

IBM z14 Model ZR1 Technical Guide

Octavian Lascu
Hervey Kamga
Frank Packheiser
Martijn Raave
John Troy
Bill White



IBM Z



International Technical Support Organization

IBM 14 Model ZR1 Technical Guide

October 2018

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

Second Edition (October 2018)

This edition applies to IBM Z[®]: IBM z14[™] Model ZR1, IBM z13[™], IBM z13s[™], IBM zEnterprise EC12 (zEC12), and IBM zEnterprise BC12 (zBC12).

© Copyright International Business Machines Corporation 2018. 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	xiii
Trademarks	xiv
Preface	xv
Authors	xv
Now you can become a published author, too!	xvii
Comments welcome	xvii
Stay connected to IBM Redbooks	xviii
Chapter 1. Introducing the new IBM Z family member: IBM z14 Model ZR1	1
1.1 A digital transformation pillar	2
1.2 z14 ZR1 highlights	3
1.2.1 Models and upgrade paths	4
1.2.2 Rack and cabling	5
1.2.3 CPC drawer	6
1.2.4 PCIe+ I/O drawer	8
1.2.5 I/O subsystem and I/O features	10
1.3 z14 ZR1 capacity and performance	12
1.4 z14 ZR1 virtualization	13
1.4.1 PR/SM mode	13
1.4.2 Dynamic Partition Manager mode	14
1.4.3 LPAR types on z14 ZR1	14
1.4.4 Coupling facility	15
1.4.5 z/VM-mode	15
1.4.6 IBM Secure Service Container	15
1.4.7 GDPS Virtual Appliance	16
1.5 z14 ZR1 RAS	16
1.6 Hardware Management Consoles and Support Elements	17
1.7 Supported operating systems and compilers	17
1.7.1 Operating systems summary	17
1.7.2 IBM Z compilers	19
Chapter 2. Central processor complex hardware components	21
2.1 System overview: Frame and drawers	22

2.1.1	z14 ZR1 configurations	23
2.1.2	PCIe+ I/O drawer	24
2.2	16U Reserved feature (FC 0617)	26
2.3	CPC drawer	27
2.3.1	Oscillator cards	31
2.3.2	System control	32
2.3.3	CPC drawer power	33
2.4	Single chip modules	33
2.4.1	Processor Unit Single Chip Module	34
2.4.2	Processor Unit (Core)	36
2.4.3	PU characterization	37
2.4.4	System Controller SCM (chip)	37
2.5	Memory	38
2.5.1	Memory subsystem topology	39
2.5.2	Redundant array of independent memory	39
2.5.3	Memory configurations	40
2.5.4	Memory upgrades	44
2.5.5	Virtual Flash Memory	45
2.5.6	Preplanned memory	45
2.6	Reliability, availability, and serviceability	46
2.6.1	RAS in the CPC memory subsystem	47
2.6.2	General z14 ZR1 RAS features	47
2.7	Connectivity	49
2.7.1	Redundant I/O interconnect	49
2.7.2	CPC drawer upgrades	51
2.7.3	System upgrades	52
2.7.4	Concurrent PU conversions	53
2.7.5	Model capacity identifier	53
2.7.6	Capacity Backup Upgrade	55
2.7.7	On/Off Capacity on Demand and CPs	57
2.8	Power and cooling	58
2.8.1	Considerations	58
2.8.2	Power and weight estimation tool	59
2.8.3	Cooling requirements	59
2.9	Summary	60
Chapter 3. Central processor complex system design		63
3.1	Overview	64
3.2	Design highlights	64
3.3	CPC drawer design	66
3.3.1	Cache levels and memory structure	67
3.3.2	CPC drawer topology	70
3.4	Processor unit design	70
3.4.1	Simultaneous multithreading	70
3.4.2	Single-instruction multiple-data (enhanced for z14 ZR1)	72
3.4.3	Out-of-Order execution	74
3.4.4	Superscalar processor	76
3.4.5	Compression and cryptography accelerators on a chip	76
3.4.6	Decimal floating point accelerator	78
3.4.7	IEEE floating point	79
3.4.8	Processor error detection and recovery	79
3.4.9	Branch prediction	79
3.4.10	Wild branch	80

3.4.11	Translation lookaside buffer	81
3.4.12	Instruction fetching, decoding, and grouping	81
3.4.13	Extended Translation Facility	82
3.4.14	Instruction set extensions	82
3.4.15	Transactional Execution	82
3.4.16	Runtime Instrumentation	82
3.5	Processor unit functions	83
3.5.1	Overview	83
3.5.2	Central processors	84
3.5.3	Integrated Facility for Linux	86
3.5.4	Internal Coupling Facility	86
3.5.5	IBM Z Integrated Information Processor	88
3.5.6	System assist processors	92
3.5.7	Reserved processors	93
3.5.8	Integrated firmware processor	93
3.5.9	Sparing rules	93
3.6	Memory design	94
3.6.1	Overview	94
3.6.2	Main storage	96
3.6.3	Hardware system area	96
3.6.4	Virtual Flash Memory	97
3.7	Logical partitioning	97
3.7.1	Overview	97
3.7.2	Storage operations	103
3.7.3	Reserved storage	105
3.7.4	Logical partition storage granularity	106
3.7.5	LPAR dynamic storage reconfiguration	106
3.8	Intelligent Resource Director	107
3.9	Clustering technology	109
3.9.1	Coupling Facility Control Code	110
3.9.2	Coupling Thin Interrupts	113
3.9.3	Dynamic CF dispatching	113
3.10	Virtual Flash Memory	115
3.10.1	IBM Z Virtual Flash Memory overview	115
3.10.2	VFM feature	115
3.10.3	VFM administration	115
	Chapter 4. Central processor complex I/O system structure	117
4.1	Introduction to I/O infrastructure	118
4.1.1	I/O infrastructure	118
4.1.2	PCIe Generation 3	119
4.2	I/O system overview	120
4.2.1	Characteristics	120
4.2.2	Supported I/O features	121
4.3	PCIe+ I/O drawer	122
4.3.1	PCIe+ I/O drawer offerings	125
4.4	CPC drawer fanouts	125
4.4.1	PCIe Generation 3 fanout (FC 0173)	126
4.4.2	Integrated Coupling Adapter (FC 0172)	126
4.4.3	Fanout considerations	127
4.5	I/O features (cards)	128
4.5.1	I/O feature card ordering information	129
4.5.2	Physical channel ID report	130

4.6	Connectivity	133
4.6.1	I/O feature support and configuration rules	133
4.6.2	Storage connectivity	136
4.6.3	Network connectivity	142
4.6.4	Parallel Sysplex connectivity	153
4.7	Cryptographic functions	159
4.7.1	CPACF functions (FC 3863)	159
4.7.2	Crypto Express6S feature (FC 0893)	159
4.7.3	Crypto Express5S feature (FC 0890)	159
4.8	Integrated Firmware Processor	160
4.9	zEDC Express	160
Chapter 5.	Central processor complex channel subsystem	161
5.1	Channel subsystem	162
5.1.1	Multiple logical channel subsystems	163
5.1.2	Multiple subchannel sets	164
5.1.3	Channel path spanning	167
5.2	I/O configuration management	170
5.3	Channel subsystem summary	171
Chapter 6.	Cryptographic features	173
6.1	Cryptography enhancements on IBM z14 ZR1	174
6.2	Cryptography overview	175
6.2.1	Modern cryptography	175
6.2.2	Kerckhoffs' principle	176
6.2.3	Keys	176
6.2.4	Algorithms	178
6.3	Cryptography on IBM z14 ZR1	179
6.4	CP Assist for Cryptographic Functions	182
6.4.1	Cryptographic synchronous functions	184
6.4.2	CPACF protected key	185
6.5	Crypto Express6S	188
6.5.1	Cryptographic asynchronous functions	190
6.5.2	Crypto Express6S as a CCA coprocessor	191
6.5.3	Crypto Express6S as an EP11 coprocessor	196
6.5.4	Crypto Express6S as an accelerator	197
6.5.5	Managing Crypto Express6S	197
6.6	TKE workstation	200
6.6.1	Logical partition, TKE host, and TKE target	203
6.6.2	Optional smart card reader	203
6.6.3	TKE hardware support and migration information	203
6.7	Cryptographic functions comparison	205
6.8	Cryptographic operating system support for z14 ZR1	207
Chapter 7.	Operating system support	209
7.1	Operating systems summary	210
7.2	Support by operating system	211
7.2.1	z/OS	211
7.2.2	z/VM	211
7.2.3	z/VSE	212
7.2.4	z/TPF	212
7.2.5	Linux on Z	212
7.2.6	KVM hypervisor	213
7.3	z14 ZR1 features and function support overview	214

7.3.1 Supported CPC functions	214
7.3.2 Coupling and clustering	217
7.3.3 Network connectivity	221
7.3.4 Cryptographic functions	225
7.3.5 Special purpose features	227
7.4 Support by features and functions	227
7.4.1 LPAR configuration and management	227
7.4.2 Base CPC features and functions	231
7.4.3 Coupling and clustering features and functions	240
7.4.4 Storage connectivity-related features and functions	244
7.4.5 Networking features and functions	255
7.4.6 Cryptography features and functions support	265
7.4.7 Special-purpose features and functions	270
7.5 z/OS migration considerations	271
7.5.1 General guidelines	272
7.5.2 Hardware Fix Categories (FIXCATs)	272
7.5.3 Coupling links	273
7.5.4 z/OS XL C/C++ considerations	274
7.5.5 z/OS V2.3	274
7.6 z/VM migration considerations	275
7.6.1 z/VM 7.1	275
7.6.2 z/VM 6.4	275
7.6.3 ESA/390-compatibility mode for guests	276
7.6.4 Capacity	276
7.7 z/VSE migration considerations	276
7.8 Software licensing	277
7.9 References	280
Chapter 8. System upgrades	281
8.1 Upgrade types	283
8.1.1 Overview of upgrade types	283
8.1.2 CoD for z14 ZR1 systems terminology	284
8.1.3 Permanent upgrades	286
8.1.4 Temporary upgrades	287
8.2 Concurrent upgrades	287
8.2.1 Upgrades	288
8.2.2 Customer Initiated Upgrade facility	290
8.2.3 Concurrent upgrade functions summary	294
8.3 Miscellaneous equipment specification upgrades	294
8.3.1 MES upgrade for PUs	295
8.3.2 MES upgrades for memory	296
8.3.3 MES upgrades for I/O	296
8.3.4 Feature on Demand	297
8.3.5 Plan-ahead feature	298
8.4 Permanent upgrade through the CIU facility	298
8.4.1 Ordering	300
8.4.2 Retrieval and activation	301
8.5 On/Off Capacity on Demand	302
8.5.1 Overview	302
8.5.2 Capacity Provisioning Manager	303
8.5.3 Ordering	304
8.5.4 On/Off CoD testing	307
8.5.5 Activation and deactivation	309

8.5.6 Termination	309
8.5.7 z/OS capacity provisioning	310
8.6 Capacity for Planned Event	314
8.7 Capacity Backup	316
8.7.1 Ordering	316
8.7.2 CBU activation and deactivation	318
8.7.3 Automatic CBU enablement for GDPS	319
8.8 Nondisruptive upgrades	320
8.8.1 Components	320
8.8.2 Concurrent upgrade considerations	321
8.9 Summary of Capacity on-Demand offerings	325
Chapter 9. Reliability, availability, and serviceability	327
9.1 RAS strategy	328
9.2 Structure change	328
9.3 Technology change	329
9.4 Reducing complexity	331
9.5 Reducing touches	331
9.6 z14 ZR1 availability characteristics	333
9.7 z14 ZR1 RAS functions	335
9.7.1 Scheduled outages	336
9.7.2 Unscheduled outages	337
9.8 z14 ZR1 Enhanced Driver Maintenance	338
9.8.1 Resource Group and native PCIe MCLs	339
9.9 RAS capability for the HMC and SE	340
Chapter 10. Environmental requirements	343
10.1 Power and cooling	344
10.1.1 Power requirements and consumption	344
10.1.2 Cooling requirements	345
10.2 Physical specifications	347
10.3 Physical planning	348
10.3.1 Raised floor or non-raised floor	348
10.3.2 Top Exit cabling feature (optional)	350
10.3.3 Top or bottom exit cables	351
10.3.4 Bottom Exit cabling feature	352
10.3.5 Frame Bolt-down kit	352
10.3.6 Service clearance areas	352
10.4 Energy management	353
10.4.1 Environmental monitoring	354
Chapter 11. Hardware Management Console and Support Elements	357
11.1 Introduction to the HMC and SE	358
11.2 HMC and SE changes and new features	358
11.2.1 Driver Level 36 HMC and SE new features	358
11.2.2 Driver Level 32 HMC and SE changes and features	359
11.2.3 Firmware Integrity Monitoring and z14 HMC	360
11.2.4 Rack-mounted HMC	361
11.2.5 New SEs	362
11.2.6 New backup options for HMCs and primary SEs	362
11.2.7 SE driver support with the HMC driver	365
11.2.8 HMC feature codes	366
11.2.9 User interface	367
11.2.10 Customize Product Engineering Access: Best practice	367

11.3 HMC and SE connectivity	367
11.3.1 Network planning for the HMC and SE	369
11.3.2 Hardware prerequisite changes	371
11.3.3 TCP/IP Version 6 on the HMC and SE	372
11.3.4 OSA Support Facility changes	372
11.3.5 Assigning addresses to the HMC and SE	373
11.3.6 HMC Multi-factor authentication	373
11.4 Remote Support Facility	374
11.4.1 Security characteristics	374
11.4.2 RSF connections to IBM and Enhanced IBM Service Support System	375
11.4.3 HMC and SE remote operations	375
11.5 HMC and SE capabilities	377
11.5.1 Central processor complex management	377
11.5.2 LPAR management	377
11.5.3 Operating system communication	379
11.5.4 HMC and SE microcode	380
11.5.5 Monitoring	383
11.5.6 Capacity on-demand support	385
11.5.7 Server Time Protocol support	386
11.5.8 CTN Split and Merge	388
11.5.9 NTP client and server support on the HMC	389
11.5.10 Security and user ID management	390
11.5.11 HMC 2.14.1 Enhancements	391
11.5.12 System Input/Output Configuration Analyzer on the SE and HMC	392
11.5.13 Automated operations	393
11.5.14 Cryptographic support	393
11.5.15 Installation support for z/VM that uses the HMC	395
11.5.16 Dynamic Partition Manager	395
Chapter 12. Performance	397
12.1 IBM z14 ZR1 performance characteristics	398
12.2 LSPR workload suite	399
12.3 Fundamental components of workload performance	399
12.3.1 Instruction path length	399
12.3.2 Instruction complexity	400
12.3.3 Memory hierarchy and memory nest	400
12.4 Relative Nest Intensity	401
12.5 LSPR workload categories based on relative nest intensity	403
12.6 Relating production workloads to LSPR workloads	403
12.7 Workload performance variation	404
12.7.1 Main performance improvement drivers with z14 ZR1 servers	405
Appendix A. IBM Secure Service Container framework	409
A.1 What is IBM Secure Service Container?	410
A.1.1 SSC framework	410
A.2 SSC LPAR	410
A.3 Why Secure Service Container?	411
A.4 IBM Z and Secure Service Container	411
Appendix B. Channel options	415
Appendix C. Native Peripheral Component Interconnect Express	419
C.1 Design of native PCIe adapter management	420
C.1.1 Native PCIe adapter	420

C.1.2 Integrated firmware processor	420
C.1.3 Resource groups	421
C.1.4 Service and management tasks	421
C.2 Native PCIe adapters plugging rules	422
C.3 Native PCIe adapter definitions	422
C.3.1 FUNCTION identifier	424
C.3.2 Virtual function number	424
C.3.3 Physical network identifier	424
Appendix D. Shared Memory Communications	425
D.1 Overview	426
D.2 Shared Memory Communication over RDMA	426
D.2.1 RDMA technology overview	426
D.2.2 Shared Memory Communications over RDMA	427
D.2.3 Single Root I/O virtualization	429
D.2.4 Hardware	429
D.2.5 RoCE Express/Express2 adapter	430
12.7.2 RoCE Express/Express2 configuration example	431
D.2.6 Hardware configuration definitions	433
D.2.7 Software use of SMC-R	433
D.2.8 SMC-R support overview	434
D.2.9 SMC-R use cases for z/OS to z/OS	435
D.2.10 Enabling SMC-R support in z/OS Communications Server	437
D.3 Shared Memory Communications - Direct Memory Access	438
D.3.1 Concepts	438
D.3.2 Internal Shared Memory technology overview	439
D.3.3 SMC-D over Internal Shared Memory	439
D.3.4 Internal Shared Memory introduction	441
D.3.5 Virtual PCI Function (vPCI Adapter)	441
D.3.6 Planning considerations	444
D.3.7 Hardware configuration definitions	444
D.3.8 Sample IOCP FUNCTION statements	445
D.3.9 Software use of ISM	446
D.3.10 SMC-D over ISM prerequisites	447
D.3.11 Enabling SMC-D support in z/OS Communications Server	448
D.3.12 SMC-D support overview	448
Appendix E. IBM Dynamic Partition Manager	451
E.1 Introduction	452
E.2 Why use DPM?	452
E.3 DPM overview	453
E.4 Setting up the DPM environment	454
E.4.1 Defining partitions in DPM mode	457
E.4.2 Summary	460
Appendix F. IBM zEnterprise Data Compression Express	461
F.1 Overview	462
F.2 zEDC Express	462
F.3 Software support	463
F.3.1 IBM Z Batch Network Analyzer	464
Appendix G. 16U Reserved feature	465
G.1 General rules	466
G.2 Basic physical requirements	467

Related publications 473

IBM Redbooks 473

Other publications 473

Online resources 473

Help from IBM 474

Notices

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

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

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

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

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

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

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

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

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

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

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

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


COPYRIGHT LICENSE:

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

Trademarks

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

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

AIX®	IBM z Systems®	S/390®
Bluemix®	IBM z13®	System Storage®
CICS®	IBM z13s®	System z®
Cognos®	IBM z14™	System z10®
Db2®	Interconnect®	System z9®
Distributed Relational Database Architecture™	Language Environment®	VIA®
DS8000®	MVS™	VTAM®
ECKD™	OMEGAMON®	WebSphere®
FICON®	Parallel Sysplex®	z Systems®
FlashCopy®	Passport Advantage®	z/Architecture®
GDPS®	PowerPC®	z/OS®
Geographically Dispersed Parallel Sysplex™	PR/SM™	z/VM®
HiperSockets™	Processor Resource/Systems Manager™	z/VSE®
HyperSwap®	RACF®	z10™
IA®	Redbooks®	z13®
IBM®	Redbooks (logo)  ®	z13s®
IBM Cloud™	Resource Link®	z9®
IBM Z®	Resource Measurement Facility™	zEnterprise®
	RMF™	

The following terms are trademarks of other companies:

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

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

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

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

Preface

This IBM® Redbooks® publication describes the new member of the IBM Z® family, IBM z14™ Model ZR1 (Machine Type 3907). It includes information about the Z environment and how it helps integrate data and transactions more securely, and can infuse insight for faster and more accurate business decisions.

The z14 ZR1 is a state-of-the-art data and transaction system that delivers advanced capabilities, which are vital to any digital transformation. The z14 ZR1 is designed for enhanced modularity, in an industry standard footprint.

A data-centric infrastructure must always be available with a 99.999% or better availability, have flawless data integrity, and be secured from misuse. It also must be an integrated infrastructure that can support new applications. Finally, it must have integrated capabilities that can provide new mobile capabilities with real-time analytics that are delivered by a secure cloud infrastructure.

IBM z14 ZR1 servers are designed with improved scalability, performance, security, resiliency, availability, and virtualization. The superscalar design allows z14 ZR1 servers to deliver a record level of capacity over the previous IBM Z platforms. In its maximum configuration, z14 ZR1 is powered by up to 30 client characterizable microprocessors (cores) running at 4.5 GHz. This configuration can run more than 29,000 million instructions per second and up to 8 TB of client memory. The IBM z14 Model ZR1 is estimated to provide up to 54% more total system capacity than the IBM z13s® Model N20.

This Redbooks publication provides information about IBM z14 ZR1 and its functions, features, and associated software support. More information is offered in areas that are relevant to technical planning. It is intended for systems engineers, consultants, planners, and anyone who wants to understand the IBM Z servers functions and plan for their usage. It is *not* intended as an introduction to mainframes. Readers are expected to be generally familiar with IBM 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.

Octavian Lascu is a Senior IT Consultant for IBM Romania with over 25 years of experience. He specializes in designing, implementing, 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 materials for and taught workshops for technical audiences around the world for the past 18 years. He has written several IBM publications.

Hervey Kamga is an IBM Z Product Engineer with the EMEA I/O Connectivity Team in Montpellier, France. Before serving in his current role, he was a Support Engineer and Engineer On Site for 13 years with Sun Microsystems and Oracle in EMEA. Hervey's areas of expertise include Oracle Solaris (Operating System and hardware products), virtualization (VMware, virtualBox), Linux (Ubuntu), and IBM Z I/O features and protocols (IBM FICON® and OSA).

Frank Packheiser is a Senior zIT Specialist at the Field Technical Sales Support office in Germany. He has 26 years of experience in IBM Z platform. Frank has worked for 10 years in the IBM Education Center in Germany, developing and providing professional training. He also provides professional services to IBM Z and mainframe clients. In 2008 and 2009, Frank supported clients in Middle East/North Africa (MENA) as a zIT Architect. In addition to co-authoring several IBM Redbooks publications since 1999, he has been an official ITSO presenter at ITSO workshops for the last four years.

Martijn Raave is a certified IBM Z & LinuxONE Client Architect for IBM Netherlands. Over a period of 20 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 IBM z/OS®, IBM Parallel Sysplex®, and hardware. Later, he joined STG as a Client Technical Specialist, supporting several Dutch IBM System z® clients, IBM Business Partners, and IBM sales representatives. In 2016, the LinuxONE brand was added to his portfolio and he is now working as a Client Architect with a broad range of customers on different engagements. He's also a board member of Guide Share Europe (GSE) Netherlands.

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

Bill White is an IBM Redbooks Project Leader and Senior Networking and Connectivity Specialist at IBM Redbooks, Poughkeepsie Center.

Thanks to Tom Ambrosio and William Lamastro of the Poughkeepsie Competitive Center (PCC), and Robert Haimowitz from IBM Redbooks publications for their support and contributions to this publication.

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

Patty Driever
Dave Surman
Harry Yudenfriend
Dale Riedy
Ellen Carbarnes
John Eells
Diana Henderson
Barbara Sannerud
Anthony Saporito
Garth Godfrey
Darelle Gent
Parwez Hamid
Gary King
Jeff Kubala
Philip Sciuto
Rhonda Sundlof
Barbara Weiler
Luis Cruz
IBM Poughkeepsie

Christine Smith
Amanda Stapels
Bill Bitner
Leslie Geer III
Romney White
IBM Endicott

Monika Zimmermann
Carl Mayer
Walter Niklaus
Angel Nunes Mencias
IBM Germany

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 new IBM Z family member: IBM z14 Model ZR1

This chapter describes the basic concepts of IBM z14 Model ZR1 (Machine Type 3907) and includes the following topics:

- ▶ 1.1, “A digital transformation pillar” on page 2
- ▶ 1.2, “z14 ZR1 highlights” on page 3
- ▶ 1.3, “z14 ZR1 capacity and performance” on page 12
- ▶ 1.4, “z14 ZR1 virtualization” on page 13
- ▶ 1.5, “z14 ZR1 RAS” on page 16
- ▶ 1.6, “Hardware Management Consoles and Support Elements” on page 17
- ▶ 1.7, “Supported operating systems and compilers” on page 17

1.1 A digital transformation pillar

Businesses and organizations of every size are experiencing a time of exponential growth in data and transaction volumes that are driven by digital transformation. New dynamics in the market provide opportunities for businesses to grab market share and win. Leaders are being asked to add value by opening their enterprises to new ways of doing business. Organizations must give their clients and partners peace of mind that no matter what device is being used, data is protected.

In this challenging climate, businesses must manage, protect, store, and most importantly, use their data for gaining competitive advantage. This challenge is creating the need to apply intelligence and insight to data for building new services that are wrapped for a customized user experience.

In addition, businesses are experiencing increased pressure from internal and external sources to protect and govern data. They are required to reduce potential data breach risks and comply with complex regulatory mandates. These demands are changing the perspective around securely handling data.

A key to digital transformation is the ability to accelerate the innovation of new business. IT environments must have an architecture that scales for modern day cloud computing, which allows for rapid development and secure delivery of new services.

As your business technology needs evolve to compete in today's digital economy, IBM Z provides intelligent, robust, and comprehensive technology solutions. The IBM approach integrates Z hardware, software, and storage solutions to ensure that each component of the stack is tightly integrated and optimized. The IBM z14 Model ZR1 leads that approach by delivering the power and speed users demand, the security users and regulators require, and the operational efficiency that maximizes your bottom line.

One of the most impactful ways to protect data is by encrypting as much of your data and transactional pipeline as possible. Cryptography has always been in the DNA of IBM Z family. As the newest member of that family, the IBM z14 Model ZR1 continues that tradition with pervasive encryption to defend and protect your critical assets with unrivaled encryption and intelligent data monitoring without compromising transactional throughput or response times. Most importantly, this pervasive encryption requires no application changes. Pervasive encryption can dramatically simplify data protection and reduce the costs of regulatory compliance. By using simple policy controls, z14 ZR1 pervasive encryption streamlines data protection for mission critical IBM Db2® for z/OS, IBM IMS, and Virtual Storage Access Method (VSAM) datasets.

The Central Processor Assist for Cryptographic Function (CPACF), which is standard on every core, supports pervasive encryption and provides hardware acceleration for encryption operations. The Crypto Express6S gets a performance boost on the z14 ZR1. Combined, these enhancements perform encryption more efficiently on than on earlier IBM Z platforms.

The z14 ZR1 is designed specifically to meet the demand for new services and customer experiences, while securing the growing amounts of data and complying with increasingly intricate regulations. With up to 30 configurable cores, the z14 ZR1 has performance and scaling advantage over previous generations and 54% more capacity than the 20-way IBM z13s Model N20.

The FICON Express16S+ delivers an increase in I/O rates and in-link bandwidth. It also reduces single-stream latency, which provides the system the ability to absorb large applications and transaction spikes that are driven by unpredictable mobile devices.

Next-generation SMT in the z14 ZR1 delivers improved virtualization performance to benefit Linux. High-speed connectivity out to the data is critical in achieving exceptional levels of transaction throughput. The IBM zHyperLink Express introduces disk I/O technology for accessing the IBM DS8880 storage system with low latency, which enables shorter batch windows and a more resilient I/O infrastructure with predictable and repeatable I/O performance.

With up to 8 TB of customer memory, the z14 ZR1 can open opportunities, such as in-memory data marts and in-memory analytics, while giving you the necessary room to tune applications for optimal performance. By using the Vector Packed Decimal Facility that allows packed decimal operations to be performed in registers rather than memory, and by using new fast mathematical computations, compilers (such as Enterprise COBOL for z/OS, V6.2, Enterprise PL/I for z/OS, V5.2, and z/OS V2.3 XL C/C++), the COBOL optimizer, Automatic Binary Optimizer for z/OS, V1.3, and Java, are optimized on the z14 ZR1. These compilers and optimizer are designed to improve application performance, reduce CPU usage, and reduce operating costs. Java improvements and the use of crypto acceleration deliver more improvements in throughput per core, which gives a natural boost to z/OS Connect EE, IBM WebSphere® Liberty in IBM CICS®, Spark for z/OS, and IBM Java for Linux on IBM Z.

Linux on IBM Z, which is optimized for open source software, brings more value to the platform. Linux on IBM Z supports a wealth of new products that are familiar to application developers, such as Python, Scala, Spark, MongoDB, PostgreSQL, and MariaDB. Access to data that was unavailable without the need for Extract Transform and Load (ETL) allows for the development of intelligent transactions and intuitive business processes.

The advanced technology in the z14 ZR1 supports your digital transformation needs. The platform gives your IT teams the ability to:

- ▶ Create a strong and reliable cloud infrastructure to support rapid development and deployment of services
- ▶ Enable the adeptness to make consistently optimal business decisions and gain operational data insights so you get the most value from your IT investment
- ▶ Fully protect your data with encryption, while facilitating regulatory compliance
- ▶ Support open source software to aid developers in infusing value into services that ultimately improve user experiences

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

1.2 z14 ZR1 highlights

The z14 ZR1 is a highly scalable symmetric multiprocessor (SMP) system, and the architecture ensures continuity and upgradeability from its predecessor, the IBM z13s. The z14 ZR1 is housed in an industry-standard 19-inch rack that can be easily installed in any data center. It includes one model (ZR1) with four CPC drawer size features: Max4, Max12, Max24, and Max30. The z14 ZR1 can have up to four PCIe+ I/O drawers to support various I/O features for network, storage, and coupling connectivity.

This section reviews some of the following most important features and capabilities of the z14:

- ▶ Models and upgrade paths
- ▶ Rack and cabling

- ▶ CPC drawer
- ▶ PCIe+ I/O drawer
- ▶ I/O subsystem and I/O features

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

1.2.1 Models and upgrade paths

The z14 ZR1 has an assigned machine type (MT) of 3907, which uniquely identifies the central processor complex (CPC). All z14 ZR1 use five, six, seven, eight, or nine processor unit cores in for the processor unit single chip modules (up to four). Spare processor units, system assist processors (SAPs), and one integrated firmware processor (IFP) are integral to the system and are present in the z14 ZR1.

The number of characterizable processor units, SAPs, and spare processor units for the various models are listed in Table 1-1. Spare processor units are used to replace defective processor units and one spare processor unit (core) is always available on a z14 ZR1. In the rare event of a processor unit failure, the spare processor unit is immediately and transparently activated and assigned the characteristics of the failing processor unit.

Table 1-1 z14 ZR1 summary (machine type 3907)

Feature name	Feature code	Characterizable processor units	Standard SAPs	Spares	Integrated firmware processor
Max4	0636	1 - 4	2	1	1
Max12	0637	1 - 12	2	1	1
Max24	0638	1 - 24	2	1	1
Max30	0639	1 - 30	2	1	1

The z14 ZR1 offers 156 subcapacity levels for up to 6 CPs (subcapacity levels A - Z). One model for all Integrated Facility for Linux (IFL) or all Internal Coupling Facility (ICF) configurations also is available.

The upgrade paths for the z14 ZR1 are shown in Figure 1-1.

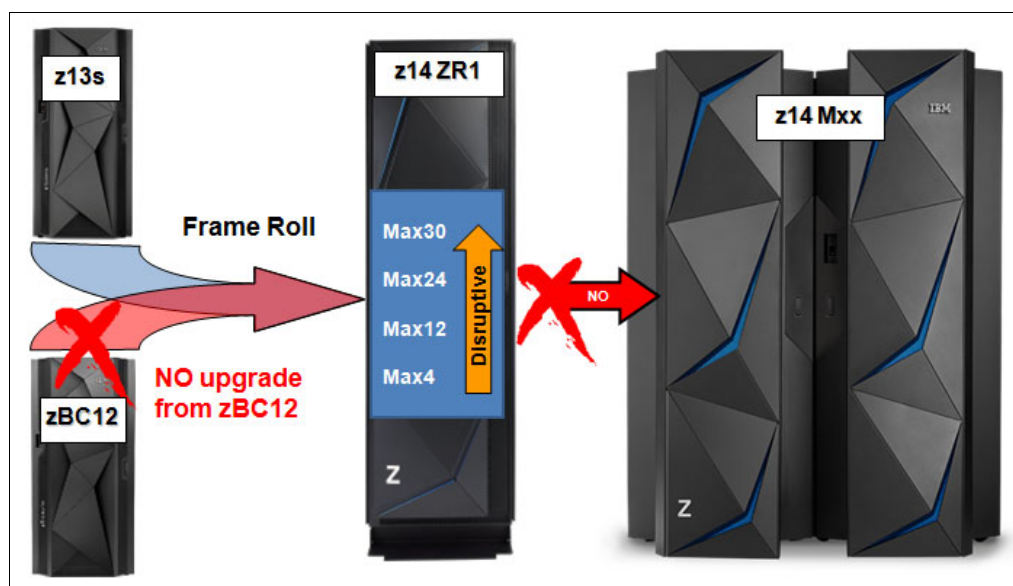


Figure 1-1 z14 upgrade paths

On the z14 ZR1, concurrent upgrades are available for CPs, IFLs, ICFs, Z Integrated Information Processors (zIIPs), and SAPs. However, concurrent processor unit upgrades require that more processor units are physically installed, but not activated previously.

1.2.2 Rack and cabling

The z14 ZR1 is designed in an industry standard 19-inch rack. The z14 ZR1 is a single rack, air-cooled system.

The rack forms the z14 ZR1 CPC and contains one CPC drawer. The number of PCIe+ I/O drawers can vary based on the number of I/O features. Up to four PCIe+ I/O drawers can be installed. PCIe I/O+ drawers can be added concurrently¹.

In addition, the z14 ZR1 (new builds and MES orders) offers top-exit options for the fiber optic and copper cables (used for I/O and power). These options (*Top Exit Power* and *Top Exit I/O Cabling*) give you more flexibility in planning where the system is installed. This flexibility potentially frees you from running cables under a raised floor, which increases air flow over the system.

The z14 ZR1 supports installation on raised floor and non-raised floor environments.

¹ The number of available PCIe fanout slots depends on the CPC drawer feature (Max4, Max12, Max24, and Max30).

The internal, front, and rear views of the z14 ZR1 system with the maximum four PCIe+ I/O drawers are shown in Figure 1-2.

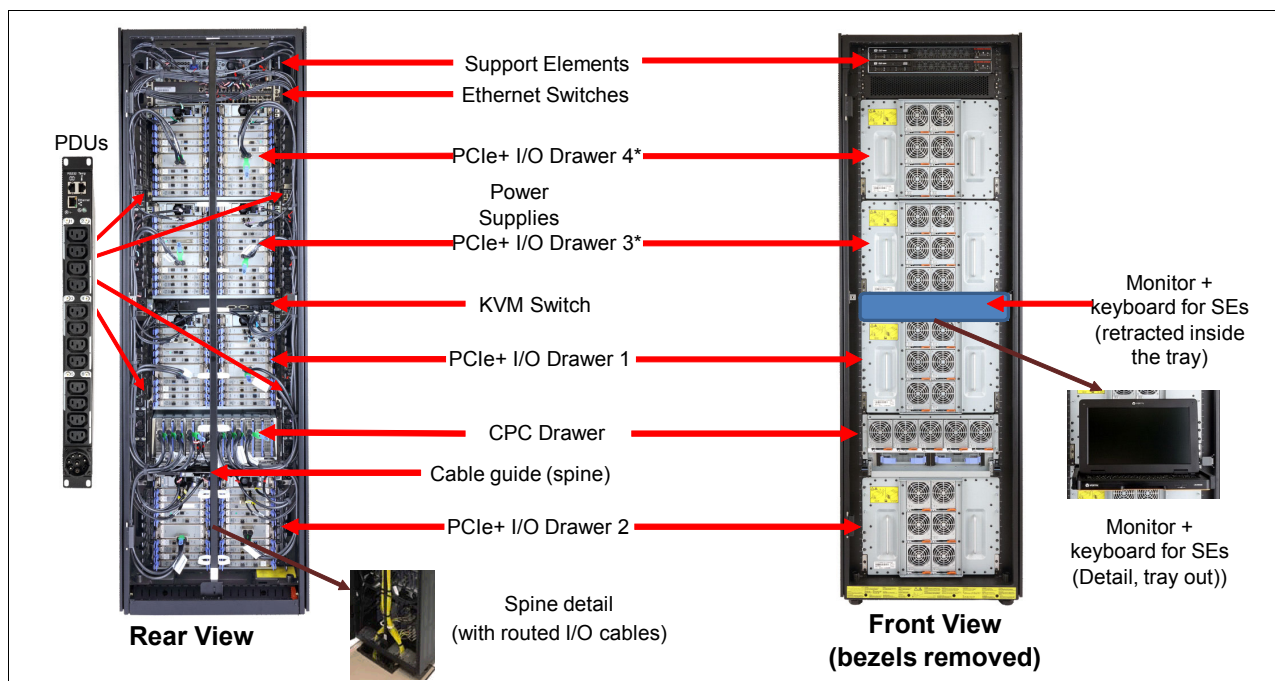


Figure 1-2 z14 ZR1 front and rear views: Configuration with four PCIe+ I/O drawers

Note: The asterisk that is shown in the Figure 1-2 highlights the fact that 3 and 4 cannot be used if FC 0617 (16U Reserved feature) is ordered.

1.2.3 CPC drawer

The z14 ZR1 is a single CPC drawer that contains the following elements:

- ▶ Single chip modules:
 - One to four PU single chip modules, each containing five, six, seven, eight, or nine processor unit cores (air-cooled). The processor unit cores run at 4.5 GHz each.
 - One System Controller single chip module, with a total of 672 MB L4 cache.
- ▶ Memory:
 - A minimum of 64 GB and a maximum of 8 TB of memory (excluding 64 GB HSA) is available for client use.
 - A total of 5, 10, 15, or 20 memory DIMMs are plugged in a CPC drawer.
 - The number of memory DIMMs that can be plugged into the CPC drawer depends on the CPC drawer feature.

▶ Fanouts:

The CPC drawer provides up to eight PCIe Gen3 fanout adapters to connect to the PCIe+ I/O drawers and Integrated Coupling Adapter Short Reach (ICA SR) coupling links. The number of fanouts that can be installed depends on the CPC drawer feature.

Each fanout includes on the following configurations:

- One-port PCIe 16 GBps I/O fanout, each supporting one domain in a 16-slot PCIe+ I/O drawer.
 - Two-port ICA SR PCIe fanout for coupling links (two links, 8 GBps each).
 - ▶ Two or four Power Supply Units (PSUs) that provide power to the CPC drawer, hot swappable that accessible form the rear.
- The loss of one PSU leaves enough power to satisfy the power requirements of the entire drawer. The PSUs can be concurrently maintained.
- ▶ Two Flexible Support Processors (FSPs) that provide redundant interfaces to the internal management network.
 - ▶ Two Oscillator Cards (OSCs) that provide clock synchronization to the CPC.

The logical diagram of a fully populated CPC drawer (for example, a Max24 or Max30 feature) is shown in Figure 1-3. The z14 ZR1 has 672 MB of L4 cache. The SC chip provides X-Bus connectivity for PUS SCMs in the adjacent logical cluster. Each PU chip has two PCIe bus interfaces (for PCIe fanouts). The GX Bus is not used for z14 ZR1.

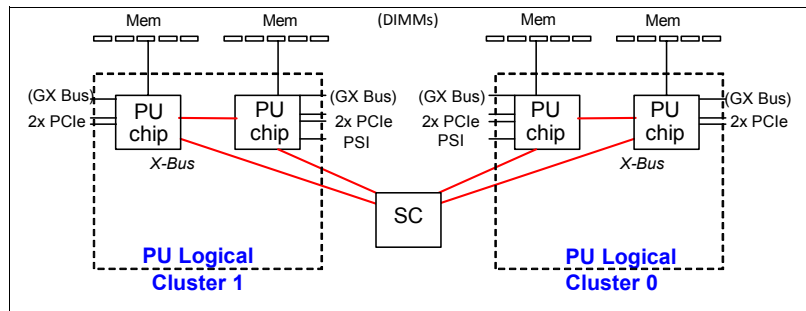


Figure 1-3 z14 ZR1 CPC drawer communication topology (fully populated drawer)

The design that is used to connect the processor unit and storage control allows the system to be operated and controlled by the IBM Processor Resource/Systems Manager™ (PR/SM™) facility as a memory-coherent symmetrical multiprocessor (SMP) system.

1.2.4 PCIe+ I/O drawer

The z14 ZR1 supports Generation 3 PCIe-based infrastructure by using PCIe+ I/O drawers (PCIe Gen3) for PCIe features (adapters). The number of supported PCIe+ I/O drawers is listed in Table 1-2.

Table 1-2 z14 ZR1 CPC drawer fanouts per feature

Feature name	PU SCMs	Max. PCIe fanouts	Max. PCIe+ I/O drawers ^a
Max4 (FC 0636)	1	2	1
Max12 (FC 0637)	2	4	2
Max24 (FC 0638)	4	8	4
Max30 (FC 0639)	4	8	4

a. If the 16U Reserved feature (FC 0617) is ordered, the maximum number of PCIe+ I/O drawers is two. For more information, see 2.2, “16U Reserved feature (FC 0617)” on page 26.

The PCIe I/O infrastructure consists of PCIe Gen3 fanouts in the CPC drawer that support 16 GBps connectivity to the PCIe+ I/O drawer.

Note: Ordering of I/O feature types determines the appropriate number of PCIe+ I/O drawers. Older PCIe I/O drawers are not supported on z14 ZR1.

The PCIe+ I/O drawer (see Figure 1-4) is a 19-inch single side drawer that is 8U high. I/O features are installed horizontally, with cooling air flow from front to rear. The drawer contains 16 slots and two switch cards. These features support two I/O domains that each contain eight features. Two PSUs provide redundant power, and six front-side fans provide redundant cooling to the PCIe+ I/O Drawer.

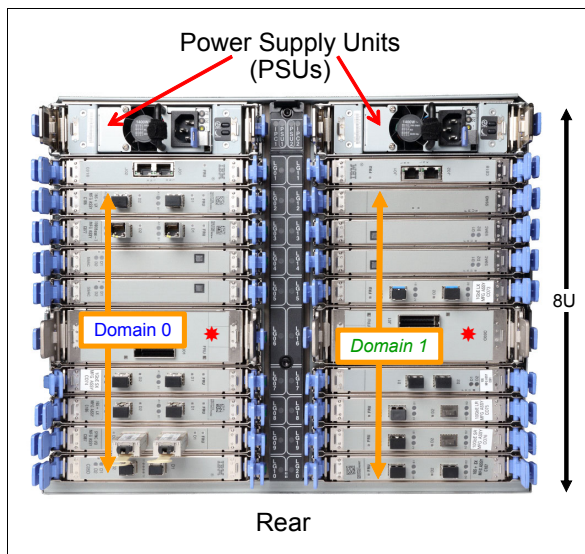


Figure 1-4 PCIe I/O drawer - rear view

A high-level view of the I/O system structure for the z14 ZR1 is shown in Figure 1-5.

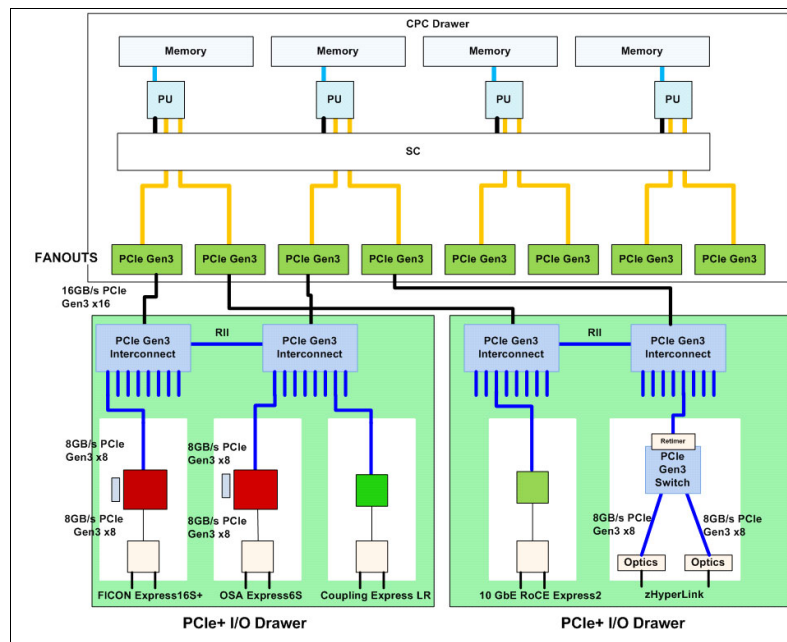


Figure 1-5 z14 ZR1 I/O system structure

The z14 ZR1 supports two fanout types (for fanout location, see Figure 1-6), which are at the front of the CPC drawer:

- ▶ Integrated Coupling Adapter Short Reach (ICA SR)
- ▶ PCIe Gen3 (for PCIe+ I/O drawer)

The PCIe Gen3 fanout has one port; ICA SR has two ports.

The PCIe connections to the PCIe+ I/O drawers type of internal I/O connectivity supports the PCIe I/O drawer.

For coupling link connectivity, the z14 ZR1 supports the following link types:

- ▶ ICA SR
- ▶ Coupling Express LR

All coupling adapters support parallel sysplex and Server Time Protocol (STP) connectivity.

CPC drawer I/O

The z14 ZR1 CPC drawer (see Figure 1-6) can include a combination of up to eight 1-port PCIe fanouts and 2-port ICA SR PCIe coupling fanouts (numbered LG01 - LG04 and LG07 - LG10).

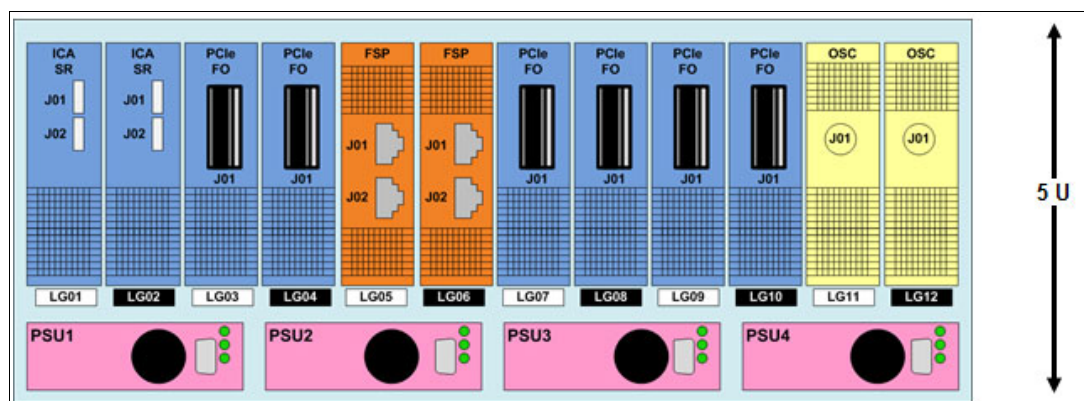


Figure 1-6 z14 ZR1 CPC drawer, front view

1.2.5 I/O subsystem and I/O features

The IBM z14 Model ZR1 offers a PCIe I/O infrastructure for its PCIe features that are installed in PCIe+ I/O drawers. Up to four PCIe+ I/O drawers per z14 ZR1 are supported, which provide space for up to 64 PCIe I/O features. Previous IBM Z server I/O infrastructures, such as PCIe I/O drawers or I/O drawers², are not supported on the z14 ZR1 and cannot be carried forward during an upgrade from a z13s.

Coupling connectivity migration: IBM z14 Model ZR1 (machine type 3907) does not support InfiniBand coupling infrastructure. The HCA3-O fanouts for 12x IFB (FC 0171) and HCA3-O LR fanouts for 1x IFB (FC 0170) are *not* supported.

You should migrate from HCA3-O to Integrated Coupling Adapter (ICA SR) and from HCA3-O LR to Coupling Express Long Reach (CE LR).

For coupling connectivity requiring a DWDM, ensure that your DWDM equipment is qualified to support the coupling links and timing only links.

The IBM z14 ZR1 uses PCIe fanout features for data communications between the CPC drawer and the I/O infrastructure, and for coupling (a fully configured z14 ZR1 supports eight PCIe fanout features). The multiple channel subsystem (CSS) architecture allows up to three CSSs, each with 256 channels.

Three *subchannel sets* are available per CSS, which allows access to many logical volumes. The third subchannel set allows extending the amount of addressable external storage for Parallel Access Volumes (PAVs), Peer-to-Peer Remote Copy (PPRC) secondary devices, and IBM FlashCopy® devices. The z14 ZR1 supports Initial Program Load (IPL) from subchannel set 1 (SS1), or subchannel set 2 (SS2), and subchannel set 0 (SS0). For more information, see “Initial program load from an alternative subchannel set” on page 166.

The system I/O buses use the Peripheral Component IBM Interconnect® Express (PCIe) technology, which are also used in coupling links.

² I/O drawers were introduced with the IBM z10™ BC and could be carried forward on earlier platform upgrades.

z14 ZR1 connectivity supports the following I/O or special purpose features:

► Storage connectivity:

– Fibre Channel connection (IBM FICON):

- FICON Express16S+ 10 KM long wavelength (LX) and short wavelength (SX)
- FICON Express16S 10 KM LX and SX (carry forward only)
- FICON Express8S 10 KM LX and SX (carry forward only)

For more information about FICON features, see “Storage connectivity” on page 136.

– IBM zHyperLink Express

For more information about zHyperLink Express, see “” on page 141.

► Network connectivity:

– Open Systems Adapter (OSA):

- OSA-Express7S 25GbE short reach (SR)
- OSA-Express6S 10GbE long reach (LR) and short reach (SR)
- OSA-Express6S GbE LX and SX
- OSA-Express6S 1000BASE-T Ethernet
- OSA-Express5S 10 GbE LR and SR (carry forward only)
- OSA-Express5S GbE LX and SX (carry forward only)
- OSA-Express5S 1000BASE-T Ethernet (carry forward only)
- OSA-Express4S 10 GbE LR and SR (carry forward only)
- OSA-Express4S GbE LX and SX (carry forward only)

For more information about OSA features, see “Network connectivity” on page 142.

– IBM HiperSockets™

For more information about the HiperSockets, see “HiperSockets” on page 151.

– Shared Memory Communication - Remote Direct Memory Access (SMC-R):

- 25GbE RoCE (RDMA over Converged Ethernet) Express2
- 10GbE RoCE Express2
- 10GbE RoCE Express (carry forward only)

– Shared Memory Communication - Direct Memory Access (SMC-D) through Internal Shared Memory (ISM)

For more information, see “25GbE RoCE Express2” on page 148.

► Coupling and Server Time Protocol connectivity:

- Internal Coupling (IC) links
- Integrated Coupling Adapter Short Reach (ICA SR)
- Coupling Express Long Reach (CE LR)

For more information about coupling and Server Time Protocol connectivity, see “Parallel Sysplex connectivity” on page 153.

► Cryptography:

- Crypto Express6S
- Crypto Express5S (carry forward only)
- Regional Crypto Enablement

For more information about the cryptographic features, see Chapter 6, “Cryptographic features” on page 173.

► IBM zEnterprise® Data Compression (zEDC) Express features, which are installed in the PCIe I/O drawers (new build and carry forward).

For more information about the zEDC feature, see Appendix F, “IBM zEnterprise Data Compression Express” on page 461.

1.3 z14 ZR1 capacity and performance

The z14 ZR1 provides increased processing capabilities and enhanced I/O infrastructure over its predecessor, the z13s. This capacity is achieved by increasing the performance of the individual PUs, increasing the number of PUs per system, redesigning the system cache, increasing the amount of memory, and using new I/O technologies.

The increased performance³ and the total system capacity that is available (with possible energy savings) allow consolidating diverse applications on a single platform with significant financial savings. The introduction of new technologies and an expanded and enhanced instruction set ensure that the z14 ZR1 is a high-performance, reliable, and rich-security platform.

The z14 ZR1 is designed to maximize the use of resources and allows you to integrate and consolidate applications and data across the enterprise IT infrastructure.

z14 ZR1 is offered as one model with four CPC drawer features. The z14 ZR1 Max4 feature (FC 0636) can have up to four customer configurable PUs, while the Max30 feature (FC 0639) can have up to 30 customer characterizable cores. z14 ZR1 Max30 is estimated to provide up to 54% more total system capacity than the z13s Model N20, with twice the amount of memory (up to 8 TB versus and lower power requirements. With enhanced SMT, co-processor features and SIMD, the performance of the z14 ZR1 delivers considerable improvement. Uniprocessor performance also was increased significantly. A z14 ZR1 uni-processor offers 10% average performance improvement over the z13s uni-processor.

The IFL and zIIP processor units on the z14 ZR1 server can be configured to run two simultaneous threads per clock cycle in a single processor (SMT). This feature increases the capacity of these processors with 25% in average over processors that are running single thread. SMT is also enabled by default on SAPs.

The z14 ZR1 provides 156 subcapacity settings, for up to six processors that are characterized as CPs (the same as z13s). The z14 ZR1 delivers scalability and granularity to meet the needs of small and medium-sized enterprises, while also satisfying the requirements for demanding, mission-critical transaction, and data processing requirements.

This comparison is based on the Large System Performance Reference (LSPR) mixed workload analysis. For more information about performance and workload variation on z14 ZR1 servers, see Chapter 12, “Performance” on page 397.

The z14 ZR1 continues to offer all the specialty engines that are available on z13s.

Workload variability

Consult the LSPR when considering performance on the z14 ZR1. 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) is available when an increased number of partitions and more PUs are available. For more information, see Chapter 12, “Performance” on page 397.

For more information about performance, [see the LSPR website](#).

³ Observed performance increases vary depending on the workload types.

Capacity on demand

Capacity on demand (CoD) enhancements enable clients to have more flexibility in managing and administering their temporary capacity requirements. The z14 ZR1 supports the same architectural approach for CoD offerings as the z13s (temporary or permanent). Within the z14 ZR1, 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 to handle many scenarios. Up to eight of these records can be installed on the server at any time. After the records are installed, the activation of the records can be done manually, or the z/OS Capacity Provisioning Manager can automatically start the activation when Workload Manager (WLM) policy thresholds are reached. Tokens are available that can be purchased for On/Off CoD before or after workload execution (pre- or post-paid).

LPAR capping

IBM Processor Resource/Systems Manager (IBM PR/SM) offers different options to limit the amount of capacity that is assigned to and used by an LPAR or a group of LPARs. By using the Hardware Management Console (HMC), a user can define an absolute or a relative capping value for LPARs that are running on the system.

1.4 z14 ZR1 virtualization

Virtualization is a key strength of Z platforms. It is embedded in the architecture and built into the hardware, firmware, and operating systems. For decades, Z platforms were designed based on the concept of partitioning resources (such as CPU, memory, storage, and network resources) so that each set of features can be used independently with its own operating environment.

This section describes built-in virtualization capabilities of z14 ZR1 supporting operating systems, hypervisors, and available virtual appliances.

1.4.1 PR/SM mode

PR/SM is Licensed Internal Code (LIC) that manages and virtualizes all of the installed and enabled system resources as a single large symmetric multiprocessor (SMP) system. This virtualization enables full sharing of the installed resources with high security and efficiency.

On z14 ZR1, the PR/SM supports configuring up to 40 LPARs, each of which includes logical processors, memory, and I/O resources. Resources of these LPARs are assigned from the installed CPC drawers and features. For more information about PR/SM functions, see 3.7, “Logical partitioning” on page 97.

LPAR configurations can be dynamically adjusted to optimize the virtual servers’ workloads. z14 ZR1 servers provide improvements to the PR/SM HiperDispatch function. HiperDispatch provides alignment of logical processors to physical processors that ultimately improves cache utilization and optimizes operating system work dispatching, which combined results in increased throughput. For more information, see “HiperDispatch” on page 69.

1.4.2 Dynamic Partition Manager mode

DPM is an administrative mode (front end panel driven interface to PR/SM) that is supported by the z14 ZR1. A system can be configured in DPM mode supports the following functions:

- ▶ Create, provision, and manage partitions (processor, memory, and adapters)
- ▶ Monitor and troubleshoot the environment

For more information, see Appendix E, “IBM Dynamic Partition Manager” on page 451.

1.4.3 LPAR types on z14 ZR1

The following LPAR types with corresponding operating systems and firmware appliances are supported:

- ▶ General:
 - z/OS
 - IBM z/VM®
 - IBM z/VSE®
 - z/TPF
 - Linux on Z (also used for the KVM Hypervisor)
- ▶ Coupling Facility: Coupling Facility Control Code (CFCC)
- ▶ LINUX only:
 - Linux on Z (also used for the KVM Hypervisor)
 - z/VM
- ▶ z/VM
- ▶ Secure Service Container:
 - VNA (z/VSE Network Appliance)
 - IBM High Security Business Network (HSBN)⁴

For DPM, the following LPAR modes are available:

- ▶ z/VM
- ▶ Secure Service Container
- ▶ Linux on Z (also used for the KVM Hypervisor)

IBM Z platforms also offer other virtual appliance-based solutions and support other the following hypervisors and containerization:

- ▶ IBM GDPS® Virtual Appliance
- ▶ The KVM hypervisor (included with supported Linux on Z distributions)
- ▶ Docker Enterprise Edition for Linux on IBM Systems⁵

⁴ IBM HSBN is a cloud service plan that is available on IBM Bluemix® for Blockchain.

⁵ For more information, see the [IBM to Deliver Docker Enterprise Edition for Linux on IBM Systems](#) topic of the IBM News releases website.

1.4.4 Coupling facility

Parallel sysplex is a synergy between hardware and software. The parallel sysplex is a highly advanced clustering solution that is designed to enable the aggregate capacity of multiple z/OS systems to be applied against common workloads. To use this technology, a special LIC is used, which is called CFCC. To activate the CFCC, a special logical partition must be defined. Only PUs that are characterized as CPs or Internal Coupling Facilities (ICFs) can be used for Coupling Facility (CF) partitions. For a production CF workload, it is recommended to use dedicated processors (ICFs or CPs).

1.4.5 z/VM-mode

The z14 ZR1 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 V6R4 or later, and allows z/VM to use a wider variety of specialty processors in a single LPAR, which increases flexibility and simplifying system management.

For example, in a z/VM-mode LPAR, z/VM can manage Linux on IBM Z guests that are running on IFL processors while also managing z/VSE and z/OS guests on CPs. It also allows z/OS to fully use zIIPs.

1.4.6 IBM Secure Service Container

IBM Secure Service Container (SSC) is an enabling technology for building virtual appliances (exploiters). It provides the base infrastructure to build and host virtual appliances on IBM Z.

SSC can be used to create isolated partitions for protecting data and applications automatically, which helps keep them safe from insider threats and external cyber criminals. SSC offers the following benefits:

- ▶ Streamlines the IBM Z Application experience so it is comparable to installing an application on a mobile device.
- ▶ Deploys an appliance in minutes, instead of days.
- ▶ Protect the workload from being accessed by a sysadmin or external attacker.

For more information, see Appendix A, “IBM Secure Service Container framework” on page 409.

IBM z/VSE Network Appliance

The z/VSE Network Appliance builds on the z/VSE Linux Fast Path (LFP) function and provides Internet Protocol network access without requiring a TCP/IP stack in z/VSE. Compared to a TCP/IP stack in z/VSE, this network appliance can support higher TCP/IP traffic throughput while reducing the processing resource consumption in z/VSE.

The z/VSE Network Appliance is an extension of the z/VSE - z/VM IP Assist (IBM VIA®) function provides network access for TCP/IP socket applications that run on z/VSE as a z/VM guest. With the new z/VSE Network Appliance, this function is available for z/VSE systems that are running in an LPAR. The z/VSE Network Appliance is provided as a downloadable package that can then be deployed with the SSC Installer and Loader.

The VIA function is available for z/VSE systems that run as z/VM guests. The z/VSE Network Appliance is available for z/VSE systems that run without z/VM in LPARs. Both functions provide network access for TCP/IP socket applications that use the LFP without the requirement of TCP/IP stack on the z/VSE system and installing Linux on IBM Z.

1.4.7 GDPS Virtual Appliance

The GDPS Virtual Appliance solution implements GDPS/PPRC Multiplatform Resilience for IBM Z (xDR). xDR coordinates near-continuous availability and a disaster recovery (DR) solution through the following features:

- ▶ Disk error detection
- ▶ Heartbeat for smoke tests
- ▶ Re-IPL in place
- ▶ Coordinated site takeover
- ▶ Coordinated IBM HyperSwap®
- ▶ Single point of control

1.5 z14 ZR1 RAS

System reliability, availability, and serviceability (RAS) is an area of continuous IBM focus and a defining IBM Z platform characteristic. The RAS objective is to reduce, or eliminate if possible, all sources of planned and unplanned outages while providing adequate service information if an issue occurs. Adequate service information is required to determine the cause of an issue without the need to reproduce the context of an event.

IBM Z servers are designed to enable highest availability and lowest downtime. These facts are recognized by various IT analysts, such as ITIC⁶ and IDC⁷. Comprehensive, multi-layered strategy includes the following features:

- ▶ Error Prevention
- ▶ Error Detection and Correction
- ▶ Error Recovery

With a properly configured z14 ZR1, further reduction of outages can be attained through First Failure Data Capture (FFDC), which is designed to reduce service times and avoid subsequent errors, and improve nondisruptive replace, repair, and upgrade functions for memory, drawers, and I/O adapters. In addition, the z14 ZR1 extended nondisruptive capability to download and install LIC updates.

z14 ZR1 RAS features provide unique high-availability and nondisruptive operational capabilities that differentiate the Z platform in the marketplace. z14 RAS enhancements are made on many components of the CPC (processor chip, memory subsystem, I/O, and service) in areas, such as error checking, error protection, failure handling, error checking, faster repair capabilities, sparing, and cooling.

The z14 ZR1 processor builds upon the RAS of the z13s with the following RAS improvements:

- ▶ The level 3 cache added powerful symbol ECC, which makes it resistant to more failures (the z13s hardened the level 4 cache and the main memory was hardened with RAIM and ECC before that addition).
- ▶ The main memory added preemptive DRAM marking to isolate and recover failures faster.
- ▶ Small array error handling was improved in the processor cores.
- ▶ Error thresholding was added to the processor core to isolate “sick but not dead” failure scenarios.

⁶ For more information, see [ITIC Global Server Hardware, Server OS Reliability Report](#).

⁷ For more information, see [Quantifying the Business Value of IBM Z](#).

- ▶ The number of Resource Groups for supporting native PCIe features increased to four from two to reduce the effect of firmware updates and failures.
- ▶ OSA-Express6S added TCP checksum on large send offload.

For more information, see Chapter 9, “Reliability, availability, and serviceability” on page 327.

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

1.6 Hardware Management Consoles and Support Elements

The HMCs and SEs are appliances that together provide platform management for IBM Z. The HMC is a workstation that is designed to provide a single point of control for managing local or remote hardware elements.

HMC is offered as a Tower (FC 0082) and a Rack Mount (FC 0083) feature. Rack Mount HMC can be placed in a customer-supplied 19-inch rack and occupies 1U rack space. z14 includes driver level 32 and HMC application Version 2.14.0.

For more information, see Chapter 11, “Hardware Management Console and Support Elements” on page 357.

1.7 Supported operating systems and compilers

The z14 ZR1 is supported by a large set of software products and programs, including independent software vendor (ISV) applications. (This section lists only the supported operating systems and compilers.) Use of various features might require the latest releases. For more information, see Chapter 7, “Operating system support” on page 209.

1.7.1 Operating systems summary

The current and minimum operating system levels that are required to support the z14 are listed in Table 1-3 on page 18. Operating system levels that are no longer in service are not covered in this publication. These older levels can support certain features.

Table 1-3 z14 - supported operating systems

Operating system	End of service	Notes
z/OS V2R3	September 2022 ^a	See: z/OS fix category (FIXCAT) IBM.Device.Server.z14ZR1-3907, z/VM, z/VSE, and z/TPF subsets of the 3907DEVICE Preventive Service Planning (PSP) buckets before installing the z14 Model ZR1.
z/OS V2R2	September 2020 ^a	
z/OS V2R1 ^b	September 2018	
z/OS V1R13 ^b	September 2016	
z/VM V7R1	Not announced	
z/VM V6R4	Not announced	
z/VSE V6R2	Not announced	
z/VSE V6R1	June 2019 ^a	
z/VSE V5R2	October 2018 ^a	
z/TPF V1R1	Not announced	
Linux on Z	Support information is available for SUSE ^c , Red Hat ^d and Ubuntu ^e	
The KVM hypervisor ^f	Offered with the following Linux distributions SLES-12 SP2 or higher, Ubuntu 16.04 LTS or higher, and Red Hat 7.5 or higher.	

- a. Planned date. All statements regarding IBM plans, directions, and intent are subject to change or withdrawal without notice. Any reliance on these Statements of Direction is at the relying party's sole risk and will not create liability or obligation for IBM.
- b. Compatibility only. The IBM Software Support Services for z/OS V1.13, offered as of October 1, 2016, provides the ability for customers to purchase extended defect support service for z/OS V1.13.
- c. For more information, see <https://www.suse.com/support/>.
- d. For more information, see <https://www.redhat.com/security/updates/errata/>.
- e. For more information, see <https://www.ubuntu.com/download/server/linuxone>.
- f. For more information about minimal and recommended distribution levels, see the distributors' websites.

For more information about supported Linux on Z distribution levels, see the [Tested platforms for Linux page](#) of the IBM Z website.

1.7.2 IBM Z compilers

IBM compilers and programming tools for Z that can be used with the z14 ZR1 include the following examples:

- ▶ Enterprise COBOL for z/OS
- ▶ Enterprise PL/I for z/OS
- ▶ Automatic Binary Optimizer
- ▶ z/OS XL C/C++
- ▶ XL C/C++ for Linux on IBM z Systems®⁸

The compilers increase the return on your investment in IBM Z hardware by maximizing application performance by using the compilers' advanced optimization technology for IBM z/Architecture®. Through their support of web services, XML, and Java, they allow for the modernization of assets in web-based applications. They also support the latest IBM middleware products (CICS, Db2, and IMS), which allows applications to use their latest capabilities.

To fully use the capabilities of z14 ZR1 servers, you must compile it by using the minimum level of each compiler. To obtain the best performance, you must specify an architecture level of 12 by using the **ARCH(12)** option.

For more information see “7.5.4, “z/OS XL C/C++ considerations” on page 274.

⁸ For more information, see <https://www.ibm.com/developerworks/downloads/r/xlcppluslinuxonz/>.



Central processor complex hardware components

This chapter introduces the z14 ZR1 central processor complex (CPC) hardware components. It also describes the significant features and functions with their characteristics and options.

This chapter describes the z14 ZR1 hardware building blocks and how these components interconnect.

This chapter includes the following topics:

- ▶ 2.1, “System overview: Frame and drawers” on page 22
- ▶ 2.2, “16U Reserved feature (FC 0617)” on page 26
- ▶ 2.3, “CPC drawer” on page 27
- ▶ 2.4, “Single chip modules” on page 33
- ▶ 2.5, “Memory” on page 38
- ▶ 2.6, “Reliability, availability, and serviceability” on page 46
- ▶ 2.7, “Connectivity” on page 49
- ▶ 2.8, “Power and cooling” on page 58
- ▶ 2.9, “Summary” on page 60

2.1 System overview: Frame and drawers

The IBM z14 Model ZR1 system is designed in an industry standard 19-inch form factor rack (frame) that can be easily installed in any data center. The design uses power distribution unit (PDU)-based power along with redundant power, cooling, and power cords.

Redesigned CPC drawer I/O infrastructure also lowers power costs, reduces the footprint, and allows installation in virtually any data center. The z14 ZR1 server is rated at ASHRAE class A3¹ data center operating environment.

The system is designed with one CPC Drawer and up to four new Peripheral Component Interconnect Express Generation 3 (PCIe Gen3) I/O drawers (named PCIe+ I/O drawers). The components that are included in the rack are described in the following sections.

The z14 ZR1 server is an air-cooled system (see Figure 2-1) without front covers and doors.

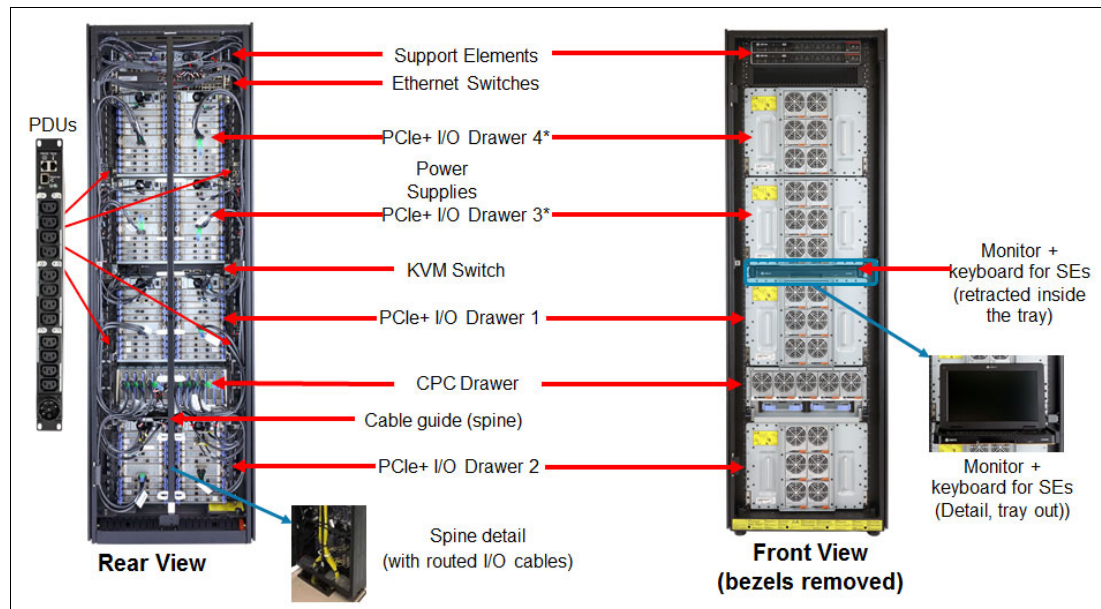


Figure 2-1 z14 ZR1 server

The single rack includes the following major components (from top to bottom of the rack):

- ▶ Two redundant Support Element (SE) 1U servers that are installed at the top of the rack.
- ▶ Two redundant Ethernet switches that provide external and internal communications to manage the server.
- ▶ Up to four new PCIe+ I/O drawers. The number of PCIe+ I/O drawers depends on the I/O configuration of the z14 ZR1 server.
- ▶ A pullout Keyboard, Mouse, Monitor (KMM) tray (new design for z14 ZR1), accessible at the front of the rack. The KMM devices are connected to an internal Keyboard, Video, Mouse (KVM) switch that is mounted in the rear of the rack, which is used to alternate the console between the two Support Elements.
- ▶ A new CPC drawer, which houses the PU and SC Single Chip Modules (SCMs), Memory, PCIe fanouts for I/O drawer connectivity, ICA SR adapters, and the necessary power elements and cooling fans.

¹ For more information, see Chapter 2, "Environmental specifications" in the *IBM 3907 Installation Manual for Physical Planning*, GC28-6974-00.

- ▶ Depending on the configuration, two or four intelligent Power Distribution Units (PDUs) are mounted vertically on each side at the rear of the rack. All of the internal components receive their power from these intelligent PDUs. The PDUs are cabled for redundancy.
- ▶ A newly designed vertical cable management guide (“spine”) can assist with proper cable management for fiber, copper, and coupling cables. The spine is shipped with configurations that contain three or four PCIe+ I/O drawers or with the 16U Reserved feature (FC 0617). All external cabling to the system (from top or bottom) can use the spine to minimize interference with the PDUs mounted on the sides of the rack.
- ▶ The z14 ZR1 server (as its predecessor, the z13s) has the option of ordering the infrastructure to support the top and bottom exit of fiber optic and copper cables.

The rack with the spine mounted is shown in Figure 2-2. It includes removable hooks that can be placed in appropriate slots throughout the length of the spine.

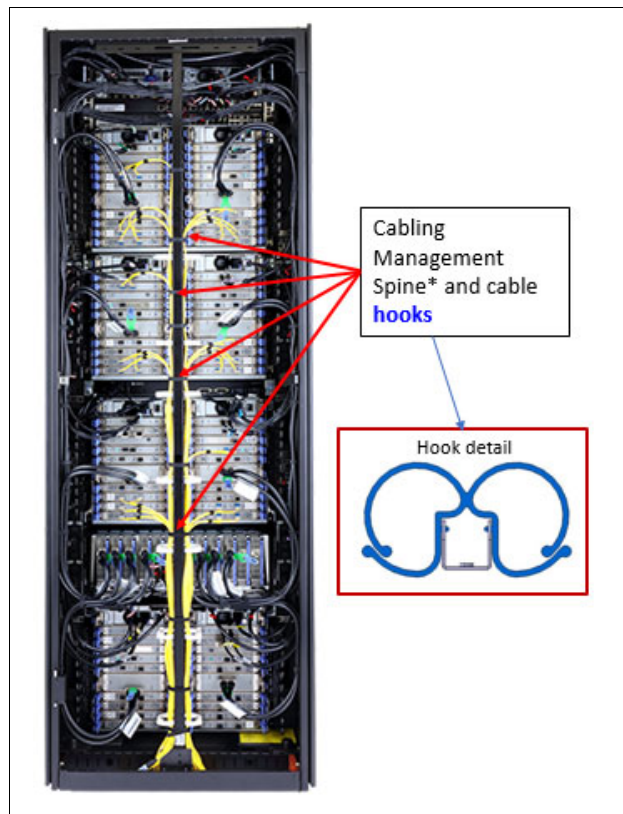


Figure 2-2 Cable management spine

2.1.1 z14 ZR1 configurations

All system components are designed and integrated in an IBM 19-inch frame (rack)². The possible configurations of the system (rear view) are shown in Figure 2-3 on page 24. Consider the following points:

- ▶ The z14 ZR1 system is built in a 42 EIA units rack (A-frame). The base rack is 40 EIA units high with a 2U removable top. FC 9975 is available if a height reduction is necessary.

² The components cannot be installed in a customer supplied rack. The term “frame” can also be used interchangeably with “rack”.

- ▶ PCIe+ I/O drawers are provided as required by the number of I/O adapters ordered. The PCIe+ I/O drawers (1, 2, 3, and 4) and are installed in order in the following EIA locations: A14B, A01B, A23B, and A31B.
- ▶ The five possible configurations without FC 0617³ with the view from the rear of the system (from left to right in Figure 2-3):
 - The first configuration can be a coupling facility with coupling fanouts that are installed in the CPC Drawer only.
 - The second configuration includes a single PCIe+ I/O drawer that is installed at EIA A14B and along with Configuration 1, requires only the upper pair of PDUs, which are mounted vertically at the rear of the system in columns N and Z.
 - The third configuration includes the second PCIe+ I/O drawer that is installed at EIA A01B and now requires the extra lower pair of PDUs installed to provide necessary power.
 - The fourth configuration includes the third PCIe+ I/O drawer that is installed at EIA A23B.
 - The fifth configuration includes the fourth PCIe+ I/O drawer that is installed at EIA A31B (maximum I/O configuration).

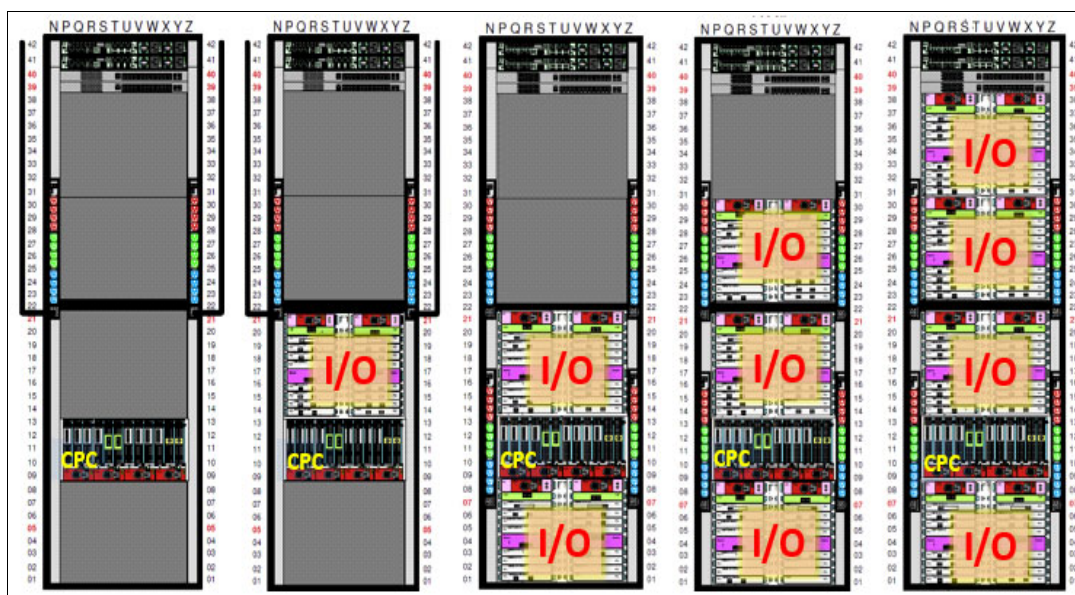


Figure 2-3 ZR1 Rack configurations: Rear view (without FC 0617)

2.1.2 PCIe+ I/O drawer

A new PCIe+ I/O drawer is introduced with the z14 ZR1 (Machine Type 3907). The new drawer is shown in Figure 2-4 on page 25. Each PCIe+ I/O drawer (up to four) has 16 PCIe slots each to support the PCIe I/O infrastructure with a bandwidth of 16 GBps and includes the following features:

- ▶ Two I/O domains (0 and 1), each capable of hosting eight PCIe adapters, for a total of 16 I/O adapters.
- ▶ Two PCIe switch cards provide connectivity to the PCIe fanouts that are installed in the CPC drawer. Each I/O domain features the PCIe slots that are allocated over two PCIe support partitions to manage the native PCIe adapters.

³ See 2.2, “16U Reserved feature (FC 0617)” on page 26.

- ▶ Two Flexible Support Processor (FSP) cards that are used to control the drawer.
- ▶ Two Power Supply Units (PSUs) in a redundant configuration.
- ▶ Six hot-swappable cooling fan modules at the front of the drawer.

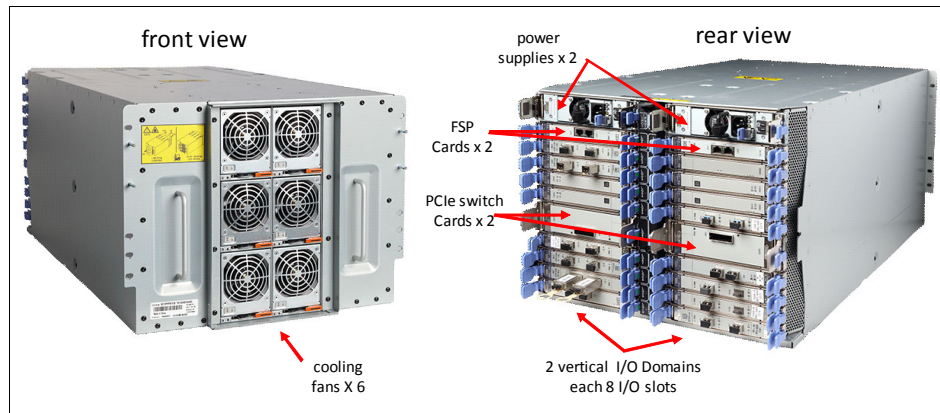


Figure 2-4 PCIe+ I/O drawer front and rear view

PCIe I/O infrastructure

Terminology: Throughout this chapter, the terms *adapter* and *card* are interchangeable and refer to a feature that is installed in a PCIe+ I/O drawer.

The PCIe I/O infrastructure uses the PCIe fanouts that are installed in the processor (CPC) drawer to connect to the PCIe+ I/O drawer. The PCIe adapters include the following features:

- ▶ FICON Express16S+ (two port card), long wavelength (LX) or short wavelength (SX), which contains two physical channel IDs (PCHIDs)
- ▶ FICON Express16S (two port card), long wavelength (LX) or short wavelength (SX), which contain two PCHIDs (only for carry-forward MES)
- ▶ FICON Express8S (two port card), long wavelength (LX) or short wavelength (SX), which contain two PCHIDs (only for carry-forward MES)
- ▶ Open Systems Adapter (OSA)-Express7S 25GbE Short Reach (SR) - New feature
- ▶ Open System Adapter (OSA)-Express6S:
 - OSA-Express6S 10 Gb Ethernet (single port card, Long Reach (LR) or Short Reach (SR), one PCHID)
 - OSA-Express6S Gb Ethernet (two port card, LX or SX, one PCHID)
 - OSA-Express6S 1000BASE-T Ethernet (two port card, RJ-45, one PCHID)
- ▶ Open System Adapter (OSA)-Express5S and 4S features (for carry-forward MES only):
 - OSA-Express5S and 4S 10 Gb Ethernet (one port card, LR or SR, and one PCHID)
 - OSA-Express5S and 4S Gb Ethernet (two port card, LX or SX, and one PCHID)
 - OSA-Express5S 1000BASE-T Ethernet (two port card, RJ-45, and one PCHID)
- ▶ Crypto Express6S (new build) and Crypto Express5s (only for carry-forward MES). Each feature holds one PCIe cryptographic adapter. Each adapter can be configured as:
 - Secure IBM Common Cryptographic Architecture (CCA) coprocessor
 - Secure IBM Enterprise Public Key Cryptography Standards (PKCS) #11 (EP11) coprocessor
 - Accelerator

- ▶ zHyperLink Express is a two-port card that is directly connected to storage subsystem controller (supported DS8880 series) that is designed to increase the scalability of IBM Z transaction processing through increased throughput and lower I/O latency. This feature is installed in the PCIe+ I/O drawer.
- ▶ Coupling Express LR (CE LR) is a two-port card that is used for long-distance coupling connectivity and uses CL5 coupling channel type. The card uses 10GbE RoCE technology and is designed to drive distances up to 10 km (6.21 miles) and support a link data rate of 10 Gigabits per second (Gbps). This feature is installed in the PCIe+ I/O drawer. Each port supports up to four CHPIDs per port and use point-to-point connectivity only.
- ▶ New feature - 25GbE RoCE Express2 - Remote Direct Memory Access (RDMA) over Converged Ethernet. The next generation RoCE with improved virtualization scale, better performance, and RAS. It is a two-port card and supports up to 31 virtual functions per port and 62 VFs per PCHID (feature)
- ▶ 10GbE RoCE Express2 - Remote Direct Memory Access (RDMA) over Converged Ethernet. The next generation RoCE with improved virtualization scale, better performance, and RAS. It is a two-port card and supports up to 31 virtual functions per port and 62 VFs per PCHID (feature).
- ▶ 10GbE RoCE Express - Remote Direct Memory Access (RDMA) over Converged Ethernet (only for a carry-forward MES). It is a two-port card and supports up to 31 virtual functions per adapter.
- ▶ zEnterprise Data Compression (zEDC) Express. The zEnterprise Data Compression Express card occupies one I/O slot, but it does not have a CHPID type. Up to 15 partitions can share the feature concurrently.

2.2 16U Reserved feature (FC 0617)

The z14 ZR1 can be ordered with FC 0617 (16U Reserved), which provides a Fit-for-Purpose solution in a single footprint by allowing z14 ZR1 configurations with two or less PCIe+ I/O drawers the use of open space in the rack for non-Z hardware⁴.

The 16U Reserved feature allows clients to create all-in-one solutions to run their entire business or independent application or cloud solutions. Consider the following points:

- ▶ FC 0617 is “16U Reserved” space in the rack for non-Z equipment (components) to be installed in the Z rack. It allows for clients to integrate other hardware into the single 19-inch rack, which reduces data center footprint requirements.
- ▶ Reserved space can be especially useful for space-constrained data centers or to keep rack-mounted HMC and other rack-mounted equipment together with the Z server.

⁴ Examples of hardware that can be installed in the unused space are: IBM rack-mounted HMC (1U), IBM rack-mounted TKE (1U), IBM V7000 storage, IBM V9000 storage, SAN switches, Network switches or other equipment that is designed to fit in a 19-inch rack. Non-IBM equipment can also be installed in the space reserved.

EIA units 22 - 38 space is designated for the 16U Reserved (FC 0617), as shown in Figure 2-5.

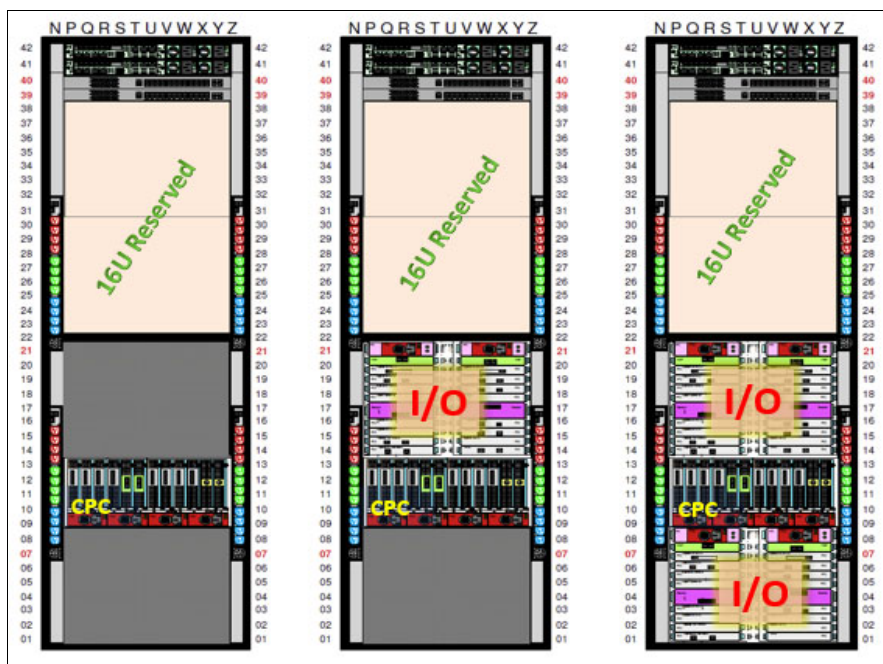


Figure 2-5 Possible configurations with the 16U Reserved feature

For more information about the 16U Reserved feature, see Appendix G, “16U Reserved feature” on page 465.

2.3 CPC drawer

The z14 ZR1 server continues the design of z13s by packaging processors in drawers. Unlike z13s, the processor drawer was designed to fit in the 19-inch rack. The single z14 ZR1 Central Processor Complex (CPC) drawer includes the following features:

- ▶ Up to four Processor Unit (PU) Single Chip Modules (SCMs)
- ▶ One System Control (SC) SCM
- ▶ Memory DIMMs (up to four banks or five DIMMs each)
- ▶ Connectors to support PCIe+ I/O drawers (through PCIe fanout hubs)
- ▶ Coupling links to other CPCs by using the Integrated Coupling Adapter Short Reach fanout

The z14 ZR1 CPC drawer and its components are shown in Figure 2-6.

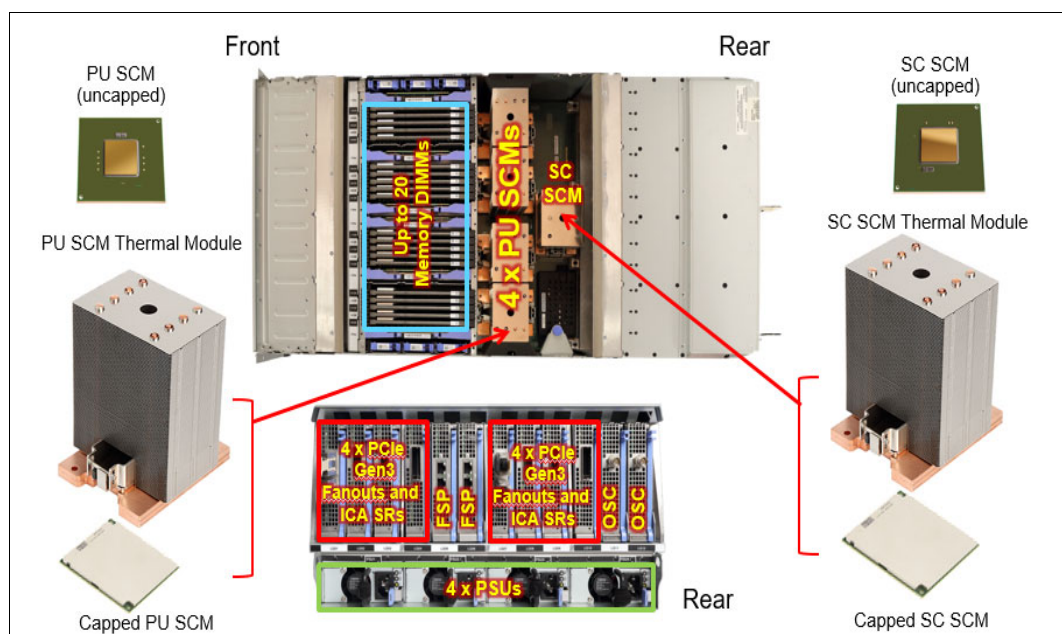


Figure 2-6 z14 ZR1 CPC drawer layout details

The z14 ZR1 CPC drawer size is determined by the number of PU SCMs that is installed and is feature-driven (four CPC drawer size features).

Depending on the CPC drawer feature, the CPC drawer contains the following components:

- Four CPC Drawer size configurations (CPC drawer has always ONE SC SCM):
 - FC 0636 = 1 PU SCM (maximum 4 characterizable PUs)
 - FC 0637 = 2 PU SCMs (maximum 12 characterizable PUs)
 - FC 0638 = 4 PU SCMs (maximum 24 characterizable PUs)
 - FC 0639 = 4 PU SCMs (maximum 30 characterizable PUs)

The following SCMs are available:

- PU SCM uses 14nm SOI technology, 17 layers of metal, 14.4 miles of wire, core running at 4.5GHz: 10 PUs/SCM (with 5, 6, 7, 8, 9 active cores).
- System Controller (SC) SCM, 17 layers of metal, 13.8 miles of wire, 672 MB L4 cache.
- Each PU SCM has one memory controller that drives five DDR4 dual inline memory modules (DIMMs), for a maximum of 20 DIMMs per drawer.

DIMMs are plugged into 5, 10, 15, or 20 DIMM groups, which provides 160 - 10240 GB of physical memory (RAIM protected) that results in 128 - 8192 GB of addressable memory (64 - 8128 customer usable memory).
- Up to eight PCIe Generation 3 I/O slots that can host two, four, or eight PCIe Gen3 x16 fanouts (16 GBps bandwidth) populated by:
 - PCIe Gen3 I/O fanout for PCIe+ I/O drawer (always ordered and used in pairs for availability)
 - ICA SR PCIe fanout for coupling
- Management elements:
 - Two flexible service processor (FSP) cards for system control (N+1 redundancy).
 - Two Oscillator cards to provide system clocking (N+1 redundancy)

- CPC drawer power infrastructure consists of:
 - Two or four Power Supply Units (PSUs) that provide power to the CPC drawer. The loss of one power supply leaves enough power to satisfy the drawer's power requirements (N+1 redundancy). The power supplies can be concurrently removed and replaced (one at a time).
 - Three to six Voltage Regulator Modules that plug next to the memory DIMMs.
 - Two Power Control cards to control the five CPC fans at the front of the CPC drawer.

The front view of the CPC drawer, which includes the cooling fans and power control cards, is shown in Figure 2-7.

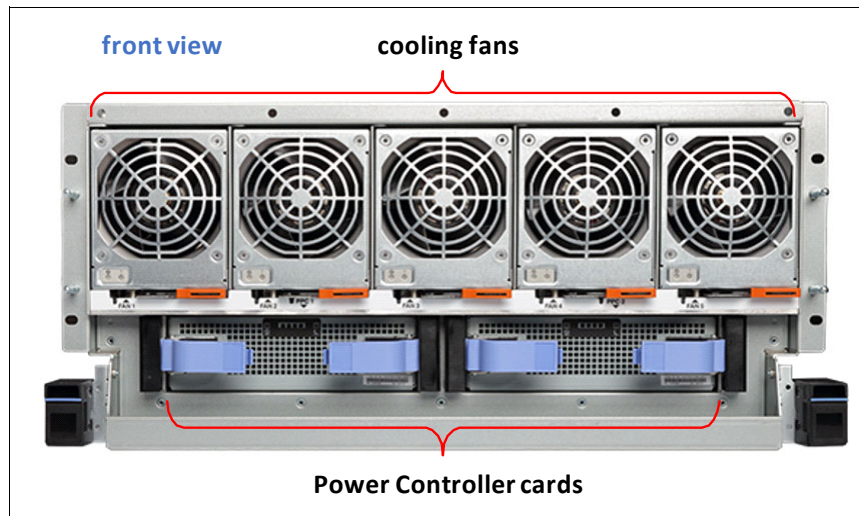


Figure 2-7 Front view of the CPC drawer

The rear view of a fully populated CPC Drawer is shown in Figure 2-8. PCIe I/O fanouts are plugged in specific slots for best performance and availability. Redundant FSP cards (two) and Oscillator cards (two) are always installed.

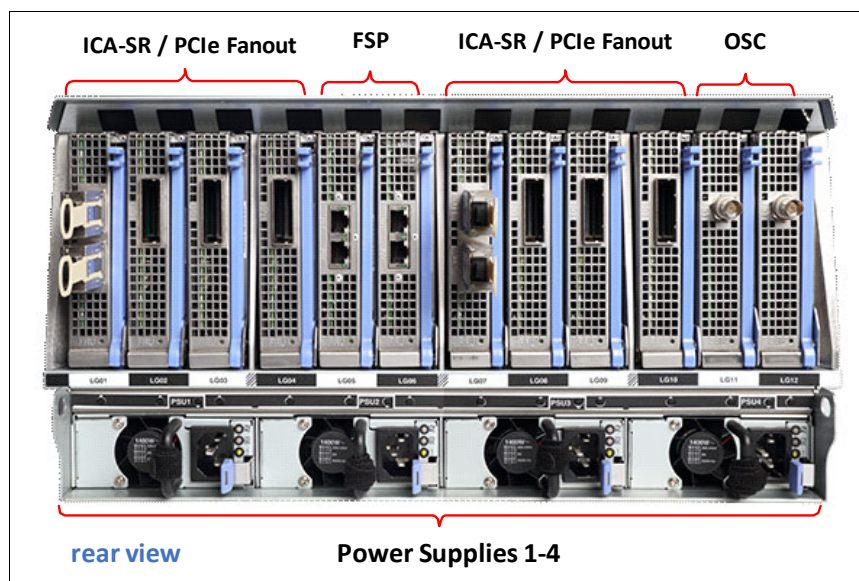


Figure 2-8 Rear view of a fully populated CPC Drawer

A top view of the CPC drawer is shown in Figure 2-9.

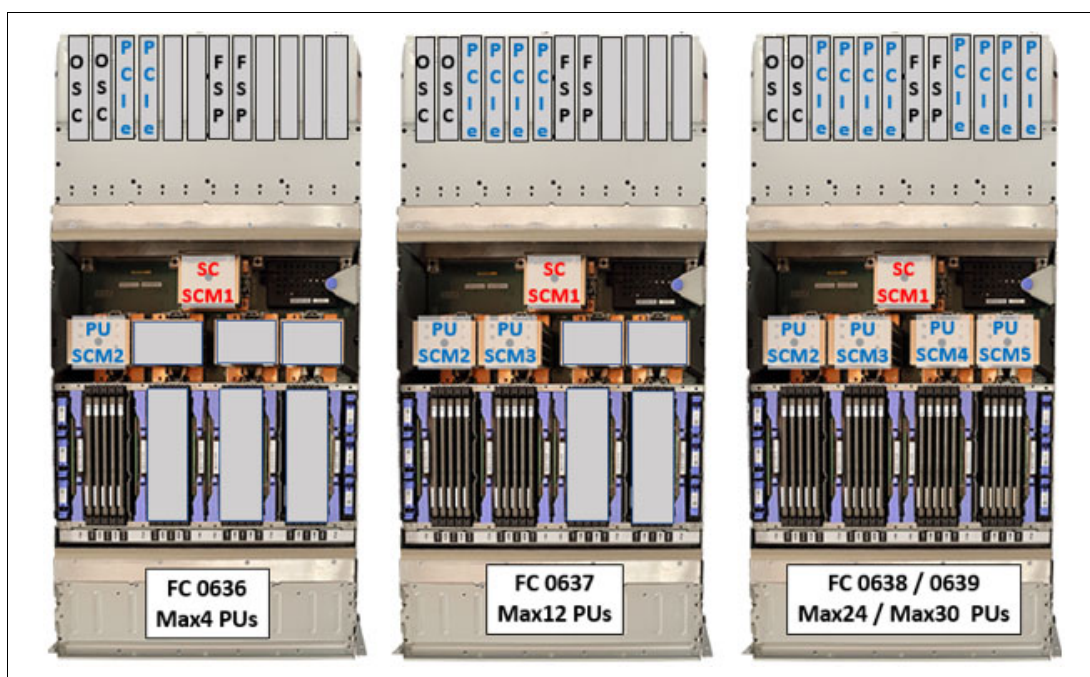


Figure 2-9 CPC drawer top view and feature code comparison

The following PU SCM feature codes relate to the resources that are enabled:

- ▶ FC 0636: One PU SCM, five DIMMs, two PCIe slots enabled
This configuration supports one PCIe+ I/O drawer or two ICA SR adapters.
- ▶ FC 0637: Two PU SCMs, up to 10 DIMMs, four PCIe slots enabled
This configuration supports up to two PCIe+ I/O drawers or four ICA SR adapters or one PCIe+ I/O Drawer and two ICA SR Adapters.
- ▶ FC 0638 and FC 0639: Four PU SCMs, up to 20 DIMMs, eight PCIe slots enabled
This configuration supports up to four PCIe+ I/O drawers or eight ICA SR adapters or a combination not to exceed eight PCIe slots in the CPC drawer.

Memory is connected to the SCMs through memory control units (MCUs). Up to four MCUs are available in a drawer (one per PU SCM) to provide the interface to the controller on memory DIMM. A memory control unit drives five DIMM slots.

The CPC drawer logical diagram is shown in Figure 2-10 on page 31.

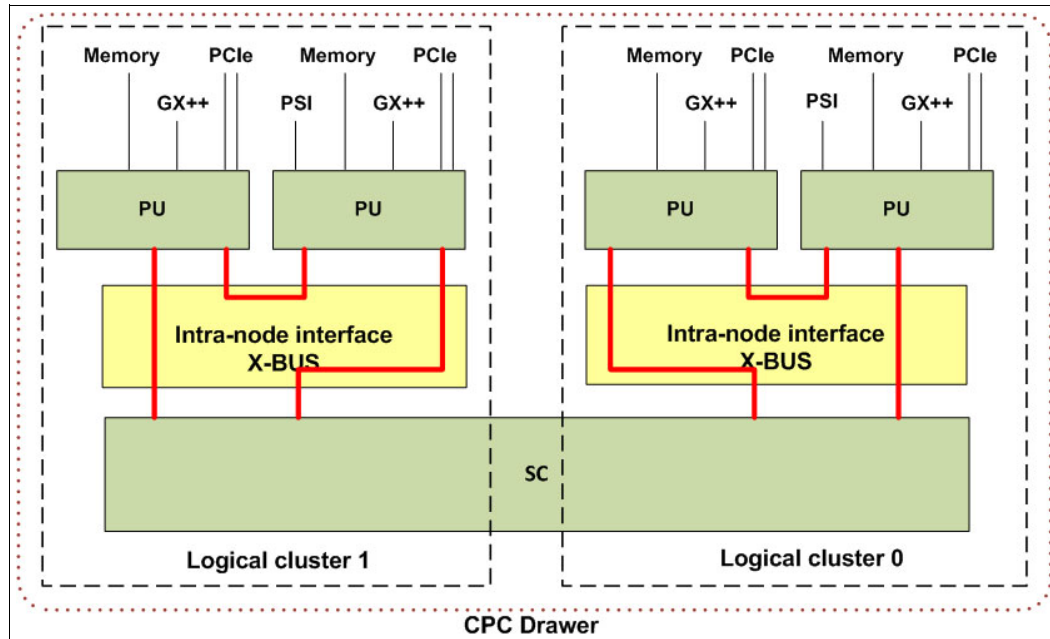


Figure 2-10 CPC drawer logical diagram

The buses are organized in the following configurations:

- ▶ The PCIe I/O buses provide connectivity for PCIe fanouts and can sustain up to 16 Gbps data traffic per bus direction.
- ▶ The X-bus provides interconnects between SC SCM to PU SCM and PU SCMs to each other, in the same node.
- ▶ Processor support interfaces (PSIs) are used to communicate with FSP cards for system control.
- ▶ Configurations with four PU SCMs operate as two Logical PU clusters with two PU SCMs per logical cluster.
- ▶ Configurations with two PU SCMs or one PU SCM operate in one logical PU cluster.

2.3.1 Oscillator cards

The z14 ZR1 CPC drawer contains the two oscillator cards (OSCs): One primary and one backup. If the primary OSC fails, the secondary detects the failure, takes over transparently, and continues to provide the clock signal to the CPC. The two oscillators have Bayonet Neill-Concelman (BNC) connectors that provide pulse per second signal (PPS) input for synchronization to an external time source with PPS output.

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

The accuracy of an STP-only CTN is improved by using an NTP server with the PPS output signal as the External Time Source (ETS). NTP server devices with PPS output are available from several vendors that offer network timing solutions. A cable connection from the PPS port on the OSC to the PPS output of the NTP server is required when z14 ZR1 uses STP and is configured in an STP-only CTN that uses NTP with PPS as the external time source.

The z14 ZR1 server cannot participate in a mixed CTN; it can participate in an STP-only CTN only.

STP with PPS timing signal accuracy

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 z14 ZR1 server.

STP without PPS timing signal accuracy

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

The OSCs cards are plugged into the rear of the CPC drawer in slots LG11 and LG12.

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

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

2.3.2 System control

The various system elements are managed through the Flexible Server Processors (FSPs). An FSP is based on the IBM PowerPC® microprocessor technology. Each FSP card has two ports to connect to two internal Ethernet LANs through the internal network switches (SW1 and SW2). The FSPs communicate with the SEs and provide a subsystem interface (SSI) for controlling components.

An overview of the system control design is shown in Figure 2-11.

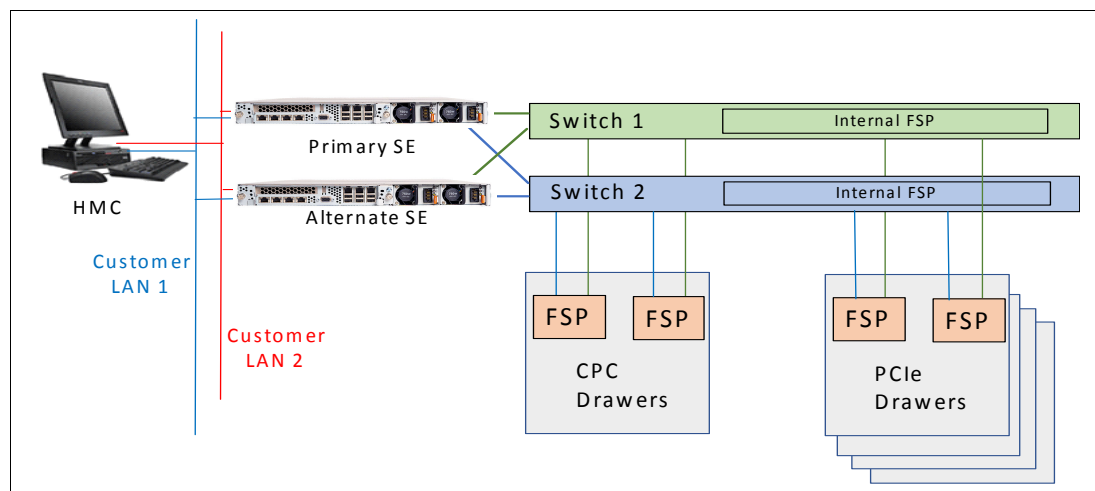


Figure 2-11 Conceptual overview of system control elements

Note: The maximum z14 ZR1 configuration has one CPC drawer and four PCIe+ I/O drawers. The various supported FSP connections are referenced in Figure 2-11.

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 (by 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 at least one 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 Hardware Management Consoles (HMCs) and SEs are connected directly to one or two Ethernet Customer LANs. One or more HMCs can be used.

2.3.3 CPC drawer power

The power for the CPC drawer is a new design. It uses the following combinations of Power Supply Units (PSUs), POL⁵s, VRMs, and Power Control Cards:

- ▶ PSUs: Provide AC to 12V DC bulk/standby power and are installed at the rear of the CPC. The quantity that is installed depends on the following configurations:
 - Four PSUs for configurations with four PU SCMs
 - Two PSUs for configurations with one or two PU SCMs
- ▶ POLs: Point of Load N+2 Redundant cards are installed next to the Memory DIMMs.
- ▶ VRMs: Voltage Regulator Modules are derivative of z13s design (N+2 redundancy).
- ▶ Power Control card: Redundant processor power and control cards connect to the CPC trail board. The control function is powered from 12V standby that is provided by the PSU. The Power Control card also includes pressure, temperature, and humidity sensors.

2.4 Single chip modules

The Single Chip Module (SCM) is a multi-layer metal substrate module that holds one PU chip or an SC chip. Both PU and SC chip size is 696 mm² (25.3 mm x 27.5 mm). Each CPC drawer has one, two or four PU SCMs (6.1 billion transistors each), and one SC SCM (9.7 billion transistors).

⁵ POL - Point of Load, VRM - Voltage Regulator Module.

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

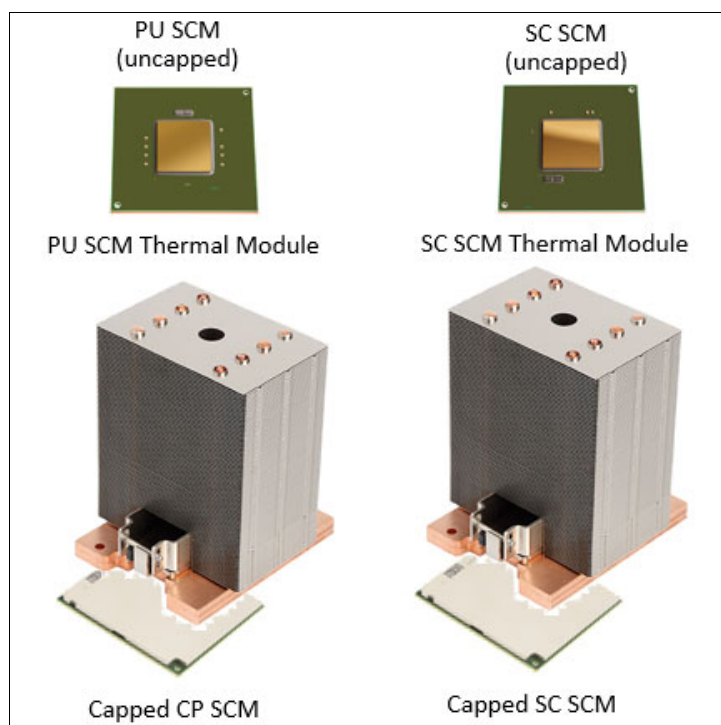


Figure 2-12 Single Chip Modules (PU SCM and SC SCM)

Both PU and SC SCMs use CMOS 14 nm process, 17 layers of metal, and state-of-the-art Silicon-On-Insulator (SOI) technology.

The SCMs are plugged into a socket that is part of the CPC drawer packaging.

2.4.1 Processor Unit Single Chip Module

Note: The terms *PU SCM* (packaged PU chip) and *PU chip* are used interchangeably in this section.

The z14 ZR1 PU chip shares the design with the z14 M0x PU chip and is an evolution of the z13s design. It includes the following features:

- ▶ CMOS 14nm SOI technology
- ▶ Pipeline enhancements, dynamic improved simultaneous multithreading (SMT), enhanced single-instruction multiple-data (SIMD), and redesigned, larger on-chip caches

Each PU chip includes up to nine active cores (10 cores by design) that run at 4.5 GHz, which means that the cycle time is 0.222 ns. The PU SCMs come in five versions: 5, 6, 7, 8 or 9 active cores. A schematic representation of the PU chip is shown in Figure 2-13 on page 35.

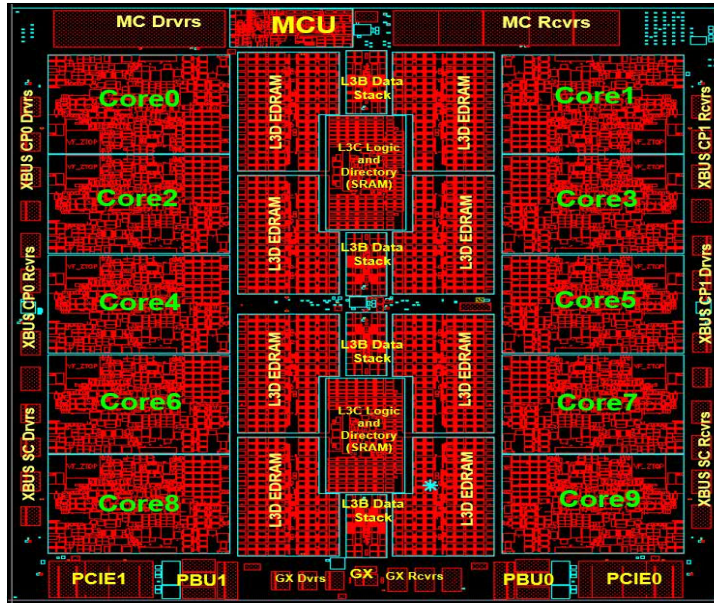


Figure 2-13 PU SCM floor plan

The PU chip contains the following enhancements:

- ▶ Cache Improvements:
 - New power efficient logical directory design
 - 33% larger L1 I\$ (128 KB), private
 - 2x larger L2 D\$ (4 MB), private
 - 2x larger L3 cache with symbol ECC, shared
- ▶ New Translation/TLB2 design:
 - Four concurrent translations
 - Reduced latency
 - Lookup that is integrated into L2 access pipe
 - 2x Consolidated region and segment table entries (CRSTE) growth
 - 1.5x Page table entry (PTE) growth
 - New 64 entry 2 GB TLB2
- ▶ Pipeline Optimizations:
 - Improved instruction delivery
 - Faster branch wake-up
 - Reduced execution latency
 - Improved Operand store compare (OSC) avoidance
 - Optimized second-generation SMT2
- ▶ Better Branch Prediction:
 - 33% Larger BTB1 & BTB2
 - New Perceptron Predictor
 - New Simple Call Return Stack

2.4.2 Processor Unit (Core)

Each processor unit (see Figure 2-14) or core is a superscalar and out-of-order processor that includes 10 execution units.

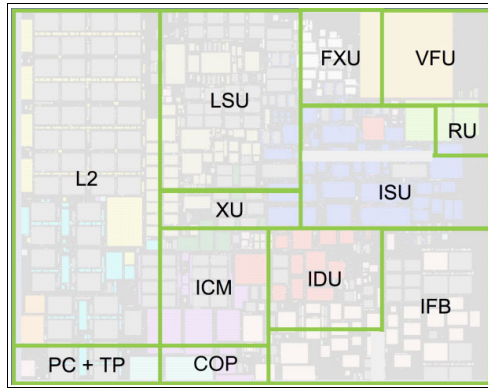


Figure 2-14 PU Core layout

Consider the following points:

- ▶ Fixed-point unit (FXU): The FXU handles fixed-point arithmetic.
- ▶ 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.
- ▶ The instruction fetch and branch (IFB) (prediction) and instruction cache and merge (ICM) sub units contain the instruction cache, branch prediction logic, instruction fetching controls, and buffers. Its relative size is the result of the elaborate branch prediction.
- ▶ 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.
- ▶ 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.
- ▶ 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.
- ▶ 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.
- ▶ Core pervasive unit (PC) for instrumentation, error collection.
- ▶ Vector and Floating point Units (VFU).
- ▶ 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 floating-point, decimal fixed-point, and fixed-point division operations.
- ▶ Vector execution unit (VXU).
- ▶ Level 2 cache (L2) for instructions and data (L2I/L2D).

2.4.3 PU characterization

The PUs are characterized for client use. The characterized PUs can be used in general to run supported operating systems, such as z/OS, z/VM, and Linux on Z. They also can run specific workloads, such as Java, XML services, IPSec, and some Db2 workloads, or functions, such as the Coupling Facility Control Code (CFCC). For more information about PU characterization, see 3.5, “Processor unit functions” on page 83.

The maximum number of characterizable PUs depends on the ZR1 CPC drawer feature code. Some PUs are characterized for system use; some are characterized for client workload use.

By default, one spare PU is available to assume the function of a failed PU. The maximum number of PUs that can be characterized for client use are listed in Table 2-1.

Table 2-1 Number of PUs per z14 model

Feature	CPs	IFLs	zIIPs	ICFs	IFPs	Standard SAPs	Add'l SAPs	Spares
Max30	0 - 6	0 - 30	0 - 12	0 - 30	1	2	0 - 2	1
Max24	0 - 6	0 - 24	0 - 12	0 - 24	1	2	0 - 2	1
Max12	0 - 6	0 - 12	0 - 8	0 - 12	1	2	0 - 2	1
Max4	0 - 4	0 - 4	0 - 2	0 - 4	1	2	0 - 2	1

2.4.4 System Controller SCM (chip)

The System Controller (SC) SCM uses the CMOS 14nm SOI chip technology, with 17 layers of metal. It measures 25.3 x 27.5 mm, and has 9.7 billion transistors. One SC SCM is available per system for all CPC drawer features.

A schematic representation of the SC chip is shown in Figure 2-15 on page 38. Consider the following points:

- ▶ X-Bus⁶ (PU-PU and PU-SC): Significant changes allow SC to fit more X-Bus connections
- ▶ 672 MB shared eDRAM L4 Cache is available
- ▶ L4 Directory is built with eDRAM
- ▶ New L4 Cache Management:
 - L3 to L4 cache capacity ration was increased with the z14 ZR1 processor design
 - New on-drawer Cluster-to-Cluster (topology change) management

⁶ z14 ZR1 is a single CPC drawer system; therefore, the A-Bus is not used. Also, the S-Bus (SC-to-SC connectivity) that was available in z13s was eliminated because of single SC SCM that incorporates a single L4 cache.

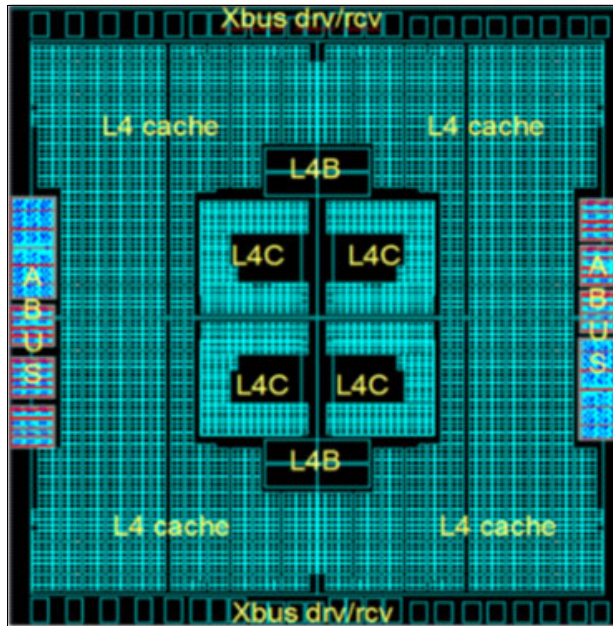


Figure 2-15 SC chip layout

2.5 Memory

The maximum physical memory size is feature-dependent. The orderable memory sizes for z14 ZR1 are listed in Table 2-2.

Table 2-2 z14 ZR1 Memory sizes

Feature	Number of PU SCMs	Standard memory (GB) ^a
Max4	1	64-1984
Max12	2	64-4032
Max24	4	64-8128
Max30	4	64-8128

a. Actual amount of customer usable memory.

The minimum physical installed memory is 160 GB. The minimum initial amount of memory that can be ordered for z14 ZR1 is 64 GB. The maximum customer memory size depends on the number of PU SCMs installed (CPC drawer feature) and is based on the physical installed memory minus the RAIM (20% of physical memory) and minus the hardware system area (HSA) memory, which has a fixed amount of 64 GB.

The memory ordering granularity (installed customer memory) is listed in Table 2-3.

Table 2-3 Memory ordering granularity

Memory increment (GB)	Offered memory sizes (GB)
8	64, 72, 80, 88, 96
32	128, 160, 192, 256, 288, 320, 352, 384
64	448, 512, 576

Memory increment (GB)	Offered memory sizes (GB)
128	704, 832, 960
256	1216, 1472, 1728, 1984, 2240 ... 4032
512	4544, 5056 ... 8128

2.5.1 Memory subsystem topology

The z14 ZR1 memory subsystem uses high-speed, differential-ended communications memory channels. An overview of the CPC drawer memory topology of a z14 ZR1 server is shown in Figure 2-16.

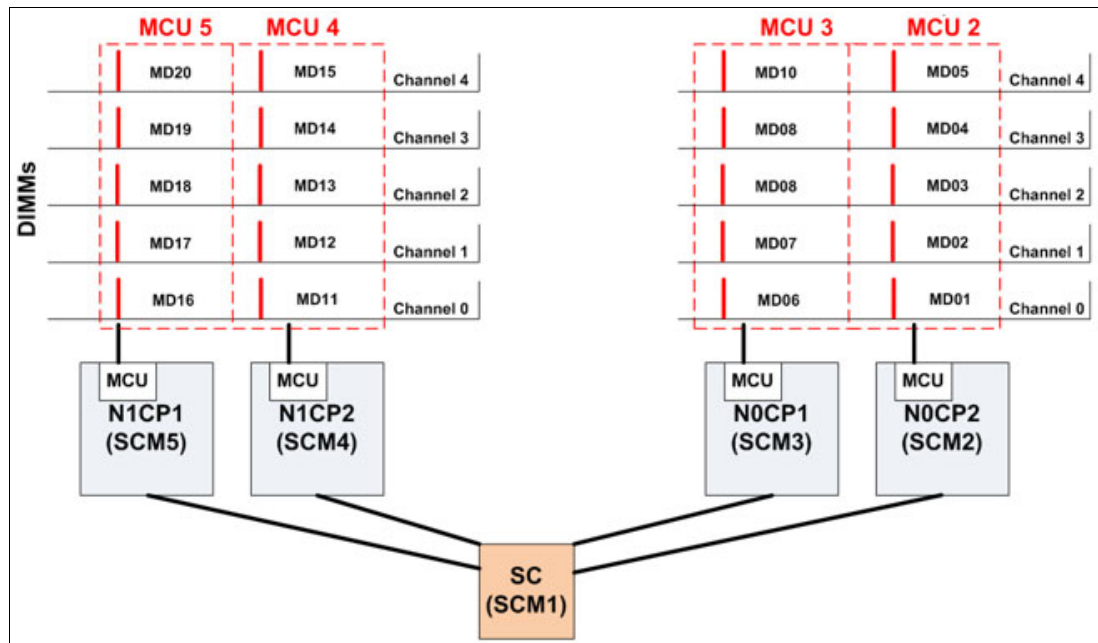


Figure 2-16 CPC drawer memory topology

The CPC drawer includes 5, 10, 15, or 20 DIMMs (up to four populated memory banks). DIMMs are connected to the memory control unit (MCU) on each PU SCM. Each PU SCM has one MCU, which uses five channels: one DIMM per channel, which implements a RAIM protection scheme (4 + 1 design). Each CPC drawer can have one, two, or four populated memory banks.

DIMMs are used in 32, 64, 128, 256, and 512 GB sizes with five DIMMs of the same size included in a memory feature (160, 320, 640, 1280, and 2560 GB RAIM array size).

2.5.2 Redundant array of independent memory

The z14 ZR1 server uses the RAIM protection scheme against hardware memory failures. The RAIM design detects and recovers from failures of dynamic random access memory (DRAM), sockets, memory channels, or DIMMs.

The RAIM design requires the addition of one memory channel that is dedicated for reliability, availability, and serviceability (RAS), as shown in Figure 2-17.

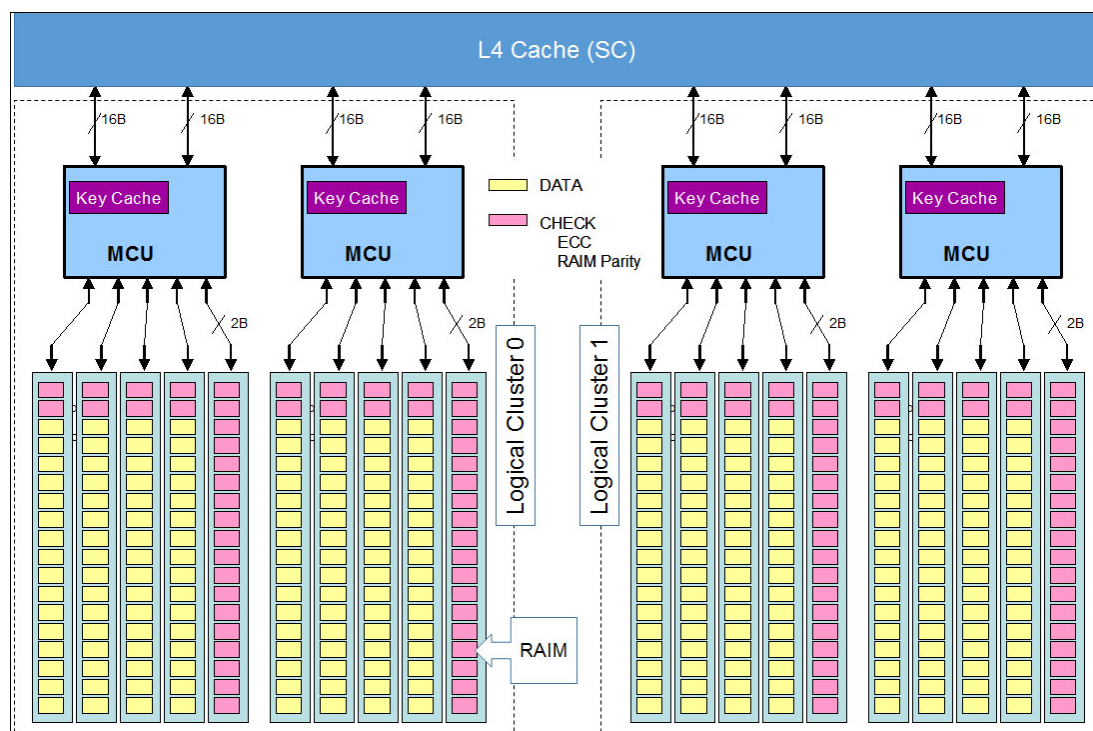


Figure 2-17 RAIM configuration (four PU SCMs)

The fifth channel in each MCU enables memory to be implemented as a Redundant Array of Independent Memory (RAIM). RAIM features significant error detection and correction capabilities: bit, lane, DRAM, DIMM, socket, and complete memory channel failures can be detected and corrected, including many types of multiple failures.

The RAIM design provides the following layers of memory recovery:

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

2.5.3 Memory configurations

Consider the following general plugging rules for memory:

- ▶ 5, 10, 15, or 20 DIMMs are plugged, depending on the configuration (minimum 5 DIMMs: one bank must be installed)
- ▶ No mixing of DRAM technology within a Memory Controller (that is, all DDR4)
- ▶ Each five slot DIMM bank must have the same DIMM size
- ▶ A mix of DIMM sizes can be used between different DIMM banks

- ▶ Another 64 GB of memory is added to the total amount of memory that is specified by the customer. The extra 64 GB of memory is reserved for use by HSA.
- ▶ The CPC drawer can have available unused memory, which can be ordered as a memory upgrade and enabled by LIC without DIMM changes.
- ▶ DIMM changes are disruptive (require machine power off) on z14 ZR1.

Memory location plugging by CPC drawer feature

Memory plugging rules for CPC drawer Max4 feature are listed in Table 2-4.

Table 2-4 Max4 FC 0636 physical memory configurations

	Physical	PU SCM2 MD01-05	PU SCM3 MD06-10	PU SCM4 MD11-15	PU SCM5 MD16-20	GB total	Customer capacity	Hardware increment
1	160	32	N/A	N/A	N/A	128	64	N/A
2	320	64	N/A	N/A	N/A	256	192	128
3	640	128	N/A	N/A	N/A	512	448	256
4	1280	256	N/A	N/A	N/A	1024	960	512
5	2560	512	N/A	N/A	N/A	2048	1984	1024

Memory plugging rules for CPC drawer Max12 feature are listed in Table 2-5.

Table 2-5 Max12 FC 0637 physical memory configurations

	Physical	PU SCM2 MD01-05	PU SCM3 MD06-10	PU SCM4 MD11-15	PU SCM5 MD16-20	GB total	Customer capacity	Hardware increment
1	320	32	32	N/A	N/A	256	192	N/A
2	640	64	64	N/A	N/A	512	448	256
3	1280	128	128	N/A	N/A	1024	960	512
4	2560	256	256	N/A	N/A	2048	1984	1024
5	3840	256	512	N/A	N/A	3072	3008	1024
6	5120	512	512	N/A	N/A	4096	4032	1024

Memory plugging rules for CPC drawer Max24 and Max30 features are listed in Table 2-6.

Table 2-6 Max24 FC 0638 and Max30 FC 0639 physical memory configurations

	Physical	PU SCM2 MD01-05	PU SCM3 MD06-10	PU SCM4 MD11-15	PU SCM5 MD16-20	GB total	Customer capacity	Hardware increment
1	320	32	32	N/A	N/A	256	192	N/A
2	480	32	32	32	N/A	384	320	128
3	640	32	32	32	32	512	448	128
4	960	64	64	32	32	768	704	256
5	1280	64	64	64	64	1024	960	256
6	1600	128	128	32	32	1280	1088	256
7	2560	128	128	128	128	2048	1984	768

	Physical	PU SCM2 MD01-05	PU SCM3 MD06-10	PU SCM4 MD11-15	PU SCM5 MD16-20	GB total	Customer capacity	Hardware increment
8	3840	128	128	256	256	3072	3008	1024
9	5120	256	256	256	256	4096	4032	1024
10	6400	256	256	256	512	5120	5056	1024
11	7680	256	256	512	512	6144	6080	1024
12	8960	256	256	512	512	7168	7104	1024
13	10240	512	512	512	512	8192	8128	1024

The View Hardware Configuration task on the Support Element lists all the hardware components in the system. It can be used to view the memory DIMM capacity that is installed by location in the system. An example of a system with 256 GB DIMMs installed in all 20 slots in the CPC Drawer is shown in Figure 2-18.

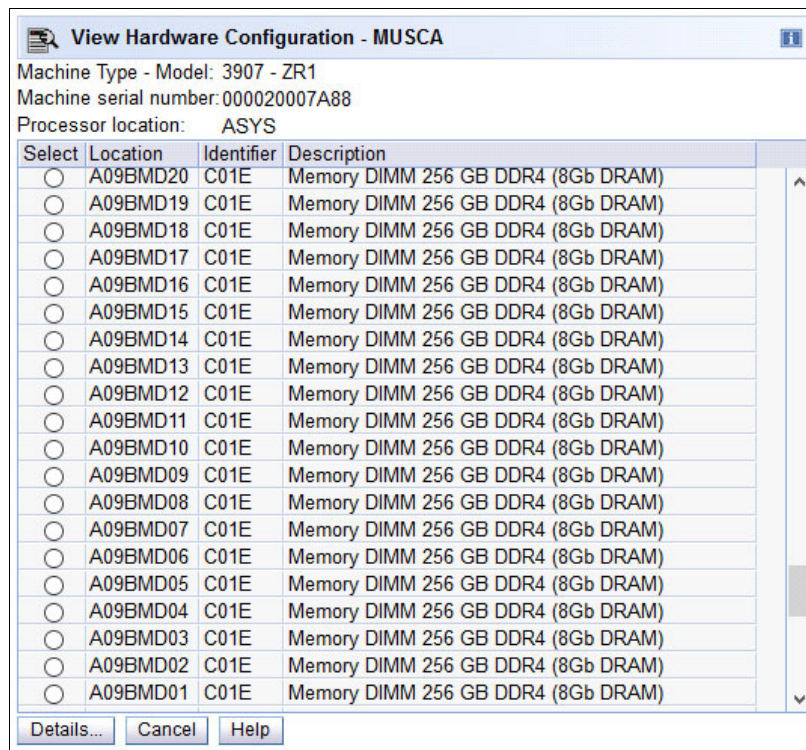


Figure 2-18 View Hardware Configuration task on the Support Element

The CPC drawer and DIMM locations for a z14 ZR1 are shown in Figure 2-19.

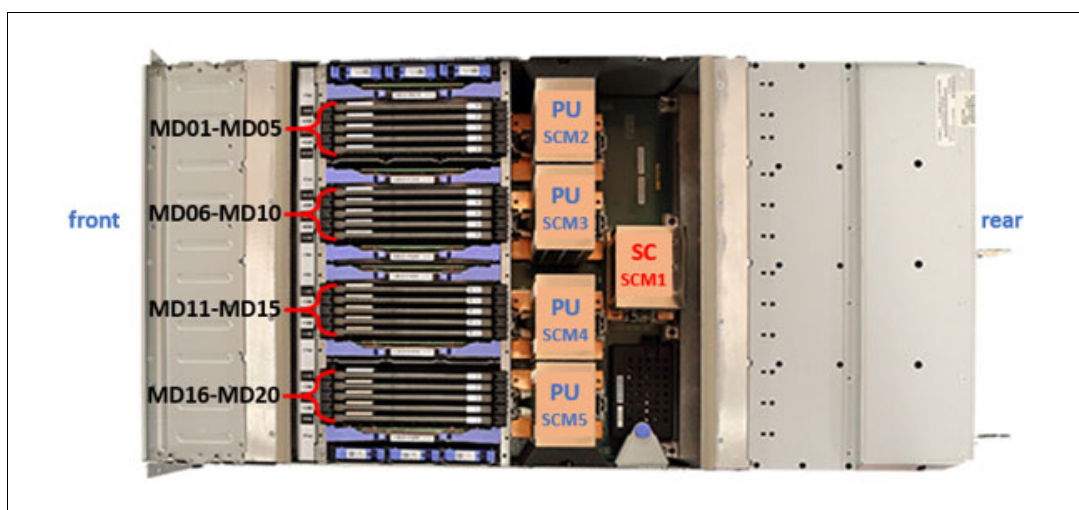


Figure 2-19 CPC drawer and DIMM locations

The physical memory DIMM plugging configurations by feature code from manufacturing when the system is ordered are listed in Table 2-7. The drawer columns for the specific model contain the memory configuration number for the specific drawer. Use available unused memory that can be enabled by LIC, when required.

If more storage is ordered by using other feature codes, such as Virtual Flash Memory or Preplanned memory, the extra storage is installed and plugged as necessary.

Table 2-7 Memory features

FC	Incr.	Mem. Incr. GB	Max4		Max12			Max24 / Max30				
			SCM2 Qty 5	Dial Max	SCM2 Qty 5	SCM2 Qty 5	Dial Max	SCM2 Qty 5	SCM2 Qty 5	SCM2 Qty 5	SCM2 Qty 5	Dial Max
3539	8	64	32GB	64	32GB	32GB	192	32GB	32GB	X	X	192
3540		72	64GB	192	32GB	32GB		32GB	32GB	X	X	
3541		80	64GB		32GB	32GB		32GB	32GB	X	X	
3542		88	64GB		32GB	32GB		32GB	32GB	X	X	
3543		96	64GB		32GB	32GB		32GB	32GB	X	X	
3544	32	128	64GB		32GB	32GB		32GB	32GB	X	X	
3545		160	64GB		32GB	32GB		32GB	32GB	X	X	
3546		192	64GB		32GB	32GB		32GB	32GB	X	X	
3547		224	128GB	448	64GB	64GB	448	32GB	32GB	32GB	X	320
3548		256	128GB		64GB	64GB		32GB	32GB	32GB	X	
3549		288	128GB		64GB	64GB		32GB	32GB	32GB	X	
3550		320	128GB		64GB	64GB		32GB	32GB	32GB	X	
3551		352	128GB		64GB	64GB		32GB	32GB	32GB	32GB	448
3552		384	128GB		64GB	64GB		32GB	32GB	32GB	32GB	
3553	64	448	128GB		64GB	64GB		32GB	32GB	32GB	32GB	
3554		512	256GB	960	128GB	128GB	960	64GB	64GB	32GB	32GB	704

FC	Incr.	Mem. Incr. GB	Max4		Max12			Max24 / Max30				
			SCM2 Qty 5	Dial Max	SCM2 Qty 5	SCM2 Qty 5	Dial Max	SCM2 Qty 5	SCM2 Qty 5	SCM2 Qty 5	SCM2 Qty 5	Dial Max
3555		576	256GB		128GB	128GB		64GB	64GB	32GB	32GB	
3556	128	704	256GB		128GB	128GB		64GB	64GB	64GB	64GB	
3557		832	256GB		128GB	128GB		64GB	64GB	64GB	64GB	960
3558		960	256GB		128GB	128GB		64GB	64GB	64GB	64GB	
3559	256	1216	512GB	1984	256GB	256GB	1984	128GB	128GB	128GB	128GB	1984
3560		1472	512GB		256GB	256GB		128GB	128GB	128GB	128GB	
3561		1728	512GB		256GB	256GB		128GB	128GB	128GB	128GB	
3562		1984	512GB		256GB	256GB		128GB	128GB	128GB	128GB	
3563		2240			256GB	256GB	3008	128GB	128GB	256GB	256GB	3008
3564		2496			256GB	256GB		128GB	128GB	256GB	256GB	
3565		2752			256GB	256GB		128GB	128GB	256GB	256GB	
3566		3008			256GB	256GB		128GB	128GB	256GB	256GB	
3567		3264			512GB	512GB	4032	256GB	256GB	256GB	256GB	4032
3568		3520			512GB	512GB		256GB	256GB	256GB	256GB	
3569		3776			512GB	512GB		256GB	256GB	256GB	256GB	
3570		4032			512GB	512GB		256GB	256GB	256GB	256GB	
3571	512	4544						256GB	256GB	256GB	512GB	5056
3572		5056						256GB	256GB	256GB	512GB	
3573		5568						256GB	256GB	512GB	512GB	6080
3574		6080						256GB	256GB	512GB	512GB	
3575		6592						256GB	512GB	512GB	512GB	7104
3576		7104						256GB	512GB	512GB	512GB	
3577		7616						512GB	512GB	512GB	512GB	8128
3578		8128						512GB	512GB	512GB	512GB	

2.5.4 Memory upgrades

Memory upgrades can be ordered and enabled by LICCC, by upgrading (replacing with higher capacity) the DIMM cards, or by adding DIMM cards.

