Draft Document for Review August 21, 2017 1:18 pm SG24-8451-00

Redbooks
ibm.com/redbooks

IBM z14 Technical Guide

Octavian Lascu
Bo Xu
Esra Ufacik
Frank Packheiser
Hervey Kamga
John Troy
Michal Kordyzon
William G White

IBM Z

In partnership with
IBM **Academy of Technology**

IBM

Redbooks

International Technical Support Organization

IBM z14 Technical Guide

July 2017

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

First Edition (July 2017)

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

This document was created or updated on August 21, 2017.

© Copyright International Business Machines Corporation 2017. All rights reserved.
Note to U.S. Government Users Restricted Rights -- Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

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.

© Copyright IBM Corp. 2017. All rights reserved.

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®
Bluemix®
CICS®
Cognitive Era™
Cognos®
DB2®
DB2 Connect™
developerWorks®
Distributed Relational Database Architecture™
DS8000®
ECKD™
FICON®
FlashCopy®
GDPS®
Geographically Dispersed Parallel Sysplex™
Global Technology Services®
HiperSockets™
HyperSwap®
IA®
IBM®
IBM LinuxONE™
IBM z®
IBM z Systems®
IBM z13®
IBM z13s™
IMS™
Language Environment®
MVS™
OMEGAMON®
Parallel Sysplex®
Passport Advantage®
PowerPC®
PR/SM™
Processor Resource/Systems Manager™
RACF®
Redbooks®
Redbooks (logo) ®
Resource Link®
Resource Measurement Facility™
RMF™
S/390®
System Storage®
System z®
System z10®
System z9®
VIA®
VTAM®
WebSphere®
z Systems®
z/Architecture®
z/OS®
z/VM®
z/VSE®
z10™
z13®
z13s™
z9®
zEnterprise®
IBM Z®

The following terms are trademarks of other companies:

Evolution, and Workforcebydesign are trademarks or registered trademarks of Kenexa, an IBM Company.

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

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

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

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

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

Contents

Notices iii
Trademarks iv

Preface xv
Authors xv
Now you can become a published author, too! xvii
Comments welcome xvii
Stay connected to IBM Redbooks xvii

Chapter 1. Introducing the IBM z14 1
1.1 Design considerations for the IBM z14 2
1.2 z14 server highlights 3
1.2.1 Processor and memory 4
1.2.2 Capacity and performance 4
1.2.3 Virtualization 6
1.2.4 I/O subsystem and I/O features 8
1.2.5 Reliability, availability, and serviceability design 10
1.3 z14 server technical overview 11
1.3.1 Models 11
1.3.2 Model upgrade paths 12
1.3.3 Frames 13
1.3.4 CPC drawer 13
1.3.5 I/O connectivity: PCIe Generation 3 17
1.3.6 I/O subsystem 17
1.3.7 I/O and special purpose features in the PCIe I/O drawer 18
1.3.8 Storage connectivity 19
1.3.9 Network Connectivity 20
1.3.10 Coupling and Server Time Protocol connectivity 22
1.3.11 Cryptography 25
1.3.12 zEDC Express 28
1.4 Reliability, availability, and serviceability 28
1.5 Hardware Management Consoles and Support Elements 29
1.6 Operating systems 29
1.6.1 Supported operating systems 29

© Copyright IBM Corp. 2017. All rights reserved.

1.6.2 IBM Compilers 31

Chapter 2. Central processor complex hardware components 33
2.1 Frames and drawers 34
2.1.1 The A Frame 35
2.1.2 Z Frame 36
2.1.3 z14 Cover (door) design 36
2.1.4 Top exit I/O cabling 37
2.1.5 PCIe I/O drawer 37
2.2 CPC drawer 39
2.2.1 CPC drawer interconnect topology 42
2.2.2 Oscillator 42
2.2.3 System control 44
2.2.4 CPC drawer power 44
2.3 Single chip modules 45
2.3.1 Processor unit chip 47
2.3.2 Processor unit (core) 49
2.3.3 PU characterization 49
2.3.4 System Controller (SC) chip 50
2.3.5 Cache level structure 51
2.4 Memory 52
2.4.1 Memory subsystem topology 53
2.4.2 Redundant array of independent memory 54
2.4.3 Memory configurations 55
2.4.4 Memory upgrades 60
2.4.5 Drawer replacement and memory 60
2.4.6 Virtual Flash Memory 60
2.4.7 Flexible Memory Option 61
2.4.8 Pre-planned memory 61
2.5 Reliability, availability, and serviceability 62
2.5.1 RAS in the CPC memory subsystem 63
2.5.2 General z14 RAS features 63
2.6 Connectivity 64
2.6.1 Redundant I/O interconnect (RII) 65
2.6.2 Enhanced drawer availability 67
2.6.3 CPC drawer upgrade 67
2.7 Model configurations 67
2.7.1 Upgrades 69
2.7.2 Concurrent PU conversions 70
2.7.3 Model capacity identifier 70
2.7.4 Model capacity identifier and MSU value 71
2.7.5 Capacity Backup 71
2.7.6 On/Off Capacity on Demand and CPs 74
2.8 Power and cooling 75
2.8.1 Power and cooling 75
2.8.2 High Voltage Direct Current power feature 76
2.8.3 Internal Battery Feature 76
2.8.4 Power estimation tool 76
2.8.5 Cooling 77
2.8.6 Radiator Unit 78
2.8.7 Water-cooling unit 80
2.9 Summary of z14 structure 83

Chapter 3. Central processor complex system design 85
3.1 Overview 86
3.2 Design highlights 86
3.3 CPC drawer design 87
3.3.1 Cache levels and memory structure 88
3.3.2 CPC drawer interconnect topology 92
3.4 Processor unit design 93
3.4.1 Single-instruction multiple-data 95
3.4.2 Out-of-order execution 96
3.4.3 Superscalar processor 99
3.4.4 Compression and cryptography accelerators on a chip 99
3.4.5 Decimal floating point accelerator 101
3.4.6 IEEE floating point 102
3.4.7 Processor error detection and recovery 102
3.4.8 Branch prediction 102
3.4.9 Wild branch 103
3.4.10 Translation lookaside buffer 103
3.4.11 Instruction fetching, decoding, and grouping 104
3.4.12 Extended Translation Facility 104
3.4.13 Instruction set extensions 104
3.4.14 Transactional Execution 105
3.4.15 Runtime Instrumentation 105
3.5 Processor unit functions 105
3.5.1 Overview 105
3.5.2 Central processors 107
3.5.3 Integrated Facility for Linux 107
3.5.4 Internal Coupling Facility 108
3.5.5 IBM Z Integrated Information Processor 110
3.5.6 System assist processors 114
3.5.7 Reserved processors 114
3.5.8 Integrated firmware processor 115
3.5.9 Processor unit assignment 115
3.5.10 Sparing rules 116
3.6 Memory design 117
3.6.1 Overview 117
3.6.2 Main storage 119
3.6.3 Hardware system area 120
3.6.4 Virtual Flash Memory 120
3.7 Logical partitioning 120
3.7.1 Overview 120
3.7.2 Storage operations 127
3.7.3 Reserved storage 129
3.7.4 Logical partition storage granularity 130
3.7.5 LPAR dynamic storage reconfiguration 130
3.8 Intelligent Resource Director (IRD) 130
3.9 Clustering technology 132
3.9.1 Coupling Facility Control Code 133
3.9.2 Coupling Thin Interrupts 135
3.9.3 Dynamic CF dispatching 135
3.10 Virtual Flash Memory 137
3.10.1 What is IBM Z Virtual Flash Memory 137
3.10.2 Virtual Flash Memory Feature 137
3.10.3 Virtual Flash Memory administration 137

Chapter 4. Central processor complex I/O system structure 139
4.1 Introduction to I/O infrastructure 140
4.1.1 I/O infrastructure 140
4.1.2 PCIe Generation 3 140
4.2 I/O system overview 141
4.2.1 Characteristics 141
4.2.2 Summary of supported I/O features 142
4.3 PCIe I/O drawer 143
4.4 PCIe I/O drawer offerings 146
4.5 Fanouts 147
4.5.1 PCIe Generation 3 fanout (FC #0173) 148
4.5.2 Integrated Coupling Adapter (FC #0172) 148
4.5.3 HCA3-O (12x IFB) fanout (FC #0171). 149
4.5.4 HCA3-O LR (1x IFB) fanout (FC #0170) 150
4.5.5 Fanout considerations. 151
4.6 I/O features (cards) 153
4.6.1 I/O feature card ordering information 153
4.6.2 Physical channel (PCHID) report 154
4.7 Connectivity. 156
4.7.1 I/O feature support and configuration rules. 157
4.7.2 Storage connectivity 159
4.7.3 Network Connectivity 166
4.7.4 Parallel Sysplex connectivity. 175
4.8 Cryptographic functions 183
4.8.1 CPACF functions (FC #3863) 183
4.8.2 Crypto Express6S feature (FC #0893) 183
4.8.3 Crypto Express5S feature (FC #0890) 183
4.9 Integrated firmware processor 184
4.10 zEDC Express 184

Chapter 5. Central processor complex channel subsystem 185
5.1 Channel subsystem. 186
5.1.1 Multiple logical channel subsystems (LCSSs) 187
5.1.2 Multiple subchannel sets. 188
5.1.3 Channel path spanning 191
5.2 I/O configuration management 193
5.3 Channel subsystem summary 194

Chapter 6. Cryptographic features 195
6.1 Cryptography enhancements on IBM z14 servers 196
6.2 Short introduction to cryptography 196
6.2.1 Modern cryptography 197
6.2.2 Kerckhoffs' principle 197
6.2.3 Keys 198
6.2.4 Algorithms 199
6.3 Cryptography on IBM z14 servers. 200
6.4 CP Assist for Cryptographic Functions 203
6.4.1 Cryptographic synchronous functions. 204
6.4.2 CPACF protected key 205
6.5 Crypto Express6S 207
6.5.1 Cryptographic asynchronous functions. 209
6.5.2 Crypto Express6S as a CCA coprocessor 210
6.5.3 Crypto Express6S as an EP11 coprocessor. 214
6.5.4 Crypto Express6S as an accelerator. 215

6.5.5 Management of Crypto Express6S . . . 215
6.6 TKE workstation . . . 218
6.6.1 Logical partition, TKE host, and TKE target . . . 220
6.6.2 Optional smart card reader . . . 221
6.6.3 TKE hardware support and migration information. . . . 221
6.7 Cryptographic functions comparison . . . 222
6.8 Cryptographic Operating system support . . . 224

