 2827 H20
Model: 710
Serial number: 2827A

Current configuration

Model capacity: 10 CPs
ICF: 1
zAAP: 1
zIIP: 1
IFL: 1
SAP: 4
Available engines: 6

Supported upgrades

Default is to renew records automatically

Figure 9-10 Order On/Off CoD record window

9.5.3 On/Off CoD testing

Each On/Off CoD-enabled system is entitled to one no-charge 24-hour test. No IBM charges are assessed for the test, including no IBM charges 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.

This test can have a maximum duration of 24 hours, commencing upon 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 terminates at the end of the 24-hour period.

In addition, you can perform administrative testing. No additional capacity is added to the system, but you can test all the procedures and automation for the management of the On/Off CoD facility.

Figure 9-11 is an example of an On/Off CoD order on the Resource Link web page.

The screenshot shows the IBM Resource Link web interface for configuring an On/Off CoD order. The breadcrumb trail indicates the path: IBM Systems > System z > Resource Link > Customer Initiated Upgrade > Machine profiles > Machine 2827 - 2827A >. The main heading is 'Order On/Off CoD record' with the sub-heading 'Step 1 of 2: Configure the record'. A left-hand navigation menu includes options like Site search, Planning, Education, Library, Fixes, Problem solving, Services, Tools, Customer Initiated Upgrade, and Feedback. The main content area contains a descriptive paragraph about the upgrade options, a 'Replenishment due date' field set to 12/09/2012 with a 'Renew automatically' checkbox, and a section titled 'Enable upgrades for up to:' with dropdown menus for Model capacity (100%), ICF (1), zAAP (1), zIIP (1), IFL (1), and SAP (4). On the right, there are two summary tables: 'Machine summary' (Type: 2827 H20, Model: 710, Serial number: 2827A) and 'Current configuration' (Model capacity: 10 CPs, ICF: 1, zAAP: 1, zIIP: 1, IFL: 1, SAP: 4, Available engines: 6). Below these is a 'Supported upgrades' section with links to 'Show upgrades' and 'Show upgrade prices'. A 'Continue' button is located at the bottom of the configuration area.

Figure 9-11 On/Off CoD order example

The example order in Figure 9-11 is an On/Off CoD order for 100% more CP capacity, and for one ICF, one zAAP, one zIIP, and one SAP. The maximum number of CPs, ICFs, zAAPs, zIIPs, and IFLs is limited by the current number of available unused PUs of the installed books. The maximum number of SAPs is determined by the model number and the number of available PUs on the already installed books.

9.5.4 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 deactivate the temporary capacity manually. Deactivation is accomplished from the Support Element, 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 it. For example, you might have to configure offline any CPs that were added to the partition, deactivate 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 different 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 allows the activation of the On/Off CoD records. The activation is performed from the HMC, and requires specifying the order number. With this API, automation code can be used to send an activation command along with the order number to the HMC to enable the order.

9.5.5 Termination

A client is contractually obligated to terminate the On/Off CoD right-to-use feature when a transfer in asset ownership occurs. A client 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. Any time the CIU enablement feature is removed, the On/Off CoD right-to-use is simultaneously removed. Reactivating the right-to-use feature subjects the client to the terms and fees that apply at that time.

Upgrade capability during On/Off CoD

Upgrades involving physical hardware are supported while an On/Off CoD upgrade is active on a particular zEC12. LICCC-only upgrades can be ordered and retrieved from Resource Link, and can be applied while an On/Off CoD upgrade is active. LICCC-only memory upgrades can be retrieved and applied while an On/Off CoD upgrade is active.

Repair capability during On/Off CoD

If the zEC12 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 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) clients are billed at the MSU level represented by the combined permanent and temporary capacity. All PSLC products are billed at the peak MSUs enabled during the month, regardless of usage. Clients 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 used.

Results from the STSI instruction reflect the current permanent and temporary CPs. For more information, see “Store System Information (STSI) instruction” on page 380.

9.5.6 z/OS capacity provisioning

The zEC12 provisioning capability combined with Capacity Provisioning Manager (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 9-12.

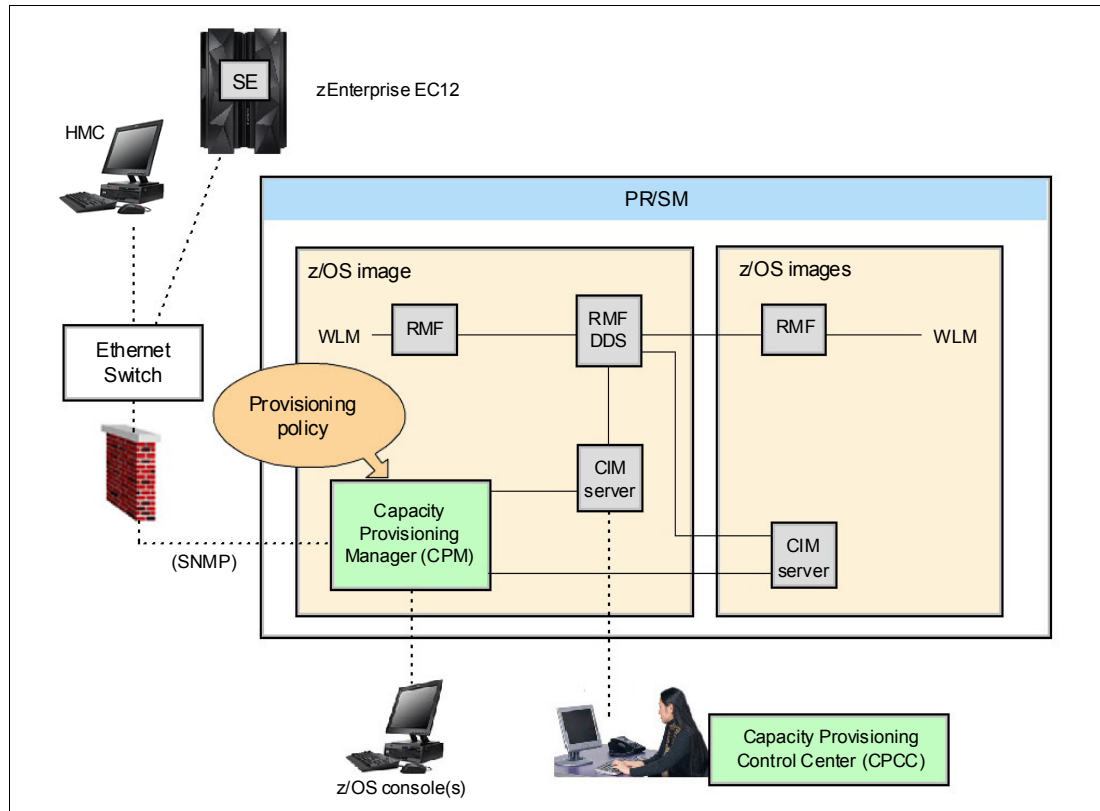


Figure 9-12 The capacity provisioning infrastructure

The z/OS WLM manages the workload by 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 Capacity Provisioning Manager (CPM), a function inside z/OS, retrieves critical metrics from one or more z/OS systems' CIM structures and protocol. CPM communicates to local or remote Support Elements 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 activates resources, a check is run 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 consumed during an activation that is driven by the CPM, the corresponding On/Off CoD record is deactivated prematurely by the system. This process occurs even if the CPM has activated this record, or parts of it. You do, however, receive warning messages if capacity tokens are getting close to being fully consumed. You receive the messages 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. You need to put operational procedures in place to handle these situations. You can either deactivate the record manually, allow it happen automatically, or replenish the specified capacity token by using the Resource Link application.

The Capacity Provisioning Control Center (CPCC), which is 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 are included in z/OSMF V1R13.

Capacity Provisioning Domain

The control over the provisioning infrastructure is run by the CPM through the *Capacity Provisioning Domain* (CPD) controlled by the *Capacity Provisioning Policy* (CPP). The Capacity Provisioning Domain is shown in Figure 9-13.

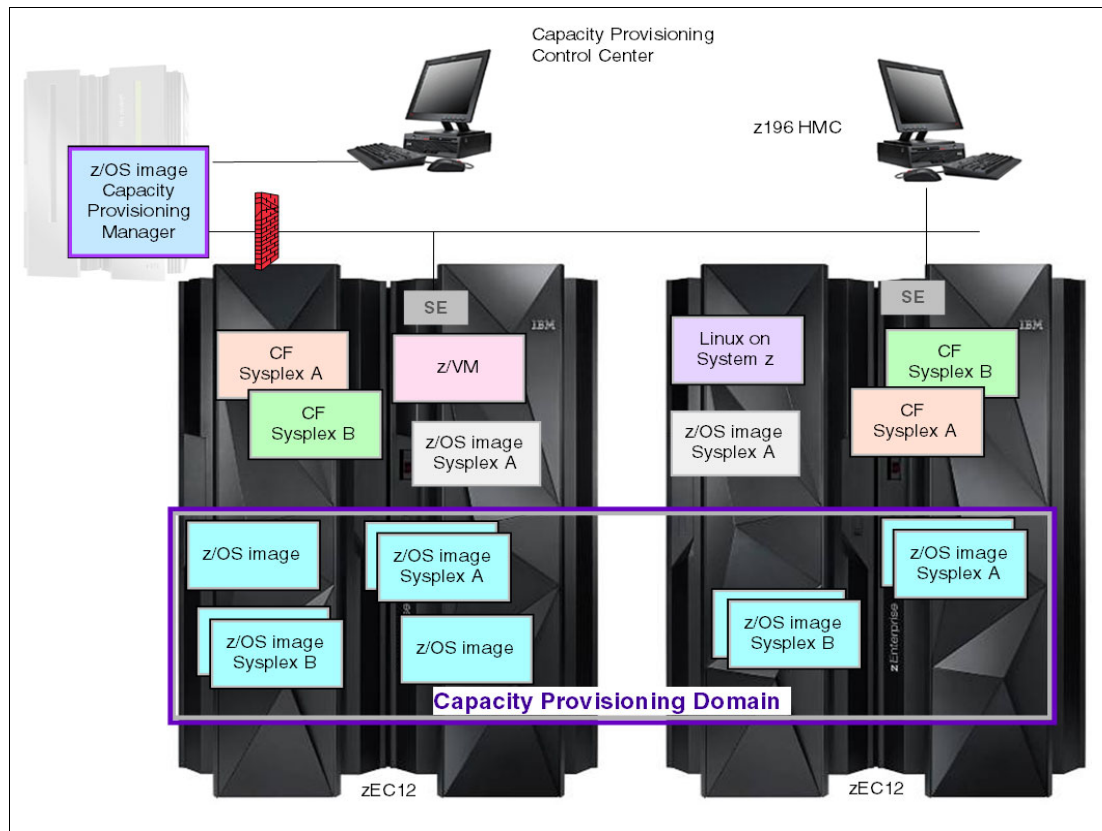


Figure 9-13 The Capacity Provisioning Domain

The Capacity Provisioning Domain (CPD) configuration defines the central processor complexes (CPCs) and z/OS systems that are controlled by an instance of the CPM. One or more CPCs, sysplexes, and z/OS systems can be defined into a domain. Sysplexes and CPCs do not have to be completely contained in a domain, but must not belong to more than one domain. Each domain has one active capacity provisioning policy. The CPCC is the CPM user interface component. Administrators work through this interface to define the domain configuration and provisioning policies. The CPCC is installed on a Microsoft Windows workstation.

CPM operates in four modes, allowing four different 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 policy criteria.
 - In analysis mode, the operator determines whether to ignore the information or to manually upgrade or downgrade the system by using the HMC, the 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 must 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 that contain information about workload and provisioning status, and the rationale for provisioning guidelines. User interfaces are provided through the z/OS console and the CPCC application.

The provisioning policy defines the circumstances under which more capacity can be provisioned (when, which, and how). There are three elements in the criteria:

- ▶ A time condition is when provisioning is allowed:
 - Start time indicates when provisioning can begin
 - Deadline indicates that provisioning of more capacity is no longer allowed
 - End time indicates that deactivation of more capacity needs to begin
- ▶ A workload condition is which work qualifies for provisioning. It can have these parameters:
 - The z/OS systems that can run eligible work
 - Importance filter indicates eligible service class periods, which are identified by WLM importance
 - Performance Index (PI) criteria:
 - Activation threshold: PI of service class periods must exceed the activation threshold for a specified duration before the work is considered to be suffering.
 - Deactivation threshold: PI of 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.

Tip: If no workload condition is specified, the full capacity described in the policy is activated and deactivated at the start and end times specified in the policy.

- ▶ Provisioning scope is how much more capacity can be activated, expressed in MSUs. The number of zAAPs and number of zIIPs must be one specification per CPC that is part of the Capacity Provisioning Domain. They are specified in MSUs. The maximum provisioning scope is the maximum additional capacity that can be activated for all the rules in the Capacity Provisioning Domain.

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 Capacity Provisioning Policy.

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 system. Changing from one to another requires stopping the active one 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, create only one On/Off CoD offering on the system by specifying the maximum allowable capacity. The CPM can then, when an activation is needed, activate a subset of the contents of the offering sufficient to satisfy the demand. If more capacity is needed later, the Provisioning Manager can activate more capacity up to the maximum allowed increase.

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.

Remember: The CPM has control over capacity tokens for the On/Off CoD records. In a situation where a capacity token is consumed, the system deactivates the corresponding offering record. Therefore, you must prepare routines for catching the warning messages about capacity tokens being consumed, and have administrative routines in place for such a situation. The messages from the system begin five days before a capacity token is fully consumed. To avoid capacity records being deactivated in this situation, replenish the necessary capacity tokens before they are consumed.

In a situation where a CBU offering is active on a system and that CBU offering is 100% or more of the base capacity, activating any On/Off CoD is not possible. This restriction is because the On/Off CoD offering is limited to 100% of the base configuration.

The Capacity Provisioning Manager operates based on Workload Manager (WLM) indications, and the construct that is used is the performance index (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 is the important one. The application might be defined as importance level 2 or 3, but might depend on other work that is running with importance level 1. Therefore, it is important to consider which workloads to control and which service class periods to specify.

9.6 Capacity for Planned Event (CPE)

CPE is offered with the zEC12 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 zEC12 in the client's environment.

Important: CPE is for planned replacement capacity only, and cannot be used for peak workload management.

CPE includes 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. The configuration consists of a model identifier, a number of speciality engines, or both. Based on the target configuration, a number of feature codes from the list are calculated automatically, and a CPE offering record is constructed.

CPE is intended to replace capacity that is 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 a maximum of three days.* Each CPE record, after it is activated, gives you access to dormant PUs on the system for which you have a contract as described by the feature codes. Processor units can be configured in any combination of CP or specialty engine types (zIIP, zAAP, SAP, IFL, and ICF). At the time of CPE activation, the contracted configuration is activated. *The general rule of one zIIP and one zAAP for each configured CP is enforced for the contracted configuration.*

The processors that can be activated by CPE come from the available unassigned PUs on any installed book. CPE features can be added to an existing zEC12 nondisruptively. *A one-time fee is applied for each CPE event.* This fee depends on the contracted configuration and its resulting feature codes. Only one CPE contract can be ordered at a time.

The base system configuration must have sufficient memory and channels to accommodate the potential requirements of the large CPE-configured system. Ensure that all required functions and resources are available on the system where CPE is activated. These functions and resources include CF LEVELs for coupling facility partitions, memory, and cryptographic functions, and include connectivity capabilities.

The CPE configuration is activated temporarily and provides more PUs in addition to the system's original, permanent configuration. The number of additional PUs is predetermined by the number and type of feature codes that are configured as described by the feature codes. The number of PUs that can be activated is limited by the unused capacity available on the system:

- ▶ A model H43 with 16 CPs, no IFLs, ICFs, or zAAPs, has 27 unassigned PUs available.
- ▶ A model H66 with 28 CPs, one IFL, and one ICF has 36 unassigned PUs available.

When the planned event is over, the system 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 on the system. CPE features can be added to an existing zEC12 nondisruptively.

9.7 Capacity Backup (CBU)

CBU provides reserved emergency backup processor capacity for unplanned situations in which capacity is lost in another part of your enterprise. It allows you to recover by adding the reserved capacity on a designated zEC12.

CBU is the quick, temporary activation of PUs and is available in these options:

- ▶ For up to 90 contiguous 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 or running the production workload. This option requires that an amount of System z workload capacity equivalent to the CBU upgrade capacity is shut down or otherwise made unusable during CBU test.⁶

Important: CBU is for disaster and recovery purposes only, and *cannot* be used for peak workload management or for a planned event.

9.7.1 Ordering

The CBU process allows for 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 PU that you require. Use the following feature codes:

- ▶ 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

⁶ All new CBU contract documents contain new CBU test terms to allow execution of production workload during CBU test. Existing CBU clients must run IBM client Agreement Amendment for IBM System z Capacity Backup Upgrade Tests (US form #Z125-8145).

- ▶ 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 the order. This date is dependent on the quantity of CBU years (FC 6817). You can extend your CBU entitlements through the purchase of additional CBU years. The number of FC 6817 per instance of FC 6818 remains limited to five. Fractional years are rounded up to the near whole integer when calculating this limit. If there are two years and eight months to the expiration date at the time of the order, the expiration date can be extended by no more than two years. One test activation is provided for each additional CBU year added to the CBU entitlement record.

FC 6805 allows for ordering more tests in increments of one. The total number of tests that is allowed is 15 for each FC 6818.

The processors that can be activated by CBU come from the available unassigned PUs on any installed book. The maximum number of CBU features that can be *ordered* is 101. The number of features that can be *activated* is limited by the number of unused PUs on the system:

- ▶ A model H20 with Capacity Model Identifier 410 can activate up to 20 CBU features: 10 to change the capacity setting of the existing CPs, and 10 to activate unused PUs.
- ▶ A model H43 with 15 CPs, four IFLs, and one ICF has 23 unused PUs available. It can *activate* up to 23 CBU features.

However, the ordering system allows for over-configuration in the order itself. You can *order* up to 101 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 additional capacity needed. If the amount of necessary CBU capacity is equal to four CPs, the CBU configuration is four CBU CPs.

The subcapacity models have multiple capacity settings of 4xx, 5yy, or 6yy. The standard models have the capacity setting 7nn. The number of CBU CPs must be equal to or greater than the number of CPs in the base configuration. All the CPs in the CBU configuration must have the same capacity setting. For example, if the base configuration is a 2-way 402, providing a CBU configuration of a 4-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 402 to a CBU configuration of a 4-way 504 requires four CBU feature codes with a capacity setting of 5yy.

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 zEC12 must have the correct number of books installed to allow the required upgrade. CBU can change the model capacity identifier to a *higher* value than the base setting (4xx, 5yy, or 6yy), but does not change the system 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 system. CBU features can be added to an existing zEC12 nondisruptively. For each system enabled for CBU, the authorization to use CBU is available for a 1 - 5 year period.

The alternate configuration is activated *temporarily*, and provides additional capacity greater than the system's original, *permanent* configuration. At activation time, determine the capacity that you require for that situation. You can decide to activate only a subset of the capacity that is specified in the CBU contract.

The base system configuration must have sufficient memory and channels to accommodate the potential requirements of the large CBU target system. Ensure that all required functions and resources are available on the backup systems. These include CF LEVELs for coupling facility partitions, memory, and cryptographic functions, as well as connectivity capabilities.

When the emergency is over (or the CBU test is complete), the system must be returned to its original configuration. The CBU features can be deactivated at any time before the expiration date. Failure to deactivate the CBU feature before the expiration date can cause the system to downgrade resources gracefully to the 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. This upgrade can result in more enabled processors, changed capacity settings available to a system configuration, or both. You can activate a subset of the CBU features ordered for the system. Therefore, more planning and tasks are required for *nondisruptive* logical upgrades. For more information, see "Guidelines to avoid disruptive upgrades" on page 382.

For more information, see the *System z Capacity on Demand User's Guide*, SC28-6846.

9.7.2 CBU activation and deactivation

The activation and deactivation of the CBU function is your responsibility and does not require the onsite presence of IBM SSRs. The CBU function is activated/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 CBU activation, the zEC12 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 LPAR weight can be changed or the number of logical processors can be concurrently increased by configuring reserved processors online. The operating system must be able to concurrently configure more processors online. If necessary, more 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 operating systems. In some cases, this process is a matter of varying the resources offline. In other cases, it can mean shutting down operating systems or deactivating LPARs. After the resources are released, the same facility on the SE is used to turn off CBU. To deactivate CBU, select the **Undo temporary upgrade** option 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 initiate a 10-day test period. A standard contract allows one test per CBU year. 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, using the same SE Perform a Model Conversion panel and then selecting the Temporary upgrades option. The HMC panels have been changed to avoid real CBU activations by setting the test activation as the default option.

The test CBU must be deactivated in the same way as the regular CBU. 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.

CBU example

An example of a capacity backup operation is 12 CBU features that are installed on a backup model H43 with model capacity identifier 708. When a production model H20 with model capacity identifier 708 has an unplanned outage, the backup system can be temporarily upgraded from model capacity identifier 708 to 720. This process allows the capacity to take over the workload from the failed production system.

Furthermore, you can configure systems to back up each other. For example, if you use two models of H20 model capacity identifier 705 for the production environment, each can have five or more features installed. If one system suffers an outage, the other one uses a temporary upgrade to recover the approximate original total capacity.

9.7.3 Automatic CBU enablement for GDPS

The IBM Geographically Dispersed Parallel Sysplex™ (GDPS) CBU enables automatic management of the PUs provided by the CBU feature in a system or site failure. Upon detection of a site failure or planned disaster test, GDPS concurrently add CPs to the systems in the take-over site to restore processing power for mission-critical production workloads. GDPS automation runs the following tasks:

- ▶ Runs the analysis that is required to determine the scope of the failure. This process minimizes operator intervention and the potential for errors.
- ▶ Automates 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 System z.

9.8 Nondisruptive upgrades

Continuous availability is an increasingly important requirement for most clients, 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 system can avoid system outages and are suitable to more operating system environments.

The zEC12 allows *concurrent* upgrades, which mean that dynamically adding more capacity to the system is possible. If the operating system images running on the upgraded system do not require disruptive tasks to use the new capacity, the upgrade is also *nondisruptive*. This process type means that power-on reset (POR), LPAR deactivation, and IPL do not have to occur.

If the concurrent upgrade is intended to satisfy an *image* upgrade to an LPAR, the operating system running in this partition must be able to concurrently configure more capacity online. z/OS operating systems have this capability. 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 operating system images, more LPARs can be created *concurrently* on the zEC12 system. These include all resources that are needed by such LPARs. These additional LPARs can be activated concurrently.

These enhanced configuration options are available through the separate HSA, which was introduced on the zEnterprise 196.

Linux operating systems, in general, cannot add more resources concurrently. However, Linux, and other types of virtual machines that run under z/VM, can benefit from the z/VM capability to nondisruptively 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 that are based on the Linux average load value. The daemon is available in Linux SLES 10 SP2 and up, as well as in Red Hat Enterprise Linux (RHEL) V5R4 and up.

9.8.1 Components

The following components can be added, depending on considerations that are described here.

Processors

CPs, ICFs, zAAPs, zIIPs, IFLs, and SAPs can be concurrently added to a zEC12 if unassigned PUs are available on any installed book. The number of zAAPs cannot exceed the number of CPs plus unassigned CPs. The same is true for the zIIPs. Additional books can also be installed concurrently, allowing further processor upgrades.

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 coupling facility LPARs by using the CFCC image operations window.

Memory

Memory can be concurrently added up to the physical installed memory limit. Additional books can also be installed concurrently, allowing further memory upgrades by LICCC, enabling memory capacity on the new books.

Using the previously defined reserved memory, z/OS operating system images, and z/VM partitions, you can dynamically configure more memory online. This process allows nondisruptive memory upgrades. Linux on System z supports Dynamic Storage Reconfiguration.

I/O

I/O features can be added concurrently if all the required infrastructure (I/O slots and HCAs) is present on the configuration. I/O drawers and PCIe I/O drawers can be added concurrently without planning if free space is available in one of the frames and the configuration permits.

Dynamic I/O configurations are supported by certain operating systems (z/OS and z/VM), allowing nondisruptive I/O upgrades. However, having dynamic I/O reconfiguration on a stand-alone coupling facility system is not possible because there is no operating system with this capability running on this system.

Cryptographic adapters

Crypto Express4S and Crypto Express3 features can be added concurrently if all the required infrastructure is in the configuration.

Special features

Special features, such as Flash Express, zEnterprise Data Compression (zEDC) Express, and RoCE, can also be added concurrently if all infrastructure is available in the configuration.

9.8.2 Concurrent upgrade considerations

By using an MES upgrade, On/Off CoD, CBU, or CPE, a zEC12 can be concurrently upgraded from one model 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, such as independent software vendor (ISV) products. Consider the effect on the software that is running on a zEC12 when you perform any of these configuration upgrades.

Processor identification

Two instructions are used to obtain processor information:

- ▶ Store System Information (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.
- ▶ Store CPU ID instruction (STIDP)
 - STIDP is provided for compatibility with an earlier version.

Store System Information (STSI) instruction

Figure 9-14 shows the relevant output from the STSI instruction. The STSI instruction returns the model capacity identifier for the permanent configuration and the model capacity identifier for any temporary capacity. This data is key to the functioning of Capacity on Demand offerings.

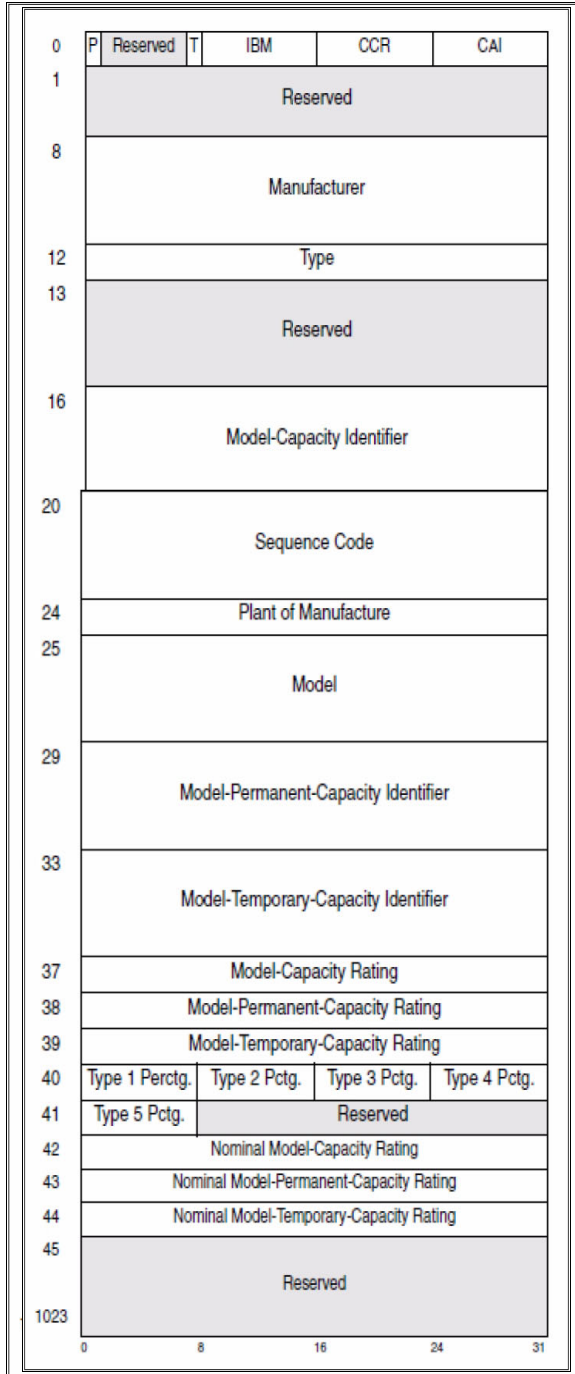


Figure 9-14 STSI output on zEC12

The model capacity identifier contains the base capacity, On/Off CoD, and CBU. The Model Permanent Capacity Identifier and the Model Permanent Capacity Rating contain the base capacity of the system. The Model Temporary Capacity Identifier and Model Temporary Capacity Rating contain the base capacity and On/Off CoD.

Store CPU ID (STIDP) instruction

The STIDP instruction provides information about the processor type, serial number, and LPAR identifier as shown in Table 9-6. The LPAR identifier field is a full byte to support more than 15 LPARs.

Table 9-6 STIDP output for zEC12

Description	Version code	CPU identification number		Machine type number	Logical partition 2-digit indicator
Bit position	0 - 7	8 - 15	16 - 31	32 - 48	48 - 63
Value	x'00' ^a	LPAR ID ^b	4-digit number that is derived from the CPC serial number	x'2827'	x'8000' ^c

a. The version code for zEC12 is x00.

b. The LPAR identifier is a two-digit number in the range of 00 - 3F. It is assigned by the user on the image profile through the SE 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 operating system that is 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 exactly as they were without running as a z/VM guest.
- ▶ If the **SET CPUID** command is 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 were without running as a z/VM guest.

Table 9-7 lists the possible output that is returned to the issuing program for an operating system that runs as a guest under z/VM.

Table 9-7 z/VM guest STIDP output for zEC12

Description	Version code	CPU identification number		Machine type number	Logical partition 2-digit indicator
Bit position	0 - 7	8 - 15	16 - 31	32 - 48	48 - 63
Without SET CPUID command	x'FF'	LPAR ID	4-digit number that is derived from the CPC serial number	x'2827'	x'8000'
With SET CPUID command	x'FF'	6-digit number as entered by the command SET CPUID = nnnnnn		x'2827'	x'8000'

Planning for nondisruptive upgrades

Online permanent upgrades, On/Off CoD, CBU, and CPE can be used to concurrently upgrade a zEC12. However, certain situations require a disruptive task to enable capacity that was recently added to the system. Some of these situations can be avoided if planning is done in advance. Planning ahead is a key factor for nondisruptive upgrades.

The following list describes the main reasons for disruptive upgrades. However, by carefully planning and reviewing “Guidelines to avoid disruptive upgrades” on page 382, you can minimize the need for these outages.

- ▶ LPAR memory upgrades when reserved storage was not previously defined are disruptive to image upgrades. z/OS and z/VM support this function.
- ▶ Installation of an I/O cage is disruptive.
- ▶ An I/O upgrade when the operating system cannot use the dynamic I/O configuration function is disruptive to that partition. Linux, z/VSE, z/TPF, and CFCC do not support dynamic I/O configuration.

Guidelines to avoid disruptive upgrades

Based on reasons for disruptive upgrades (“Planning for nondisruptive upgrades” on page 382), here are guidelines for avoiding or at least minimizing these situations, increasing the possibilities for nondisruptive upgrades:

- ▶ Using an SE function called Logical Processor add under Operational Customization tasks, CPs, zIIPs, and zAAPs can be added concurrently to a running partition. The CP, zIIP, and zAAP initial or reserved number of processors can be dynamically changed.
- ▶ The operating system 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 operating system. z/OS V1R11, z/OS V1R12, and z/OS V1R13 with PTFs support up to 100 processors. z/OS V2R1 also supports 100 processors. These processors include CPs, zAAPs, and zIIPs. z/VM supports up to 32 processors.

- ▶ Configure reserved storage to LPARs.

Configuring reserved storage for all LPARs before their activation enables them to be nondisruptively upgraded. The operating system that is running in the LPAR must be able to configure memory online. The amount of reserved storage can be above the book threshold limit, even if no other book is already installed. The current partition storage limit is 1 TB. z/OS and z/VM support this function.

- ▶ Consider the flexible and plan-ahead memory options.

Use a convenient entry point for memory capacity, and select memory options that allow future upgrades within the memory cards installed on the books. For more information about the offerings, see this information:

- 2.5.6, “Flexible Memory Option” on page 51
- 2.5.7, “Pre-planned memory” on page 52

Considerations when installing additional books

During an upgrade, more books can be installed concurrently. Depending on the number of additional books in the upgrade and your I/O configuration, a fanout rebalancing might be needed for availability reasons.

9.9 Summary of Capacity on Demand offerings

The capacity on-demand infrastructure and its offerings are major features that were introduced with the zEC12 system. These features are based on numerous client requirements for more flexibility, granularity, and better business control over the System z infrastructure, operationally and financially.

One major client requirement is to eliminate the need for a client authorization connection to the IBM Resource Link system when activating an offering. This requirement is being met by the z196 and zEC12. After the offerings are installed on the zEC12, they can be activated at any time, completely at the client's discretion. No intervention by IBM or IBM personnel is necessary. In addition, the activation of the Capacity Backup does not require a password.

The zEC12 can have up to eight offerings installed at the same time, with the limitation that only *one* of them can be an On/Off Capacity on Demand offering. The others can be any combination. 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. IBM applications, ISV programs, and client-written applications can control the usage of the offerings.

Resource consumption (and therefore financial exposure) can be controlled by using capacity tokens in the 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 workload. The CPM cannot control other offerings.

9.10 Reference

For more information, see the *IBM zEnterprise System Capacity on Demand User's Guide*, SC28-2605.



Reliability, availability, and serviceability

This chapter describes a few of the reliability, availability, and serviceability (RAS) features of the IBM zEnterprise EC12 (zEC12).

The zEC12 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, books, 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 cooling requirements.

The design of the memory on the zEC12 is implemented based on a fully redundant memory infrastructure, Redundant Array of Independent Memory (RAIM). This concept is similar to the RAID design used in external disk storage systems. RAIM was first introduced with the z196. The zEnterprise CPCs are the only systems in the industry that offer this level of memory design.

RAS also provides digitally signed delivery and transmission of microcode (LIC), fixes, and restoration/backup files. Any data that is transmitted to IBM Support is encrypted.

The design goal for the zEC12 is to remove all sources of planned outages.

This chapter includes the following sections:

- ▶ zEC12 availability characteristics
- ▶ zEC12 RAS functions
- ▶ zEC12 enhanced book availability (EBA)
- ▶ zEC12 enhanced driver maintenance (EDM)
- ▶ RAS capability for the HMC and SE
- ▶ RAS capability for zBX
- ▶ Considerations for PowerHA in zBX environment
- ▶ IBM System z Advanced Workload Analysis Reporter (IBM zAware)
- ▶ RAS capability for Flash Express

10.1 zEC12 availability characteristics

The following functions include availability characteristics on the zEC12:

- ▶ Enhanced book availability (EBA)

EBA is a *procedure* under which a book in a multibook system can be removed and reinstalled during an upgrade or repair action with no impact on the workload.

- ▶ Concurrent memory upgrade or replacement

Memory can be upgraded concurrently by using Licensed Internal Code Configuration Control (LICCC) if physical memory is available on the books. If the physical memory cards must be changed in a multibook configuration, requiring the book to be removed, the enhanced book availability function can be useful. It requires the availability of more resources on other books or reducing the need for resources during this action. To help ensure that the appropriate level of memory is available in a multiple-book configuration, select the flexible memory option. This option provides more resources to use EBA when repairing a book or memory on a book. They are also available when upgrading memory where larger memory cards might be required.

Memory can be upgraded concurrently by using LICCC if physical memory is available. The plan-ahead memory function available with the zEC12 allows you to plan for nondisruptive memory upgrades by having the system pre-plugged based on a target configuration. You can enable the pre-plugged memory by placing an order through LICCC.

- ▶ Enhanced driver maintenance (EDM)

One of the greatest contributors to downtime during planned outages is LIC driver updates that are performed in support of new features and functions. The zEC12 is designed to support the concurrent activation of a selected new driver level.

- ▶ IBM zAware

IBM System z Advanced Workload Analysis Reporter (IBM zAware) is an availability feature designed to use near real-time continuous learning algorithms, providing a diagnostics capability intended to help you quickly pinpoint problems, which in turn, can help you to more rapidly address service disruptions. zAware uses analytics to examine z/OS messages to find unusual patterns, inconsistencies, and variations. For more information about zAware, see 10.8, “IBM System z Advanced Workload Analysis Reporter (IBM zAware)” on page 405.

- ▶ Flash Express

Internal flash storage is spread over two Peripheral Component Interconnect Express (PCIe) adapters, which mirror to 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 10.9, “RAS capability for Flash Express” on page 406.

- ▶ Redundant IBM zEnterprise BladeCenter Extension (zBX) Model 003 configurations

Redundant hardware configurations within the 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. This redundancy allows 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 the zBX racks, ToRs, and BladeCenters.

- ▶ Plan Ahead for Balanced Power (FC 3003)

The *Plan Ahead for Balanced Power feature* allows 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 six BPR pairs will be shipped, installed, and activated.

Note: When this feature is ordered, a corequisite feature, the Plan Ahead for Line Cords feature (FC 1901) is automatically selected.

- ▶ Concurrent fanout addition or replacement

A PCIe, host channel adapter (HCA), or Memory Bus Adapter (MBA) fanout card provides the path for data between memory and I/O through InfiniBand (IFB) cables or PCIe cables. With the zEC12, a hot-pluggable and concurrently upgradeable fanout card is available. Up to eight fanout cards per book are available if the zEC12 has one or two books installed (models H20 and H43); therefore, a one book server has eight fanouts and two books, totaling 16. In a three-book machine (model H66), the maximum number of fanouts is reduced in two of the books to provide additional cooling. A total of 20 fanouts is supported: eight on the first book, and then six on each of the two remaining books. Finally, a four-book model (H89 or H101) supports only six fanouts per book, totaling 24 fanouts per server. During an outage, a fanout card that is used for I/O can be concurrently repaired while redundant I/O interconnect ensures that no I/O connectivity is lost.

- ▶ Redundant I/O interconnect

Redundant I/O interconnect helps maintain critical connections to devices. The zEC12 allows a single book, in a multibook system, to be concurrently removed and reinstalled during an upgrade or repair. Connectivity to the system I/O resources is maintained through a second path from a different book.

- ▶ Dynamic oscillator switch-over

The zEC12 has 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.

- ▶ Processor unit (PU) sparing

The zEC12 has two spare PUs to maintain performance levels if an active central processor (CP), Internal Coupling Facility (ICF), Integrated Facility for Linux (IFL), System z Application Assist Processor (zAAP), System z Integrated Information Processor (zIIP), integrated firmware processor (IFP), or System Assist Processor (SAP) fails. Transparent sparing for failed processors is supported. There are two spare PUs per system and sparing is supported across the books in the unlikely event that the book with the failure has no spares available.

- ▶ Application preservation

Application preservation is used where a CP fails and there are no spares left. The state of the failing CP is passed to another active CP, where the operating system uses it to successfully resume the task, in most cases without client intervention.

- ▶ Cooling improvements

The zEC12 comes with a newly designed radiator cooling system to replace the modular refrigeration unit (MRU). The radiator cooling system can support all four books simultaneously with a redundant design that consists of two pumps and two blowers. One active pump and blower can support the entire system load. The replacement of pump or blower is concurrent with no performance impact.

A water cooling system is also an option in zEC12, with water cooling unit (WCU) technology. Two redundant WCUs run with two independent chilled water feeds. Like the radiator cooling system, one WCU and one water feed can support the entire system load. Both radiator and water cooling systems are backed up by an air cooling system in the rare event of a cooling system problem.

10.2 zEC12 RAS functions

Hardware RAS function improvements focus on addressing all sources of outages. Sources of outages have these classifications:

- Unscheduled** This outage occurs because of an unrecoverable malfunction in a hardware component of the system.
- Scheduled** This outage is caused by changes or updates that must be done to the system 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 system. A planned outage can be caused by a capacity upgrade or a driver upgrade. A planned outage is usually requested by the client, and often requires pre-planning. The zEC12 design phase focuses on enhancing planning to simplify or eliminate planned outages.

The difference between scheduled outages and planned outages is, perhaps, not obvious. The general consensus is that scheduled outages are considered to take place somewhere soon. The time frame is approximately two weeks. Planned outages are outages that are planned well in advance and go beyond this approximate 2-week time frame. This chapter does not distinguish between scheduled and planned outages.

Preventing unscheduled, scheduled, and planned outages has been addressed by the IBM System z[®] system design for many years.

The zEC12 introduces a fixed size hardware system area (HSA) of 32 GB. This size helps eliminate pre-planning requirements for HSA and provides flexibility to dynamically update the configuration.

You can perform the following tasks dynamically¹:

- ▶ Add a logical partition (LPAR)
- ▶ Add a logical channel subsystem (LCSS)
- ▶ Add a subchannel set
- ▶ Add a logical CP to an LPAR
- ▶ Add a cryptographic coprocessor
- ▶ Remove a cryptographic coprocessor
- ▶ Enable I/O connections
- ▶ Swap processor types
- ▶ Add memory
- ▶ Add a physical processor

¹ Some pre-planning considerations might exist. For more information, see Chapter 9, “System upgrades” on page 339.