If all or part of the added memory is enabled for use, it might become available to an active LPAR if the partition includes defined reserved storage. (For more information, see 3.7.3, “Reserved storage” on page 105.) Alternatively, the added memory can be used by an already-defined LPAR that is activated after the memory addition.

Note: Memory downgrades by way of LICCC are always disruptive.

2.5.5 Virtual Flash Memory

IBM Virtual Flash Memory (VFM) FC 0614 replaces the Flash Express features (0402 and 0403) that were available on the IBM z13s. It offers up to 2.0 TB of virtual flash memory in up to four 512 MB increments for improved application availability and to handle paging workload spikes.

No application changes are required to change from IBM Flash Express to VFM. Consider the following points:

- ▶ Dialed memory + zVFM = total hardware plugged
- ▶ Dialed memory + Plan Ahead memory + VFM = total hardware plugged

VFM is offered as a dialed 512GB memory increment size. Feature code 0614 represents one 512GB zVFM increment, as shown in the following example:

FC 0614 - Min=0, Max=4

VFM is designed to help improve availability and handling of paging workload spikes when z/OS V2.1, V2.2, or V2.3, or on z/OS V1.13⁷ is run. With this support, z/OS is designed to help improve system availability and responsiveness by using VFM across transitional workload events, such as market openings and diagnostic data collection. z/OS is also designed to help improve processor performance by supporting middleware use of pageable large (1 MB) pages.

VFM can also be used in coupling facility images to provide extended capacity and availability for workloads that use IBM WebSphere MQ Shared Queues structures. The use of VFM can help availability by reducing latency from paging delays that can occur at the start of the workday or during other transitional periods. It is also designed to help eliminate delays that can occur when collecting diagnostic data during failures.

VFM can help organizations meet their most demanding service level agreements and compete more effectively. VFM is easy to configure, and provides rapid time to value.

2.5.6 Preplanned memory

Preplanned memory helps plan for nondisruptive memory upgrades. The required hardware is pre-plugged based on a target capacity that is specified by the customer. This pre-plugged hardware is enabled by way of an LICCC order that is placed by the customer when they determine that more memory capacity is needed. The pre-plugged memory can be made available through a LICCC update.

For more information about ordering this LICCC, see the following resources:

- ▶ The [IBM Resource Link](#) website (login required)
- ▶ Your IBM representative

⁷ z/OS V1.13 includes more requirements. For more information, see Chapter 7, “Operating system support” on page 209.

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

Table 2-8 Feature codes for plan-ahead memory

Memory	ZR1 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 1993 - 8 GB ▶ FC 1996 - 16 GB
Virtual Flash Memory (VFM) Pre-planned Memory Charged when physical memory is installed. Used for tracking the quantity of physical increments of plan-ahead VFM memory capacity.	FC 1999 - 64 GB
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 1739 (8 GB memory capacity Increments <128GB) ▶ FC 1740 (8 GB memory capacity Increments >=128GB) ▶ FC 1741(16 GB memory capacity Increments >1=28GB) ▶ FC 1742 (32GB memory capacity increments >=128GB)

The payment for plan-ahead memory is a two-phase process. One charge occurs when the plan-ahead memory is ordered. Another charge occurs when the prepaid memory is activated for use. For more information about the exact terms and conditions, contact your IBM representative.

Pre-planned memory is installed by ordering FC 1993 (8 GB) or FC 1996 (16 GB). The ordered amount of plan-ahead memory is charged at a reduced price compared to the normal price for memory. One FC 1993 is needed for each 8 GB of usable memory (10 GB RAIM), or one FC 1996 is needed for each 16 GB of usable memory (20GB RAIM).

Reminder: Maximum amount of preplanned memory is limited to 2TB

2.6 Reliability, availability, and serviceability

IBM Z servers continue to deliver enterprise class RAS with IBM z14 ZR1 servers. The main philosophy behind RAS is about preventing or tolerating (masking) outages. It is also about providing the necessary instrumentation (in hardware, LIC and microcode, and software) to capture or collect the relevant failure information to help identify an issue without requiring a reproduction of the event. These outages can be planned or unplanned. Planned and unplanned outages can include the following situations (examples are not related to the RAS features of IBM Z servers):

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

The IBM Z hardware has decades of intense engineering behind it, which results in a robust and reliable platform. The hardware has many RAS features that are built into it. For more information, see Chapter 9, “Reliability, availability, and serviceability” on page 327.

2.6.1 RAS in the CPC memory subsystem

Patented error correction technology in the memory subsystem continues to provide the most robust error correction from IBM to date. Two full DRAM failures per rank can be spared and a third full DRAM failure can be corrected.

DIMM level failures, including components, such as the memory controller application-specific integrated circuit (ASIC), power regulators, clocks, and system board, can be corrected. Memory channel failures, such as signal lines, control lines, and drivers and receivers on the MCM, can be corrected.

Upstream and downstream data signals can be spared by using two spare wires on the upstream and downstream paths. One of these signals can be used to spare a clock signal line (one upstream and one downstream). The following improvements were also added in the z14 ZR1 server:

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

2.6.2 General z14 ZR1 RAS features

The z14 ZR1 server includes the following RAS features:

- ▶ The Power/Thermal Subsystem is new for z14 ZR1. It uses switchable, intelligent Power Distribution Units (PDUs) instead of Bulk Power Assemblies in past generations. The z14 ZR1 server provides a true N+1 cooling function with fans.
- ▶ Redundant (N+1), number of PDUs is configuration-dependent (2 or 4).
Input power Single Phase, not cross-coupled. The loss of a single phase puts power subsystem into N-mode.
- ▶ CPC drawer is packaged to fit in the 19-inch rack. CPC drawer power and cooling includes:
 - PSUs: AC to 12V bulk/standby (N+1 redundant). PSU Fans are not separate FRUs and are available in quantities of 2 or 4.
 - POLs: N+2 Phase and Master Redundant and are available in quantities of 3 - 6.
 - VRM sticks⁸: Derivative of z13s design (N+2 Phase and Master redundancy) and are available in quantities of 6.
 - Power Control Card: New power control card to control CPC fans (N+1 redundant) and are available in quantities of 2.
 - SCMs are all air-cooled and used new heat sinks for PU SCMs on z14 ZR1. The SC heat sink on the SC SCM is the same from z14 M0x.

⁸ Voltage Regulator Module stick converts the DC bulk power that is delivered by the PSUs (12V) into localized low voltage that is used by the installed components (for example, PU SCMs, SC SCM, memory DIMMs, and other circuitry).

- Fans: Drawer has five fans and are N+1 redundant.
- FSPs: Redundant (N+1).
- ▶ PCIe+ I/O Drawer Power/Thermal is all new for z14 ZR1:
 - Two Power Supply assemblies: Power supply with dedicated on board fan, combined with I/O Power Control (Power Supply and I/O Power Control are separate FRUs) in N+1 configuration
 - Fans: Drawer has six fans, N+1 redundant
 - FSPs: Redundant (N+1)

The internal intelligent Power Distribution Unit (iPDU) provide the following capabilities:

- ▶ Switchable PDUs provide outlet control by way of Ethernet:
 - Provide a System Reset capability
 - Power cycle an SE if a hang occurs
 - Verify a power cable at installation
- ▶ System Reset Function:
 - No EPO switch is on the z14 ZR1. This function provides a means to put a server into a known state similar to past total power reset.
 - This function does not provide the option to power down and keep the power down to the system. The power must be unplugged or the customer-supplied power is turned off at the panel.
- ▶ Other characteristics:
 - iPDU Firmware can be concurrently updated
 - Concurrently repairable
 - Power redundancy check
- ▶ Cable verification test by way of IPDU:
 - By power cycling individual iPDU outlets, the system can verify proper cable connectivity
 - Power cable test runs during system Power On
 - Runs at system Installation and at every system Power On until the test passes and erases the cable test file

The power service and control network (PSCN) is used to control and monitor the elements in the system and include the following components:

- ▶ Ethernet Top of Rack (TOR) switches provide the internal PSCN connectivity:
 - Switches are redundant (N+1)
 - Concurrently maintainable
 - Each switch has on integrated power supply
 - FSPs are cross wired to the Ethernet switches
- ▶ Redundant SEs

Each SE has two power supplies (N+1) and input power is cross-coupled from the PDUs.

IBM z14 ZR1 servers continue to deliver robust server designs through new technologies, hardening both new and classic redundancy.

For more information, see Chapter 9, “Reliability, availability, and serviceability” on page 327.

2.7 Connectivity

Connections to PCIe+ I/O drawers and Integrated Coupling Adapters (ICAs) are driven from the CPC drawer fanout cards. These fanouts are on the front of the CPC drawer.

The location of the fanouts for the CPC drawer is shown in Figure 2-20.

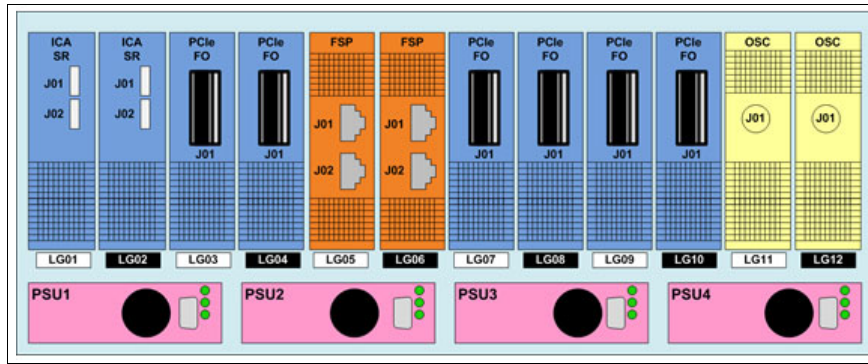


Figure 2-20 Location of the PCIe, FSP and OSC adapters

The number of available fanouts depends on the CPC drawer feature:

- ▶ Eight PCIe fanout slots are available for systems with Max24/Max30 CPC drawer. PCIe fanout locations are LG01 - LG04 and LG07 - LG10.
- ▶ Two PCIe fanouts are available for systems with Max4 CPC drawer. PCIe fanout locations are LG09 and LG10.
- ▶ Four PCIe fanouts are available for systems with Max12 CPC drawer. PCIe fanout locations are LG07- LG10.

The CPC drawer has two FSPs for system control. The location codes for the FSPs are LG05 and LG06.

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 49.

The following types of fanouts are available:

- ▶ PCIe Generation3 fanout card: This copper fanout provides connectivity to the PCIe switch cards in the PCIe+ I/O drawer.
- ▶ Integrated Coupling Adapter (ICA SR): This adapter provides coupling connectivity between z14 ZR1 and z14 M0x / z13 / z13s servers.

When you are configuring for availability, balance coupling links across adapters, and I/O features across PCIe+ I/O drawers. In a system that is configured for maximum availability, alternative paths maintain access to critical I/O devices, such as disks and networks. The CHPID Mapping Tool can be used to assist with configuring a system for high availability.

2.7.1 Redundant I/O interconnect

Redundancy is provided for PCIe I/O interconnects.

The PCIe+ I/O drawer supports up to 16 PCIe cards, which are organized in two hardware domains per drawer, as shown in Figure 2-21 on page 50.

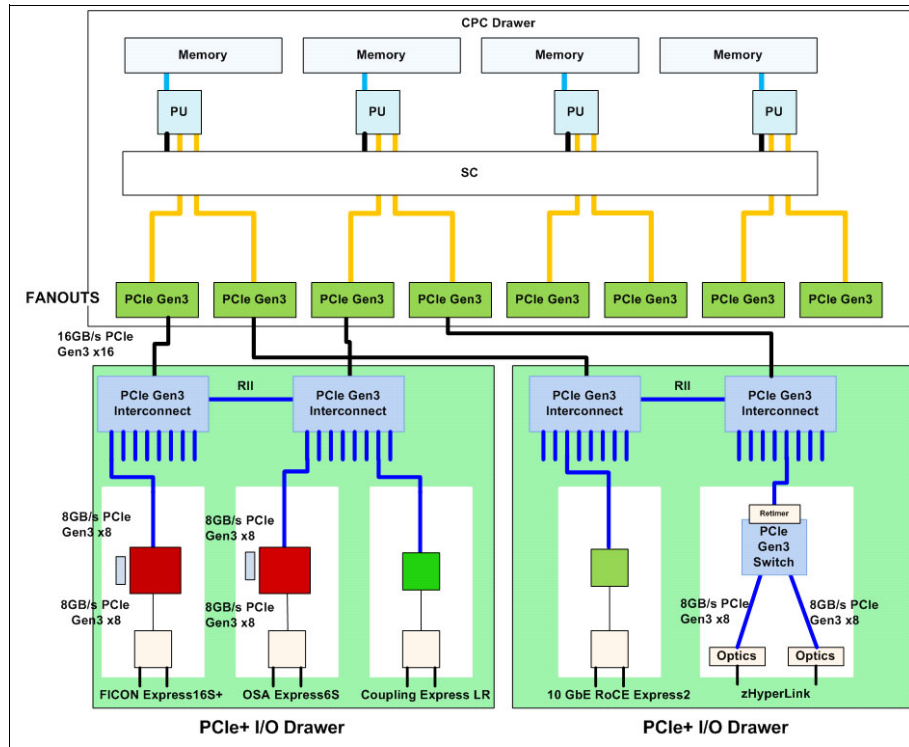


Figure 2-21 Infrastructure for PCIe+ I/O drawer (system with two PCIe+ I/O drawers)

Each domain is driven through a PCIe Gen3 switch. The two PCIe switch cards (LG06 and LG16) provide a backup path (Redundant I/O Interconnect - RII) for each other through the passive connection in the PCIe+ I/O drawer backplane. During a PCIe fanout or cable failure, all 16 PCIe cards in the two domains can be driven through a single PCIe switch card (see Figure 2-22).

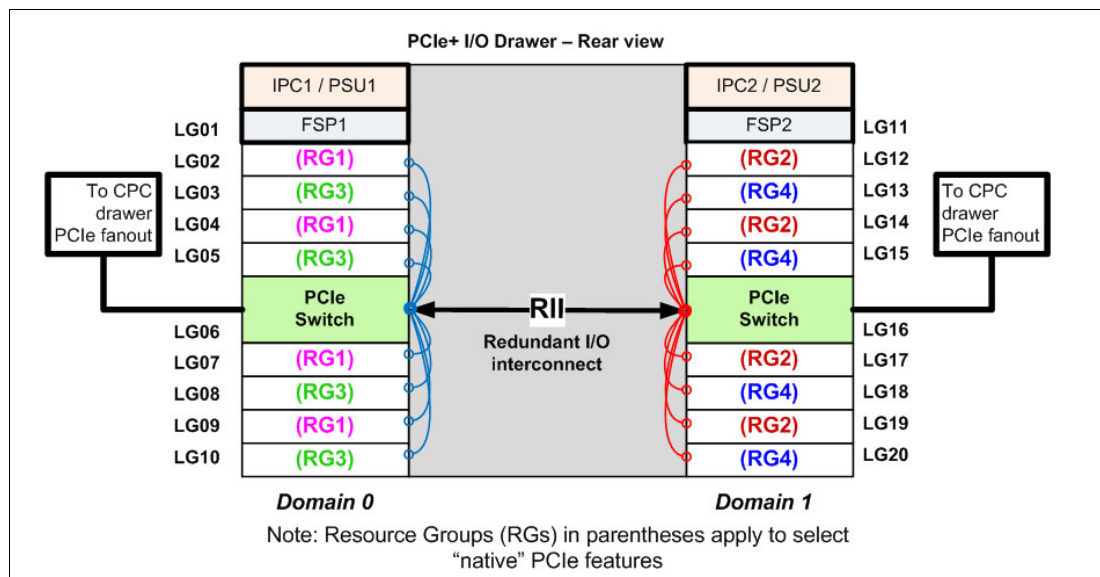


Figure 2-22 Redundant I/O Interconnect

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

Note: The PCIe Gen3 Interconnect (switch) adapter must be installed in the PCIe+ I/O drawer to maintain the interconnect across I/O domains. If the adapter is removed (for a service operation), the I/O cards in that domain (up to eight) become unavailable.

2.7.2 CPC drawer upgrades

All fanouts that are used for I/O and coupling links are rebalanced if the upgrade does not involve adding components inside the CPC drawer. If MES involves extra components, such as PU SCMs and Memory DIMMs, the change is disruptive.

When a z14 ZR1 is ordered, the PUs are characterized according to their intended usage. The PUs can be ordered as any of the following items:

CP	The processor is purchased and activated. PU supports the z/OS, z/VSE, z/VM, z/TPF, and Linux on Z ⁹ operating systems. It can also run Coupling Facility Control 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 PU is installed. It does not affect software licenses or maintenance charges.
IFL	The Integrated Facility for Linux (IFL) is a processor that is purchased and activated for use by z/VM for Linux guests and Linux on Z ⁹ operating systems.
Unassigned IFL	A processor that is purchased for future use as an IFL. It is offline and cannot be used until an upgrade for the IFL is installed. It does not affect software licenses or maintenance charges.
ICF	An internal coupling facility (ICF) processor that is purchased and activated for use by the Coupling Facility Control Code.
zIIP	An “Off Load Processor” for workload that is restricted to Db2 type applications.
Additional SAP	An optional processor that is purchased and activated for use as an SAP (System Assist Processor).

A minimum of one PU that is characterized as a CP, IFL, or ICF is required per system. The maximum number of characterizable PUs is 30 (any combination of 6 CPs, 30 IFLs, 30 ICF, two more SAP¹⁰s, and 12 zIIPs). The maximum number of zIIPs is up to twice the number of PU that are characterized as CP.

Remainder: Not all PUs on a model must be characterized.

⁹ The KVM hypervisor is part of select Linux on Z distributions.

¹⁰ Two standard SAPs with every ZR1 system are always featured. Up to two more SAPs can be ordered. The extra SAPs are part of the customer characterizable PUs.

The following items are present in the z14 ZR1 server, but they are not part of the PUs that clients purchase and require no characterization:

- ▶ Standard SAP to be used by the channel subsystem. The number of standard SAPs is always two.
- ▶ One IFP (Integrated Firmware Processor), which is used in the support of “native” PCIe features.
- ▶ One spare PU, which can transparently assume any characterization if another PU a permanent fails.

The various feature code driven CPC drawer sizes are listed in Table 2-1 on page 37.

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

2.7.3 System upgrades

Concurrent upgrades of CPs, IFLs, ICFs, zIIPs, or SAPs are available for the z14 ZR1 server. However, concurrent PU upgrades require that more PUs are installed, but not activated.

The spare PU is used to replace defective PUs. In the rare event of a PU failure, a spare PU is activated concurrently and transparently and is assigned the characteristics of the failing PU.

If an upgrade request cannot be accomplished within the configuration, a hardware upgrade is required. The upgrade that requires the addition of one or more PU SCMs and DIMMs per to accommodate the wanted capacity.

All upgrades for z14 ZR1 that involve adding PU SCMs or memory are disruptive.

The upgrade paths for the z14 are shown in Figure 2-23.

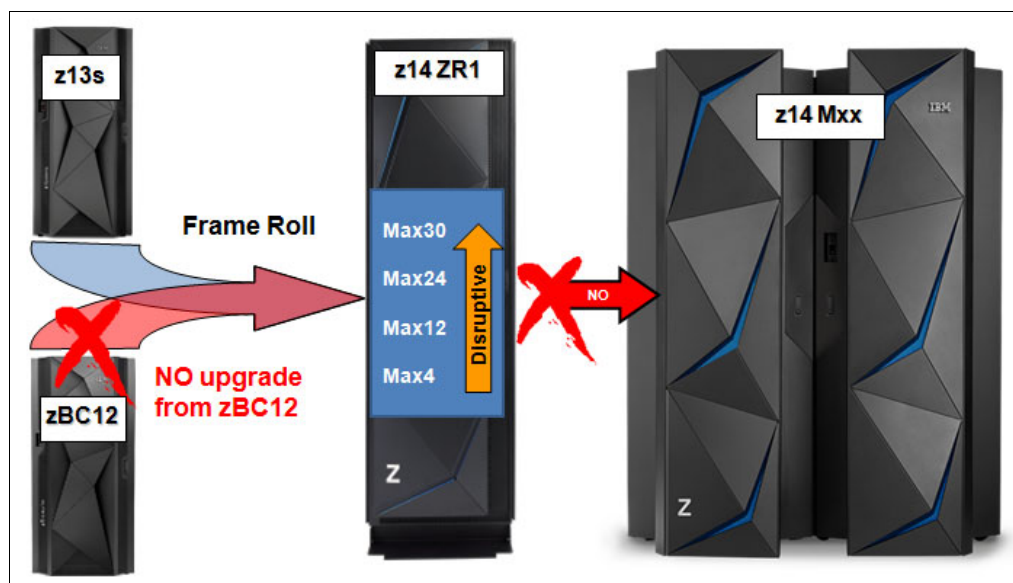


Figure 2-23 z14 ZR1 system upgrade paths

Consider the following points regarding upgrades:

- ▶ Upgrade from a z14 LR1 to a z14 ZR1 *is* supported
- ▶ Upgrade from z14 ZR1 to z14 LR1 is *not* supported
- ▶ Upgrade from z14 ZR1 to z14 M0x is *not* supported
- ▶ Upgrade from z14 LR1 to z14 LM1 is *not* supported

You can upgrade a z13s (2965) server and preserve the CPC serial number (S/N). The I/O cards can also be carried forward (with certain restrictions) to the z14 ZR1 server. For frame roll MES from z13s to z14 ZR1, new frames are shipped. New PCIe+ I/O drawers are supplied with the MES for z13s to replace the PCIe I/O drawers.

Important: Upgrades from IBM z13s are always disruptive.

2.7.4 Concurrent PU conversions

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

Most conversions are nondisruptive. In exceptional cases, the conversion might be disruptive; for example, when a model ZR1 with 6 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 listed in Table 2-9.

Table 2-9 Concurrent PU conversions

To\From	CP	IFL	Unassigned IFL	ICF	zIIP	Additional SAP
CP	-	Yes	Yes	Yes	Yes	Yes
IFL	Yes	-	Yes	Yes	Yes	Yes
Unassigned IFL	Yes	Yes	-	Yes	Yes	Yes
ICF	Yes	Yes	Yes	-	Yes	Yes
zIIP	Yes	Yes	Yes	Yes	-	Yes
Additional SAP	Yes	Yes	Yes	Yes	Yes	-

2.7.5 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, ICF, or zIIP is not reflected in the output of the STSI instruction because characterization has no effect on software charging. For more information about STSI output, see “Processor identification” on page 321.

Capacity identifiers: Within a z14 ZR1 server, all CPs feature the same capacity identifier. Specialty engines (IFLs, zIIPs, and ICFs) operate at full speed.

Model capacity identifiers

All model capacity identifiers feature a related MSU value that is used to determine the software license charge for MLC software, as listed in Table 2-10 on page 54.

Table 2-10 Model capacity identifier and MSU values

Model cap ID	MSU	Model cap ID	MSU	Model cap ID	MSU	Model cap ID	MSU	Model cap ID	MSU	Model cap ID	MSU
A01	11	B01	12	C01	14	D01	16	E01	19	F01	21
A02	21	B02	23	C02	26	D02	30	E02	35	F02	40
A03	30	B03	33	C03	38	D03	44	E03	51	F03	59
A04	39	B04	44	C04	49	D04	58	E04	67	F04	77
A05	48	B05	53	C05	60	D05	71	E05	82	F05	94
A06	56	B06	63	C06	70	D06	83	E06	96	F06	110
G01	24	H01	27	I01	30	J01	34	K01	38	L01	42
G02	45	H02	51	I02	57	J02	64	K02	71	L02	80
G03	64	H03	74	I03	83	J03	93	K03	104	L03	116
G04	86	H04	96	I04	108	J04	121	K04	135	L04	150
G05	105	H05	118	I05	132	J05	147	K05	164	L05	184
G06	123	H06	138	I06	154	J06	173	K06	193	L06	216
M01	48	N01	53	O01	60	P01	67	Q01	75	R01	83
M02	90	N02	100	O02	112	P02	125	Q02	140	R02	156
M03	130	N03	145	O03	162	P03	181	Q03	203	R03	225
M04	168	N04	189	O04	211	P04	235	Q04	263	R04	294
M05	206	N05	231	O05	258	P05	288	Q05	322	R05	360
M06	241	N06	271	O06	303	P06	339	Q06	379	R06	423
S01	93	T01	104	U01	117	V01	130	W01	145	X01	162
S02	175	T02	195	U02	218	V02	243	W02	272	X02	304
S03	253	T03	282	U03	316	V03	353	W03	395	X03	442
S04	328	T04	367	U04	411	V04	459	W04	514	X04	575
S05	402	T05	450	U05	504	V05	563	W05	629	X05	705
S06	473	T06	529	U06	592	V06	663	W06	741	X06	828
Y01	178	Z01	195								
Y02	333	Z02	365								
Y03	484	Z03	531								
Y04	631	Z04	693								
Y05	772	Z05	848								
Y06	909	Z06	998								

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

2.7.6 Capacity Backup Upgrade

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

- ▶ CBU for CP
- ▶ CBU for IFL
- ▶ CBU for ICF
- ▶ CBU for zIIP
- ▶ CBU for more (optional) SAPs

When CBU for CP is added within the same capacity setting range (indicated by the model capacity indicator) as the currently assigned PUs, the total number of active PUs (the sum of all assigned CPs, IFLs, ICFs, zIIPs, and optional SAPs) plus the number of CBUs cannot exceed the total number of PUs that are 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, and zIIPs) always run at full capacity, and when running as a CBU to replace lost capacity for disaster recovery.

When you order CBU, specify the maximum number of CPs, ICFs, IFLs, zIIPs, and SAPs to be activated for disaster recovery. If a disaster occurs, 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 various CBU activation variations that apply to the installation.

The number of CBU test activations that you can run for no extra 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 includes 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 1 - 5 years.

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

- ▶ Number of CP CBUs that are allowed to be activated
- ▶ Number of IFL CBUs that are allowed to be activated
- ▶ Number of ICF CBUs that are allowed to be activated
- ▶ Number of zIIP CBUs that are allowed to be activated
- ▶ Number of SAP CBUs that are allowed to be activated
- ▶ Number of extra 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 this example, one CBU record is made for a five-year CBU contract without more 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, and two zIIP CBUs, with five more tests and a five-year CBU contract. The result is that 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 were requested. Because a total of five years are contracted for a total of three CP CBUs (two IFL CBUs 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) - zIIP(2)
Total Tests(5) - Years(5)

Resulting cumulative feature numbers in configuration:

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

CBU for CP rules

Consider the following guidelines when 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 two CBU CPs are added to the current model D03, and the capacity level does not change, the D03 becomes D05, as shown in the following example:

$(D03 + 2 = D05)$

If the capacity level changes to a E06, the number of extra CPs (three) is added to the three CPs of the D03, which results in a total number of CBU CP capacity features of six:

$(3 + 3 = 6)$

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

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

CBU for specialty engines

Specialty engines (ICFs, IFLs, and zIIPs) run at full capacity for all capacity settings. This fact also applies to CBU for specialty engines. The minimum and maximum (min-max) numbers of all types of CBUs that can be activated on each of the models are listed in Table 2-11. The CBU record can contain larger numbers of CBUs than can fit in the current model.

Table 2-11 Capacity Backup matrix

Model ZR1	Total PUs available	CBU CPs min - max	CBU IFLs min - max	CBU ICFs min - max	CBU zIIPs min - max	CBU SAPs min - max
Max30	30	0-6	0 - 30	0 - 30	0 - 12	0 - 2
Max24	24	0-6	0 - 24	0 - 24	0 - 12	0 - 2
Max12	12	0-6	0 - 12	0 - 12	0 - 8	0 - 2
Max4	4	0-4	0 - 4	0 - 4	0 - 2	0 - 2

2.7.7 On/Off Capacity on Demand and CPs

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

On/Off CoD and granular capacity

When temporary capacity that is 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 D03 has two CPs added temporarily, it becomes a model capacity identifier D05.

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

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

On/Off CoD guidelines

When you request temporary capacity, consider the following guidelines:

- ▶ Temporary capacity must be greater than permanent capacity.
- ▶ Temporary capacity cannot be more than double the purchased capacity.
- ▶ On/Off CoD cannot decrease the number of engines on the CPC.
- ▶ The number of engines cannot be increased to more than what is installed.

For more information about temporary capacity increases, see Chapter 8, “System upgrades” on page 281.

2.8 Power and cooling

The z14 ZR1 power and cooling system is a change from previous systems because the system is packaged in an industry standard 19-inch rack form factor for all the internal system elements. The power subsystem is based on Power Distribution Units (PDUs) that are mounted at the rear of the system in pairs. The new PSCN structure uses industry standard Ethernet TOR switches that replace the previous IBM System Control Hubs (SCHs).

2.8.1 Considerations

The IBM Z systems operate with redundant power infrastructure. The z14 ZR1 is designed with a new power infrastructure that is based on intelligent (PDUs that are mounted vertically on the rear side of the 19-inch rack) and Power Supply Units for the internal components. The PDUs are single phase, 200 - 240 VAC and are controlled by using an Ethernet port.

The power supply units convert the AC power to DC power that is used as input for the Points of Load (POLs) in the CPC drawer and the PCIe+ I/O drawers.

The power requirements depend on the number of PU SCMs that is installed in the CPC drawer and number of PCIe+ I/O drawers (1 - 4) and I/O features that are installed in the PCIe+ I/O drawers.

The power subsystem in a z14 ZR1 includes the following main characteristics:

- ▶ Two or four single phase PDUs (200 - 240 VAC, 50/60 Hz), each with a 30A power cord.
- ▶ Maximum supported rack power is 9600W.
- ▶ No High-Voltage DC power option.
- ▶ No three-phase power, no 480 VAC.
- ▶ No Emergency Power Off (EPO) switch.
- ▶ No balanced power option or plan ahead power cables (installation of extar PDUs and PCIe+ I/O drawers is nondisruptive if enough PCIe fanouts are available in the CPC drawer).
- ▶ No Internal Battery Feature (IBF). The standard data center power protection (uninterruptible power supply based) can be used.
- ▶ If the 16U Reserved feature (FC 0617) is installed, the total power consumption for equipment that is installed under this feature is limited to 3400W. For more information, see Appendix G, “16U Reserved feature” on page 465.

The following previous features are not available on the ZR1:

- ▶ Internal Battery Feature (IBF)
- ▶ 480V AC
- ▶ High-Voltage DC Power
- ▶ Balanced Power, no plan ahead power cords
- ▶ external EPO (Emergency Power Off) switch
- ▶ Three-Phase Power
- ▶ Single SE Display and keyboard, which includes a KVM switch

2.8.2 Power and weight estimation tool

By using the power and weight estimation tool for the z14 ZR1 server, you can enter your precise server configuration to obtain an *estimate* of power consumption. Log in to the Resource link with your user ID. Click **Planning** → **Tools** → **Power and weight estimation Tools**. Specify your server and 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.

The exact power consumption for your system varies. The object of the tool is to estimate the power and weight 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 and the CPC details panel (Energy Management tab), as shown in Figure 2-24.

Instance Information	Acceptable Status	Product Information	Network Information	STP Information	Energy Management	Security
CPC						
Power rating: 9600 W						
Power consumption: 2797 W						
Power saving: Not supported						
Power save profile: Custom energy management						
Power capping: Not supported						
zCPC						
Power rating: 9600 W						
Power consumption: 2797 W						
Ambient temperature: 27.5°C (81.5°F)						
Exhaust temperature: 35.0°C (95.0°F)						
Humidity: 13 %						
Dew point: .0°C (32.0°F)						
Heat load: 9550 BTU/hr.						
Heat load (forced-air): 9550 BTU/hr.						
Heat load (water): 0 BTU/hr.						
Maximum potential power: 4939 W						
Maximum potential heat load: 16864 BTU/hr.						
Power saving: High performance						
Power capping: Not supported						

OK Apply Change Options... Cancel Help

Figure 2-24 CPC Details Panel: Energy Management tab

2.8.3 Cooling requirements

The z14 ZR1 is an air-cooled system. With the new design, no cables are used in the front of the system. This configuration results in better air flow. Air flow is front-to-rear (front is cold air input, rear is warm air exhaust). Chilled air, ideally coming from under a raised floor, is required to fulfill the cooling requirements. The chilled air is often provided through perforated floor tiles.

For more information about the amount of chilled air that is required for various temperatures under the floor of the computer room, see *IBM 3907 Installation Manual for Physical Planning*, GC28-6974.

The z14 ZR1 is classified as ASHRAE class A3 compliant.

If the 16U Reserved feature is present, the non-Z equipment that is installed in the space that is provided by the feature must conform with the specifications that are described in *IBM 3907 Installation Manual for Physical Planning*, GC28-6974.

The 16U Reserved feature includes the following basic requirements:

- ▶ Front-to-rear airflow
- ▶ No cables in front of the rack
- ▶ Visible conformity labels (according to country of installation requirements)
- ▶ Unused rack space must be covered with fillers to provide proper air flow

2.9 Summary

All aspects of the z14 ZR1 structure are listed in Table 2-12.

Table 2-12 System structure summary

Description	Max4	Max12	Max24	Max30
Number of CPC drawers	1	1	1	1
Number of SCMs	2	3	5	5
Total number of PU SCMs	1	2	4	4
Total number of SC SCMs	1	1	1	1
Total number of PUs	8	16	28	34
Maximum number of characterized PUs	4	12	24	30
Number of CPs	0 - 4	0 - 6	0 - 6	0 - 6
Number of IFLs	0 - 4	0 - 12	0 - 24	0 - 30
Number of ICFs	0 - 4	0 - 12	0 - 24	0 - 30
Number of zIIPs	0 - 2	0 - 8	0 - 12	0 - 12
Standard SAPs	2	2	2	2
Additional SAPs	0 - 2	0 - 2	0 - 2	0 - 2
Number of IFP	1	1	1	1
Standard spare PUs	1	1	1	1
Enabled memory sizes GB	64 - 1984	64 - 4032	64 - 8182	64 - 8182
L1 cache per PU (I/D)	128/128 KB	128/128 KB	128/128 KB	128/128 KB
L2 cache per PU	2/4 MB (I/D)	2/4 MB (I/D)	2/4 MB (I/D)	2/4 MB (I/D)
L3 shared cache per PU SCM	128 MB	128 MB	128 MB	128 MB
L4 shared cache per node	672 MB	672 MB	672 MB	672 MB
Cycle time (ns)	0.888	0.888	0.888	0.888
Clock frequency	4.5 GHz	4.5 GHz	4.5 GHz	4.5 GHz

Description	Max4	Max12	Max24	Max30
Maximum number of PCIe fanouts	2	4	8	8
I/O interface per PCIe cable	16 GBps	16 GBps	16 GBps	16 GBps
Number of support elements	2	2	2	2
External AC power	Single-phase	Single-phase	Single-phase	Single-phase



Central processor complex system design

This chapter describes the design of the IBM z14 ZR1 processor. By understanding the processor design, users become familiar with the functions that make the z14 ZR1 server a system that accommodates a broad mix of workloads for enterprises of all sizes.

This chapter includes the following topics:

- ▶ 3.1, “Overview” on page 64
- ▶ 3.2, “Design highlights” on page 64
- ▶ 3.3, “CPC drawer design” on page 66
- ▶ 3.4, “Processor unit design” on page 70
- ▶ 3.5, “Processor unit functions” on page 83
- ▶ 3.6, “Memory design” on page 94
- ▶ 3.7, “Logical partitioning” on page 97
- ▶ 3.8, “Intelligent Resource Director” on page 107
- ▶ 3.9, “Clustering technology” on page 109
- ▶ 3.10, “Virtual Flash Memory” on page 115

3.1 Overview

The z14 ZR1 symmetric multiprocessor (SMP) system is the next step in an evolutionary trajectory that began with the introduction of the IBM System/360 in 1964. Over time, the design was adapted to the changing requirements that were dictated by the shift toward new types of applications on which clients depend.

z14 ZR1 servers offer high levels of performance, reliability, availability, serviceability (RAS), resilience, and security. The z14 ZR1 server fits into the IBM strategy in which mainframes play a central role in creating an infrastructure for cloud, analytics, and mobile, which is underpinned by security. The z14 ZR1 server is designed so that everything around it, such as operating systems, middleware, storage, security, and network technologies that support open standards, helps you achieve your business goals.

For more information about the z14 ZR1 RAS features, see Chapter 9, “Reliability, availability, and serviceability” on page 327.

The z14 ZR1 processor includes the following features:

- ▶ Ultra-high frequency, large, high-speed buffers (caches) and memory
- ▶ Superscalar processor design
- ▶ Out-of-order core execution
- ▶ Simultaneous multithreading (SMT)
- ▶ Single-instruction multiple-data (SIMD)
- ▶ Flexible configuration options

The z14 ZR1 processor is the next implementation of IBM Z servers to address the ever-changing IT environment.

3.2 Design highlights

The physical packaging of z14 ZR1 server is new, with one CPC drawer that fits the 19-inch form factor rack (frame). The CPC drawer has a modular design that uses higher density single chip modules (SCM) for processors and system controller. Higher chip density addresses thermal design complexity that is related to building systems with ever-increasing capacities. The modular CPC drawer design is flexible and expandable, which offers unprecedented capacity to meet consolidation needs.

The microprocessor of the z14 ZR1 uses the same design as z14 M0x. The difference is the frequency: 4.5G Hz for z14 ZR1 versus 5.2 GHz for z14 M0x.

The Processor Unit (PU) SCMs are air-cooled (versus water-cooled for z14 M0x) because the lower frequency reduces the generated heat.

z14 ZR1 servers continue the line of mainframe processors that are compatible with an earlier version. The current evolution brings the following processor design enhancements:

- ▶ The processor chip is designed with 10 cores, with 5, 6, 7, 8, or 9 active cores
- ▶ Pipeline optimization
- ▶ Improved SMT and SIMD
- ▶ Better branch prediction
- ▶ Improved co-processor functionality

The z14 ZR1 processor uses 24-bit, 31-bit, and 64-bit addressing modes, multiple arithmetic formats, and multiple address spaces for robust interprocess security.

The z14 ZR1 system design has the following main objectives:

- ▶ Offer a data-centric approach to information (data) security that is simple, transparent, and consumable (extensive data encryption from inception to archive, in flight and at rest).
- ▶ Offer a *flexible infrastructure* to concurrently accommodate a wide range of operating systems and applications, from the traditional systems (for example, z/OS and z/VM) to the world of Linux, cloud, analytics, and mobile computing.
- ▶ Offer state-of-the-art *integration* capability for server consolidation by using virtualization capabilities in a highly *secure environment*:
 - Logical partitioning, which allows 40 independent logical servers (logical partitions).
 - z/VM, which can virtualize hundreds to thousands of servers as independently running virtual machines (guests).
 - HiperSockets, which implement virtual LANs between logical partitions (LPARs) within the system.
 - Efficient data transfer that uses direct memory access (SMC-D), Remote Direct Memory Access (SMC-R), and reduced storage access latency for transactional environments - zHyperLink Express.
 - The IBM Z PR/SM is designed for Common Criteria Evaluation Assurance Level 5+ (EAL 5+) certification for security, so an application that is running on one partition (LPAR) cannot access another application on a different partition (essentially the same security as an air-gapped separated system).

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

- ▶ Offer *high-performance computing* to achieve the outstanding response times that are required by new workload-type applications. This performance is achieved by high-frequency, enhanced superscalar processor technology, out-of-order core execution, large high-speed buffers (cache) and memory, an architecture with multiple complex instructions, and high-bandwidth channels.
- ▶ Offer the *processing capacity* and *scalability* that are required by the most demanding applications, from the single-system and clustered-systems points of view.
- ▶ Offer the capability of *concurrent upgrades* for processors, memory, and I/O connectivity, which prevents system outages in planned situations.
- ▶ Implement a system with *high availability* and *reliability*. These goals are achieved with redundancy of critical elements and sparing components of a single system, and the clustering technology of the Parallel Sysplex environment.
- ▶ Include 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) includes a dedicated and optimized CP Assist for Cryptographic Function (CPACF).
- ▶ Optional Crypto Express features with cryptographic coprocessors provide the highest standardized security certification.¹ These optional features can also be configured as Cryptographic Accelerators to enhance the performance of Secure Sockets Layer/Transport Layer Security (SSL/TLS) transactions.
- ▶ Be *self-managing* and *self-optimizing* by adjusting itself when the workload changes to achieve the best system throughput. This process can be done by using the Intelligent Resource Director or the Workload Manager functions, which are assisted by HiperDispatch.

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

- Have a *balanced system* design with pervasive encryption, which provides large data rate bandwidths for high-performance connectivity along with processor and system capacity, while protecting every byte that enters and exits the z14 ZR1.

The remaining sections describe the z14 ZR1 system structure, showing a logical representation of the data flow from PUs, caches, memory cards, and various interconnect capabilities.

3.3 CPC drawer design

A z14 ZR1 system has one CPC drawer, with up to 34 PUs that can be characterized for customer use, and up to 8128 GB of customer usable memory capacity. The CPC drawer is logically divided in two clusters to improve the processor and memory affinity and availability.

The following types of CPC drawer² configurations are available for z14 ZR1 system:

- One PU SCM (8 PUs), 2 PCIe Fanouts, up to 2 TB memory
- Two PU SCMs (16 PUs), 4 PCIe Fanouts, up to 4 TB memory
- Four PU SCMs (28 PUs), 8 PCIe Fanouts, up to 8 TB memory
- Four PU SCMs (34 PUs), 8 PCIe Fanouts, up to 8 TB memory

Table 3-1 z14 ZR1 features (PU and memory)

Feature	CP max	IFL max	I/O fanouts	Memory
Max30 (FC 0639)	6	30	8	64GB-8TB
Max24 (FC 0638)	6	24	8	64GB-8TB
Max12 (FC 0637)	6	12	4	64GB-4TB
Max4 (FC 0636)	4	4	2	64GB-2TB

The z14 ZR1 has up to four memory controller units (MCUs). The configuration uses five-channel redundant array of independent memory (RAIM) protection, with dual inline memory modules (DIMM) bus cyclic redundancy check (CRC) error retry.

The cache hierarchy (L1, L2, L3, and L4) is implemented with embedded dynamic random access memory (eDRAM) caches. Until recently, eDRAM was considered to be too slow for this use. However, a breakthrough in technology that was made by IBM eliminated that limitation. In addition, eDRAM offers higher density, less power utilization, fewer soft errors, and better performance.

z14 ZR1 servers use CMOS Silicon-on-Insulator (SOI) 14 nm chip technology, with advanced low latency pipeline design, which creates high-speed yet power-efficient circuit designs. The PU SCM has 17 layers of metal. For more information, see 2.8.1, “Considerations” on page 58.

² The CPC drawer always includes one System Controller Single Chip Module (SC SCM).

3.3.1 Cache levels and memory structure

The z14 ZR1 memory subsystem focuses on keeping data “closer” to the PU core. With the current processor configuration, all on chip cache levels increased.

Although L1, L2, and L3 caches are implemented on the PU SCM, the fourth cache level (L4) is implemented within the system controller (SC) SCM. One L4 cache is present in each CPC drawer, which is shared by all PU SCMs. The cache structure of the z14 ZR1 has the following characteristics:

- ▶ Larger L1, L2, and L3 caches (more data closer to the core).
- ▶ L1 and L2 caches use eDRAM, and are private for each PU core.
- ▶ L2-L3 interface has a new *Fetch cancel* protocol, a revised *L2 Least Recent Used* (LRU) demote handling.
- ▶ L3 cache also uses eDRAM and is shared by all activated cores within the PU chip. The CPC drawer has up to four L3 caches, depending on CPC drawer feature. Therefore, a Max24 and Max30 CPC drawer feature four L3 caches, which results in 512 MB (4 x 128 MB) of this shared PU chip-level cache. For availability and reliability, L3 cache now implements symbol ECC.
- ▶ L4 cache also uses eDRAM, and is shared by all PU chips. L4 cache has 672 MB inclusive of L3's, 42w Set Associative and 256 bytes cache line size.

In most real-world situations, several cache lines exist in multiple L3s underneath L4. The L4 does not contain the same line multiple times, but rather once with an indication of all the cores that have a copy of that line. As such, 672 MB of inclusive L4 can easily cover 512 MB of underlying L3 caches.

- ▶ Main storage has up to 8 TB addressable memory in the CPC drawer, which uses 20 DIMMs.

Considerations

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

Also, the distance to remote caches as seen from the microprocessor becomes a significant factor. Although the L4 cache is rather large, several cycles are needed to travel the distance to the cache. The node-cache topology of z14 ZR1 servers is shown in Figure 3-1.

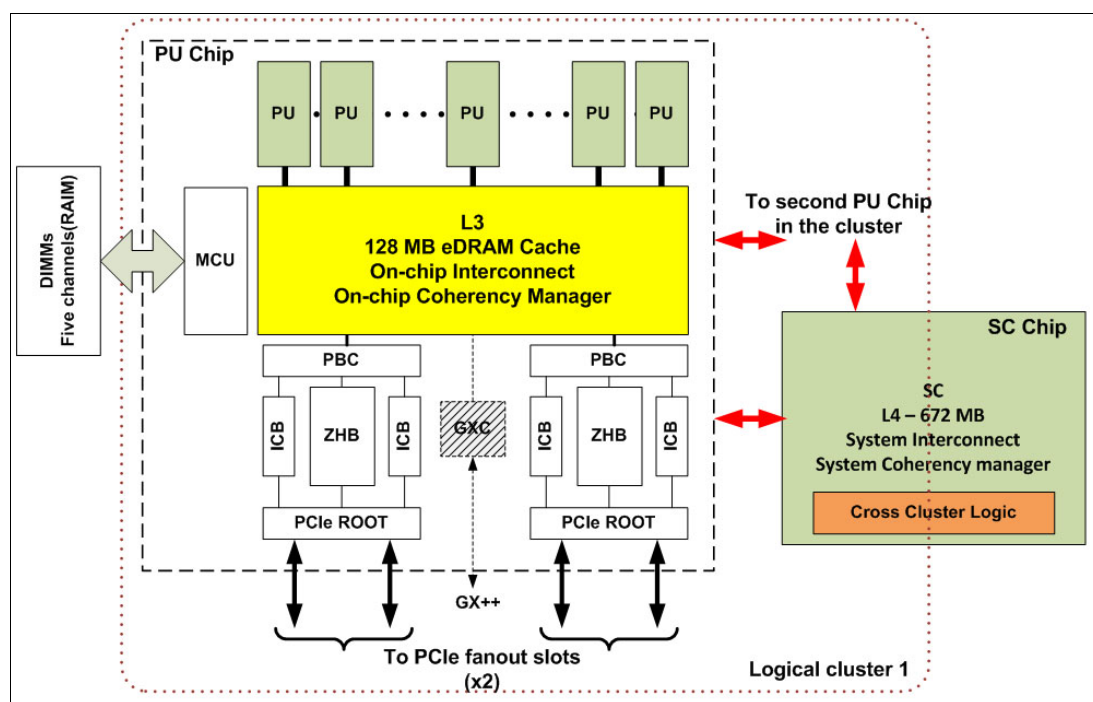


Figure 3-1 z14 ZR1 cache topology

Although large caches mean increased access latency, the new technology of CMOS 14S0 (14 nm chip lithography) and the lower cycle time allows z14 ZR1 servers to increase the size of cache levels (L1, L2, and L3) within the PU chip by using denser packaging. This design reduces traffic to and from the shared L4 cache, which is on another chip (SC chip).

Only when a cache miss occurs in L1, L2, or L3 is a request sent to L4. L4 is the coherence manager, which means that all memory fetches must be in the L4 cache before that data can be used by the processor. However, in the z14 ZR1 cache design, some lines of the L3 cache are not included in the L4 cache.

The cache structure of z14 ZR1 servers is compared with the previous generation of IBM Z servers (z13s) in Figure 3-2.

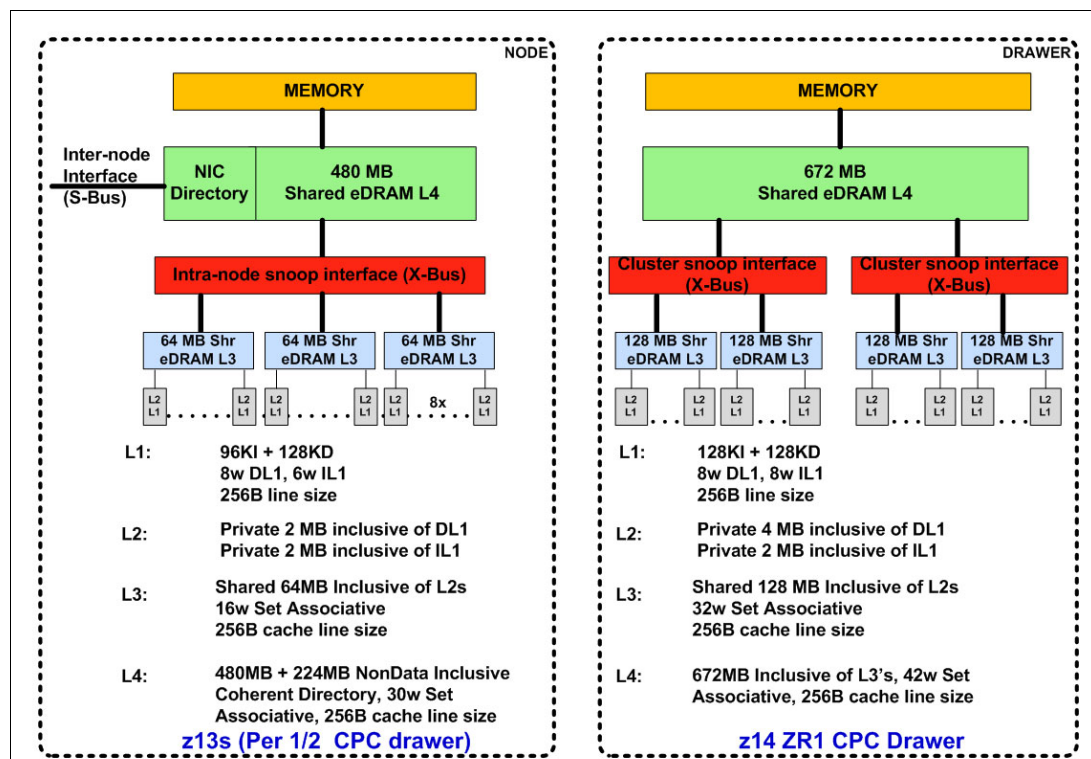


Figure 3-2 z14 ZR1 and z13s cache levels comparison

Compared to z13s, the z14 ZR1 cache design has larger L1, L2, and L3 cache sizes. In z14 ZR1 servers, more affinity exists between the memory of a partition, the L4 cache in the SC (which is accessed by the two logical clusters in the same CPC drawer), and the cores in the PU.

The access time of the private cache often occurs in one cycle. The z14 ZR1 cache level structure is focused on keeping more data closer to the PU. This design can improve system performance on many production workloads.

HiperDispatch

To help avoid latency in a high-frequency processor design, PR/SM and the dispatcher must be prevented from scheduling and dispatching a workload on *any* processor available, which keeps the workload in as small a portion of the system as possible. The cooperation between z/OS and PR/SM is bundled in a function called *HiperDispatch*. HiperDispatch uses the z14 ZR1 cache topology, which features reduced cross-cluster “help” and better locality for multi-task address spaces.

PR/SM can use dynamic PU reassignment to move processors (CPs, ZIIPs, IFLs, ICFs, SAPs, and spares) to a different chip to improve the reuse of shared caches by processors of the same partition. For more information about HiperDispatch, see 3.7, “Logical partitioning” on page 97.

3.3.2 CPC drawer topology

The z14 ZR1 CPC drawer topology with the interconnection between CP and SC is shown in Figure 1-3 on page 7. The SC regulates coherent cluster-to-cluster traffic.

3.4 Processor unit design

Processor cycle time is especially important for processor-intensive applications. Current systems design is driven by processor cycle time, although improved cycle time does not automatically mean that the performance characteristics of the system improve.

Through innovative processor design (pipeline and cache management redesigns), the IBM Z processor performance continues to evolve. With the introduction of out-of-order execution, ever improving branch prediction mechanism, and simultaneous multi-threading, the processing performance was enhanced beyond the slight frequency increase (z13s core runs at 4.3 GHz).

z14 ZR1 core frequency is 4.5 GHz, which allows the increased number of processors that share larger caches to have quick processing times for improved capacity and performance. Although the cycle time of the z14 ZR1 processor frequency was only slightly increased (4%) compared to z13s, the processor performance was increased even further (z14 ZR1 uni-processor PCI up 10% compared to z13s) through improved processor design, such as pipeline enhancements, out-of-order execution design, branch prediction, time of access to high-speed buffers (caches), and the relative nest intensity (RNI) redesigns. For more information about RNI, see 12.4, “Relative Nest Intensity” on page 401.

z13s servers introduced architectural extensions with instructions that reduce processor quiesce effects, cache misses, and pipeline disruption, and increase parallelism with instructions that process several operands in a single instruction (SIMD). The processor architecture was further developed for z14 ZR1 and includes the following features:

- ▶ Optimized second-generation SMT
- ▶ Enhanced SIMD instructions set
- ▶ Improved Out-of-Order core execution
- ▶ Improvements in branch prediction and handling
- ▶ Pipeline optimization
- ▶ Enhanced branch prediction structure and sequential instruction fetching

The z14 ZR1 enhanced Instruction Set Architecture (ISA) includes a set of instructions that is added to improve compiled code efficiency. These instructions optimize PUs to meet the demands of various business and analytics workload types without compromising the performance characteristics of traditional workloads.

3.4.1 Simultaneous multithreading

Aligned with industry directions, z14 ZR1 servers can process up to two simultaneous threads in a single core while sharing certain resources of the processor, such as execution units, translation lookaside buffers (TLBs), and caches. When one thread in the core is waiting for other hardware resources, the second thread in the core can use the shared resources rather than remaining idle. This capability is known as *simultaneous multithreading* (SMT).

SMT is supported only for Integrated Facility for Linux (IFL) and IBM Z Integrated Information Processor (zIIP) speciality engines on z14³ ZR1 servers, and it requires operating system support.

An operating system with SMT support can be configured to dispatch work to a thread on a zIIP (for eligible workloads in z/OS) or an IFL (for z/VM and Linux on Z) core in single thread or SMT mode so that HiperDispatch cache optimization can be considered. For more information about operating system support, see Chapter 7, “Operating system support” on page 209.

SMT technology allows instructions from more than one thread to run in any pipeline stage at a time. SMT can handle up to four pending translations.

Each thread has its own unique state information, such as Program Status Word - S/360 Architecture (PSW) and registers. The simultaneous threads cannot necessarily run instructions instantly and must at times compete to use certain core resources that are shared between the threads. In some cases, threads can use shared resources that are not experiencing competition.

Two threads (A and B) that are running on the same processor core on different pipeline stages and sharing the core resources are shown in Figure 3-3.

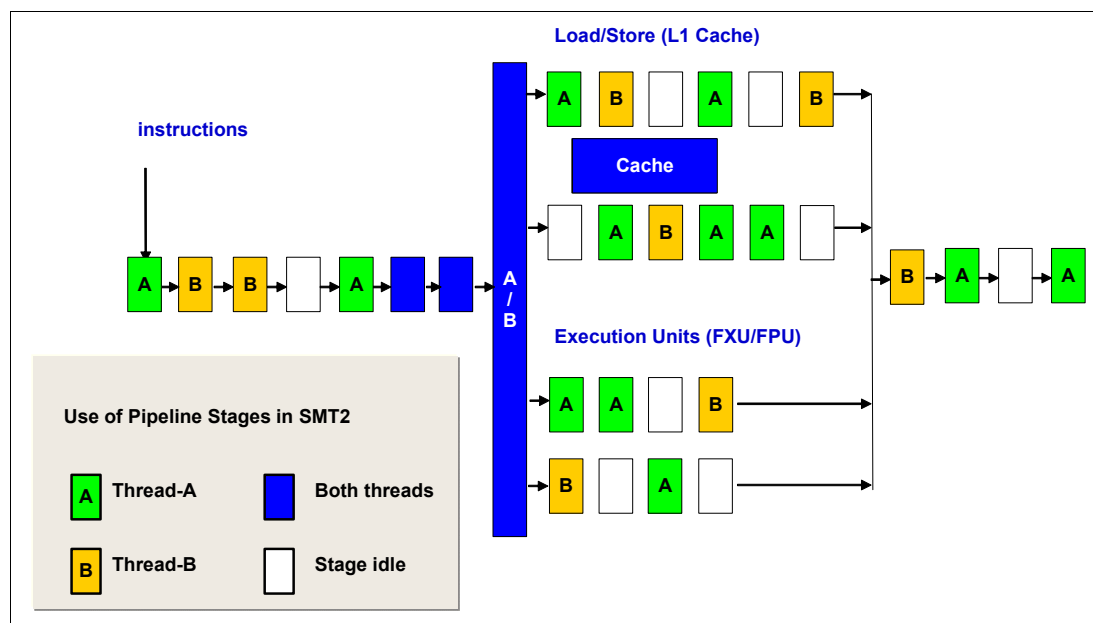


Figure 3-3 Two threads running simultaneously on the same processor core

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

Although SMT increases the processing capacity, the performance in some cases might be superior if a single thread is used. Enhanced hardware monitoring supports measurement through CPUMF for thread usage and capacity.

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

³ In addition to optional SMT support for zIIPs and IFLs, z14 introduced SMT as default for SAPs (not user controllable).

SMT use is functionally transparent to middleware and applications. No changes are required to run them in an SMT-enabled partition.

3.4.2 Single-instruction multiple-data (enhanced for z14 ZR1)

The z14 ZR1 superscalar processor has 32 vector registers and an instruction set architecture that includes a subset of 139 new instructions (known as SIMD) that were added to improve the efficiency of complex mathematical models and vector processing. These new instructions allow many operands to be processed with a single instruction. The SIMD instructions use the superscalar core to process operands in parallel.

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

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

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

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

A schematic representation of a SIMD instruction with 16-byte size elements in each vector operand is shown in Figure 3-4.

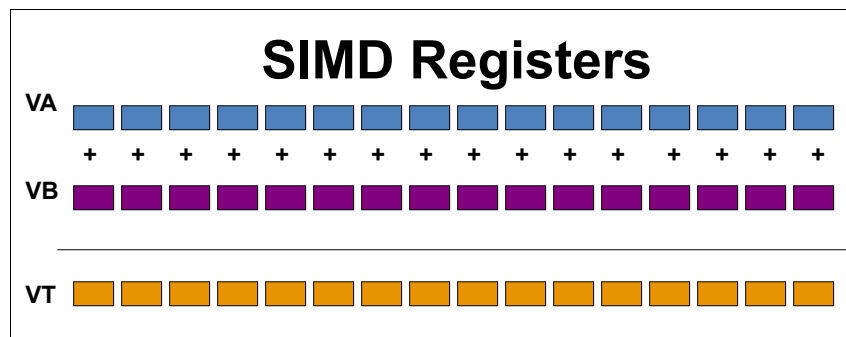


Figure 3-4 Schematic representation of add SIMD instruction with 16 elements in each vector

The vector register file overlays the floating-point registers (FPRs), as shown in Figure 3-5. The FPRs use the first 64 bits of the first 16 vector registers, which saves hardware area and power, and makes it easier to mix scalar and SIMD codes. Effectively, the core gets 64 FPRs, which can further improve FP code efficiency.

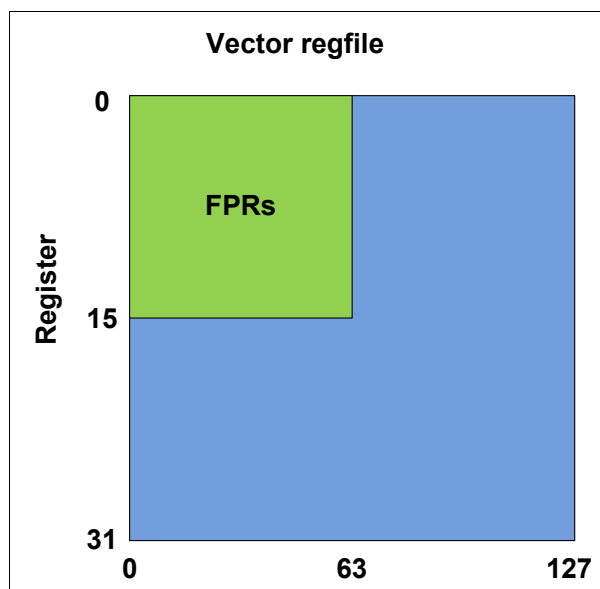


Figure 3-5 Floating point registers overlaid by vector registers

SIMD instructions include the following examples:

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

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

z14 ZR1 SIMD features the following enhancements (compared to z13s):

- ▶ Doubled vector double precision Binary Floating Point (BFP) operations throughput (2x 64b)
- ▶ Added vector single precision BFP (4x 32b)
- ▶ Added vector quad precision BFP (128b)
- ▶ Added binary Fixed Multiply Add (FMA) operations to speed up code
- ▶ Vector Single Precision/ Double Precision/ Quad Precision (SP/DP/QP) compare/min/max with programming language support
- ▶ Enhanced to Storage-to-Storage Binary Coded Decimal (BCD)
- ▶ Vector load/store right-most with length

3.4.3 Out-of-Order execution

z14 ZR1 servers have an Out-of-Order core, much like the z13s. Out-of-Order yields significant performance benefits for compute-intensive applications. It does so by reordering instruction execution, which allows later (younger) instructions to be run ahead of a stalled instruction, and reordering storage accesses and parallel storage accesses. Out-of-Order maintains good performance growth for traditional applications.

Out-of-Order execution can improve performance in the following ways:

- ▶ Reordering instruction execution

Instructions stall in a pipeline because they are waiting for results from a previous instruction or the execution resource that they require is busy. In an in-order core, this stalled instruction stalls all later instructions in the code stream. In an out-of-order core, later instructions are allowed to run ahead of the stalled instruction.

- ▶ Reordering storage accesses

Instructions that access storage can stall because they are waiting on results that are needed to compute the storage address. In an in-order core, later instructions are stalled. In an out-of-order core, later storage-accessing instructions that can compute their storage address are allowed to run.

- ▶ Hiding storage access latency

Many instructions access data from storage. Storage accesses can miss the L1 and require 7 - 50 more clock cycles to retrieve the storage data. In an in-order core, later instructions in the code stream are stalled. In an out-of-order core, later instructions that are not dependent on this storage data are allowed to run.

The z14 ZR1 processor includes pipeline enhancements that benefit Out-of-Order execution. The IBM Z processor design features advanced micro-architectural innovations that provide the following benefits:

- ▶ Maximized instruction-level parallelism (ILP) for a better cycles per instruction (CPI) design.
- ▶ Maximized performance per watt. Two cores are added (as compared to the z13/z13s chip) at only slightly higher chip power.
- ▶ 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/TLB miss penalty.
- ▶ Improved execution of D-Cache store and reload and new Fixed-point divide.
- ▶ New Operand Store Compare (OSC) (load-hit-store conflict) avoidance scheme.
- ▶ Enhanced branch prediction structure and sequential instruction fetching.

Program results

The Out-of-Order execution does not change any program results. Execution can occur out of (program) order, but all program dependencies are honored. The same results occur as 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 z14 ZR1 core is shown in Figure 3-6 on page 75.

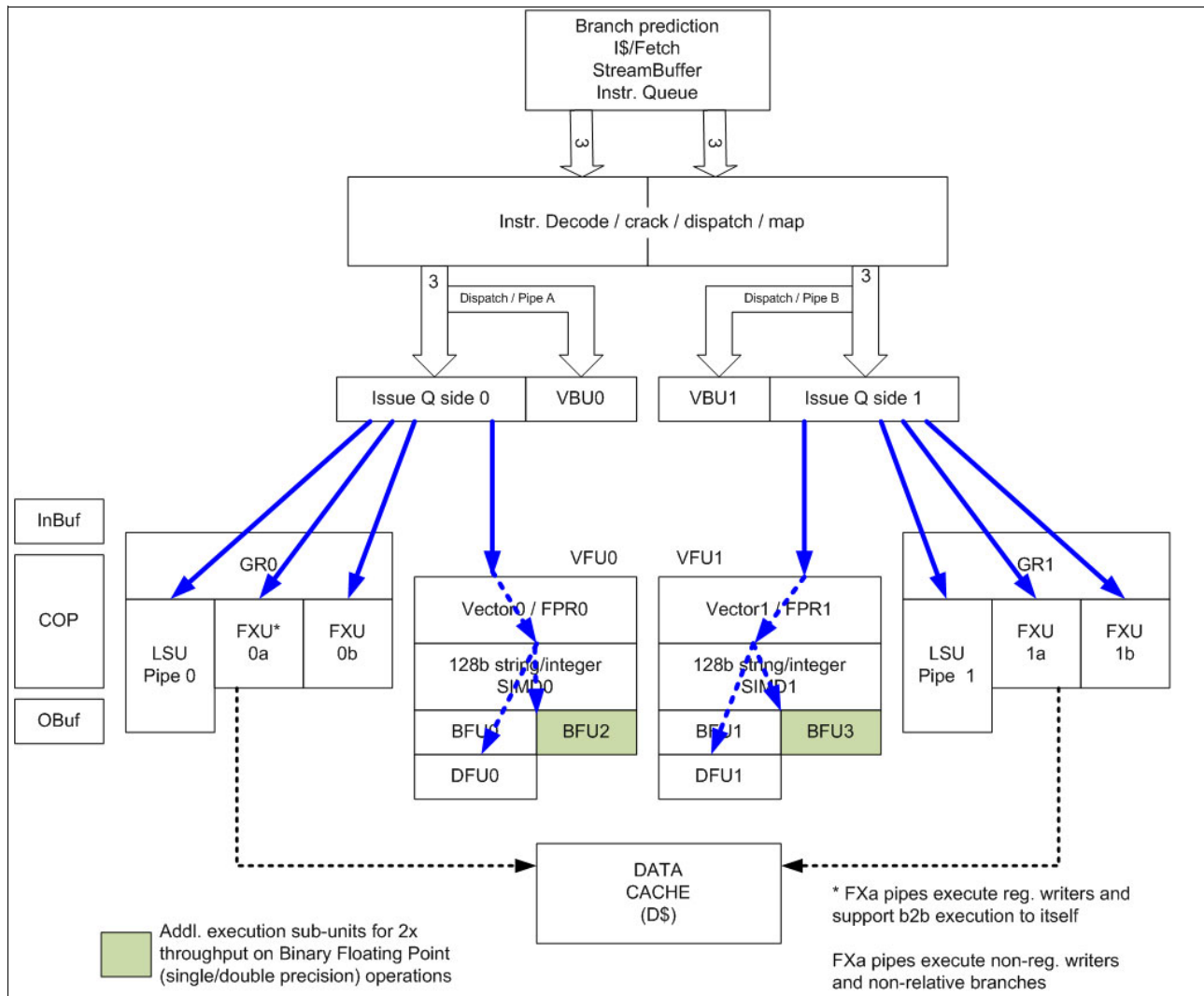


Figure 3-6 z14 ZR1 PU core logical diagram

Memory address generation and memory accesses can occur out of (program) order. This capability can provide a greater use of the z14 ZR1 superscalar core, and can improve system performance.

The z14 ZR1 processor unit core is a superscalar, out-of-order, SMT processor with 10 execution units. Up to six instructions can be decoded per cycle, and up to 10 instructions or operations can be started to run per clock cycle.

The execution of the instructions can occur out of program order. 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. This technology results in shorter workload runtime.

Branch prediction

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

For this reason, various history-based branch prediction mechanisms are used, as shown on the in-order part of the z14 PU core logical diagram in Figure 3-6 on page 75. The branch target buffer (BTB) runs ahead of instruction cache pre-fetches to prevent branch misses in an early stage. Also, a branch history table (BHT), in combination with a pattern history table (PHT) and the use of tagged multi-target prediction technology branch prediction, offers a high branch prediction success rate.

The z14 ZR1 microprocessor improves the branch prediction throughput by using the new branch prediction and instruction fetch front end.

3.4.4 Superscalar processor

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

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

On z14 ZR1 servers, up to six instructions can be decoded per cycle and up to 10 instructions or operations can be in execution per cycle. Execution can occur out of (program) order. These improvements also make possible the simultaneous execution of two threads in the same processor.

Many challenges exist in creating an efficient superscalar processor. The superscalar design of the PU made significant strides in avoiding address generation interlock 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.

3.4.5 Compression and cryptography accelerators on a chip

This section introduces the CPACF enhancements for z14 ZR1.

Coprocessor units

One coprocessor unit is available for compression and cryptography on each core in the chip. The compression engine uses static dictionary compression and expansion. The compression dictionary uses the L1-cache (instruction cache).

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

The coprocessors feature the following characteristics:

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

The location of the coprocessor on the chip is shown in Figure 3-7.

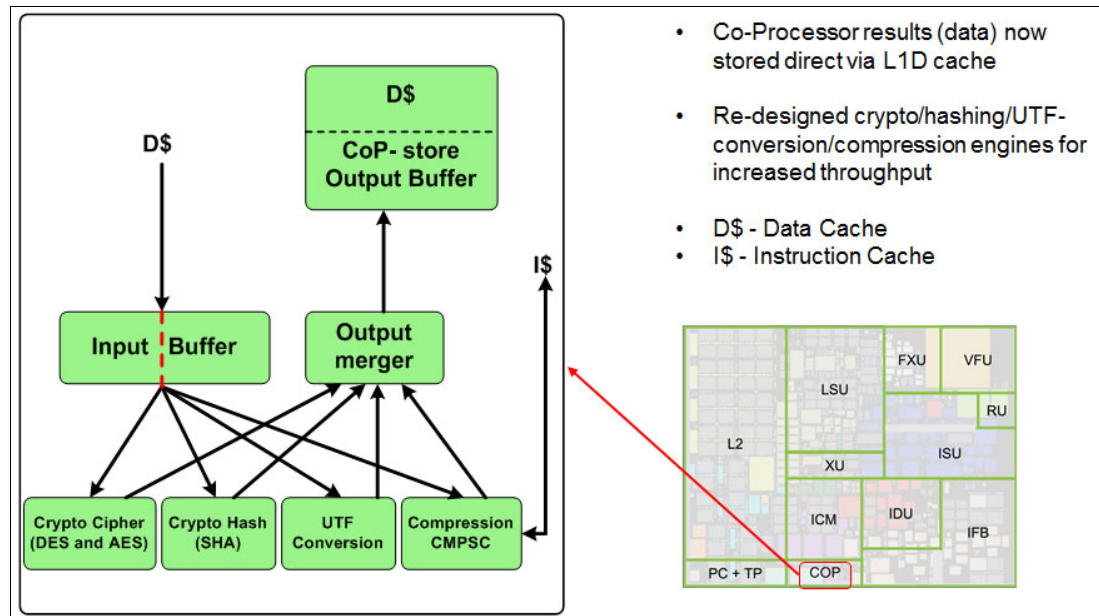


Figure 3-7 Compression and cryptography accelerators on a core in the chip

Compression enhancements

The compression features the following enhancements:

- ▶ Huffman compression on top of CMPSC compression (embedded in dictionary, reuse of generators)
- ▶ Order Preserving compression in B-Trees and other index structures
- ▶ Faster expansion algorithms
- ▶ Reduced overhead on short data

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 FIPS 140-2 level 4 security. The assist function uses a special instruction set for symmetrical clear key cryptographic encryption and decryption, and for hash operations. This group of instructions is known as the Message-Security Assist (MSA). For more information about these instructions, see *z/Architecture Principles of Operation*, SA22-7832.

Crypto functions enhancements

The crypto functions include the following enhancements:

- ▶ Reduced overhead on short data (hashing and encryption)
- ▶ 4x throughput for AES
- ▶ Special instructions for elliptic curve crypto/RSA
- ▶ New hashing algorithms; for example, SHA-3
- ▶ Support for authenticated encryption (combined encryption and hashing; for example, AES-GCM)
- ▶ True random number generator (for example, for session keys)

For more information about cryptographic functions on z14 ZR1 servers, see Chapter 6, “Cryptographic features” on page 173.

3.4.6 Decimal floating point accelerator

The decimal floating point (DFP) accelerator function is present on each of the microprocessors (cores) on the 10-core chip. Its implementation meets business application requirements for better performance, precision, and function.

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

Hardware DFP computational instructions provide the following features:

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

Benefits of the DFP accelerator

The DFP accelerator offers the following benefits:

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

z14 ZR1 servers have two DFP accelerator units per core, which improve the decimal floating point execution bandwidth. The floating point instructions operate on newly designed vector registers (32 new 128-bit registers).

z14 ZR1 servers include new decimal floating point-packed conversion facility support with the following benefits:

- ▶ Reduces code path length because extra instructions to format conversion are no longer needed.
- ▶ Operates packed data in memory by all decimal instructions without general-purpose registers, which were required only to prepare for decimal floating point packed conversion instruction.
- ▶ Converting from packed can now force the input packed value to positive instead of requiring a separate OI, OILL, or load positive instruction.
- ▶ Converting to packed can now force a positive zero result instead of requiring ZAP instruction.

Software support

DFP is supported in the following programming languages and products:

- ▶ Release 4 and later of the High Level Assembler
- ▶ C/C++, which 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 that use the BigDecimal Class Library
- ▶ SQL support as of Db2 Version 9 and later

3.4.7 IEEE floating point

Binary and hexadecimal floating-point instructions are implemented in z14 ZR1 servers. They incorporate IEEE standards into the system.

The z14 ZR1 core implements two other execution subunits for 2x throughput on BFP (single/double precision) operations (see Figure 3-6 on page 75).

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

3.4.8 Processor error detection and recovery

The PU core 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 core. Relocation under hardware control is possible because the R-unit has the full designed state in its buffer. PU error detection and recovery are shown in Figure 3-8.

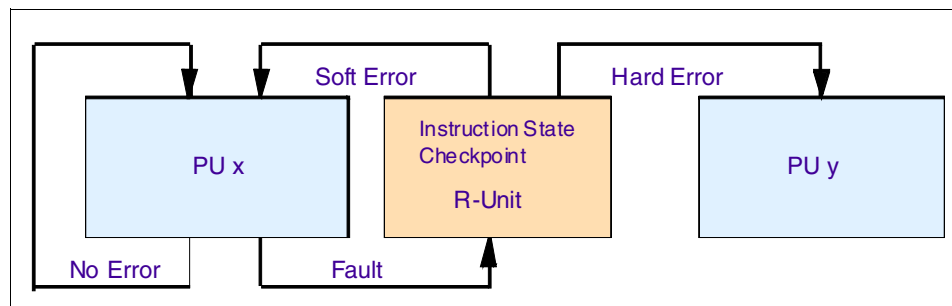


Figure 3-8 PU core error detection and recovery

3.4.9 Branch prediction

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

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

The BHT offers significant branch performance benefits. The BHT allows each PU core to take instruction branches that are based on a stored BHT, which improves processing times for calculation routines. In addition to the BHT, z14 ZR1 servers use the following techniques to improve the prediction of the correct branch to be run:

- ▶ BTB
- ▶ PHT
- ▶ BTB data compression

The success rate of branch prediction contributes significantly to the superscalar aspects of z14 ZR1 servers. This success is because the architecture rules prescribe that, for successful parallel execution of an instruction stream, the correctly predicted result of the branch is essential.

The z14 ZR1 branch prediction includes the following enhancements over z13s:

- ▶ Branch prediction search pipeline extended from five to six cycles to accommodate new predictors for increased accuracy/performance.
- ▶ New predictors:
 - Perceptron (neural network direction predictor)
 - SSCRS (hardware-based super simple call-return stack)
- ▶ Capacity increases:
 - Level 1 Branch Target Buffer (BTB1): 1 K rows x 6 sets → 2 K rows x 4 sets
 - Level 2 Branch Target Buffer (BTB2): 16 K rows x 6 sets → 32 K rows x 4 sets
- ▶ Better power efficiency: Several structures were redesigned to maintain their accuracy while less power is used through smart access algorithms.
- ▶ New static IBM IA® regions expanded from four to eight. To conserve space, prediction structures do not store full target addresses. Instead, they use the locality and limited ranges of “4gig regions” of virtual instruction addresses - IA(0:31).

3.4.10 Wild branch

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

With the wild branch hardware facility (named Breaking Event Address Register - BEAR), 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 can also collect data from that storage location. This approach decreases the number of debugging steps that are necessary when you want to know from where the branch came.

3.4.11 Translation lookaside buffer

The translation lookaside buffer (TLB) in the instruction and data L1 caches use a secondary TLB to enhance performance.

The size of the TLB is kept as small as possible because of its short access time requirements and hardware space limitations. Because memory sizes recently increased significantly as a result of the introduction of 64-bit addressing, a smaller working set is represented by the TLB.

To increase the working set representation in the TLB without enlarging the TLB, large (1 MB) page and giant page (2 GB) support is available and can be used when appropriate. For more information, see “Large page support” on page 95.

With the enhanced DAT-2 (EDAT-2) improvements, the IBM Z servers support 2 GB page frames.

z14 ZR1 TLB enhancements

IBM z14 Model ZR1 switches to a logical-tagged L1 directory and inline TLB2. Each L1 cache directory entry contains the virtual address and Address Space Control Element (ASCE) because it no longer must access TLB for L1 cache hit. TLB2 is accessed in parallel to L2, which saves significant latency compared to TLB1-miss.

The new translation engine allows up to four translations pending concurrently. Each translation step is ~2x faster, which helps level 2 guests.

3.4.12 Instruction fetching, decoding, and grouping

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

Instruction fetching

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

Instruction decoding

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

Instruction grouping

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

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 IBM Z compilers and JVMs are constantly updated to benefit from new instructions and advances in microprocessor designs.

3.4.13 Extended Translation Facility

The z/Architecture instruction set includes instructions in support of the Extended Translation Facility. They are used in data conversion operations for Unicode data, which causes 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 in which XML and SOAP technologies are used. The High Level Assembler supports the Extended Translation Facility instructions.

3.4.14 Instruction set extensions

The processor supports the following 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. The use of these instructions improves performance.
- ▶ Load instructions for handling unsigned halfwords, such as those unsigned halfwords that are used for Unicode.
- ▶ Cryptographic instructions, which are known as the MSA, offer the full complement of the AES, SHA-1, SHA-2, and DES algorithms. They also include functions for random number generation.
- ▶ Extended Translate Facility-3 instructions, which are enhanced to conform with the current Unicode 4.0 standard.
- ▶ Assist instructions that help eliminate hypervisor processor usage.
- ▶ SIMD instructions, which allow the parallel processing of multiple elements in a single instruction.

3.4.15 Transactional Execution

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

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

3.4.16 Runtime Instrumentation

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

3.5 Processor unit functions

The PU functions are described in this section.

3.5.1 Overview

All PUs on a z14 ZR1 server are physically identical. When the system is started, one integrated firmware processor (IFP) is allocated from the pool of PUs that is available for the entire system. The other PUs can be characterized to specific functions (CP, IFL, ICF, zIIP, or SAP).

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

Only characterized PUs include a designated function. Non-characterized PUs are considered spares. Order at least one CP, IFL, or ICF on a z14 ZR1 server.

This design brings outstanding flexibility to z14 ZR1 servers because any PU can assume any available characterization. The design also plays an essential role in system availability because PU characterization can be done dynamically, with no system outage.

For more information about software level support of functions and features, see Chapter 7, “Operating system support” on page 209.

Concurrent PU upgrades

Concurrent upgrades can be done by the LIC, which assigns a PU function to a previously non-characterized PU. The upgrade can be done concurrently through the following facilities:

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

If the PU SCMs in the CPC drawer have no available (unused) PUs, an upgrade results in a CPC feature upgrade (Max4 => Max12 ==> Max24 ==> Max30) which means the installation or replacement of PU SCMs (the maximum is four in the CPC drawer). This operation is not concurrent.

For more information about Capacity on Demand, see Chapter 8, “System upgrades” on page 281.

PU sparing

In the rare event of a PU failure, the failed PU’s characterization is dynamically and transparently reassigned to a spare PU. z14 ZR1 servers have one spare PU. PUs that are not characterized on a CPC configuration can also be used as extra spare PUs. For more information about PU sparing, see 3.5.9, “Sparing rules” on page 93.

PU pools

PUs that are defined as CPs, IFLs, ICFs, and zIIPs are grouped in their own pools from where they can be managed separately. This configuration significantly simplifies capacity planning and management for LPARs. The separation also affects weight management because CP and zIIP weights can be managed separately. For more information, see “PU weighting” on page 84.

All assigned PUs are grouped in the PU pool. These PUs are dispatched to online logical PUs. As an example, consider a z14 ZR1 server with 4 CPs, 2 IFLs, 2 zIIPs, and 1 ICF. This system has a PU pool of 9 PUs, called the *pool width*. Subdivision defines the following pools:

- ▶ A CP pool of four CPs
- ▶ An ICF pool of one ICF
- ▶ An IFL pool of two IFLs
- ▶ A zIIP pool of two zIIPs

PUs are placed in the pools in the following circumstances:

- ▶ When the system is PORed
- ▶ At the time of a concurrent upgrade
- ▶ As a result of adding 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 occurs as the result of the removal of a CBU. They are also removed through the On/Off CoD process and the conversion of a PU. When a dedicated LPAR is activated, its PUs are taken from the correct pools. This process is also the case when an LPAR logically configures a PU as on, if the width of the pool allows for it.

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

PU weighting

Because CPs, zIIPs, IFLs, and ICFs have their own pools from where they are dispatched, they can be given their own weights. For more information about PU pools and processing weights, see the *IBM Z Processor Resource/Systems Manager Planning Guide*, SB10-7169.

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 30 PUs can be characterized as CPs, depending on the configuration.

The z14 ZR1 server can be started in LPAR (PR/SM) mode or in Dynamic Partition Manager (DPM) mode. For more information, see Appendix E, “IBM Dynamic Partition Manager” on page 451.

CPs are defined as 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 adds function, reserved processors are no longer needed. Regardless of the CPC installed feature, an LPAR can have up to 170 logical CPs that are defined (the sum of active and reserved logical CPs). In practice, define no more CPs than the operating system supports.

All PUs that are characterized as CPs within a configuration are grouped into the CP pool. The CP pool can be seen on the Hardware Management Console (HMC) workplace. Any z/Architecture operating system, CFCCs, and appliances (Secure Service Container) can run on CPs that are assigned from the CP pool.

The z14 zR1 server can be configured with 26 distinct capacity settings for each CP (156 subcapacity settings). The capacity setting for one CP is listed in Table 3-2.

Table 3-2 Capacity settings for one CP

CP capacity	Feature code
CP-A	1069
CP-B	1070
CP-C	1071
CP-D	1072
CP-E	1073
CP-F	1074
CP-G	1075
CP-H	1076
CP-I	1077
CP-J	1078
CP-K	1079
CP-L	1080
CP-M	1081
CP-N	1082
CP-O	1083
CP-P	1084
CP-Q	1085
CP-R	1086
CP-S	1087
CP-T	1088
CP-U	1089
CP-V	1090
CP-W	1091
CP-X	1092
CP-Y	1093
CP-Z	1094

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 2.7.5, “Model capacity identifier” on page 53.

3.5.3 Integrated Facility for Linux

An IFL is a PU that can be used to run Linux on Z, Linux guests on z/VM operating systems, and Secure Service Container (SSC). Up to 30 PUs can be characterized as IFLs, depending on the configuration. IFLs can be dedicated to a Linux, z/VM, or Secure Service Container LPAR, or can be shared by multiple Linux guests, z/VM LPARs, or SSC that are running on the same z14 server. Only z/VM, Linux on Z operating systems, appliances that are running in a Secure Service Container LPAR, and designated software products can run on IFLs. IFLs are orderable by using FC 1064.

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 z14 ZR1 server. Software product license charges that are based on the model capacity identifier are not affected by the addition of IFLs.

Unassigned IFLs

An IFL that is purchased but not activated is registered as an unassigned IFL (FC 1068). When the system is later upgraded with another IFL, the system recognizes that an IFL was purchased and is present.

3.5.4 Internal Coupling Facility

An Internal Coupling Facility (ICF) is a PU that is used to run the CFCC for Parallel Sysplex environments. Within the sum of all unassigned PUs in the CPC drawer, up to 30 ICFs can be characterized, depending on CPC drawer feature. 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 1065.

ICFs exclusively run CFCC. ICFs do not change the model capacity identifier of the z14 server. Software product license charges that are based on the model capacity identifier are not affected by the addition of ICFs.

All ICFs within a configuration are grouped into the ICF pool. The ICF pool can be seen on the HMC workplace.

The ICFs can be used by coupling facility LPARs only. ICFs are 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 zBC12), the IBM 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 coupling facility (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.
 - The completion of a CF signal that was previously sent by the CF occurs (completion of a CF-to-CF duplexing signal).

- ▶ On the z/OS side:
 - CF signal is received by a shared-engine z/OS image (arrival of a List Notification signal).
 - An asynchronous CF operation completes.

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

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

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

CF processor combinations

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

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

Shared ICFs add flexibility. However, running only with shared coupling facility PUs (ICFs or CPs) is not a preferable production configuration. It is preferable for a production CF to operate by using dedicated ICFs. With CFCC Level 19 (and later; z14 ZR1 servers run CFCC level 22), Coupling Thin Interrupts are available, and dedicated engines continue to be recommended to obtain the best coupling facility performance.

The CPC on the left side of Figure 3-9 has two environments that are defined (production and test), and each has one z/OS and one coupling facility image. The coupling facility images share an ICF.

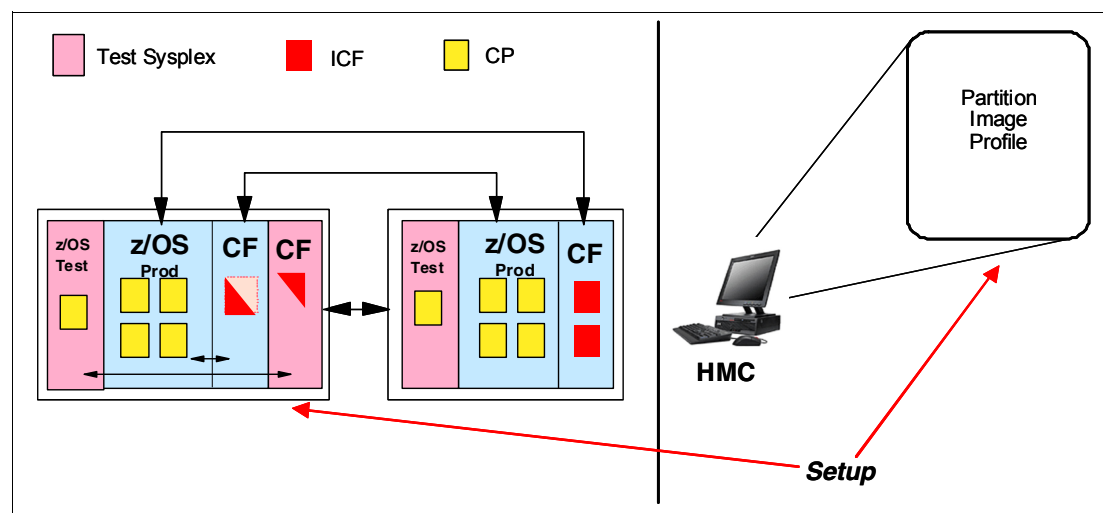


Figure 3-9 ICF options: Shared ICFs

The LPAR processing weights are used to define the amount of processor capacity of each CF image. The capped option can also be set for a test CF image to protect the production environment.

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

Dynamic CF dispatching

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

CFCC Level 19 introduced Coupling Thin Interrupts and the new DYNDISP specification. It allows more environments with multiple CF images to coexist in a server, and to share CF engines with reasonable performance. For more information, see 3.9.3, “Dynamic CF dispatching” on page 113.

Coupling Facility Processor Scalability

CF work management and dispatcher changed to allow improved efficiency as processors are added to scale up the capacity of a CF image.

CF images support up to 16 processors. To obtain sufficient CF capacity, customers might be forced to split the CF workload across more CF images. However, this change brings more configuration complexity and granularity (more, smaller CF images, more coupling links, and logical CHPIDs to define and manage for connectivity, and so on).

To improve CF processor scaling for the customer's CF images and to make effective use of more processors as the sysplex workload increases, CF work management and dispatcher provide the following improvements (z14 ZR1):

- ▶ Non-prioritized (FIFO-based) work queues, which avoids overhead of maintaining ordered queues in the CF.
- ▶ Streamlined system-managed duplexing protocol, which avoids costly latching deadlocks that can occur between primary and secondary structure.
- ▶ “Functionally specialized” ICF processors that operate for CF images with dedicated processors defined under certain conditions that realize the following benefits:
 - One “functionally specialized” processor for inspecting suspended commands
 - One “functionally specialized” processor for pulling in new commands
 - The remaining processors are non-specialized for general CF request processing
 - Avoids many inter-processor contentions that were associated with CF dispatching

3.5.5 IBM Z Integrated Information Processor

An IBM Z Integrated Information Processor (zIIP)⁴ reduces the standard processor (CP) capacity requirements for z/OS Java, XML system services applications, and a portion of work of z/OS Communications Server and Db2 UDB for z/OS Version 8 or later, which frees up capacity for other workload requirements.

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

⁴ IBM z Systems Application Assist Processors (zAAPs) are not available on z14 ZR1 servers. A zAAP workload is dispatched to available zIIPs (zAAP on zIIP capability).

z14 ZR1 processors support SMT. z14 ZR1 servers implement two threads per core on IFLs and zIIPs. SMT must be enabled at the LPAR level and supported by the z/OS operating system. SMT was enhanced for z14 ZR1 and it is enabled for SAPs by default (no customer intervention required).

How zIIPs work

zIIPs are designed for supporting designated z/OS workloads. One of the workloads is Java code execution. When Java code must be run (for example, under control of IBM WebSphere), the z/OS JVM calls the function of the zIIP. The z/OS dispatcher then suspends the JVM task on the CP that it is running on and dispatches it on an available zIIP. 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, which frees that capacity for other workloads.

The logical flow of Java code that is running on a z14 ZR1 server that has a zIIP available is shown in Figure 3-10. When JVM starts running a Java program, it passes control to the z/OS dispatcher that verifies the availability of a zIIP.

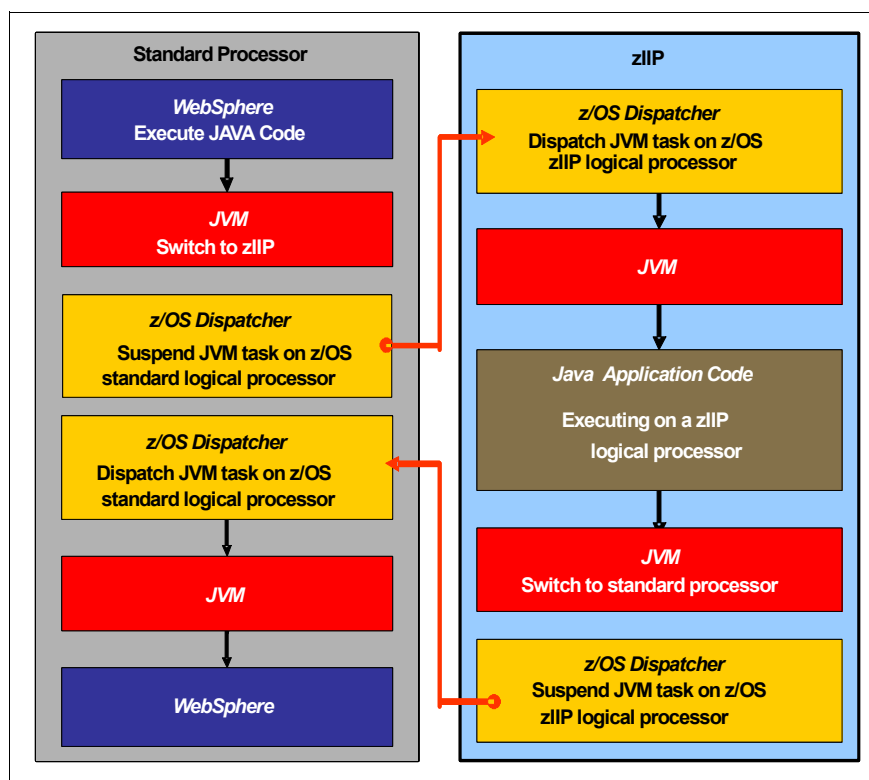


Figure 3-10 Logical flow of Java code execution on a zIIP

The availability is treated in the following manner:

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

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

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

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

- ▶ Option 1: Java dispatching by priority (IIPHONORPRIORITY=YES)

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

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

- ▶ Option 2: Java dispatching by priority (IIPHONORPRIORITY=NO)

zIIP-eligible work runs on zIIPs only while at least one zIIP engine is online. zIIP-eligible work is not normally dispatched on a CP, even if the zIIPs are overcommitted and CPs are unused. The exception is that zIIP-eligible work can sometimes run on a CP to resolve resource conflicts.

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

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

zIIPs provide the following benefits:

- ▶ Potential cost savings.
- ▶ Simplification of infrastructure as a result of the colocation and integration of new applications with their associated database systems and transaction middleware, such as Db2, IMS, or CICS. Simplification can happen, for example, by introducing a uniform security environment, and by reducing the number of TCP/IP programming stacks and system interconnect links.
- ▶ Prevention of processing latencies that occur if Java application servers and their database servers are deployed on separate server platforms.

The following Db2 UDB for z/OS V8 or later workloads are eligible to run in Service Request Block (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, which reduces the need for more gateway products that can affect performance and availability. The application uses the DRDA requester or server to access a remote database. IBM Db2 Connect is an example of a DRDA application requester.
- ▶ Star schema query processing, which is 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. For example, a star schema query 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, the databases become less efficient and must be maintained as data in large databases is manipulated.

The zIIP runs portions of eligible database workloads, which helps to free computer capacity and lower software costs. Not all Db2 workloads are eligible for zIIP processing. Db2 UDB for z/OS V8 and later gives z/OS the information to direct portions of the work to the zIIP. The result is that in every user situation, different variables determine how much work is redirected to the zIIP.

On a z14 server, the following workloads can also benefit from zIIPs:

- ▶ z/OS Communications Server uses the zIIP for eligible Internet Protocol Security (IPSec) network encryption workloads. This configuration requires z/OS V1R10 or later. Portions of IPSec processing use 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), also uses the zIIP. Most z/OS Data Facility Storage Management Subsystem (DFSMS) system data mover (SDM) processing that is associated with z/OS Global Mirror can run on the zIIP. This configuration requires z/OS V1R10 or later releases.
- ▶ The first IBM user of z/OS XML system services is Db2 V9. For Db2 V9 before the z/OS XML System Services enhancement, z/OS XML System Services non-validating parsing was partially directed to zIIPs when used as part of a distributed Db2 request through DRDA. This enhancement benefits Db2 V9 by making all z/OS XML System Services non-validating parsing eligible to zIIPs. This configuration is possible when processing is used as part of any workload that is running in enclave SRB mode.
- ▶ z/OS Communications Server also allows the HiperSockets Multiple Write operation for outbound large messages (originating from z/OS) to be run by a zIIP. Application workloads that are based on XML, HTTP, SOAP, and Java, and traditional file transfer can benefit.
- ▶ For BI, 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.

zIIP installation

One CP must be installed with or before any zIIP is installed. In z14 ZR1, the zIIP-to-CP ratio is 2:1, which means that up to 12 zIIPs can be characterized. The allowable number of zIIPs for each feature is listed in Table 3-3.

Table 3-3 Number of zIIPs per model

z14 ZR1 feature	Max4	Max12	Max24	Max30
Maximum zIIPs	0 - 2	0 - 8	0 - 12	0 - 12

zIIPs are orderable by using FC 1067. Up to two zIIPs can be ordered for each CP or marked CP configured in the system.

PUs that are 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 dedicated or shared, depending on whether they are part of an LPAR with dedicated or shared CPs. In an LPAR, at least one CP must be defined before zIIPs for that partition can be defined. The number of zIIPs that are available in the system is the number of zIIPs that can be defined to an LPAR.

LPAR: In an LPAR, as many zIIPs as are available can be defined together with at least one CP.

3.5.6 System assist processors

A system assist processor (SAP) is a PU that runs the channel subsystem LIC to control I/O operations. All SAPs run I/O operations for all LPARs. All models feature standard SAPs configured. The number of standard SAPs on the z14 ZR1 is two, regardless of the CPC drawer feature.

SAP configuration

A standard SAP configuration provides a well-balanced system for most environments. However, some application environments have high I/O rates, typically Transaction Processing Facility (TPF) environments. In this case, more SAPs can be ordered. Assigning of more SAPs can increase the capability of the channel subsystem to run I/O operations.

Optional other orderable (extra) SAPs

The option to order more SAPs is available on all z14 ZR1 features (FC 1066). These extra SAPs increase the capacity of the channel subsystem to run I/O operation, which is suggested for TPF environments. In z14 ZR1 systems, the maximum number of optional (extra) orderable SAPs is two, regardless of the CPC drawer feature.

3.5.7 Reserved processors

Reserved processors are defined by the PR/SM to allow for a nondisruptive capacity upgrade. Reserved processors are similar to spare logical processors, and can be shared or dedicated. Reserved PUs 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 the request. The PR/SM rules that govern logical processor activation remain unchanged.

By using reserved processors, you can define more logical processors than the number of available CPs, IFLs, ICFs, and zIIPs in the configuration to an LPAR. This process makes it possible to nondisruptively configure more logical processors online after more CPs, IFLs, ICFs, and zIIPs are made available concurrently. They can be made available with one of the Capacity on-demand options.

The maximum number of reserved processors that can be defined to an LPAR depends on the number of logical processors that are defined. The maximum number of logical processors plus reserved processors is 170. If the operating system in the LPAR supports the logical processor add function, reserved processors are no longer needed.

Do not define more active and reserved processors than the operating system for the LPAR can support. For more information about logical processors and reserved processors and their definitions, see 3.7, “Logical partitioning” on page 97.

3.5.8 Integrated firmware processor

An integrated firmware processor (IFP) is allocated from the pool of PUs and is available for the entire system. Unlike other characterized PUs, the IFP is standard and its definition is not controlled by the client. It is a single PU that is dedicated solely to supporting the management and service operations of the following *native* Peripheral Component Interconnect Express (PCIe) features:

- ▶ 10GbE Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE) Express
- ▶ 25GbE and 10GbE RoCE Express2
- ▶ zEnterprise Data Compression (zEDC) Express
- ▶ IBM zHyperLink Express
- ▶ Coupling Express Long Reach

The IFP is started at POR. The IFP supports Resource Group (RG) LIC to provide native PCIe I/O feature virtualization and service functions. For more information, see Appendix C, “Native Peripheral Component Interconnect Express” on page 419.

3.5.9 Sparing rules

On a z14 ZR1 system, one PU is reserved as spare. The reserved spare is available to replace any characterized PU, whether they are CP, IFL, ICF, zIIP, SAP, or IFP.

Systems with a failed PU for which no spare is available *call home* for a replacement. A system with a failed PU that is spared and requires an SCM to be replaced (referred to as a *pending repair*) can still be upgraded when sufficient PUs are available.

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

Depending on the feature, sparing of CP, IFL, ICF, zIIP, SAP, and IFP is transparent and does not require operating system or operator intervention.

With *transparent sparing*, the status of the application that was running on the failed processor is preserved. The application continues processing on a newly assigned CP, IFL, ICF, zIIP, SAP, or IFP (allocated to the spare PU) without client intervention.

Application preservation

If no spare PU is available, *application preservation* (z/OS only) is started. The state of the failing processor is passed to another active processor that is used by the operating system. Through operating system recovery services, the task is resumed successfully (in most cases, without client intervention).

Dynamic SAP and IFP sparing and reassignment

Dynamic recovery is provided if a failure of the SAP or IFP occurs. 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.

3.6 Memory design

This section describes design and implementation considerations for the z14 ZR1 memory.

3.6.1 Overview

The z14 ZR1 system has only one CPC drawer. As such, not all memory upgrade scenarios are nondisruptive. Therefore, memory upgrades require a different approach than in the multi-CPC drawer systems. Considering the z14 ZR1 plan ahead memory, the system flexibility, high availability can be achieved for memory upgrades.

Concurrent memory upgrades are supported up to the physical memory that is installed.

z14 ZR1 servers be configured with more physically installed memory than the initial client 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.

Any other memory upgrade is disruptive. Therefore, plan ahead memory is important. With memory plan ahead option, memory DIMMs are preinstalled to support a specified target planned memory size.

Note: The pre-planned (preinstalled, available for LICCC activation) memory amount cannot exceed 2TB.

Physical memory upgrades require processor drawer removal and reinstall after adding or replacing memory DIMMs. Because z14 ZR1 is a single CPC drawer system, physical memory upgrades are always disruptive.

When the total installed memory amount is larger than the customer usable memory required for a configuration, the LIC Configuration Control (LICCC) determines how much memory is used.

Memory allocation

When system is activated (POR), PR/SM determines the total installed memory and the client enabled memory. Later in the process during LPAR activation, PR/SM assigns and allocates memory to each partition according to partition image profile.

Large page support

By default, page frames are allocated with a 4 KB size. z14 ZR1 servers also support large page sizes of 1 MB or 2 GB. The first z/OS release that supports 1 MB pages is z/OS V1R9. Linux on Z large pages support (1 MB) is available in SUSE Linux Enterprise Server 10 SP2 or later, Red Hat Enterprise Linux (RHEL) 5.2 or later, and Ubuntu 16.04 LTS or later.

The TLB reduces 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) mechanism when it must find the correct page for the correct address space. Each TLB entry represents one page. As with other buffers or caches, lines are discarded from the TLB on a least recently used (LRU) basis.

The worst-case translation scenario is encountered when a TLB miss occurs and the segment table (which is needed to find the page table) and the page table (which is needed to find the entry for the particular page in question) are not in cache. This case involves two complete real memory access delays plus the address translation delay. Because the duration of a processor cycle is much shorter than the duration of a memory cycle, a TLB miss is relatively costly.

It is preferable to have addresses in the TLB. With 4 K pages, holding all of the addresses for 1 MB of storage takes 256 TLB lines. When 1 MB pages are used, it takes only one TLB line. Therefore, large page size users have a much smaller TLB footprint.

Large pages allow the TLB to better represent a large working set and suffer fewer TLB misses by allowing a single TLB entry to cover more address translations.

Users of large pages are better represented in the TLB and are expected to see performance improvements in elapsed time and processor usage. These improvements occur 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 also must maintain frequent memory access to keep the pertinent addresses in the TLB.

Short-running work does not overcome the processor usage. Short processes with small working sets are expected to receive little or no improvement. Long-running work with high memory-access frequency is the best candidate to benefit from large pages.

Long-running work with low memory-access frequency is less likely to maintain its entries in the TLB. However, when it does run, few address translations are required to resolve all of the memory it needs. Therefore, a long-running process can benefit even without frequent memory access.

Weigh the benefits of whether something in this category must use large pages as a result of the system-level costs of tying up real storage. A balance exists between the performance of a process that uses large pages and the performance of the remaining work on the system.

On z14 ZR1 servers, 1 MB large pages become pageable if Virtual Flash Memory⁵ is available and enabled. They are available only for 64-bit virtual private storage, such as virtual memory that is greater than 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 tradeoff between size and performance.

3.6.2 Main storage

Main storage is addressable by programs and storage that is not directly addressable by programs. Non-addressable storage includes the hardware system area (HSA).

Main storage provides the following functions:

- ▶ Data storage and retrieval for PUs and I/O
- ▶ Communication with PUs and I/O
- ▶ Communication with and control of optional expanded storage
- ▶ Error checking and correction

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

3.6.3 Hardware system area

The HSA is a reserved storage area that contains system LIC and configuration-dependent control blocks. On z14 ZR1 servers, the HSA has a fixed size of 64 GB and is not part of the client purchased memory.

The fixed size of the HSA eliminates planning for future expansion of the HSA because the hardware configuration definition (HCD) and input/output configuration program (IOCP) always reserves space for the following items:

- ▶ Three channel subsystems (CSSs)
- ▶ A total of 15 LPARs in the first two CSSs and 10 LPARs for the third CSS for a total of 40 LPARs per system
- ▶ 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 includes sufficient reserved space to allow for dynamic I/O reconfiguration changes to the maximum capability of the processor.

⁵ Virtual Flash Memory replaced IBM zFlash Express for z14. No carry forward of zFlash Express exists.

3.6.4 Virtual Flash Memory

IBM Virtual Flash Memory (VFM) is the replacement for the Flash Express features that were available on the IBM zEC12, zBC12, IBM z13®, and IBM z13s. No application changes are required to change from IBM Flash Express to VFM.

On z14 ZR1, up to four Virtual Flash Memory features (FC 0614) can be ordered. One VFM feature has 512 GB on z14 ZR1.

3.7 Logical partitioning

The logical partitioning features are described in this section.

3.7.1 Overview

Logical partitioning is a function that is implemented by the PR/SM⁶. z14 ZR1 can be managed in standard or Dynamic Partition Manager (DPM) modes. DPM provides a dynamic LPAR and resource management of the z14 ZR1 and uses a graphical, interactive interface to the PR/SM.

HiperDispatch

PR/SM and z/OS work in tandem to use processor resources more efficiently. HiperDispatch is a function that combines the dispatcher actions and the knowledge that PR/SM has about the topology of the system.

Performance can be optimized by redispersing units of work to the same processor group, which keeps processes running near their cached instructions and data, and minimizes transfers of data ownership among processors in different PU SCMs.

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

z/OS dispatcher manages multiple queues, called *affinity queues*, with a target number of eight processors per queue, which fits well onto a single PU chip. These queues are used to assign work to as few logical processors as are needed for an LPAR workload. Therefore, even if the LPAR is defined with many logical processors, HiperDispatch optimizes this number of processors to be near the required capacity.

Tip: z/VM V6.3 and later also support HiperDispatch, which is required for activating SMT. (z14 ZR1 supports z/VM V6.4 or newer.)

Logical partitions

PR/SM enables z14 ZR1 servers to be started for a logically partitioned operation, supporting up to 40 LPARs. Each LPAR can run its own operating system image in any image mode, independently from the other LPARs.

⁶ PR/SM -Processor Resource/Systems Manager.

An LPAR can be added, removed, activated, or deactivated at any time. Changing the number of LPARs is not disruptive and does not require a POR. Certain facilities might not be available to all operating systems because the facilities might include software corequisites.

Each LPAR has the following resources that are the same as a real CPC:

► Processors

Known as *logical processors*, they can be defined as CPs, IFLs, ICFs, or zIIPs. They can be dedicated to an LPAR or shared among LPARs. When shared, a processor weight can be defined to provide the required level of processor resources to an LPAR. Also, the capping option can be turned on, which prevents an LPAR from acquiring more than its defined weight and limits its processor consumption.

LPARs for z/OS can have CP and zIIP logical processors. The two logical processor types can be defined as all dedicated or all shared. The zIIP support is available in z/OS.

The weight and number of online logical processors of an LPAR can be dynamically managed by the LPAR CPU Management function of the Intelligent Resource Director (IRD). These functions can be used to achieve the defined goals of this specific partition and of the overall system. The provisioning architecture of z14 ZR1 servers (as described in Chapter 8, “System upgrades” on page 281) adds another dimension to the dynamic management of LPARs.

PR/SM is enhanced to support an option to limit the amount of physical processor capacity that is used by an individual LPAR when a PU is defined as a general-purpose processor (CP) or an IFL that is shared across a set of LPARs.

This enhancement is designed to provide a physical capacity limit that is enforced as an absolute (versus relative) limit. It is not affected by changes to the logical or physical configuration of the system. This physical capacity limit can be specified in units of CPs or IFLs. The Change LPAR Controls and Customize Activation Profiles tasks on the HMC were enhanced to support this new function.

For the z/OS Workload License Charges (WLC) pricing metric and metrics that are based on it, such as Advanced Workload License Charges (AWLC), an LPAR *defined capacity* can be set. This defined capacity enables the soft capping function.

Workload charging introduces the capability to pay software license fees that are based on the processor use of the LPAR on which the product is running, rather than on the total capacity of the system. Consider the following points:

- 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 four-hour rolling average of the processor usage of the LPAR. When the four-hour average processor consumption exceeds the defined capacity limit, WLM dynamically activates LPAR capping (soft capping). When the rolling four-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 more information about software licensing, see 7.8, “Software licensing” on page 277.

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 that corresponds to the relative weight settings, and no capping at all. It is best to avoid this scenario and instead attempt to establish a defined capacity that is equal or close to the relative weight.

► Memory

Memory (main storage) must be dedicated to an LPAR. The defined storage must be available during the LPAR activation; otherwise, the LPAR activation fails.

Reserved storage can be defined to an LPAR, which enables nondisruptive memory addition to and removal from an LPAR by using the LPAR dynamic storage reconfiguration (z/OS and z/VM). For more information, see 3.7.5, “LPAR dynamic storage reconfiguration” on page 106.

► Channels

Channels can be shared between LPARs by including the partition name in the partition list of a channel-path identifier (CHPID). I/O configurations are defined by the IOCP or the HCD with the CHPID mapping tool (CMT). The CMT is an optional tool that is used to map CHPIDs onto physical channel IDs (PCHIDs). PCHIDs represent the physical location of a port on a card in a 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. For more information, see *IBM Z Input/Output Configuration Program User's Guide for ICP IOCP*, SB10-7172-01.

HCD is available on the z/OS and z/VM operating systems. Consult the appropriate 3907DEVICE Preventive Service Planning (PSP) buckets before implementation.

Fibre Channel connection (FICON) channels can be managed by the Dynamic CHPID Management (DCM) function of the Intelligent Resource Director. DCM enables the system to respond to ever-changing channel requirements by moving channels from lesser-used control units to more heavily used control units, as needed.

Modes of operation

The modes of operation are listed in Table 3-4. All available mode combinations, including their operating modes and processor types, operating systems, and addressing modes, also are listed. Only the currently supported versions of operating systems are considered.

Table 3-4 z14 modes of operation

Image mode	PU type	Operating system	Addressing mode
z/Architecture (General) ^a	CP and zIIP	z/OS z/VM	64-bit
	CP	z/VSE Linux on Z z/TPF	64-bit
Coupling facility	ICF or CP	CFCC	64-bit

Image mode	PU type	Operating system	Addressing mode
Linux only	IFL <i>or</i> CP	Linux on Z (64-bit)	64-bit
		z/VM	
		Linux on Z (31-bit)	31-bit
z/VM	CP, IFL, zIIP, or ICF	z/VM	64-bit
SSC ^b	IFL or CP	z/VSE Network Appliance ^c	64 bit

a. Formerly ESA/390 mode

b. Secure Service Container

c. More appliances to be announced and supported in the future

For more information about operating system support, see Chapter 7, “Operating system support” on page 209.

Logically partitioned mode

If the z14 ZR1 server runs in LPAR mode, each of the 40 LPARs can be defined to operate in one of the following image modes:

- ▶ z/Architecture (General) mode to run the following systems:
 - A z/Architecture operating system, on dedicated or shared CPs
 - ESA/390 operating systems
 - A Linux on 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 zIIPs
 - Shared CPs *and* shared zIIPs

zIIP usage: zIIPs can be defined to an z/Architecture mode or z/VM mode image, as listed in Table 3-4 on page 99. However, zIIPs are used only by z/OS. Other operating systems cannot use zIIPs, even if they are defined to the LPAR. z/VM^a supports real and virtual zIIPs to guest z/OS systems.

a. z/VM V6R4 or newer is supported on IBM z14 ZR1.

- ▶ z/Architecture (General) mode is also used 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 the following types:
 - Dedicated or shared CPs
 - Dedicated or shared ICFs
- ▶ LINUX only mode to run the following systems:
 - A Linux on Z operating system, on either of the following types:
 - Dedicated or shared IFLs
 - Dedicated or shared CPs
 - A z/VM operating system, on either of the following types:
 - Dedicated or shared IFLs
 - Dedicated or shared CPs
- ▶ z/VM mode to run z/VM on dedicated or shared CPs or IFLs, plus zIIPs and ICFs

- SSC (Secure Service Container) mode LPAR can run on:
 - Dedicated or shared CPs
 - Dedicated or shared IFLs

All LPAR modes, required characterized PUs, operating systems, and the PU characterizations that can be configured to an LPAR image are listed in Table 3-5. The available combinations of dedicated (DED) and shared (SHR) processors are also included. For all combinations, an LPAR also can include reserved processors that are defined, which allows for nondisruptive LPAR upgrades.

Table 3-5 LPAR mode and PU usage

LPAR mode	PU type	Operating systems	PUs usage
z/Architecture (General)	CPs	z/Architecture operating systems (z/OS, z/VSE, z/TPF) Linux on Z	CPs DED or CPs SHR
	CPs <i>and</i> zIIPs	z/OS z/VM (V6R3 and later for guest exploitation). ESA/390 operating systems ^a	CPs DED or zIIPs DED or CPs SHR or zIIPs SHR
Coupling facility	ICFs <i>or</i> CPs	CFCC	ICFs DED or ICFs SHR or CPs DED or CPs SHR
LINUX only	IFLs <i>or</i> CPs	Linux on Z z/VM	IFLs DED or IFLs SHR or CPs DED or CPs SHR
z/VM	CPs, IFLs, zIIPs, or ICFs	z/VM (V6R4 and later)	All PUs must be SHR or DED
SSC ^b	IFLs, <i>or</i> CPs	IBM zAware z/VSE Network Appliance	IFLs DED or IFLs SHR or CPs DED or CPs SHR

a. ESA/390 operating systems cannot be IPL'ed on z14 ZR1. Limited support for z/VM guests.

b. Secure Service Container

Dynamically adding or deleting a logical partition name

Dynamically adding or deleting an LPAR name is the ability to add or delete LPARs and their associated I/O resources to or from the configuration without a POR.

The extra channel subsystem and multiple image facility (MIF) image ID pairs (CSSID/MIFID) can be assigned later to an LPAR for use (or later removed). This process can be done through dynamic I/O commands by using the 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 numbers are assigned to a partition profile of a specific name. The client assigns these AP numbers and domains to the partitions and continues to have the responsibility to clear them out when their profiles change.

Adding logical processors to a logical partition

Logical processors can be concurrently added to an LPAR by defining them as reserved in the image profile and later configuring them online to the operating system by using the appropriate console commands. Logical processors also can be concurrently added to a logical partition dynamically by using the Support Element (SE) “Logical Processor Add” function under the CPC Operational Customization task. This SE function allows the initial and reserved processor values to be dynamically changed. The operating system must support the dynamic addition of these resources. In z/OS, this support is available since Version 1 Release 10 (z/OS V1.10), while z/VM supports this addition since z/VM V5.4, and z/VSE since V4.3.

Adding a crypto feature to a logical partition

You can plan the addition of Crypto Express6S(5S) features to an LPAR on the crypto page in the image profile by defining the Cryptographic Candidate List, and the Usage and Control Domain indexes, in the partition profile. By using the Change LPAR Cryptographic Controls task, you can add crypto adapters dynamically to an LPAR without an outage of the LPAR. Also, dynamic deletion or moving of these features does not require pre-planning. Support is provided in z/OS, z/VM, z/VSE, Secure Service Container (based on appliance requirements), and Linux on Z.

LPAR dynamic PU reassignment

The system configuration is enhanced to optimize the PU-to-CPC drawer assignment of physical processors dynamically. The initial assignment of client-usable physical processors to PU SCMs can change dynamically to better suit the LPAR configurations that are in use.

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 PU SCMs.

LPAR dynamic PU reassignment can swap client processors of different types between PU SCMs. For example, reassignment can swap an IFL on a PU SCM 1 with a CP on PU SCM 2. Swaps can also occur between PU SCMs within different PU clusters and can include spare PU. The goals are to pack the LPAR on fewer PU chips, based on the z14 ZR1 PU SCMs 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 (LPAR group absolute capping)

The group capacity limit feature allows the definition of a group of LPARs on a z14 ZR1 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 use this feature, you must be running z/OS V1.10 or later in the all LPARs in the group.

PR/SM and WLM work together to enforce the capacity that is defined for the group and the capacity that is optionally defined for each individual LPAR.

LPAR absolute capping

Absolute capping is a logical partition control that was made available with zEC12 and is supported on z14 ZR1 servers. With this support, PR/SM and the HMC are enhanced to support a new option to limit the amount of physical processor capacity that is used by an individual LPAR when a PU is defined as a general-purpose processor (CP), zIIP, or an IFL processor that is shared across a set of LPARs.

Unlike traditional LPAR capping, absolute capping is designed to provide a physical capacity limit that is enforced as an absolute (versus relative) value that is not affected by changes to the virtual or physical configuration of the system.

Absolute capping provides an optional maximum capacity setting for logical partitions that is specified in the absolute processors capacity (for example, 5.00 CPs or 2.75 IFLs). This setting is specified independently by processor type (namely CPs, zIIPs, and IFLs) and provides an enforceable upper limit on the amount of the specified processor type that can be used in a partition.

Absolute capping is ideal for processor types and operating systems that the z/OS WLM cannot control. Absolute capping is not intended as a replacement for defined capacity or group capacity for z/OS, which are managed by WLM.

Absolute capping can be used with any z/OS, z/VM, or Linux on z LPAR that is running on an IBM Z server. If specified for a z/OS LPAR, it can be used concurrently with defined capacity or group capacity management for z/OS. When used concurrently, the absolute capacity limit becomes effective before other capping controls.

Dynamic Partition Manager mode

Dynamic Partition Manager (DPM) is an IBM Z server operation mode that provides a simplified approach to create and manage virtualized environments, which reduces the barriers of its adoption for new and existing customers.

The implementation provides built-in integrated capabilities that allow advanced virtualization management on IBM Z servers. With DPM, you can use your Linux and virtualization skills while taking advantage of the full value of IBM Z hardware, robustness, and security in a workload optimized environment.

DPM provides facilities to define and run virtualized computing systems by using a firmware-managed environment that coordinates the physical system resources that are shared by the partitions. The partitions' resources include processors, memory, network, storage, crypto, and accelerators.

DPM provides a new mode of operation for IBM Z servers that provide the following services:

- ▶ Facilitates defining, configuring, and operating PR/SM LPARs in a similar way to how someone performs these tasks on another platform.
- ▶ Lays the foundation for a general IBM Z new user experience.

DPM is *not* another hypervisor for IBM Z servers. DPM uses the PR/SM hypervisor infrastructure and provides an intelligent interface that allows customers to define, use, and operate the platform virtualization without IBM Z experience or skills. For more information about DPM, see Appendix E, "IBM Dynamic Partition Manager" on page 451.

3.7.2 Storage operations

In z14 ZR1 servers, memory can be assigned as main storage supporting up to 40 LPARs. Before you activate an LPAR, main storage must be defined to the LPAR. All installed storage can be configured as main storage. Each z/OS individual LPAR can be defined with a maximum of 4 TB of main storage. z/VM V6R4 supports 2 TB of main storage.

Main storage can be dynamically assigned to expanded storage and back to main storage as needed without a POR.

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. z/VM, in turn, virtualizes this support to its guests. For more information, see 3.7.5, “LPAR dynamic storage reconfiguration” on page 106.

Operating systems that run as guests of z/VM can use the z/VM capability of implementing virtual memory to guest virtual machines. The z/VM dedicated real storage can be shared between guest operating systems.

The z14 ZR1 storage allocation and usage possibilities, depending on the image mode, are listed in Table 3-6.

Table 3-6 Main storage definition and usage possibilities

Image mode	Architecture mode (addressability)	Maximum main storage	
		Architecture	z14 ZR1 definition
z/Architecture (General)	z/Architecture (64-bit)	16 EB	4 TB
Coupling facility	CFCC (64-bit)	1.5 TB	1 TB
Linux only	z/Architecture (64-bit)	16 EB	2 TB
z/VM	z/Architecture (64-bit)	16 EB	2 TB
SSC ^a	z/Architecture (64-bit)	16 EB	2 TB

a. Secure Service Container

The following modes are provided:

► z/Architecture mode

In z/Architecture (General, formerly ESA/390 or ESA/390-TPF) mode, storage addressing is 64-bit, which allows for virtual addresses up to 16 exabytes (16 EB). The 64-bit architecture theoretically allows a maximum of 16 EB to be used as main storage. However, the current main storage limit for LPARs is 8 TB for z14 ZR1. The operating system that runs in z/Architecture mode must support the real storage. Currently, z/OS supports up to 4 TB⁷ of real storage (z/OS V2R1 and later releases).

► CF mode

In CF mode, storage addressing is 64 bit for a CF image that runs at CFCC Level 12 or later. This configuration allows for an addressing range up to 16 EB. However, the current z14 ZR1 definition limit for CF LPARs is 1 TB of storage. The following CFCC levels are supported in a Sysplex with IBM z14 ZR1:

- CFCC Level 23, available on z14 (Driver level 36)
- CFCC Level 22, available on z14 ZR1 (Driver level 32)
- CFCC Level 21, available on z13 and z13s (Driver Level 27)
- CFCC Level 20, available for z13 servers with Driver Level 22

Restriction: z14 ZR1 does not support direct coupling connectivity to zEC12/zBC12 systems.

For more information, see 3.9.1, “Coupling Facility Control Code” on page 110.

⁷ 1 TB for z/OS V1R13.

Expanded storage cannot be defined for a CF image. Only IBM CFCC can run in CF 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.

Only Linux and z/VM operating systems can run in Linux only mode. Linux on Z 64-bit distributions (SUSE Linux Enterprise Server 10 and later, Red Hat RHEL 5 and later, and Ubuntu 16.04 LTS 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 feature increases flexibility and simplifies systems management by allowing z/VM to run the following tasks in the same z/VM LPAR:

- Manage guests to operate Linux on Z on IFLs
- Operate z/VSE and z/OS on CPs
- Offload z/OS system software processor usage, such as Db2 workloads on zIIPs
- Provide an economical Java execution environment under z/OS on zIIPs

- ▶ Secure Service Container (SSC) mode

In SSC mode, storage addressing is 64 bit for an embedded product. This configuration allows for an addressing range up to 16 EB. However, the current z14 ZR1 definition limit for LPARs is 8 TB of storage (physical memory limit).

Currently, the z/VSE Network Appliance (available on z14 ZR1) runs in an SSC LPAR.

3.7.3 Reserved storage

Reserved storage can be optionally defined to an LPAR, which allows a nondisruptive image memory upgrade for this partition. Reserved storage can be defined to central and expanded storage, and to any image mode except CF mode.

An LPAR must define an amount of main storage and optionally (if not a CF image), an amount of expanded storage. Main storage and expanded storage can have the following storage sizes defined:

- ▶ The *initial value* is the storage size that is allocated to the partition when it is activated.
- ▶ The *reserved value* is another 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 need be available when the partition is activated. They are predefined storage sizes to allow a storage increase, from an LPAR point of view.

Without the reserved storage definition, an LPAR storage upgrade is a disruptive process that requires the following steps:

1. Partition deactivation.
2. An initial storage size definition change.
3. Partition activation.

The extra storage capacity for an LPAR upgrade can come from the following sources:

- ▶ Any unused available storage
- ▶ Another partition that features 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 V5R4⁸ 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. Removing storage from the z/VM LPAR is disruptive. Removing memory from a z/VM guest is not disruptive to the z/VM LPAR.

SLES 11 and later supports concurrent add and remove.

3.7.4 Logical partition storage granularity

Granularity of main storage for an LPAR depends on the largest main storage amount that is defined for initial or reserved main storage, as listed in Table 3-7.

Table 3-7 Logical partition main storage granularity (z14)

Largest main storage amount	Main storage granularity
Main storage amount <= 512 GB	1 GB
512 GB < main storage amount <= 1 TB	2 GB
1 TB < main storage amount <= 2 TB	4 GB
2 TB < main storage amount <= 4 TB	8 GB
4 TB < main storage amount <= 8 TB	16 GB

LPAR storage granularity information is required for LPAR image setup and for z/OS RSU definition. LPARs are limited to a maximum size of 8 TB of main storage. However, the maximum amount of memory that is supported by z/OS V2.3 at the time of this writing is 4 TB; for z/VM V6R4 and V7R1, the limit is 2 TB.

3.7.5 LPAR dynamic storage reconfiguration

Dynamic storage reconfiguration on z14 ZR1 servers allows an operating system that is running on an LPAR to add (nondisruptively) its reserved storage amount to its configuration. This process can occur only if unused storage exists. This unused storage can be obtained when another LPAR releases storage, or when a concurrent memory upgrade occurs.

With dynamic storage reconfiguration, the unused storage need not be continuous.

When an operating system that is 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 that is running on an LPAR releases a storage increment.

⁸ z14 ZR1 supports z/VM 6.4 or newer.

3.8 Intelligent Resource Director

Intelligent Resource Director (IRD) is a z14 ZR1 and IBM Z capability that is used by z/OS only. IRD is a function that optimizes processor and channel resource utilization across LPARs within a single IBM Z server.

This feature extends the concept of goal-oriented resource management. It does so by grouping system images that are on the same z14 ZR1 or Z servers that are running in LPAR mode, and in the same Parallel Sysplex, into an *LPAR cluster*. This configuration allows WLM to manage resources (processor and I/O) across the entire cluster of system images and not only in one single image.

An LPAR cluster is shown in Figure 3-11. 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 CF image. In this example, the scope over which IRD has control is the defined LPAR cluster.

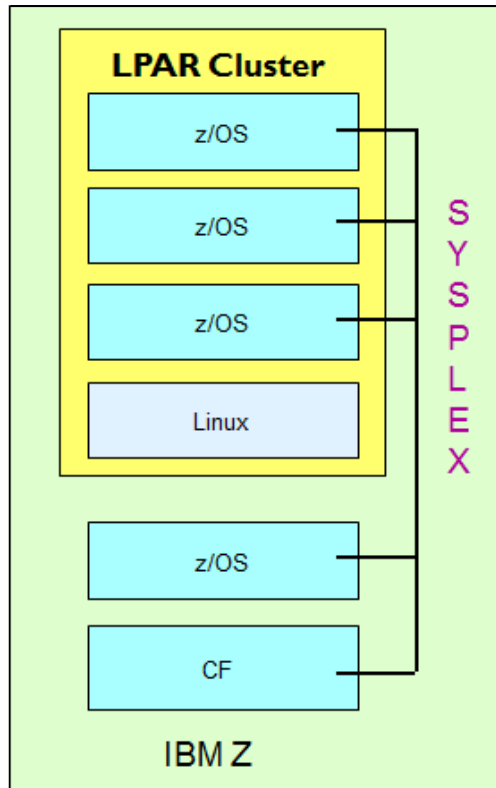


Figure 3-11 IRD LPAR cluster example

IRD features the following characteristics:

- IRD processor management

WLM dynamically adjusts the number of logical processors within an LPAR and the processor weight that is 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 97.

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. z14 ZR1 servers support DCM within a channel subsystem.

- Channel subsystem priority queuing

This function on z14 ZR1 and Z servers 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 is the clustering technology that is used with z14 ZR1 servers. The components of a Parallel Sysplex as implemented within the z/Architecture are shown in Figure 3-12, which is one of many possible Parallel Sysplex configurations.

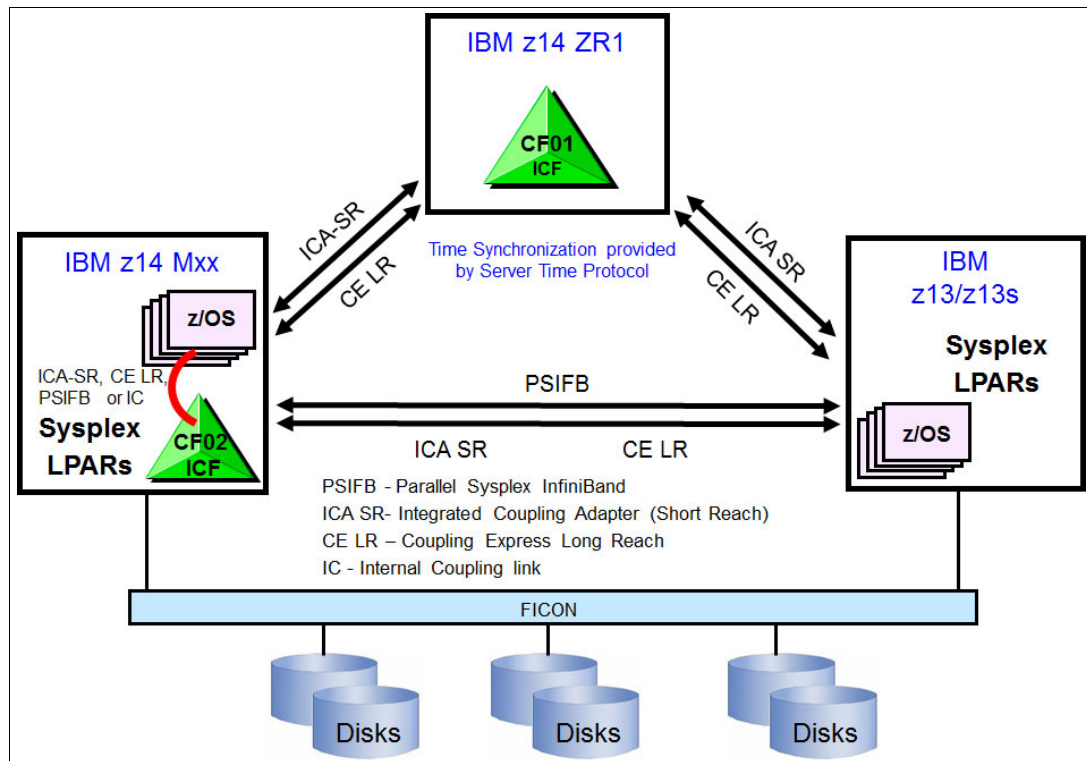


Figure 3-12 Sysplex hardware overview

A z14 M0x system that contains multiple z/OS sysplex partitions also is shown in Figure 3-12. It contains an internal CF (CF02), a z14 ZR1 system that contains a stand-alone CF (CF01), and a z13/z13s that contains multiple z/OS sysplex partitions.

STP over coupling links provides time synchronization to all systems. The appropriate CF link technology (Integrated Coupling Adapter, Coupling Express Long reach, 1x InfiniBand, or 12x InfiniBand) selection, depends on the system configuration and how distant they are physically. The ISC-3 coupling link is not supported since z13 servers. For more information about link technologies, see “Coupling links” on page 153.

Important: New for z14 ZR1, the z14 ZR1 supports only PCIe-based coupling technology (ICA SR and Coupling Express Long Reach). As a consequence, the z14 ZR1 cannot be connected to a zEC12 or zBC12 server directly; therefore, they cannot be part of the same Parallel Sysplex cluster.

Parallel Sysplex technology is an enabling technology that allows highly reliable, redundant, and robust IBM Z technology to achieve near-continuous availability. A Parallel Sysplex makes up one or more (z/OS) operating system images that are coupled through one or more Coupling Facilities. The images can be combined to form clusters.

A correctly configured Parallel Sysplex cluster maximizes availability in the following 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: 2 - 32 z/OS images in a sysplex.
- ▶ Dynamic workload balancing: Because it is viewed as a single logical resource, work can be directed to any similar operating system image in a Parallel Sysplex cluster that has available capacity.
- ▶ Systems management: The architecture provides the infrastructure to satisfy client requirements for continuous availability. It also provides techniques for achieving simplified systems management consistent with this requirement.
- ▶ Resource sharing: Several base (z/OS) components use the CF 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, user, and database administrator. A single system image ensures reduced complexity from operational and definition perspectives.
- ▶ N-1 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. However, a direct connection between z14 ZR1 with a N-2 Z servers is *not* supported because z14 ZR1 does not have InfiniBand coupling.