Chapter 7. Operating System Support . . . 227
7.1 Operating systems summary . . . 228
7.2 Support by operating system . . . 229
7.2.1 z/OS . . . 229
7.2.2 z/VM . . . 229
7.2.3 z/VSE . . . 230
7.2.4 z/TPF . . . 230
7.2.5 Linux on z Systems . . . 231
7.2.6 KVM for IBM Z . . . 231
7.3 z14 features and function support summary . . . 232
7.3.1 Supported CPC functions . . . 233
7.3.2 Coupling and clustering . . . 235
7.3.3 Network connectivity . . . 239
7.3.4 Cryptographic functions . . . 243
7.3.5 Special purpose features . . . 245
7.4 Support by features and functions . . . 245
7.4.1 LPAR Configuration and Management . . . 246
7.4.2 Base CPC Features and Functions. . . . 249
7.4.3 Coupling and clustering Features and Functions . . . 258
7.4.4 Storage Connectivity Related Features and Functions. . . . 262
7.4.5 Networking Features and Functions . . . 274
7.4.6 Cryptography Features and Functions Support . . . 284
7.4.7 Special-purpose Features and Functions . . . 288
7.5 z/OS migration considerations . . . 289
7.5.1 General guidelines . . . 289
7.5.2 Hardware PSP Buckets and Fix Categories . . . 290
7.5.3 Coupling links . . . 290
7.5.4 z/OS XL C/C++ considerations . . . 291
7.5.5 z/OS V2.3 Preview . . . 291
7.6 z/VM migration considerations . . . 292
7.6.1 ESA/390-compatibility mode for guests: . . . 293
7.6.2 Capacity . . . 293
7.7 z/VSE migration considerations . . . 293
7.8 Software licensing . . . 294
7.9 References . . . 297

Chapter 8. System upgrades . . . 299
8.1 Upgrade types . . . 300
8.1.1 Overview of upgrade types . . . 300
8.1.2 Terminology that is related to CoD for z14 systems . . . 301
8.1.3 Permanent upgrades . . . 303
8.1.4 Temporary upgrades. . . . 304
8.2 Concurrent upgrades . . . 304
8.2.1 Model upgrades . . . 305
8.2.2 Customer Initiated Upgrade facility . . . 307

8.2.3 Summary of concurrent upgrade functions . . . 311
8.3 Miscellaneous equipment specification upgrades . . . 312
8.3.1 MES upgrade for processors . . . 313
8.3.2 MES upgrades for memory . . . 314
8.3.3 MES upgrades for I/O . . . 315
8.3.4 Feature on Demand . . . 316
8.3.5 Summary of plan-ahead features . . . 316
8.4 Permanent upgrade through the CIU facility . . . 317
8.4.1 Ordering . . . 319
8.4.2 Retrieval and activation . . . 320
8.5 On/Off Capacity on Demand . . . 321
8.5.1 Overview . . . 321
8.5.2 Capacity Provisioning Manager . . . 322
8.5.3 Ordering . . . 323
8.5.4 On/Off CoD testing . . . 326
8.5.5 Activation and deactivation . . . 328
8.5.6 Termination . . . 328
8.5.7 z/OS capacity provisioning . . . 329
8.6 Capacity for Planned Event . . . 333
8.7 Capacity Backup . . . 334
8.7.1 Ordering . . . 334
8.7.2 CBU activation and deactivation . . . 336
8.7.3 Automatic CBU enablement for GDPS . . . 337
8.8 Nondisruptive upgrades . . . 337
8.8.1 Components . . . 338
8.8.2 Concurrent upgrade considerations . . . 339
8.9 Summary of Capacity on Demand offerings . . . 343

Chapter 9. Reliability, availability, and serviceability . . . 345
9.1 The RAS strategy . . . 346
9.2 Technology change . . . 346
9.3 Structure change . . . 349
9.4 Reducing complexity . . . 349
9.5 Reducing touches . . . 349
9.6 z14 availability characteristics . . . 350
9.7 z14 RAS functions . . . 354
9.7.1 Scheduled outages . . . 355
9.7.2 Unscheduled outages . . . 356
9.8 z14 enhanced drawer availability . . . 358
9.8.1 EDA planning considerations . . . 358
9.8.2 Enhanced drawer availability processing . . . 360
9.9 z14 Enhanced Driver Maintenance . . . 365
9.9.1 Resource Group and native PCIe MCLs . . . 366
9.10 RAS capability for the Hardware Management Console and Support Element . . . 368

Chapter 10. Environmental requirements . . . 371
10.1 Power and cooling . . . 372
10.1.1 Rear cover design for vectored air output . . . 372
10.1.2 Power requirements and consumption . . . 372
10.1.3 Cooling requirements . . . 375
10.1.4 Internal Battery Feature (IBF) . . . 378
10.1.5 Emergency power-off switch . . . 379
10.2 Physical specifications . . . 379

10.3 Physical planning 380
10.3.1 Raised floor or non-raised floor 380
10.3.2 Top Exit Power option 381
10.3.3 Top Exit I/O Cabling feature 382
10.3.4 Weight distribution plate 383
10.3.5 Bolt-down kit for raised floor 383
10.3.6 Nonraised floor frame tie-down kit 384
10.3.7 Service clearance areas 384
10.4 Energy management 385
10.4.1 Environmental monitoring 385

Chapter 11. Hardware Management Console and Support Elements 389
11.1 Introduction to the HMC and SE 390
11.2 HMC and SE changes and new features 390
11.2.1 Driver Level 32 HMC and SE changes and new features 390
11.2.2 Enhanced Computing and z14 HMC 392
11.2.3 Rack-mounted HMC 393
11.2.4 New Support Elements 394
11.2.5 New backup options for HMCs and primary SEs 395
11.2.6 SE driver support with the HMC driver 397
11.2.7 HMC feature codes 398
11.2.8 User Interface 399
11.2.9 Customize Product Engineering Access - Best practice 399
11.3 HMC and SE connectivity 399
11.3.1 Network planning for the HMC and SE 401
11.3.2 Hardware prerequisite changes 404
11.3.3 TCP/IP Version 6 on the HMC and SE 404
11.3.4 OSA Support Facility changes 404
11.3.5 Assigning addresses to the HMC and SE 405
11.3.6 HMC Multi-factor authentication 406
11.4 Remote Support Facility 406
11.4.1 Security characteristics 407
11.4.2 RSF connections to IBM and Enhanced IBM Service Support System 407
11.4.3 HMC and SE remote operations 408
11.5 HMC and SE key capabilities 409
11.5.1 Central processor complex management 409
11.5.2 Logical partition management 410
11.5.3 Operating system communication 411
11.5.4 HMC and SE microcode 412
11.5.5 Monitoring 415
11.5.6 Capacity on demand support 417
11.5.7 Server Time Protocol support 418
11.5.8 NTP client and server support on the HMC 420
11.5.9 Security and user ID management 421
11.5.10 System Input/Output Configuration Analyzer on the SE and HMC 423
11.5.11 Automated operations 423
11.5.12 Cryptographic support 424
11.5.13 Installation support for z/VM using the HMC 425
11.5.14 Dynamic Partition Manager 425

Chapter 12. Performance 429
12.1 IBM z14 performance characteristics 430
12.2 LSPR workload suite 431

12.3 Fundamental components of workload performance 431
12.3.1 Instruction path length. 431
12.3.2 Instruction complexity 432
12.3.3 Memory hierarchy and memory nest. 432
12.4 Relative Nest Intensity 433
12.5 LSPR workload categories based on relative nest intensity 435
12.6 Relating production workloads to LSPR workloads 435
12.7 Workload performance variation 436
12.7.1 Main performance improvement drivers with z14 servers. 437

Appendix A. IBM Secure Service Container Infrastructure 441
A.1 What is SSC? 442
A.2 Why SSC? 443
A.3 IBM Z servers and SSC 443

Appendix B. Channel options 447

Appendix C. Native Peripheral Component Interconnect Express (PCIe) 451
C.1 Design of native PCIe adapter management 452
C.1.1 Native PCIe adapter 452
C.1.2 Integrated firmware processor 452
C.1.3 Resource groups 453
C.1.4 Management tasks 453
C.2 Native PCIe feature plugging rules 454
C.3 Native PCIe feature definitions 454
C.3.1 FUNCTION identifier 456
C.3.2 Virtual function number 456
C.3.3 Physical network identifier 456

Appendix D. Shared Memory Communications 457
D.1 Shared Memory Communications overview 458
D.2 Shared Memory Communication over RDMA. 458
D.2.1 RDMA technology overview 458
D.2.2 Shared Memory Communications over RDMA. 459
D.2.3 Single Root I/O virtualization (SR-IOV) 460
D.2.4 Hardware 461
D.2.5 10GbE RoCE Express/Express2 features 461
D.2.6 10GbE RoCE Express/Express2 configuration example 463
D.2.7 Hardware configuration definitions 464
D.2.8 Software exploitation of SMC-R 466
D.2.9 SMC-R support overview 466
D.2.10 SMC-R use cases for z/OS to z/OS 467
D.2.11 Enabling SMC-R support in z/OS Communications Server 469
D.3 Shared Memory Communications - Direct Memory Access 470
D.3.1 Internal Shared Memory technology overview 471
D.3.2 SMC-D over Internal Shared Memory 471
D.3.3 Internal Shared Memory introduction 473
D.3.4 Virtual PCI Function (vPCI Adapter). 473
D.3.5 Planning considerations 476
D.3.6 Hardware configuration definitions 476
D.3.7 Sample IOCP FUNCTION statements 477
D.3.8 Software exploitation of ISM. 478
D.3.9 SMC-D over ISM prerequisites. 478
D.3.10 Enabling SMC-D support in z/OS Communications Server 479

D.3.11 SMC-D support overview 480

Appendix E. IBM Dynamic Partition Manager 483
E.1 Introduction to IBM Dynamic Partition Manager 484
E.2 Reasons to use DPM 484
E.3 IBM Z servers and DPM 485
E.4 Setting up the DPM environment 486
E.4.1 Defining partitions in DPM mode 489
E.4.2 Summary 492

Appendix F. IBM zEnterprise Data Compression Express 493
F.1 Overview 494
F.2 zEDC Express 494
F.3 Software support 495
F.3.1 IBM Z Batch Network Analyzer (zBNA) 496

Related publications 497
IBM Redbooks 497
Other publications 497
Online resources 497
Help from IBM 498

Preface

This IBM® Redbooks® publication describes the new member of the IBM Z family, IBM z14. IBM z14 is the trusted enterprise platform for pervasive encryption, integrating data, transactions, and insight. A data-centric infrastructure must always be available with a 99.999% or better availability, have flawless data integrity, and be secured from misuse. It needs to be an integrated infrastructure that can support new applications. It needs to have integrated capabilities that can provide new mobile capabilities with real-time analytics delivered by a secure cloud infrastructure.