In addition, by addressing the elimination of planned outages, the following tasks are also possible:

- ▶ Concurrent driver upgrades
- ▶ Concurrent and flexible customer-initiated upgrades

For more information about the flexible upgrades that are initiated by clients, see 9.2.2, “Customer Initiated Upgrade (CIU) facility” on page 347.

10.2.1 Scheduled outages

Concurrent hardware upgrades, concurrent parts replacement, concurrent driver upgrades, and concurrent firmware fixes, available with the zEC12, all address elimination of scheduled outages. Furthermore, the following indicators and functions that address scheduled outages are included:

- ▶ Double memory data bus lane sparing
This feature reduces the number of repair actions for memory.
- ▶ Single memory clock sparing
- ▶ Double DRAM chipkill tolerance
- ▶ Field repair of the cache fabric bus
- ▶ Book power distribution $N+2$ design
This feature uses Voltage Transformation Modules (VTMs) in a highly redundant $N+2$ configuration.
- ▶ Redundant ($N+2$) humidity sensors
- ▶ Redundant ($N+2$) altimeter sensors
- ▶ Redundant ($N+2$) ambient temperature sensors
- ▶ Corrosion sensor²
- ▶ Unified support for the zBX
The zBX is supported like any other feature on the zEC12.
- ▶ Dual inline memory module (DIMM) field-replaceable unit (FRU) indicators
These indicators imply that a memory module is not error free and might fail sometime in the future. This indicator gives IBM a warning and the potential and time to concurrently repair the storage module if the zEC12 is a multibook system. First, isolate or “fence off” the book, remove the book, replace the failing storage module, and then add the book. The flexible memory option might be necessary to maintain sufficient capacity while repairing the storage module.
- ▶ Single processor core checkstop and sparing
This indicator shows that a processor core has malfunctioned and is *spared*. IBM determines what to do based on the system and the history of that system.
- ▶ Point-to-point fabric for symmetric multiprocessing (SMP)
Having fewer components that can fail is an advantage. In a multibook system, the ring connection between all of the books was replaced by point-to-point connections. A book can always be added concurrently.

² The current implementation is only for collecting field data for analysis. System operation will not be affected by the availability or functionality of this sensor.

- ▶ Redundant ($N+1$) radiators and Water Cooling Units (WCUs)

zEC12 implements true $N+1$ redundancy on pumps and blowers for radiators and WCUs. The radiator cooling system can support all four books simultaneously with a redundant design that consists of two pumps and two blowers. One active pump and blower can support the entire system load. The replacement of the pump or blower causes no performance impact. A water cooling system is also an option in zEC12, with WCU technology. Two redundant WCUs run with two independent chilled water feeds. Like the radiator cooling system, one WCU and one water feed can support the entire system load. Both radiator and water cooling systems are backed up by an air cooling system in the rare event of a cooling system problem.
- ▶ Hot-swap IFB hub cards

When correctly configured for redundancy, hot swapping (replacing) the IFB (HCA2-O (12xIFB) or HCA3-O (12xIFB)) hub cards is possible. This process avoids any interruption when you must replace these types of cards.
- ▶ Redundant 1-Gbps Ethernet service network with virtual LAN (VLAN)

The service network in the system gives the machine code the capability to monitor each internal function in the system. This process helps to identify problems, maintain the redundancy, and helps concurrently replacing a part. Through the implementation of the VLAN to the redundant internal Ethernet service network, these advantages are improved, making the service network easier to handle and more flexible.
- ▶ The PCIe I/O drawer is available for the zEC12. It and all of the PCIe I/O drawer-supported features can be installed concurrently.

10.2.2 Unscheduled outages

An *unscheduled outage* occurs because of an unrecoverable malfunction in a hardware component of the system.

The following improvements can minimize unscheduled outages:

- ▶ Continued focus on firmware quality

For LIC and hardware design, failures are eliminated through rigorous design rules; design walk-through; peer reviews; element, subsystem, and system simulation; and extensive engineering and manufacturing testing.
- ▶ Memory subsystem improvements

z196 introduced RAIM on System z systems, which is a concept similar to the known disk industry Redundant Array of Independent Disks (RAID). RAIM design detects and recovers from DRAM, socket, memory channel, or DIMM failures. The RAIM design requires the addition of one memory channel that is dedicated for RAS. The parity of the four “data” DIMMs is stored in the DIMMs attached to the fifth memory channel. Any failure in a memory component can be detected and corrected dynamically. The zEC12 inherited this memory architecture.

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 the zEC12 is implemented with an enhanced version of the Reed-Solomon error-correcting code (ECC) that is known as 90B/64B. It provides protection against memory channel and DIMM failures. A precise marking of faulty chips helps ensure timely DRAM replacements. The design of the zEC12 further improved this chip marking technology. The key cache on the zEC12 memory is mirrored. For more information about the memory system on the zEC12, see 2.5, “Memory” on page 45.

- ▶ Improved thermal, altitude, and condensation management
- ▶ Soft-switch firmware

zEC12 is equipped with the capabilities of soft-switching firmware. Enhanced logic in this function ensures that every affected circuit is powered off during the soft-switching of firmware components. For example, when you are upgrading the microcode of a FICON feature, enhancements are implemented to avoid any unwanted side effects detected on previous systems.

- ▶ STP recovery enhancement

When HCA3-O (12xIFB) or HCA3-O long reach (LR) (1xIFB) coupling links are used, an unambiguous “going away signal” is sent when the system is about to enter a failed (check-stopped) state. If the “going away signal” is sent by the Current Time Server (CTS) in a Server Time Protocol (STP)-only Coordinated Timing Network (CTN), the receiving end (the Backup Time Server (BTS)) can safely take over as the CTS. BTS does not have to rely on the Offline Signal (OLS) in a two-server CTN, or on the Arbiter in a CTN with three or more servers.

10.3 zEC12 enhanced book availability (EBA)

EBA is a *procedure* under which a book in a multibook system can be removed and reinstalled during an upgrade or repair action. This procedure has no impact on the running workload.

The EBA procedure and careful planning help ensure that all the resources are still available to run critical applications in a ($n-1$) book configuration. This process allows you to avoid planned outages. Consider the flexible memory option to provide more memory resources when you are replacing a book.

To minimize affecting current workloads, ensure that there are sufficient inactive physical resources on the remaining books to complete a book removal. Also, consider deactivating non-critical system images, such as test or development LPARs. After you stop these non-critical LPARs and free their resources, you might find sufficient inactive resources to contain critical workloads while completing a book replacement.

10.3.1 EBA planning considerations

To use the enhanced book availability function, configure enough physical memory and engines so that the loss of a single book does not result in any degradation to critical workloads during the following occurrences:

- ▶ A degraded restart in the rare event of a book failure
- ▶ A book replacement for repair or physical memory upgrade

The following configurations especially enable use of the enhanced book availability function. These zEC12 models need enough spare capacity so that they can cover the resources of the fenced or isolated book. This configuration imposes limits on the number of the client-owned processor units (PUs) that can be activated when one book within a model is fenced:

- ▶ A maximum of 21 client PUs are configured on the H43.
- ▶ A maximum of 44 client PUs are configured on the H66.
- ▶ A maximum of 66 client PUs are configured on the H89.
- ▶ A maximum of 75 client PUs are configured on the HA1.
- ▶ No special feature codes are required for PU and model configuration.

- ▶ For all zEC12 models, there are four SAPs in every book.
- ▶ The flexible memory option delivers physical memory so that 100% of the purchased memory increment can be activated even when one book is fenced.

The system configuration must have sufficient dormant resources on the remaining books in the system for the *evacuation* of the book that is to be replaced or upgraded. Dormant resources include the following possibilities:

- ▶ Unused PUs or memory that is not enabled by LICCC
- ▶ Inactive resources that are enabled by LICCC (memory that is not being used by any activated LPARs)
- ▶ Memory that is purchased with the flexible memory option
- ▶ Additional books

The I/O connectivity must also support book removal. Most of the paths to the I/O have redundant I/O interconnect support in the I/O infrastructure (drawers and cages) that enable connections through multiple fanout cards.

If sufficient resources are not present on the remaining books, certain non-critical LPARs might have to be deactivated. One or more CPs, specialty engines, or storage might have to be configured offline to reach the required level of available resources. Plan to address these possibilities to help reduce operational errors.

Exception: Single-book systems cannot use the EBA procedure.

Include the planning as part of the initial installation and any follow-on upgrade that modifies the operating environment. A client can use the Resource Link machine information report to determine the number of books, active PUs, memory configuration, and the channel layout.

If the zEC12 is installed, click **Prepare for Enhanced Book Availability** in the Perform Model Conversion window of the EBA process on the Hardware Management Console (HMC). This task helps you determine the resources that are required to support the removal of a book with acceptable degradation to the operating system images.

The EBA process determines which resources, including memory, PUs, and I/O paths, are freed to allow for the removal of a book. You can run this preparation on each book to determine which resource changes are necessary. Use the results as input in the planning stage to help identify critical resources.

With this planning information, you can examine the LPAR configuration and workload priorities to determine how resources might be reduced and allow for the book to be removed.

Include the following tasks in the planning process:

- ▶ Review of the zEC12 configuration to determine the following values:
 - Number of books that are installed and the number of PUs enabled. Note the following information:
 - Use the Resource Link machine information or the HMC to determine the model, number, and types of PUs (CPs, IFLs, ICFs, zAAPs, and zIIPs).
 - Determine the amount of memory, both physically installed and LICCC-enabled.

- Work with your IBM service support representative (SSR) to determine the memory card size in each book. The memory card sizes and the number of cards that are installed for each book can be viewed from the SE under the CPC configuration task list. Use the View Hardware Configuration option.
- Channel layouts, and host channel adapter (HCA) to channel connections
 - Use the Resource Link machine information to review the channel configuration, including the HCA paths. This process is a normal part of the I/O connectivity planning. The alternate paths must be separated as far into the system as possible.
- ▶ Review the system image configurations to determine the resources for each.
- ▶ Determine the importance and relative priority of each LPAR.
- ▶ Identify the LPAR or workloads and the actions to be taken:
 - Deactivate the entire LPAR.
 - Configure PUs.
 - Reconfigure memory, which might require the use of Reconfigurable Storage Unit (RSU) values.
 - Vary off the channels.
- ▶ Review the channel layout and determine whether any changes are necessary to address single paths.
- ▶ Develop the plan to address the requirements.

When you perform the review, document the resources that can be made available if the EBA is to be used. The resources on the books are allocated during a power-on reset (POR) of the system and can change after that process. Perform a review when changes are made to the zEC12, such as adding books, CPs, memory, or channels. Also, perform a review when workloads are added or removed, or if the HiperDispatch feature was enabled and disabled since the last time you performed a POR.

10.3.2 Enhanced book availability processing

To use the EBA, first ensure that the following conditions are satisfied:

- ▶ Free the used processors (PUs) on the book that will be removed.
- ▶ Free the used memory on the book.
- ▶ For all I/O domains connected to the book, ensure that alternate paths exist. Otherwise, place the I/O paths offline.

For the EBA process, this is the preparation phase. It is started from the SE, either directly or on the HMC by using the **Single object operation** option on the Perform Model Conversion window from the CPC configuration task list. See Figure 10-1 on page 395.

Processor availability

Processor resource availability for reallocation or deactivation is affected by the type and quantity of the resources in use:

- ▶ Total number of PUs that are enabled through LICCC
- ▶ PU definitions in the profiles that can be dedicated and dedicated reserved or shared
- ▶ Active LPARs with dedicated resources at the time of the book repair or replacement

To maximize the PU availability option, ensure that there are sufficient inactive physical resources on the remaining books to complete a book removal.

Memory availability

Memory resource availability for reallocation or deactivation depends on these factors:

- ▶ Physically installed memory
- ▶ Image profile memory allocations
- ▶ Amount of memory that is enabled through LICCC
- ▶ Flexible memory option

For more information, see 2.7.2, “Enhanced book availability” on page 58.

Fanout card to I/O connectivity requirements

The optimum approach is to maintain maximum I/O connectivity during book removal. The *redundant I/O interconnect (RII)* function provides for redundant HCA connectivity to all installed I/O domains in the PCIe I/O drawers, I/O cage, and I/O drawers.

Preparing for enhanced book availability

The Prepare Concurrent Book replacement option validates that enough dormant resources exist for this operation. If enough resources are not available on the remaining books to complete the EBA process, the process identifies those resources. It then guides you through a series of steps to select and free up resources. The preparation process does not complete until all processors, memory, and I/O conditions are successfully resolved.

Preparation: The preparation step does not reallocate any resources. It is only used to record client choices and produce a configuration file on the SE that is used to run the concurrent book replacement operation.

The preparation step can be done in advance. However, if any changes to the configuration occur between the preparation and the physical removal of the book, you must rerun the preparation phase.

The process can be run multiple times because it does not move any resources. To view results of the last preparation operation, select **Display Previous Prepare Enhanced Book Availability Results** from the Perform Model Conversion window in SE.

The preparation step can be run a few times without actually performing a book replacement. You can use it to dynamically adjust the operational configuration for book repair or replacement before IBM SSR activity. Figure 10-1 on page 395 shows the Perform Model Conversion window where you select **Prepare for Enhanced Book Availability**.

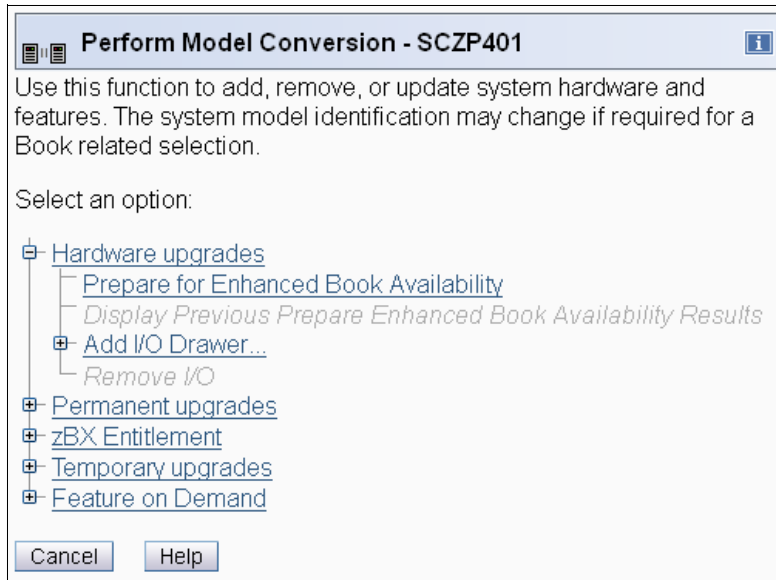


Figure 10-1 Perform Model Conversion: Select Prepare for Enhanced Book Availability

After you select **Prepare for Enhanced Book Availability**, the Enhanced Book Availability window opens. Select the book that is to be repaired or upgraded, then select **OK** as shown in Figure 10-2. Only one target book can be selected at a time.

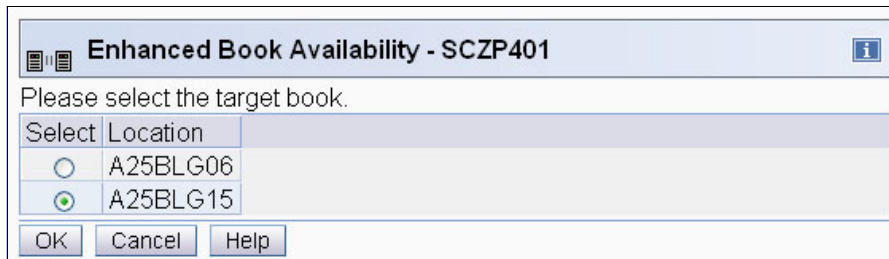


Figure 10-2 Enhanced Book Availability, selecting the target book

The system verifies the resources that are required for the removal, determines the required actions, and presents the results for review. Depending on the configuration, the task can take from a few seconds to several minutes.

The preparation step determines the readiness of the system for the removal of the targeted book. The configured processors and the memory in the selected book are evaluated against unused resources available across the remaining books. The system also analyzes I/O connections that are associated with the removal of the targeted book for any single path I/O connectivity.

If insufficient resources are available, the system identifies the conflicts so that you can free up other resources.

Three states can result from the preparation step:

- ▶ The system is ready to run the enhanced book availability for the targeted book with the original configuration.
- ▶ The system is not ready to run the enhanced book availability because of conditions that are indicated by the preparation step.

- ▶ The system is ready to run the enhanced book availability for the targeted book. However, to continue with the process, processors are reassigned from the original configuration. Review the results of this reassignment relative to your operation and business requirements. The reassignments can be changed on the final window that is presented. However, before making changes or approving reassignments, ensure that the changes are reviewed and approved by the correct level of support based on your organization's business requirements.

Preparation tabs

The results of the preparation are presented for review in a tabbed format. Each tab indicates conditions that prevent the EBA option from being run. Tabs are for processors, memory, and various single path I/O conditions. See Figure 10-3. The following tab selections are available:

- ▶ Processors
- ▶ Memory
- ▶ Single I/O
- ▶ Single Domain I/O
- ▶ Single Alternate Path I/O

Only the tabs that have conditions that prevent the book from being removed are displayed. Each tab indicates the specific conditions and possible options to correct them.

Example window from the preparation phase

Figure 10-3 shows the Single I/O tab. The preparation identified single I/O paths that are associated with the removal of the selected book. The paths must be placed offline to perform the book removal. After you address the condition, rerun the preparation step to ensure that all the required conditions are met.

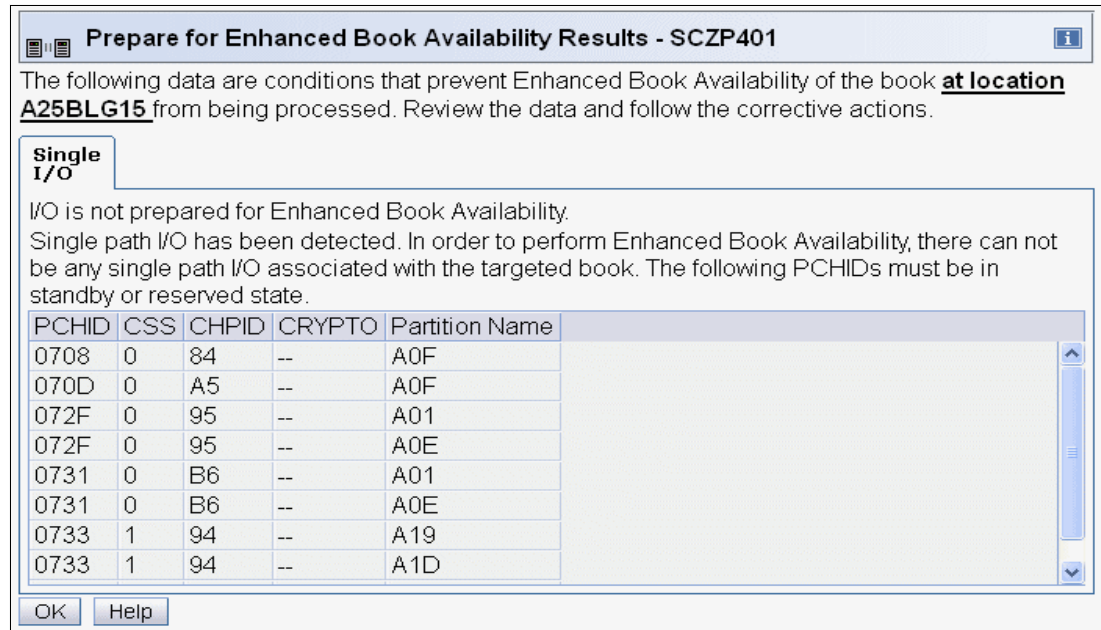


Figure 10-3 Prepare for EBA: Single I/O conditions

Preparing the system to perform enhanced book availability

During the preparation, the system determines the CP configuration that is required to remove the book. Figure 10-4 on page 397 shows the results and provides the option to change the assignment on non-dedicated processors.

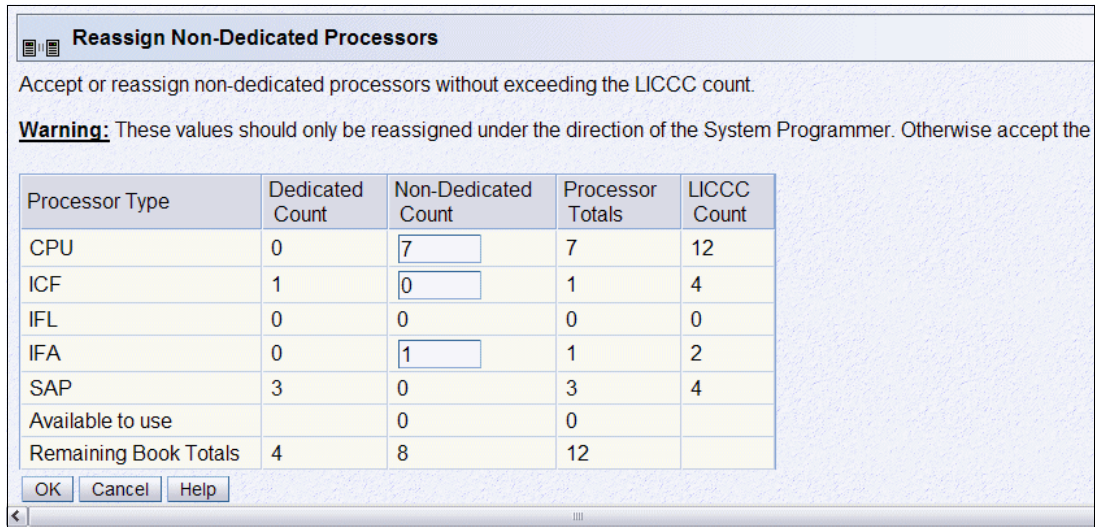


Figure 10-4 Reassign Non-Dedicated Processors results

Important: Consider the results of these changes relative to the operational environment. Understand the potential impact of making such operational changes. Changes to the PU assignment, although technically correct, can result in constraints for critical system images. In certain cases, the solution might be to defer the reassignments to another time that might have less impact on the production system images.

After you review the reassignment results, and make any necessary adjustments, click **OK**.

The final results of the reassignment, which include changes made as a result of the review, are displayed as shown in Figure 10-5. These results are the assignments when the book removal phase of the EBA is completed.

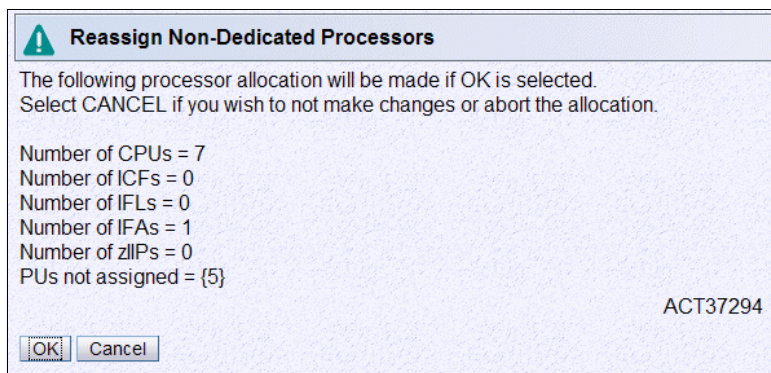


Figure 10-5 Reassign Non-Dedicated Processors, message ACT37294

Summary of the book removal process steps

To remove a book, the following resources must be moved to the remaining active books:

- ▶ PUs: Enough PUs must be available on the remaining active books, including all types of PUs that can be characterized (CPs, IFLs, ICFs, zAAPs, zIIPs, the IFP³, and SAPs).
- ▶ Memory: Enough installed memory must be available on the remaining active books.

³ If any native PCIe features are installed on the machine

- ▶ I/O connectivity: Alternate paths to other books must be available on the remaining active books, or the I/O path must be taken offline.

By understanding both the system configuration and the LPAR allocation for memory, PUs, and I/O, you can make the best decision about how to free necessary resources and allow for book removal.

To concurrently replace a book, perform these steps:

1. Run the preparation task to determine the necessary resources.
2. Review the results.
3. Determine the actions to perform to meet the required conditions for EBA.
4. When you are ready for the book removal, free the resources that are indicated in the preparation steps.
5. Rerun the step in Figure 10-1 on page 395 (the Prepare for Enhanced Book Availability task) to ensure that the required conditions are all satisfied.
6. Upon successful completion (see Figure 10-6), the system is ready for the removal of the book.

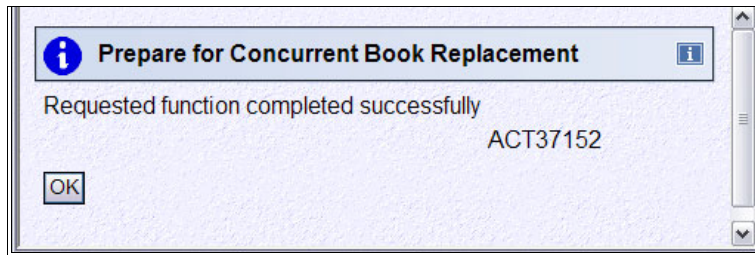


Figure 10-6 Preparation completed successfully, message ACT37152

The preparation process can be run multiple times to ensure that all conditions are met. It does not reallocate any resources. It only produces a report. The resources are not reallocated until the Perform Book Removal process is started.

Rules during EBA

During EBA, the following processor, memory, and single I/O rules are enforced:

- ▶ Processor rules

All processors in any remaining books are available to be used during EBA. This requirement includes the two spare PUs or any available PU that is non-LICCC.

The EBA process also allows conversion of one PU type to another PU type. One example is converting a zAAP to a CP during the EBA function. The preparation for the concurrent book replacement task indicates whether any SAPs must be moved to the remaining books.

- ▶ Memory rules

All physical memory that is installed in the system, including flexible memory, is available during the EBA function. Any physical installed memory, whether purchased or not, is available to be used by the EBA function.

- ▶ Single I/O rules

Alternate paths to other books must be available, or the I/O path must be taken offline.

Review the results. The result of the preparation task is a list of resources that must be made available before the book replacement can take place.

Free any resources

At this stage, create a plan to free up these resources. The following list shows the resources and actions that are necessary to free them:

- ▶ To free any PUs:
 - Vary the CPs off by using the Perform a Model Conversion window, reducing the number of CPs in the shared CP pool.
 - Deactivate the LPARs.
- ▶ To free memory:
 - Deactivate an LPAR.
 - Vary offline a portion of the reserved (online) memory. For example, in z/OS, issue the command:

```
CONFIG_STOR(E=1), <OFFLINE/ONLINE>
```

This command enables a storage element to be taken offline. The size of the storage element depends on the RSU value. In z/OS, the following command configures offline smaller amounts of storage than the amount that was set for the storage element:

```
CONFIG_STOR(nnM), <OFFLINE/ONLINE>
```
 - A combination of both LPAR deactivation and varying memory offline.

Reserved storage: If you plan to use the EBA function with z/OS LPARs, set up reserved storage and an RSU value. Use the RSU value to specify the number of storage units that are to be kept free of long-term fixed storage allocations. This configuration allows for storage elements to be varied offline.

10.4 zEC12 enhanced driver maintenance (EDM)

EDM is one more step toward reducing both the necessity for and the eventual duration of a scheduled outage. One of the contributors to planned outages is LIC Driver updates that are run in support of new features and functions.

When correctly configured, the zEC12 supports 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. There are certain LIC updates where a concurrent update/upgrade might not be possible.

Consider the following key points of EDM:

- ▶ The HMC can query whether a system is ready for a concurrent driver upgrade.
- ▶ Previous firmware updates, which require an initial machine load (IML) of zEC12 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 SSR to define the concurrent driver upgrade sync point to be used for an EDM.
- ▶ The ability to concurrently install and activate a new driver can eliminate or reduce a planned outage.
- ▶ The zEC12 introduces Concurrent Driver Upgrade (CDU) cloning support to other CPCs for CDU preload and activate.

- ▶ 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 completely 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 CHPID code changes might require CHPID Vary OFF/ON to activate new code.
- ▶ Changes to the code of native PCIe features might require additional action from the client if the specific feature needs to 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.

10.5 RAS capability for the HMC and SE

The HMC and the SE have the following RAS capabilities:

- ▶ Back up from HMC and SE

On a scheduled basis, the HMC and SE hard disks are backed up on the HMC backup USB media.

- ▶ 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 12.5, “Remote Support Facility (RSF)” on page 440.

- ▶ 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 hiper MCLs warrants.

For more information, see 12.6.4, “HMC and SE microcode” on page 445.

- ▶ Support Element (SE)

The zEC12 is provided with two notebook computers inside the System z 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.

For more information, see 12.1, “Introduction to the HMC and SE” on page 432.

- ▶ 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 System z systems. From a serviceability point of view, all the components of the ensemble, including the zBX, are treated as zEC12 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 the zEC12.

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 zEC12 infrastructure. Therefore, the HMC is in a stateful state that needs high availability features to ensure the survival of the system in a failure. Each ensemble must therefore be equipped with two HMC workstations: a primary and an alternate. The primary HMC can perform all HMC activities (including Unified Resource Manager activities). The alternate can only be the backup. The alternate 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 12.7, “HMC in an ensemble” on page 461.

- ▶ 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 performing an HMC switchover.

10.6 RAS capability for zBX

The zBX was built with traditional System z quality of service (QoS) to include RAS capabilities. The zBX offering provides extended service capability with the zEC12 hardware management structure. The HMC/SE functions of the zEC12 CPC provide management and control functions for the zBX solution.

Apart from a zBX configuration with one chassis 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 Top of Rack switches (ToRs). These switches provide $N + 1$ connectivity for the private networks between the zEC12 CPC and the zBX 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.
- ▶ ($N + 1$) PDUs. Provided that the Power Distribution Units (PDUs) have power inputs from two separate sources, in case of a single source failure, the second PDU will take 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.
- ▶ 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 they might have reached other maximums, such as power usage.

zBX firmware

The testing, delivery, installation, and management of the zBX firmware is handled in exactly the same way as the zEC12 CPC. The same processes and controls are used. All fixes to the zBX are downloaded to the controlling zEC12's SE and applied to the zBX.

The Microcode Change Levels (MCLs) for the zBX are designed to be concurrent and their status can be viewed at the zEC12's HMC.

zBX RAS and the Unified Resource Manager

The Hypervisor Management function of 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 the Smart Analytics Optimizer or the DataPower solutions, the solution handles the complete end-to-end management for these blades and their operating systems, middleware, and applications.

For blades that are deployed by the client, the Unified Resource Manager handles 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 will power up the blade, verify that there is an entitlement for the blade, and verify that the blade can participate in an ensemble. If these two conditions are not met, the Unified Resource Manager powers down the blade.
- ▶ The blade will be populated with the necessary microcode and firmware.
- ▶ The appropriate hypervisor will be loaded on the blade.
- ▶ The management scope will be deployed according to which management enablement level is present in the configuration.
- ▶ The administrator can define the blade profile, as well as the profiles for virtual servers to execute 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 will be handled by the client's usual system management disciplines.

zBX Model 003: 2458-003

The zEC12 only supports a zBX Model 003. When upgrading a z196 to a zEC12, the zBX is also upgraded from a Model 002 to a Model 003.

The zBX Model 003 is based on the BladeCenter and blade hardware offerings that contain IBM certified components. zBX Model 003 BladeCenter and blade RAS features are extended considerably for IBM System z®:

- ▶ 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
- ▶ 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 firmware fixes being installed or activated, a deactivation of a target blade might be required.

The zBX offering provides extended service capabilities with the zEC12 hardware management structure. The HMC/SE functions of the zEC12 system provide management and control functions for the zBX solution.

The zBX has two pairs of Top of Rack (ToR) switches, the intra-node management network (INMN) *N*+1 pair and the intra-ensemble data network (IEDN) *N*+1 switch pair. The management switch pair (INMN) provides *N*+1 connectivity for the private networks between the zEC12 system and the zBX. 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.

Not only hardware and firmware provide RAS capabilities. The operating system can also contribute significantly to improving RAS. IBM PowerHA® SystemMirror® for AIX (PowerHA) supports the zBX PS701 blades⁴. PowerHA enables setting up a PowerHA environment on the zEC12 controlled zBX. Table 10-1 provides more detail about PowerHA and the required AIX⁵ levels that are needed for a PowerHA environment on zBX.

Table 10-1 PowerHA and required AIX levels

IBM zBX Model 003	AIX V5.3	AIX V6.2	AIX V7.1
PowerHA V5.5	AIX V5.3 TL12 Reliable Scalable Cluster Technology (RSCT) 2.4.13.0	AIX V6.1 TL05 RSCT 2.5.5.0	PowerHA V5.5 SP8 AIX V7.1 RSCT V3.1.0.3
PowerHA V6.1	AIX V5.3 TL12 RSCT 2.4.13.0	AIX V6.1 TL05 RSCT 2.5.5.0	PowerHA V6.1 SP3 AIX V7.1 RSCT V3.1.0.3
PowerHA V7.1	Not supported	AIX V6.1 TL06 RSCT V3.1.0.3	AIX V7.1 RSCT V3.1.0.3

⁴ PS701 8406-71Y blades.

⁵ 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.

zEnterprise BladeCenter Extension (zBX) Model 003 introduces a new version for the advanced management module (AMM). It also includes major firmware changes compared to the zBX Model 002. zBX Model 003 takes the RAS concept of the zBX Model 003 to higher levels.

10.7 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^{™6}). 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 24x7. 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 servers⁷ (up to 32) that have their resources managed by PowerHA cluster services. The configuration provides automated service recovery for the applications managed. 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 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 failover inside the same ensemble. All zBXs participating in the PowerHA cluster must have access to the same storage.

⁶ High Availability Cluster Multi-Processing.

⁷ Servers can be also virtual servers; one server = one instance of the AIX Operating System.

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)
- ▶ 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 10-7.

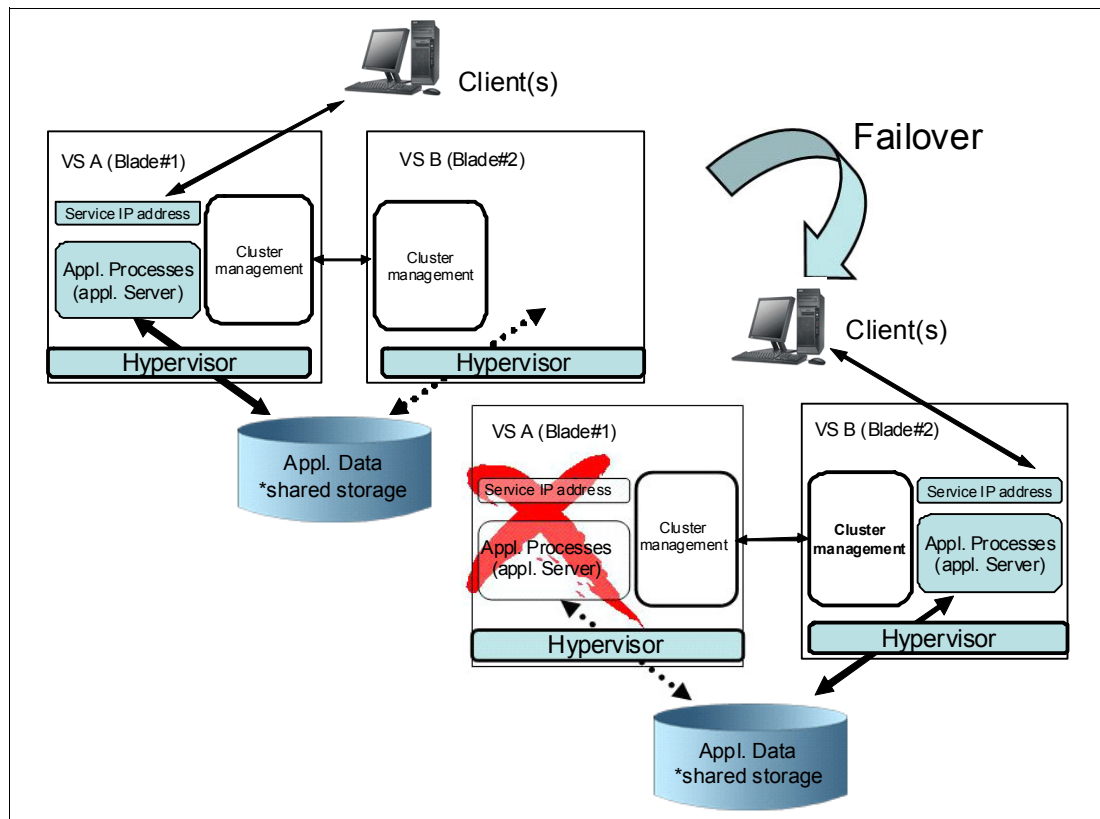


Figure 10-7 Typical PowerHA cluster diagram

For more information about IBM PowerHA SystemMirror for AIX, see this website:

<http://www-03.ibm.com/systems/power/software/availability/aix/index.html>

10.8 IBM System z Advanced Workload Analysis Reporter (IBM zAware)

IBM zAware provides a smart solution for detecting and diagnosing anomalies in z/OS 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 zEC12 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 systems 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).

Statement of Direction: IBM plans to provide new capability within the Tivoli Integrated Service Management family of products. This capacity takes advantage of analytics information from IBM zAware to provide alert and event notification.

IBM zAware improves the overall RAS capability of zEC12 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 A, “IBM System z Advanced Workload Analysis Reporter (IBM zAware)” on page 479.

10.9 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 disks (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 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 only usable on the system with the key that encrypted it. For key management, both the Primary and Alternate Support Elements (SE) have a smart card reader installed.

Flash Express cards support concurrent firmware upgrades.

Figure 10-8 shows the various components that support Flash Express RAS functions.

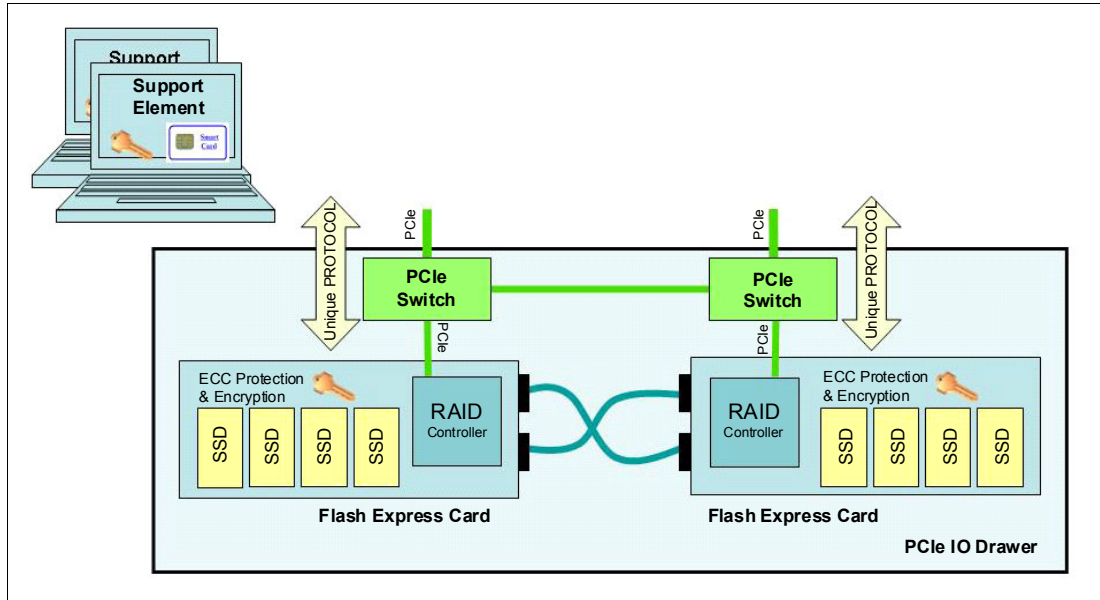


Figure 10-8 Flash Express RAS components

For more information about Flash Express, see Appendix C, “Flash Express” on page 501.



Environmental requirements

This chapter addresses the environmental requirements for the IBM zEnterprise EC12 (zEC12). It lists the dimensions, weights, power, and cooling requirements needed to plan for the installation of an IBM zEnterprise EC12 and IBM zEnterprise BladeCenter Extension (zBX) Model 003.