Through state-of-the-art cluster technology, the power of multiple images can be harnessed to work together on common workloads. The IBM 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

The LPAR that is running the Coupling Facility Control Code (CFCC) can be on z14, z14 ZR1, z13, z13s, zEC12, and zBC12 systems. For more information about CFCC requirements for supported systems, see “Coupling facility and CFCC considerations” on page 240.

Consideration: z14 ZR1 servers cannot coexist in the same sysplex with zEC12/zBC12 and previous systems. The introduction of z14 ZR1 servers into existing installations might require more planning.

CFCC Level 23

CFCC level 23 is delivered on the z14 ZR1 with driver level 36. CFCC Level 23 introduces the following enhancements:

- ▶ Asynchronous cross-invalidate (XI) of CF cache structures. Requires PTF support for z/OS and explicit data manager support (Db2 V12 with PTFs):
 - Instead of performing XI signals synchronously on every cache update request that causes them, data managers can “opt in” for the CF to perform these XIs asynchronously (and then sync them up with the CF at or before transaction completion). Data integrity is maintained if all XI signals complete by the time transaction locks are released.

- Results in faster completion of cache update CF requests, especially with cross-site distance that is involved.
- Provides improved cache structure service times and coupling efficiency
- Coupling Facility hang detect enhancements provide a significant reduction in failure scope and client disruption (CF-level to structure-level), with no loss of FFDC collection capability:
 - When a hang is detected, the CF confines the scope of the failure in most cases to “structure damage” for the single CF structure the hung command was processing against, capture diagnostics with a non-disruptive CF dump, and continue operating without aborting or rebooting the CF image.
 - Provides a significant reduction in failure scope and client disruption (CF-level to structure-level), with no loss of FFDC collection capability.
- Coupling Facility ECR granular latching
 - With this support, most CF list and lock structure ECR processing no longer uses structure-wide latching. It serializes its execution by using the normal structure object latches that all mainline commands use.
 - Eliminates the performance degradation caused by structure-wide latching.
 - A small number of “edge conditions” in ECR processing still require structure-wide latching to be used to serialize them.
 - Cache structure ECR processing continues to require and use structure-wide latches for its serialization.

z14 ZR1 servers with CFCC Level 23 require z/OS V1R13 or later, and z/VM V6R4 or later for virtual guest coupling.

CFCC Level 22

CFCC level 22 is delivered on the z14 ZR1 servers with driver level D32. CFCC Level 22 introduces the following enhancements:

- CF Enhancements:
 - CF structure encryption.
 CF Structure encryption is transparent to CF-using middleware and applications, while CF users are unaware of and not involved in the encryption. All data and adjunct data that flows between z/OS and the CF is encrypted. The intent is to encrypt all data that might be sensitive.

 Internal control information and related request metadata are not encrypted, including locks and lock structures.

 z/OS generates the required structure-related encryption keys and does much of the key management automatically by using CFRM that uses secure, protected keys (never clear keys). Secure keys maintained in CFRM couple dataset.
- CF Asynchronous Duplexing for Lock Structures:
 - New asynchronous duplexing protocol for lock structures:
 - z/OS sends command to primary CF only
 - Primary CF processes command and returns result
 - Primary CF forwards description of required updates to secondary CF
 - Secondary CF updates secondary structure instance asynchronously

- Provided for lock structures only:
 - z/OS V2.2 SPE with PTFs for APAR OA47796
 - Db2 V12 with PTFs
 - Most performance-sensitive structures for duplexing
- Benefit/Value:
 - Db2 locking receives performance similar to simplex operations
 - Reduces CPU and CF link overhead
 - Avoids the overhead of synchronous protocol handshakes on every update
 - Duplexing failover much faster than log-based recovery
- Targeted at multi-site clients who run split workloads at distance to make duplexing lock structures at distance practical.
- CF Processor Scalability:
 - CF work management and dispatcher changes to allow improved efficiency as processors are added to scale up the capacity of a CF image.
 - Functionally specialized ICF processors that operate for CF images having more than a threshold number of dedicated processors defined for them:
 - One functionally specialized processor for inspecting suspended commands.
 - One functionally specialized processor for pulling in new commands.
 - The remaining processors are non-specialized for general CF request processing.
 - Avoids many inter-processor contentions that were associated with CF dispatching.
- Enable systems management applications to collect valid CF LPAR information through z/OS BCPII:
 - System Type (CFCC)
 - System Level (CFCC LEVEL)
 - Dynamic Dispatch settings to indicate CF state (dedicated, shared, and thin interrupt), which are useful when investigating functional performance problems

z14 ZR1 systems with CFCC Level 22 require z/OS V1R13 with PTFs or later, and z/VM V6R4 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 CF LPARs are running on different servers. CF LPARs that run on the same server share the CFCC level.

z14 ZR1 servers (CFCC level 22) can coexist in a sysplex with CFCC levels 20 and 21.

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 setting also means that the CF Control Code uses all the processor capacity (cycles) that are available for the CF LPAR. If the LPAR that is running the CFCC includes only dedicated processors (CPs or ICFs), all processor capacity (cycles) can be used. However, this configuration can be an issue if the LPAR that is running the CFCC also includes shared processors. Therefore, enable dynamic dispatching on the CF LPAR.

Starting with CFCC Level 19 and Coupling Thin Interrupts, shared-processor CF can provide more consistent CF service time and acceptable usage in a broader range of configurations. For more information, see 3.9.3, “Dynamic CF dispatching” on page 113.

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

CF structure sizing changes are expected when moving from CFCC Level 17 (or earlier) to CFCC Level 20 or later. Review the CF structure size by using [the CFSizer tool](#).

For more information about the recommended CFCC levels, see the [current exception letter that is published on Resource Link](#) (login required).

3.9.2 Coupling Thin Interrupts

CFCC Level 19 introduced Coupling Thin Interrupts to improve performance in environments that share CF engines. Although dedicated engines are preferable to obtain the best CF performance, Coupling Thin Interrupts can help facilitate the use of a shared pool of engines, which helps to lower hardware acquisition costs.

The interrupt causes a shared logical processor CF partition to be dispatched by PR/SM (if it is not already dispatched), which allows the request or signal to be processed in a more timely manner. The CF relinquishes control when work is exhausted or when PR/SM takes the physical processor away from the logical processor.

The use of Coupling Thin Interrupts is controlled by the new DYNDISP specification.

You can experience CF response time improvements or more consistent CF response time when CFs are used with shared engines. This improvement can 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 the use of Coupling Thin Interrupts on the z/OS host system, regardless of whether the CF is using shared or dedicated engines.

3.9.3 Dynamic CF dispatching

Dynamic CF dispatching uses the following process on a CF:

1. If no work is available, CF enters a wait state (by time).
2. After an elapsed time, CF wakes up to see whether any new work is available (that is, if any requests are in the CF Receiver buffer).
3. If no work exists, CF sleeps again for a longer period.
4. If new work is available, CF enters the normal active wait until no other work is available. After all work is complete, the process starts again.

With the introduction of the Coupling Thin Interrupt support, which is used only when the CF partition is using shared engines and the new **DYNDISP=THININTERRUPT** parameter, the CFCC code is changed to handle these interrupts correctly. CFCC was also changed to relinquish voluntarily control of the processor whenever it runs out of work to do. It relies on Coupling Thin Interrupts to dispatch the image again in a timely fashion when new work (or new signals) arrives at the CF to be processed.

3.10 Virtual Flash Memory

Flash Express is not supported on z14 ZR1. This feature was replaced by IBM Z Virtual Flash Memory (zVFM), FC 0614.

3.10.1 IBM Z Virtual Flash Memory overview

Virtual Flash Memory (VFM) is an IBM solution to replace the external zFlash Express feature with support that is based on main memory.

The “storage class memory” that is provided by Flash Express adapters is replaced with memory allocated from main memory (VFM).

VFM is designed to help improve availability and handling of paging workload spikes when z/OS V2.1, V2.2, or V2.3 is running. With this support, z/OS is designed to help improve system availability and responsiveness by using VFM across transitional workload events, such as market openings and diagnostic data collection. z/OS is also designed to help improve processor performance by supporting middleware use of pageable large (1 MB) pages.

VFM can also be used in CF images to provide extended capacity and availability for workloads that use IBM WebSphere MQ Shared Queues structures. The use of VFM can help availability by reducing latency from paging delays that can occur at the start of the workday or during other transitional periods. It is also designed to eliminate delays that can occur when collecting diagnostic data during failures.

3.10.2 VFM feature

A VFM feature (FC 0614) has 512 GB on z14 ZR1. The maximum number of VFM features is *four* per z14 ZR1 server.

3.10.3 VFM administration

The allocation and definition information of VFM for all partitions is viewed through the Storage Information panel under the Operational Customization panel.

Tip: For LPARs that require VFM, the maximum amount (2 TB for z14 ZR1) should be defined for every LPAR where a supporting OS might run, even if the initial amount is zero. Defining the maximum amount of supported VFM allows the dynamic addition of the VFM later without requiring an LPAR reactivation.

VFM is much simpler to manage (HMC task) and no hardware repair and verify (no cables and no adapters) are needed. Also, because this feature is part of internal memory, VFM is protected by RAIM and ECC and can provide better performance because no I/O to an attached adapter occurs.

Note: Use cases for FlashExpress did not change (for example, z/OS paging and CF shared queue overflow). Instead, they transparently benefit from the changes in the hardware implementation. No option is available for VFM plan ahead.



Central processor complex I/O system structure

This chapter describes the I/O system structure and connectivity options that are available on the IBM z14 Model ZR1 servers.

Note: Throughout this chapter, “z14” refers to IBM z14 Model M0x (Machine Type 3906).

This chapter includes the following topics:

- ▶ 4.1, “Introduction to I/O infrastructure” on page 118
- ▶ 4.2, “I/O system overview” on page 120
- ▶ 4.3, “PCIe+ I/O drawer” on page 122
- ▶ 4.4, “CPC drawer fanouts” on page 125
- ▶ 4.5, “I/O features (cards)” on page 128
- ▶ 4.6, “Connectivity” on page 133
- ▶ 4.7, “Cryptographic functions” on page 159
- ▶ 4.8, “Integrated Firmware Processor” on page 160
- ▶ 4.9, “zEDC Express” on page 160

4.1 Introduction to I/O infrastructure

This section describes the I/O features that are available on the IBM z14 Model ZR1 server. The z14 ZR1 servers support PCIe+ I/O drawers only. I/O cage, I/O drawer, or PCIe I/O drawer are not supported.

Note: Throughout this chapter, the terms *adapter* and *card* refer to a PCIe I/O feature that is installed in a PCIe+ I/O drawer.

4.1.1 I/O infrastructure

IBM extends the use of industry standards on the IBM Z platform by offering a Peripheral Component Interconnect Express Generation 3 (PCIe Gen3) I/O infrastructure. The PCIe I/O infrastructure that is provided by the central processor complex (CPC) improves I/O capability and flexibility, while allowing for the future integration of PCIe adapters and accelerators.

The PCIe I/O infrastructure in z14 ZR1 consists of the following components:

- ▶ PCIe fanouts that support 16 GBps I/O bus for CPC drawer connectivity to the PCIe+ I/O drawers.
- ▶ PCIe fanouts that support Coupling Link Feature Integrated Coupling Adapter Short Reach (ICA SR).
- ▶ The 8U, 16-slot, and 2-domain PCIe+ I/O drawer for PCIe I/O features.

The z14 ZR1 I/O infrastructure provides the following benefits:

- ▶ The bus connecting the CPC drawer to the I/O domain in the PCIe+ I/O drawer bandwidth is 16 GBps.
- ▶ The PCIe+ I/O drawer also doubles the number of I/O ports compared to an I/O drawer (z13s or earlier only). Up to 32 channels (16 PCIe I/O cards) are supported in the PCIe+ I/O drawer.
- ▶ Granularity for the storage area network (SAN) and the local area network (LAN):
 - The FICON Express16S+ features two channels per feature for Fibre Channel connection (FICON), High Performance FICON on Z (zHPF), and Fibre Channel Protocol (FCP) storage area networks.
 - The Open Systems Adapter (OSA)-Express6S GbE and the OSA-Express6S 1000BASE-T features includes two ports each (LAN connectivity); the OSA-Express7s 25GbE and the OSA-Express6S 10 GbE features have one port each (LAN connectivity).
- ▶ Native PCIe features (plugged into the PCIe+ I/O drawer):
 - IBM zHyperLink Express (new for z14 ZR1)
 - 25GbE Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE) Express2 (new)
 - 10 GbE RoCE Express2 (introduced with z14)
 - Coupling Express Long Reach (CE LR) (available on z14 M0x, z14 ZR1, z13, and z13s)
 - zEnterprise Data Compression (zEDC) Express
 - 10 GbE RDMA over RoCE Express (carry forward only)

- Crypto Express6s (introduced with z14)
- Crypto Express5s (carry forward only)

4.1.2 PCIe Generation 3

The PCIe Generation 3 uses 128b/130b encoding for data transmission. This encoding reduces the encoding processor usage to approximately 1.54%, compared to the PCIe Generation 2, which features an encoding processor usage of 20% by using 8b/10b encoding.

The PCIe standard uses a low-voltage differential serial bus. Two wires are used for signal transmission, and a total of four wires (two for transmit and two for receive) form a *lane* of a PCIe link, which is full duplex. Multiple lanes can be aggregated into a larger link width. PCIe supports link widths of 1, 2, 4, 8, 12, 16, and 32 lanes (x1, x2, x4, x8, x12, x16, and x32).

The data transmission rate of a PCIe link is determined by the link width (numbers of lanes), the signaling rate of each lane, and the signal encoding rule. The signaling rate of one PCIe Generation 3 lane is 8 gigatransfers per second (GTps), which means that nearly 8 gigabits are transmitted per second (Gbps).

A PCIe Gen3 x16 link features the following data transmission rates:

- The maximum theoretical data transmission rate per lane:
 $8 \text{ Gbps} * 128/130 \text{ bit (encoding)} = 7.87 \text{ Gbps} = 984.6 \text{ MBps}$
- The maximum theoretical data transmission rate per link:
 $984.6 \text{ MBps} * 16 \text{ (lanes)} = 15.75 \text{ GBps}$

Considering that the PCIe link is full-duplex mode, the data throughput rate of a PCIe Gen3 x16 link is 31.5 GBps (15.75 GBps in both directions).

Link performance: The link speeds do not represent the actual performance of the link. The actual performance depends on many factors, including latency through the adapters, cable lengths, and the type of workload.

PCIe Gen3 x16 links are used in z14 ZR1 servers for driving the PCIe I/O drawers, and for coupling links for CPC to CPC communications.

Note: Unless specified otherwise, *PCIe* refers to PCIe Generation 3 in remaining sections of this chapter.

4.2 I/O system overview

The z14 ZR1 I/O characteristics and supported features are described in this section.

4.2.1 Characteristics

The z14 ZR1 I/O subsystem is designed to provide great flexibility, high availability, and the following excellent performance characteristics:

- ▶ High bandwidth

Link performance: The link speeds do not represent the actual performance of the link. The actual performance depends on many factors that include latency through the adapters, cable lengths, and the type of workload.

z14 ZR1 servers use PCIe as an internal interconnect protocol to drive PCIe+ I/O drawers and CPC to CPC connections. The I/O bus infrastructure data rate of up to 128 GBps per system (eight PCIe Gen3 fanout slots). For more information about coupling link connectivity, see 4.6.4, “Parallel Sysplex connectivity” on page 153.

- ▶ Connectivity options:
 - z14 ZR1 servers can be connected to an extensive range of interfaces, such as FICON/FCP for SAN connectivity, 10 Gigabit Ethernet, Gigabit Ethernet, and 1000BASE-T Ethernet for LAN connectivity, zHyperLink Express for storage connectivity (low latency compared to FICON).
 - For CPC to CPC connections, z14 ZR1 servers use Integrated Coupling Adapter (ICA SR) and the Coupling Express Long Reach (CE LR). The Parallel Sysplex InfiniBand is not supported.
 - The 25GbE RoCE Express2, 10GbE RoCE Express2, and 10GbE RoCE Express features provide high-speed memory-to-memory data exchange to a remote CPC by using the Shared Memory Communications over RDMA (SMC-R) protocol for TCP (socket-based) communications.
- ▶ Concurrent I/O upgrade

You can concurrently add I/O features to z14 ZR1 servers if unused I/O slot positions are available.
- ▶ Concurrent PCIe+ I/O drawer upgrade

Extra PCIe+ I/O drawers can be installed concurrently if free frame slots for the PCIe+ I/O drawers and PCIe fanouts in the CPC drawer are available.
- ▶ Dynamic I/O configuration

Dynamic I/O configuration supports the dynamic addition, removal, or modification of the channel path, control units, and I/O devices without a planned outage.
- ▶ Pluggable optics:
 - The FICON Express16S+ FICON Express16S and FICON Express8S, OSA Express7S, OSA Express6S, OSA Express5S, RoCE Express2, and RoCE Express features include Small Form-Factor Pluggable (SFP) optics.¹ These optics allow each channel to be individually serviced in a fiberoptic module failure. The traffic on the other channels on the same feature can continue to flow if a channel requires servicing.

¹ OSA-Express4S, 5S and 6S 1000BASE-T features do not have optics (copper only; RJ45 connectors).

- For zHyperLink Express, it uses the cable with MTP² connector and the cable goes to a CXP³ optics are provided with the adapter.
- ▶ Concurrent I/O card maintenance

Every I/O card that is plugged in PCIe+ I/O drawer supports concurrent card replacement during a repair action.

4.2.2 Supported I/O features

The following I/O features are supported:

- ▶ Up to 128 FICON Express16S+ channels
- ▶ Up to 128 FICON Express16S channels
- ▶ Up to 128 FICON Express8S channels
- ▶ Up to 48 OSA-Express7S 25GbE SR ports
- ▶ Up to 96 OSA-Express6S ports
- ▶ Up to 96 OSA-Express5S ports
- ▶ Up to 8 zEDC Express features
- ▶ Up to four 25GbE RoCE Express2 features
- ▶ Up to four 10GbE RoCE Express2 features
- ▶ Up to four 10GbE RoCE Express features
- ▶ Up to 16 zHyperLink Express features
- ▶ Up to 8 ICA SR features with up to 16 coupling links
- ▶ Up to 16 CE LR features with up to 32 coupling links

Notes: The maximum number of coupling CHPIDs on a z14 ZR1 server is 176, which is a combination of the following ports (not all combinations are possible; subject to I/O configuration options):

- ▶ Up to 8 ICA SR ports
- ▶ Up to 16 CE LR ports

IBM Virtual Flash Memory replaces IBM zFlash Express feature on z14 ZR1 servers.

RoCE Express2 (new) can be mixed with RoCE Express (carry forward only), but the maximum (combined) number of RoCE and RoCE2 features is four.

² Multifiber Termination Push-On.

³ For more information, see:

http://www.infinibandta.org/content/pages.php?pg=technology_public_specification

4.3 PCIe+ I/O drawer

The PCIe+ I/O drawers (see Figure 4-1) are attached to the CPC drawer through a PCIe cable and use PCIe as the infrastructure bus within the drawer. The PCIe I/O bus infrastructure data rate is up to 16 GBps.

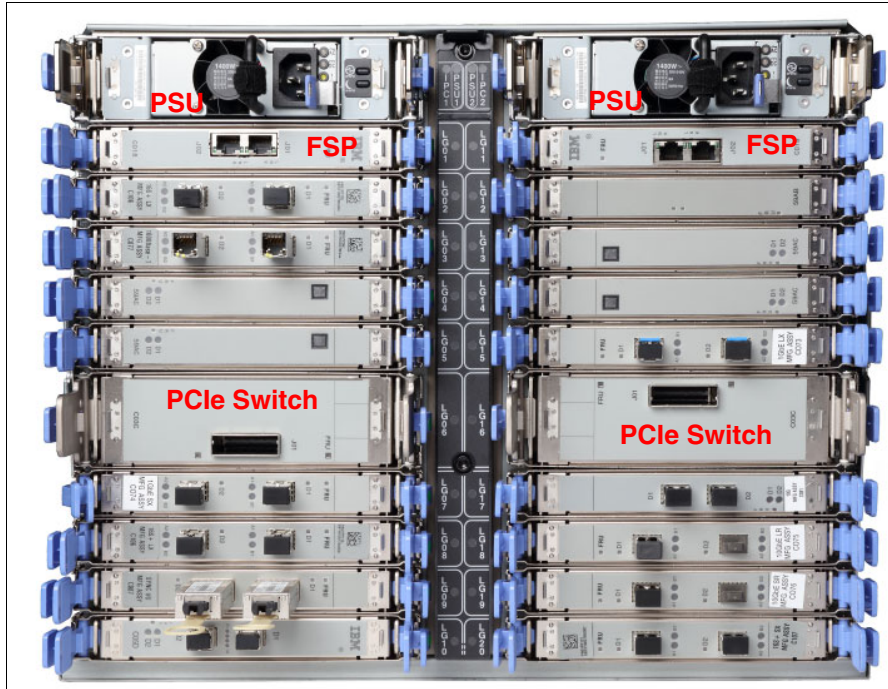


Figure 4-1 Rear view of PCIe+ I/O drawer

PCIe switch application-specific integrated circuits (ASICs) are used to fan out the host bus from the CPC drawer through the PCIe+ I/O drawer to the individual I/O features. Maximum 16 PCIe I/O features (up to 32 channels) per PCIe+ I/O drawer are supported.

The PCIe+ I/O drawer is a one-sided drawer (all I/O cards on one side, in the rear of the drawer) that is 8U high. The PCIe+ I/O drawer contains the 16 I/O slots for PCIe features, two switch cards, and two power supply units (PSUs) to provide redundant power, as shown in Figure 4-1.

The PCIe I/O drawer slots numbers are shown in Figure 4-2.

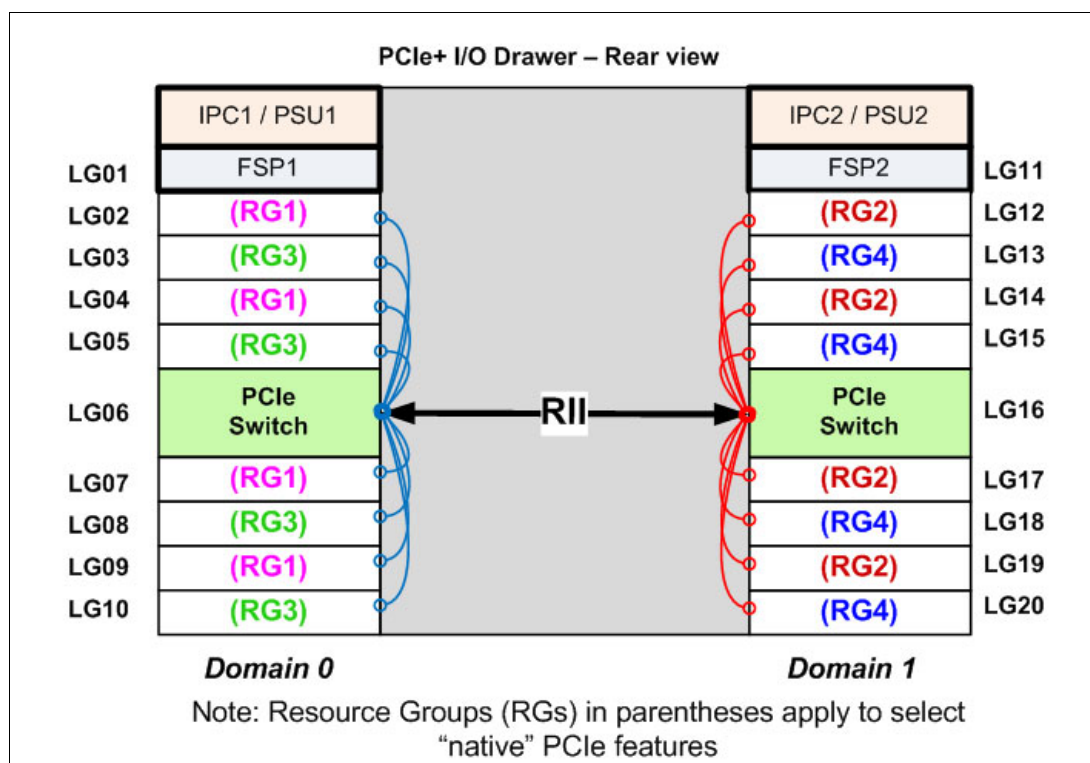


Figure 4-2 PCIe+ I/O drawer slots numbers

The I/O structure in a z14 ZR1 CPC is shown in Figure 4-3 on page 124. 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 CPC drawer through a single x16 PCIe Gen 3 bus from a PCIe fanout card.

In the PCIe+ I/O drawer, the eight I/O feature cards that directly attach to the switch card constitute an I/O domain. The PCIe+ I/O drawer supports concurrent add and replace I/O features to with which you can increase I/O capability as needed depending on the number PU SCM. Therefore, consider planning ahead.

Note: The number of fanout cards is related to the number of installed PU SCMs:

- ▶ FC 0636: 1 PU + 1 SC SCMs → 2 PCIe Fanouts
- ▶ FC 0637: 2 PU + 1 SC SCMs → 4 PCIe Fanouts
- ▶ FC 0638: 4 PU + 1 SC SCMs → 8 PCIe Fanouts
- ▶ FC 0639: 4 PU + 1 SC SCMs → 8 PCIe Fanouts

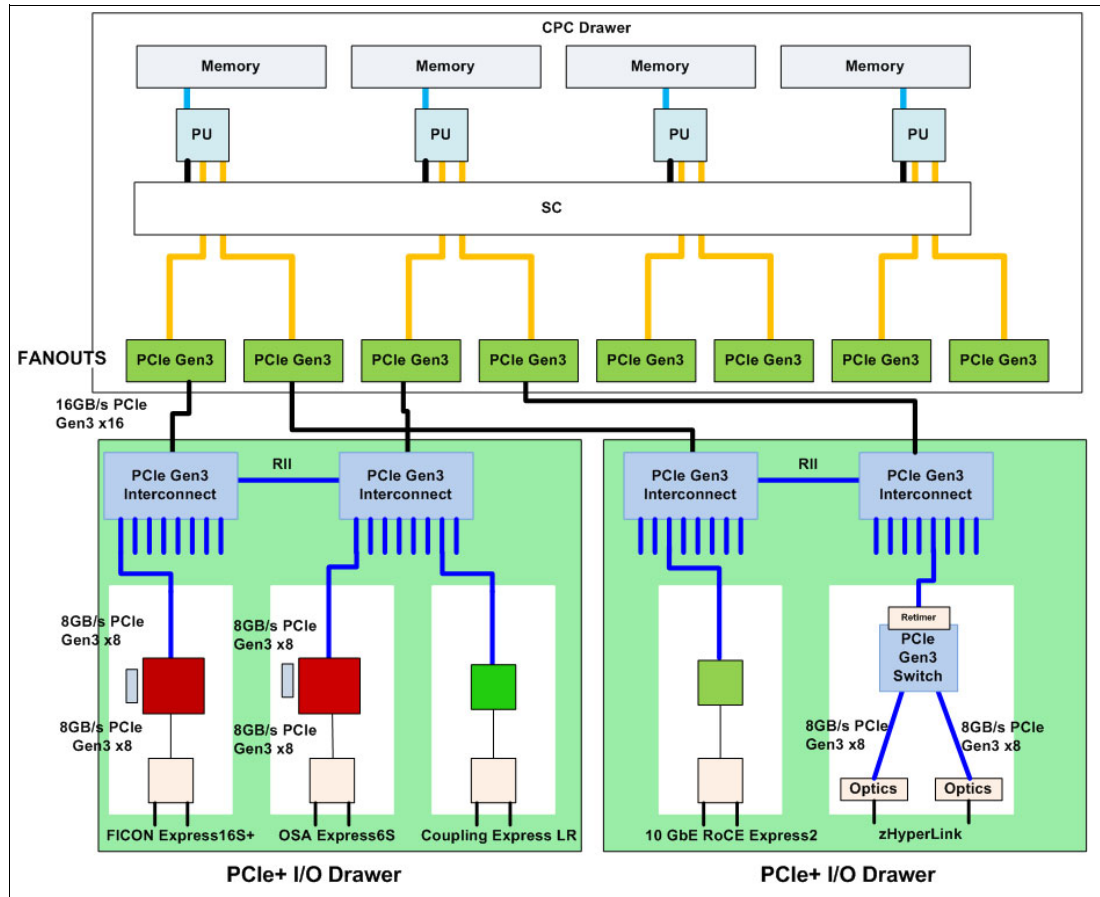


Figure 4-3 z14 ZR1 I/O connectivity

The PCIe slots in a drawer are organized into two I/O domains. Each I/O domain supports up to eight features and is driven through a PCIe switch card. Two PCIe switch cards always provide a backup path for each other through the passive connection in the PCIe+ I/O drawer backplane. During a PCIe fanout card or cable failure, 16 I/O cards in two domains can be driven through a single PCIe switch card.

The two switch cards are interconnected through the PCIe+ I/O drawer board (Redundant I/O Interconnect, or RII). In addition, switch cards in same PCIe+ I/O drawer are connected to PCIe fanouts across clusters in CPC drawer for higher availability.

The RII design provides a failover capability during a PCIe fanout card failure. Both domains in one of these PCIe+ I/O drawers are activated with two fanouts. The flexible service processors (FSPs) are used for system control.

The domains and their related I/O slots are shown in Figure 4-2 on page 123.

Each I/O domain supports up to eight features (FICON, OSA, Crypto, and so on.) All I/O cards connect to the PCIe switch card through the backplane board. The I/O domains and slots are listed in Table 4-1 on page 125.

Table 4-1 I/O domains of PCIe+ I/O drawer

Domain	I/O slot in domain
0	LG02, LG03, LG04, LG05, LG07, LG08, LG09, and LG10
1	LG12, LG13, LG14, LG15, LG17, LG18, LG19, and LG20

4.3.1 PCIe+ I/O drawer offerings

A maximum of four PCIe+ I/O drawers can be installed for supporting up to 64 PCIe I/O features.

For an upgrade to z14 ZR1 servers, only the following PCIe features can be carried forward:

- ▶ FICON Express16S
- ▶ FICON Express8S
- ▶ OSA-Express5S (all 5S features)
- ▶ OSA-Express4S (all, except for 1000BASE-T)
- ▶ 10GbE RoCE Express
- ▶ Crypto Express5S
- ▶ zEDC Express
- ▶ Coupling Express Long Reach (CE LR)

Consideration: On a z14 ZR1 server, only PCIe+ I/O drawers are supported. No older generation drawers can be carried forward.

A new build IBM z14 ZR1 server supports the following PCIe I/O features that are hosted in the PCIe+ I/O drawers:

- ▶ FICON Express16S+
- ▶ OSA-Express7S 25GbE SR
- ▶ OSA-Express6S
- ▶ 25GbE RoCE Express2
- ▶ 10GbE RoCE Express2
- ▶ Crypto Express6S
- ▶ zEDC Express
- ▶ Coupling Express Long Reach (CE LR)
- ▶ zHyperLink Express

4.4 CPC drawer fanouts

The z14 ZR1 server uses fanout cards to connect the I/O subsystem to the CPC drawer. The fanout cards also provide the ICA SR coupling links for Parallel Sysplex. All fanout cards support concurrent add, delete, and move.

The z14 ZR1 CPC drawer I/O infrastructure consists of the following features:

- ▶ The PCIe Generation 3 fanout cards: One port card (feature) that connects to a PCIe+ I/O drawer supporting an eight-slot I/O domain. This card is always installed in pairs to support I/O domain redundant connectivity.
- ▶ ICA SR fanout cards: two ports per card (feature) that connect to other (external) CPC drawers.

Note: IBM z14 ZR1 does not support HCA3-O and HCA3-O LR adapters.

The PCIe fanouts cards are installed in the rear of the CPC drawer (rack location A09B). The CPC drawer has eight PCIe Gen3 fanout slots.

Important: The number of available (usable) PCIe fanout slots depends on the number of PU SCMs that is installed in the CPC drawer (CPC drawer features 0636, 0637, 0638, and 0639).

The PCIe fanout and ICA SR fanouts are installed in locations LG01 - LG04, and LG07 - LG10 left to right in the CPC drawer (see Figure 2-20 on page 49). Slots LG05 and LG06 are used for FSP. Slots LG11 and LG12 are used for Oscillator cards.

Two types of fanout cards are supported by z14 ZR1 servers. Each CPC drawer fanout slot can hold one of the following fanouts:

- ▶ PCIe Gen3 fanout card: This copper fanout provides connectivity to the PCIe switch card in the PCIe I/O drawer.
- ▶ Integrated Coupling Adapter (ICA SR): This adapter provides coupling connectivity between z14 ZR1, z14, z13 and z13s servers, up to 150 meters (492 feet), 8 GBps link rate.

The PCIe Gen3 fanout card features one port. The ICA SR fanout card includes two ports.

An I/O connection scheme is shown in Figure 4-3 on page 124.

4.4.1 PCIe Generation 3 fanout (FC 0173)

The PCIe Gen3 fanout card provides connectivity to an PCIe+ I/O drawer by using a copper cable. One port on the fanout card is dedicated for PCIe I/O. The bandwidth of this PCIe fanout card supports a link rate of 16 GBps.

A 16x PCIe copper cable of 1.5 meters (4.92 feet) to 4.0 meters (13.1 feet) is used for connection to the PCIe switch card in the PCIe+ I/O drawer. PCIe fanout cards are always plugged in pairs and provide redundancy for I/O domains within the PCIe+ I/O drawer.

PCIe fanout: The PCIe fanout is used exclusively for I/O and cannot be shared for any other purpose.

4.4.2 Integrated Coupling Adapter (FC 0172)

Introduced with IBM z13, the IBM ICA SR is a two-port fanout feature that is used for short distance coupling connectivity and uses channel type CS5.

The ICA SR uses PCIe Gen3 technology, with x16 lanes that are bifurcated into x8 lanes for coupling. No performance degradation is expected compared to the coupling over InfiniBand 12x IFB3 protocol.

The ICA SR is designed to drive distances up to 150 meters (492 feet) with a link data rate of 8 GBps. ICA SR supports up to four channel-path identifiers (CHPIDs) per port and eight subchannels (devices) per CHPID.

The coupling links can be defined as shared between images within a CSS. They also can be spanned across multiple CSSs in a CPC. Unlike the HCA3-O 12x InfiniBand links, the ICA SR cannot define more than four CHPIDS per port. When STP is enabled, ICA SR coupling links can be defined as timing-only links to other z14 ZR1, z14, and z13/z13s CPCs.

The ICA SR fanout is housed in the PCIe I/O fanout slot on the z14 ZR1 CPC drawer. Up to eight ICA SR fanouts and up to 16 ICA SR ports are supported on a z14 ZR1 CPC drawer. This configuration enables greater connectivity for short distance coupling on a single processor node compared to previous generations. The maximum number of ICA SR fanout features is 16 per system for z14 ZR1 servers.

The ICA SR can be used for coupling connectivity between z14/z14 ZR1 and z13/z13s servers. It does not support connectivity to zEC12, zBC12 servers. It also cannot be connected to HCA3-O or HCA3-O LR coupling fanouts.

The ICA SR fanout requires cabling that is different from the 12x IFB cables. For distances up to 100 meters (328 feet), OM3 fiber optic can be used. For distances up to 150 meters (492 feet), OM4 or OM5 fiber optic cables can be used. For more information, see the following resources:

- ▶ *Planning for Fiber Optic Links*, GA23-1407
- ▶ *IBM 3907 Installation Manual for Physical Planning*, GC28-6974

4.4.3 Fanout considerations

Fanout slots in the CPC drawer can be used to plug different fanouts. The CPC drawer can hold up to eight PCIe fanout cards.

Adapter ID number assignment

PCIe fanouts and ports are identified by an Adapter ID (AID) that is initially dependent on their physical locations, which is unlike channels that are installed in a PCIe+ I/O drawer. Those channels are identified by a physical channel ID (PCHID) number that is related to their physical location. This AID must be used to assign a CHPID to the fanout in the IOCDS definition. The CHPID assignment is done by associating the CHPID to an AID port (see Table 4-2).

Table 4-2 AIDs and fanout locations

	Fanout Slots							
Features	LG01	LG02	LG03	LG04	LG07	LG08	LG09	LG10
Max4	Not populated						16	17
Max12	Not populated				14	15	16	17
Max24	10	11	12	13	14	15	16	17
Max30	10	11	12	13	14	15	16	17

Fanout slots

The fanout slots are numbered LG01 - LG04 and LG07 - LG10, from left to right (LG05 and LG06 are used for FSPs), as shown in Figure 4-2 on page 127. All fanout locations and their AIDs for the CPC drawer are shown for reference only. Slots LG05 and LG06 never include a fanout that is installed because they are dedicated for FSPs.

Important: The AID numbers that are listed in Table 4-2 on page 127 are valid only for a new build system. If a fanout is moved, the AID follows the fanout to its new physical location.

The AID assignment is listed in the PCHID REPORT that is provided for each new server or for an MES upgrade on existing servers. Part of a PCHID REPORT for a z14 ZR1 is shown in Example 4-1. In this example, four fanout cards are installed at locations LG01, LG03, LG07, and LG09 with AIDs 10, 12, 14, and 16.

Example 4-1 AID assignment in PCHID REPORT

CHPIDSTART				PCHID REPORT		Mar 14,2018	
07356479							
Machine: 3907-ZR1 NEW1							

Source	Drwr	Slot	F/C	PCHID/Ports	or AID	Comment	
A09/LG01	A09B	LG01	0172	AID=10			
A09/LG03	A09B	LG03	0172	AID=12			
A09/LG07	A09B	LG07	0172	AID=14			
A09/LG09	A09B	LG09	0172	AID=16			

Fanout features that are supported by the z14 ZR1 server are listed in Table 4-3, which includes the feature type, feature code, and information about the link supported by the fanout feature.

Table 4-3 Fanout summary

Fanout feature	Feature code	Use	Cable type	Connector type	Maximum distance	Link data rate ^a
PCIe fanout	0173	Connect to PCIe I/O drawer	Copper	N/A	4 m (13.1 ft)	16 Gbps
ICA SR	0172	Coupling link	OM4, OM5	MTP	150 m (492 ft)	8 Gbps
			OM3	MTP	100 m (328 ft)	8 Gbps

a. The link data rates do not represent the actual performance of the link. The actual performance depends on many factors, including latency through the adapters, cable lengths, and the type of workload.

4.5 I/O features (cards)

I/O features (adapters) include ports⁴ to connect the z14 ZR1 server to external devices, networks, or other servers. I/O features are plugged into the PCIe+ I/O drawer, based on the configuration rules for the server. Different types of I/O cards are available, one for each channel or link type. I/O cards can be installed or replaced concurrently.

⁴ Certain I/O features do not have external ports, such as Crypto Express and zEDC.

4.5.1 I/O feature card ordering information

The I/O features that are supported by z14 ZR1 servers and the ordering information for them are listed in Table 4-4.

Table 4-4 I/O features and ordering information

Channel feature	Feature code	New build	Carry-forward
FICON Express16S+ LX	0427	Y	N/A
FICON Express16S+ SX	0428	Y	N/A
FICON Express16S 10KM LX	0418	N	Y
FICON Express16S SX	0419	N	Y
FICON Express8S 10KM LX	0409	N	Y
FICON Express8S SX	0410	N	Y
OSA-Express7S 25GbE SR	0429	Y	N/A
OSA-Express6S 10GbE LR	0424	Y	N/A
OSA-Express6S 10GbE SR	0425	Y	N/A
OSA-Express6S GbE LX	0422	Y	N/A
OSA-Express6S GbE SX	0423	Y	N/A
OSA-Express6S 1000BASE-T Ethernet	0426	Y	N/A
OSA-Express5S 10GbE LR	0415	N	Y
OSA-Express5S 10GbE SR	0416	N	Y
OSA-Express5S GbE LX	0413	N	Y
OSA-Express5S GbE SX	0414	N	Y
OSA-Express5S 1000BASE-T Ethernet	0417	N	Y
OSA-Express4S 10GbE LR	0406	N	Y
OSA-Express4S 10GbE SR	0407	N	Y
OSA-Express4S GbE LX	0404	N	Y
OSA-Express4S GbE SX	0405	N	Y
Integrated Coupling Adapter (ICA SR)	0172	Y	Y
Coupling Express LR	0433	Y	Y
Crypto Express6S	0893	Y	N/A
Crypto Express5S	0890	N	Y
25GbE RoCE Express2	0430	Y	N/A
10GbE RoCE Express2	0412	Y	N/A
10GbE RoCE Express	0411	N	Y
zEDC Express	0420	Y	Y
zHyperLink Express	0431	Y	N/A

Important: z14 ZR1 servers do not support the ISC-3, HCA2-O (12x), HCA3-O (12x), HCA2-O LR (1x), or HCA3-O (1x) features and cannot participate in a Mixed Coordinated Timing Network (CTN). z196, z114, and older CPCs *cannot* coexist in the same Parallel Sysplex or STP CTN with z14 ZR1 (no coupling connectivity). zEC12 or zBC12 can coexist in the same Parallel Sysplex with z14 ZR1 only if the CPC hosting the CFs includes coupling connectivity to the zEC12/zBC12 and z14 ZR1 CPCs (the CF LPAR cannot be on zEC12/BC12 or the z14 ZR1 CPCs).

4.5.2 Physical channel ID report

A physical channel ID (PCHID) reflects the physical location of a channel-type interface. A PCHID number is based on the following factors:

- ▶ PCIe+ I/O drawer location
- ▶ Channel feature slot number
- ▶ Port number of the channel feature

A CHPID does not directly correspond to a hardware channel port. Instead, it is assigned to a PCHID in the hardware configuration definition (HCD) or IOCP.

A PCHID REPORT is created for each new build server and for upgrades on servers. The report lists all I/O features that are installed, the physical slot location, and the assigned PCHID. A portion of a sample PCHID REPORT is shown in Example 4-2. For more information about the AID numbering rules for coupling links, see Example 4-2.

Example 4-2 PCHID REPORT

CHPIDSTART				PCHID REPORT		Apr 23,2018
11357270						
Machine: 3907-ZR1 SN1						

Source	Drwr	Slot	F/C	PCHID/Ports or AID		Comment
A09/LG01	A09B	LG01	0172	AID=10		
A09/LG07	A09B	LG07	0172	AID=14		
A09/LG04/J01	A14B	02	0428	100/D1 101/D2		
A09/LG04/J01	A14B	05	0425	10C/D1		
A09/LG04/J01	A14B	07	0431	110/D1D2		
A09/LG04/J01	A14B	08	0893	114/P00		
A09/LG04/J01	A14B	09	0412	118/D1D2		RG1
A09/LG04/J01	A14B	10	0433	11C/D1D2		RG3
A09/LG10/J01	A14B	12	0428	120/D1 121/D2		
A09/LG10/J01	A14B	13	0426	124/D1D2		
A09/LG10/J01	A14B	14	0431	128/D1D2		
A09/LG10/J01	A14B	15	0423	12C/D1D2		

A09/LG10/J01	A14B	17	0420	130	RG2
A09/LG10/J01	A14B	18	0425	134/D1	
A09/LG10/J01	A14B	19	0433	138/D1D2	RG2
A09/LG10/J01	A14B	20	0412	13C/D1D2	RG4
A09/LG02/J01	A01B	02	0428	140/D1 141/D2	
A09/LG02/J01	A01B	03	0428	144/D1 145/D2	
A09/LG02/J01	A01B	05	0423	14C/D1D2	
A09/LG02/J01	A01B	07	0425	150/D1	
A09/LG02/J01	A01B	08	0420	154	RG3
A09/LG02/J01	A01B	09	0433	158/D1D2	RG1
A09/LG02/J01	A01B	10	0433	15C/D1D2	RG3
A09/LG08/J01	A01B	12	0426	160/D1D2	
A09/LG08/J01	A01B	13	0428	164/D1 165/D2	
A09/LG08/J01	A01B	14	0423	168/D1D2	
A09/LG08/J01	A01B	17	0431	170/D1D2	
A09/LG08/J01	A01B	18	0893	174/P00	
A09/LG08/J01	A01B	19	0412	178/D1D2	RG2
A09/LG08/J01	A01B	20	0433	17C/D1D2	RG4
A09/LG03/J01	A23B	02	0426	180/D1D2	
A09/LG03/J01	A23B	03	0428	184/D1 185/D2	
A09/LG03/J01	A23B	04	0423	188/D1D2	
A09/LG03/J01	A23B	07	0420	190	RG1
A09/LG03/J01	A23B	08	0425	194/D1	
A09/LG03/J01	A23B	09	0433	198/D1D2	RG1
A09/LG03/J01	A23B	10	0412	19C/D1D2	RG3
A09/LG09/J01	A23B	12	0428	1A0/D1 1A1/D2	
A09/LG09/J01	A23B	13	0428	1A4/D1 1A5/D2	

A09/LG09/J01	A23B	15	0431	1AC/D1D2	
A09/LG09/J01	A23B	17	0426	1B0/D1D2	
A09/LG09/J01	A23B	18	0420	1B4	RG4
A09/LG09/J01	A23B	19	0433	1B8/D1D2	RG2
A09/LG09/J01	A23B	20	0433	1BC/D1D2	RG4

Legend:

Source	Book Slot/Fanout Slot/Jack
A09B	CEC Drawer 1 in A frame
A14B	PCIe Drawer 1 in A frame
A01B	PCIe Drawer 2 in A frame
A23B	PCIe Drawer 3 in A frame
0426	OSA Express6S 1000BASE T 2 Ports
0428	16GB FICON Express16S+ SX 2 Ports
0423	OSA Express6S GbE SX 2 Ports
RG1	Resource Group 1
0420	zEDC Express
0425	OSA Express6S 10 GbE SR 1 Ports
0433	Coupling Express LR
RG3	Resource Group 3
0412	10GbE RoCE Express
0431	zHyperLink Express
RG4	Resource Group 4
RG2	Resource Group 2
0893	Crypto Express6S
0172	ICA SR 2 Links

The PCHID REPORT that is shown in Example 4-2 on page 130 includes the following components (among others):

- ▶ Feature codes 0172 (Integrated Coupling Adapters (ICA SR) is installed in the CPC drawer (location A09B, slots LG01 and LG07), and have AIDs 10 and 14 assigned.
- ▶ Feature codes 0425 (OSA-Express6S 10GbE SR) are installed in PCIe+ I/O drawer 1 (location A14B, slots LG05 and LG18 with PCHIDs 10C/D1 and 134/D1 assigned), PCIe+ drawer 2 (location A01B, slot LG07 with PCHID 150/D1 assigned) and drawer 3 (location A23B, slot LG08 with PCHID 194/D1 assigned).
- ▶ Feature code 0428 (FICON Express16S+ short wavelength (SX) 150 m / 492 ft.) are installed in PCIe+ I/O drawer 1 (location A14B, slots LG01 and LG12, with PCHIDs 100/D1, 101/D2, 120/D1, and 121/D2 assigned), PCIe+ I/O drawer 2 (location A01B, slots LG02, LG03, and LG13, with PCHIDs 140/D1, 141/D2, 144/D1, 145/D2, 164/D1, and 165/D2 assigned), and PCIe+ I/O drawer 3 (location A23B, slots LG02, LG12, and LG13, with PCHIDs 184/D1, 185/D2, 1A0/D1, 1A1/D2, 1A4/D1, and 1A5/D2 assigned).

A resource group (RG) parameter is shown in the PCHID REPORT for native PCIe features. A balanced plugging of native PCIe features exists between four resource groups (RG1, RG2, RG3, and RG4).

For more information about resource groups, see Appendix C, “Native Peripheral Component Interconnect Express” on page 419.

The preassigned PCHID number of each I/O port relates directly to its physical location (jack location in a specific slot).

4.6 Connectivity

I/O channels are part of the CSS. They provide connectivity for data exchange between servers, between servers and external control units (CUs) and devices, or between networks.

For more information about connectivity to external I/O subsystems (for example, disks), see “Storage connectivity” on page 136.

For more information about communication to LANs, see “Network connectivity” on page 142.

Communication between servers is implemented by using CE LR, ICA SR, or channel-to-channel (CTC) connections. For more information, see “Parallel Sysplex connectivity” on page 153.

4.6.1 I/O feature support and configuration rules

The supported I/O features are listed in Table 4-5. Also listed in Table 4-5 are the number of ports per card, port increments, the maximum number of feature cards, and the maximum number of channels for each feature type. The CHPID definitions that are used in the IOCDs also are listed.

Table 4-5 z14 ZR1 supported I/O features

I/O feature	Ports per card	Port increments	Max. ports	Max. I/O slots	PCHID	CHPID definition
FICON Express16S+ LX/SX	2	2	128	64	Yes	FC, FCP ^a
FICON Express16S LX/SX	2	2	128	64	Yes	FC, FCP
FICON Express8S LX/SX	2	2	128	64	Yes	FC, FCP
OSA-Express6S 25GbE SR	1	1	48	48	Yes	OSD, OSX
OSA-Express6S 10 GbE LR/SR	1	1	48	48	Yes	OSD, OSX
OSA-Express6S GbE LX/SX	2	2	96	48	Yes	OSD
OSA-Express6S 1000BASE-T	2	2	96	48	Yes	OSC, OSD, OSE, OSM
OSA-Express5S 10 GbE LR/SR	1	1	48	48	Yes	OSD, OSX
OSA-Express5S GbE LX/SX	2	2	96	48	Yes	OSD
OSA-Express5S 1000BASE-T	2	2	96	48	Yes	OSC, OSD, OSE, OSM
OSA-Express4S 10 GbE LR/SR	1	1	48	48	Yes	OSD, OSX

I/O feature	Ports per card	Port increments	Max. ports	Max. I/O slots	PCHID	CHPID definition
OSA-Express4S GbE LX/SX	2	2	96	48	Yes	OSD
25GbE RoCE Express2	2	2	8	4	Yes	N/A ^b
10GbE RoCE Express2	2	2	8	4	Yes	N/A ^b
10GbE RoCE Express	2	2	8	4	Yes	N/A ^b
Coupling Express LR	2	2	32	16	Yes	CL5
Integrated Coupling Adapter (ICA SR)	2	2	16	8	N/A	CS5
zHyperLink Express	2	2	32	16	Yes	N/A ^b

a. Both ports must be defined with the same CHPID type.

b. These features are defined by using Virtual Functions IDs (FIDs).

At least one I/O feature (FICON) or one coupling link feature (ICA SR or CE LR) must be present in the minimum configuration.

The following features can be shared and spanned:

- ▶ FICON channels that are defined as FC or FCP
- ▶ OSA-Express features that are defined as OSC, OSD, OSE, OSM, or OSX
- ▶ Coupling links that are defined as CS5 or CL5
- ▶ 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 (5S/6S) feature occupies one I/O slot, but does not include a CHPID type. However, LPARs in all CSSs can access the features. Each Crypto Express adapter can be defined to up to 40 LPARs.
- ▶ Each RoCE Express/Express2 feature occupies one I/O slot but does not include a CHPID type. However, LPARs in all CSSs can access the feature. The 10GbE RoCE Express can be defined to up to 31 LPARs per feature (port is defined in z/OS Communications Server). The 25 GbE RoCE Express2 and the 10GbE RoCE Express2 features support up to 31 LPARs per port (up to 62 LPARs per feature).
- ▶ Each zEDC Express feature occupies one I/O slot but does not include a CHPID type. However, LPARs in all CSSs can access the feature. The zEDC feature can be defined to up to 15 LPARs.
- ▶ Each zHyperLink Express feature occupies one I/O slot but does not include a CHPID type. However, LPARs in all CSSs can access the feature. The zHyperLink Express adapter works as native PCIe adapter and can be shared by multiple LPARs. Each port supports up to 127 Virtual Functions (VFs), with one or more VFs/PFIDs being assigned to each LPAR. This support gives a maximum of 254 VFs per adapter.

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, SAN, LAN, or the end-to-end enterprise.

Cables: All fiber optic cables, cable planning, labeling, and installation are client responsibilities for new z14 ZR1 installations and upgrades. Fiber optic conversion kits and mode conditioning patch cables are not orderable as features on z13 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 z14 ZR1 server for quick connection. The FQC is offered as a feature on z13 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 the need to extend the maintenance window.

The required connector and cable type for each I/O feature on z14 ZR1 servers are listed in Table 4-6.

Table 4-6 I/O feature connector and cable types

Feature code	Feature name	Connector type	Cable type
0427	FICON Express16S+ LX 10 km	LC Duplex	9 µm SM
0428	FICON Express16S+ SX	LC Duplex	50, 62.5 µm MM
0418	FICON Express16S LX 10 km	LC Duplex	9 µm SM
0419	FICON Express16S SX	LC Duplex	50, 62.5 µm MM
0409	FICON Express8S LX 10 km	LC Duplex	9 µm SM
0410	FICON Express8S SX	LC Duplex	50, 62.5 µm MM
0429	OSA-Express7S 25GbE SR	LC Duplex	50 µm MM OM4 ^b
0424	OSA-Express6S 10GbE LR	LC Duplex	9 µm SM
0425	OSA-Express6S 10 GbE SR	LC Duplex	50, 62.5 µm MM
0422	OSA-Express6S GbE LX	LC Duplex	9 µm SM
0423	OSA-Express6S GbE SX	LC Duplex	50, 62.5 µm MM
0426	OSA-Express6S 1000BASE-T	RJ-45	Category 5 UTP ^a
0415	OSA-Express5S 10 GbE LR	LC Duplex	9 µm SM
0416	OSA-Express5S 10 GbE SR	LC Duplex	50, 62.5 µm MM
0413	OSA-Express5S GbE LX	LC Duplex	9 µm SM
0414	OSA-Express5S GbE SX	LC Duplex	50, 62.5 µm MM
0417	OSA-Express5S 1000BASE-T	RJ-45	Category 5 UTP
0406	OSA-Express4S 10 GbE LR	LC Duplex	9 µm SM
0407	OSA-Express4S 10 GbE SR	LC Duplex	50, 62.5 µm MM
0404	OSA-Express4S GbE LX	LC Duplex	9 µm SM
0405	OSA-Express4S GbE SX	LC Duplex	50, 62.5 µm MM

Feature code	Feature name	Connector type	Cable type
0430	25GbE RoCE Express2	LC Duplex	50 µm MM OM4 ^b
0412	10GbE RoCE Express2	LC Duplex	50, 62.5 µm MM
0411	10GbE RoCE Express	LC Duplex	50, 62.5 µm MM
0433	CE LR	LC Duplex	9 µm SM
0172	Integrated Coupling Adapter (ICA SR)	MTP	50 µm MM OM4 or OM5 ^b
0431	zHyperLink Express	MPO	50 µm MM OM4 or OM5 ^b

a. UTP is unshielded twisted pair. Consider the use of category 6 UTP for 1000 Mbps connections.

b. Or 50 µm MM OM3, but OM4 or OM5 is highly recommended.

MM = Multi-Mode

SM = Single-Mode

4.6.2 Storage connectivity

Connectivity to external I/O subsystems (for example, disks) is provided by FICON channels and zHyperLink⁵.

FICON channels

z14 ZR1 supports the following FICON features:

- ▶ FICON Express16S+
- ▶ FICON Express16S (carry-forward only)
- ▶ FICON Express8S (carry-forward only)

The FICON Express16S+, FICON Express16S, and FICON Express8S features conform to the following architectures:

- ▶ Fibre Connection (FICON)
- ▶ High Performance FICON on Z (zHPF)
- ▶ Fibre Channel Protocol (FCP)

The FICON features provide connectivity between any combination of servers, directors, switches, and devices (control units, disks, tapes, and printers) in a SAN.

Each FICON Express16S+, FICON Express16S, and FICON Express 8S feature occupies one I/O slot in the PCIe+ I/O drawer. Each feature includes two ports, each supporting an LC Duplex connector, with one PCHID and one CHPID that is associated with each port.

All FICON Express16S+, FICON Express16S, and FICON Express8S features use SFP optics that allow for concurrent repairing or replacement for each SFP. The data flow on the unaffected channels on the same feature can continue. A problem with one FICON port no longer requires replacement of a complete feature.

All FICON Express16S+, FICON Express16S, and FICON Express8S 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.

⁵ zHyperLink feature operates with a FICON channel.

z14 ZR1 servers support 32K devices per FICON channel for all FICON features.

Each FICON Express16S+, FICON Express16S, and FICON Express8S channel can be defined independently for connectivity to servers, switches, directors, disks, tapes, and printers, by using the following CHPID types:

- ▶ CHPID type FC: The FICON, zHPF, and FCTC 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 defined as spanned. All ports on a FICON feature must be of the same type (LX or SX). The features are connected to a FICON capable control unit (point-to-point or switched point-to-point) through a Fibre Channel switch.

FICON Express16S+

The FICON Express16S+ feature is installed in the PCIe+ I/O drawer. Each of the two independent ports is capable of 4 Gbps, 8 Gbps, or 16 Gbps. The link speed depends on the capability of the attached switch or device. The link speed is auto-negotiated, point-to-point, and is transparent to users and applications.

The following types of FICON Express16S+ optical transceivers are supported (no mix on same card):

- ▶ FICON Express16S+ LX feature, FC 0427, with two ports per feature, supporting LC Duplex connectors
- ▶ FICON Express16S+ SX feature, FC 0428, with two ports per feature, supporting LC Duplex connectors

Each port of the FICON Express16S+ LX feature uses an optical transceiver that supports an unrepeated distance of 10 kilometers (6.2 miles) by using 9 µm single-mode fiber.

Each port of the FICON Express16S+ SX feature uses an optical transceiver that supports to up to 125 meters (410 ft.) of distance variable with link data rate and fiber type.

Consideration: FICON Express16S+ features do not support auto-negotiation to a data link rate of 2 Gbps (only 4, 8, or 16 Gbps).

FICON Express16S

The FICON Express16S feature is installed in the PCIe+ I/O drawer. Each of the two independent ports is capable of 4 Gbps, 8 Gbps, or 16 Gbps. The link speed depends on the capability of the attached switch or device. The link speed is auto-negotiated, point-to-point, and is transparent to users and applications.

The following types of FICON Express16S optical transceivers are supported:

- ▶ FICON Express16S LX feature, FC 0418, with two ports per feature, supporting LC Duplex connectors
- ▶ FICON Express16S SX feature, FC 0419, with two ports per feature, supporting LC Duplex connectors

Each port of the FICON Express16S LX feature uses an optical transceiver that supports an unrepeated distance of 10 kilometers (6.2 miles) by using 9 µm single-mode fiber.

Each port of the FICON Express16S SX feature uses an optical transceiver that supports to up to 125 meters (410 ft.) of distance depending on the fiber that is used.

Consideration: FICON Express16S features do not support auto-negotiation to a data link rate of 2 Gbps (only 4, 8, or 16 Gbps).

FICON Express8S

The FICON Express8S feature is installed in the PCIe I/O drawer. Each of the two independent ports is capable of 2 Gbps, 4 Gbps, or 8 Gbps. The link speed depends on the capability of the attached switch or device. The link speed is auto-negotiated, point-to-point, and is transparent to users and applications.

The following types of FICON Express8S optical transceivers are supported:

- ▶ FICON Express8S 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 LX feature uses an optical transceiver that supports an unrepeated distance of 10 kilometers (6.2 miles) by using 9 µm single-mode fiber.

Each port of the FICON Express8S SX feature uses an optical transceiver that supports up to 150 meters (492 feet) of distance depending on the fiber used.

FICON enhancements

Together with the FICON Express16S+, z14 ZR1 servers provide enhancements for FICON in functional and performance aspects.

Forward Error Correction

Forward Error Correction (FEC) is a technique that is used for reducing data errors when transmitting over unreliable or noisy communication channels (improving signal to noise ratio). By adding redundancy error-correction code (ECC) to the transmitted information, the receiver can detect and correct several errors without requiring retransmission. This process features improve signal reliability and bandwidth use by reducing retransmissions because of bit errors, especially for connections across long distance, such as an inter-switch link (ISL) in a GDPS Metro Mirror environment.

The FICON Express16S+ and FICON Express16S are designed to support FEC coding on top of its 64b/66b data encoding for 16Gbps connections. This design can correct up to 11 bit errors per 2112 bits transmitted. Therefore, while connected to devices that support FEC at 16 Gbps connections, the FEC design allows FICON Express16S+ and FICON Express16S channels to operate at higher speeds, over longer distances, with reduced power and higher throughput while retaining the same reliability and robustness for which FICON channels are traditionally known.

With the IBM DS8870 or newer, IBM z14 ZR1 servers can extend the use of FEC to the fabric N_Ports for a completed end-to-end coverage of 16 Gbps FC links. For more information, see the *IBM DS8884 and z13s: A new cost optimized solution*, REDP-5327.

FICON dynamic routing

With the IBM z14 ZR1, IBM z14 M0x, IBM z13, and IBM z13s servers, FICON channels are no longer restricted to the use of static SAN routing policies for ISLs for cascaded FICON directors. The Z servers now support dynamic routing in the SAN with the FICON Dynamic Routing (FIDR) feature. It is designed to support the dynamic routing policies that are provided by the FICON director manufacturers; for example, Brocade's exchange-based routing (EBR) and Cisco's originator exchange ID (OxID)⁶ routing.

A static SAN routing policy normally assigns the ISL routes according to the incoming port and its destination domain (port-based routing), or the source and destination ports pairing (device-based routing).

The port-based routing (PBR) assigns the ISL routes statically that is based on “first come, first served” when a port starts a fabric login (FLOGI) to a destination domain. The ISL is round-robin that is selected for assignment. Therefore, I/O flow from same incoming port to same destination domain always is assigned the same ISL route, regardless of the destination port of each I/O. This setup can result in some ISLs overloaded while some are under-used. The ISL routing table is changed whenever Z server undergoes a power-on-reset (POR), so the ISL assignment is unpredictable.

Device-based routing (DBR) assigns the ISL routes statically that is based on a hash of the source and destination port. That I/O flow from same incoming port to same destination is assigned to same ISL route. Compared to PBR, the DBR is more capable of spreading the load across ISLs for I/O flow from the same incoming port to different destination ports within a destination domain.

When a static SAN routing policy is used, the FICON director features limited capability to assign ISL routes based on workload. This limitation can result in unbalanced use of ISLs (some might be overloaded, while others are under-used).

The dynamic routing ISL routes are dynamically changed based on the Fibre Channel exchange ID, which is unique for each I/O operation. ISL is assigned at I/O request time, so different I/Os from same incoming port to same destination port are assigned different ISLs.

With FIDR, z14 ZR1 servers feature the following advantages for performance and management in configurations with ISL and cascaded FICON directors:

- ▶ Support sharing of ISLs between FICON and FCP (PPRC or distributed)
- ▶ I/O traffic is better balanced between all available ISLs
- ▶ Improved use of FICON director and ISL
- ▶ Easier to manage with a predictable and repeatable I/O performance

FICON dynamic routing can be enabled by defining dynamic routing-capable switches and control units in HCD. Also, z/OS implemented a health check function for FICON dynamic routing.

Improved zHPF I/O execution at distance

By introducing the concept of pre-deposit writes, zHPF reduces the number of round trips of standard FCP I/Os to a single round trip. Originally, this benefit is limited to writes that are less than 64 KB. zHPF on z14 ZR1, z14 M0x, z13s, and z13 servers were enhanced to allow all large write operations (> 64 KB) at distances up to 100 kilometers to be run in a single round trip to the control unit. This improvement avoids elongating the I/O service time for these write operations at extended distances.

Read Diagnostic Parameter Extended Link Service support

To improve the accuracy of identifying a failed component without unnecessarily replacing components in a SAN fabric, a new Extended Link Service (ELS) command called Read Diagnostic Parameters (RDP) was added to the Fibre Channel T11 standard to allow Z servers to obtain extra diagnostic data from the SFP optics that are throughout the SAN fabric.

⁶ Check with the switch provider for their support statement.

z14 ZR1, z14 M0x, z13s, and z13 servers now can read this extra diagnostic data for all the ports that are accessed in the I/O configuration and make the data available to an LPAR. For z/OS LPARs that use FICON channels, z/OS displays the data with a new message and display command. For Linux on Z, z/VM, and z/VSE, and LPARs that use FCP channels, this diagnostic data is available in a new window in the SAN Explorer tool.

N_Port ID Virtualization enhancement

N_Port ID Virtualization (NPIV) allows multiple system images (in LPARs or z/VM guests) to use a single FCP channel as though each were the sole user of the channel. First introduced with IBM z9® EC, this feature can be used with earlier FICON features that were carried forward from earlier servers.

By using the FICON Express16S (or newer) as an FCP channel with NPIV enabled, the maximum numbers of the following aspects for one FCP physical channel are doubled:

- ▶ Maximum number of NPIV hosts defined: Increased from 32 to 64
- ▶ Maximum number of remote N_Ports communicated: Increased from 512 to 1024
- ▶ Maximum number of addressable LUNs: Increased from 4096 to 8192
- ▶ Concurrent I/O operations: Increased from 764 to 1528

For more information about operating systems that support NPIV, see “N_Port ID Virtualization” on page 254.

Export/import physical port WWPNs for FCP Channels

IBM Z systems automatically assign worldwide port names (WWPNs) to the physical ports of an FCP channel that is based on the PCHID. This WWPN assignment changes when an FCP channel is moved to a different physical slot position.

z14 ZR1, z14 M0x, z13, and z13s servers allow for the modification of these default assignments, which also allows FCP channels to keep previously assigned WWPNs, even after being moved to a different slot position. This capability can eliminate the need for reconfiguration of the SAN in many situations, and is especially helpful during a system upgrade (FC 0099 - WWPN Persistence).

Note: For more information about the FICON enhancement of z14 ZR1 servers, see *Get More Out of Your IT Infrastructure with IBM z13 I/O Enhancements*, REDP-5134.

FICON support for multiple-hop cascaded SAN configurations

Before the introduction of z13 and z13s servers, IBM Z FICON SAN configurations supported a single ISL (a single hop) in a cascaded FICON SAN environment only. The z14 ZR1, z14 M0x, z13, and z13s servers now support up to three hops in a cascaded FICON SAN environment. This support allows clients to more easily configure a three- or four-site disaster recovery solution.

For more information about the FICON multi-hop, see the [FICON Multihop: Requirements and Configurations white paper](#) at the IBM Techdocs Library website.

FICON feature summary

The FICON feature codes, cable type, maximum unrepeated distance, and the link data rate on a z14 ZR1 server are listed in Table 4-7. All FICON features use LC Duplex connectors.

Table 4-7 FICON Features

Channel feature	Feature codes	Bit rate	Cable type	Maximum unrepeated distance ^a (MHz -km)
FICON Express16S+ 10KM LX	0427	4, 8, or 16 Gbps	SM 9 µm	10 km
FICON Express16S+ SX	0428	16 Gbps	MM 50 µm	35 m (500) 100 m (2000) 125 m (4700)
		8 Gbps	MM 62.5 µm MM 50 µm	21 m (200) 50 m (500) 150 m (2000) 190 m (4700)
		4 Gbps	MM 62.5 µm MM 50 µm	70 m (200) 150 m (500) 380 m (2000) 400 m (4700)
FICON Express16S 10KM LX	0418	4, 8, or 16 Gbps	SM 9 µm	10 km
FICON Express16S SX	0419	16 Gbps	MM 50 µm	35 m (500) 100 m (2000) 125 m (4700)
		8 Gbps	MM 62.5 µm MM 50 µm	21 m (200) 50 m (500) 150 m (2000) 190 m (4700)
		4 Gbps	MM 62.5 µm MM 50 µm	70 m (200) 150 m (500) 380 m (2000) 400 m (4700)
FICON Express8S 10KM LX	0409	2, 4, or 8 Gbps	SM 9 µm	10 km
FICON Express8S SX		8 Gbps	MM 62.5 µm MM 50 µm	21 m (200) 50 m (500) 150 m (2000) 190 m (4700)
		4 Gbps	MM 62.5 µm MM 50 µm	70 m (200) 150 m (500) 380 m (2000) 400 m (4700)
		2 Gbps	MM 62.5 µm MM 50 µm	150 m (200) 300 m (500) 500 m (2000) N/A (4700)

a. Minimum fiber bandwidths in MHz/km for multimode fiber optic links are included in parentheses, where applicable.

zHyperLink Express (FC 0431)

zHyperLink is a new technology that provides up to 5x reduction in I/O latency times for Db2 read requests with the qualities of service IBM Z clients expect from I/O infrastructure for Db2 v12 with z/OS 2.1 with patches.

The zHyperLink Express feature (FC 0431) provides a low latency direct connection between z14 ZR1 and DS8880 I/O Port.

The zHyperLink Express is the result of new business requirements that demand fast and consistent application response times. It dramatically reduces latency by interconnecting the z14 ZR1 directly to I/O Bay of the DS8880 by using PCIe Gen3 x 8 physical link (up to 150-meter [492-foot] distance). A new transport protocol is defined for reading and writing IBM CKD data records⁷, as documented in the zHyperLink interface specification.

On z14 ZR1, zHyperLink Express card is a new PCIe adapter, which installed in the PCIe+ I/O drawer. HCD definition support was added for new PCIe function type with PORT attributes.

Requirements of zHyperLink

The zHyperLink Express feature is available on z14 ZR1 servers, and includes the following requirements:

- ▶ z/OS 2.1 or later
- ▶ DS888x with I/O Bay Planar board and firmware level 8.3
- ▶ z14 with zHyperLink Express adapter (FC 0431) installed
- ▶ FICON channel as a driver
- ▶ Only ECKD supported
- ▶ z/VM is not supported

Up to 16 zHyperLink Express adapters can be installed in a z14 ZR1 (up to 32 links).

The zHyperLink Express is virtualized as a native PCIe adapter and can be shared by multiple LPARs. Each port can support up to 127 Virtual Functions (VFs), with one or more VFs/PFIDs being assigned to each LPAR. This configuration gives a maximum of 254 VFs per adapter. The zHyperLink Express requires the following components:

- ▶ zHyperLink connector on DS8880 I/O Bay
 - For DS8880 firmware R8.3 or newer, the I/O Bay planar is updated to support the zHyperLink interface. This update includes the update of the PEX 8732 switch to PEX8733 that includes a DMA engine for the zHyperLink transfers, and the upgrade from a copper to optical interface by a CXP connector (provided).
- ▶ Cable
 - The zHyperLink Express uses optical cable with MTP connector. Maximum supported cable length is 150 meters (492 feet).

4.6.3 Network connectivity

Communication for LANs is provided by the OSA-Express7S, OSA-Express6S, OSA-Express5S, OSA-Express4S (except 1000BASE-T), 10GbE RoCE Express2, and 10GbE RoCE Express features.

OSA-Express7S

OSA-Express7S 25 Gigabit Ethernet SR (FC 0429) is installed in the PCIe+ I/O Drawer.

⁷ CKD data records are handled by using IBM Enhanced Count Key Data (ECKD™) command set.

The OSA-Express7S 25GbE Short Reach (SR) feature includes one PCIe adapter and one port per feature. The port supports CHPID types OSD and OSX. The 25GbE feature is designed to support attachment to a multimode fiber 25 Gbps Ethernet LAN or Ethernet switch that is capable of 25 Gbps. The port can be defined as a spanned channel and shared among LPARs within and across logical channel subsystems.

The OSA-Express7S 25GbE 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-Express7S 25GbE SR feature does not support auto-negotiation to any other speed and runs in full duplex mode only.

A 50 μ m multimode fiber optic cable that ends with an LC Duplex connector is required for connecting each port on this feature to the selected device.

OSA-Express6S

The OSA-Express6S feature is installed in the PCIe+ I/O drawer. The following OSA-Express6S features can be installed on z14 ZR1 servers:

- ▶ OSA-Express6S 10 Gigabit Ethernet LR, FC 0424
- ▶ OSA-Express6S 10 Gigabit Ethernet SR, FC 0425
- ▶ OSA-Express6S Gigabit Ethernet LX, FC 0422
- ▶ OSA-Express6S Gigabit Ethernet SX, FC 0423
- ▶ OSA-Express6S 1000BASE-T Ethernet, FC 0426

The supported OSA-Express6S features are listed in Table 4-5 on page 133.

OSA-Express6S 10 Gigabit Ethernet LR (FC 0424)

The OSA-Express6S 10 Gigabit Ethernet (GbE) Long Reach (LR) feature includes one PCIe adapter and one port per feature. The port supports CHPID types OSD and OSX. 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.

Note: zBX Model 004 can be carried forward during an upgrade from z13 to IBM z14 (as the z/BX is an independent Ensemble node, not tied to any IBM Z CPC); however, ordering any zBX features was withdrawn from marketing as of March 31, 2017.

The OSA-Express6S 10 GbE LR feature supports the use of an industry standard small form factor LC Duplex connector. Ensure that the attaching or downstream device includes an LR transceiver. The transceivers at both ends must be the same (LR to LR).

The OSA-Express6S 10 GbE LR feature does not support auto-negotiation to any other speed and runs in full duplex mode only.

A 9 μ m single-mode fiber optic cable that ends with an LC Duplex connector is required for connecting this feature to the selected device.

OSA-Express6S 10 Gigabit Ethernet SR (FC 0416)

The OSA-Express6S 10 GbE Short Reach (SR) feature includes one PCIe adapter and one port per feature. The port supports CHPID types OSD and OSX. 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 shared among LPARs within and across logical channel subsystems.

The OSA-Express6S 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-Express6S 10 GbE SR feature does not support auto-negotiation to any other speed and runs in full duplex mode only.

A 50 or a 62.5 μm multimode fiber optic cable that ends with an LC Duplex connector is required for connecting each port on this feature to the selected device.

OSA-Express6S Gigabit Ethernet LX (FC 0422)

The OSA-Express6S GbE LX feature includes one PCIe adapter and two ports. The two ports share a channel path identifier (CHPID type OSD). 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-Express6S GbE LX feature supports the use of an LC Duplex connector. Ensure that the attaching or downstream device has an LX transceiver. The sending and receiving transceivers must be the same (LX to LX).

A 9 μm single-mode fiber optic cable that ends with an LC Duplex connector is required for connecting each port on this feature to the selected device. If multimode fiber optic cables are being reused, a pair of Mode Conditioning Patch cables is required, with one cable for each end of the link.

OSA-Express6S Gigabit Ethernet SX (FC 0423)

The OSA-Express6S GbE SX feature includes one PCIe adapter and two ports. The two ports share a channel path identifier (CHPID type OSD). The ports support attachment to a 1 Gbps Ethernet LAN. Each port can be defined as a spanned channel and shared among LPARs and across logical channel subsystems.

The OSA-Express6S GbE SX feature supports the use of an LC Duplex connector. Ensure that the attaching or downstream device has an SX transceiver. The sending and receiving transceivers must be the same (SX to SX).

A multi-mode fiber optic cable that ends with an LC Duplex connector is required for connecting each port on this feature to the selected device.

OSA-Express6S 1000BASE-T Ethernet feature (FC 0426)

Feature code 0426 occupies one slot in the PCIe+ I/O drawer. It features two ports that connect to a 1000 Mbps (1 Gbps) or 100 Mbps Ethernet LAN. Each port has an SFP with an RJ-45 receptacle for cabling to an Ethernet switch. The RJ-45 receptacle is required to be attached by using an EIA/TIA Category 5 or Category 6 UTP cable with a maximum length of 100 meters (328 feet). The SFP allows a concurrent repair or replace action.

OSA-Express6S 1000BASE-T adapters^a: OSA-Express6S 1000BASE-T adapters (#0426) are the last generation of OSA 1000BASE-T adapters to support connections operating at 100 Mbps link speed. Future OSA-Express 1000BASE-T adapter generations will support operation at 1000 Mbps (1Gbps) link speed only.

a. IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM's sole discretion.

The OSA-Express6S 1000BASE-T Ethernet feature supports auto-negotiation when attached to an Ethernet router or switch. If you allow the LAN speed and duplex mode to default to auto-negotiation, the OSA-Express port and the attached router or switch auto-negotiate the LAN speed and duplex mode settings between them. They then connect at the highest common performance speed and duplex mode of interoperation. If the attached Ethernet router or switch does not support auto-negotiation, the OSA-Express port examines the signal that it is receiving and connects at the speed and duplex mode of the device at the other end of the cable.

The OSA-Express6S 1000BASE-T Ethernet feature can be configured as CHPID type OSC, OSD, OSE, or OSM. Non-QDIO operation mode requires CHPID type OSE.

Note: CHPID type OSN is not supported on OSA-Express6S 1000BASE-T Ethernet feature for NCP (LP to LP).

The following settings are supported on the OSA-Express6S 1000BASE-T Ethernet feature port:

- ▶ Auto-negotiate
- ▶ 100 Mbps half-duplex or full-duplex
- ▶ 1000 Mbps full-duplex

If auto-negotiate is not used, the OSA-Express port attempts to join the LAN at the specified speed and duplex mode. If this specified speed and duplex mode do not match the speed and duplex mode of the signal on the cable, the OSA-Express port does not connect.

OSA-Express5S

The OSA-Express5S feature is installed in the PCIe I/O drawer. The following OSA-Express5S features can be installed on z14 servers (carry forward only):

- ▶ 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

The OSA-Express5S features are listed in Table 4-5 on page 133.

OSA-Express5S 10 Gigabit Ethernet LR (FC 0415)

The OSA-Express5S 10 Gigabit Ethernet (GbE) Long Reach (LR) feature includes one PCIe adapter and one port per feature. The port supports CHPID types OSD and OSX.

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 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 includes an LR transceiver. The transceivers at both ends must be the same (LR to LR).

The OSA-Express5S 10 GbE LR feature does not support auto-negotiation to any other speed and runs in full duplex mode only.

A 9 μ m single-mode fiber optic cable that ends 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 GbE Short Reach (SR) feature includes 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 IBM z14 ZR1, IBM z14, or z13 servers to zBX.

The 10 GbE feature is designed to support attachment to a multimode fiber 10 Gbps Ethernet LAN or Ethernet switch that is capable of 10 Gbps. The port can be defined as a spanned channel and 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 includes an SR transceiver. The sending and receiving transceivers must be the same (SR to SR).

The OSA-Express5S 10 GbE SR feature does not support auto-negotiation to any other speed and runs in full duplex mode only.

A 50 or a 62.5 μ m multimode fiber optic cable that ends 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 GbE LX feature includes 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 shared among LPARs and across logical channel subsystems.

The OSA-Express5S GbE LX feature supports the use of an LC Duplex connector. Ensure that the attaching or downstream device has an LX transceiver. The sending and receiving transceivers must be the same (LX to LX).

A 9 μ m single-mode fiber optic cable that ends with an LC Duplex connector is required for connecting each port on this feature to the selected device. If multimode fiber optic cables are being reused, a pair of Mode Conditioning Patch cables is required, with one cable for each end of the link.

OSA-Express5S Gigabit Ethernet SX (FC 0414)

The OSA-Express5S GbE SX feature includes one PCIe adapter and two ports. The two ports share a channel path identifier (CHPID type OSD exclusively). The ports support attachment to a 1 Gbps Ethernet LAN. Each port can be defined as a spanned channel and can be shared among LPARs and across logical channel subsystems.

The OSA-Express5S GbE SX feature supports the use of an LC Duplex connector. Ensure that the attaching or downstream device has an SX transceiver. The sending and receiving transceivers must be the same (SX to SX).

A multi-mode fiber optic cable that ends with an LC Duplex connector is required for connecting each port on this feature to the selected device.

OSA-Express5S 1000BASE-T Ethernet feature (FC 0417)

Feature code 0417 occupies one slot in the PCIe I/O drawer. It has two ports that connect to a 1000 Mbps (1 Gbps) or 100 Mbps Ethernet LAN. Each port has an SFP with an RJ-45 receptacle for cabling to an Ethernet switch. The RJ-45 receptacle is required to be attached by using an EIA/TIA Category 5 or Category 6 UTP cable with a maximum length of 100 meters (328 feet). The SFP allows a concurrent repair or replace action.

The OSA-Express5S 1000BASE-T Ethernet feature supports auto-negotiation when attached to an Ethernet router or switch. If you allow the LAN speed and duplex mode to default to auto-negotiation, the OSA-Express port and the attached router or switch auto-negotiate the LAN speed and duplex mode settings between them. They then connect at the highest common performance speed and duplex mode of interoperation. If the attached Ethernet router or switch does not support auto-negotiation, the OSA-Express port examines the signal that it is receiving and connects at the speed and duplex mode of the device at the other end of the cable.

The OSA-Express5S 1000BASE-T Ethernet feature can be configured as CHPID type OSC, OSD, OSE, 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 auto-negotiate is not used, the OSA-Express port attempts to join the LAN at the specified speed and duplex mode. If this specified speed and duplex mode do not match the speed and duplex mode of the signal on the cable, the OSA-Express port does not connect.

OSA-Express4S features

This section describes the characteristics of all OSA-Express4S features that are supported on z14 servers.

The OSA-Express4S feature is installed in the PCIe I/O drawer. Only OSA-Express4S 1000BASE-T Ethernet, FC 0408 is supported on IBM z14 servers as a carry forward during an MES.

The characteristics of the OSA-Express4S features that are supported on z14 ZR1 are listed in Table 4-5 on page 133.

OSA-Express4S 10 Gigabit Ethernet LR (FC 0406)

The OSA-Express4S 10 Gigabit Ethernet (GbE) Long Reach (LR) feature includes one PCIe adapter and one port per feature. The port supports CHPID types OSD and OSX.

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 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 includes an LR transceiver. The transceivers at both ends must be the same (LR to LR).

The OSA-Express4S 10 GbE LR feature does not support auto-negotiation to any other speed and runs in full duplex mode only.

A 9 μ m single-mode fiber optic cable that ends 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 GbE Short Reach (SR) feature includes one PCIe adapter and one port per feature. The port supports CHPID types OSD and OSX. 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 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.

A 50 or a 62.5 μ m multimode fiber optic cable that ends with an LC Duplex connector is required for connecting each port on this feature to the selected device.

OSA-Express4S Gigabit Ethernet LX (FC 0404)

The OSA-Express4S GbE LX feature includes one PCIe adapter and two ports. The two ports share a channel path identifier (CHPID type OSD). 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 ends with an LC Duplex connector is required for connecting each port on this feature to the selected device. If multimode fiber optic cables are being reused, a pair of Mode Conditioning Patch cables is required, with one cable for each end of the link.

OSA-Express4S Gigabit Ethernet SX (FC 0405)

The OSA-Express4S GbE SX feature includes one PCIe adapter and two ports. The two ports share a channel path identifier (CHPID type OSD). The ports support attachment to a 1 Gbps Ethernet LAN. Each port can be defined as a spanned channel and shared among LPARs and across logical channel subsystems.

The OSA-Express4S GbE SX feature supports the use of an LC Duplex connector. Ensure that the attaching or downstream device has an SX transceiver. The sending and receiving transceivers must be the same (SX to SX).

A multi-mode fiber optic cable that ends with an LC Duplex connector is required for connecting each port on this feature to the selected device.

Note: OSA-Express4S BASE-T Ethernet is *not* supported on z14 ZR1.

25GbE RoCE Express2

25GbE RoCE Express2 (FC 0430) is installed in the PCIe I/O drawer and is supported on IBM z14™. The 25GbE RoCE Express2 is a native PCIe feature. It does not use a CHPID and is defined by using the IOCP **FUNCTION** statement or in the hardware configuration definition (HCD).

On IBM z14™ servers, both ports are supported by z/OS and can be shared by up to 126 partitions (LPARs) per PCHID. The 25GbE RoCE Express2 feature uses SR optics and supports the use of a multimode fiber optic cable that ends with an LC Duplex connector. Both point-to-point connections and switched connections with an enterprise-class 25GbE switch are supported.

Switch configuration for RoCE Express2: If the IBM 25GbE RoCE Express2 features are connected to 25GbE switches, the switches must meet the following requirements:

- ▶ Global Pause function enabled
- ▶ Priority flow control (PFC) disabled
- ▶ No firewalls, no routing, and no IEDN

The 25GbE RoCE Express feature does not support auto-negotiation to any other speed and runs in full duplex mode only.

10GbE and 25GbE RoCE features should *not* be mixed in a z/OS SMC-R Link Group.

The maximum supported unrepeat distance, point-to-point, is 100 meters (328 feet). A client-supplied cable is required. Two types of cables can be used for connecting the port to the selected 25GbE switch or to the 25GbE RoCE Express2 feature on the attached server:

- ▶ OM3 50-micron multimode fiber optic cable that is rated at 2000 MHz-km that ends with an LC Duplex connector, which supports 70 meters (229 feet)
- ▶ OM4 50-micron multimode fiber optic cable that is rated at 4700 MHz-km that ends with an LC Duplex connector, which supports 100 meters (328 feet)

For more information about the management and definition of the 25GbE RoCE Express2 feature, see Appendix D, “Shared Memory Communications” on page 425, and Appendix C, “Native Peripheral Component Interconnect Express” on page 419.

10GbE RoCE Express2

RoCE Express2 (FC 0412) is installed in the PCIe+ I/O drawer and is supported on z14 ZR1. The 10GbE RoCE Express2 is a native PCIe feature. It does not use a CHPID and is defined by using the IOCP **FUNCTION** statement or in the hardware configuration definition (HCD).

On z14 ZR1 servers, both ports are supported by z/OS and can be shared by up to 126 partitions (LPARs) per PCHID. The 10GbE RoCE Express2 feature uses SR optics and supports the use of a multimode fiber optic cable that ends with an LC Duplex connector. Both point-to-point connections and switched connections with an enterprise-class 10 GbE switch are supported.

The virtualization capabilities for z14 ZR1 are 31 Virtual Functions per port (62 VFs per feature). The RAS was improved and ECC double bit correction added.

Switch configuration for RoCE Express2: If the IBM 10GbE RoCE Express2 features are connected to 10 GbE switches, the switches must meet the following requirements:

- ▶ Global Pause function enabled
- ▶ Priority flow control (PFC) disabled
- ▶ No firewalls, no routing, and no IEDN

The maximum supported unrepeat distance, point-to-point, is 300 meters (984 feet). A client-supplied cable is required. The following types of cables can be used for connecting the port to the selected 10 GbE switch or to the 10GbE RoCE Express2 feature on the attached server:

- ▶ OM3 50-micron multimode fiber optic cable that is rated at 2000 MHz-km that ends with an LC Duplex connector; supports 300 meters (984 feet)
- ▶ OM2 50-micron multimode fiber optic cable that is rated at 500 MHz-km that ends with an LC Duplex connector; supports 82 meters (269 feet)
- ▶ OM1 62.5-micron multimode fiber optic cable that is rated at 200 MHz-km that ends with an LC Duplex connector; supports 33 meters (108 feet)

For more information about the management and definition of the 10GbE RoCE2, see Appendix D, “Shared Memory Communications” on page 425, and Appendix C, “Native Peripheral Component Interconnect Express” on page 419.

10GbE RoCE Express

The 10GbE RoCE Express feature (FC 0411) is installed in the PCIe+ I/O drawer. This feature is supported on z13, z13s, zEC12, and zBC12 servers and can be carried forward during an MES upgrade to a z14 ZR1.

The 10GbE RoCE Express is a native PCIe feature. It does not use a CHPID and is defined by using the IOCP **FUNCTION** statement or in the hardware configuration definition (HCD).

For zEC12 and zBC12, each feature can be dedicated to an LPAR only, and z/OS can use only one of the two ports. Both ports are supported by z/OS and can be shared by up to 31 partitions (LPARs) per PCHID on z14 ZR1, z14 M0x, z13s, and z13.

The 10GbE RoCE Express feature uses SR optics and supports the use of a multimode fiber optic cable that ends with an LC Duplex connector. Point-to-point connections and switched connections with an enterprise-class 10 GbE switch are supported.

Switch configuration for RoCE: If the IBM 10GbE RoCE Express features are connected to 10 GbE switches, the switches must meet the following requirements:

- ▶ Global Pause function enabled
- ▶ Priority flow control (PFC) disabled
- ▶ No firewalls, no routing, and no IEDN

The maximum supported unrepeat distance, point-to-point, is 300 meters (984 feet). A client-supplied cable is required. The following 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 that is rated at 2000 MHz-km that ends with an LC Duplex connector; supports 300 meters (984 feet)
- ▶ OM2 50-micron multimode fiber optic cable that is rated at 500 MHz-km that ends with an LC Duplex connector; supports 82 meters (269 feet)
- ▶ OM1 62.5-micron multimode fiber optic cable that is rated at 200 MHz-km that ends with an LC Duplex connector; supports 33 meters (108 feet)

For more information about the management and definition of the 10GbE RoCE, see Appendix D, “Shared Memory Communications” on page 425, and Appendix C, “Native Peripheral Component Interconnect Express” on page 419.

Shared Memory Communications functions

The Shared Memory Communication (SMC) capabilities of the z14 ZR1 help optimize the communications between applications for server-to-server (SMC-R) or LPAR-to-LPAR (SMC-D) connectivity.

SMC-R

SMC-R provides application transparent use of the RoCE-Express feature. This feature reduces the network overhead and latency of data transfers, which effectively offers the benefits of optimized network performance across processors.

SMC-D

SMC-D was used with the introduction of the Internal Shared Memory (ISM) virtual PCI function. ISM is a virtual PCI network adapter that enables direct access to shared virtual memory, which provides a highly optimized network interconnect for IBM Z intra-CPC communications.

SMC-D maintains the socket-API transparency aspect of SMC-R so that applications that use TCP/IP communications can benefit immediately without requiring any application software or IP topology changes. SMC-D completes the overall SMC solution, which provides synergy with SMC-R.

SMC-R and SMC-D use shared memory architectural concepts, which eliminates the TCP/IP processing in the data path, yet preserves TCP/IP Qualities of Service for connection management purposes.

Internal Shared Memory (ISM)

ISM is a function that is supported by z14 ZR1, z14 M0x, z13, and z13s machines. It is the firmware that provides connectivity by using shared memory access between multiple operating system images within the same CPC. ISM creates virtual adapters with shared memory that is allocated for each OS image.

ISM is defined by the FUNCTION statement with a virtual CHPID (VCHID) in hardware configuration definition (HCD)/IOCDS. Identified by the PNETID parameter, each ISM VCHID defines an isolated, internal virtual network for SMC-D communication, without any hardware component required. Virtual adapters are defined by virtual function (VF) statements. Multiple LPARs can access the same virtual network for SMC-D data exchange by associating their VF with same VCHID.

Applications that use HiperSockets can realize network latency and CPU reduction benefits and performance improvement by using the SMC-D over ISM.

z14 ZR1 servers support up to 32 ISM VCHIDs per CPC. Each VCHID supports up to 255 VFs, with a total maximum of 8,000 VFs.

For more information about the SMC-D and ISM, see Appendix D, “Shared Memory Communications” on page 425.

HiperSockets

The HiperSockets function of z14 ZR1 servers provides up to 32 high-speed virtual LAN attachments.

HiperSockets IOCP definitions on z14 ZR1: A parameter was added for HiperSockets IOCP definitions on z14 ZR1, z14 M0x, z13, and z13s servers. Therefore, the IOCP definitions must be migrated to support the HiperSockets definitions (CHPID type IQD).

On z14 ZR1, z14 M0x, z13, and z13s servers, the CHPID statement of HiperSockets devices require the keyword VCHID. VCHID specifies the virtual channel identification number that is associated with the channel path. The vSalid range is 7E0 - 7FF.

VCHID is not valid on Z servers before z13.

For more information, see *IBM Z Input/Output Configuration Program User's Guide for ICP IOCP*, SB10- 7172.

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 the need to use I/O subsystem operations and traverse an external network connection to communicate between LPARs in the same z14 ZR1 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 the following 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) features its own Media Access Control (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 also can be used to connect to the HiperSockets Layer 2 networks of different servers.

HiperSockets Layer 2 is supported by Linux on Z, and by z/VM for Linux guest use.

z14 ZR1 supports 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, data can be temporarily held until the receiver has buffers that are available in its inbound queue during high volume situations. The HiperSockets Completion Queue function requires the following minimum applications⁸:

- ▶ z/OS V1.13
- ▶ Linux on Z distributions:
 - Red Hat Enterprise Linux (RHEL) 6.2
 - SUSE Linux Enterprise Server (SLES) 11 SP2
 - Ubuntu server 16.04 LTS
- ▶ z/VSE V5.1.1⁹
- ▶ z/VM V6.2¹⁰ with maintenance

In z/VM V6.4 and newer, the virtual switch function transparently bridges a guest virtual machine network connection on a HiperSockets LAN segment. This bridge allows a single HiperSockets guest virtual machine network connection to communicate directly with the following systems:

- ▶ Other guest virtual machines on the virtual switch
- ▶ External network hosts through the virtual switch OSA UPLINK port

4.6.4 Parallel Sysplex connectivity

Coupling links are required in a Parallel Sysplex configuration to provide connectivity from the z/OS images to the coupling facility (CF). A properly configured Parallel Sysplex provides a highly reliable, redundant, and robust IBM Z technology solution to achieve near-continuous availability. A Parallel Sysplex is composed of one or more z/OS operating system images that are coupled through one or more CFs.

Coupling links

The type of coupling link that is used to connect a CF to an operating system LPAR is important. The link performance significantly affects 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.

The following links are available to connect an operating system LPAR to a CF running on z14 ZR1:

- ▶ Integrated Coupling Adapter (ICA SR) for short distance connectivity, which is defined as CHPID type CS5. The ICA SR can be used only for coupling connectivity between z14 ZR1, z14, z13, and z13s servers. It does not support connectivity to zEC12 or zBC12 servers, and it cannot be connected to HCA3-O or HCA3-O LR coupling fanouts.

⁸ Minimum OS support for z14 ZR1 can differ. For more information, see Chapter 7, “Operating system support” on page 209.

⁹ z/VSE 5.1.1 is end of support.

¹⁰ z/VM V6.2 is not supported on z14 ZR1. z/VM V6.4 or newer is needed.

The ICA SR supports distances up to 150 meters (492 feet) and a link data rate of 8 Gbps. OM3 fiber optic cable is used for distances up to 100 meters (328 feet), and OM4 or OM5 for distances up to 150 meters (492 feet). ICA SR supports four CHPIDs per port and seven subchannels (devices) per CHPID. ICA SR supports transmission of Server Time Protocol (STP) messages.

- **Coupling Express Long Reach:** Coupling Express LR (FC 0433) is needed for Long Distance Coupling z14 ZR1/z14 M0x/z13/z13s to z13 and above. It supports a maximum unrepeatd distance to 10 kilometers (6.2 miles) and up to 100 kilometers (62 miles) with a qualified DWDM. CE LR coupling links are defined as CHPID CL5. CE LR uses same 9 µm single mode fiber cable as 1x IFB.

The coupling link options are listed in Table 4-8.

Table 4-8 Coupling link options that are supported on z14 ZR1

Type	Description	Use for connecting	Link rate	Distance	z14 ZR1 maximum number of ports
CE LR	Coupling Express LR	z14/z13/z13s to z14/z13/z13s	10 Gbps	10 km unrepeatd (6.2 miles) 100 km repeated (62 miles)	32
ICA SR	Integrated Coupling Adapter	z14/z13/z13s to z14/z13/z13s	8 Gbps	150 meters (492 feet)	16
IC	Integrated Coupling Adapter	Internal communication	Internal speeds	N/A	32

The maximum number of combined external coupling links (active CE LR, ICA SR links) is 44 per z14 ZR1 server. z14 ZR1 servers support up to 176 coupling CHPIDs per CPC. A z14 ZR1 coupling link support summary is shown in Figure 4-4.

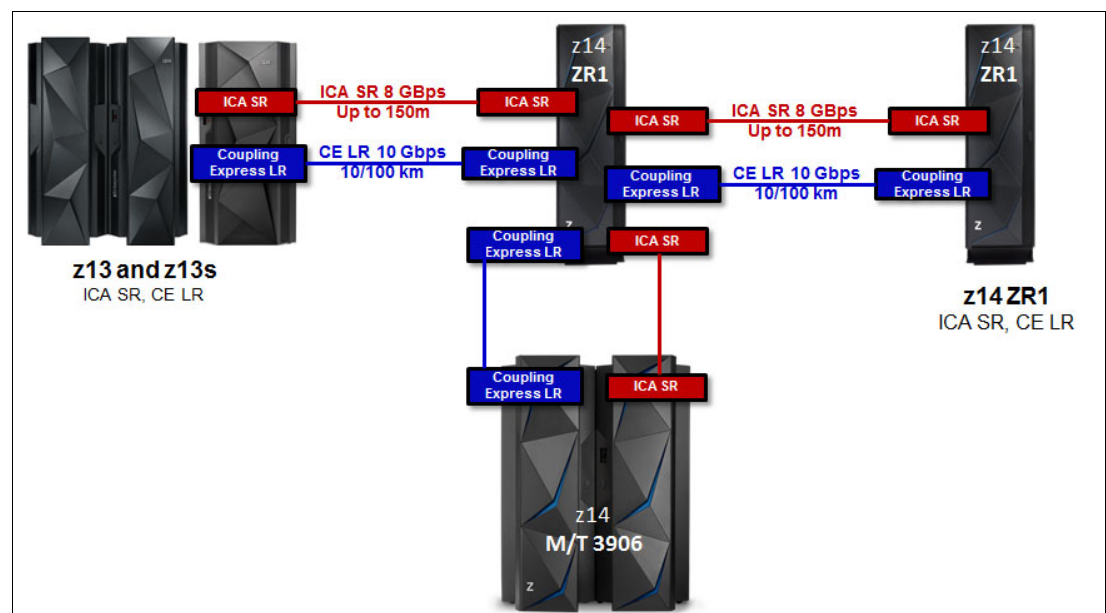


Figure 4-4 z14 ZR1 Parallel Sysplex coupling connectivity

In a Parallel Sysplex configuration, z/OS and CF images can run on the same or on separate servers. At least one CF that is connected to all z/OS images must exist, even though other CFs can be connected to selected z/OS images only. Two CF 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, the following components must be used at minimum:

- ▶ Two coupling links between the z/OS and CF images.
- ▶ Two CF images not running on the same server.
- ▶ One stand-alone CF. If system-managed CF structure duplexing is used or is running with *resource sharing* only, a stand-alone CF is not mandatory.

Coupling link features

z14 ZR1 server supports the following coupling link features:

- ▶ ICA SR fanout, FC 0172
- ▶ CE LR adapter, FC 0433

Note: z14 ZR1 does not support the following coupling link features:

- ▶ HCA3-O fanout for 12x InfiniBand, FC 0171
- ▶ HCA3-O LR fanout for 1x InfiniBand, FC 0170

If a Parallel Sysplex configuration uses these coupling links, other extensions must be considered.

Extended distance support

For more information about extended distance support, see *System z End-to-End Extended Distance Guide*, SG24-8047.

Internal coupling links

IC links are LIC-defined links that connect a CF to a z/OS LPAR in the same server. These links are available on all IBM Z systems. The IC link is a 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, which is implemented in Licensed Internal Code (LIC), and so does not require any hardware or cabling.

An IC link is a fast coupling link that uses memory-to-memory data transfers. Although IC links do not have PCHID numbers, they 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. The links are always defined and connected in pairs. The IC link operates in peer mode, and its existence is defined in HCD/IOCP.

IC links feature the following attributes:

- ▶ Operates in peer mode (channel type ICP) on IBM Z systems.
- ▶ Provides the fastest connectivity, which is faster than any external link alternatives.
- ▶ Results in better coupling efficiency than with external links, which effectively reduces the server cost that is associated with Parallel Sysplex technology.
- ▶ Can be used in test or production configurations, and reduces 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). The use of ICFs with IC channels results in considerable cost savings when you are configuring a cluster.

IC links are enabled by defining channel type ICP. A maximum of 32 IC channels can be defined on a Z server.

Migration considerations

Upgrading from previous generations of IBM Z systems in a Parallel Sysplex to z14 ZR1 servers in that same Parallel Sysplex requires proper planning for coupling connectivity. Planning is important because of the change in the supported type of coupling link adapters and the number of available fanout slots of the z14 ZR1 CPC drawer, as compared to the number of available fanout slots of the processor books of the previous generation Z servers, such as zEC12.

Coupling links connectivity support: z196 and z114 are *not* supported in same Parallel Sysplex or STP CTN with z14 ZR1.

zEC12 or zBC12 can coexist in the same Parallel Sysplex with z14 ZR1 only if the CPC hosting the CFs includes coupling connectivity to the zEC12/zBC12 and z14 ZR1 CPCs.

The ICA SR fanout provides short-distance connectivity to another z14 ZR1, z14 M0x, z13s, or z13 server.

The CE LR adapter provides long-distance connectivity to another z14 ZR1, z14 M0x, z13s, or z13 server. For more information, see “Coupling links” on page 153.

The z14 ZR1 server fanout slots in the CPC drawer provide coupling links connectivity through the ICA SR fanout cards. In addition to coupling links for Parallel Sysplex, the fanout cards that the fanout slots provide allow connectivity for the PCIe+ I/O drawer (PCIe fanout).

Up to eight PCIe fanout cards can be installed in CPC drawer.

It is beyond the scope of this book to describe all possible migration scenarios. Always consult with subject matter experts to help you to develop your migration strategy.

The following considerations can help you assess possible migration scenarios. The objective of this list is to enable migration to z14 ZR1 servers, support legacy coupling where essential, and adopt ICA SR where possible to avoid the need for more CPC drawers and other possible migration issues:

- ▶ The IBM zEnterprise EC12 and BC12 are the last generation of Z servers to support ISC-3, 12x HCA2-O, and 1x HCA2-O LR. They also are the last Z servers that can be part of a Mixed Coordinated Timing Network (CTN).
- ▶ Consider the following coupling requirements for migration:
 - CE LR is the only long-distance coupling link available on z14 ZR1 servers.
 - ICA SR should be used for short distance coupling requirements.
 - The use of ICA SR cards affects the number of PCIe+ I/O drawers, and the number of CE LR coupling links. PCIe I/O slots availability versus the number of ICA SR fanouts installed are listed in Table 4-9 on page 157.

Table 4-9 Effect of ICA SR cards on maximum of I/O slots

ICA SR	Z14 ZR1
	Max I/O slots
0	64
1, 2	48
3, 4	32
5, 6	16
7, 8	0

- ▶ Evaluate configurations for opportunities to eliminate or consolidate links:
 - Eliminate any redundant links. Two physical links between CPCs is the minimum requirement from a reliability, availability, and serviceability (RAS) perspective.
 - Coupling Link Analysis: Capacity Planning tools and services can help.
- ▶ Depending on your I/O requirements, you might need to use Coupling Express LR instead of ICA SR. This configuration allows the system to include more PCIe+ I/O drawers.

Coupling links and Server Time Protocol

All external coupling links can be used to pass time synchronization signals by using Server Time Protocol (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 CF messages in a Parallel Sysplex.

The use of the coupling links to exchange STP messages has the following advantages:

- ▶ By using the same links to exchange STP messages and CF messages in a Parallel Sysplex, STP can scale with distance. Servers that are exchanging messages over short distances, such as IFB or ICA SR links, can meet more stringent synchronization requirements than servers that exchange messages over long IFB LR links, with distances up to 100 kilometers (62 miles). This advantage is an enhancement over the IBM Sysplex Timer implementation, which does not scale with distance.
- ▶ Coupling links also provide the connectivity that is necessary in a Parallel Sysplex. Therefore, a potential benefit can be realized of minimizing the number of cross-site links that is required in a multi-site Parallel Sysplex.

Between any two servers that are intended to exchange STP messages, configure each server so that at least two coupling links exist for communication between the servers. This configuration prevents the loss of one link from causing the loss of STP communication between the servers. If a server does not have a CF LPAR, timing-only links can be used to provide STP connectivity.

The z14 ZR1 server does not support attachment to the IBM Sysplex Timer. A z14 ZR1 server cannot be added into a Mixed CTN; it can participate in an STP-only CTN only.

STP enhancements on z14 ZR1

Important: For more information about configuring an STP CTN with three or more servers, see the [Important Considerations for STP server role assignments](#) white paper that is available at the IBM Techdocs Library website.

If the guidelines are not followed, it might result in all the servers in the CTN becoming unsynchronized. This condition results in a sysplex-wide outage.

STP on z14 ZR1 features the following enhancements:

- ▶ Additional stratum level

The limit was STP Stratum 3 before z14. The extra stratum allows CPCs to operate as part of CTN at STP stratum level 4, which can avoid the extra complexity and expense of system reconfiguration.

Warning: This extra stratum level should be used only as a temporary state during reconfiguration. Customer should not run with machines at stratum level 4 for extended periods because of the lower quality of the time synchronization.

- ▶ Graphical display of a Coordinated Timing Network (CTN)

This graphical display improved the user interface to STP controls. This type of visual display of the CTN status, which provided a clearer view of CTNs, can avoid outages, such as a user intentionally took down the CTS, but did not realize the BTS was down.

For more information about STP configuration, 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

Pulse per second input

A pulse per second (PPS) signal can be received from an external time source (ETS) device. One PPS port is available on each of the two oscillator cards. These cards are installed into slots LG11 and LG12 of the CPC drawer rear of the frame of z14 ZR1 servers (see Figure 2-20 on page 49).

Connections to the CPC drawer provide redundancy for continued operation and concurrent maintenance when a single oscillator card fails. Each oscillator card includes a Bayonet Neill-Concelman (BNC) connector for PPS connection support, which attaches to two different ETSs. Two PPS connections from two different ETSs are preferable for redundancy.

The time accuracy of an STP-only CTN is improved by adding an ETS device with the PPS output signal. STP tracks the highly stable accurate PPS signal from ETSs. It maintains accuracy of 10 μ s as measured at the PPS input of the z14 ZR1 server. If STP uses an NTP server without PPS, a time accuracy of 100 meters (328 feet) to the ETS is maintained. ETSs with PPS output are available from various vendors that offer network timing solutions.

4.7 Cryptographic functions

Cryptographic functions are provided by the CP Assist for Cryptographic Function (CPACF) and the PCI Express cryptographic adapters. z14 ZR1 servers support the Crypto Express6S feature.

4.7.1 CPACF functions (FC 3863)

FC 3863¹¹ is required to enable CPACF functions.

4.7.2 Crypto Express6S feature (FC 0893)

Crypto Express6S is a new feature on z14 ZR1 servers. On the initial configuration, a minimum of two features are installed. The number of features then increases one at a time up to a maximum of 16 features.

Each Crypto Express6S 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.

The tamper-resistant hardware security module, which is contained on the Crypto Express6S feature, conforms to the Federal Information Processing Standard (FIPS) 140-2 Level 4 Certification. It supports User Defined Extension (UDX) services to implement cryptographic functions and algorithms (when defined as an IBM CCA coprocessor).

The following CCA compliance levels are available:

- ▶ Non-compliant (default)
- ▶ PCI-HSM 2016
- ▶ PCI-HSM 2016 (migration, key tokens while migrating to compliant)

The following EP11 compliance levels are available (Crypto Express6S and Crypto Express5S):

- ▶ FIPS 2009 (default)
- ▶ FIPS 2011
- ▶ BSI 2009
- ▶ BSI 2011

Each Crypto Express6S feature occupies one I/O slot in the PCIe I/O drawer, and features no CHPID assigned. However, it includes one PCHID.

4.7.3 Crypto Express5S feature (FC 0890)

Crypto Express5S was introduced from z13 servers. On the initial configuration, a minimum of two features are installed. The number of features then increases one at a time up to a maximum of 16 features.

Each Crypto Express5S feature holds one PCI Express cryptographic adapter. Each adapter can be configured by the installation as a Secure IBM CCA coprocessor, as a Secure IBM Enterprise Public Key Cryptography Standards (PKCS) #11 (EP11) coprocessor, or as an accelerator.

¹¹ Subject to export regulations.

Each Crypto Express5S feature occupies one I/O slot in the PCIe I/O drawer, and features no CHPID assigned. However, it includes one PCHID.

4.8 Integrated Firmware Processor

The Integrated Firmware Processor (IFP) was introduced with the zEC12 and zBC12 servers. The IFP is dedicated for managing a new generation of PCIe features. The following features are installed in the PCIe+ I/O drawer:

- ▶ zEDC Express
- ▶ 25GbE RoCE Express2
- ▶ 10GbE RoCE Express2
- ▶ 10GbE RoCE Express
- ▶ IBM zHyperlink Express

All native PCIe features should be ordered in pairs for redundancy. The features are assigned to one of the four resource groups (RGs) that are running on the IFP according to their physical location in the PCIe+ I/O drawer, which provides management functions and virtualization functions.

If two features of the same type are installed, one always is managed by resource group 1 (RG 1) or resource group 3 (RG3) while the other feature is managed by resource group 2 (RG 2) or resource group 4 (RG 4). This configuration provides redundancy if one of the features or resource groups needs maintenance or fails.

The IFP and RGs support the following infrastructure management functions:

- ▶ Firmware update of adapters and resource groups
- ▶ Error recovery and failure data collection
- ▶ Diagnostic and maintenance tasks

For more information about the IFP and RGs, see Appendix C, “Native Peripheral Component Interconnect Express” on page 419.

4.9 zEDC Express

zEDC Express is an optional feature (FC 0420) that is available on z14 ZR1, z14 M0x, z13, z13s, zEC12, and zBC12 servers. It is designed to provide hardware-based acceleration for data compression and decompression.

The IBM zEnterprise Data Compression (zEDC) acceleration capability in z/OS and the zEDC Express feature helps to improve cross-platform data exchange, reduce CPU consumption, and save disk space.

The feature installs exclusively on the PCIe I/O drawer. Up to 16 features can be installed on the system. One PCIe adapter or compression coprocessor is available per feature, which implements compression as defined by RFC1951 (DEFLATE).

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

For more information about the management and definition of the zEDC feature, see Appendix F, “IBM zEnterprise Data Compression Express” on page 461, and Appendix C, “Native Peripheral Component Interconnect Express” on page 419.



Central processor complex channel subsystem

This chapter describes the concepts of the z14 ZR1 channel subsystem, including multiple channel subsystems and multiple subchannel sets. It also describes the technology, terminology, and implementation aspects of the channel subsystem.

This chapter includes the following topics:

- ▶ 5.1, “Channel subsystem” on page 162
- ▶ 5.2, “I/O configuration management” on page 170
- ▶ 5.3, “Channel subsystem summary” on page 171

5.1 Channel subsystem

Channel subsystem (CSS) is a collective name of facilities that Z servers use to control I/O operations.

The channel subsystem directs the flow of information between I/O devices and main storage. It allows data processing to proceed concurrently with I/O processing, which relieves data processors (central processor [CP], Integrated Facility for Linux [IFL]) of the task of communicating directly with I/O devices.

The channel subsystem includes subchannels, I/O devices that are attached through control units, and channel paths between the subsystem and control units. For more information about the channel subsystem, see 5.1.1, “Multiple logical channel subsystems”.

The design of IBM Z platform offers considerable processing power, memory size, and I/O connectivity. In support of the larger I/O capability, the CSS structure is scaled up by introducing the multiple logical channel subsystem (LCSS) since z990, and multiple subchannel sets (MSS) since z9.

An overview of the channel subsystem for z14 ZR1 servers is shown in Figure 5-1. z14 ZR1 servers are designed to support up to three logical channel subsystems, each with three subchannel sets and up to 256 channels.

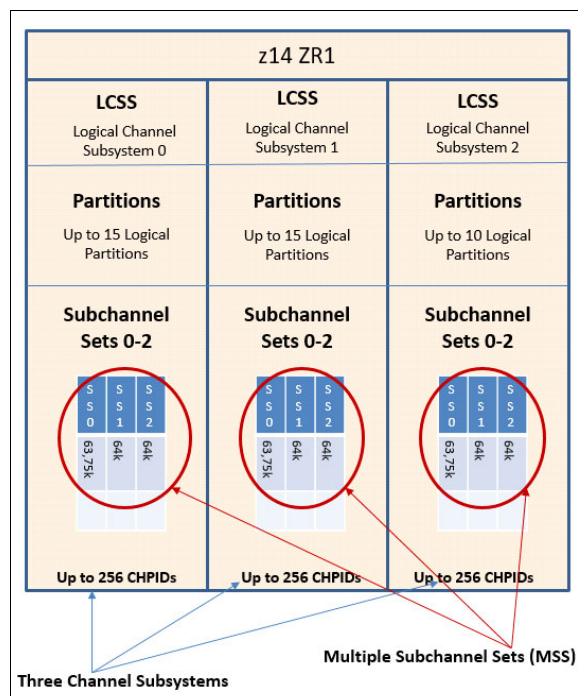


Figure 5-1 Multiple channel subsystems and multiple subchannel sets

All channel subsystems are defined within a single configuration, which is called I/O configuration data set (IOCDS). The IOCDS is loaded into the hardware system area (HSA) during a central processor complex (CPC) power-on reset (POR) to start all of the channel subsystems.

On z14 ZR1 servers, the HSA is pre-allocated in memory with a fixed size of 64 GB, which is in addition to the customer-purchased memory. This fixed size memory for HSA eliminates the requirement for more planning of the initial I/O configuration and pre-planning for future I/O expansions.

The following objects are always reserved in the z14 ZR1 HSA during POR, whether they are defined in the IOCDS for use:

- ▶ Three CSSs
- ▶ A total of 15 LPARs in each CSS0 to CSS1
- ▶ A total of 10 LPARs in CSS2
- ▶ Subchannel set 0 with 63.75 K devices in each CSS
- ▶ Subchannel set 1 with 64 K minus one device in each CSS
- ▶ Subchannel set 2 with 64 K minus one device in each CSS

5.1.1 Multiple logical channel subsystems

In the z/Architecture, a *single channel subsystem* can have up to 256 channel paths that are defined, which limited the total numbers of I/O connectivities on older Z servers to 256.

The introduction of *multiple LCSSs* enabled an IBM Z system to have more than one channel subsystems logically, while each logical channel subsystem maintains the same manner of I/O processing. Also, a logical partition (LPAR) is now attached to a specific logical channel subsystem, which makes the extension of multiple logical channel subsystems not apparent to the operating systems and applications. The multiple image facility (MIF) in the structure enables resource sharing across LPARs within a single LCSS or across the LCSSs.

The multiple LCSS structure extended the Z servers' total number of I/O connectivities to support a balanced configuration for the growth of processor and I/O capabilities.

A one-digit number ID starting from 0 (CSSID) is assigned to an LCSS, and a one-digit hexadecimal ID (MIF ID) starting from 0 is assigned to an LPAR within the LCSS.

Note: The phrase *channel subsystem* has same meaning as *logical channel subsystem* in this section, unless otherwise stated.

Subchannels

A *subchannel* provides the logical appearance of a device to the program and contains the information that is required for sustaining a single I/O operation. Each device is accessible by using one subchannel in a channel subsystem to which it is assigned according to the active IOCDS of the Z server.

A subchannel set (SS) is a collection of subchannels within a channel subsystem. The maximum number of subchannels of a subchannel set determines how many devices are accessible to a channel subsystem.

In z/Architecture, the first subchannel set of an LCSS can have 63.75 K subchannels (with 0.25 K reserved), with a subchannel set ID (SSID) of 0. By enabling the multiple subchannel sets, which are described in 5.1.2, "Multiple subchannel sets" on page 164, extra subchannel sets are available to increase the device addressability of a channel subsystem.

Channel paths

A *channel path* provides a connection between the channel subsystem and control units (CUs) that allows the channel subsystem to communicate with I/O devices. Depending on the type of connections, a channel path might be a physical connection to a control unit with I/O devices, such as FICON, or an internal logical control unit, such as HiperSockets.

Each channel path in a channel subsystem features a unique 2-digit hexadecimal identifier that is known as a channel-path identifier (CHPID), which ranges 00 - FF. Therefore, a total of 256 CHPIDs are supported by a CSS, and a maximum of 768 CHPIDs are available on a z14 ZR1 server with three logical channel subsystems.

By assigning a CHPID to a physical port of an I/O feature adapter, such as FICON Express16S+, or a fanout adapter (ICA SR) port, the channel subsystem connects to the I/O devices through these physical ports.

A port on an I/O feature card includes a unique physical channel identifier (PCHID) according to the physical location of this I/O feature adapter, and the sequence of this port on the adapter.

In addition, a port on a fanout adapter has a unique adapter identifier (AID), according to the physical location of this fanout adapter, and the sequence of this port on the adapter.

A CHPID is assigned to a physical port by defining the corresponding PCHID or AID in the I/O configuration definitions.

Control units

A *control unit* provides the logical capabilities that are necessary to operate and control an I/O device. It adapts the characteristics of each device so that it can respond to the standard form of control that is provided by the CSS.

A control unit can be housed separately or physically and logically integrated with the I/O device, channel subsystem, or within the Z server.

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 accessible through one or more channel paths that are connected to the control unit.

5.1.2 Multiple subchannel sets

A subchannel set is a collection of subchannels within a channel subsystem. The maximum number of subchannels of a subchannel set determines how many I/O devices that a channel subsystem can access. This number also determines the number of addressable devices to the program (for example, an operating system) that is running in the LPAR.

Each subchannel has a unique four-digit hexadecimal number 0x0000 - 0xFFFF. Therefore, a single subchannel set can address and access up to 64 K I/O devices.

MSS was introduced in z9 to extend the maximum number of addressable I/O devices for a channel subsystem.

As with the z13s server, the z14 ZR1 support three subchannel sets for each logical channel subsystem. It can access a maximum of 191.74 K devices for a logical channel subsystem and a logical partition and the programs that are running on it.

Note: Do not confuse the multiple subchannel sets function with multiple channel subsystems.

Subchannel number

The subchannel number is a four-digit hexadecimal number 0x0000 - 0xFFFF, which is assigned to a subchannel within a subchannel set of a channel subsystem. Subchannels in each subchannel set are always assigned subchannel numbers within a single range of contiguous numbers.

The lowest-numbered subchannel is subchannel 0, and the highest-numbered subchannel includes a subchannel number equal to one less than the maximum numbers of subchannels that are supported by the subchannel set. Therefore, a subchannel number is always unique within a subchannel set of a channel subsystem and depends on the sequence of assigning.

With the subchannel numbers, a program that is running on an LPAR (for example, an operating system) can specify all I/O functions relative to a specific I/O device by designating a subchannel that is assigned to the I/O devices.

Normally, subchannel numbers are used only in communication between the programs and the channel subsystem.

Subchannel set identifier

While introducing the MSS, the channel subsystem is extended to assign a value 0 - 2 for each subchannel set, which is the SSID. A subchannel can be identified by its SSID and subchannel number.

Device number

A device number is an arbitrary number 0x0000 - 0xFFFF, which is defined by a system programmer in an I/O configuration for naming an I/O device. The device number must be unique within a subchannel set of a channel subsystem. It is assigned to the corresponding subchannel by channel subsystem when an I/O configuration is activated. Therefore, a subchannel in a subchannel set of a channel subsystem includes a device number together with subchannel number for designating an I/O operation.

The device number provide a means to identify a device, independent of any limitations that are imposed by the system model, configuration, or channel-path protocols.

A device number also can be used to designate an I/O function to a specific I/O device. Because it is an arbitrary number, it can easily be fit into any configuration management and operating management scenarios.

With multiple subchannel sets, a subchannel is assigned to a specific I/O device by the channel subsystem with an automatically assigned subchannel number and a device number that is defined by user. An I/O device can always be identified by an SSID with a subchannel number or a device number. For example, a second PAV exposure for a device with device number AB00 in subchannel set zero can be designated as 1AB00 in subchannel set 1.

Normally, the subchannel number is used by the programs to communicate with the channel subsystem and I/O device, whereas the device number is used by a system programmer, operator, and administrator.

Device in subchannel set 0 and extra subchannel sets

An LCSS always includes the first subchannel set (SSID 0), which can have up to 63.75 K subchannels with 256 subchannels that are reserved by the channel subsystem. Users can always define their I/O devices in this subchannel set for general use.

For the extra subchannel sets that are enabled by the MSS facility, each has 65535 subchannels (64 K minus one) for specific types of devices. These extra subchannel sets are referred as *alternative subchannel sets* in z/OS. Also, a device that is defined in an alternative subchannel set is considered a *special device*, which often features a special device type in the I/O configuration.

Currently, a z14 ZR1 server that is running z/OS defines the following types of devices in another subchannel set, with the minimum release or required PTF installed:

- ▶ Alias devices of the parallel access volumes (PAV).
- ▶ Secondary devices of GDPS Metro Mirror Copy Service, which is formerly known as Peer-to-Peer Remote Copy (PPRC).
- ▶ FlashCopy SOURCE and TARGET devices.
- ▶ Db2 data backup volumes.

The use of another subchannel set for these special devices helps reduce the number of devices in the subchannel set 0, which increases the growth capability for accessing more devices.

Initial program load from an alternative subchannel set

z14 ZR1 servers support initial program load (IPL) from alternative subchannel sets in addition to subchannel set 0. Devices that are used early during IPL processing now can be accessed by using subchannel set 1, or subchannel set 2 on a z14 ZR1 server. This configuration allows the users of Metro Mirror (formerly PPRC) secondary devices that are defined by using the same device number and a new device type in an alternative subchannel set to be used for IPL, an I/O definition file (IODF), and stand-alone memory dump volumes, when needed.

IPL from an alternative subchannel set is supported by z/OS V1.13 or later.

z/OS display ios,config command

The z/OS **display ios,config(a11)** command that is shown in Figure 5-2 includes information about the MSSs.

```
-D IOS,CONFIG(ALL)
IOS506I 14.10.48 I/O CONFIG DATA 329
ACTIVE IODF DATA SET = SYS6.IODF78
CONFIGURATION ID = ITS0          EDT ID = 01
TOKEN:  PROCESSOR DATE      TIME      DESCRIPTION
SOURCE: MUSCA    18-03-02 09:47:04 SYS9    IODF78
ACTIVE CSS:  1    SUBCHANNEL SETS CONFIGURED: 0, 1, 2
CHANNEL MEASUREMENT BLOCK FACILITY IS ACTIVE
SUBCHANNEL SET FOR PPRC PRIMARY: INITIAL = 0    ACTIVE = 0
HYPERSWAP FAILOVER HAS OCCURRED: NO
LOCAL SYSTEM NAME (LSYSTEM): MUSCA
HARDWARE SYSTEM AREA AVAILABLE FOR CONFIGURATION CHANGES
PHYSICAL CONTROL UNITS                8041
CSS 0 - LOGICAL CONTROL UNITS          3947
SS 0  SUBCHANNELS                      50000
SS 1  SUBCHANNELS                      65055
SS 2  SUBCHANNELS                      65535
CSS 1 - LOGICAL CONTROL UNITS          3962
SS 0  SUBCHANNELS                      50112
SS 1  SUBCHANNELS                      65055
SS 2  SUBCHANNELS                      65535
CSS 2 - LOGICAL CONTROL UNITS          4088
SS 0  SUBCHANNELS                      65280
SS 1  SUBCHANNELS                      65535
SS 2  SUBCHANNELS                      65535
```

Figure 5-2 Output for display ios,config(all) command with MSS

5.1.3 Channel path spanning

With the implementation of multiple LCSSs, a channel path can be available to LPARs as dedicated, shared, and spanned.

Although a shared channel path can be shared by LPARs within a same LCSS, a spanned channel path can be shared by LPARs within and across LCSSs.

By assigning the same CHPID from different LCSSs to the same channel path (for example, a PCHID), the channel path can be accessed by any LPARs from these LCSSs at the same time. The CHPID is spanned across those LCSSs. The use of spanned channels paths decreases the number of channels that are needed in an installation of Z servers.

A sample of channel paths that are defined as dedicated, shared, and spanned is shown in Figure 5-3.

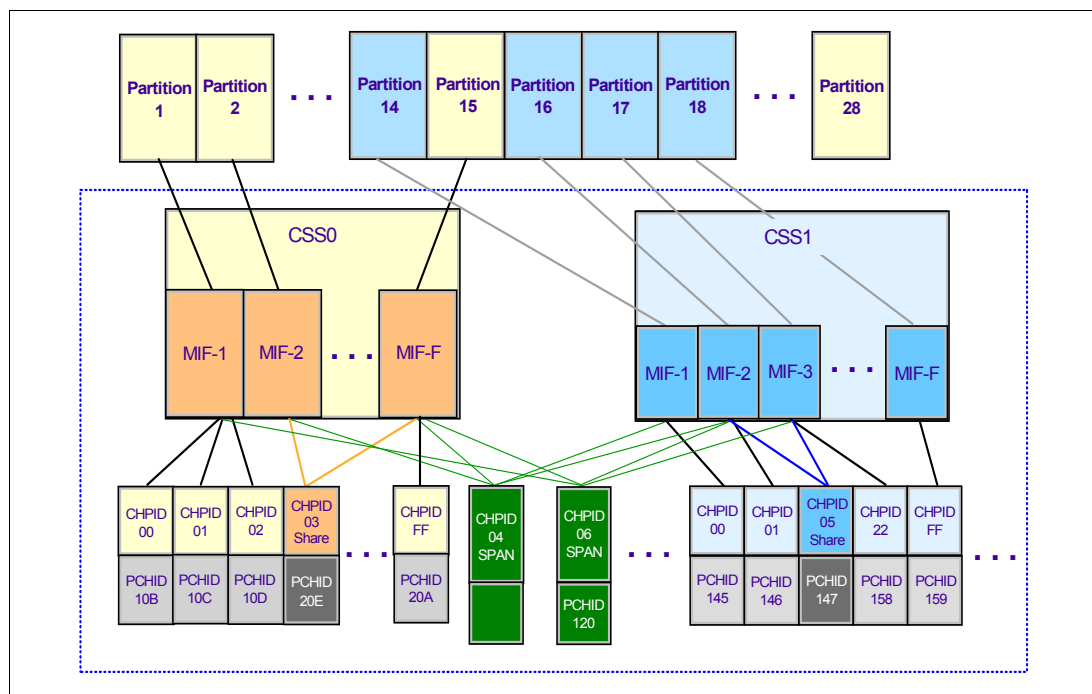


Figure 5-3 IBM Z CSS: Channel subsystems with channel spanning

The following definitions of a channel path are shown in Figure 5-3:

- ▶ CHPID FF, which is assigned to PCHID 20A is dedicated access for partition 15 of LCSS0. The same applies to CHPID 00,01,02 of LCSS0, and CHPID 00,01,FF of LCSS1.
- ▶ CHPID 03, which is assigned to PCHID 20E is shared access for partition 2, and 15 of LCSS0. The same applies to CHPID 05 of LCSS1.
- ▶ CHPID 06, which is assigned to PCHID 120 is spanned access for partition 1, 15 of LCSS0, and partition 16, 17 of LCSS1. The same applies to CHPID 04.

Channel spanning is supported for internal links (HiperSockets and IC links) and for certain types of external links. External links that are supported on z14 ZR1 servers include FICON Express16S+, FICON Express16S, FICON Express8S, OSA-Express6S, OSA-Express5S, and Coupling Links.

The definition of LPAR name, MIF image ID, and LPAR ID are used to identify an LPAR by the channel subsystem to identify I/O functions from different LPARs of multiple LCSSs, which support the implementation of these dedicated, shared, and spanned paths.

An example of definition of these LPAR-related identifications is shown in Figure 5-4.

CSS0			CSS1			CSS2		Specified in HCD / IOCP
Logical Partition Name			Logical Partition Name			LPAR Name		Specified in HCD / IOCP
TST1	PROD1	PROD2	TST2	PROD3	PROD4	TST3	TST4	
Logical Partition ID			Logical Partition ID			LPAR ID		Specified in Image Profile
02	04	0A	14	16	1D	22	26	
MIF ID			MIF ID			MIF ID		Specified in HCD / IOCP
2	4	A	4	6	D	2	6	

Figure 5-4 CSS, LPAR, and identifier example

LPAR name

The LPAR name is defined as partition name parameter in the **RESOURCE** statement of an I/O configuration. The LPAR name must be unique across the server.

MIF image ID

The MIF image ID is defined as a parameter for each LPAR in the **RESOURCE** statement of an I/O configuration. It ranges 1 - F, and must be unique within an LCSS. However, duplicates are allowed in different LCSSs.

If a MIF image ID is not defined, an arbitrary ID is assigned when the I/O configuration activated. The z14 ZR1 server supports a maximum of three LCSSs, with a total of 40 LPARs that can be defined. Each LCSS of a z14 ZR1 server can support the following numbers of LPARs:

- ▶ LCSS0 - LCSS1 support 15 LPARs each, and the MIF image ID is 1 - F.
- ▶ LCSS2 supports 10 LPARs, and the MIF image IDs are 1 - A.

LPAR ID

The LPAR ID is defined by a user in an image activation profile for each LPAR. It is a 2-digit hexadecimal number 00 - 7F. The LPAR ID must be unique across the server.

Although it is arbitrarily defined by the user, an LPAR ID often is the CSS ID concatenated to its MIF image ID, which makes the value more meaningful for the system administrator. For example, an LPAR with LPAR ID 1A that is defined in this manner means that the LPAR is defined in LCSS1, with the MIF image ID A.

5.2 I/O configuration management

The following tools are available to help maintain and optimize the I/O configuration:

- ▶ IBM Configurator for e-business (eConfig)

The eConfig tool is available from your IBM representative. It is used to create configurations or upgrades of a configuration, and maintains tracking to the installed features of those configurations. eConfig produces reports that help you understand the changes that are being made for a new system or a system upgrade, and the components of the target configuration.

- ▶ Hardware configuration definition (HCD)

HCD supplies an interactive dialog to generate the IODF, and later the IOCDS. Generally, use HCD or Hardware Configuration Manager (HCM) to generate the I/O configuration rather than writing input/output configuration program (IOCP) statements. The validation checking that HCD runs against a IODF source file helps minimize the risk of errors before an I/O configuration is activated.

HCD support for multiple channel subsystems is available with z/VM and z/OS. HCD provides the capability to make dynamic hardware and software I/O configuration changes.

Note: Certain functions might require specific levels of an operating system, PTFs, or both.

- ▶ Consult the appropriate fix categories:

- z14 M0x: IBM.Device.Server.z14-3906
- z14 ZR1: IBM.Device.Server.z14ZR1-3907
- z13: IBM.Device.Server.z13-2964
- z13s: IBM.Device.Server.z13s-2965
- zEC12: IBM.Device.Server.zEC12-2827
- zBC12: IBM.Device.Server.zBC12-2828

- ▶ HCM

HCM is a priced optional feature that supplies a graphical interface of HCD. It is installed on a PC and allows you to manage the physical and logical aspects of a mainframe's hardware configuration.

- ▶ CHPID Mapping Tool (CMT)

The CMT helps to map CHPIDs onto PCHIDs that are based on an IODF source file and the eConfig configuration file of a mainframe. It provides a CHPID to PCHID mapping with high availability for the targeted I/O configuration. It also features built-in mechanisms to generate a mapping according to customized I/O performance groups. More enhancements are implemented in CMT to support z14 ZR1 servers. The CMT is available for download from the [IBM Resource Link website](#) (login is required).

5.3 Channel subsystem summary

z14 ZR1 servers support the channel subsystem features of multiple LCSS, MSS, and the channel spanning that is described in this chapter. The channel subsystem capabilities of z14 ZR1 servers are listed in Table 5-1.

Table 5-1 z14 ZR1 CSS overview

Maximum number of CSSs	3
Maximum number of LPARs per CSS	CSS0 - CSS1: 15 CSS2: 10
Maximum number of LPARs per system	40
Maximum number of subchannel sets per CSS	3
Maximum number of subchannels per CSS	191.74 K SS0: 65280 SS1 - SS2: 65535
Maximum number of CHPIDs per CSS	256



Cryptographic features

This chapter describes the hardware cryptographic functions that are available on IBM z14 ZR1. The CP Assist for Cryptographic Function (CPACF), together with the Peripheral Component Interconnect Express (PCIe) cryptographic coprocessors, offer a balanced use of processing resources and unmatched scalability for fulfilling pervasive encryption demands.

The z14 ZR1 is designed for delivering a transparent and consumable approach that enables extensive (pervasive) encryption of data in flight and at rest, with the goal of substantially simplifying data security and reducing the costs that are associated with protecting data while achieving compliance mandates.

This chapter also introduces the principles of cryptography and describes the implementation of cryptography in the hardware and software architecture of IBM Z. It also describes the features that IBM z14 ZR1 offers. Finally, the chapter summarizes the cryptographic features and required software.

Naming: The IBM z14 server generation is available as the following machine types and models:

- ▶ Machine Type 3906 (M/T 3906), Models M01, M02, M03, M04, and M05 → further identified as *IBM z14 Model M0x*, or *z14 M0x*, unless otherwise specified.
- ▶ Machine Type 3907 (M/T 3907), Model ZR1 → further identified as *IBM z14 Model ZR1*, or *z14 ZR1*, unless otherwise specified.