IBM z14 servers are designed with improved scalability, performance, security, resiliency, availability, and virtualization. The superscalar design allows z14 servers to deliver a record level of capacity over the prior IBM Z platforms. In its maximum configuration, z14 is powered by up to 170 client characterizable microprocessors (cores) running at 5.2 GHz. This configuration can run more than 146,000 millions of instructions per second (MIPS) and up to 32 TB of client memory. The IBM z14 Model M05 is estimated to provide up to 35% more total system capacity than the IBM z13® Model NE1.

This book provides information about IBM z14 and its functions, features, and associated software support. Greater detail is offered in areas relevant to technical planning. It is intended for systems engineers, consultants, planners, and anyone who wants to understand the IBM Z servers functions and plan for their usage. It is not intended as an introduction to mainframes. Readers are expected to be generally familiar with existing IBM Z 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 17 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. Prior to that, 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 (FICON® and OSA).

Esra Ufacik is an IBM Z Lead Architect in IBM Systems covering Asia Pacific. Throughout her career, she has held various roles in different companies with Z products, such as IBM z/OS® and IBM Parallel Sysplex®. These roles include implementation, migration, SAN, storage, software problem support, solution designing, Linux on z Systems, IT Optimization, TCO assessments, technical authoring, and guest lecturing. In March 2016, she was assigned as IOT lead for Blockchain in IBM Systems division

© Copyright IBM Corp. 2017. All rights reserved.

Bo Xu is a Senior System Service Representative in China. He has more than 20 years of experience with IBM Z platform maintenance support. He has been working in IBM's Global Technology Services® department to provide IBM Z platform support to clients as a local Top Gun, and as second-level support for DS8000® as one of country Top Gun & Skill Owner. His areas of expertise include IBM Z platform hardware, channel connectivity, and IBM DS8000 storage.

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.

Frank Packheiser is a 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. Besides co-authoring several IBM Redbooks publications since 1999, he has been an official ITSO presenter at ITSO workshops for the last four years.

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

Michal Kordyzon is an IBM Z Client Technical Specialist at IBM Poland with 12 years of experience with the IBM Z platform. He also has expertise in LinuxONE, His areas of expertise include mainframe architecture, Linux systems, Hyperledger fabric, Node.js, and Machine Learning algorithms.

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

Patty Driever
Dave Surman
Harry Yudenfriend
Ellen Carbarnes
Diana Henderson
Robert Haimowitz
Anthony Saporito
Garth Godfrey
Darelle Gent
Parwez Hamid
Gary King
Jeff Kubala
Philip Sciuto
Rhonda Sundlo
Barbara Weiler
IBM Poughkeepsie

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

Barbara Sannerud
IBM White Plains

Monika Zimmermann
Carl Mayer
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

1

Introducing the IBM z14

This section covers the basic concepts and design considerations around IBM z14 servers.

This chapter includes the following sections:

- Design considerations for the IBM z14
- z14 server highlights
- z14 server technical overview
- Reliability, availability, and serviceability
- Hardware Management Consoles and Support Elements
- Operating systems

© Copyright IBM Corp. 2017. All rights reserved.

1.1 Design considerations for the IBM z14

Digital technologies are driving major industry shifts, ushering in a new age. In the Cognitive Era™[1], it is either disrupting or being disrupted. Tomorrow's disrupters will be organizations that can converge digital business with a new level of digital intelligence. A key in digital transformation is the ability to accelerate the innovation of new business (Internet of Things, Cognitive, and Virtual Reality), while keeping complexity in check.

Information Technology (IT) today is experiencing a time of exponential growth in data and transaction volumes driven by digital transformation. New dynamics in the market provide opportunities for businesses to grab market share and win. As companies struggle to meet the demands of the new IT, dominated by cloud, analytics, mobile, social, and internet of things, IT leaders are now being challenged to add value by opening their enterprises to new ways of doing business.

In this challenging landscape; businesses must be able to manage, store, protect, and, most important, use the information for gaining competitive advantage. This is creating the demand to apply intelligence and insight to the data to build new services wrapped for a customized user experience. By creating applications that provide intelligent user experiences, companies can provide value to their partners and clients, which ultimately preserves loyalty. And most of all, to succeed organizations must provide their users with the peace of mind that no matter what device is being used, data is protected.

Organizations in every industry and sector must secure that growing data and comply with increasingly intricate regulations. This operational environment increases the pressure on IT to securely deliver services and support on time and on budget. By encrypting as much of your data and transactional pipeline as possible, you can reduce potential data breach risks and financial losses, and comply with complex regulatory mandates.

Cryptography has always been in the DNA of IBM Z family. And as the newest member of that family, the IBM z14, 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. Using simple policy controls, z14 pervasive computing streamlines data protection for mission critical DB2® for z/OS®, IMS™ and virtual storage access method (VSAM) datasets.

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

The new IBM z14 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 170 configurable cores z14 has performance and scaling advantage over prior generations, and 31% more capacity than the 141-way z13. The new FICON Express16S+ delivers an increase in I/O rates and in link bandwidth, and a reduction in single-stream latency, providing the system the ability to absorb large applications and transaction spikes driven by unpredictable mobile and IoT[2] devices. Next-generation SMT on z14 delivers improved virtualization performance to benefit Linux. High-speed connectivity out to the data is critical in achieving exceptional levels of transaction throughput. The new IBM zHyperLink Express introduces disk I/O technology for accessing the IBM DS8880 storage

[1] See: https://www.ibm.com/cognitive/uk-en/outthink/?lnk=hm
[2] Internet of Things

system with very low latency, enabling shorter batch windows and a more resilient I/O infrastructure with predictable and repeatable I/O performance.

With up to 32 TB, z14 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 exploiting the Vector Packed Decimal Facility that allows packed decimal operations to be performed in registers rather than memory, using new fast mathematical computations, compilers, such as Enterprise COBOL for z/OS, V6.2, Enterprise PL/I for z/OS, V5.2, z/OS V2.3 XL C/C++, the COBOL optimizer, Automatic Binary Optimizer for z/OS, V1.3, and Java, are optimized on z14. 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 will deliver additional improvements in throughput per core, which gives a natural boost to z/OS Connect EE, WebSphere® Liberty in CICS®, Spark for z/OS, and IBM Java for Linux on z Systems®.

Smoothly handling the data tsunami requires robust infrastructure designed specifically for high volume data transactions. To take advantage of new unstructured data[3], businesses on IBM z® can leverage the use of application programming interfaces (APIs) that can help with the creation and delivery of innovative new services. Linux on IBM Z, which is optimized for open source software, brings additional value to the platform. Linux on z Systems supports a wealth of new products familiar to application developers, such as Python, Scala, Spark, MongoDB, PostgreSQL, and MariaDB. Access to data previously unavailable, without the need for Extract Transform and Load (ETL) will allow development of intelligent transactions and intuitive business processes.

As your business technology needs evolve to compete in today's digital economy, IBM stands ready to help with intelligent, robust and comprehensive technology solutions. The IBM approach integrates server, software and storage solutions to ensure each member of the stack is designed and optimized to work together. The new IBM z14 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.

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

1.2 z14 server highlights

This section reviews some of the most important features and functions of z14 (Driver 32) servers:

- Processor and memory
- Capacity and performance
- Virtualization
- I/O subsystem and I/O features
- Reliability, availability, and serviceability design

[3] accounts for 80% of all data generated today and is expected to grow to over 93% by 2020.

1.2.1 Processor and memory

IBM continues its technology leadership with the z14 server. The z14 server is built using the IBM modular multi-drawer design that supports 1 - 4 processor drawers per CPC. Each processor drawer contains either five or six CP (Central Processor) SCM (Single-chip modules) and one SC (Storage Controller) SCM. Both SCMs are redesigned using 14nm FINFET SOI technology[4]. Each CP SCM has ten processor units (PUs, or cores). In addition to SCMs, CPC drawers host memory DIMMs, connectors for I/O, oscillator interface, Flexible Service Processors (FSPs), and cooling manifolds.

The superscalar processor implements second generation SMT[5] now enabled for SAPs (System Assist Processors) also, redesigned caches and TLB (translation lookaside buffer), optimized pipeline, better branch prediction, an expanded instruction set with Vector Packed Decimal Facility, Guarded Storage Facility, Vector Facility enhancements, Semaphore Assist Facility, Order Preserving Compression and Entropy Encoding for Co-processor Compression, for better performance in several different areas.

Depending on the model, the z14 server can support from 256 GB to a maximum of 32TB of usable memory, with up to 8 TB of usable memory per CPC drawer. In addition, a fixed amount of 192 GB is reserved for the hardware system area (HSA) and is not part of customer-purchased memory. Memory is implemented as a redundant array of independent memory (RAIM) and uses extra physical memory as spare memory. The RAIM function accounts for 20% of the physical installed memory in each CPC drawer.

New with z14, VFM (Virtual Flash Memory) feature is offered from the main memory capacity in 1.5 TB units and replaces the Flash Express adapters which were available on the zEC12 and z13. VFM provides much simpler management and better performance by eliminating the I/O the adapters located in the PCIe drawers. VFM does not require any application changes when moving from IBM Flash Express.

1.2.2 Capacity and performance

The z14 server provides increased processing and enhanced I/O capabilities over its predecessor, the z13 system. This capacity is achieved by increasing the performance of the individual PUs (processor units), increasing the number of PUs per system, redesigning the system cache, increasing the amount of memory, and introducing new and I/O technologies. The increased performance and the total system capacity available, with possible energy savings, allow consolidating diverse applications on a single platform, with significant financial savings. The introduction of new technologies and an expanded and enhanced instruction set ensure that the z14 server is a high performance, reliable, and rich-security platform. The z14 server is designed to maximize resource exploitation and utilization, and allows you to integrate and consolidate applications and data across the enterprise IT infrastructure.

z14 servers are offered in five models, with 1 - 170 configurable PUs. Models M01, M02, M03, and M04 have up to 41 PUs per CPC drawer, while the high-capacity model (the M05) has four processor (CPC) drawers with 49 PU per drawer. Model M05 is estimated to provide up to 35% more total system capacity than the z13 Model NE1, with the same amount of memory and power requirements. With up to 32TB of main storage and enhanced SMT, the performance of the z14 processors deliver considerable improvement. Uniprocessor performance has also increased significantly. A z14 Model 701 offers average performance improvements of 10%[6] over the z13 Model 701.