There are several options for the physical installation of the server:

- ▶ Air or water cooling
- ▶ Installation on raised floor or non-raised floor
- ▶ I/O and power cables exiting under the raised floor or off the top of the server frames
- ▶ Having a high-voltage DC power supply as an alternative to the usual AC power supply

For more information about physical planning, see *Installation Manual - Physical Planning 2827*, GC28-6914, for the zEC12 and *zBX Model 003 Installation Manual - Physical Planning*, GC27-2619, for the IBM zEnterprise BladeCenter Extension (zBX) Model 003.

This chapter includes the following sections:

- ▶ zEC12 power and cooling
- ▶ IBM zEnterprise EC12 physical specifications
- ▶ IBM zEnterprise EC12 physical planning
- ▶ zBX environmental requirements
- ▶ Energy management

11.1 zEC12 power and cooling

The zEC12 is always a two-frame system. The frames are shipped separately and are bolted together during the installation procedure. The zEC12 supports installation on a raised floor or non-raised floor. However, the zEC12 with the water cooling feature must be installed on a raised floor because the water hoses must attach to the server from underneath the raised floor. Power and I/O cables also exit from the bottom of the server frames unless the Top Exit I/O Cabling feature code (FC 7942) or Top Exit Power feature code (FC 7901) is installed. These options allow I/O cables and power cables to exit from the top of the server into overhead cabling rails.

11.1.1 Power consumption

The system operates with two fully redundant power supplies. Each power supply has an individual power cord or a pair of power cords, depending on the configuration.

For redundancy, the server needs to have two power feeds. Each power feed is either one or two power cords. The number of power cords that are required depends on system configuration. Power cords attach to either 3-phase, 50/60 Hz, 200 - 480 V AC power, or 380 - 520 V DC power. The total loss of one power feed has no impact on system operation.

For ancillary equipment, such as the Hardware Management Console (HMC), its display and switch, more single-phase outlets are required.

The power requirements depend on the cooling facility that is installed, and on the number of books and I/O units installed. I/O power units are values for I/O cages (an I/O cage equals two I/O units) and I/O drawers or Peripheral Component Interconnect Express (PCIe) drawers (both drawer types equal one I/O unit).

Heat output, which is expressed in kBTU per hour, can be derived by multiplying the table entries by a factor of 3.4.

Table 11-1 lists the absolute maximum power requirements for the air-cooled models in a warm room (≥ 28 degrees Centigrade).

Table 11-1 Power requirements: Air-cooled models

Power requirement kVA	Number of I/O units						
	0	1	2	3	4	5	6
H20	5.8	7.7	9.7	11.3	13.2	13.4	13.4
H43	9.7	11.6	13.4	15.2	17.1	19.0	19.8
H66	13.2	15.0	16.9	18.8	20.5	22.4	23.3
H89/HA1	17.6	19.5	21.4	23.2	24.9	26.8	27.6

Consideration: Power will be lower in a normal ambient temperature room and for configurations that do not have every I/O slot plugged (maximum memory and maximum configured processors). Power will also be slightly lower for DC input voltage.

Actual power for any configuration, power source, and room condition can be obtained by using the power estimation tool at the IBM Resource Link:

<https://www-304.ibm.com/servers/resourceLink/>

Table 11-2 lists the maximum power requirements for the water-cooled models.

Table 11-2 Power requirements: Water-cooled models

Power requirement kVA	Number of I/O units						
	0	1	2	3	4	5	6
H20	5.5	7.4	9.4	11.0	12.9	13.1	13.4
H43	9.1	10.9	12.7	14.6	16.4	18.3	19.1
H66	12.4	14.3	16.1	18.0	19.7	21.6	22.5
H89/HA1	17.7	18.6	20.4	22.3	24.1	25.8	26.6

Table 11-3 lists the Bulk Power Regulator (BPR) requirements for books and I/O units. A second pair of power cords is installed if the number of BPR pairs is four or higher.

If your initial configuration needs one power cord pair, but for growth will need a second pair, you can order the power cord Plan Ahead feature (FC 2000). This feature installs four power cords at the initial configuration. Also, if Balanced Power Plan Ahead (FC 3003) is ordered, four power cords are shipped and all 12 possible BPRs are installed. If the zEC12 is configured with the Internal Battery Feature (IBF), Balanced Power Plan Ahead automatically supplies the maximum number of batteries, six IBFs, with the system.

Table 11-3 Requirements for the number of BPRs

Number of BPRs per side	Number of I/O units						
	0	1	2	3	4	5	6
H20	1	1	1	2	3	3	3
H43	2	3	3	3	3	4	4
H66	3	3	4	4	4	4	4
H89/HA1	4	4	5	5	5	5	5

Systems that specify two power cords can be brought up with one power cord and continue to run. The larger systems that have a minimum of four BPR pairs installed must have four power cords installed. Four power cords offer power redundancy, so that when a power cord fails, the remaining cords deliver sufficient power to keep the system running.

11.1.2 Internal Battery Feature

The optional Internal Battery Feature (IBF) provides sustained system operations for a relatively short time, allowing for an orderly shutdown. In addition, an external uninterrupted power supply system can be connected, allowing for longer periods of sustained operation.

The IBF can provide emergency power for the estimated time that is listed in Table 11-4 on page 412. The batteries are only connected to the BPRs associated with section one, so one pair of batteries is connected to BPR 1 and BPR 2. See Table 11-3 for the number of BPRs installed in relation to I/O units and number of books. The number of batteries depends on the number of BPRs. They are installed in pairs. You can have two, four, or six batteries (odd numbers not allowed).

Table 11-4 Battery holdup times

Internal battery holdup times in minutes	Number of I/O units						
	0	1	2	3	4	5	6
H20	7.7	5.0	4.0	7.9	11.1	11.0	11.0
H43	9.7	13.6	11.0	9.1	7.9	7.0	6.7
H66	11.1	9.3	8.0	7.1	6.4	5.2	4.9
H89/HA1	7.6	6.8	6.1	5.0	4.4	4.0	3.8

Consideration: The system holdup times in Table 11-4 assume that both sides are functional and have fresh batteries under normal room ambient conditions. Holdup times are greater for configurations that do not have every I/O slot plugged (maximum memory and maximum configured processors). Holdup times for actual configurations are provided in the power estimation tool. See the IBM Resource Link (authentication required):

<https://www-304.ibm.com/servers/resourceLink/>

11.1.3 Emergency power-off switch

On the front of frame A is an emergency power-off switch that, when activated, immediately disconnects utility *and battery power* from the server. This process causes all volatile data in the server to be lost.

If the server is connected to a room's emergency power-off switch, and the Internal Battery Feature (IBF) is installed, the batteries take over if the switch is engaged.

To avoid the takeover, connect the room emergency power-off switch to the server power-off switch. Then, when the room emergency power-off switch is engaged, all power is disconnected from the power cords and the Internal Battery Features. However, all volatile data in the server will be lost.

11.1.4 Cooling requirements

The zEC12 cooling system is a combination of an air cooling system and water cooling system. In normal working conditions, the zEC12 Multi-chip module (MCM) is cooled by a water cooling system with radiator or water cooling units (WCUs). I/O drawers, power enclosures, and books are cooled by chilled air with blowers.

Air cooling system requirements

The air cooling system requires chilled air to fulfill the air cooling requirements. Normal air exhausts from the front to the rear of the frames. The chilled air is usually provided through perforated floor panels in front of the system.

Figure 11-1 on page 413 does not represent any particular server system type, and is intended only to show hot and cold airflow and the arrangement of server aisles.

Typically, the zEC12 air-cooled models use chilled air, which is provided from under the raised floor, to cool the system. As shown in Figure 11-1, rows of servers must face front-to-front. Chilled air is usually provided through perforated floor panels that are placed in rows between the fronts of servers (the cold aisles that are shown in the figure). Perforated tiles generally are not placed in the hot aisles. If your computer room causes the temperature in the hot aisles to exceed a comfortable temperature, add as many perforated tiles as necessary to create a satisfactory comfort level. Heated exhaust air exits the computer room above the computing equipment.

For more information about the requirements for air-cooling options, see *Installation Manual - Physical Planning 2827*, GC28-6914.

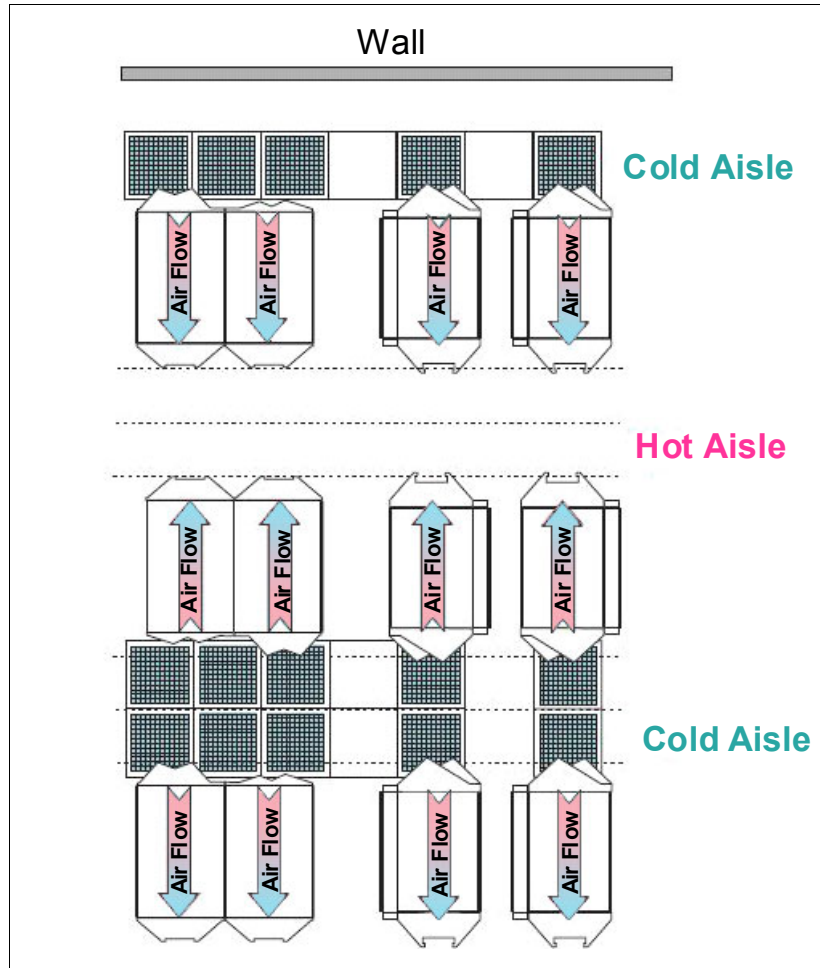


Figure 11-1 Hot and cold aisles

Cycle steering mode: In specific cases, when one of the modular refrigeration units (MRU) fails, the system (either air cooled or water cooled) can enter an *air cooling mode* or *cycle steering mode* when that system becomes too hot. In this situation, the system might also vent hot air through the top of the machine to lower the temperature.

Water cooling system requirements

The water cooling system requires chilled building water to be supplied to the zEC12 WCUs. The zEC12 requires four connections to the facility water: two feeds and two returns.

These connections are made by using hoses that are fixed to the facility plumbing and are routed up through the front tailgate of the system. They terminate with quick connect couplings.

Before you install the EC12, several conditions for water-cooled systems are required for your facility:

- ▶ Total Hardness must not exceed 200 mg/L as calcium carbonate.
- ▶ pH must be 7 - 9.
- ▶ Turbidity must be less than 10 Nephelometric Turbidity Units (NTUs).
- ▶ Bacteria must be less than 1000 colony-forming units (CFUs)/ml.
- ▶ Water needs to be as free of particulate matter as feasible.
- ▶ The allowable system inlet water temperature range is 6°C - 20°C (43°F - 68°F), using standard building chilled water. A special water system is typically not required.
- ▶ The required flow rate to the frame is 3.7 - 79.4 lpm (1 - 21 gpm), depending on the inlet water temperature and the number of nodes in the server. Colder inlet water temperatures require less flow than warmer water temperatures. Fewer nodes require less flow than the maximum populated processors.
- ▶ The minimum water pressure that is required across the IBM hose ends is 0.34 - 2.32 BAR (5 - 33.7 psi), depending on the minimum flow required.
- ▶ The maximum water pressure that is supplied at the IBM hose connections to the client's water supply cannot exceed 6.89 BAR (100 psi).

For more information about the requirements for water cooling options, see *Installation Manual - Physical Planning 2827*, GC28-6914, and Figure 11-2 on page 415.

Supply hoses

The zEC12 water cooling system includes 4.2-m (13.7-ft) water hoses. One set includes one supply and one return for use water supply connections that are under the floor. Multiple sets of hoses can be ordered based on your requirements. They can be pre-ordered as M/T 2819-W00.

Figure 11-2 shows the WCU water supply connections.

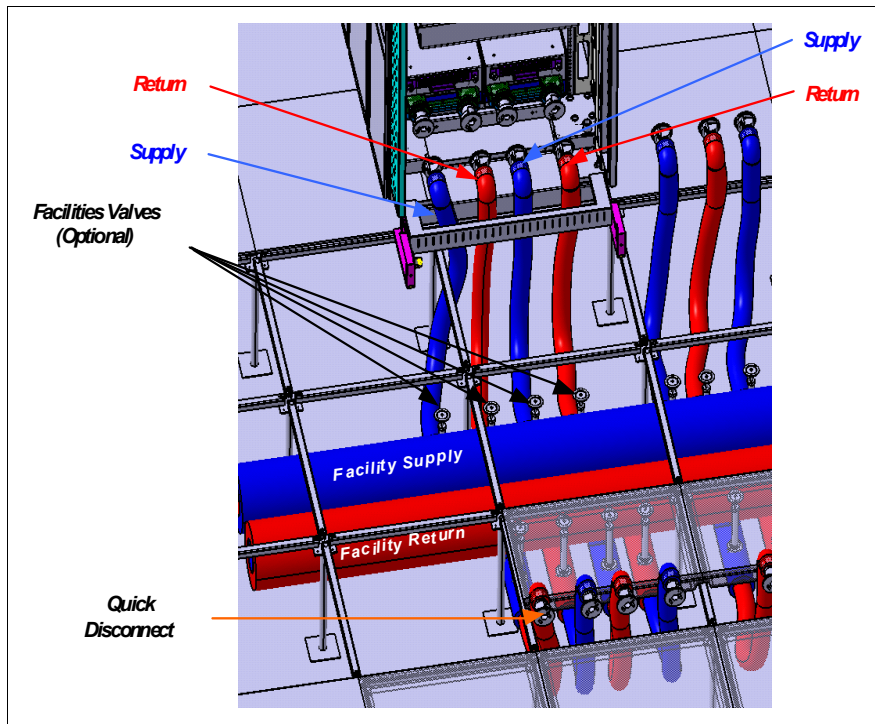


Figure 11-2 WCU water supply connections

The client's ends of the hoses are left open, allowing you to cut the hose to the length that you need. An insulation clamp is provided to secure the insulation and protective sleeving after you cut the hose to the correct length and install it onto your plumbing.

Raised floor: The minimum raised floor height for water-cooled system is 22.86 centimeters (8.6 inches).

Generally, use shut-off valves in front of the hoses. This configuration allows for removal of the hoses for a service procedure or relocation. Valves are not included in the order. A stainless steel fitting is available for ordering. Fitting is barbed on one side and has a 2.54 cm (1 in) male national pipe thread (NPT). This fitting is not supplied with the original ship group, and must be ordered separately.

11.2 IBM zEnterprise EC12 physical specifications

This section describes weights and dimensions of the zEC12.

The zEC12 can be installed on a raised or non-raised floor. For more information about weight distribution and floor loading tables, see the *Installation Manual - Physical Planning 2827*, GC28-6914. This data is to be used together with the maximum frame weight, frame width, and frame depth to calculate the floor loading.

Table 11-5 on page 416 indicates the maximum system dimension and weights for the HA1 model. The weight ranges are based on configuration models with five PCIe I/O drawers, IBF, and with the top exit cable features.

Table 11-5 System dimensions and weights

Maximum	A and Z frames with IBF (FC 3212)	A and Z frames with IBF (FC 3212) and Top Exit Cabling Features (FC 7942 and FC 7901)
Radiator-cooled servers		
Weight kg (lbs)	2430 (5358)	2516.5 (5548)
Width mm (in)	1568 (61.7)	1847 (72.7)
Depth mm (in)	1869 (73.6)	1806 (71.1)
Height mm (in)	2015 (79.3)	2154 (84.8)
Height reduction mm (in)	1803 (71.0)	1803 (71.0)
Water-cooled servers		
Weight kg (lbs)	2473 (5453)	2660 (5643)
Width mm (in)	1568 (61.7)	1847 (72.7)
Depth mm (in)	1971 (77.7)	1908 (75.1)
Height mm (in)	2015 (79.3)	2154 (84.8)
Height reduction mm (in)	1809 (71.2)	1809 (71.2)
Notes:		
1. Weight does not include covers. Covers add 68 kg (150 lbs) to each frame. Width, depth, and height are also indicated without covers.		
2. Weight is based on the maximum system configuration, not the addition of the maximum weight of each frame.		

11.3 IBM zEnterprise EC12 physical planning

This section describes the floor mounting options, and power and I/O cabling options.

11.3.1 Raised floor or non-raised floor

The zEC12 can be installed on a raised or non-raised floor. The water-cooled models require a raised floor.

Raised floor

If the zEC12 server is installed in a raised floor environment, both air-cooled and water-cooled models are supported. Typically, I/O cables, power cables, and water hoses connect to the server from underneath the raised floor. For the zEC12, you can select top exit features to route I/O cables and power cables from the top frame of the zEC12 server. The following additional options are available for the zEC12:

- ▶ Top Exit I/O Cabling feature code (FC 7942)
- ▶ Top Exit Power feature code (FC 7901)

Figure 11-3 on page 417 shows the top exit feature options of the zEC12 in a raised floor environment.

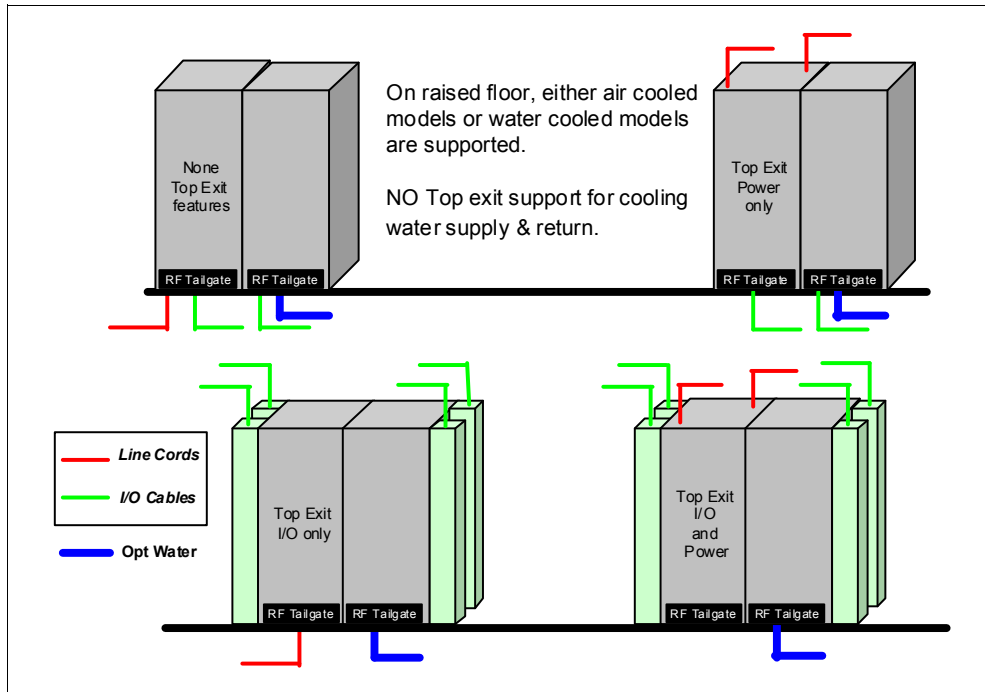


Figure 11-3 Raised floor options

Remember: There is no top exit feature support of water hoses, which must go through the system from underneath the raised floor.

Non-raised floor

If you install the zEC12 server in a non-raised floor environment, you can select only air-cooled models. The Non-Raised Floor Support feature code (FC 7998) is required. The Top Exit I/O Cabling feature code (FC 7942) and Top Exit Power feature code (FC 7901) must be ordered, as well. All cables must exit from the top frame of the zEC12 server as shown in Figure 11-4.

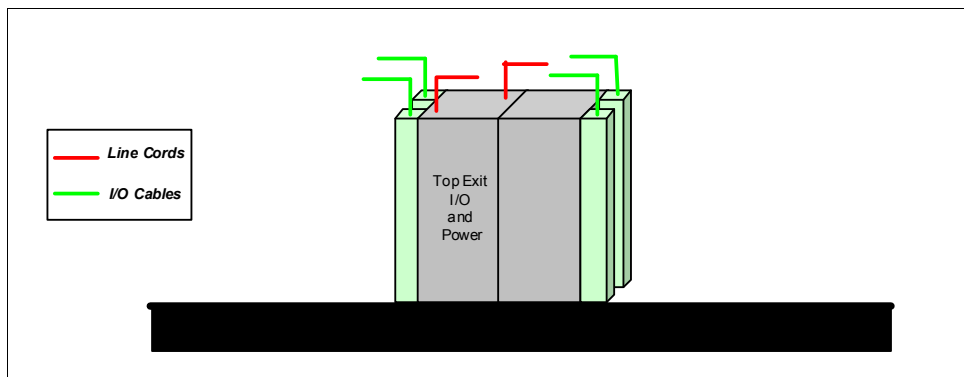


Figure 11-4 Non-raised floor options

11.3.2 Top Exit Power feature

Top Exit Power is a new feature of the zEC12. The Top Exit Power feature (FC 7901) is designed to provide you with an additional option. Instead of all of your power cables exiting under the raised floor, you can route your power cables from the top of the frame.

The Top Exit Power feature (FC 7901) is shipped separately from the system and installed onsite. It is installed on the top of the z frame, and increases the height of the frame from 17.78 cm (7 inches) - 30.48 cm (1 ft) based on the selected power cords.

Two types of power cords are offered in this feature:

- ▶ Cut cords

These cords are 4.3 m (14 ft) long from the exit point of the frame with an attached mount bracket that you can use to fix power cords on the top of frame as shown in Figure 11-4 on page 417.

- ▶ Plugged cords

These cords are all new for the zEC12. On the z114, the plugged cords are 1.8 m (6 ft) long from the frame exit. They must be plugged within 15 cm - 20 cm (6 - 8 in) from the top of the frame. Because these cords have 30A fittings, which are not difficult to plug in, this is a reasonable solution. For the zEC12, the fittings are 60A and require much more force to plug in successfully. For the 60A plugs, the “power cord” is a short connection from the power enclosure to the top of the frame. The plug is rigidly fixed to the frame. The client drop must come down to the frame to meet the system input plug.

Figure 11-5 shows the difference between cut cords and plugged cords.

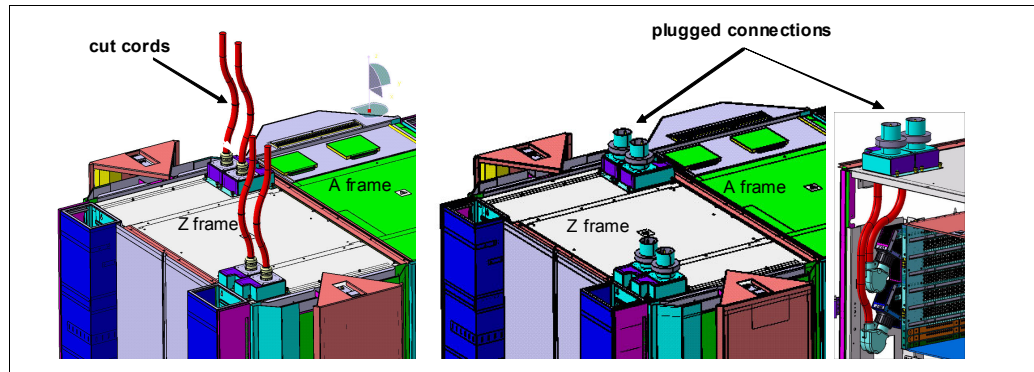


Figure 11-5 Top Exit Power feature

11.3.3 Top Exit I/O Cabling feature

Like the z196, the zEC12 supports the Top Exit I/O Cabling feature (FC 7942). This feature routes all coupling links and all I/O cables, including 1000BASE-T Ethernet cable from I/O cages or drawers, through four more frame extensions, out the top of the frame.

Figure 11-6 on page 419 shows the frame extensions, also called *chimneys*, that are installed to each corner of the frames (A frame and Z frame) when the Top Exit I/O Cabling feature (FC 7942) is ordered. Only Coupling Link, Ethernet, and Fibre Cabling can enter the system through the chimneys. The bottom of the chimney is closed with welded sheet metal.

The Top Exit I/O Cabling feature adds 15 cm (6 in) to the width of each frame and about 95 lbs (43 kg) to the weight.

In the zEC12, the Top Exit I/O Cabling feature (FC 7942) is available for both air-cooled models and water-cooled models.

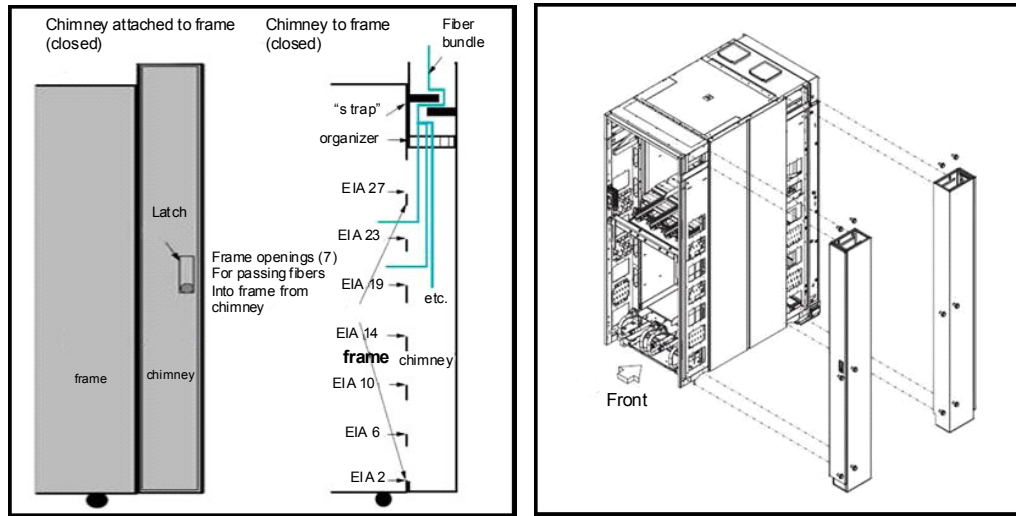


Figure 11-6 Top Exit I/O Cabling feature

11.3.4 Weight distribution plate

The *weight distribution plate* is designed to distribute the weight of a frame onto two floor panels in a raised-floor installation. As listed in Table 11-5 on page 416, the weight of a frame can be substantial. A concentrated load on a caster or leveling foot can be half of the total frame weight. In a multiple system installation, one floor panel can have two casters from two adjacent systems on it, potentially inducing a highly concentrated load on a single floor panel. The weight distribution plate distributes the weight over two floor panels. The weight distribution kit is ordered and delivered by using FC 9970.

Always consult the floor tile manufacturer to determine the load rating of the tile and pedestal structure. Additional panel support might be required to improve the structural integrity because cable cutouts significantly reduce the floor tile rating.

11.3.5 Three-in-one bolt-down kit for raised floor

A bolt-down kit for raised floor environments can be ordered for the EC12 frames. The kit provides hardware to enhance the ruggedness of the frames and to tie down the frames to a concrete floor beneath a raised floor of 15 cm - 91 cm (5.9 in - 35.8 in). The kit is offered in the following configurations:

- ▶ The Bolt-Down Kit for an air-cooled system (FC 8000) provides frame stabilization and bolt-down hardware for securing a frame to a concrete floor beneath the raised floor.
- ▶ The Bolt-Down Kit for a water-cooled system (FC 8001) provides frame stabilization and bolt-down hardware for securing a frame to a concrete floor beneath the raised floor.

Each server needs two features, one for each of the frames. The kits help secure the frames and their contents from damage when exposed to shocks and vibrations such as those generated by a seismic event. The frame tie-downs are intended for securing a frame that weighs less than 1632 kg (3600 lbs).

11.4 zBX environmental requirements

The following sections address the environmental requirements in summary for the zEnterprise BladeCenter Extension (zBX). For more information about the environmental requirements for the zBX, see *zBX Installation Manual for Physical Planning 2458-003*, GC27-2619.

11.4.1 zBX configurations

The zBX can have 1 - 4 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 blades that are required determines the actual components that are required for each configuration. The number of blades determines the number of BladeCenters and racks as shown in Table 11-6.

Table 11-6 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 POWER7 blades. A maximum of 56 IBM BladeCenter HX5 blades can be installed in a zBX. For DataPower blades, the maximum number is 28 because they are double-wide.

11.4.2 zBX power components

The zBX has its own power supplies and cords that are independent of the zEC12 server power. Depending on the configuration of the zBX, up to 16 client-supplied power feeds might be required. A fully configured four-rack zBX has 16 Power Distribution Units (PDUs). The zBX operates with the following characteristics:

- ▶ 50/60Hz AC power
- ▶ Voltage (240 V)
- ▶ Both single-phase and three-phase wiring

PDUs and power cords

The zBX has these available PDU options:

- ▶ FC 0520 - 7176 Model 3NU with attached power cord (US)
- ▶ FC 0521 - 7176 Model 2NX (WW)

The following power cord options are available for the zBX:

- ▶ FC 0531 - 4.3 meter, 60A/208V, US power cord, Single Phase
- ▶ FC 0532 - 4.3 meter, 63A/230V, non-US power cord, Single Phase
- ▶ FC 0533 - 4.3 meter, 32A/380V-415V, non-US power cord, Three Phase. 32A WYE 380V provides 220 V line to neutral, and 32A WYE 415V provides 240 V line to neutral. This setting ensures that the BladeCenter maximum of 240 V is not exceeded.

Power installation considerations

Each zBX BladeCenter operates from two fully redundant PDUs installed in the rack with the BladeCenter. These PDUs each have their own power cords, as shown in Table 11-7. This configuration allows the system to survive the loss of the client's 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 11-7 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

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.

DataPower blades: The DataPower blade is a double-wide blade.

For maximum availability, attach the power cords on each side of the racks to different building power distribution units.

Actual power consumption is dependent on the zBX configuration in terms of the number of BladeCenters and blades installed. Input power in kVA is equal to the output power in 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.

11.4.3 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 internal BladeCenter components:

- ▶ If the ambient temperature is 25°C (77°F) or below, the BladeCenter unit blowers run at their minimum rotational speed. They increase their speed as required to control the 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 to cool the BladeCenter unit and blade servers.

Typical heat output

Table 11-8 shows the typical heat that is released by various zBX solution configurations.

Table 11-8 zBX power consumption and heat output

Number of blades	Max 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 with limited cooling capacity, use the Rear Door Heat eXchanger (FC 0540) as shown in Figure 11-7 on page 423. It is a more cost-effective solution than adding another air conditioning unit.

Rear Door Heat eXchanger: The Rear Door Heat eXchanger is not a requirement for BladeCenter cooling. It is a solution for clients that cannot upgrade a data center's air conditioning units because of 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 client-supplied water, using industry standard fittings and couplings.
- ▶ Up to 50,000 BTUs (or approximately 15 kW) of heat can be removed from the air that is exiting the back of a zBX rack.

The IBM Rear Door Heat eXchanger details are shown in Figure 11-7 on page 423.

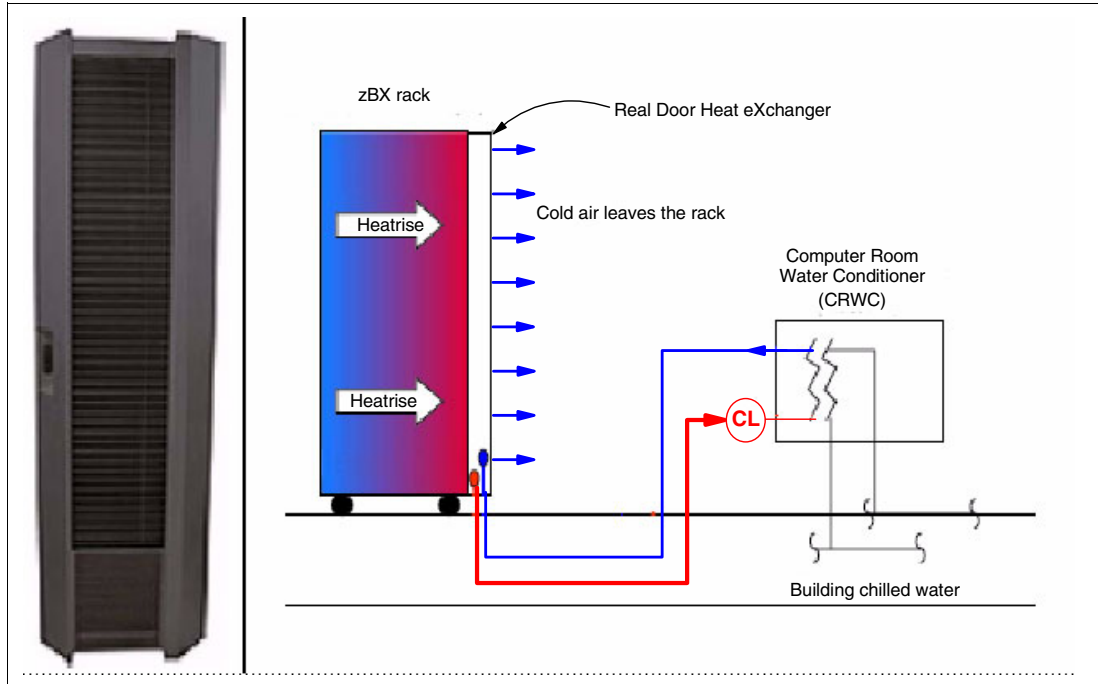


Figure 11-7 Rear Door Heat eXchanger (left) and functional diagram

The IBM Rear Door Heat eXchanger also offers a convenient way to handle hazardous “hot spots”, which can help you lower the total energy cost of the data center.

11.4.4 zBX physical specifications

The zBX solution is delivered either with one rack (Rack B) or four racks (Rack B, C, D, and E). Table 11-9 shows the physical dimensions of the zBX minimum and maximum solutions.

Table 11-9 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 feature (FC 0545)

This feature (Figure 11-8 on page 424) allows you to route I/O and power cabling through the top of the zBX rack. The feature will add 177 mm (7 in) to the height and 9.75 kg (21.5 lbs) to the weight of the zBX rack once installed. It can be ordered with a new zBX, but also added later. You require one feature per installed rack.

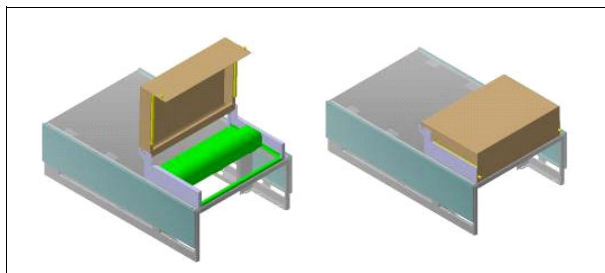


Figure 11-8 Top Exit Support feature for the zBX

Height Reduction feature

This feature (FC 0570) is required if you must reduce the shipping height for the zBX. Order it if you have doorways with openings less than 1941 mm (76.4 inches) high. It accommodates doorway openings as low as 1832 mm (72.1 inches).

zBX weight

Table 11-10 lists the maximum weights of fully populated zBX racks and BladeCenters.

Table 11-10 Weights of zBX racks

Rack description	Weight kgs (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)

Remember: A fully configured Rack B is heavier than a fully configured Rack C, D, or E because Rack B has the Top of Rack (ToR) switches installed.

For more information about the physical requirements for zBX, see *zBX Model 003 Installation Manual - Physical Planning*, GC27-2619.

11.5 Energy management

This section addresses the elements of energy management in areas of tooling to help you understand the requirements for power and cooling, monitoring and trending, and reducing power consumption. The energy management structure for the server is shown in Figure 11-9 on page 425.

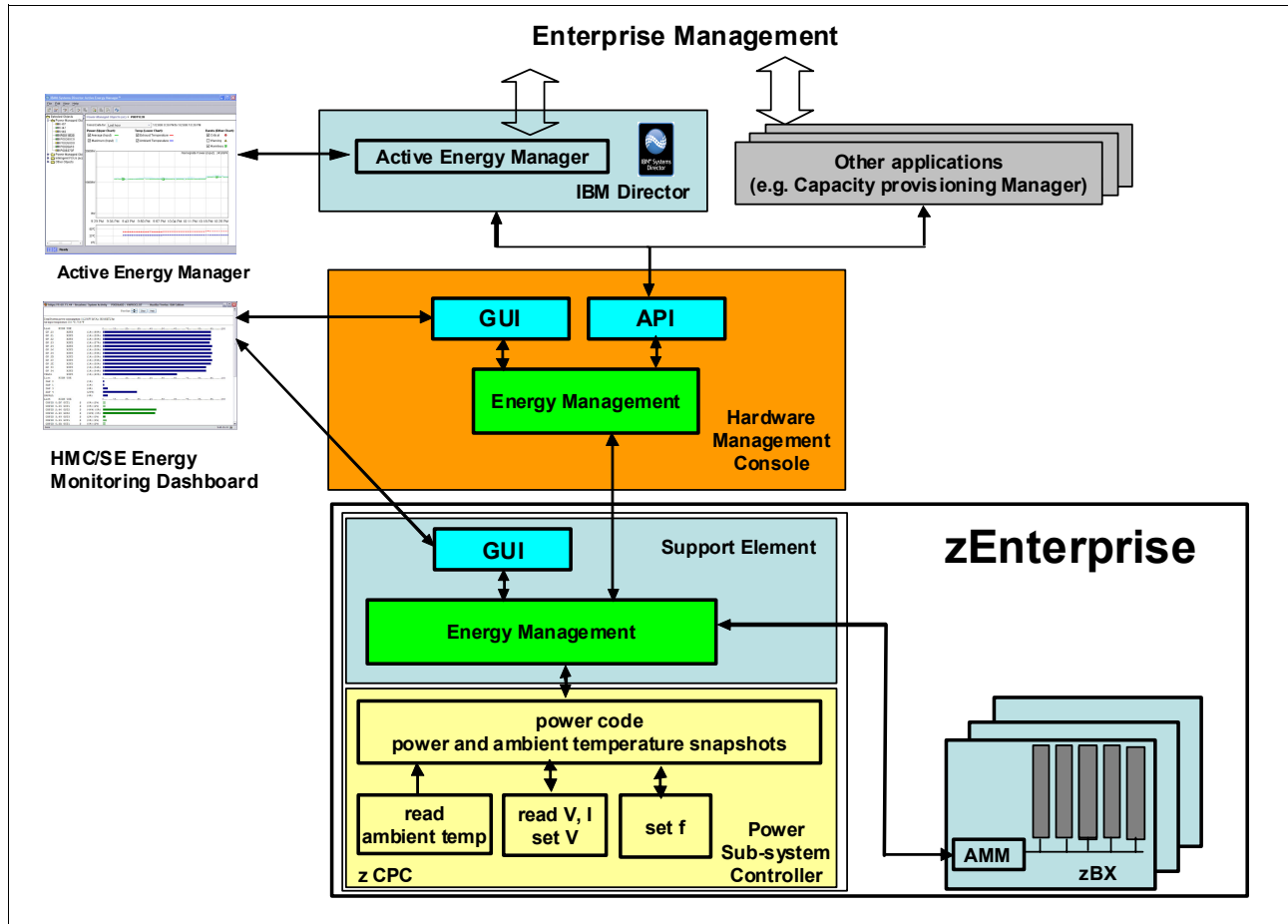


Figure 11-9 zEC12 energy management

The hardware components in the zEC12 and the optional zBX are monitored and managed by the Energy Management component in the Support Element (SE) and Hardware Management Console (HMC). The GUIs of the SE and HMC provide views, such as the System Activity Display or Monitors Dashboard. For example, through a Simple Network Management Protocol (SNMP) API, energy information is available to IBM Systems Director Active Energy Manager, a plug-in of IBM Systems Director. For more information, see 11.5.4, “IBM Systems Director Active Energy Manager” on page 427.