In the remainder of this chapter, *IBM z14 (z14)* refers to both machine types.

This chapter includes the following topics:

- ▶ 6.1, “Cryptography enhancements on IBM z14 ZR1” on page 174
- ▶ 6.2, “Cryptography overview” on page 175
- ▶ 6.3, “Cryptography on IBM z14 ZR1” on page 179
- ▶ 6.4, “CP Assist for Cryptographic Functions” on page 182
- ▶ 6.5, “Crypto Express6S” on page 188
- ▶ 6.6, “TKE workstation” on page 200
- ▶ 6.7, “Cryptographic functions comparison” on page 205
- ▶ 6.8, “Cryptographic operating system support for z14 ZR1” on page 207

6.1 Cryptography enhancements on IBM z14 ZR1

IBM z14 introduced the new PCI Crypto Express6S feature, together with an improved CPACF Coprocessor, that is managed by a new Trusted Key Entry (TKE) workstation. In addition, the IBM Common Cryptographic Architecture (CCA) and the IBM Enterprise PKCS #11 (EP11) Licensed Internal Code (LIC) were enhanced.

The new functions support new standards and are designed to meet the following compliance requirements:

- ▶ Payment Card Industry (PCI) Hardware Security Module (HSM) certification to strengthen the cryptographic standards for attack resistance in the payment card systems area.
PCI HSM certification is exclusive for Crypto Express6S.
- ▶ National Institute of Standards and Technology (NIST) through the Federal Information Processing Standard (FIPS) standard to implement guidance requirements.
- ▶ Common Criteria EP11 EAL4.
- ▶ German Banking Industry Commission (GBIC).
- ▶ VISA Format Preserving Encryption (VFPE) for credit card numbers.
- ▶ Enhanced public key Elliptic Curve Cryptography (ECC) for users such as Chrome, Firefox, and Apple's iMessage.

These enhancements are described in this chapter.

IBM z14 ZR1 includes standard cryptographic hardware and optional cryptographic features for flexibility and growth capability. IBM has a long history of providing hardware cryptographic solutions. This history stretches from the development of the Data Encryption Standard (DES) in the 1970s to the Crypto Express tamper-sensing and tamper-responding programmable features.

Crypto Express is designed to meet the US Government's highest security rating of FIPS 140-2 Level 4¹. It also meets several other security ratings, such as the Common Criteria for Information Technology Security Evaluation, the PCI HSM, criteria, and the criteria for German Banking Industry Commission (formerly known as Deutsche Kreditwirtschaft evaluation).

The cryptographic functions include the full range of cryptographic operations that are necessary for local and global business and financial institution applications. User Defined Extensions (UDX) allow you to add custom cryptographic functions to the functions that z14 systems offer.

¹ FIPS 140-2 Security Requirements for Cryptographic Modules.

6.2 Cryptography overview

From the early beginning of human history, when two people were communicating with each other, the demand always existed to keep certain messages secret so that a third person cannot understand what the sender is telling to the receiver.

Also, it is necessary to ensure that a message cannot be corrupted, while ensuring that the sender and the receiver really are the persons who they claim to be. Over time, several methods were used to achieve these objectives, with more or less success. Many procedures and algorithms for encrypting and decrypting data were developed that are increasingly complicated and time-consuming.

6.2.1 Modern cryptography

With the development of computing technology, the encryption and decryption algorithms can be performed by computers, which enables the use of complicated mathematical algorithms. Most of these algorithms are based on the prime factorization of large numbers.

Modern cryptography involves the following purposes for protecting information:

- Data protection

The protection of data usually is the main concept that is associated with cryptography. Only authorized persons should be able to read the message or get information about it. Data is encrypted by using a known algorithm and secret keys, such that the intended party can de-scramble the data, but an interloper cannot. This concept is also referred to as *confidentiality*.

- Authentication (identity validation)

This process decides whether the communication partners are who they claim to be, which can be done by using certificates and signatures. It must be possible to clearly identify the owner of the data or the sender and the receiver of the message.

- Integrity

The verification of data ensures that what was received is identical to what was sent. It must be proven that the data is complete and was not altered during the moment it was transmitted (by the sender) and the moment it was received (by the receiver).

- Non-repudiation

It must be impossible for the owner of the data or the sender of the message to deny authorship. Non-repudiation ensures that both sides of a communication know that the other side agreed to what was exchanged, and not someone else. This specification implies a legal liability and contractual obligation, which is the same as a signature on a contract.

These goals should all be possible without unacceptable overhead to the communication. The goal is to keep the system secure, manageable, and productive.

The basic method data protection is to encrypt and decrypt it, while hash algorithms, message authentication codes (MACs), digital signatures, and certificates are used for authentication, integrity, and non-repudiation.

When encrypting a message, the sender transforms the clear text into a secret text. Doing so requires the following main elements:

- ▶ The *algorithm* is the mathematical or logical formula that is applied to the key and the clear text to deliver a ciphered result, or to take a ciphered text and deliver the original clear text.
- ▶ The *key* ensures that the result of the encrypting data transformation by the algorithm is only the same when the same key is used. That decryption of a ciphered message results only in the original clear message when the correct key is used. Therefore, the receiver of a ciphered message must know which algorithm and which key must be used to decrypt the message.

6.2.2 Kerckhoffs' principle

In modern cryptography, the algorithm is published and known to everyone, whereas the keys are kept secret. This configuration corresponds to Kerckhoffs' principle, which is named after Auguste Kerckhoffs, a Dutch cryptographer, who formulated it in 1883:

"A system should not depend on secrecy, and it should be able to fall into the enemy's hands without disadvantage."

In other words, the security of a cryptographic system should depend on the security of the key, so the key must be kept secret. Therefore, the secure management of keys is the primal task of modern cryptographic systems.

Adhering to Kerckhoffs' Principle is done for the following reasons:

- ▶ It is much more difficult to keep an algorithm secret than a key.
- ▶ It is harder to exchange a compromised algorithm than to exchange a compromised key.
- ▶ Secret algorithms can be reconstructed by reverse engineering software or hardware implementations.
- ▶ Errors in public algorithms can generally be found more easily, when many experts look into it.
- ▶ In history, most secret encryption methods proved to be weak and inadequate.
- ▶ When a secret encryption method is used, it is possible that a back door was built in.
- ▶ If an algorithm is public, many experts can form an opinion about it. Also, the method can be more thoroughly investigated for potential weaknesses and vulnerabilities.

6.2.3 Keys

The keys that are used for the cryptographic algorithms often are sequences of numbers and characters, but can also be any other sequence of bits. The length of a key influences the security (strength) of the cryptographic method. The longer the used key, the more difficult it is to compromise a cryptographic algorithm.

For example, the DES (symmetric key) algorithm uses keys with a length of 56 bits, Triple-DES (TDES) uses keys with a length of 112 bits, and Advanced Encryption Standard (AES) uses keys of 128, 192, 256, or 512 bits. The asymmetric key RSA algorithm (named after its inventors Rivest, Shamir, and Adleman) uses keys with a length of 1024 - 4096 bits.

In modern cryptography, keys must be kept secret. Depending on the effort that is made to protect the key, keys are classified into the following levels:

- ▶ A *clear key* is a key that is transferred from the application in clear text to the cryptographic function. The key value is stored in the clear (at least briefly) somewhere in unprotected memory areas. Therefore, the key can be made available to someone under certain circumstances who is accessing this memory area.

This risk must be considered when clear keys are used. However, many applications exist where this risk can be accepted. For example, the transaction security for the (widely used) encryption methods Secure Sockets Layer (SSL) and Transport Layer Security (TLS) is based on clear keys.

- ▶ The value of a *protected key* is stored only in clear in memory areas that cannot be read by applications or users. The key value does not exist outside of the physical hardware, although the hardware might not be tamper-resistant. The principle of protected keys is unique to IBM Z systems. For more information, see 6.4.2, “CPACF protected key” on page 185.
- ▶ For a *secure key*, the key value does not exist in clear format outside of a special hardware device (HSM), which must be secured and tamper-resistant. A secure key is protected from disclosure and misuse, and can be used for the trusted execution of cryptographic algorithms on highly sensitive data. If used and stored outside of the HSM, a secure key must be encrypted with a *master key*, which is created within the HSM and never leaves the HSM.

Because a secure key must be handled in a special hardware device, the use of secure keys usually is far slower than the use of clear keys, as shown in Figure 6-1.

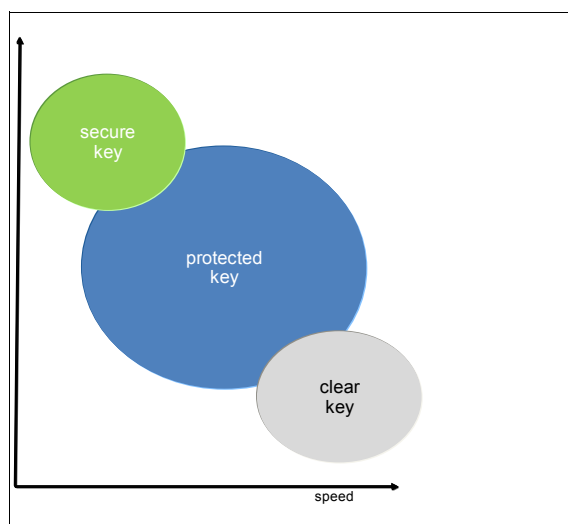


Figure 6-1 Three levels of protection with three levels of speed

6.2.4 Algorithms

The following algorithms of modern cryptography are differentiated based on whether they use the same key for the encryption of the message as for the decryption:

- *Symmetric algorithms* use the same key to encrypt and to decrypt data. The function that is used to decrypt the data is the opposite of the function that is used to encrypt the data. Because the same key is used on both sides of an operation, it must be negotiated between both parties and kept secret. Therefore, symmetric algorithms are also known as *secret key algorithms*.

The main advantage of symmetric algorithms is that they are fast and therefore can be used for large amounts of data, even if they are not run on specialized hardware. The disadvantage is that the key must be known by both sender and receiver of the messages, which implies that the key must be exchanged between them. This key exchange is a weak point that can be attacked.

Prominent examples for symmetric algorithms are DES, TDES, and AES.

- *Asymmetric algorithms* use two distinct but related key: the *public key* and the *private key*. As the names imply, the private key must be kept secret, whereas the public key is shown to everyone. However, with asymmetric cryptography, it is not important who sees or knows the public key. Whatever is done with one key can be undone by the other key only.

For example, data that is encrypted by the public key can be decrypted by the associated private key only, and vice versa. Unlike symmetric algorithms, which use distinct functions for encryption and decryption, only one function is used in asymmetric algorithms. Depending on the values that are passed to this function, it encrypts or decrypts the data. Asymmetric algorithms are also known as *public key algorithms*.

Asymmetric algorithms use complex calculations and are relatively slow (about 100 - 1000 times slower than symmetric algorithms). Therefore, such algorithms are not used for the encryption of bulk data.

Because the private key is never exchanged between the parties in communication, they are less vulnerable than symmetric algorithms. Asymmetric algorithms mainly are used for authentication, digital signatures, and for the encryption and exchange of secret keys, which in turn are used to encrypt bulk data with a symmetric algorithm.

Examples for asymmetric algorithms are RSA and the elliptic curve algorithms.

- *One-way algorithms* are not cryptographic functions. They do not use keys, and they can scramble data only, not de-scramble it. These algorithms are used extensively within cryptographic procedures for digital signing and tend to be developed and governed by using the same principles as cryptographic algorithms. One-way algorithms are also known as *hash algorithms*.

The most prominent one-way algorithms are the Secure Hash Algorithms (SHA).

6.3 Cryptography on IBM z14 ZR1

In principle, cryptographic algorithms can run on processor hardware. However, these workloads are compute-intensive, and the handling of secure keys also requires special hardware protection. Therefore, IBM Z offer several cryptographic hardware features, which are specialized to meet the requirements for cryptographic workload.

The cryptographic hardware that is supported on IBM z14 ZR1 is shown in Figure 6-2. These features are described in this chapter.

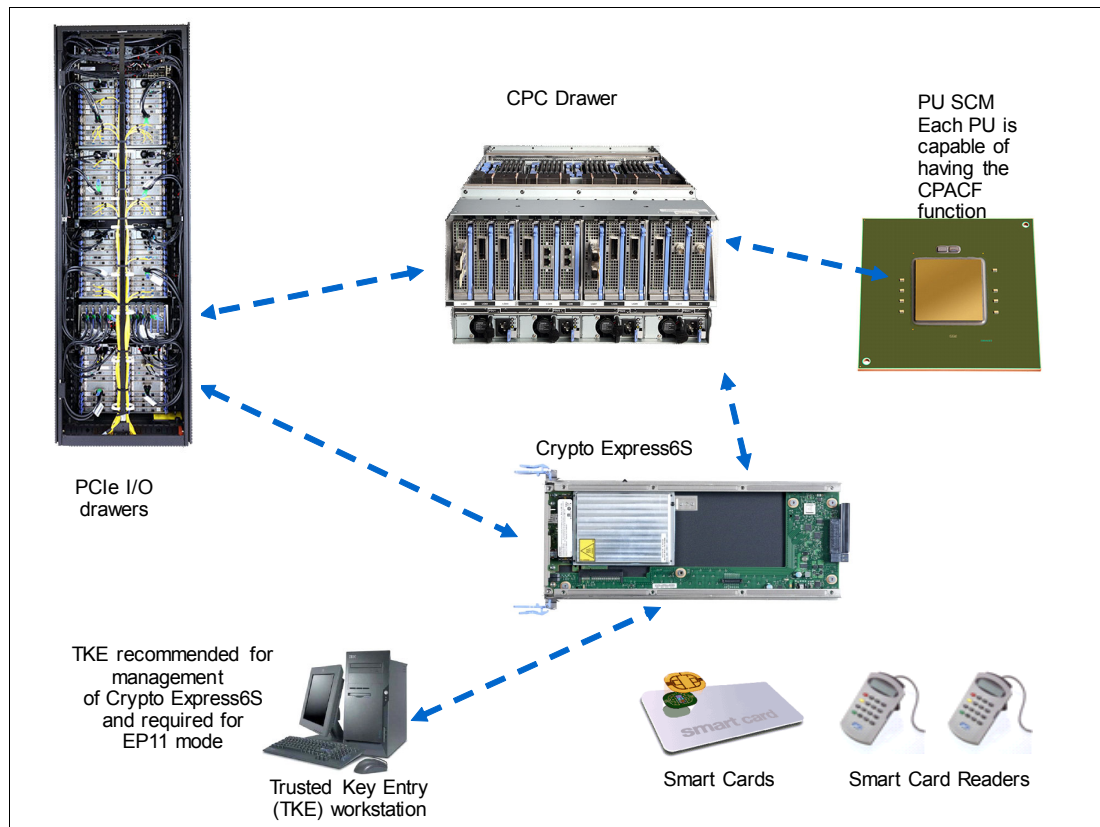


Figure 6-2 Cryptographic hardware that is supported in IBM z14 ZR1

Implemented in every processor unit (PU) or core in a central processor complex (CPC) is a cryptographic coprocessor that can be used² for cryptographic algorithms that uses clear keys or protected keys. For more information, see 6.4, “CP Assist for Cryptographic Functions” on page 182.

The Crypto Express6S adapter is an HSM that is placed in the PCIe+ I/O drawer of z14 ZR1. It also supports cryptographic algorithms by using secret keys. For more information, see 6.5, “Crypto Express6S” on page 188.

Finally, a TKE workstation is required for entering keys in a secure way into the Crypto Express6S HSM, which often also is equipped with smart card readers. For more information, see 6.6, “TKE workstation” on page 200.

² CPACF enablement feature must be ordered (FC 3863).

The feature codes and purpose of the cryptographic hardware features that are available for IBM z14 ZR1 are listed in Table 6-1.

Table 6-1 Cryptographic features for IBM z14 ZR1

Feature code	Description
3863	<p>CP Assist for Cryptographic Function (CPACF) enablement</p> <p>This feature is a prerequisite to use CPACF (except for SHA-1, SHA-224, SHA-256, SHA-384, and SHA-512) and the Crypto Express6S and Crypto Express5S feature.</p>
0893	<p>Crypto Express6S adapter</p> <p>A maximum of 16 features can be ordered (minimum of two features). This feature is optional and each feature of which contains one PCI Express cryptographic adapter (adjunct processor). This feature is supported only in z14.</p>
0890	<p>Crypto Express5S adapter</p> <p>This feature is available as a carry forward MES from z13s. The maximum supported number of Crypto Express5S and Crypto Express6S together is 16. This feature is optional and each feature of which contains one PCI Express cryptographic adapter (adjunct processor). This feature is supported only in z14, z13, and z13s systems.</p>
0086	<p>TKE tower workstation</p> <p>A TKE provides basic key management (key identification, exchange, separation, update, and backup) and security administration. It is optional for running a Crypto Express6S or Crypto Express5S adapter in CCA mode and required for running it in EP11 mode. The TKE workstation has one Ethernet port, and supports connectivity to an Ethernet local area network (LAN) operating at 10, 100, or 1000 Mbps. Up to 10 features per z14 ZR1 can be ordered.</p>
0085	<p>TKE rack-mounted workstation</p> <p>The rack-mounted version of the TKE, which needs a customer-provided standard 19-inch rack or the use of the new 16U Reserved feature (FC 0617) for the z14 ZR1 rack. It features a 1u TKE unit and an (optional) 1u console tray (screen, keyboard, and pointing device). When smart card readers are used, another customer-provided tray is needed. Up to 10 features per z14 ZR1 can be ordered.</p>
0880	<p>TKE 9.1 Licensed Internal Code (LIC)</p> <p>Included with the TKE tower workstation FC 0847 and the TKE rack-mounted workstation FC 0085 for z14 ZR1. Earlier versions of TKE features (feature codes: 0080, 0081, 0081, 0085, 0086, and 0849) can also be upgraded to TKE 9.1 LIC.</p>
0879	<p>TKE 9.0 Licensed Internal Code (LIC)</p> <p>Included with the TKE tower workstation FC 0847 and the TKE rack-mounted workstation FC 0085 for z14. Earlier versions of TKE features (feature codes 0080, 0081, 0081, 0085, 0086, and 0849) can also be upgraded to TKE 9.0 LIC.</p>
0891	<p>TKE Smart Card Reader</p> <p>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.</p>

Feature code	Description
0900	New TKE smart cards This will allow the TKE to support zones with EC 521 key strength (EC 521 strength for Logon Keys, Authority Signature Keys, and EP11 signature keys).
0892	More TKE smart cards When one feature code is ordered, 10 smart cards are included. The order increment is 1 - 99 (990 blank smart cards).

A TKE includes support for the AES encryption algorithm with 256-bit master keys and key management functions to load or generate master keys to the cryptographic coprocessor.

If the TKE workstation is chosen to operate the Crypto Express5S adapter in a z14 ZR1, TKE workstation with the TKE 9.x LIC is required. For more information, see 6.6, “TKE workstation” on page 200.

Important: Products that include any of the cryptographic feature codes contain cryptographic functions that are subject to special export licensing requirements by the United States Department of Commerce. It is your responsibility to understand and adhere to these regulations when you are moving, selling, or transferring these products.

To access and use the cryptographic hardware devices that are provided by z14 ZR1, the application must use an application programming interface (API) that is provided by the operating system. In z/OS, the Integrated Cryptographic Service Facility (ICSF) provides the APIs and is managing the access to the cryptographic devices, as shown in Figure 6-3.

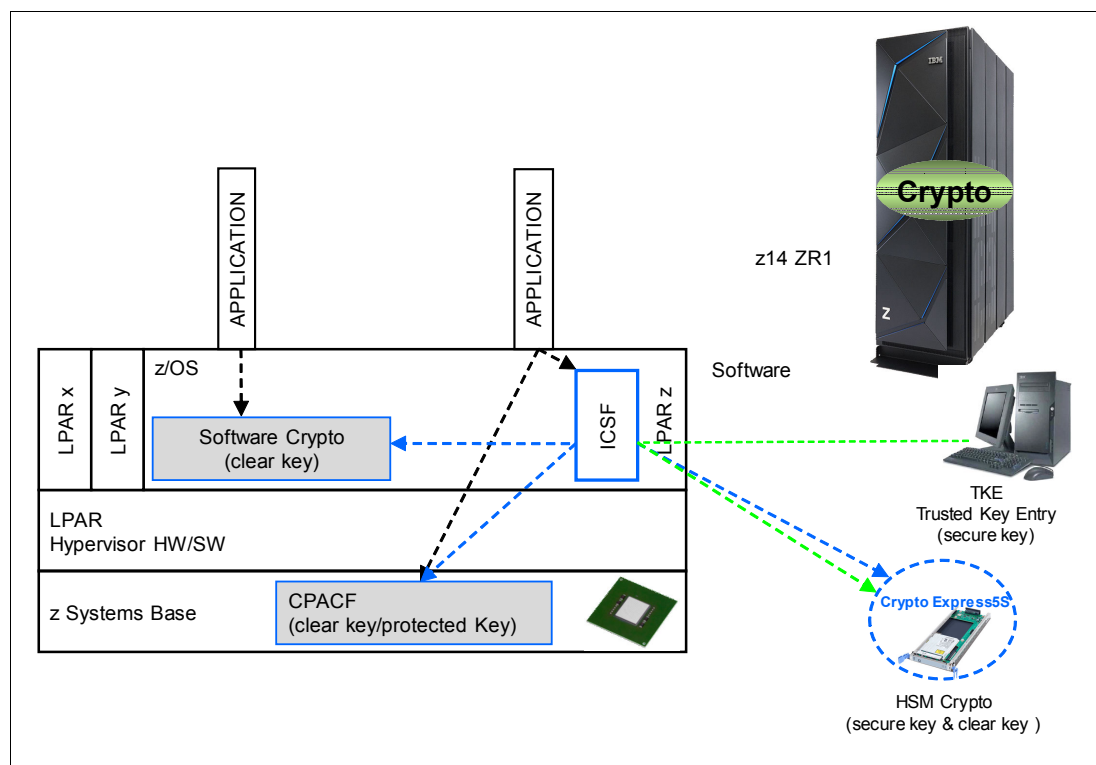


Figure 6-3 z14 ZR1 cryptographic support in z/OS

ICSF is a software component of z/OS. ICSF works with the hardware cryptographic features and the Security Server (IBM Resource Access Control Facility [IBM RACF®] element) to provide secure, high-speed cryptographic services in the z/OS environment. ICSF provides the APIs by which applications request the cryptographic services, and from the CPACF and the Crypto Express6S and Crypto Express5S adapter.

ICSF transparently routes application requests for cryptographic services to one of the integrated cryptographic engines (CPACF or a Crypto Express6S/Crypto Express5S adapter), depending on performance or requested cryptographic function. ICSF is also the means by which the secure Crypto Express6S/Crypto Express5S adapters are loaded with master key values, which allows the hardware features to be used by applications.

The cryptographic hardware that is installed in z14 ZR1 determines the cryptographic features and services that are available to the applications.

The users of the cryptographic services call the ICSF API. Some functions are performed by the ICSF software without starting the cryptographic hardware features. Other functions result in ICSF going into routines that contain proprietary IBM Z crypto instructions. These instructions are run by a CPU engine and result in a work request that is generated for a cryptographic hardware feature.

6.4 CP Assist for Cryptographic Functions

Attached to every PU (core) of a z14 ZR1 system are two independent engines, one for compression and one for cryptographic functions, as shown in Figure 6-4.

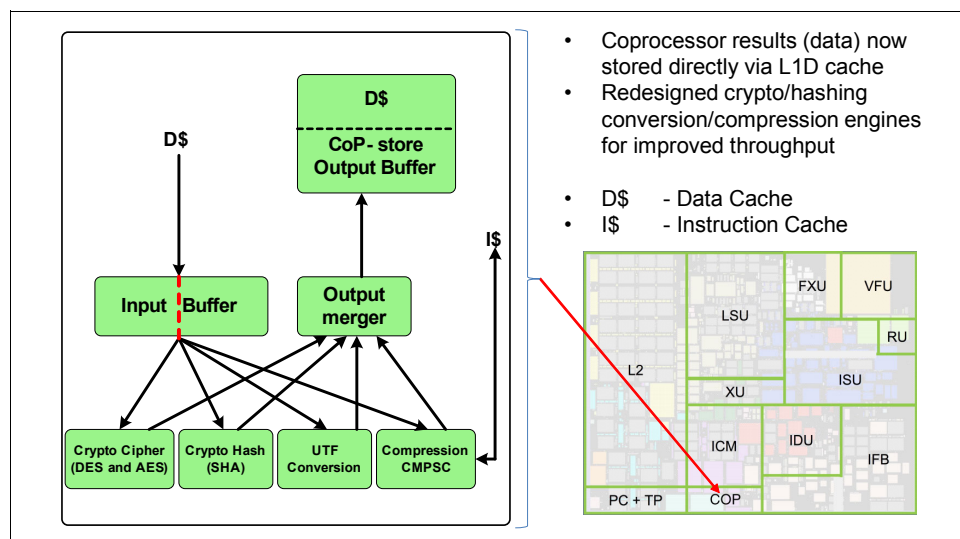


Figure 6-4 The cryptographic coprocessor CPACF

This cryptographic coprocessor, known as the CPACF, is not qualified as an HSM; therefore, it is not suitable for handling algorithms that use secret keys. However, the coprocessor can be used for cryptographic algorithms that use clear keys or protected keys. The CPACF works synchronously with the PU, which means that the owning processor is busy when its coprocessor is busy. This setup provides a fast device for cryptographic services.

CPACF supports pervasive encryption. Simple policy controls allow business to enable encryption to protect data in mission-critical databases without the need to stop the database or re-create database objects. Database administrators can use z/OS Dataset Encryption, z/OS Coupling Facility Encryption, z/VM encrypted hypervisor paging, and z/TPF transparent database encryption, which use performance enhancements in the hardware.

The CPACF offers a set of symmetric cryptographic functions that enhances the encryption and decryption performance of clear key operations. These functions are for SSL, virtual private network (VPN), and data-storing applications that do not require FIPS 140-2 Level 4 security.

CPACF is designed to facilitate the privacy of cryptographic key material when used for data encryption through key wrapping implementation. It ensures that key material is not visible to applications or operating systems during encryption operations. For more information, see 6.4.2, “CPACF protected key” on page 185.

The CPACF feature provides hardware acceleration for the following cryptographic services:

- ▶ DES
- ▶ Triple-DES
- ▶ AES-128
- ▶ AES-192
- ▶ AES-256 (all for clear and protected keys)
- ▶ SHA-1
- ▶ SHA-256 (SHA-2 or SHA-3 standard)
- ▶ SHA-384 (SHA-2 or SHA-3 standard)
- ▶ SHA-512 (SHA-2 or SHA-3 standard)
- ▶ SHAKE-128
- ▶ SHAKE-256
- ▶ PRNG
- ▶ DRNG
- ▶ TRNG

It provides high-performance hardware encryption, decryption, hashing, and random number generation support. The following instructions support the cryptographic assist function:

- ▶ KMAC: Compute Message Authentic Code
- ▶ KM: Cipher Message
- ▶ KMC: Cipher Message with Chaining
- ▶ KMF: Cipher Message with CFB
- ▶ KMCTR: Cipher Message with Counter
- ▶ KMO: Cipher Message with OFB
- ▶ KIMD: Compute Intermediate Message Digest
- ▶ KLMD: Compute Last Message Digest
- ▶ PCKMO: Provide Cryptographic Key Management Operation

These functions are provided as problem-state z/Architecture instructions that are directly available to application programs. These instructions are known as Message-Security Assist (MSA). When enabled, the CPACF runs at processor speed for every CP, IFL, and zIIP. For more information about MSA instructions, see *z/Architecture Principles of Operation*, SA22-7832.

The CPACF must be enabled by using an enablement feature (feature code 3863), which is available for no extra charge. The exception is support for the hashing algorithms SHA-1, SHA-256, SHA-384, and SHA-512, which is always enabled.

6.4.1 Cryptographic synchronous functions

Because the CPACF works synchronously with the PU, it provides cryptographic synchronous functions. For IBM and client-written programs, CPACF functions can be started by using the MSA instructions. z/OS ICSF callable services on z/OS, in-kernel crypto APIs, and a *libica* cryptographic functions library that is running on Linux on Z can also start CPACF synchronous functions.

The CPACF coprocessor in z14 ZR1 is redesigned for improved performance compared to the z13s, depending on the function that is being used. The following tools might benefit from the throughput improvements:

- ▶ Db2/IMS encryption tool
- ▶ Db2 built-in encryption
- ▶ z/OS Communication Server: IPsec/IKE/AT-TLS
- ▶ z/OS System SSL
- ▶ z/OS Network Authentication Service (Kerberos)
- ▶ DFDSS Volume encryption
- ▶ z/OS Java SDK
- ▶ z/OS Encryption Facility
- ▶ Linux on Z: Kernel, openssl, openCryptoki, and GSKIT

The z14 ZR1 hardware includes the implementation of algorithms as hardware synchronous operations. This configuration holds the PU processing of the instruction flow until the operation completes.

z14 ZR1 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. New for z14 ZR1 is SHA-3 support for SHA-224, SHA-256, SHA-384, and SHA-512 and the two extendable output functions as described by the standard SHAKE-128 and SHAKE-256.
- ▶ Message authentication code (MAC):
 - Single-length key MAC
 - Double-length key MAC
- ▶ Pseudo-Random Number Generator (PRNG), Deterministic Random Number Generation (DRNG), and True Random Number Generation (TRNG) for cryptographic key generation.
- ▶ Galois Counter Mode (GCM) encryption, which is enabled by a single hardware instruction.

For the SHA hashing algorithms and the random number generation algorithms, only clear keys are used. For the symmetric encryption and decryption DES and AES algorithms and clear keys, protected keys can also be used. Protected keys require a Crypto Express6S or a Crypto Express5S adapter that is running in CCA mode. For more information, see 6.5.2, “Crypto Express6S as a CCA coprocessor” on page 191.

The hashing algorithms SHA-1, SHA-2, and SHA-3 support for SHA-224, SHA-256, SHA-384, and SHA-512, are enabled on all systems and do not require the CPACF enablement feature. For all other algorithms, the no-charge CPACF enablement feature (FC 3863) is required.

The CPACF functions are implemented as processor instructions and require operating system support for use. Operating systems that use the CPACF instructions include z/OS, z/VM, z/VSE, z/TPF, and Linux on Z.

6.4.2 CPACF protected key

z14 ZR1 supports the protected key implementation. Since PCI-XCC³ deployment, secure keys are processed on the PCI-X and PCIe adapters. This process requires an asynchronous operation to move the data and keys from the general-purpose central processor (CP) to the crypto adapters.

Clear keys process faster than secure keys because the process is done synchronously on the CPACF. Protected keys blend the security of Crypto Express6S or Crypto Express5S coprocessors and the performance characteristics of the CPACF. This process allows it to run closer to the speed of clear keys.

CPACF facilitates the continued privacy of cryptographic key material when used for data encryption. In Crypto Express6S or Express5S coprocessors, a secure key is encrypted under a master key. However, a protected key is encrypted under a wrapping key that is unique to each LPAR.

Because the wrapping key is unique to each LPAR, a protected key cannot be shared with another LPAR. By using key wrapping, CPACF ensures that key material is not visible to applications or operating systems during encryption operations.

CPACF code generates the wrapping key and stores it in the protected area of the hardware system area (HSA). The wrapping key is accessible only by firmware. It cannot be accessed by operating systems or applications. DES/T-DES and AES algorithms are implemented in CPACF code with the support of hardware assist functions. Two variations of wrapping keys are generated: One for DES/T-DES keys and another for AES keys.

³ IBM 4764 PCI-X cryptographic coprocessor.

Wrapping keys are generated during the clear reset each time an LPAR is activated or reset. No customizable option is available at Support Element (SE) or Hardware Management Console (HMC) that permits or avoids the wrapping key generation. This function flow for the Crypto Express6S adapter is shown in Figure 6-5.

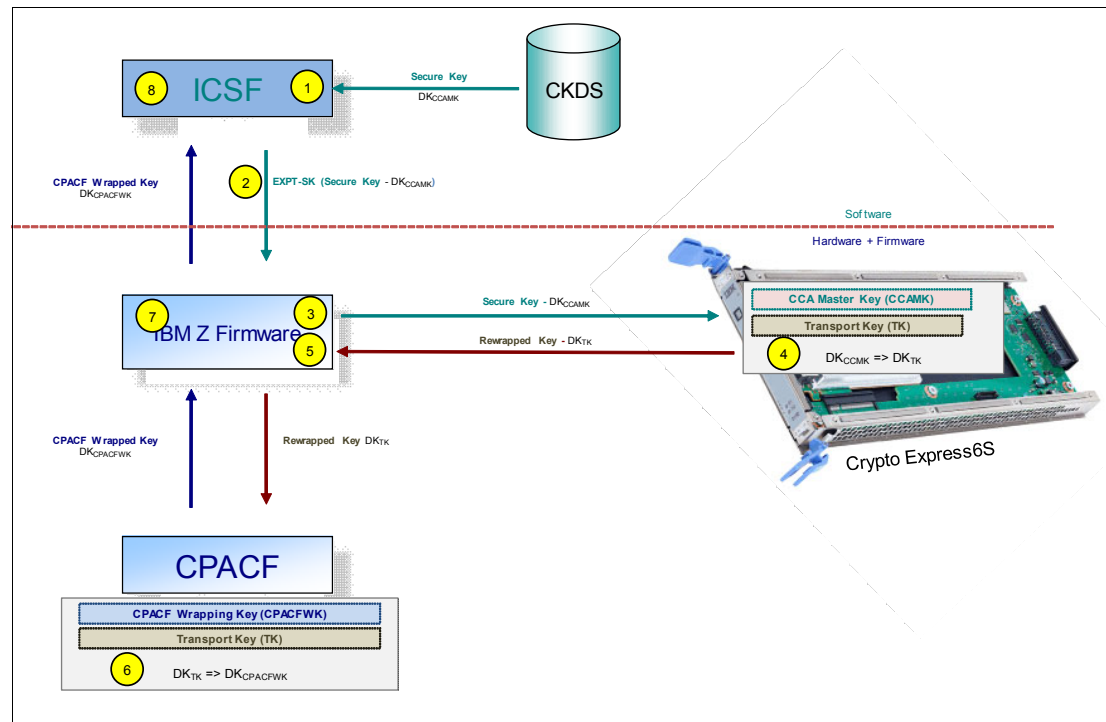


Figure 6-5 CPACF key wrapping for Crypto Express6S

The key wrapping for Crypto Express5S is similar to Crypto Express6S; however, the Data Key that is exchanged between the Crypto Express5S and the CPACF is not wrapped by way of a Transport Key.

The CPACF Wrapping Key and the Transport Key for use with Crypto Express6S, are in a protected area of the HSA that is not visible to operating systems or applications.

The function flow for Crypto Express5S is shown in Figure 6-6.

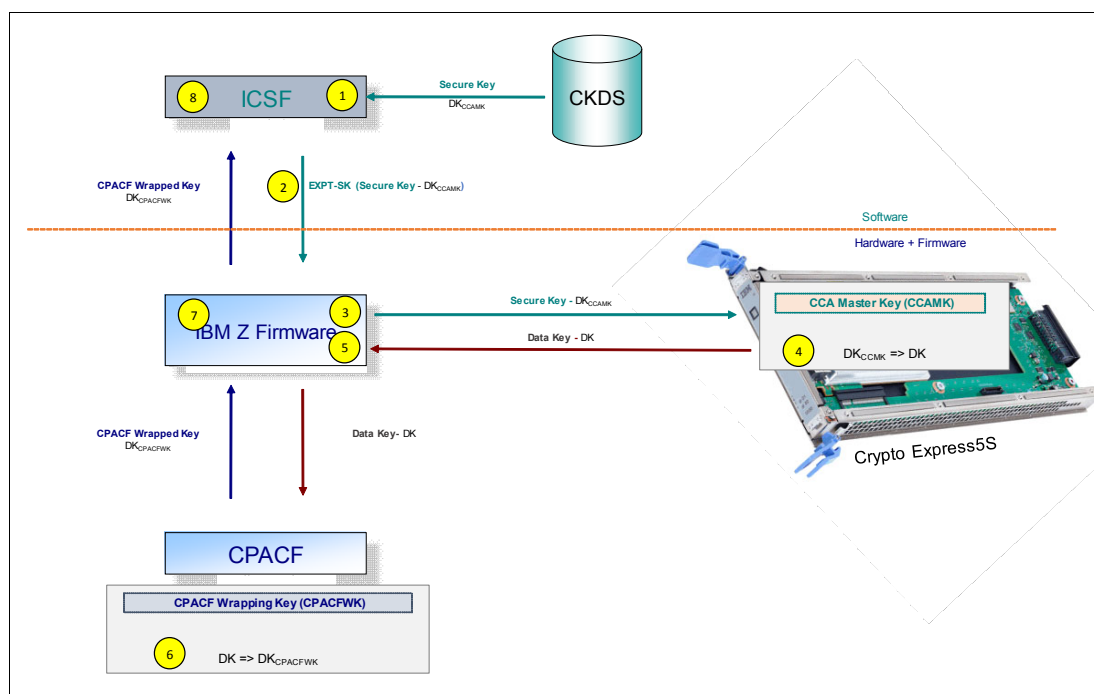


Figure 6-6 CPACF key wrapping for Crypto Express5S

If a Crypto Express6S coprocessor (CEX6C) or Express5S coprocessor (CEX5C) is available, a protected key can begin its life as a secure key. Otherwise, an application is responsible for creating or loading a clear key value, and then uses the PCKMO instruction to wrap the key. ICSF is not called by the application if the CEX6C or CEX5C is not available.

A new segment in the profiles of the CSFKEYS class in IBM RACF restricts which secure keys can be used as protected keys. By default, all secure keys are considered not eligible to be used as protected keys. The process that is shown in Figure 6-5 considers a secure key as the source of a protected key.

The source key in this case is stored in the ICSF Cryptographic Key Data Set (CKDS) as a secure key, which was encrypted under the master key. This secure key is sent to CEX6C or CEX5C to be deciphered and then, sent to the CPACF in clear text. At the CPACF, the key is wrapped under the LPAR wrapping key, and is then returned to ICSF. After the key is wrapped, ICSF can keep the protected value in memory. It then passes it to the CPACF, where the key is unwrapped for each encryption or decryption operation.

The protected key is designed to provide substantial throughput improvements for a large volume of data encryption and low latency for encryption of small blocks of data. A high-performance secure key solution, also known as a protected key solution, requires the ICSF HCR7770 as a minimum release.

6.5 Crypto Express6S

The Crypto Express6S feature (FC 0893) is an optional feature that is exclusive to z14 systems. Each feature has one PCIe cryptographic adapter. The Crypto Express6S (CEX6S) feature occupies one I/O slot in a z14 ZR1 PCIe+ I/O drawer. This feature is an HSM and provides a secure programming and hardware environment on which crypto processes are run.

Each cryptographic coprocessor includes a general-purpose processor, non-volatile storage, and specialized cryptographic electronics. The Crypto Express6S feature provides tamper-sensing and tamper-responding, high-performance cryptographic operations.

Each Crypto Express6S PCI Express adapter is available in one of the following configurations:

- ▶ Secure IBM CCA coprocessor (CEX6C) for FIPS 140-2 Level 4 certification. This configuration includes secure key functions. It is optionally programmable to deploy more functions and algorithms by using UDX. For more information, see 6.5.2, “Crypto Express6S as a CCA coprocessor” on page 191.
- ▶ Secure IBM Enterprise PKCS #11 (EP11) coprocessor (CEX6P) implements an industry-standardized set of services that adheres 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. For more information, see 6.5.3, “Crypto Express6S as an EP11 coprocessor” on page 196.

A TKE workstation is required to support the administration of the Crypto Express5S when it is configured in EP11 mode.

- ▶ Accelerator (CEX6A) for acceleration of public key and private key cryptographic operations that are used with SSL/TLS processing. For more information, see 6.5.4, “Crypto Express6S as an accelerator” on page 197.

These modes can be configured by using the SE. The PCIe adapter must be configured offline to change the mode.

Attention: Switching between configuration modes erases all adapter secrets. The exception is when you are switching from Secure CCA to accelerator, and vice versa.

The Crypto Express6S feature was released for enhanced cryptographic performance. Clients who migrated to variable-length AES key tokens cannot take advantage of faster encryption speeds by using CPACF. Support is being added to translate a secure variable-length AES CIPHER token to a protected key token (protected by the system wrapping key). This support allows for faster AES encryption speeds when variable-length tokens are used while maintaining strong levels of security.

The Crypto Express6S feature does not include external ports and does not use optical fiber or other cables. It does not use channel path identifiers (CHPIDs), but requires one slot in the PCIe I/O drawer and one physical channel ID (PCHID) for each PCIe cryptographic adapter. Removal of the feature or adapter *zeroizes* its content. Access to the PCIe cryptographic adapter is controlled through the setup in the image profiles on the SE.

Adapter: Although PCIe cryptographic adapters include no CHPID type and are not identified as external channels, all logical partitions (LPARs) in all channel subsystems can access to the adapter. In z14 systems, up to 85 LPARs (40 LPAR for z14 ZR1) are available per adapter. Accessing the adapter requires a setup in the image profile for each partition. The adapter must be in the candidate list.

Each z14 ZR1 supports up to 16 Crypto Express6S and Crypto Express 5S features in total. Crypto Express5S features are not orderable for a new build system, but can be carried forward from a z13s by using an MES. Configuration information for Crypto Express6S is listed in Table 6-2.

Table 6-2 Crypto Express6S features

Feature	Quantity
Minimum number of orderable features for z14 ZR1 ^a	2
Order increment above two features	1
Maximum number of features for z14 ZR1 (CEX6S and CEX5S in total)	16
Number of PCIe cryptographic adapters for each feature (coprocessor or accelerator)	1
Number of cryptographic domains at z14 ZR1 for each PCIe adapters ^b	40

- a. The minimum initial order of Crypto Express6S features is two. After the initial order, more Crypto Express6S features can be ordered one feature at a time, up to a maximum of 16.
- b. More than one partition, which is defined to the same channel subsystem (CSS) or to different CSSs, can use the same domain number when assigned to different PCIe cryptographic adapters.

The concept of *dedicated processor* does not apply to the PCIe cryptographic adapter. Whether configured as a coprocessor or an accelerator, the PCIe cryptographic adapter is made available to an LPAR. It is made available as directed by the domain assignment and the candidate list in the LPAR image profile. This availability is not changed by the shared or dedicated status that is given to the PUs in the partition.

When installed non-concurrently, Crypto Express6S features are assigned PCIe cryptographic adapter numbers sequentially during the power-on reset (POR) that follows the installation. When a Crypto Express6S feature is installed concurrently, the installation can select an out-of-sequence number from the unused range. When a Crypto Express6S feature is removed concurrently, the PCIe adapter numbers are automatically freed.

The definition of domain indexes and PCIe cryptographic adapter numbers in the candidate list for each LPAR must be planned to allow for nondisruptive changes. Consider the following points:

- ▶ Operational changes can be made by using the Change LPAR Cryptographic Controls task from the SE, which reflects the cryptographic definitions in the image profile for the partition. With this function, adding and removing the cryptographic feature 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 that is coupled with the usage domain index that is specified must be unique across all active LPARs.

The same PCIe cryptographic adapter number and usage domain index combination can be defined for more than one LPAR (up to 40 for z14 ZR1). For example, you might define a configuration for backup situations. However, only one of the LPARs can be active at a time.

For more information, see 6.5.5, “Managing Crypto Express6S” on page 197.

6.5.1 Cryptographic asynchronous functions

The optional PCIe cryptographic coprocessors Crypto Express6S provides asynchronous cryptographic functions to z14 ZR1. Over 300 Cryptographic algorithms and modes are supported, including the following algorithms and modes:

- ▶ **DES/TDES w DES/TDES MAC/CMAC:** The Data Encryption Standard is a widespread symmetrical encryption algorithm. DES, along with its double-length and triple length variations, TDES today are considered to be not sufficient secure for many applications. They were replaced by the AES as the official US standard, but it is still used in the industry with the MAC and the Cipher-based Message Authentication Code (CMAC) for verifying the integrity of messages.
- ▶ **AES, AESKW, AES GMAC, AES GCM, AES XTS, AES CIPHER mode, and CMAC:** AES replaced DES as the official US standard in October 2000. The enhanced standards for AES Key Warp (AESKW), the AES Galois Message Authentication Code (AES GMAC) and Galois/Counter Mode (AES GCM), the XEX-based tweaked-codebook mode with ciphertext stealing (AES XTS), and CMAC are supported.
- ▶ **MD5, SHA-1, SHA-2, or SHA-3⁴ (224, 256, 384, 512), and HMAC:** The Secure Hash Algorithm (SHA-1 and the enhanced SHA-2 or SHA-3 for different block sizes), the older message-digest (MD5) algorithm, and the advanced keyed-hash method authentication code (HMAC) are used for verifying the data integrity and the authentication of a message.
- ▶ **Visa Format Preserving Encryption (VFPE):** A method of encryption in which the resulting cipher text features the same form as the input clear text, which is developed for use with credit cards.
- ▶ **RSA (512, 1024, 2048, and 4096):** RSA was published in 1977. It is widely used asymmetric public-key algorithm, which means that the encryption key is public whereas the decryption key is kept secret. It is based on the difficulty of factoring the product of two large prime numbers. The number describes the length of the keys.
- ▶ **ECDSA (192, 224, 256, 384, and 521 Prime/NIST):** ECC is a family of asymmetric cryptographic algorithms that are based on the algebraic structure of elliptic curves. ECC can be used for encryption, pseudo-random number generation, and digital certificates. The Elliptic Curve Digital Signature Algorithm (ECDSA) Prime/NIST method is used for ECC digital signatures, which are recommended for government use by NIST.
- ▶ **ECDSA (160, 192, 224, 256, 320, 384, and 512 BrainPool):** ECC Brainpool is a workgroup of companies and institutions that collaborate on developing ECC algorithms. The ECDSA algorithms that are recommended by this group are supported.
- ▶ **ECDH (192, 224, 256, 384, and 521 Prime/NIST):** Elliptic Curve Diffie-Hellman (ECDH) is an asymmetric protocol that is used for key agreement between two parties by using ECC-based private keys. The recommendations by NIST are supported.
- ▶ **ECDH (160, 192, 224, 256, 320, 384, and 512 BrainPool):** ECDH according to the Brainpool recommendations.
- ▶ **Montgomery Modular Math Engine:** The Montgomery Modular Math Engine is a method for fast modular multiplication. Many crypto systems, such as RSA and Diffie-Hellman key Exchange, can use this method.
- ▶ **Random Number Generator (RNG):** The generation of random numbers for cryptographic key generation is supported.

⁴ SHA-3 was standardized by NIST in 2015. SHA-2 is still acceptable and no indication exists that SHA-2 is vulnerable or that SHA-3 is more or less vulnerable than SHA-2.

- ▶ Prime Number Generator (PNG): The generation of prime numbers is also supported.
- ▶ Clear Key Fast Path (Symmetric and Asymmetric): This mode of operation gives a direct hardware path to the cryptographic engine and provides high performance for public-key cryptographic functions.

Several of these algorithms require a secure key and must run on an HSM. Some of these algorithms can also run with a clear key on the CPACF. Many standards are supported only when Crypto Express6S is running in CCA mode. Others are supported only when the adapter is running in EP11 mode.

The three modes for Crypto Express6S are described next. For more information, see 6.7, “Cryptographic functions comparison” on page 205.

6.5.2 Crypto Express6S as a CCA coprocessor

A Crypto Express6S adapter that is running in CCA mode supports IBM CCA. CCA is an architecture and a set of APIs. It provides cryptographic algorithms, secure key management, and many special functions that are required for banking. Over 129 APIs with more than 600 options are provided, with new functions and algorithms always being added.

The IBM CCA provides functions for the following tasks:

- ▶ Encryption of data (DES/TDES/AES)
- ▶ Key management:
 - Using TDES or AES keys
 - Using RSA or Elliptic Curve keys
- ▶ Message authentication for MAC/HMAC/AES-CMAC
- ▶ Key generation
- ▶ Digital signatures
- ▶ Random number generation
- ▶ Hashing (SHA, MD5, and others)
- ▶ ATM PIN generation and processing
- ▶ Credit card transaction processing
- ▶ Visa Data Secure Platform (DSP) Point to Point Encryption (P2PE)
- ▶ Europay, MasterCard, and Visa (EMV) card transaction processing
- ▶ Card personalization
- ▶ Other financial transaction processing
- ▶ Integrated role-based access control system

User-defined extensions support

User-defined extension (UDX) allows a developer to add customized operations to IBM's CCA Support Program. UDXs to the CCA support customized operations that run within the Crypto Express features when defined as a coprocessor.

UDX is supported under a special contract through an IBM or approved third-party service offering. The Crypto Cards website directs your request to an IBM Global Services location for your geographic location. A special contract is negotiated between IBM Global Services and you for the development of the UDX code by IBM Global Services according to your specifications and an agreed-upon level of the UDX.

A UDX toolkit for IBM Z systems is tied to specific versions of the CCA code and the related host code. UDX is available for the Crypto Express6S (Secure IBM CCA coprocessor mode only) features. An UDX migration is no more disruptive than a normal Microcode Change Level (MCL) or ICSF release migration.

In z14 ZR1, up to four UDX files can be imported. These files can be imported from a DVD only. The UDX configuration window is updated to include a Reset to IBM Default button.

Consideration: CCA features a new code level starting with z13 systems, and the UDX clients require a new UDX.

On z14 ZR1, Crypto Express6S is delivered with CCA Level 6.0 firmware. A new set of cryptographic functions and callable services is provided by the IBM CCA LIC to enhance the functions that secure financial transactions and keys. The Crypto Express6S includes the following features:

- ▶ Greater than 16 domains support up to 40 LPARs on z14 ZR1
- ▶ Payment Card Industry (PCI) PIN Transaction Security (PTS) HSM Certification exclusive to CEX6S and z14
- ▶ Visa Format Preserving Encryption (VFPE) support, which was introduced with z13/z13s systems
- ▶ AES PIN support for the German banking industry
- ▶ PKA Translate UDX function into CCA
- ▶ Verb Algorithm Currency

Greater than 16 domains support

z14 ZR1 supports up to 40 LPARs. The IBM Z crypto architecture was designed to support 16 domains, which matched the LPAR maximum at the time. Before z13 systems, crypto workload separation can be complex in customer environments where the number of LPARs was larger than 16. These customers mapped a large set of LPARs to a small set of crypto domains.

Now in z14 systems, the IBM Z crypto architecture can support up to 256 domains in an adjunct processor (AP) with the AP extended addressing (APXA) facility that is installed. As such, the Crypto Express adapters are enhanced to handle 256 domains. The IBM Z firmware provides up to 85 domains for z14 Mxx and 40 domains for z14 ZR1 to customers (to match the current LPAR maximum). Customers can map individual LPARs to unique crypto domains or continue to share crypto domains across LPARs.

The following requirements must be met to support 40 domains:

- ▶ Hardware: z14 ZR1 and Crypto Express6S (or Crypto Express5S)
- ▶ Operating systems:
 - z/OS all functions require ICSF WD17 (HCR77C1), unless otherwise noted. WD17 supports z/OS V2R1, V2R2, and V2R3.
 - z/VM Version 6.4 with PTFs or newer for guest use.

Payment Card Industry-HSM certification

Payment Card Industry (PCI) standards are developed to help ensure security in the PCI. PCI defines their standards as a set of security standards that is designed to ensure that all companies that accept, process, store, or transmit credit card information that is maintained a secure environment.

Compliance with the PCI-HSM standard is valuable for customers, particularly those customers who are in the banking and finance industry. This certification is important to clients for the following fundamental reasons:

- ▶ Compliance is increasingly becoming mandatory.
- ▶ The requirements in PCI-HSM make the system more secure.

Industry requirements for PCI-HSM compliance

The PCI organization cannot require compliance with its standards. Compliance with PCI standards is enforced by the payment card brands, such as Visa, MasterCard, American Express, JCB International, and Discover.

If you are a bank, acquirer, processor, or other participant in the payment card systems, the card brands can impose requirements on you if you want to process their cards. One set of requirements they are increasingly enforcing is the PCI standards.

The card brands work with PCI in developing these standards, and they focused first on the standards they considered most important, particularly the PCI Data Security Standard (PCI-DSS). Some of the other standards were written or required later, and PCI-HSM is one of the last standards to be developed. In addition, the standards themselves were increasing the strength of their requirements over time. Some requirements that were optional in earlier versions of the standards are now mandatory.

In general, the trend is for the card brands to enforce more of the PCI standards and to enforce them more rigorously. The trend in the standards is to impose more and stricter requirements in each successive version. The net result is that companies subject to these requirements can expect that they eventually must comply with all of the requirements.

Improved security through use of PCI-HSM

PCI-HSM was developed primarily to improve security in payment card systems. It imposes requirements in key management, HSM API functions, and device physical security. It also controls during manufacturing and delivery, device administration, and several other areas. It prohibits many things that were in common use for many years, but are no longer considered secure.

The result of these requirements is that applications and procedures often must be updated because they used some of the things that are now prohibited. While this issue is inconvenient and imposes some costs, it does truly increase the resistance of the systems to attacks of various kinds. Updating a system to use PCI-HSM compliant HSMs is expected to reduce the risk of loss for the institution and its clients.

The following requirements must be met to use PCI-HSM:

- ▶ Hardware: z14⁵ systems and Crypto Express6S
- ▶ Operating systems:
 - z/OS - ICSF WD17 (HCR77C1), unless otherwise noted. WD17 supports z/OS V2R1, V2R2, and V2R3
 - z/VM Version 6.4 with PTFs or newer for guest use

⁵ Always check the latest information about security certification status for your specific model.

VISA Format Preserving Encryption

VFPE refers to a method of encryption in which the resulting cipher text features the same form as the input clear text. The form of the text can vary according to use and application. One of the classic examples is a 16-digit credit card number. After VFPE is used to encrypt a credit card number, the resulting cipher text is another 16-digit number. This process helps older databases contain encrypted data of sensitive fields without having to restructure the database or applications.

VFPE allows customers to add encryption to their applications in such a way that the encrypted data can flow through their systems without requiring a massive redesign of their application. In our example, if the credit card number is VFPE-encrypted at the point of entry, the cipher text still behaves as a credit card number. It can flow through business logic until it meets a back-end transaction server that can VFPE-decrypt it to get the original credit card number to process the transaction.

Note: VISA Format Preserving Encryption (VFPE) technology forms part of Visa, Inc.'s Data Secure Platform (DSP). The use of this function requires a service agreement with Visa. You must maintain a valid service agreement with Visa when you use DSP/FPE.

The FPE features the following requirements:

- ▶ Hardware: z14 systems and Crypto Express 6S (or Crypto Express5S with CCA V5.2 firmware).
- ▶ Operating systems:
 - z/OS: All functions require ICSF WD17 (HCR77C1), unless otherwise noted. WD17 supports z/OS V2R1, V2R2, and V2R3.
 - z/OS V2.1 and z/OS V1.13 with the Cryptographic Support for z/OS V1R13-z/OS V2R1 web deliverable (FMID HCR77B0).
 - z/VM Version 6.4 with PTFs or newer for guest use.

AES PIN support for the German banking industry

The German banking industry organization, DK, defined a new set of PIN processing functions to be used on the internal systems of banks and their servers. CCA is designed to support the functions that are essential to those parts of the German banking industry that are governed by DK requirements. The functions include key management support for new AES key types, AES key derivation support, and several DK-specific PIN and administrative functions.

This support includes PIN method APIs, PIN administration APIs, new key management verbs, and new access control points support that is needed for DK-defined functions.

The following requirements must be met to use AES PIN support:

- ▶ Hardware:
 - z14 systems and Crypto Express6S with CCA V6.0 firmware or Crypto Express5S with CCA V5.2 firmware
 - z13s systems and Crypto Express5S with CCA V5.2 firmware
 - z13 systems and Crypto Express5S with CCA V5.0 or later firmware
 - zEC12 or zBC12 and Crypto Express4S with CCA V4.4 firmware
 - zEC12, zBC12, z196 or z114 and Crypto Express3 with CCA V4.4 firmware

- ▶ Operating systems requirements for z14:
 - z/OS: All functions require ICSF WD17 (HCR77C1), unless otherwise noted. WD17 supports z/OS V2R1, V2R2, and V2R3.
 - z/VM Version 6.3, and Version 6.4 with PTFs for guest use.

Support for the updated German Banking standard (DK)

Update support requires ICSF WD18 (HCR77D0) for z/OS V2R2 and V2R3 for:

- ▶ CCA 5.4 & 6.1⁶:
 - ISO-4 PIN Blocks (ISO-9564-1)
 - Directed keys: A key can encrypt or decrypt data, but not both.
 - Allow AES transport keys to be used to export/import *DES* keys in a standard ISO 20038 key block. This feature helps with interoperability between CCA and non-CCA systems.
 - Allow AES transport keys to be used to export/import a small subset of *AES* keys in a standard ISO 20038 key block. This feature helps with interoperability between CCA and non-CCA systems.
 - Triple-length TDES keys with Control Vectors for increased data confidentiality
- ▶ CCA 6.2: PCI HSM 3K DES: Support for triple length DES keys (standards compliance).

PKA Translate UDX function into CCA

UDX is custom code that allows the client to add unique operations or extensions to the CCA firmware. Certain UDX functions are integrated into the base CCA code over time to accomplish the following tasks:

- ▶ Remove headaches and challenges that are associated with UDX management and currency.
- ▶ Make available popular UDX functions to a wider audience to encourage adoption.

UDX is integrated into the base CCA code to support translating an external RSA CRT key into new formats. These new formats use tags to identify key components. Depending on which new rule array keyword is used with the PKA Key Translate callable service, the service TDES encrypts those components in CBC or ECB mode. In addition, AES CMAC support is delivered.

The following requirements must be met to use this function:

- ▶ Hardware:
 - z14 systems and Crypto Express6S (or Crypto Express5S with CCA V5.2 firmware)
 - z13s systems and Crypto Express5S with CCA V5.2 firmware
 - z13 systems and Crypto Express5S with CCA V5.0 or later firmware
- ▶ Operating systems requirements:
 - z/OS: ICSF WD17 (HCR77C1), otherwise noted. WD17 supports z/OS V2R1, V2R2, and V2R3.
 - z/VM Version 6.3, and Version 6.4 with PTFs for guest use.

Note: Although older IBM Z systems and operating systems also are supported, they are beyond the scope of this IBM Redbooks publication.

⁶ CCA 5.4 and 6.1 enhancements are also supported for z/OS V2R1 with ICSF HCR77C1 (WD17) with Small Program Enhancements (SPEs) (z/OS continuous delivery model).

Verb Algorithm Currency

Verb Algorithm Currency is a collection of CCA verb enhancements that are related to customer requirements, with the intent of maintaining currency with cryptographic algorithms and standards. It is also intended for customers who want to maintain the following latest cryptographic capabilities:

- ▶ Secure key support AES GCM encryption
- ▶ Key Check Value (KCV) algorithm for service CSNBKYT2 Key Test 2
- ▶ Key derivation options for CSNDEDH EC Diffie-Hellman service

The following requirements must be met to use this function:

- ▶ Hardware:
 - z14 systems and Crypto Express6S (or Crypto Express5S with CCA V5.2 firmware)
 - z13s or z13 systems and Crypto Express5S with CCA V5.2 firmware
- ▶ Software:
 - z/OS: ICSF WD17 (HCR77C1), otherwise noted. WD17 supports z/OS V2R1, V2R2, and V2R3.
 - z/VM 6.3, and 6.4 with PTFs for guest use.

6.5.3 Crypto Express6S as an EP11 coprocessor

A Crypto Express6S adapter that is configured in Secure IBM Enterprise PKCS #11 (EP11) coprocessor mode provides PKCS #11 secure key support for public sector requirements. Before EP11, the ICSF PKCS #11 implementation supported only clear keys. In EP11, keys can now be generated and securely wrapped under the EP11 Master Key. The secure keys never leave the secure coprocessor boundary decrypted.

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, SHA2 or SHA3 [SHA224, SHA256, SHA384, and SHA512])
- ▶ Digest (SHA1, SHA2 or SHA3 [SHA224, SHA256, SHA384, and SHA512])
- ▶ Wrap and unwrap keys
- ▶ Random number generation
- ▶ Get mechanism list and information
- ▶ Attribute values
- ▶ Key Agreement (Diffie-Hellman)

The function extension capability through UDX is not available to the EP11.

When defined in EP11 mode, the TKE workstation is required to manage the Crypto Express6S feature.

z/OS V2.2 and V2.3 require ICSF Web Deliverable WD18 (HCR77D0) to support the following new features:

- ▶ EP11 Stage 4:
 - New elliptic curve algorithms for PKCS#11 signature, key derivation operations
 - Ed448 elliptic curve
 - EC25519 elliptic curve

- EP11 Concurrent Patch Apply: Allows service to be applied to the EP11 coprocessor dynamically without taking the crypto adapter offline (already available for CCA coprocessors).
- eIDAS compliance: eIDAS: Cross-border EU regulation for portable recognition of electronic identification.

6.5.4 Crypto Express6S as an accelerator

A Crypto Express6S adapter that is running in accelerator mode supports only RSA clear key and SSL Acceleration. A request is processed fully in hardware. The Crypto Express accelerator is a coprocessor that is reconfigured by the installation process so that it uses only a subset of the coprocessor functions at a higher speed. Reconfiguration is disruptive to coprocessor and accelerator operations. The coprocessor or accelerator must be deactivated before you begin the reconfiguration.

FIPS 140-2 certification is not relevant to the accelerator because it operates with clear keys only. The function extension capability through UDX is not available to the accelerator.

The functions that remain available when the Crypto Express6S 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 - 4,096 bits in the Modulus-Exponent (ME) and Chinese Remainder Theorem (CRT) formats.

6.5.5 Managing Crypto Express6S

With zEC12 and older systems, each cryptographic coprocessor has 16 physical sets of registers or queue registers. Starting with z13 systems, this number was raised to 85. This increase corresponds to the maximum number of LPARs that are running on a z14 Mxx, which is also 85 (40 LPARs for z14 ZR1). Each of these sets belongs to the following domains:

- ▶ A cryptographic domain index, in the range of 0 - 39 for z14 ZR1, is allocated to a logical partition in its image profile. The same domain must also be allocated to the ICSF instance that is running in the logical partition that uses the Options data set.
- ▶ Each ICSF instance accesses only the Master Keys or queue registers that correspond to the domain number that is specified in the logical partition image profile at the SE and in its Options data set. Each ICSF instance sees a logical cryptographic coprocessor that consists of the physical cryptographic engine and the unique set of registers (the domain) that is allocated to this logical partition.

The installation of CP Assist for Cryptographic Functions (CPACF) DES/TDES enablement (FC 3863) is required to use the Crypto Express6S feature.

Each Crypto Express6S feature includes one PCI-X adapter. The adapter is available in the following configurations:

- ▶ IBM Enterprise Common Cryptographic Architecture (CCA) Coprocessor (CEX6C)
- ▶ IBM Enterprise Public Key Cryptography Standards #11 (PKCS) Coprocessor (CEX6P)
- ▶ IBM Crypto Express6S Accelerator (CEX6A)

During the feature installation, the PCI-X adapter is configured by default as the CCA coprocessor.

The configuration of the Crypto Express6S adapter as EP11 coprocessor requires a TKE tower workstation (FC 0086) or a TKE rack-mounted workstation (FC 0085) with TKE 9.0 (FC 0879) LIC.

The Crypto Express6S feature does not use CHPIDs from the channel subsystem pool. However, the Crypto Express6S feature requires one slot in a PCIe I/O drawer, and one PCHID for each PCIe cryptographic adapter.

For enabling an LPAR to use a Crypto Express6S adapter, the following cryptographic resources in the image profile must be defined for each partition:

- ▶ Usage domain index
- ▶ Control domain index
- ▶ PCI Cryptographic Coprocessor Candidate List
- ▶ PCI Cryptographic Coprocessor Online List

This task is accomplished by using the Customize/Delete Activation Profile task, which is in the Operational Customization Group, from the HMC or from the SE. Modify the cryptographic initial definition from the Crypto option in the image profile, as shown in Figure 6-7 on page 198.

Important: After this definition is modified, any change to the image profile requires a DEACTIVATE and ACTIVATE of the logical partition for the change to take effect. Therefore, this cryptographic definition is disruptive to a running system.

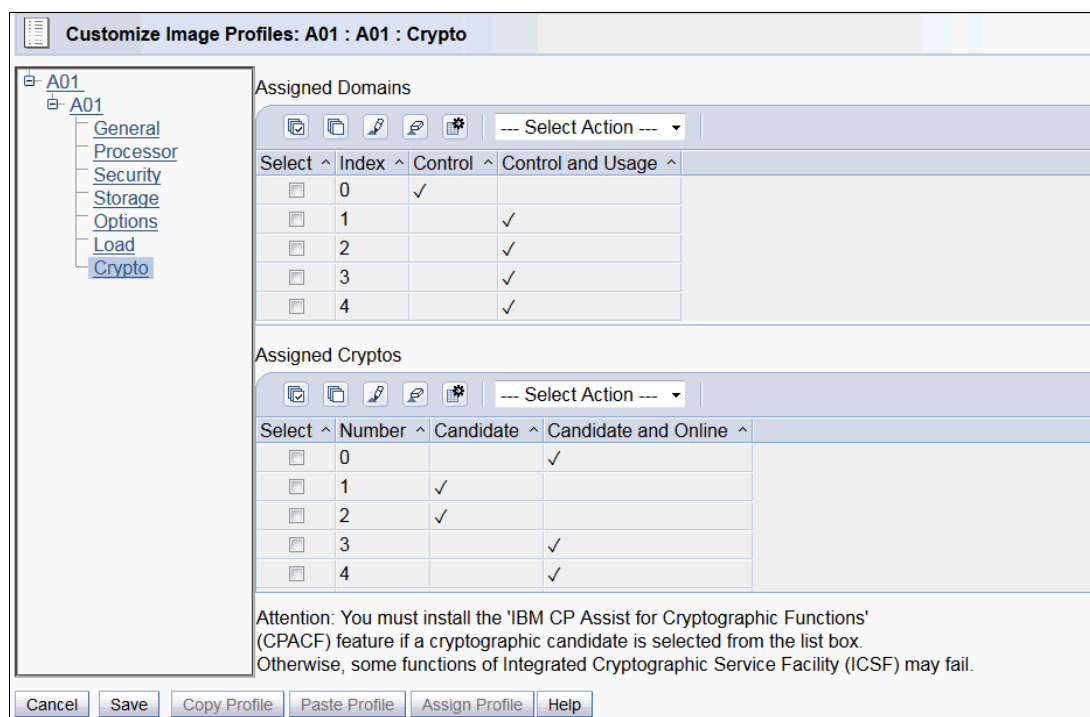


Figure 6-7 Customize Image Profiles: Crypto

The following cryptographic resource definitions are used:

- **Control Domain**

Identifies the cryptographic coprocessor domains that can be administered from this logical partition if it is set up as the TCP/IP host for the TKE.

If you are setting up the host TCP/IP in this logical partition to communicate with the TKE, the partition is used as a path to other domains' Master Keys. Indicate all the control domains that you want to access (including this partition's own control domain) from this partition.

- **Control and Usage Domain**

Identifies the cryptographic coprocessor domains that are assigned to the partition for all cryptographic coprocessors that are configured on the partition. The usage domains cannot be removed if they are online. The numbers that are selected must match the domain numbers that are entered in the Options data set when you start this partition instance of ICSF.

The same usage domain index can be used by multiple partitions, regardless to which CSS they are defined. However, the combination of PCIe adapter number and usage domain index number must be unique across all active partitions.

- **Cryptographic Candidate list**

Identifies the cryptographic coprocessor numbers that can be accessed by this logical partition. From the list, select the coprocessor numbers (in the range 0 - 15) that identify the PCIe adapters to be accessed by this partition.

- **Cryptographic Online list**

Identifies the cryptographic coprocessor numbers that are automatically brought online during logical partition activation. The numbers that are selected in the online list must also be part of the candidate list.

After they are activated, the active partition cryptographic definitions can be viewed from the SE only. Select the CPCs, and click **View LPAR Cryptographic Controls** in the CPC Operational Customization window. The resulting window displays the definition of Usage and Control domain indexes, and PCI Cryptographic candidate and online lists, as shown in Figure 6-8 on page 200. Information is provided only for active logical partitions.

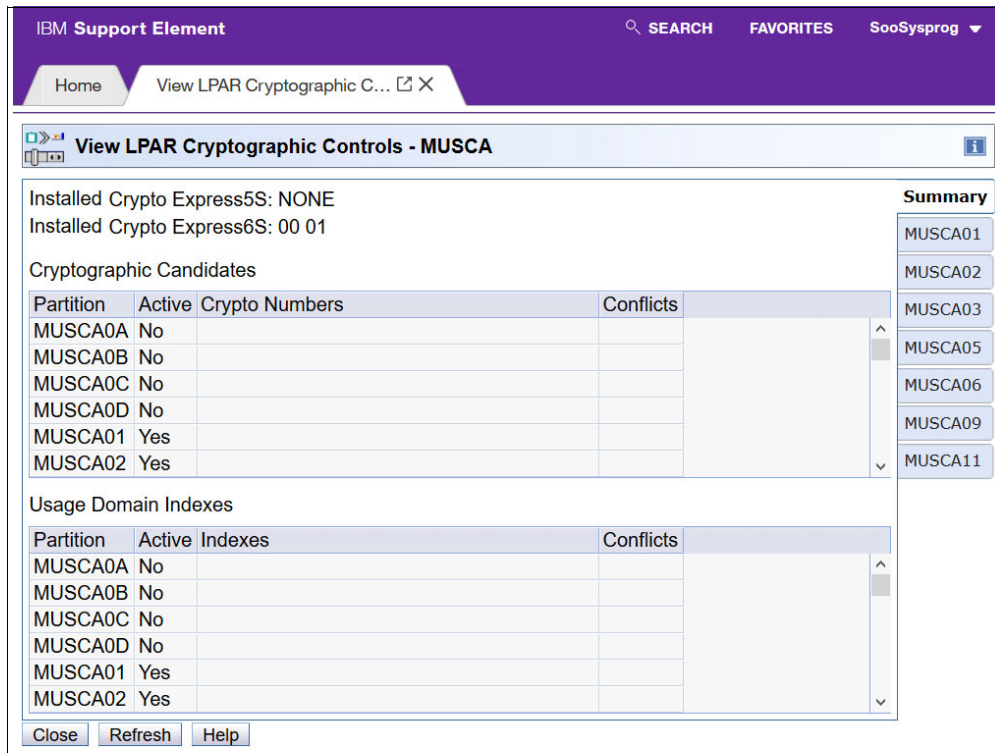


Figure 6-8 SE: View LPAR Cryptographic Controls

Operational changes can be made by using the Change LPAR Cryptographic Controls task from the SE, which reflects the cryptographic definitions in the image profile for the partition. With this function, the cryptographic feature can be added and removed dynamically, without stopping a running operating system.

For more information about the management of Crypto Express6S, see *IBM 14 Model ZR1 Configuration Setup*, SG24-8560.

6.6 TKE workstation

The TKE workstation is an optional feature that offers key management functions. It can be a TKE tower workstation (FC 0085) or TKE rack-mounted workstation (FC 0086) for z14 systems to manage Crypto Express6S or Crypto Express5S.

The TKE contains a combination of hardware and software. A mouse, keyboard, flat panel display, PCIe adapter, and a writable USB media to install the TKE LIC are included with the system unit. The TKE workstation requires an IBM 4768 crypto adapter.

A TKE workstation is part of a customized solution for the use of the Integrated Cryptographic Service Facility for z/OS (ICSF for z/OS) or Linux for z Systems. This program provides a basic key management system for the cryptographic keys of a z14 system that has Crypto Express features installed.

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 run by one PCIe cryptographic coprocessor. The TKE workstation communicates with the IBM Z system through a TCP/IP connection. The TKE workstation is available with Ethernet LAN connectivity only. Up to 10 TKE workstations can be ordered.

TKE FCs 0085 and 0086 can be used to control Crypto Express6S or Crypto Express5S on z14 ZR1. They also can be used to control the Crypto Express5S on z13 and z13s systems, and the Crypto adapters on older still supported systems.

The new TKE 9.1 LIC (FC 0880) features the following enhancements:

- ▶ TKE 9.1 License Internal Code enhancements for support EC521 strength TKE and Migration zones. An EC521 Migration zone is required if you want to use the migration wizard to collect and apply PCI-compliant domain information.
- ▶ TKE 9.1 also has a new family of wizards that makes it easy to create EC521 zones on all of its smart cards. This feature simplifies the process of deploying a TKE for the first time or moving data from a weaker TKE zone to a new EC521 zone.
- ▶ A new smart card for the Trusted Key Entry (TKE) allows stronger Elliptic Curve Cryptography (ECC) levels. Other TKE Smart Cards (FC 0900, packs of 10, FIPS certified blanks) require TKE 9.1 LIC.

The TKE 9.0 LIC (FC 0879) features the following enhancements:

- ▶ Key material copy to alternative zone
By using TKE 9.0, key material can be copied from smart cards in one TKE zone to smart cards in another zone. You might have old 1024-bit strength TKE zones, and might want to move or copy the key material in those zones into a new, stronger TKE zone. To use this new feature, you create TKE or EP11 smart cards on your TKE 9.0 system. You then enroll the new TKE or EP11 smart cards in an alternative zone. This process allows you to copy smart card content from a smart card that is enrolled in the alternative zone.
- ▶ Save TKE data directory structure with files to USB
TKE data can be saved to, or restored from, removable media in the same directory structure they were found on the TKE.
- ▶ Create key parts without opening a host
Administrators can now use the TKE application to create key parts without opening a host. This ability allows the key administrator to create key parts while being offline or before any hosts are defined. This feature can be found in the TKE application under the Utilities → Create CCA key parts pull-down menu.
- ▶ New TKE Audit Log application
A new TKE Audit Log application is available for the Privileged Mode Access ID of AUDITOR. This application provides an easy-to-use interface to view the TKE workstation security audit records from the TKE workstation.
- ▶ Heartbeat audit record
TKE workstations cut an audit record when the TKE starts or when no audit events occurred during a client-configured duration. The record shows the serial number of the TKE local crypto adapter and indicates whether the local crypto adapter was changed since the last check.

- Performance improvements for domain groups

Depending on the size of a domain group, you might experience performance improvements with CCA version 5.3 when a Load, Set, or Clear operation is performed from inside a domain group. For example, if you group all 85 domains on a Host Crypto Express 5 and issue a Clear New Master Key register operation, the number of commands that is issued to the module drops from 85 to 1.

- Secure key entry on EP11

TKE 9.0 EP11 smart card applet now supports secure key entry of EP11 master key parts.

- New certificate manager for domains

Every domain now can manage a set of parent X.509 certificates for validating operating X.509 certificates that are used by applications that are running in the domain.

The following features are related to support for the Crypto Express6S with CCA 6.0. The Crypto Express6S with CCA 6.0 is designed to meet the PCI-HSM PIN Transaction Security v3.0, 2016 standard:

- Domain mode management

With CCA 6.0, individual domains are in one of the following modes:

- Normal Mode
- Imprint Mode
- Compliant Mode

Imprint and compliant mode were added to indirectly and directly meet the PCI-HSM PIN Transaction Security v3.0, 2016 requirement. TKE is required to manage Host Crypto Module domains in imprint and compliant mode.

- Set clock

With TKE 9.0, the host crypto module's clock can be set. The clock must be set before a domain can be placed in imprint mode.

- Domain-specific Host Crypto Module Audit Log management

Domains in imprint mode or compliant mode on a Crypto Express6S maintain a domain-specific module audit log. The TKE provides a feature for downloading the audit records so they can be viewed.

- Domain-specific roles and authorities

Domains in imprint mode or compliant mode on a Crypto Express6S must be managed by using domain-specific roles and authorities. The TKE provides new management features for the domain-specific roles and authorities. The roles are subject to forced dual control policies that prevent roles from issuing and co-signing a command. For information about how to manage imprint and compliant mode domains, see the TKE User's Guide.

- Setup PCI Environment Wizard

To simplify the management of a compliant domain, the TKE provides a setup wizard that creates a minimum set of forced dual control roles and authorities that are needed to manage a compliant domain. For more information about how to manage imprint and compliant mode domains, see the TKE User's Guide.

Tip: For more information about handling a TKE, see the [TKE Introduction video](#) that is available on YouTube.

6.6.1 Logical partition, TKE host, and TKE target

If one or more LPARs are configured to use Crypto Express6S or Crypto Express5 S coprocessors, the TKE workstation can be used to manage DES, AES, ECC, and PKA master keys. This management can be done for all cryptographic domains of each Crypto Express coprocessor feature that is assigned to the LPARs that are defined to the TKE workstation.

Each LPAR in the same system that uses a domain that is managed through a TKE workstation connection is a TKE host or TKE target. An LPAR with a TCP/IP connection to the TKE is referred to as the *TKE host*; all other partitions are *TKE targets*.

The cryptographic controls that are set for an LPAR through the SE determine whether the workstation is a TKE host or a TKE target.

6.6.2 Optional smart card reader

An optional smart card reader (FC 0895) can be added to the TKE workstation. One FC 0895 includes two smart card readers, two cables to connect them to the TKE workstation, and 20 smart cards. The reader supports the use of smart cards that contain an embedded microprocessor and associated memory for data storage. The memory can contain the keys to be loaded into the Crypto Express features. These readers can be used with smart cards only that have applets that were loaded from a TKE 8.1 or later. These cards are FIPS certified.

Smart card readers from feature code 0885 or 0891 can be carried forward. Smart cards can be used on TKE 9.0 with these readers. Access to and use of confidential data on the smart card are protected by a user-defined PIN. Up to 990 other smart cards can be ordered for backup. (The extra smart card feature code is FC 0892.) When one feature code is ordered, 10 smart cards are included. The order increment is 1 - 99 (10 - 990 blank smart cards).

If smart cards with applets that are not supported by the new smart card reader are reused, new smart cards on TKE 8.1 or later must be created and the content from the old smart cards to the new smart cards must be copied. The new smart cards can be created and copied on a TKE 8.1 system. If the copies are done on TKE 9.0, the source smart card must be placed in an older smart card reader from feature code 0885 or 0891.

A new smart card for the Trusted Key Entry (TKE) allows stronger Elliptic Curve Cryptography (ECC) levels. More TKE Smart Cards (FC 0900, packs of 10, FIPS certified blanks) require TKE 9.1 LIC.

6.6.3 TKE hardware support and migration information

The new TKE 9.0 LIC (FC 0879) is originally shipped with a new z14 server. As of December, 2018, a new level of TKE LIC, 9.1 (FC 0880) is available and shipped with any new z14 server. If a new TKE is purchased, the following versions are available:

- ▶ TKE 9.1 tower workstation (FC 0086)
- ▶ TKE 9.1 rack-mounted workstation (FC 0085)

Note: Several options for ordering the TKE with or without ordering Keyboard, Mouse, and Display are available. Ask your IBM Representative for more information about which option is the best option for you.

The TKE 9.x⁷ LIC requires the 4768 crypto adapter. The TKE 8.0 and TKE 8.1 workstations can be upgraded to the TKE 9.x tower workstation by purchasing a 4768 crypto adapter.

The Omnikey Cardman 3821 smart card readers can be carried forward to any TKE 9.x workstation. Smart cards 45D3398, 74Y0551, and 00JA710 can be used on TKE 9.0

When performing a MES upgrade from TKE 7.3, TKE 8.0, or TKE 8.1 to a TKE 9.x installation, the following steps must be completed:

1. Save Upgrade Data on an old TKE to USB memory to save client data.
2. Replace the 4767 crypto adapter with the 4768 crypto adapter.
3. Upgrade the firmware to TKE 9.0.
4. Install the Frame Roll to apply Save Upgrade Data (client data) to the TKE 9.1 system.
5. Run the TKE Workstation Setup wizard.

TKE upgrade considerations

If you are migrating your configuration with Crypto Express5S and TKE Release 8.x to a z14 ZR1, you do not need to upgrade the TKE LIC.

Note: A workstation that was upgraded to TKE V8.x includes the 4767 cryptographic adapter that is required to manage Crypto Express5S; however, it cannot be used to manage the Crypto Express6s.

If your z14 ZR1 includes Crypto Express6S, you must upgrade to TKE V9.0, which requires the 4768 cryptographic adapter.

Upgrading to TKE V9.0 requires that your TKE hardware is compatible with the 4768 cryptographic adapter. The following older TKE hardware features are compatible 4768 cryptographic adapters:

- ▶ FC 0842
- ▶ FC 0847
- ▶ FC 0097

Important: TKE workstations that are at FC 0841 or older do not support the 4767 or 4768 cryptographic adapters.

For more information about TKE hardware support, see Table 6-3. For some functionality, requirements must be considered; for example, the characterization of a Crypto Express adapter in EP 11 mode always requires the use of a TKE.

Table 6-3 TKE Compatibility Matrix

TKE workstation	TKE Release LIC	7.2	7.3	8.0	8.1	9.0	9.1
	HW Feature Code	0814	0842 ^a	0847	0847 or 0097	0085 or 0086	0085 or 0086
	LICC	0850	0872	0877	0878	0879	0880
	Smart Card Reader	0885	0885	0891	0891	0895	0895
	Smart Card	0884	0884	0892	0892	0892	0892

⁷ TKE 9.x represents 9.0 and 9.1 LIC.

Supported systems	z14 (any model)		See ^a	Yes	Yes	Yes	Yes
	z13		See ^a	Yes	Yes	Yes	Yes
	z13s		See ^a	Yes	Yes	Yes	Yes
	zEC12	Yes	Yes	Yes	Yes	Yes	Yes
	zBC12		Yes	Yes	Yes	Yes	Yes
Manage Host Crypto Module	CEC3C (CCA)	Yes	Yes	Yes	Yes	Yes	Yes
	CEX4C (CCA)	Yes	Yes	Yes	Yes	Yes	Yes
	CEX4P (EP11)	Yes	Yes	Yes	Yes	Yes	Yes
	CEX5C (CCA)		See ^a	Yes	Yes	Yes	Yes
	CEX5P (EP11)		See ^a	Yes	Yes	Yes	Yes
	CEX6C (CCA)		See ^a	Yes	Yes	Yes	Yes
	CEX6P (EP11)		See ^a	Yes	Yes	Yes	Yes

a. The TKE workstation FC 0842 that is running LIC V7.3 can be upgraded to TKE LIC V8.x by adding a 4767 cryptographic adapter or to TKE LIC V9.0 by adding a 4786 cryptographic adapter.

Attention: The TKE is unaware of the CPC type where the host crypto module is installed. That is, the TKE does not consider whether a Crypto Express is running on a zEC12, zBC12, z13, 13s, or z14 system. Therefore, the LIC can support any CPC where the coprocessor is supported, but the TKE LIC must support the specific crypto module.

6.7 Cryptographic functions comparison

The functions or attributes on z14 ZR1 for the two cryptographic hardware features are listed in Table 6-4, where “X” indicates that the function or attribute is supported.

Table 6-4 Cryptographic functions on z14 ZR1

Functions or attributes	CPACF	CEX6C	CEX6P	CEX6A
Supports z/OS applications that use CSF	X	X	X	X
Supports Linux on Z CCA applications	X	X	-	X
Encryption and decryption by using secret-key algorithm	-	X	X	-
Provides the highest SSL/TLS handshake performance	-	-	-	X
Supports SSL/TLS functions	X	X	-	X
Provides the highest symmetric (clear key) encryption performance	X	-	-	-
Provides the highest asymmetric (clear key) encryption performance	-	-	-	X
Provides the highest asymmetric (encrypted key) encryption performance	-	X	X	-

Functions or attributes	CPACF	CEX6C	CEX6P	CEX6A
Nondisruptive process to enable	-	X ^a	X ^a	X ^a
Requires IOCDs definition	-	-	-	-
Uses CHPID numbers	-	-	-	-
Uses PCHIDs (one PCHID)	-	X	X	X
Requires CPACF enablement (FC 3863)	X ^b	X ^b	X ^b	X ^b
Requires ICSF to be active	-	X	X	X
Offers UDX	-	X	-	-
Usable for data privacy: Encryption and decryption processing	X	X	X	-
Usable for data integrity: Hashing and message authentication	X	X	X	-
Usable for financial processes and key management operations	-	X	X	-
Crypto performance IBM RMF™ monitoring	-	X	X	X
Requires system master keys to be loaded	-	X	X	-
System (master) key storage	-	X	X	-
Retained key storage	-	X	-	-
Tamper-resistant hardware packaging	-	X	X	X ^c
Designed for FIPS 140-2 Level 4 certification	-	X	X	X
Supports Linux applications that perform SSL handshakes	-	-	-	X
RSA functions	-	X	X	X
High-performance SHA-1, SHA-2, and SHA-3	X	X	X	-
Clear key DES or triple DES	X	-	-	-
Advanced Encryption Standard (AES) for 128-bit, 192-bit, and 256-bit keys	X	X	X	-
True random number generator (TRNG)	X	X	X	-
Deterministic random number generator (DRNG)	X	X	X	-
Pseudo random number generator (PRNG)	X	X	X	-
Clear key RSA	-	-	-	X
Payment Card Industry (PCI) PIN Transaction (PTS) Hardware Security Module (HSM) PCI-HSM		X		
Europay, MasterCard, and Visa (EMV) support	-	X	-	-
Public Key Decrypt (PKD) support for Zero-Pad option for clear RSA private keys	-	X	-	-
Public Key Encrypt (PKE) support for Mod_Raised_to Power (MRP) function	-	X	X	-

Functions or attributes	CPACF	CEX6C	CEX6P	CEX6A
Remote loading of initial keys in ATM	-	X	-	-
Improved key exchange with non-CCA systems	-	X	-	-
ISO 16609 CBC mode triple DES message authentication code (MAC) support	-	X	-	-
AES GMAC, AES GCM, AES XTS mode, CMAC	-	X	-	-
SHA-2, SHA-3 (384,512), HMAC	-	X	-	-
Visa Format Preserving Encryption	-	X	-	-
AES PIN support for the German banking industry				
ECDSA (192, 224, 256, 384, 521 Prime/NIST)	-	X	-	-
ECDSA (160, 192, 224, 256, 320, 384, 512 BrainPool)	-	X	-	-
ECDH (192, 224, 256, 384, 521 Prime/NIST)	-	X	-	-
ECDH (160, 192, 224, 256, 320, 384, 512 BrainPool)	-	X	-	-
PNG (Prime Number Generator)	-	X	-	-

- To make adding the Crypto Express features nondisruptive, the logical partition must be predefined with the appropriate PCI Express cryptographic adapter number. This number must be selected from its candidate list in the partition image profile.
- 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.
- This feature is physically present, but is not used when configured as an accelerator (clear key only).

6.8 Cryptographic operating system support for z14 ZR1

The following section gives an overview of the operating systems requirements in relation to cryptographic elements.

Crypto Express6S (0893) Toleration

Crypto Express6S (0893) Toleration treats Crypto Express6S cryptographic coprocessors and accelerators as Crypto Express5 coprocessors and accelerators. The following minimum prerequisites must be met:

- ▶ z/OS V2.3 with PTFs
- ▶ z/OS V2.2 with PTFs
- ▶ z/OS V2.1 with PTFs
- ▶ z/VM V7.1 for guest use
- ▶ z/VM V6.4 with PTFs for guest use
- ▶ z/VSE V6.2 with PTFs
- ▶ z/VSE V6.1 with PTFs
- ▶ z/VSE V5.2 with PTFs
- ▶ zTPF V1.1 with PTFs
- ▶ Linux on Z: IBM is working with its Linux distribution partners to provide support by way of maintenance or future releases for the following distributions:
 - SUSE Linux Enterprise Server 12 and SLES 11

- Redhat Enterprise Linux (RHEL) 7 and Redhat Enterprise Linux 6
- Ubuntu 16.04 LTS (or higher)
- ▶ The KVM hypervisor, which is offered with SLES 12 SP2 with service, RHEL 7.5 with kernel-alt package (kernel 4.14) and Ubuntu 16.04 LTS with service and Ubuntu 18.04 LTS with service Linux distributions. For more information about minimal and recommended distribution levels, see the [Tested platforms for Linux webpage](#) of the IBM IT infrastructure website.

Crypto Express6S (0893) support of VISA Format Preserving Encryption

The following minimum prerequisites must be met to use this element:

- ▶ z/OS V2.3 with PTFs
- ▶ z/OS V2.2 with the Enhanced Cryptographic Support for z/OS V1.13-V2.2 web deliverable installed
- ▶ z/VM V7.1 for guest use
- ▶ z/VM V6.4 with PTFs for guest use
- ▶ The KVM hypervisor, which is offered with the following Linux distributions:
 - SLES-12 SP2 or higher
 - Ubuntu 16.04 LTS or higher
 - RHEL 7.5 or higher

For more information about minimal and recommended distribution levels, see the [Tested platforms for Linux page](#) of the IBM IT infrastructure website.

Crypto Express6S (0893) support of greater than 16 domains

The following prerequisites must be met to support more than 16 domains:

- ▶ z/OS V2.3 with PTFs
- ▶ z/OS V2.2 with the Enhanced Cryptographic Support for z/OS V1.13-V2.2 Web deliverable installed
- ▶ z/VM V7.1 for guest use
- ▶ z/VM V6.4 with PTFs for guest use
- ▶ z/VSE V6.2 with PTFs
- ▶ z/VSE V6.1 with PTFs
- ▶ z/VSE V5.2 with PTFs
- ▶ Linux on Z: IBM is working with its Linux distribution partners to provide support by way of maintenance or future releases for the following distributions:
 - SLES 12 and SLES 11
 - RHEL 7 and RHEL 6
 - Ubuntu 16.04 LTS (or higher)

For more information about the software support levels for cryptographic functions, see Chapter 7, “Operating system support” on page 209.



Operating system support

This chapter describes the minimum operating system requirements and support considerations for the IBM z14 Model ZR1 servers and their features. It addresses z/OS, z/VM, z/VSE, z/TPF, and Linux on IBM Z (Linux on Z).

Because this information is subject to change, see the hardware fix categories (IBM.Device.Server.z14-3906.*) for the most current information.

Support of z14 ZR1 functions depends on the operating system, its version, and release.

This chapter includes the following topics:

- ▶ 7.1, “Operating systems summary” on page 210
- ▶ 7.2, “Support by operating system” on page 211
- ▶ 7.3, “z14 ZR1 features and function support overview” on page 214
- ▶ 7.4, “Support by features and functions” on page 227
- ▶ 7.5, “z/OS migration considerations” on page 271
- ▶ 7.6, “z/VM migration considerations” on page 275
- ▶ 7.7, “z/VSE migration considerations” on page 276
- ▶ 7.8, “Software licensing” on page 277
- ▶ 7.9, “References” on page 280

7.1 Operating systems summary

The minimum operating system levels that are required on z14 ZR1 servers are listed in Table 7-1.

End of service operating systems: Operating system levels that are no longer in service are not covered in this publication. These older levels might support some features.

Table 7-1 z14 ZR1 minimum operating systems requirements

Operating systems ^a	Supported version and release on z14 ZR1 ^b
z/OS	V1R13 ^c
z/VM	V6R4
z/VSE	V5R2 ^d
z/TPF	V1R1
Linux on Z	See Table 7-2 on page 213
KVM Hypervisor ^e	Offered with the following Linux distributions: SLES-12 SP2 or higher, Ubuntu 16.04 LTS or higher, and Redhat Enterprise Linux 7.5 or higher.

- a. Only z/Architecture mode is supported. For more information, see the shaded box titled “z/Architecture mode” that follows this table.
- b. Service is required. For more information, see the shaded box that is titled “Features” on page 229.
- c. z/OS V1R13 and V2R1 - Compatibility only. The IBM Software Support Services for z/OS V1.13, offered as of October 1, 2016, and V2R1 offered as October 1st, 2018, provide the ability for customers to purchase extended defect support service for z/OS V1.13, V2.1, respectively.
- d. End of service date for z/VSE V5R2 is October 31, 2018.
- e. For more information about distribution levels, see [the Linux on Z website](#).

z/Architecture mode: As announced on January 14, 2015 with Announcement letter 115-001, beginning with IBM z14™, all IBM Z servers support operating systems that are running in z/Architecture mode only. This support applies to operating systems that are running native on PR/SM and operating systems that are running as second-level guests.

IBM operating systems that run in ESA/390 mode are no longer in service or currently available only with extended service contracts, and they are not usable on systems beginning with IBM z14. However, IBM z14 Model ZR1 does provide ESA/390-compatibility mode, which is an environment that supports a subset of DAT-off ESA/390 applications in a hybrid architectural mode.

Programs (24-bit and 31-bit) are unaffected by this change.

The use of certain features depends on the operating system. In all cases, program temporary fixes (PTFs) might be required with the operating system level that is indicated.

Check the z/OS Fix categories (FIXCAT IBM.Device.Server.z14ZR1-3907), or the subsets of the 3907DEVICE PSP buckets for z/VM and z/VSE. The FIXCATs and 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 more information about Linux on Z distributions and KVM hypervisor, see the distributor's support information.

7.2 Support by operating system

z14 ZR1 servers introduce several new functions. This section describes the support of those functions by the current operating systems. Also included are some of the functions that were introduced in previous generations of the IBM Z platform and carried forward or enhanced in z14 ZR1 servers. Features and functions that are available on previous servers but no longer supported by z14 ZR1 servers were removed.

For more information about supported functions that are based on operating systems, see 7.3, “z14 ZR1 features and function support overview” on page 214. Tables are built by function and feature classification to help you determine, by a quick scan, what is supported and the minimum operating system level that is required.

7.2.1 z/OS

z/OS Version 2 Release 2 is the earliest in-service release that supports z14 ZR1 servers. Consider the following points:

- ▶ Service support for z/OS Version 1 Release 13 ended in September of 2016; however, a fee-based extension for defect support (for up to three years) can be obtained by ordering IBM Software Support Services - Service Extension¹ for z/OS 1.13.
- ▶ Service support for z/OS Version 2 Release 1 ended in September of 2018; however, a fee-based extension for defect support (for up to three years) can be obtained by ordering IBM Software Support Services - Service Extension for z/OS 2.1.

z14 ZR1 capabilities differ depending on the z/OS release. Toleration support is provided on z/OS V1R13 and V2R1. Usage support is provided only on z/OS V2R2 and later.

For more information about supported functions and their minimum required support levels, see 7.3, “z14 ZR1 features and function support overview” on page 214.

7.2.2 z/VM

z/VM V6R4 and z/VM V7R1 provide support that enables guests to use the following features that are supported by z/VM on IBM z14 Model ZR1:

- ▶ z/Architecture support
- ▶ New hardware facilities
- ▶ ESA/390-compatibility mode for guests
- ▶ Crypto Clear Key ECC operations
- ▶ RoCE Express2 support

¹ Beginning with z/OS V1.12, IBM Software Support Services replaced the IBM Lifecycle Extension for z/OS offering with a service extension for extended defect support.

- Dynamic I/O support

Provided for managing the configuration of OSA-Express7S and OSA-Express6S OSD CHPIDs, FICON Express16S+ FC and FCP CHPIDs, zHyperLink Express, RoCE Express2 features, and Regional Crypto Enablement (RCE)

- Improved memory management

For more information about supported functions and their minimum required support levels, see 7.3, “z14 ZR1 features and function support overview” on page 214.

Statements of Directions^a: Consider the following points:

- Future z/VM release guest support: z/VM V6.4 is the last release that is supported as a guest of z/VM V6.2 or older releases.
- Disk-only support for z/VM dumps: z/VM V6.4 is the last z/VM release to support tape as a media option for stand-alone, hard abend, and snap dumps. Subsequent releases will support dumps to ECKD DASD or FCP SCSI disks only.

a. All statements regarding IBM plans, directions, and intent are subject to change or withdrawal without notice. Any reliance on these statements of general direction is at the relying party's sole risk and will not create liability or obligation for IBM.

7.2.3 z/VSE

z14 ZR1 support is provided by z/VSE V5R2² and later with PTFs, with the following considerations:

- z/VSE runs in z/Architecture mode only.
- z/VSE supports 64-bit real and virtual addressing.

For more information about supported functions and their minimum required support levels, see 7.3, “z14 ZR1 features and function support overview” on page 214.

7.2.4 z/TPF

z14 ZR1 support is provided by z/TPF V1R1 with PTFs. For more information about supported functions and their minimum required support levels, see 7.3, “z14 ZR1 features and function support overview” on page 214.

7.2.5 Linux on Z

Generally, a new machine is not apparent to Linux on Z. For z14 ZR1, toleration support is required for the following functions and features:

- IPL in “z/Architecture” mode
- Crypto Express6S cards
- RoCE Express cards
- 8-byte LPAR offset

² z/VSE End of Support date is October 31, 2018.

The service levels of SUSE, Red Hat, and Ubuntu releases that are supported at the time of this writing are listed in Table 7-2.

Table 7-2 Linux on Z distributions

Linux on Z distribution ^a	Supported version and release on z14 ZR1 ^b
SUSE Linux Enterprise Server	12 SP2 or higher
SUSE Linux Enterprise Server	11 SP4
Red Hat RHEL	7.3 or higher
Red Hat RHEL	6.8
Ubuntu	16.04 LTS (or higher)

a. Only z/Architecture (64-bit mode) is supported. IBM testing identifies the minimum required level and the recommended levels of the tested distributions.

b. Fix installation is required for toleration.

For more information about supported Linux distributions on IBM Z platform, see the [Tested platforms for Linux](#) page of the IBM IT infrastructure website.

IBM is working with Linux distribution Business Partners to provide further use of selected z14 ZR1 functions in future Linux on Z distribution releases.

Consider the following guidelines:

- ▶ Use SUSE Linux Enterprise Server 12, Red Hat RHEL 7, or Ubuntu 16.04 LTS in any new projects for z14 ZR1 servers.
- ▶ Update any Linux distribution to the latest service level before migrating to z14 ZR1 servers.
- ▶ Adjust the capacity of any z/VM and Linux on 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 z14 ZR1 servers.

7.2.6 KVM hypervisor

KVM is now offered through our Linux distribution partners for IBM Z and LinuxONE to help simplify delivery and installation. Linux and KVM is provided from a single source, and with KVM being included in the Linux distribution, it should make ordering and installing KVM easier.

The KVM hypervisor is supported with the following minimum Linux distributions:

- ▶ SLES 12 SP2 with service.
- ▶ RHEL 7.5 with kernel-alt package (kernel 4.14).
- ▶ Ubuntu 16.04 LTS with service and Ubuntu 18.04 LTS with service.

For more information about minimal and recommended distribution levels, see [the IBM Z website](#).

7.3 z14 ZR1 features and function support overview

The following tables summarize the z14 ZR1 features and functions and their minimum required operating system support levels:

- ▶ Table 7-3, “Supported Base CPC Functions or z/OS and z/VM” on page 214
- ▶ Table 7-4, “Supported base CPC functions for z/VSE, z/TPF, and Linux on Z” on page 216
- ▶ Table 7-5, “Supported coupling and clustering functions for z/OS and z/VM” on page 217
- ▶ Table 7-6, “Supported storage connectivity functions for z/OS and z/VM” on page 218
- ▶ Table 7-7, “Supported storage connectivity functions for z/VSE, z/TPF, and Linux on Z” on page 219
- ▶ Table 7-8, “Supported network connectivity functions for z/OS and z/VM” on page 221
- ▶ Table 7-9, “Supported network connectivity functions for z/VSE, z/TPF, and Linux on Z” on page 223
- ▶ Table 7-10, “Supported cryptography functions for z/OS and z/VM” on page 225
- ▶ Table 7-11, “Supported cryptography functions for z/VSE, z/TPF, and Linux on Z” on page 226
- ▶ Table 7-12, “Supported Special-purpose feature for z/OS and z/VM” on page 227

Information about Linux on Z refers exclusively to the appropriate distributions of SUSE, Red Hat, and Ubuntu.

All 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.

Note: The following tables list but do not explicitly mark all the features that require fixes that are required by the corresponding operating system for toleration or exploitation. For more information, see the PSP bucket for 3907DEVICE.

7.3.1 Supported CPC functions

The supported Base CPC Functions or z/OS and z/VM are listed in Table 7-3.

Table 7-3 Supported Base CPC Functions or z/OS and z/VM

Function ^a	z/OS V2R3	z/OS V2R2	z/OS V2R1	z/OS V1R13	z/VM V7R1	z/VM V6R4
z14 ZR1 servers	Y	Y	Y	Y	Y	Y
Maximum processor unit (PUs) per system image	170 ^{b,c}	170 ^{b,c}	170 ^{b,c}	100 ^{b,c}	64 ^d	64 ^d
Maximum main storage size ^e	4 TB	4 TB	4 TB	1 TB	2 TB	2 TB
40 LPARs	Y	Y	Y	Y	Y	Y
Separate LPAR management of PUs	Y	Y	Y	Y	Y	Y
Dynamic PU add	Y	Y	Y	Y	Y	Y
Dynamic LPAR memory upgrade	Y	Y	Y	Y	Y	Y

Function ^a	z/OS V2R3	z/OS V2R2	z/OS V2R1	z/OS V1R13	z/VM V7R1	z/VM V6R4
LPAR group absolute capping	Y	Y	Y	Y	Y	Y
Capacity Provisioning Manager	Y	Y	Y	Y	N	N
Program-directed re-IPL	-	-	-	-	-	Y
HiperDispatch	Y	Y	Y	Y	Y	Y
IBM Z Integrated Information Processors (zIIPs)	Y	Y	Y	Y	Y ^g	Y ^g
Transactional Execution	Y	Y	Y	Y	Y ^{fg}	Y ^{fg}
Java Exploitation of Transactional Execution	Y	Y	Y	Y	Y ^g	Y ^g
Simultaneous multithreading (SMT)	Y ^h	Y ^h	Y ^h	N	Y ⁱ	Y ⁱ
Single Instruction Multiple Data (SIMD)	Y	Y	Y	N	Y ^{gj}	Y ^{gj}
Hardware decimal floating point ^k	Y	Y	Y	Y	Y	Y
2 GB large page support	Y	Y	Y	Y	N	N
Large page (1 MB) support	Y	Y	Y	Y	Y ^g	Y ^g
Out-of-order execution	Y	Y	Y	Y	Y	Y
CPUMF (CPU measurement facility) for z14 ZR1	Y	Y	Y	N	Y	Y
Enhanced flexibility for Capacity on Demand (CoD)	Y	Y	Y	Y	Y	Y
IBM Virtual Flash Memory (VFM)	Y	Y	Y	Y ^l	N	N
1 MB pageable large pages ^m	Y	Y	Y	Y	N	N
Guarded Storage Facility (GSF)	Y	Y	N	N	Y ^g	Y ^g
Instruction Execution Protection (IEP)	Y	Y	N	N	Y ^g	Y ^g
Co-processor Compression Enhancements	Y	Y	Y	N	Y ^g	Y ^g

a. PTFs might be required for toleration support or exploitation of z14 ZR1 features and functions.

b. 170-way without multithreading; 128-way with multithreading enabled.

c. z14 ZR1 supports maximum six CPs for z/OS and up to 12 zIIPs.

d. 64-way without multithreading; 32-way with multithreading enabled; however, z14 ZR1 supports up to 30 IFLs for running z/VM.

e. A total of 8 TB of real storage is supported per server.

f. Guests are informed that TX facility is available for use.

g. Guest use support.

h. For zIIPs only.

i. Dynamic SMT with z14 ZR1.

j. Guests are informed that SIMD is available for use.

k. Packed decimal conversion support.

l. A web deliverable is required, which is available from [the z/OS downloads website](#).

m. With IBM Virtual Flash Memory for middleware use.

The supported base CPC functions for z/VSE, z/TPF, and Linux on Z are listed in Table 7-4.

Table 7-4 Supported base CPC functions for z/VSE, z/TPF, and Linux on Z

Function ^a	z/VSE V6R2	z/VSE V6R1	z/VSE V5R2	z/TPF V1R1	Linux on Z ^b
z14 ZR1 servers	Y	Y	Y	Y	Y
Maximum processor unit (PUs) per system image	10	10	10	86 ^c	170 ^{d,c}
Maximum main storage size ^e	32 GB	32 GB	32 GB	4 TB	16 TB ^f
40 LPARs	Y	Y	Y	Y	Y
Separate LPAR management of PUs	Y	Y	Y	Y	Y
Dynamic PU add	Y	Y	Y	N	Y
Dynamic LPAR memory upgrade	N	N	N	N	Y
LPAR group absolute capping	Y	Y	Y	N	N
Capacity Provisioning Manager	-	-	-	-	-
Program-directed re-IPL	Y ^g	Y ^g	Y ^g	N	Y
HiperDispatch	N	N	N	N ^h	Y
IBM Z Integrated Information Processors (zIIPs)	N	N	N	N	N
Transactional Execution	N	N	N	N	Y
Java Exploitation of Transactional Execution	N	N	N	N	Y
Simultaneous multithreading (SMT)	N	N	N	N	Y
Single Instruction Multiple Data (SIMD)	Y	N	N	N	Y
Hardware decimal floating point ⁱ	N	N	N	N	Y
2 GB large page support	N	N	N	Y	Y
Large page (1 MB) support	Y ^j	Y ^j	Y ^j	Y	Y
Out-of-order execution	Y	Y	Y	Y	Y
CPUMF (CPU measurement facility) for z14 ZR1	N	N	N	Y	N ^k
Enhanced flexibility for CoD	N	N	N	N ^h	N
IBM Virtual Flash Memory (VFM)	N	N	N	N	Y
1 MB pageable large pages ^l	N	N	N	N	N
Guarded Storage Facility (GSF)	N	N	N	N	Y
Instruction Execution Protection (IEP)	N	N	N	N	Y
Co-processor Compression Enhancements	N	N	N	N	N

a. PTFs might be required for toleration support or use of z14 ZR1 features and functions.

b. Support statement varies based on Linux on Z distribution and release.

c. z14 ZR1 supports up to 6 CPs and up to 30 IFLs.

d. For SLES12/RHEL7/Ubuntu 16.10, Linux supports 256 cores without SMT and 128 cores with SMT (=256 threads).

e. A total of 8 TB of real storage is supported per server.

f. Linux on Z releases can support up to 64 TB of memory.

- g. On SCSI disks.
- h. Availability expected in fourth quarter of 2017.
- i. Packed decimal conversion support.
- j. Supported for data spaces.
- k. IBM is working with its Linux distribution Business Partners to provide this feature.
- l. With IBM Virtual Flash Memory0 for middleware exploitation.

7.3.2 Coupling and clustering

The supported coupling and clustering functions for z/OS and z/VM are listed in Table 7-5.

Table 7-5 Supported coupling and clustering functions for z/OS and z/VM

Function ^a	z/OS V2R3	z/OS V2R2	z/OS V2R1	z/OS V1R13	z/VM V7R1	z/VM V6R4
Server Time Protocol (STP)	Y	Y	Y	Y	Y	Y
CFCC Level 23 ^b	Y	Y	Y	Y	Y ^d	Y ^d
CFCC Level 22 ^c	Y	Y	Y	Y	Y ^d	Y ^d
CFCC Level 22 Coupling Thin Interrupts	Y	Y	Y	Y	N	N
CFCC Level 22 Large Memory support	Y	Y	Y	Y	N	N
CFCC Level 22 Support for 256 Coupling CHPIDs per CPC ^e	Y	Y	Y	Y	Y ^d	Y ^d
CFCC Level 22 Coupling Facility Processor Scalability	Y	Y	Y	N	Y ^d	Y ^d
CFCC Level 22 List Notification Enhancements	Y	Y	N	N	Y ^d	Y ^d
CFCC Level 22 Encryption Support	Y	N ^f	N ^f	N	Y ^d	Y ^d
CFCC Level 22 Exploitation of VFM (Virtual Flash Memory)	Y	Y	Y	Y	N	N
RMF coupling channel reporting	Y	Y	Y	Y	N	N
Coupling over InfiniBand CHPID type CIB	-	-	-	-	N	N
InfiniBand coupling links 12x at a distance of 150 m (492 ft.)	-	-	-	-	N	N
InfiniBand coupling links 1x at an unrepeated distance of 10 km (6.2 miles)	-	-	-	-	N	N
Integrated Coupling Adapter (ICA SR) links CHPID CS5	Y	Y	Y	Y	Y ^g	Y ^g
Coupling Express LR (CE LR) CHPID CL5	Y	Y	Y	N	Y ^g	Y ^g
z/VM Dynamic I/O support for InfiniBand CHPIDs	-	-	-	-	N	N
z/VM Dynamic I/O support for ICA CHPIDs	-	-	-	-	Y ^g	Y ^g
Asynchronous CF Duplexing for lock structures	Y	Y	N	N	Y ^d	Y ^d
Asynchronous cross-invalidate (XI) for CF cache structures ^h	Y ⁱ	Y ⁱ	Y ^j	Y ^j	Y ^d	Y ^d
Dynamic I/O activation for standalone CF CPCs	Y ^k	Y ^k	Y ^k	Y ^k	Y ^k	Y ^k

a. PTFs might be required for toleration support or use of z14 ZR1 features and functions.

b. CFCC Level 23 with Driver 36.

c. CFCC Level 22 with Driver 32.

d. Virtual guest coupling.

e. z14 ZR1 supports up to 176 coupling CHPIDs per CPC.

- f. Toleration support ("locking out" down level systems that cannot use encrypted structure) will be provided for z/OS 2.2 and z/OS 2.1.
- g. To define, modify, and delete CHPID type CS5/CL5 when z/VM is the controlling LPAR for dynamic I/O. APAR VM65865 is required for defining CL5 CHPIDs.
- h. Requires data manager support (Db2 fixes).
- i. Requires fixes for APAR OA54688 for exploitation.
- j. Toleration support only; requires fixes for APAR OA54985. Functional support in z/OS 2.2 and later
- k. Requires HMC 2.14.1, Driver level 36 and various OS fixes (HCD, HCM, IOS, IOCP)

In addition to operating system support that is listed in Table 7-5 on page 217, Server Time Protocol is supported on z/TPF V1R1 and Linux on Z. CFCC Level 22 and Level 23 are supported for z/TPF V1R1.

Storage connectivity

The supported storage connectivity functions for z/OS and z/VM are listed Table 7-6.

Table 7-6 Supported storage connectivity functions for z/OS and z/VM

Function^a	z/OS V2R3	z/OS V2R2	z/OS V2R1	z/OS V1R13	z/VM V7R1	z/VM V6R4
zHyperLink Express	Y	Y	Y	N	N	N
The 63.75K subchannels	Y	Y	Y	Y	Y	Y
Three logical channel subsystems (LCSSs)	Y	Y	Y	Y	Y	Y
Three subchannel set per LCSS	Y	Y	Y	Y	Y ^b	Y ^b
Health Check for FICON Dynamic routing	Y	Y	Y	Y	N	N
z/VM Dynamic I/O support for FICON Express16S+ FC and FCP CHPIDs	-	-	-	-	Y	Y
CHPID (Channel-Path Identifier) type FC						
Extended distance FICON ^c	Y	Y	Y	Y	Y	Y
FICON Express16S+ for support of zHPF (z Systems High-Performance FICON)	Y	Y	Y	Y	Y	Y
FICON Express16S for support of zHPF	Y	Y	Y	Y	Y	Y
FICON Express8S for support of zHPF	Y	Y	Y	Y	Y	Y
MIDAW (Modified Indirect Data Address Word)	Y	Y	Y	Y	Y ^e	Y ^e
zDAC (z/OS Discovery and Auto-Configuration)	Y	Y	Y	Y	N	N
FICON Express16S+ when using FICON or CTC (channel-to-channel)	Y	Y	Y	Y	Y ^d	Y ^d
FICON Express16S when using FICON or CTC	Y	Y	Y	Y	Y ^d	Y ^d
FICON Express8S when using FICON or CTC	Y	Y	Y	Y	Y ^d	Y ^d
Global resource serialization (GRS) FICON CTC toleration	Y	Y	Y	Y	N	N
IPL from an alternative subchannel set	Y	Y	Y	Y	Y	Y
32 K subchannels for the FICON Express16S+	Y	Y	Y	Y	Y	Y
32 K subchannels for the FICON Express16S	Y	Y	Y	Y	Y	Y
Request node identification data	Y	Y	Y	Y	N	N

Function ^a	z/OS V2R3	z/OS V2R2	z/OS V2R1	z/OS V1R13	z/VM V7R1	z/VM V6R4
FICON link incident reporting	Y	Y	Y	Y	N	N
CHPID (Channel-Path Identifier) type FCP						
FICON Express16S+ for support of SCSI devices	-	-	-	-	Y	Y
FICON Express16S for support of SCSI devices	-	-	-	-	Y	Y
FICON Express8S for support of SCSI devices	-	-	-	-	Y	Y
FICON Express16S+ support of hardware data router	-	-	-	-	Y ^e	Y ^e
FICON Express16S support of hardware data router	-	-	-	-	Y ^e	Y ^e
FICON Express8S support of hardware data router	-	-	-	-	Y ^e	Y ^e
FICON Express16S+ T10-DIF support	-	-	-	-	Y ^e	Y ^e
FICON Express16S T10-DIF support	-	-	-	-	Y ^e	Y ^e
FICON Express8S T10-DIF support	-	-	-	-	Y ^e	Y ^e
Increased performance for the FCP protocol	-	-	-	-	Y	Y
N_Port ID Virtualization (NPIV)	-	-	-	-	Y	Y
Worldwide port name tool	-	-	-	-	Y	Y

a. PTFs might be required for toleration support or use of z14 ZR1 features and functions.

b. For specific Geographically Dispersed Parallel Sysplex (GDPS) usage only.

c. Transparent to operating systems.

d. CTC channel type not supported for CPC in DPM (Dynamic Partition Manager) mode.

e. For guest use.

The supported storage connectivity functions for z/VSE, z/TPF, and Linux on Z are listed in Table 7-7.

Table 7-7 Supported storage connectivity functions for z/VSE, z/TPF, and Linux on Z

Function ^a	z/VSE V6R2	z/VSE V6R1	z/VSE V5R2	z/TPF V1R1	Linux on Z ^b
zHyperLink Express	-	-	-	-	-
The 63.75-K subchannels	N	N	N	N	Y
Three logical channel subsystems (LCSSs)	Y	Y	Y	N	Y
Three subchannel set per LCSS	Y	Y	Y	N	Y
Health Check for FICON Dynamic routing	N	N	N	N	N
z/VM Dynamic I/O support for FICON Express16S+ FC and FCP CHPIDs	-	-	-	-	-
CHPID (Channel-Path Identifier) type FC					
Extended distance FICON ^c	Y	Y	Y	Y	Y
FICON Express16S+ for support of zHPF (IBM Z High-Performance FICON) ^d	Y	N	N	Y	Y
FICON Express16S for support of zHPF	Y	N	N	Y	Y

Function ^a	z/VSE V6R2	z/VSE V6R1	z/VSE V5R2	z/TPF V1R1	Linux on Z ^b
FICON Express8S for support of zHPF	Y	N	N	Y	Y
MIDAW (Modified Indirect Data Address Word)	N	N	N	N	N
zDAC (z/OS Discovery and Auto-Configuration)	-	-	-	-	-
FICON Express16S+ when using FICON or CTC (channel-to-channel)	Y	Y	Y	Y	Y ^e
FICON Express16S when using FICON or CTC	Y	Y	Y	Y	Y ^e
FICON Express8S when using FICON or CTC	Y	Y	Y	Y	Y
Global resource serialization (GRS) FICON CTC toleration	-	-	-	-	-
IPL from an alternative subchannel set	N	N	N	N	N
32 K subchannels for the FICON Express16S+	N	N	N	N	Y
32 K subchannels for the FICON Express16S	N	N	N	N	Y
Request node identification data	N	N	N	N	N
FICON link incident reporting	N	N	N	N	N
CHPID (Channel-Path Identifier) type FCP					
FICON Express16S+ for support of SCSI devices	Y	Y	Y	-	Y
FICON Express16S for support of SCSI devices	Y	Y	Y	-	Y
FICON Express8S for support of SCSI devices	Y	Y	Y	-	Y
FICON Express16S+ support of hardware data router	N	N	N	N	Y
FICON Express16S support of hardware data router	N	N	N	N	Y
FICON Express8S support of hardware data router	N	N	N	N	Y
FICON Express16S+ T10-DIF support	N	N	N	N	Y
FICON Express16S T10-DIF support	N	N	N	N	Y
FICON Express8S T10-DIF support	N	N	N	N	Y
Increased performance for the FCP protocol	Y	Y	Y	-	Y
N_Port ID Virtualization (NPIV)	Y	Y	Y	N	Y
Worldwide port name tool	-	-	-	-	Y

a. PTFs might be required for toleration support or use of z14 ZR1 features and functions.

b. Support statement varies based on Linux on Z distribution and release.

c. Transparent to operating systems.

d. Will be supported on z/VSE V6.2 with PTFs.

e. CTC channel type not supported for CPC in DPM (Dynamic Partition Manager) mode.

7.3.3 Network connectivity

The supported network connectivity functions for z/OS and z/VM are listed in Table 7-8.

Table 7-8 Supported network connectivity functions for z/OS and z/VM

Function ^a	z/OS V2R3	z/OS V2R2	z/OS V2R1	z/OS V1R13	z/VM V7R1	z/VM V6R4
Checksum offload for IPV6 packets	Y	Y	Y	Y	Y ^b	Y ^b
Checksum offload for LPAR-to-LPAR traffic with IPv4 and IPv6	Y	Y	Y	Y	Y ^b	Y ^b
Querying and displaying an OSA configuration	Y	Y	Y	Y	N	N
QDIO data connection isolation for z/VM	-	-	-	-	Y	Y
QDIO interface isolation for z/OS	Y	Y	Y	Y	-	-
QDIO OLM (Optimized Latency Mode)	Y	Y	Y	Y	-	-
Adapter interruptions for QDIO	N	N	N	N	Y	Y
QDIO Diagnostic Synchronization	Y	Y	Y	Y	N	N
IWQ (Inbound Workload Queuing) for OSA	Y	Y	Y	Y	Y ^b	Y ^b
VLAN management enhancements	Y	Y	Y	Y	Y ^c	Y ^c
GARP VLAN Registration Protocol	Y	Y	Y	Y	Y	Y
Link aggregation support for z/VM	-	-	-	-	Y	Y
Multi-vSwitch Link Aggregation	-	-	-	-	Y	Y
Large send for IPV6 packets	Y	Y	Y	Y	Y ^b	Y ^b
z/VM Dynamic I/O Support for OSA-Express6S OSD CHPIDs	-	-	-	-	Y	Y
OSA Dynamic LAN idle	Y	Y	Y	Y	N	N
OSA Layer 3 virtual MAC for z/OS environments	Y	Y	Y	Y	-	-
Network Traffic Analyzer	Y	Y	Y	Y	N	N
Hipersockets						
HiperSockets ^d	Y	Y	Y	Y	Y	Y
32 HiperSockets	Y	Y	Y	Y	Y	Y
HiperSockets Completion Queue	Y	Y	Y	Y	Y	Y
HiperSockets Virtual Switch Bridge	-	-	-	-	Y	Y
HiperSockets Multiple Write Facility	Y	Y	Y	Y	N	N
HiperSockets support of IPV6	Y	Y	Y	Y	Y	Y
HiperSockets Layer 2 support	Y	Y	Y	Y	Y	Y
HiperSockets Network Traffic Analyzer for Linux on Z	-	-	-	-	-	-
SMC-D and SMC-R						
SMC-D ^e over ISM (Internal Shared Memory)	Y	Y	N	N	Y ^b	Y ^b

Function ^a	z/OS V2R3	z/OS V2R2	z/OS V2R1	z/OS V1R13	z/VM V7R1	z/VM V6R4
10GbE RoCE ^f Express	Y	Y	Y	Y	Y ^b	Y ^b
25GbE & 10GbE RoCE Express2 for SMC-R	Y	Y	Y	N	Y ^b	Y ^b
25GbE & 10GbE RoCE Express2 for Ethernet communications ^g including Single-Root I/O Virtualization (SR-IOV)	N	N	N	N	Y ^b	Y ^b
z/VM Dynamic I/O support for RoCE Express2	-	-	-	-	Y	Y
Shared RoCE environment	Y	Y	Y	N	Y	Y
Open Systems Adapter (OSA)^{h,i}						
OSA-Express6S 1000BASE-T Ethernet CHPID type OSC	Y	Y	Y	Y	Y	Y
OSA-Express5S 1000BASE-T Ethernet CHPID type OSC	Y	Y	Y	Y	Y	Y
OSA-Express7S ^j 25-Gigabit Ethernet Short Reach (SR) CHPID type OSD	Y	Y	Y	Y	Y ^k	Y ^k
OSA-Express6S 10-Gigabit Ethernet Long Reach (LR) and Short Reach (SR) CHPID type OSD	Y	Y	Y	Y	Y	Y
OSA-Express5S 10-Gigabit Ethernet Long Reach (LR) and Short Reach (SR) CHPID type OSD	Y	Y	Y	Y	Y	Y
OSA-Express4S 10-Gigabit Ethernet Long Reach (LR) and Short Reach (SR) CHPID type OSD	Y	Y	Y	Y	Y	Y
OSA-Express6S Gigabit Ethernet LX and SX CHPID type OSD	Y	Y	Y	Y	Y	Y
OSA-Express5S Gigabit Ethernet LX and SX CHPID type OSD	Y	Y	Y	Y	Y	Y
OSA-Express4S Gigabit Ethernet LX and SX CHPID type OSD	Y	Y	Y	Y	Y	Y
OSA-Express6S 1000BASE-T Ethernet CHPID type OSD	Y	Y	Y	Y	Y	Y
OSA-Express5S 1000BASE-T Ethernet CHPID type OSD	Y	Y	Y	Y	Y	Y
OSA-Express6S 1000BASE-T Ethernet CHPID type OSE	Y	Y	Y	Y	Y	Y
OSA-Express5S 1000BASE-T Ethernet CHPID type OSE	Y	Y	Y	Y	Y	Y

a. PTFs might be required for toleration support or exploitation of z14 ZR1 features and functions.

b. For guest use or exploitation.

c. Support of guests is transparent to z/VM if the device is directly connected to the guest (pass through).

d. On z14 ZR1, the CHPID statement of HiperSockets devices requires the keyword VCHID. Therefore, the IOCP definitions must be migrated to support the HiperSockets definitions (CHPID type IQD). VCHID specifies the virtual channel identification number that is associated with the channel path (valid range is 7E0 - 7FF). VCHID is not valid on Z servers before z13.

e. Shared Memory Communications - Direct Memory Access.

f. Remote Direct Memory Access (RDMA) over Converged Ethernet.

- g. Does not require a peer OSA.
- h. Supported CHPID types: OSC, OSD, OSE, and OSM.
- i. OSC, OSX, and OSE channel types not supported on a CPC running in DPM mode.
- j. Requires PTFs for APARs OA55256 (VTAM®) and PI95703 (TCP/IP).
- k. Requires PTF for APAR PI99085.

The supported network connectivity functions for z/VSE, z/TPF, and Linux on Z are listed in Table 7-9.

Table 7-9 Supported network connectivity functions for z/VSE, z/TPF, and Linux on Z

Function ^a	z/VSE V6R2	z/VSE V6R1	z/VSE V5R2	z/TPF V1R1	Linux on Z ^b
Checksum offload for IPV6 packets	N	N	N	N	Y
Checksum offload for LPAR-to-LPAR traffic with IPv4 and IPv6	N	N	N	N	Y
Querying and displaying an OSA configuration	N	-	-	-	-
QDIO data connection isolation for z/VM	-	-	-	-	-
QDIO interface isolation for z/OS	-	-	-	-	-
QDIO OLM (Optimized Latency Mode)	-	-	-	-	-
Adapter interruptions for QDIO	Y	Y	Y	N	Y
QDIO Diagnostic Synchronization	N	N	N	N	N
IWQ (Inbound Workload Queuing) for OSA	N	N	N	N	N
VLAN management enhancements	N	N	N	N	N
GARP VLAN Registration Protocol	N	N	N	N	Y ^c
Link aggregation support for z/VM	N	N	N	N	N
Multi-vSwitch Link Aggregation	N	N	N	N	N
Large send for IPV6 packets	N	N	N	N	Y
z/VM Dynamic I/O Support for OSA-Express6S OSD CHPIDs	N	N	N	N	N
OSA Dynamic LAN idle	N	N	N	N	N
OSA Layer 3 virtual MAC for z/OS environments	-	-	-	-	-
Network Traffic Analyzer	-	-	-	-	-
Hipersockets					
HiperSockets ^d	Y	Y	Y	N	Y
32 HiperSockets	Y	Y	Y	N	Y
HiperSockets Completion Queue	Y	Y	Y	N	Y
HiperSockets Virtual Switch Bridge	-	-	-	-	Y ^e
HiperSockets Multiple Write Facility	N	N	N	N	N
HiperSockets support of IPV6	Y	Y	Y	N	Y
HiperSockets Layer 2 support	N	N	N	N	Y

Function ^a	z/VSE V6R2	z/VSE V6R1	z/VSE V5R2	z/TPF V1R1	Linux on Z ^b
HiperSockets Network Traffic Analyzer for Linux on Z	N	N	N	N	Y
SMC-D and SMC-R					
SMC-D ^f over ISM (Internal Shared Memory)	N	N	N	N	N
10GbE RoCE ^g Express	N	N	N	N	Y ^h
25GbE & 10GbE RoCE Express2 for SMC-R	N	N	N	N	N
25GbE & 10GbE RoCE Express2 for Ethernet communications ⁱ including Single Root I/O Virtualization (SR-IOV)	N	N	N	N	Y
z/VM Dynamic I/O support for RoCE Express2	-	-	-	-	-
Shared RoCE environment	N	N	N	N	Y
Open Systems Adapter (OSA)^{j,k}					
OSA-Express6S 1000BASE-T Ethernet CHPID type OSC	Y	Y	Y	Y	-
OSA-Express5S 1000BASE-T Ethernet CHPID type OSC	Y	Y	Y	Y	-
OSA-Express7S 25-Gigabit Ethernet Short Reach (SR) CHPID type OSD	Y	Y	Y	Y	Y
OSA-Express6S 10-Gigabit Ethernet Long Reach (LR) and Short Reach (SR) CHPID type OSD	Y	Y	Y	Y	Y
OSA-Express5S 10-Gigabit Ethernet Long Reach (LR) and Short Reach (SR) CHPID type OSD	Y	Y	Y	Y	Y
OSA-Express4S 10-Gigabit Ethernet Long Reach (LR) and Short Reach (SR) CHPID type OSD	Y	Y	Y	Y	Y
OSA-Express6S Gigabit Ethernet LX and SX CHPID type OSD	Y	Y	Y	Y	Y
OSA-Express5S Gigabit Ethernet LX and SX CHPID type OSD	Y	Y	Y	Y	Y
OSA-Express4S Gigabit Ethernet LX and SX CHPID type OSD	Y	Y	Y	Y	Y
OSA-Express6S 1000BASE-T Ethernet CHPID type OSD	Y	Y	Y	Y	Y
OSA-Express5S 1000BASE-T Ethernet CHPID type OSD	Y	Y	Y	Y	Y
OSA-Express6S 1000BASE-T Ethernet CHPID type OSE	Y	Y	Y	N	N
OSA-Express5S 1000BASE-T Ethernet CHPID type OSE	Y	Y	Y	N	N

a. PTFs might be required for toleration support or use of z14 ZR1 features and functions.

b. Support statement varies based on Linux on Z distribution and release.

c. By using VLANs.

- d. On z14 ZR1, the CHPID statement of HiperSockets devices requires the keyword VCHID. Therefore, the IOCP definitions must be migrated to support the HiperSockets definitions (CHPID type IQD). VCHID specifies the virtual channel identification number that is associated with the channel path (valid range is 7E0 - 7FF). VCHID is not valid on Z servers before z13.
- e. Applicable to guest operating systems.
- f. Shared Memory Communications - Direct Memory Access.
- g. Remote Direct Memory Access (RDMA) over Converged Ethernet.
- h. Linux can use RocE Express as a standard NIC (Network Interface Card) for Ethernet.
- i. Does not require a peer OSA.
- j. Supported CHPID types: OSC, OSD, OSE, and OSM.
- k. OSC, OSX, and OSE channel types not supported on a CPC running in DPM mode.

7.3.4 Cryptographic functions

The supported cryptography functions for z/OS and z/VM are listed in Table 7-10.

Table 7-10 Supported cryptography functions for z/OS and z/VM

Function ^a	z/OS V2R3	z/OS V2R2	z/OS V2R1	z/OS V1R13	z/VM V7R1	z/VM V6R4
CP Assist for Cryptographic Function (CPACF)	Y	Y	Y	Y	Y ^b	Y ^b
CPACF greater than 16 Domain Support	Y	Y	Y	Y	Y ^b	Y ^b
CPACF AES-128, AES-192, and AES-256	Y	Y	Y	Y	Y ^b	Y ^b
CPACF SHA-1, SHA-224, SHA-256, SHA-384, and SHA-512	Y	Y	Y	Y	Y ^b	Y ^b
CPACF protected key	Y	Y	Y	Y	Y ^b	Y ^b
Crypto Express6S	Y	Y ^c	Y ^c	Y ^c	Y ^b	Y ^b
Crypto Express6S Support for Visa Format Preserving Encryption	Y	Y ^c	Y ^c	Y ^c	Y ^b	Y ^b
Crypto Express6S Support for Coprocessor in PCI-HSM Compliance Mode ^d	Y	Y ^c	Y ^c	N	Y ^b	Y ^b
Crypto Express6S spouting up to 40 domains	Y	Y ^c	Y ^c	Y	Y ^b	Y ^b
Crypto Express5S	Y	Y	Y	Y	Y ^b	Y ^b
Crypto Express5S spouting up to 40 domains	Y	Y	Y	Y	Y ^b	Y ^b
Elliptic Curve Cryptography (ECC)	Y	Y	Y	Y	Y ^b	Y ^b
Secure IBM Enterprise PKCS #11 (EP11) coprocessor mode	Y	Y	Y	Y	Y ^b	Y ^b
RCE (Regional Crypto Enablement)	Y	Y	N	N	Y ^b	Y ^b
z/VM Dynamic I/O Support for RCE	Y	-	-	-	Y	Y
z/OS Data Set Encryption	Y	Y	N	N	-	-
z/VM Encrypted paging support	-	-	-	-	Y	Y
RMF Support for Crypto Express6	Y	Y	Y	N	-	-
z/OS encryption readiness technology (zERT)	Y	Y	Y	N	-	-
z/TPF transparent database encryption	-	-	-	-	-	-

- a. PTFs might be required for toleration support or use of z14 ZR1 features and functions.
- b. For guest use.
- c. A web deliverable is required. For more information and to download the deliverable, see [the z/OS downloads page](#) of the IBM IT infrastructure website.
- d. Requires TKE 9.0.

The supported cryptography functions for z/VSE, z/TPF, and Linux on Z are listed in Table 7-11.

Table 7-11 Supported cryptography functions for z/VSE, z/TPF, and Linux on Z

Function ^a	z/VSE V6R2	z/VSE V6R1	z/VSE V5R2	z/TPF V1R1	Linux on Z ^b
CP Assist for Cryptographic Function (CPACF)	Y	Y	Y	Y	Y
CPACF greater than 16 Domain Support	Y	Y	Y	N	Y
CPACF AES-128, AES-192, and AES-256	Y	Y	Y	Y ^c	Y
CPACF SHA-1, SHA-224, SHA-256, SHA-384, and SHA-512	Y	Y	Y	Y ^d	Y
CPACF protected key	N	N	N	N	N
Crypto Express6S	Y	Y	Y	Y	Y
Crypto Express6S Support for Visa Format Preserving Encryption	N	N	N	N	N
Crypto Express6S Support for Coprocessor in PCI-HSM Compliance Mode ^e	N	N	N	N	N
Crypto Express6S spouting up to 40 domains	Y	Y	Y	N	Y
Crypto Express5S	Y	Y	Y	Y	Y
Crypto Express5S spouting up to 40 domains	Y	Y	Y	N	Y
Elliptic Curve Cryptography (ECC)	Y	N	N	N	Y
Secure IBM Enterprise PKCS #11 (EP11) coprocessor mode	N	N	N	N	Y
RCE (Regional Crypto Enablement)	N	N	N	N	N
z/VM Dynamic I/O Support for RCE	N	N	N	-	N
z/OS Data Set Encryption	-	-	-	-	-
z/VM Encrypted paging support	N	N	N	-	N
RMF Support for Crypto Express6	-	-	-	-	-
z/OS encryption readiness technology (zERT)	-	-	-	-	-
z/TPF transparent database encryption	-	-	-	Y	-

- a. PTFs might be required for toleration support or use of z14 ZR1 features and functions.
- b. Support statement varies based on Linux on Z distribution and release.
- c. z/TPF supports only AES-128 and AES-256.
- d. z/TPF supports only SHA-1 and SHA-256.
- e. Requires TKE 9.0.