[4] FINFET is the industry solution, SOI is the IBM solution for SER
[5] Simultaneous Multi-Threading - two threads per core
[6] Observed performance increases vary depending on the workload types

The IFL, zIIP processor units on the z14 server can be configured to run two simultaneous threads per clock cycle in a single processor (SMT), increasing the capacity of these processors with 25% in average[7] over processors running single thread. SMT is also enabled by default on SAPs.

The z14 server expands the subcapacity settings, offering three subcapacity levels (via models 4xx, 5xx and 6xx) for up to 33 processors that are characterized as CPs (compared to up to 30 for z13). This configuration gives a total of 269 distinct capacity settings. The z14 servers deliver scalability and granularity to meet the needs of medium-sized enterprises, while also satisfying the requirements of large enterprises that have demanding, mission-critical transaction and data processing requirements.

This comparison is based on the Large System Performance Reference (LSPR) mixed workload analysis. For a description of performance and workload variation on z14 servers, see Chapter 12, “Performance” on page 429.

z14 servers continue to offer all the specialty engines[8] that are available on z13.

Workload variability

Consult the LSPR when considering performance on z14 servers. The range of performance ratings across the individual LSPR workloads is likely to have a large spread. There is more performance variation of individual logical partitions (LPARs) when an increased number of partitions and more PUs is available. For more information, see Chapter 12, “Performance” on page 429.

For detailed performance information, see the LSPR website:

https://www.ibm.com/servers/resourcelink/lib03060.nsf/pages/lsprindex

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

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

Capacity on demand

Capacity on demand (CoD) enhancements enable clients to have more flexibility in managing and administering their temporary capacity requirements. The z14 server supports the same architectural approach for CoD offerings as the z13 (temporary or permanent). Within the z14 server, 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 either before or after workload execution (pre- or post-paid).

LPAR capping

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

---

[7] Observed performance increases vary depending on the workload types.

[8] Workloads qualified to run on zAAP (System z® Application Assist Processor) engines that were available on IBM Z server generations z990 to zEC12, run on a zIIP processor thus reducing the complexity of the IBM z/Architecture®

1.2.3 Virtualization

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

z14 servers support z/Architecture mode only and can be initialized either in LPAR mode (also known as PR/SM) or DPM (Dynamic Partition Manager) mode.

PR/SM mode

PR/SM is Licensed Internal Code (LIC) that manages and virtualizes all the installed and enabled system resources as a single large symmetric multiprocessor (SMP) system. This virtualization enables full sharing of the installed resources with high security and efficiency. PR/SM supports configuring up to 85 LPARs, each of which has 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 120.

LPAR configurations can be dynamically adjusted to optimize the virtual servers’ workloads. z14 servers provide improvements to the PR/SM HiperDispatch function. HiperDispatch provides alignment of logical processors to physical processors that ultimately improves cache utilization, minimizes inter-CPC drawer communication, and optimizes operating system work dispatching; all combined resulting increased throughput. For more information, see “HiperDispatch” on page 91.

HiperSockets

z14 servers support the definition of up to 32 IBM HiperSockets™. *HiperSockets* provide for memory to memory communication across LPARs without the need for any I/O adapters, and have virtual LAN (VLAN) capability.

Dynamic Partition Manager mode

DPM is an administrative mode (front-end to PR/SM), introduced for Linux only systems for IBM z14, IBM z13, IBM z13s™, and IBM LinuxONE™ servers. A system can be configured either in DPM mode or in PR/SM mode (POR is required to switch modes). DPM supports the following:

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

LPAR Modes on z14

Listed below are PR/SM LPAR *modes* with corresponding operating systems and firmware appliances supported:

- General
  - z/OS
  - z/VM®
  - z/VSE®
  - z/TPF
  - Linux on z Systems
- Coupling Facility
  - CFCC (Coupling Facility Control Code)
- Linux only
  - Linux on z Systems
  - z/VM
- z/VM
  - z/VM

- Secure Service Container
  - VNA (z/VSE Network Appliance)
  - IBM HSBN (High Security Business Network)[9]

For DPM, LPAR modes are as follows:

- z/VM V6R4
- KVM for IBM Z
- Linux on z Systems

See “Dynamic Partition Manager mode” on page 6 for more information.

IBM Z servers also offer other virtual appliance-based solutions, and support additional hypervisors and containerization like:

- GDPS® Virtual Appliance
- KVM for IBM Z
- Docker Enterprise Edition for Linux on IBM Systems[10]

Coupling Facility mode logical partition

Parallel Sysplex® is a synergy between hardware and software - a highly advanced technology for clustering 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 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 ICFs.

The z/VM-mode LPAR

z14 servers support 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 V6R3 or later, and allows z/VM to use a wider variety of specialty processors in a single LPAR; increasing flexibility and simplifying system management.

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

SSC (Secure Service Container)

IBM Secure Service Container 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 helping to keep them safe from insider threats as well as external cyber criminals. SSC offers the following benefits:

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

For more information, refer to Appendix A, “IBM Secure Service Container Infrastructure” on page 441.

[9] IBM HSBN is a cloud service plan available on IBM Bluemix® for Blockchain.
[10] See https://www-03.ibm.com/press/us/en/pressrelease/52096.wss

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. The appliance uses the Secure Service Container infrastructure introduced on z13 and z13s servers. Compared to a TCP/IP stack in z/VSE, this 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 introduced on z114 and z196 servers. VIA 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 running in an LPAR. The z/VSE Network Appliance is provided as a downloadable package and it can then be deployed with the SSC Installer and Loader.

In summary, 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 z Systems.

IBM zAware: With *Announcement Letter 916-201* dated November 1, 2016; IBM has changed how IBM System z Advanced Workload Analysis Reporter (IBM zAware) is delivered. IBM zAware has been available as a firmware feature on zEC12, zBC12, z13 and z13s is now offered as a software feature with IBM Operations Analytics for Z.

IBM Operations Analytics for Z brings new capabilities and functions to the product, or applies maintenance, based on user schedules, not tied to IBM Z firmware updates. Integration with IBM Operational Analytics for Z Problem Insights dashboard eliminates the need for tedious searching through volumes of operational data, putting key operational issues at your fingertips. New functions, such as proactive outage avoidance with e-mail alerts, improve users' ability to respond to identified anomalies.

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 HyperSwap®
- Single point of control

1.2.4 I/O subsystem and I/O features

The z14 server supports both PCIe and InfiniBand I/O infrastructure. PCIe features are installed in PCIe I/O drawers. Up to five PCIe I/O drawers per z14 server are supported, providing space for up to 160 PCIe I/O features. I/O drawers[11] are not supported on z14 servers and cannot be carried forward during an upgrade from either a z13 or zEC12 server.

[11] I/O drawers were introduced with the IBM z10™ BC.

Statement of Direction[a]**:** IBM z14 will be the last z Systems and IBM Z server to support HCA3-O and HCA3-O LR adapters: z14 will be last z Systems and IBM Z server to support HCA3-O fanout for 12x IFB (#0171) and HCA3-O LR fanout for 1x IFB (#0170). As announced previously, z13s is the last mid-range IBM Z server to support these adapters. Enterprises should begin migrating from HCA3-O and HCA3-O LR adapters to ICA SR and/or Coupling Express Long Reach (CE LR) adapters on z14, 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 requiring a DWDM, enterprises will need to determine their desired DWDM vendor's plan to qualify the planned replacement long-range coupling link.

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.

For a four CPC drawer system, up to 40 PCIe and 16 InfiniBand fanout slots can be configured for data communications between the CPC drawers and the I/O infrastructure, and for coupling. The multiple channel subsystem (CSS) architecture allows up to six CSSs, each with 256 channels.

For I/O constraint relief, four subchannel sets are available per CSS, allowing access to a larger number of logical volumes. The fourth subchannel set allows extending the amount of addressable external storage for PAVs (Parallel Access Volumes), PPRC (Peer-to-Peer Remote Copy) secondary devices, and IBM FlashCopy® devices. z14 supports IPL (Initial Program Load) from subchannel set 1 (SS1), subchannel set 2 (SS2), or subchannel set 3 (SS3), in addition to subchannel set 0 (SS0). For more information, see "IPL from an alternate subchannel set" on page 190.

The system I/O buses take advantage of the Peripheral Component Interconnect Express (PCIe) technology and the InfiniBand technology, which are also used in coupling links.

z14 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)
  - IBM zHyperLink Express
- Network Connectivity:
  - Open Systems Adapter (OSA):
    - OSA-Express6S 10 GbE 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 1000BASE-T Ethernet (carry forward only)
  - IBM HiperSockets
  - Shared Memory Communication - Remote Direct Memory Access (SMC-R):
    - 10 GbE RoCE (RDMA over Converged Ethernet) Express2

    - 10 GbE RoCE Express (carry forward only)
  - Shared Memory Communication - Direct Memory Access (SMC-D) through Internal Shared Memory (ISM)
- Coupling and Server Time Protocol connectivity:
  - Internal Coupling (IC) links
  - Integrated Coupling Adapter Short Reach (ICA SR)
  - Coupling Express Long Reach (CE LR)
  - HCA3-O, 12x Parallel Sysplex InfiniBand (IFB) coupling links
  - HCA3-O, 1x Parallel Sysplex InfiniBand (IFB) coupling links
- Cryptography:
  - Crypto Express6S
  - Crypto Express5S (carry forward only)
  - Regional Crypto Enablement
- zEnterprise® Data Compression (zEDC) Express features, which are installed in the PCIe I/O drawers (new build and carry forward).

1.2.5 Reliability, availability, and serviceability design

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 something happens. 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. This is recognized by various IT analysts like ITIC[12] and IDC[13]. Comprehensive, multi-layered strategy includes:

- Error Prevention
- Error Detection and Correction
- Error Recovery

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

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

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.

[12] ITIC Global Server Hardware, Server OS Reliability Report, available at http://itic-corp.com/category/servers/
[13] Quantifying the Business Value of IBM Z, available at http://idcdocserv.com/download/258434.pdf

1.3 z14 server technical overview

This section briefly reviews the major elements of z14 servers:

- Models
- Model upgrade paths
- Frames
- CPC drawer
- I/O connectivity: PCIe Generation 3
- I/O subsystem
- Coupling and Server Time Protocol connectivity
- Cryptography
- zEDC Express
- Reliability, availability, and serviceability

1.3.1 Models

The IBM z14 server has a machine type of 3906. Five models are offered: M01, M02, M03, M04, and M05. The model name indicates the number of CPC drawers for models M01 thru M04. Model M05 also has four CPC drawers but with more PU per drawer than models M01 thru M04. A PU is the generic term for the IBM z/Architecture processor unit (processor core) on the CP SCM.