When Unified Resource Manager features are installed, several monitoring and control functions can be used to run Energy Management. For more information, see 12.7.1, “Unified Resource Manager” on page 461 and 11.5.5, “Unified Resource Manager: Energy management” on page 428.

A few aids are available to plan and monitor the power consumption and heat dissipation of the zEC12. The following tools are available to plan and monitor the energy consumption of the zEC12:

- ▶ Power estimation tool
- ▶ Query maximum potential power
- ▶ System Activity Display and Monitors Dashboard
- ▶ IBM Systems Director Active Energy Manager™

11.5.1 Power estimation tool

The power estimation tool for System z servers is available through the IBM Resource Link:
<http://www.ibm.com/servers/resourceLink>

The tool provides an estimate of the anticipated power consumption of a system model based on its configuration. You enter the system model, memory size, number of I/O cages, 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 system in the IBM Resource Link, you can access the Power Estimator tool through the system information page of that particular system. In the Tools section of the Resource Link, you also can enter the Power Estimator and enter any system configuration for which you want to calculate power requirements.

11.5.2 Query maximum potential power

The maximum potential power that is used by the system is less than the *Label Power*, as depicted in the atypical power usage report in Figure 11-10. The *Query maximum potential power* function shows your systems' maximum potential power usage and heat dissipation. By using this function, you can 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 CPC Details view of the HMC.

This function enables operations personnel with no System z knowledge to query the maximum power draw of the system. The implementation helps you avoid capping enforcement through dynamic capacity reduction. The client controls are implemented in the HMC, SE, and Active Energy Manager. Use this function with the Power Estimation tool that allows for pre-planning for power and cooling requirements. For more information, see 11.5.1, "Power estimation tool" on page 426.

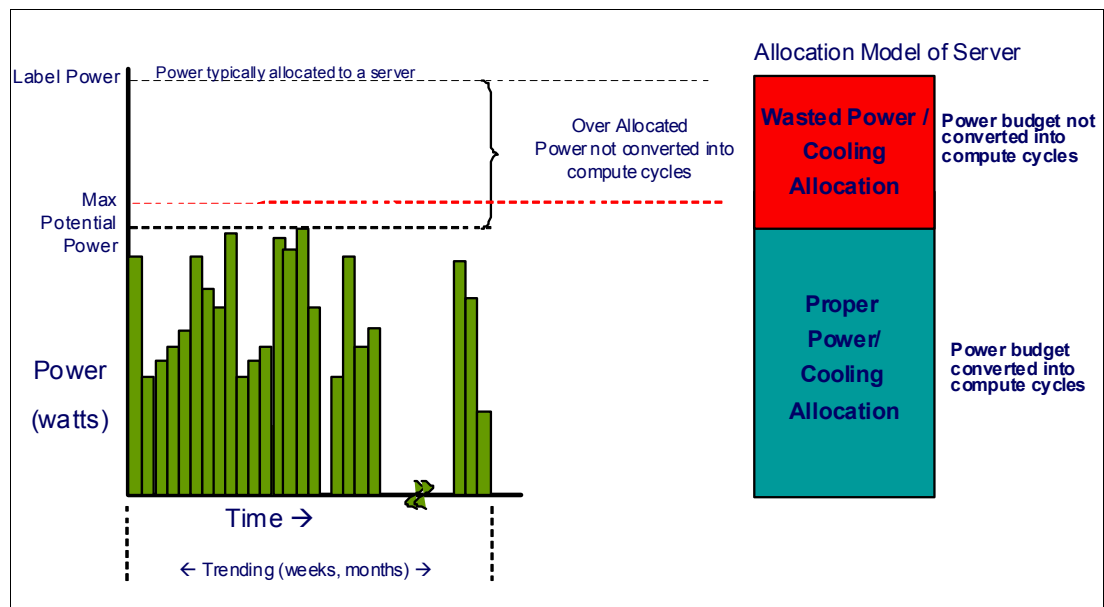


Figure 11-10 Maximum potential power

11.5.3 System Activity Display and Monitors Dashboard

The System Activity Display shows you, among other information, the current power usage as shown in Figure 11-11.

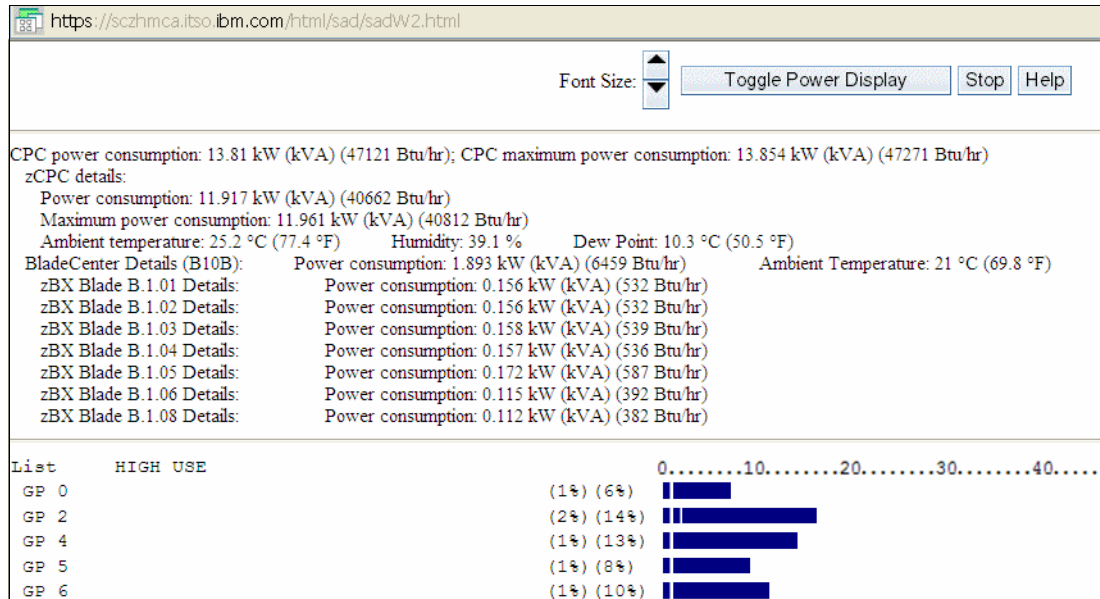


Figure 11-11 Power usage on the System Activity Display (SAD)

The Monitors Dashboard of the HMC displays power and other environmental data. It also provides a Dashboard Histogram Display, where you can see the trend for a particular interesting value, such as the power consumption of a blade or the ambient temperature of the zEC12.

11.5.4 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 client. Active Energy Manager is an industry-leading cornerstone of the IBM energy management framework.

Active Energy Manager Version 4.3.1 is a plug-in to IBM Systems Director Version 6.2.1, and is available for installation on Linux on System z. It can also run on Microsoft 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.

Use Active Energy Manager to monitor the power and environmental values of resources. It supports System z and other IBM products, such as IBM Power Systems and IBM System x. It also supports devices and hardware 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 is not directly connected to the System z servers. It is attached through a LAN connection to the HMC as shown in Figure 11-9 on page 425. For more information, see 12.4, "HMC and SE connectivity" on page 436. Active Energy Manager discovers the HMC that manages the server by using a discovery profile and by specifying the HMC's IP address and the SNMP credentials for that System z HMC. As the system is discovered, the System z 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 systems 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, you can more effectively manage the power of the systems while lowering the cost of computing.

The following data is available through Active Energy Manager:

- ▶ System name, system type, model, serial number, and firmware level of System z servers and optional zBXs attached to IBM zEnterprise Systems.
- ▶ Ambient temperature
- ▶ Exhaust temperature
- ▶ Average power usage
- ▶ Peak power usage
- ▶ Limited status and configuration information. This information helps explain the following types of changes, called *Events*, to the power consumption:
 - Changes in fan speed
 - Radiator and WCU failures
 - Changes between power off, power on, and initial machine load (IML) complete states
 - Number of books and I/O cages
 - Capacity Backup (CBU) record expirations

IBM Systems Director Active Energy Manager provides you with the data necessary to effectively manage power consumption in the data center. Active Energy Manager, an extension to IBM Director systems management software, monitors actual power usage and trend data for any single physical system or group of systems. Active Energy Manager uses monitoring circuitry, developed by IBM, to help identify how much actual power is being used and the temperature of the system.

11.5.5 Unified Resource Manager: Energy management

This section addresses the energy management capabilities of Unified Resource Manager.

Choice of suites

The energy management capabilities for Unified Resource Manager that can be used in an ensemble depend on which suite is installed in the ensemble:

- ▶ Manage suite (FC 0019)
- ▶ Automate/advanced management suite (FC 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 demands of the zEC12 system. Unified Resource Manager provides complete monitoring and trending capabilities for the zEC12 and the zBX by using one or more of the following options:

- ▶ Monitor dashboard
- ▶ Environmental Efficiency Statistics
- ▶ Details view

Automate/advanced management suite

The Unified Resource Manager offers multiple energy management tasks as part of the automate/advanced management suite. These tasks allow you to actually change the systems' behavior for optimized energy usage and energy savings:

- ▶ Power Cap
- ▶ Group Power Cap
- ▶ Power Save
- ▶ Group Power Save

Depending on the scope that is selected inside the Unified Resource Manager GUI, different options are available.

Set Power Cap function

The Set 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 actually throttling the processors in the system.

The Unified Resource Manager shows all components of an ensemble in the Set Power Cap window, as seen in Figure 11-12. Not all components used in a specific environment necessarily support power capping. Only those marked as “enabled” can actually run power capping functions.

A zEC12 does not support power capping, as opposed to specific blades, which can be power-capped. When capping is enabled for a zEC12, this capping level is used as a threshold. When the threshold is reached, a warning message is sent that informs you that the CPC went above the set cap level. The lower limit of the cap level is equal to the maximum potential power value. For more information, see 11.5.2, “Query maximum potential power” on page 426.

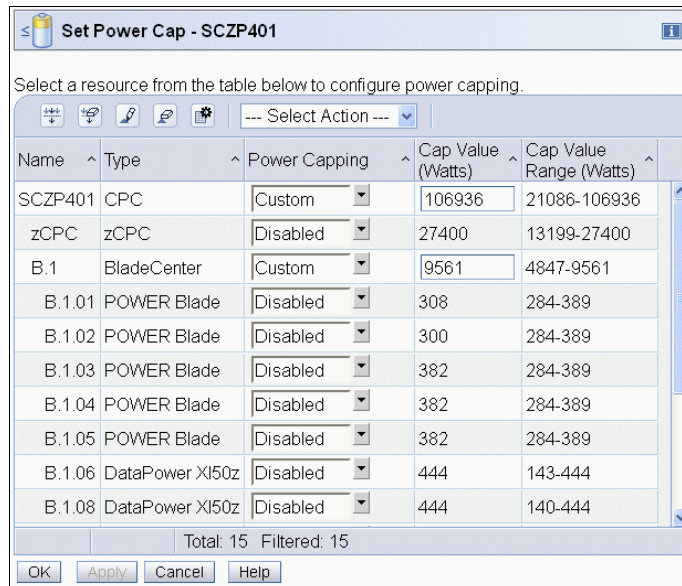


Figure 11-12 Set Power Cap window

Static power-saving mode

The server has a mechanism to vary frequency and voltage, originally developed to help avoid interruptions because of cooling failures. The mechanism can be used to reduce the energy consumption of the system in periods of low utilization and to partially power off systems designed mainly for disaster recovery. The mechanism is under the full control of the client. There is no autonomous function to run changes. The client controls are implemented in the HMC, SE, and Active Energy Manager with one power-saving mode. The expectation is that the frequency change is 15%, the voltage change is 8%, and the total power savings is 6% - 16%, depending on the configuration.

Figure 11-13 shows the Set Power Saving window.

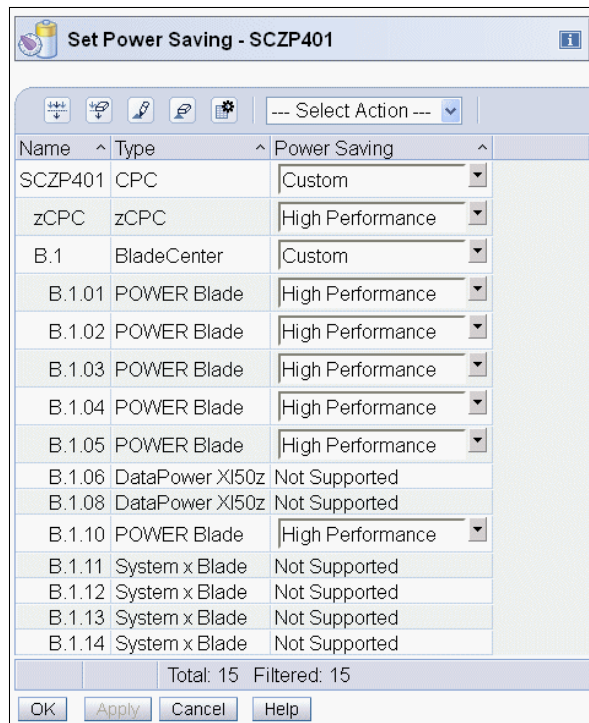


Figure 11-13 Set Power Saving window

For more information about Energy Management with Unified Resource Manager, see *IBM zEnterprise Unified Resource Manager, SG24-7921*.



Hardware Management Console and Support Element

The Hardware Management Console (HMC) supports many functions and tasks to extend the management capabilities of the IBM zEnterprise EC12 (zEC12). 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 zEnterprise BladeCenter Extension (zBX).

This chapter addresses the HMC and SE in general, and adds relevant information for HMCs that manage ensembles with the IBM zEnterprise Unified Resource Manager.

This chapter includes the following sections:

- ▶ Introduction to the HMC and SE
- ▶ HMC and SE enhancements and changes
- ▶ Remote Support Facility (RSF)
- ▶ HMC and SE remote operations
- ▶ HMC and SE key capabilities
- ▶ HMC in an ensemble

12.1 Introduction to the HMC and SE

The Hardware Management Console (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 System z CPCs. It manages System z hardware, its logical partitions (LPARs), and provides support applications. At least one HMC is required to operate an IBM System z[®] server. An HMC can manage multiple System z CPCs and can be at a local or a remote site.

If the zEC12 is defined as a member of an ensemble, a pair of HMCs is required (a primary and an alternate). When a zEC12 is defined as a member of an ensemble, certain restrictions apply. For more information, see 12.7, “HMC in an ensemble” on page 461.

The Support Elements (SEs) are two integrated notebook computers that are supplied with the zEC12. One is the primary SE and the other is the alternate SE. The primary SE is the active one. The alternate acts as the backup. The SEs are closed systems, just like the HMCs, 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 System z CPC. The SE then issues those commands to their CPC and controlled zBX (if any). One HMC can control up to 100 SEs and one SE can be controlled by up to 32 HMCs.

The microcode for the System z and zBX is managed through the HMC.

Some functions are only available on the SE. With Single Object Operations (SOO), these functions can be used from the HMC. See “Single Object Operating (SOO)” on page 442 for further details.

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 12.5, “Remote Support Facility (RSF)” on page 440.

12.2 SE driver support with the new HMC

The driver of the HMC and SE is always equivalent to a specific HMC and SE version, as illustrated in these examples:

- ▶ Driver 15 is equivalent to Version 2.12.1
- ▶ Driver 86 is equivalent to Version 2.11.0
- ▶ Driver 79 is equivalent to Version 2.10.2

At the time of this writing, a zEC12 is shipped with HMC Version 2.12.1, which can support different System z types. Some functions that are available on Version 2.12.1 and later are only supported when connected to a zEC12 with 2.12.1.

Table 12-1 shows a summary of the SE minimum drivers and versions that are supported by the new HMC Version 2.12.1 (driver 15).

Table 12-1 zEC12 HMC: System z support summary

System z family name	Machine type	Minimum SE driver	Minimum SE version
zBC12	2828	15	2.12.1
zEC12	2827	12	2.12.0
z114	2818	93	2.11.1
z196	2817	86	2.11.0
z10 BC	2098	79	2.10.2
z10 EC	2097	79	2.10.2
z9 BC	2096	67	2.9.2
z9 EC	2094	67	2.9.2
z890	2086	55	1.8.2
z990	2084	55	1.8.2

12.2.1 HMC Feature Code 0092 changes

The New Build Feature Code (FC) 0092 is an HMC that contains 16 GB of memory. The previous FC 0091 can be carried forward, but an HMC for the zEC12 needs 16 GB of memory. Some FC 0091s that shipped before the zEC12 have only 8 GB of memory. When Driver 15 is ordered for an existing FC 0091 HMC, the additional 8 GB of memory is provided if the HMC has only 8 GB of memory. HMCs that are older than FC 0091 are not supported for the zEC12 at Driver 12 or Driver 15.

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, 95 mm (3.75 in) longer.

12.3 HMC and SE enhancements and changes

The zEC12 comes with the new HMC application Version 2.12.1. Generally, use the “What’s New Wizard” to explore the new features that are available for each release. For a complete list of HMC and SE functions, see the System z HMC and SE (Version 2.12.1) Information Center:

<http://pic.dhe.ibm.com/infocenter/hwmca/v2r12m1/index.jsp>

The HMC and SE with Driver 15 has several enhancements and changes for the zEC12:

- Tasks and panels are updated to support the configuration and management of the new features introduced with the zEC12: Flash Express, IBM System z Advanced Workload Analysis Reporter (IBM zAware), zEnterprise Data Compression (zEDC) Express, and 10GbE Remote Direct Memory Access over Converged Ethernet (RoCE) Express.

- ▶ Updates to x3270 Support (also known as “Secure 3270”)

The “Configure 3270 Emulators” task, on the HMC and Trusted Key Entry (TKE) workstation consoles, is enhanced to verify the authenticity of the certificate returned by the 3270 server when a secure and encrypted Secure Sockets Layer (SSL) connection is established to an IBM host. For more information, see “Updates to x3270 support” on page 444.
- ▶ Enhanced IBM Service Support System

If the HMC and SE are at Driver 15, they can use a new remote infrastructure to connect via Remote Support Facility (RSF) for some tasks. This might require the client to change the network settings (proxy, firewall, and so on) for the RSF infrastructure. For more information, see 12.5.2, “RSF connections to IBM and the Enhanced IBM Service Support System” on page 441.
- ▶ Audit log changes

With Driver 15, the RSF security events moved from security to audit logs. The SSL connection information is logged, including the host name, host name on the certificate, and the cipher used.
- ▶ 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 ability to change the roles for a default user ID, create your own version by copying an existing default user ID.
- ▶ OSA/SF now available on the HMC

The Open Systems Adapter Support Facility (OSA/SF) is a component of z/OS, z/VM, and z/VSE.

With Driver 15, the task “OSA Advanced Facilities” on the HMC is enhanced to provide configuration, validation, activation, and display support exclusively for the OSA-Express5S and OSA-Express4S features.

OSA/SF on the HMC is required for the OSA-Express5S feature. Either OSA/SF on the HMC or the OSA/SF in the operating system component can be used for the OSA-Express4S features.

For detailed information, see *OSA/SF on the HMC*, SC14-7580.
- ▶ Help infrastructure updates

The content from the following publications is now incorporated into the Hardware Management Console (HMC) and Support Element (SE) (Version 2.12.1) help system:

 - *System z Hardware Management Console Operations Guide*
 - *zEnterprise System Hardware Management Console Operations Guide for Ensembles*
 - *zEnterprise System Support Element Operations Guide*

This information is also in the System z HMC and SE (Version 2.12.1) Information Center:
<http://pic.dhe.ibm.com/infocenter/hwmca/v2r12m1/index.jsp>
- ▶ Defined capacity of LPARs (absolute physical HW LPAR capacity setting)

Driver 15 introduces the ability to define, in the image profile for shared processors, the absolute processor capacity that the image is allowed to use (independent of partition 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. The absolute capping value can either be “None” or a value for the number of processors (value 0.01 - 255.0 can be specified).

► zBX firmware management

The zBX Model 003 is managed from the HMC and owning processor Support Element (SE) by using the zEnterprise Unified Resource Manager. The following support is provided:

- zBX firmware upgrades are downloaded from IBM RETAIN by using the HMC broadband Remote Support Facility (RSF) connection. Firmware updates are saved locally to be installed during a scheduled Microcode Change Level (MCL) apply session.
- Firmware updates are installed from the HMC and SEs using the same process and controls currently in use for System z.
- zBX hardware-related and firmware-related failures are reported to IBM and the IBM support structure is engaged, using the HMC RSF. This is the same process used for reporting System z problems.

► zBX lifecycle management:

- zBX Model 003 supports the same System x, POWER7, and DataPower X150z blade types supported in zBX Model 002.
- The new System x blades use a new 10Gb internal Ethernet adapter.
- The BladeCenter H advanced management module (AMM) is no longer used on a zBX Model 003. A new improved version of the AMM, called *AMMe*, replaced the hardware used in the zBX Model 002.
- All zBX Model 003 component firmware is upgraded.

► Server Time Protocol (STP): Improved SE time accuracy before initial microcode load (IML)

With Driver 15, if the CPC has not run the IML, the SE will take the time from the external time source (ETS) every hour, if the Network Time Protocol (NTP) servers are configured in the “ETS Configuration” tab in the task “System (Sysplex) Time”. This helps to ensure that the SE time is accurate before the IML, which is also known as the *power-on reset* (POR).

► STP NTP broadband security

Authentication is added to the HMC NTP communication with NTP time servers. For more information, see “HMC NTP broadband authentication support for the zEC12” on page 454.

► The environmental task has usability improvements regarding the time frame. For more information, see “Environmental Efficiency Statistics task” on page 449.

► Crypto Function integration in the Monitors Dashboard

For more information, see “The Monitors Dashboard task” on page 448.

► Removal of modem support from the HMC

This change affects clients that have set up the modem for RSF or for STP NTP access. For more information, see 12.5, “Remote Support Facility (RSF)” on page 440 and 12.6.10, “NTP client and server support on the HMC” on page 454.

► Installation and activation by maintenance change level (MCL) bundle target

For more information, see “Microcode installation by MCL bundle target” on page 446.

► A confirmation panel before processing an Alt-Ctrl-Delete request is added.

Note: If an HMC must be rebooted, always use the **Shutdown and Restart** task on the HMC to avoid any file corruption.

- ▶ The capability to modify the time of the SE mirror scheduled operation is added.
- ▶ It is now possible to allow the mass deletion of messages from the Operating System Messages task.
- ▶ The Network Settings task is updated to clearly show the ordering of the routing table entries.
- ▶ Remove support for the Coprocessor Group Counter Sets

In the zEC12, each physical processor has its own crypto coprocessor. They no longer must share this coprocessor with another PU. The Coprocessor Group Counter Sets of the counter facilities will not be available. All of the necessary crypto counter information is available in the crypto activity counter sets directly. The check box selection for the Coprocessor Group Counter Sets is removed from the Image profile definition and the Change Logical Partition Security task.

12.3.1 HMC media support

The HMC provides a DVD-RAM drive and, with HMC Version 2.11.0, a USB flash memory drive (UFD) was introduced as an alternative. The tasks that require access to a DVD-RAM drive now can access a UFD. There can be more than one UFD inserted into the HMC.

12.3.2 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 popular in newer operating systems, and features context-based task launching. The *Classic Style User Interface* uses the drag-and-drop interface style.

Tutorials: The IBM Resource Link^a provides tutorials that demonstrate how to change from the Classic to the Tree Style Interface, and introduce the function of the Tree Style Interface on the HMC:

<https://www-304.ibm.com/servers/resourceLink/hom03010.nsf/pages/education?OpenDocument>

a. Registration is required to access the IBM Resource Link.

12.4 HMC and SE connectivity

The HMC has two Ethernet adapters, which are supported by HMC Version 2.12.1 for connectivity to up to two different Ethernet LANs.

The SEs on the zEC12 are connected to the Bulk Power Hub (BPH). The HMC to BPH communication is only possible through an Ethernet switch. Other System z servers and HMCs can also be connected to the switch. To provide redundancy for the HMCs, install two switches.

Only the switch (and not the HMC directly) can be connected to the BPH ports J02 and J01 for the customer networks 1 and 2.

Table 12-1 on page 433 shows the connectivity between the HMC and the SE.

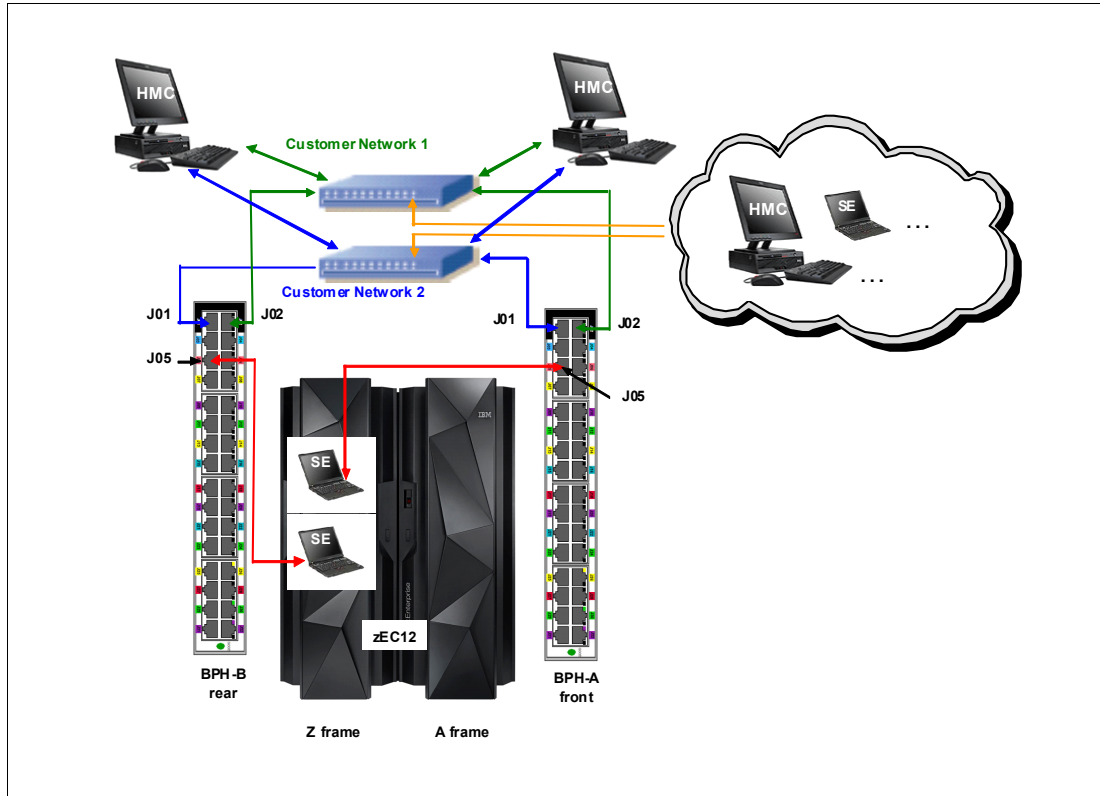


Figure 12-1 HMC to SE connectivity

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 document. Whenever the networks are interconnected, security exposures can exist. The document “IBM System z HMC Security” provides information about HMC security. It is available on the IBM Resource Link^a:

[https://www-304.ibm.com/servers/resourceLink/lib03011.nsf/pages/zHmcSecurity/\\$file/zHMCSecurity.pdf](https://www-304.ibm.com/servers/resourceLink/lib03011.nsf/pages/zHmcSecurity/$file/zHMCSecurity.pdf)

For more information about the HMC settings related to access and security, see the System z HMC and SE (Version 2.12.1) Information Center:

<http://pic.dhe.ibm.com/infocenter/hwmca/v2r12m1/index.jsp>

a. Registration is required to access the IBM Resource Link.

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 System z capabilities benefit from the various 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 that can be used for HMC user authentication
- ▶ NTP client/server support
- ▶ RSF through broadband
- ▶ HMC access via remote web browser
- ▶ Enablement of the Simple Network Management Protocol (SNMP) and Common Information Model (CIM) APIs to support automation or management applications, such as IBM System Director Active Energy Manager (AEM)

These examples are shown in Figure 12-2.

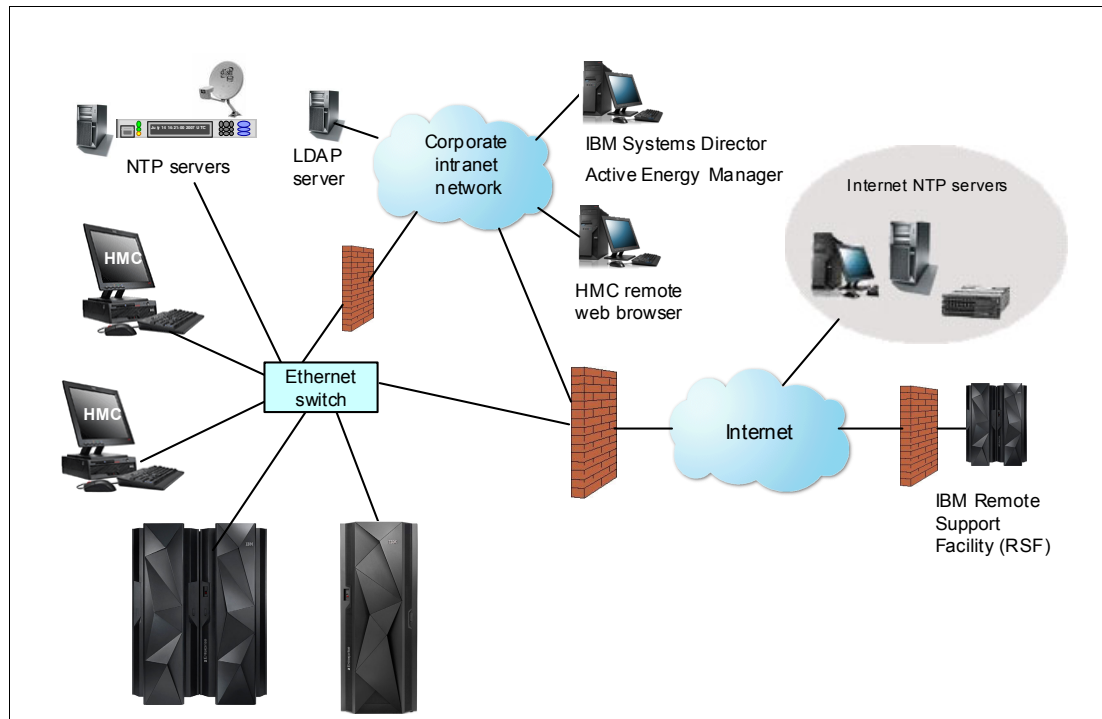


Figure 12-2 HMC connectivity examples

For more information, see the following documentation:

- ▶ System z HMC and SE (Version 2.12.1) Information Center:
<http://pic.dhe.ibm.com/infocenter/hwmca/v2r12m1/index.jsp>
- ▶ 11.5.4, “IBM Systems Director Active Energy Manager” on page 427
- ▶ *Installation Manual - Physical Planning 2827*, GC28-6914

12.4.1 Hardware prerequisite changes

The following HMC changes are important for the zEC12:

- ▶ No HMC LAN switches can be ordered from IBM
- ▶ RSF is broadband-only

No HMC LAN switches can be ordered from IBM

You can no longer order the Ethernet switches that are required by the HMCs to connect to the zEC12. You must provide them yourself. Existing supported switches can still be used, however.

Ethernet switches/hubs typically have these characteristics:

- ▶ Sixteen auto-negotiation ports
- ▶ 10/100/1000 Mbps data rate
- ▶ Full or half duplex operation
- ▶ Auto-medium-dependent interface crossover (MDIX) on all ports
- ▶ Port status LEDs

RSF is broadband-only

RSF through a modem *is not supported* on the zEC12 HMC. Broadband is needed for hardware problem reporting and service. For more information, see 12.5, “Remote Support Facility (RSF)” on page 440.

12.4.2 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 an SE is unnecessary if the SE only communicates 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.

12.4.3 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 (HCP)-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.

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 can also 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 browsers.

12.5 Remote Support Facility (RSF)

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

Restriction: RSF through a modem *is not supported* on the zEC12 HMC. Broadband connectivity is needed for hardware problem reporting and service. Future HMC hardware will not include modem hardware. Modems on installed HMC FC 0091 hardware will not work with HMC Version 2.12.0, which is required to support the zEC12.

12.5.1 Security characteristics

The following security characteristics are in effect:

- ▶ RSF requests are always 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 in high-grade Transport Layer Security (TLS)/Secure Sockets Layer (SSL) encryption.
- ▶ When starting the SSL/TLS-encrypted connection, the HMC validates the trusted host with its 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.

¹ For a client-supplied switch, multicast must be enabled at the switch level.

Additional resource: For more information about the benefits of broadband RSF and SSL/TLS-secured protocol, and a sample configuration for the Broadband RSF connection, see the IBM Resource Link^a:

<https://www-304.ibm.com/servers/resourceLink/lib03011.nsf/pages/zHmcBroadbandRsF0verview>

a. Registration is required to access the IBM Resource Link.

12.5.2 RSF connections to IBM and the Enhanced IBM Service Support System

If the HMC and SE are at Driver 15, they can use a new remote infrastructure at IBM when the HMC connects via RSF for certain tasks. To use the Enhanced IBM Service Support System and the current available connections, it is required that you check the network infrastructure settings.

At the time of this writing, RSF still uses the “traditional” RETAIN connection. We suggest that you 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, we suggest that the HMC has access to the Internet to IBM via RSF in the following manner:

- ▶ Via IP labels, which are also known as host names (preferred method):
 - www-945.ibm.com, port 443
 - esupport.ibm.com, port 443
- ▶ For IP addresses (for redundancy if the domain name server (DNS) is not available). Also, check the *Installation Manual - Physical Planning 2827*, GC28-6914.
- ▶ 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
 - 129.42.54.129:443
 - 129.42.56.129:443
 - 129.42.58.129:443
 - 129.42.60.129:443
 - 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
 - 2620:0:6C0:200:129:42:56:189:443
 - 2630:0:6C1:200:129:42:58:189:443
 - 2630:0:6C2:200:129:42:60:189:443
 - 2620:0:6C4:200:129:42:58:189:443

Note: All other previous existing IP addresses are no longer supported.

Host name resolving and domain name server (DNS)

If the HMC initiates an RSF Internet connection to IBM and the DNS name needs to be resolved, there are two methods:

- ▶ The HMC has a DNS configured and the option “Resolve IBM IP addresses on console” is selected.
- ▶ The configured SSL proxy has access to a DNS server.

12.5.3 HMC and SE remote operations

There are two ways 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 client-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 12.5, “Remote Support Facility (RSF)” on page 440.

- ▶ Using a web browser to connect to an HMC

The zEC12 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. USB flash media drive (UFD) access needs physical access, and you cannot shut down or restart the HMC from a remote location. 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 Goggle Chrome were tested as remote browsers. For detailed web browser requirements, see the System z HMC and SE (Version 2.12.1) Information Center:

<http://pic.dhe.ibm.com/infocenter/hwmca/v2r12m1/index.jsp>

Single Object Operating (SOO)

It is not necessary to be physically close to an SE to use it. The HMC can be used to access the SE remotely by using the Single Object Operation (SOO). The interface is the same as the one on the SE. For more information, see the System z HMC and SE (Version 2.12.1) Information Center:

<http://pic.dhe.ibm.com/infocenter/hwmca/v2r12m1/index.jsp>

12.6 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 System z HMC and SE (Version 2.12.1) Information Center:

<http://pic.dhe.ibm.com/infocenter/hwmca/v2r12m1/index.jsp>

12.6.1 Central processor complex (CPC) 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 zEC12 is powered on and off 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.

12.6.2 Logical partition management

Use the HMC to define logical partition (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 a Load task on the HMC to IPL an operating system. It causes a program to be read from a designated device, and starts that program. You can IPL the operating system from storage, the HMC DVD-RAM drive, the USB flash memory drive (UFD), or a File Transfer Protocol (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 be also 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 the zEC12 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 introduces 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. The absolute capping value can either be "None" or a value for the number of processors (0.01 - 255.0 can be specified).

12.6.3 Operating system communication

The Operating System Messages task displays messages from an LPAR. You can also enter operating system commands and interact with the system. This task is especially valuable to enter 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 is enhanced with Driver 15 to verify the authenticity of the certificate returned by the 3270 server when a secure and encrypted SSL connection is established to an IBM host. This is also known as *Secure 3270*.

Use the Certificate Management task if the certificates returned by the 3270 server are not signed by a well-known trusted certificate authority (CA) certificate, such as VeriSign or Geotrust. The advanced action, Manage Trusted Signing Certificates, within the Certificate Management task, is used to add trusted signing certificates.

For example, if the certificate associated with the 3270 server on the IBM host is signed and issued by a corporate certificate, it needs to be imported as shown in Figure 12-3.

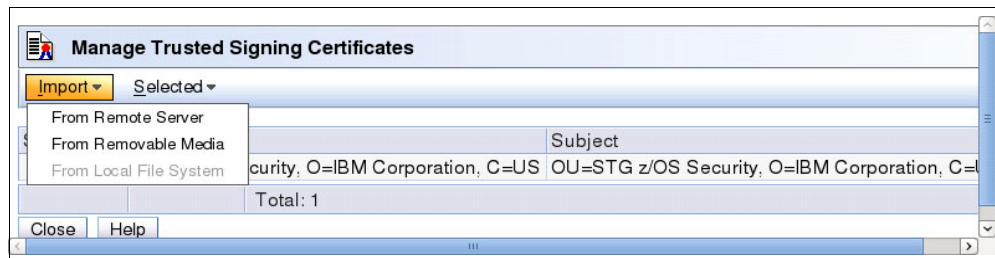


Figure 12-3 Manage Trusted Signing Certificates

If the connection between the console and the IBM host can be trusted at the time of importing the certificate, the import from the remote server option can be used as you can see in Figure 12-4. Otherwise, import the certificate from removable media.

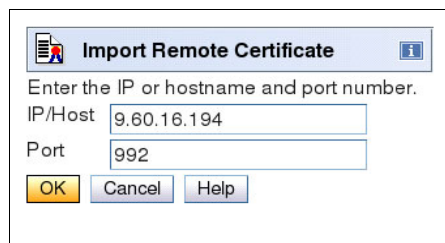


Figure 12-4 Import Remote Certificate example

A secure Telnet connection is established by prepending L: to the IP address:port of the IBM host as shown in Figure 12-5 on page 445.

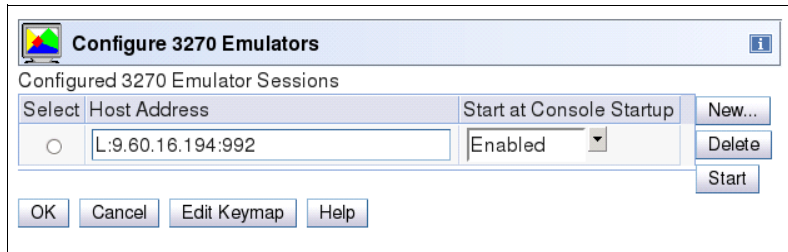


Figure 12-5 Configure 3270 Emulators

12.6.4 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, enhanced driver maintenance (EDM). EDM provides also the installation of the latest functions and the patches (MCLs) of the new driver.

When you perform a driver upgrade, always check the “Driver xx Customer Exception Letter” in the “Fixes” section on IBM Resource Link.

For more information, see 10.4, “zEC12 enhanced driver maintenance (EDM)” on page 399.

Microcode Change Level (MCL)

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 following link in IBM Resource Link^a provides access to the system information for your System z according to the system availability data that is sent on a scheduled basis. It provides more information about the MCL status of your zEC12:

<https://www-304.ibm.com/servers/resourceLink/svc0303a.nsf/fwebsearchstart?openform>

a. Registration is required to access the IBM Resource Link.

Microcode terms

The microcode has these characteristics:

- ▶ The driver contains engineering change (EC) streams.
- ▶ Each EC stream covers the code for a specific component of the zEC12. It has a specific name and an ascending number.
- ▶ The EC stream name and a specific number are one Microcode Change Level (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 12-6 shows how the driver, bundle, EC stream, MCL, and MCFs interact with each other.