7.3.5 Special purpose features

The supported Special-purpose feature for z/OS, z/VM, and Linux on Z is listed in Table 7-12.

Table 7-12 Supported Special-purpose feature for z/OS and z/VM

Function ^a	z/OS V2R3	z/OS V2R2	z/OS V2R1	z/OS V1R13	z/VM V7R1	z/VM V6R4	Linux on Z
zEDC ^b Express	Y	Y	Y	N	Y ^c	Y ^c	Y ^d

a. PTFs might be required for toleration support or use of z14 ZR1 features and function.

b. zEnterprise Data Compression.

c. For guest use.

d. See the IBM support site for Linux on Z.

7.4 Support by features and functions

This section addresses operating system support by function. Only the currently in-support releases are included.

Tables in this section use the following convention:

- ▶ N/A: Not applicable
- ▶ NA: Not available

7.4.1 LPAR configuration and management

A single system image can control several processor units, such as CPs, zIIPs, or IFLs.

Maximum number of PUs per system image

The maximum number of PUs that is supported by each operating system image and by special-purpose LPARs are listed in Table 7-13.

Important: For more information about z14 ZR1 processor characterization support, see 3.5, “Processor unit functions” on page 83.

Table 7-13 Maximum number of PUs per system image

Operating system	Maximum number of PUs per system image ^a
z/OS V2R3	256 ^{a,b,c}
z/OS V2R2	256 ^{a,b,c}
z/OS V2R1	256 ^{a,b,c}
z/OS V1R13	100 ^a
z/VM V6R4	64 ^{a,d}
z/VSE V5R2 and later	z/VSE Turbo Dispatcher can use up to 4 CPs, and tolerates up to 10-way LPARs
z/TPF V1R1	86 ^a CPs
CFCC Level 22 and 23	16 CPs or ICFs CPs and ICFs cannot be mixed

Operating system	Maximum number of PUs per system image ^a
Linux on Z ^a	<ul style="list-style-type: none"> ▶ SUSE Linux Enterprise Server 12: 256 CPs or IFLs ▶ SUSE Linux Enterprise Server 11: 64 CPs or IFLs ▶ Red Hat RHEL 7: 256 CPs or IFLs ▶ Red Hat RHEL 6: 64 CPs or IFLs ▶ Ubuntu 16.04 LTS and 18.04 LTS: 256 CPs or IFLs
KVM Hypervisor ^a	The KVM hypervisor is offered with the following Linux distributions -- 256CPs or IFLs--: <ul style="list-style-type: none"> ▶ SLES 12 SP2. ▶ RHEL 7.5 with kernel-alt package (kernel 4.14). ▶ Ubuntu 16.04 LTS and 18.04 LTS.
Secure Service Container	80 ^a
GDPS Virtual Appliance	80 ^a

a. z14 ZR1 supports up to 6 CPs and up to 12 ZIIPs in a LPAR for z/OS, and up to 30 IFLs or up to 30 ICFs for other LPARs.

b. A 256-way without multithreading; 128-way with multithreading.

c. Total characterizable PUs, including zIIPs and CPs.

d. A 64-way without multithreading and 32-way with multithreading enabled.

Maximum main storage size

The maximum amount of main storage that is supported by current operating systems is listed in Table 7-14. A maximum of 8 TB of main storage can be defined for an LPAR on a z14 ZR1 server.

Table 7-14 Maximum memory that is supported by the operating system

Operating system ^a	Maximum supported main storage ^b
z/OS	z/OS V2R1 and later support 4 TB
z/VM	z/VM V6R4 supports 2 TB
z/VSE	z/VSE V5R2 and later support 32 GB
z/TPF	z/TPF supports 4 TB
CFCC	Level 22 and 23 support up to 3 TB
Secure Service Container	Supports up to 3 TB
Linux on Z (64-bit)	<ul style="list-style-type: none"> ▶ SUSE Linux Enterprise Server 12 supports 10 TB ▶ SUSE Linux Enterprise Server 11 supports 4 TB ▶ Red Hat RHEL 7 supports 10 TB ▶ Red Hat RHEL 6 supports 4TB ▶ Ubuntu 16.04 LTS and 18.04 LTS support 10 TB

a. An LPAR on z14 ZR1 supports up to 8 TB of memory.

b. z14 ZR1 servers support 8 TB user configurable memory per server.

Up to 40 LPARs

This feature was first made available on z13s servers and allows the system to be configured with up to 85 LPARs. Because channel subsystems can be shared by up to 15 LPARs, it is necessary to configure three channel subsystems to reach the 40 LPARs limit.

The supported operating systems are listed in Table 7-3 on page 214 and Table 7-4 on page 216.

Remember: A software appliance that is deployed in a Secure Service Container runs in a dedicated LPAR. When activated, it reduces the maximum number of available LPARs by one.

Separate LPAR management of PUs

z14 ZR1 servers use 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.

The supported operating systems are listed in Table 7-3 on page 214 and Table 7-4 on page 216.

Dynamic PU add

Planning an LPAR configuration includes defining reserved PUs that can be brought online when extra capacity is needed. Operating system support is required to use this capability without an IPL; that is, nondisruptively. This support is available in z/OS for some time.

The dynamic PU add function enhances this support by allowing you to dynamically define and change the number and type of reserved PUs in an LPAR profile, which removes any planning requirements. The new resources are immediately made available to the operating system and in the case of z/VM, to its guests.

The supported operating systems are listed in Table 7-3 on page 214 and Table 7-4 on page 216.

Dynamic LPAR memory upgrade

An LPAR can be defined with 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. Although these two memory zones do not have to be contiguous in real memory, they appear as logically contiguous to the operating system that runs in the LPAR.

z/OS can take advantage of this support and nondisruptively acquire and release memory from the reserved area. z/VM can acquire memory nondisruptively and immediately make it available to guests. z/VM virtualizes this support to its guests, which now also can increase their memory nondisruptively if supported by the guest operating system. Releasing memory from z/VM is not supported. Releasing memory from the z/VM guest depends on the guest's operating system support.

Linux on Z also supports acquiring and releasing memory nondisruptively. This feature is enabled for SUSE Linux Enterprise Server 11 and RHEL 6 and later releases and for supported Ubuntu versions.

LPAR group absolute capping

On z13s servers, PR/SM is enhanced to support an option to limit the amount of physical processor capacity that is used by an individual LPAR when a PU that is defined as a CP or an IFL is shared across a set of LPARs. This enhancement is designed to provide a physical capacity limit that is enforced as an absolute (versus a relative) limit. It is not affected by changes to the logical or physical configuration of the system. This physical capacity limit can be specified in units of CPs or IFLs.

The supported operating systems are listed in Table 7-3 on page 214 and Table 7-4 on page 216.

Capacity Provisioning Manager

The provisioning architecture enables clients to better control the configuration and activation of the On/Off CoD. For more information, see 8.8, “Nondisruptive upgrades” on page 320. 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 interfaces with z/OS Workload Manager (WLM) and implements capacity provisioning policies. Several implementation options are available, from an analysis mode that issues only guidelines, to an autonomic mode that provides fully automated operations.

Replacing manual monitoring with autonomic management or supporting manual operation with guidelines can help ensure that sufficient processing power is available with the least possible delay. The supported operating systems are listed in Table 7-3 on page 214.

Program-directed re-IPL

Program directed re-IPL allows an operating system on a z14 ZR1 to IPL again without operator intervention. This function is supported for SCSI and IBM extended count key data (IBM ECKD) devices.

The supported operating systems are listed in Table 7-3 on page 214 and Table 7-4 on page 216.

IOCP

All IBM Z servers require a description of their I/O configuration. This description is stored in I/O configuration data set (IOCDs) files. The I/O configuration program (IOCP) allows for the creation of the IOCDs file from a source file that is known as the I/O configuration source (IOCS).

The IOCS file contains definitions of LPARs and channel subsystems. It also includes detailed information for each channel and path assignment, each control unit, and each device in the configuration.

IOCP for z14 ZR1 supports the following features:

- ▶ z14 ZR1 Base machine definition
- ▶ New hardware (announced with Driver 36)
- ▶ IOCP support for Dynamic I/O for standalone CF (Driver 36)
- ▶ New PCI function adapter for zHyperLink (HYL)
- ▶ New PCI function adapter for RoCE Express2 (CX4)
- ▶ New IOCP Keyword MIXTYPE required for pervious FICON³ cards

IOCP required level for z14 ZR1 servers: The required level of IOCP for the z14 ZR1 is IOCP 5.4.0 with PTFs. For more information, see the following publications:

- ▶ *IBM Z Stand-Alone Input/Output Configuration Program User's Guide*, SB10-7173
- ▶ *IBM Z Input/Output Configuration Program User's Guide for ICP IOCP*, SB10-7172

Dynamic Partition Manager V3.2: At the time of this writing, the Dynamic Partition Manager V3.2 is available for managing IBM Z systems that are running Linux. DPM 3.2 is available with HMC Driver Level 36. IOCP does not need to configure a server that is running in DPM mode. For more information, see *IBM Dynamic Partition Manager (DPM) Guide*, SB10-7170-02.

³ FICON Express16S+ does not allow mixing of CHPID types on the same cards.

7.4.2 Base CPC features and functions

In this section, we describe the features and functions of Base CPC.

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.

Beginning with z/OS V1R13, the IEAOPTxx keyword **HIPERDISPATCH** defaults to YES when it is running on a z14 M0x, z14 ZR1, z13, z13s, zEC12, or zBC12 server. If HIPERDISPATCH=NO is specified, the specification is accepted as it was on previous z/OS releases.

The use of SMT on z14 ZR1 servers requires that HiperDispatch is enabled on the operating system. For more information, see “Simultaneous multithreading” on page 233.

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 logical processors are added to an LPAR that has 64 or fewer logical processors and the extra logical processors raise the number of logical processors to more than 64, the LPAR automatically operates in HiperDispatch Management Mode, regardless of the HIPERDISPATCH=YES/NO specification. That is, even if the LPAR has the HIPERDISPATCH=NO specification, that LPAR is converted to operate in HiperDispatch Management Mode.
- ▶ An LPAR with more than 64 logical processors that are running in HiperDispatch Management Mode cannot be reverted to run in non-HiperDispatch Management Mode.

HiperDispatch on z14 ZR1 servers uses a new chip and CPC drawer configuration to improve the access cache performance. Beginning with z/OS V2R1, HiperDispatch was changed to use the new node cache structure of z14 ZR1 servers. The base support is provided by PTFs that are identified by `IBM.device.server.z14-3907.requiredservice`.

The PR/SM on z14 ZR1 servers seeks to assign all logical processors that are packed into PU chips in cooperation with operating system HiperDispatch to optimize shared cache usage.

To use HiperDispatch effectively, WLM goal adjustment might be required. Review the WLM policies and goals and update them as necessary. 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 z14 ZR1 servers.

z/VM V7R1 and V6R4⁴

z/VM also uses the HiperDispatch facility for improved processor efficiency by better use of the processor cache to take advantage of the cache-rich processor, node, and drawer design of the z14 ZR1 system. The supported processor limit was increased to 64, whereas it remains at 32 with SMT and supports up to 64 threads that are running simultaneously.

CPU polarization support in Linux on Z

You can optimize the operation of a vertical SMP environment by adjusting the SMP factor based on the workload demands. For more information about CPU polarization support in Linux on Z, see the [CPU polarization page](#) of IBM Knowledge Center.

⁴ z14 ZR1 supports up to 30 IFLs (up to 60 threads with SMT enabled).

z/TPF

z/TPF on z14 ZR1 can use more processors immediately without reactivating the LPAR or IPLing the z/TPF system.

In installations older than z14 ZR1, z/TPF workload is evenly distributed across all available processors, even in low-utilization situations. This configuration causes cache and core contention with other LPARs. When z/TPF is running in a shared processor configuration, the achieved MIPS is higher when z/TPF uses a minimum set of processors.

In low-utilization periods, z/TPF now minimizes the processor footprint by compressing TPF workload onto a minimal set of I-streams (engines), which reduces the effect on other LPARs and allows the entire CPC to operate more efficiently.

As a consequence, z/OS and z/VM experience less contention from the z/TPF system when the z/TPF system is operating at periods of low demand.

The supported operating systems are listed in Table 7-3 on page 214 and Table 7-4 on page 216.

zIIP support

zIIPs do not change the model capacity identifier of z14 ZR1 servers. IBM software product license charges that are based on the model capacity identifier are not affected by the addition of zIIPs. On a z14 ZR1 server, z/OS Version 1 Release 13 is the minimum level for supporting zIIPs.

No changes to applications are required to use zIIPs. They can be used by the following applications:

- ▶ Db2 V8 and later for z/OS data serving for applications that use data Distributed Relational Database Architecture (DRDA) over TCP/IP, such as data serving, data warehousing, and selected utilities.
- ▶ z/OS XML services.
- ▶ z/OS CIM Server.
- ▶ z/OS Communications Server for network encryption (Internet Protocol Security [IPSec]) and for large messages that are sent by HiperSockets.
- ▶ IBM GBS Scalable Architecture for Financial Reporting.
- ▶ IBM z/OS Global Mirror (formerly XRC) and System Data Mover.
- ▶ IBM OMEGAMON® XE on z/OS, OMEGAMON XE on Db2 Performance Expert, and Db2 Performance Monitor.
- ▶ Any Java application that uses the current IBM SDK.
- ▶ WebSphere Application Server V5R1 and later, and products that are based on it, such as WebSphere Portal, WebSphere Enterprise Service Bus (WebSphere ESB), and WebSphere Business Integration (WBI) for z/OS.
- ▶ CICS/TS V2R3 and later.
- ▶ Db2 UDB for z/OS Version 8 and later.
- ▶ IMS Version 8 and later.
- ▶ zIIP Assisted HiperSockets for large messages.
- ▶ z/OSMF (z/OS Management Facility).
- ▶ IBM z/OS Platform for Apache Spark.
- ▶ IBM Machine Learning for z/OS.

The functioning of a zIIP is transparent to application programs. The supported operating systems are listed in Table 7-3 on page 214.

On z14 ZR1 servers, the zIIP processor is designed to run in SMT mode, with up to two threads per processor. This new function is designed to help improve throughput for zIIP workloads and provide appropriate performance measurement, capacity planning, and SMF accounting data. This support is available for z/OS V2.1 with PTFs and higher.

Use the **PROJECTCPU** option of the IEAOPTxx parmlib member to help determine whether zIIPs can be beneficial to the installation. Setting PROJECTCPU=YES directs z/OS to record the amount of eligible work for zIIPs in SMF record type 72 subtype 3. The field APPL% IIPCP of the Workload Activity Report listing by WLM service class indicates the percentage of a processor that is zIIP eligible. Because of the zIIP's lower price as compared to a CP, even a utilization as low as 10% can provide cost benefits.

Transactional Execution

The IBM zEnterprise EC12 introduced an architectural feature called Transactional Execution (TX). This capability is known in academia and industry as *hardware transactional memory*. Transactional execution is also implemented on z14 (models M0x and ZR1), z13, z13s, and zBC12 servers.

This feature enables software to indicate to the hardware the beginning and end of a group of instructions that must be treated in an atomic way. All of their results occur or none occur, 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 occurs. 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 increases the software's efficiency by providing a way to avoid locks (lock elision). This advantage is of special importance for speculative code generation and highly parallelized applications.

TX is used by IBM Java virtual machine (JVM) and might be used by other software. The supported operating systems are listed in Table 7-3 on page 214 and Table 7-4 on page 216.

Simultaneous multithreading

SMT is the hardware capability to process up to two simultaneous threads in a single core, sharing the resources of the superscalar core. This capability improves the system capacity and efficiency in the usage of the processor, which increases the overall throughput of the system.

The z14 ZR1 can run up two threads simultaneously in the same processor, which dynamically shares resources of the core, such as cache, translation lookaside buffer (TLB), and execution resources. It provides better utilization of the cores and more processing capacity.

SMT⁵ is supported for zIIPs and IFLs.

Note: For zIIPs and IFLs, SMT must be enabled on z/OS, z/VM, or Linux on Z instances. An operating system with SMT support can be configured to dispatch work to a thread on a zIIP (for eligible workloads in z/OS) or an IFL (for z/VM) core in single-thread or SMT mode.

⁵ On IBM z14 Model ZR1, SMT is also enabled (not user configurable) by default for SAPs.

The supported operating systems are listed in Table 7-3 on page 214 and Table 7-4 on page 216.

An operating system that uses SMT controls each core and is responsible for maximizing their throughput and meeting workload goals with the smallest number of cores. In z/OS, HiperDispatch cache optimization should be considered when you must choose the two threads to be dispatched in the same processor.

HiperDispatch attempts to dispatch guest virtual CPUs on the same logical processor on which they ran. PR/SM attempts to dispatch a vertical low logical processor in the same physical processor. If that process is not possible, it attempts to dispatch it in the same node, or then the same CPC drawer where it was dispatched before to maximize cache reuse.

From the perspective of an application, SMT is transparent and no changes are required in the application for it to run in an SMT environment, as shown in Figure 7-1.

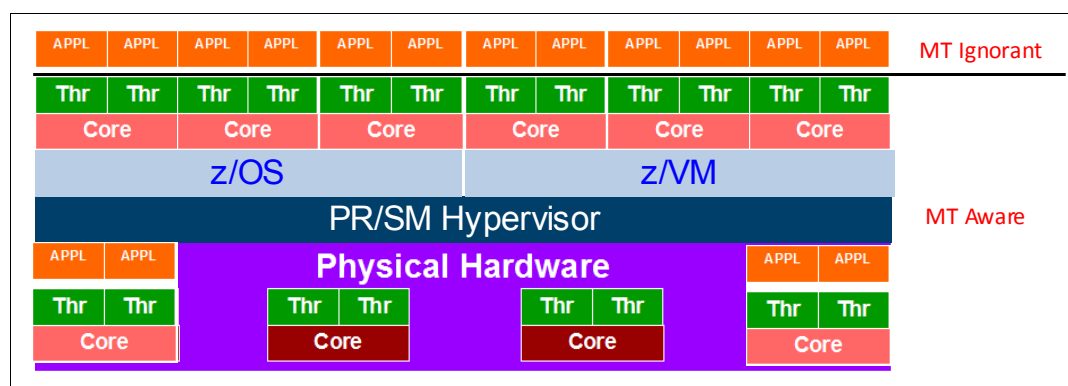


Figure 7-1 Simultaneous multithreading

z/OS

The following APARs must be applied to z/OS V2R1 to use SMT:

- ▶ OA43366 (BCP)
- ▶ OA43622 (WLM)
- ▶ OA44439 (XCF)

The use of SMT on z/OS V2R1 requires enabling HiperDispatch, and defining the processor view (**PROCVIEW**) control statement in the LOADxx parmlib member and the **MT_ZIIP_MODE** parameter in the IEAOPTxx parmlib member.

The **PROCVIEW** statement is defined for the life of IPL, and can have the following values:

- ▶ **CORE**: This value specifies that z/OS should configure a processor view of core, in which a core can include one or more threads. The number of threads is limited by z14 ZR1 to two threads. If the underlying hardware does not support SMT, a core is limited to one thread.
- ▶ **CPU**: This value is the default. It specifies that z/OS should configure a traditional processor view of CPU and not use SMT.
- ▶ **CORE,CPU_OK**: This value specifies that z/OS should configure a processor view of core (as with the **CORE** value) but the **CPU** parameter is accepted as an alias for applicable commands.

When **PROCVIEW CORE** or **CORE,CPU_OK** are specified in z/OS that is running in z14 ZR1, HiperDispatch is forced to run as enabled, and you cannot disable HiperDispatch. The **PROCVIEW** statement cannot be changed dynamically; therefore, you must run an IPL after changing it to make the new setting effective.

The **MT_ZIIP_MODE** parameter in the IEAOPTxx controls zIIP SMT mode. It can be 1 (the default), where only one thread can be running in a core, or 2, where up to two threads can be running in a core. If **PROCVIEW CPU** is specified, the **MT_ZIIP_MODE** is always 1. Otherwise, the use of SMT to dispatch two threads in a single zIIP logical processor (**MT_ZIIP_MODE=2**) can be changed dynamically by using the **SET OPT=xx** setting in the IEAOPTxx parmlib. Changing the MT mode for all cores can take some time to complete.

The activation of SMT mode also requires that the HMC Customize/Delete Activation Profiles task “Do not end the time slice if a partition enters a wait state” must not be selected. This setting is the recommended default setting.

PROCVIEW CORE requires **DISPLAY M=CORE** and **CONFIG CORE** to display the core states and configure an entire core.

With the introduction of Multi-Threading support for SAPs in z14 ZR1, RMF is updated to support this change by implementing page break support in the I/O Queuing Activity report that is generated by the RMF Post processor.

z/VM V7R1 and V6R4⁶

The use of SMT in z/VM is enabled by using the **MULTITHREADING** statement in the system configuration file. Multithreading is enabled only if z/VM is configured to run with the HiperDispatch vertical polarization mode enabled and with the dispatcher work distribution mode set to reshuffle.

The default in z/VM is multithreading disabled. With the addition of dynamic SMT capability to z/VM V6R4 through an SPE, the number of active threads per core can be changed without a system outage and potential capacity gains going from SMT-1 to SMT-2 (one to two threads per core) can now be achieved dynamically. Dynamic SMT requires applying PTFs that are running in SMT enabled mode and enables dynamically varying the active threads per core.

z/VM supports up to 32 multithreaded cores (64 threads) for IFLs, and each thread is treated as an independent processor. z/VM dispatches virtual IFLs on the IFL logical processor so that the same or different guests can share a core. Each core has a single dispatch vector, and z/VM attempts to place virtual sibling IFLs on the same dispatch vector to maximize cache reuses.

The guests have no awareness of SMT, and cannot use it. z/VM SMT exploitation does not include guest support for multithreading. The value of this support for guests is that the first-level z/VM hosts under the guests can achieve higher throughput from the multi-threaded IFL cores.

Linux on Z and the KVM hypervisor

The upstream kernel 4.0 features SMT functionality that was developed by the Linux on Z development team. SMT is supported on LPAR only (not as a second-level guest). For more information, see the [Kernel 4.0 page of the developerWorks website](#).

The following minimum releases of Linux on Z distributions natively support SMT:

- ▶ Red Hat RHEL 7.2
- ▶ SUSE Linux Enterprise Server12 SP1
- ▶ Ubuntu 16.04 LTS
- ▶ KVM hypervisor, which is offered with the following Linux distributions:
 - SLES 12 SP2 with service.

⁶ z14 ZR1 supports up to 30 IFLs (up to 60 threads with SMT enabled).

- RHEL 7.5 with kernel-alt package (kernel 4.14).
- Ubuntu 16.04 LTS with service and Ubuntu 18.04 LTS with service

Single-instruction multiple-data

The SIMD feature introduces a new set of instructions to enable parallel computing that can accelerate code with string, character, integer, and floating point data types. The SIMD instructions allow a larger number of operands to be processed with a single complex instruction.

z14 ZR1 is equipped with new set of instructions to improve the performance of complex mathematical models and analytic workloads through vector processing and new complex instructions, which can process much data with a single instruction. This new set of instructions, which is known as SIMD, enables more consolidation of analytic workloads and business transactions on Z servers.

SIMD on z14 ZR1 includes support for 32-bit floats and enhanced math libraries that provide performance improvements for analytical workloads by processing more information with a single CPU instruction.

The supported operating systems are listed in Table 7-3 on page 214 and Table 7-4 on page 216. Operating System support includes the following items⁷:

- ▶ Enablement of vector registers.
- ▶ Use of vector registers that use XL C/C++ ARCH(11) and TUNE(11).
- ▶ A math library with an optimized and tuned math function (Mathematical Acceleration Subsystem, or MASS) that can be used in place of some of the C standard math functions. It includes a SIMD vectorized and non-vectorized version.
- ▶ A specialized math library, which is known as Automatically Tuned Linear Algebra Software (ATLAS), that is optimized for the hardware.
- ▶ IBM Language Environment® for C runtime function enablement for ATLAS.
- ▶ DBX to support the disassembly of the new vector instructions, and to display and set vector registers.
- ▶ XML SS exploitation to use new vector processing instructions to improve performance.

MASS and ATLAS can reduce the time and effort for middleware and application developers. IBM provides compiler built-in functions for SIMD that software applications can use as needed, such as for using string instructions.

The use of new hardware instructions through XL C/C++ ARCH(12) and TUNE(12) or SIMD usage by MASS and ATLAS libraries requires the z14 support for z/OS V2R1 XL C/C++ web deliverable.

The followings compilers include built-in functions for SIMD:

- ▶ IBM Java
- ▶ XL C/C++
- ▶ Enterprise COBOL
- ▶ Enterprise PL/I

⁷ These features might not be available on all operating systems that are listed.

Code must be developed to take advantage of the SIMD functions. Applications with SIMD instructions abend if they run on a lower hardware level system. Some mathematical function replacement can be done without code changes by including the scalar MASS library before the standard math library.

The MASS and standard math library include different accuracies, so assess the accuracy of the functions in the context of the user application before deciding whether to use the MASS and ATLAS libraries.

The SIMD functions can be disabled in z/OS partitions at IPL time by using the **MACHMIG** parameter in the LOADxx member. To disable SIMD code, use the MACHMIG VEF hardware-based vector facility. If you do not specify a **MACHMIG** statement, which is the default, the system is unlimited in its use of the Vector Facility for z/Architecture (SIMD).

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. z14 ZR1 servers inherited the decimal floating point accelerator feature that was introduced with z10 EC.

z14 ZR1 features a new decimal architecture with Vector Enhancements Facility and Vector Packed Decimal Facility for Data Access Accelerator. Vector Packed Decimal Facility introduces a set of instructions that perform operations on decimal types that use vector registers to improve performance.

The supported operating systems are listed in Table 7-3 on page 214 and Table 7-4 on page 216. For more information, see 7.5.4, “z/OS XL C/C++ considerations” on page 274.

Out-of-order execution

Out-of-order (OOO) execution yields significant performance benefits for compute-intensive applications by reordering instruction execution, which allows later (newer) instructions to be run ahead of a stalled instruction. Storage accesses and parallel storage accesses also are reordered. OOO maintains good performance growth for traditional applications.

The supported operating systems are listed in Table 7-3 on page 214 and Table 7-4 on page 216. For more information, see “3.4.3, “Out-of-Order execution” on page 74.

CPU Measurement Facility

Also known as Hardware Instrumentation Services (HIS), CPU Measurement Facility (CPUMF) was introduced with z10 EC to gain insight into the interaction of workload and hardware it runs on. CPU MF data can be collected by z/OS System Measurement Facility on SMF 113 records. The supported operating systems are listed in Table 7-3 on page 214.

For more information about this function, see [The Load-Program-Parameter and the CPU-Measurement Facilities](#).

For more information about the CPU Measurement Facility, see the [CPU MF - Update and WSC Experiences page](#) of the IBM Techdocs Library website.

For more information, see “12.2, “LSPR workload suite” on page 399.

Large page support

In addition to the 1-MB large pages, 4-KB pages, and page frames, z14 ZR1 servers support pageable 1-MB large pages (large pages that are 2 GB) and large page frames. The supported operating systems are listed in Table 7-3 on page 214 and Table 7-4 on page 216.

Virtual Flash Memory

IBM Virtual Flash Memory (VFM) is the replacement for the Flash Express features (FC 0402, FC 0403), which were available on the IBM zBC12 and IBM z13s. No application changes are required to change from IBM Flash Express to VFM because it implements EADM Architecture by using HSA-like memory instead of Flash card pairs.

IBM Virtual Flash Memory (FC 0614) offers up to 2.0 TB of memory in 512 GB increments for improved application availability and to handle paging workload spikes.

IBM Virtual Flash Memory is designed to help improve availability and handling of paging workload spikes when running z/OS V2.1, V2.2, or V2.3, or on z/OS V1.13. With this support, z/OS is designed to help improve system availability and responsiveness by using VFM across transitional workload events, such as market openings, and diagnostic data collection. z/OS is also designed to help improve processor performance by supporting middleware exploitation of pageable large (1 MB) pages.

Therefore, VFM can help organizations meet their most demanding service level agreements and compete more effectively. VFM is easily configurable, and to provide rapid time to value.

The supported operating systems are listed in Table 7-3 on page 214 and Table 7-4 on page 216.

Guarded Storage Facility

Also known as less-pausing garbage collection, Guarded Storage Facility (GSF) is a new architecture that was introduced with z14 ZR1 to enable enterprise scale Java applications to run without periodic pause for garbage collection on larger heaps.

z/OS

GSF support allows an area of storage to be identified such that an Exit routine assumes control if a reference is made to that storage. GSF is managed by new instructions that define Guarded Storage Controls and system code to maintain that control information across undispach and redispach.

Enabling a less-pausing approach can improve Java garbage collection. Function is provided on z14 ZR1 that is running z/OS 2.2 (with APAR OA51643 installed) and later. The **MACHMIG** statement in **LOADxx** of **SYS1.PARMLIB** provides disables the function.

z/VM

With the PTF for APAR VM65987, z/VM V6.4 supports guest exploitation of the z14 ZR1 guarded storage facility. This facility is designed to improve the performance of garbage-collection processing by various languages, in particular Java.

The supported operating systems are listed in Table 7-3 on page 214 and Table 7-4 on page 216.

Instruction Execution Protection

Instruction Execution Protection (IEP) is a new hardware function that was introduced with z14 ZR1 that enables software, such as Language Environment, to mark certain memory regions (for example, a heap or stack), as non-executable to improve the security of programs running on IBM Z servers against stack-overflow or similar attacks.

Through enhanced hardware features (based on DAT table entry bit) and specific software requests to obtain memory areas as non-executable, areas of memory can be protected from unauthorized execution. A Protection Exception occurs if an attempt is made to fetch an instruction from an address in such an element or if an address in such an element is the target of an execute-type instruction.

z/OS

To use IEP, Real Storage Manager (RSM) is enhanced to request non-executable memory allocation. Use new keyword **EXECUTABLE=YES|NO** on **STORAGE OBTAIN** or **IARV64** to indicate whether memory to be used contains executable code. Recovery Termination Manager (RTM) writes LOGREC record of any program-check that results from IEP.

IEP support is for z/OS 2.2 with APARs OA51030 and OA51643 installed and z/OS 2.3 running on a z14 processor.

z/VM

Guest exploitation support for the Instruction Execution Protection Facility is provided with APAR VM65986.

The supported operating systems are listed in Table 7-3 on page 214 and Table 7-4 on page 216.

Co-processor compression enhancements

Since z196, IBM Z processors include on-chip co-processors with built-in capabilities, such as compression, expansion, and translation. Each follow-on server generation improved on-chip co-processors functionality and performance and z14 ZR1 is taking this idea even further with following enhancements:

- Entropy Encoding for CMPSC with Huffman Coding

z14 ZR1 enables increased compression ratio (by using Huffman coding) for the on-chip compression coprocessor, which results in fewer CPU cycles to enable further data compression. This feature improves memory, transfer, and disk efficiency. Software-based expansion of Huffman-encoded text is used when hardware support is not present. Support is provided by z/OS 2.1 and later running on z14 ZR1 with OA49967 installed.

- Order Preserving Compression

Order Preserving Compression allows comparisons against compressed data. It helps to achieve further disk and memory savings by compressing search trees, index, or sort files that were impractical to compress previously.

Combining Order Preserving Compression (individual compression of index entries) with Entropy Encoding (to compress non-index data) can result up to 35% data reduction⁸ for Db2 considering that significant portion of Db2 disk space is in indexes for fast access to data.

The supported operating systems are listed in Table 7-3 on page 214 and Table 7-4 on page 216.

⁸ Results are based on modeling, not actual measurements.

7.4.3 Coupling and clustering features and functions

In this section, we describe the coupling and cluster features.

Coupling facility and CFCC considerations

Coupling facility (CF) connectivity to a z14 ZR1 is supported on the z14 M0x, z13, z13s, or another z14 ZR1. The CFCC levels that are supported on Z servers are listed in Table 7-15.

Table 7-15 IBM Z CFCC code-levels

IBM Z system	Code level
z14 M0x and z14 ZR1	CFCC Level 22 or CFCC Level 23
z13	CFCC Level 20 or CFCC Level 21
z13s	CFSSCC Level 21
zEC12	CFCC Level 18 or CFCC Level 19
zBC12	CFCC Level 19

Consideration: Because coupling link connectivity to zEC12/zBC12 and previous systems is not supported, introducing z14 ZR1 into an installation requires extra planning. Consider the level of CFCC. For more information, see “Migration considerations” on page 156.

CFCC Level 23

CFCC Level 23 is delivered on z14 ZR1 servers with driver level 36. In addition to CFCC Level 22 enhancements, it introduces the following enhancements:

- ▶ Asynchronous cross invalidation (XI) for CF cache structures
This enhancement requires z/OS fixes for APARs OA54688 (exploitation) and OA54985 (toleration). It also requires explicit data manager support (Db2 V12 with PTFs).
- ▶ Coupling Facility hang detection enhancements
This enhancement provide a significant reduction in failure scope and client disruption (CF-level to structure-level), with no loss of FFDC collection capability. With this support, the CFCC dispatcher significantly reduces the CF hang detection interval to only 2 seconds, which allows more timely detection and recovery from such events.
When a hang is detected, the CF confines the scope of the failure in most cases to “structure damage” for the single CF structure the hung command was processing against, captures diagnostics with a non-disruptive CF dump, and continues operating without aborting or rebooting the CF image.
- ▶ Coupling Facility granular latching
This enhancement eliminates the performance degradation that is caused by structure-wide latching. With this support, most CF list and lock structure ECR processing no longer uses structure-wide latching; instead, it serializes its execution by using the normal structure object latches that all mainline commands use. However, a small number of “edge conditions” in ECR processing still require structure-wide latching.

Before you begin the migration process, install the compatibility and coexistence PTFs. A planned outage is required when you upgrade the CF or CF LPAR to CFCC Level 23.

CFCC Level 22

CFCC Level 22 is delivered on z14 ZR1 servers with driver level 32. CFCC Level 22 introduces the following enhancements:

- ▶ *Coupling Express Long Range (CE LR)*: A new link type that was introduced with z14 ZR1 for long-distance coupling connectivity.
- ▶ *Coupling Facility (CF) Processor Scalability*: CF work management and dispatching changes for IBM z14 Model ZR1 allow improved efficiency and scalability for coupling facility images.

First, ordered work queues were eliminated from the CF in favor of first-in/first-out queues, which avoids the overhead of maintaining ordered queues.

Second, protocols for system-managed duplexing were simplified to avoid the potential for latching deadlocks between duplexed structures.

Third, the CF image can now use its processors to perform specific work management functions when the number of processors in the CF image exceeds a threshold. Together, these changes improve the processor scalability and throughput for a CF image.

- ▶ *CF List Notification Enhancements*: Significant enhancements were made to CF notifications that inform users about the status of shared objects within in a Coupling Facility.

First, structure notifications can use a round-robin scheme for delivering immediate and deferred notifications that avoids excessive “shotgun” notifications, which reduces notification overhead.

Second, an option is now available for delivering “aggressive” notifications, which can drive a notification when new elements are added to a queue. This feature provides initiative to get new work processed in a timely manner.

Third, notifications can now be driven when a queue transitions between full and not-full, which allows users to redrive messages that could not previously be written to a “full” queue. The combination of these notification enhancements provides flexibility to accommodate notification preferences among various CF users and yields more consistent, timely notifications.

- ▶ *CF Encryption*: z/OS 2.3 supports end-to-end encryption for CF data in flight and data at rest in CF structures (as a part of the Pervasive Encryption solution). Host-based CPACF encryption is used for high performance and low latency. IBM z14 Model ZR1 CF images are not required, but are recommended to simplify some sysplex recovery and reconciliation scenarios involving encrypted CF structures. (The CF image never decrypts or encrypts any data). IBM z14 Model ZR1 z/OS images are not required, but are recommended for the improved AES CBC encrypt/decrypt performance that z14 ZR1 provides.

The supported operating systems are listed in Table 7-5 on page 217.

For more information about the latest CFCC code levels, see [the current exception letter](#) that is published on IBM Resource Link® website (login is required).

CF structure sizing changes are expected when upgrading from a previous CFCC Level to CFCC Level 21. Review the CF LPAR size by using the available CFSizer tool, which is available for [download from the IBM Systems support website](#).

Sizer Utility, an authorized z/OS program download, is useful when you are upgrading a CF. The tool is available [for download from the IBM Systems support website](#).

Before you begin the migration process, install the compatibility and coexistence PTFs. A planned outage is required when you upgrade the CF or CF LPAR to CFCC Level 22.

Coupling links support

Integrated Coupling Adapter (ICA) Short Reach and Coupling Express Long Reach (CE LR) coupling link options provide high-speed connectivity at short and longer distances over fiber optic interconnections. Several areas of this book address CE LR and ICA SR characteristics and support. For more information, see 4.6.4, “Parallel Sysplex connectivity” on page 153.

Integrated Coupling Adapter

PCIe Gen3 fanout, which is also known as Integrated Coupling Adapter Short Range (ICA SR), supports a maximum distance of 150 meters (492 feet) and is defined as CHPID type CS5 in IOCP.

Coupling Express Long Reach

The Coupling Express Long Reach (CE LR) link provides point-to-point coupling connectivity at distances of 10 kilometers (6.21 miles) unrepeatable and defined as CHPID type CL5 in IOCP. The supported operating systems are listed in Table 7-5 on page 217.

Note: IBM z14 ZR1 server does *not* support HCA3-O fanout for 12x IFB (FC 0171) and HCA3-O LR fanout for 1x IFB (FC 0170).^a As announced previously, z14 Model M0x is the last IBM Z server to support these adapters.

Enterprises should migrate from HCA3-O and HCA3-O LR adapters to ICA SR or Coupling Express Long Reach (CE LR) adapters on z14 ZR1, z14 M0x, z13, and z13s. For high-speed short-range coupling connectivity, enterprises should migrate to the Integrated Coupling Adapter (ICA SR).

For long-range coupling connectivity, enterprises should migrate to the new Coupling Express LR coupling adapter. For long-range coupling connectivity that requires a DWDM, enterprises must determine their needed DWDM vendor's plan to qualify the planned replacement long-range coupling links.

IBM Z enterprises should plan to migrate from InfiniBand coupling links. For high-speed short-range coupling connectivity, enterprises should migrate to the Integrated Coupling Adapter (ICA SR).

For long-range coupling connectivity, enterprises should migrate to the new CE LR coupling link. For long-range coupling connectivity that requires a DWDM, enterprises must determine their wanted DWDM vendor's plan to qualify the CE LR. For more information, see Hardware Announcement 117-031, dated March 2017.

a. Per previous Statement of Direction.

Virtual Flash Memory use by CFCC

VFM can be used in coupling facility images to provide extended capacity and availability for workloads that use WebSphere MQ Shared Queues structures. The use of VFM can help availability by reducing latency from paging delays that can occur at the start of the workday or during other transitional periods. It is also designed to help eliminate delays that can occur when diagnostic data during failures are collected.

CFCC Coupling Thin Interrupts

The Coupling Thin Interrupts enhancement is delivered with CFCC 19. It improves the performance of a CF partition and improves the dispatching of z/OS LPARs that are awaiting the arrival of returned asynchronous CF requests when used in a shared engine environment. For more information, see “Coupling Thin Interrupts” on page 86. The supported operating systems are listed in Table 7-5 on page 217.

Asynchronous CF Duplexing for lock structures

Asynchronous CF Duplexing enhancement is a general-purpose interface for any CF Lock structure user. It enables secondary structure updates to be performed asynchronously regarding primary updates. Initially delivered with CFCC 21 on z13 as an enhanced continuous availability solution, it offers performance advantages for duplexing lock structures and avoids the need for synchronous communication delays during the processing of every duplexed update operation.

Asynchronous CF Duplexing for lock structures requires the following software support:

- ▶ z/OS V2.3
- ▶ z/OS V2.2 SPE with PTFs for APAR OA47796 and OA49148
- ▶ z/VM V6.4 with PTFs for z/OS exploitation of guest coupling environment
- ▶ Db2 12 with PTFs for APAR PI66689
- ▶ IRLM V2.3 with PTFs for APAR PI68378

The supported operating systems are listed in Table 7-5 on page 217.

Asynchronous cross-invalidate (XI) for CF cache structures

Asynchronous XI for CF cache structures enables improved efficiency in CF data sharing by adopting a more transactional behavior for cross-invalidate (XI) processing, which is used to maintain coherency and consistency of data managers' local buffer pools across the sysplex. Instead of performing XI signals synchronously on every cache update request that causes them, data managers can "opt in" for the CF to perform these XIs asynchronously (and then sync them up with the CF at or before transaction completion). Data integrity is maintained if all XI signals complete by the time transaction locks are released.

The feature enables faster completion of cache update CF requests (especially with cross-site distance that is involved) and provides improved cache structure service times and coupling efficiency. It requires explicit data manager exploitation and participation, which is not transparent to the data manager. No SMF data changes were made for CF monitoring and reporting.

The following requirements must be met:

- ▶ CFCC Level 23 support, plus
- ▶ z/OS PTFs on every exploiting system in the sysplex:
- ▶ Fixes for APAR OA54688: Exploitation support z/OS 2.2 and 2.3
- ▶ Fixes for APAR OA54985: Toleration support for z/OS 1.13 and 2.1
- ▶ Db2 V12 with PTFs for exploitation

z/VM Dynamic I/O support for ICA and CE LR CHPIDs

z/VM dynamic I/O configuration support allows you to add, delete, and modify the definitions of channel paths, control units, and I/O devices to the server and z/VM without shutting down the system.

This function refers exclusively to the z/VM dynamic I/O support of ICA and CE LR coupling links. Support is available for CS5/CL5 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 the ICA or CE LR features, and does not support the use of ICA or CE LR coupling links by guests. The supported operating systems are listed in Table 7-5 on page 217.

7.4.4 Storage connectivity-related features and functions

In this section, we describe the storage connectivity-related features and functions.

zHyperlink Express

z14 ZR1 introduces IBM zHyperLink Express as a new IBM Z input/output (I/O) channel link technology since FICON. zHyperLink Express is designed to help bring data closer to processing power, increase the scalability of Z transaction processing, and lower I/O latency.

zHyperLink Express is designed for up to 5x lower latency than High-Performance FICON for Z (zHPF) by directly connecting the Z Central Processor Complex (CPC) to the I/O Bay of the DS8880. This short distance (up to 150 meters), direct connection is intended to speed Db2 for z/OS transaction processing and improve active log throughput.

The improved performance of zHyperLink Express allows the Processing Unit (PU) to make a synchronous request for the data that is in the DS8880 cache. This feature eliminates the undispatch of the running request, queuing delays to resume the request, and PU cache disruption.

Support for zHyperLink Writes can accelerate Db2 log writes to help deliver superior service levels by processing high-volume Db2 transactions at speed. IBM zHyperLink Express (FC 0431) requires compatible levels of DS8880/F hardware, firmware R8.5.1, and Db2 12 with PTFs.

The supported operating systems are listed in Table 7-6 on page 218 and Table 7-7 on page 219.

FICON Express16S+

FICON Express16S+ supports a link data rate of 16 gigabits per second (Gbps) and autonegotiation to 4 or 8 Gbps for synergy with switches, directors, and storage devices. With support for native FICON, High-Performance FICON for Z (zHPF), and Fibre Channel Protocol (FCP), the IBM z14 Model ZR1 server enables you to position your SAN for even higher performance, which helps you to prepare for an end-to-end 16 Gbps infrastructure to meet the lower latency and increased bandwidth demands of your applications.

The new FICON Express16S+ channel works with your fiber optic cabling environment (single mode and multimode optical cables). The FICON Express16S+ feature running at end-to-end 16 Gbps link speeds provides reduced latency for large read/write operations and increased bandwidth compared to the FICON Express8S feature.

The supported operating systems are listed in Table 7-6 on page 218 and Table 7-7 on page 219.

FICON Express16S

FICON Express16S supports a link data rate of 16 Gbps and autonegotiation to 4 or 8 Gbps for synergy with switches, directors, and storage devices. With support for native FICON, zHPF, and FCP, the z14 ZR1 server enables SAN for even higher performance, which helps to prepare for an end-to-end 16 Gbps infrastructure to meet the increased bandwidth demands of your applications.

The new features for the multimode and single mode fiber optic cabling environments reduce latency for large read/write operations and increase bandwidth compared to the FICON Express8S features.

The supported operating systems are listed in Table 7-6 on page 218 and Table 7-7 on page 219.

FICON Express8S

The FICON Express8S 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 include the same type).

Statement of Direction^a: IBM z14 Model M0x is the last z Systems and IBM Z high-end server to support FICON Express8S (FC 0409 and FC 0410) channels. Enterprises should begin migrating from FICON Express8S channels to FICON Express16S+ channels (FC 0427 and FC 0428). FICON Express8S is supported on future high-end IBM Z servers as carry forward on an upgrade.

- a. All statements regarding IBM plans, directions, and intent are subject to change or withdrawal without notice. Any reliance on these statements of general direction is at the relying party's sole risk and will not create liability or obligation for IBM.

FICON Express8S introduced a hardware data router for more efficient zHPF data transfers. It is the first channel with hardware that is designed to support zHPF, as compared to FICON Express8, FICON Express4, and FICON Express2, which include a firmware-only zHPF implementation.

The supported operating systems are listed in Table 7-6 on page 218 and Table 7-7 on page 219.

Extended distance FICON

An enhancement to the industry-standard FICON architecture (FC-SB-3) helps avoid degradation of performance at extended distances by implementing a new protocol for persistent IU pacing. Extended distance FICON is transparent to operating systems and applies to all FICON Express16S+, FICON Express16S, and FICON Express8S features that carry native FICON traffic (CHPID type FC).

To use this enhancement, the control unit must support the new IU pacing protocol. IBM System Storage® DS8000® series supports extended distance FICON for IBM Z environments. The channel defaults to current pacing values when it operates with control units that cannot use extended distance FICON.

The supported operating systems are listed in Table 7-6 on page 218 and Table 7-7 on page 219.

High-performance FICON

High-performance FICON (zHPF) was first provided on IBM System z10®, and is a FICON architecture for protocol simplification and efficiency. It reduces the number of information units (IUs) that are processed. Enhancements were made to the z/Architecture and the FICON interface architecture to provide optimizations for online transaction processing (OLTP) workloads.

zHPF is available on z14 Model M0x, z14 Model ZR1, z13, z13s, zEC12, and zBC12 servers. The FICON Express16S+, FICON Express16S, and FICON Express8S (CHPID type FC) concurrently support the FICON protocol and the zHPF protocol in the server LIC.

When used by the FICON channel, the z/OS operating system, and the DS8000 control unit or other subsystems, the FICON channel processor usage can be reduced and performance improved. Appropriate levels of Licensed Internal Code (LIC) are required.

Also, the changes to the architectures provide end-to-end system enhancements to improve reliability, availability, and serviceability (RAS).

zHPF is compatible with the following 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

For example, the zHPF channel programs can be used by the z/OS OLTP I/O workloads, Db2, VSAM, the partitioned data set extended (PDSE), and the z/OS file system (zFS).

At the zHPF announcement, zHPF supported the transfer of small blocks of fixed size data (4 K) from a single track. This capability was 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, which results in higher throughputs and lower response times.

The multitrack operations extension applies exclusively to the FICON Express16S+, FICON Express16S, and FICON Express8S, on the z14 M0x, z14 ZR1, z13, z13s, zEC12, and zBC12, 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.

zHPF is enhanced to allow all large write operations (greater than 64 KB) at distances up to 100 kilometers (62.13 miles) to be run in a single round trip to the control unit. This process does not increase the I/O service time for these write operations at extended distances. This enhancement to zHPF removes a key inhibitor for clients that are adopting zHPF over extended distances, especially when the IBM HyperSwap capability of z/OS is used.

From the z/OS perspective, the FICON architecture is known as *command mode* and the zHPF architecture is known as *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) that are defined to the same LCU must support zHPF. The inclusion of any non-compliant zHPF features in the path group causes the entire path group to support command mode only.

The mode that is used for an I/O operation depends on the control unit that supports zHPF and its settings in the z/OS operating system. For z/OS use, a parameter is available 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 (CCWs). How zHPF (transport mode) manages channel program operations is different from the CCW operation for the existing FICON architecture (command mode).

While in command mode, each CCW is sent to the control unit for execution. In transport mode, multiple channel commands are packaged together and sent over the link to the control unit in a single control block. Fewer processors are used compared to the existing FICON architecture. Certain complex CCW chains are not supported by zHPF.

The supported operating systems are listed in Table 7-6 on page 218 and Table 7-7 on page 219.

For more information about FICON channel performance, see the performance technical papers that are available [at the IBM Z I/O connectivity page](#) of the IBM IT infrastructure website.

Modified Indirect Data Address Word facility

The Modified Indirect Data Address Word (MIDAW) facility improves FICON performance. It provides a more efficient channel command word (CCW)/indirect data address word (IDAW) structure for certain categories of data-chaining I/O operations.

The MIDAW facility is a system architecture and software feature that is designed to improve FICON performance. This facility was first made available on IBM System z9® servers, and is used by the Media Manager in z/OS.

The MIDAW facility provides a more efficient CCW/IDAW structure for certain categories of data-chaining I/O operations.

MIDAW can improve FICON performance for extended format data sets. Non-extended data sets can also benefit from MIDAW.

MIDAW can improve channel utilization and I/O response time. It also 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. The supported operating systems are listed in Table 7-6 on page 218 and Table 7-7 on page 219.

MIDAW technical description

An IDAW is used to specify data addresses for I/O operations in a virtual environment.⁹ The 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 manage complete 2 K or 4 K units of data.

⁹ Exceptions are made to this statement, and many details are omitted in this description. In this section, we assume that you can merge this brief description with an existing understanding of I/O operations in a virtual memory environment.

A single CCW that controls the transfer of data that spans non-contiguous 4 K frames in main storage is shown in Figure 7-2. When the IDAW flag is set, the data address in the CCW points to a list of words (IDAWs). Each IDAW contains an address that designates a data area within real storage.

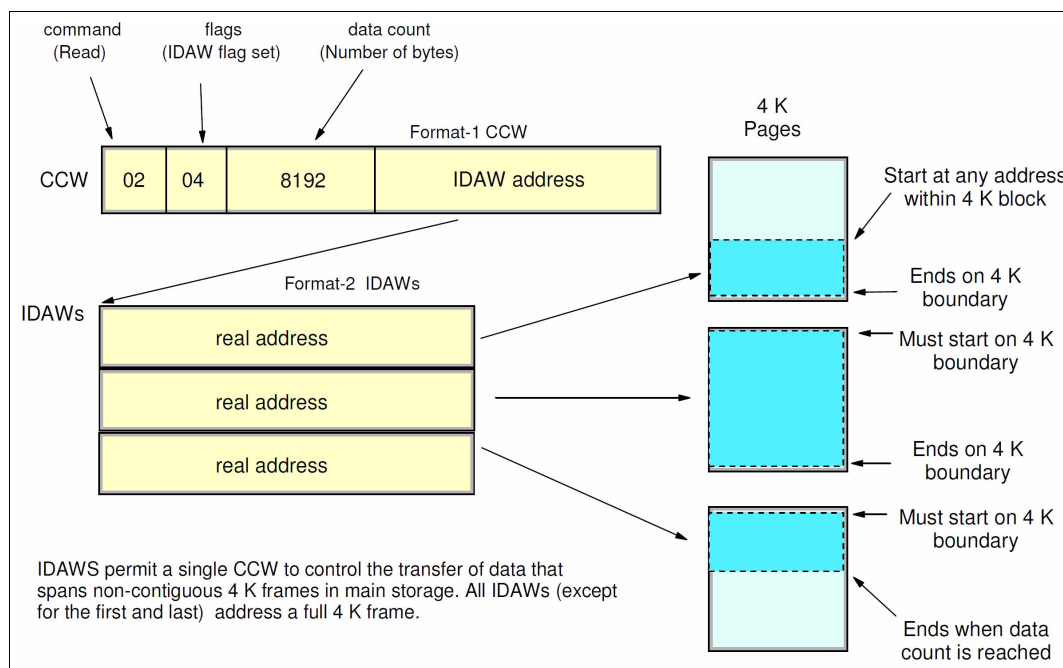


Figure 7-2 IDAW usage

The number of required IDAWs for a CCW is determined by the following factors:

- ▶ IDAW format as specified in the operation request block (ORB)
- ▶ Count field of the CCW
- ▶ Data address in the initial IDAW

For example, three IDAWS are required when the following events occur:

- ▶ The ORB specifies format-2 IDAWs with 4 KB blocks.
- ▶ The CCW count field specifies 8 KB.
- ▶ The first IDAW designates a location in the middle of a 4 KB block.

CCWs with data chaining can be used to process I/O data blocks that have a more complex internal structure, in which portions of the data block are directed into separate buffer areas. This process is sometimes known as *scatter-read* or *scatter-write*. However, as technology evolves and link speed increases, data chaining techniques become less efficient because of switch fabrics, control unit processing and exchanges, and other issues.

The MIDAW facility is a method of gathering and scattering data from and into discontinuous storage locations during an I/O operation. The MIDAW format is shown in Figure 7-3. It is 16 bytes long and is aligned on a quadword.

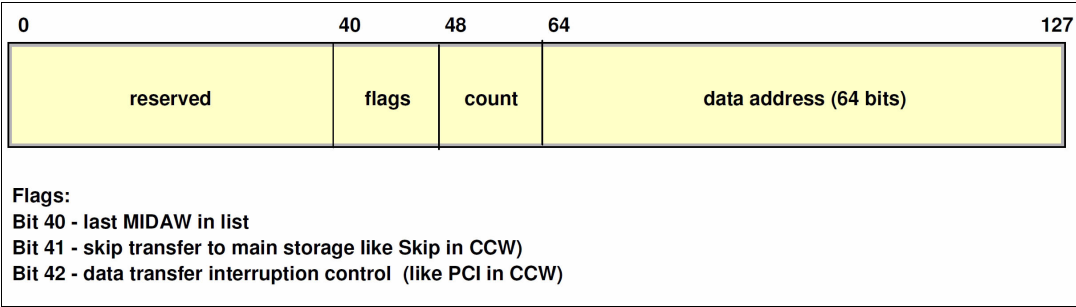


Figure 7-3 MIDAW format

An example of MIDAW usage is shown in Figure 7-4.

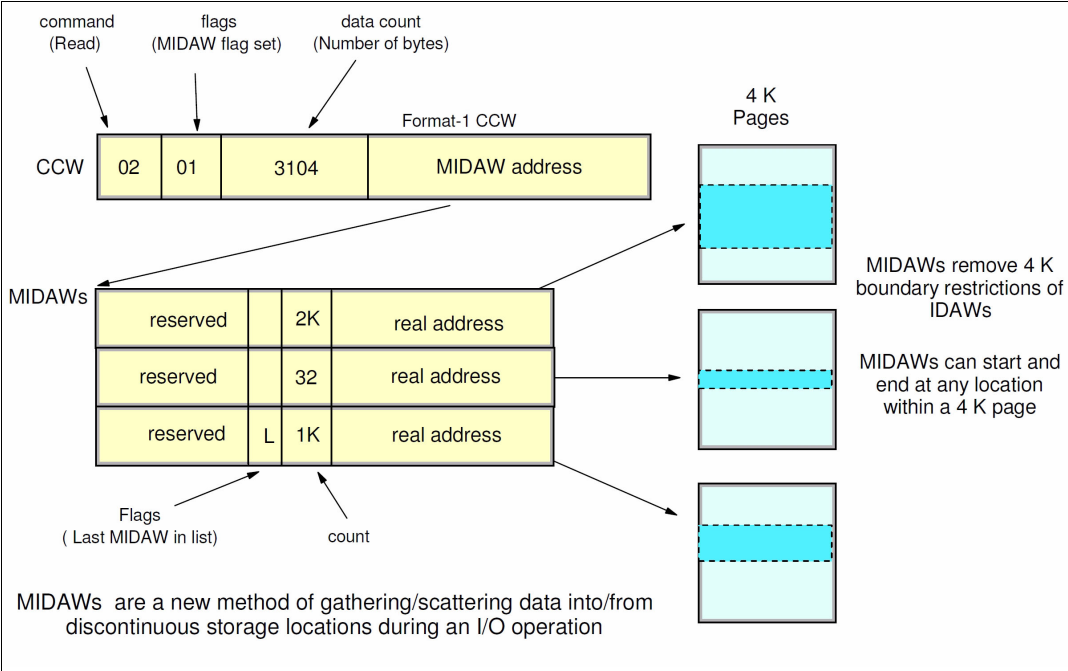


Figure 7-4 MIDAW usage

The use of MIDAWs is indicated by the MIDAW bit in the CCW. If this bit is set, the skip flag cannot be set in the CCW. The skip flag in the MIDAW can be used instead. The data count in the CCW must equal the sum of the data counts in the MIDAWs. The CCW operation ends when the CCW count goes to zero or the last MIDAW (with the last flag) ends.

The combination of the address and count in a MIDAW cannot cross a page boundary. Therefore, 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 buffers that are 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.

Extended format data sets

z/OS extended format (EF) data sets use internal structures (usually not visible to the application program) that require a scatter-read (or scatter-write) operation. Therefore, CCW data chaining is required, which produces less than optimal I/O performance. Because the most significant performance benefit of MIDAWs is achieved with EF data sets, a brief review of the EF data sets is included in this section.

VSAM and non-VSAM (DSORG=PS) sets can be defined as EF data sets. For non-VSAM data sets, a 32-byte suffix is appended to the end of every physical record (that is, block) on disk. VSAM appends the suffix to the end of every control interval (CI), which normally corresponds to a physical record.

A 32 K CI is split into two records to span tracks. This suffix is used to improve data reliability, and facilitates other functions that are described next. 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 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 the following 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 the use of 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 extra CCW for the EF data set suffix is eliminated.

MIDAWs benefit EF and non-EF data sets. For example, to read 12 4 K records from a non-EF data set on a 3390 track, Media Manager chains together 12 CCWs by using data chaining. To read 12 4 K records from an EF data set, 24 CCWs are chained (two CCWs per 4 K record). By using Media Manager track-level command operations and MIDAWs, an entire track can be transferred by using a single CCW.

Performance benefits

z/OS Media Manager features I/O channel program support for implementing EF data sets, and automatically uses MIDAWs when appropriate. Most disk I/Os in the system are generated by using Media Manager.

Users of the Executing Fixed Channel Programs in Real Storage (EXCPVR) instruction can construct channel programs that contain MIDAWs. However, doing so requires that they construct an IOBE with the IOBEMIDA bit set. Users of the EXCP instruction cannot construct channel programs that contain MIDAWs.

The MIDAW facility removes the 4 K boundary restrictions of IDAWs and, for 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, and use the channel resources more efficiently.

The performance of a specific workload can vary based on 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 page](#) of the IBM IT infrastructure website, which includes information about FICON channel performance.
- ▶ *DS8000 Performance Monitoring and Tuning*, SG24-7146.

ICKDSF

Device Support Facilities, ICKDSF, Release 17 is required on all systems that share disk subsystems with a z14 ZR1 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 determine all sharing systems that might run ICKDSF. Therefore, this support is required for z14 ZR1.

Remember: The need for ICKDSF Release 17 also applies to systems that are not part of the same sysplex, or are running an operating system other than z/OS, such as z/VM.

z/OS Discovery and Auto-Configuration

z/OS Discovery and Auto Configuration (zDAC) is designed to automatically run several I/O configuration definition tasks for new and changed disk and tape controllers that are connected to a switch or director, when attached to a FICON channel.

The zDAC function is integrated into the 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 (IODFs) 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 z14 ZR1 systems that run z/OS, and reduces complexity and setup time.

zDAC applies to all FICON features that are supported on z14 ZR1 when configured as CHPID type FC. The supported operating systems are listed in Table 7-6 on page 218 and Table 7-7 on page 219.

Platform and name server registration in FICON channel

The FICON Express16S+, FICON Express16S, and FICON Express8S features 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 z14 ZR1 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.

The 63.75K subchannels

Servers before z9 BC reserved 1024 subchannels for internal system use, out of a maximum of 64K subchannels. Starting with z9 BC, the number of reserved subchannels was reduced to 256, which increased the number of available subchannels. Reserved subchannels exist in subchannel set 0 only. One subchannel is reserved in each of subchannel sets 1 and 2.

The informal name, 63.75K subchannels, represents 65280 subchannels, as shown in the following equation:

$$63 \times 1024 + 0.75 \times 1024 = 65280$$

This equation is applicable for subchannel set 0. For subchannel sets 1 and 2, the available subchannels are derived by using the following equation:

$$(64 \times 1024) - 1 = 65535$$

The supported operating systems are listed in Table 7-6 on page 218 and Table 7-7 on page 219.

Multiple subchannel sets

First introduced in z9 EC, multiple subchannel sets (MSS) provide a mechanism for addressing more than 63.75K I/O devices and aliases for FICON (CHPID types FC) on the z14 M0x, z14 ZR1, z13, z13s, zEC12, and zBC12. z196 introduced the third subchannel set (SS2). With z13, one more subchannel set (SS3) was introduced, which expands the alias addressing by 64K more I/O devices.

z/VM V6R3 and later MSS support for mirrored direct access storage device (DASD) provides a subset of host support for the MSS facility to allow the use of an alternative subchannel set for Peer-to-Peer Remote Copy (PPRC) secondary volumes.

The supported operating systems are listed in Table 7-6 on page 218 and Table 7-7 on page 219. For more information about channel subsystem, see Chapter 5, “Central processor complex channel subsystem” on page 161.

Third subchannel set

With z13s, a *third subchannel set* (SS2) was introduced. Together with the second subchannel set (SS1), SS2 can be used for disk alias devices of 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).

For more information about supported operating systems, see Table 7-6 on page 218 and Table 7-7 on page 219.

IPL from an alternative subchannel set

z14 ZR1 supports IPL from subchannel set 1 (SS1), or subchannel set 2 (SS2), in addition to subchannel set 0.

For more information about supported operating systems, see Table 7-6 on page 218 and Table 7-7 on page 219. For more information, see “Initial program load from an alternative subchannel set” on page 166.

32K subchannels for the FICON Express16S+ and FICON Express16S

To help facilitate growth and continue to enable server consolidation, the z14 ZR1 supports up to 32K subchannels per FICON Express16S+ and FICON Express16S channels (CHPID). More devices can be defined per FICON channel, which includes primary, secondary, and alias devices. The maximum number of subchannels across all device types that are addressable within an LPAR remains at 63.75K for subchannel set 0 and 64K (64 X 1024)-1 for subchannel sets 1 and 2.

This support is exclusive to the z14 M0x, z14 ZR1, z13, and z13s servers and applies to the FICON Express16S+ and FICON Express16S features (defined as CHPID type FC). FICON Express8S remains at 24 subchannel support when defined as CHPID type FC.

The supported operating systems are listed in Table 7-6 on page 218 and Table 7-7 on page 219.

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. The supported operating systems are listed in Table 7-6 on page 218.

FICON link incident reporting

FICON link incident reporting allows an operating system image (without operator intervention) to register link incident reports. The supported operating systems are listed in Table 7-6 on page 218.

Health check for FICON dynamic routing

Starting with z13, the channel microcode was changed to support FICON dynamic routing. Although change is required in z/OS to support dynamic routing, I/O errors can occur if the FICON switches are configured for dynamic routing despite the missing support in the processor or storage controllers. Therefore, a health check is provided that interrogates the switch to determine whether dynamic routing is enabled in the switch fabric.

No action is required on z/OS to enable the health check; it is automatically enabled at IPL and reacts to changes that might cause problems. The health check can be disabled by using the **PARMLIB** or **SDSF** modify commands.

The supported operating systems are listed in Table 7-6 on page 218. For more information about FICON Dynamic Routing (FIDR), see “Central processor complex I/O system structure” on page 117.

Global resource serialization FICON CTC toleration

For some configurations that depend on ESCON CTC definitions, global resource serialization (GRS) FICON CTC toleration that is provided with APAR OA38230 is essential, especially after ESCON channel support was removed from IBM Z starting with zEC12.

The supported operating systems are listed in Table 7-6 on page 218.

Increased performance for the FCP protocol

The FCP LIC is modified to help increase I/O operations per second for small and large block sizes, and to support 16-Gbps link speeds.

For more information about FCP channel performance, see [the performance technical papers that are available](#) at the IBM Z I/O connectivity page of the IBM IT infrastructure website.

The FCP protocol is supported by z/VM, z/VSE, and Linux on Z. The supported operating systems are listed in Table 7-6 on page 218 and Table 7-7 on page 219.

T10-DIF support

American National Standards Institute (ANSI) T10 Data Integrity Field (DIF) standard is supported on IBM Z for SCSI end-to-end data protection on fixed block (FB) LUN volumes. IBM Z provides added end-to-end data protection between the operating system and the DS8870 unit. This support adds protection information that consists of Cyclic Redundancy Checking (CRC), Logical Block Address (LBA), and host application tags to each sector of FB data on a logical volume.

IBM Z support applies to FCP channels only. The supported operating systems are listed in Table 7-6 on page 218 and Table 7-7 on page 219.

N_Port ID Virtualization

N_Port ID Virtualization (NPIV) allows multiple system images (in LPARs or z/VM guests) to use a single FCP channel as though each were the sole user of the channel. First introduced with z9 EC, this feature can be used with supported FICON features on z14 ZR1 servers. The supported operating systems are listed in Table 7-6 on page 218 and Table 7-7 on page 219.

Worldwide port name tool

Part of the z14 ZR1 system installation is the pre-planning of the SAN environment. IBM includes a stand-alone tool to assist with this planning before the installation.

The capabilities of the WWPN are extended to calculate and show WWPNs for virtual and physical ports ahead of system installation.

The tool assigns WWPNs to each virtual FCP channel or port by using the same WWPN assignment algorithms that a system uses when assigning WWPNs for channels that use NPIV. Therefore, the SAN can be set up in advance, which allows 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 be created manually or exported from the HCD/HCM. The supported operating systems are listed in Table 7-6 on page 218 and Table 7-7 on page 219.

The WWPN tool is applicable to all FICON channels that are defined as CHPID type FCP (for communication with SCSI devices) on z14 ZR1. It is available [for download at the Resource Link](#) at the following website (login required).

Note: An optional feature can be ordered for WWPN persistency before shipment to keep the same I/O serial number on the new CPC. Current information must be provided during the ordering process.

7.4.5 Networking features and functions

In this section, we describe the networking features and functions.

25GbE RoCE Express2

Based on the RoCE Express2 generation hardware, the 25GbE RoCE Express2 (FC 0430), which was introduced with October 2nd, 2018 Announcement, provides two 25GbE physical ports and requires 25GbE optics and Ethernet switch 25GbE support. The switch port must support 25GbE (negotiation down to 10GbE is not supported).

The 25GbE RoCE Express2 has one PCHID and the same virtualization characteristics and the 10GbE RoCE Express2 (FC 0412) - 62 Virtual Functions per PCHID.

z/OS requires fixes for APAR OA55686. RMF 2.2 and later is also enhanced to recognize the CX4 card type and properly display CX4 cards in the PCIe Activity reports.

25GbE RoCE Express2 feature also are exploited by Linux on Z for applications that are coded to the native RoCE verb interface or use Ethernet (such as TCP/IP). This native exploitation does not require a peer OSA (see Table 7-8 on page 221 and Table 7-9 on page 223).

10GbE RoCE Express2

z14 ZR1 introduces the next generation of RoCE technology with the IBM 10GbE RoCE Express2, which provides a natively attached PCIe+ I/O drawer-based Ethernet feature that supports 10 Gbps Converged Enhanced Ethernet (CEE) and RDMA over CEE (RoCE). The RoCE feature, with an OSA feature, enables shared memory communications between two CPCs by using a shared switch.

On z14 ZR1, RoCE Express2 provides increased virtualization (sharing capability) by supporting 31 Virtual Functions (VFs) per physical port for a total of 62 VFs per PCHID. This configuration allows RoCE to be extended to more workloads.

z/OS Communications Server (CS) provides a new software device driver ConnectX4 (CX4) for RoCE Express2. The device driver is not apparent to both upper layers of the CS (the SMC-R and TCP/IP stack) and application software (by using TCP sockets). RoCE Express2 introduces a minor change in how the physical port is configured.

RMF 2.2 and later is also enhanced to recognize the new CX4 card type and properly display CX4 cards in the PCIe Activity reports.

10GbE RoCE Express2 feature also are used by Linux on Z for applications that are coded to the native RoCE verb interface or use Ethernet (such as TCP/IP). This native exploitation does not require a peer OSA.

The supported operating systems are listed in Table 7-8 on page 221 and Table 7-9 on page 223.

10GbE RoCE Express

z14 ZR1 servers support carrying forward the 10GbE RoCE Express feature. This feature provides support to the second port on the adapter and sharing the ports to up to 31 partitions (per adapter) by using both ports.

The 10-Gigabit Ethernet (10GbE) RoCE Express feature is designed to help reduce consumption of CPU resources for applications that use the TCP/IP stack (such as WebSphere accessing a Db2 database). Use of the 10GbE RoCE Express feature also can help reduce network latency with memory-to-memory transfers by using Shared Memory Communications over Remote Direct Memory Access (SMC-R) in z/OS V2R1 or later.

It is transparent to applications and can be used for LPAR-to-LPAR communication on a single z14 ZR1 server or for server-to-server communication in a multiple CPC environment.

The supported operating systems are listed in Table 7-8 on page 221 and Table 7-9 on page 223. For more information, see Appendix D, “Shared Memory Communications” on page 425.

Shared Memory Communication - Direct Memory Access

First introduced with z13 servers, the Shared Memory Communication - Direct Memory Access (SMC-D) feature maintains the socket-API transparency aspect of SMC-R so that applications that use TCP/IP communications can benefit immediately without requiring application software to undergo IP topology changes. Similar to SMC-R, this protocol uses shared memory architectural concepts that eliminate TCP/IP processing in the data path, yet preserve TCP/IP Qualities of Service for connection management purposes.

The supported operating systems are listed in Table 7-8 on page 221 and Table 7-9 on page 223. For more information, see Appendix D, “Shared Memory Communications” on page 425.

HiperSockets Completion Queue

The HiperSockets Completion Queue function is implemented on z14 M0x, z14 ZR1, z13, z13s, zEC12, and zBC12. The HiperSockets Completion Queue function is designed to allow HiperSockets to transfer data synchronously (if possible) and asynchronously, if necessary. Therefore, it combines ultra-low latency with more tolerance for traffic peaks. HiperSockets Completion Queue can be especially helpful in burst situations. The supported operating systems are listed in Table 7-8 on page 221 and Table 7-9 on page 223.

HiperSockets Virtual Switch Bridge

The HiperSockets Virtual Switch Bridge is implemented on z14 M0x, z14 ZR1, z13, z13s, zEC12, and zBC12. With the HiperSockets Virtual Switch Bridge, z/VM virtual switch 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 the following components:

- ▶ Other guest virtual machines on the virtual switch
- ▶ External network hosts through the virtual switch OSA UPLINK port

The supported operating systems are listed in Table 7-8 on page 221 and Table 7-9 on page 223.

HiperSockets Multiple Write Facility

The HiperSockets Multiple Write Facility allows the streaming of bulk data over a HiperSockets link between two LPARs. Multiple output buffers are supported on a single Signal Adapter (SIGA) write instruction. The key advantage of this enhancement is that it allows the receiving LPAR to process a much larger amount of data per I/O interrupt. This process is transparent to the operating system in the receiving partition. HiperSockets Multiple Write Facility with fewer I/O interrupts is designed to reduce processor utilization of the sending and receiving partitions.

Support for this function is required by the sending operating system. For more information, see “HiperSockets” on page 151. The supported operating systems are listed in Table 7-8 on page 221.

HiperSockets support of IPV6

IPv6 is expected to be a key element in the future of networking. The IPv6 support for HiperSockets allows compatible implementations between external networks and internal HiperSockets networks. The supported operating systems are listed in Table 7-8 on page 221 and Table 7-9 on page 223.

HiperSockets Layer 2 support

For flexible and efficient data transfer for IP and non-IP workloads, the HiperSockets internal networks on z14 ZR1 can support two transport modes: 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 features 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.

The supported operating systems are listed in Table 7-8 on page 221 and Table 7-9 on page 223.

HiperSockets network traffic analyzer for Linux on Z

Introduced with IBM System z10, HiperSockets network traffic analyzer (HS NTA) supports tracing Layer2 and Layer3 HiperSockets network traffic in Linux on Z. This support allows Linux on Z to control the trace for the internal virtual LAN to capture the records into host memory and storage (file systems).

Linux on Z tools can be used to format, edit, and process the trace records for analysis by system programmers and network administrators.

OSA-Express7S 25 Gigabit Ethernet SR

With October 2nd, 2018 Announcement¹⁰, IBM z14 ZR1 introduces a new generation of OSA: OSA-Express7S 25GbE (FC 0429). This feature includes one 25GbE physical port and requires 25GbE optics and Ethernet switch 25GbE support (negotiation down to 10GbE is not supported).

Consider the following operating system support requirements:

- ▶ z/OS V2R1, V2R2, and V2R3 require fixes for the OA55256 (VTAM) and PI95703 (TCP/IP) APARs.
- ▶ z/VM V6R4 and V7R1 require PTF for APAR PI99085.

The supported operating systems are listed in Table 7-8 on page 221 and Table 7-9 on page 223.

OSA-Express6S 10-Gigabit Ethernet LR and SR

z14 ZR1 introduces an Ethernet technology refresh with OSA-Express6S 10-Gigabit Ethernet features to be installed in the PCIe+ I/O drawer, which is supported by the 16 GBps PCIe Gen3 host bus. The performance characteristics are comparable to the OSA-Express5S features and they also retain the same form factor and port granularity.

The supported operating systems are listed in Table 7-8 on page 221 and Table 7-9 on page 223.

OSA-Express5S 10-Gigabit Ethernet LR and SR

Introduced with the zEC12 and zBC12, the OSA-Express5S 10-Gigabit Ethernet feature is installed exclusively in the PCIe+ I/O drawer. Each feature includes one port, which is defined as CHPID type OSD that supports the queued direct input/output (QDIO) architecture for high-speed TCP/IP communication.

The supported operating systems are listed in Table 7-8 on page 221 and Table 7-9 on page 223.

OSA-Express6S Gigabit Ethernet LX and SX

z14 ZR1 introduces an Ethernet technology refresh with OSA-Express6S Gigabit Ethernet features to be installed in the PCIe+ I/O drawer, which is supported by the 16 GBps PCIe Gen3 host bus. The performance characteristics are comparable to the OSA-Express5S features and they also retain the same form factor and port granularity.

Note: Operating system support is required to recognize and use the second port on the OSA-Express6S Gigabit Ethernet feature.

The supported operating systems are listed in Table 7-8 on page 221 and Table 7-9 on page 223.

¹⁰ Check the Announcement letter and the Driver Exception Letter for feature availability.

OSA-Express5S Gigabit Ethernet LX and SX

The OSA-Express5S Gigabit Ethernet feature is installed exclusively in the PCIe+ I/O drawer. Each feature includes 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 LAN. The ports can be defined as a spanned channel, and can be shared among LPARs and across logical channel subsystems.

Note: Operating system support is required to recognize and use the second port on the OSA-Express5S Gigabit Ethernet feature.

The supported operating systems are listed in Table 7-8 on page 221 and Table 7-9 on page 223.

OSA-Express4S Gigabit Ethernet LX and SX

The OSA-Express5S Gigabit Ethernet feature is installed in the PCIe+ I/O drawer. Each feature includes 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 LAN. The ports can be defined as a spanned channel, and can be shared among LPARs and across logical channel subsystems.

Note: Operating system support is required to recognize and use the second port on the OSA-Express5S Gigabit Ethernet feature.

The supported operating systems are listed in Table 7-8 on page 221 and Table 7-9 on page 223.

OSA-Express6S 1000BASE-T Ethernet

z14 ZR1 introduces an Ethernet technology refresh with OSA-Express6S 1000BASE-T Ethernet features to be installed in the PCIe+ I/O drawer, which is supported by the 16 GBps PCIe Gen3 host bus. The performance characteristics are comparable to the OSA-Express5S features and they also retain the same form factor and port granularity.

Note: Operating system support is required to recognize and use the second port on the OSA-Express6S 1000BASE-T Ethernet feature.

The supported operating systems are listed in Table 7-8 on page 221 and Table 7-9 on page 223.

Statement of Direction^a: OSA-Express6S 1000BASE-T adapters (#0426) is the last generation of OSA 1000BASE-T adapters to support connections operating at 100 Mbps link speed. Future OSA-Express 1000BASE-T adapter generations will support operation only at 1000 Mbps (1Gbps) link speed.

a. All statements regarding IBM plans, directions, and intent are subject to change or withdrawal without notice. Any reliance on these statements of general direction is at the relying party's sole risk and will not create liability or obligation for IBM.

OSA-Express5S 1000BASE-T Ethernet

The OSA-Express5S 1000BASE-T Ethernet feature is installed exclusively in the PCIe+ I/O drawer. Each feature includes one PCIe adapter and two ports. The two ports share a CHPID, which can be defined as OSC, OSD or OSE. The ports can be defined as a spanned channel, and can be shared among LPARs and across logical channel subsystems.

Each adapter can be configured in the following modes:

- ▶ QDIO mode, with CHPID types OSD
- ▶ Non-QDIO mode, with CHPID type OSE
- ▶ Local 3270 emulation mode, including OSA-ICC, with CHPID type OSC

Note: Operating system support is required to recognize and use the second port on the OSA-Express5S 1000BASE-T Ethernet feature.

The supported operating systems are listed in Table 7-8 on page 221 and Table 7-9 on page 223.

OSA-Express NCP support

Removal of support for configuring OSA-Express NCP support OSN CHPID types:

The IBM z13 and z13s is the last z Systems and IBM Z servers to support configuring OSN CHPID types. IBM z14 Model ZR1 servers do not support CHPID Type = OSN.

OSN CHPIDs were used to communicate between an operating system instance that is running in one logical partition and the IBM Communication Controller for Linux on Z (CCL) product in another logical partition on the same CPC. For more information about withdrawal from marketing for the CCL product, see announcement letter #914-227 that is dated 12/02/2014.

OSA-Integrated Console Controller

The OSA-Express 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 includes multiple LPAR support as shared or spanned channels.

With the OSA-ICC function, 3270 emulation for console session connections is integrated in the z14 ZR1 through a port on the OSA-Express6S 1000BASE-T or the OSA-Express5S 1000BASE-T.

OSA-ICC can be configured on a PCHID-by-PCHID basis, and is supported at any of the feature settings. Each port can support up to 120 console session connections.

To improve security of console operations and to provide a secure, validated connectivity, OSA-ICC supports Transport Layer Security/Secure Sockets Layer (TLS/SSL) with Certificate Authentication starting with z13 GA2 (Driver level 27).

Note: OSA-ICC supports up to 48 *secure* sessions per CHPID (the overall maximum of 120 connections is unchanged).

OSA-ICC enhancements with HMC 2.14.1

The following enhancements were introduced with HMC 2.14.1:

- ▶ The IPv6 communications protocol is supported by OSA-ICC 3270 so that clients can comply with regulations that require all computer purchases to support IPv6.
- ▶ TLS negotiation levels (the supported TLS protocol levels) for the OSA-ICC 3270 client connection can now be specified:
 - TLS 1.0 OSA-ICC 3270 server permits TLS 1.0, TLS 1.1 and TLS 1.2 client connections.

- TLS 1.1 OSA-ICC 3270 server permits TLS 1.1 and TLS 1.2 client connections.
- TLS 1.2 OSA-ICC 3270 server permits only TLS 1.2 client connections.
- Separate and unique OSA-ICC 3270 certificates are supported (for each PCHID), for the benefit of customers who host workloads across multiple business units or data centers where cross-site coordination is required. Customers can avoid interruption of all the TLS connections at the same time when they must renew expired certificates. OSA-ICC continues to also support a single certificate for all OSA-ICC PCHIDs in the system.

The supported operating systems are listed in Table 7-8 on page 221 and Table 7-9 on page 223.

Checksum offload for in QDIO mode (CHPID type OSD)

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.

Checksum offload provide checksum offload for several types of traffic and is supported by OSA-Express6S GbE, OSA-Express6S 1000BASE-T Ethernet, OSA-Express5S GbE, OSA-Express5S 1000BASE-T Ethernet, and OSA-Express4S 1000BASE-T Ethernet features when configured as CHPID type OSD (QDIO mode only).

When checksum is offloaded, the OSA-Express feature runs the checksum calculations for Internet Protocol version 4 (IPv4) and Internet Protocol version 6 (IPv6) packets. The checksum offload function applies to packets that go to or come from the LAN.

When multiple IP stacks share an OSA-Express, and an IP stack sends a packet to a next hop address that is owned by another IP stack that is sharing the OSA-Express, OSA-Express sends the IP packet directly to the other IP stack. The packet does not have to be placed out on the LAN, which is termed LPAR-to-LPAR traffic. Checksum offload is enhanced to support the LPAR-to-LPAR traffic, which was not originally available.

The supported operating systems are listed in Table 7-8 on page 221 and Table 7-9 on page 223.

Querying and displaying an OSA configuration

OSA-Express3 introduced the capability for the operating system to query and display directly the current OSA configuration information (similar to OSA/SF). z/OS uses this OSA capability by introducing the TCP/IP operator command **display OSAINFO**. z/VM provides this function with the **NETSTAT OSAINFO TCP/IP** command.

The use of **display OSAINFO** (z/OS) or **NETSTAT OSAINFO** (z/VM) allows the operator to monitor and verify the current OSA configuration and helps improve the overall management, serviceability, and usability of OSA-Express cards.

These commands apply to CHPID type OSD. The supported operating systems are listed in Table 7-8 on page 221.

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.

QDIO data connection isolation is supported by all OSA-Express features on z14 ZR1. The supported operating systems are listed in Table 7-8 on page 221 and Table 7-9 on page 223.

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 on all OSA-Express features on z14 ZR1. The supported operating systems are listed in Table 7-8 on page 221 and Table 7-9 on page 223.

QDIO optimized latency mode

QDIO optimized latency mode (OLM) can help improve performance for applications that feature 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-Express features without needing more program-controlled interrupts (PCIs).
- ▶ For outbound processing, the OSA-Express cards also look more frequently for available data to process from the TCP/IP stack. Therefore, the process does not require a Signal Adapter (SIGA) instruction to determine whether more data is available.

The supported operating systems are listed in Table 7-8 on page 221.

QDIO Diagnostic Synchronization

QDIO Diagnostic Synchronization enables system programmers and network administrators to coordinate and simultaneously capture software and hardware traces. It allows z/OS to signal OSA-Express features (by using a diagnostic assist function) to stop traces and capture the current trace records.

QDIO Diagnostic Synchronization is supported by the OSA-Express features on z14 ZR1 when in QDIO mode (CHPID type OSD). The supported operating systems are listed in Table 7-8 on page 221.

Adapter interruptions for QDIO

Linux on Z and z/VM work together to provide performance improvements by using extensions to the QDIO architecture. First added to z/Architecture with HiperSockets, adapter interruptions provide an efficient, high-performance technique for I/O interruptions to reduce path lengths and processor usage. These reductions are in the host operating system and the adapter (supported OSA-Express cards when CHPID type OSD is used).

In extending the use of adapter interruptions to OSD (QDIO) channels, the processor utilization to handle a traditional I/O interruption is reduced. This configuration benefits OSA-Express TCP/IP support in z/VM, z/VSE, and Linux on Z. The supported operating systems are listed in Table 7-8 on page 221 and Table 7-9 on page 223.

Inbound workload queuing for OSA

OSA-Express3 introduced inbound workload queuing (IWQ), which creates multiple input queues and allows OSA to differentiate workloads “off the wire.” It then assigns work to a specific input queue (per device) to z/OS.

Each input queue is a unique type of workload, and includes 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), which avoids traditional resource contention.

IWQ reduces the conventional z/OS processing that is required to identify and separate unique workloads. This advantage results in improved overall system performance and scalability.

A primary objective of IWQ is to provide improved performance for business-critical interactive workloads by reducing contention that is created by other types of workloads. In a heavily mixed workload environment, this “off the wire” network traffic separation is provided by OSA-Express6S, OSA-Express5S, or OSA-Express4S¹¹ features that are defined as CHPID type OSD. OSA IWQ is shown in Figure 7-5.

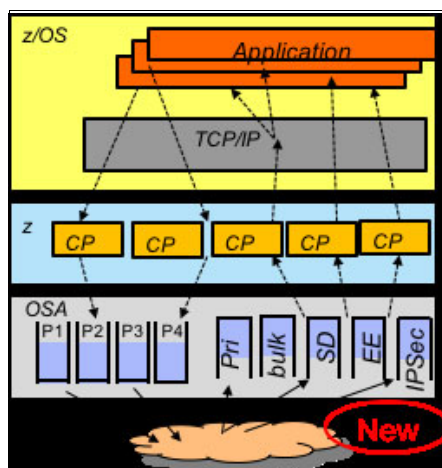


Figure 7-5 OSA inbound workload queuing

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.

- EE (Enterprise Extender / SNA traffic)

IWQ for the OSA-Express features is enhanced to differentiate and separate inbound Enterprise Extender traffic to a dedicated input queue.

¹¹ Only OSA-Express4S GBitEthernet SX and LX cards are supported on z14 ZR1 as carry forward.

The supported operating systems are listed in Table 7-8 on page 221 and Table 7-9 on page 223.

VLAN management enhancements

VLAN management enhancements are valid for supported OSA-Express features on z14 ZR1 defines as CHPID type OSD. The supported operating systems are listed in Table 7-8 on page 221 and Table 7-9 on page 223.

GARP VLAN Registration Protocol

All OSA-Express¹² features support VLAN prioritization, which is a component of the IEEE 802.1 standard. GARP VLAN Registration Protocol (GVRP) support allows an OSA-Express port to register or unregister its VLAN IDs with a GVRP-capable switch and dynamically update its table as the VLANs change. This process simplifies the network administration and management of VLANs because manually entering VLAN IDs at the switch is no longer necessary. The supported operating systems are listed in Table 7-8 on page 221 and Table 7-9 on page 223.

Link aggregation support for z/VM

Link aggregation (IEEE 802.3ad) that is controlled by the z/VM Virtual Switch (VSWITCH) allows the dedication of an OSA-Express^{11,12} 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-Express ports into a single logical link. This configuration increases throughput, and provides nondisruptive failover if a port becomes unavailable. The target links for aggregation must be of the same type.

Link aggregation is applicable to CHPID type OSD (QDIO). The supported operating systems are listed in Table 7-8 on page 221 and Table 7-9 on page 223.

Multi-VSwitch Link Aggregation

Multi-VSwitch Link Aggregation support allows a port group of OSA-Express features to span multiple virtual switches within a single z/VM system or between multiple z/VM systems. Sharing a Link Aggregation Port Group (LAG) with multiple virtual switches increases optimization and utilization of the OSA-Express features when handling larger traffic loads.

Higher adapter utilization protects customer investments, which is increasingly important as 10 GbE deployments become more prevalent. The supported operating systems are listed in Table 7-8 on page 221 and Table 7-9 on page 223.

Large send for IPv6 packets

Large send for IPv6 packets improves performance by offloading outbound TCP segmentation processing from the host to an OSA-Express feature by employing a more efficient memory transfer into it.

Large send support for IPv6 packets applies to the OSA-Express^{11,12} features (CHPID type OSD) on z14 M0x, z14 ZR1, z13, z13s, zEC12, and zBC12.

z13 added support of large send for IPv6 packets (segmentation offloading) for LPAR-to-LPAR traffic. OSA-Express6S on z14 ZR1 added TCP checksum on large send, which reduces the cost (CPU time) of error detection for large send.

The supported operating systems are listed in Table 7-8 on page 221 and Table 7-9 on page 223.

¹² OSA-Express4S or newer.

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.

The blocking algorithm is modified based on the following application requirements:

- ▶ For latency-sensitive applications, the blocking algorithm is modified considering latency.
- ▶ 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.

Supported OSA-Express features adapt to the changes, which avoids thrashing and frequent updates to the OAT. Based on the TCP/IP settings, OSA holds the packets before presenting them to the host. A dynamic setting is designed to avoid or minimize host interrupts.

OSA Dynamic LAN idle is supported by the OSA-Express^{11,12} features on z14 ZR1 when in QDIO mode (CHPID type OSD). The supported operating systems are listed in Table 7-8 on page 221.

OSA Layer 3 virtual MAC for z/OS environments

To help simplify the infrastructure and facilitate load balancing when an LPAR is sharing an OSA MAC address with another LPAR, each operating system instance can have its own unique logical or virtual MAC (VMAC) address. All IP addresses that are associated with a TCP/IP stack are accessible by using their own VMAC address instead of sharing the MAC address of an OSA port. This situation also applies to Layer 3 mode and to an OSA port spanned among channel subsystems.

OSA Layer 3 VMAC is supported by the OSA-Express^{11,12} features on z14 ZR1 when in QDIO mode (CHPID type OSD). The supported operating systems are listed in Table 7-8 on page 221.

Network Traffic Analyzer

The z14 ZR1 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-Express^{11,12} features on z14 ZR1 when in QDIO mode (CHPID type OSD). The supported operating systems are listed in Table 7-8 on page 221.

7.4.6 Cryptography features and functions support

IBM z14 Model ZR1 provides the following major groups of cryptographic functions:

- ▶ Synchronous cryptographic functions, which are provided by CPACF
- ▶ Asynchronous cryptographic functions, which are provided by the Crypto Express6S feature

The minimum software support levels are described in the following sections. Review the current PSP buckets to ensure that the latest support levels are known and included as part of the implementation plan.

CP Assist for Cryptographic Function

Central Processor Assist for Cryptographic Function (CPACF), which is standard¹³ on every z14 ZR1 core, now supports pervasive encryption. Simple policy controls allow business to enable encryption to protect data in mission-critical databases without stopping the database or re-create database objects. Database administrators can use z/OS Dataset Encryption, z/OS Coupling Facility Encryption, z/VM encrypted hypervisor paging, and z/TPF transparent database encryption, which use the performance enhancements in the hardware.

CPACF supports the following features in z14 ZR1:

- ▶ Advanced Encryption Standard (AES, symmetric encryption)
- ▶ Data Encryption Standard (DES, symmetric encryption)
- ▶ Secure Hash Algorithm (SHA, hashing)
- ▶ SHAKE Algorithms
- ▶ True Random Number Generation (TRNG)
- ▶ Improved GCM (Galois Counter Mode) encryption (enabled by a single hardware instruction)

CPACF also is used by several IBM software product offerings for z/OS, such as IBM WebSphere Application Server for z/OS. For more information, see 6.4, “CP Assist for Cryptographic Functions” on page 182.

The supported operating systems are listed in Table 7-10 on page 225 and Table 7-11 on page 226.

Crypto Express6S

Introduced with z14 ZR1, Crypto Express6S complies with the following Physical Security Standards:

- ▶ FIPS 140-2 level 4
- ▶ Common Criteria EP11 EAL4
- ▶ Payment Card Industry (PCI) HSM
- ▶ German Banking Industry Commission (GBIC, formerly DK)

Support of Crypto Express6S functions varies by operating system and release and by the way the card is configured as a coprocessor or an accelerator. For more information, see 6.5, “Crypto Express6S” on page 188. The supported operating systems are listed in Table 7-10 on page 225 and Table 7-11 on page 226.

Crypto Express5S (carry forward on z14 ZR1)

Support of Crypto Express5S functions varies by operating system and release and by the way the card is configured as a coprocessor or an accelerator. The supported operating systems are listed in Table 7-10 on page 225 and Table 7-11 on page 226.

¹³ CPACF hardware is implemented on each z14 core. CPACF functionality is enabled with FC 3863.

Regional Crypto Enablement

Starting with z13 GA2, IBM enabled geo-specific cryptographic support that is supplied by IBM approved vendors. China is the first geography to use this support to meet the cryptography requirements of Chinese clients that are required to comply with the People's Bank of China Financial IC Card Specifications (PBOC 3.0) for payment card processing. When ordered, the Regional Crypto Enablement (RCE) support reserves the I/O slot or slots for the IBM approved vendor-supplied cryptographic card or cards. Clients must contact the IBM approved vendor directly for purchasing information.

RCE is a framework to enable the integration of IBM certified third-party cryptographic hardware for regional or industry encryption requirements. It also supports the use of cryptography algorithms and equipment from selected providers with IBM Z in specific countries. Support for the use of international algorithms (AES, DES, RSA, and ECC) with regional crypto devices (supporting regional algorithms, such as SMx) is added to the ICSF PKCS#11 services.

The supported operating systems are listed in Table 7-10 on page 225 and Table 7-11 on page 226.

Web deliverables

For more information about web-deliverable code on z/OS, see [the z/OS downloads website](#).

For Linux on Z, support is delivered through IBM and the distribution partners. For more information, see [Linux on Z on the IBM developerWorks website](#).

z/OS Integrated Cryptographic Service Facility

Integrated Cryptographic Service Facility (ICSF) is a base component of z/OS. It is designed to transparently use the available cryptographic functions, whether CPACF or Crypto Express, to balance the workload and help address the bandwidth requirements of the applications.

Despite being a z/OS base component, ICSF functions are generally made available through web deliverable support a few months after a new z/OS release. Therefore, new functions are related to an ICSF function modification identifier (FMID) instead of a z/OS version.

ICSF HCR77D0 - Cryptographic Support for z/OS V2R2 and z/OS V2R3

z/OS V2.2 and V2.3 require ICSF Web Deliverable WD18 (HCR77D0) to support the following features:

- ▶ Support for the updated German Banking standard (DK):
 - CCA 5.4 & 6.1¹⁴:
 - ISO-4 PIN Blocks (ISO-9564-1)
 - Directed keys: A key can encrypt or decrypt data, but not both.
 - Allow AES transport keys to be used to export or import *DES* keys in a standard ISO 20038 key block. This feature helps with interoperability between CCA and non-CCA systems.
 - Allow AES transport keys to be used to export or import a small subset of *AES* keys in a standard ISO 20038 key block. This feature helps with interoperability between CCA and non-CCA systems.
 - Triple-length TDES keys with Control Vectors for increased data confidentiality.

¹⁴ CCA 5.4 and 6.1 enhancements are also supported for z/OS V2R1 with ICSF HCR77C1 (WD17) with SPEs (Small Program Enhancements (z/OS continuous delivery model)).

- CCA 6.2: PCI HSM 3K DES: Support for triple length DES keys (standards compliance).
- ▶ EP11 Stage 4:
 - New elliptic curve algorithms for PKCS#11 signature, key derivation operations
 - Ed448 elliptic curve
 - EC25519 elliptic curve
 - EP11 Concurrent Patch Apply: Allows service to be applied to the EP11 coprocessor dynamically without taking the crypto adapter offline (already available for CCA coprocessors).
 - eIDAS compliance: eIDAS: Cross-border EU regulation for portable recognition of electronic identification.

ICSF HCR77C1 - Cryptographic Support for z/OS V2R1 - z/OS V2R3

ICSF Web Deliverable HCR77C1 supports the following features:

- ▶ Usage and administration of Crypto Express6S

This feature might be configured as an accelerator (CEX6A), a CCA coprocessor (CEX6C), or an EP-11 coprocessor (CEX6P).
- ▶ Coprocessor in PCI-HSM Compliance Mode (enablement requires TKE 9.0 or newer).
- ▶ z14 ZR1 CPACF support. For more information, see “CP Assist for Cryptographic Function” on page 266.

The following software enhancements are available in ICSF Web Deliverable HCR77C1 when running on z14 ZR1 server:

- ▶ Crypto Usage Statistics: When enabled, ICSF aggregates statistics that are related to crypto workloads and logs to an SMF record.
- ▶ Panel-based CKDS Administration: ICSF added an ISPF, panel-driven interface that allows interactive administration (View, Create, Modify, and Delete) of CKDS keys.
- ▶ CICS End User Auditing: When enabled, ICSF retrieves the CICS user identity and includes it as a log string in the SAF resource check. The user identity is not checked for access to the resource. Instead, it is included in the resource check (SMF Type 80) records that are logged for any of the ICSF SAF classes protecting crypto keys and services (CSFKEYS, XCSFKEY, CRYPTOZ, and CSFSERV).

For more information about ICSF versions and FMID cross-references, see the abstract [z/OS: ICSF Version and FMID Cross Reference](#), TD103782, which is available at the IBM Techdocs website.

For PTFs that allow previous levels of ICSF to coexist with the Cryptographic Support for z/OS 2.1 - z/OS V2R3 (HCR77C1) web deliverable, check below FIXCAT, as shown in the following example:

```
IBM.Coexistence.ICSF.z/OS_V2R1-V2R3-HCR77C1
```

RMF Support for Crypto Express6

RMF enhances the Monitor I Crypto Activity data gatherer to recognize and use performance data for the new Crypto Express6 (CEX6) card. RMF supports all valid card configurations on z14 ZR1 and provides CEX6 crypto activity data in the SMF type 70 subtype 2 records and RMF Postprocessor Crypto Activity Report.

With October 2nd, 2018 Announcement, reporting can be done at an LPAR/domain level to provide more granular reports for capacity planning and diagnosing problems. This feature requires fix for APAR OA54952.

The supported operating systems are listed in Table 7-10 on page 225.

z/OS Data Set Encryption

Aligned with IBM Z Pervasive Encryption initiative, IBM provides application-transparent, policy-controlled dataset encryption in IBM z/OS.

Policy driven z/OS Data Set Encryption enables users to perform the following tasks:

- ▶ De-couple encryption from data classification; encrypt data automatically independent of labor-intensive data classification work.
- ▶ Encrypt data immediately and efficiently at the time it is written.
- ▶ Reduce risks that are associated with mis-classified or undiscovered sensitive data.
- ▶ Help protect digital assets automatically.
- ▶ Achieve application transparent encryption.

IBM Db2 for z/OS and IBM Information Management System (IMS) intend to use z/OS Data Set Encryption.

With z/OS Data Set Encryption, DFSMS enhances data security with support for data set level encryption by using DFSMS access methods. This function is designed to give users the ability to encrypt their data sets without changing their application programs. DFSMS users can identify which data sets require encryption by using JCL, Data Class, or the RACF data set profile. Data set level encryption can allow the data to remain encrypted during functions, such as backup and restore, migration and recall, and replication.

z/OS Data Set Encryption requires CP Assist for Cryptographic Functions (CPACF). For protected keys, it requires z196 or later Z servers with CEX3 or later. The degree of encryption performance improvement is based on the encryption mode that is used.

Considering the significant enhancements that were introduced with z14, the encryption mode of XTS is used by access method encryption to obtain the best performance possible. It is not recommended to enable z/OS data set encryption until all sharing systems, fallback, backup, and DR systems support encryption.

In addition to applying PTFs enabling the support, ICSF configuration is required. The supported operating systems are listed in Table 7-10 on page 225.

Crypto Analytics Tool for Z

The IBM CAT is an analytics solution that collects data about your z/OS cryptographic infrastructure, presents reports, and analyzes to determine whether vulnerabilities exist. CAT collects cryptographic information from across the enterprise and provides reports to help users better manage the crypto infrastructure and ensure it follows best practices. The use of CAT can help you deal with managing complex cryptography resources across your organization.

z/VM encrypted hypervisor paging (encrypted paging support)

With the PTF for APAR VM65993, z/VM V6.4 supports encrypted paging in support of the z14 ZR1 pervasive encryption philosophy of encrypting all data in flight and at rest. Ciphering occurs as data moves between active memory and a paging volume that is owned by z/VM.

Included in this support is the ability to dynamically control whether a running z/VM system is encrypting this data. This support protects guest paging data from administrators or users with access to volumes. Enabled with AES encryption, z/VM Encrypted Paging includes low overhead by using CPACF.

The supported operating systems are listed in Table 7-10 on page 225.

z/TPF transparent database encryption

Shipped in August 2016, z/TPF at-rest Data Encryption provides the following features and benefits:

- ▶ Automatic encryption of at-rest data by using AES CBC (128 or 256).
- ▶ No application changes required.
- ▶ Database level encryption by using highly efficient CPACF.
- ▶ Inclusion of data on disk and cached in memory.
- ▶ Ability to include data integrity checking (optionally by using SHA-256) to detect accidental or malicious data corruption.
- ▶ Tools to migrate a database from unencrypted to encrypted state or change the encryption key/algorithm for a specific DB while transactions are flowing (no database downtime).

Pervasive encryption for Linux on Z

Pervasive encryption for Linux on Z combines the full power of Linux with z14 ZR1 capabilities by using the support of the following features:

- ▶ Kernel Crypto: z14 ZR1 CPACF
- ▶ LUKS dm-crypt Protected-Key CPACF
- ▶ Libica and openssl: z14 ZR1 CPACF and acceleration of RSA handshakes by using SIMD
- ▶ Secure Service Container: High security virtual appliance deployment infrastructure

Protection of data at-rest

By using the integration of industry-unique hardware accelerated CPACF encryption into the standard Linux components, users can achieve optimized encryption transparently to prevent raw key material from being visible to OS and applications.

Protection of data in-flight

Because of the potential costs and overhead, most of the organizations avoid the use of host-based network encryption today. By using enhanced CPACF and SIMD on z14 ZR1, TLS and IPSec can use hardware performance gains while benefitting from transparent enablement. Reduced cost of encryption enables broad use of network encryption.

7.4.7 Special-purpose features and functions

This section describes the zEnterprise Data Compression Express.

zEnterprise Data Compression Express

The growth of data that must be captured, transferred, and stored for extended periods is unrelenting. Software-implemented compression algorithms are costly in terms of processor resources, and storage costs are not negligible.

zEnterprise Data Compression (zEDC) Express is an optional feature that is available on z14 M0x, z14 ZR1, z13, z13s, zEC12, and zBC12 servers that addresses those requirements by providing hardware-based acceleration for data compression and decompression. zEDC provides data compression with lower CPU consumption than the compression technology that was available on Z servers.

Support for data recovery (decompression) when the zEDC is not installed, or installed but not available on the system, is provided through software on z/OS V2R2, z/OS V2R1, and V1R13 with required PTFs applied. Because software decompression is slow and uses considerable processor resources, it is not recommended for production environments.

zEDC supports QSAM/BSAM (non-VSAM) data set compression by using any of the following methods:

- ▶ Data class level: Two new values, zEDC Required (ZR) and zEDC Preferred (ZP), can be set by using the **COMPACTION** option in the data class.
- ▶ System Level: Two new values, zEDC Required (ZEDC_R) and zEDC Preferred (ZEDC_P), can be specified by using the **COMPRESS** parameter that is found in the IGDSMSXX member of the SYS1.PARMLIB data set.

In z/OS V2R1 and later, SMF can be configured to use zEDC Express for increased throughput of SMF record logging. This change can increase the recording throughput and enable the following functions:

- ▶ Capture extra SMF data that is uncollected because of System Logger constraints. Coupling facility (CF) and storage management subsystem (SMS) direct access storage device (DASD) are examples of such constraints.
- ▶ Mitigate z/OS image growth because of consolidation, new workloads, or growing workloads, which cause more SMF data to be generated.

Data class takes precedence over system level. The supported operating systems are listed in Table 7-12 on page 227.

For more information about zEDC Express, see Appendix F, “IBM zEnterprise Data Compression Express” on page 461, and *Reduce Storage Occupancy and Increase Operations Efficiency with IBM zEnterprise Data Compression*, SG24-8259.

7.5 z/OS migration considerations

Except for base processor support, z/OS releases do not require any of the functions that are introduced with the z14. Minimal toleration support that is needed depends on z/OS release.

Although z14 ZR1 servers do *not* require any “functional” software, it is recommended to install all z14 ZR1 services before upgrading to the new server. The support matrix for z/OS releases and the Z servers that support them are listed in Table 7-16.

Table 7-16 z/OS support summary

z/OS Release	z9 EC z9 BC WDFM ^a	z10 EC z10 BC WDFM ^a	z196 z114 WDFM ^a	zEC12 zBC12 WDFM ^a	z13 z13s	z14 (M0x and ZR1)	End of Service	Extended Defect Support ^b
V1R13	X	X	X	X	X	X ^b	09/2016	09/2019 ^c
V2R1	X	X	X	X	X	X	09/2018 ^d	09/2021 ^c

z/OS Release	z9 EC z9 BC WDFM ^a	z10 EC z10 BC WDFM ^a	z196 z114 WDFM ^a	zEC12 zBC12 WDFM ^a	z13 z13s	z14 (M0x and ZR1)	End of Service	Extended Defect Support ^b
V2R2		X	X	X	X	X	09/2020 ^c	09/2023 ^c
V2R3				X	X	X	09/2022 ^c	09/2025 ^c

a. Server was withdrawn from marketing.

b. The IBM Software Support Services for z/OS V1.13, which was offered as of October 1, 2016, provides the ability for customers to purchase extended defect support service for z/OS V1.13.

c. Planned. All statements regarding IBM's plans, directions, and intent are subject to change or withdrawal without notice.

d. The IBM Software Support Services for z/OS V2.1 provides the ability for customers to purchase extended defect support service for z/OS V2.1.

7.5.1 General guidelines

The IBM z14 Model ZR1 introduces the latest IBM Z technology. Although support is provided by z/OS starting with z/OS V1R13, the capabilities and use of z14 ZR1 depends on the z/OS release. Also, web deliverables¹⁵ are needed for some functions on some releases. In general, consider the following guidelines:

- ▶ Do not change software releases and hardware at the same time.
- ▶ Keep members of the sysplex at the same software level, except during brief migration periods.
- ▶ Migrate to an STP-only network before introducing a z14 ZR1 into a sysplex.
- ▶ More planning consideration is required for z14 ZR1 because it does not support InfiniBand coupling features.
- ▶ Review any restrictions and migration considerations before creating an upgrade plan.
- ▶ Acknowledge that some hardware features cannot be ordered or carried forward for an upgrade from an earlier server to z14 ZR1 and plan accordingly.
- ▶ Determine the changes in IOCP, HCD, and HCM to support defining z14 ZR1 configuration and the new features and functions it introduces.
- ▶ Ensure that none of the new z/Architecture Machine Instructions (mnemonics) that were introduced with z14 ZR1 are colliding with the names of Assembler macro instructions you use¹⁶.
- ▶ Check the use of **MACHMIG** statements in **LOADxx PARMLIB** commands.

7.5.2 Hardware Fix Categories (FIXCATs)

Base support includes fixes that are required to run z/OS on the IBM z14 Model ZR1 server. They are identified by:

IBM.Device.Server.z14ZR1-3907.RequiredService

¹⁵ For example, the use of Crypto Express6S requires the Cryptographic Support for z/OS V2R1 - z/OS V2R3 web deliverable.

¹⁶ For more information, see the [Tool to Compare IBM z14 Instruction Mnemonics with Macro Libraries](#) IBM technote.

The exploitation of many functions covers fixes that are required to use the capabilities of the IBM z14™ server. They are identified by:

IBM.Device.Server.z14ZR1-3907.Exploitation

Recommended service is identified by:

IBM.Device.Server.z14ZR1-3907.RecommendedService

Support for z14 ZR1 is provided by using a combination of web deliverables and PTFs, which are documented in PSP Bucket Upgrade = 3907DEVICE, Subset = 3907/ZOS.

Consider the following other Fix Categories of Interest:

- Fixes that are required to use Parallel Sysplex InfiniBand Coupling links:

IBM.Function.ParallelSysplexInfiniBandCoupling

- Fixes that are required to use the Server Time Protocol function:

IBM.Function.ServerTimeProtocol

- Fixes that are required to use the High-Performance FICON function:

IBM.Function.zHighPerformanceFICON

- PTFs that allow previous levels of ICSF to coexist with the latest Cryptographic Support for z/OS 2.2 - z/OS V2R3 (HCR77D0) web deliverable:

IBM.Coexistence.ICSF.z/OS_V2R2-V2R3-HCR77D0

- PTFs that allow previous levels of ICSF to coexist with the Cryptographic Support for z/OS 2.1 - z/OS V2R3 (HCR77C1) web deliverable:

IBM.Coexistence.ICSF.z/OS_V2R1-V2R3-HCR77C1

Use the SMP/E **REPORT MISSINGFIX** command to determine whether any FIXCAT APARs exist that are applicable and are not yet installed, and whether any SYSMODs are available to satisfy the missing FIXCAT APARs.

For more information about IBM Fix Category Values and Descriptions, see the [IBM Fix Category Values and Descriptions page](#) of the IBM IT infrastructure website.

7.5.3 Coupling links

z14 ZR1 servers support only active participation in the same Parallel Sysplex with z13, z13s, and z14 M0x. Configurations with z/OS on one of these servers can add a z14 ZR1 server to their Sysplex for a z/OS or a Coupling Facility image.

Configurations with a Coupling Facility on one of these servers can add a z14 ZR1 server to their Sysplex for a z/OS or a Coupling Facility image. z14 ZR1 does not support participating in a Parallel Sysplex with System zEC12, zBC12, and earlier systems.

Each system can use, or not use, internal coupling links, Coupling Express Long Reach (CE LR) coupling links, or ICA coupling links independently of what other systems are using.

Coupling connectivity is available only when other systems also support the same type of coupling. For more information about supported coupling link technologies on z14 ZR1, see 4.6.4, “Parallel Sysplex connectivity” on page 153, and the [Coupling Facility Configuration Options](#) white paper.

7.5.4 z/OS XL C/C++ considerations

z/OS V2R1 with PTFs or higher and z14 are required to run code at the latest level (12) for the following C/C++ compiler options:

- ▶ **ARCHITECTURE**: This option selects the minimum level of system architecture on which the program can run. Certain features that are provided by the compiler require a minimum architecture level. **ARCH(12)** uses instructions that are available on the z14 ZR1.
- ▶ **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.

The following new functions provide performance improvements for applications by using new z14 ZR1 instructions:

- ▶ Vector Programming Enhancements
- ▶ New z14 ZR1 hardware instruction support
- ▶ Packed Decimal support that uses vector registers
- ▶ Auto-SIMD enhancements to use new data types

To enable the use of new functions, specify **ARCH(12)** and **VECTOR** for compilation. The binaries that are produced by the compiler on z14 ZR1 can be run on z14 M0x and z14 ZR1 only because it uses the vector facility on z14 ZR1 for new functions. The use of older versions of the compiler on z14 ZR1 do not enable new functions.

For more information about the **ARCHITECTURE**, **TUNE**, and **VECTOR** compiler options, see *z/OS V2R2.0 XL C/C++ User's Guide*, SC09-4767.

Important: Use the previous **Z ARCHITECTURE** or **TUNE** options for C/C++ programs if the same applications run on the z14 ZR1 and on previous IBM Z servers. However, if C/C++ applications run on z14 ZR1 servers only, use the latest **ARCHITECTURE** and **TUNE** options to ensure that the best performance possible is delivered through the latest instruction set additions.

For more information, see *Migration from z/OS V2R1 to z/OS V2R2*, GA32-0889.

7.5.5 z/OS V2.3

IBM announced z/OS Version 2 Release 3 - Engine for digital transformation through Announcement letter 217-246 on July 17, 2017. Focusing on three critical areas (Security, Simplification, and Cloud), z/OS V2.3 provides a simple and transparent approach to enable extensive encryption of data and to simplify the overall management of the z/OS system to increase productivity. Focus is also given to providing a simple approach for self-service provisioning and rapid delivery of software as a service, while enabling for the API economy.

Consider the following points before migrating z/OS 2.3 to IBM z14 Model ZR1:

- ▶ IBM z/OS V2.3 with z14 ZR1 requires a minimum of 8 GB of memory. When running as a z/VM guest or on an IBM System z Personal Development Tool, a minimum of 2 GB is required for z/OS V2.3. If the minimum is not met, a warning WTOR is issued at IPL.

Continuing with less than the minimum memory might affect availability. A migration health check will be introduced at z/OS V2.1 and z/OS V2.2 to warn if the system is configured with less than 8 GB.

- ▶ Dynamic splitting and merging of Coordinated Timing Network (CTN) is available with z14 ZR1.

- ▶ The z/OS V2.3 real storage manager (RSM) is planned to support a new asynchronous memory clear operation to clear the data from 1M page frames by using I/O processors (SAPs) on next generation processors. The new asynchronous memory clear operation eliminates the CPU cost for this operation and help improve performance of RSM first reference page fault processing and system services, such as IARV64 and STORAGE OBTAIN.
- ▶ RMF support is provided to collect SMC-D related performance measurements in SMF 73 Channel Path Activity and SMF 74 subtype 9 PCIE Activity records. It also provides these measurements in the RMF Postprocessor and Monitor III PCIE and Channel Activity reports. This support is also available on z/OS V2.2 with PTF UA80445 for APAR OA49113.
- ▶ HyperSwap support is enhanced to allow RESERVE processing. When a system runs a request to swap to secondary devices that are managed by HyperSwap, z/OS detects when RESERVEs are held and ensures that the devices that are swapped also hold the RESERVE. This enhancement is provided with collaboration from z/OS, GDPS HyperSwap, and CSM HyperSwap.

7.6 z/VM migration considerations

IBM z14 ZR1 supports z/VM 7.1 and z/VM 6.4. z/VM is moving to continuous delivery model. For more information, see [this web page](#).

7.6.1 z/VM 7.1

z/VM 7.1 can be installed directly on IBM z14 ZR1. z/VM V7R1 includes the following new features:

- ▶ Includes Single System Image and Live Guest Relocation in the base. In z/VM 6.4, this feature was the VMSSI-priced feature.
- ▶ Enhances the dump process to reduce the time that is required to create and process dumps.
- ▶ Upgrades to a new Architecture Level Set. This upgrade requires an IBM zEnterprise EC12 or BC12, or later.
- ▶ Provides the base for more functionality to be delivered as service Small Program Enhancements (SPE) after general availability.

z/VM 7.1 includes SPEs shipped for z/VM 6.4, including Virtual Switch Enhanced Load Balancing, DS8K z-Thin Provisioning, and Encrypted Paging.

7.6.2 z/VM 6.4

z/VM V6.4 can be installed on a z14 ZR1 server with an image that is obtained from IBM after August 25, 2017. The PTF for APAR VM65942 must be applied immediately after installing z/VM V6.4 and before configuring any part of the new z/VM system.

A z/VM Release Status Summary is listed in Table 7-17.

Table 7-17 z/VM Release Status Summary

z/VM Level ^a	General Availability	End of Marketing	End of Service	Minimum Processor Level	Maximum Processor Level
7.1	September, 2018	Not announced	Not announced	zEC12 & zBC12	-
6.4	November, 2016	Not announced	Not announced	z196 & z114	-

a. Older z/VM versions (6.3, 6.2, 5.4 are End Of Support)

7.6.3 ESA/390-compatibility mode for guests

IBM z14 Model ZR1 no longer supports the full ESA/390 architectural mode (the z14 ZR1 does not support ESA/390 mode IPL). However, IBM z14 Model ZR1 does provide ESA/390-compatibility mode, which is an environment that supports a subset of DAT-off ESA/390 applications in a hybrid architectural mode.

z/VM provides the support necessary for DAT-off guests to run in this new compatibility mode. This support allows guests, such as CMS, GCS, and those guests that start in ESA/390 mode briefly before switching to z/Architecture mode to continue to run on IBM z14 Model ZR1.

The available PTF for APAR VM65976 provides infrastructure support for ESA/390 compatibility mode within z/VM V6.4. It must be installed on all members of an SSI cluster before any z/VM V6.4 member of the cluster is run on an IBM z14 Model ZR1 server.

In addition to OS support, all the stand-alone utilities a client uses must be at a minimum level or need a PTF.

7.6.4 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 might want to adjust the number to accommodate the processor unit (PU) capacity of z14 ZR1 servers.

7.7 z/VSE migration considerations

As described in “z/VSE” on page 212, IBM z14 ZR1 supports z/VSE 6.2, z/VSE 6.1, z/VSE 5.2, and z/VSE 5.1¹⁷.

Consider the following general guidelines when you are migrating z/VSE environment to z14 ZR1 servers:

- Collect reference information before migration

This information includes baseline data that reflects the status of, for example, performance data, CPU utilization of reference workload, I/O activity, and elapsed times.

¹⁷ z/VSE 5.1 is end of support since June 2016. It can be IPL'ed on z14 after applying APAR DY47654 (PTF UD54170).

This information is required to size z14 ZR1 and is the only way to compare workload characteristics after migration.

For more information, see the *z/VSE Release and Hardware Upgrade* document.

- ▶ Apply required maintenance for z14 ZR1

Review the Preventive Service Planning (PSP) bucket 3907DEVICE for z14 ZR1 and apply the required PTFs for IBM and independent software vendor (ISV) products.

Note: z14 ZR1 supports z/Architecture mode only.

7.8 Software licensing

The IBM z14 Model ZR1 software portfolio includes operating system software (that is, z/OS, z/VM, z/VSE, and z/TPF) and middleware that runs on these operating systems. The portfolio also includes middleware for Linux on Z environments.

For the z14 ZR1, the following metric groups for software licensing are available from IBM, depending on the software product:

- ▶ Monthly license charge (MLC)

MLC pricing metrics feature a recurring charge that applies each month. In addition to the permission to use the product, the charge includes access to IBM product support during the support period. MLC pricing applies to z/OS, z/VSE, and z/TPF operating systems. Charges are based on processor capacity, which is measured in millions of service units (MSU) per hour.

- ▶ IPLA

IPLA metrics have a single, up-front charge for an entitlement to use the product. An optional and separate annual charge (called *subscription and support*) entitles clients to access IBM product support during the support period. With this option, you can also receive future releases and versions at no extra charge.

Software licensing references

For more information about software licensing, see the following resources:

- ▶ [Learn about Software licensing](#)
- ▶ [Base license agreements](#)
- ▶ [IBM Z Software Pricing reference guide](#)
- ▶ [IBM Z Software Pricing](#)
- ▶ [The IBM International Passport Advantage® Agreement](#) can be downloaded from the [Learn about Software licensing website](#).

Subcapacity license charges

For eligible programs, subcapacity licensing allows software charges that are based on the measured utilization by logical partitions instead of the total number of MSUs of the CPC. Subcapacity licensing removes the dependency between the software charges and CPC (hardware) installed capacity.

The subcapacity licensed products are charged monthly based on the highest observed 4-hour rolling average utilization of the logical partitions in which the product runs. The exception is products that are licensed by using the Select Application License Charge (SALC) pricing metric. This type of charge requires measuring the utilization and reporting it to IBM.

The 4-hour rolling average utilization of the logical partition can be limited by a defined capacity value on the image profile of the partition. This value activates the soft capping function of the PR/SM, which limits the 4-hour rolling average partition utilization to the defined capacity value. Soft capping controls the maximum 4-hour rolling average usage (the last 4-hour average value at every 5-minute interval), but does not control the maximum instantaneous partition use.

You can also use an LPAR group capacity limit, which sets soft capping by PR/SM for a group of logical partitions that are running z/OS.

Even by using the soft capping option, the use of the partition can reach up to 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, which allows utilization peaks above the defined capacity value.

Some pricing metrics apply to stand-alone Z servers. Others apply to the aggregation of multiple Z server workloads within the same Parallel Sysplex.

For more information about WLC and how to combine logical partition utilization, see *z/OS Planning for Sub-Capacity Pricing*, SA23-2301.

Key MLC metrics and offerings

MLC metrics include various offerings. The following metrics and pricing schemes are available. Offerings often are tied to or made available to only on certain Z servers:

- ▶ Key MLC Metrics:
 - WLC (Workload License Charges)
 - AWLC (Advanced Workload License Charges)
 - CMLC (Country Multiplex License Charges)
 - VWLC (Variable Workload License Charges)
 - FWLC (Flat Workload License Charges)
 - AEWLC (Advanced Entry Workload License Charges)
 - EWLC (Entry Workload License Charges)
 - TWLC (Tiered Workload License Charges)
 - zNALC (System z New Application License Charges)
 - PSLC (Parallel Sysplex License Charges)
 - MWLC (Midrange Workload License Charges)
 - zELC (zSeries Entry License Charges)
 - GOLC (Growth Opportunity License Charges)
 - SALC (Select Application License Charges)
- ▶ Pricing:
 - GSSP (Getting Started Sub-Capacity Pricing)
 - IWP (Integrated Workload Pricing)
 - MWP (Mobile Workload Pricing)
 - zCAP (Z Collocated Application Pricing)
 - Parallel Sysplex Aggregated Pricing
 - CMP (Country Multiplex Pricing)
 - ULC (IBM S/390® Usage Pricing)

One of the recent changes in software licensing for z/OS and z/VSE is Multi-Version Measurement (MVM), which replaced Single Version Charging (SVC), Migration Pricing Option (MPO), and the IPLA Migration Grace Period.

MVM for z/OS and z/VSE removes time limits for running multiple eligible versions of a software program. Clients can run different versions of a program simultaneously for an unlimited duration during a program version upgrade.

Clients can also choose to run multiple different versions of a program simultaneously for an unlimited duration in a production environment. MVM allows clients to selectively deploy new software versions, which provides more flexible control over their program upgrade cycles. For more information, see *Software Announcement 217-093*, dated February 14, 2017.

Technology transition offerings with z14 ZR1

Complementing the announcement of the z14 ZR1 server, IBM introduced the following Technology Transition Offerings (TTOs):

- ▶ Technology Update Pricing for the IBM z14 Model ZR1.
- ▶ New and revised Transition Charges for Sysplexes or Multiplexes TTOs for actively coupled Parallel Sysplexes (z/OS), Loosely Coupled Complexes (z/TPF), and Multiplexes (z/OS and z/TPF).

Technology Update Pricing for the IBM z14 Model ZR1 extends the software price and performance that is provided by AWLC and CMLC for z14 ZR1 servers. The new and revised Transition Charges for Sysplexes or Multiplexes offerings provide a transition to Technology Update Pricing for the IBM z14 Model ZR1 for customers who did not yet fully migrate to z14 ZR1 servers. This transition ensures that aggregation benefits are maintained and also phases in the benefits of Technology Update Pricing for the IBM z14 Model ZR1 pricing as customers migrate.

When a z14 ZR1 server is in an actively coupled Parallel Sysplex or a Loosely Coupled Complex, you might choose aggregated Advanced Workload License Charges (AWLC) pricing or aggregated Parallel Sysplex License Charges (PSLC) pricing (subject to all applicable terms and conditions).

When a z14 ZR1 server is part of a Multiplex under Country Multiplex Pricing (CMP) terms, Country Multiplex License Charges (CMLC), Multiplex zNALC (MzNALC), and Flat Workload License Charges (FWLC) are the only pricing metrics available (subject to all applicable terms and conditions).

For more information about software pricing for the z14 ZR1 server, see *Software Announcement 217-273*, dated July 17, 2017, *Technology Transition Offerings for the IBM z14 Model ZR1* offer price-performance advantages.

When a z14 ZR1 server is running z/VSE, you can choose Mid-Range Workload License Charges (MWLC) (subject to all applicable terms and conditions).

For more information about AWLC, CMLC, MzNALC, PSLC, MWLC, or the Technology Update Pricing and Transition Charges for Sysplexes or Multiplexes TTO offerings, see the [IBM z Systems Software Pricing page](#) of the IBM IT infrastructure website.

7.9 References

For more information about planning, see the home pages for each of the following operating systems:

- ▶ [z/OS](#)
- ▶ [z/VM](#)
- ▶ [z/VSE](#)
- ▶ [z/TPF](#)
- ▶ [Linux on Z](#)
- ▶ [KVM for IBM Z](#)



System upgrades

This chapter provides an overview of z14 ZR1 upgrade capabilities and procedures, with an emphasis on capacity on demand (CoD) offerings. The upgrade offerings to the z14 ZR1 systems were developed from previous IBM Z servers.

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 more information about all aspects of system upgrades, see the [IBM Resource Link website](#) (login required). At the website, click **Resource Link** → **Client Initiated Upgrade Information**, and then, select **Education**. Select your product from the list of available systems.

The growth capabilities that are provided by the z14 ZR1 servers include the following benefits:

- ▶ Enabling the use of new business opportunities
- ▶ Supporting the growth of dynamic, smart, and cloud environments
- ▶ Managing the risk of volatile, high-growth, and high-volume applications
- ▶ Supporting 24 x 7 application availability
- ▶ Enabling capacity growth during lockdown periods
- ▶ Enabling planned-downtime changes without availability effects

Naming: The IBM z14 server generation is available as the following machine types and models:

- ▶ Machine Type 3906 (M/T 3906), Models M01, M02, M03, M04, and M05, which is further identified as *IBM z14 Model M0x*, or *z14 M0x*, unless otherwise specified.
- ▶ Machine Type 3907 (M/T 3907), Model ZR1, which is further identified as *IBM z14 Model ZR1*, or *z14 ZR1*, unless otherwise specified.

In the remainder of this chapter, IBM z14 (z14) refers to both machine types.

This chapter includes the following topics:

- ▶ 8.1, “Upgrade types” on page 283
- ▶ 8.2, “Concurrent upgrades” on page 287
- ▶ 8.3, “Miscellaneous equipment specification upgrades” on page 294
- ▶ 8.4, “Permanent upgrade through the CIU facility” on page 298
- ▶ 8.5, “On/Off Capacity on Demand” on page 302
- ▶ 8.6, “Capacity for Planned Event” on page 314
- ▶ 8.7, “Capacity Backup” on page 316
- ▶ 8.8, “Nondisruptive upgrades” on page 320
- ▶ 8.9, “Summary of Capacity on-Demand offerings” on page 325

8.1 Upgrade types

The types of upgrades for a z14 ZR1 server are described in this section.

8.1.1 Overview of upgrade types

Upgrades can be categorized as described in this section.

Permanent and temporary upgrades

Permanent and temporary upgrades are different types of upgrades that can be used in different situations. For example, a growing workload might require more memory, I/O cards, or 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. IBM z14 Model ZR1 servers offer the following solutions for such situations:

- Permanent upgrades:

- Miscellaneous equipment specification (MES)

The MES upgrade order is always performed by IBM personnel. The result can be 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)

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 supports only LICCC upgrades.

- Temporary upgrade

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 8.1.2, “CoD for z14 ZR1 systems terminology” on page 284.

Tip: An MES provides system upgrades that can result in more enabled PUs, a different central processor (CP) capacity level, in more PU SCMs in the drawer, memory, PCIe+ I/O drawers, and I/O features (physical upgrade). Extra planning tasks are required for nondisruptive logical upgrades. An MES is ordered through your IBM representative and installed by IBM service support representatives (IBM SSRs).

Concurrent and nondisruptive upgrades

Depending on the affect 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 turned off during the upgrade. For more information, see 8.2, “Concurrent upgrades” on page 287.

- Non-concurrent

This type of upgrade requires turning off the hardware that is being upgraded. Examples include CPC feature upgrades from a z14 ZR1Max4 to a z14 ZR1 Max12, and physical memory capacity upgrades.

► Disruptive

An upgrade is considered *disruptive* when resources that are modified or added to an operating system image require that the operating system be restarted 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 8.8, “Nondisruptive upgrades” on page 320.

8.1.2 CoD for z14 ZR1 systems terminology

The most frequently used terms that are related to CoD for z14 ZR1 systems are listed in Table 8-1.

Table 8-1 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 (expected or unexpected). The one billable offering that is available is On/Off Capacity on Demand (OOCOD).
Capacity	Hardware resources (processor and memory) that can 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 systems between data centers, upgrades, and other routine management tasks. CPE is an offering of CoD.
Capacity levels	Can be full capacity or subcapacity. For the z14 ZR1 system, capacity levels for the CP engines are A - Z (26 subcapacity levels): <ul style="list-style-type: none"> ► A full capacity CP engine is indicated by Z. ► A subcapacity CP engine is indicated by A - Y.
Capacity setting	Derived from the capacity level and the number of processors. For the z14 ZR1 CPC, the capacity levels are A01 - Z06, where the last digit indicates the number of active CPs, and the letter A - Z indicates the processor capacity level. An all IFL or all ICF system has a capacity setting of A00.
Customer Initiated Upgrade (CIU)	A web-based facility in which 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 z14 ZR1 systems.
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.

Term	Description
Full capacity CP feature	For z14 ZR1 servers, capacity settings Z0n 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 installed on the central processor complex (CPC). A maximum of eight different records can be concurrently installed and active.
Model capacity identifier (MCI)	Shows the current active capacity on the server, including all replacement and billable capacity. For z14 ZR1 servers, the model capacity identifier is in the form of A0x - Z0x, where x indicates the number of active CPs (xx can have a range of 1 - 6).
Model Permanent Capacity Identifier (MPCI)	Keeps information about the capacity settings that are 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, MTCI equals the MPCI.
On/Off Capacity on Demand (CoD)	Represents a function that allows 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.
Features on Demand (FoD)	FoD is a new centralized way to entitle flexibly features and functions on the system. On z196 and z114, the HWMs are stored in the processor and memory LICCC record. On z14, z14 ZR1, z13, z13s, zBC12 and zEC12 servers, the HWMs are stored in the FoD 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 that is 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.
Processor Unit (PU)	A characterizable core.
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 that is 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 Systems technical support website that provides a comprehensive set of tools and resources (login required).
Secondary approval	An option that is selected by the client that requires second approver control for each CoD order. When a secondary approval is required, the request is sent for approval or cancellation to the Resource Link secondary user ID.
Staged record	The point when a record that represents a temporary or permanent capacity upgrade is retrieved and loaded on the SE disk.
Subcapacity	For z14 ZR1 servers, CP features A01 to Y06 represent subcapacity configurations, and CP features Z01 to Z06 represent full capacity configurations.
Temporary capacity	An optional capacity that is added to the current system capacity for a limited amount of time. It can be capacity that is owned or not owned by the client.

Term	Description
Vital product data (VPD)	Information that uniquely defines system, hardware, software, and microcode elements of a processing system.

8.1.3 Permanent upgrades

Permanent upgrades can be obtained by using the following 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 is enabled through the permanent upgrade authorization feature code (FC 9898).

Permanent upgrades that are ordered through an IBM representative

Through a permanent upgrade, you can accomplish the following tasks:

- ▶ Add PU SCMs to the CPC drawer
- ▶ 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)

Consideration: Most of the MESs can be concurrently applied without disrupting the workload. For more information, see 8.2, “Concurrent upgrades” on page 287. However, certain MES changes are non-concurrent; for example, CPC feature upgrades such as from a z14 ZR1 Max4 to a z14 ZR1 Max12/Max24/Max30.

Memory upgrades are only concurrent when the required memory capacity is already physical available (for example, by way of Plan Ahead Memory) and can be activated through LICCC.

Permanent upgrades by using CIU on the IBM Resource Link

Ordering the following permanent upgrades by using the CIU application through Resource Link allows you to add capacity to fit within your hardware:

- ▶ Add model capacity
- ▶ Add specialty engines
- ▶ Add memory
- ▶ Activate unassigned model capacity or IFLs
- ▶ Deactivate activated model capacity or IFLs

8.1.4 Temporary upgrades

z14 ZR1 servers offer the following types of temporary upgrades:

- ▶ On/Off Capacity on Demand (On/Off CoD)

This offering allows you to temporarily add capacity or specialty engines to cover seasonal activities, period-end requirements, peaks in workload, or application testing. This temporary upgrade can be ordered by using the CIU application through Resource Link only.

- ▶ CBU

This offering allows you to replace model capacity or specialty engines in a backup system that is used in an unforeseen loss of system capacity because of a disaster.

- ▶ 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 daily.

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 replacement capacity offerings are available:

- ▶ Capacity Backup
- ▶ Capacity for Planned Event

8.2 Concurrent upgrades

Concurrent upgrades on z14 ZR1 servers can provide more capacity with no system outage. In most cases, a concurrent upgrade can be nondisruptive to the operating system with planning and operating system support.

The concurrent capacity growth capabilities that are provided by z14 ZR1 servers include, but are not limited to, the following benefits:

- ▶ Enabling the meeting of new business opportunities
- ▶ Supporting the growth of smart and cloud environments
- ▶ Managing the risk of volatile, high-growth, and high-volume applications
- ▶ Supporting 24 x 7 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.

Subcapacity models provide for a CP capacity increase in two dimensions that can be used together to deliver configuration granularity. The first dimension is by adding CPs to the configuration, and the second dimension is by changing the capacity setting of the CPs currently installed to a higher MCI. In addition, a capacity increase can be delivered by increasing the CP capacity setting, and at the same time decreasing the number of active CPs.