On z14 servers, some PUs are part of the system base, that is, they are not part of the PUs that can be purchased by clients. They are characterized by default as follows:

- System assist processor (SAP) that is used by the channel subsystem. The number of predefined SAPs depends on the z14 model.
- One integrated firmware processor (IFP). The IFP is used in support of select features, such as zEDC and 10GbE RoCE.
- Two spare PUs that can transparently assume any characterization during a permanent failure of another PU.

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

- Central processor (CP) for general-purpose use.
- Integrated Facility for Linux (IFL) for the use of Linux on z Systems.
- IBM z Integrated Information Processor (zIIP[14]) is designed to help free-up general computing capacity and lower overall total cost of computing for select data and transaction processing workloads.

  zIIPs: At least one CP must be purchased with, or before, a zIIP can be purchased. Clients can purchase up to two zIIPs for each purchased CP (assigned or unassigned) on the system (2:1). However, for migrations from zEC12 with zAAPs, the ratio (CP:zIIP) can go up to 4:1.

- Internal Coupling Facility (ICF) is used by the CFCC.
- Additional SAPs are used by the channel subsystem.

[14] IBM zEnterprise Application Assist Processors (zAAPs) are not available on z14 servers. The zAAP workload is run on zIIPs.

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

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

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

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

1.3.2 Model upgrade paths

Upgrades from Models M01, M02, and M03 to M02, M03, and M04 are concurrent. Any z13 or zEC12 model can be upgraded to z14 models M01, M02, M03, M04 and M05 disruptively. Upgrades from z14 M01, M02, M03, or M04 to z14 model M05 are not supported. M05 is offered only as factory build. Figure 1-1 shows the upgrade paths supported.

Consideration: An air-cooled z14 server cannot be converted to a water-cooled z14 server, and vice versa.

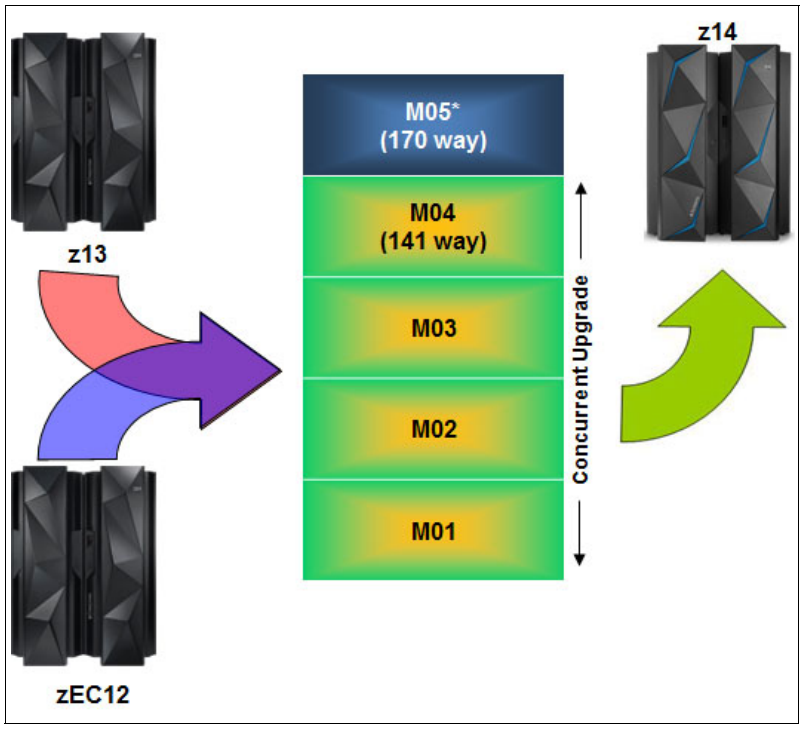


Figure 1-1 z14 upgrades

z13 upgrade to z14

When a z13 is upgraded to a z14, the z13 driver level must be at least 27. Upgrading from z13 to z14 server is disruptive.

zEC12 upgrade to z14

When a zEC12 is upgraded to a z14, the zEC12 must be at least at Driver level 15. Upgrading from zEC12 to z14 servers is disruptive. zBX Model 003 cannot be upgraded to zBX Model 004 during an upgrade from zEC12 to z14.

The following processes are not supported:

- Downgrades within the z14 models
- Upgrade from a z13s or zBC12 to z14 servers
- Upgrades from z196 or earlier systems
- Attachment of a zBX Model 002 or model 003 to a z14 server

1.3.3 Frames

z14 servers have two frames that are bolted together and are known as the A frame and the Z frame. The frames contain the following components:

- Up to four CPC drawers in Frame A
- Up to five PCIe I/O drawers (up to one in Frame A and up to four in Frame Z) which hold I/O features and special purpose features
- Power supplies in Frame Z
- Optional Internal Battery Feature (IBF)
- Cooling units for either air or water cooling
- Two System Control Hubs (SCHs) to interconnect the CPC components through Ethernet
- Two 1U rack-mounted Support Elements (mounted in A frame) with their keyboards, pointing devices, and displays mounted on a tray in the Z frame.

1.3.4 CPC drawer

Up to four CPC drawers are installed in frame A of z14 server. Each CPC drawer houses the SCMs, memory, and I/O interconnects.

Single Chip Module (SCM) technology

z14 servers are built on the superscalar microprocessor architecture of its predecessor, and provide various enhancements over the z13. Each CPC drawer has two logical CP clusters and one System Controller (SC) SCM. There are two CPC drawer sizes in z14, depending on the number of CP SCMs. The z14 model M05 has six CP SCMs containing 49 active cores. All other z14 models have five CP SCMs containing 41 active cores. The CP SCM has ten cores by design, with seven, eight, nine or ten active cores, which can be characterized as CPs, IFLs, ICFs, zIIPs, SAPs, or IFPs.

The SCM provides a significant increase in system scalability and an extra opportunity for server consolidation. All CPC drawers are fully interconnected by using high-speed communication links, through the L4 cache (in the SC SCM). This configuration allows the z14 server to be controlled by the PR/SM facility as a memory-coherent and cache-coherent SMP system.

The PU configuration includes two designated spare PUs per CPC and a variable number of SAPs. The SAPs scale with the number of CPC drawers that are installed in the server. For example, there are five standard SAPs with one CPC drawer installed, and up to 23 when four CPC drawer are installed. In addition, one PU is used as an IFP and is not available for client use. The remaining PUs can be characterized as CPs, IFL processors, zIIPs, ICF processors, or extra SAPs. For z14 the SAPs have Simultaneous Multi-Threading enabled by default (cannot be changed).

The SCMs are cooled by a cold plate that is connected to an internal water cooling loop. In an air-cooled system, the radiator units (RUs) exchange the heat from the internal water loop with air. The RU has N+1 availability for pumps and blowers.

The z14 server offers also a water-cooling option for increased system and data center energy efficiency. The water cooling units (WCUs) are fully redundant in an N+1 arrangement.

Processor features

The processor core operates at 5.2 GHz. Depending on the z14 model, 41 - 196 active PUs are available on 1 - 4 CPC drawers.

Each core on the CP SCM includes an enhanced dedicated coprocessor for data compression (CMPSC)[15] and cryptographic functions (Central Processor Assist for Cryptographic Functions - CPACF)[16]. Having standard clear key cryptographic coprocessors that are integrated with the processor provides high-speed cryptography for protecting data.

CPACF on z14 is improved to support pervasive encryption which provides an envelope of protection around the data that resides on IBM Z servers. CPACF encryption rates for like modes and data sizes on z14 are up to six times faster[17] than z13. Enhancements include support of SHA-3 and SHAKE hashing algorithms, True Random-number generation (TRNG), addition of a new hardware instruction to perform Galois Counter Mode (GCM) encryption.

Hardware data compression can play a significant role in improving performance and saving costs over performing compression in software. z14 is the fourth IBM Z generation having CMPSC - the on-chip compression co-processor. New compression ratio with Entropy Encoding using Huffman coding and Order Preserving Compression in z14 results in fewer CPU cycles to enable further compression of data including DB2 indices, improving memory, transfer, and disk efficiency.

The zEDC Express feature complements the functionality of the coprocessor (CPACF). Their functions are not interchangeable.

The micro-architecture of the core has been improved in several ways in order to increase parallelism and pipeline efficiency. z14 cores have 2x more on-chip cache, compared to z13 per core, to minimize memory waits while maximizing the throughput of concurrent workloads, making it perfectly optimized for data serving. z14 comes with a new translation lookaside buffer (TLB2) design with four hardware-implemented translation engines reducing latency when compared with one pico-coded engine on z13. Pipeline optimization presents enhancements including improved instruction delivery, faster branch wake-up, reduced execution latency, improved OSC (Operand Store Compare) prediction.

Processor RAS

In the unlikely case of a permanent core failure, each core can be individually replaced by one of the available spares. Core sparing is transparent to the operating system and applications.

[15] CMPSC - Compression Coprocesor
[16] Feature code (FC) 3863 must be ordered to enable CPACF. This feature code is available for no extra fee.
[17] Based on preliminary internal IBM lab measurements on a standalone dedicated system in a controlled environment and compared to the z13. Results may vary.

Vector and Floating point Unit Enhancements

Each core has two Vector and Floating point Units (VFU) containing sub-units like Single instruction multiple data (SIMD), Binary floating-point unit (BFU), Decimal floating-point unit (DFU) that are enhanced with z14 in different ways.

z14 has a new decimal architecture with Vector Enhancements Facility and Vector Packed Decimal Facility for Data Access Accelerator. Vector Packed Decimal Facility introduces set of instructions that perform operation on decimal type using vector registers in order to improve performance. z14 offers up to 2x throughput of vector Binary Floating Point double precision operations and RSA/ECC acceleration (Long Multiply Support).

By exploiting the Vector Packed Decimal Facility that allows packed decimal operations to be performed in registers rather than memory - using new fast mathematical computations - compilers, such as Enterprise COBOL for z/OS,V6.2, Enterprise PL/I for z/OS, V5.2, z/OS V2.3 XL C/C++, the COBOL optimizer, Automatic Binary Optimizer for z/OS, V1.3, and Java, are optimized on z14.

Much of today's commercial computing uses decimal floating point calculus, so on-core hardware decimal floating point units meet the requirements of business and user applications, which provides greater floating point execution throughput with improved performance and precision.

Hot Line Speculation Avoidance

The new mechanism allows full out-of-order / branch speculation for any instruction hitting in L1/L2 caches, and also allow for out-of-order TLB-miss handling. For known hot cache lines, the implementation does not allow sending requests to L3 speculatively and Load Store Unit (LSU) rejects the request until the instruction is next-to-complete

Guarded Storage Facility

Also known as less-pausing garbage collection, Guarded Storage Facility is a new architecture introduced with z14 to enable enterprise scale Java applications to run without periodic pause for garbage collection on larger heaps. This facility improves Java performance by for reducing program pauses during Java Garbage Collection.