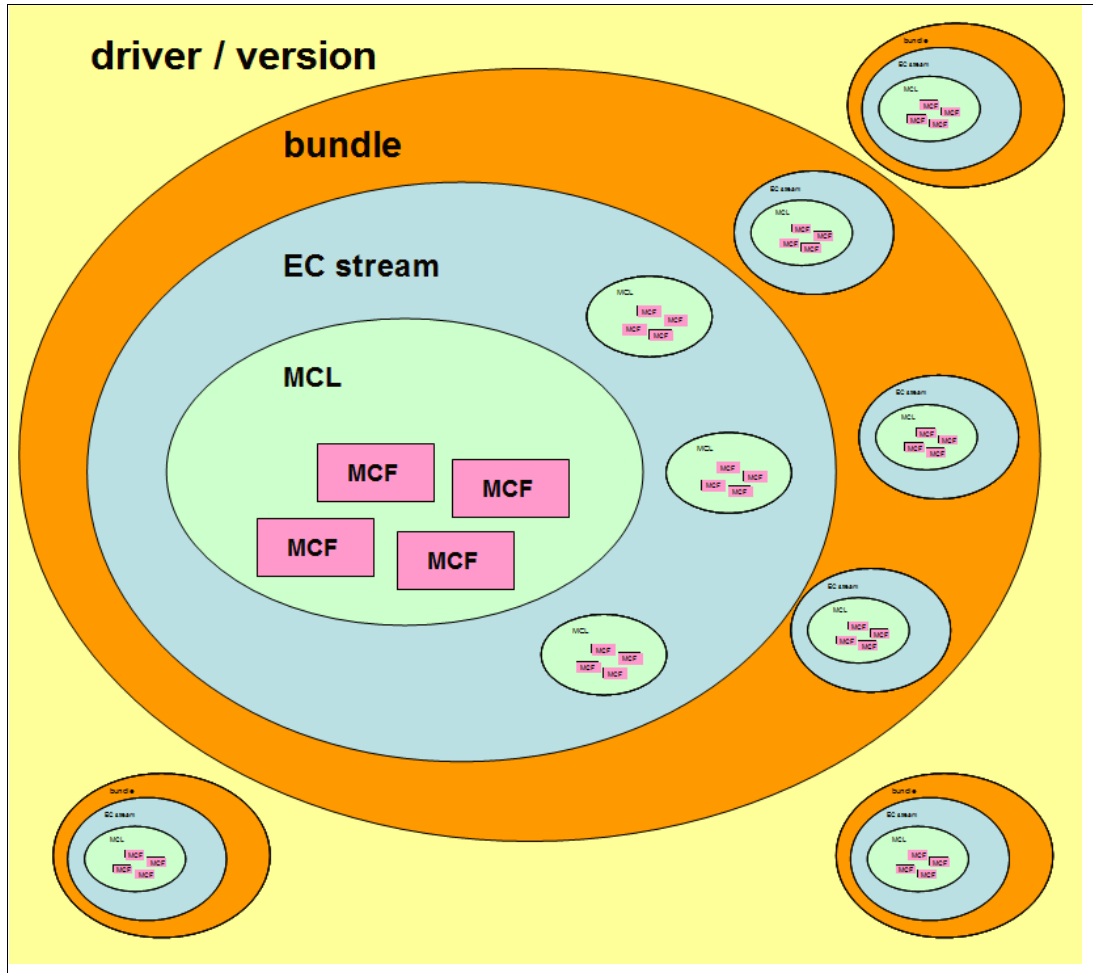


Figure 12-6 Microcode terms and interaction

Microcode installation by MCL bundle target

A *bundle* is a set of MCLs 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 12-7 on page 447.

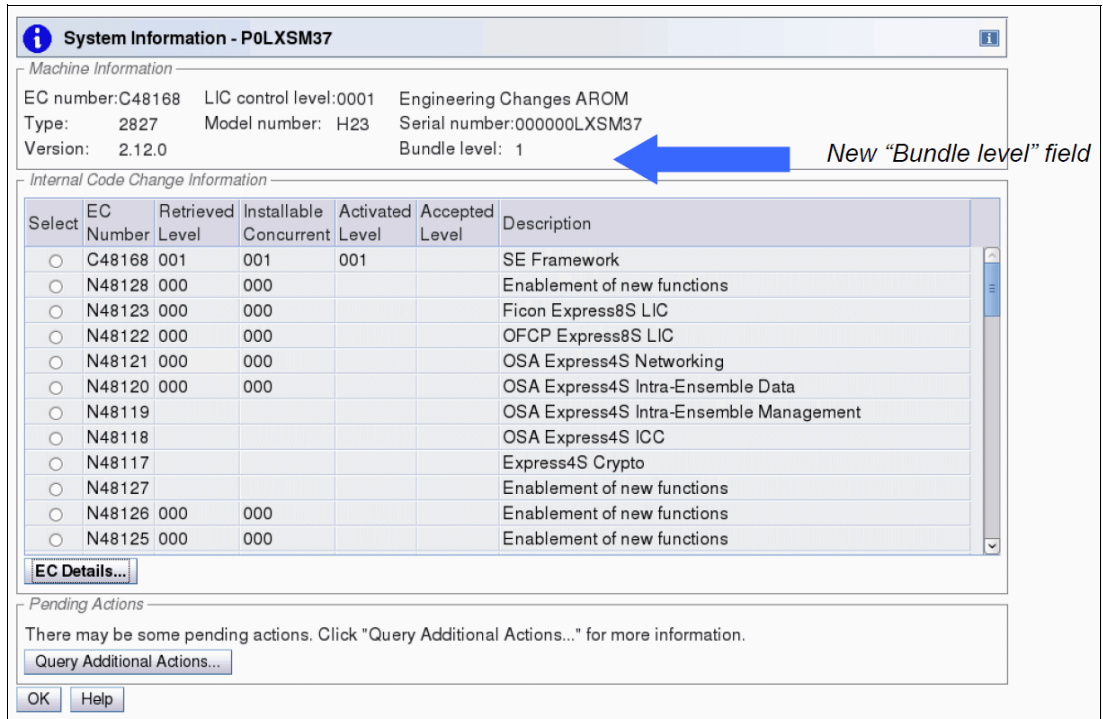


Figure 12-7 System Information: Bundle level

12.6.5 Monitoring

This section addresses monitoring considerations.

Monitor Task Group

The task group named Monitor on the HMC and SE includes monitoring-related tasks for the zEC12 as shown in Figure 12-8.

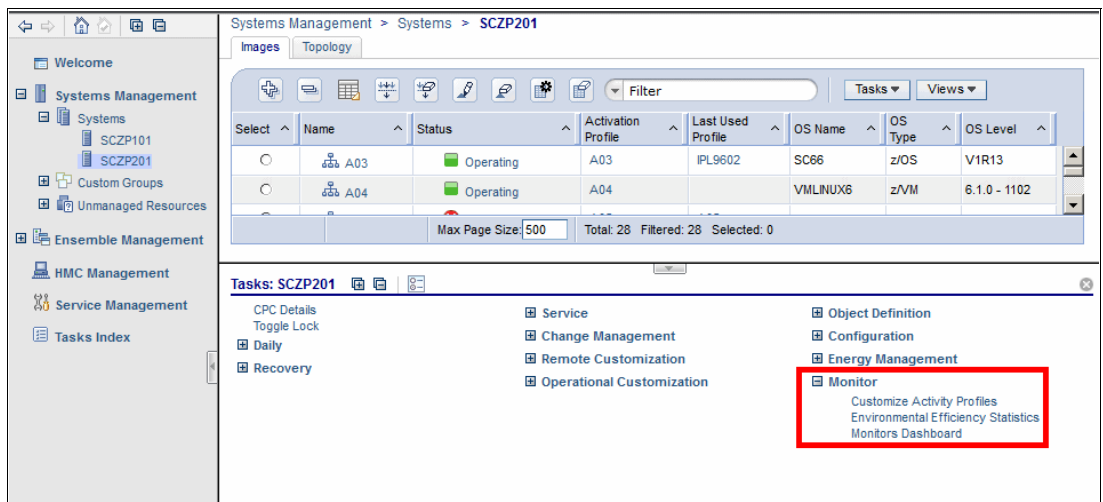


Figure 12-8 HMC Monitor Task Group

Customize Activity Profiles

Use the Customize Activity Profiles task to set profiles that are based on your monitoring requirements. Multiple activity profiles can be defined.

The Monitors Dashboard task

The Monitors Dashboard supersedes the System Activity Display (SAD). In the zEC12, 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 12-9 shows an example of the Monitors Dashboard task.

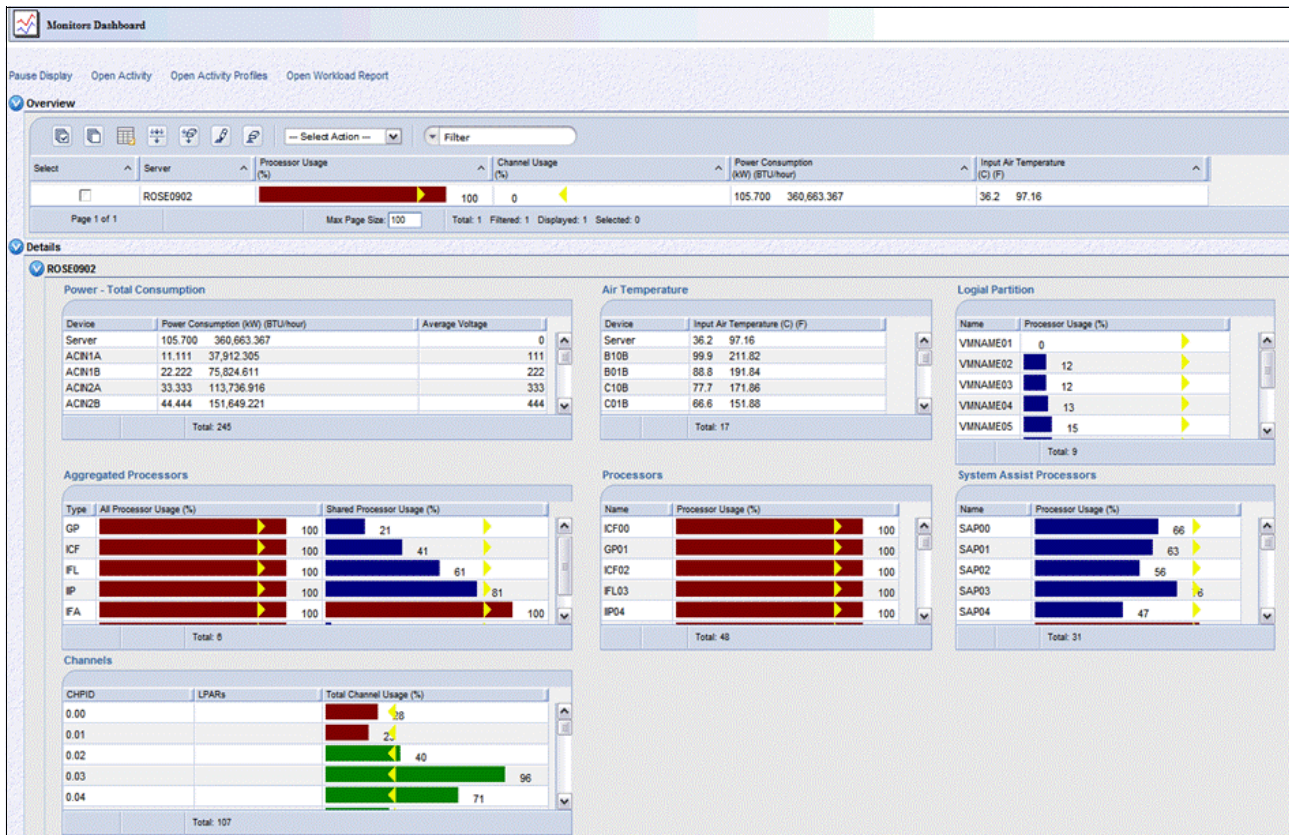


Figure 12-9 Monitors Dashboard task

With the zEC12, the Monitors Dashboard is enhanced with an adapters table. 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 utilization rate on a system-wide basis, not per LPAR, as shown in Figure 12-10 on page 449.

Select	Channel ID	Type	Adapter Usage (%)
<input type="checkbox"/>	0500	Crypto (ID = 0)	81
<input type="checkbox"/>	0501	Crypto (ID = 1)	97
<input type="checkbox"/>	0280	Crypto (ID = 3)	100
<input type="checkbox"/>	0281	Crypto (ID = 4)	30
<input type="checkbox"/>	032C	Crypto (ID = 5)	0

Page 1 of 1 Max Page Size: 100 Total: 6 Filtered: 6 Displayed: 6

Figure 12-10 Monitors Dashboard: Crypto function integration

For Flash Express, a new window is added as shown in Figure 12-11.

Select	Channel ID	Type	Adapter Usage (%)
<input type="checkbox"/>	0500	Flash Express	0
<input type="checkbox"/>	052C	Flash Express	0
<input type="checkbox"/>	0580	Flash Express	0
<input type="checkbox"/>	05AC	Flash Express	0

Page 1 of 1 Max Page Size: 100 Total: 4 Filtered: 4 Displayed: 4 Selected: 0

Figure 12-11 Monitors Dashboard: Flash Express function integration

Environmental Efficiency Statistics task

The Environmental Efficiency Statistics task (Figure 12-12 on page 450) 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.

Before the zEC12, when the data is first shown (the default is one day), the chart displayed data from midnight of the prior day to midnight of the current day. In the zEC12, the initial chart display shows the 24 hours before the current time so that a full 24 hours of recent data is displayed.

The panel is enhanced with the capability to specify a starting time.

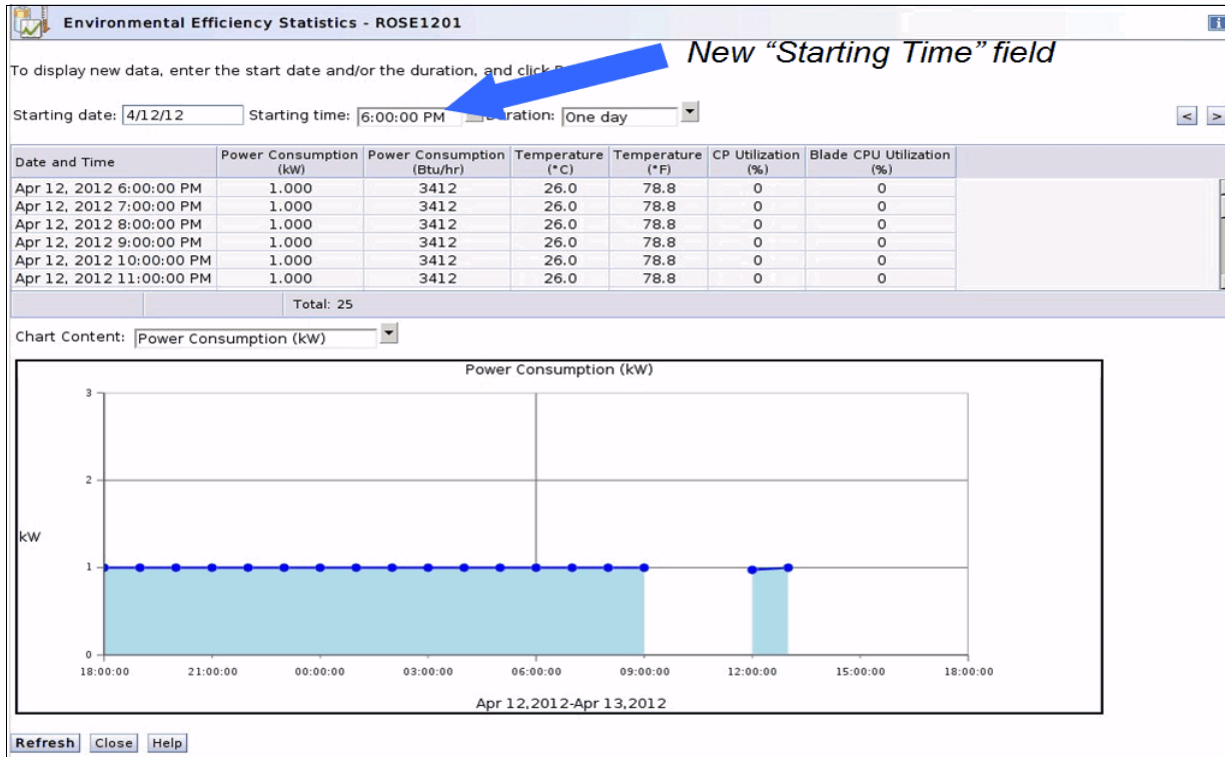


Figure 12-12 Environmental Efficiency Statistics

12.6.6 IBM Mobile Systems Remote

IBM Remote is a no-charge mobile application developed by IBM, which is now also able to help you monitor and manage your zEnterprise environment by using your mobile device (smartphone or tablet) as shown in Figure 12-13 on page 451.

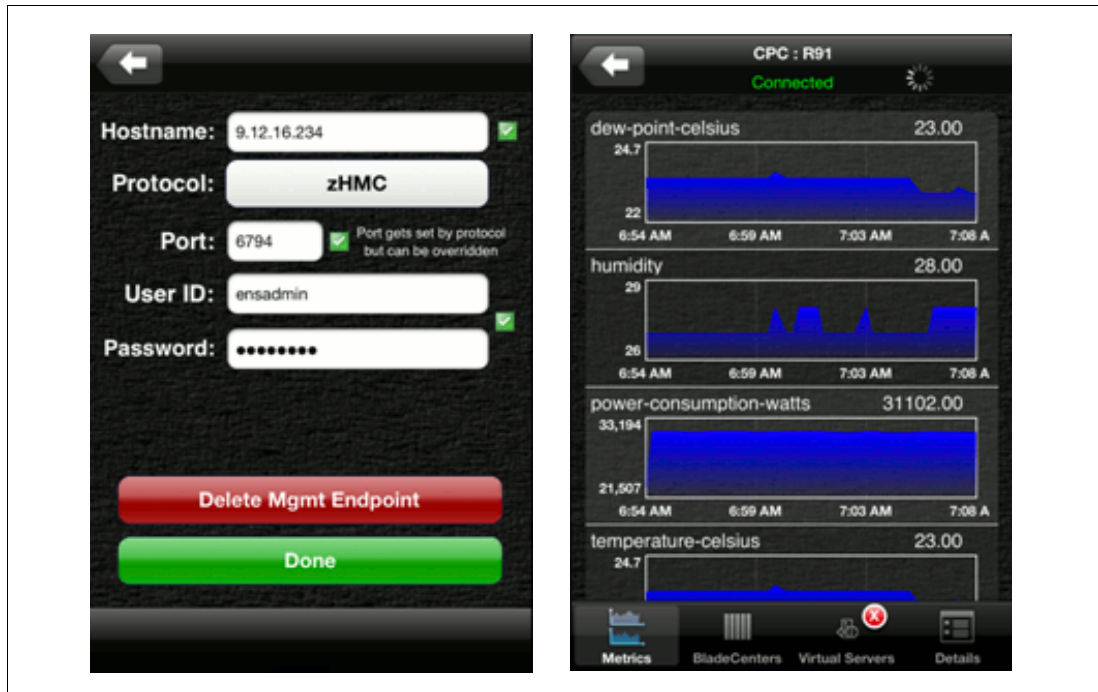


Figure 12-13 Sample screen captures of the application

By interacting with the zEnterprise HMC, the application allows you to access almost all the information that you normally view on the HMC in the palm of your hand. You can monitor your zEnterprise CPC. If you have an ensemble, you can monitor the IBM BladeCenters and installed blades in your zBX.

You can monitor these entities:

- ▶ Ensemble
- ▶ zEnterprise CPC
- ▶ BladeCenter
- ▶ Individual blade
- ▶ Workload
- ▶ Virtual server

Depending on the type of entity, you can display its health, details, and metrics, such as the power consumption and ambient temperature.

For more information and links to the application stores, see the IBM Mobile Systems Remote website:

<http://ibmremote.com/>

12.6.7 Capacity on Demand (CoD) support

All CoD upgrades are performed from 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 IBM zEnterprise EC12 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 panel features (accessed through HMC Single Object Operations):
 - Panel controls for granular activation and deactivation
 - History panel for all CoD actions
 - Description editing of CoD records
- ▶ HMC/SE Version 2.12.0 provides the following CoD information:
 - Millions of service units (MSU) and processor tokens
 - Last activation time
 - Pending resources shown by processor type instead of only a total count
 - Option to show details of installed and staged permanent records
 - More details for the Attention state by providing seven more flags

New since SE Version 2.12.0: Some preselected defaults are removed. Specifying each selection in the window is required.

HMC and SE are a part of the z/OS Capacity Provisioning environment. The Capacity Provisioning Manager (CPM) communicates with the HMC through System z APIs, and enters CoD requests. For this reason, SNMP must be configured and enabled 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*
- ▶ *zEnterprise System Capacity on Demand User's Guide, SC28-2605*

12.6.8 Feature on Demand (FoD) support

FoD is a new centralized way to flexibly entitle 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 client has purchased. On the z196/z114, the zBX HWMs are stored in the processor and memory LICCC record. On the zEC12, they are in the Feature on Demand record.

FoD allows separate LICCC controls for System z processors (central processors (CPs), Integrated Facility for Linux (IFL), System z Application Assist Processors (zAAPs), and System z Integrated Information Processors (zIIPs)), and zBX high water marks (HWMs), providing entitlement controls for each blade type. It is also used as LICCC support for the following features:

- ▶ zAware: Enablement/max connections
- ▶ Base/proprietary service: Expiration date
- ▶ New features: Yet to be announced or developed

12.6.9 Server Time Protocol support

With the Server Time Protocol (STP) functions, the role of the HMC is extended to provide the user interface for managing the Coordinated Timing Network (CTN):

- ▶ The zEC12 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 zEC12 server. If STP uses an NTP server without PPS, a time accuracy of 100 milliseconds to the ETS is maintained.
- ▶ You can have a zEC12 as a Stratum 2 or Stratum 3 server in a Mixed CTN linked to z10s (STP configured) attached to the Sysplex Timer operating as Stratum 1 servers. In this configuration, use two Stratum 1 servers to provide redundancy and avoid a single point of failure.
- ▶ The zEC12 cannot be in the same CTN with a System z9 ($n-2$) or earlier systems.

Figure 12-14 shows the coexistence that is supported by the zEC12 and previous System z servers in relation to the sysplex and STP.

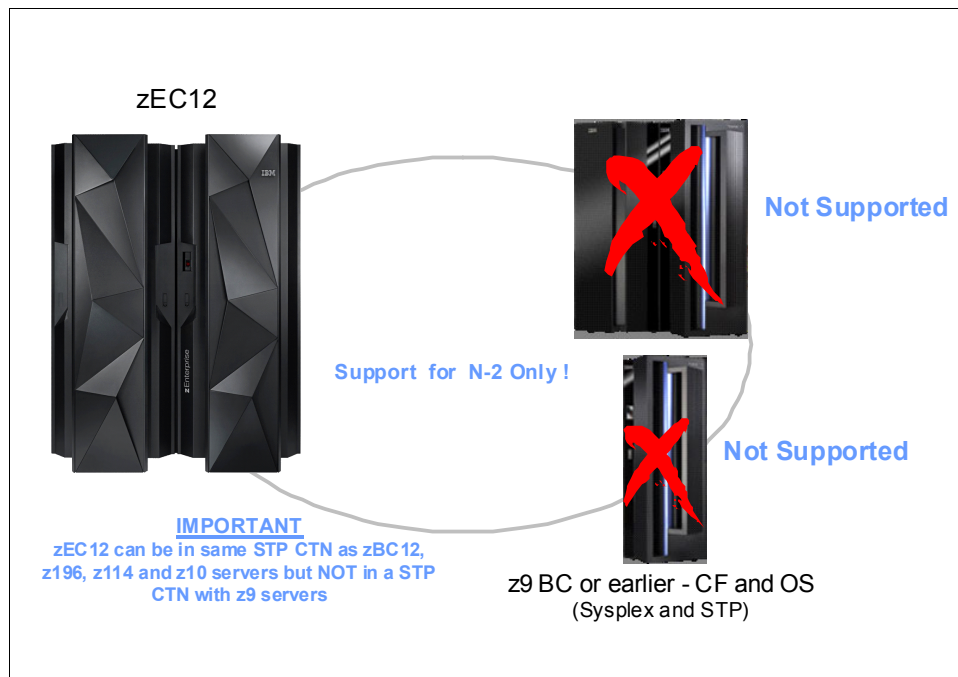


Figure 12-14 Parallel Sysplex System z coexistence

In an 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, as well as 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 initialized for STP message exchanges.

- ▶ For diagnostic purposes, the Pulse per Second port state on a zEC12 can be displayed and fenced ports can be reset individually.

STP recovery is enhanced since zEnterprise. For more information, see “STP recovery enhancement” on page 178.

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

12.6.10 NTP client and server support on the HMC

The Network Time Protocol (NTP) client support allows an STP-only Coordinated Timing Network (CTN) to use an NTP server as an External Time Source (ETS).

Restriction: The ETS connection through a modem is not supported on the zEC12 HMC.

This capability addresses the following requirements:

- ▶ Clients that want time accuracy for the STP-only CTN
- ▶ Clients that 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 System z (such as AIX, Microsoft Windows, and others) that have NTP clients.

The HMC can act as an NTP server. With this support, the zEC12 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 separate from the HMC/SE LAN can be used.

HMC NTP broadband authentication support for the zEC12

HMC NTP authentication can now be used with HMC Level 2.12.1. 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 provides 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 12-15.

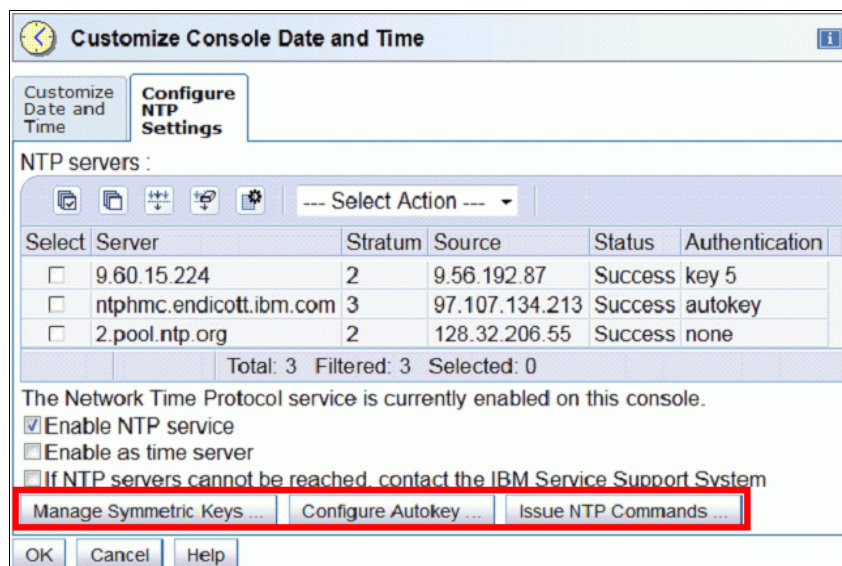


Figure 12-15 HMC NTP broadband authentication support

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 exchanging data keep this key to themselves. Messages encrypted with a secret key can be only decrypted with the same secret key. Symmetric key authentication does support 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 Network Address Translation (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 zEC12 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 configured as the ETS in an STP-only CTN. This configuration is shown in Figure 12-16. For more information, see the *Server Time Protocol Planning Guide*, SG24-7280.

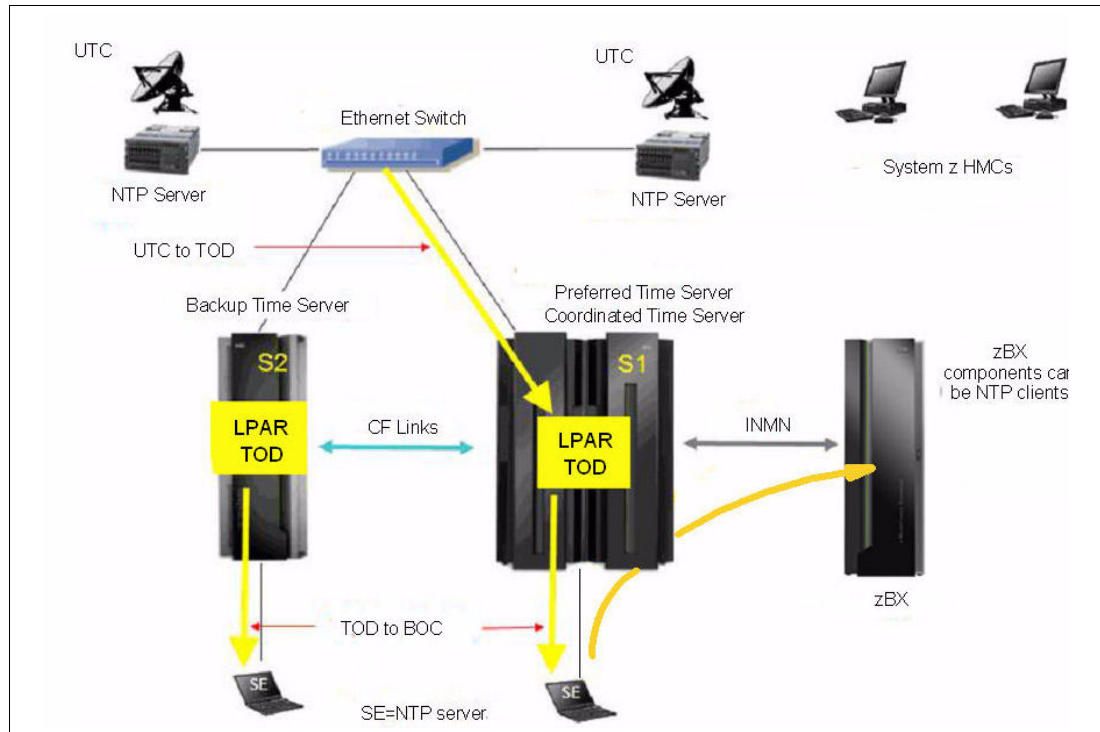


Figure 12-16 Time coordination for zBX components

12.6.11 Security and user ID management

This section addresses security and user ID management considerations.

HMC and SE security audit improvements

With the Audit & 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 used for both hardware and operating system messages.

With the zEC12, you can offload the following HMC and SE log files for client audit:

- ▶ Console event log
- ▶ Console service history
- ▶ Tasks performed log
- ▶ Security logs
- ▶ System log

Full logoff load and delta logoff load (since 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 & Log Management task or scheduled by the Scheduled Operations task. The data can be offloaded in the HTML and XML formats.

HMC user ID templates and LDAP user authentication

Lightweight Directory Access Protocol LDAP user authentication and HMC user ID templates enable the addition and removal of HMC users according to your own 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 that 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 ID’s 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.

Secure FTP infrastructure allows HMC and SE applications to query if 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 box is displayed that points to the Manage SSH Keys task to input one. The following tasks provide this support:

- ▶ Import/Export IOCDS
- ▶ Advanced Facilities FTP IBM Content Collector (ICC) Load
- ▶ Audit and Log Management (Scheduled Operations only)

12.6.12 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 necessary 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 options. With the tool, data is formatted and displayed in five different views. The tool provides various sort options, and data can be exported to a USB flash memory drive (UFD) for later viewing.

The following five views are available:

- ▶ PCHID Control Unit View shows PCHIDs, channel subsystems (CSS), 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.

12.6.13 Automated operations

As an alternative to manual operations, an application can interact with the HMC and SE through an application programming interface (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 Simple Network Management Protocol (SNMP) and the Common Information Model (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 additional systems management API. The focus is on attribute query and operational management functions for System z, such as CPCs, images, and activation profiles. The zEC12 contains 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 the *System z Application Programming Interfaces*, SB10-7030.

12.6.14 Cryptographic support

This section lists the cryptographic management and control functions available in the HMC and the SE.

Cryptographic hardware

The IBM zEnterprise EC12 includes 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 Express4S, a new Peripheral Component Interconnect Express (PCIe) Cryptographic Coprocessor, is an optional zEC12 exclusive feature. Crypto Express4S provides a secure programming and hardware environment on which crypto processes are run. Each Crypto Express4S adapter can be configured by the installation as a Secure IBM CCA coprocessor, a Secure IBM Enterprise Public Key Cryptography Standards (PKCS) #11 (EP11) coprocessor, or an accelerator.

When EP11 mode is selected, a unique Enterprise PKCS #11 firmware is loaded into the cryptographic coprocessor. It is separate from the Common Cryptographic Architecture (CCA) firmware that is loaded when 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 Express4S when configured as an Enterprise PKCS #11 coprocessor.

Crypto Express3 is also available in a carry-forward only basis when you upgrade from earlier generations to the zEC12.

To support the new Crypto Express4S card, the Cryptographic Configuration window was changed to support the following card modes:

- ▶ Accelerator mode (CEX4A)
- ▶ CCA Coprocessor mode (CEX4C)
- ▶ PKCS #11 Coprocessor mode (CEX4P)

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 the zEC12.
- ▶ UDX import now supports importing from DVD only.

Figure 12-17 shows an example of the Cryptographic Configuration window.

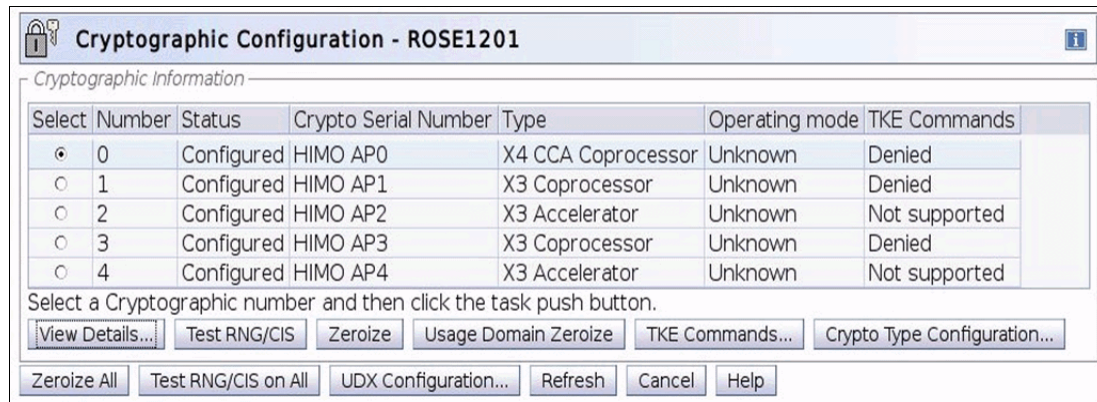


Figure 12-17 Cryptographic Configuration window

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 Express4S in EP11 mode will be configured to the standby state after Zeroize.

For more information, see [IBM zEnterprise Ec12 Configuration Setup](#), SG24-8034.

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 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 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 System z products during firmware updates. It enables the zEC12 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 System z focus of security for the HMC and the SE.

12.6.15 z/VM virtual machine management

The HMC can be used for basic management of z/VM and its virtual machines. The HMC uses the z/VM Systems Management Application Programming Interface (SMAPI), which provides a graphical user interface (GUI)-based alternative to the 3270 interface.

Monitoring the status information and changing the settings of z/VM and its virtual machines are possible. From the HMC interface, virtual machines can be activated, monitored, and deactivated.

Authorized HMC users can obtain various status information, including the following information:

- ▶ Configuration of the particular z/VM virtual machine
- ▶ z/VM image-wide information about virtual switches and guest LANs
- ▶ Virtual Machine Resource Manager (VMRM) configuration and measurement data

The activation and deactivation of z/VM virtual machines are integrated into the HMC interface. You can select the Activate and Deactivate tasks on CPC and CPC image objects, and for virtual machine management.

An event monitor is a trigger that monitors events from objects that are managed by the HMC. When z/VM virtual machines change their status, they generate these events. You can create event monitors to handle these events. For example, selected users can be notified by an email message if the virtual machine changes status from Operating to Exception, or any other state.

In addition, in z/VM V5R4 (or later releases), the APIs can run the following functions:

- ▶ Create, delete, replace, query, lock, and unlock directory profiles
- ▶ Manage and query LAN access lists (granting and revoking access to specific user IDs)
- ▶ Define, delete, and query virtual processors within an active virtual image and in a virtual image's directory entry
- ▶ Set the maximum number of virtual processors that can be defined in a virtual image's directory entry

12.6.16 Installation support for z/VM using the HMC

Starting with z/VM V5R4 and System z10, Linux on System z can be installed in a z/VM virtual machine from an HMC workstation media. This Linux on System z installation can use the existing communication path between the HMC and the SE. No external network or additional network setup is necessary for the installation.

12.7 HMC in an ensemble

An *ensemble* is a platform systems management domain that consists of up to eight zEC12 or zEnterprise nodes. Each node comprises a zEnterprise CPC and its optional attached IBM zEnterprise BladeCenter Extension (zBX). The ensemble provides an integrated way to manage virtual server resources and the workloads that can be deployed on those resources. The IBM zEnterprise System (zEnterprise) is a workload-optimized technology system that delivers a multiple platform, integrated hardware system. This system spans System z, System p, and System x blade server technologies.

Management of the ensemble is provided by the IBM zEnterprise Unified Resource Manager.

Restriction: The ensemble HMC mode is only available for managing IBM zEnterprise Systems (the z196, z114, zEC12, and zBC12).

12.7.1 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 12-18 shows the Unified Resource Manager functions and suites.

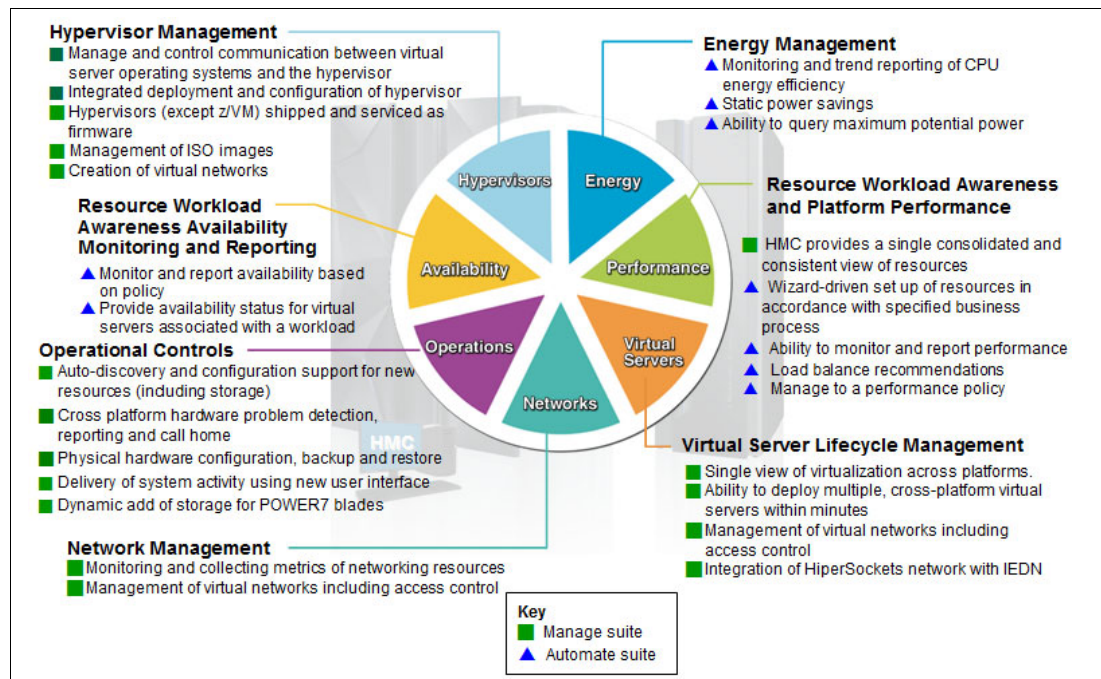


Figure 12-18 Unified Resource Manager functions and suites

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 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

► Ensemble Availability Management (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 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
- 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 attached zBX:

► **Manage suite**

Provides Unified Resource Manager's 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 functionality for IBM System x blades delivers 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 functionality.

Table 12-2 lists the feature codes that must be ordered to enable Unified Resource Manager. To get ensemble membership, ensure that you also order FC 0025 for the zEC12.