Consideration: A z14 ZR1 Max4 has a maximum of four PUs available, so it can concurrently be upgraded to models A04 - Z04. If a capacity setting with more than four CPs is required or the combination of CPs and specialty engines exceeds four, a non-concurrent upgrade to a CPC drawer Max12 feature (FC 0637)^a is required.

a. If more specialty engines are needed, Max24 or Max30 feature might be required.

z14 ZR1 servers allow 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. This addition is made by using one of the following methods:

- ▶ With planning ahead for the future need of extra processors. Reserved processors can be specified in the LPAR's profile. 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 initial program load (IPL).
- ▶ Another (easier) way is to enable the dynamic addition of processors through the z/OS LOADxx member. Set the **DYNCPADD** parameter in member LOADxx to ENABLE. z14 ZR1 servers support dynamic processor addition in the same way that the z14 M0x, z13, z13s, zEC12, and zBC12 support it. The operating system must be z/OS V1R10 or later.

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.

8.2.1 Upgrades

z14 ZR1 servers feature the following machine type and model, CPC drawer features, and model capacity identifiers:

- ▶ Machine type and model 3907-ZR1

The 3907-ZR1 is available in the following CPC drawer features:

- Feature Max4 (one PU SCM installed) can have a maximum of 4 PUs for client characterization.
- Feature Max12 (two PU SCMs) can have a maximum of 12 PUs
- Feature Max24 (four PU SCMs) can have a maximum of 24 PUs
- Feature Max30 (four PU SCMs) can have a maximum of 30 PUs
- ▶ Model capacity identifiers (MCI) are A01 to Z06¹. The MCI described how many CPs are characterized (01 - 06) and the capacity setting (A to Z) of the CPs).

¹ or an all IFL or all ICF system the capacity identifier is A00.

A hardware configuration upgrade always requires more physical hardware (PU SCMs, PCIe+ I/O drawers, or both). A system upgrade can change either (or both) of the system model and the MCI.

Consider the following points regarding upgrades:

- ▶ LICCC upgrade:
 - Can add memory or Virtual Flash Memory (VFM) up to the amount physically installed
 - Can change the model capacity identifier, the capacity setting, or both
- ▶ Hardware installation upgrade:
 - Can change the CPC drawer feature by adding one or more PU SCMs
 - Can change the model capacity identifier, the capacity setting, or both
 - Can add physical memory, PCIe+ I/O drawers, and other hardware features

The model capacity identifier can be concurrently changed. Concurrent upgrades can be performed for permanent and temporary upgrades.

CPC drawer feature upgrades: All upgrades from a CPC feature to another CPC feature are disruptive because the extra PU SCMs must be physically installed or changed (for example, upgrade from Max24 to Max30).

Licensed Internal Code upgrades (MES ordered)

The LICCC provides for system upgrades without hardware changes by activating extra (previously installed) unused capacity. Concurrent upgrades through LICCC can be performed for the following resources:

- ▶ Processors, such as CPs, ICFs, IBM z Integrated Information Processors (zIIPs), IFLs, and SAPs, if unused PUs are available in the CPC drawer, or if the model capacity identifier for the CPs can be increased.
- ▶ Memory and VFM, when unused capacity is available on the installed memory cards. Plan-ahead memory is available to give you better control over future memory upgrades. For more information, see 2.5.6, “Preplanned memory” on page 45.

Concurrent hardware installation upgrades (MES ordered)

Configuration upgrades can be concurrent when installing the following resources:

- ▶ ICA SR or PCIe fanout cards
- ▶ I/O cards, when slots are available on the installed PCIe+ I/O drawers
- ▶ PCIe+ I/O drawers, when fanout slots are available in the CPC drawer

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)

z14 ZR1 servers support concurrent conversion between all PU types, which includes SAPs, to provide flexibility to meet changing business requirements.

Important: The LICCC-based PU conversions require that at least one PU (CP, ICF, or IFL), remains unchanged. Otherwise, the conversion is disruptive. The PU conversion generates an 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 also might have to free the PUs to be converted. The operating systems must include 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 points 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.

8.2.2 Customer Initiated Upgrade facility

The Customer Initiated Upgrade (CIU) facility is an IBM online system through which you can order, download, and install permanent and temporary upgrades for IBM Z servers. 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 z14 ZR1 servers at any time, often it is added when ordering a z14 ZR1 server. The CIU facility is controlled through the permanent upgrade authorization feature code, FC 9898.

After you place an order through the CIU facility, you receive a notice that the order is ready for download. You can then download and apply the upgrade by using functions that are available through the Hardware Management Console (HMC), along with the RSF. After all of 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 other capacity that is required for an upgrade must be previously installed. Extra processor drawers 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, zIIPs, IFLs, and unassigned IFLs cannot exceed the client (characterized) PU count of the installed processor drawers. The total number of zIIPs can be twice 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](#) (login required).

As part of the setup, 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](#) (login required) and start the CIU application to upgrade a system for processors or memory. You can request a client order approval to conform to your operational policies. You also can allow the definition of more IDs to be authorized to access the CIU. More IDs can be authorized to enter or approve CIU orders, or only view orders.

Permanent upgrades

Permanent upgrades can be ordered by using the CIU facility. Through the CIU facility, you can generate online permanent upgrade orders to concurrently add processors (CPs, ICFs, zIIPs, IFLs, and SAPs) and memory, or change the model capacity identifier. These upgrades and changes are available up to the limits of the installed processor drawers on a system.

Temporary upgrades

The base model z14 ZR1 server describes permanent and dormant capacity by using the capacity marker and the number of PU features that are installed on the system. Up to eight temporary offerings can be present. Each offering includes 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, only one On/Off CoD offering can be active at any time if enough resources are available to fulfill the offering specifications.

Temporary upgrades are represented in the system by a *record*. All temporary upgrade records are on the SE hard disk drive (HDD). The records can be downloaded from the RSF or installed from portable media. At the time of activation, you can control everything locally.

The provisioning architecture is shown in Figure 8-1.

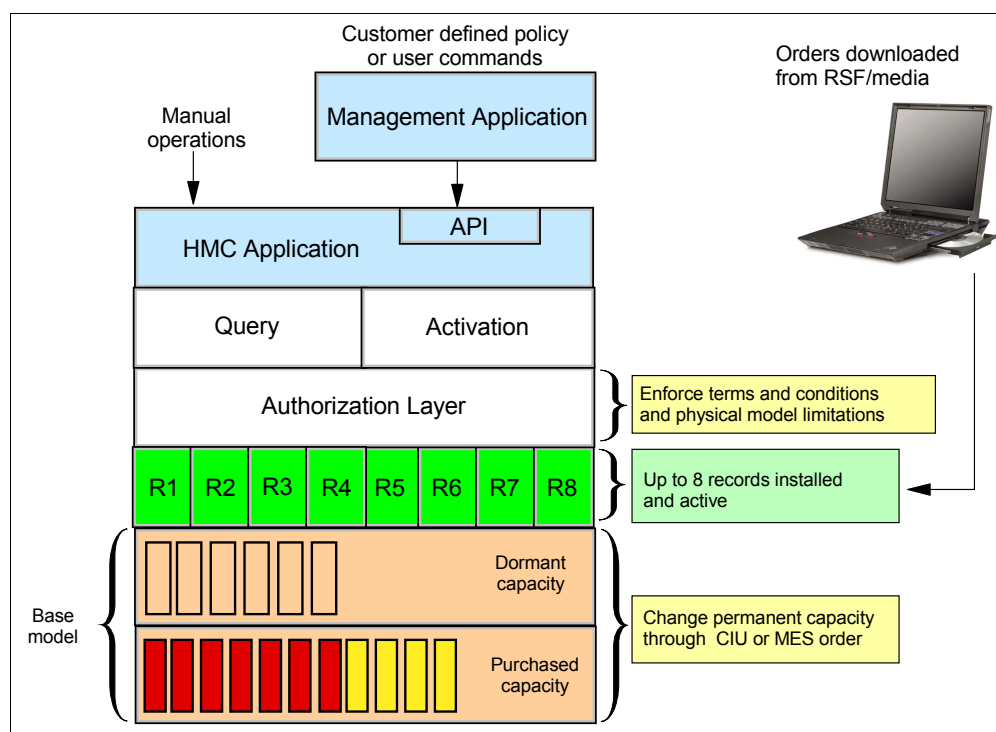


Figure 8-1 Provisioning architecture

The authorization layer enables administrative control over the temporary offerings. The activation and deactivation can be driven manually or under the control of an application through a documented application programming interface (API).

By using the API approach, you can customize at activation time the resources that are necessary to respond to the current situation, up to the maximum that is 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 specifications for all possible scenarios. However, the ordered configuration is the only possible activation for CPE.

This approach also enables you to update and replenish temporary upgrades, even in situations where the upgrades are active. Likewise, depending on the configuration, permanent upgrades can be performed while temporary upgrades are active. Examples of the activation sequence of multiple temporary upgrades are shown in Figure 8-2.

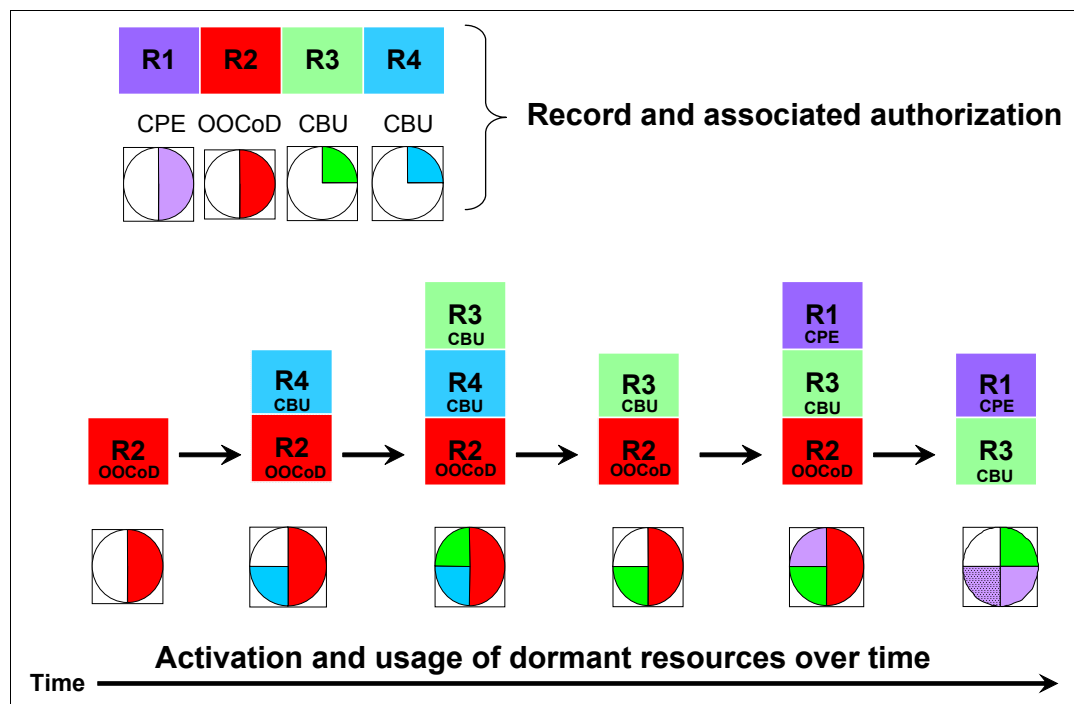


Figure 8-2 Example of temporary upgrade activation sequence

As shown in Figure 8-2, 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. Consider the following points:

- 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 includes a daily hardware and maintenance charge. The software charges can vary according to the license agreement for the individual products. For more information, contact your IBM Software Group representative.

On/Off CoD can concurrently add processors (CPs, ICFs, zIIPs, IFLs, and SAPs), increase the model capacity identifier, or both. It can do so up to the limit of the installed processor drawers of a system. It is restricted to twice the installed capacity. On/Off CoD requires a contractual agreement between you and IBM.

You decide whether to pre-pay or post-pay On/Off CoD. Capacity tokens that are inside the records are used to control activation time and resources.

- CBU is a concurrent and temporary activation of more CPs, ICFs, 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 method to use for workload management. A CBU activation² 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 processor drawers of the backup system. They can be available as unused PUs, an increase in the model capacity identifier, or both.

The CBU contract provides for one 90-day activation for 1 - 5 years, depending on the customers requirement. For every CBU year ordered, a 10-day CBU test (the *CBU test activation*) is part of the CBU contract. In addition, customers can order up to 10 more CBU tests within the same CBU record.

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. Terms and conditions are stated in the associated CBU contracts. For more information, contact you IBM representative.

- CPE is a concurrent and temporary activation of extra CPs, ICFs, 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, such as 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 processor drawers of the backup system. The capacity must be available 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. For more information, contact your IBM representative.

² z14 ZR1 servers provide more improvements in the CBU activation windows. These windows were improved to prevent inadvertent CBU activation.

8.2.3 Concurrent upgrade functions summary

The possible concurrent upgrades combinations are listed in Table 8-2.

Table 8-2 Concurrent upgrade summary

Type	Name	Upgrade	Process
Permanent	MES	CPs, ICFs, zIIPs, IFLs, SAPs, PU SCMs, memory, and I/O	Installed by IBM SSRs
	Online permanent upgrade	CPs, ICFs, zIIPs, IFLs, SAPs, and memory	Performed through the CIU facility
Temporary	On/Off CoD	CPs, ICFs, zIIPs, IFLs, and SAPs	Performed through the OOCOD facility
	CBU	CPs, ICFs, zIIPs, IFLs, and SAPs	Performed through the CBU facility
	CPE	CPs, ICFs, zIIPs, IFLs, and SAPs	Performed through the CPE facility

8.3 Miscellaneous equipment specification upgrades

MES upgrades enable concurrent and permanent capacity growth. MES upgrades allow the concurrent adding of processors (CPs, ICFs, zIIPs, IFLs, and SAPs), memory capacity, and I/O features.

For subcapacity models, MES upgrades allow the concurrent adjustment of the number of processors and capacity level. The MES upgrade can be performed by using LICCC only, installing more PU SCMs, adding PCIe+ I/O drawers, adding PCIe I/O³ features, or by using the following combinations:

- ▶ MES upgrades for PUs are done by any of the following methods:
 - LICCC assigning and activating unassigned PUs up to the limit of the installed PU SCMs in the CPC drawer.
 - LICCC to adjust the number and types of PUs, to change the capacity setting, or both.
 - Installing more PU SCMs, and LICCC assigning and activating unassigned PUs on the installed CPC drawer.
- ▶ MES upgrades for memory are done by one of the following methods:
 - By using LICCC to activate more memory capacity up to the limit of the memory cards on the CPC drawer. The Plan-ahead feature enables you to implement better control over future memory upgrades. For more information about this memory feature, see 2.5.6, “Preplanned memory” on page 45.
 - Installing more physical memory⁴ and the use of LICCC to activate more memory capacity.
- ▶ MES upgrades for I/O⁵ are done by installing more I/O⁵ features and supporting infrastructure (if required) on PCIe+ I/O drawers that are installed, or installing more PCIe+ I/O drawers to hold the new cards.

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.

³ Other adapter types, such as zHyperlink, Coupling Express LR, zEDC, and RoCE Express, also can be added to the PCIe+ I/O drawers through an MES.

⁴ When more memory is required than supported by the current CPC drawer feature, an upgrade of the CPC drawer feature (by adding one or more PU SCMs) is required.

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 used.

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 instruction” on page 322.

Upgrades: The MES provides the physical upgrade, which results in more installed PU SCMs, enabled PUs, different capacity settings for the CPs, and more memory, I/O ports, I/O adapters, and I/O drawers. Extra planning tasks are required for nondisruptive logical upgrades. For more information, see “Guidelines to avoid disruptive upgrades” on page 324.

8.3.1 MES upgrade for PUs

An MES upgrade for PUs can concurrently add CPs, ICFs, zIIPs, IFLs, and SAPs to a z14 ZR1 server by assigning available PUs on the CPC drawer through LICCC. Depending on the quantity of the extra PUs in the upgrade, more PU SCMs might be required, which is a disruptive upgrade. With the subcapacity models, more capacity can be provided by adding CPs, changing the capacity identifier on the current CPs, or both.

Limits: The sum of CPs, inactive CPs, ICFs, zIIPs, IFLs, unassigned IFLs, and SAPs cannot exceed the maximum limit of PUs available for client use. The number of zIIPs cannot exceed twice the number of purchased CPs.

An example of an MES upgrade for PUs (with two upgrade steps) is shown in Figure 8-3.

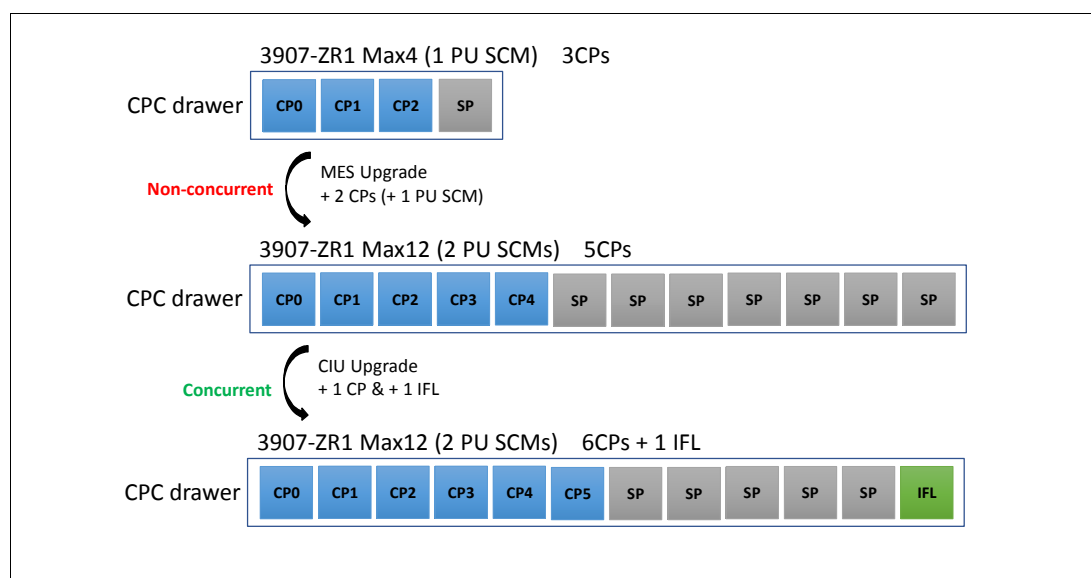


Figure 8-3 PU upgrade example

In the example that is shown in Figure 8-3 on page 295, a z14 ZR1 Max4 with three active CPs is upgraded with two more CPs. Because the Max4 can have only four active PUs, the z14 ZR1 is upgraded from a Max4 to a Max12 by physically adding a PU SCM to the CPC drawer. This upgrade is a *non-concurrent* upgrade.

For the second upgrade, which is adding one CP and one IFL by way of CIU, enough unassigned PUs are available to assign and activate the one CP and one IFL *concurrently*. If needed, more LPARs can be created concurrently to use the newly added processors.

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 usage, not on the system total capacity. For more information about WLC, see 7.8, “Software licensing” on page 277.

8.3.2 MES upgrades for memory

A MES upgrade for memory can concurrently add memory by enabling, through LICCC, extra capacity up to the limit of the currently installed memory cards. Physically installing more memory cards and then enabling (part of) the new memory capacity is a non-concurrent upgrade.

The Preplanned Memory Feature is available to allow better control over future memory upgrades. For more information see 2.5.6, “Preplanned memory” on page 45.

With proper 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.

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 were 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 extra storage resources to be available to the operating system image.

8.3.3 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 installed PCIe+ I/O drawer.
- ▶ By using the installed PCIe I/O drawer that provides 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 if not enough slots are available in the existing PCIe+ I/O drawer configuration.⁵

⁵ Adding a PCIe+ drawer concurrently is possible only when the CPC drawer has free PCIe fanout slots. If not, first a non-current CPC drawer feature upgrade is required.

For more information about PCIe+ I/O drawers, see 4.2, “I/O system overview” on page 120.

The number of PCIe+ I/O drawers that can be present in a z14 ZR1 server is listed in Table 8-3.

Table 8-3 PCIe+ drawer summary

Description	New build	MES add
PCIe+ I/O drawer	0 - 4	0 - 4
PCIe+ I/O drawer with 16U Reserved (FC 0617) ^a	0 - 2	0 - 2

a. When FC 0617 is ordered for the machine, it cannot be removed at a later stage; for example, in favor of adding a third PCIe+ I/O drawer.

Depending on the number of I/O features that are carried forward on an upgrade, the configurator determines the number of PCIe+ I/O drawers.

z/VSE, z/TPF, Linux on Z, and CFCC do *not* provide dynamic I/O configuration support. Although installing the new hardware is done concurrently, defining the new hardware to these operating systems requires an IPL.

Tip: z14 ZR1 servers feature a hardware system area (HSA) of 64 GB. HSA is *not* part of the client-purchased memory.

8.3.4 Feature on Demand

Only one FoD LICCC record is installed or staged at any time in the system. Its contents can be viewed under the Manage window, as shown in Figure 8-4.

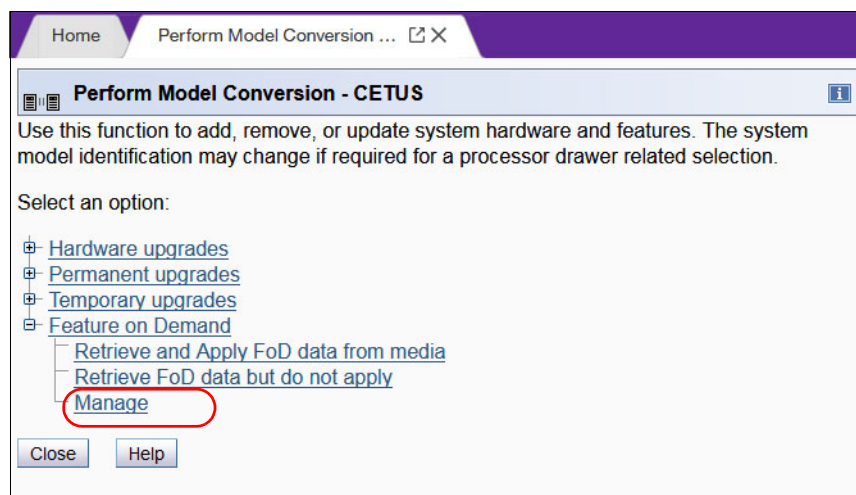


Figure 8-4 Features on-Demand window for zBX blade feature HWMs

A staged record can be removed without installing it. An FoD record can be installed only completely; no selective feature or partial record installation is available. The features that are installed are merged with the CPC LICCC after activation.

An FoD record can be installed only once. If it is removed, a new FoD record is needed to reinstall. A remove action cannot be undone.

8.3.5 Plan-ahead feature

The pre-planned memory plan-ahead feature is available for z14 ZR1 servers.

Pre-planned memory allows you to plan for nondisruptive memory upgrades. Any hardware that is required is pre-plugged, based on a target capacity that is specified in advance. Pre-plugged hardware can be enabled by using an LICCC order when more memory capacity is needed. FC 1993 provides 8 GB of pre-planned memory and FC 1996 provides 16 GB of pre-planned memory. The feature codes that are required to activate previously installed pre-planned memory are listed in Table 8-4.

Limit: The maximum configurable amount of pre-planned memory is 2048 GB. For example, with 1472 GB active memory, you can concurrently upgrade to 3520 GB active memory with the maximum amount of pre-planned memory configured.

Table 8-4 Feature codes for pre-planned memory activation

FC	Capacity increment (GB)	Notes
1739	8	When target < 128 GB
1740	8	When target >= 128 GB
1741	16	When target >= 128 GB
1742	32	When target >= 128 GB

Tip: Accurate planning and the definition of the target configuration allows you to maximize the value of plan-ahead features.

8.4 Permanent upgrade through the CIU facility

By using the CIU facility (through [the IBM Resource Link](#)), you can start a permanent upgrade for CPs, ICFs, zIIPs, IFLs, SAPs, or memory. When performed through the CIU facility, you add the resources without the need to have IBM personnel present at your location. You can also unassign previously purchased CPs and IFLs through the CIU facility.

Adding permanent upgrades to a system through the CIU facility requires that the Online CoD Buying feature (FC 9900) and the permanent upgrade enablement feature (FC 9898) are installed on the system. A permanent upgrade might change the system model capacity identifier (A0x - Z0x) if more CPs are requested, or if the capacity identifier is changed as part of the permanent upgrade. If necessary, more LPARs can be created concurrently to use the newly added PUs.

Consideration: A permanent upgrade of PUs can provide a physical concurrent upgrade (for example, from a Z03 to a Z04), which results in more enabled PU that are 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 324.

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 usage rather than system total capacity. For more information about WLC, see 7.8, “Software licensing” on page 277.

The CIU facility process on IBM Resource Link is shown in Figure 8-5.

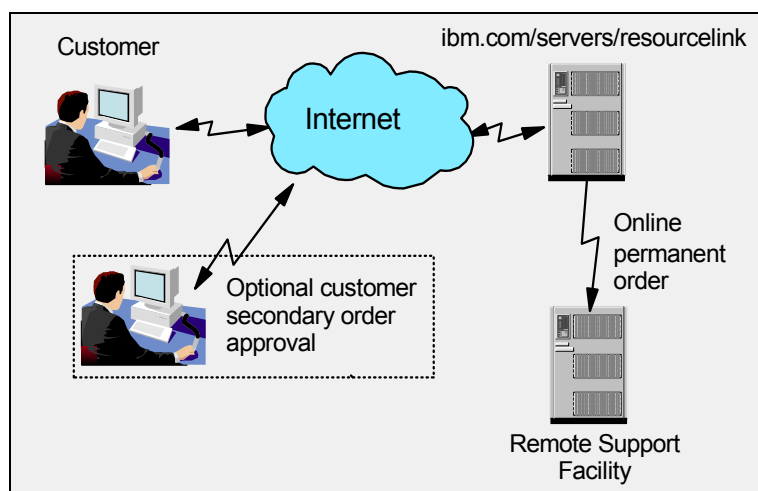


Figure 8-5 Permanent upgrade order example

The following sample sequence shows how to start an order on the IBM Resource Link:

1. Sign on to Resource Link.
2. Select **Customer Initiated Upgrade** from the main Resource Link page. Client and system information that is 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 options 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 is available within hours.

The order activation process for a permanent upgrade is shown in Figure 8-6. When the LICCC is passed to the Remote Support Facility, you are notified through an email that the upgrade is ready to be downloaded.

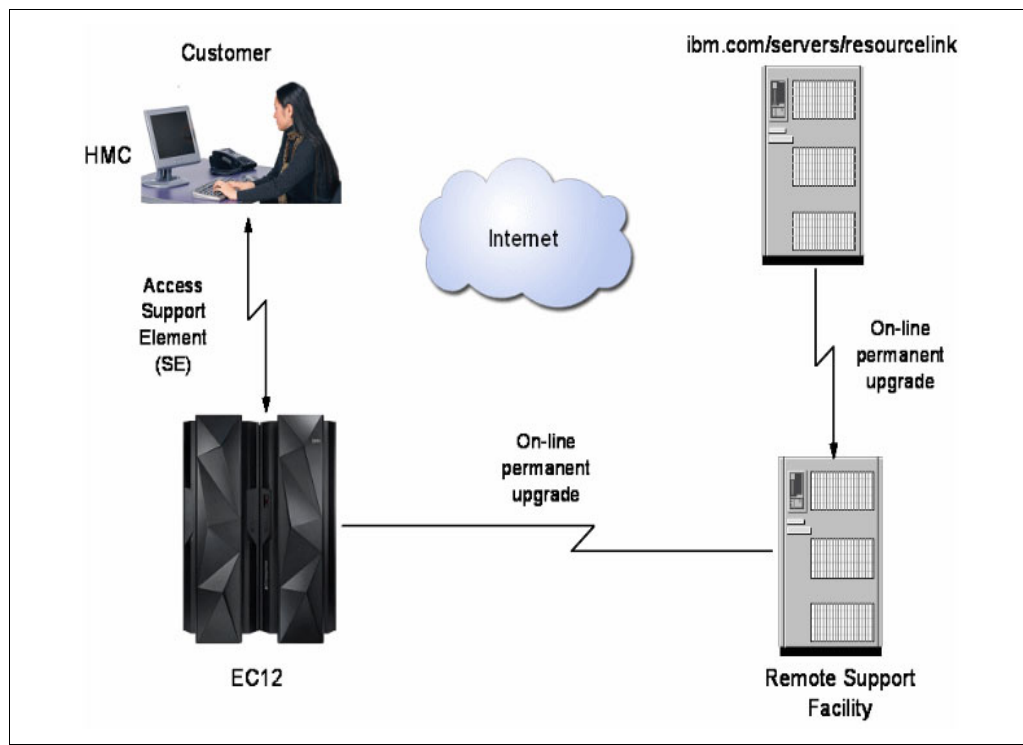


Figure 8-6 CIU-eligible order activation example

8.4.1 Ordering

Resource Link provides the interface that enables you to order a concurrent upgrade for a system. You can create, cancel, or 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 more information, see the [Resource Link website](#) (login required).

The initial view of the Machine profile on Resource Link is shown in Figure 8-7.

IBM Systems > System z > Resource Link > Customer Initiated Upgrade >

Machine profile

2818 - CEC04 - 5555556

Current configuration	
Model Capacity:	W02 (2 CPs)
ICF:	0
zAAP:	0
zIIP:	0
IFL:	1
SAP:	2
Memory:	24
Unassigned IFLs:	0
Management enablement level:	1. Manage

Current configuration as of 27 May 2011 12:16:48

Machine summary	
Type, model, serial:	2818 - M05 - CEC04

Customer summary	
Company name:	IBM
Customer number:	5555556
GEO, country:	Americas - zDutchy of Merwyn

Ordering options	
Order permanent upgrade	<input type="checkbox"/>
Order On/Off CoD record	<input type="checkbox"/>
Order On/Off CoD test record	<input type="checkbox"/>
Order On/Off CoD record with prepaid upgrades	<input type="checkbox"/>
Order On/Off CoD record with spending limits	<input type="checkbox"/>
Order administrative On/Off CoD test record	<input type="checkbox"/>
Order Capacity Backup (CBU) record	<input type="checkbox"/>
Order Capacity for Planned Events (CPE) record	<input type="checkbox"/>
Display upgrade matrix	

About ordering	
Authorization to create orders	
User ID:	ciutestuser@us.ibm.com
Name:	ciutestuser
Authorization to approve orders	
Not required	
Notes:	
<ul style="list-style-type: none"> A pre-negotiated price agreement exists for this machine. On/Off CoD Test: 0 staged out of 1 remaining 	

Ordering options	
CIU Permanent:	Enabled
On/Off CoD:	Enabled
CBU:	Enabled
CPE:	Enabled

To update profile	
Upload VPD	<input type="button" value="Upload"/>
Upload upgrade billing XML data	<input type="button" value="Upload"/>

For more information	
View machine's On/Off CoD order billing history	<input type="button" value="View"/>
Download upgrade history CSV (2KB)	<input type="button" value="Download"/>
Users authorized to order upgrades	<input type="button" value="View"/>
Users authorized to view orders	<input type="button" value="View"/>
Order status definitions	<input type="button" value="View"/>
Customer Initiated Upgrade information	<input type="button" value="View"/>

Permanent upgrades	
Open orders	<input type="button" value="Open"/>
Complete orders	<input type="button" value="Complete"/>
All orders	<input type="button" value="All"/>
There are no open orders for this machine.	

Figure 8-7 Machine profile window

The number of CPs, ICFs, zIIPs, IFLs, SAPs, memory size, and unassigned IFLs on the current configuration are displayed on the left side of the 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 also allows upgrades only within the bounds of the currently installed hardware.

8.4.2 Retrieval and activation

After an order is placed and processed, the appropriate upgrade record is passed to the IBM support system for download.

When the order is available for download, you receive an email that contains an activation number. You can then retrieve the order by using the Perform Model Conversion task from the SE, or through the Single Object Operation to the SE from an HMC.

In the Perform Model Conversion window, select **Permanent upgrades** to start the process, as shown in Figure 8-8.

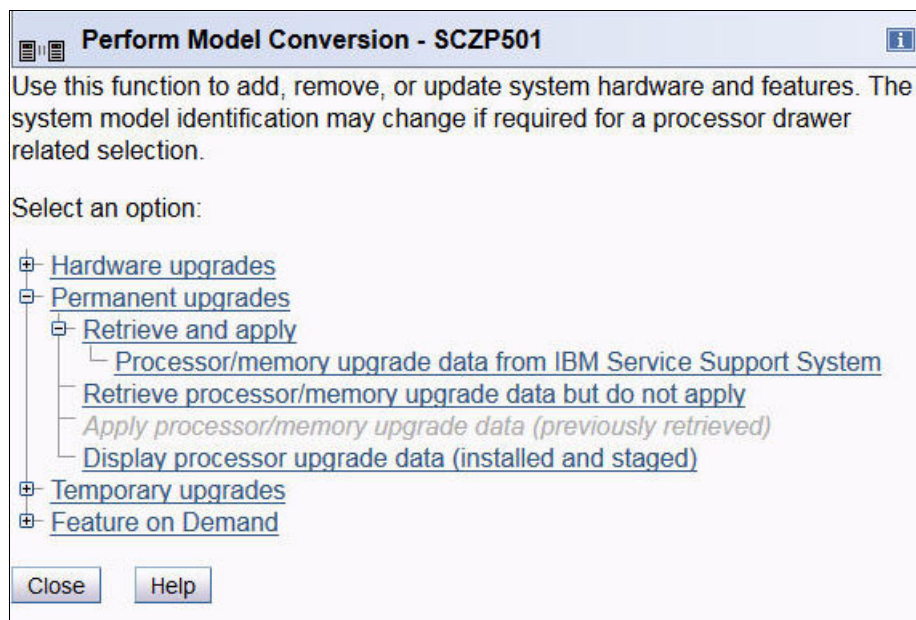


Figure 8-8 z14 ZR1 Perform Model Conversion window

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 start the permanent upgrade, as shown in Figure 8-9.

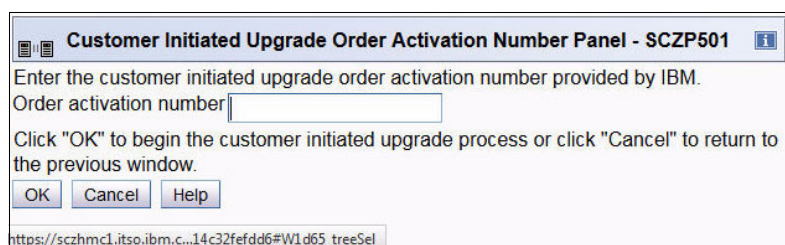


Figure 8-9 Customer Initiated Upgrade Order Activation Number window

8.5 On/Off Capacity on Demand

On/Off CoD allows you to enable temporarily PUs and unassigned IFLs that are available within the current hardware model. You can also use it to change capacity settings for CPs to help meet your peak workload requirements.

8.5.1 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 features its own tokens, and each On/Off CoD record includes 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 include 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. Billing occurs at the end of each billing window.

The resources that are billed are the highest resource usage inside each billing window for each capacity type. An activation period is one or more complete billing windows. 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.

Note: On/Off CoD (FC 9896) requires that the Online CoD Buying feature (FC 9900) is installed on the system that you want to upgrade.

The On/Off CoD to Permanent Upgrade Option is a new offering. It is an offshoot of On/Off CoD that takes advantage of 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 eligible resources for temporary use are CPs, ICFs, 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 PU SCMs can be temporarily and concurrently activated as CPs, ICFs, 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, zIIPs, or IFLs. Therefore, an On/Off CoD upgrade cannot change the system model. The addition of new processor drawers is not supported. However, the activation of an On/Off CoD upgrade can increase the model capacity identifier (A0x - Z0x).

8.5.2 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. For more information, see *z/OS MVS Capacity Provisioning User's Guide*, SA33-8299.

The use of the capacity provisioning function requires the following prerequisites:

- ▶ TCP/IP connectivity to observed systems
- ▶ RMF Distributed Data Server must be active
- ▶ CIM server must be active
- ▶ Security and CIM customization
- ▶ Capacity Provisioning Manager customization.

In addition, the Capacity Provisioning Control Center must be downloaded from the host and installed on a PC server. This application is used only to define policies. It is not required for regular operation.

Customizing the capacity provisioning function is required on the following systems:

- ▶ Observed z/OS systems

These systems are in one or multiple sysplexes that are to be monitored. For more information about the capacity provisioning domain, see 8.8, “Nondisruptive upgrades” on page 320.

- ▶ Runtime systems

These systems are where the Capacity Provisioning Manager is running, or to which the server can fail over after server or system failures.

8.5.3 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
- ▶ zIIP features up to the number of installed zIIPs
- ▶ SAPs up to two for all CPC drawer capacities

On/Off CoD can provide CP temporary capacity in the following 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 includes the following limits that are associated with its use:

- The number of CPs cannot be reduced.
- The target configuration capacity is limited to the following amounts:
 - Twice the currently installed capacity, expressed in MSUs for CPs.
 - Twice the number of installed IFLs, ICFs, and zIIPs (the number of extra SAPs that can be activated is two). For more information, see 8.2.1, “Upgrades” on page 288.

On/Off CoD can be ordered as prepaid or postpaid. A prepaid On/Off CoD offering record contains resource descriptions, MSUs, several 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 uses all of its capacity tokens. Then, 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 all of the resource tokens are used, or until the record expires. The record usually expires 180 days after its installation. If one capacity token type is used, resources from the entire record are deactivated.

For example, for a z14 ZR1 server with capacity identifier D02 (two CPs), a capacity upgrade through On/Off CoD can be delivered in the following ways:

- ▶ Add CPs of the same capacity setting. With this option, the model capacity identifier can be changed to a D03, which adds one more CP to make it a 3-way CP. It can also be changed to a D04, which adds two CPs and makes it a 4-way CP.
- ▶ Change to a different capacity level of the current CPs and change the model capacity identifier to a E02 or F02. The capacity level of the CPs is increased, but no other CPs are added. The D02 also can be temporarily upgraded to a E03, which increases the capacity level and adds another processor.

Use the Large System Performance Reference (LSPR) information to evaluate the capacity requirements according to your workload type. For more information about LSPR data for current IBM processors, see the [Large Systems Performance Reference for IBM Z page](#) of the IBM Systems website.

The On/Off CoD hardware capacity is charged on a 24-hour basis. A grace period is granted at the end of the On/Off CoD day. This grace period allows up to an hour after the 24-hour billing period to 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 z14 ZR1 server and sent back to the support systems.

If On/Off capacity is active, 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 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 extra 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, which results in more enabled processors that are available to a system configuration. Extra planning tasks are required for nondisruptive upgrades. For more information, see “Guidelines to avoid disruptive upgrades” on page 324.

To participate in this offering, you must accept contractual terms for purchasing capacity through the Resource Link, establish a profile, and install 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. An invoice is generated if the capacity is enabled during the calendar month. You are billed for the use of temporary capacity until the system is returned to the original configuration. Remove the enablement code if the On/Off CoD support is no longer needed.

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 that 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 the order sequence. If a staged order is installed out of sequence and later a higher-priced order is staged, the daily cost is based on the lower price.

Another possibility is to store unlimited On/Off CoD LICCC records on the SE with the same or different capacities, which gives you greater flexibility to enable quickly 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 can 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, 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 be activated only once. Subsequent sessions require a new order to be generated, which produces 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, as shown in Figure 8-10.

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: 12/09/2012 (mm/dd/yyyy) ☒ Renew automatically

Enable upgrades for up to:

Model capacity: 100%

ICF: 1

zAAP: 1

zIIP: 1

IFL: 1

SAP: 4

Default is to renew records automatically

Figure 8-10 Order On/Off CoD record window

8.5.4 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 charges that are associated with temporary hardware capacity, IBM software, and 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, which commences 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 stops at the end of the 24-hour period.

You also can perform administrative testing. No capacity is added to the system, but you can test all the procedures and automation for the management of the On/Off CoD facility.

An example of an On/Off CoD order on the Resource Link web page is shown in Figure 8-11.

Order On/Off CoD record

Step 2 of 2: Review and submit your order

Review the range of upgrades you selected on the previous page. The On/Off CoD record you are about to order will be configured to support activating any configurations within the range.

(*) indicates accepting the [Terms and Conditions of this order](#) is required to submit it. Mark the check box to indicate acceptance.

Expiration date:11 Oct 2015

Renew automatically: Yes

	Enable upgrades up to	Daily hardware prices	Daily maintenance prices (estimated) ¹
Model capacity:	0% more model capacity		
ICF:	2 more ICF engines	\$0.00	⌘12.00
zIIP:	2 more zIIP engines	\$0.00	⌘12.00
IFL:	0 more IFL engines		
SAP:	0 more SAP engines		

Description:

+0% model capacity, +2 ICF, +2 zIIP, +0 IFL, +0 SAP

Notes:

1. Reflects current established prices for the selected machine. Prices are subject to change; the actual prices in effect at the time of use will apply.

2. Daily prices for ICF, zIIP, IFL, and SAP upgrades are **per engine**.

3. The IFL upgrade daily hardware price includes per IFL for the management enablement level in effect for this machine.

Figure 8-11 On/Off CoD order example

The example order that is shown in Figure 8-11 is an On/Off CoD order for 0% more CP capacity (so the MCI remains the same), and for two more ICFs and two more zIIPs. The maximum number of CPs, ICFs, zIIPs, and IFLs is limited by the current number of available unused PUs of the CPC drawer. The maximum number of extra SAPs for any model is 2, but it also depends on the number of available PUs on the CPC drawer.

To finalize the order, you must accept Terms and Conditions for the order, as shown in Figure 8-12.

Terms of Order

You have requested an On/Off Capacity on Demand, or Temporary Capacity upgrade. Your enterprise has previously accepted the Temporary Capacity terms, restated here. In the event there is a conflict between the terms shown on this website and the terms specified in your contract with IBM, the terms of such contract prevail:

1) upon download and installation of this Temporary Capacity Upgrade, IBM grants you only a temporary license to use the LIC enabling such Temporary Capacity Upgrade. You may use such Temporary Capacity Upgrade only on the TC Eligible Machine for which such LIC is provided, and only to the extent of the authorization identified via the CIU Facility.

☒ I accept the Terms and Conditions of this order*

Submit

Figure 8-12 CIU order Terms and Conditions

8.5.5 Activation and deactivation

When a previously ordered On/Off CoD is retrieved from Resource Link, it is downloaded and stored on the SE HDD. You can activate the order manually or through automation when the capacity is needed.

If the On/Off CoD offering record does not contain resource tokens, you must deactivate the temporary capacity manually. Deactivation is done from the SE and is nondisruptive. Depending on how the 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 that were 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 be downloaded and activated only once. 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, which 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.

8.5.6 Termination

A client is contractually obligated to stop the On/Off CoD right-to-use feature when a transfer in asset ownership occurs. A client also can choose to stop the feature without transferring ownership.

Applying FC 9898 stops 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. When the CIU enablement feature is removed, the On/Off CoD right-to-use feature is simultaneously removed. Reactivating the right-to-use feature subjects the client to the terms and fees that apply then.

Upgrade capability during On/Off CoD

Upgrades that involve physical hardware are supported while an On/Off CoD upgrade is active on a particular z14 ZR1 server. 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 z14 ZR1 server requires service while an On/Off CoD upgrade is active, the repair can take place without affecting the temporary capacity.

Monitoring

When you activate an On/Off CoD upgrade, an indicator is set in 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 the 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 that is represented by the combined permanent and temporary capacity. All PSLC products are billed at the peak MSUs that are 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 increase the software bill until that capacity is allocated to LPARs and used.

Results from the STSI instruction reflect the current permanent and temporary CPs. For more information, see “Store System Information instruction” on page 322.

8.5.7 z/OS capacity provisioning

The z14 ZR1 provisioning capability that is combined with CPM functions in z/OS provides a flexible, automated process to control the activation of On/Off Capacity on Demand. The z/OS provisioning environment is shown in Figure 8-13.

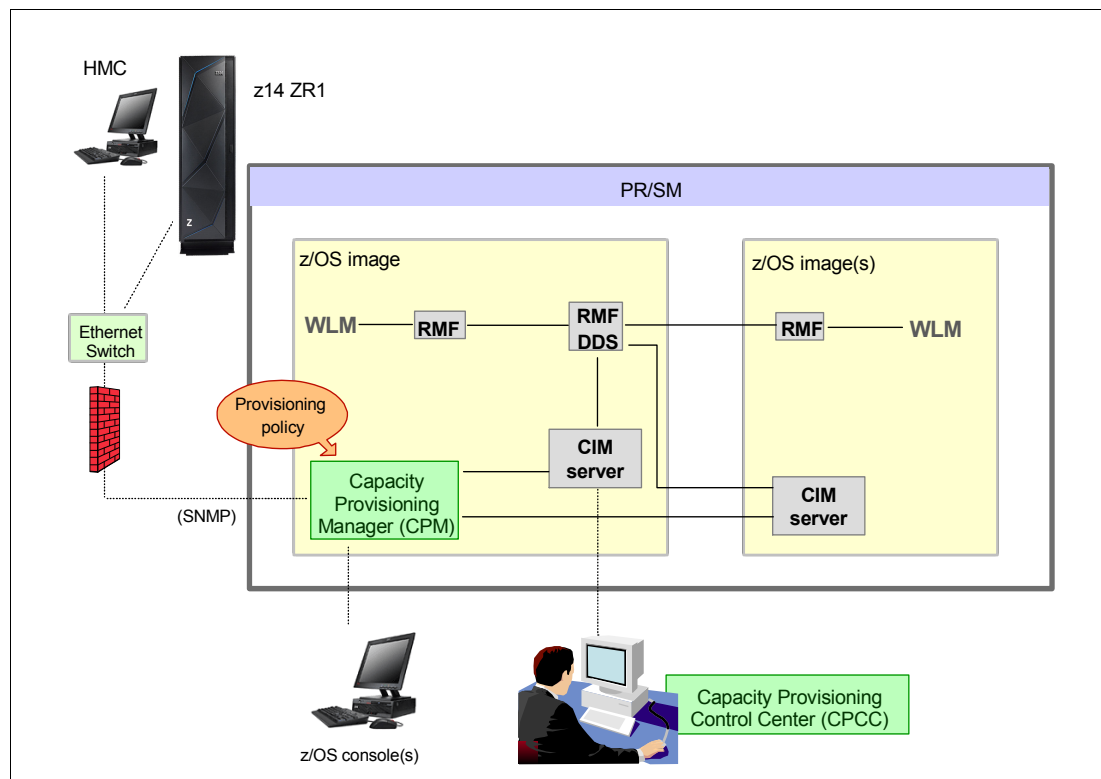


Figure 8-13 Capacity provisioning infrastructure

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.

A function inside z/OS, the CPM retrieves critical metrics from one or more z/OS systems' CIM structures and protocols. CPM communicates to local or remote SEs and HMCs by using the Simple Network Management Protocol (SNMP).

CPM can see 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 used 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 activates this record, or parts of it. However, you do receive warning messages if capacity tokens are close to being fully used.

You receive the messages five days before a capacity token is fully used. The five days are based on the assumption that the usage is constant for the five days. You must put operational procedures in place to handle these situations. You can deactivate the record manually, allow it occur 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 is moved over time into the z/OS Management Facility (z/OSMF). Parts of the CPCC are moved starting with z/OSMF V1R13, and are included in z/OSMF V2R1 and z/OS V2R2 z/OSMF.

Capacity Provisioning Domain

The provisioning infrastructure is managed by the CPM through the Capacity Provisioning Domain (CPD), which is controlled by the Capacity Provisioning Policy (CPP). The CPD is shown in Figure 8-14.

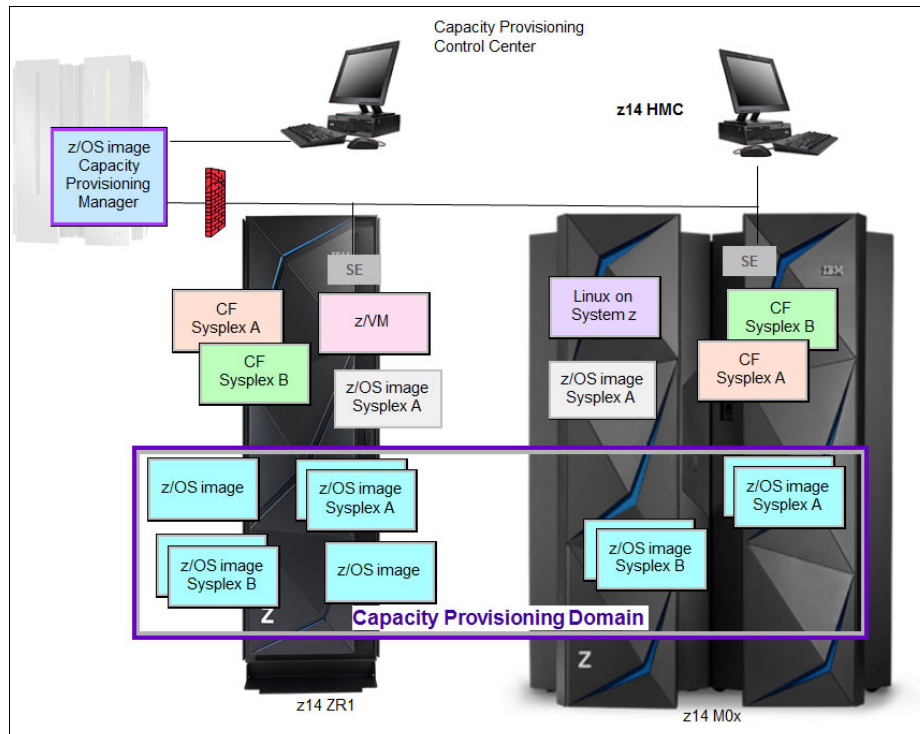


Figure 8-14 Capacity Provisioning Domain

The CPD configuration defines the 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. Although sysplexes and CPCs do not have to be contained in a domain, they 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 the following modes, which allows 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.

Also, the operator determines whether to ignore the information or to manually upgrade or downgrade the system by using the HMC, SE, or available CPM commands.

- Confirmation mode

In this mode, CPM processes capacity provisioning policies and interrogates the installed temporary offering records. Every action that is proposed by the CPM must be confirmed by the operator.

► Autonomic mode

This mode is similar to the confirmation mode, but no operator confirmation is required.

Several reports are available in all modes that contain information about the workload, 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). The criteria features the following elements:

- 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 must begin.
- A workload condition is which work qualifies for provisioning. It can have the following parameters:
 - The z/OS systems that can run eligible work.
 - The 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 that is described in the policy is activated and deactivated at the start and end times that are specified in the policy.

- Provisioning scope is how much more capacity can be activated and is expressed in MSUs.

The number of zIIPs must be one specification per CPC that is part of the CPD and are specified in MSUs.

The maximum provisioning scope is the maximum extra capacity that can be activated for all the rules in the CPD.

The provisioning rule is that if the specified workload is behind its objective in the specified time interval, up to the defined extra 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*, SC34-2661.

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 controls capacity tokens for the On/Off CoD records. In a situation where a capacity token is used, the system deactivates the corresponding offering record. Therefore, you must prepare routines for catching the warning messages about capacity tokens being used, and have administrative procedures in place for such a situation.

The messages from the system begin five days before a capacity token is fully used. To avoid capacity records being deactivated in this situation, replenish the necessary capacity tokens before they are used.

The Capacity Provisioning Manager operates based on Workload Manager (WLM) indications, and the construct that is used is the PI of a service class period. It is important to select service class periods that are appropriate for the business application that needs more capacity. For example, the application in question might be running through several service class periods, where the first period 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.

8.6 Capacity for Planned Event

CPE is offered with z14 servers to provide replacement backup capacity for planned downtime events. For example, if a server room requires an extension or repair work, replacement capacity can be installed temporarily on another z14 server in the client's environment.

Important: CPE is for planned replacement capacity only, and cannot be used for peak workload management.

CPE includes the following 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 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 the eConfig tool or the Resource Link is used, a target configuration must be ordered. The configuration consists of a model identifier, several speciality engines, or both. Based on the target configuration, several 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.

Note: CPE is intended for short duration events that last a maximum of three days.

After each CPE record is activated, you can access 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, SAP, IFL, and ICF). At the time of CPE activation, the contracted configuration is activated. The general rule of two zIIPs for each configured CP is enforced for the contracted configuration.

The PUs that can be activated by CPE come from the available unassigned PUs on any installed CPC drawer. CPE features can be added to a z14 ZR1 server 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 extra 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 that is available on the system. When the planned event ends, 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 a z14 ZR1 server nondisruptively.

8.7 Capacity Backup

CBU provides reserved emergency backup processor capacity for unplanned situations in which capacity is lost in another part of your enterprise. It allows you to recover by adding the reserved capacity on a designated z14 server.

CBU is the quick, temporary activation of PUs and is available in the following options:

- ▶ For up to 90 contiguous days, for 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 IBM Z workload capacity that is equivalent to the CBU upgrade capacity is shut down or otherwise made unusable during the CBU test.⁶

Important: CBU is for disaster and recovery purposes only. It *cannot* be used for peak workload management or for a planned event.

8.7.1 Ordering

The CBU process allows for CBU to activate CPs, ICFs, zIIPs, IFLs, and SAPs. 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 by using the following feature codes:

- ▶ FC 6805: More CBU tests
- ▶ FC 6817: Total CBU years ordered
- ▶ FC 6818: CBU records that are ordered
- ▶ FC 6820: Single CBU CP-year
- ▶ FC 6821: 25 CBU CP-year
- ▶ FC 6822: Single CBU IFL-year
- ▶ FC 6823: 25 CBU IFL-year
- ▶ FC 6824: Single CBU ICF-year
- ▶ FC 6825: 25 CBU ICF-year
- ▶ FC 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 depends on the quantity of CBU years (FC 6817). You can extend your CBU entitlements through the purchase of more CBU years.

The number of FC 6817 per instance of FC 6818 remains limited to five. Fractional years are rounded up to the nearest whole integer when calculating this limit. If there are two years and eight months before 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 CBU year that is 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.

⁶ All new CBU contract documents contain new CBU test terms to allow execution of production workload during CBU test. CBU clients must run the IBM client Agreement Amendment for IBM Z Capacity Backup Upgrade Tests.

The PUs that can be activated by CBU come from the available unassigned PUs on the CPC drawer. The maximum number of CBU features that can be *ordered* is 30. The number of features that can be *activated* is limited by the number of unused PUs on the system. However, the ordering system allows for over-configuration in the order.

You can order up to 30 CBU features regardless of the current configuration. However, at activation, only the capacity that is installed can be activated. 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 completed. On the full-capacity models, the CBU features indicate the amount of extra capacity that is needed. If the amount of necessary CBU capacity is equal to four CPs, the CBU configuration is four CBU CPs.

The number of CBU CPs must be equal to or greater than the number of CPs in the base configuration. Also, all of the CPs in the CBU configuration must have the same capacity setting. For example, if the base configuration is a two-way D02, providing a CBU configuration of a four-way of the same capacity setting requires two CBU feature codes.

If the required CBU capacity changes the capacity setting of the CPs, going from model capacity identifier D02 to a CBU configuration of a four-way E04 requires four CBU feature codes: two to upgrade from a D02 to a E02 and two to upgrade from an E02 to a E04.

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 when the capacity setting of the CPs is changed.

CBU can add CPs through LICCC only, and the z14 ZR1 server must have the correct number of processor drawers that are installed to allow the required upgrade. CBU can change the model capacity identifier to a *higher* value than the base setting, 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 a z14 ZR1 server nondisruptively. For each system enabled for CBU, the authorization to use CBU is available for a 1 - 5-year period.

The alternative configuration is activated *temporarily*, and provides more capacity that is 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 functions include CF LEVELs for coupling facility partitions, memory, and cryptographic functions, and 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 PUs, changed capacity settings that are available to a system configuration, or both. You can activate a subset of the CBU features that are 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 324.

For more information, see the *IBM Z Capacity on Demand User's Guide*, SC28-6846.

8.7.2 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 or deactivated concurrently from the HMC by using the API. On the SE, CBU is activated 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 HMC and SSO to 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 z14 ZR1server can have more capacity, more active PUs, or both. The extra 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 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 extra 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 HMC/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 start 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; that is, by using the same SE Perform a Model Conversion window and selecting the Temporary upgrades option. The HMC windows were changed to avoid accidental 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 CBU operation is shown in Figure 8-15. The permanent configuration is a B02, and a record contains four CP CBU features. During an activation, many target configurations are available. With four CP CBU features, you can add up to 4 CPs within the same MCI, which enables the activation of a B03, B04, B05, or a B06 (the blue path). Alternatively, two CP CBU features can be used to change the MCI (in the example from a B02 to a E02) and then add the remaining two CP CBU features to upgrade to a E04 (the red path).

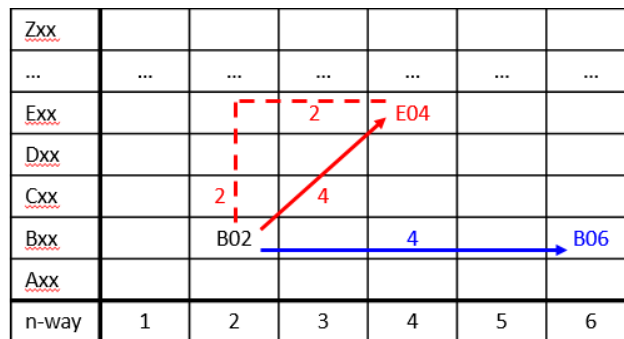


Figure 8-15 CBU example of a B02 with four CP CBU features

8.7.3 Automatic CBU enablement for GDPS

The IBM Geographically Dispersed Parallel Sysplex (GDPS) CBU enables automatic management of the PUs that are provided by the CBU feature during a system or site failure. Upon detection of a site failure or planned disaster test, GDPS concurrently adds CPs to the systems in the take-over site to restore processing power for mission-critical production workloads. GDPS automation runs the following tasks:

- ▶ 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 Z.

8.8 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.

z14 ZR1 servers allow *concurrent* upgrades, which means that dynamically adding capacity to the system is possible. If the operating system images that run 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 are not required to occur.

If the concurrent upgrade is intended to satisfy an *image* upgrade to an LPAR, the operating system that is running in this partition must concurrently configure more capacity online. z/OS operating systems include 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 z14 ZR1 system. These LPARs include all needed resources. These extra 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 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 *load average* value. The daemon is available in Linux SLES 10 SP2 and later, and in Red Hat Enterprise Linux (RHEL) V5R4 and up.

8.8.1 Components

The following components can be added, depending on the considerations that are described in this section:

- ▶ PUs
- ▶ Memory
- ▶ I/O
- ▶ Cryptographic adapters
- ▶ Special features

PUs

CPs, ICFs, zIIPs, IFLs, and SAPs can be added concurrently to a z14 ZR1 server if unassigned PUs are available on the CPC drawer. The number of zIIPs cannot exceed twice the number of CPs plus unassigned CPs.

If necessary, more LPARs can be created concurrently to use the newly added processors.

The Coupling Facility Control Code (CFCC) can also configure more processors online to coupling facility LPARs by using the CFCC image operations window.

Memory

Memory can be added concurrently up to the physical installed memory limit. By 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 Z supports Dynamic Storage Reconfiguration.

I/O

I/O features can be added concurrently if all the required infrastructure (I/O slots and PCIe Fanouts) is present in the configuration. 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), which allows nondisruptive I/O upgrades. However, having dynamic I/O reconfiguration on a stand-alone coupling facility system is not possible because no operating system with that capability is running on the system.

Cryptographic adapters

Crypto Express6S features can be added concurrently if all the required infrastructure is in the configuration.

Special features

Special features, such as zHyperlink, Coupling Express LR, zEnterprise Data Compression (zEDC) Express, and RoCE Express features, also can be added concurrently if all infrastructure is available in the configuration.

8.8.2 Concurrent upgrade considerations

By using an MES upgrade, On/Off CoD, CBU, or CPE, a z14 ZR1 server can be upgraded concurrently from one model to another (temporarily or permanently).

Enabling and using the extra processor capacity is not apparent to most applications. However, certain programs depend on processor model-related information, such as ISV products. Consider the effect on the software that is running on a z14 ZR1 server when you perform any of these configuration upgrades.

Processor identification

The following 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 other 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 (STIDP) instruction
STIDP is provided for compatibility with an earlier version.

Store System Information instruction

The relevant output from the STSI instruction is shown in Figure 8-16. 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 CoD offerings.

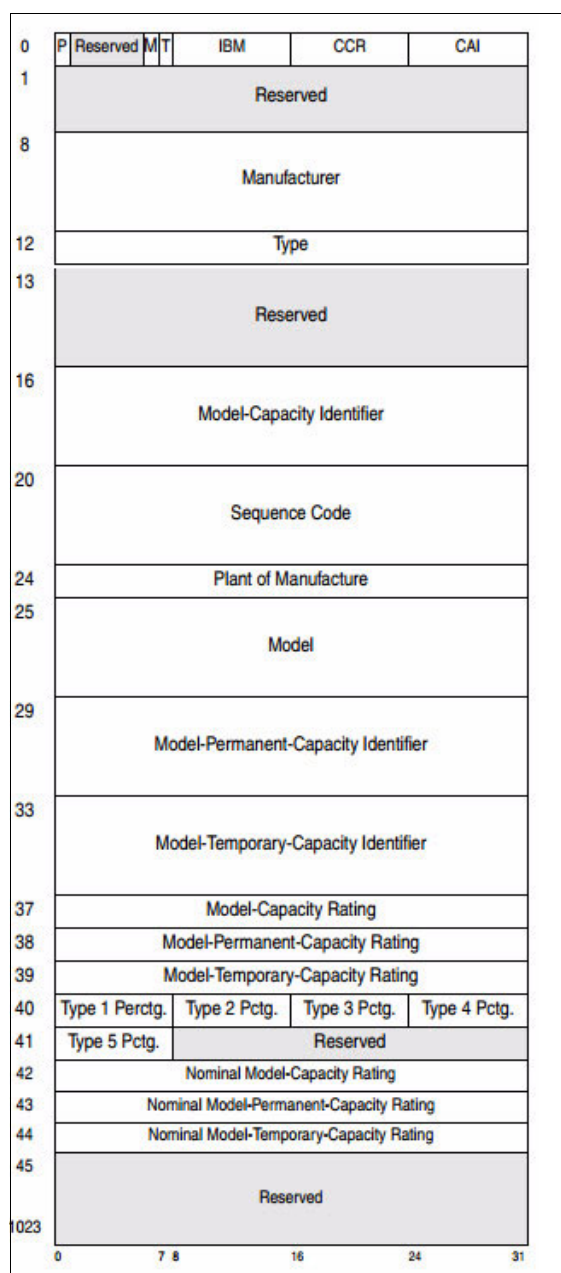


Figure 8-16 STSI output on a z14 ZR1 server

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 instruction

The Store CPU ID (STIDP) instruction provides information about the processor type, serial number, and LPAR identifier, as listed in Table 8-5. The LPAR identifier field is a full byte to support more than 15 LPARs.

Table 8-5 STIDP output for z14 servers

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'3907'	x'8000' ^c

a. The version code for z14 ZR1 servers 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. A high-order bit that is on indicates that the LPAR ID value that is 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. Consider the following points:

- ▶ 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 that is entered in the **SET CPUID** command. Bits 32 - 63 are the same as they were without running as a z/VM guest.

The possible output that is returned to the issuing program for an operating system that runs as a guest under z/VM is listed in Table 8-6.

Table 8-6 z/VM guest STIDP output for z14 ZR1 servers

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'3907'	x'8000'
With SET CPUID command	x'FF'	6-digit number as entered by the command SET CPUID = <i>nnnnnn</i>		x'3907'	x'8000'

Planning for nondisruptive upgrades

Online permanent upgrades, On/Off CoD, CBU, and CPE can be used to upgrade concurrently a z14 ZR1 server. 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. Planning ahead is a key factor for nondisruptive upgrades.

Disruptive upgrades are performed for the following reasons:

- ▶ LPAR memory upgrades when reserved storage was not previously defined are disruptive to image upgrades. z/OS and z/VM support this function.
- ▶ Upgrading from one CPC drawer feature to another (for example, from a Max4 to a Max12) by adding one or more PU SCMs is disruptive. Reasons for such an upgrade might be when:
 - More PU capacity is required
 - More physical memory is required
 - More PCIe Fanouts are required to install an ICA SR card or extra PCIe+ drawer
- ▶ Any installation of physical memory, also within the same CPC drawer feature, 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.

You can minimize the need for these outages by carefully planning and reviewing “Guidelines to avoid disruptive upgrades” on page 324.

Guidelines to avoid disruptive upgrades

Based on the reasons for disruptive upgrades (see “Planning for nondisruptive upgrades” on page 324), you can use the following guidelines to avoid or at least minimize these situations, which increases the chances for nondisruptive upgrades:

- ▶ By using an SE function that is called Logical Processor add (which is under Operational Customization tasks), CPs and zIIPs can be added concurrently to a running partition. The CP and zIIP and initial or reserved number of processors can be changed dynamically.
- ▶ The operating system that runs in the targeted LPAR must support the dynamic addition of resources and to configure processors online. The total number of defined and reserved CPs cannot exceed the number of CPs that are supported by the operating system. z/OS V1R13 with PTFs supports up to 100 processors. z/OS V2R1 and later supports 256 PUs per z/OS LPAR in non-SMT mode and 128 PUs per z/OS LPAR in SMT mode. For both, the PU total is the sum of CPs and zIIPs. z/VM supports up to 64 processors.
- ▶ Configure reserved storage to LPARs.

Configuring reserved storage for all LPARs before their activation enables them to be nondisruptively upgraded. The operating system that is running in the LPAR must configure memory online. The amount of reserved storage can be above the CPC drawer threshold limit, even if no other CPC drawer is already installed. With z14 servers, the current partition storage limit is 4 TB for z/OS V2.3, V2.2, and V2.1. z/VM 6.4 supports 2 TB memory partitions.
- ▶ Consider the plan-ahead memory option.

Use a convenient entry point for memory capacity, and select memory options that allow future upgrades within the memory cards that are installed on the CPC drawer. For more information about the offering, see 2.5.6, “Preplanned memory” on page 45.

8.9 Summary of Capacity on-Demand offerings

The CoD infrastructure and its offerings are major features that were introduced with the z13 system. These features are based on numerous client requirements for more flexibility, granularity, and better business control over the IBM Z infrastructure, operationally, and financially.

One major client requirement was to eliminate the need for a client authorization connection to the IBM Resource Link system when activating an offering. This requirement is met by the z196, zEC12, z13, and z14 servers.

After the offerings are installed on the z14 ZR1 server, they can be activated at any time at the client's discretion. No intervention by IBM or IBM personnel is necessary. In addition, the activation of CBU does not require a password.

The z14 ZR1 server can have up to eight offerings that are installed at the same time, with the limitation that only *one* of them can be an On/Off CoD offering. The others can be any combination. 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 usage (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 that is based on the requirements of the workload. The CPM cannot control other offerings.

For more information about any of the topics in this chapter, see *IBM Z Capacity on Demand User's Guide*, SC28-6943.



Reliability, availability, and serviceability

From the quality perspective, the z14 RAS design is driven by a set of high-level program reliability, availability, and serviceability (RAS) objectives. The IBM Z platform continues to drive toward Continuous Reliable Operation (CRO) at the single footprint level.

In order of priority, the key objectives are to ensure data and computational integrity, reduce or eliminate unscheduled outages, and reduce scheduled outages, planned outages, and the number of Repair Actions.

RAS can be accomplished with improved concurrent replace, repair, and upgrade functions for processors, memory, drawers, and I/O. RAS also extends to the nondisruptive capability for installing Licensed Internal Code (LIC) updates. In most cases, a capacity upgrade can be concurrent without a system outage. As an extension to the RAS capabilities, environmental controls are implemented in the system to help reduce power consumption and meet cooling requirements.

This chapter includes the following topics:

- ▶ 9.1, “RAS strategy” on page 328
- ▶ 9.2, “Structure change” on page 328
- ▶ 9.3, “Technology change” on page 329
- ▶ 9.4, “Reducing complexity” on page 331
- ▶ 9.5, “Reducing touches” on page 331
- ▶ 9.6, “z14 ZR1 availability characteristics” on page 333
- ▶ 9.7, “z14 ZR1 RAS functions” on page 335
- ▶ 9.8, “z14 ZR1 Enhanced Driver Maintenance” on page 338
- ▶ 9.9, “RAS capability for the HMC and SE” on page 340

9.1 RAS strategy

The RAS strategy is to manage change by learning from previous generations and investing in new RAS function to eliminate or minimize all sources of outages. Enhancements to z13s RAS designs are implemented on the z14 ZR1 system through the introduction of new technology, structure, and requirements. Continuous improvements in RAS are associated with new features and functions to ensure that IBM Z servers deliver exceptional value to clients.

The following overriding RAS requirements are principles as shown in Figure 9-1:

- ▶ Inclusion of existing (or equivalent) RAS characteristics from previous generations.
- ▶ Learn from current field issues and addressing the deficiencies.
- ▶ Understand the trend in technology reliability (hard and soft) and ensure that the RAS design points are sufficiently robust.
- ▶ Invest in RAS design enhancements (hardware and firmware) that provide IBM Z and Customer valued differentiation.

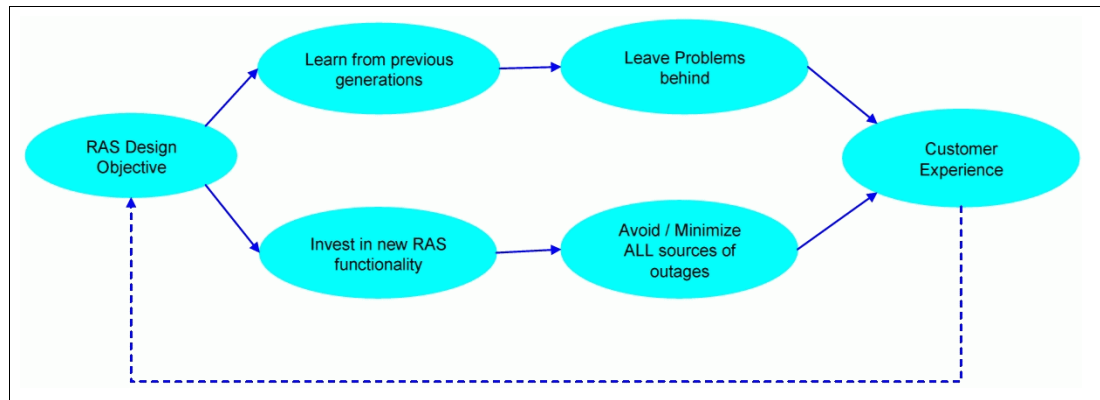


Figure 9-1 Overriding RAS requirements

9.2 Structure change

The z14 ZR1 is a transitional machine, not just a simplified version of z14 M0x with the following Goals:

- ▶ Enhanced system modularity
- ▶ Standardization to enable rapid integration
- ▶ Platform simplification

The z14 ZR1 is built up in a new form factor that is an industry standard 19-inch rack. It is an air cooled system that fulfills the requirements for an ASHRAE A3 environment (as the z13s). The CPC Drawer and the PCIe+ I/O Drawer uses the same Chipsets and PCIe I/O adapters as the z14 M0x models.

The power subsystem is completely redesigned and is now based on 200 - 240 V AC power distribution units (PDU). This configuration uses a Power System Control Network (PSCN) structure, which is the industry standard in data centers.

With z14 ZR1, the processing infrastructure within the CPC is designed by using drawer technology. To improve field-replaceable unit (FRU) isolation, Time Domain Reflectometry (TDR) techniques are applied between SCMs (CP-CP, and CP-SC), and between the CP SCM and dual inline memory modules (DIMMs).

Enhancements to thermal RAS also were introduced; for example, two redundant oscillator cards are installed into the central processor complex (CPC) drawer. The following z13s characteristics are continued with z14 ZR1:

- ▶ Keyed cables and plugging detection
- ▶ Master-master redundant oscillator design in the main memory
- ▶ Processor and nest chips are separate FRUs
- ▶ Point-of-load cards are separate FRUs
- ▶ Oscillator redundancy and concurrent oscillator card repair capability
- ▶ Built-in time domain reflectometer for FRU isolation in interface errors

9.3 Technology change

The IBM z14 Model ZR1 builds upon the RAS of the z13s with the following RAS improvements for the PU/Cache/Memory structure (as shown in Figure 9-2 on page 330):

- ▶ Symbol ECC on L3 data cache for better availability. This layer is now another layer of ECC protection that was with the main memory RAIM and L4 symbol ECC.
- ▶ New PU sparing algorithm, which gives one single dedicated spare PU for all configurations.
- ▶ The single CPC drawer-only configuration simplifies the structure and avoids CPC drawer to CPC drawer connectivity (no SMP cables and no SC SCM to SC SCM).
- ▶ L3 ability to monitor (dynamically) fenced macros and allow integrated sparing.
- ▶ Ability to spare the PU upon non-L2 cache/DIR Core Array delete.
- ▶ Improved error thresholding on PU, which avoids continuous recovery.
- ▶ Dynamic L3 cache monitor (“stepper”) to find and demote HSA lines in the cache.
- ▶ Symbol ECC on the L4 cache data, directory, configuration array and on the store protects key cache data.
- ▶ L4 ability to monitor (dynamically) fenced macros and allow integrated sparing.
- ▶ MCU (memory control unit) Cache Symbol ECC.
- ▶ L1 and L1+; L2 protected by PU sparing.
- ▶ PU mandatory address checking.
- ▶ Redundant parity on error in recovery unit (RU) bit to protect wordline (WL).

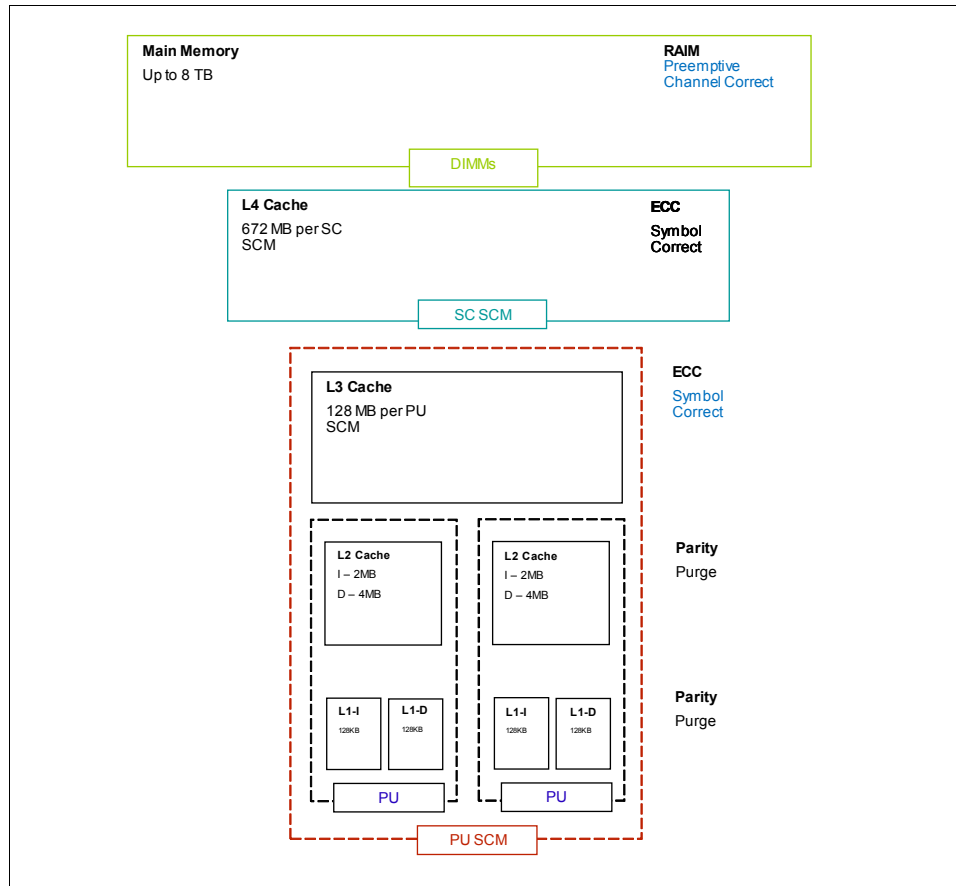


Figure 9-2 Memory and cache structure

The following storage controller (SC) single chip module (SCM) and main memory improvements also were made:

- ▶ Preemptive memory channel marking:
 - Analysis of uncorrectable errors considers pattern of previous correctable errors
 - More robust uncorrectable error handling
 - Simplified repair action
- ▶ Improved resilience in KUE¹ repair capability
- ▶ Virtual Flash Memory (Flash Express replacement) solution is moved to CPC drawer memory

Solution is moved to more robust storage RAIM protected (same function that main memory uses)

I/O and service

The following I/O and service improvements were made:

- ▶ OSA-express6S adds TCP checksum on large send:
 - Now on large and small packet sizes
 - Reduced the cost (CPU time) of error detection for large send
- ▶ Increased number of PSP, PCIe support partitions for managing native PCIe I/O adapters:
 - Now four partitions (was two on previous systems)

¹ Key in storage error uncorrected: Indicates that the hardware cannot repair a storage key that was in error.

- Reduced effect on MCL updates
- Better availability
- ▶ Faster Dynamic Memory Relocation engine:
 - Enables faster reallocation of memory (2x faster) that is used for LPAR activations, CDR, and concurrent upgrade
 - Provides faster, more robust service actions
- ▶ Dynamic Time Domain Reflectometry (TDR), which was static

Hardware facility that is used to isolate failures on wires provides better FRU isolation and improved service actions.
- ▶ Universal spare for PU SCMs and SC SCMs and processor drawer

9.4 Reducing complexity

z14 ZR1 continues the z13s enhancements that reduced system complexity. Specifically, simplifications were made in CPC drawer technology, which reduces the number of CPC drawers to one and RAIM recovery in the memory subsystem design. Memory DIMMs are not cascaded, which eliminates the double FRU call for DIMM errors.

Independent channel recovery with replay buffers on all interfaces allows recovery of a single DIMM channel, while other channels remain active. More redundancies are incorporated in I/O pins for clock lines to main memory, which eliminates the loss of memory clocks because of connector (pin) failure.

The following RAS enhancements reduce service complexity:

- ▶ Continued use of RAIM with ECC.
- ▶ No cascading of memory DIMM to simplify the recovery design.
- ▶ Replay buffer for hardware retry on soft errors on the main memory interface.
- ▶ Redundant I/O pins for clock lines to main memory.

9.5 Reducing touches

IBM Z RAS efforts focus on the reduction of unscheduled, scheduled, planned, and unplanned outages. IBM Z technology has a long history of demonstrated RAS improvements, and this effort continues with changes that reduce service *touches* on the system.

Firmware was updated to improve filtering and resolution of errors that do not require action. Enhanced integrated sparing in processor cores, cache relocates, N+1 SEEPROM and POL N+2 redundancies, and DRAM marking also are incorporated to reduce touches. The following RAS enhancements reduce service touches:

- ▶ Improved error resolution to enable filtering
- ▶ Enhanced integrated sparing in processor cores
- ▶ Cache relocates
- ▶ N+1 SEEPROM
- ▶ N+2 POL
- ▶ DRAM marking

- ▶ (Dynamic) Spare lanes for CP-SC, CP-CP, CP-mem
- ▶ N+1 controllers, blowers, and sensors
- ▶ N+1 Support Element (SE) (with N+1 SE power supplies)
- ▶ Redundant SEEPROM on memory DIMM
- ▶ Redundant temperature sensor (one SEEPROM and one temperature sensor per I2C bus)
- ▶ FICON forward error correction

Table 9-1 compares the integrated sparing functions for zBC12, z13s, and z14 ZR1 and shows the improvements that are archived.

Table 9-1 Integrated Sparing functions

Subsystem	Integrated sparing RAS function	zBC12	z13s	z14ZR1
PU				
	Dynamic PU Sparing	y	y	y
	Dynamic cache line delete/relocate	y	y	y
	L4 Dynamic subarray delete	n/a	y	y
	L3 Dynamic subarray delete	n	n	y
	Dual module SEEPROM	y	y	y
Memory				
	Dynamic DRAM Marking - N+2 DRAMS per rank	y	y	y
	DIMM temperature sensors	y	y	y
	Dual DIMM SEEPROM	y	y	y
	DIMM sockets - dual clock tabs	n/a	y	y
Power				
	Point of Load (POL) - N+2, SEEPROM	y	y	y
	Voltage regulator module (VRM) N+2	n	y	y
Thermal				
	Corrosion sensor	y	y	y
Fabric Interconnect				
	Dynamic PU/SC bus lane sparing	n	y	y
	Dynamic PU/PU bus lane sparing	n	y	y
	Dynamic memory bus data lane sparing	n	y	y

9.6 z14 ZR1 availability characteristics

The following functions include availability characteristics on z14 ZR1:

- Concurrent memory upgrade

Memory can be upgraded concurrently by using Licensed Internal Code Configuration Control (LICCC) if physical memory is available on the drawer. Memory plan ahead can be used at the time of initial configuration to provides more resources for future use.

- 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. z14 ZR1 is designed to support the concurrent activation of a selected new driver level.

- Concurrent fanout addition or replacement

A PCIe fanout card provides the path for data between memory and I/O through PCIe cables. With z14 ZR1, hot-pluggable and concurrently upgradeable fanouts are available. Up to eight PCIe fanout are available to the CPC drawer for z14 ZR1.

- Redundant I/O interconnect

Redundant I/O interconnect helps maintain critical connections to devices. z14 ZR1 allows a single PCIe+ I/O drawer adapter a fanout card, or even a PCIe+ I/O drawer in a multi PCIe I/O drawer system to be removed and reinstalled concurrently during a repair action. Connectivity to the system I/O resources is maintained through a second path when planned thoroughly.

- Dynamic oscillator switch-over

z14 ZR1 has two oscillator cards: a primary and a backup. During a primary card failure, the backup card transparently detects the failure, switches over, and provides the clock signal to the system.

- Processor unit (PU) sparing

z14 ZR1 has one spare PU to maintain performance levels if an active CP, Internal Coupling Facility (ICF), Integrated Facility for Linux (IFL), IBM z Integrated Information Processor (zIIP), integrated firmware processor (IFP), or system assist processor (SAP) fails. Transparent integrated sparing for failed processors is supported. One spare PU is available per system.

- Application preservation

This function is used when a CP fails and no spares are available. 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 z14 air-cooled configuration includes a newly designed front to rear air cooling system. The fans, controls, and sensors are N+1 redundant.

- FICON Express16S+ with Forward Error Correction (FEC)

FICON Express16S+ features continue to provide a new standard for transmitting data over 16 Gbps links by using 64b/66b encoding. The new standard that is defined by T11.org FC-FS-3 is more efficient than the current 8b/10b encoding.

FICON Express16S+ channels that are running at 16 Gbps can use FEC capabilities when connected to devices that support FEC.

FEC allows FICON Express16S+ channels to operate at higher speeds, over longer distances, with reduced power and higher throughput. They also retain the same reliability and robustness for which FICON channels are traditionally known.

FEC is a technique that is used for controlling errors in data transmission over unreliable or noisy communication channels. When running at 16 Gbps link speeds, clients often see fewer I/O errors, which reduces the potential effect to production workloads from those I/O errors.

Read Diagnostic Parameters (RDP) improve Fault Isolation. After a link error is detected (for example, IFCC, CC3, reset event, or a link incident report), link data that is returned from Read Diagnostic Parameters is used to differentiate between errors that result from failures in the optics versus failures because of dirty or faulty links.

Key metrics can be displayed on the operator console. The results of a display matrix command with the **LINKINFO=FIRST** parameter that collects information from each device in the path from the channel to the I/O device is shown in Figure 9-3 (a z14 M02 is used in this example). The following output is displayed:

- Transmit (Tx) and Receive (Rx) optic power levels from the PCHID, Switch Input and Output, and I/O device
- Capable and Operating speed between the devices
- Error counts
- Operating System requires new function APAR OA49089

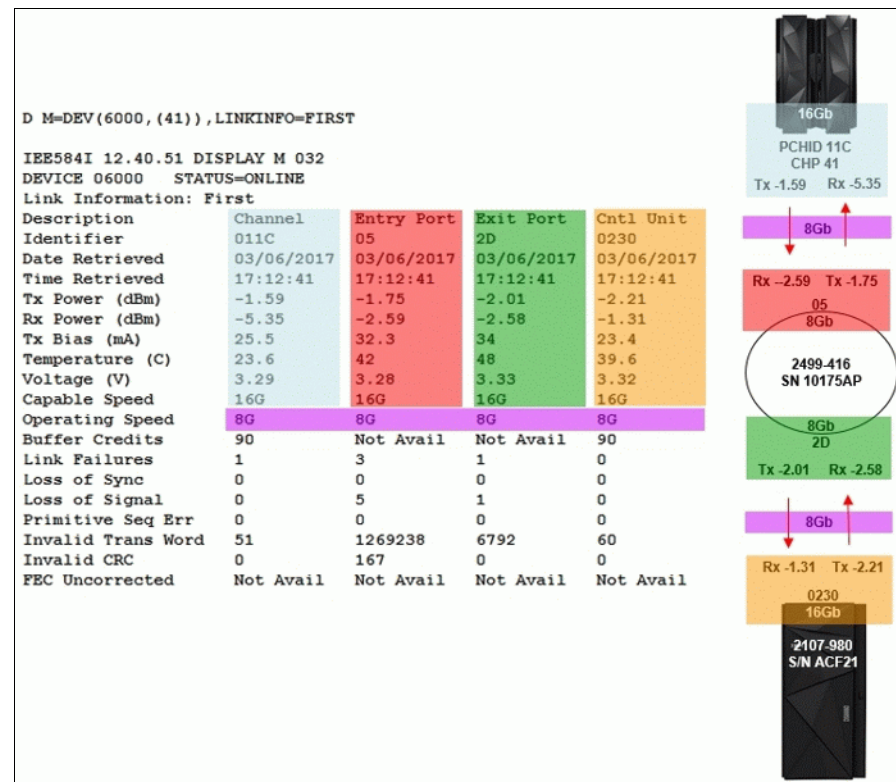


Figure 9-3 Read Diagnostic Parameters function

The new IBM Z Channel Subsystem Function performs periodic polling from the channel to the end points for the logical paths that are established and reduces the number of useless Repair Actions (RAs).

The RDP data history is used to validate Predictive Failure Algorithms and identify Fibre Channel Links with degrading signal strength before errors start to occur. The new Fibre Channel Extended Link Service (ELS) retrieves signal strength.

► **FICON Dynamic Routing**

FICON Dynamic Routing (FIDR) enables the use of storage area network (SAN) dynamic routing policies in the fabric. With the z14 systems, FICON channels are no longer restricted to the use of static routing policies for inter-switch links (ISLs) for cascaded FICON directors.

FICON Dynamic Routing dynamically changes the routing between the channel and control unit that is based on the Fibre Channel Exchange ID. Each I/O operation has a unique exchange ID. FIDR supports static SAN routing policies and dynamic routing policies.

FICON Dynamic Routing can help clients reduce costs by providing the following features:

- Share SANs between their FICON and FCP traffic.
- Improve performance because of SAN dynamic routing policies that better use all the available ISL bandwidth through higher use of the ISLs.
- Simplify management of their SAN fabrics by using static routing policies that assign different ISL routes with each power-on-reset (POR), which makes the SAN fabric performance difficult to predict.

Clients must ensure that all devices in their FICON SAN support FICON Dynamic Routing before they implement this feature.

9.7 z14 ZR1 RAS functions

Hardware RAS function improvements focus on addressing all sources of outages. Sources of outages feature the following 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 z14 ZR1 design phase focuses on enhancing planning to simplify or eliminate planned outages.

The difference between scheduled outages and planned outages might not be obvious. The general consensus is that scheduled outages occur sometime soon. The time frame is approximately two weeks.

Planned outages are outages that are planned well in advance and go beyond this approximate two-week time frame. The distinction between scheduled and planned outages is beyond the scope of this chapter.

Preventing unscheduled, scheduled, and planned outages was addressed by the IBM Z system design for many years.

z14 ZR1 introduces a fixed size HSA of 64 GB. This size helps eliminate planning requirements for HSA and provides the flexibility to update dynamically the configuration. You can perform the following tasks dynamically:²

- ▶ 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

By addressing the elimination of planned outages, the following tasks also are possible:

- ▶ Concurrent driver upgrades
- ▶ Concurrent and flexible customer-initiated upgrades

For more information about the flexible upgrades that are started by clients, see 8.2.2, “Customer Initiated Upgrade facility” on page 290.

9.7.1 Scheduled outages

Concurrent hardware upgrades, parts replacement, driver upgrades, and firmware fixes that are available with z14 ZR1 all address the elimination of scheduled outages. Also, the following indicators and functions that address scheduled outages are included:

- ▶ Double memory data bus lane sparing.
- ▶ Redundant N+1 Power Distribution Units (PDU).

The bulk power hub (BPH) in former Z systems is repacked into a new design that is based on switchable standard PDU. This design power cycles the SEs and PSCN Ethernet switch and avoids cable misplugging. The number of PDUs is configuration-dependent.
- ▶ CPC drawer power distribution N+1 design by way of Power Supply Units (PSU).
- ▶ The CPC Drawer uses point of load (POL) cards in a highly redundant N+2 configuration. POL regulators are daughter cards that contain the voltage regulators for the principle logic voltage boundaries in the z14 CPC drawer. They plug onto the CPC drawer system board and are nonconcurrent FRUs for the affected drawer, similar to the memory DIMMs.
- ▶ Redundant (N+2) ambient temperature, pressure, and humidity sensors.
- ▶ 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 provides scheduled time to repair the storage module.

- ▶ Single PU checkstop and sparing.

This indicator shows that a PU malfunctioned and is spared. IBM determines what course of action to take based on the system and the history of that system.

² Some planning considerations might be necessary. For more information, see Chapter 8, “System upgrades” on page 281.

- ▶ Air-cooled system, which features fans with N+1 redundancy and a new designed front-to-rear cooling system.
- ▶ 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 concurrently replace a part. Through the implementation of the VLAN to the redundant internal Ethernet service network, these advantages are improved, which makes the service network easier to handle and more flexible.
- ▶ The PCIe+ I/O drawer is available for z14 ZR1. It and all of the PCIe I/O drawer-supported adapters can be installed concurrently.
- ▶ Memory interface logic to maintain channel synchronization when one channel goes into replay. z14 ZR1 can isolate recovery to only the failing channel.
- ▶ PCIe redrive hub cards plug straight in (no blind mating of connector). Simplified plugging that is more reliable is included.
- ▶ ICA (short distance) coupling cards plug straight in (no blind mating of connector). Simplified plugging that is more reliable is included.
- ▶ Coupling Express LR (CE LR) coupling cards plug into the PCIe I/O Drawer, which allows more connections with the same bandwidth.

9.7.2 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
RAIM on IBM Z systems is a concept similar to the concept of Redundant Array of Independent Disks (RAID). The RAIM design detects and recovers from dynamic random access memory (DRAM), socket, memory channel, or DIMM failures. The RAIM design requires the adding one memory channel that is dedicated for RAS.
The parity of the four data DIMMs is stored in the DIMMs that are attached to the fifth memory channel. Any failure in a memory component can be detected and corrected dynamically. z14 ZR1 inherited this memory architecture.
The memory system on z14 ZR1 is implemented with an enhanced version of the Reed-Solomon 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 DIMM replacements. The design of the z14 ZR1 further improved this chip marking technology. Graduated DRAM marking is available, and channel marking and scrubbing calls for replacement on the third DRAM failure is available. For more information about the memory system on z14 ZR1, see 2.5, “Memory” on page 38.

- ▶ Improved thermal, altitude, and condensation management
- ▶ Server Time Protocol (STP) recovery enhancement

Enhanced Console Assisted Recovery (ECAR) was new with z13s and z13 GA2 and carried forward to z14. It contains better recovery algorithms during a failing Primary Time Server (PTS) and uses communication over the HMC/SE network to assist with BTS takeover. For more information, see Chapter 11, “Hardware Management Console and Support Elements” on page 357.
- ▶ Design of pervasive infrastructure controls in processor chips in memory ASICs.
- ▶ Improved error checking in the processor recovery unit (RU) to better protect against word line failures in the RU arrays.

9.8 z14 ZR1 Enhanced Driver Maintenance

Enhanced Driver Maintenance (EDM) is one more step toward reducing the necessity for and the duration of a scheduled outage. One of the components to planned outages is LIC Driver updates that are run in support of new features and functions.

When correctly configured, z14 ZR1 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. Certain LIC updates do not allow a concurrent update or upgrade.

Consider the following key points regarding 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 the z14 ZR1 to be activated, can block the ability to run a concurrent driver upgrade.
- ▶ An icon on the 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 driver can eliminate or reduce a planned outage.
- ▶ Concurrent crossover from Driver level N to Driver level $N+1$, then 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 started. Unrecoverable errors during an update might require a scheduled outage to recover.

The EDM function does not eliminate the need for planned outages for driver-level upgrades. Upgrades might require a system level or a functional element scheduled outage to activate the new LIC. The following circumstances require a scheduled outage:

- ▶ Specific complex code changes might dictate a disruptive driver upgrade. You are alerted in advance so that you can plan for the following changes:
 - Design data or hardware initialization data fixes
 - CFCC release level change

- OSA CHPID code changes might require PCHID Vary OFF/ON to activate new code.
z14 introduced the support to concurrently activate an MCL on an OSA-ICC channel to improve the availability and simplification of the firmware maintenance. The OSD channels already feature this capability.
- Crypto code changes might require PCHID Vary OFF/ON to activate new code.

Note: zUDX clients should contact their User Defined Extensions (UDX) provider before installing Microcode Change Levels (MCLs). Any changes to Segments 2 and 3 from a previous MCL level might require a change to the client's UDX. Attempting to install an incompatible UDX at this level results in a Crypto checkstop.

9.8.1 Resource Group and native PCIe MCLs

Microcode fixes (referred to as *individual MCLs* or *packaged in Bundles*) might be required to update the Resource Group code and the native PCIe features. Although the goal is to minimize changes or make the update process concurrent, the maintenance updates at times can require the Resource Group or the affected native PCIe to be toggled offline and online to implement the updates. The native PCIe features (managed by Resource Group code) are listed in Table 9-2.

Table 9-2 Native PCIe cards for z14 ZR1

Native PCIe adapter type	Feature code	Resource required to be offline
25GbE RoCE Express2	0430	FIDs/PCHID
10GbE RoCE Express	0412	FIDs/PCHID
zEDC Express	0420	FIDs/PCHID
Coupling Express LR	0433	CHPIDs/PCHID
IBM zHyperLink Express	0431	FIDs/PCHID

Consider the following points for managing native PCIe adapters microcode levels:

- Updates to the Resource Group require all native PCIe adapters that are installed in that RG to be offline. For more information about this requirement, see Appendix C, "Native Peripheral Component Interconnect Express" on page 419.
- Updates to the native PCIe adapter require the adapter to be offline. If the adapter is not defined, the MCL session automatically installs the maintenance that is related to the adapter.

The PCIe native adapters are configured with Function IDs (FIDs) and might need to be configured offline when changes to code are needed. To help alleviate the number of adapters (and FIDs) that are affected by the Resource Group code update, z14 ZR1 increased the number of Resource Groups from two per system (for previous systems) to four per system (CPC).

Note: Other adapter types, such as FICON Express, OSA Express, and Crypto Express that are installed in the PCIe I/O drawer are not effected because they are not managed by the Resource Groups.

The rear view of the PCIe I/O drawer and the Resource Group assignment by card slot are shown in Figure 9-4. All PCIe I/O drawers that are installed in the system feature the same Resource Group assignment.

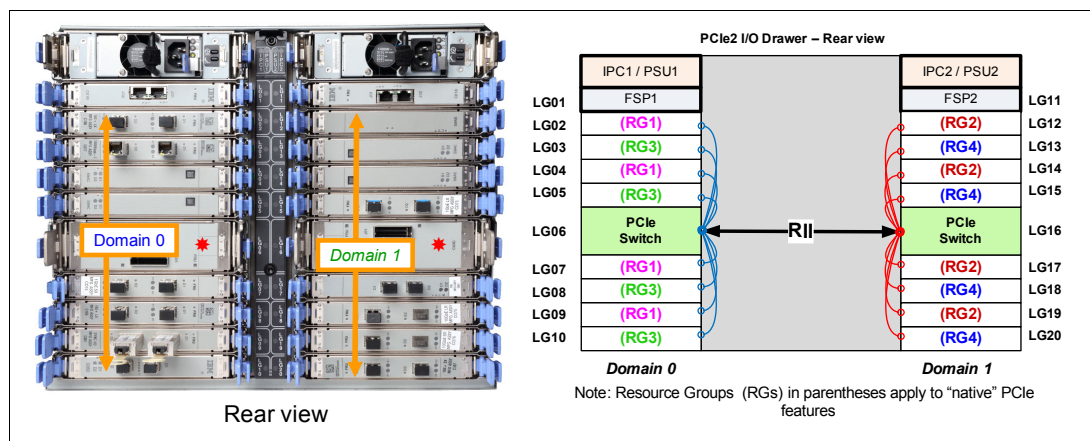


Figure 9-4 Resource Group slot assignment

9.9 RAS capability for the HMC and SE

The HMC and the SE include the following RAS capabilities:

- ▶ Back up from HMC and SE

For the customers who do not have an FTP server that is defined for backups, the HMC can be configured as an FTP server, which is new with z14.

On a scheduled basis, the HMC hard disk drive (HDD) is backed up to the USB flash memory drive (UFD), a defined FTP server, or both.

SE HDDs are backed up on to the primary SE HDD and an alternative SE HDD. In addition, you can save the backup to a defined FTP server.

For more information, see 11.2.6, “New backup options for HMCs and primary SEs” on page 362.

- ▶ Remote Support Facility (RSF)

The HMC RSF provides the important communication to a centralized IBM support network for hardware problem reporting and service. For more information, see 11.4, “Remote Support Facility” on page 374.

- ▶ Microcode Change Level (MCL)

Regular installation of MCLs is key for RAS, optimal performance, and new functions. Generally, plan to install MCLs quarterly at a minimum. Review hiper MCLs continuously. You must decide whether to wait for the next scheduled apply session, or schedule one earlier if your risk assessment of the new hiper MCLs warrants.

For more information, see 11.5.4, “HMC and SE microcode” on page 380.

- ▶ SE

z14 ZR1 is provided with two 1U trusted servers inside the rack: one is always the primary SE and the other is the alternative SE. The primary SE is the active SE. The alternative acts as the backup. Information is mirrored once per day. The SE servers include N+1 redundant power supplies.

For more information, see 11.2.5, “New SEs” on page 362.

- HMC in an ensemble

The serviceability function for the components of an ensemble is delivered through the traditional HMC/SE constructs, as for earlier Z servers. The primary HMC for the ensemble is where portions of the Unified Resource Manager routines run. The Unified Resource Manager is an active part of the ensemble and z14 infrastructure. Therefore, the HMC is in a stateful state that needs high availability features to ensure the survival of the system during a failure. Each ensemble must be equipped with two HMCs: a primary and an alternative. The primary HMC performs all HMC activities (including Unified Resource Manager activities). The alternative is only the backup and cannot be used for tasks or activities.

Failover: The primary HMC and its alternative must be connected to the same LAN segment. This configuration allows the alternative HMC to take over the IP address of the primary HMC during failover processing.

- Alternative HMC preload function

The Manage Alternate HMC task allows you to reload internal code onto the alternative HMC to minimize HMC downtime during an upgrade to a new driver level. After the new driver is installed on the alternative HMC, it can be made active by running an HMC switchover.



Environmental requirements

This chapter describes the environmental requirements for IBM z14 Model ZR1 servers. It also lists the dimensions, weights, power, and cooling requirements that are needed to plan for the installation of an z14 ZR1 server.

The following features are available for physically installing the server:

- ▶ Air cooling
- ▶ Installation on a raised floor or non-raised floor
- ▶ I/O and power cables can exit under the raised floor or off the top of the server frame
- ▶ AC power supply

For more information about physical planning, see *IBM 3907 Installation Manual for Physical Planning*, GC28-6974.

This chapter includes the following topics:

- ▶ 10.1, “Power and cooling” on page 344
- ▶ 10.2, “Physical specifications” on page 347
- ▶ 10.3, “Physical planning” on page 348
- ▶ 10.4, “Energy management” on page 353

10.1 Power and cooling

The z14 ZR1 server is a 19-inch, single-frame system that supports installation on a raised floor or non-raised floor.

The following features are available to allow I/O cables and line cords to exit the frame from the top, bottom, or both sides of the frame:

- ▶ Top Exit cabling feature (FC 7917). For more information, see 10.3.2, “Top Exit cabling feature (optional)” on page 350.
- ▶ Bottom Exit cabling feature (FC 7919). For more information, see 10.3.4, “Bottom Exit cabling feature” on page 352.

10.1.1 Power requirements and consumption

This section describes the power requirements and consumption for z14 ZR1 servers.

Power requirements

The z14 ZR1 is designed with a fully redundant power system. It has one or two intelligent Power Distribution Unit (PDU) pairs to power the components of the server. To make full use of the redundancy that is built into the server, the PDUs within one pair must be powered from different power distribution panels. In that case, if one PDU in a pair fails, the second PDU ensures continued operation of the server without interruption.

The second PDU pair is installed when the second PCIe+ I/O drawer or the 16U Reserved feature is ordered. For more information, see Figure 10-1 and Table 10-1 on page 345.

Power cords for the PDUs are attached to 1-phase, 50/60 Hz, 200 - 240 V AC power.

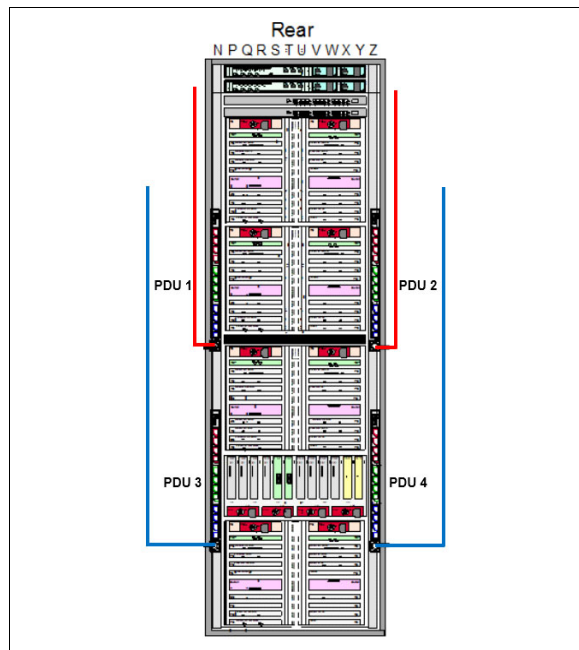


Figure 10-1 The PDU pairs in the rear of the frame

For non-IBM Z components in the optional 16U Reserved rack space, such as the rack-mounted Hardware Management Console (HMC), and its Keyboard/Mouse/Monitor (KMM) assembly, use the appropriate PDU outlets.

For more information, see *IBM 3907 Installation Manual for Physical Planning*, GC28-6974.

Table 10-1 PDU pairs driven by I/O Drawers and 16U Reserved feature

PCIe+ I/O drawers	0	1	2	3	4
PDU pairs	1	1	2	2	2
<p>Notes: Adding the 16U Reserved feature (FC 0617) to the rack always drives the second PDU pair.</p> <p>Each PDU features one 30 A line cord. Therefore, the number of line cord pairs is the same as the number of PDU pairs.</p>					

Power consumption

The maximum power consumption for the z14 ZR1 is listed in Table 10-2.

Table 10-2 Power consumption (kW)

CPC drawer feature	Number of PCIe+ I/O drawers				
	0	1	2	3	4
Max4 (FC 0636)	1.36	2.26	-	-	-
Max12 (FC 0637)	1.77	2.67	3.58	-	-
Max24 (FC 0638)	2.59	3.48	4.39	5.29	6.12
Max30 (FC 0639)	2.59	3.48	4.39	5.29	6.12
<p>Note: The power consumption numbers that are listed in this table assume that the CPC drawer and PCIe+ I/O drawers include the maximum power features (that is, memory and I/O adapters and fanouts). Also assumed is that the system is running at the maximum allowable ambient temperature.</p>					

Considerations: The total power capacity that is available for the non-IBM Z components in the 16U space is 3400 W (for more information, see Appendix G, “16U Reserved feature” on page 465).

Power consumption is lower in a normal ambient temperature room, and for configurations that feature a lesser number of I/O slots, smaller amount of memory, and fewer PUs.

Power estimation for any configuration, power source, and room condition can be obtained by using the power estimation tool that is available at the [IBM Resource Link website](#) (login required).

On the Resource Link page, click **Tools** → **Power and weight estimation**.

10.1.2 Cooling requirements

The z14 ZR1 servers are air-cooled. They require chilled air, ideally coming from under a raised floor, to fulfill the air-cooling requirements. However, a non-raised floor option is available.

The front-to-rear airflow within the system is regulated by intake fans in the front of the machine for the CPC and the PCIe+ drawers. Therefore, all cabling exits the z14 ZR1 from the rear of the machine.

z14 ZR1 servers include a recommended (long-term) ambient temperature range of 18°C (64.4°F) - 27°C (80.6°F). The minimum allowed ambient temperature is 5°C (41°F) and the maximum allowed temperature is 40°C (104°F).

Consideration: When the 16U Reserved feature (FC 0617) is installed, all customer-installed hardware in those 16U must meet the front-to-rear airflow requirement.

All z14 ZR1 components meet ASHRAE A3 environmental class level. Therefore, any non-IBM Z components that are installed in the 16U Reserve space of the rack with ASHRAE class below A3 lower or restrict the ASHRAE class of the full rack.

When two or more PCIe+ drawers are installed, or the 16U Reserved is ordered, more hardware in the form of a cable management spine is installed to route cabling through the frame without blocking the front-to-rear airflow.

For more information about the environmental specifications, see *IBM 3907 Installation Manual for Physical Planning*, GC28-6974.

Rack placement

The z14 ZR1 system is built in an IBM 19-inch rack that fits most standardized data centers and simplify installation.

A generic example of hot and cold airflow and the arrangement of server aisles is shown in Figure 10-2.

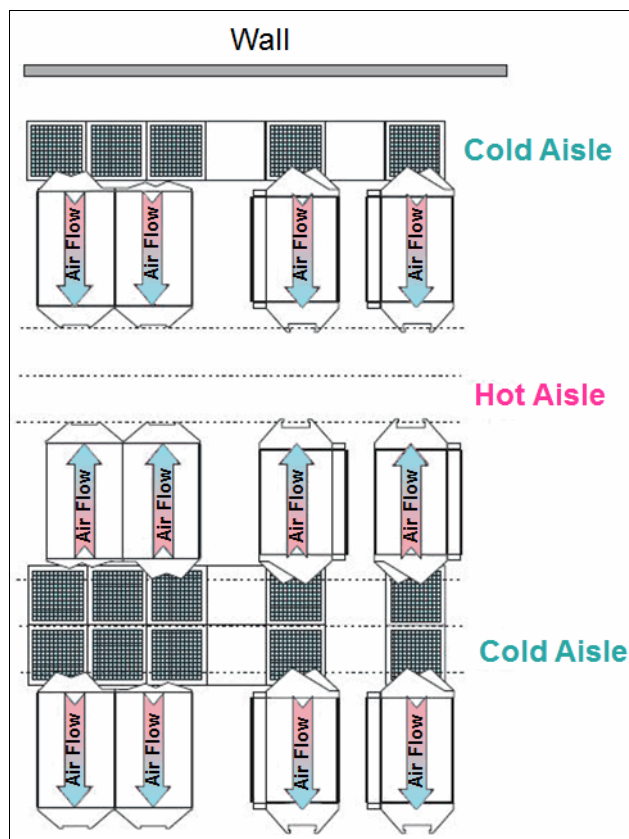


Figure 10-2 Hot and cold aisles

As shown in Figure 10-2, rows of servers must be placed front-to-front. Chilled air is provided through perforated floor panels that are placed in rows between the fronts of servers (the cold aisles). 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 *IBM 3907 Installation Manual for Physical Planning*, GC28-6974.

10.2 Physical specifications

This section describes the weights and dimensions of z14 ZR1 server. The z14 ZR1 is the first IBM Z server that is based on a 19-inch rack; therefore, dimensions and weights are significantly different from previous generations of IBM Z servers.

The z14 ZR1 can be installed on a raised or non-raised floor. For more information about weight distribution and floor loading tables, see the *IBM 3907 Installation Manual for Physical Planning*, GC28-6974. This data is used with the maximum frame weight, frame width, and frame depth to calculate the floor loading.

Note: Components that are installed in the 16U Reserve feature space must not weigh more than 20.4 kg (45 lbs) per EIA location. For example, a 4U unit can weigh up to 81.65 kg (180 lbs).

The power and weight estimation tool for Z servers on Resource Link covers the estimated weight for your designated configuration. The tool is available for download from the [IBM Resource Link website](#) (login required).

On the Resource Link page, click **Tools** → **Power and weight estimation**.

10.3 Physical planning

This section describes the floor mounting, power, and I/O cabling options. For more information, see the *IBM 3907 Installation Manual for Physical Planning*, GC28-6974.

10.3.1 Raised floor or non-raised floor

z14 ZR1 servers can be installed on a raised or non-raised floor. The following options and features are available for I/O cabling and line cords.

Note: On the z14 ZR1, all I/O cabling and line cords come from the rear of the machine; therefore, all related features for Bottom and Top Exit cabling are in the rear of the frame.

Raised floor

If the z14 ZR1 server is installed in a raised floor environment, the following top and bottom exit features or options are available to route I/O cables and line cords:

- ▶ Optional¹ Top Exit cabling feature (FC 7917)
- ▶ Route cabling directly through the top of the frame
- ▶ Bottom Exit cabling feature (FC 7919)

¹ For more information, see 10.3.2, “Top Exit cabling feature (optional)” on page 350.

The Top Exit and Bottom Exit cabling feature options of z14 ZR1 servers in a raised floor environment are shown in Figure 10-3.

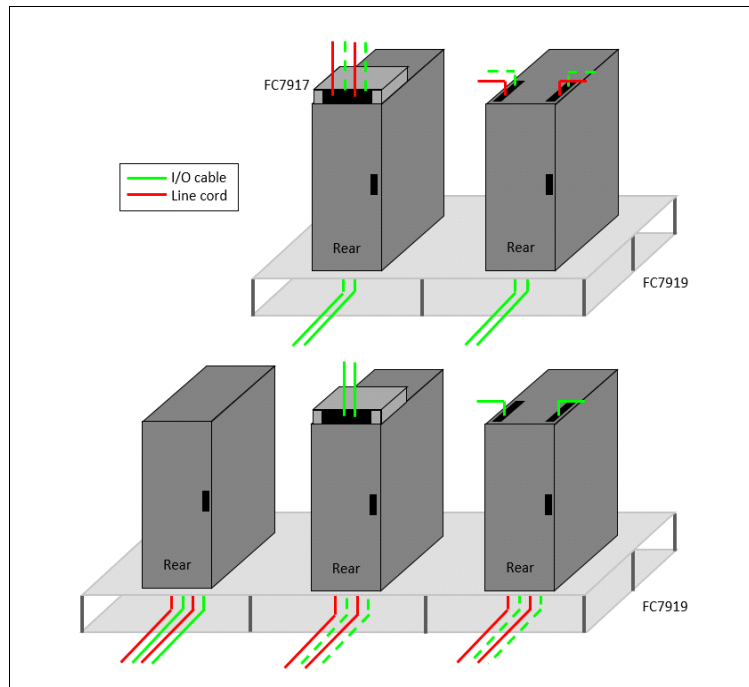


Figure 10-3 Raised floor options

Non-raised floor

If you install the z14 ZR1 server in a non-raised floor environment, you can select the optional² Top Exit cabling feature (FC 7917) or you can directly route the cabling through the top of the frame. All cables must exit from the top of the z14 ZR1 server, as shown in Figure 10-4.

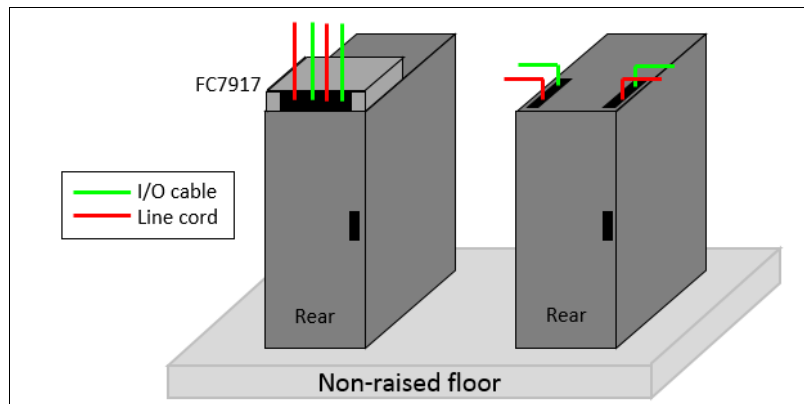


Figure 10-4 Non-raised floor options

² For more information, see 10.3.2, “Top Exit cabling feature (optional)”.

10.3.2 Top Exit cabling feature (optional)

The optional Top Exit cabling feature (FC 79176) allows for I/O cabling and line cords to exit the top of the frame. This feature adds cable management options, such as trunking and retainer brackets, as shown in Figure 10-5. The Top Exit cabling feature can be placed as shown in Figure 10-5, with the exit area towards the front of the frame, or with the exit area towards the rear of the frame.

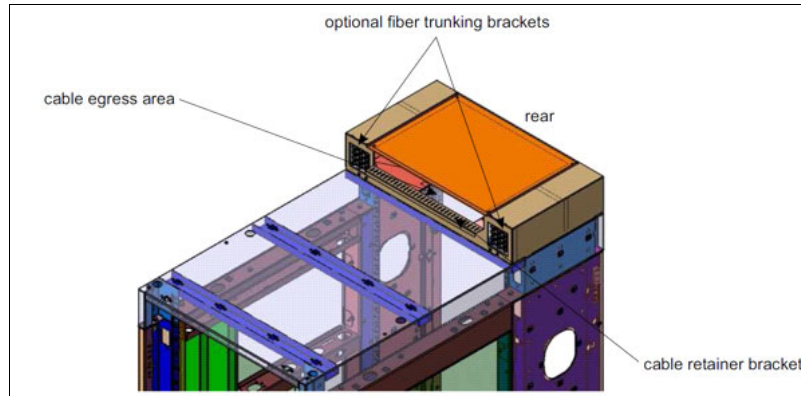


Figure 10-5 Top Exit cabling feature

The Top Exit cabling feature adds 117.5 mm (4.63 in.) to the height of the frame and approximately 5.4 kg (12 lbs) to the weight.

If the Top Exit cabling feature is not ordered, two sliding plates are available on the top of the frame (one on each side of the rear of the frame) that can be partially opened. By opening these plates, I/O cabling and line cords can exit the frame, as shown in Figure 10-6.



Figure 10-6 Sliding panels (no Top Exit cabling feature)

10.3.3 Top or bottom exit cables

Features allow for Top Exit Cabling (FC 7917) or Bottom Exit Cabling (FC 7919) cabling, or a combination of both. These features are independent of raised floor or non-raised floor installations and offer flexible possibilities for the data center.

All external cabling enters the rear of the rack from under floor or from above the rack. Different from previous Z Systems, no cabling access or cable plugging is available at the front of the rack. The top view of the rack with and without FC 7917 is shown in Figure 10-7.

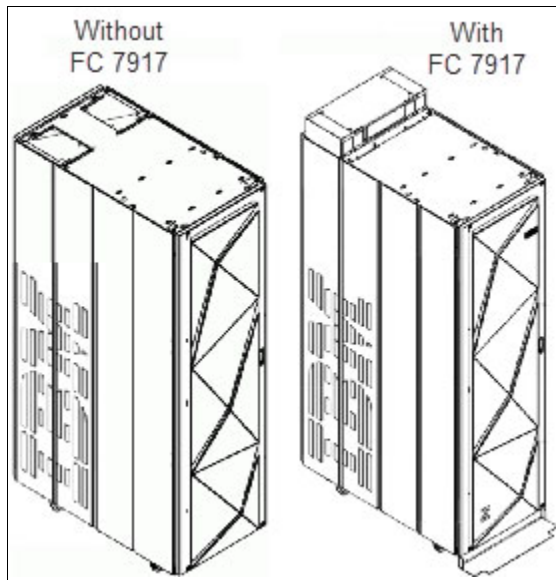


Figure 10-7 ZR1 rack with and without the Top I/O Exit feature

The Top Exit Cabling feature provides new hardware. The new hardware resembles a rectangular box with an open side that faces the front or rear of the rack. It includes other hardware to organize and fasten cables.

The Top Exit Cabling option can be used for routing power cables and IO cables out the top of the machine.

Without the Top Exit Cabling feature, power and cables still can be run out the top of the rack through two adjustable openings at the top rear of the rack, as shown on the left side of Figure 10-7.

The Bottom Exit Cabling feature provides tailgate hardware for routing power cables or IO cables out the bottom of the machine.

For more information, see *IBM 3907 Installation Manual for Physical Planning*, GC28-6974, and 10.3, “Physical planning” on page 348.

10.3.4 Bottom Exit cabling feature

The Bottom Exit cabling feature (FC 7919) is required for raised floor environments, where I/O cabling or line cords must exit from the bottom of the frame. This feature includes the hardware to allow bottom exit, and other components for cable management and filler plates to preserve the recommended air circulation, as shown in Figure 10-8.

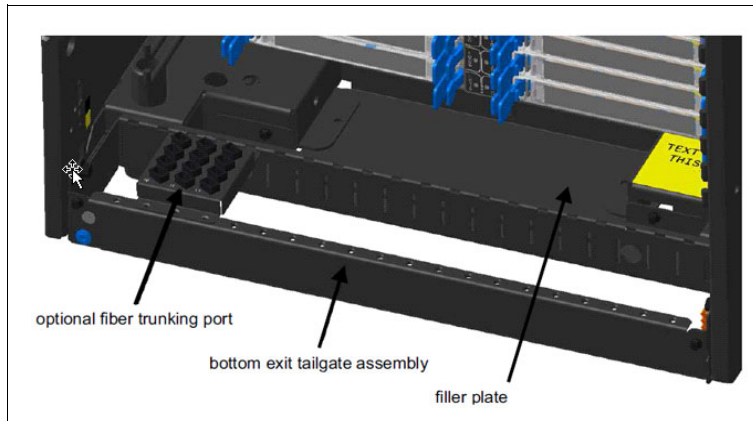


Figure 10-8 Bottom Exit cabling feature

10.3.5 Frame Bolt-down kit

A bolt-down kit (FC 8006) is available for the z14 ZR1 servers. The kit provides hardware to enhance the ruggedness of the frame, the frame stiffener, and to tie down the frame to a concrete floor.

The frame tie-down kit can be used on a non-raised floor where the frame is secured directly to a concrete floor, or on a raised floor where the frame is secured to the concrete floor underneath the raised floor. Raised floors 241.3 mm (9.5 inches) - 1270 mm (50 inches) are supported.

The kits help secure the frames and their contents from damage when they are exposed to shocks and vibrations, such as in a seismic event. The frame tie-downs are intended for securing a frame that weighs up to 1308 kg (2885 lbs).

For more information see *IBM 3907 Installation Manual for Physical Planning*, GC28-6974.

10.3.6 Service clearance areas

z14 ZR1 servers require specific service clearance to ensure the fastest possible repair in the unlikely event that a part must be replaced. Failure to provide enough clearance to open the front and rear covers results in extended service times or outages.

For more information, see *IBM 3907 Installation Manual for Physical Planning*, GC28-6974.

10.4 Energy management

This section describes the elements of energy management 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 10-9.

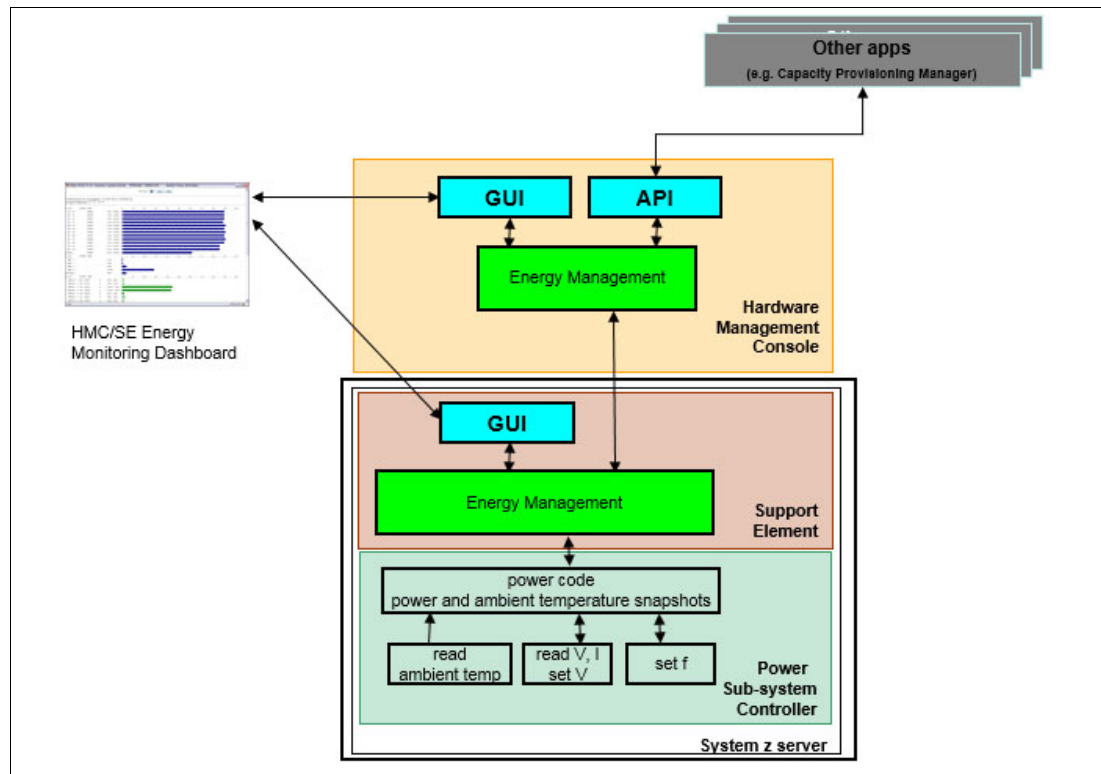


Figure 10-9 z14 energy management

The hardware components in the z14 ZR1 server are monitored and managed by the energy management component in the Support Element (SE) and HMC. The graphical user interfaces (GUIs) of the SE and HMC provide views, such as the Monitors Dashboard and Environmental Efficiency Statistics Monitor Dashboard.

The following tools are available to plan and monitor the energy consumption of z14 ZR1 servers:

- Power estimation tool on Resource Link
- Energy Management task for maximum potential power on HMC and SE
- Monitors Dashboard and Environmental Efficiency Statistics tasks on HMC and SE

10.4.1 Environmental monitoring

This section describes energy monitoring HMC and SE tasks.

Monitor task group

The Monitor task group on the HMC and SE includes monitoring-related tasks for z14 ZR1 servers, as shown in Figure 10-10.

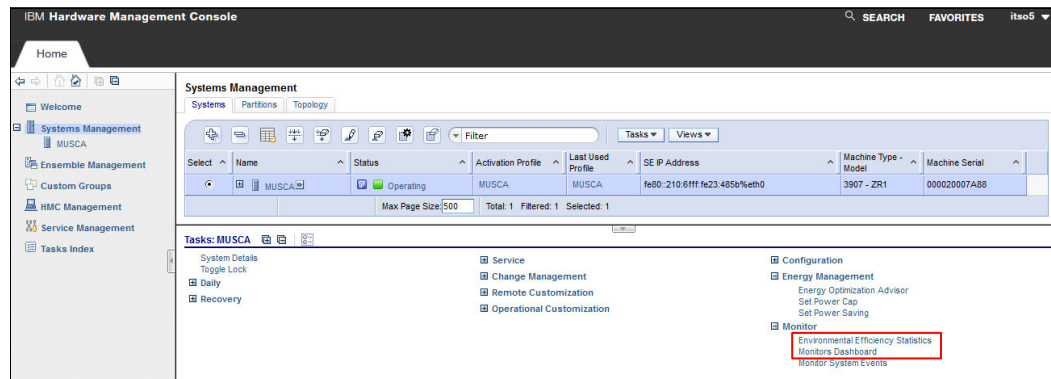


Figure 10-10 HMC Monitor task group

Monitors Dashboard task

In z14 ZR1 servers, the Monitors Dashboard task in the Monitor task group provides a tree-based view of resources. Multiple graphical views display data, including history charts. This task monitors processor and channel usage. It produces data that includes power monitoring information, power consumption, and the air input temperature for the server.

An example of the Monitors Dashboard task is shown in Figure 10-11.

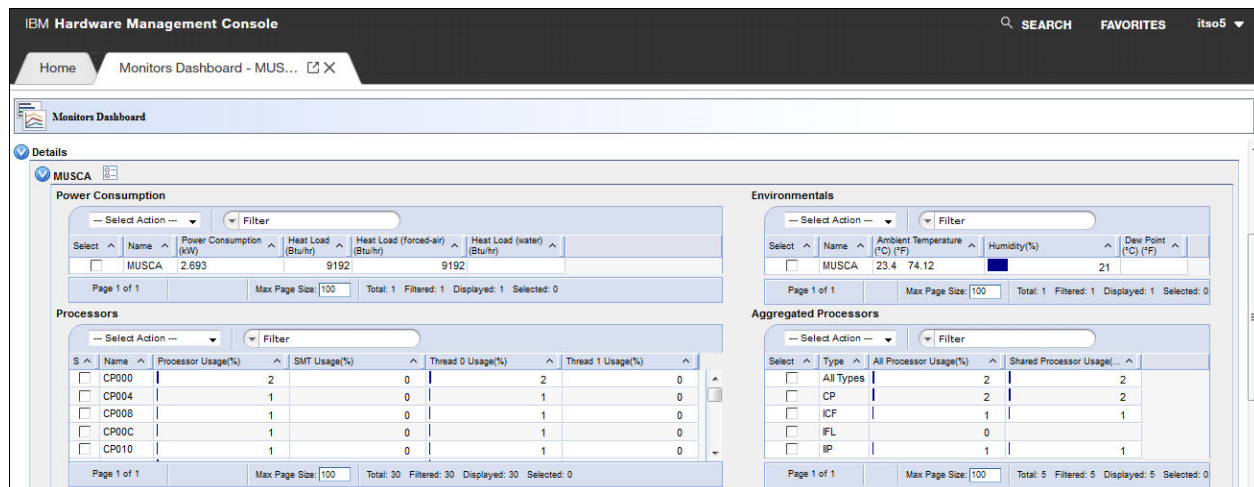


Figure 10-11 Monitors Dashboard task

Environmental Efficiency Statistics task

The Environmental Efficiency Statistics task (see Figure 10-12) is part of the Monitor task group. It provides historical power consumption and thermal information for the CPC.

The data is presented in table format and graphical “histogram” format. The data can also be exported to a .csv-formatted file so that the data can be imported into a spreadsheet. For this task, you must use a web browser to connect to an HMC.

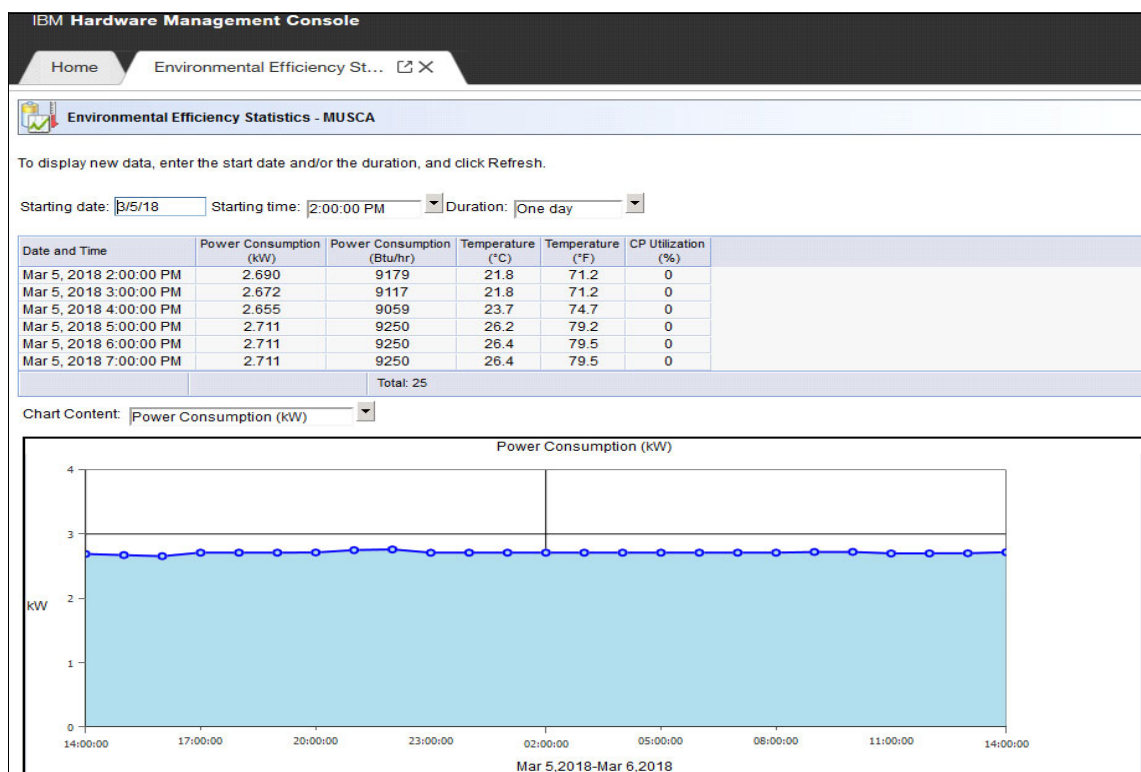


Figure 10-12 Environmental Efficiency Statistics task



Hardware Management Console and Support Elements

The Hardware Management Console (HMC) supports many functions and tasks to extend the management capabilities of IBM z14 ZR1. 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 complex (CPC).

This chapter describes the newest elements for the HMC and SE.

Note: The Help function is a good starting point to get more information about all of the functions that can be used by the HMC and SE. The Help feature is available by clicking **Help** from a drop-down menu that appears when you click your user ID.

For more information, see [IBM Knowledge Center](#).

This chapter includes the following topics:

- ▶ 11.1, “Introduction to the HMC and SE” on page 358
- ▶ 11.2, “HMC and SE changes and new features” on page 358
- ▶ 11.3, “HMC and SE connectivity” on page 367
- ▶ 11.4, “Remote Support Facility” on page 374
- ▶ 11.5, “HMC and SE capabilities” on page 377

11.1 Introduction to the HMC and SE

The HMC is a stand-alone computer that runs a set of management applications. The HMC is a closed system, which means that no other applications can be installed on it.

The HMC is used to set up, manage, monitor, and operate one or more CPCs. It manages IBM Z hardware, its logical partitions (LPARs), and provides support applications. At least one HMC is required to operate an IBM Z. An HMC can manage multiple Z CPCs, and can be at a local or a remote site.

The SEs are two integrated servers in the z14 ZR1 frame. One SE is the primary SE (active) and the other is the alternative SE (backup). As with the HMCs, the SEs are closed systems, and no other applications can be installed on them.

When tasks are performed at the HMC, the commands are routed to the active SE of the z14 ZR1. The SE then issues those commands to their CPC. One HMC can control up to 100 SEs and one SE can be controlled by up to 32 HMCs.

Some functions are available only on the SE. With Single Object Operations (SOOs), these functions can be used from the HMC. For more information, see “Single Object Operations” on page 376.

With Driver 27 (Version 2.13.1), the IBM Dynamic Partition Manager (DPM) was introduced for CPCs that are running Linux only with Fibre Channel Protocol (FCP) attached storage. HMC Driver 32 (Version 2.14.0) with MCLs added support for ECKD FICON disks to the DPM (Release 3.1). HMC 2.14.1 includes DPM 3.2, with enhanced storage management capabilities. DPM is a mode of operation that enables customers with little or no knowledge of IBM Z technology to set up the system efficiently and with ease.

For more information, see [IBM Knowledge Center](#). At IBM Knowledge Center, click the search engine window and enter DPM.

The HMC Remote Support Facility (RSF) provides an important communication to a centralized IBM support network for hardware problem reporting and service. For more information, see 11.4, “Remote Support Facility” on page 374.

11.2 HMC and SE changes and new features

The initial release that is included with z14 ZR1 is HMC application Version 2.14.0. Use the “What’s New” task to examine the new features that are available for each release. For more information about HMC and SE functions, use the HMC and SE (Version 2.14.0) console help system or see [IBM Knowledge Center](#).

At IBM Knowledge Center, search for “z14 HMC”.