Instruction Execution Protection (IEP)

Instruction Execution Protection is a new hardware function introduced with z14 that enables software like LE (Language Environment®) to mark certain memory regions (E.g. a heap or stack) as non-executable to improve the security of programs running on IBM Z servers against stack-overflow or similar attacks.

Simultaneous multithreading

z/Architecture introduced SMT support with z13; allowing simultaneous execution of two threads (SMT) in the same zIIP or IFL core, dynamically sharing processor resources such as execution units and caches. SMT in z13 allowed a more efficient utilization of the core and increased capacity because while one of the threads is waiting for a storage access (cache miss), the other thread that is running simultaneously in the core can use the shared resources rather than remain idle.

Adjusted with the growth in the core cache and TLB2, second generation SMT on z14 improves thread balancing, supports multiple outstanding translations, optimizes hang avoidance mechanisms and delivers improved virtualization performance to benefit Linux on z Systems. z14 provides economies of scale with next generation multithreading (SMT) for Linux and zIIP-eligible workloads. In z14, the I/O System Assist Processor (SAP) has SMT enabled by default.

Single instruction multiple data instruction set

Introduced with the IBM z13, the SIMD unit and instructions use the superscalar core to process operands in parallel, which enables more processor throughput.

z14 introduces a new decimal architecture and new SIMD (vector) instruction set which are designed to boost performance for traditional workloads using COBOL and new applications like analytics. The SIMD unit in z14 now supports 32-bit floating point operations. Leveraging enhanced mathematical libraries like OpenBLAS provides performance improvements for analytical workloads.

Transactional Execution facility

The IBM z14 has a set of instructions (like its predecessor z13,) that allows defining groups of instructions that are run atomically, that is, either all the results are committed or none are. The facility provides for faster and more scalable multi-threaded execution, and is known as *hardware transactional memory*.

Out-of-order execution

Like its predecessor z13, z14 has an enhanced superscalar microprocessor with Out-of-Order execution to achieve faster throughput. With Out-of-Order, instructions might not run in the original program order, although results are presented in the original order. For example, Out-of-Order allows a few instructions to complete while another instruction is waiting. Up to six instructions can be decoded per system cycle, and up to 10 instructions can be in execution.

Concurrent processor unit conversions

The z14 supports concurrent conversion between various PU types, which provides the flexibility to meet the requirements of changing business environments. CPs, IFLs, zIIPs, ICFs, and optional SAPs can be converted to CPs, IFLs, zIIPs, ICFs, and optional SAPs.

Memory subsystem and topology

The z14 servers use DDR4 (double data rate fourth-generation) DIMM (dual inline memory module) technology. For this purpose, IBM has developed a chip that controls communication with the PU (System Controller - SC - chip), and drives address and control from DIMM to DIMM. The DIMM capacities are 32 GB, 64 GB, 128 GB, 256 GB, and 512 GB.

Memory topology provides the following benefits:

- A RAIM for protection at the dynamic random access memory (DRAM), DIMM, and memory channel levels
- A maximum of 32 TB of user configurable memory with a maximum of 40 TB of physical memory (with a maximum of 16 TB configurable to a single LPAR)
- One memory port for each CP SCM, and up to five independent memory ports per CPC drawer
- Increased bandwidth between memory and I/O
- Asymmetrical memory size and DRAM technology across CPC drawers
- Large memory pages (1 MB and 2 GB)
- Key storage
- Storage protection key array that is kept in physical memory
- Storage protection (memory) key that is also kept in every L2 and L3 cache directory entry
- A larger (192 GB) fixed-size HSA that eliminates having to plan for HSA

PCIe fanout hot-plug

The *PCIe fanout* provides the path for data between memory and the PCIe features through the PCIe 16 GBps bus and cables. The PCIe fanout is hot-pluggable. During an outage, a redundant I/O interconnect allows a PCIe fanout to be concurrently repaired without loss of access to its associated I/O domains. Up to 10 PCIe fanouts are available per CPC drawer. The PCIe fanout can also be used for the ICA SR. If redundancy in coupling link connectivity is ensured, the PCIe fanout can be concurrently repaired.

Host channel adapter fanout hot-plug

The HCA fanout also is used for the InfiniBand coupling links (HCA3-O and HCA3-O LR) which provide the path for data between memory and the remote sysplex members, and it is connected to through 6 GBps or 5 Gbps IFB cables respectively. If redundancy in coupling link connectivity is ensured, the HCA fanout can be concurrently repaired.

1.3.5 I/O connectivity: PCIe Generation 3

The z14 server offers new and improved I/O features and uses PCIe Gen 3 for I/O connectivity. This section briefly reviews the most relevant I/O capabilities.

The z14 takes advantage of PCIe Generation 3 to implement the following features:

- PCIe Generation 3 (Gen3) fanouts that provide 16 GBps connections to the PCIe I/O features in the PCIe I/O drawers.
- PCIe Gen3 fanouts that provide 8 GBps coupling link connections through the new IBM ICA SR.

1.3.6 I/O subsystem

The z14 I/O subsystem is similar to the one on z13 and includes the PCIe Gen3 infrastructure. This infrastructure is designed to reduce processor usage and I/O latency, and provide increased throughput.

z14 servers offer PCIe I/O drawers hosting PCIe features. I/O drawers used in previous IBM Z servers are not supported on z14.

PCIe I/O drawer

The *PCIe I/O drawer*, together with the PCIe features, offers finer granularity and capacity over previous I/O infrastructures. It can be concurrently added and removed in the field, easing planning. Only PCIe cards (features) are supported, in any combination. Up to five PCIe I/O drawers can be installed on a z14 server.

Native PCIe and Integrated Firmware Processor

Native PCIe support was introduced with the zEDC and RoCE Express features, which are managed differently from the traditional PCIe I/O (FICON Express and OSA-Express) features. The device drivers for the native features are provided in the operating system. The diagnostic tests for the adapter layer functions of the native PCIe features are managed by LIC that is running as a resource group (RG), which runs on the IFP.

With z14, the number of Resource Groups are increased from two to four in order to add more granularity thus helping mitigate the impact of the disruptive Resource Group MCL (Microcode Change Level) installations. This firmware management enhancement contributes to the RAS of the server.

During the ordering process of the native PCIe features, features of the same type are evenly spread across the four resource groups (RG1, RG2, RG3 and RG4) for availability and serviceability reasons. Resource groups are automatically activated when these features are present in the CPC.

In addition to the zEDC and 10 GbE RoCE Express features, the z14 introduces the following native PCIe I/O features:

- Coupling Express Long Reach (CE LR)
- zHyperLink Express
- 10 GbE RoCE Express2

1.3.7 I/O and special purpose features in the PCIe I/O drawer

The z14 server (new build) supports the following PCIe features that are installed in the PCIe I/O drawers:

- Storage connectivity
  - FICON Express16S+ Short Wave (SX)
  - FICON Express16S+ Long Wave (LX) 10 km (6.2 miles)
  - zHyperLink Express
- Network Connectivity
  - OSA-Express6S 10 GbE Long Reach (LR)
  - OSA-Express6S 10 GbE Short Reach (SR)
  - OSA-Express6S GbE LX
  - OSA-Express6S GbE SX
  - OSA-Express6S 1000BASE-T
  - 10GbE RoCE Express2
- Coupling and Server Time Protocol connectivity
  - Coupling Express LR
- Cryptography
  - Crypto Express6S
  - Regional Crypto Enablement
- zEDC Express

When carried forward on an upgrade, the z14 server also supports the following features in the PCIe I/O drawers:

- Storage connectivity
  - FICON Express 16S SX
  - FICON Express 16S LX
  - FICON Express 8S SX
  - FICON Express 8S LX
- Network Connectivity
  - OSA-Express5S 10 GbE Long Reach (LR)
  - OSA-Express5S 10 GbE Short Reach (SR)
  - OSA-Express5S GbE LX
  - OSA-Express5S GbE SX
  - OSA-Express5S 1000BASE-T
  - OSA-Express 4S 1000BASE-T
  - 10GbE RoCE Express
- Cryptography
  - Crypto Express5S
  - Regional Crypto Enablement
- zEDC Express

Used for coupling connectivity, the IBM Integrated Coupling Adapter (ICA SR), the InfiniBand coupling links HCA3-O and HCA3-O LR are other z14 supported features not listed in the previous list as they are attached directly to the CPC drawer.

1.3.8 Storage connectivity

z14 supports the new IBM zHyperLink Express and FICON channels for storage connectivity.

IBM zHyperLink Express

zHyperLink Express is a new short-distance mainframe attach link designed to increase the scalability of IBM Z transaction processing and lower I/O latency than High Performance FICON (HPF) by for bringing data close to processing power. It is the new mainframe input/output (I/O) channel link technology since FICON.

zHyperLink Express feature directly connects the z14 Central Processor Complex (CPC) to the I/O Bay of the DS8880 (R8.3). This short distance (up to 150m), direct connection is intended to reduce I/O latency and improve storage I/O throughput. The improved performance of zHyperLink Express allows the z14 PU to make a synchronous request for the data that resides in the DS8880 cache. This eliminates the un-dispatch of the running request, the queuing delays to resume the request, and the PU cache disruption.

The IBM zHyperLink Express is a two-port feature residing in the PCIe I/O drawer. Up to 16 features with up to 32 zHyperLink Express ports are supported in a z14 CPC. The zHyperLink Express feature utilizes PCIe Gen3 technology, with x16 lanes that are bifurcated into x8 lanes for storage connectivity. It is designed to support a link data rate of 8 GigaBytes per second (GBps)[18].

The point-to-point link is established by 24x fiber optic cable with Multi-fiber Termination Push-on (MTP) connectors. For more information, see “zHyperLink Express (FC #0431)” on page 165.

FICON channels

Up to 160 features with up to 320 FICON Express16S+ channels are supported on a new build z14. FICON Express 16S+ and FICON Express 16S (carry forward only) support 4, 8, or 16 Gbps, while FICON Express8S features (carry forward only) support link data rates of 2, 4, or 8 Gbps.

FICON Express16S+ offers increased performance compared to FICON Express16S with its new IBM I/O ASIC that supports up to 3x the I/O start rate of previous FICON/FCP solution. Another distinction for FICON Express16S+ is that both ports of a feature must be defined as the same CHPID type (no mix of FC and FCP CHPID for the same feature).