Table 12-2 Unified Resource Manager feature codes and charge indicators

Unified Resource Manager-managed component	Manage ^a (per connection)	Advanced Management ^a (per connection)	Automate ^a (per connection)
Base features	FC 0019 ^d - N/C	N/A	FC 0020 ^c - N/C
IFL	N/C	N/A	FC 0054 - Yes
POWER7 Blade	FC 0048 - Yes	N/A	FC 0051 - Yes
DataPower Blade	FC 0047 - Yes	N/A	FC 0050 - N/C
IBM System x blades	FC 0049 - Yes	FC 0053 - Yes Available on Driver 12 only	FC 0071 - Yes

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 functionality of the Managed suite.

b. Feature code 0019 is a prerequisite for FC 0020, FC 0047, FC 0048, and FC 0049.

c. Feature code 0020 is a prerequisite for FC 0050, FC 0051, FC 0053, FC 0054, and 0071.

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, and workload-based optimization, among others.

The Web Services API consists of two major components that are accessed by client applications through TCP/IP network connections with the HMC.

For more information about the API and the Unified Resource Manager, see *System z Hardware Management Console Web Services API*, SC27-2616, 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 will continue to play a distinct and strategic role in the management of virtualized environments created by the integrated firmware hypervisors (PR/SM, PowerVM, and x hypervisor based on KVM) of zEnterprise.

12.7.2 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.

Feature code

The Ensemble Membership Flag feature, FC 0025, is associated with an HMC when a zEC12 is ordered. This feature code is required on the *controlling* zEC12 to be able to attach a zBX.

The new Create Ensemble task allows the Access Administrator to create an ensemble that contains CPCs, images, workloads, virtual networks, and storage pools. This ensemble can be created with or without an optional zBX.

If a zEC12 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 connected to the ensemble are able to run system management tasks (but not ensemble management tasks) for any CPC within the ensemble. The primary HMC can also 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 to allow you to configure certain ensemble-specific properties. You can, for example, set the virtual network 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 and you can modify and delete those roles.
- ▶ **Event monitoring:** Displays ensemble-related events, but you cannot change or delete the event.

HMC considerations when used 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 0092 or FC 0091 can be primary or alternate HMCs for the zEC12.
- ▶ The primary HMC and the alternate HMC must be the same machine type and feature code.
- ▶ There is a single HMC pair that 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.
- ▶ The HMC that ran the Create Ensemble wizard becomes the primary HMC. An alternate HMC is elected and paired with the primary.
- ▶ The Primary Hardware Management Console (Version 2.12.0 or later) and Alternate Hardware Management Console (Version 2.12.0 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 eight nodes, and is managed by one primary HMC and its alternate. Each node comprises a zEnterprise CPC and its optional attached IBM zEnterprise BladeCenter Extension (zBX).
- ▶ 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 Support Elements.

12.7.3 HMC availability

The HMC is attached to the same LAN as the server's SE. This LAN is referred to as the *Customer Managed Management Network*. The HMC communicates with each CPC, and optionally to one or more zEnterprise BladeCenter Extensions (zBXs), through the SE.

If the zEC12 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.

12.7.4 Considerations for multiple HMCs

Clients often deployed 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 multiple primary and alternate HMC pairs to manage them. If a zEC12 or zEnterprise System node is added to an ensemble, management actions that target that system can be done only from the managing (primary) HMC for that ensemble.

12.7.5 HMC browser session to a primary HMC

A remote HMC browser session to the primary HMC that manages an ensemble allows a user who is logged on to another HMC or a workstation to perform ensemble-related actions.

12.7.6 HMC ensemble topology

The system management functions that pertain to an ensemble use the virtual server resources and the intraensemble management network (IEDN). They are provided by the HMC and the SE through the internode management network (INMN).

Figure 12-19 depicts an ensemble with two zEC12s and a zBX that are managed by the Unified Resource Manager in the primary and alternate HMCs. CPC1 controls the zBX, and CPC2 is a stand-alone CPC. CPC1 controls the zBX, and CPC2 is a stand-alone CPC.

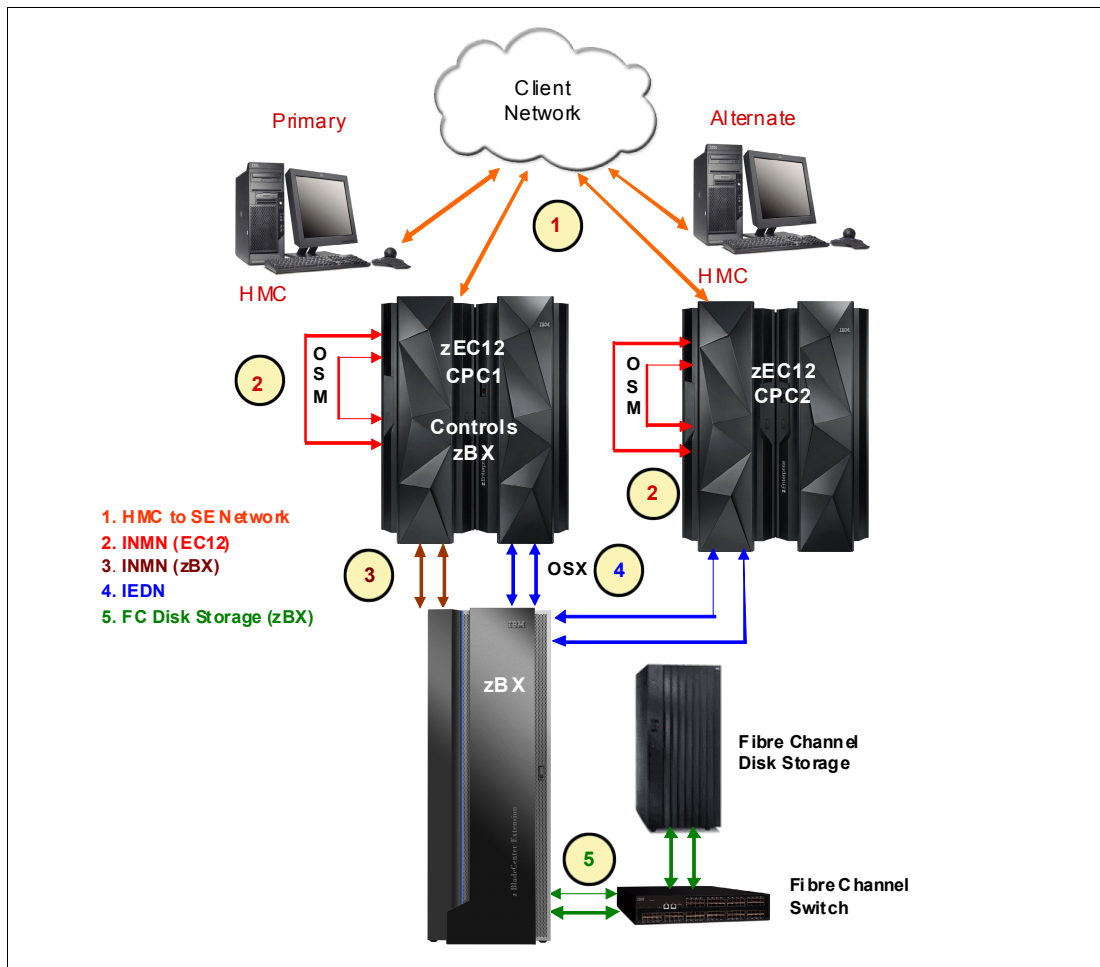


Figure 12-19 Ensemble example with primary and alternate HMCs

For the stand-alone CPC ensemble node (CPC2), two OSA-Express4S 1000BASE-T ports (CHPID type OSM) connect to the Bulk Power Hubs (port J07) with 3.2-meter (10.49 ft) Category 6 Ethernet cables. The HMCs also communicate with all the components of the ensemble by the BPHs in the CPC.

The OSA-Express4S 10 GbE ports (CHPID type OSX) are plugged with client-provided 10 GbE cables. These cables are either short reach (SR) or long reach (LR), depending on the OSA feature.

For more information about zBX, see Chapter 7, “zEnterprise BladeCenter Extension (zBX) Model 003” on page 227.

Performance

The zEC12 Model HA1 is designed to offer approximately 1.5 times more capacity than the z196 Model M80 system. Uniprocessor performance has also increased significantly. A zEC12 Model 701 offers, on average, performance improvements of about 1.25 times over the z196 Model 701. Figure 13-1 shows the estimated capacity ratios for zEC12, z196, z10 EC, and z9 EC. The Large System Performance Reference (LSPR) numbers given for z196, z10 EC, and z9 EC systems were obtained with the z/OS V1R11 operating system. For the zEC12 system, z/OS V1R13 was used.

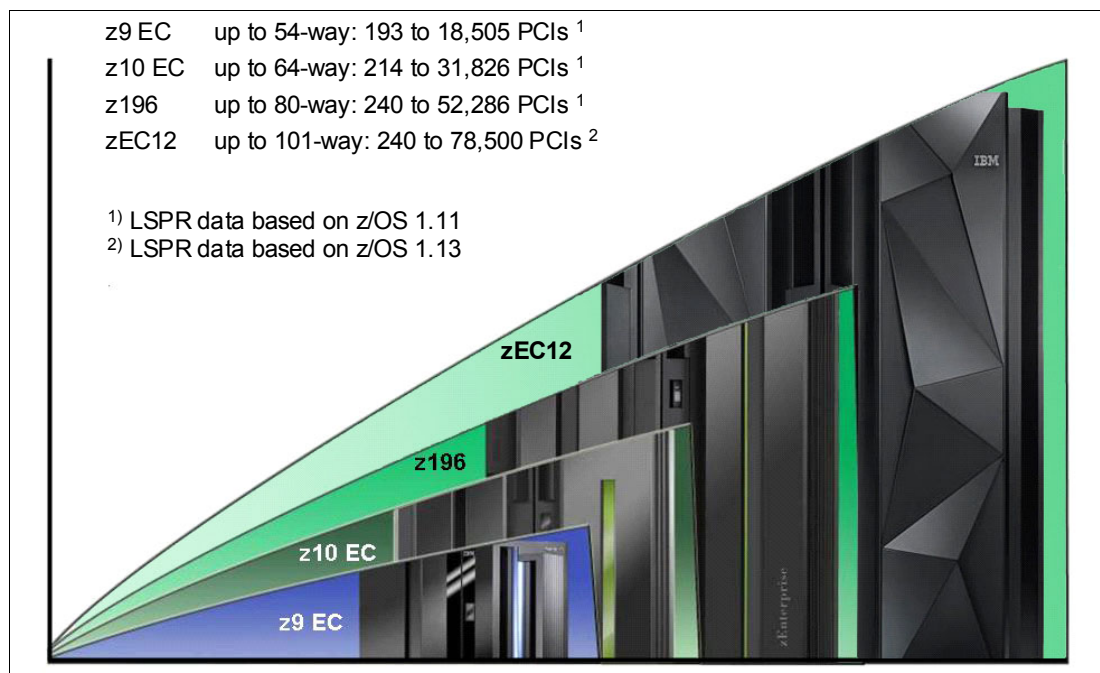


Figure 13-1 zEC12 to z196, z10 EC, and z9 EC performance comparison

On average, the zEC12 can deliver up to 50% more performance in a 101-way configuration than an IBM System zEnterprise 196 (z196) 80-way. However, variations on the observed performance increase are dependent on the workload type.

Consult Large Systems Performance Reference (LSPR) when you consider performance on the zEC12. The range of performance ratings across the individual LSPR workloads is 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 13.6, “Workload performance variation” on page 476.

For detailed performance information, see the LSPR website:

<https://www-304.ibm.com/servers/resourceLink/lib03060.nsf/pages/lsprindex>

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

<http://www-03.ibm.com/systems/z/resources/swprice/reference/exhibits/>

13.1 LSPR workload suite

Historically, LSPR capacity tables, including pure workloads and mixes, have been identified with application names or a *software* characteristic. Examples are CICS, IMS, OLTP-T¹, CB-L², LoIO-mix³, and TI-mix⁴. However, capacity performance is more closely associated with how a workload uses and interacts with a particular processor *hardware* design. With the CPU Measurement Facility (CPU MF) data that was introduced on the z10, the ability to gain insight into the interaction of workload and *hardware design* in production workloads has arrived. 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 zEC12 and the previous generation processor families. These ratios are based upon measurements and projections that use standard IBM benchmarks in a controlled environment. The actual 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 performance ratios stated.

13.2 Fundamental components of workload capacity performance

Workload capacity performance is sensitive to three major factors: instruction path length, instruction complexity, and memory hierarchy. Let us examine each of these three factors.

¹ Traditional online transaction processing workload (formerly known as IMS)

² Commercial batch with long-running jobs

³ Low I/O Content Mix Workload

⁴ Transaction Intensive Mix Workload

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 depending 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 presuming 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 logical partition (LPAR). As the number of logical processors grows, more instructions are needed to manage resources that are serialized by latches and locks.

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 we can define as *instruction complexity*. There are many design alternatives that 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 often varies. However, when on a processor, this component tends to be similar across all models of that processor.

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.

There are many design choices that affect this component:

- ▶ Cache size
- ▶ Latencies (sensitive to distance from the microprocessor)
- ▶ Number of levels, Modified, Exclusive, Invalid, Shared (MESI) (management) protocol, controllers, switches, 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 System z processor refers to the shared caches and memory along with the data buses that interconnect them.

Figure 13-2 on page 472 shows a memory nest in a zEC12 two-book system.

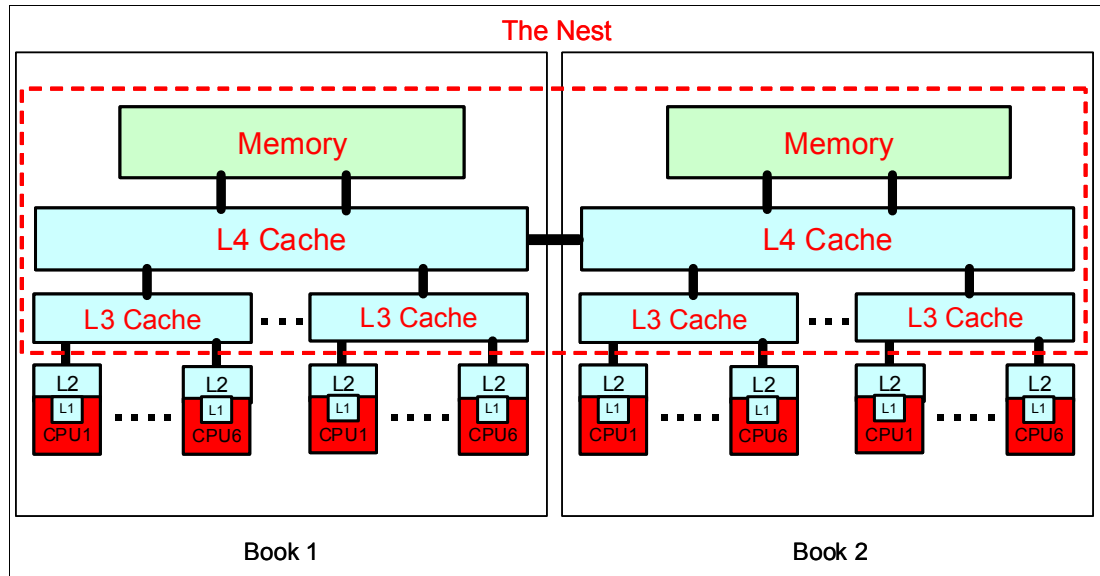


Figure 13-2 Memory hierarchy on the zEC12 two-book system

Workload capacity performance is sensitive to how deep into the memory hierarchy the processor must go to retrieve the workload instructions and data for execution. 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 execution. If 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 significantly. This variation is because the location of a workload's instructions and data within the memory hierarchy is affected by many factors including, but not limited to, these factors:

- ▶ Locality of reference
- ▶ I/O rate
- ▶ Competition from other applications and LPARs

13.3 Relative nest intensity

The most performance-sensitive area of the memory hierarchy is the activity to the memory nest. This is the distribution of activity to the shared caches and memory. We introduce a term, *Relative Nest Intensity (RNI)*, to indicate 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.

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.

We emphasize that these factors are simply tendencies and not 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 have been used to categorize workloads in the past are listed along with their RNI tendency in Figure 13-3.

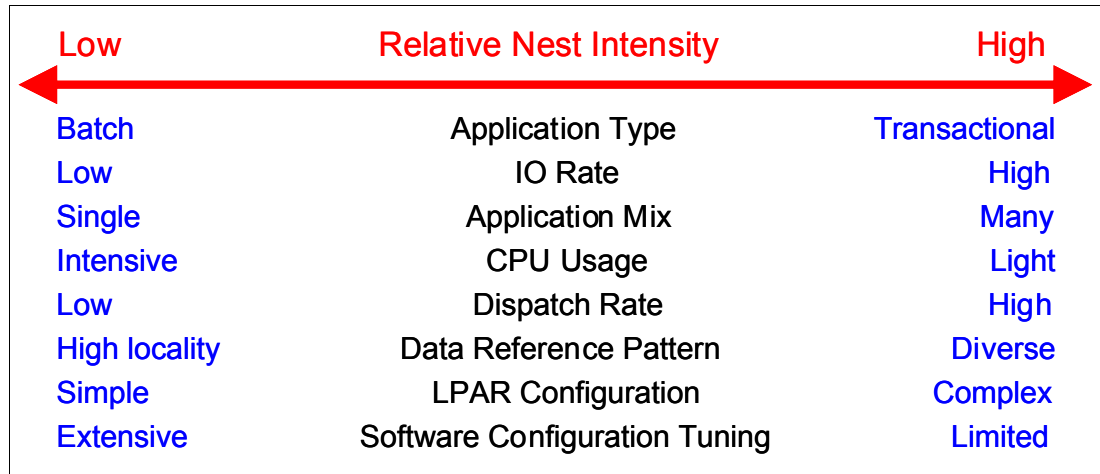


Figure 13-3 The traditional factors that have been used to categorize workloads

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 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 has always existed but its sensitivity is higher with the current high frequency microprocessors. Spreading the same workload over a larger number of 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 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 bigger or faster processor might be needed to achieve the published LSPR ratios.

13.4 LSPR workload categories based on relative nest intensity

A workload's relative nest intensity 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 heavy use of the memory hierarchy. This category is similar to the past TI-mix workload.

These categories are based on the relative nest intensity. The RNI is influenced by many variables, such as application type, I/O rate, application mix, processor usage, data reference patterns, LPAR configuration, and software configuration running. CPU MF data can be collected by z/OS System Measurement Facility on SMF 113 records. On zEC12, the number of extended counters is increased to 183. The structure of the SMF records does not change.

13.5 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 made up of the following workloads:

- ▶ Traditional online transaction processing workload 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 (commercial batch with long-running jobs or 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 13-3 on page 473 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 13-4. The IBM zPCR tool⁵ is changed to support the new z/OS workload categories.

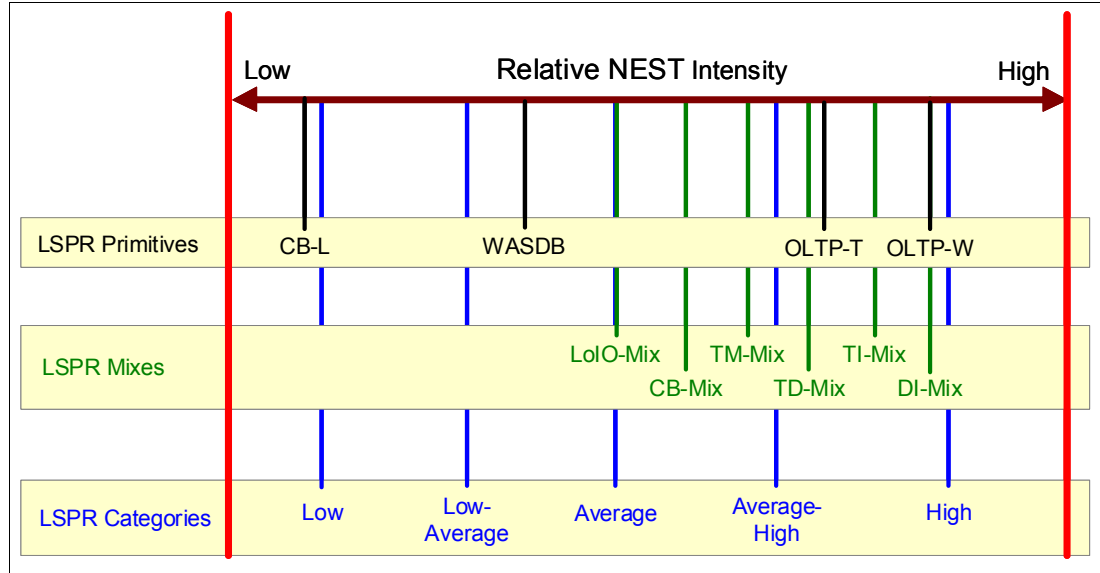


Figure 13-4 New z/OS workload categories defined

However, as addressed in 13.4, “LSPR workload categories based on relative nest intensity” on page 474, the underlying performance sensitive factor is how a workload interacts with the processor hardware. These past techniques were approximating 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. To reflect the memory hierarchy changes in the new zEC12 system, the number of counters is increased to 183. A production workload can now be matched to an LSPR workload category through these hardware characteristics. For more information about RNI, see 13.4, “LSPR workload categories based on relative nest intensity” on page 474.

The AVERAGE RNI LSPR workload is intended to match most client workloads. When no other data is available, use it for a capacity analysis.

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) <5), and 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 Processor Capacity Reference for System z (zPCR). Low-Average and Average-High categories allow better granularity for workload characterization.

⁵ The IBM Processor Capacity Reference tool reflects the latest IBM LSPR measurements. It is available at no extra charge at <http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/PRS1381>.

For z10 and newer processors, the CPU MF data can be used to provide an additional 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 hint are automated in the zPCR tool. It is best to use zPCR for capacity sizing.

The LSPR workloads that are updated for EC12 are considered to reasonably reflect the current and growth workloads of the client. The set contains three generic workload categories that are based on z/OS V1R13 supporting up to 101 processors in a single image.

13.6 Workload performance variation

Because of the nature of the zEC12 multi-book system and resource management across those books, performance variability from application to application is expected. This variation is similar to that seen on the z196, z10 EC, and z9 EC. 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 EC.

The memory and cache designs affect various workloads in a number of ways. All workloads are improved, with cache-intensive loads benefiting the most. When comparing moving from z9 EC to z10 EC with moving from z10 EC to z196 or from z196 to zEC12, it is likely that the relative benefits per workload will vary. Those workloads that benefited more than the average when moving from z9 EC to z10 EC will benefit less than the average when moving from z10 EC to z196. Nevertheless, the workload variability for moving from z196 to zEC12 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 MIPS, MSUs, and CPU time charge-back algorithms.

Experience demonstrates that System z 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. Be careful when deciding the number of processors and the uniprocessor capacity that you remember both the workload characteristics and LPAR configuration. For these reasons, when you plan capacity, we recommend using zPCR and involving IBM technical support.

Main performance improvement drivers with zEC12

The zEC12 is designed to deliver new levels of performance and capacity for large-scale consolidation and growth. The following attributes and design points of the zEC12 contribute to overall performance and throughput improvements as compared to the z196.

The z/Architecture implementation has the following enhancements:

- ▶ Transactional Execution (TX) designed for z/OS, Java, DB2, and other exploiters
- ▶ 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 zEC12 microprocessor design has the following enhancements:

- ▶ Six processor cores per chip
- ▶ Second generation Out Of Order (OOO) execution design
- ▶ Improved pipeline balance
- ▶ Enhanced branch prediction latency and instruction fetch throughput
- ▶ Improvements in execution bandwidth and throughput
- ▶ New design for Level 2 private cache with separation of cache structures for instructions and L2 operands
- ▶ Reduced access latency for most Level 1 cache misses
- ▶ Larger Level 2 cache with shorter latency
- ▶ Third level on-chip shared cache is doubled
- ▶ Fourth level book-shared cache is doubled
- ▶ Hardware and software prefetcher handling improvements
- ▶ Increased execution/completion throughput
- ▶ Improved fetch and store conflict scheme
- ▶ Enhance branch prediction structure and sequential instruction fetching
- ▶ Millicode performance improvements
- ▶ Optimized floating-point performance
- ▶ Faster engine for fixed-point division
- ▶ New second-level branch prediction array
- ▶ One cryptographic/compression co-processor per core
- ▶ Cryptography support of UTF8<>UTF16 conversions
- ▶ Higher clock frequency at 5.5 GHz
- ▶ IBM CMOS 13S 32nm SOI technology with IBM eDRAM technology

The zEC12 design has the following enhancements:

- ▶ Increased total number of PUs available on the system, from 96 to 120, and number of characterizable cores, from 80 to 101
- ▶ Hardware System Area (HSA) increased from 16 GB to 32 GB
- ▶ Increased default number of SAP processors per book
- ▶ New Coupling Facility Control Code (CFCC) available for improved performance:
 - Elapsed time improvements when dynamically altering the size of a cache structure
 - DB2 conditional write 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 zEC12:

- ▶ Open Systems Adapter (OSA)-Express5S family of features
- ▶ 10GbE Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE) Express feature
- ▶ Crypto Express4S performance enhancements
- ▶ Flash Express Peripheral Component Interconnect Express (PCIe) cards to handle paging workload spikes and improve performance
- ▶ zEnterprise Data Compression (zEDC) Express feature



IBM System z Advanced Workload Analysis Reporter (IBM zAware)

This appendix introduces *IBM System z Advanced Workload Analysis Reporter* (IBM zAware), the next generation of system monitoring. It is a new feature that is designed to offer a near real-time, continuous learning, diagnostic, and monitoring capability. 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
- ▶ Learning 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 *Advanced Workload Analysis Reporter (IBM zAware)*, SC27-2623.

A.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 the 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 ability to remain competitive in the marketplace.

However, as systems become more complex, errors can occur anywhere. Some problems begin with symptoms that go undetected for long periods of time. Systems often experience "soft failures" (sick but not dead) that are much more difficult or unusual to detect. Moreover, problems can grow, cascade, and get out of control.

Many everyday activities can introduce system anomalies and initiate 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 a challenging and skills-intensive task, as well as an error-prone task.

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.

A.2 Introducing IBM zAware

IBM zAware is an integrated expert solution that contains sophisticated analytics, IBM insight into the problem domain, and web-browser-based visualization.

IBM zAware 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.

Statement of Direction fulfillment:

IBM zAware and Tivoli Service Management are a powerful combination. You can get more from the zAware feature by integrating with Tivoli Service Management. Tivoli Service Management uses the zAware API to integrate log analysis with existing service management capabilities:

- ▶ Provide visibility into IBM zAware anomalies via 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

IBM zAware runs on a client-visible logical partition (LPAR) as a virtual appliance and provides out-of-band monitoring. It converts data into information and provides visualization to help you gain insight into the behavior of complex systems like a z/OS sysplex. It reduces problem determination time and improves service availability even beyond what it is in z/OS today.

Figure A-1 on page 482 shows how IBM zAware complements an existing environment.

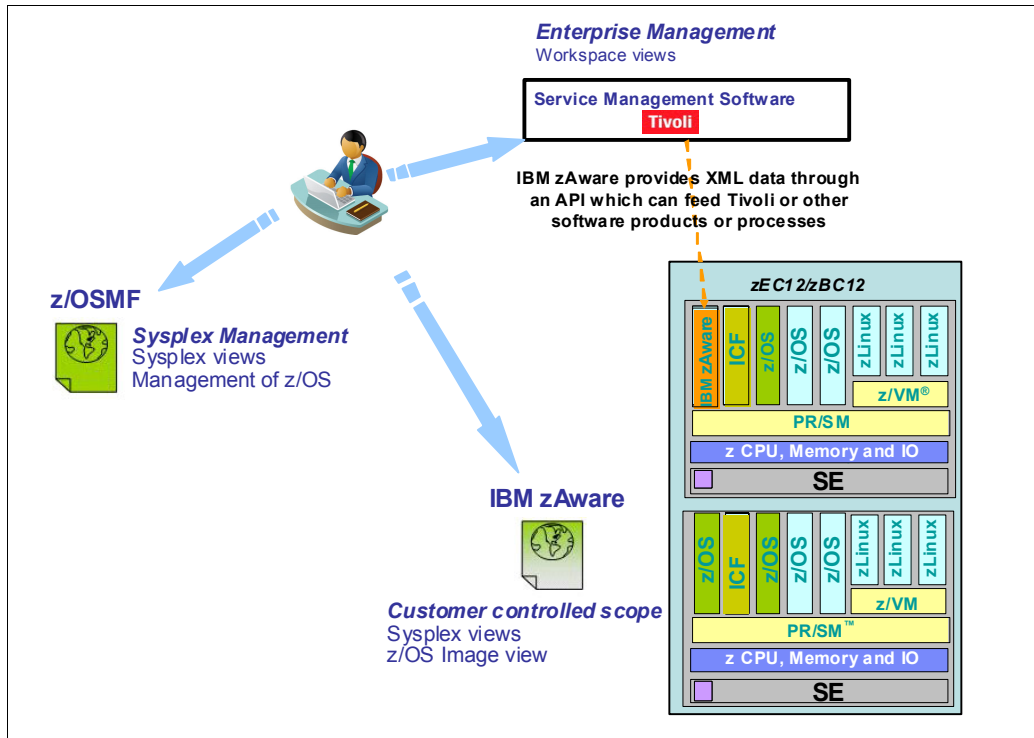


Figure A-1 IBM zAware complements an existing environment

A.2.1 Value of IBM zAware

Early detection and focused diagnosis can help improving time to recover from complex z/OS problems. These problems can be cross sysplex, across a set of System z servers, and beyond central processing complex (CPC) boundaries.

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 hard to spot. 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 A-2 on page 483 depicts how IBM zAware shortens the business impact of a problem.

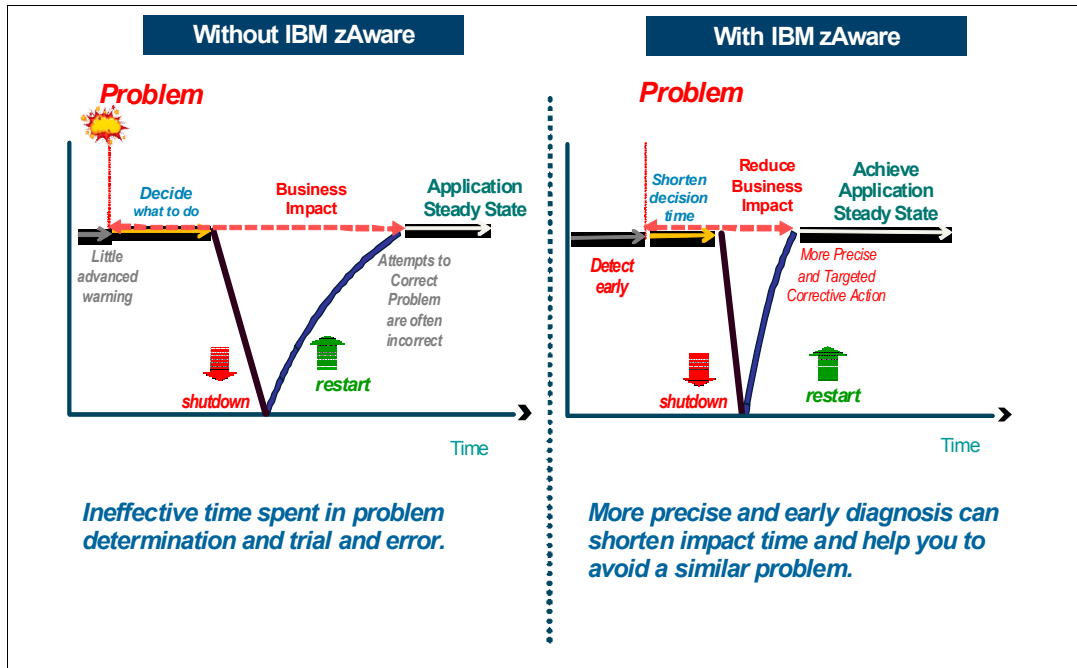


Figure A-2 IBM zAware shortens the business impact of a problem

IBM zAware also provides an easy-to-use graphical user interface (GUI) with 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 they can take corrective action for faster problem resolution.

A.2.2 IBM z/OS Solutions to improve problem diagnostic procedures

Table A-1 shows why IBM zAware is a more effective monitoring tool among all other problem diagnostic solutions for IBM z/OS.

Table A-1 Positioning IBM zAware

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
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

Solution	Available functions	Rules based	Analytics/ Statistical model	Examines message traffic	Self-learning	Method
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.

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.

Note:

- ▶ 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.

A.3 Understanding IBM zAware technology

IBM zAware runs analytics in firmware and intelligently examines OPERLOG data for potential deviations, inconsistencies, or variations from the normal behavior. It automatically manages the creation of the behavioral model that is used to compare current message log data from the connected z/OS systems.

Historical data, machine learning, mathematical modeling, statistical analysis, and cutting-edge pattern recognition techniques combine to uncover unusual patterns and understand the nuances of your unique environment.

Figure A-3 on page 485 depicts the basic components of an IBM zAware environment.

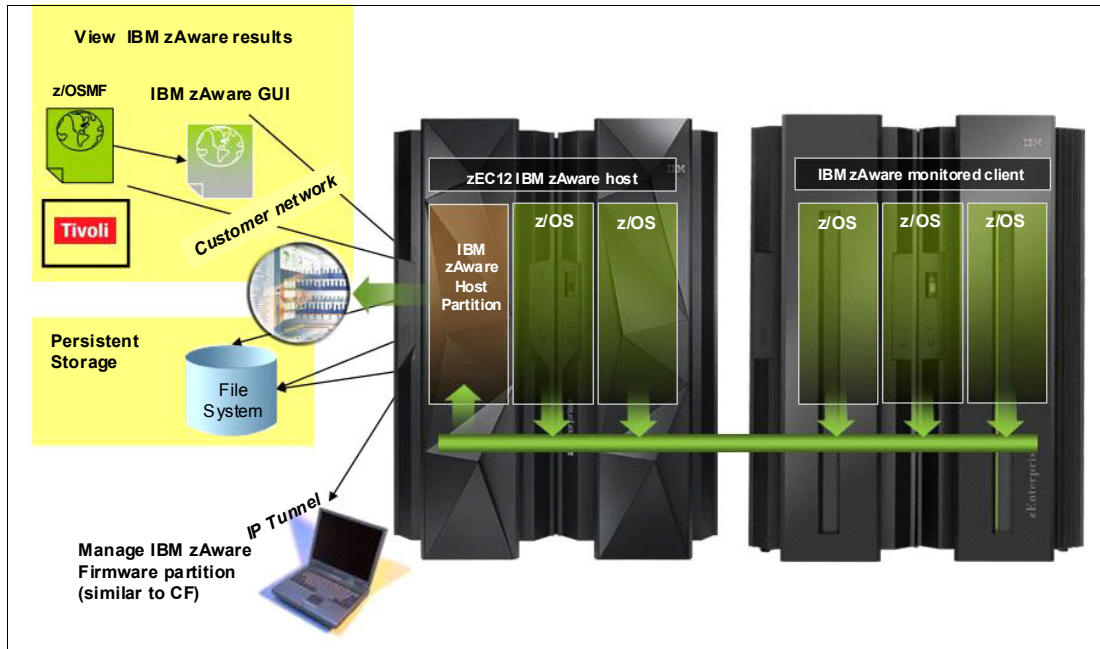


Figure A-3 Elements of an IBM zAware configuration

IBM zAware runs in an LPAR as firmware. IBM zAware has the following characteristics:

- ▶ Requires the **zEC12** configuration to have a priced feature code.
- ▶ Needs processor, memory, disk, and network resources to be assigned to the LPAR it runs. The requirements are similar to Coupling Facility LPARs.
- ▶ Is updated like all other firmware, with a separate engineering change stream.
- ▶ Is loaded from the Support Element (SE) hard disk.
- ▶ Employs out-of-band monitoring with minimal effect on z/OS product workloads.

Figure A-4 on page 486 shows IBM zAware Image Profile on the Hardware Management Console (HMC).

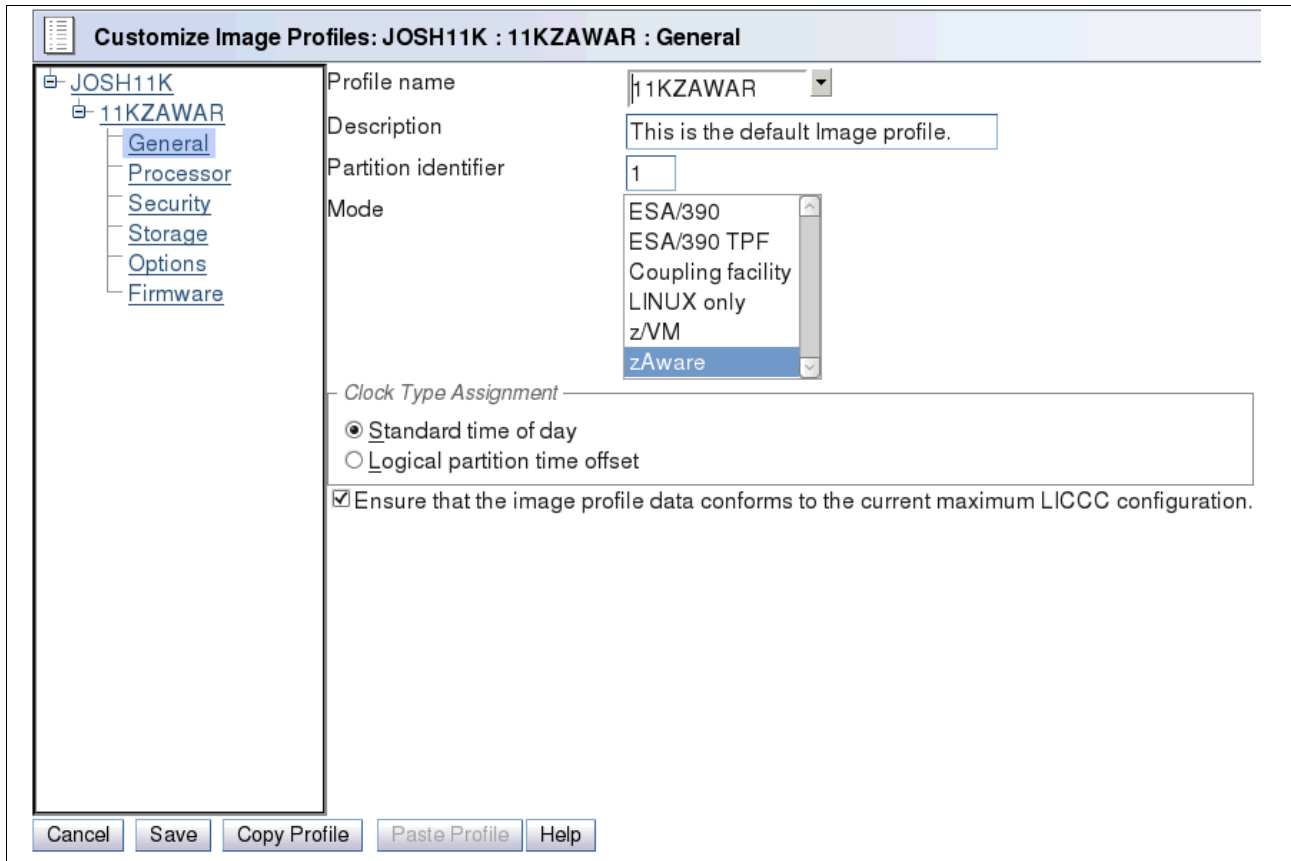


Figure A-4 HMC Image Profile for an IBM zAware LPAR

IBM zAware analyzes massive amounts of OPERLOG messages, including all z/OS console messages, including ISV and application-generated messages, to build sysplex and LPAR detailed views in the IBM zAware GUI. Figure A-5 shows a sample of the sysplex view.



Figure A-5 IBM zAware sysplex view that shows all connected, managed z/OS clients

Figure A-6 show a sample of the detailed view.