11.2.1 Driver Level 36 HMC and SE new features

The following support has been added with Driver 36:

- ▶ Dynamic I/O for Standalone CF CPCs (requires z/OS or z/VM support)
- ▶ CTN Split and CTN merge
- ▶ Coupling Facility Control Code 23 (enhancements and new features)

- ▶ Various OSA-ICC 3270 enhancements:
 - IPv6 support
 - TLS level limits negotiation for secure OSA-ICC connections
 - Separate security certificates management (per PCHID)
- ▶ Support Element remote logging from HMC handling (Single Object Operations -SOO) enhancements
- ▶ Help infrastructure updates

The content from the following publications is incorporated into the HMC and SE help system:

 - *IBM Z Hardware Management Console Operations Guide Version 2.14.1*
 - *IBM Z Hardware Management Console Operations Guide for Ensembles Version 2.14.1*
 - *IBM Z Support Element Operations Guide Version 2.14.1*

11.2.2 Driver Level 32 HMC and SE changes and features

The z14 HMC and SE with Driver Level 32 features the following enhancements and changes:

- ▶ Classic User Interface style no longer supported. The z14 HMC and SE support only tree Style User Interface.
- ▶ New set of Enhanced Computing features is implemented for tampering protection. For more information see , “For more information, see the HMC and SE (Version 2.14.0) console help system or see IBM Knowledge Center. At IBM Knowledge Center, click IBM Z. Then, click z14.” on page 360.
- ▶ Starting with version 2.13.1, HMC Tasks no longer include Java Applets-based implementations. Java Applets were used in Operating System Messages, Integrated 3270 Console, Integrated ASCII, and Text Console.
- ▶ Version 2.14.0 implements new IOCDS Source option on Input/Output configuration task. This option enables you to edit IOCDS source directly on HMC Console. However, an alternative method with remote browsing on Support Element is still available.
- ▶ Starting with z14, all FTP operations that originate from the SE are proxied through a managing HMC. This change allows the FTP SE-originated operations to follow our security recommendation.

In addition, all HMC/SE tasks that support FTP provide three options of FTP: FTP, FTPS, and SFTP. For more information, see 11.3.1, “Network planning for the HMC and SE” on page 369.

- ▶ Secure console-to-console communication was established for z14 HMC consoles, which instituted new security standards (even on internal communication). For more information, see 11.3.1, “Network planning for the HMC and SE” on page 369.
- ▶ Functional enhancements are included in SNMP/BCPii API interfaces, such as queries to Virtual Flash Memory or queries to Secure Service Container. The security of the BCPii interface was also enhanced. You can disable BCPii's sending or receiving capability for each partition. The cross-partition authority setting on SE remains the same.
- ▶ Remote Browser IP Address Limiting function was implemented because of security reasons. It allows you to specify a valid remote browser IP or valid mask for a group of IP addresses. Global settings to enable and disable remote access are still available.

- ▶ Multi-factor authentication was implemented for z14 HMC/SE/TKE. The feature enables you to log in with higher security levels by using two factors: traditional login and password and a passcode that is sent on your smartphone. For more information, see 11.3.6, “HMC Multi-factor authentication” on page 373.
- ▶ The HMC Global OSA/SF now provides a global view of all OSA PCHIDs and the monitoring and diagnostic information that was available in the Query Host command. For more information, see 11.3.4, “OSA Support Facility changes” on page 372.
- ▶ Compliance mode for CCA PCI-HSM and EP11 and other certificates is now displayed on the SE. Setup and administration tasks are done on Trusted Key Entry (TKE). For more information, see 11.5.14, “Cryptographic support” on page 393.
- ▶ Enhancements and changes were made for the Server Time Protocol (STP), which are described in 11.5.7, “Server Time Protocol support” on page 386.
- ▶ A new Mobile application interface is provided for the HMC 2.14.0 and systems, including z14, z13/z13s, and zEC12/zBC12, which includes security technology. For more information, see 11.4.3, “HMC and SE remote operations” on page 375.

For more information, see the HMC and SE (Version 2.14.0) console help system or see [IBM Knowledge Center](#). At IBM Knowledge Center, click **IBM Z**. Then, click **z14**.

11.2.3 Firmware Integrity Monitoring and z14 HMC

Firmware Integrity Monitoring is a set of security features that were implemented with z14 HMC to improve its resistance to security attacks. z14 HMC consoles, TKE devices, and SEs are firmware-compliant with NIST Computer Security Standard 800-147, which results in the following benefits:

- ▶ Current IBM Z Firmware is protected during delivery by using Digital Signatures.
- ▶ BIOS Secure boot function is used by SE/HMC/TKE.
- ▶ Signature and hash verification of SE/HMC IBM Z firmware.
- ▶ Non-stop monitoring and checking the integrity of files.
- ▶ Code measurements are stored in Trusted Platform Module (TPM) on SE and HMC.
- ▶ Provides security logs for internal analysis.
- ▶ Trusted third-Party Validation (IBM Resource Link by using zRSF data).
- ▶ Analyzes periodic call home measurement data.
- ▶ Initiates challenge and response to verify authenticity of the data.
- ▶ Display of local console data analysis, Resource Link analysis, and notification of lack of receiving console data (console locked or blocked network reporting of data).

Firmware tamper detection

z14 ZR1 also offers an enhancement on the SE that provides notification if tampering with booting of firmware on the system (CPC) is detected. This enhancement meets the BIOS Protection Guidelines recommended and published by the National Institute of Standards and Technology (NIST) in Special Publication 800-147B. If tampering is detected, the SE issues a customer alert by using a warning or a lock of the SE, depending on the configuration.

11.2.4 Rack-mounted HMC

Feature code FC 0083 provides a rack-mounted HMC.

The HMC is a 1U IBM server and an optional IBM 1U standard tray that features a monitor and a keyboard. The system unit and tray must be mounted in the rack in two adjacent 1U locations in the “ergonomic zone” between 21U and 26U in a standard 19-inch rack.

The customer must provide the rack or it can be installed in the 16U reserved space of the z14 ZR1. Three C13 power receptacles are required: two for the system unit and one for the display and keyboard, as shown in Figure 11-1.

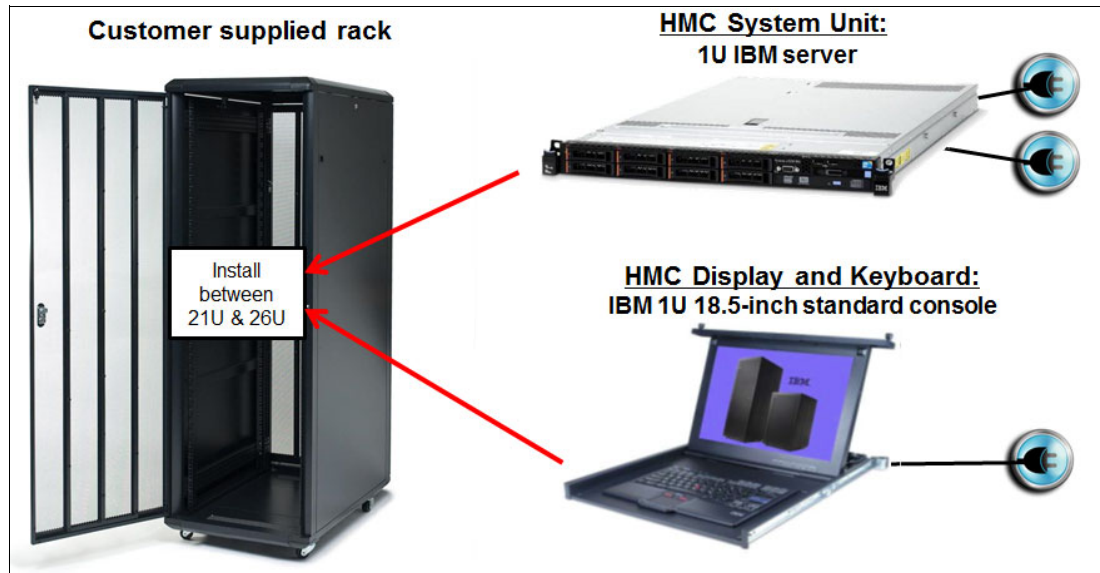


Figure 11-1 Rack-mounted HMC installed in an extra rack

11.2.5 New SEs

The SEs are no longer two notebook computers in the z14 ZR1. Instead, two servers are now installed at the top of the frame. They are managed by the keyboard, pointing device, and display that are installed in the EIA unit 22 in the front of the frame, as shown in Figure 11-2. The connection between the two SEs and the Monitor/Keyboard unit is created by way of a KVM switch that is installed at the same location (EIA unit 22) as Monitor/Keyboard, but in the rear.

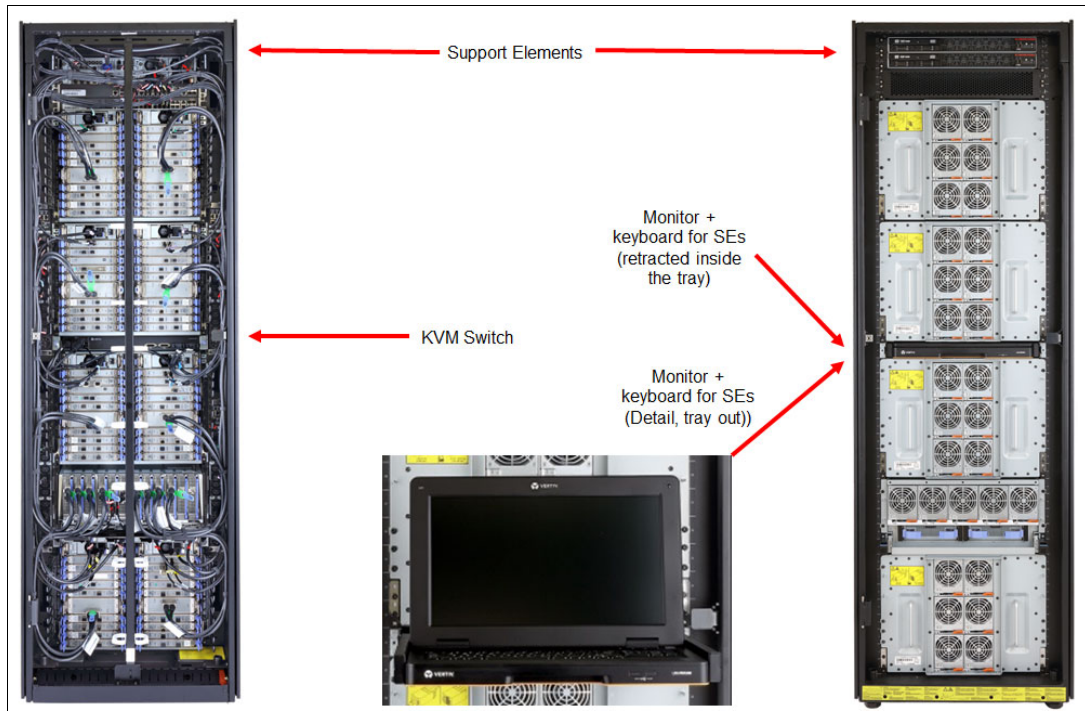


Figure 11-2 SEs location

11.2.6 New backup options for HMCs and primary SEs

This section describes the new backup options that are available for HMC Version 2.14.0.

Backup of primary SEs or HMCs to an FTP server

With Driver 32 or later, you can perform a backup of primary SEs or HMCs to a File Transfer Protocol (FTP) server. Starting with z14 systems, three FTP options are available: FTP, FTPS, or SFTP. For more information, see 11.3.1, “Network planning for the HMC and SE” on page 369.

Note: If you back up to an FTP server for a z14 ZR1, ensure that you set up a connection to the FTP server by using the Configure Backup Setting task. If a connection to the FTP server is not set up, a message appears that prompts you to configure the connection.

The FTP server must be supplied by the customer. You can enable a secure FTP connection to your server.

The information that is required to configure your backup FTP server is shown in Figure 11-3.

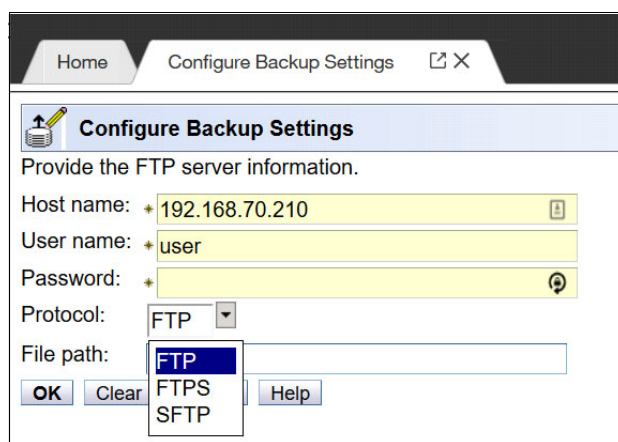


Figure 11-3 Configure Backup Settings

Note: Backup FTP site is a static setting for an HMC. If an alternative FTP site is needed to perform a backup, this process is done from another HMC.

Backing up HMCs

A backup of the HMC can be performed to the following media:

- ▶ USB flash memory drive (UFD)
- ▶ FTP server
- ▶ UFD and FTP server

The destination options of the Backup Critical Console Data task are shown in Figure 11-4.

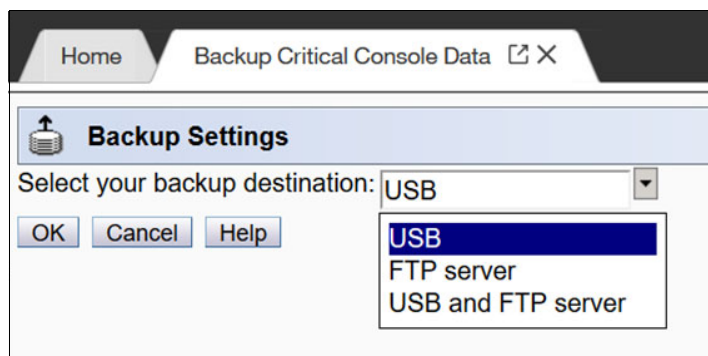


Figure 11-4 Backup Critical Console Data destinations

Optional 32 GB UFD FC 0848

With z14 ZR1, optional 32 GB UFDs are available for backups. An 8 GB UFD is included by default with the system. SE and HMC backup files are usually larger in later IBM Z, but depend also on configuration.

Backup of primary SEs

The backup for the primary SE of a z14 ZR1 can be made to the following media:

- ▶ Primary SE HDD and alternative SE HDD
- ▶ Primary SE HDD and alternative SE HDD and FTP server

It is no longer possible to complete the primary SE backup to an UFD of a z14 ZR1. The SE Backup options for external media are listed in Table 11-1.

Table 11-1 SE Backup options

System type	UFD media	FTP server
z14 (3906/3907)	No	Yes
z13/z13s (2964/2965)	No	Yes
zBX 004	No	Yes
zEC12/zBC12	Yes	No
z196/z114	Yes	No
z10EC/z10BC	Yes	No
z9EC/z9BC	Yes	No

Examples of the different destination options of the SE Backup Critical Data for different CPC machine types are shown in Figure 11-4.

For more information, see the HMC and SE console help function or [IBM Knowledge Center](#).

Scheduled operations for the backup of HMCs and SEs

The Scheduled Operation task with the new backup options for HMC changed are shown in Figure 11-5.

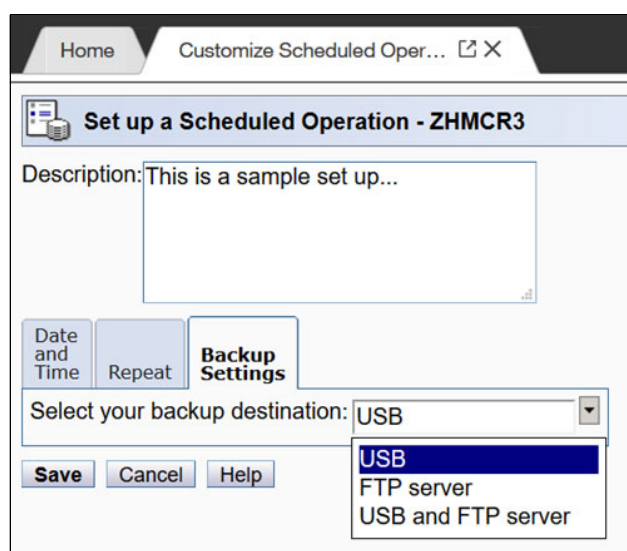


Figure 11-5 Scheduled Operation for HMC backup

The Scheduled Operation task with the new backup options for the SEs changed are shown in Figure 11-6.

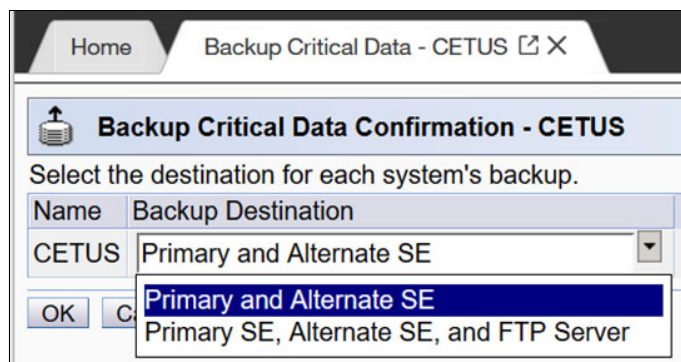


Figure 11-6 Scheduled Operation for SEs backup

11.2.7 SE driver support with the HMC driver

HMC legacy systems support (Statement of Direction^a): IBM z14 is planned to be the last release that will allow HMC support across the prior four generations of server (N through N-4).

Future HMC releases are intended to be tested for support of the prior two generations (N through N-2). For example, the next HMC release would support the zNext generation, plus z14 generation and z13/z13s generation.

This change will improve the number and extent of new features and functions that can be pre-tested and maintained in a given release with IBM's continued high-reliability qualification procedures.

- a. All statements regarding IBM plans, directions, and intent are subject to change or withdrawal without notice. Any reliance on these statements of general direction is at the relying party's sole risk and will not create liability or obligation for IBM.

The driver of the HMC and SE is equivalent to a specific HMC and SE version, as shown in the following examples:

- ▶ Driver 79 is equivalent to Version 2.10.2
- ▶ Driver 86 is equivalent to Version 2.11.0
- ▶ Driver 93 is equivalent to Version 2.11.1
- ▶ Driver 15 is equivalent to Version 2.12.1
- ▶ Driver 22 is equivalent to Version 2.13.0
- ▶ Driver 27 is equivalent to Version 2.13.1
- ▶ Driver 32 is equivalent to Version 2.14.0
- ▶ Driver 36 is equivalent to Version 2.14.1

An HMC with Version 2.14.1 or Version 2.14.0 can support different IBM Z types. Some functions that are available on Version 2.14.1 and later are supported only when the HMC is connected to an IBM Z system with Version 2.14.1.

The SE drivers and versions that are supported by the z14 ZR1 HMC Version 2.14.0 (Driver 32) and earlier versions are listed in Table 11-2.

Table 11-2 Summary of SE drivers

IBM Z family name	Machine type	SE driver	SE version	Ensemble node potential
z14	3906, 3907	32, 36	2.14.0, 2.14.1	Yes ^a
z13s	2965	27	2.13.1	Yes ^a
z13	2964	22, 27	2.13.0, 2.13.1	Yes ^a
zBX Node	2458 Model 004	22	2.13.0	Required
zBC12	2828	15	2.12.1	Yes
zEC12	2827	15	2.12.1	Yes
z114	2818	93	2.11.1	Yes
z196	2817	93	2.11.1	Yes
z10 BC	2098	79	2.10.2	No
z10 EC	2097	79	2.10.2	No
z9 BC	2096	67	2.9.2	No
z9 EC	2094	67	2.9.2	No

a. A CPC in DPM mode cannot be a member of an ensemble; however, the CPC can still be managed by the ensemble HMC.

Note: The z9 EC / z9 BC (Driver 67, SE version 2.9.2), the z900/z800 (Driver 3G, SE Version 1.7.3) and z990/z890 (Driver 55, SE Version 1.8.2) systems are no longer supported. If you are using these older systems, consider managing these systems by using separate HMCs that are running older drivers.

11.2.8 HMC feature codes

HMCs that are older than FC 0091 are not supported for z14 ZR1 at Driver 32 or 36.

The following HMC feature codes are available:

- ▶ FC 0082 M/T 2461-TW2

This feature is a tower model HMC that works with z14, z13, and z13s systems.

- ▶ FC 0083 M/T 2461-SE1

This feature is the new rack-mounted HMC that works with z14, z13, and z13s systems.

The following previous HMCs can be carried forward (the carry forward HMCs do not provide the Enhanced feature):

- ▶ Tower FC 0092
- ▶ Tower FC 0095
- ▶ 1U Rack FC 0094
- ▶ 1U Rack FC 0096

11.2.9 User interface

Starting with HMC Version 2.14.0, only Tree Style User Interface is available. The Classic Style User Interface was discontinued.

11.2.10 Customize Product Engineering Access: Best practice

At times, the HMC or the SE must be accessed in a support role to perform problem determination.

The task to set the authorization for IBM Product Engineering access to the console is shown in Figure 11-7. When access is authorized, an IBM product engineer can use an exclusive user ID and reserved password to log on to the console that provides tasks for problem determination.

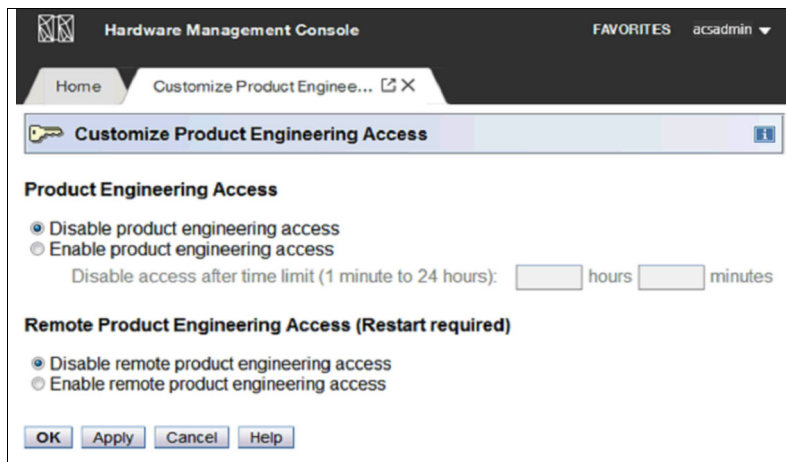


Figure 11-7 Customize Product Engineering Access tab

As shown in Figure 11-7, the task is available only to users with ACSADMIN authority. Consider the following points:

- ▶ Customers must ensure that they have redundant administrator users for each console.
- ▶ Customers must document contact information and procedures.
- ▶ The “Welcome Text” task can be used to identify contact information so that IBM Service personnel know how to engage customer administrators if HMC or SE access is needed.
- ▶ The options are disabled by default.

11.3 HMC and SE connectivity

The HMC has two Ethernet adapters that are supported by HMC Driver 22 or later for connectivity to up to two different Ethernet LANs.

The SEs on z14 ZR1 are connected to the Ethernet switches that are installed at the top of the z14 ZR1 rack, under the SEs. In previous IBM Z systems, the customer network was connected to the bulk power hub (BPH). Now, the SEs are directly connected to the customer network.

The HMC-to-CPC communication is now possible through only an Ethernet switch that is connected to the J03 or J04 port on the SEs. Other IBM Z systems and HMCs also can be connected to the switch. To provide redundancy, install two Ethernet switches.

Only the switch (and not the HMC directly) can be connected to the SEs.

The connectivity between HMCs and the SEs is shown in Figure 11-8.

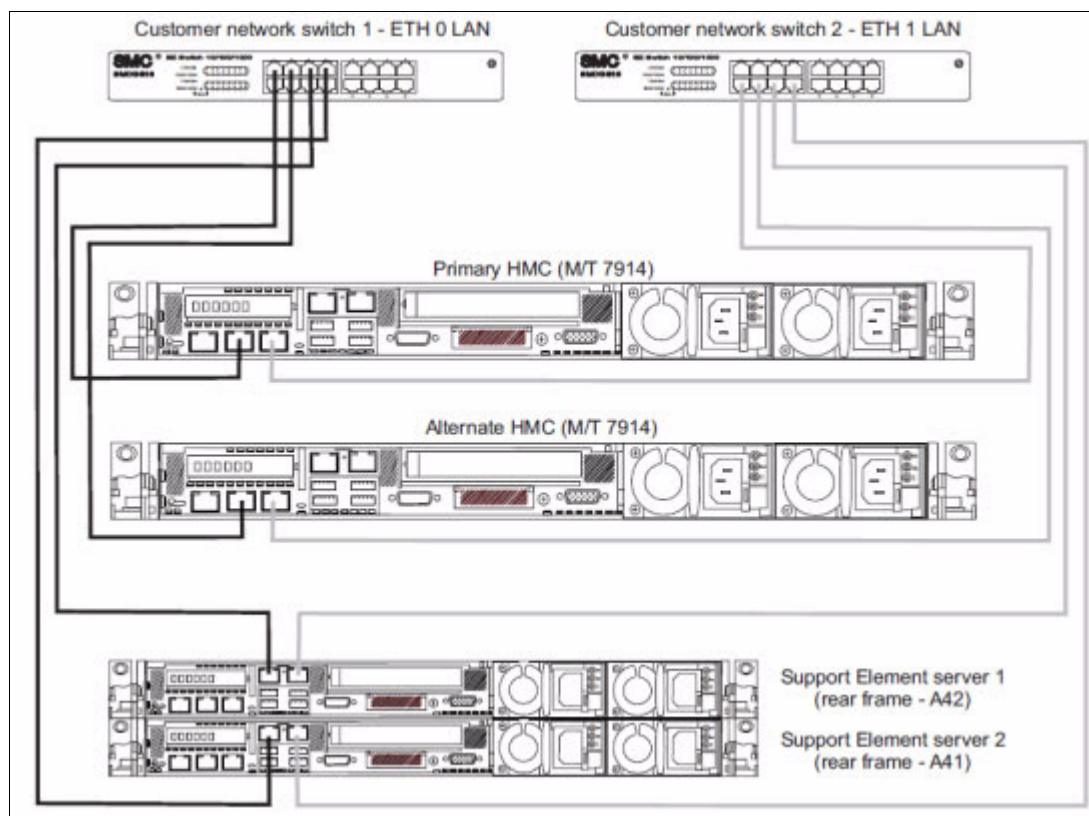


Figure 11-8 HMC and SE connectivity

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 that is based on your connectivity and security requirements.

Security: The configuration of network components, such as routers or firewall rules, is beyond the scope of this book. Whenever the networks are interconnected, security exposures can exist. For more information about HMC security, see *Integrating the Hardware Management Console's Broadband Remote Support Facility into your Enterprise*, SC28-6951.

For more information about the HMC settings that are related to access and security, see the HMC and SE console help function or [IBM Knowledge Center](#).

11.3.1 Network planning for the HMC and SE

Plan the HMC and SE network connectivity carefully to allow for current and future use. Many of the IBM Z capabilities benefit from the various network connectivity options that are available. The following functions, which depend on the HMC connectivity, are available to the HMC:

- ▶ Lightweight Directory Access Protocol (LDAP) support, which can be used for HMC user authentication
- ▶ Network Time Protocol (NTP) client/server support
- ▶ RSF through broadband
- ▶ HMC access through a remote web browser
- ▶ Enablement of the SNMP and CIM¹ APIs to support automation or management applications, such as IBM System Director Active Energy Manager (AEM)

These examples are shown in Figure 11-9.

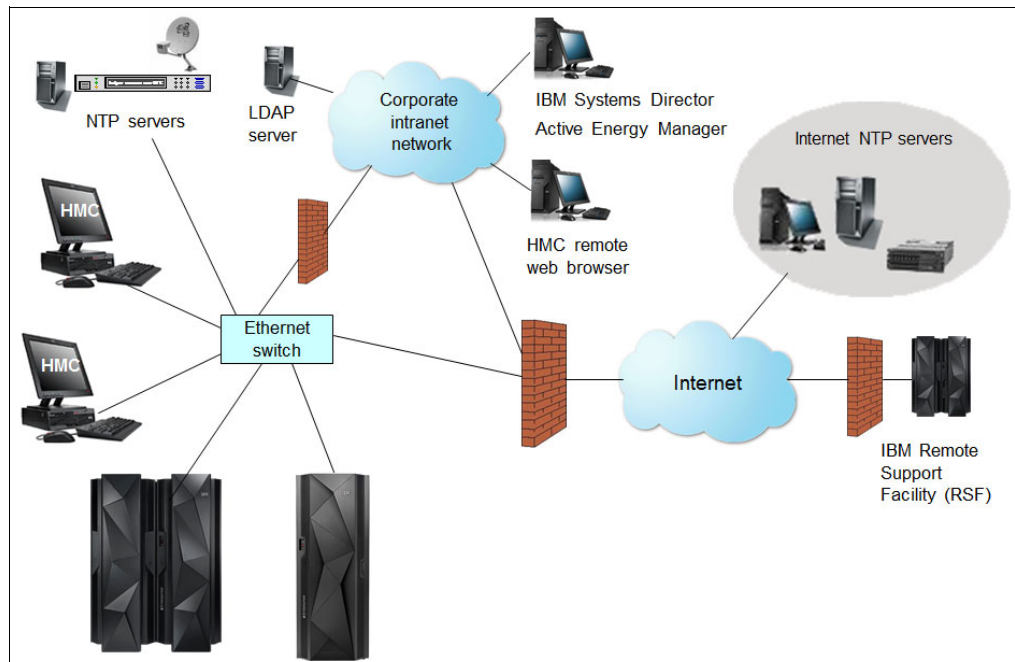


Figure 11-9 HMC connectivity

FTP, FTPS, and SFTP support

FTP, FTPS, and SFTP protocols are now supported on HMC/SE environment. All three FTP protocols require login and password credentials.

FTPS is based on Secure Sockets Layer cryptographic protocol (SSL) and requires certificates to authenticate servers. SFTP is based on Secure Shell cryptographic protocol (SSH) and requires SSH keys to authenticate servers. Required certificates and key pairs are hosted on the z14 HMC Console.

All three protocols are supported for tasks that previously used only FTP. Although several tasks used only removable media, FTP connections are used with z14 HMC console. The recommended network topology for HMC, SE, and FTP server is shown in Figure 11-10 on page 370.

¹ CIM support was removed from the HMC with Version 2.14.0.

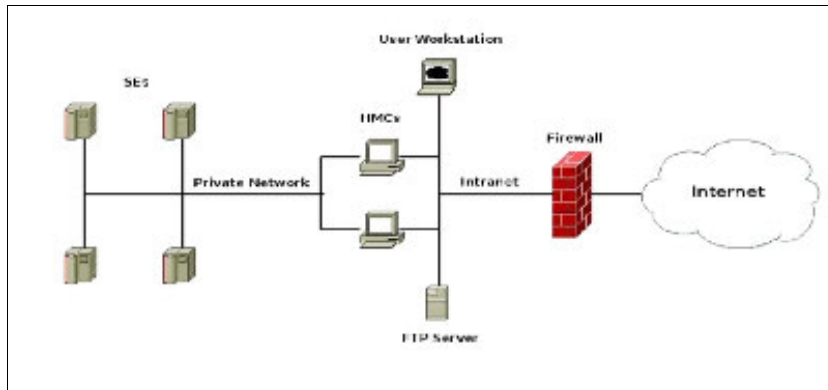


Figure 11-10 Recommended Network Topology for HMC, SE, and FTP server

The following FTP server requirements must be met:

- ▶ Support “passive” data connections
- ▶ A server configuration that allows the client to connect on an ephemeral port

The following FTPS server requirements must be met:

- ▶ Operate in “explicit” mode
- ▶ Allows a server to offer secure and unsecured connections
- ▶ Must support “passive” data connections
- ▶ Must support secure data connections

SFTP server requirements must support password-based authentication.

The FTP server choices for HMC are shown in Figure 11-11.

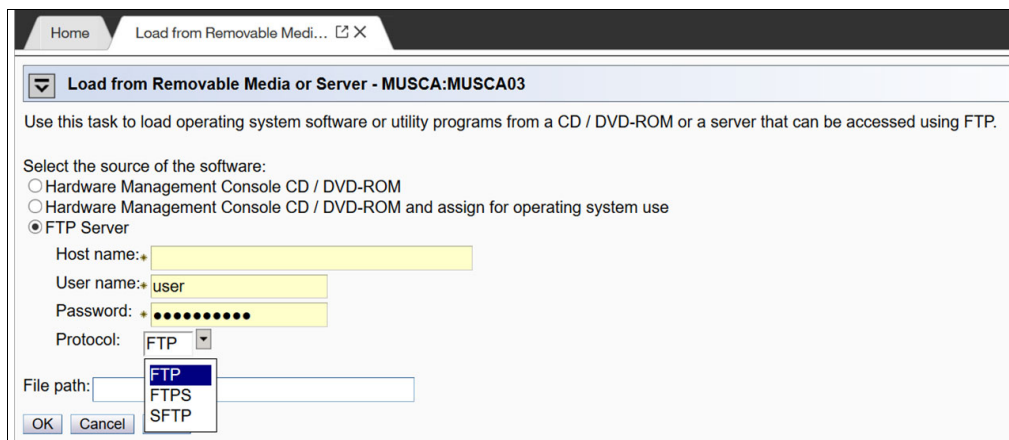


Figure 11-11 New FTP protocols drop-down list

FTP through HMC

It is highly recommended to keep IBM Z systems, HMC consoles, and SEs on an isolated network. This approach prevents SEs making FTP connections with outside networks and applies to all FTP supported protocols: FTP, FTPS, and SFTP.

With z14 HMC, all FTP connections that originate from SEs are taken to HMC consoles. Secure FTP server credentials must be imported to one or more managing HMC consoles.

After the HMC console completes all FTP operations, the HMC console performs the FTP operation on the SE's behalf and returns the results. The IBM Z platform must be managed by at least one HMC to allow FTP operations to work.

Secure console-to-console communication

Before z14, IBM Z HMC consoles used anonymous cipher suites to set console-to-console communication. Anonymous cipher suite is a part of SSL/TLS protocol and it can be used to create internal point-to-point connections. Anonymous cipher suite does not exchange certificates, which can be the source of a security breach.

Therefore, z14 HMC consoles abandon anonymous cipher suite and implement an industry standard-based, password-driven cryptography system. The Domain Security Settings are used to provide authentication and high-quality encryption. Because of these changes, we now recommend that customers use unique Domain Security settings to provide maximum security. The new system provides greater security than anonymous cipher suites, even if the default settings are used.

To allow greater flexibility in password selection, the password limit was increased to 64 characters and special characters are allowed for z14 only installations. If communication with older systems is needed, the previous password limits must be followed (6 - 8 characters, only uppercase and number characters allowed).

For more information about HMC networks, see the following resources:

- ▶ The HMC and SE (Version 2.14.0) console help system, or [IBM Knowledge Center](#)
- ▶ *IBM 3907 Installation Manual for Physical Planning*, GC28-6974

11.3.2 Hardware prerequisite changes

The following HMC changes are important for z14 ZR1:

- ▶ IBM does not provide Ethernet switches with the system.
- ▶ RSF is broadband-only.

Ethernet switches

Ethernet switches for HMC and SE connectivity are provided by the customer. Existing supported switches can still be used.

Ethernet switches and hubs often include the following characteristics:

- ▶ A total of 16 auto-negotiation ports
- ▶ 100/1000 Mbps data rate
- ▶ Full or half duplex operation
- ▶ Auto medium-dependent interface crossover (MDIX) on all ports
- ▶ Port status LEDs

Note: The recommendation is to use 1000 Mbps/Full duplex.

RSF is broadband-only

RSF through a modem is *not* supported on the z14 ZR1 HMC. Broadband is needed for hardware problem reporting and service. For more information, see 11.4, "Remote Support Facility" on page 374.

11.3.3 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 communicates only with HMCs on the same subnet. The HMC and SE can use IPv6 link-local addresses to communicate with each other.

IPv6 link-local addresses feature the following characteristics:

- ▶ Every IPv6 network interface is assigned a link-local IP address.
- ▶ A link-local address is used on a single link (subnet) only and is never routed.
- ▶ Two IPv6-capable hosts on a subnet can communicate by using link-local addresses, without having any other IP addresses assigned.

11.3.4 OSA Support Facility changes

Since OSA/SF was moved from z/OS to HMC/SE environment, it was noted that it is no longer easy to obtain a global view of all OSA PCHIDs and the monitoring and diagnostic information that was previously available in the Query Host command.

To address this issue, the following changes were made:

- ▶ If a CPC is targeted, the initial panel provides a global view of all OSA PCHIDs.
- ▶ The user can browse to various OSA Advanced Facilities subtasks from the initial panel, which makes the process of getting to them less cumbersome.
- ▶ Today's View Port Parameters and Display OAT entries support exporting data of one OSA PCHID. A new action is added to the initial panel that exports the data for all OSA PCHIDs.
- ▶ The initial panel was changed to display status information of all OSA PCHID (see Figure 11-12).

Select	PCHID	Hardware Type	Status	CHPID Type	Code Level	Port 0 Status	Port 0 MAC Address	Port 1 Status	Port 1 MAC Address
<input checked="" type="radio"/>	010C	OSA-Express6S 10Gb SR Ethernet	Operating	OSD	0184	Enabled	98BE94797504		
<input type="radio"/>	0110	OSA-Express6S 1000Base-T Ethernet	Operating	OSC	0064	Enabled	98BE94793CB2	Enabled	98BE94793CB3
<input type="radio"/>	0128	OSA-Express6S 10Gb SR Ethernet	Operating	OSD	0184	Enabled	98BE947957F4		
<input type="radio"/>	012C	OSA-Express6S 1000Base-T Ethernet	Operating	OSD	0184	Enabled	98BE94797926	Enabled	98BE94797927
<input type="radio"/>	0130	OSA-Express6S 1000Base-T Ethernet	Operating	OSD	0184	Enabled	98BE947978B4	Enabled	98BE947978B5
<input type="radio"/>	0148	OSA-Express6S 10Gb SR Ethernet	Operating	OSD	0184	Enabled	98BE94795290		
<input type="radio"/>	014C	OSA-Express6S 1000Base-T Ethernet	Operating	OSD	0184	Enabled	98BE94797876	Enabled	98BE94797877
<input type="radio"/>	0150	OSA-Express6S 1000Base-T Ethernet	Operating	OSD	0184	Enabled	98BE947978C2	Enabled	98BE947978C3
<input type="radio"/>	016C	OSA-Express6S 10Gb SR Ethernet	Operating	OSD	0184	Enabled	98BE94794DC2		
<input type="radio"/>	0170	OSA-Express6S 1000Base-T Ethernet	Operating	OSD	0184	Enabled	98BE9479754E	Enabled	98BE9479754F

Total: 10 Filtered: 10 Selected: 1

Figure 11-12 OSA Advanced Facilities panel

11.3.5 Assigning addresses to the HMC and SE

An HMC can have the following IP configurations:

- ▶ Statically assigned IPv4 or statically assigned IPv6 addresses
- ▶ Dynamic Host Configuration Protocol (DHCP)-assigned IPv4 or DHCP-assigned IPv6 addressees
- ▶ 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 the use of 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 multicasting² to discover automatically 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 features 16 bytes. It is written as eight 16-bit hexadecimal blocks that are separated by colons, as shown in the following example:

```
2001:0db8:0000:0000:0202:b3ff:fe1e:8329
```

Because many IPv6 addresses are not fully qualified, shorthand notation can be used. In shorthand notation, the leading zeros can be omitted, and a series of consecutive zeros can be replaced with a double colon. The address in the previous example also can be written in the following manner:

```
2001:db8::202:b3ff:fe1e:8329
```

If an IPv6 address is assigned to the HMC for remote operations that use a web browser, browse to it by specifying that address. The address must be surrounded with square brackets in the browser's address field, as shown in the following example:

```
https://[fdab:1b89:fc07:1:201:6cff:fe72:ba7c]
```

The use of link-local addresses must be supported by your browser.

11.3.6 HMC Multi-factor authentication

Multi-factor authentication is an optional feature, which is configured on per-user, per-template basis. It enhances security of the z14 models by requiring not only what you know, which is first factor, but also what you have available, which means that only person who owns a specific phone number can log in.

² For a customer-supplied switch, multicast must be enabled at the switch level.

Multi-factor authentication first factor is login and password; the second factor is a time-based, one-time password that is sent to your smartphone. This password is defined in RFC 6238 standard and uses a cryptographic hash function that combines a secret key with the current time to generate a one-time password.

The secret key is generated by HMC/SE/TKE while the user is performing first factor logon. The secret key is known only to HMC/SE/TKE and to the user's smartphone. For that reason, it must be protected as much as your first factor password.

Multi-factor authentication code (MFA code) that was generated as a second factor is time-sensitive. Therefore, it is important to remember that it should be used soon after it is generated.

The algorithm within the HMC that is responsible for MFA code generation changes the code every 30 seconds. However, to make things easier, the HMC and SE console accepts current, previous, and next MFA codes. It is also important to have HMC, SE, TKE, and smartphone clocks synced. If the clocks are not synced, the MFA logon attempt fails. Another important fact is that time zone differences are irrelevant because the MFA code algorithm uses UTC.

11.4 Remote Support Facility

The HMC Remote Support Facility (RSF) provides important communication to a centralized IBM support network for hardware problem reporting and service. The following types of communication are provided:

- ▶ Problem reporting and repair data
- ▶ Microcode Change Level (MCL) delivery
- ▶ Hardware inventory data, which is also known as vital product data (VPD)
- ▶ On-demand enablement

Consideration: RSF through a modem is *not* supported on the z14 ZR1 HMC. Broadband connectivity is needed for hardware problem reporting and service.

11.4.1 Security characteristics

The following security characteristics are in effect:

- ▶ RSF requests always are started from the HMC to IBM. An inbound connection is never started from the IBM Service Support System.
- ▶ All data that is transferred between the HMC and the IBM Service Support System is encrypted with high-grade SSL/Transport Layer Security (TLS) encryption.
- ▶ When starting the SSL/TLS-encrypted connection, the HMC validates the trusted host with the digital signature that is issued for the IBM Service Support System.
- ▶ Data that is sent to the IBM Service Support System consists of hardware problems and configuration data.

More information: For more information about the benefits of Broadband RSF and the SSL/TLS-secured protocol, and a sample configuration for the Broadband RSF connection, see *Integrating the HMC Broadband Remote Support Facility into Your Enterprise*, SC28-6986.

11.4.2 RSF connections to IBM and Enhanced IBM Service Support System

If the HMC and SE are at Driver 22 or later, the driver uses a new remote infrastructure at IBM when the HMC connects through RSF for certain tasks. Check your network infrastructure settings to ensure that this new infrastructure works.

At the time of this writing, RSF still uses the “traditional” RETAIN connection. You must add access to the new Enhanced IBM Service Support System to your current RSF infrastructure (proxy, firewall, and so on).

To have the best availability and redundancy and to be prepared for the future, the HMC must access IBM by using the internet through RSF in the following manner: Transmission to the enhanced IBM Support System requires a domain name server (DNS). The DNS must be configured on the HMC if a proxy for RSF is not used. If a proxy for RSF is used, the proxy must provide the DNS.

The following host names and IP addresses are used and your network infrastructure must allow the HMC to access the following host names:

- ▶ www-945.ibm.com on port 443
- ▶ esupport.ibm.com on port 443

The following IP addresses (IPv4, IPv6, or both) can be used:

- ▶ IBM Enhanced support facility:
 - IPV4:
 - 129.42.56.189
 - 129.42.60.189
 - 129.42.54.189
 - IPV6:
 - 2620:0:6c0:200:129:42:56:189
 - 2620:0:6c2:200:129:42:60:189
 - 2620:0:6c4:200:129:42:54:189
- ▶ Legacy IBM support Facility:
 - IPV4:
 - 129.42.26.224
 - 129.42.42.224
 - 129.42.50.224
 - IPV6:
 - 2620:0:6c0:1::1000
 - 2620:0:6c2:1::1000
 - 2620:0:6c4:1::1000

Note: All other previous IP addresses are no longer supported.

11.4.3 HMC and SE remote operations

You can use the following methods to perform remote manual operations on the HMC:

- ▶ Use of 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 customer-installed firewalls between the remote HMC and its managed objects must permit communication between the HMC and the SE. For service and support, the remote HMC also requires connectivity to IBM, or to another HMC with connectivity to IBM through RSF. For more information, see 11.4, “Remote Support Facility” on page 374.

► Use of a web browser to connect to an HMC

The z14 ZR1 HMC application simultaneously supports one local user and any number of remote users. The user interface in the web browser is the same as the local HMC and has the same functions. Some functions are not available.

Access by the UFD requires physical access to the HMC. Logon security for a web browser is provided by the local HMC user logon procedures. Certificates for secure communications are provided, and can be changed by the user. A remote browser session to the primary HMC that is managing an ensemble allows a user to perform ensemble-related actions, such as limiting remote web browser access.

You can now limit remote web browser access by specifying an IP address from the Customize Console Services task. To enable or disable the Remote operation service, click **Change...** in the Customize Console Services window, as shown in Figure 11-13.

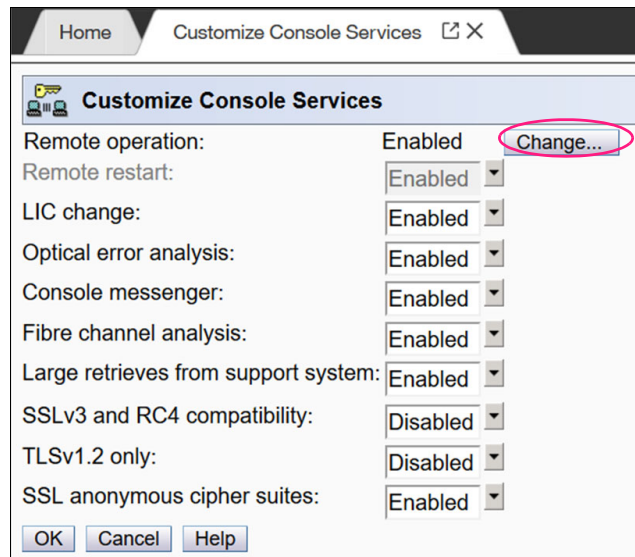


Figure 11-13 Customizing HMC remote operation

Microsoft Internet Explorer, Mozilla Firefox, and Google Chrome were tested as remote browsers. For more information about web browser requirements, see the HMC and SE console help system or [IBM Knowledge Center](#).

Single Object Operations

It is not necessary to be physically close to a SE to use it. The HMC can be used to access the SE remotely by using the SOO task. The interface is the same as the interface that is used on the SE. For more information, see the HMC and SE console help system or [IBM Knowledge Center](#).

HMC mobile interface

The new mobile application interface allows HMC users to securely monitor and manage systems from anywhere. iOS and Android HMC applications are available to provide system and partition views, monitor the status and Hardware and Operating System Messages, and receive mobile push notifications from the HMC by using the zRSF (IBM Z Remote Support Facility) connection.

A full set of granular security controls are provided from the HMC console, to the user, monitor only, and mobile app password, including multi-factor authentication. This mobile interface is optional and is disabled by default.

11.5 HMC and SE capabilities

The HMC and SE feature many capabilities. This section describes the key areas. For more information about these capabilities, see the HMC and SE (Version 2.14.0) console help system or [IBM Knowledge Center](#).

With the introduction of the DPM mode for Linux on Z, only CPCs the user interface and user interaction with the HMC changed dramatically; the capabilities underneath are still the same. The figures and command examples that are shown in this section were taken in PR/SM mode.

11.5.1 Central processor complex management

The HMC is the primary place for CPC control. For example, the input/output configuration data set (IOCDS) includes definitions of LPARs, channel subsystems, control units, and devices, and their accessibility from LPARs. IOCDS can be created and put into production from the HMC.

The HMC is used to start the power-on reset (POR) of the system. 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 started into the hardware system area (HSA).

The hardware messages task displays hardware-related messages at the CPC, LPAR, or SE level. It also displays hardware messages that relate to the HMC.

11.5.2 LPAR management

Use the HMC to define LPAR properties, such as the number of processors of each type, how many are reserved, and how much memory is assigned to it. These parameters are defined in LPAR profiles and stored on the SE.

Because Processor Resource/Systems Manager (PR/SM) must manage LPAR access to processors and the initial weights of each partition, weights are used to prioritize partition access to processors.

You can use the Load task on the HMC to perform an IPL of an operating system. This task causes a program to be read from a designated device, and starts that program. You can perform the IPL of the operating system from storage, the HMC DVD-RAM drive, the USB flash memory drive (UFD), or an FTP server.

When an LPAR is active and an operating system is running in it, you can use the HMC to dynamically change certain LPAR parameters. The HMC provides an interface to change partition weights, add logical processors to partitions, and add memory.

LPAR weights can also be changed through a scheduled operation. Use the Customize Scheduled Operations task to define the weights that are set to LPARs at the scheduled time.

Channel paths can be dynamically configured on and off (as needed for each partition) from an HMC.

The Change LPAR Controls task for z14 ZR1 can export the Change LPAR Controls table data to a comma-separated value (.csv)-formatted file. This support is available to a user when they are 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 more information about a Change LPAR Controls scheduled operation is available on the SE.

One example of managing the LPAR settings is the absolute physical HW LPAR capacity setting. Driver 15 introduced the capability to define (in the image profile for shared processors) the absolute processor capacity that the image is allowed to use (independent of the image weight or other cappings).

To indicate that the LPAR can use the undedicated processors absolute capping, select **Absolute capping** on the Image Profile Processor settings to specify an absolute number of processors at which to cap the LPAR's activity. The absolute capping value can be "None" or a value for the number of processors (0.01 - 255.0).

The LPAR group absolute capping was the next step in partition capping options that are available on z14 M0x, z14 ZR1, z13s, and z13 CPCs at Driver level 27 and greater. Following on to LPAR absolute capping, LPAR group absolute capping uses a similar methodology to enforce the following components:

- ▶ Customer licensing
- ▶ Non-z/OS partitions where group soft capping is not an option
- ▶ z/OS partitions where ISV does not support software capping

A group name, processor capping value, and partition membership are specified at the hardware console, along with the following properties:

- ▶ Set an absolute capacity cap by CPU type on a group of LPARs.
- ▶ Allow each of the partitions to use capacity up to their individual limits if the group's aggregate consumption does not exceed the group absolute capacity limit.
- ▶ Include updated SysEvent QVS support (used by vendors who implement software pricing).
- ▶ Only shared partitions are managed in these groups.
- ▶ Specify caps for one or more processor types in the group.
- ▶ Specify in absolute processor capacity (for example, 2.5 processors).

- Use Change LPAR Group Controls (as with windows that are used for software group-defined capacity), as shown in Figure 11-14 (snapshot on a z13).

Group Name	Member Partitions	Group Capacity Value	Absolute Capping for CPs	Absolute Capping for ICFs	Absolute Capping for IFLs	Absolute Capping for zIIPs
DEFAULT	CETUS05 CETUS14	0	None	None	None	None
DEVGROU	CETUS0F	50	67.44	128.54	None	None
PRVEPOOL	CETUS0C	0	None	None	None	None
ROBSPool	CETUS0D CETUS02	0	8.00	None	None	None

Figure 11-14 Change LPAR Group Controls: Group absolute capping

Absolute capping is specified as an absolute number of processors to which the group's activity is capped. The value is specified to hundredths of a processor (for example, 4.56 processors) worth of capacity.

The value is not tied to the Licensed Internal Code (LIC) configuration code (LICCC). Any value 0.01 - 255.00 can be specified. This configuration makes the profiles more portable, which means that you do not have issues in the future when profiles are migrated to new machines.

Although the absolute cap can be specified to hundredths of a processor, the exact amount might not be that precise. The same factors that influence the “machine capacity” also influence the precision with which the absolute capping works.

11.5.3 Operating system communication

The Operating System Messages task displays messages from an LPAR. You also can enter operating system commands and interact with the system. This task is especially valuable for entering Coupling Facility Control Code (CFCC) commands.

The HMC also provides integrated 3270 and ASCII consoles. These consoles allow an operating system to be accessed without requiring other network or network devices, such as TCP/IP or control units.

Updates to x3270 support

The Configure 3270 Emulators task on the HMC and TKE consoles was enhanced with Driver 15 to verify the authenticity of the certificate that is returned by the 3270 server when a secure and encrypted SSL connection is established to an IBM host. This 3270 Emulator with encrypted connection is also known as *Secure 3270*.

Use the Certificate Management task if the certificates that are returned by the 3270 server are not signed by a well-known trusted certificate authority (CA) certificate, such as VeriSign or Geotrust. An advanced action within the Certificate Management task, Manage Trusted Signing Certificates, is used to add trusted signing certificates.

For example, if the certificate that is associated with the 3270 server on the IBM host is signed and issued by a corporate certificate, it must be imported, as shown in Figure 11-15.

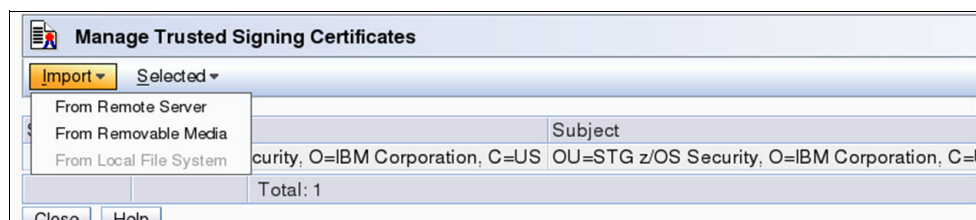


Figure 11-15 Manage Trusted Signing Certificates

The import from the remote server option can be used if the connection between the console and the IBM host can be trusted when the certificate is imported, as shown in Figure 11-16. Otherwise, import the certificate by using removable media.

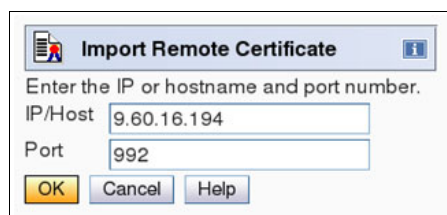


Figure 11-16 Import Remote Certificate example

A secure Telnet connection is established by adding the prefix L: to the IP address:port of the IBM host, as shown in Figure 11-17.

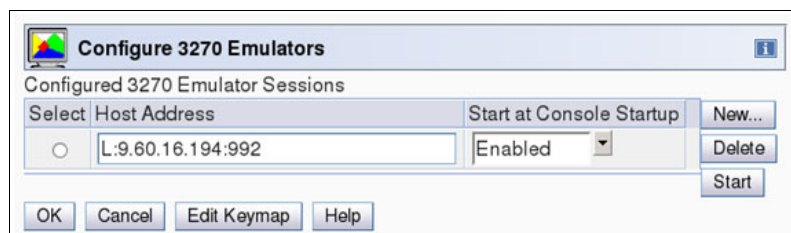


Figure 11-17 Configure 3270 Emulators

11.5.4 HMC and SE microcode

The microcode for the HMC, SE, and CPC is included in the driver or version. The HMC provides the management of the driver upgrade through Enhanced Driver Maintenance (EDM). EDM also provides the installation of the latest functions and the patches (MCLs) of the new driver.

When you perform a driver upgrade, always check the Driver (xx) Customer Exception Letter option in the Fixes section at the IBM Resource Link.

For more information, see 9.8, “z14 ZR1 Enhanced Driver Maintenance” on page 338.

Microcode Change Level

Regular installation of Microcode Change Levels (MCLs) is key for reliability, availability, and serviceability (RAS), optimal performance, and the following new functions:

- ▶ Install MCLs on a quarterly basis at a minimum.
- ▶ Review hiper MCLs continuously to decide whether to wait for the next scheduled fix application session or to schedule one earlier if the risk assessment warrants.

Tip: The IBM Resource Link provides access to the system information for your IBM Z system according to the system availability data that is sent on a scheduled basis. It provides more information about the MCL status of your z14 systems.

For more information about accessing the Resource Link, see [the IBM Resource Link website](#) (login required)

At the Resource Link website, click **Tools** → **Machine Information**, choose your IBM Z system, and then, click **EC/MCL**.

Microcode terms

The microcode features the following characteristics:

- ▶ The driver contains engineering change (EC) streams.
- ▶ Each EC stream covers the code for a specific component of z14 ZR1. It includes a specific name and an ascending number.
- ▶ The EC stream name and a specific number are one MCL.
- ▶ MCLs from the same EC stream must be installed in sequence.
- ▶ MCLs can include 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).

How the driver, bundle, EC stream, MCL, and MCFs interact with each other is shown in Figure 11-18.

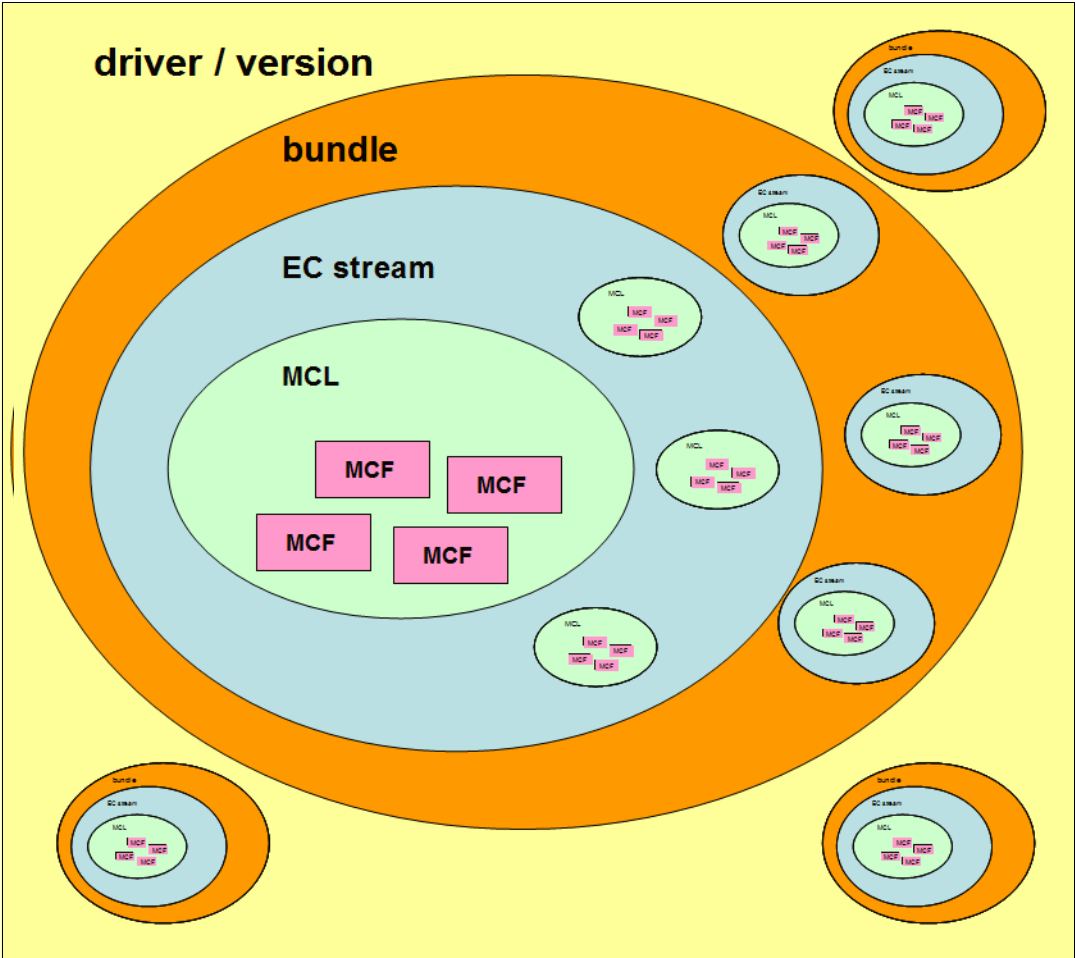
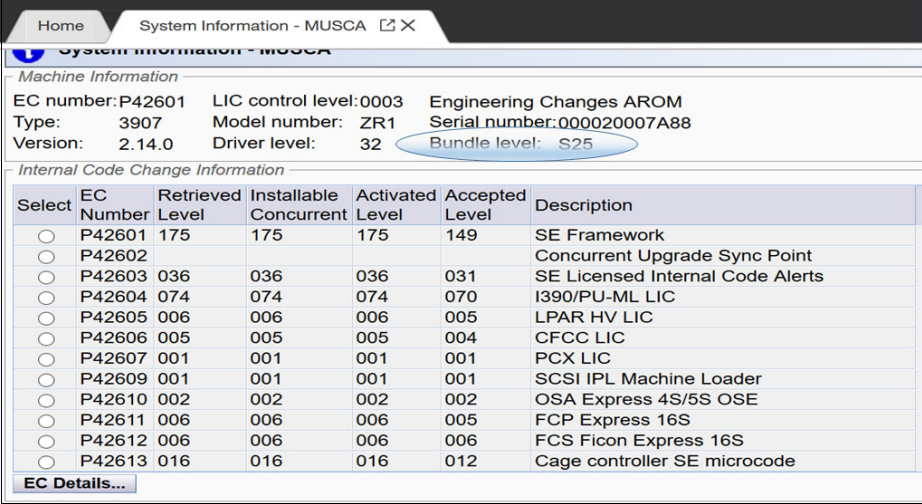


Figure 11-18 Microcode terms and interaction

Microcode installation by MCL bundle target

A *bundle* is a set of MCLs that are grouped during testing and released as a group on the same date. You can install an MCL to a specific target bundle level. The System Information window is enhanced to show a summary bundle level for the activated level, as shown in Figure 11-19.



The screenshot shows the 'System Information - MUSCA' window. The 'Machine Information' section displays: EC number: P42601, Type: 3907, Version: 2.14.0, LIC control level: 0003, Model number: ZR1, Driver level: 32, Engineering Changes AROM, Serial number: 000020007A88, and Bundle level: S25 (highlighted with a blue oval). The 'Internal Code Change Information' section contains a table with columns: Select, EC Number, Retrieved Level, Installable Concurrent, Activated Level, Accepted Level, and Description.

Select	EC Number	Retrieved Level	Installable Concurrent	Activated Level	Accepted Level	Description
<input type="radio"/>	P42601	175	175	175	149	SE Framework
<input type="radio"/>	P42602					Concurrent Upgrade Sync Point
<input type="radio"/>	P42603	036	036	036	031	SE Licensed Internal Code Alerts
<input type="radio"/>	P42604	074	074	074	070	I390/PU-ML LIC
<input type="radio"/>	P42605	006	006	006	005	LPAR HV LIC
<input type="radio"/>	P42606	005	005	005	004	CFCC LIC
<input type="radio"/>	P42607	001	001	001	001	PCX LIC
<input type="radio"/>	P42609	001	001	001	001	SCSI IPL Machine Loader
<input type="radio"/>	P42610	002	002	002	002	OSA Express 4S/5S OSE
<input type="radio"/>	P42611	006	006	006	005	FCP Express 16S
<input type="radio"/>	P42612	006	006	006	006	FCS Ficon Express 16S
<input type="radio"/>	P42613	016	016	016	012	Cage controller SE microcode

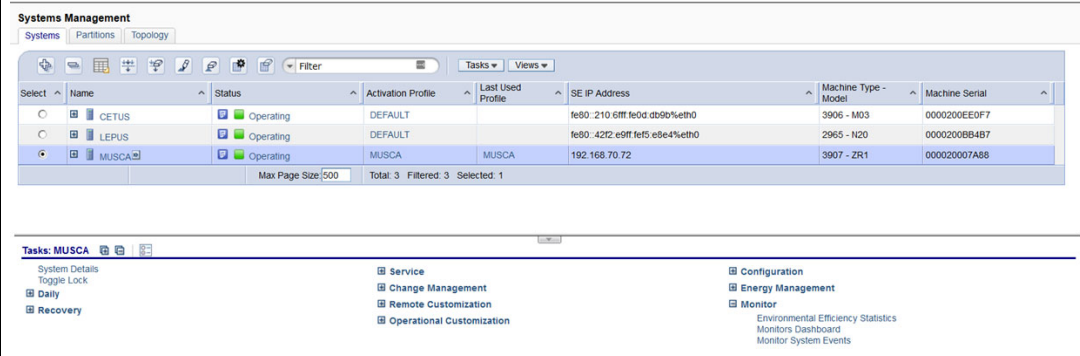
Figure 11-19 System Information: Bundle level

11.5.5 Monitoring

This section describes monitoring considerations.

Monitor task group

The Monitor task group on the HMC and SE includes monitoring-related tasks for IBM Z CPCs, as shown in Figure 11-20.



The screenshot shows the 'Systems Management' console with the 'Monitor' task group selected. The main table lists systems: CETUS, LEPUS, and MUSCA. MUSCA is selected, and its details are shown in the 'Tasks: MUSCA' section below.

Select	Name	Status	Activation Profile	Last Used Profile	SE IP Address	Machine Type - Model	Machine Serial
<input type="radio"/>	CETUS	Operating	DEFAULT		fe80: 210: 6fff: fe0d: cb9b%eth0	3906 - M03	0000200E0F7
<input type="radio"/>	LEPUS	Operating	DEFAULT		fe80: 42f2: e9ff: fe15: e8e4%eth0	2965 - N20	0000200B84B7
<input checked="" type="radio"/>	MUSCA	Operating	MUSCA	MUSCA	192.168.70.72	3907 - ZR1	000020007A88

Tasks: MUSCA

- System Details
- Toggle Lock
- Daily
- Recovery
- Service
- Change Management
- Remote Customization
- Operational Customization
- Configuration
- Energy Management
- Monitor
 - Environmental Efficiency Statistics
 - Monitors Dashboard
 - Monitor System Events

Figure 11-20 HMC Monitor Task Group

Monitors Dashboard task

The Monitors Dashboard task supersedes the System Activity Display (SAD). In the z14 ZR1, the Monitors Dashboard task in the Monitor task group provides a tree-based view of resources.

Multiple graphical views are available for displaying data, including history charts. The Open Activity task (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 system.

An example of the Monitors Dashboard task is shown in Figure 11-21.

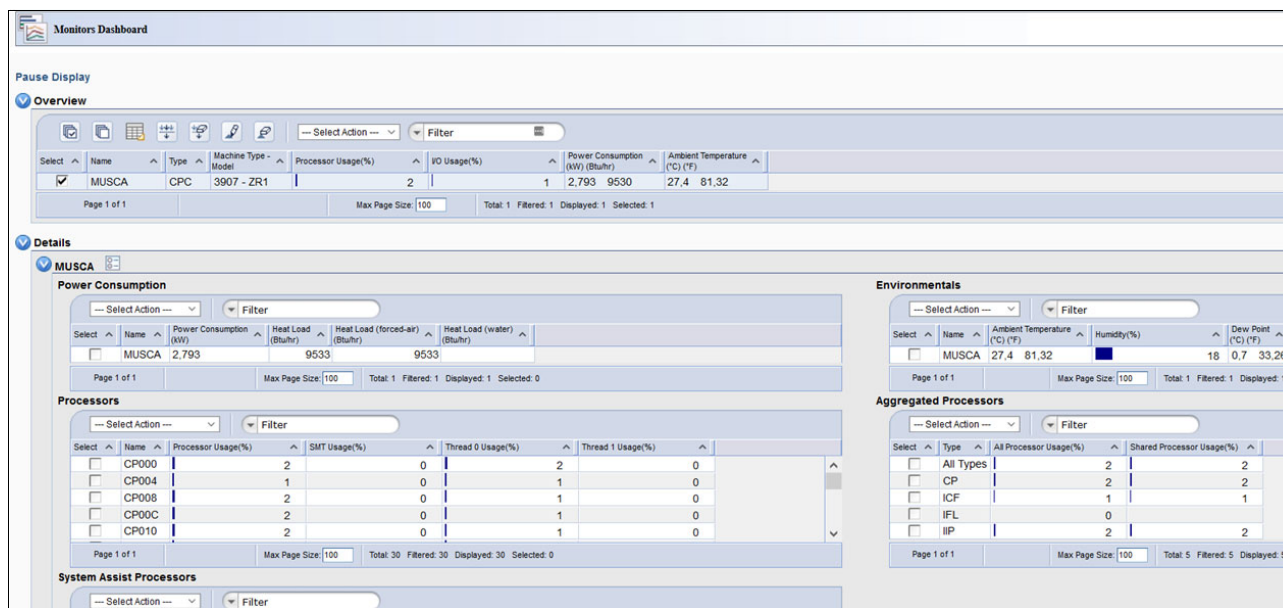


Figure 11-21 Monitors Dashboard task

You can display more information for the following components (see Figure 11-22 on page 385):

- ▶ Power consumption
- ▶ Environmentals
- ▶ Aggregated processors
- ▶ Processors (with SMT information)
- ▶ System Assist Processors
- ▶ Logical Partitions
- ▶ Channels
- ▶ Adapters: Crypto use percentage is displayed according to the physical channel ID (PCHID number).

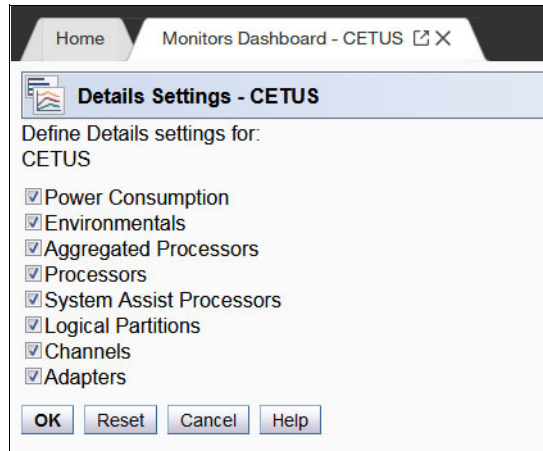


Figure 11-22 Monitors dashboard Detailed settings

Environmental Efficiency Statistics task

The Environmental Efficiency Statistics task 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 also can 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.

11.5.6 Capacity on-demand support

All capacity on demand (CoD) upgrades are performed by using the SE Perform a Model Conversion task. Use the task to retrieve and activate a permanent upgrade, and to retrieve, install, activate, and deactivate a temporary upgrade. The task shows a list of all installed or staged LICCC records to help you manage them. It also shows a history of recorded activities.

The HMC for IBM z14 ZR1 features the following 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 Back Up (CBU)
- ▶ SE window features (accessed through HMC Single Object Operations):
 - Window controls for granular activation and deactivation
 - History window for all CoD actions
 - Description editing of CoD records
- ▶ HMC/SE provides the following CoD information:
 - Millions of service units (MSU) and processor tokens
 - Last activation time
 - Pending resources that are shown by processor type instead of only a total count
 - Option to show more information about installed and staged permanent records
 - More information for the Attention state by providing seven more flags

HMC and SE are a part of the z/OS Capacity Provisioning environment. The Capacity Provisioning Manager (CPM) communicates with the HMC through IBM Z APIs, and enters CoD requests. For this reason, SNMP must be configured and enabled by using the Customize API Settings task on the HMC.

For more information about using and setting up CPM, see [IBM Knowledge Center](#) or the following publications:

- ▶ *z/OS MVS™ Capacity Provisioning User's Guide*, SC33-8299
- ▶ *IBM Z System Capacity on Demand User's Guide*, SC28-6943

11.5.7 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). Consider the following points:

- ▶ IBM Z rely on STP for time synchronization, and continue to provide support of a pulse per second (PPS) port. STP with PPS maintains an accuracy of 10 microseconds as measured at the PPS input of the z14 ZR1. If STP uses a Network Time Protocol (NTP) server without PPS, time accuracy of 100 milliseconds to the External Time Source (ETS) is maintained.
- ▶ The z14 ZR1 cannot be in the same CTN with a z114 or z196 or earlier systems and cannot become member of a mixed CTN.
- ▶ An STP-only CTN can be managed by using different HMCs. However, the HMC must be at the same driver level (or later) than any SE that is to be managed. Also, all SEs to be managed must be known (defined) to that HMC.

In a STP-only CTN, the HMC can be used to perform the following tasks:

- ▶ Start or modify the CTN ID.
- ▶ Start the time (manually or by contacting an NTP server).
- ▶ Start the time zone offset, Daylight Saving Time offset, and leap second offset.
- ▶ Assign the roles of preferred, backup, and current time servers, and arbiter.
- ▶ Adjust time by up to plus or minus 60 seconds.
- ▶ Schedule changes to the offsets listed. STP can automatically schedule Daylight Saving Time, based on the selected time zone.
- ▶ Monitor the status of the CTN.
- ▶ Monitor the status of the coupling links that are started for STP message exchanges.
- ▶ For diagnostic purposes, the PPS port state on a z14 ZR1 can be displayed and fenced ports can be reset individually.

STP changes and enhancements

The STP-related functions included dramatic changes and gained a new, intuitive GUI. Administrators of z14 systems are guided through a system time management workflow, which reduces the need to refer to external documentation.

The inline definition of technical terms eliminates the need to look up documentation to determine definitions. Detailed instructions and guidelines are provided within task workflow.

New tasks provide a visual representation of STP topology. Current system time networks are shown in topological display. A preview of any configuration action is also shown in topological display. Those changes make administrator more confident and prevent from errors. An example of the topology view is shown in Figure 11-23.

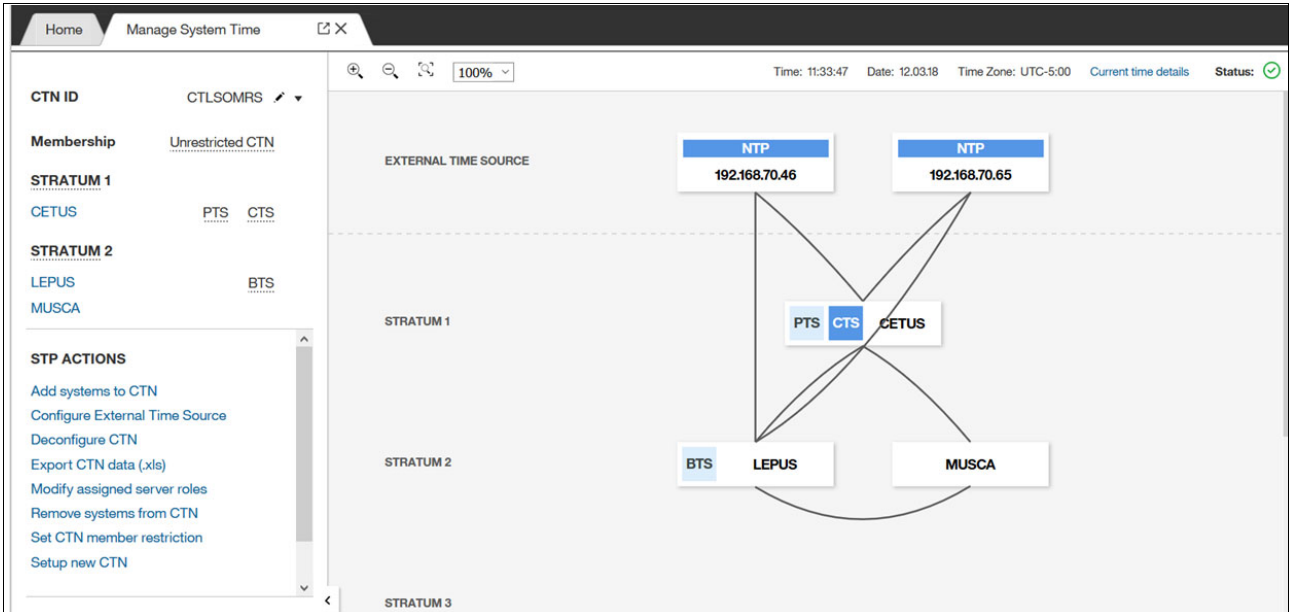


Figure 11-23 CTN topology visible on HMC Manage System Time window

Click **Current time details** for more information and available options (for example, adjusting time zone and leap second offset), as shown in Figure 11-24.

System Time for CTLSOMRS

Current network configuration

Configured at (UTC):22.02.18 19:49:16

Coordinated Server Time

Time zone:(UTC-05:00) Eastern Time (US & Canada) (EST/EDT)

Currently:EDT

Adjust time

Adjust time zone offset

Offsets

Leap second:0

Time zone offset from UTC:-5:00

Daylight saving time (hours:minutes):1:00

Scheduled daylight saving time:EST (0:00) at 04.11.18 06:00:00 UTC

Scheduled leap second offset:No update scheduled

Adjust leap second offset

Figure 11-24 Current time details

Enhanced Console Assisted Recovery

Enhanced Console Assisted Recovery (ECAR) speeds up the process of BTS takeover by performing the following steps:

- 1. When the Primary Time Server (PTS/CTS) detects a checkstop condition, the CEC informs its SE and HMC.

2. The PTS SE recognizes the checkstop pending condition, and calls the PTS SE STP code.
3. The PTS SE sends an ECAR request thorough HMC to the Backup Time Server (BTS) SE.
4. The BTS SE communicates with the BTS to start the takeover.

ECAR support is faster than the original CAR support because the console path changes from a 2-way path to a 1-way path. Also, almost no lag time is incurred between the system checkstop and the start of CAR processing. Because the request is generated from the PTS before system logging, it avoids the potential of recovery being held up.

Requirements

ECAR is available on z14 M0x, z14 ZR1, and z13/z13s systems on Driver 27 and later only. In a mixed environment with previous generation machines, you should define a z14 M0x, z13, or z13s system as the PTS and CTS.

Attention: z14 ZR1 does not support InfiniBand connectivity; therefore, it cannot be connected by using coupling/timing links to a zEC12 or zBC12 CPC. As such, in a CTN with zEC12 or zBC12, the z14 ZR1 cannot be assigned a role (PTS, CTS or Arbiter; its failure affects the time synchronization functionality of other servers in the CTN).

For more information about planning and setup, see the following publications:

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

11.5.8 CTN Split and Merge

With HMC 2.14.1, STP management was enhanced with two new actions: CTN split and CTN merge.

CTN Split

The HMC menus for Server Time Protocol (STP) were enhanced to provide support when one or more systems must be split in to a separate CTN without interruption in the clock source.

The task is available under the Advanced Actions menu in the Manage System Time task. Several checks are performed to avoid potential disruptive actions. If targeted CTN only has members with the roles, task launch fails with error message. If targeted CTN has at least one system without any roles, task launches. An informational warning is presented to the user to acknowledge that sysplex workloads are divided appropriately.

Merging two CTNs

When two separate CTNs must be merged in to the single CTN without interruption in the clock source, the system administrator must perform the Join existing CTN action, which is available in the Advanced Actions menu.

Note: After joining the selected CTN, all systems within the current CTN are synchronized with the Current Time Server of the selected CTN. A coupling link must be in place that connects the CTS of the selected CTN and the CTS of the current CTN.

During the transition state, most of the STP actions for the two affected CTNs are disabled. After the merge is completed, STP actions are enabled again.

For more information about planning and understanding STP server roles, see the following publications:

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

11.5.9 NTP client and server support on the HMC

The NTP client support allows a STP-only CTN to use an NTP server as an ETS. This capability addresses the following requirements:

- ▶ Clients who want time synchronization for the servers members of the STP-only CTN
- ▶ Clients who use a common time reference across heterogeneous systems

The NTP server becomes the single time source (the ETS) for STP and other servers that are not IBM Z systems (such as AIX®, and Microsoft Windows) that include NTP clients.

The HMC can act as an NTP server. With this support, the z14 ZR1 can receive the time from the HMC without accessing a LAN other than the HMC and SE network. When the HMC is used as an NTP server, it can be configured to receive the NTP source from the internet. For this type of configuration, a LAN that is separate from the HMC/SE LAN can be used.

HMC NTP broadband authentication support

HMC NTP authentication can be used since HMC Driver 15. The SE NTP support is unchanged. To use this option on the SE, configure the HMC with this option as an NTP server for the SE.

Authentication support with a proxy

Some client configurations use a proxy for external access outside the corporate data center. NTP requests are User Datagram Protocol (UDP) socket packets and cannot pass through the proxy. The proxy must be configured as an NTP server to get to target servers on the web. Authentication can be set up on the client's proxy to communicate with the target time sources.

Authentication support with a firewall

If you use a firewall, HMC NTP requests can pass through it. Use HMC authentication to ensure untampered time stamps.

NTP symmetric key and autokey authentication

With symmetric key and autokey authentication, the highest level of NTP security is available. HMC Level 2.12.0 and later provide windows that accept and generate key information to be configured into the HMC NTP configuration. They can also issue NTP commands.

The HMC offers the following 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 encryption and decryption. Users that are exchanging data keep this key to themselves. Messages encrypted with a secret key can be decrypted only with the same secret key. Symmetric key authentication supports network address translation (NAT).

- ▶ Symmetric key autokey (NTP V4) authentication
This autokey uses public key cryptography, as described in RFC 5906, which was made available in NTP Version 4. You can generate keys for the HMC NTP by clicking **Generate Local Host Key** in the Autokey Configuration window. This option issues the **ntp-keygen** command to generate the specific key and certificate for this system. Autokey authentication is not available with the NAT firewall.
- ▶ Issue NTP commands
NTP command support is added to display the status of remote NTP servers and the current NTP server (HMC).

For more information about planning and setup for STP and NTP, see the following publications:

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

11.5.10 Security and user ID management

This section addresses security and user ID management considerations.

HMC and SE security audit improvements

With the Audit and Log Management task, audit reports can be generated, viewed, saved, and offloaded. The Customize Scheduled Operations task allows you to schedule audit report generation, saving, and offloading. The Monitor System Events task allows Security Logs to send email notifications by using the same type of filters and rules that are used for hardware and operating system messages.

With z14 ZR1, you can offload the following HMC and SE log files for customer audit:

- ▶ Console event log
- ▶ Console service history
- ▶ Tasks performed log
- ▶ Security logs
- ▶ System log

Full log offload and delta log offload (since the last offload request) are provided. Offloading to removable media and to remote locations by FTP is available. The offloading can be manually started by the new Audit and Log Management task or scheduled by the Customize Scheduled Operations task. The data can be offloaded in the HTML and XML formats.

HMC user ID templates and LDAP user authentication

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 who is mapped to that template. The HMC user is mapped to a specific user ID template by user ID pattern matching. The system then obtains the name of the user ID template from content in the LDAP server schema data.

Default HMC user IDs

It is no longer possible to change the Managed Resource or Task Roles of the default user ID's operator, advanced, sysprog, acsadmin, and service.

If you want to change the roles for a default user ID, create your own version by copying a 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 include “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 the HMC and SE, to import public keys that are associated with a host address.

The Secure FTP infrastructure allows HMC and SE applications to query whether a public key is associated with a host address and to use the Secure FTP interface with the appropriate public key for a host. Tasks that use FTP now provide a selection for the secure host connection.

When selected, the task verifies that a public key is associated with the specified host name. If a public key is not provided, a message window opens that points to the Manage SSH Keys task to enter a public key. The following tasks provide this support:

- ▶ Import/Export IOCDs
- ▶ Advanced Facilities FTP IBM Content Collector Load
- ▶ Audit and Log Management (Scheduled Operations only)
- ▶ FCP Configuration Import/Export
- ▶ OSA view Port Parameter Export
- ▶ OSA-Integrated Console Configuration Import/Export

11.5.11 HMC 2.14.1 Enhancements

HMC Version 2.14.1 introduces the following enhancements:

- ▶ IPv6 Support for OSA-ICC 3270

In addition to IPV4 protocol, HMC 2.14.1 added IPv6 support for OSA-ICC 3270 for compliance with regulations that require all computer purchases to support IPv6.

- ▶ TLS level negotiation limits for OSA-ICC 3270

The supported TLS protocol levels for the OSA-ICC 3270 client connection can now be specified. Supported protocol levels are TLS 1.0, TLS 1.1, and TLS 1.2:

- TLS 1.0 → OSA-ICC 3270 server permits TLS 1.0, TLS 1.1, and TLS 1.2 client connections.
- TLS 1.1 → OSA-ICC 3270 server permits TLS 1.1 and TLS 1.2 client connections.
- TLS 1.2 → OSA-ICC 3270 server permits only TLS 1.2 client connections.

TLS 1.2 was introduced for z13 Driver level 27 (HMC 2.13.1) for OSA-Express4S and OSA-Express5S.

► Separate Security Certificates for OSA-ICC 3270

Separate and unique OSA-ICC 3270 certificates are now supported for clients who host workloads across multiple business units or data centers where cross-site coordination is required. Clients can avoid interruption of all the TLS connections at the same time, when they renew expired certificates. The certificate for the PCHID is independently managed with respect to expiry/renewal and other properties (such as self-signed or CA signed).

OSA-ICC also continues to support a single certificate for all OSA-ICC PCHIDs in the system.

► SCSI load normal enhancements

Before z14 GA2, when performing a standard (CCW-type) load, the user can choose to clear memory or not clear memory (that is, “normal”). However, memory is always cleared during a SCSI load.

For z14 with HMC 2.14.1 and Driver level 36, SCSI load can be performed without clearing memory first (that is, “*SCSI load normal*”). Faster load times when the loaded program does not require memory to be cleared for proper operation.

Notes: Memory is always cleared as part of activating an image before any load is performed. Therefore, not clearing the memory is not an option when activating with an image profile.

When managed by HMC version 2.14.1, a z14 Driver level 32 or older system cannot take advantage of the *SCSI load normal* option.

► Support Element login enhancements for HMC connections

With HMC 2.14.1, when attempting to access an SE by using Single Object Operations (SOO) from an HMC, and an SOO session exists, the following results occur:

- Access is still initially denied, but the denial panel now offers an option to *disconnect* the remote user ID.
- When selecting this option, a confirmation panel is displayed, which requires the HMC user to confirm before proceeding with the *disconnect*.
- If confirmed, the existing SOO session to the SE from the other HMC is ended and its associated user is *disconnected*. Establishment of the new SOO session proceeds immediately.
- Security log entry is written on the SE to record information about both the disconnected HMC/session and the disconnecting HMC/session.

Note: When the user is physically logged in (that is, using the SE’s keyboard/display), sessions are *not* disconnected; only the “Chat” option is available.

11.5.12 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 that is needed to manage a system’s I/O configuration must be obtained from many separate sources. The System Input/Output Configuration Analyzer task enables the system hardware administrator to access, from one location, the information from those sources. Managing I/O configurations then becomes easier, particularly across multiple systems.

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. By using the tool, data is formatted and displayed in five different views. The tool provides various sort options, and data can be exported to a UFD for later viewing.

The following 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.

11.5.13 Automated operations

As an alternative to manual operations, an application can interact with the HMC and SE through an API. The interface allows a program to monitor and control the hardware components of the system in the same way a user performs these tasks. The HMC APIs provide monitoring and control functions through SNMP and the CIM. These APIs can get and set a managed object's attributes, issue commands, receive asynchronous notifications, and generate SNMP traps.

The HMC supports the CIM as an extra systems management API. The focus is on attribute query and operational management functions for IBM Z systems, such as CPCs, images, and activation profiles. z14 systems contain several enhancements to the CIM systems management API. The function is similar to the function that is provided by the SNMP API.

For more information about APIs, see *IBM Z Application Programming Interfaces*, SB10-7164.

11.5.14 Cryptographic support

This section describes the cryptographic management and control functions that are available in the HMC and the SE.

Cryptographic hardware

z14 systems include 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 Express6S, which is a new Peripheral Component Interconnect Express (PCIe) cryptographic coprocessor, is an optional z14 exclusive feature. Crypto Express6S provides a secure programming and hardware environment on which crypto processes are run. Each Crypto Express6S 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 a CCA coprocessor is selected. CCA firmware and PKCS #11 firmware cannot coexist in a card.

The Trusted Key Entry (TKE) Workstation with smart card reader feature is required to support the administration of the Crypto Express6S when configured as an Enterprise PKCS #11 coprocessor.

To support the new Crypto Express6S card, the Cryptographic Configuration window was changed to support the following card modes:

- Accelerator mode (CEX6A)
- CCA Coprocessor mode (CEX6C)
- PKCS #11 Coprocessor mode (CEX6P)

An example of the Cryptographic Configuration window is shown in Figure 11-25.

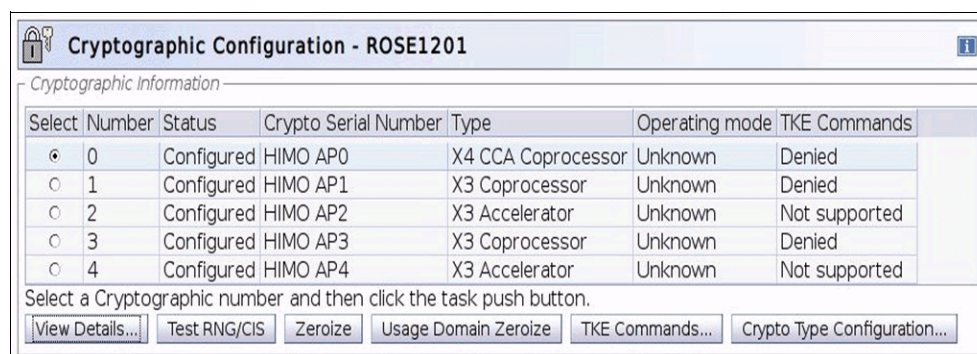


Figure 11-25 Cryptographic Configuration window

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 Express6/5S in EP11 mode is configured to the standby state after the zeroize process.

For more information, see *IBM z14 Model ZR1 Configuration Setup*, SG24-8560.

Digitally signed firmware

Security and data integrity are critical issues with firmware upgrades. Procedures are in place to use a process to digitally sign the firmware update files that are sent to the HMC, SE, and TKE. By using a hash algorithm, a message digest is generated that is then encrypted with a private key to produce a digital signature.

This operation ensures that any changes that are made to the data are detected during the upgrade process by verifying the digital signature. It helps ensure that no malware can be installed on IBM Z products during firmware updates. It also enables the z14 ZR1 Central Processor Assist for Cryptographic Function (CPACF) functions to comply with Federal Information Processing Standard (FIPS) 140-2 Level 1 for Cryptographic LIC changes. The enhancement follows the IBM Z focus of security for the HMC and the SE.

z14 Crypto Card CEX6S is compliant with CCA PCI HSM.

The following CCA compliance levels for Crypto Express6S are available on SE:

- ▶ CCA: Non-compliant (default)
- ▶ CCA: PCI-HSM 2016
- ▶ CCA: PCI-HSM 2016

The following EP11 compliance levels (Crypto Express5S and Crypto Express6S) are available:

- ▶ FIPS 2009 (default)
- ▶ FIPS 2011
- ▶ BSI 2009
- ▶ BSI 2011

11.5.15 Installation support for z/VM that uses the HMC

Starting with z/VM V5R4 and z10, Linux on Z can be installed in a z/VM virtual machine from HMC workstation media. This Linux on Z installation can use the communication path between the HMC and the SE. No external network or extra network setup is necessary for the installation.

11.5.16 Dynamic Partition Manager

DPM is an IBM Z mode of operation that provides a simplified approach to create and manage virtualized environments, which reduces the barriers of its adoption for new and existing customers.

Setting up is a disruptive action. The selection of the DPM mode of operation is done by using the Enable Dynamic Partition Manager function, which is available in the SE CPC Configuration menu.

Enabling DPM by using the SE interface is shown in Figure 11-26.

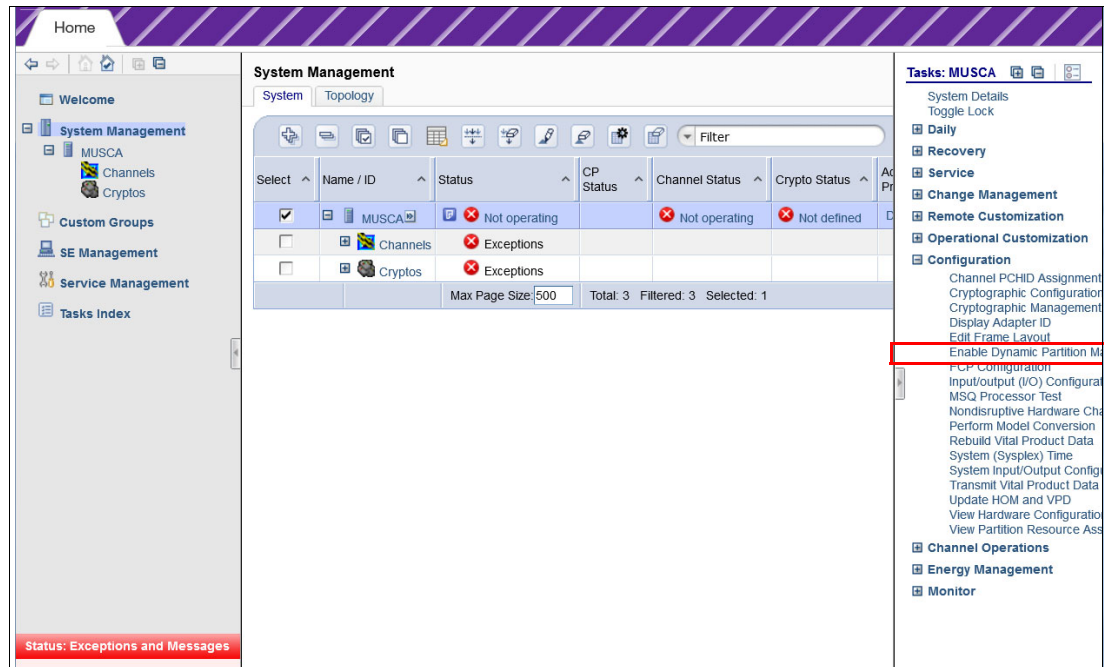


Figure 11-26 Enabling DPM (SE interface)

Attention: The Enabling Dynamic Partition Manager task is run on the SE and can be performed by your IBM system service representative (SSR).

After the CPC is restarted and you log on to the HMC in which this CPC is defined, the HMC shows the Welcome window that is shown in Figure 11-27.

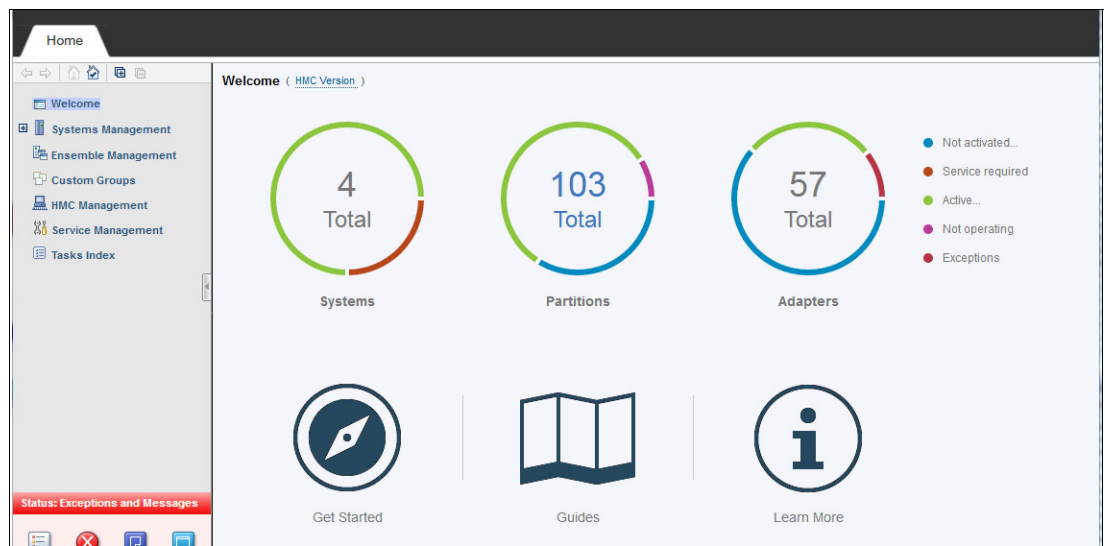


Figure 11-27 HMC welcome window

New LPARs can be added by selecting **Get Started**. For more information, see [IBM Knowledge Center](#).



Performance

This chapter describes the performance considerations for IBM z14 ZR1 servers.

This chapter includes the following topics:

- ▶ 12.1, “IBM z14 ZR1 performance characteristics” on page 398
- ▶ 12.2, “LSPR workload suite” on page 399
- ▶ 12.3, “Fundamental components of workload performance” on page 399
- ▶ 12.4, “Relative Nest Intensity” on page 401
- ▶ 12.5, “LSPR workload categories based on relative nest intensity” on page 403
- ▶ 12.6, “Relating production workloads to LSPR workloads” on page 403
- ▶ 12.7, “Workload performance variation” on page 404

12.1 IBM z14 ZR1 performance characteristics

z14 ZR1 Z06 is designed to offer up to 12% more capacity and twice the amount of memory than the z13s Z06 system.

Uniprocessor performance also increased. On average, a z14 ZR1 model Z01 offers performance improvements of up to 9% over the z13s model Z01 (see Figure 12-1).

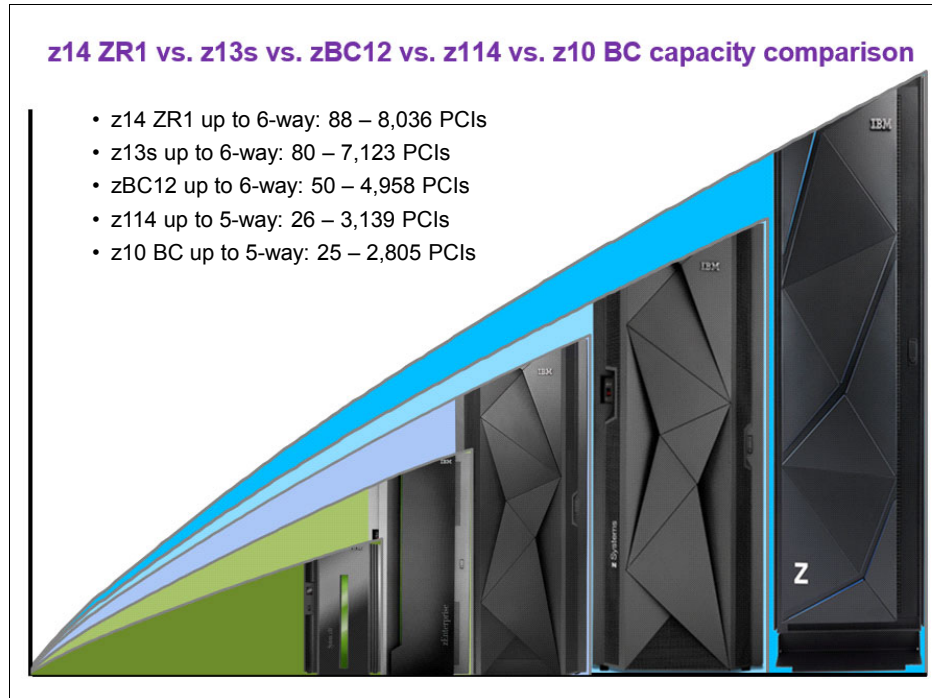


Figure 12-1 IBM Z generations capacity comparison

The Large System Performance Reference (LSPR) provides capacity ratios among various processor families that are based on various measured workloads. It is a common practice to assign a capacity scaling value to processors as a high-level approximation of their capacities. The numbers for z14 ZR1 were obtained with z/OS V2R2.

For z/OS V2R2 studies, the capacity scaling factor that is commonly associated with the reference processor is set to a 2094-701 with a Processor Capacity Index (PCI) value of 593. This value is unchanged since z/OS V1R11 LSPR. The use of the same scaling factor across LSPR releases minimizes the changes in capacity results for an older study and provides more accurate capacity view for a new study.

On average, z14 ZR1 servers can deliver up to 12% more performance in a 6-way configuration than an z13s 6-way. However, the observed performance increase varies depending on the workload type.

Consult the LSPR when you consider performance on the z14 ZR1. The range of performance ratings across the individual LSPR workloads is likely to include a large spread. Performance of the individual logical partitions (LPARs) varies depending on the fluctuating resource requirements of other partitions and the availability of processor units (PUs). For more information, see 12.7, “Workload performance variation” on page 404.

For more information about performance, see the [Large Systems Performance Reference for IBM Z page](#) of the Resource Link website.

For more information about millions of service units (MSU) ratings, see the [IBM z Systems Software Contracts page](#) of the IBM IT infrastructure website.

12.2 LSPR workload suite

Historically, LSPR capacity tables, including pure workloads and mixes, were identified with application names or a *software* characteristic; for example, CICS, IMS, OLTP-T,¹ CB-L,² LoIO-mix,³ and TI-mix.⁴ However, capacity performance is more closely associated with how a workload uses and interacts with a particular processor *hardware* design.

The CPU Measurement Facility (CPU MF) data that was introduced on the z10 provides insight into the interaction of workload and *hardware design* in production workloads. CPU MF data helps LSPR to adjust workload capacity curves that are based on the underlying hardware sensitivities; in particular, the processor access to caches and memory. This processor access to caches and memory is called *nest*. By using this data, LSPR introduces three workload capacity categories that replace all older primitives and mixes.

LSPR contains the internal throughput rate ratios (ITRRs) for the z14 ZR1 and the previous generation processor families. These ratios are based on measurements and projections that use standard IBM benchmarks in a controlled environment.

The throughput that any user experiences can vary depending on the amount of multiprogramming in the user's job stream, the I/O configuration, and the workload processed. Therefore, no assurance can be given that an individual user can achieve throughput improvements that are equivalent to the performance ratios that are stated.

12.3 Fundamental components of workload performance

Workload performance is sensitive to the following major factors:

- ▶ Instruction path length
- ▶ Instruction complexity
- ▶ Memory hierarchy and memory nest

These factors are described in this section.

12.3.1 Instruction path length

A transaction or job runs a set of instructions to complete its task. These instructions are composed of various paths through the operating system, subsystems, and application. The total count of instructions that are run across these software components is referred to as the *transaction or job path length*.

The path length varies for each transaction or job, and depends on the complexity of the tasks that must be run. For a particular transaction or job, the application path length tends to stay the same, assuming that the transaction or job is asked to run the same task each time.

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.

However, the path length that is associated with the operating system or subsystem can vary based on the following 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 number of logical processors (*n-way*) of the image or LPAR. As the number of logical processors grows, more instructions are needed to manage resources that are serialized by latches and locks.

12.3.2 Instruction complexity

The type of instructions and the sequence in which they are run interacts with the design of a microprocessor to affect a performance component. This factor is defined as *instruction complexity*. The following design alternatives affect this component:

- ▶ Cycle time (GHz)
- ▶ Instruction architecture
- ▶ Pipeline
- ▶ Superscalar
- ▶ Out-of-order execution
- ▶ Branch prediction
- ▶ Transaction Lookaside Buffer (TLB)
- ▶ Transactional Execution (TX)
- ▶ Single instruction multiple data instruction set (SIMD)
- ▶ Simultaneous multithreading (SMT)⁵

Performance varies as workloads are moved between microprocessors with various designs. However, when on a processor, this component tends to be similar across all models of that processor.

12.3.3 Memory hierarchy and memory nest

The *memory hierarchy* of a processor generally refers to the caches, data buses, and memory arrays that stage the instructions and data that must be run on the microprocessor to complete a transaction or job.

The following design choices affect this component:

- ▶ Cache size.
- ▶ Latencies (sensitive to distance from the microprocessor).
- ▶ Number of levels, the Modified, Exclusive, Shared, Invalid (MESI) protocol, controllers, switches, the number and bandwidth of data buses, and so on.

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 an IBM Z processor refers to the shared caches and memory along with the data buses that interconnect them.

⁵ Available for IFL, zIIP, and SAP processors only.

A memory nest in a fully populated z14 ZR1 CPC drawer is shown in Figure 12-2.

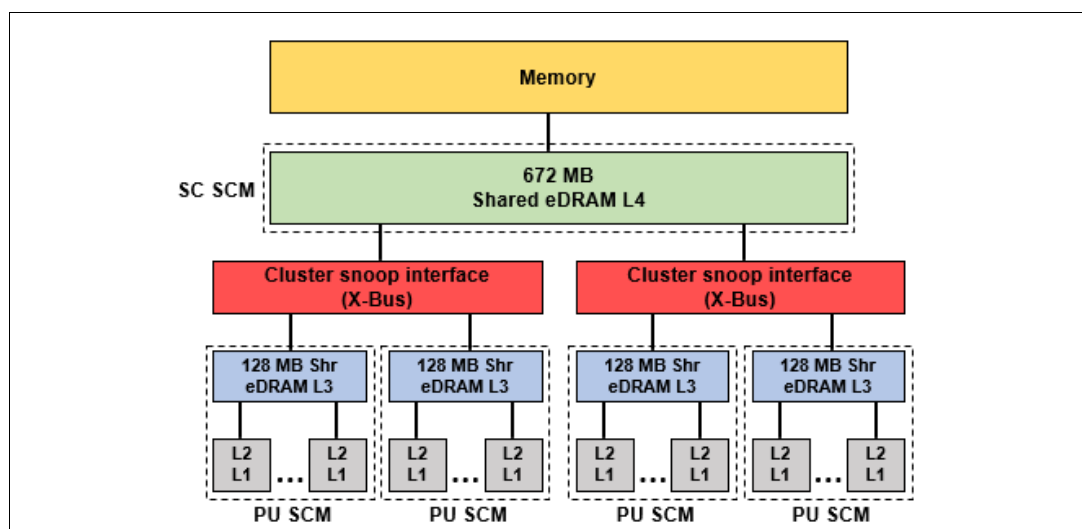


Figure 12-2 Memory hierarchy in a fully populated z14 ZR1 CPC drawer

Workload performance is sensitive to how deep into the memory hierarchy the processor must go to retrieve the workload instructions and data for running. The best performance occurs when the instructions and data are in the caches nearest the processor because little time is spent waiting before running. If the instructions and data must be retrieved from farther out in the hierarchy, the processor spends more time waiting for their arrival.

As workloads are moved between processors with various memory hierarchy designs, performance varies because the average time to retrieve instructions and data from within the memory hierarchy varies. Also, when on a processor, this component continues to vary because the location of a workload's instructions and data within the memory hierarchy is affected by several factors that include, but are not limited to, the following factors:

- ▶ Locality of reference
- ▶ I/O rate
- ▶ Competition from other applications and LPARs

12.4 Relative Nest Intensity

The most performance-sensitive area of the memory hierarchy is the activity to the memory nest. This area is the distribution of activity to the shared caches and memory.

The term Relative Nest Intensity (RNI) indicates the level of activity to this part of the memory hierarchy. By using data from CPU MF, the RNI of the workload that is running in an LPAR can be calculated. The higher the RNI, the deeper into the memory hierarchy the processor must go to retrieve the instructions and data for that workload.

RNI reflects the distribution and latency of sourcing data from shared caches and memory, as shown in Figure 12-3.

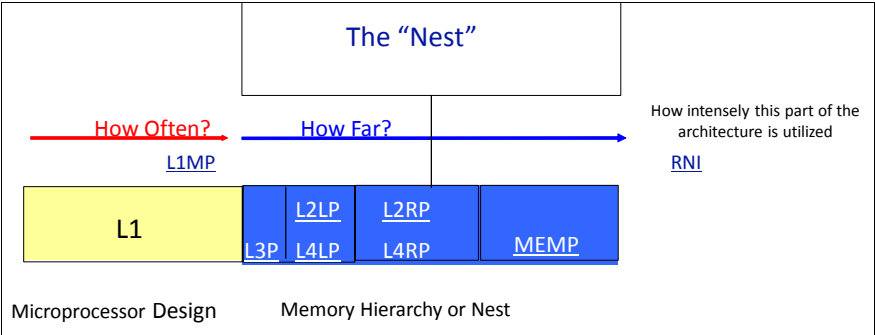


Figure 12-3 Relative Nest Intensity

Many factors influence the performance of a workload. However, these factors often are influencing is the RNI of the workload. The interaction of all these factors results in a net RNI for the workload, which in turn directly relates to the performance of the workload.

These factors are tendencies, 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 tends to create a higher RNI. It is the net effect of the interaction of all these factors that determines the RNI.

The traditional factors that were used to categorize workloads in the past are shown with their RNI tendency in Figure 12-4.

Relative Nest Intensity		
Low		High
Batch	Application Type	Transactional
Low	IO Rate	High
Single	Application Mix	Many
Intensive	CPU Usage	Light
Low	Dispatch Rate	High
High locality	Data Reference Pattern	Diverse
Simple	LPAR Configuration	Complex
Extensive	Software Configuration Tuning	Limited

Figure 12-4 Traditional factors that were used to categorize workloads

Little can be done to affect most of these factors. An application type is whatever is necessary to do the job. The data reference pattern and processor usage tend to be inherent to the nature of the application. The LPAR configuration and application mix are mostly a function of what must be supported on a system. The I/O rate can be influenced somewhat through buffer pool tuning.

However, one factor, *software configuration tuning*, is often overlooked but can have a direct effect on RNI. This term refers to the number of address spaces (such as CICS application-owning regions [AORs] or batch initiators) that are needed to support a workload.

This factor always existed, but its sensitivity is higher with the current high-frequency microprocessors. Spreading the same workload over more address spaces than necessary can raise a workload's RNI. This increase occurs because the working set of instructions and data from each address space increases the competition for the processor caches.

Tuning to reduce the number of simultaneously active address spaces to the optimum number that is needed to support a workload can reduce RNI and improve performance. In the LSPR, the number of address spaces for each processor type and *n-way* configuration is tuned to be consistent with what is needed to support the workload. Therefore, the LSPR workload capacity ratios reflect a presumed level of software configuration tuning. Retuning the software configuration of a production workload as it moves to a larger or faster processor might be needed to achieve the published LSPR ratios.

12.5 LSPR workload categories based on relative nest intensity

A workload's RNI is the most influential factor in determining workload performance. Other more traditional factors, such as application type or I/O rate, have RNI tendencies. However, it is the net RNI of the workload that is the underlying factor in determining the workload's 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.

The following workload categories are represented in the LSPR tables:

- ▶ **LOW** (relative nest intensity)
A workload category that represents light use of the memory hierarchy.
- ▶ **AVERAGE** (relative nest intensity)
A workload category that represents average use of the memory hierarchy. This category is expected to represent most production workloads.
- ▶ **HIGH** (relative nest intensity)
A workload category that represents a heavy use of the memory hierarchy.

These categories are based on the RNI. The RNI is influenced by many variables, such as application type, I/O rate, application mix, processor usage, data reference patterns, LPAR configuration, and the software configuration that is running. CPU MF data can be collected by z/OS System Measurement Facility on SMF 113 records or z/VM Monitor starting with z/VM V5R4.

12.6 Relating production workloads to LSPR workloads

Historically, the following techniques were used to match production workloads to LSPR workloads:

- ▶ Application name (a client that is 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 that are used a mix of low I/O rate LSPR workloads)

The IBM Processor Capacity Reference for IBM Z (zPCR) tool supports the following workload categories:

- ▶ Low
- ▶ Low-Average

- ▶ Average
- ▶ Average-high
- ▶ High

For more information about the no-charge IBM zPCR tool (which reflects the latest IBM LSPR measurements), see the [Getting Started with zPCR \(IBM's Processor Capacity Reference\)](#) page of the IBM Techdocs Library website.

As described in 12.5, “LSPR workload categories based on relative nest intensity” on page 403, the underlying performance sensitive factor is how a workload interacts with the processor hardware.

Beginning with the z10 processor, the hardware characteristics can be measured by using CPU MF (SMF 113) counters data. A production workload can be matched to an LSPR workload category through these hardware characteristics. For more information about RNI, see 12.5, “LSPR workload categories based on relative nest intensity” on page 403.

The AVERAGE RNI LSPR workload is intended to match most client workloads. When no other data is available, use the AVERAGE RNI LSPR workload for capacity analysis.

Low-Average and Average-High categories allow better granularity for workload characterization.

For z10 and newer processors, the CPU MF data can be used to provide an extra hint as to workload selection. When available, this data allows the RNI for a production workload to be calculated.

By using the RNI and another factor from CPU MF, the L1MP (percentage of data and instruction references that miss the L1 cache), a workload can be classified as LOW, AVERAGE, or HIGH RNI. This classification and resulting hit are automated in the zPCR tool. It is preferable to use zPCR for capacity sizing.

12.7 Workload performance variation

As the size of transistors approaches the size of atoms that stand as a fundamental physical barrier, a processor chip's performance can no longer double every two years (known as the Moore's Law⁶).

A holistic performance approach is required when the performance gains because of frequency are reduced today. Therefore, hardware and software synergy becomes an absolute requirement.

Starting with z13, Instructions Per Cycle (IPC) improvements in core and cache became the driving factor for performance gains. As these microarchitectural features increase (which contributes to instruction parallelism), overall workload performance variability also increases because not all workloads react the same way to these enhancements. Also, the memory and cache designs affect various workloads in many ways. All workloads are improved, with cache-intensive loads expected to benefit the most.

The workload variability for moving from z13s to z14 ZR1 is expected to be stable. Workloads that are migrating from zBC12 and previous generations to z14 ZR1 can expect to see similar results with slightly less variability than the typical z13s experience.

⁶ For more information, see the [Moore's Law website](#).

The effect of this variability is increased deviations of workloads from single-number metric-based factors, such as millions of instructions per second (MIPS), MSUs, and CPU time charge-back algorithms.

Experience demonstrates that IBM Z servers can be run at up to 100% utilization levels, sustained. However, most clients prefer to leave some room and run at 90% or slightly under. For any capacity comparison exercise that uses a single metric, such as MIPS or MSU, is not a valid method. When deciding the number of processors and the uniprocessor capacity, consider the workload characteristics and LPAR configuration. For these reasons, the use of zPCR and involving IBM technical support are recommended when you plan capacity.

12.7.1 Main performance improvement drivers with z14 ZR1 servers

z14 ZR1 servers deliver new levels of performance and capacity for large-scale consolidation and growth. The attributes and design points of z14 ZR1 servers contribute to overall performance and throughput improvements as compared to the z13s.

The z/Architecture implementation includes the following enhancements:

- ▶ Guarded Storage Facility: An enhancement for garbage collected languages.
- ▶ Vector Packed Decimal Facility: An enhancement of packed decimal operations.
- ▶ Vector Enhancements Facility: Includes several vector enhancements, such as adding support for single precision floating point and VMSL for cryptographic computations.
- ▶ Order Preserving Compression and Entropy Encoding for CMPSC: Allows comparisons against compressed data. Entropy Encoding increases compression ratio.
- ▶ Miscellaneous New General Instructions: An enhancement of 64-bit halfword operations and new multiply instructions.
- ▶ Removal of Runtime Instrumentation External Interruption: Avoids RI Buffer Full interrupt.
- ▶ Semaphore Assist Facility, Enhanced NIAI (next instruction access intent) code points.
- ▶ (MSA 6, 7, 8), adding SHA-3 hash, true random number generation, and AES-GCM mode.

The z14 ZR1 microprocessor includes the following design enhancements:

- ▶ 14nm FINFET SOI technology with IBM embedded dynamic static random access memory (eDRAM) technology
- ▶ Up to nine active processor cores per chip
- ▶ Clock frequency at 4.5 GHz
- ▶ A new translation/TLB2 design with four hardware-implemented translation engines reducing latency when compared with one pico-coded engine on z13s
- ▶ Branch prediction improvements:
 - 33% Branch Target Buffer (BTB)1-and-2 growth
 - New perceptron predictor
 - Simple call-return stack
- ▶ Pipeline optimization:
 - Improved instruction delivery
 - Faster branch wake-up
 - Reduced execution latency
 - Improved OSC prediction

- ▶ Second generation of Simultaneous multithreading (SMT):
 - Includes SAPs, zIIPs, and IFLs
 - Improved thread balancing
 - Multiple outstanding translations
 - Optimized hang avoidance mechanisms
- ▶ Improved Hot Cache Line handling; dynamic throttling
- ▶ Cache improvements:
 - New power efficient logical directory design
 - L1 I-Cache increased from 96 K to 128 K per Core (33%)
 - L2 D-Cache increased from 2 MB to 4 MB per Core (2x)
 - L3 Cache (shared) increased from 64 MB to 128 MB per PU SCM (2x)
 - New L4 Cache design with 672 MB (shared) per drawer and L4 sequential prefetch
- ▶ Enhanced binary coded decimal architecture (full set of register-to-register BCD instructions)
- ▶ New instructions for Single-instruction multiple-data (SIMD) operations
- ▶ One cryptographic/compression co-processor per core, redesigned:
 - CP Assist for Cryptographic Functions (CPACF) (hardware) runs more operations, such as SHA-3, SHAKE hashing algorithms, and True Random-number generation (TRNG)
 - Improved Galois Counter Mode (GCM) performance
 - Entropy-Encoding Compression Enhancement with Huffman encoding
 - Order-Preserving compression
- ▶ Adjusted Hiperdispatch to use new chip configuration

The z14 ZR1 design features the following enhancements as compared with the z13s:

- ▶ Increased number of characterizable cores, from 20 to 30
- ▶ Hardware system area (HSA) increased from 40 GB to 64 GB
- ▶ A total of 8 TB of addressable memory (configurable to LPARs)
- ▶ PR/SM enhancements:
 - Improved memory affinity
 - Optimized LPAR placement algorithms
- ▶ Dynamic Partition Manager Version 3.2:
 - FC (with z14 hardware) and FCP storage support
 - Storage Groups management enhancements
- ▶ SMT enablement for system assist processor (SAP) processors
- ▶ New Coupling Facility Control Code (CFCC) with improved performance and following enhancements:
 - Asynchronous Cross-Invalidate (XI) of CF cache Structures
 - Coupling Facility (CF) Processor Scalability
 - CF List Notification Enhancements
 - CF Request Diagnostics
 - Coupling Link Constraint Relief
 - CF Encryption
 - Asynchronous duplexing of CF lock structures

The following new features are available on z14 ZR1 servers:

- Dynamic I/O for Standalone Coupling Facility CPCs
- Coupling Express Long Reach
- zHyperlink Express
- OSA-Express7S⁷ 25GbE SR
- FICON Express16S+
- 25GbE RoCE Express2
- 10GbE RoCE Express2
- Crypto Express6S with up to 40 domains

⁷ Check the IBM United States Hardware Announcement 118-075 and Driver Exception Letter for feature availability.



IBM Secure Service Container framework

Naming: The IBM z14 server generation is available as the following machine types and models:

- ▶ Machine Type 3906 (M/T 3906), Models M01, M02, M03, M04, and M05 → further identified as *IBM z14 Model M0x*, or *z14 M0x*.
- ▶ Machine Type 3907 (M/T 3907), Model ZR1 → further identified as *IBM z14 Model ZR1*, or *z14 ZR1*.

Unless otherwise specified, *IBM z14 (z14)* refers to both machine types in the remainder of this appendix.

In this appendix, the IBM Secure Service Container¹ (SSC) framework is described. The SSC framework is available on IBM z14, and IBM z13 and z13s (Driver level 27) systems.

This appendix also briefly describes the reason why IBM created the SSC framework and how the SSC environment is intended to be used.

This appendix includes the following topics:

- ▶ A.1, “What is IBM Secure Service Container?” on page 410
- ▶ A.2, “SSC LPAR” on page 410
- ▶ A.3, “Why Secure Service Container?” on page 411
- ▶ A.4, “IBM Z and Secure Service Container” on page 411

¹ Secure Service Container is the infrastructure required to deploy appliances (framework) in a secure container on supported IBM Z hardware. With *IBM United States Software Announcement 218-152*, dated October 2, 2018, IBM introduces *IBM Secure Service Container for IBM Cloud Private*. IBM Cloud™ Private is a Platform as a Service (PaaS) environment for developing and managing containerized applications.

A.1 What is IBM Secure Service Container?

An appliance is an application (software) that provides a specified function or set of functions (service). It is packaged and deployed with a specific (trimmed) operating system in a virtual machine or a dedicated commodity of the shelf (COTS) hardware (physical server). It requires little to no intervention from a system administrator (software update, OS update, and maintenance).

An appliance must satisfy various requirements, such as certified functionality and security (the function it provides must be tamper-resistant, even from system administrators or other privileged users) and simple deployment and maintenance.

In the current IT deployments, various components that serve the business processes (databases, middleware, applications, and so on) require specialized management functions (such as access management, enterprise directories, secure key management, backup and restore). The development requirements of the management functions do not follow the dynamic of the actual business functions.

Because of the diversity of the platforms on which the business applications run, the management function must be maintained (updated, tested, or even certified) if the management functions are deployed alongside the mainstream business applications when the platform must be maintained or upgraded. However, the complexity and associated IT spending is increased.

As such, these management functions can be deployed by using an appliance model in which the functions that are provided are available and accessible through standardized methods.

Many appliances are available from various suppliers. Each appliance includes the following features:

- ▶ Separate administration and deployment process
- ▶ Different hardware configuration requirements
- ▶ Different performance profile and management requirements
- ▶ Different security characteristics that require alignment with enterprise requirements

A.1.1 SSC framework

IBM developed the SSC framework. This framework provides the base infrastructure to create and deploy an appliance, including operating system, middleware, Software Development Kit (SDK), and firmware support. A special feature of the IBM SSC framework is that it protects the deployed workload from being accessed by a system administrator or an external attacker.

A.2 SSC LPAR

For IBM Z, the SSC Partition is an LPAR type that runs an appliance based on Secure Service Container framework.

Multiple virtual appliances that are integrated into IBM Secure Service Container can be deployed on IBM z14 (z13 and z13s also). These virtual appliances include the following common features:

- ▶ Administration (deployment)
- ▶ Hardware configuration

- ▶ Managed performance profiles
- ▶ Security characteristics (aligned with enterprise requirements)

At the time of this writing, the following appliances are available from IBM:

- ▶ z/VSE Network Appliance.
- ▶ IBM Z Advanced Workload Analysis Reporter (IBM zAware), which is now deployed as a software appliance and integrated with IBM Operations Analytics for Z.

More appliances are expected in the future. Appliances can be implemented as firmware or software, depending on the environment on which the appliance runs and the function it must provide.

The SSC framework is available on IBM z14, z13, and z13s.

A.3 Why Secure Service Container?

The SSC framework simplifies the process that a team must apply to create an appliance. It also enforces a common set of behaviors for operations that all appliances must perform.

The SCC framework also provides a set of utilities that is used to implement the common functions that all appliances need (FFDC, network setup, appliance configuration, and so on.). An application developer can use the SSC framework to turn a solution into a stand-alone appliance that is easily installed onto the IBM Z platform.

The SSC framework enables the release a product as software or firmware that is based on a business decision, not on a technical decision.

Deploying an appliance takes minutes. Appliances do not require any operating system knowledge or middleware knowledge. They allow users to focus on the core services they deliver.

A.4 IBM Z and Secure Service Container

Appliances that are based on the SSC framework share the following features and characteristics:

- ▶ Encapsulated operating systems
- ▶ Services that are provided by using Remote APIs (RESTful) and web interfaces
- ▶ Embedded monitoring and self-healing
- ▶ End-to-end tamper-protection
- ▶ Protected intellectual property
- ▶ Tested and qualified by IBM for a specific use case
- ▶ Can be delivered as firmware or software

The deployment model for an appliance is shown in Figure A-1.

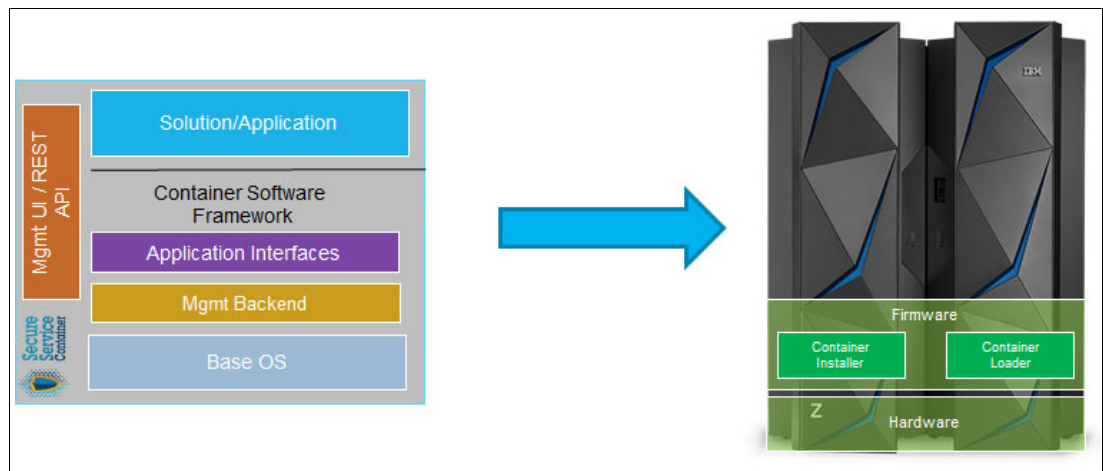


Figure A-1 Appliance deployment in an SSC LPAR on IBM Z

SSC provides a highly secure context (see Figure A-2) for deploying appliances that include the following features:

- ▶ Allows no system admin access:
 - After the appliance image is built, OS access is not possible
 - Only Remote APIs are available
 - Memory access of system admin is disabled
- ▶ Data storage uses encrypted disk
- ▶ Debug data (dumps) is encrypted
- ▶ Strong isolation between container instances
- ▶ High assurance isolation

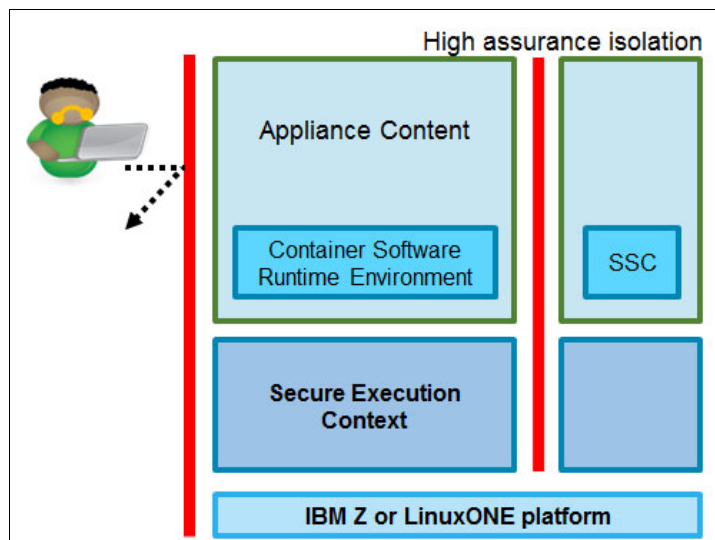


Figure A-2 Secure Service Container protection

The process that is used to deploy an appliance includes the following steps (see Figure A-3):

1. Purchasing the software appliance.
2. Downloading the appliance image.
3. Creating and activating the SSC LPAR.
4. Deploying the appliance by using the appliance installer.
5. Configuring and using the appliance through REST API or a web UI.

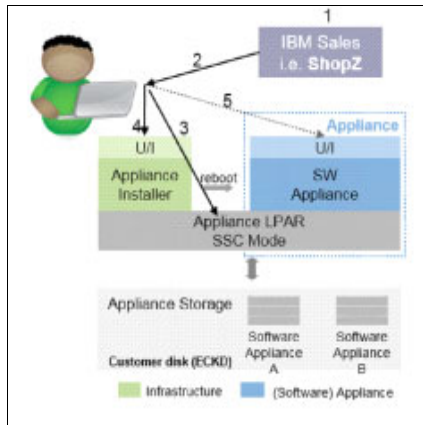


Figure A-3 Deployment in five steps

The SSC framework provides following appliance management controls for appliance administrators:

- ▶ View messages and events
- ▶ Manage network, users, and disks
- ▶ View appliance status
- ▶ Export and import data
- ▶ Apply services and updates
- ▶ Support for software license

At the time of this writing, the SSC software framework supports the following components:

- ▶ FCP and ECKD storage
- ▶ Dynamic Partition Manager
- ▶ User management within appliance with LDAP
- ▶ Enhanced network and storage management user interface (UI)
- ▶ File system with embedded CRC checking
- ▶ Include KVM, qemu, and virsh packages
- ▶ Embedded OS upgrades
- ▶ Support smart card machine unique key handling



Channel options

This appendix describes all of the channel attributes, required cable types, maximum unrepeated distance, and bit rate for z14 ZR1. The features that are hosted in the PCIe drawer for Cryptography and Compression are also listed.

For all optical links, the connector type is LC Duplex, except for the zHyperLink, the 12xIFB, and the ICA SR connections, which are established with multifiber push-on (MPO) connectors.

The MPO connector of the zHyperLink and the ICA connection feature two rows of 12 fibers and are interchangeable. The electrical Ethernet cable for the Open Systems Adapter (OSA) connectivity is connected through an RJ45 jack.

The attributes of the channel options that are supported on z14 ZR1 are listed in Table B-1.

Table B-1 z14 ZR1 channel feature support

Channel feature	Feature codes	Bit rate ^a in Gbps (or stated)	Cable type	Maximum unrepeated distance ^b	Ordering information
zHyperLink and Fiber Connection (FICON)					
zHyperlink Express	0431	8 Gbps	OM3,	100 m	New build
			OM4	150 m	
FICON Express16S+ 10KM LX	0427	4, 8, or 16	SM 9 μ m	10 km (6.2 miles)	New build
FICON Express16S+ SX	0428	4, 8, or 16	OM2, OM3, OM4	See Table B-2 on page 417.	New build
FICON Express16S 10KM LX	0418	4, 8, or 16	SM 9 μ m	10 km (6.2 miles)	Carry forward
FICON Express16S SX	0419	4, 8, or 16	OM2, OM3, OM4	See Table B-2 on page 417.	Carry forward

Channel feature	Feature codes	Bit rate ^a in Gbps (or stated)	Cable type	Maximum unrepeat distance ^b	Ordering information
FICON Express8S 10KM LX	0409	2, 4, or 8	SM 9 µm	10 km (6.2 miles)	Carry forward
FICON Express8S SX	0410	2, 4, or 8	OM2, OM3, OM4	See Table B-2 on page 417.	Carry forward
Open Systems Adapter (OSA) and Remote Direct Memory over Converged Ethernet (RoCE)					
OSA-Express7S 25GbE SR	0429	25	MM 50 µm	70 m (2000) 100 m (4700)	New build
OSA-Express6S 10 GbE LR	0424	10	SM 9 µm	10 km (6.2 miles)	New build
OSA-Express5S 10 GbE LR	0415				Carry forward
OSA-Express6S 10 GbE SR	0425	10	MM 62.5 µm	33 m (200)	New build
OSA-Express5S 10 GbE SR	0416		MM 50 µm	82 m (500) 300 m (2000)	Carry forward
OSA-Express6S GbE LX	0422	1.25	SM 9 µm	5 km (3.1 miles)	New build
OSA-Express5S GbE LX	0413				Carry forward
OSA-Express4S GbE LX	0404				
OSA-Express6S GbE SX	0423	1.25	MM 62.5 µm	275 m (200)	New build
OSA-Express5S GbE SX	0414			550 m (500)	Carry forward
OSA-Express4S GbE SX	0405		MM 50 µm		
OSA-Express6S 1000BASE-T	0426	100 or 1000 Mbps	Cat 5, Cat 6 unshielded twisted pair (UTP)	100 m	New build
OSA-Express5S 1000BASE-T	0417				Carry forward
25GbE RoCE Express2	0430	25	OM4	100 m ^c	New build
10GbE RoCE Express2	0412	10	OM4	300 m ^c	New build
10GbE RoCE Express	0411	10	OM3	300 m ^c	Carry forward
Parallel Sysplex					
CE LR	0433	10 Gbps	SM 9 µm	10 km (6.2 miles)	New build
ICA SR	0172	8 GBps	OM4	150 m	New build or Carry forward
			OM3	100 m	
IC	N/A		N/A	N/A	N/A
Cryptography and Compression					
Crypto Express6S	0893	N/A	N/A	N/A	New build
Crypto Express5S	0890	N/A	N/A	N/A	Carry forward
zEDC Express	0420	N/A	N/A	N/A	New build or Carry forward

- a. The link data rate does not represent the actual performance of the link. The actual performance depends on many factors, including latency through the adapters, cable lengths, and the type of workload.
- b. Where applicable, the minimum fiber bandwidth distance in MHz·km for multi-mode fiber optic links is included in parentheses.
- c. A 600 meters maximum when sharing the switch across 2 RoCE Express features.

The maximum unrepeated distances for FICON short wavelength (SX) features are listed in Table B-2.

Table B-2 Maximum unrepeated distance for FICON SX features

Cable type/bit rate	1 Gbps	2 Gbps	4 Gbps	8 Gbps	16 Gbps
OM1 (62.5 µm at 200 MHz·km)	300 meters	150 meters	70 meters	21 meters	N/A
	984 feet	492 feet	230 feet	69 feet	N/A
OM2 (50 µm at 500 MHz·km)	500 meters	300 meters	150 meters	50 meters	35 meters
	1640 feet	984 feet	492 feet	164 feet	115 feet
OM3 (50 µm at 2000 MHz·km)	860 meters	500 meters	380 meters	150 meters	100 meters
	2822 feet	1640 feet	1247 feet	492 feet	328 feet
OM4 ^a (50 µm at 4700 MHz·km)	N/A	500 meters	400 meters	190 meters	125 meters
	N/A	1640 feet	1312 feet	693 feet	410 feet

- a. Fibre Channel Standard (not certified for Ethernet)



Native Peripheral Component Interconnect Express

This appendix introduces the native Peripheral Component Interconnect Express (PCIe) adapters management on IBM Z servers (IBM z14, z13, and z13s)¹. The appendix also includes concepts of the integrated firmware processor (IFP) and resource groups (RG).