The FICON features on z14 support the following protocols:

- FICON (CHPID type FC) and High Performance FICON for Z (zHPF). zHPF offers improved performance for data access, which is of special importance to online transaction processing (OLTP) applications.
- FICON channel-to-channel (CHPID type CTC).
- Fibre Channel Protocol (CHPID type FCP).

FICON also offers the following capabilities:

[18] The link data rates do not represent the performance of the links. The actual performance is dependent upon many factors including latency through the adapters, cable lengths, and the type of workload.

- Modified Indirect Data Address Word (MIDAW) facility: Provides more capacity over native FICON channels for programs that process data sets that use striping and compression, such as DB2, VSAM, partitioned data set extended (PDSE), hierarchical file system (HFS), and z/OS file system (zFS). It does so by reducing channel, director, and control unit processor usage.
- Enhanced problem determination, analysis, and manageability of the storage area network (SAN) by providing registration information to the fabric name server for both FICON and FCP.
- An Extended Link Service command, Read Diagnostic Parameters (RDP) is used to obtain extra diagnostic data from the Small Form Factor Pluggable optics that are located throughout the SAN fabric to improve the accuracy of identifying a failing component.

1.3.9 Network Connectivity

The IBM z14 supports the following technologies for network connectivity:

- Open Systems Adapter (OSA)
- HiperSockets
- Shared Memory Communication - Remote Direct Memory Access over Converged Ethernet (SMC-R)
- Shared Memory Communications - Direct Memory Access over Internal Shared Memory (SMC-D)

Open Systems Adapter

z14 allows any mix of the supported OSA Ethernet features listed in 1.3.7, “I/O and special purpose features in the PCIe I/O drawer” on page 18. OSA-Express6 features are a technology refresh of the OSA-Express5S features. Up to 48 OSA-Express6S features, with a maximum of 96 ports, are supported. The maximum number of combined OSA-Express features cannot exceed 48.

OSA-Express features provide important benefits for TCP/IP traffic by reducing latency and improving throughput for standard and jumbo frames. Data router function present in all OSA-Express features enables performance enhancements. With OSA-Express6S, OSA-Express5S and OSA-Express4S; the functions that used to be performed in firmware are performed in the hardware. Extra logic in the IBM application-specific integrated circuit (ASIC) that is included with these features handle packet construction, inspection, and routing; allowing packets to flow between host memory and the LAN at line speed without firmware intervention.

On both z14 and z13, OSA configured as an integrated console controller CHPID type (OSC) supports the configuration and enablement of secure connections by using the Transport Layer Security (TLS) protocol versions 1.0, 1.1 and 1.2.

Important: Fulfilling the related Statement of Direction, z14 has removed the support for configuring OSN CHPID types.

For more information about the OSA features, see 4.7, “Connectivity” on page 156.

HiperSockets

The HiperSockets function - also known as internal queued direct input/output (internal QDIO or iQDIO) - is an integrated function of the z14 server that provides users with attachments to up to 32 high-speed virtual LANs with minimal system and network processor usage.

For communications between LPARs in the same z14 server, HiperSockets eliminates the need to use I/O subsystem features to traverse an external network. Connection to HiperSockets offers significant value in server consolidation by connecting many virtual servers.

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

HiperSockets can also be used for Dynamic cross-system coupling, a z/OS Communications Server feature that creates trusted, internal links to other stacks within a Parallel Sysplex.

Shared Memory Communication - Remote Direct Memory Access

zEC12 GA2 was the first IBM Z server generation to support RoCE (Remote Direct Memory Access over Converged Ethernet) technology which is designed to provide fast, reduced CPU consumption, memory-to-memory communications between two IBM Z CPCs.

RoCE Express features reduce CPU consumption for applications that use the TCP/IP stack (sockets communication) such as IBM WebSphere Application Server accessing a DB2 database. It is transparent to applications and also might help to reduce network latency with memory-to-memory transfers that use SMC-R in supported z/OS releases.

IBM Z server generations continue to enhance the RoCE architecture. While the z13 enabled both ports of the 10GbE RoCE Express feature and sharing among 31 LPARs running z/OS, z14 has refreshed the technology with the new 10GbE RoCE Express2 which provides support for 4x the number of LPAR and performance improvements. RoCE Express2 increases virtualization by supporting 63 Virtual Functions (VFs) per port - for up to 126 VFs per PCHID (physical channel ID).

Both 10GbE RoCE Express2 and 10GbE RoCE Express features use SR optics, and support the use of a multimode fiber optic cable that terminates with an LC Duplex connector. Both support point-to-point and switched connections with an enterprise-class 10 GbE switch. A maximum of 8 RoCE Express features can be installed in PCIe I/O drawers of z14.

For more information, see “Shared Memory Communications over RDMA” on page 459.

Shared Memory Communications - Direct Memory Access

SMC-D enables low processor usage and low latency communications within a CPC using an Direct Memory Access connection over Internal Shared Memory (ISM). SMC-D implementation is Similar to SMC-R over RoCE; SMC-D over ISM extends the benefits of SMC-R to operating system instances running on the same CPC, without requiring physical resources (RoCE adapters, PCI bandwidth, ports, I/O slots, network resources, 10GbE switches).

Introduced with z13 GA2 and z13s, SMC-D enables high bandwidth LPAR-to-LPAR TCP/IP traffic (sockets communication) using the direct memory access software protocols over virtual Internal Shared Memory PCIe devices (vPCIe). 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.

z14 continues to support SMC-D with its lightweight design improving throughput, latency, and CPU consumption, complementing HiperSockets, OSA, or RoCE without sacrificing quality of service.

SMC-D requires either an OSA or a HiperSockets connection to establish the initial TCP communications and can coexist with them. SMC-D uses a virtual PCIe adapter and is

configured like a physical PCIe device. There are up to 32 ISM adapters, each with a unique Physical Network ID per CPC.

Notes:

- SMC-D does not support coupling facilities (z/OS to z/OS only).
- Shared Memory Communication protocols (SMC-R and SMC-D) do not currently support multiple IP subnets.

1.3.10 Coupling and Server Time Protocol connectivity

IBM z14 support for Parallel Sysplex includes the Coupling Facility (running the Coupling Facility Control Code - CFCC) and coupling links.

Coupling links support

Coupling connectivity in support of Parallel Sysplex environments is provided on the z14 server by the following features:

- Internal Coupling (IC) links, operating at memory speed.
- Integrated Coupling Adapter Short Reach (ICA SR)
- Coupling Express Long Reach (CE LR)
- InfiniBand (IFB) Coupling links
  - HCA3-O, 12x Parallel Sysplex InfiniBand (IFB) coupling links.
  - HCA3-O, 1x Parallel Sysplex InfiniBand (IFB) coupling links.

Note[a]: z14 is the last high-end IBM Z server to support Parallel Sysplex InfiniBand (PSIFB) coupling links, while z13s is the last midrange IBM Z server to support them. IBM Z enterprises should plan to migrate off of PSIFB links.

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.

All coupling link types can be used to carry STP messages.

Integrated Coupling Adapter Short Reach (ICA SR)

The ICA SR feature (introduced with IBM z13) is a two-port fanout used for short distance coupling connectivity. It uses PCIe Gen3 technology, with x16 lanes that are bifurcated into x8 lanes for coupling. The ICA SR is designed to drive distances up to 150m and support a link data rate of 8 gigabytes per second (GBps). The ICA SR fanout takes one PCIe I/O fanout slot in the z14 CPC drawer. It is used for coupling connectivity between z14, z13 and z13s CPCs, and cannot be connected to HCA3-O or HCA3-O LR coupling fanouts. The ICA SR is only compatible with another ICA SR.

Coupling Express Long Reach (CE LR)

z14 introduces the new Coupling Express Long Range (CE LR) which is a two port feature used for long distance coupling connectivity and defined as coupling channel type, CL5. Unlike the 1x IFB Infiniband coupling links, which were plugged into the CPC drawer, the CE LR link is plugged in a PCIe I/O drawer slot, taking advantage industry standard I/O technology. Compared to the HCAO-3 LR feature, which has 4-port and 4-link increments, the CE LR link allows for more granularity when scaling up or completing maintenance. Link performance is similar to the InfiniBand 1x coupling link and uses identical Single Mode fiber.

The CE LR link provides point-to-point coupling connectivity at distances of 10 km unrepeated and 100 km with a qualified dense wavelength division multiplexing (DWDM) device.

InfiniBand (IFB) Coupling links

Parallel Sysplex coupling links using IFB:

- 12x InfiniBand coupling links (HCA3-O) for local connections (up to 150 meters / 492 feet). The 12x IFB link (HCA3-O) connects between z14, z13, z13s, zEC12 and zBC12 servers and. It cannot be connected to an ICA SR.
- 1x InfiniBand coupling links (HCA3-O LR) for extended distance connections between z14, z13, z13s, zEC12 and zBC12 servers).The HCA3-O LR 1X InfiniBand links have a bandwidth of 5 Gbps, utilize single mode fiber and cannot be connected to a CE LR. The HCA3-O LR link provides point-to-point coupling connectivity at distances of 10 km unrepeated and 100 km with a qualified dense wavelength division multiplexing (DWDM) device.

CFCC Level 22

CFCC level 22 is delivered on the z14 with driver level 32. CFCC Level 22 introduces the following enhancements:

- Support for up to 170 ICF processors per z14:
  - The maximum number of logical processors in a Coupling Facility Partition remains 16.
- Support for new Coupling Express LR links
- *Coupling Facility (CF) Processor Scalability*: Coupling Facility work management and dispatching changes for z14 allow improved efficiency and scalability for coupling facility images.
- *CF List Notification Enhancements*: There are significant enhancements to CF notifications that inform users about the status of shared objects within a Coupling Facility.
- *Coupling Link Constraint Relief*: z14 provides additional physical coupling link connectivity compared to z13.
- *CF Encryption*: z/OS 2.3 provides support for end-to-end encryption for both 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.

z14 servers with CFCC Level 22 require z/OS V1R13 or later, and z/VM V6R3 or later for virtual guest coupling.

While the coupling facility LPARs are running on different server generations, different levels of CFCC can coexist in the same sysplex, enabling upgrade from one CFCC level to the next. CF LPARs running on the same server share a single CFCC level.

A CF running on a z14 server (CFCC level 22) can coexist in a sysplex with CFCC levels 21 and 19. Review the CF LPAR size by using the CFSizer tool:

`http://www.ibm.com/systems/z/cfsizer`

Server Time Protocol facility

Time synchronization for Parallel Sysplex Server Time Protocol (STP) is designed to ensure events occurring in different servers are properly sequenced in time. STP is designed for servers that have been configured in a Parallel Sysplex or a basic sysplex (without a Coupling Facility), as well as servers that are not in a sysplex but need time synchronization.