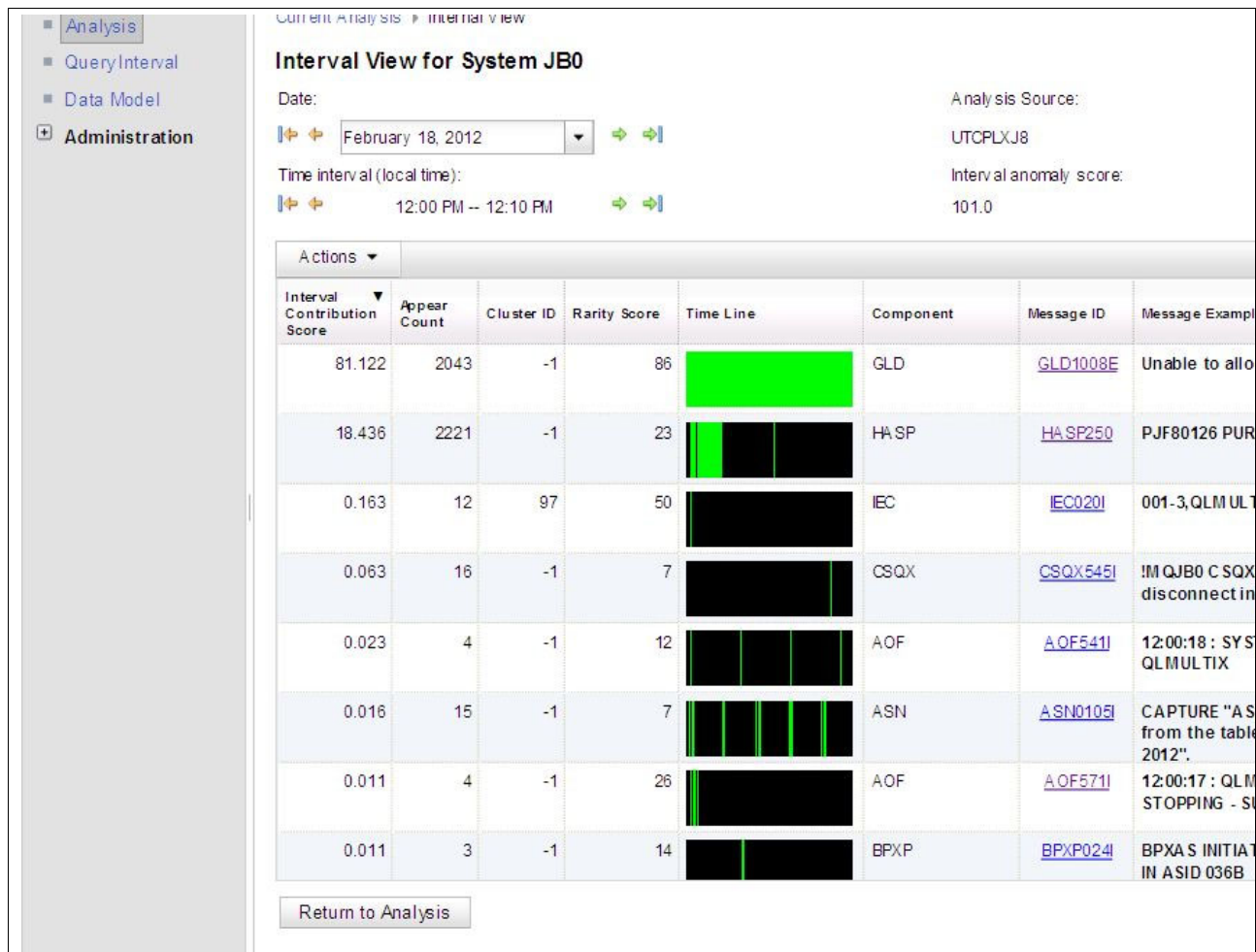


Figure A-6 IBM zAware detailed view, drilled down to a single z/OS image

The analytics create a statistical model of the normal message traffic that is generated by each z/OS. This model is stored in a database and 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 performs the following analysis on bursts of messages:

- ▶ Whether the unusual message IDs are coming from a single component
- ▶ Whether the message is a critical z/OS 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

The choice of unique message IDs is embedded in the domain knowledge of IBM zAware. IBM zAware detects things 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 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
- ▶ Orange (Interesting)
Interval score between 99.5 - 100
- ▶ Red (Rare)
An interval score of 101

A.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. 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.

A.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.

A.3.3 IBM zAware ignore message support

When a new workload is added to a system being monitored by zAware, or moved to a different system, it often generates messages that are not recognized by zAware. These messages are subsequently flagged as anomalous and cause orange bars to appear on the zAware analysis panel.

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 on zEC12 and zBC12 with Driver level 15 introduces the ignore message support to give the user input into the zAware rules. It allows you to mark desired messages as “ignore”. An ignored message is not part of zAware analysis and scoring.

The first iteration of this work requires the user to mark each message to be ignored on a per system basis. In other words, for each message you want to ignore, you have to mark the particular message on each system for which zAware is to ignore it. You can choose from one of two types of ignore message, until the next training period occurs (automatic or manual train) or forever.

A.3.4 IBM zAware graphical user interface

IBM zAware creates XML data with the status of the z/OS 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 using a standard web browser (Internet Explorer 8, Mozilla Firefox, or Chrome).

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. There also is a z/OSMF link on the navigation bar.

A.3.5 IBM zAware is complementary to 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 can be queued by existing system monitoring tools, such as Tivoli, by using published APIs.

A.4 Learning IBM zAware prerequisites

This section describes the hardware and software requirements for IBM zAware.

A.4.1 IBM zAware features and ordering

IBM zAware is available with IBM zEnterprise EC12 (zEC12) and zEnterprise BC12 (zBC12) models. IBM zAware feature-related definitions are listed in Table A-2.

Table A-2 IBM zAware feature code definitions

Name	Related feature code	Description
IBM zAware host system	FC0011	Represents the zEC12 or zBC12 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 zEC12 or zBC12, but there is 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.
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 (CPs) 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 zEC12 or zBC12 with no-charge firmware to run IBM zAware in a disaster situation.

a.

FC0101: IBM zAware CP 10 pack (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.

FC0102: IBM zAware DR CP 10 pack(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)

Feature on Demand (FoD)

FoD is a new centralized way to flexibly entitle features and functions on the system. FoD contains, for example, the IBM zEnterprise BladeCenter Extension (zBX) Model 003 *High Water Marks* (HWMs). HWMs refer to 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. On the zEC12, the HWMs are in the Feature on Demand record.

The zAware feature availability and installed capacity are also controlled by the FoD LICCC record. The current 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 A-7 shows the panel for FoD zAware feature status and value shown under the Perform Model Conversion, Feature on Demand Manage function.

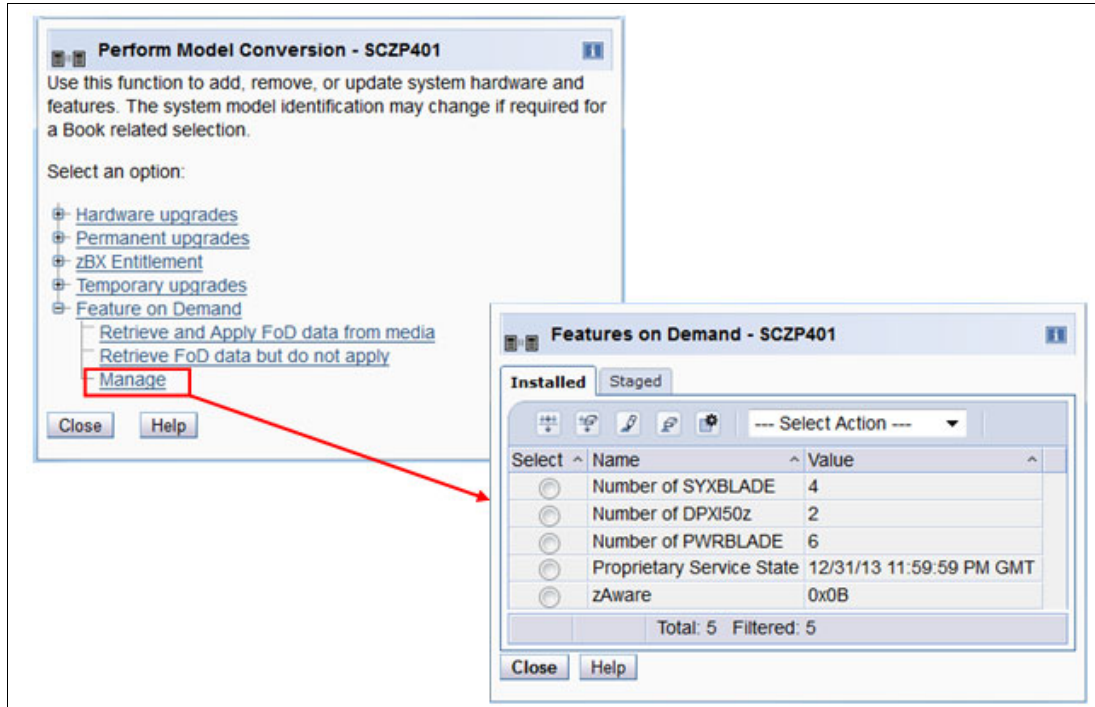


Figure A-7 Feature on Demand panel for zAware feature

There is only one FoD LICCC record installed or staged at any time in the system and its contents can be viewed under the Manage panel as shown in Figure A-7. A staged record can be removed without installing it. An FoD record can only be installed completely. There is no selective feature or partial record installation, and the features installed will be merged with the CEC LICCC after activation.

An FoD record can only be installed once and 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 zEC12 or zBC12 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 performing the following steps:

1. Determine which machines have z/OS images to be monitored by IBM zAware, including the zEC12 or zBC12 where the IBM zAware LPAR resides.
2. Count the number of CPs on the machines that were identified in the previous step. Include banked CPs (High Water Mark):
 - a. Round up to the nearest factor of 10 (zEC12).
 - b. Round up to the nearest factor of 2 or 10 (zBC12).

A Disaster Recovery option (IBM zAware DR CP packs) is also available and indicates that IBM zAware is installed on a DR zEC12 or zBC12 server. This feature is available at no additional fee, but is exclusive with 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. And also 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.

A.4.2 IBM zAware operating requirements

The following components are required for IBM zAware.

IBM zAware host system

The zEC12 or zBC12 can host the IBM zAware server. The IBM zAware server requires its own LPAR and runs its own self-contained firmware stack.

Note: Host system resources (processors, memory, DASD, and so on) are dependent 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 Integrated Facility for Linux (IFL) that can be shared with other LPARs in the zEC12 or zBC12.
 - Usage estimates between a partial engine to two engines, depending on the size of the configuration.
- ▶ Memory:
 - Minimum 6 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.
 - Flash Express is not supported.
- ▶ Direct access storage devices (DASD):
 - 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 and uses Logical Volume Manager (LVM) to aggregate multiple physical devices into a single logical device.
- ▶ Network (for both instrumentation data gathering and outbound alerting/communications):
 - HiperSockets for the z/OS LPARs running on the same zEC12 or zBC12 as the IBM zAware LPAR.
 - Open Systems Adapter (OSA) ports for the z/OS LPARs running on a different CPC than where the IBM zAware LPAR runs.
 - Dedicated IP address for IBM zAware LPAR.

IBM zAware monitored client

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 is increased to 3500 km (2174.79 miles).
- ▶ IBM zAware monitored clients can be on any System z servers (IBM zEC12, zBC12, z196, z114, z10, and so on) if they fulfill z/OS requirements. Monitoring can be done by sharing log files through an Internet Protocol network with IBM zAware servers.

z/OS requirements

IBM zAware monitored clients have the following z/OS requirements:

- ▶ z/OS V2.1 or higher
- ▶ z/OS V1.13 with PTFs
- ▶ Ninety days historical SYSLOG or formatted OPERLOG data to initially prime IBM zAware

A.5 Configuring and using IBM zAware virtual appliance

The following checklist provides a task summary for configuring and using IBM zAware:

- ▶ Phase 1: Planning:
 - Plan the configuration of the IBM zAware environment.
 - Plan the LPAR characteristics of the IBM zAware partition.
 - Plan the network connections that are required for the IBM zAware partition and each z/OS monitored client.
 - Plan the security requirements for the IBM zAware server, its monitored clients, and users of the IBM zAware GUI.
 - Plan for using the IBM zAware GUI.
- ▶ Phase 2: Configuring the IBM zAware partition:
 - Verify that your installation meets the prerequisites for using the IBM zAware virtual appliance.
 - Configure network connections for the IBM zAware partition through the hardware configuration definition (HCD) or the input/output configuration program (IOCP).
 - Configure persistent storage for the IBM zAware partition through the HCD or IOCP.
 - Define the LPAR characteristics of the IBM zAware partition through the Hardware Management Console (HMC).
 - Define network settings for the IBM zAware partition through the HMC.
 - Activate the IBM zAware partition through the HMC.
- ▶ Phase 3: Configuring the IBM zAware server and its monitored clients:
 - Assign storage devices for the IBM zAware server through the IBM zAware GUI.
 - Optional: Replace the self-signed certificate authority (CA) certificate that is configured in the IBM zAware server.
 - Optional: Configure a Lightweight Directory Access Protocol (LDAP) directory or local file-based repository for authenticating users of the IBM zAware GUI.
 - Optional: Authorize users or groups to access the IBM zAware GUI.

- Optional: Modify the configuration values that control the IBM zAware analytics operation.
- Configure a network connection for each z/OS monitored client through the TCP/IP profile. If necessary, update firewall settings.
- Verify that each z/OS system meets the sysplex configuration and OPERLOG requirements for IBM zAware virtual appliance monitored clients.
- Configure the z/OS system logger to send data to the IBM zAware virtual appliance server.
- Prime the IBM zAware server with prior data from monitored clients.
- Build a model of normal system behavior for each monitored client. The IBM zAware server uses these models for analysis.
- Optional: Use the IBM zAware ignore message support to give your input to the zAware rules. It allows you to mark desired messages as “ignore”. An ignored message is not part of zAware analysis and scoring.



B

Channel options

This appendix describes all channel attributes, the required cable types, the maximum unrepeated distance, and the bit rate for the zEC12.

For all optical links, the connector type is LC Duplex except for the 12xIFB connection, which is established with a Multi-fiber Push-On (MPO) connector. The electrical Ethernet cable for the Open Systems Adapter (OSA) connectivity is connected through an RJ45 jack.

Table B-1 on page 498 lists the attributes of the channel options that are supported on zEC12.

Statements of direction:

- ▶ The zBC12 and zEC12 are planned to be the last IBM System z[®] server to support InterSystem Channel-3 (ISC-3) Links. Enterprises must continue migrating from ISC-3 features to InfiniBand Coupling Links.
- ▶ The zBC12 and zEC12 are planned to be the last System z server to support Ethernet half-duplex operation and 10-Mbps link data rate on 1000BASE-T Ethernet features. Any future 1000BASE-T Ethernet feature will support full-duplex operation and auto-negotiation to 100 or 1000 Mbps exclusively.
- ▶ The zBC12 and zEC12 are planned to be the last System z server to support the OSA-Express3 family of features. Enterprises must continue migrating from the OSA-Express3 features to the OSA-Express5S features.
- ▶ The zBC12 and zEC12 are planned to be the last System z server to support FICON Express4 features. Enterprises must continue migrating from the FICON Express4 features to the FICON Express8S features.

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

Table B-1 zEC12 channel feature support

Channel feature	Feature codes	Bit rate in Gbps (or stated)	Cable type	Maximum unrepeated distance ^a	Ordering information
Fiber Connection (FICON)					
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 Express4 10KM LX	3321	1, 2, or 4			Carry forward
FICON Express8S SX	0410	2, 4, or 8	OM1, OM2, OM3	See Table B-2 on page 499.	New build
FICON Express8 SX	3326				Carry forward
FICON Express4 SX	3322	1, 2, or 4			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-Express3 10 GbE LR	3370				Carry forward
OSA-Express5S 10 GbE SR	0416	10	MM 62.5 μm	33 m (200)	New build
OSA-Express4S 10 GbE SR	0407		MM 50 μm	300 m (2000) 82 m (500)	Carry forward
OSA-Express3 10 GbE SR	3371		Carry forward		
OSA-Express5S GbE LX	0413	1	SM 9 μm	5 km (3.1 miles)	New build
OSA-Express4S GbE LX	0404				Carry forward
OSA-Express3 GbE LX	3362				Carry forward
OSA-Express5S GbE SX	0414	1	MM 62.5 μm	220 m (166) 275 m (200)	New build
OSA-Express4S GbE SX	0405		MM 50 μm	550 m (500)	Carry forward
OSA-Express3 GbE SX	3363		Carry forward		
OSA-Express5S 1000BASE-T	0417	100 or 1000 Mbps	Cat 5, Cat 6 copper		New build
OSA-Express4S 1000BASE-T	0408	10, 100, or 1000 Mbps			Carry forward
OSA-Express3 1000BASE-T	3367	Carry forward			
10GbE Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE) Express	0411	10	OM3	300 m	New build
Parallel Sysplex					
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

Channel feature	Feature codes	Bit rate in Gbps (or stated)	Cable type	Maximum unrepeated distance ^a	Ordering information
HCA2-O (12x IFB)	0163	6 Gbps	OM3	150 m	Carry forward
HCA2-O LR (1x IFB)	0168	2.5 or 5 Gbps	SM 9 μm	10 km (6.2 miles)	Carry forward
IC	N/A		N/A	N/A	N/A
ISC-3 (peer mode)	0217 0218 0219	2	SM 9 μm	10 km (6.2 miles)	Carry forward
ISC-3 (RPQ 8P2197 Peer mode at 1 Gbps) ^b		1	SM 9 μm	20 km (12.4 miles)	Carry forward
Cryptography					
Crypto Express4s	0865	N/A	N/A	N/A	New build
Crypto Express3	0864	N/A	N/A	N/A	Carry forward
Flash Express	0402	N/A	N/A	N/A	New build
zEDC Express	0420	N/A	N/A	N/A	New build

- a. Where applicable, the minimum fiber bandwidth distance in MHz-km for multi-mode fiber optic links is included in parentheses.
- b. RPQ 8P2197 enables the ordering of a daughter card that supports 20-km (12.4-mile) unrepeated distance for 1-Gbps peer mode. RPQ 8P2262 is a requirement for that option. Other than the normal mode, the channel increment is two (that is, both ports (FC 0219) at the card must be activated).

Table B-2 shows the maximum unrepeated distances for FICON SX features.

Table B-2 Maximum unrepeated distance for FICON SX features

Cable type\bit rate	1 Gbps	2 Gbps	4 Gbps	8 Gbps
OM1 (62.5 μm at 200 MHz·km)	300 meters	150 meters	70 meters	21 meters
	984 feet	492 feet	230 feet	69 feet
OM2 (50 μm at 500 MHz·km)	500 meters	300 meters	150 meters	50 meters
	1640 feet	984 feet	492 feet	164 feet
OM3 (50 μm at 2000 MHz·km)	860 meters	500 meters	380 meters	150 meters
	2822 feet	1640 feet	1247 feet	492 feet



Flash Express

This appendix introduces the IBM Flash Express feature available 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 contains these sections:

- ▶ Flash Express overview
- ▶ Using Flash Express
- ▶ Security on Flash Express

C.1 Flash Express overview

Flash Express introduces SSD technology to the IBM zEnterprise EC12 server (zEC12), which is implemented by using Flash SSDs mounted in Peripheral Component Interconnect Express (PCIe) Flash Express feature cards.

Flash Express is an innovative solution available on zEC12 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 improve access time for critical business z/OS workloads. It can also reduce latency time during diagnostic collection (dump operations).

Flash Express introduces a new level in the zEC12 storage hierarchy as shown in Figure C-1.

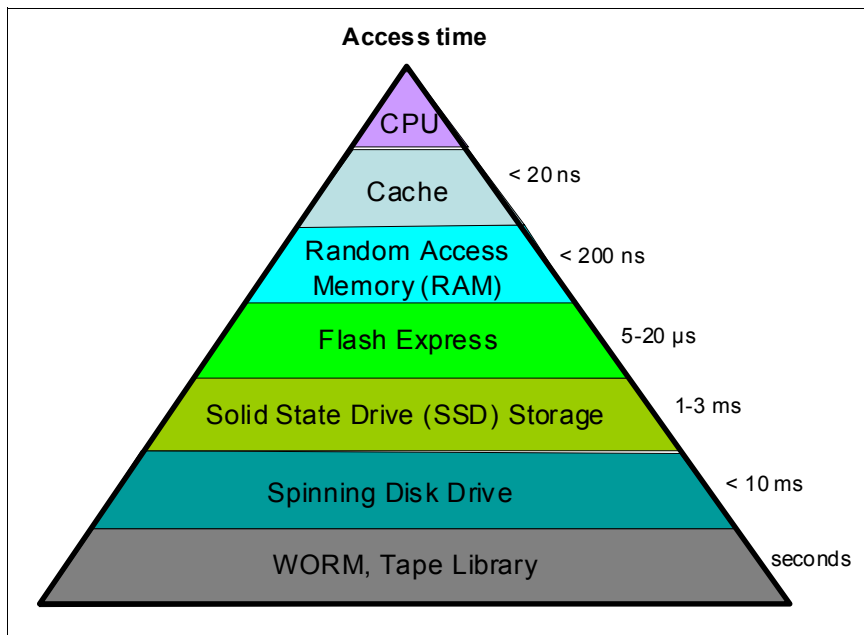


Figure C-1 zEC12 storage hierarchy

Flash Express is an optional Peripheral Component Interconnect Express (PCIe) card feature available on zEC12 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) Express8S, Crypto Express4S, and Open Systems Adapter (OSA) Express4S cards. You can order a minimum of two features (FC 0402) and a maximum of eight. The cards are ordered in increments of two.

Flash Express cards are assigned one physical channel ID (PCHID) even though they have no ports. No hardware configuration definition (HCD) or input/output configuration program (IOCP) definition is required for Flash Express installation. Flash uses subchannels that are allocated from the .25K reserved in subchannel set 0. Similar to other PCIe I/O 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 C-2.



Figure C-2 Flash Express PCIe adapter

Each card 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 utilization.

Figure C-3 shows a PCIe I/O drawer that is fully populated with Flash Express cards.

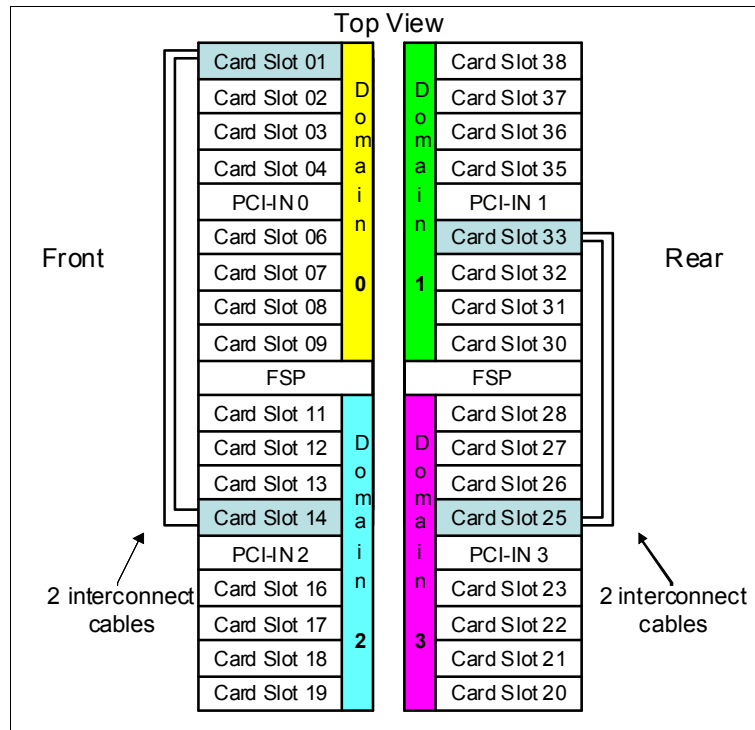


Figure C-3 PCIe I/O drawer fully populated with Flash Express cards

For higher resiliency and high availability, Flash Express cards are always installed in pairs. A maximum of four pairs are supported in a zEC12 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, and card replacement is concurrent with the client's operations.

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 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.

C.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 (SVC) dumps, and stand-alone dumps.

In z/OS, Flash Express memory is accessed by using the new System z 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 like the central 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 SVC dumps. Using Flash Express memory, 1-MB large pages become pageable. It is expected to provide applications with substantial improvement in SVC 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 CEC is assigned to a coupling facility (CF) partition via hardware definition panels the same way that it is assigned to the z/OS partitions.

Flash Express exploitation by the CF provides emergency capacity to handle WebSphere MQ shared queue buildups during abnormal situations, such as 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.

Other software subsystems might take advantage of Flash Express in the future.

Table C-1 on page 505 gives the minimum support requirements for Flash Express.

Table C-1 Minimum support requirements for Flash Express

Operating system	Support requirements
z/OS	z/OS V1R13 ^a and V2R1
CFCC	CF Level 19

a. Web delivery and 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 C-4 gives a sample SE/HMC interface that is used for Flash Express allocation.

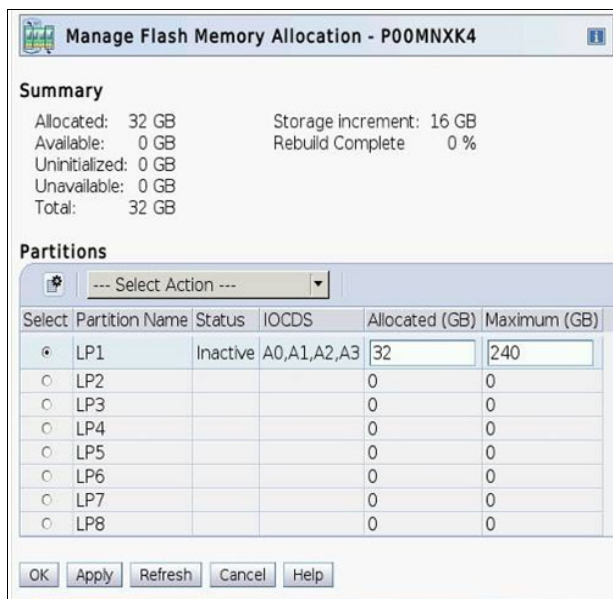


Figure C-4 Sample SE/HMC window for Flash Express allocation to LPAR

The new SE user interface for Flash Express provides four new 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 C-5 gives an example of Flash Express allocation between two z/OS LPARs.

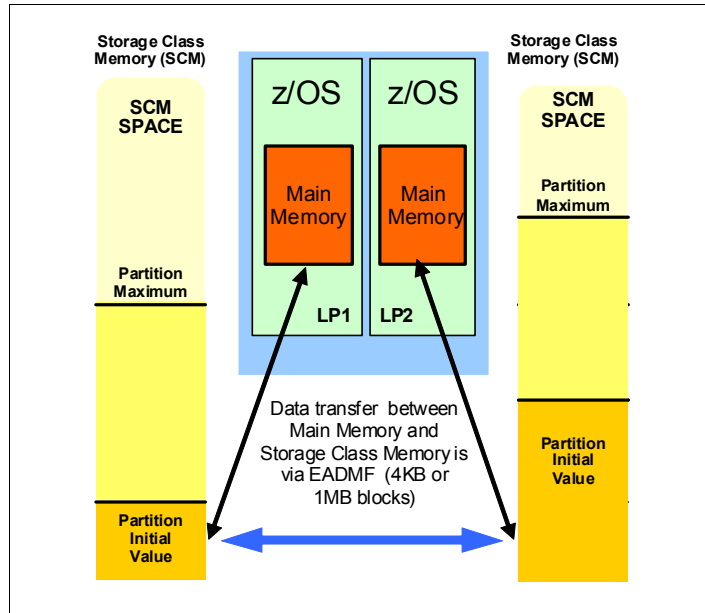


Figure C-5 Flash Express allocation in z/OS LPARs

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. Flash Express improves paging performance.

Currently, 1-MB large pages are not pageable. With the introduction of Flash Express, 1-MB large pages can be on Flash and pageable.

Table C-2 on page 507 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 C-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 I/O (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.
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 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
```

C.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 Support Element 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 physical key to the physical lock.

C.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 algorithm (RSA) 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 C-6 on page 509 shows the cryptographic keys that are involved in creating this tight-coupling design.

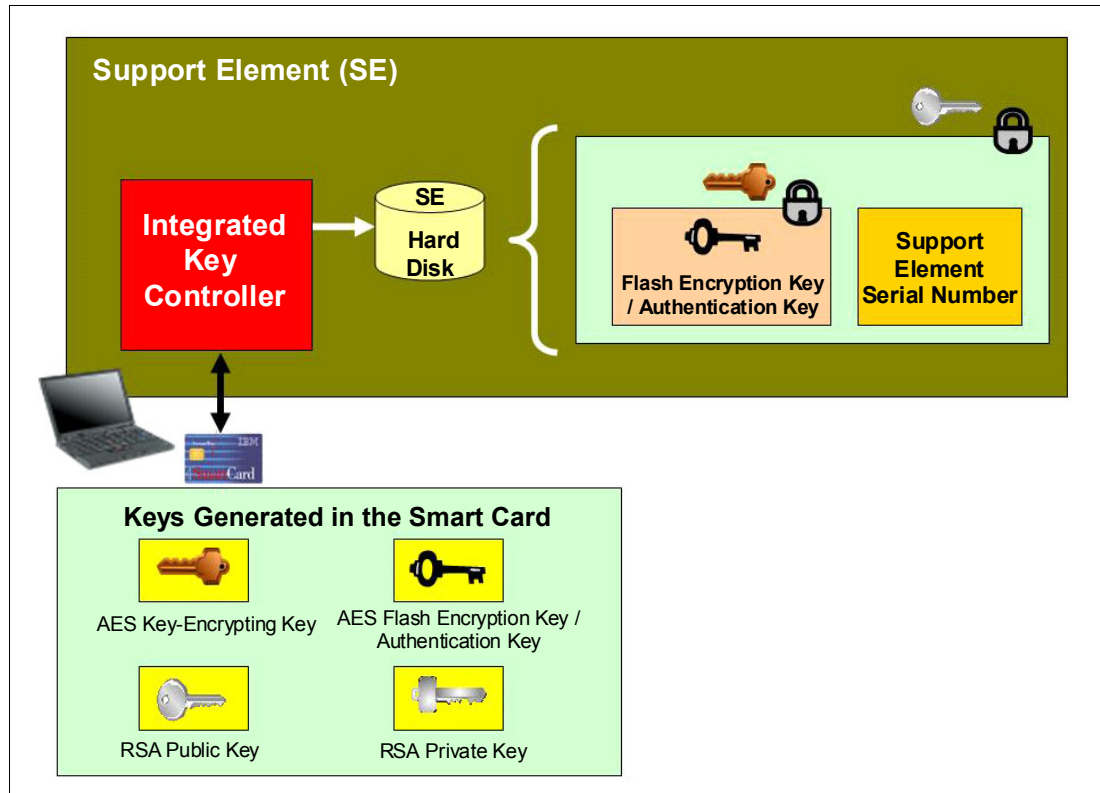


Figure C-6 Integrated Key Controller

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, 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.

C.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 C-7 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.

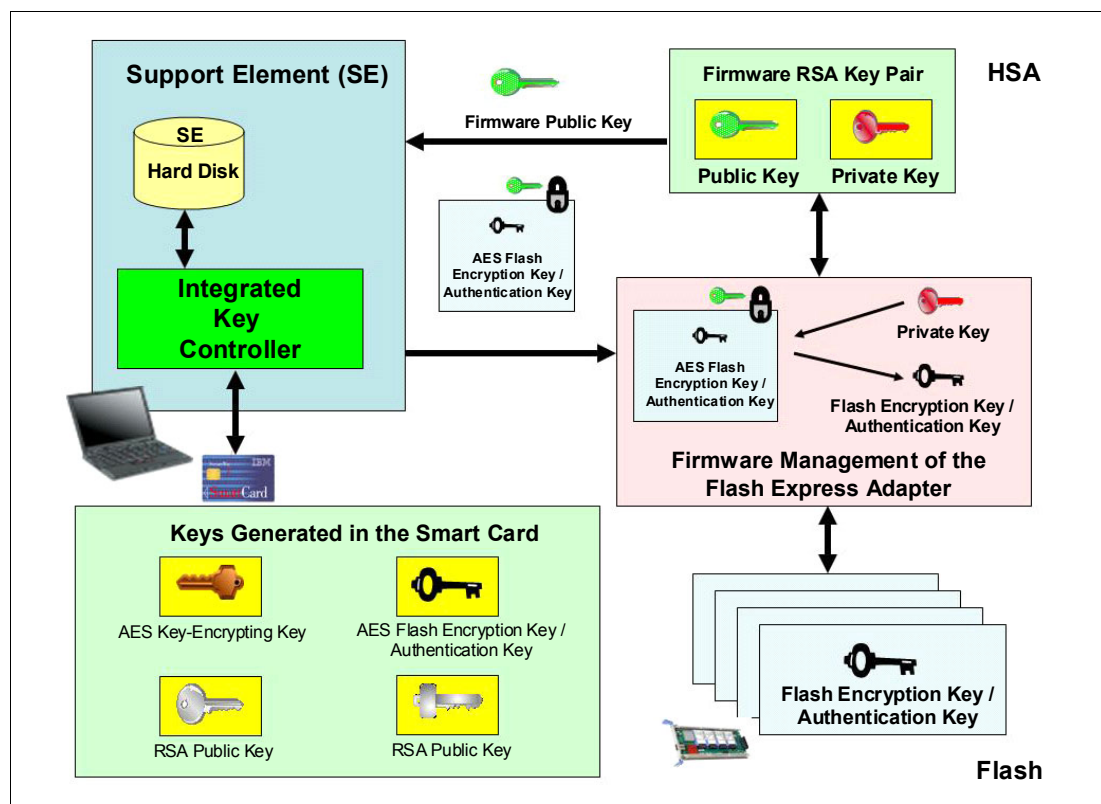


Figure C-7 Key serving topology

C.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, there is 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 during the switchover when the primary SE fails, 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 Support Elements both fail, the key cannot be served. If the devices are still up, the key is still valid. If either or both Support Elements 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 powered up.

If both Flash Express devices are still powered up, 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.



Remote Direct Memory Access over Converged Ethernet (RoCE)

This appendix briefly describes the optional RoCE feature of the IBM zEnterprise EC12 (zEC12) servers, and includes the following topics:

- ▶ Overview
- ▶ Hardware
- ▶ Software exploitation

D.1 Overview

Each generation of Open Systems Adapter (OSA) Express continues to provide significant new functionality, resiliency, and performance. HiperSockets also continue to provide significant industry unique qualities of service, virtualization, and performance (achieving natural improvements with each new System z processor). zEC12 delivers a significant paradigm shift in network communications by using existing System z and industry standard communications technology along with emerging new network technology:

- ▶ Remote Direct Memory Access (RDMA) technology provides low latency, high bandwidth, high throughput, and low processor utilization attachment between hosts.
- ▶ Shared Memory Communications (SMC) - Remote Direct Memory Access (SMC-R) is a new protocol that allows existing TCP applications to benefit transparently from RDMA for transferring data:
 - SMC-R uses RDMA over Converged Ethernet (RoCE) as the physical transport layer.
 - Initial deployment is limited to z/OS to z/OS communications with a goal to expand exploitation to additional operating systems and possibly appliances and accelerators.

D.1.1 Remote Direct Memory Access (RDMA) technology overview

Remote Direct Memory Access (RDMA) over Converged Ethernet (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 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 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.

There are two key requirements for RDMA as shown in Figure D-1:

- ▶ A reliable “lossless” network fabric (LAN for layer 2 data center network distance)
- ▶ An RDMA-capable network interface card (NIC) and Ethernet fabric

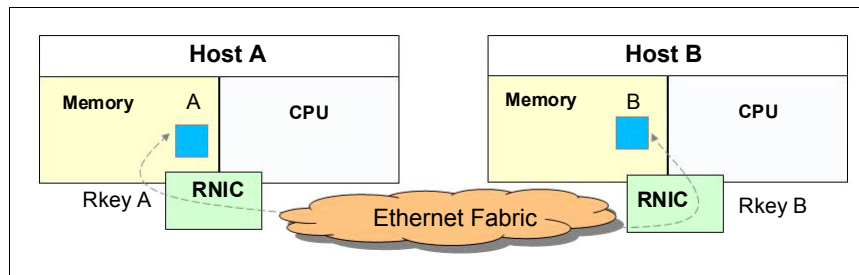


Figure D-1 RDMA technology overview

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 network interchange cards (NICs) in the host).

D.1.2 Shared Memory Communications – RDMA (SMC-R)

Shared Memory Communications - 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:

- ▶ It uses a TCP connection to establish the SMC-R connection.
- ▶ A TCP option 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 existing operational and network management features of TCP/IP.

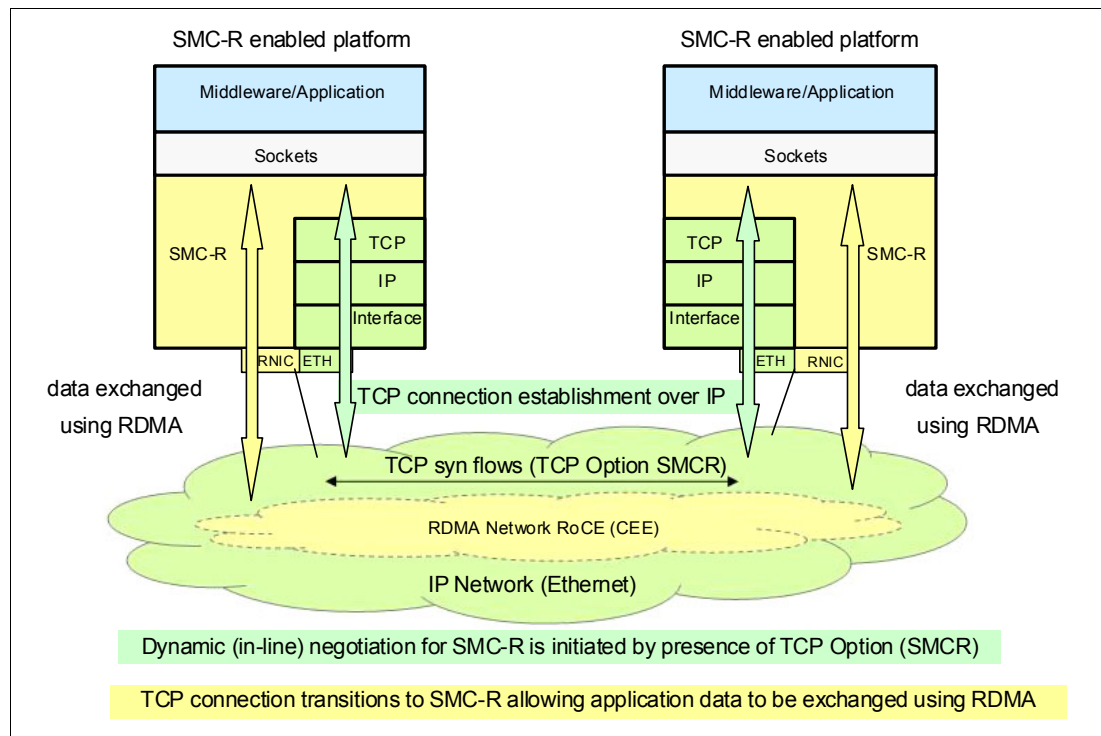


Figure D-2 Dynamic transition from TCP to SMC-R

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 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, for example, 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; all host application workloads can benefit immediately.

D.2 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) that have firmware for the 10GbE RoCE Express feature. For more detailed information about IFP and RG, see Appendix F, “Native Peripheral Component Interconnect Express (PCIe)” on page 529.

D.2.1 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. It is transparent to applications and can be used for logical partition (LPAR)-to-LPAR communication on a single z/OS system or server-to-server communication in a multiple CPC environment.

The 10GbE RoCE Express feature that is shown in Figure D-3 on page 517 is exclusive to the zEC12 and IBM zEnterprise BC12 (zBC12). It is used exclusively in the Peripheral Component Interconnect Express (PCIe) I/O drawer. Each feature has one PCIe adapter. A maximum of 16 features can be installed and only one port per feature is supported by z/OS. 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 and switched connection with an enterprise-class 10 GbE switch are supported.

If the IBM 10GbE RoCE Express features are connected to 10 GbE switches, the switches must support the following requirements:

- ▶ Global Pause function enabled
- ▶ Priority Flow Control (PFC) disabled
- ▶ No firewalls, no routing, and no intraensemble data network (IEDN)

The maximum supported unrepeatable distance, point-to-point 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))



Figure D-3 10GbE RoCE Express

D.2.2 10GbE RoCE Express configuration sample

Figure D-4 on page 518 illustrates a sample configuration that allows redundant SMC-R connectivity among LPAR A, LPAR C and LPAR 2, and LPAR 3.

Each feature must be dedicated to an LPAR. Like the sample configuration, two features per LPAR are advised for redundancy.

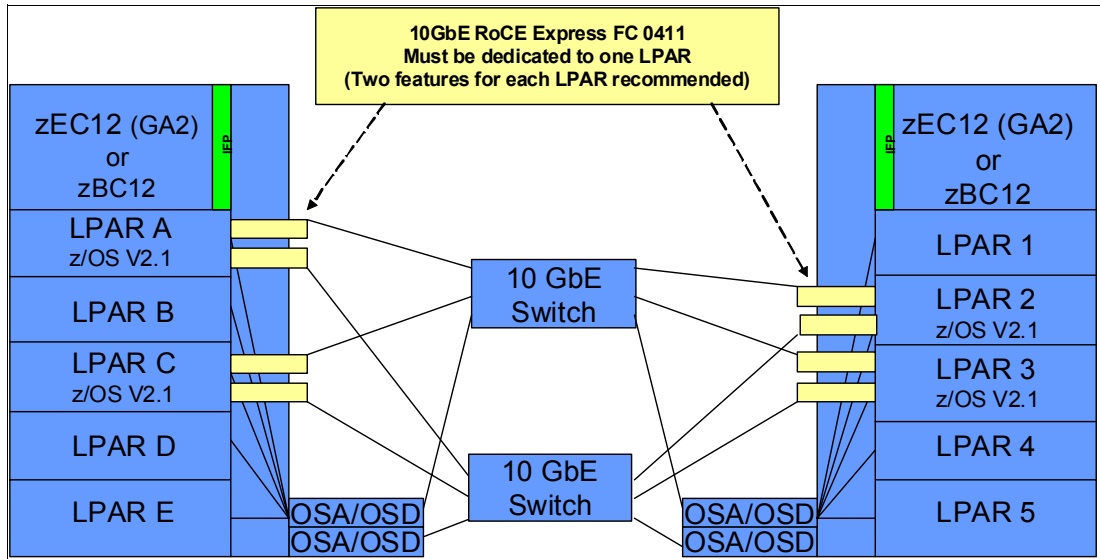


Figure D-4 10GbE RoCE Express sample configuration

An OSA-Express feature, defined as channel-path identifier (CHPID) type OSD, is required to establish SMC-R. Figure D-5 shows the interaction of OSD and the RoCE Network Interface Card (RNIC). The 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-4, 1 GbE OSD connections still can be used instead of 10 GbE and OSD connections can flow through the same 10 GbE switches.

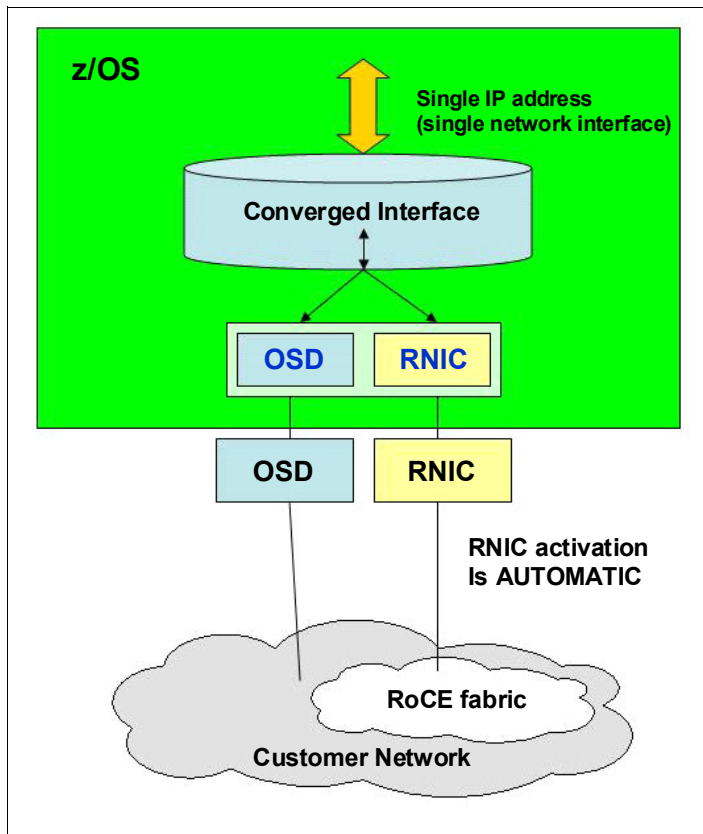


Figure D-5 RNIC and OSD interaction

The following notes refer to Figure D-5 on page 518:

- ▶ The z/OS system administrator only has to configure and manage the OSD interface.
- ▶ The Communication 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.2.3 Hardware configuration definitions

The following hardware configuration definitions (HCD) are important.

Function ID

The RoCE feature is identified by a hexadecimal FUNCTION Identifier (FID) in the range 00 - FF in the HCD or HMC to create the /input/output configuration program (IOCP) input. An FID can only be configured to one LPAR but it is reconfigurable. The RoCE feature in a specific PCIe I/O drawer and slot to be used for the defined FUNCTION can be identified by assigning a physical channel identifier (PCHID). Only one FID is supported by one PCHID.

Physical network (PNet) ID

As one parameter for the FUNCTION statement, the 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. 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. 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-6 shows the three physically separate networks defined by client.

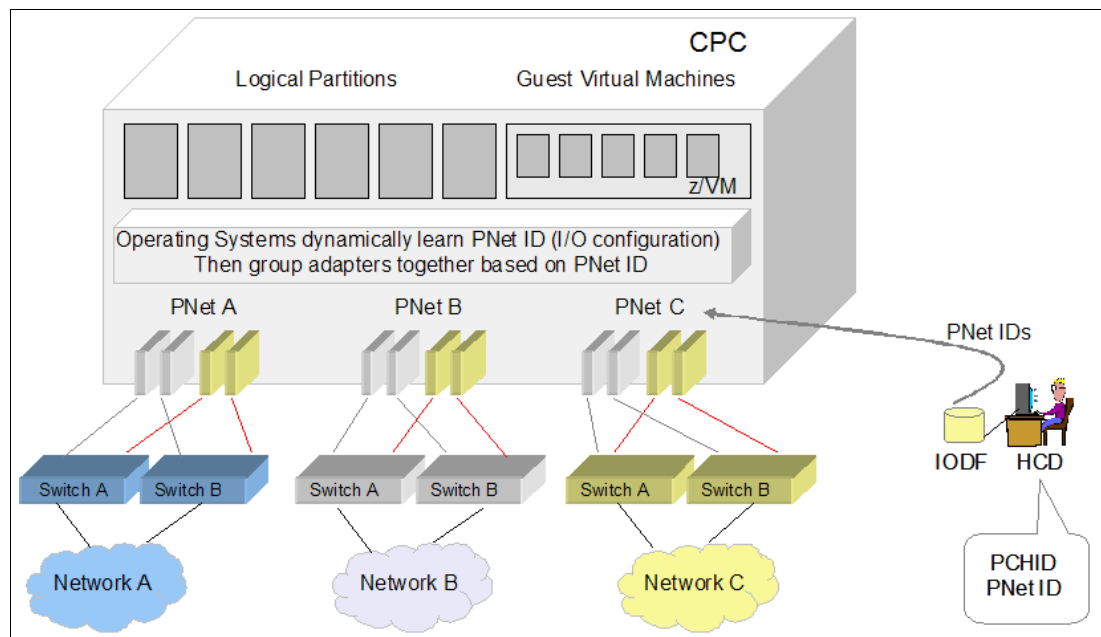


Figure D-6 Physical network ID example

Sample IOCP FUNCTION statement

Example D-1 shows one sample IOCP FUNCTION statement.

Example D-1 IOCP FUNCTION statements