This appendix includes the following topics:

- ▶ C.1, “Design of native PCIe adapter management” on page 420
- ▶ C.2, “Native PCIe adapters plugging rules” on page 422
- ▶ C.3, “Native PCIe adapter definitions” on page 422

¹ IBM zEC12 and zBC12 also support the 10GbE RoCE Express feature (FC 0411), but one feature must be dedicated to one LPAR.

C.1 Design of native PCIe adapter management

The native PCIe adapter is a new category of features that was introduced in zEC12. IBM z14 ZR1 supports the following native PCIe features:

- ▶ 25GbE RoCE Express2 (FC 0430)²
- ▶ 10GbE RoCE Express2 (FC 0412)
- ▶ 10GbE RoCE Express (FC 0411) (carry forward only)
- ▶ zEDC Express (FC 0420)
- ▶ Coupling Express Long Reach (CE LR) (FC 0433)
- ▶ zHyperLink Express (FC 0431)³

These adapters are installed into a PCIe+ I/O drawer and are identified by a physical channel ID (PCHID) that is assigned according to the physical location.

For adapters that are installed in a PCIe+ I/O drawer, management functions in the form of device drivers and diagnostic tools are always implemented to support virtualization of the adapter, service, and maintenance.

Traditionally, these management functions are integrated on the adapter with specific hardware design (FICON Express, OSA-Express). For most of the newly introduced native PCIe adapters, these functions are moved out of the adapter and are now handled by an Integrated Firmware Processor (IFP) which runs the Resource Group (RG) firmware.

C.1.1 Native PCIe adapter

For traditional I/O adapters, such as the Open Systems Adapter (OSA) and Fibre Channel connection (FICON) cards, the application-specific integrated circuit (ASIC) chip on the adapter always downloads the device drivers and diagnostic tools from the Support Element (SE) and runs the management functions on the adapter. With the native PCIe adapter design, no ASIC chip is used for management (including virtualization) function on the native PCIe adapters.

For the RoCE Express, Coupling Express Long Reach, and zEDC, device drivers and diagnostic tools are now running on the IFP and use four RGs. Management functions, including virtualization, servicing and recovery, diagnostics, failover, firmware updates against an adapter, and other functions are implemented within the RG microcode. For the zHyperlink Express adapters, only service (diagnostics) and firmware maintenance are handled for the Resource Group microcode.

C.1.2 Integrated firmware processor

The IFP is a processor unit (PU) that is exclusively used to manage native PCIe adapters that are installed in the PCIe+ I/O drawer. It is allocated from the system PU pool and is not counted in the PUs available for characterization.

If a native PCIe adapter is installed in the system, the system allocates and initializes an IFP during its power-on reset (POR) phase. Although the IFP is allocated to one of the physical PUs, it is not visible to the users. In an error or failover scenario, PU sparing also applies for IFP, with the same rules as other PUs.

² Unless otherwise specified, RoCE Express2 refers to both 25GbE and 10GbE RoCE Express2 features (FC 0430 and FC 0412, respectively) for the remainder of this appendix.

³ Although the zHyperLink Express (FC 0431) is assimilated to a native PCIe adapter, it is managed only for service actions and firmware maintenance (updates).

C.1.3 Resource groups

The IFP allocates four resource groups for running the management functions of native PCIe adapters. A native PCIe adapter is managed by one of the resource groups according to the adapter location in the PCIe+ I/O drawer.

As shown in Figure C-1, each I/O domain in a PCIe+ I/O drawer of a z14 ZR1 server is driven by a PCIe switch that is connected to the PCIe I/O feature that is in the CPC drawer. Each slot in a PCIe_ I/O drawer is attached to one resource group. The native PCIe I/O adapters are managed by their respective RG for device drivers and diagnostic tools functions.

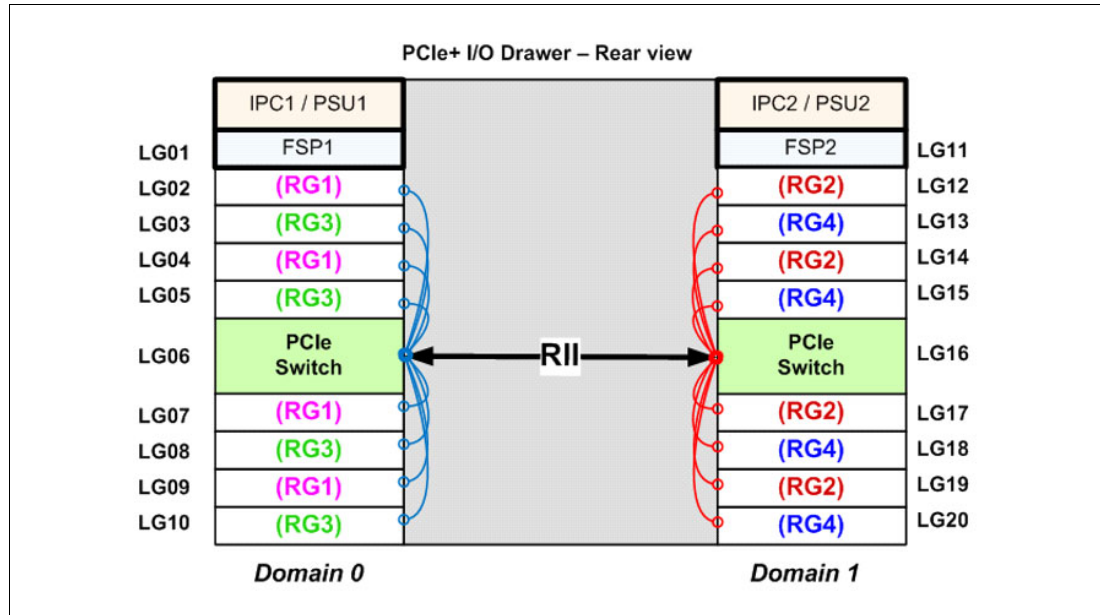


Figure C-1 I/O domains and resource groups that are managed by the IFP - z14 ZR1

Up to four PCIe+ I/O drawers are supported on z14 ZR1 servers. The same type of native PCIe adapter is always assigned to different I/O domains and in different resource groups (and different PCIe I/O drawers if the configuration includes them) to eliminate the possibility of a single point of failure.

C.1.4 Service and management tasks

The IFP and resource groups perform the following management tasks on the native PCIe adapters:

- Firmware update of adapters and resource groups
- Error recovery and failure data collection
- Diagnostic and service tasks

Firmware update of adapters and resource groups

The firmware of native PCIe adapters and resource groups is part of the system's microcode and can be updated by a Microcode Change Level (MCL) upgrade. MCL upgrades on adapters or on the code of the resource groups require the specific adapter or all native PCIe adapters that are managed by the specific resource group (depending on the type of microcode that it applied) to be offline during activation of the MCL.

However, to maintain availability, MCLs can be applied to only one resource group at a time. While one resource group is offline, the other three resource groups and all their adapters remain active. An MCL application for a native PCIe adapter or resource group is not possible if an error condition exists within the other resource groups.

Error recovery and failure data collection

If an error occurs in one of the resource groups or features that are assigned to one of the resource groups, the IFP manages error recovery and collects error data. The error data is sent by the IFP to the SE, which then provides a message on the SE and the Hardware Management Console (HMC). If an error requires maintenance, a call home to the IBM Support system is started by the HMC.

Diagnostic and service tasks

Any maintenance of a native PCIe feature is managed by the IFP, including testing or replacing a feature card. Before a feature is configured offline, the IFP ensures that the same type of feature is available in the same or one of the other resource groups (if applicable).

C.2 Native PCIe adapters plugging rules

The following maximum number of native PCIe adapters can be installed in a z14 ZR1 server:

- ▶ Up to four RoCE Express/Express2 adapters.
- ▶ Up to 8 zEDC Express adapters.
- ▶ Up to 16 zHyperLink Express adapters.
- ▶ Up to 16 Coupling Express Long Reach adapters.

Considering availability, install adapters of the same type in slots of different I/O domains, drawers, fanouts, and resource groups.

The next sections provide more information about achieving a highly available configuration.

C.3 Native PCIe adapter definitions

During the ordering process of the native PCIe adapters, such as the zEDC Express and RoCE Express/Express2, adapters of the same type are evenly spread across resource groups 1, 2, 3, and 4 for availability and serviceability.

A sample PCHID report of a z14 ZR1 configuration with four zEDC Express adapters and four 10GbE RoCE Express2 adapters is shown in Figure C-2 on page 423. The following information is listed for each adapter:

- ▶ PCHID and ports
- ▶ The Resource Group that the adapter is attached to (Comment column)
- ▶ Physical location (drawer, slot)

Source	Drwr	Slot	F/C*	PCHID/Ports or AID	Comment
A09/LG09/J01	A14B	02	0420	100	RG1
A09/LG09/J01	A14B	03	0420	104	RG3
A09/LG09/J01	A14B	04	0412	108/D1D2	RG1
A09/LG09/J01	A14B	05	0412	10C/D1D2	RG1
A09/LG10/J01	A14B	12	0420	120	RG2
A09/LG10/J01	A14B	13	0420	124	RG4
A09/LG10/J01	A14B	14	0412	128/D1D2	RG2
A09/LG10/J01	A14B	15	0412	12C/D1D2	RG2

Figure C-2 Sample output of AO data or PCHID report

The native PCIe adapters are not part of the traditional channel subsystem (CSS). Although they do not include a channel-path identifier (CHPID) assigned, they do include a PCHID that is assigned according to their physical location in the PCIe+ I/O drawer.

To define the native PCIe adapters in the HCD or HMC, a new I/O configuration program (IOCP) FUNCTION statement is introduced that includes several feature-specific parameters.

The IOCP example that is shown in Figure C-3 defines zEDC Express and 10GbE RoCE Express2 adapters to LPARs LP14 and LP15.

```

zEDC Express Functions for LPAR LP14, Reconfigurable to LP01:
FUNCTION FID=05,VF=1,PART=((LP14),(LP01)),TYPE=ZEDC,PCHID=100
FUNCTION FID=06,VF=1,PART=((LP14),(LP01)),TYPE=ZEDC,PCHID=104

zEDC Express Functions for LPAR LP15, Reconfigurable to LP02:
FUNCTION FID=07,VF=1,PART=((LP15),(LP02)),TYPE=ZEDC,PCHID=120
FUNCTION FID=08,VF=1,PART=((LP15),(LP02)),TYPE=ZEDC,PCHID=124

10GbE RoCE Express2 Functions for LPAR LP14, Reconfigurable to LP03 or LP04
FUNCTION FID=9,VF=01,PART=((LP14),(LP03,LP04)),PNETID=(NET1,NET2), *
      TYPE=ROC2,PCHID=108,PORT=1
FUNCTION FID=A,VF=01,PART=((LP14),(LP03,LP04)),PNETID=(NET1,NET2), *
      TYPE=ROC2,PCHID=10C,PORT=2

10GbE RoCE Express2 Functions for LPAR LP15, Reconfigurable to LP03 or LP04
FUNCTION FID=B,VF=01,PART=((LP15),(LP03,LP04)),PNETID=(NET1,NET2), *
      TYPE=ROC2,PCHID=128,PORT=1
FUNCTION FID=C,VF=01,PART=((LP15),(LP03,LP04)),PNETID=(NET1,NET2), *
      TYPE=ROC2,PCHID=12C,PORT=2

```

Figure C-3 Example of IOCP statements for zEDC Express and 10GbE RoCE Express2

C.3.1 FUNCTION identifier

The FUNCTION identifier (FID) is a hexadecimal number 000 - FFF that you use to assign a PCHID to the FUNCTION to identify the specific adapter in the PCIe+ I/O drawer. Because the FUNCTION is not related to a channel subsystem, all LPARs on a central processor complex (CPC) can be defined to it. However, a FUNCTION cannot be shared between LPARs. It is only dedicated or reconfigurable by using the **PART** parameter. The TYPE parameter is required for z14 ZR1.

C.3.2 Virtual function number

If you want several LPARs to use a zEDC Express or RoCE Express or Express2 adapter, you must use a Virtual Function (VF) number. A VF number is a number 1 - *nnn*, where *nnn* is the maximum number of LPARs that the feature supports. The maximum is 15 for the zEDC Express feature, 31 for the 10GbE RoCE Express (FC 0411), 62 for the RoCE Express2 adapter (31 VFs per physical port), and 254 for the zHyperLink Express feature⁴.

C.3.3 Physical network identifier

The physical network ID (PNETID) is required to set up the Shared Memory Communications over Remote Direct Memory Access (SMC-R) communication between two RoCE Express or Express2 features. Each FUNCTION definition supports up to four PNETIDs.

Notes: Consider the following points:

- ▶ For more information about FUNCTION statement, see *IBM Z Input/Output Configuration Program User's Guide for ICP IOCP*, SB10-7172.
- ▶ The definition of RoCE Express/Express2 feature is required to pair up with an OSD CHPID definition by using the parameter of PNETID. The OSD CHPID definition statement is not listed in the example.

⁴ The zHyperLink Express feature is not managed by the Resource Groups firmware.



Shared Memory Communications

Naming: The IBM z14 server generation is available as the following machine types and models:

- ▶ Machine Type 3906 (M/T 3906), Models M01, M02, M03, M04, and M05 → further identified as *IBM z14 Model M0x*, or *z14 M0x*.
- ▶ Machine Type 3907 (M/T 3907), Model ZR1 → further identified as *IBM z14 Model ZR1*, or *z14 ZR1*.

In the remainder of this document, *IBM z14 (z14)* refers to both machine types, unless otherwise specified.

This appendix briefly describes the optional Shared Memory Communications (SMC) function that is implemented on IBM Z servers as Shared Memory Communications over Remote Direct Memory Access (SMC-R) and Shared Memory Communications - Direct Memory Access (SMC-D) of IBM z14, z13, and z13s servers.

The following types adapters are available for SMC-R for physical connectivity:

- ▶ 25GbE RoCE Express2 (FC 0430)
- ▶ 10GbE RoCE Express2 (FC 0412)
- ▶ 10GbE RoCE Express (FC 0411): This adapter can be carried forward to z14 ZR1

Throughout this appendix, we use the terms *10GbE RoCE Express* and *RoCE Express* for both adapters, except when the adapter specifications differ.

This appendix includes the following topics:

- ▶ D.1, “Overview” on page 426
- ▶ D.2, “Shared Memory Communication over RDMA” on page 426
- ▶ D.3, “Shared Memory Communications - Direct Memory Access” on page 438

D.1 Overview

As the volume of data that is generated and transmitted by technologies that are driven by cloud, mobile, analytics, and social computing applications grows, pressure increases on business IT organizations to provide fast access to that data across the web, application, and database tiers that comprise most enterprise workloads. SMCs help to access data faster and with less latency. They also reduce CPU resource consumption over traditional TCP/IP for communications.

D.2 Shared Memory Communication over RDMA

SMC-R (RoCE) for IBM z14 ZR1 includes the following features:

- ▶ Remote Direct Memory Access (RDMA) technology provides low latency, high bandwidth, high throughput, and low processor utilization attachment between hosts.
- ▶ SMC-R is a protocol that allows TCP applications to benefit transparently from RDMA for transferring data. Consider the following points:
 - SMC-R uses RoCE Express adapter as the physical transport layer.
 - Initial deployment is limited to z/OS to z/OS communications with a goal to expand usage to more operating systems, and possibly appliances and accelerators.
- ▶ Single Root I/O Virtualization (SR-IOV) technology provides the capability to share the RoCE Express adapter between logical partitions (LPARs), with the following specifications for z14 ZR1:
 - 25GbE & 10GbE RoCE Express2 features support 31 Virtual Functions (VFs) per physical port for a total of 62 VFs per PCHID (z14 M0x support a different number of VFs for the FC 0412 and FC 0430).
 - 10GbE RoCE Express supports 31 VFs per PCHID.
- ▶ Maximum number of RoCE Express and RoCE Express2 adapters supported per z14 ZR1 is four (combined).

D.2.1 RDMA technology overview

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 the transfer of data to or from memory on a remote system with low latency, high throughput, and low CPU usage.

Traditional Ethernet transports, such as TCP/IP, typically use software-based mechanisms for error detection and recovery. They also are based on the underlying Ethernet fabric that uses a “best-effort” policy. With the traditional policy, the switches typically discard packets that are in congestion and rely on the upper-level transport for packet retransmission.

However, RoCE uses hardware-based error detection and recovery mechanisms that are defined by the InfiniBand specification. A RoCE transport performs best when the underlying Ethernet fabric provides a lossless capability, where packets are not routinely dropped.

This process 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 RoCE Express/Express2 adapters. This capability is enabled in the RoCE Express/Express2 adapter by default.

The following key requirements for RDMA are shown in Figure D-1:

- ▶ A reliable “lossless” Ethernet network fabric (LAN for layer 2 data center network distance)
- ▶ An RDMA network interface card (RNIC)

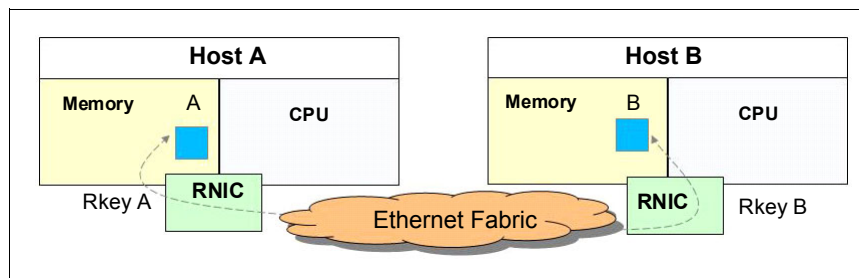


Figure D-1 RDMA technology overview

RDMA technology is now available on Ethernet. RoCE uses an Ethernet fabric (switches with Global Pause enabled) and requires advanced Ethernet hardware (RNICs on the host).

D.2.2 Shared Memory Communications over RDMA

SMC-R is a protocol that allows TCP socket applications to transparently use RDMA. It also is a “hybrid” solution (see Figure D-2 on page 428) that includes the following features:

- ▶ Uses a TCP connection to establish the SMC-R connection.
- ▶ A TCP option (SMCR) controls switching from TCP to “out-of-band” SMC-R.
- ▶ SMC-R information is exchanged within the TCP data stream.
- ▶ Socket application data is exchanged through RDMA (write operations).
- ▶ Retains TCP connection to control the SMC-R connection.
- ▶ Preserves many critical operational and network management features of TCP/IP.

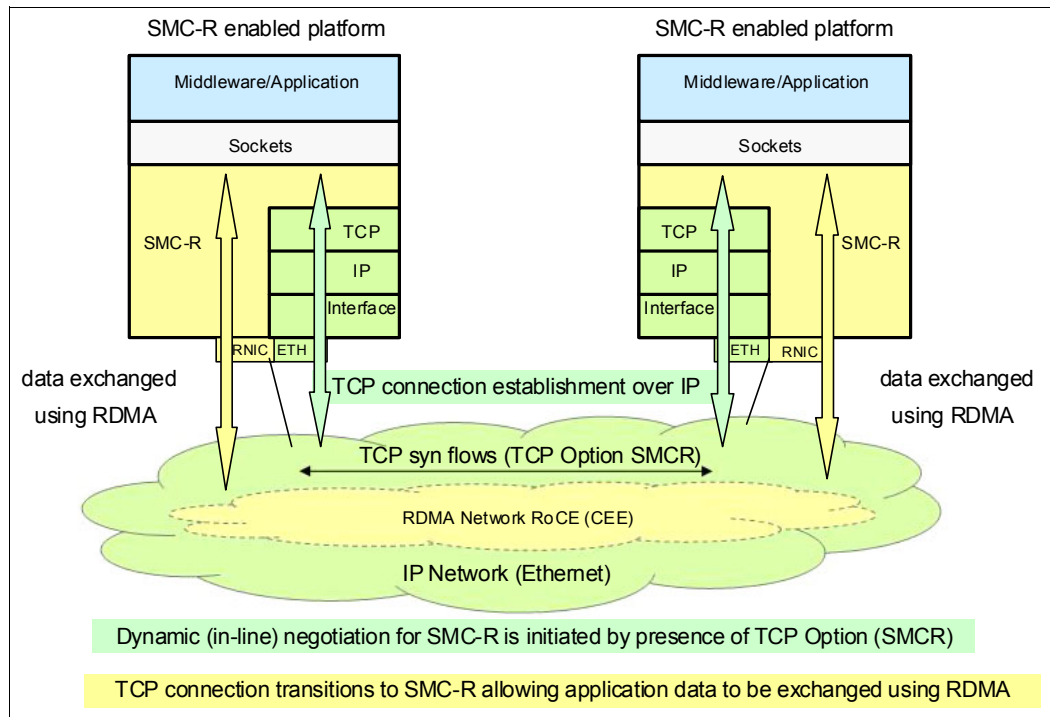


Figure D-2 Dynamic transition from TCP to SMC-R

The hybrid model of SMC-R uses the following key attributes:

- ▶ Follows the standard TCP/IP connection setup.
- ▶ Switches to RDMA (SMC-R) dynamically.
- ▶ TCP connection remains active (idle) and is used to control the SMC-R connection.
- ▶ Preserves the following critical operational and network management TCP/IP features:
 - Minimal (or zero) IP topology changes
 - Compatibility with TCP connection-level load balancers
 - Preservation of the IP security model, such as IP filters, policies, virtual LANs (VLANs), and Secure Sockets Layer (SSL)
 - Minimal network administration and management changes
- ▶ Host application software is not required to change; therefore, all host application workloads can benefit immediately.

D.2.3 Single Root I/O virtualization

Single Root I/O virtualization (SR-IOV) is a technology that provides the capability to share the adapter between LPARs. SR-IOV is also designed to provide isolation of virtual functions within the PCIe RoCE Express/Express2 adapter. For example, one LPAR cannot cause errors that are visible to other virtual functions or other LPARs. Each operating system that is running in an LPAR includes its own application queue in its own memory space.

The concept of the Shared RoCE Mode is shown in Figure D-3.

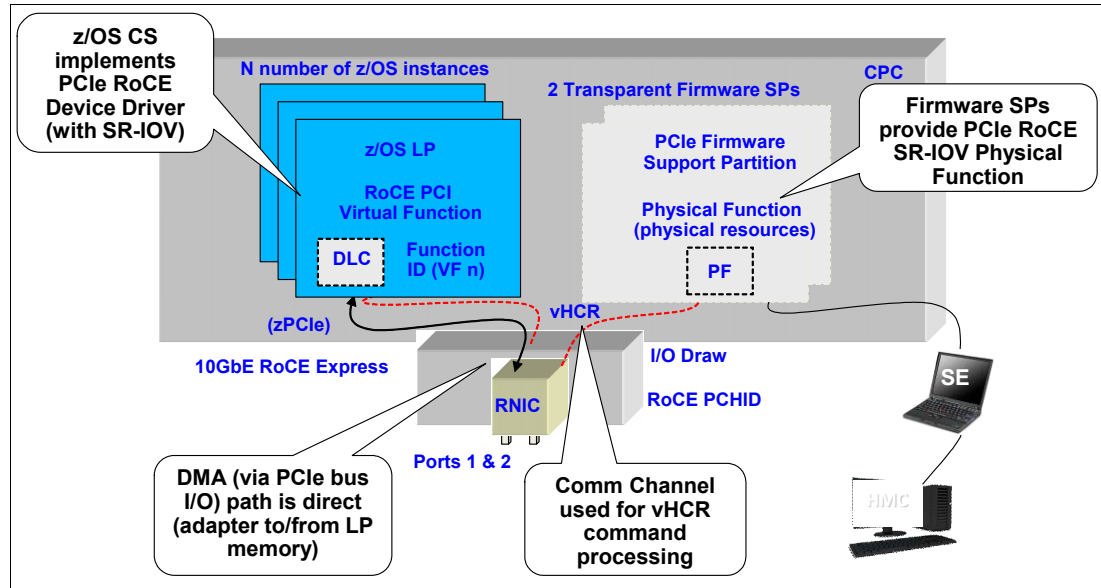


Figure D-3 Shared RoCE mode concepts

The Physical Function Driver communicates with the physical function in the PCIe adapter and is responsible for the following functions:

- Manage resource allocation
- Perform hardware error handling
- Perform code updates
- Run diagnostics

The device-specific IBM Z Licensed Internal Code (LIC) connects Physical Function Driver to Support Elements (SEs) and limited system level firmware required services.

D.2.4 Hardware

The 10GbE RoCE Express adapter (FC 0411), 25GbE RoCE Express2 (FC 0430), and 10GbE RoCE Express2 (FC 0412) are RDMA-capable NICs. The integrated firmware processor (IFP) runs four resource groups (RGs) that contain firmware for the RoCE Express adapter. For more information, see C.1.3, “Resource groups” on page 421.

D.2.5 RoCE Express/Express2 adapter

The RoCE Express adapters are designed to help reduce the consumption of CPU resources for applications that use the TCP/IP stack, such as WebSphere accessing a Db2 database. The use of RoCE Express also helps to reduce network latency with memory-to-memory transfers that use SMC-R in z/OS V2.1 or later. It is not apparent to applications and can be used for LPAR-to-LPAR communications on a single z/OS system or server-to-server communications in a multiple CPC environment.

The 10GbE RoCE Express2 adapter that is shown in Figure D-4 is installed in the PCIe+ I/O drawer.



Figure D-4 10GbE RoCE Express2

Each PCIe adapter has two ports. A maximum of four adapters can be installed in a z14 ZR1 server. The adapter use a short reach (SR) laser as the optical transceiver and support the use of a multimode fiber optic cable that ends with an LC Duplex connector. Point-to-point connection (with another RoCE Express/Express2 adapter of the SAME speed) and switched connection with an enterprise-class switch are supported.

RoCE Physical Connectivity: The 25GbE RoCE Express2 feature does *not* support negotiation (to a lower speed). Therefore, it must be connected to a 25 Gbps port of an Ethernet Switch or to another 25GbE RoCE Express2 feature.

The 10GbE RoCE Express and 10GbE RoCE Express2 features can be connected to each other in a point-to-point connection or to a 10 Gbps port of an Ethernet switch.

SMC-R can use direct RoCE Express to RoCE Express connectivity (without any switch). However, this type of direct physical connectivity forms a single physical point-to-point connection, which disallows any other connectivity with other LPARs, such as other SMC-R peers. Although this option is viable for test scenarios, it is not practical (nor recommended) for production deployment.

If the IBM RoCE Express/Express2 adapters are connected to Ethernet switches, the switches must support the following requirements:

- ▶ 10 Gbps or 25 Gbps ports (depending on the RoCE feature specifications)
- ▶ Global Pause function frame (as described in the IEEE 802.3x standard) must be enabled
- ▶ Priority flow control (PFC) disabled
- ▶ No firewalls, outing, or intraensemble data network (IEDN)

The maximum supported unrepeatd distance, point-to-point at initial introduction was 300 meters for 10 Gbps features. The 25GbE RoCE Express2 feature supports a maximum unrepeatd distance of 100 meters (328 feet). These distances can be extended across multiple cascaded switches or qualified DWDMs to even 100 km (62 miles). For more information, see the [SMC-R over distance presentation](#).

12.7.2 RoCE Express/Express2 configuration example

Mixing of RoCE Generations: Mixing generations of RoCE adapters on the same stack is supported with the following considerations:

- ▶ 25GbE RoCE Express2 should *not* be mixed with 10GbE RoCE Express2 or 10GbE RoCE Express in the same SMC-R Link Group.
- ▶ 10GbE RoCE Express2 can be mixed with 10GbE RoCE Express (that is, provisioned to the same TCP/IP stack or same SMC-R Link Group).

A sample configuration that allows redundant SMC-R connectivity among LPAR A and C, and LPAR 1, 2 and 3 is shown in Figure D-5. Each adapter can be shared or dedicated to an LPAR. As shown in Figure D-5, two adapters per LPAR are advised for redundancy.

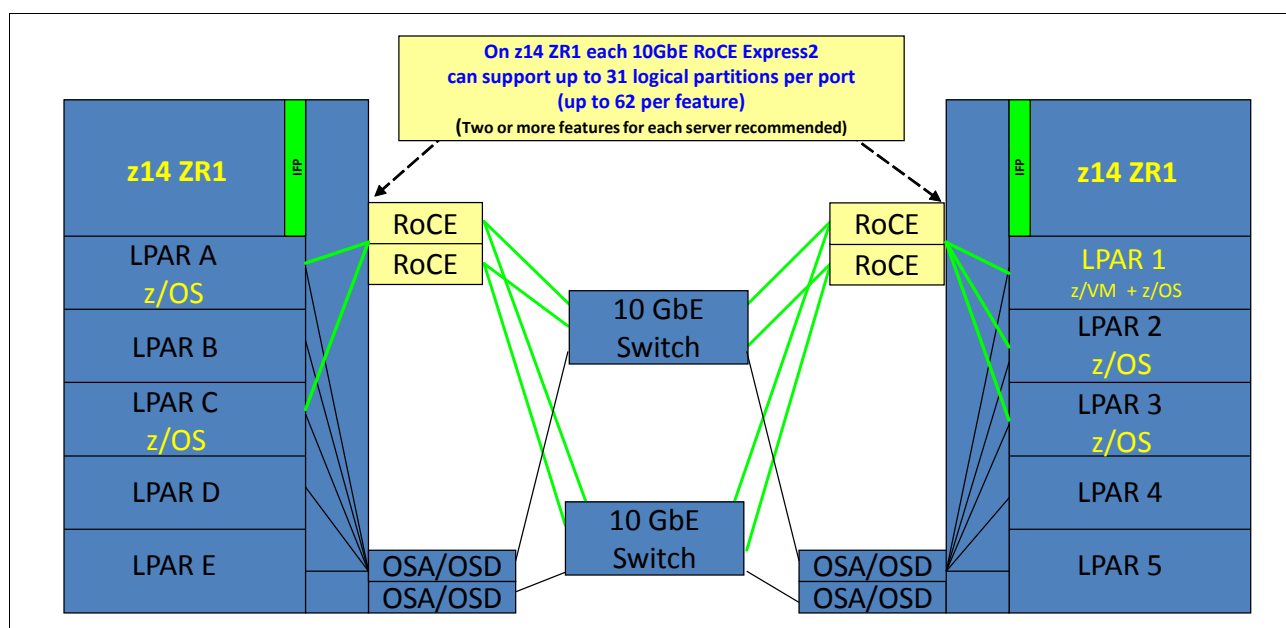


Figure D-5 10GbE RoCE Express sample configuration

The configuration that is shown in Figure D-5 allows redundant SMC-R connectivity among LPAR A, LPAR C, LPAR 1, LPAR 2, and LPAR 3. LPAR to LPAR OSD connections are required to establish the SMC-R communications. The 1 GbE OSD connections can be used. OSD connections can flow through the same switches or different switches.

Note: The OSA-Express Adapter and the RoCE Express adapter must be associated to each other by having equal PNET IDs (defined in the hardware configuration definition [HCD]).

An OSA-Express adapter, which is defined as channel-path identifier (CHPID) type OSD, is required to establish SMC-R. The interaction of OSD and the RNIC is shown in Figure D-6.

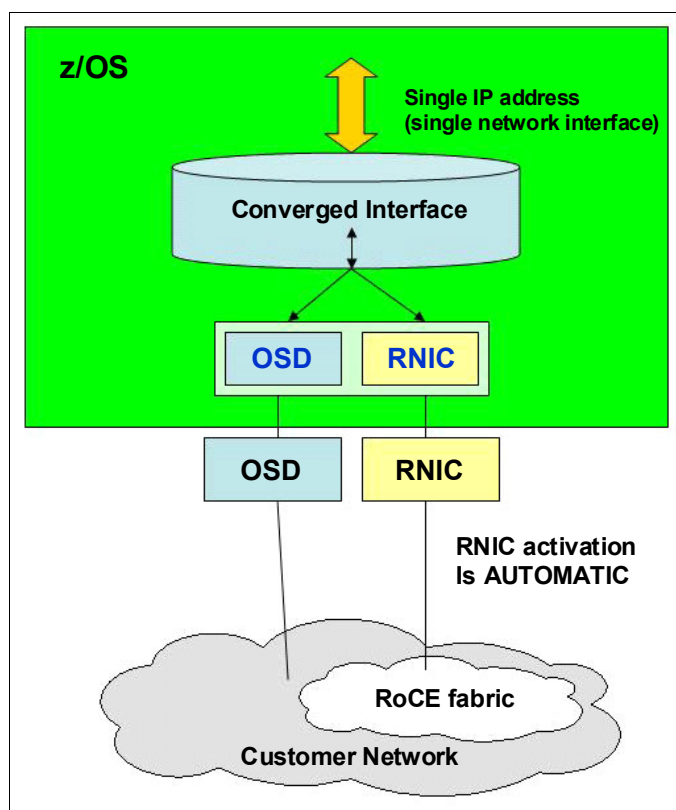


Figure D-6 RNIC and OSD interaction

The OSA adapter might be a single or pair of 10 GbE, 1 GbE, or 1000Base-T OSAs. The OSA must be connected to another OSA on the system with which the RoCE adapter is communicating. As shown in Figure D-5 on page 431, 1 GbE OSD connections can still be used instead of 10 GbE and OSD connections can flow through the same 10 GbE switches.

Consider the following points regarding Figure D-6:

- ▶ The z/OS system administrator must configure and manage the OSD interface only.
- ▶ The Communications Server transparently splits and converges network traffic to and from the converged interface.
- ▶ Only OSD connectivity must be configured.

With SMC-R, the RNIC interface is dynamically and transparently added and configured.

D.2.6 Hardware configuration definitions

The following HCDs are important:

- Function ID

The RoCE adapter is identified by a hexadecimal Function Identifier (FID). It features a dedicated limit in the range 00 - FF, and the shared limit is 000 - 0FFF in the HCD or Hardware Management Console (HMC) to create the I/O configuration program (IOCP) input.

An FID can be configured to only one LPAR, but it is reconfigurable. The RoCE adapter, as installed in a specific PCIe+ I/O drawer and slot, is to be used for the defined function. The physical installation (drawer and slot) determines the physical channel identifier (PCHID). Only one FID can be defined for dedicated mode.

- Virtual Function ID

Virtual Function ID is defined when PCIe hardware is shared between LPARs. Virtual Function ID has a decimal Virtual Function Identifier (VF=) in the range of 1 - *nn*, where *nn* is the maximum number of partitions that the PCIe adapter supports. For example, at z14 ZR1 the RoCE Express2 adapter supports up to 62 partitions, and a zEDC Express adapter supports up to 15.

- 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 OSA and RNIC interfaces in the HCD.

A PNet ID is defined for each physical port. z/OS Communications Server receives the information during the activation of the interfaces and associates the OSD interfaces with the RNIC interfaces that include matching PNet ID values.

Attention: Activation fails if you do not configure a PNet ID for the RNIC adapter. Activation succeeds if you do not configure a PNet ID for the OSA adapter; however, the interface is not eligible to use SMC-R.

D.2.7 Software use of SMC-R

SMC-R can be implemented on the RoCE and can communicate memory-to-memory, which avoids the CPU resources of TCP/IP by reducing network latency and improving wall clock time. It focuses on “time to value” and widespread performance benefits for all TCP socket-based middleware.

The following advantages are gained, as shown in Figure D-7 on page 434:

- No middleware or application changes (transparent)
- Ease of deployment (no IP topology changes)
- LPAR-to-LPAR communications on a single central processing complex (CPC)
- Server-to-server communications in a multi-CPC environment
- Retained key qualities of service that TCP/IP offers for enterprise class server deployments (high availability, load balancing, and an IP security-based framework)

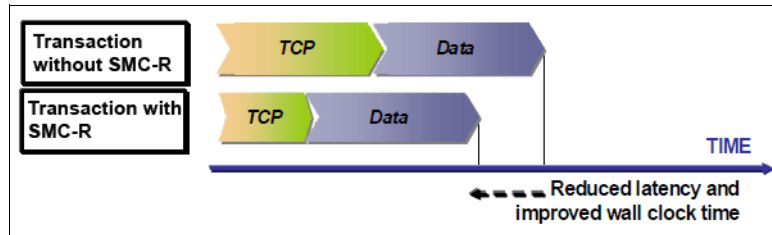


Figure D-7 Reduced latency and improved wall clock time with SMC-R

D.2.8 SMC-R support overview

SMC-R needs hardware and software support, as described in this section.

Hardware requirements

SMC-R requires the following hardware:

- ▶ PCIe-based RoCE Express2:
 - z14 systems
 - Dual port 25GbE or 10GbE adapter
- ▶ PCIe-based RoCE Express:
 - z14, z13, z13s, zEC12, and zBC12
 - Dual port 10 GbE adapter:
 - z14 ZR1 maximum 4 RoCE Express2/Express adapters per CPC
 - z14 M0x maximum 8 RoCE Express2/Express adapters per CPC
- ▶ HCD and input/output configuration data set (IOCDS): PCIe FID, VF (sharing), and RoCE configuration with PNet ID.
- ▶ Optional: Standard switch (CEE-enabled switch is not required).
- ▶ Required queued direct input/output (QDIO) Mode OSA connectivity between z/OS LPARs, as shown in Figure D-5 on page 431.
- ▶ Adapter **MUST** be dedicated to an LPAR on a zEC12 or zBC12. It *must* be shared (or at least in shared mode) to one or more LPARs on a z14, z13, or z13s systems.
- ▶ SMC-R cannot be used in IEDN.

Software requirements

SMC-R requires the following software:

- ▶ z/OS V2R1 (with PTFs) or higher are the only supported operating systems for the SMC-R protocol. You cannot roll back to previous z/OS releases.
- ▶ z/OS guests under z/VM 6.4 or later are supported to use RoCE adapters.
- ▶ IBM is working with its Linux distribution partners to include support in future Linux on Z distribution releases.

Other RoCE considerations

RoCE includes the following considerations:

- ▶ RoCE system limits:
 - 62 unique VLANs per PCHID physical port for z14 ZR1
 - Each VF ensures a minimum of 2 VLANs max of 16

- ▶ z/OS CS consumption of RoCE virtual resources:
 - One VF used per TCP stack (per PFID / port)
 - One virtual Media Access Control (VMAC) per VF (z/OS uses PF generated VMAC)
 - One VLAN ID (up to 16) per OSA VLAN (“inherited” as TCP connections occur)
- ▶ z/OS Communications Server Migration considerations:
 - RoCE HCD (IOCDS) configuration changes are required
 - z/OS RoCE users might be required to make a TCP/IP configuration change; that is, TCP/IP profiles (PFIDs) might be compatible with shared RoCE
- ▶ Changes are required for RoCE users for the following cases:
 - z/OS users who use multiple TCP/IP stacks and both stacks currently use the same RoCE adapter (single z/OS image sharing a physical adapter among multiple stacks).
 - z/OS users who need to use physical RoCE ports from the same z/OS instance (not “best practices”, but is allowed).
 - z/OS users who do not continue the use of (coordinate) the same PFID values (continue the use of the PFID value that is used in the dedicated environment for a specific z/OS instance) when multiple PFIDs and VFs are added to the same adapter (for more shared users).

D.2.9 SMC-R use cases for z/OS to z/OS

SMC-R with RoCE provides high-speed communications and “HiperSockets-like” performance across physical processors. It can help all TCP-based communications across z/OS LPARs that are in different CPCs.

The following typical communications patterns are used:

- ▶ 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 sysplex, SMC-R offers a significant performance advantage. Traffic between client and server can flow directly between the two servers without traversing 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

A traditional Sysplex Distributor is shown in Figure D-8.

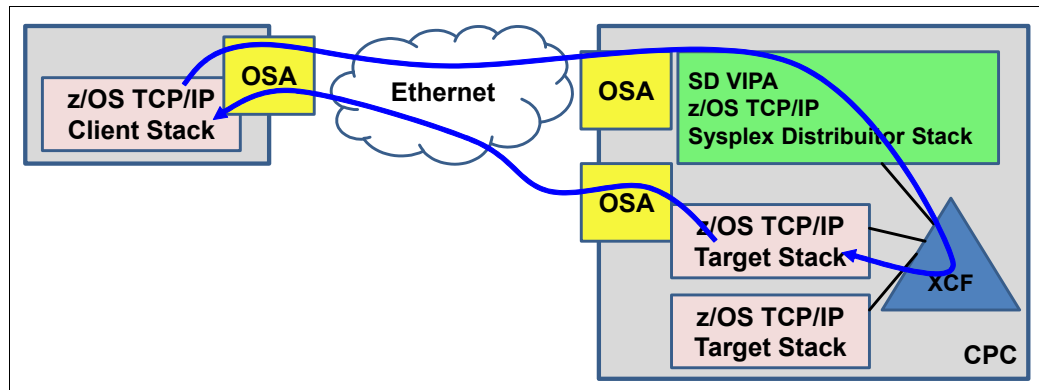


Figure D-8 Sysplex Distributor before RoCE

The traditional Sysplex Distributor features the following characteristics:

- ▶ All traffic from the client to the target application goes through the Sysplex Distributor TCP/IP stack.
- ▶ All traffic from the target application goes directly back to the client by using the TCP/IP routing table on the target TCP/IP stack.

Sysplex Distributor after RoCE

A RoCE Sysplex Distributor is shown in Figure D-9. Consider the following points:

- ▶ The initial connection request goes through the Sysplex Distributor stack.
- ▶ The session then flows directly between the client and the target over the RoCE adapters.

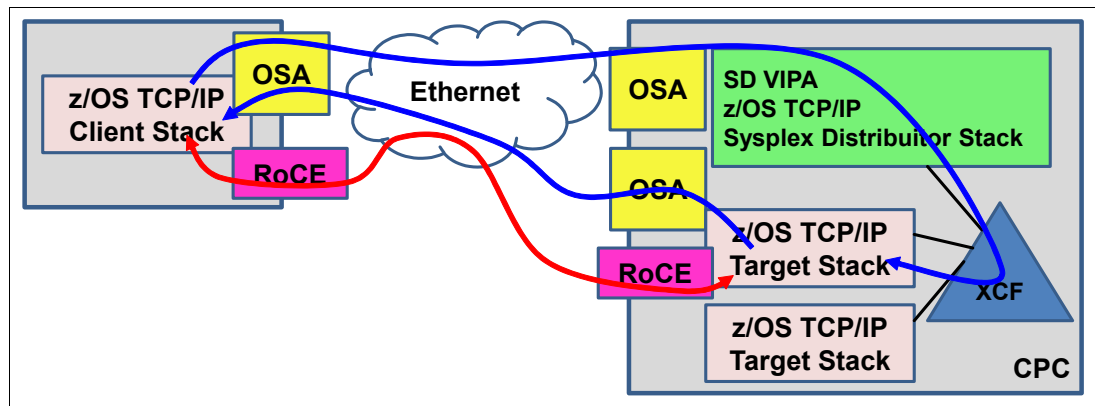


Figure D-9 Sysplex Distributor after RoCE

Note: As with all RoCE Communications, the session end also flows over OSAs.

D.2.10 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 IP configuration for LAN access that uses OSD:

- ☐ HCD definitions (install and configure RNICs in the HCD):
 - ☐ Add the PNet ID for the current OSD.
 - ☐ Define PFIDs for RoCE (with the same PNet ID).
- ☐ Specify the **GLOBALCONFIG SMCR** parameter (TCP/IP Profile):
 - ☐ Must specify at least one PCIe Function ID (PFID):
 - ☐ A PFID represents a specific RNIC adapter.
 - ☐ A maximum of 16 PFID values can be coded.
 - ☐ Up to eight TCP/IP stacks can share a RoCE PCHID (RoCE adapter) in a specific LPAR (each stack must define a unique FID value).
- ☐ Start the IPAQENET or IPAQENET6 INTERFACE with CHPIDTYPE OSD:
 - ☐ SMC-R is enabled, by default, for these interface types.
 - ☐ SMC-R is not supported on any other interface types.

Note: The IPv4 INTERFACE statement (IPAQENET) must also specify an IP subnet mask

- ☐ Repeat in each host (at least two hosts).

Start the TCP/IP traffic and monitor it with Netstat and IBM VTAM displays.

Note: For RoCE Express2, the PCI Function IDs (PFIDs) are now associated with a specific (single) physical port (that is, port 0 or port 1). The port number is now configured with the FID number in HCD (or IOCDS) and the port number must be configured (no default is available). z/OS CommServer does not learn the RoCE generation until activation. During activation, CommServer learns the port number for RoCE Express2.

Consider the following points:

- ▶ The port number for RoCE Express is configured in z/OS TCP/IP profile and does not change.
- ▶ When defining a FID in the TCP/IP profile for RoCE Express2, the port number is no longer applicable (it is ignored for RoCE Express2).
- ▶ A warning message is issued if the TCP/IP profile does not match the HCD configured value (that is, the value is ignored and it is incorrect).

D.3 Shared Memory Communications - Direct Memory Access

This section describes the new SMC-D functions that are implemented in IBM z14, and z13 and z13s (Driver Level 27) systems.

D.3.1 Concepts

The collocation of multiple tiers of a workload onto a single IBM Z physical system allows for the use of HiperSockets, which is an internal LAN technology that provides low-latency communication between virtual machines within a physical IBM Z CPC.

HiperSockets is implemented fully within IBM Z firmware; therefore, it requires no physical cabling or external network connection to purchase, maintain, or replace. The lack of external components also provides for a secure and low latency network connection because data transfer occurs, much like a cross-address-space memory move.

SMC-D maintains the socket-API transparency aspect of SMC-R so that applications that use TCP/IP communications can benefit immediately without requiring any application software or IP topology changes. SMC-D completes the overall Shared Memory Communications solution, which provides synergy with SMC-R. Both protocols use shared memory architectural concepts, which eliminate TCP/IP processing in the data path, yet preserves TCP/IP Qualities of Service for connection management purposes.

From an operations standpoint, SMC-D is similar to SMC-R. The objective is to provide consistent operations and management tasks for SMC-D and SMC-R. SMC-D uses a new virtual PCI adapter that is called Internal Shared Memory (ISM). The ISM Interfaces are associated with IP interfaces; for example, HiperSockets or OSA (ISM interfaces do not exist without an IP interface).

ISM interfaces are not defined in software. Instead, ISM interfaces are dynamically defined and created, and automatically started and stopped. You do not need to operate (Start or Stop) ISM interfaces. Unlike RoCE, ISM FIDs (PFIDs) are not defined in software. Instead, they are auto-discovered based on their PNet ID.

SMC-R uses RDMA (RoCE), which is based on Queue Pair (QP) technology. Consider the following points:

- ▶ RC-QPs represent SMC Links (logical point-to-point connection).
- ▶ RC-QPs over unique RNICs are logically bound together to form Link Groups (used for HA and load balancing).
- ▶ Link Groups (LGs) and Links are provided in many Netstat displays (for operational and various network management tasks).

SMC-D over ISM does not use QPs. Consider the following points:

- ▶ Links and LGs based on QPs (or other hardware constructs) are not applicable to ISM. Therefore, the SMC-D information in the Netstat command displays is related to ISM link information rather than LGs.
- ▶ SMC-D protocol (such as SMC-R) feature a design concept of a “logical point-to-point connection” and preserves the concept of an SMC-D Link (for various reasons that include network administrative purposes).

Note: The SMC-D information in the Netstat command displays is related to ISM link information (not LGs).

D.3.2 Internal Shared Memory technology overview

ISM is a function that is supported by the z14, z13, and z13s systems. It is the firmware that provides the connectivity for shared memory access between multiple operating systems within the same CPC. It provides the same functionality as SMC-R, but without physical adapters (such as the RoCE adapter) by using instead virtual ISM devices as SMC-R. It is a HiperSocket-like function that provides guest-to-guest communications within the same machine. A possible solution that uses only SMC-D is shown in Figure D-10.

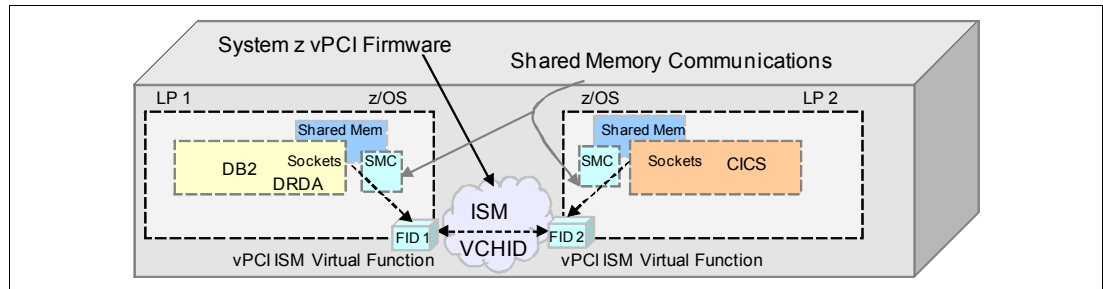


Figure D-10 Connecting two LPARs on the same CPC by using SMC-D

SMC-D and SMC-R technologies can be used at the same time on the same CPCs. A fully configured three-tier solution that uses SMC-D and SMC-R is shown in Figure D-11.

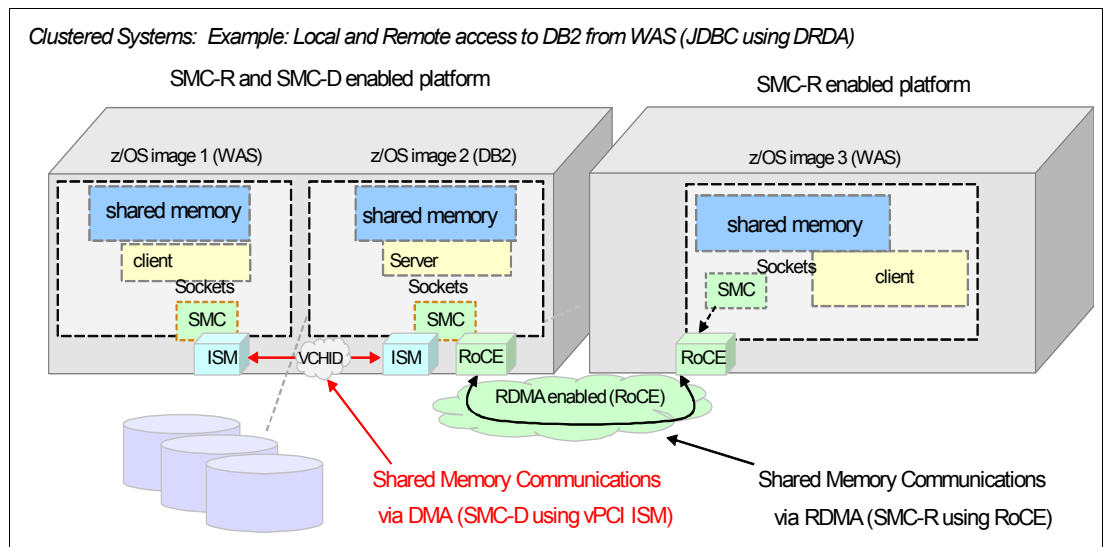


Figure D-11 Clustered systems: Multitier application solution. RDMA, and DMA

D.3.3 SMC-D over Internal Shared Memory

ISM is a virtual channel that is similar to IQD for HiperSockets. A virtual adapter is created in each OS. The memory is logically shared by using the SMC protocol. The network is firmware-provided and a new device is required to manage that virtual function. SMC is based on a TCP/IP connection and preserves the entire network infrastructure.

SMC-D is a protocol that allows TCP socket applications to transparently use ISM. It is a “hybrid” solution, as shown in Figure D-12 on page 440.

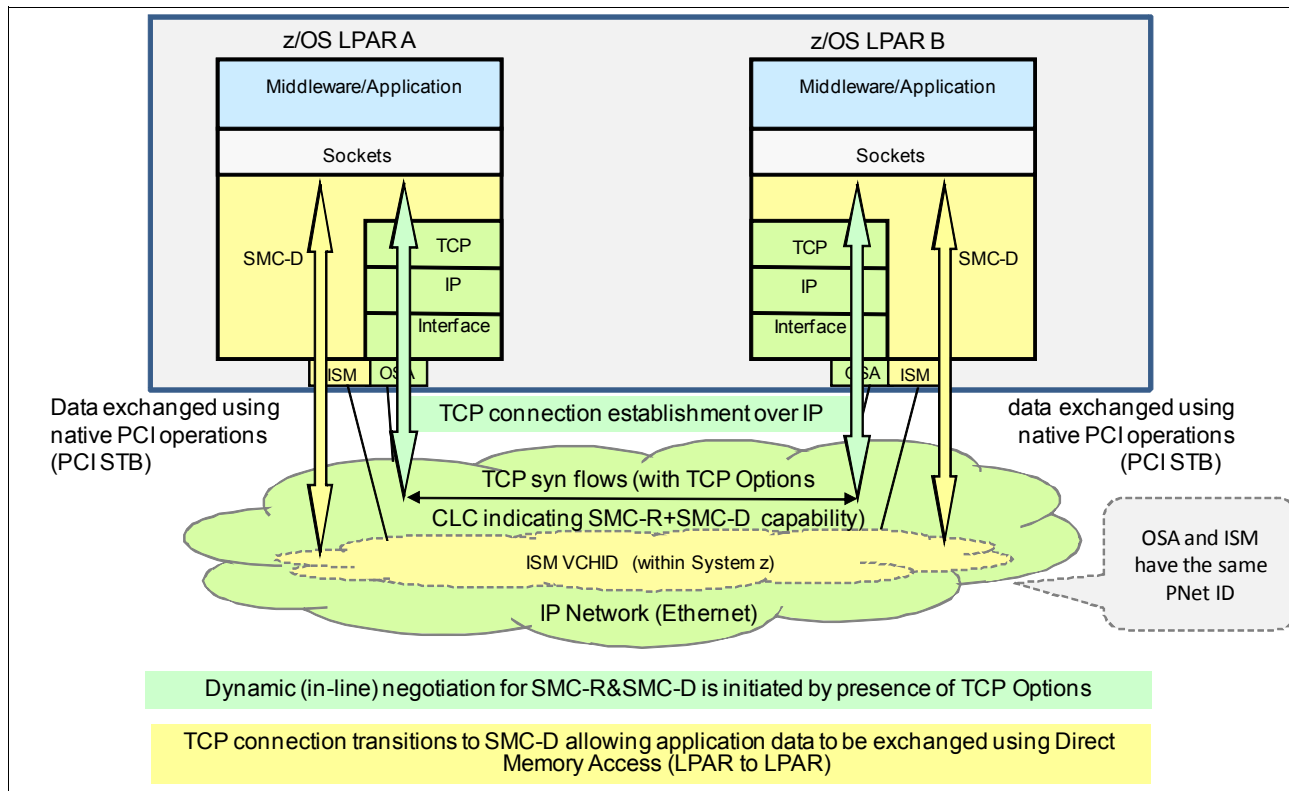


Figure D-12 Dynamic transition from TCP to SMC-D by using two OSA-Express adapters

Consider the following points:

- ▶ It uses a TCP connection to establish the SMC-D connection.
- ▶ The TCP connection can be through the OSA adapter or IQD HiperSockets.
- ▶ A TCP option (SMCD) controls switching from TCP to “out-of-band” SMC-D.
- ▶ The SMC-D information is exchanged within the TCP data stream.
- ▶ Socket application data is exchanged through ISM (write operations).
- ▶ The TCP connection remains to control the SMC-D connection.

This model preserves many critical operational and network management features of TCP/IP.

The hybrid model of SMC-D uses the following key attributes:

- ▶ It follows the standard TCP/IP connection setup.
- ▶ The hybrid model switches to ISM (SMC-D) dynamically.
- ▶ The TCP connection remains active (idle) and is used to control the SMC-D connection.
- ▶ The hybrid model preserves the following critical operational and network management TCP/IP features:
 - Minimal (or zero) IP topology changes
 - Compatibility with TCP connection-level load balancers
 - Preservation of the IP security model, such as IP filters, policies, VLANs, and SSL
 - Minimal network administration and management changes
- ▶ Host application software is not required to change; therefore, all host application workloads can benefit immediately.

D.3.4 Internal Shared Memory introduction

The IBM z14, IBM z13 (Driver 27), and z13s systems support the ISM virtual PCI function. ISM is a virtual PCI network adapter that enables direct access to shared virtual memory that provides a highly optimized network interconnect for IBM Z intra-CPC communications.

ISM introduces a new static virtual channel identifier (VCHID) Type. The VCHID is referenced in IOCDS/HCD. The ISM VCHID concepts are similar to the IQD (HiperSockets) type of virtual adapters. ISM is based on IBM Z PCIe architecture (that is, virtual PCI function or adapter). It introduces a new PCI Function Group and type (ISM virtual PCI). A new virtual adapter is scheduled for release.

The system administrator, configuration, and operations tasks follow the same process (HCD/IOCDS) as PCI functions, such as RoCE Express and zEDC Express. ISM supports dynamic I/O.

ISM Provides adapter virtualization (Virtual Functions) with high scalability. Consider the following points:

- ▶ It supports up to 32 ISM VCHIDs per CPC (z14, z13, or z13s servers, each VCHID represents a unique internal shared memory network each with a unique Physical Network ID).
- ▶ Each VCHID supports up to 255 VFs per VCHID (the maximum is 8 k VFs per CPC), which provide significant scalability.

Note: No concept of a PCI Physical Function is available to provide virtualization. No concept of MACs, MTU, or Frame size is available.

- ▶ Each ISM VCHID represents a unique and isolated internal network, each having a unique Physical Network ID (PNet IDs are configured in HCD/IOCDS).
- ▶ ISM VCHIDs support VLANs; therefore, subdividing a VCHID by using virtual LANs is supported.
- ▶ ISM provides a Global Identifier (GID) that is internally generated to correspond with each ISM FID.
- ▶ ISM is supported by z/VM in pass-through mode (PTF required).

D.3.5 Virtual PCI Function (vPCI Adapter)

Virtual Function ID is defined when PCIe hardware is shared between LPARs. Virtual Function ID includes a decimal Virtual Function Identifier (VF=) in the range 1 - *nn*, where *nn* is the maximum number of partitions that the PCIe adapter supports. For example, the SMC-D ISM supports up to 32 partitions, and a zEDC Express adapter supports up to 15.

The following basic infrastructure is available:

- ▶ zPCI architecture
- ▶ RoCE, zEDC, ISM
- ▶ zPCI layer in z/OS and Linux for z systems
- ▶ vPCI for SD queues

The basic concept vPCI adapter is shown in Figure D-13.

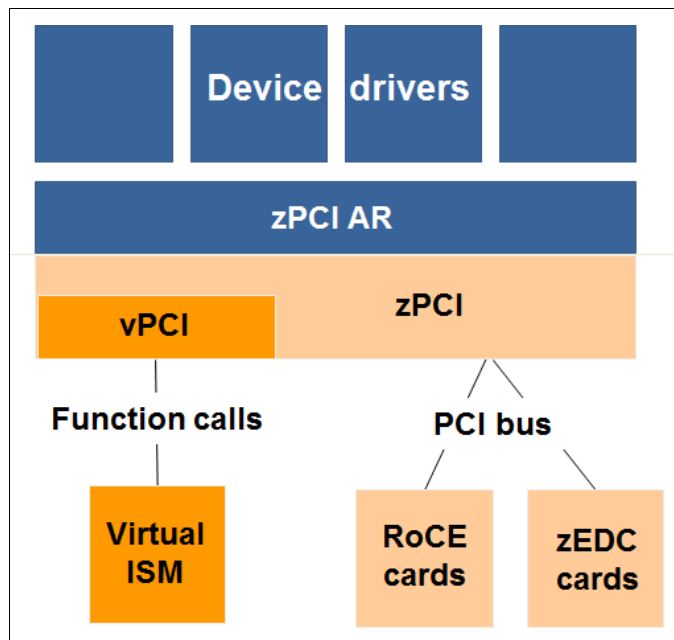


Figure D-13 Concept of vPCI adapter implementation

Note: The following basic z/VM support is available:

- ▶ Generic zPCI pass-through support starting from z/VM 6.3
- ▶ The use of the zPCI architecture remains unchanged

An SMC-D configuration in which Ethernet provides the connectivity is shown in Figure D-14.

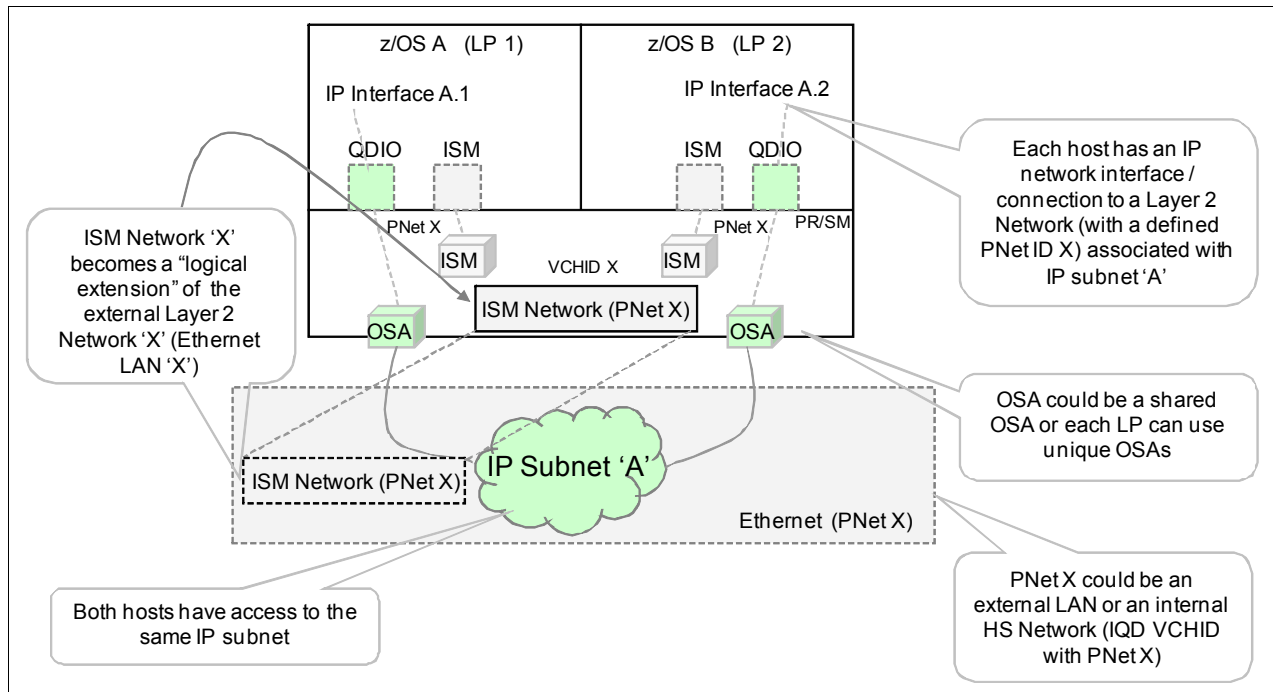


Figure D-14 SMC-D configuration that uses Ethernet to provide connectivity

An SMC-D configuration in which HyperSockets provide the connectivity is shown in Figure D-15.

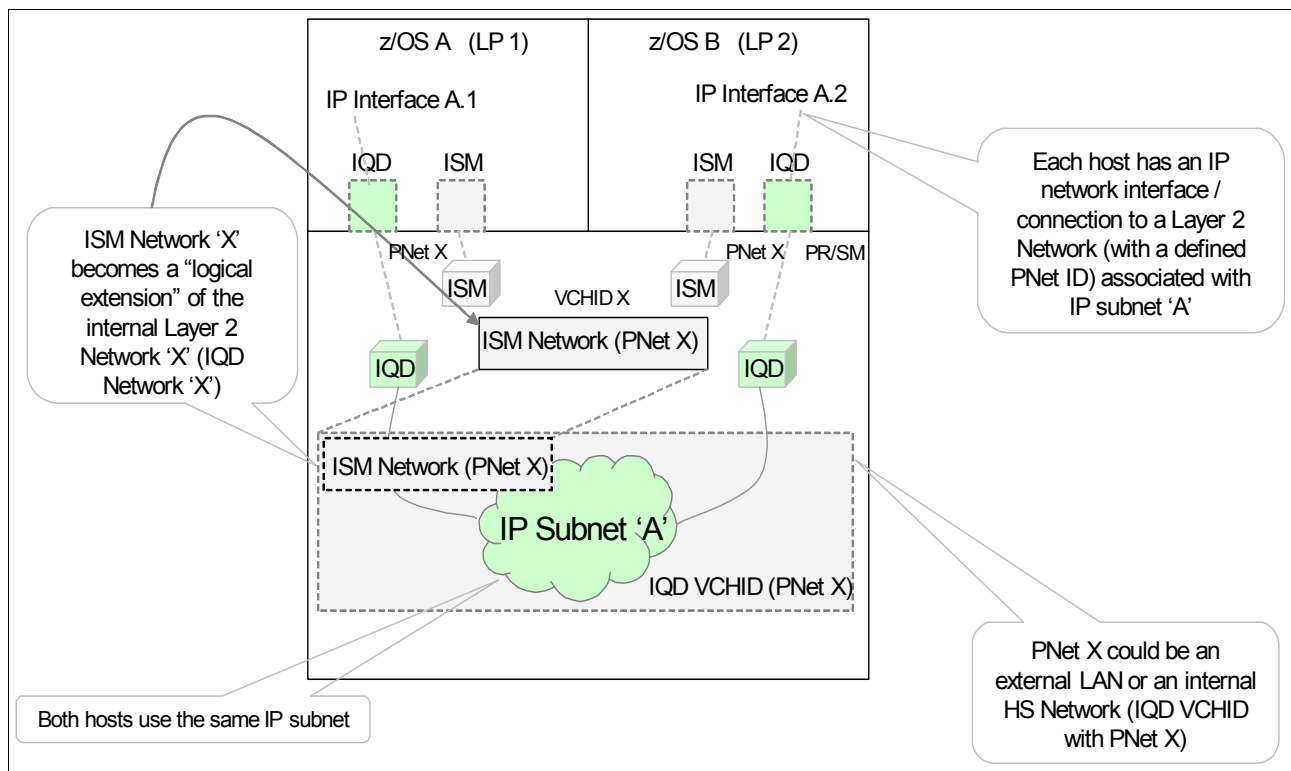


Figure D-15 SMC-D configuration that uses HyperSockets to provide connectivity

D.3.6 Planning considerations

In the z/OS SMC-D implementation, z/OS uses a single VF per ISM PNet. This configuration is true for a single VLAN or for multiple VLANs per PNet. The number of VLANs that are defined for a specific PNet does not affect the number of VFs required.

z/OS Communications Server requires one ISM FID per ISM PNet ID per TCP/IP stack. This requirement is not affected by the version of the IP (that is, it is true even if both IPv4 and IPv6 are used).

z/OS might use more ISM FIDs for the following reasons:

- ▶ IBM supports up to eight TCP/IP stacks per z/OS LPAR. SMC-D can use up to eight FIDs or VFs (one per TCP/IP stack).
- ▶ IBM supports up to 32 ISM PNet IDs per CPC. Each TCP/IP stack can have access to PNet ID that uses up to 32 FIDs (one VF per PNet ID).

D.3.7 Hardware configuration definitions

Complete the following steps to use HCDs:

1. Configure ISM vPCI Functions (HCD/HCM).
2. Define PNet IDs (OSA, HiperSockets [IQD], and ISM) in HCD/HCM.
3. Activate the definition by using HCD.
4. Enable SMC-D in at least two z/OS instances, which are a single parameter in TCP/IP Global configuration. Both z/OS instances must run on the same CPC.
5. Review and adjust as needed the available real memory and fixed memory usage limits (z/OS and CS). SMC requires fixed memory. You might need to review the limits and provision more real memory for z/OS.
6. Review the IP topology, VLAN usage considerations, and IP security. For more information, see the [IBM z/OS Shared Memory Communications: Security Considerations](#) white paper.
7. Run Shared Memory Communications Applicability Tool (SMC-AT) to evaluate applicability and potential value.
8. Review changes to messages, monitoring information, and diagnostic tools. Similar to SMC-R, many updates are made to the following items:
 - Messages (VTAM and TCP stack)
 - Netstat (status, monitoring, and display information)
 - CS diagnostic tools (VIT, Packet trace, CTRACE, and IPCS formatted dumps)

Note: No application changes (transparent to Socket applications) are made. Also, no optional operation changes are required (for example starting or stopping devices).

ISM Functions must be associated with another channel (CHID) of one of the following types:

- ▶ IQD (a single IQD HiperSockets) channel
- ▶ OSD channels

Note: A single ISM PCHID cannot be associated with both IQD and OSD.

D.3.8 Sample IOCP FUNCTION statements

The IOCP FUNCTION statements (see Example D-1) describe the configuration that defines ISM adapters that are shared between LPARs on the same CPC, as shown in Figure D-16.

Example D-1 IOCP FUNCTION statements

```
FUNCTION FID=1017,VCHID=7E1,VF=1,PART=((LP1),(LP1,LP2)),PNETID=(PNET1),TYPE=ISM
FUNCTION FID=1018,VCHID=7E1,VF=2,PART=((LP2),(LP1,LP2)),PNETID=(PNET1),TYPE=ISM
```

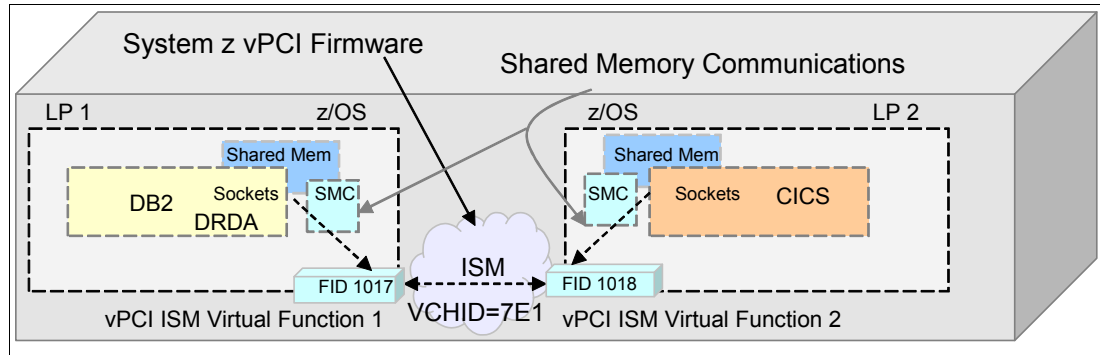


Figure D-16 ISM adapters that are shared between LPARs

Note: On the IOCDS statement, the VCHID is defined as 7E1. As shown in Figure D-16, the ISM network “PNET 1” is referenced by the IOCDS VCHID statement. ISM (as with IQD) does not use physical adapters or adapter slots (PCHID). Instead, only logical (firmware) instances that are defined as VCHIDs in IOCDS are used.

A sample IOCP FUNCTION configuration (see Example D-2) that defines ISM adapters that are shared between LPSRs and multiple VLANs on the same CPC as shown in Figure D-17 on page 446.

Example D-2 Sample IOCP Function

```
FUNCTION FID=1017,VCHID=7E1,VF=1,PART=((LPAR3),(LPAR3,LPAR4)),PNETID=(PNETA),TYPE=ISM
FUNCTION FID=1018,VCHID=7E1,VF=2,PART=((LPAR4),(LPAR3,LPAR4)),PNETID=(PNETA),TYPE=ISM
FUNCTION FID=1019,VCHID=7E1,VF=3,PART=((LPAR5),(LPAR4,LPAR5)),PNETID=(PNETA),TYPE=ISM
FUNCTION FID=1020,VCHID=7E1,VF=4,PART=((LPAR6),(LPAR5,LPAR6)),PNETID=(PNETA),TYPE=ISM
FUNCTION FID=1021,VCHID=7E1,VF=5,PART=((LPARn),(LPAR6,LPARn)),PNETID=(PNETA),TYPE=ISM
FUNCTION FID=1022,VCHID=7E2,VF=1,PART=((LPAR1),(LPAR1,LPAR2)),PNETID=(PNETB),TYPE=ISM
FUNCTION FID=1023,VCHID=7E2,VF=2,PART=((LPAR2),(LPAR1,LPAR2)),PNETID=(PNETB),TYPE=ISM
FUNCTION FID=1024,VCHID=7E2,VF=3,PART=((LPARn),(LPAR1,LPARn)),PNETID=(PNETB),TYPE=ISM
```

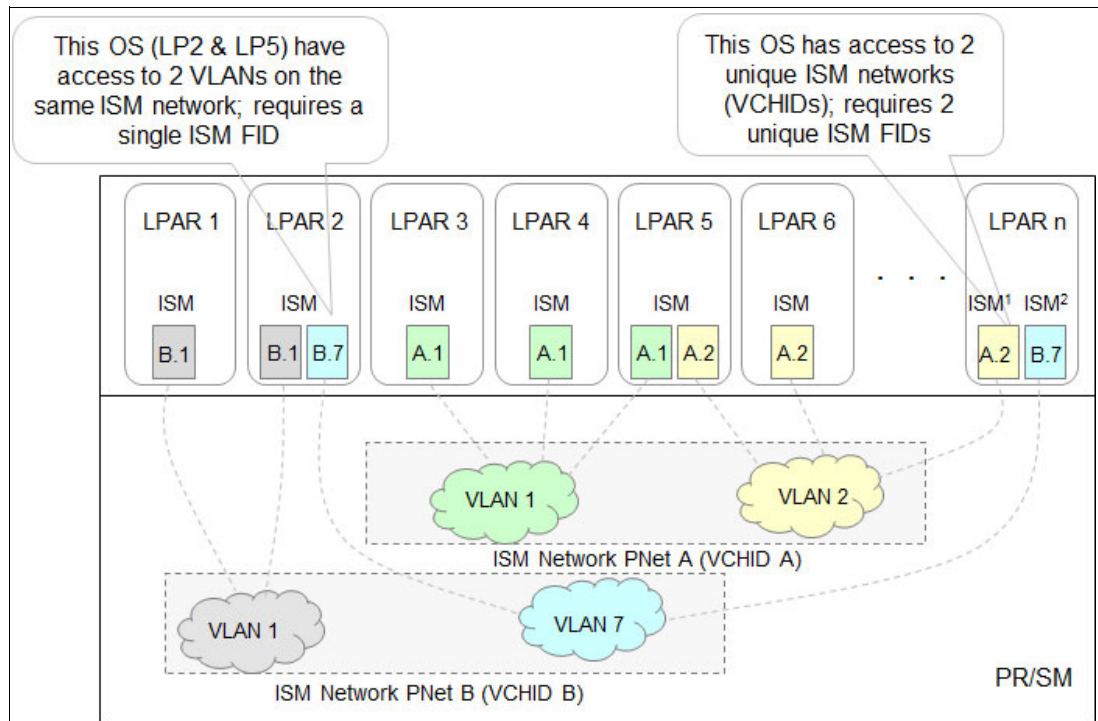


Figure D-17 Multiple LPARs connected through multiple VLANs

Workloads can be logically isolated on separate ISM VCHIDs. Alternatively, workloads can be isolated by using VLANs. The ISM VLAN definitions are inherited from the associated IP network (OSA or HiperSockets).

Configuration considerations

The IOCDS (HCD) definitions for ISM PCI VFs are not directly related to the software (SMC-D) use of ISM (that is, the z/OS TCP/IP and SMC-D implementation and usage are not directly related to the I/O definition).

The user defines a list of ISM FIDs (VFs) in IOCDS (HCD), and z/OS dynamically selects an eligible FID that is based on the required PNet ID. FIDs or VFs are *not* defined in Communications Server for z/OS TCP/IP. Instead, z/OS selects an available FID for a specific PNET. Access to more VLANs does not require configuring extra VFs.

Note: Consider over-provisioning the I/O definitions; for example, consider defining eight FIDs instead of five.

For native PCI devices, FIDs must be defined. Each FID in turn also defines a corresponding VF. In terms of operating system administration tasks, the administrator typically references FIDs. VFs (and VF numbers) often are transparent.

D.3.9 Software use of ISM

ISM enables SMC-D, which provides SMC capability within the CPC (SMC without requiring RoCE hardware or network equipment). Host virtual memory is managed by each OS (similar to SMC-R, logically shared memory) following IBM Z PCI I/O translation architecture.

Only minor changes are required for z/VM guests. An OS can be enabled for SMC-R and SMC-D. SMC-D is used when both peers are within the same CPC (and ISM PNet and VLAN). After the ISM HCD configuration is complete, SMC-D can be enabled in z/OS with a single TCP/IP parameter (GLOBALCONFIG SMCD). ISM FIDs must be associated with an IP network. The association is accomplished by matching PNet IDs (for example, HiperSockets and ISM).

Note: ISM FIDs must be associated with HiperSockets or with an OSA adapter by using a PNet ID. It cannot be associated to both.

D.3.10 SMC-D over ISM prerequisites

SMC-D over ISM features the following prerequisites:

- ▶ IBM z14, z13s, or z13 (Driver 27): HMC/SE for ISM vPCI Functions.
- ▶ At least two z/OS V2.2 systems (or later) in two LPARs on the same CPC with required service installed:
 - SMC-D can communicate with another z/OS V2.2 instance only and peer hosts must be on the same CPC and ISM PNet.
 - SMC-D requires an IP Network with access through OSA or HiperSockets that includes a defined PNet ID that matches the ISM PNet ID.
- ▶ If running as a z/OS guest under z/VM, z/VM 6.3 with APAR VM65716, including APARs is required for guest access to RoCE (Guest Exploitation only).
- ▶ Linux support is planned for a future deliverable.

The required APARs per z/OS subsystem are listed in Table D-1.

Table D-1 Prerequisite APARs for SMC-D enablement

Subsystem	F MID	APAR
IOS	HBB77A0	OA47913
Communications Server SNA VTAM	HVT6220	OA48411
Communications Server IP	HIP6220	PI45028
HCD	HCS77A0 HCS7790 HCS7780 HCS7770 HCS7760 HCS7750	OA46010
IOCP	HIO1104	OA47938
HCM	HCM1F10 HCM1E10 HCM1D10 HCM1C10 HCM1B10 HCM1A10	IO23612

Restrictions: SMC (existing architecture) cannot be used in the following circumstances:

- ▶ Peer hosts are not within the same IP subnet and VLAN
- ▶ TCP traffic requires IPsec or the server uses FRCA

D.3.11 Enabling SMC-D support in z/OS Communications Server

The new parameter SMCD (see Figure D-18) is available on the GLOBALCONFIG statement in the TCP/IP profile of the z/OS Communications Server (similar to the SMCR parameter). The SMCD parameter is the only parameter that is required to enable SMC-D.

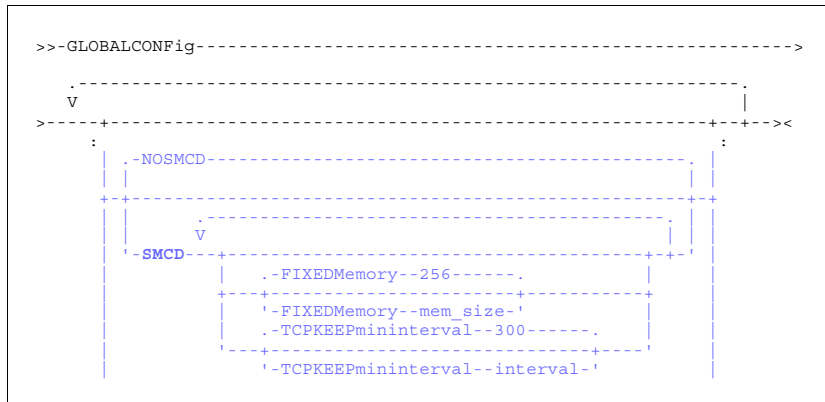


Figure D-18 SMCD parameter in GLOBALCONFIG

The key difference from the SMCR parameter is that ISM PFIDs are not defined in TCP/IP. Rather, ISM FIDs are discovered automatically based on matching PNETID that is associated with the OSD or HiperSockets. An extract from z/OS Communications Server: IP Configuration Reference is shown in Figure D-18.

D.3.12 SMC-D support overview

SMC-D requires IBM z14, or IBM z13 and IBM z13s servers at driver level 27 or later for ISM support.

IOCP required level: The required level of IOCP for z14 is V5 R4 L1 or later with PTFs. Defining ISM devices other than the z14, z13, or z13s systems is not possible. For more information, see the following publications:

- ▶ *IBM Z Stand-Alone Input/Output Configuration Program User's Guide*, SB10-7166
- ▶ *IBM Z Input/Output Configuration Program User's Guide for ICP IOCP*, SB10-7163

SMC-D requires the following software:

- ▶ z/OS V2R2 with PTFs (see Table D-1 on page 447) or later is the only supported operating system for the SMC-D protocol. Consider the following points:
 - HCD APAR (OA46010) is required.
 - You cannot roll back to previous z/OS releases.
- ▶ z/OS guests under z/VM 6.3 and later are supported to use SMC-D.
- ▶ At the time of this writing, IBM is working with its Linux distribution partners to include support in future Linux on Z distribution releases.

Other ISM considerations

ISM systems include the following limits:

- ▶ A total of 32 ISM VCHIDs (in IOCDS/HCD) per CPC. Each IOCDS/HCD VCHID represents a unique internal shared memory network each with a unique Physical Network ID.
- ▶ A total of 255 VFs per VCHID (8k VFs per CPC). For example, the maximum number of virtual servers that can communicate over the same ISM VCHID is 255.
- ▶ Each ISM VCHID in IOCDS/HCD represents a unique (isolated) internal network, each having a unique Physical Network ID (PNet IDs are configured in HCD/IOCDS).
- ▶ ISM VCHIDs support VLANs (can be subdivided into VLANs).
- ▶ ISM provides a GID (internally generated) to correspond with each ISM FID.
- ▶ All MACs (VMACs), MTU, physical ports, and Frame sizes are not applicable.
- ▶ ISM is supported by z/VM (for pass-through guest access to support the new PCI function).

More information

For more information about a configuration example for SMC-D, see *IBM z/OS V2R2 Communications Server TCP/IP Implementation - Volume 1*, SG24-8360.



E

IBM Dynamic Partition Manager

This appendix describes IBM Dynamic Partition Manager (DPM) on IBM Z and how the DPM environment can be set up and managed.

This appendix includes the following topics:

- ▶ E.1, “Introduction” on page 452
- ▶ E.2, “Why use DPM?” on page 452
- ▶ E.3, “DPM overview” on page 453
- ▶ E.4, “Setting up the DPM environment” on page 454

E.1 Introduction

DPM is a resource management and operational environment that provides a simplified approach to creating and managing virtualized IBM Z servers. DPM reduces the barriers to the adoption of IBM Z for new and existing customers.

The implementation provides built-in integrated capabilities that allow advanced virtualization management on IBM Z servers. With DPM, customers can use their Linux and virtualization skills while realizing the full value of IBM Z hardware's robustness and security in a workload optimized environment.

DPM provides facilities to define and run virtualized computing systems by using a firmware-managed environment that coordinates the physical system resources that are shared by the partitions¹. The partitions' resources include processors, memory, network, storage, Crypto, and Accelerators.

DPM provides a new mode of operation for IBM Z servers that provides the following benefits:

- ▶ Facilitates defining, configuring, and operating partitions, similar to the way these tasks are performed on other platforms.
- ▶ Lays the foundation for a general IBM Z new user experience.

DPM is not an extra hypervisor for IBM Z servers. DPM uses the PR/SM hypervisor infrastructure and provides an intelligent interface that allows customers to define, use, and operate the platform virtualization with little or no IBM Z experience.

Note: When IBM z14 servers are set to run in DPM mode, the following components are supported:

- ▶ Linux virtual servers that are running in a partition (LPAR)
- ▶ KVM hypervisor^a for Linux guests
- ▶ z/VM with Linux guests
- ▶ Virtual appliances that use the Secure Service Container (SSC) framework

a. Available with Linux distributions

E.2 Why use DPM?

DPM mode is targeted at customers (distributed market) with no specific IBM Z skills or z/VM knowledge who want to implement and use the cloud infrastructure to consolidate and integrate their IT by using IBM Z technology. DPM also eases the management and administration of their Linux environment and workloads.

DPM is of special value for customer segments with the following characteristics:

- ▶ New IBM Z, or Linux adopters, or distributed-driven:
 - Likely not z/VM² users
 - Looking for integration into their distributed business models
 - Want to ease migration of distributed environments to IBM Z servers and improve centralized management

¹ DPM uses the term "partition", which is the same as logical partition (LPAR).

² z/VM is the IBM Z solution for a second-level hypervisor and has enterprise class virtualization capabilities and efficient I/O, processor, and memory resource management. For more information, see: <https://www.vm.ibm.com>.

- ▶ Currently not running on IBM Z servers:
 - No IBM Z skills
 - Want to implement cloud
 - Have expectations that are acquired from another hypervisors' management, such as VMware, KVM, and Citrix

E.3 DPM overview

Traditional IBM Z servers are highly virtualized with the goal of maximizing the use of compute (processor and memory) and I/O (storage and network) resources, and simultaneously lowering the total amount of resources needed for workloads. For decades, virtualization was embedded in the IBM Z architecture and built into the hardware and firmware.

Virtualization requires a hypervisor, which manages resources that are required for multiple independent virtual machines. The IBM Z hardware hypervisor is known as IBM Processor Resource/Systems Manager (PR/SM). PR/SM is implemented in firmware as part of the base system. It fully virtualizes the system resources, and does not require extra software to run.

PR/SM allows the defining and managing of subsets of the IBM Z resources in LPARs. The LPAR definitions include several logical processing units (LPUs), memory, and I/O resources. LPARs can be added, modified, activated, or deactivated in IBM Z platforms by using the traditional Hardware Management Console (HMC) interface.

DPM uses all its capabilities as the foundation for the new user experience. In addition to these capabilities, DPM provides an HMC user interface that allows customers to define, implement, and run Linux partitions without requiring deep knowledge of the underlying IBM Z infrastructure management; for example, input/output configuration program (IOCP) or hardware configuration definition (HCD).

The DPM infrastructure is shown in Figure E-1.

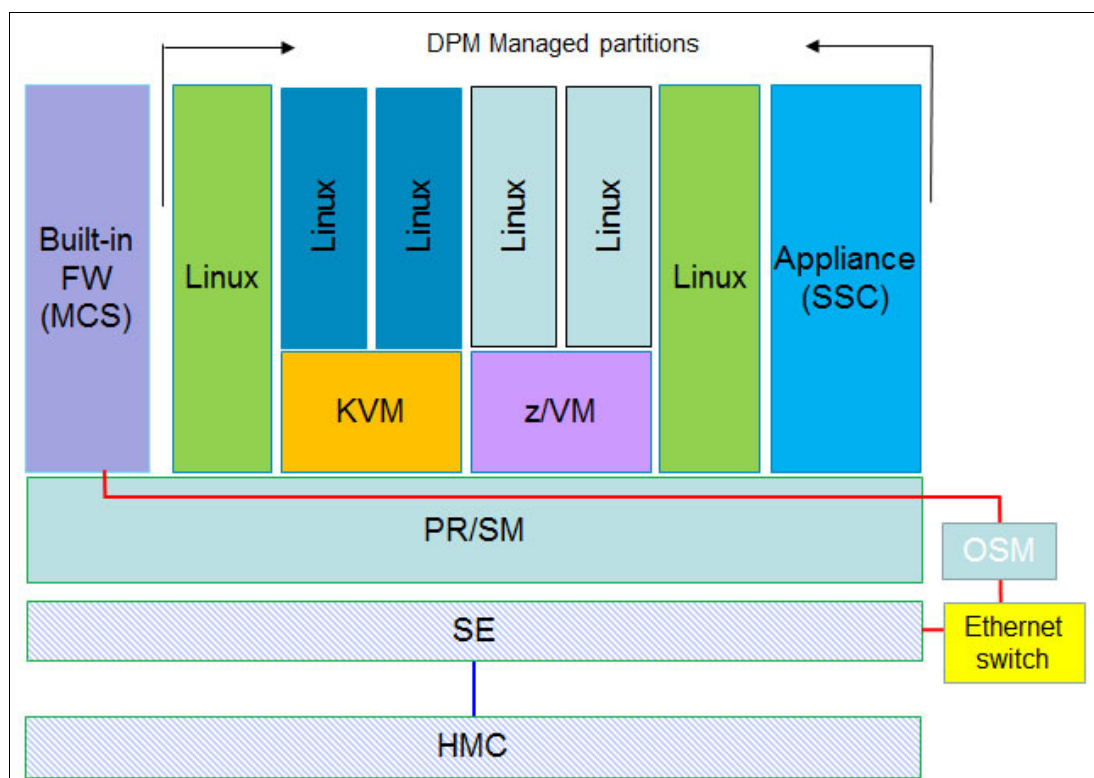


Figure E-1 High-level view of DPM implementation

The firmware partition (similar to the PCIe support partitions, which is also known as master control services [MCS] partition), along with the Support Element (SE), provides instrumentation to create and manage the Linux native partitions, or partitions that are running kernel-based virtual machine (KVM) code. The connectivity from the SE to the MCS is provided through the internal management network by two OSA-Express 1000BASE-T that are acting as OSA Management adapters.

This implementation integrates platform I/O resource management and dynamic resource management.

E.4 Setting up the DPM environment

The DPM is a mode of operation that requires IBM z14 M0x, IBM z14 ZR1, IBM z13 (driver 27), or z13s CPCs. Enabling DPM is a disruptive action. The selection of DPM mode of operation is done by using a function that is called Enable Dynamic Partition Manager, under the CPC Configuration menu in the SE interface, as shown in Figure E-2 on page 455. The DPM mode of operation setting is normally performed at machine installation time by the service support representative (SSR).

Note: DPM is a feature code (FC 0016) that can be selected during the machine order process. If selected, a pair of OSA Express 1000BASE-T adapters must be included in the configuration.

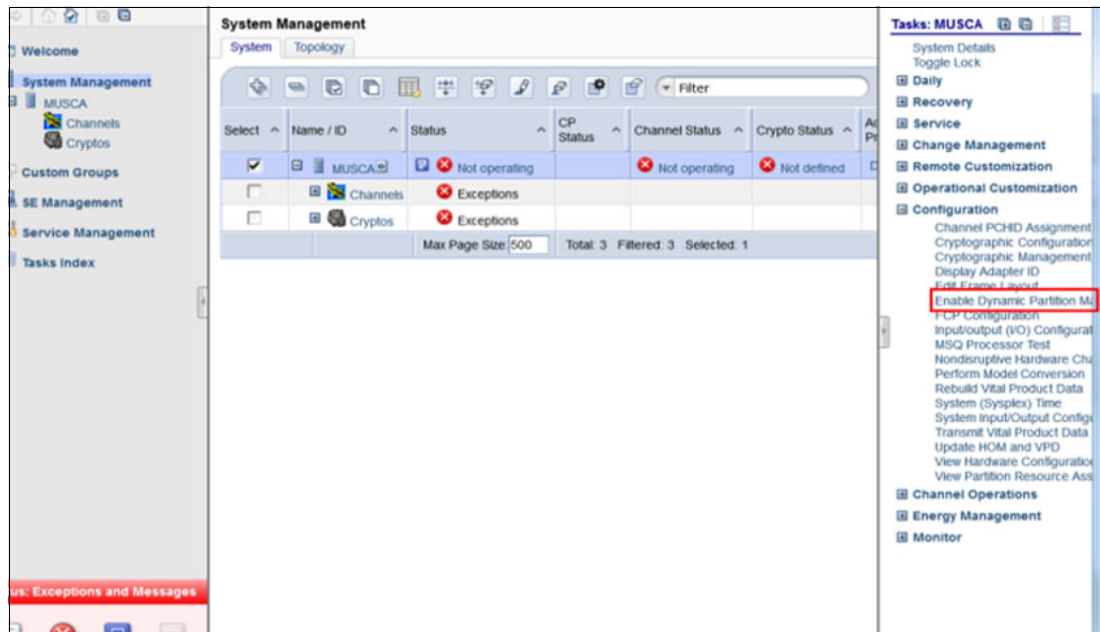


Figure E-2 Enabling DPM mode of operation from the SE CPC configuration options

After the option is selected, a new window opens (see Figure E-3) in which you enter the two OSA Express 1000BASE-T ports that are selected and cabled to the two switches in the frame during the Z server installation.

Enable DPM - MUSCA

To enable DPM on the system, the two adapters cabled to be the OSM adapters for this system must first be identified. Note that only adapters whose card type supports OSM are displayed. Adapters are identified by Adapter ID (PCHID).

OSM 1:

0130

▼

OSM 2:

014C

▼

Click **Enable** to enable DPM. The Support Element will be restarted. This might take several minutes to complete.

Enable

Cancel

Help

Figure E-3 Entering the OSA ports that are used by the management network

Note: During the machine installation process, the IBM SSR connects the two OSA Express 1000BASE-T cables to the Ethernet Top of Rack switches in the designated ports.

After entering the OSA adapter port numbers that were cabled to the switches, click **Enable**. The SE then restarts, and, when finished, the DPM mode becomes active and operational.

Important: Consider the following points:

- ▶ A CPC in DPM mode cannot be part of an Ensemble that is managed by Unified Resource Manager. The HMC that is used to enable the CPC in DPM mode must *not* be an Ensemble HMC (Primary or Backup Ensemble HMC).
- ▶ All definitions that are made for the CPC (if any) before the DPM mode is activated are saved and can be brought back if you choose to revert to standard PR/SM mode. However, when switching the CPC into standard PR/SM mode, any definitions that are made with the CPC in DPM mode are lost.

The DPM mode welcome window is shown in Figure E-4. The three options at the bottom (Getting Started, Guides, and Learn More) include mouse-over functions that briefly describe their meaning or provide more functions.

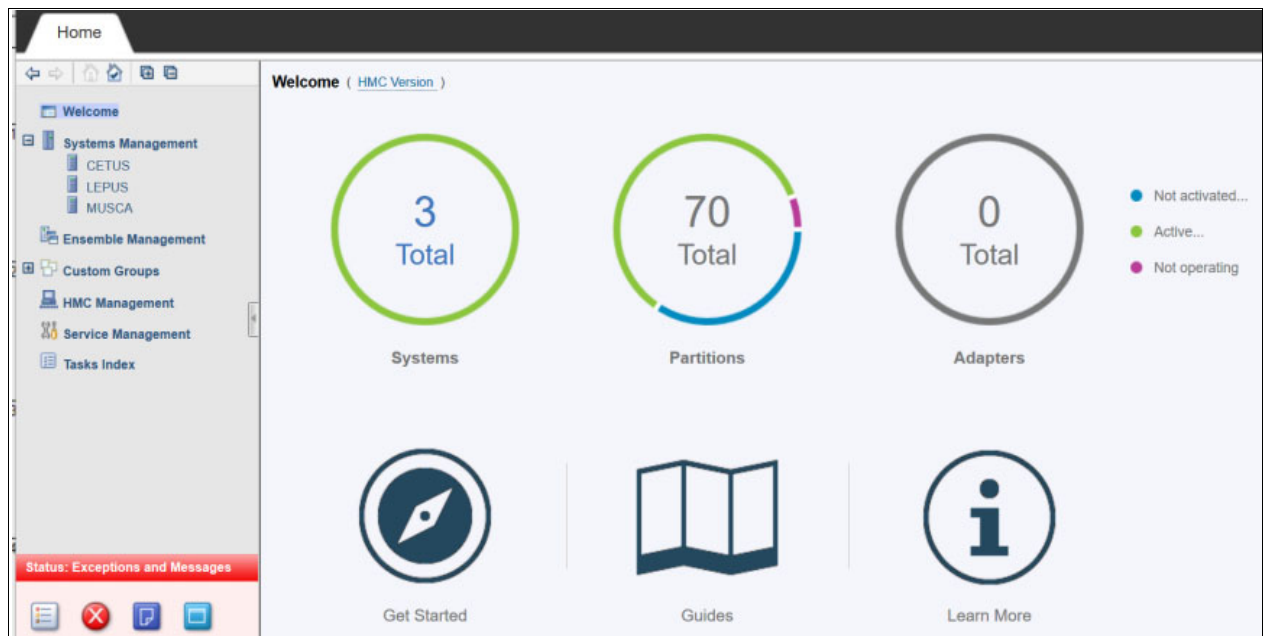


Figure E-4 DPM mode welcome window

The HMC can monitor and control up to 32 IBM Z CPCs. The monitored and controlled CPCs must be *defined* to the HMC by using the Object Definition task and adding the CPC object.

The welcome window that is shown in Figure E-4 opens only when at least one HMC defined CPC is active in DPM mode. Otherwise, the traditional HMC window is presented when you log on to the HMC.

The welcome window from a traditional HMC when none of the defined CPC objects are running in DPM mode is shown in Figure E-5.

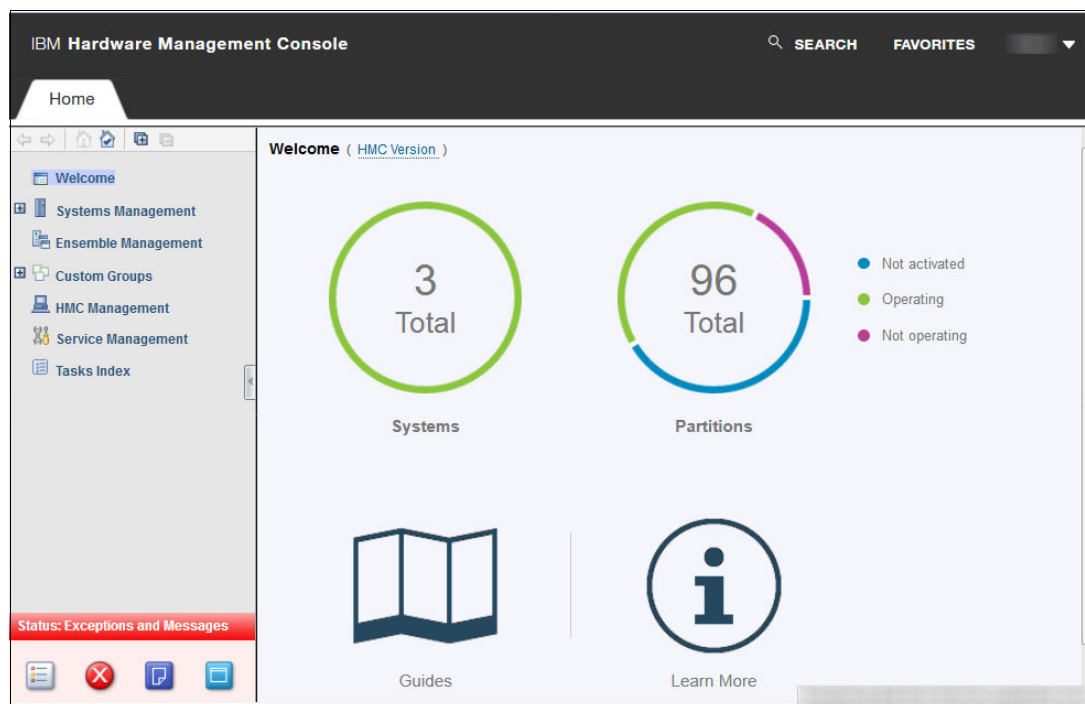


Figure E-5 Traditional HMC Welcome (2.14.0) window (no defined CPCs are running in DPM mode)

E.4.1 Defining partitions in DPM mode

After the CPC is in DPM mode, the user can choose one of the options that is provided in the welcome window to learn more about the process of defining partitions, browse the tutorials, or start creating the environment by defining and activating the partitions.

The three options that are presented to the user in the HMC welcome page when at least one CPC is running in DPM mode are shown in Figure E-6.

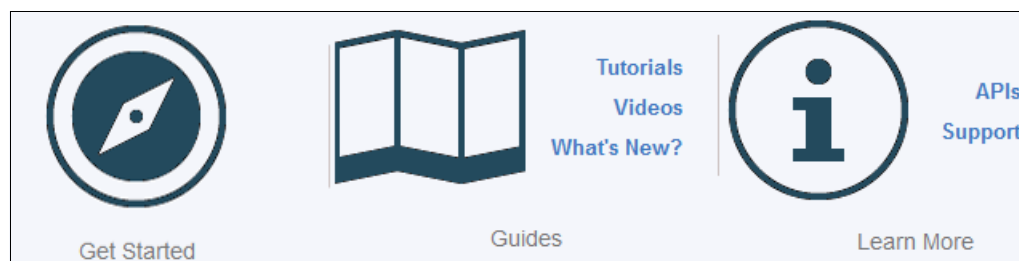


Figure E-6 User Options when the HMC presents the DPM welcome window

As shown in Figure E-6, the following options are available:

- ▶ The Guides option provides tutorials, videos, and information about What's New in DPM.
- ▶ The Learn More option covers the application programming interfaces (APIs).
- ▶ The Support option takes the user to the IBM ResourceLink website.

The first option on the left of the window that is shown on Figure E-6 on page 457, is Getting Started. This option starts the DPM wizard application on the HMC, which allows users to define their partitions and associate processor and memory resources, network and storage I/O, crypto adapters, and accelerators to them.

From the Getting Started with DPM window, users can select the Partition option that opens the Create Partition wizard. The Create Partition wizard can also be accessed clicking **Next** at the bottom of Getting Started with DPM window.

On the left banner (see Figure E-7 on page 459), the following HMC create partition wizard steps are available to define and activate a partition:

- ▶ **Welcome:** Initial window that contains basic information about the process.
- ▶ **Name:** This window is used to provide name and description for the partition being created.
- ▶ **Processors:** The partition's processing resources are defined in this window.
- ▶ **Memory:** This window is used to define partition's initial and maximum memory.
- ▶ **Network:** The window in which users define partition's network NICs resources.
- ▶ **Storage:** The Storage Groups are used to manage FC³ (CKD) and FCP (SCSI) storage.
- ▶ **Accelerators:** Partition resources, such as zEDC, can be added in this window.
- ▶ **Cryptos:** Wizard window in which users define their cryptographic resources.
- ▶ **Boot:** In this window, users define the partition's OS and their source. The following options as the source for loading an OS are available:
 - FTP Server
 - Storage Device (SAN)
 - Network Server (PXE)
 - Hardware Management Console removable media (USB key or DVD media)
 - ISO image
- ▶ **Summary:** This window provides a view of all defined partition resources.

The final step after the partition creation process is to start it. After the partition is started (Status: Active), the user can start the messages or the Integrated ASCII console interface to operate it.

³ FC (FICON) storage management requires HMC 2.14.1 or later (which includes DPM 3.2) and z14 Hardware (Driver level 36 or later). For z13/z13s, DPM supports FCP (SCSI) storage only.

An option at the lower left that is called Advanced is shown in Figure E-7. This option allows users to open a window that contains all definitions that are made for the partition. This window provides more settings for some of the definitions, such as defining a processor as shared or dedicated, and associating a weight for a partition.

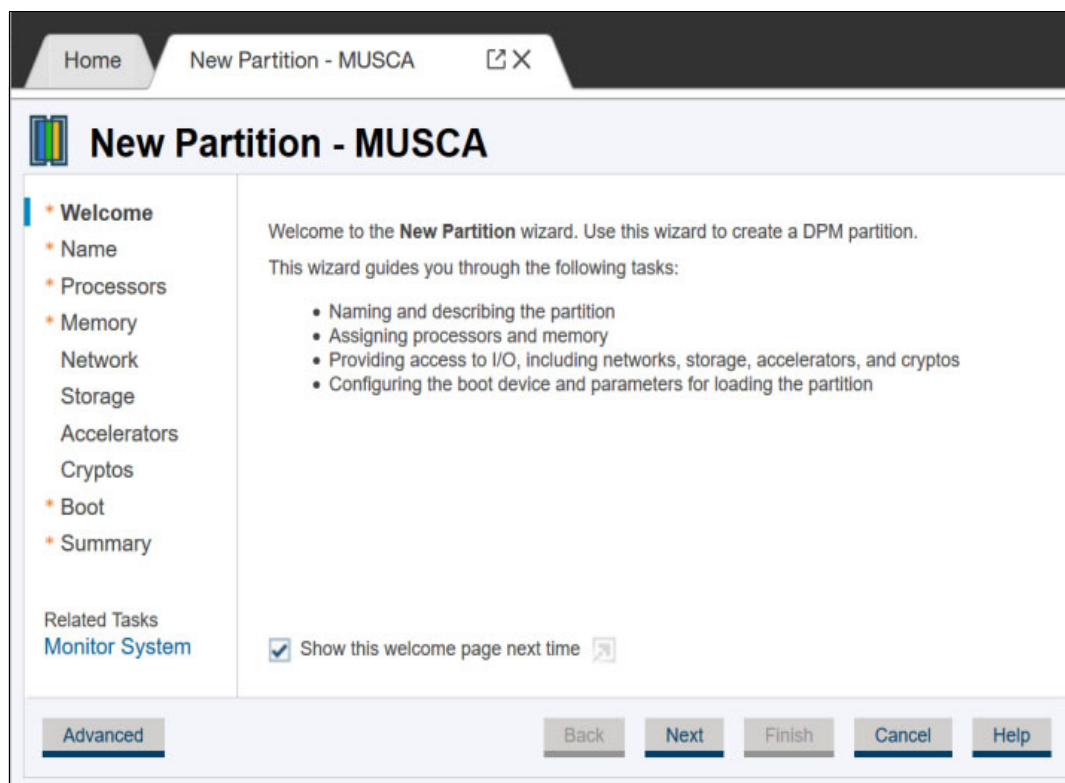


Figure E-7 DPM wizard welcome window options

An important other facility that is provided by the DPM is the Monitor System. This option allows users to monitor and manage their DPM environment. The following monitoring and management capabilities are available:

- ▶ Partition overall performance (shown in usage percentages) including:
 - Processors
 - Storage utilization
 - Network adapters
 - Storage adapters
 - Cryptos
 - Accelerators
 - Power Consumption in KW
 - Environmentals (Ambient Temperature in Fahrenheit)
- ▶ Adapters that exceed a user predefined threshold value
- ▶ Overall port usage in the last 36 hours
- ▶ Usage details are available by selecting one of the performance indicators
- ▶ Manage Adapters Task

E.4.2 Summary

DPM provides simplified IBM Z hardware and virtual infrastructure management, including integrated dynamic I/O management for users that intend to run Linux on Z and the KVM hypervisor in a partition and for z/VM.

The new mode, DPM, provides partition lifecycle and dynamic I/O management capabilities by using the HMC for the following tasks:

- ▶ Create and provision: Creating partitions, assigning processors and memory, configuring I/O Adapters (Network, FCP Storage, Crypto, and Accelerators).
- ▶ Manage the environment: Modification of system resources without disrupting running workloads.
- ▶ Monitor and troubleshoot the environment: Source identification of system failures, conditions, states, or events that might lead to workload degradation.

A CPC can be in DPM mode or standard PR/SM mode. The mode is enabled before the CPC power-on reset.

DPM mode requires two OSA-Express 1000BASE-T Ethernet features for primary and backup connectivity (OSA-Express 1000BASE-T Ethernet), along with associated cabling (hardware for DPM FC0016).



IBM zEnterprise Data Compression Express

This appendix briefly describes the optional IBM zEnterprise Data Compression (zEDC) Express feature of the z14 ZR1, z14, z13, z13s, zEC12, and IBM zBC12 servers.

This appendix includes the following topics:

- ▶ F.1, “Overview” on page 462
- ▶ F.2, “zEDC Express” on page 462
- ▶ F.3, “Software support” on page 463

F.1 Overview

The growth of data that must be captured, transferred, and stored for large periods is unrelenting. In addition, software-implemented compression algorithms can be costly in terms of processor resources and storage costs.

zEDC Express, which is an optional feature available for z14 ZR1, z14, z13, z13s, zEC12, and zBC12 servers, addresses these requirements by providing hardware-based acceleration for data compression and decompression. zEDC provides data compression with lower CPU consumption than compression technology that was available on the IBM Z server.

The use of the zEDC Express feature with the z/OS V2R1 zEnterprise Data Compression acceleration capability (or later releases) is designed to deliver an integrated solution. It helps 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.

F.2 zEDC Express

zEDC Express is an optional feature (FC 0420). It is designed to provide hardware-based acceleration for data compression and decompression.

The feature installs in the Peripheral Component Interconnect Express Plus (PCIe+) I/O drawer. Up to eight features can be installed on the system. One PCIe adapter or compression coprocessor is available per feature, which implements compression as defined by RFC1951 (DEFLATE).

A zEDC Express feature can be shared by up to 15 logical partitions (LPARs) on the same CPC.

Adapter support for zEDC is provided by Resource Group (RG) code that runs on the system-integrated firmware processor (IFP). The recommended high availability configuration per server is four features. This configuration provides continuous availability during concurrent update.

For resilience, the z14 ZR1 system always includes four independent RGs on the system, which share the IFP. Install a minimum of two zEDC features for resilience and throughput. The feature installs in the PCIe+ I/O drawer.

The PCIe I/O drawer structure and the relationships among card slots, domains, and resource groups are shown in Figure F-1.

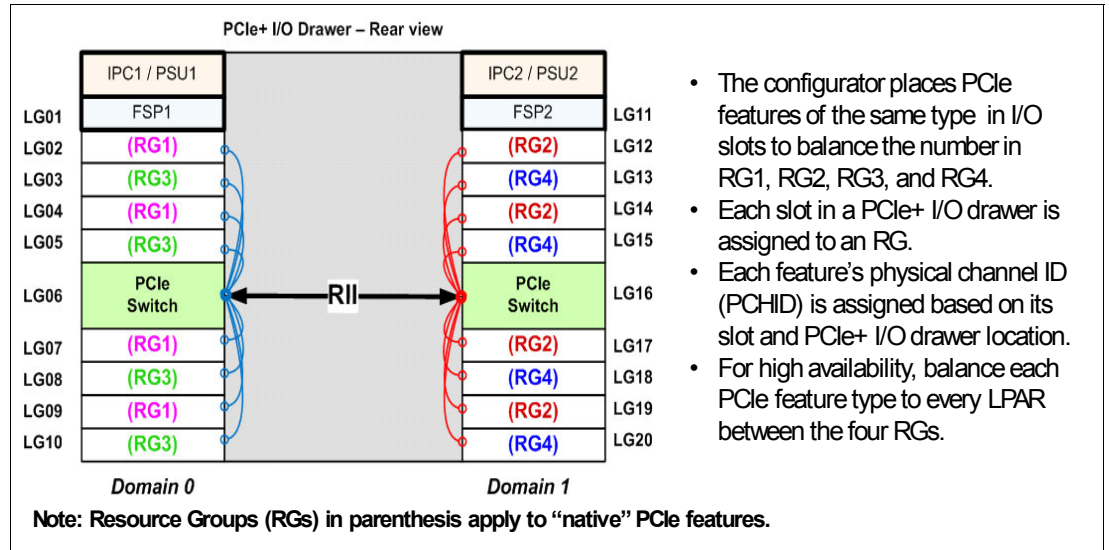


Figure F-1 Relationships among PCIe I/O drawer card slots, I/O domains, and resource groups

F.3 Software support

Support of the zEDC Express function is provided by z/OS V2R1 zEnterprise Data Compression or later for data compression and decompression. Support for data recovery (decompression) in the case that zEDC is not installed, or installed but not available, is provided through software in z/OS V2R2, V2R1, V1R13, and V1R12 with the appropriate program temporary fixes (PTFs).

Software decompression is slow and can use considerable processor resources. Therefore, it is not suggested for production environments.

A specific fix category that is named `IBM.Function.zEDC` identifies the fixes that enable or use the zEDC function.

Reference: z/OS support for the zEDC can be found by using `FIXCAT: IBM.Function.zEDC`.

z/OS guests that run under z/VM V6.3 with PTFs and later can use the zEDC Express feature. zEDC for z/OS V2.1 or later and the zEDC Express feature are designed to support a data compression function to help provide high-performance, low-latency compression without significant CPU processor usage. This feature can help to reduce disk usage, and provide optimized cross-platform exchange of data and higher write rates for SMF data.

For more information, see the [Additional Enhancements to z/VM 6.3 page](#) of the IBM Systems website.

IBM 31-bit and 64-bit SDK for z/OS Java Technology Edition, Version 7 Release 1 (5655-W43 and 5655-W44) (IBM SDK 7 for z/OS Java) now provides use of the zEDC Express feature and Shared Memory Communications-Remote Direct Memory Access (SMC-R), which is used by the 10GbE RoCE Express feature.

For more information about how to implement and use the zEDC feature, see *Reduce Storage Occupancy and Increase Operations Efficiency with IBM zEnterprise Data Compression*, SG24-8259.

F.3.1 IBM Z Batch Network Analyzer

IBM 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, and provides graphical and text reports, including Gantt charts and support for alternative processors.

zBNA can be used to analyze client-provided System Management Facilities (SMF) records to identify jobs and data sets that are candidates for zEDC compression across a specified time window (often a batch window).

zBNA can generate lists of data sets by the following jobs:

- ▶ Performs hardware compression and might be candidates for zEDC
- ▶ Might be zEDC candidates, but are not in extended format

zBNA also can help you estimate the use of zEDC features and help determine the number of features needed. The following resources are available:

- ▶ IBM Employees can obtain zBNA and other CPS tools at the [IBM Z Batch Network Analyzer \(zBNA\) Tool page](#) of the IBM Techdocs website.
- ▶ IBM Business Partners can obtain zBNA and other CPS tools at the [IBM PartnerWorld website](#) (log in required).
- ▶ IBM clients can obtain zBNA and other CPS tools at the [IBM Z Batch Network Analyzer \(zBNA\) Tool page](#) of the IBM Techdocs Library website.



16U Reserved feature

The following new 16U Reserved feature allows clients to create all-in-one solutions to run their entire business or independent application or cloud solutions:

- ▶ FC 0617 is “16U Reserved” space in the rack for client-owned components to be installed in the Z rack.
- ▶ Allows for clients to integrate other hardware into the single 19-inch rack, which lowers data center footprint requirements.
- ▶ Reserved space can be especially useful for space-constrained data centers or to keep rack-mounted HMC and other rack-mounted equipment together with the Z server.

This appendix includes the following topics:

- ▶ G.1, “General rules” on page 466
- ▶ G.2, “Basic physical requirements” on page 467

G.1 General rules

Consider the following general rules regarding the 16U Reserved feature:

- ▶ A system with FC 0617 can have 0, 1, or 2 PCIe+ I/O drawers.
- ▶ FC 0617 can be ordered as part of a factory new build or as field MES upgrade.
- ▶ Upon ordering, 16 EIA rack units are reserved to prevent further I/O upgrades that require PCIe+ I/O drawers 3 or 4.
- ▶ The feature code must be installed before any extra components are added to keep the machine under warranty.
- ▶ Feature cannot be removed after it is ordered.
- ▶ A set of ballast plates is provided and installed in the bottom of the rack (to mitigate potential tipping hazard) for configurations with 0 or 1 PCIe+ I/O drawers.
- ▶ Requires two extra PDUs to be installed in the lower vertical locations on each side of the rear of the rack:
 - If MES is ordered and second PCIe+ I/O drawer is installed, no other PDUs are required.
 - Specific PDU plugging locations must be used for the equipment installed (as described in *IBM 3907 Installation Manual for Physical Planning*, GC28-6974).
- ▶ A set of 32 1U air flow filler plates (for front and rear side of the rack) is included for cosmetics and to manage proper airflow.
- ▶ Cable management spine hardware is included with the 16U Reserved feature.

The front and rear of the system with the 16U Reserved defined location are shown in Figure G-1. The rear of the system view displays the cable management spine that is mounted at the center of the rack.

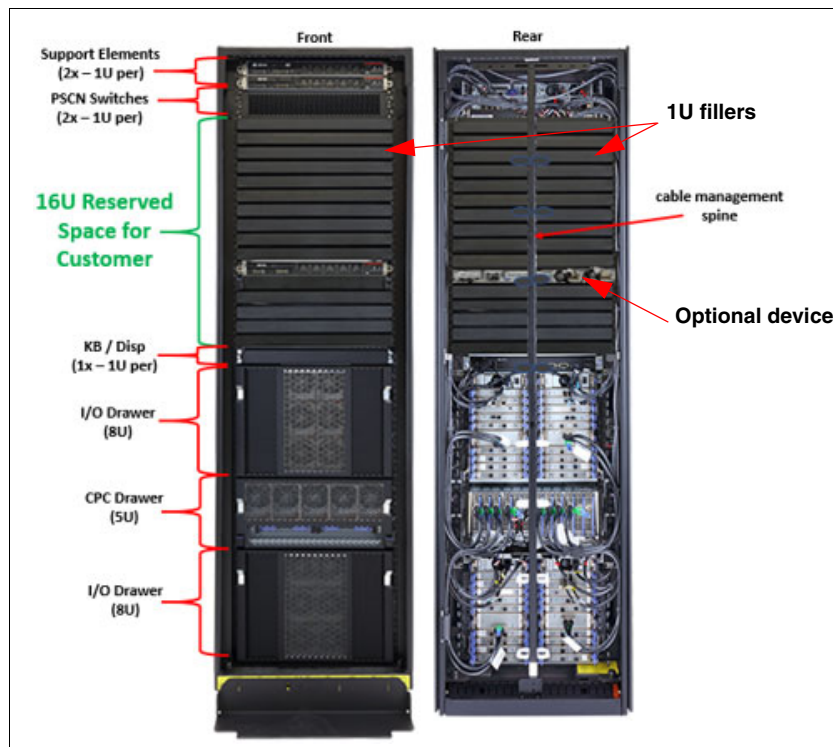


Figure G-1 16U Reserved space for customer use

Note: The configuration that is shown in Figure G-1 on page 466 is for reference only. The optional device that is installed in the rack is for presentation purposes only. Actual fillers might differ from the fillers that are shown.

G.2 Basic physical requirements

The following basic physical requirements must be met:

- ▶ Components fit within a 19-inch rack EIA rail-to-rail width (465 mm usable), and accommodate round rail holes
- ▶ Component rails fit within a 718 mm (928.25-inch) EIA rail-to-rail depth
- ▶ Content cannot extend beyond the face of the front (25 mm) and rear (100 mm) vertical EIA rails
- ▶ Components include front-to-rear airflow
- ▶ One unit cannot weigh more than 20.4 kg (45 lbs) per EIA location
Example: a 4U unit can weigh up to 81.6 kg (180 lbs)
- ▶ Component to be installed shows safety certification labels that are required for server components in your country (UL, TUV, and so on)
- ▶ Components should be populated from lowest open EIA and move up to avoid tipping hazards (similar to other racks or shelving)
- ▶ No liquid-based systems to be added because humidity monitoring is active
- ▶ FC 0167 effectively caps the number of PCIe+ I/O drawers to a quantity of two; therefore, this FC should not be ordered if it is expected that future MESs that require a third or fourth PCIe+ I/O drawer will be added

A sample usage of the 16U Reserved feature that uses various components for a single solution is shown in Figure G-2 on page 468. Equipment that can be considered to use the space include the following examples:

- ▶ Rack mount HMC
- ▶ Rack mount TKE workstation
- ▶ V7000 Storage
- ▶ V9000 Storage
- ▶ Network/SAN switches
- ▶ Future products

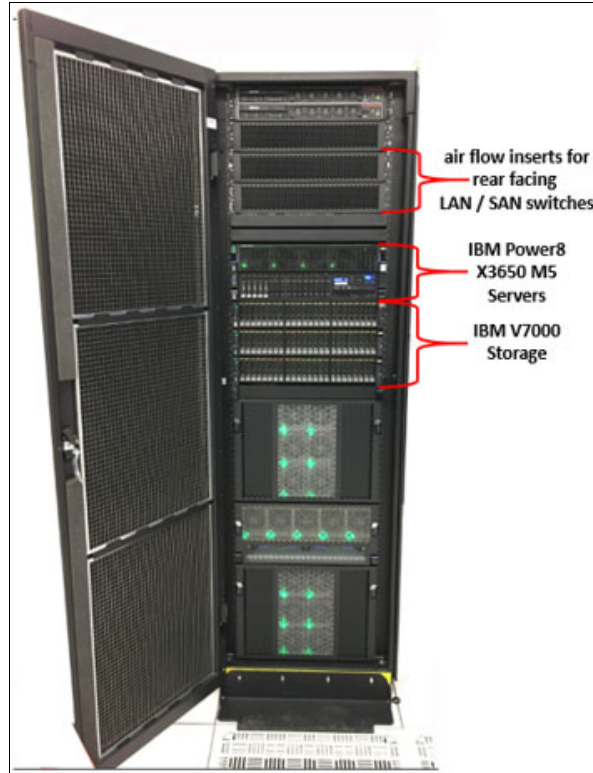


Figure G-2 Example of FC 0617 populated in the z14 ZR1 rack

- Intelligent Power Distribution Units (PDUs) two or four per system, depending on the configuration, switchable (Ethernet controlled). A breakout of a single PDU is shown in Figure G-3.

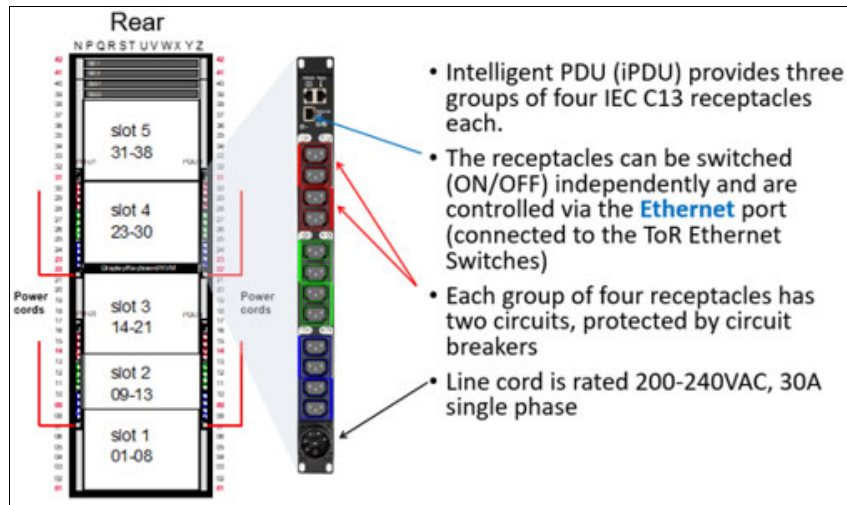


Figure G-3 Intelligent PDU breakout

The PDUs are mounted at the rear right and rear left sides of the rack and can exit the top or bottom of the rack. Each PDU features a single phase line cord. The number of line cords that are required depends on the system configuration.

The total loss of one PDU has no effect on system operation. Systems that specify two line cords can be brought up with one line cord and continue to run.

The larger systems that have a maximum of four PDUs installed must include four installed line cords. Systems that specify four line cords can be started with two line cords on the *same side* with sufficient power to keep the system running. The number of PDUs that are installed (which is dependent on the number of I/O drawers that is installed) is listed in Table G-1.

Table G-1 I/O drawers and PDU and line cords

Number of I/O drawers	Number of PDUs and line cords
0	2
1	2
2	4
3	4
4	4

Note: FC 0617 16U Reserved requires four PDUs and four line cords, regardless of the number of I/O drawers.

The following general electrical power requirements must be met:

- ▶ 50/60 HZ AC
- ▶ Voltage ranges 200 V - 240 VAC single-phase wiring
- ▶ Two or four line cords, depending on the number of PDUs that are installed
- ▶ For HMC tower models, a customer-supplied outlet must provide 100 V - 130 V or 200 V - 240 V 50/60 HZ, single-phase AC power
- ▶ For the rack-mount HMC models, the customer must supply a PDU that provides C13 outlets for the three C13/C14 power jumper cables

The internal PDU power cabling for a configuration with no PCIe+ I/O Drawers and no FC 0617 16U Reserved is shown in Figure G-4. This configuration requires one pair of PDUs.

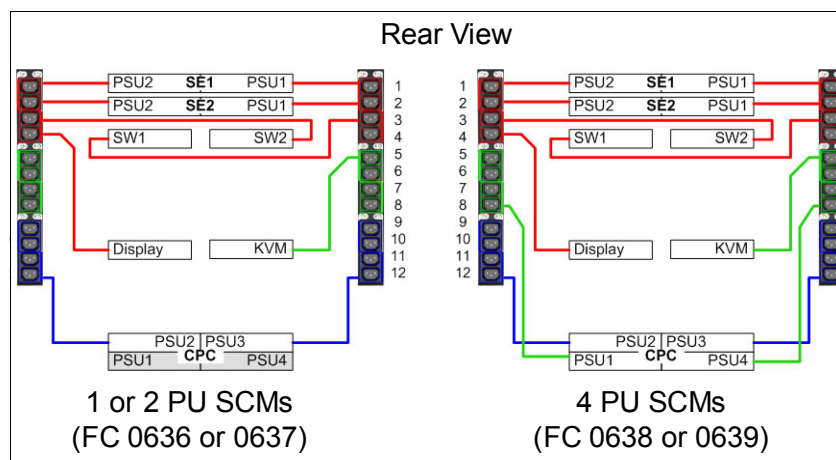


Figure G-4 PDU cabling with no PCIe+ I/O drawers installed

A configuration with one PCIe+ I/O drawer and no FC 0617 16U Reserved is shown in Figure G-5.

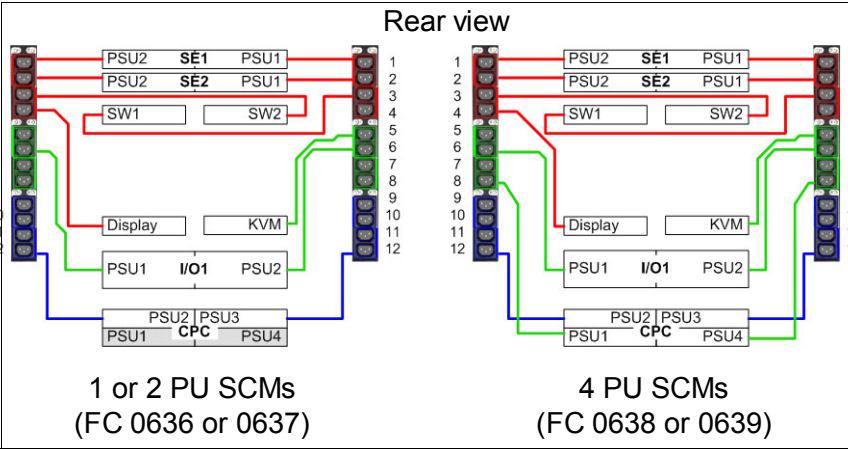


Figure G-5 PDU cabling with one PCIe+ I/O drawer installed

A configuration with two PCIe+ I/O drawers is shown in Figure G-6. The second pair of PDUs is installed and require two more power feeds from the customer.

When the FC 0617 (16U Reserved) is ordered, the bottom pair of PDUs is used to supply power outlets for the equipment that is installed in the designated space. When this feature is ordered, the maximum number of I/O drawers that is installed is two.

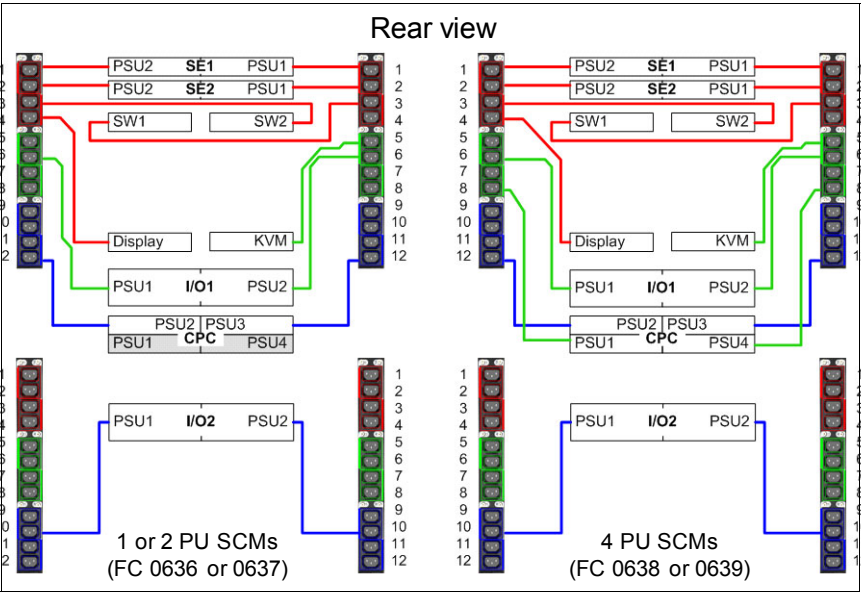


Figure G-6 PDU cabling with two PCIe+ I/O drawers installed

A configuration that includes PCIe+ I/O drawers 3 and 4 is shown in Figure G-7. This configuration requires four PU SCMs to be installed.

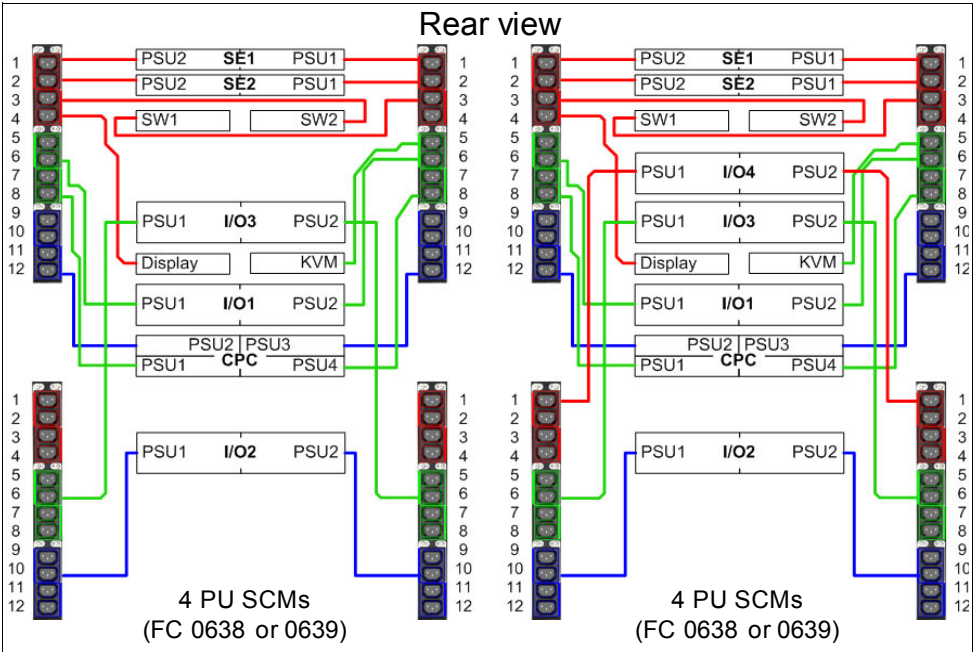


Figure G-7 PDU cabling with three or four PCIe+ I/O drawers installed

Related publications

The publications that are listed in this section are considered particularly suitable for a more detailed discussion of the topics that are covered in this book.

IBM Redbooks

The following IBM Redbooks publications provide more information about the topics in this document. Note that some publications that are referenced in this list might be available in softcopy only:

- ▶ *IBM z14 ZR1 Technical Introduction*, SG24-8550
- ▶ *IBM Z Connectivity Handbook*, SG24-5444
- ▶ *IBM Z Functional Matrix*, REDP-5157
- ▶ *IBM z14 Model ZR1 Configuration Setup*, SG24-8560
- ▶ *z Systems Simultaneous Multithreading Revolution*, REDP-5144
- ▶ *z/OS Infrastructure Optimization using Large Memory*, REDP-5146
- ▶ *IBM z14 Model Technical Introduction*, SG24-8550

You can search for, view, download, or order these documents and other Redbooks, Redpapers, Web Docs, draft, and other materials at the following website:

ibm.com/redbooks

Other publications

The following publications are also relevant as further information sources:

- ▶ *Capacity on Demand User's Guide*, SC28-6985
- ▶ *IBM 3907 Installation Manual for Physical Planning*, GC28-6974
- ▶ *PR/SM Planning Guide*, SB10-7169
- ▶ *IOCP Users Guide*, SB10-7172-01

Online resources

The following websites are also relevant as further information sources:

- ▶ IBM Resource Link:
<https://www.ibm.com/servers/resourceLink/hom03010.nsf?OpenDatabase&login>
- ▶ IBM Offering Information:
http://www.ibm.com/common/ssi/index.wss?request_locale=en

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services



IBM z14 Model ZR1 Technical Guide

SG24-8651-01
ISBN 0738456896



(1.0" spine)
0.875" <-> 1.498"
460 <-> 788 pages



SG24-8651-01

ISBN 0738457264

Printed in U.S.A.

Get connected