```
FUNCTION FID=10,PART=((LP14),(LP03,LP04,LP12,LP22)),
          PNETID=(NET1,NET2,N3,),PCHID=11C
FUNCTION FID=11,PART=((LP14),(LP03,LP04,LP12,LP22)),
          PNETID=(NET1,NET2,N3,),PCHID=144
```

D.3 Software exploitation

SMC-R can be implemented on the RoCE that can communicate memory to memory, avoiding the CPU resources of TCP/IP, and therefore reducing network latency and improving wall clock time. It focuses on “Time to Value” and widespread performance benefits for all TCP socket-based middleware as shown in Figure D-7:

- ▶ No middleware or application changes (transparent)
- ▶ Ease of deployment (no IP topology changes)
- ▶ LPAR-to-LPAR communication on a single z/OS system
- ▶ Server-to-server communication in a multiple 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)

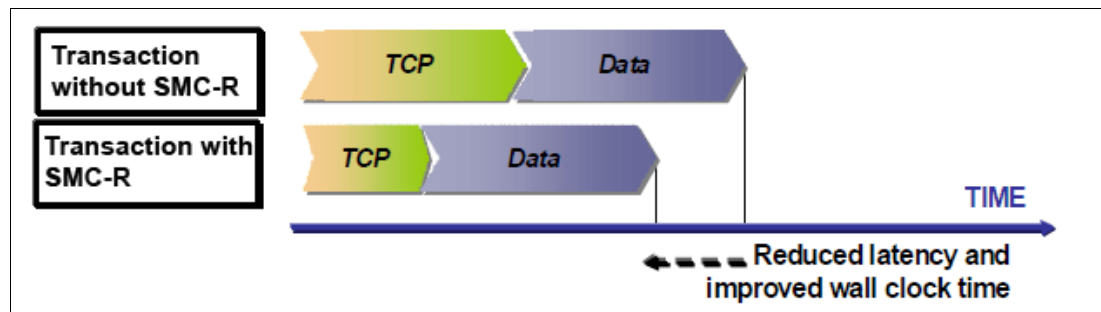


Figure D-7 Reduced latency and improved wall clock time with SMC-R

D.3.1 SMC-R support overview

SMC-R needs both hardware and software support on the zBC12 or zEC12.

Hardware

SMC-R requires the following hardware:

- ▶ PCIe-based RoCE Express (hardware dependencies for zBC12 and zEC12):
 - Dual port 10GbE adapter
 - Only one port per feature
 - Maximum of 16 RoCE Express features per CPC
- ▶ HCD and input/output configuration data set (IOCDs):
 - PCIe FID and RoCE configuration with PNet ID
- ▶ Optional: Standard 10GbE switch (CEE-enabled switch is not required)

- ▶ Required Queued Direct Input/Output (QDIO) Mode OSA connectivity between z/OS LPARs as Figure D-4 on page 518 shows.
- ▶ Adapter needs to be dedicated to a single z/OS LPAR.
- ▶ SMC-R cannot be used in IEDN due to the lack of VLAN enforcement capability.

Software

SMC-R requires the following software:

- ▶ z/OS V2R1 with PTFs is the only supported operating system for the SMC-R protocol:
 - You cannot roll back to previous z/OS releases.
 - You need IOCP 3.4.0.

Statement of direction: In a future z/VM deliverable, IBM plans to offer support for guest exploitation of the 10GbE RoCE Express feature on the IBM zEnterprise EC12 and IBM zEnterprise BC12 servers. This is designed to allow guests to use Shared Memory Communications - Remote Direct Memory Access (SMC-R) using Remote Direct Memory Access over Converged Ethernet (RoCE).

- ▶ IBM is working with its Linux distribution partners to include support in future Linux on System z distribution releases.

D.3.2 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 reside in different CPCs.

The following list shows several typical communication 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), scalability, and so on in the sysplex environment.

When the clients and servers are all in the same ensemble, 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-8 shows a traditional Sysplex Distributor.

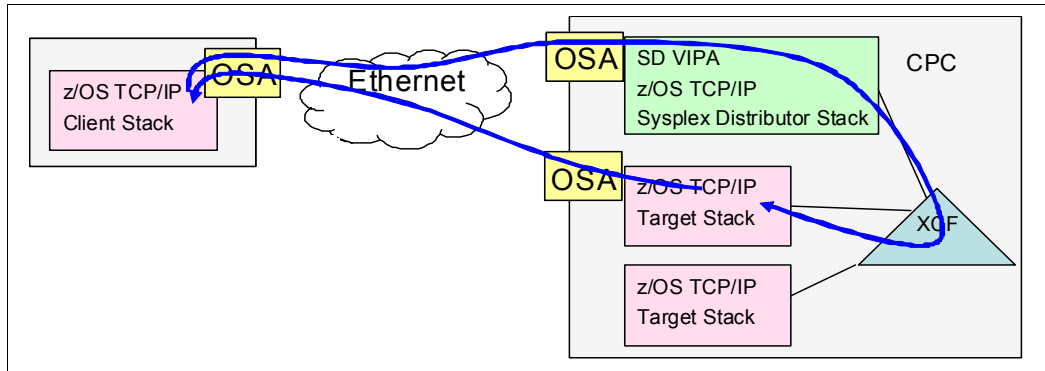


Figure D-8 Sysplex Distributor before RoCE

Figure D-8 shows a traditional Sysplex Distributor:

- ▶ 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-9 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.

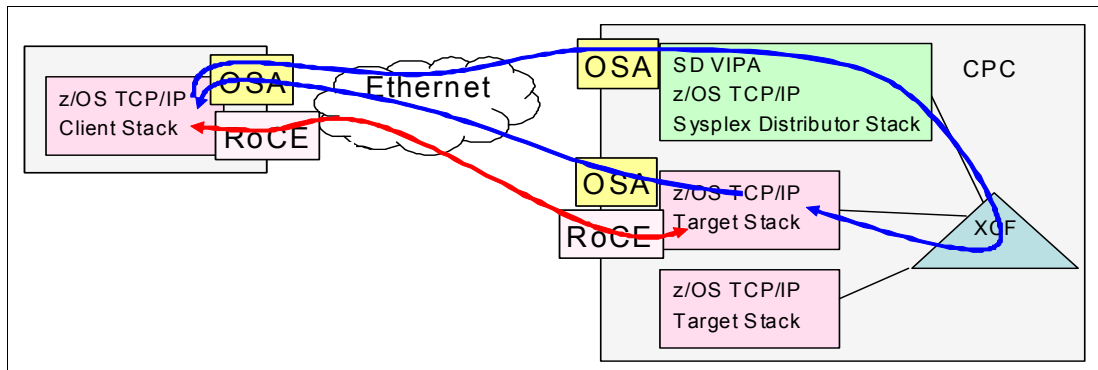


Figure D-9 Sysplex Distributor after RoCE

Note: As with all RoCE communication, the session end also flows over OSAs.

D.3.3 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 via OSD:

- ▶ HCD definitions (install and configure RNICs in the HCD):
 - Add the PNetID for the current OSD.
 - Define PFIDs for RoCE (with the same PNetID).
- ▶ Specify the **GLOBALCONFIG SMCR** parameter (TCP/IP Profile):
 - Must specify at least one PCIe Function ID (PFID):
 - A PFID represents a specific RDMA network interface card (RNIC) adapter.
 - A maximum of 16 PFID values can be coded.
 - Up to eight TCP/IP stacks can share the same PFID in a specific LPAR.
- ▶ 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.*
- ▶ Repeat in each host (at least two hosts).
- ▶ Start the TCP/IP traffic and monitor it with NetStat and IBM VTAM® displays.



zEnterprise Data Compression (zEDC) Express

This appendix briefly describes the optional zEnterprise Data Compression (zEDC) Express feature of the IBM zEnterprise EC12 (zEC12) and IBM zEnterprise BC12 (zBC12) servers, and includes the following topics:

- ▶ Overview
- ▶ zEDC Express
- ▶ Software support

E.1 Overview

The growth of data that needs to be captured, transferred, and stored for large periods of time is not relenting. On the contrary, software-implemented compression algorithms are costly in terms of processor resources and storage costs.

zEDC Express, an optional feature exclusive to zEC12 and zBC12, 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 System z server.

Using the zEDC Express feature by the z/OS V2R1 zEnterprise Data Compression acceleration capability is designed to deliver an integrated solution to help reduce CPU consumption, optimize the performance of compression-related tasks, and 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.

E.2 zEDC Express

zEDC Express is an optional feature (FC 0420), exclusive to the zEC12 and zBC12. 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. Up to two zEDC Express features can be installed per PCIe I/O drawer domain. However, if the domain contains a Flash Express or 10GbE RoCE feature, only one zEDC feature can be installed on that domain.

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).

Adapter support for zEDC is provided by Resource Group (RG) code running on the system integrated firmware processor (IFP). 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.

Consider also the total required data throughput. Consider that if one feature becomes unavailable, the other features need to be able to absorb the load. Therefore, for the best data throughput and availability, install at least two features per RG.

Figure E-1 on page 527 illustrates the PCIe I/O cage structure and the relationships among card slots, domains, and RGs.

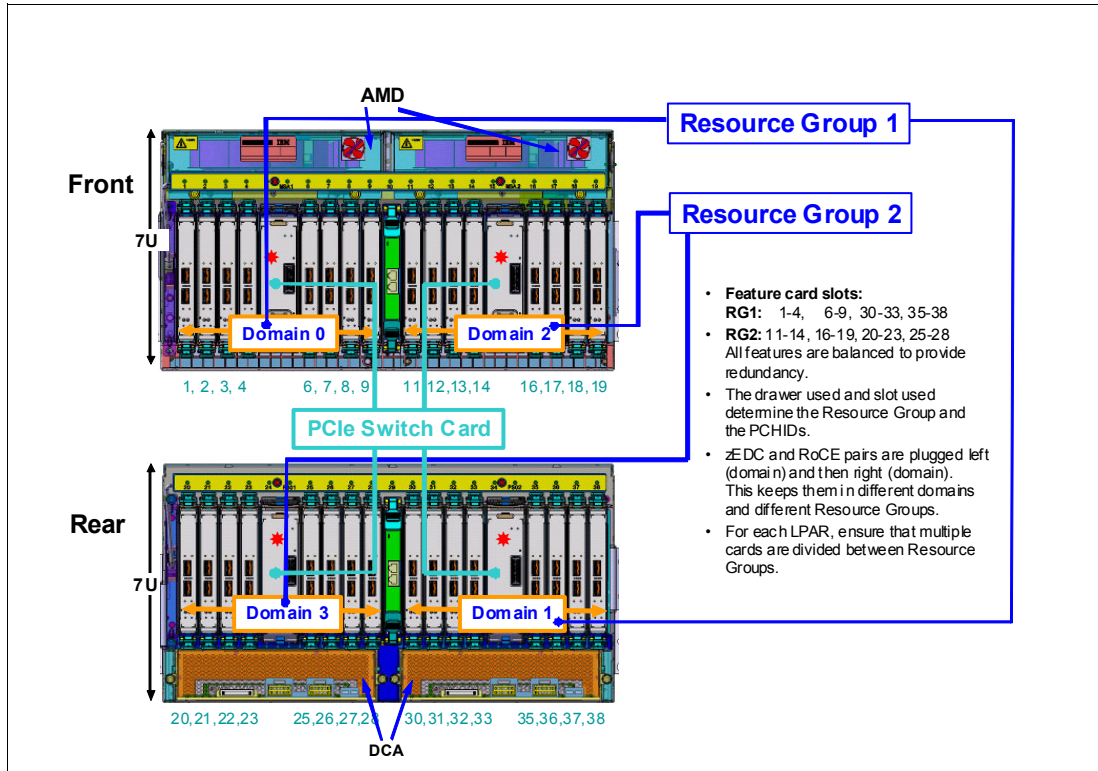


Figure E-1 Relationships among PCIe I/O cage card slots, I/O domains, and RGs

E.3 Software support

Exploitation support of zEDC Express functionality is provided exclusively by z/OS V2R1 zEnterprise Data Compression for both data compression and decompression.

Support for data recovery (decompression) in the case that zEDC is not installed, or installed but not available, on the system, is provided via software on z/OS V2R1, V1R13, and V1R12 with the appropriate PTFs. Software decompression is slow and uses considerable processor resources; therefore, it is not suggested for production environments.

Statements of direction:

z/OS

IBM plans for future updates of IBM 31-bit and 64-bit SDK7 for z/OS Java Technology Edition, Version 7 (5655-W43 and 5655-W44) (IBM SDK7 for z/OS Java) to provide exploitation of the zEDC Express feature and the Shared Memory Communications - Remote Direct Memory Access (SMC-R), which is used by the 10GbE RoCE Express feature.

z/VM V6R3

In a future z/VM deliverable, IBM plans to offer z/VM support for guest exploitation of the IBM zEnterprise Data Compression (zEDC) Express feature on the IBM zEnterprise EC12 and IBM zEnterprise BC12 systems.

IBM System z Batch Network Analyzer

The IBM System z 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 alternate processors.

zBNA can be used to analyze client-provided System Measurement Facility (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 via the IBM intranet:
<http://w3-03.ibm.com/support/techdocs/atmastr.nsf/WebIndex/PRS5126>
- ▶ IBM Business Partners can obtain zBNA and other CPS tools via the Internet:
https://www.ibm.com/partnerworld/wps/servlet/mem/ContentHandler/tech_PRS5133
- ▶ IBM clients can obtain zBNA and other CPS tools via the Internet:
<http://www-03.ibm.com/support/techdocs/atmastr.nsf/WebIndex/PRS5132>



Native Peripheral Component Interconnect Express (PCIe)

In this appendix, we introduce the concept of managing native Peripheral Component Interconnect Express (PCIe) features (10GbE Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE) Express and zEnterprise Data Compression (zEDC) Express). The new concept requires the use of an integrated firmware processor (IFP) and Resource Groups (RGs) associated to the physical location of the feature cards. We describe how these features are implemented into the PCIe I/O structure of the IBM zEnterprise EC12 (zEC12). The following components and functions manage the native PCIe features:

- ▶ Design of native PCIe I/O adapter management
- ▶ Native PCIe adapter
- ▶ Integrated firmware processor (IFP)
- ▶ Resource Groups (RGs)
- ▶ Management tasks
- ▶ zEDC Express
- ▶ 10GbE RoCE Express
- ▶ Native PCIe feature definitions

F.1 Design of native PCIe I/O adapter management

There are feature card types introduced on zEC12 and IBM zEnterprise BC12 (zBC12), which require a new design to manage these adapters. The native PCIe features are 10GbE RoCE Express and zEDC Express.

These adapters are plugged in a PCIe I/O drawer together with existing I/O features, but they are managed in a different way than the existing I/O adapters and features. The native PCIe feature cards are exclusively plugged into the PCIe I/O drawer and have a physical channel ID (PCHID) assigned according to the physical location in the PCIe I/O drawer.

On existing features that were plugged into an I/O drawer or I/O cage, all adapter layer functions are integrated into the adapter hardware. For the new features introduced by zEC12 and zBC12 CPCs, the adapter layer function is now handled by an integrated firmware processor (IFP).

We describe the following topics:

- ▶ Native PCIe adapter
- ▶ Integrated firmware processor
- ▶ Resource Groups
- ▶ Management functions

F.2 Native PCIe adapter

For traditional PCIe I/O adapters, such as the Open Systems Adapter (OSA) and Fibre Channel connection (FICON) cards, the diagnostic program and device drivers are downloaded from the Support Element (SE) to the application-specific integrated circuit (ASIC) chips that reside on those cards. With the introduction of the integrated firmware processor (IFP) and the native PCIe adapters, which do not have an ASIC chip, the device drivers for these native PCIe adapters were moved to the operating systems. The adapter layer function runs on the IFP and uses two RGs.

All virtualization, recovery, diagnostics, failover, concurrent firmware (FW) updates, and so on functions on traditional I/O features are performed on the adapter level. For the native PCIe features, these functions are done by the integrated firmware processor (IFP).

F.3 Integrated firmware processor (IFP)

The integrated firmware processor (IFP) is used to manage native PCIe adapters installed in a PCIe I/O drawer. On previous systems, this processor was not used but known as a *reserved processor*. The IFP is allocated from a pool of processor units (PUs) available for the whole system. Because IFP is exclusively used to manage native PCIe adapters, it is not taken from the pool of PUs that can be characterized for the client's use.

If a native PCIe feature is present in the system, the IFP is initialized and allocated during the systems' power-on reset (POR) phase. Although the IFP is allocated to one of the physical PUs, it is not visible to the client. In an error or failover scenario, the IFP acts like any other PU (that is, sparing).

F.4 Resource Groups (RGs)

To manage the PCIe features, the IFP has two Resource Groups (RGs) allocated. The two RGs handle the adapter layer function of the native PCIe feature cards. Each I/O domain in a PCIe I/O drawer is assigned to one of the two RGs. There are four I/O domains in the PCIe I/O drawer, where I/O domain 0 and 1 are handled by RG 1, and I/O domain 2 and 3 are handled by RG 2. Figure F-1 shows the relationships among the I/O domains, I/O slots, and the RGs that are managed by the IFP.

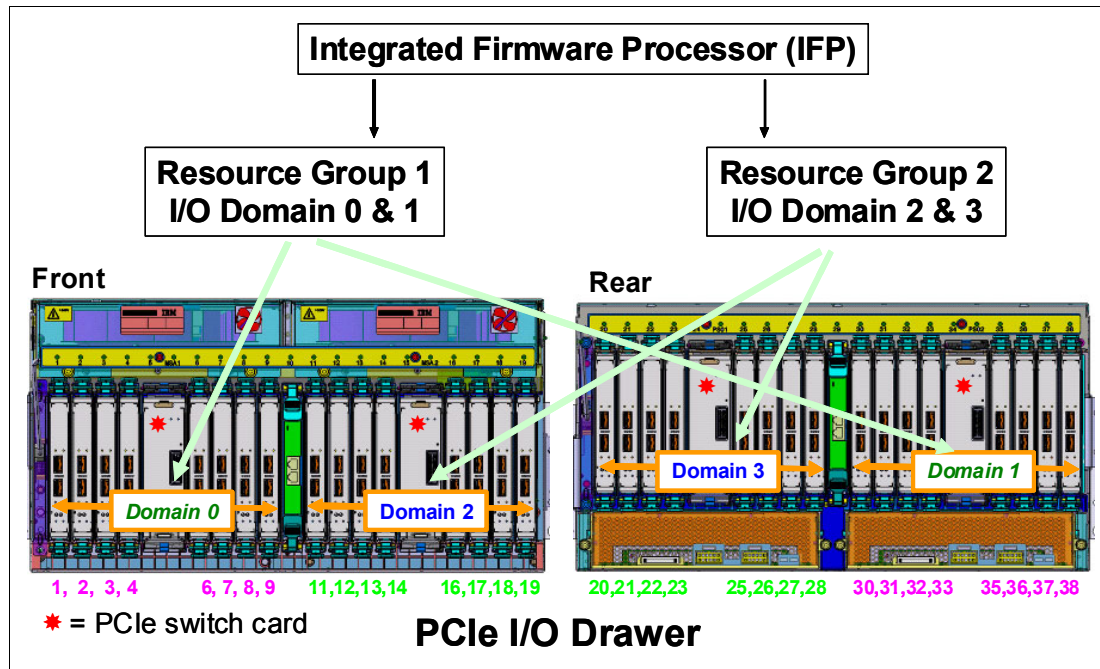


Figure F-1 I/O domains and RGs managed by the IFP

Up to five PCIe I/O drawers are supported on the zEC12. Native PCIe features of the same type are configured to different RGs, PCIe I/O drawers, and I/O domains to prevent a single point of failure. Other than the existing PCIe feature cards (FICON, OSA, and Crypto), each I/O domain supports a total of two native PCIe feature cards. The following native PCIe features are supported in the PCIe I/O drawer:

- ▶ Flash Express
- ▶ zEDC Express
- ▶ 10GbE RoCE Express

Only zEDC Express and 10GbE RoCE Express features are managed by the IFP, but the Flash Express feature is counted when configuring native PCIe features into the PCIe I/O drawer.

The following management functions are provided for the native PCIe features by the IFP:

- ▶ Firmware update
- ▶ Error recovery
- ▶ Maintenance functions

F.5 Native PCIe feature plugging rules

There is a limitation in the number of features that you can order for a specific native PCIe feature. There is also a maximum number of PCIe features. This maximum is based on the maximum number of physical native PCIe cards per PCIe I/O drawer, which is eight, but also on the presence of the Flash Express feature. The Flash Express feature is different from the two other native PCIe features because it does not use the IFP or RGs. However, remember that *one* Flash Express feature takes up *two* slots in the PCIe I/O drawer, therefore limiting the remaining slots for the other native PCIe features.

The zEC12 system can have up to five PCIe I/O drawers for a maximum of 40 slots for native PCIe features (10GbE RoCE Express and zEDC Express) and the Flash Express feature. Up to 32 slots can be used for the maximum of 16 slots for 10GbE RoCE Express, eight slots for the zEDC Express feature, and up to eight slots for Flash Express.

Table F-1 shows the dependencies and the maximum number of native PCIe features installable in the PCIe I/O drawer.

Table F-1 Maximum number of combined native PCIe features

Number of Flash Express features	Total of combined zEDC and RoCE features ^a	Minimum number of PCIe I/O drawers required
0	8	1
	16	2
	24	3
1	6	1
	14	2
	22	3
	24	4
2	4	1
	12	2
	20	3
	24	4
3	10	2
	18	3
	24	4
4 ^b	8	2
	16	3
	24	4

a. The maximum number of zEDC features is eight, and the maximum number of 10GbE RoCE features is 16, for a total of 24 combined features.

b. The maximum number of Flash Express features.

Each Flash Express feature (FC 0402) occupies two slots in the PCIe I/O drawer. Each 10GbE RoCE Express (FC 0411) or zEDC Express (FC 0420) feature occupies one slot. If one Flash Express feature is installed in a zEC12 with one PCIe drawer, it allows the installation of up to six 10GbE RoCE or zEDC features in this PCIe I/O drawer. Any additional 10GbE RoCE or zEDC feature (total of more than six) requires a second PCIe I/O drawer.

F.6 Management tasks

For earlier I/O features, parts of the management function were included on the adapter. The integrated firmware processor (IFP) performs all management tasks on the native PCIe features:

- ▶ Firmware update
- ▶ Error recovery
- ▶ Maintenance tasks

F.6.1 Firmware update

Microcode Change Level (MCL) upgrades on native PCIe adapters or on the code of the RGs require the specific adapter or all native PCIe adapters managed by the specific RG (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 RG at a time. While one RG is offline, the second RG and all adapters in it remain active at all times. An MCL application for a native PCIe adapter or RG is not possible if an error condition exists within the other RG.

F.6.2 Error recovery

In an error in one of the RGs or features assigned to one of the RGs occurs, the IFP manages error recovery and collects error data. The error data is sent by the IFP to the Support Element (SE), which then provides a message on the SE and the Hardware Management Console (HMC). In an error that requires maintenance, a call home to the IBM Support system is initiated by the HMC.

F.6.3 Maintenance tasks

Any maintenance action on a native PCIe feature is managed by the IFP. This includes 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 RG (if applicable).

F.7 zEDC Express

zEDC Express is an optional feature (FC 0420), exclusive to the zEC12 and zBC12. It is designed to provide hardware-based acceleration for data compression and decompression.

The feature installs exclusively on the PCIe I/O drawer. Up to two zEDC Express features can be installed for each PCIe I/O drawer domain. However, if the domain contains a Flash Express or 10GbE RoCE feature, only one zEDC feature can be installed on that domain.

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

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

For details about zEDC Express, see Appendix E, “zEnterprise Data Compression (zEDC) Express” on page 525.

F.8 10GbE RoCE Express

The 10 Gigabit Ethernet (10GbE) RoCE Express feature (FC 0411) is designed to help reduce the consumption of CPU resources for applications using the TCP/IP stack, such as WebSphere accessing a DB2 database. Use of the 10GbE RoCE Express feature might also help to reduce network latency with memory-to-memory transfers using Shared Memory Communications - Remote Direct Memory Access (SMC-R) in z/OS V2.1. It is transparent to applications and can be used for LPAR-to-LPAR communication on a single z/OS system or server-to-server communication in a multiple CPC environment.

This feature resides exclusively in the PCIe I/O drawer (FC 4009) and is exclusive to the zEC12 and zBC12. The 10GbE RoCE Express feature has one PCIe adapter. It does not use a channel-path identifier (CHPID). It is defined using the Input/Output Configuration Program (IOCP) FUNCTION statement or in the Hardware Configuration Definition (HCD).

Each feature must be dedicated to an LPAR. Only one of the two ports can be used at the same time.

For details about the 10GbE RoCE Express card, see Appendix D, “Remote Direct Memory Access over Converged Ethernet (RoCE)” on page 513.

F.9 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 (RG1 and RG2) for availability and serviceability. In Figure F-2 on page 535, you can see a sample of the PCHID report for a configuration with four of each of the previously mentioned features and how they are spread across RG1 and RG2.

Even though Flash Express features are counted as native PCIe cards for the total number of Native PCIe features, they are not part of any RG.

Source	Cage	Slot	F/C ^a	PCHID/Ports or AID	Comment
A21/D8/J01	A02B	01	0420	100/	RG1
A21/D8/J01	A02B	09	0411	11C/D1D2	RG1
A21/D1/J02	A02B	11	0411	120/D1D2	RG2
A21/D1/J02	A02B	14	0420	12C/	RG2
A21/D8/J02	A02B	20	0420	140/	RG2
A21/D8/J02	A02B	21	0411	144/D1D2	RG2
A21/D1/J01	A02B	37	0411	178/D1D2	RG1
A21/D1/J01	A02B	38	0420	17C/	RG1

a. Feature code (F/C) 0411 = 10GbE RoCE Express, and F/C 0420 = zEDC Express

Figure F-2 Sample output of AO data or PCHID report

The native PCIe features are not part of the traditional channel subsystem (CSS). They do not have a Channel Path ID (CHPID) assigned, but they have a Physical Channel ID (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 input/output configuration program (IOCP) FUNCTION statement is introduced, including several feature-specific parameters. In Figure F-3, you can see examples of the specific statements for the 10GbE RoCE Express and zEDC Express features. In this example, we define two zEDC features (PCHID 100 and 12C) and two 10GbE RoCE Express features (PCHID 11C and 144).

```

zEDC Express Functions for LPAR LP14, Reconfigurable to LP01:
FUNCTION FID=01,VF=1,PART=((LP14),(LP01)),PCHID=100
FUNCTION FID=02,VF=1,PART=((LP14),(LP01)),PCHID=12C

zEDC Express Functions for LPAR LP15, Reconfigurable to LP02:
FUNCTION FID=03,VF=2,PART=((LP15),(LP02)),PCHID=100
FUNCTION FID=04,VF=2,PART=((LP15),(LP02)),PCHID=12C

10GbE RoCE Express Function for LPAR LP14, Reconfigurable
FUNCTION FID=05,PART=((LP14),(LP03,LP04,LP12,LP22)),          *
      PNETID=(NET1,NET2,NET3,),PCHID=11C
FUNCTION FID=06,PART=((LP14),(LP03,LP04,LP12,LP22)),          *
      PNETID=(NET1,NET2,NET3,),PCHID=144

```

Figure F-3 Example of I/O configuration data set (IOCDs) definition for 10GbE RoCE Express feature

FUNCTION Identifier (FID)

The FUNCTION Identifier (FID) is a hexadecimal number between 00 and FF, which you use to assign a physical-channel identifier (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 zEnterprise 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.

Virtual Function number (VF)

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 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.

Physical Network Identifier (PNETID)

The PNETID is required to set up the Shared Memory Communications - Remote Direct Memory Access (SMC-R) communication between two 10GbE RoCE Express features. Each FUNCTION definition supports up to four PNETIDs.

RoCE consideration: Because the initial link setup between two 10GbE RoCE Express features is done through normal OSA Express ports, you need to add the (same) **PNETID** parameter to several OSA OSD ports between the LPARs that you want to connect via RoCE.

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 zEnterprise EC12 Technical Introduction*, SG24-8050
- ▶ *IBM System z Connectivity Handbook*, SG24-5444
- ▶ *IBM zEnterprise Ec12 Configuration Setup*, SG24-8034
- ▶ *Extending z/OS System Management Functions with IBM zAware*, SG24-8070
- ▶ *IBM zEnterprise Unified Resource Manager*, SG24-7921
- ▶ *Implementing IBM Systems Director Active Energy Manager 4.1.1*, SG24-7780
- ▶ *Using IBM System z As the Foundation for Your Information Management Architecture*, REDP-4606

You can search for, view, download or order these documents and other Redbooks, Redpapers, Web Docs, draft, and additional materials, at:

ibm.com/redbooks

Other publications

These publications are also relevant as further information sources:

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

Online resources

These websites are also relevant as further information sources:

- ▶ IBM Resource Link:
<http://www.ibm.com/servers/resourcelink/>
- ▶ IBM Communication Controller for Linux on System z:
<http://www-01.ibm.com/software/network/cc1/>
- ▶ FICON channel performance:
<http://www.ibm.com/systems/z/connectivity/>

- ▶ Large Systems Performance Reference measurements:
<https://www-304.ibm.com/servers/resourceLink/lib03060.nsf/pages/lsprindex>
- ▶ Materialized Query Tables (MQTs):
<http://www.ibm.com/developerworks/data/library/techarticle/dm-0509melnyk>
- ▶ IBM zIIP:
<http://www-03.ibm.com/systems/z/advantages/ziip/about.html>
- ▶ Parallel Sysplex coupling facility configuration:
<http://www.ibm.com/systems/z/advantages/ps0/index.html>
- ▶ Parallel Sysplex CFCC code levels:
<http://www.ibm.com/systems/z/ps0/cftable.html>
- ▶ IBM InfiniBand:
<http://www.infinibandta.org>
- ▶ ESCON to FICON migration:
<http://www-935.ibm.com/services/us/index.wss/offering/its/c337386u66547p02>
- ▶ Optica Technologies Inc.:
<http://www.opticatech.com/>
- ▶ FICON channel performance:
http://www-03.ibm.com/systems/z/hardware/connectivity/ficon_performance.html
- ▶ z/OS deliverables on the web:
<http://www.ibm.com/systems/z/os/zos/downloads/>
- ▶ Linux on System z:
<http://www.ibm.com/developerworks/linux/linux390/>
- ▶ ICSF versions and FMID cross-references:
<http://www.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/TD103782>
- ▶ z/OS:
<http://www.ibm.com/systems/support/z/zos/>
- ▶ z/VM:
<http://www.ibm.com/systems/support/z/zvm/>
- ▶ z/TPF:
<http://www.ibm.com/software/hpf/tpf/pages/maint.htm>
- ▶ z/VSE:
<http://www.ibm.com/servers/eserver/zseries/zvse/support/preventive.html>
- ▶ IBM license charges on System z:
<http://www.ibm.com/servers/eserver/zseries/swprice/zna1c.html>

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services



Redbooks

IBM zEnterprise EC12 Technical Guide

(1.0" spine)

0.875" x 1.498"

460 x 788 pages



IBM zEnterprise EC12 Technical Guide



Redbooks®

Describes the zEnterprise System and related features and functions

Discusses hardware and software capabilities

Explains virtualizing and managing the infrastructure for complex applications

This IBM Redbooks publication addresses the new IBM zEnterprise System. This system consists of the IBM zEnterprise EC12 (zEC12), an updated IBM zEnterprise Unified Resource Manager, and the IBM zEnterprise BladeCenter Extension (zBX) Model 003.

The zEC12 is designed with improved scalability, performance, security, resiliency, availability, and virtualization. The superscalar design allows the zEC12 to deliver a record level of capacity over the prior System z servers. It is powered by 120 of the world's most powerful microprocessors. These microprocessors run at 5.5 GHz and are capable of running more than 75,000 millions of instructions per second (MIPS). The zEC12 Model HA1 is estimated to provide up to 50% more total system capacity than the IBM zEnterprise 196 (z196) Model M80.

The zBX Model 003 infrastructure works with the zEC12 to enhance System z virtualization and management. It does so through an integrated hardware platform that spans mainframe, IBM POWER7, and IBM System x technologies. Through the Unified Resource Manager, the zEnterprise System is managed as a single pool of resources, integrating system and workload management across the environment.

This book provides information about the zEnterprise System 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 zEnterprise System 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 System z technology and terminology.

INTERNATIONAL TECHNICAL SUPPORT ORGANIZATION

BUILDING TECHNICAL INFORMATION BASED ON PRACTICAL EXPERIENCE

IBM Redbooks are developed by the IBM International Technical Support Organization. Experts from IBM, Customers and Partners from around the world create timely technical information based on realistic scenarios. Specific recommendations are provided to help you implement IT solutions more effectively in your environment.

**For more information:
ibm.com/redbooks**

SG24-8049-01

ISBN 0738438863