STP is a server-wide facility that is implemented in the LIC, presenting a single view of time to Processor Resource/Systems Manager (PR/SM). Any IBM Z CPC (including CPCs running as stand-alone coupling facilities) can be enabled for STP by installing the STP feature.

STP uses a message based protocol in which timekeeping information is passed over externally defined coupling links between servers. The STP design introduced a concept called Coordinated Timing Network (CTN), a collection of servers and Coupling Facilities that are time-synchronized to a time value called Coordinated Server Time (CST).

Network Time Protocol (NTP) client support is available to the STP code on the z14, z13, z13s, zEC12, and zBC12 servers. With this function, these servers can be configured to use an NTP server as an External Time Source (ETS). This implementation fulfills the need for a single time source across the heterogeneous platforms in the enterprise including IBM Z servers and others systems running Linux, UNIX and Microsoft Windows operating systems.

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

HMC can be configured as an NTP client or an NTP server. In order to ensure secure connectivity, HMC NTP broadband authentication can be enabled on z14, z13 and zEC12 servers.

STP enhancements in z14

IBM z14 introduces an additional *synchronization level* for Server Time Protocol (STP) - *STP stratum level 4*. With the additional stratum level, STP can synchronize systems up to 3 levels away from the Current Time Server (CTS). Prior systems only allowed synchronization up to level 3, or up to 2 levels from the CTS. This additional stratum level is not intended for long-term use; rather, it is specifically intended for short-term use during configuration changes for large timing networks, to avoid some of the cost and complexity caused by being constrained to a three level STP stratum configuration.

z14 also introduces a new Graphical User Display for the STP network and configuration. The new user interface has been revamped for a quick, intuitive view and management of the various pieces of the CTN, including the status of the components of the timing network. The new level of HMC allows the management of older systems using the same new interface.

Attention: A z14 server, like its predecessor z13, cannot be connected to a Sysplex Timer and cannot be a member in a Mixed CTN. An STP only CTN is required for the z14 and z13 servers.

If a current configuration consists of a Mixed CTN or a Sysplex Timer (M/T 9037), the configuration must be changed to an STP-only CTN before z14 integration. The z14 server can coexist only with IBM Z CPCs that do not have the external time reference (ETR) port capability.

Enhanced Console Assisted Recovery

Console Assisted Recovery (CAR) support is designed to help a Backup Time Server determine whether the Primary Time Server (PTS) is still up and running if coupling traffic ceases. The CAR process is started by the Backup Time Server (BTS) if there is a lack of communication between the Primary and Backup Time Servers. The BTS queries the state of the PTS/CTS SE using the SE and HMC of the Backup Time Server. If the PTS is down, the BTS starts takeover.

With the new Enhanced Console Assisted Recovery (ECAR), the process of BTS takeover is faster. When the PTS encounters a checkstop condition, the CPC informs the SE and HMC of the condition. The PTS SE recognizes the pending checkstop condition, and an ECAR request is sent directly from the HMC to the BTS SE to start the takeover. The new ECAR support is faster than the original support as there is almost no delay between the system checkstop and the start of CAR processing. ECAR is only available on z14, z13 GA2 and z13s servers. In a mixed environment with previous generation machines, define ECAR supported servers as the PTS and CTS.

1.3.11 Cryptography

There is a strong synergy between cryptography and security. Cryptography provides the primitives to support security functions. Similarly, security functions help to ensure authorized use of key material and cryptographic functions.

Cryptography on IBM Z is built on the platform with integrity. IBM Z platform offers hardware based cryptography features exploited by Java, DB2/IMS encryption tool, DB2 built in encryption z/OS Communication Server, IPsec/IKE/AT-TLS, z/OS System SSL, z/OS, z/OS Encryption Facility, Linux on z Systems and more:

- CP Assist For Cryptographic Functions (CPACF)
- Crypto Express6S
- Regional Crypto Enablement (RCE)
- Trusted Key Entry (TKE) workstation

CP Assist For Cryptographic Functions (CPACF)

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

With z14, CPACF is enhanced to support pervasive encryption to provide faster encryption and decryption than previous servers. For every Processor Unit defined as a Central Processor (CP) or an Integrated Facility for Linux (IFL) it offers the following enhancements over z13:

- Reduced overhead on short data (hashing and encryption)
- Up to 4x throughput for AES
- Special instructions for elliptic curve cryptography (ECC) / RSA
- New hashing algorithms, e.g. SHA-3
- Support for authenticated encryption (combined encryption and hashing, e.g. AES-GCM)
- True random number generator (e.g. for session keys)

The z13 CPACF provides (supported by z14 also):

- *For data privacy and confidentially*: Data Encryption Standard (DES), Triple Data Encryption Standard (TDES), Advanced Encryption Standard (AES) for 128-bit, 192-bit and 256-bit keys.

- *For data integrity*: Secure Hash Algorithm-1 (SHA-1) 160-bit, and SHA-2 for 224-, 256-, 384- and 512-bit support. SHA-1 and SHA-2 are shipped enabled on all z14's and do not require the no-charge enablement feature.
- *For key generation*: Pseudo Random Number Generation (PRNG), Random Number Generation Long (RNGL) - 1 byte to 8192 bytes, Random Number Generation Long (RNG) with up to 4096-bit key RSA support for message authentication.

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

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

Crypto Express6S

Crypto Express6S represents the newest generation of cryptographic features. Cryptographic performance improvements with new Crypto Express6S (FC #0893) allow more data to be securely transferred across the Internet. Crypto Express6S is designed to complement the cryptographic capabilities of the CPACF. It is an optional feature of the z14 server generation.

The Crypto Express6S feature is designed to provide granularity for increased flexibility with one PCIe adapter per feature. Although installed in the PCIe I/O drawer, Crypto Express6S features do not perform I/O operations. That is, no data is moved between the CPC and any externally attached devices. For availability reasons, a minimum of two features is required.

z14 servers allow sharing of a cryptographic coprocessor across 85 domains (the maximum number of LPARs on the system for z14 is 85).

The Crypto Express6S is a state-of-the-art, tamper-sensing, and tamper-responding programmable cryptographic feature that provides a secure cryptographic environment. Each adapter contains a tamper-resistant hardware security module (HSM). The HSM can be configured as a Secure IBM CCA coprocessor, as a Secure IBM Enterprise PKCS #11 (EP11) coprocessor, or as an accelerator:

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

The Crypto Express6S is designed to meet these cryptographic standards, among others:

- FIPS 140-2 Level 4[19]
- Common Criteria EP11 EAL4
- ANSI 9.97
- Payment Card Industry (PCI) HSM
- German Banking Industry Commission (GBIC), (formerly DK, Deutsche Kreditwirtschaft)

Federal Information Processing Standard (FIPS) 140-2 certification is supported only when Crypto Express6S is configured as a CCA or an EP11 coprocessor.

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

Crypto Express6S supports a number of ciphers and standards covered in the following sections. For more information about cryptographic algorithms and standards, see Chapter 6, “Cryptographic features” on page 195.

Regional Crypto Enablement (RCE)

Starting with z13 GA2, IBM has enabled geo-specific cryptographic support that is supplied by IBM approved vendors. China is the first geography to exploit 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.

RCE is a framework to enable the integration of IBM certified 3rd party cryptographic hardware for regional or industry encryption requirements and supports the use of cryptography algorithms and equipment from selected providers in conjunction with IBM Z in specific countries. Support for exploiting international algorithms (AES, DES, RSA, ECC) with regional crypto devices (supporting regional algorithms like SMx) is added to the ICSF PKCS#11 services.

When ordered, the Regional Crypto Enablement support will reserve the I/O slot(s) for the IBM approved vendor-supplied cryptographic card(s). Clients will need to directly contact the IBM approved vendor for purchasing information.

Trusted Key Entry (TKE) workstation

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

The TKE workstation offers a security-rich solution for basic local and remote key management. It provides authorized personnel with a method for key identification, exchange, separation, update, and backup, and a secure hardware-based key loading mechanism for operational and master keys. TKE also provides secure management of host cryptographic module and host capabilities.

Support for an optional smart card reader that is attached to the TKE workstation allows the use of smart cards that contain an embedded microprocessor and associated memory for data storage. Access to and the use of confidential data on the smart cards are protected by a user-defined personal identification number (PIN).

TKE workstation and the most recent TKE 9.0 LIC are optional features on the z14. TKE workstation is offered in two types; TKE Tower (FC #0086) and TKE Rack Mount (FC #0085). TKE 9.0 requires the crypto adapter FC 4768. You can use a prior TKE version to collect data from previous generations of cryptographic modules and apply the data to Crypto Express6S coprocessors.

TKE 9.0 is required if you choose to use the TKE to manage a Crypto Express6S. TKE 9.0 is also required to manage the new CCA mode PCI-HSM settings which are available on the Crypto Express6S. A TKE is required to manage any Crypto Express feature that is running in IBM Enterprise PKCS #11 (EP11) mode. If EP11 is to be defined, smart cards that are used require FIPS certification.

For more information about the cryptographic features, see Chapter 6, “Cryptographic features” on page 195. Also, see Web Deliverables download site for the most current ICSF updates available:

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

1.3.12 zEDC Express

zEDC Express, an optional feature that is available on z14, z13, z13s, zEC12, and zBC12 servers, provides hardware-based acceleration for data compression and decompression with lower CPU consumption. The IBM zEnterprise Data Compression (zEDC) acceleration capability in z/OS and the zEDC Express feature are designed to help improve cross-platform data exchange, reduce CPU consumption, and save disk space.

Although installed in the PCIe I/O drawer, zEDC Express features do not perform I/O operations. That is, no data is moved between the CPC and externally attached devices. There is one PCIe adapter/compression coprocessor per feature. The zEDC Express feature can be shared by up to 15 LPARs. Up to 16 features can be installed on z14.

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

http://www.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/PRS5132

For more information, see Appendix F, “IBM zEnterprise Data Compression Express” on page 493.

1.4 Reliability, availability, and serviceability

The z14 RAS strategy employs a building-block approach, which is developed to meet the client's stringent requirements for achieving continuous reliable operation. Those building blocks are error prevention, error detection, recovery, problem determination, service structure, change management, measurement, and analysis.

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

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

The IBM z14 processor builds upon the Reliability, Availability and Serviceability (RAS) of the z13 family; with the following RAS improvements:

- The level 3 cache added powerful symbol ECC making it resistant to more failures (the z13 hardened the level 4 cache and the main memory was hardened with RAIM and ECC prior to that).
- The main memory added preemptive DRAM marking to isolate and recover failures faster.
- Small array error handling is improved in the processor cores.
- Additional error thresholding was added to the processor core to isolate “sick but not dead” failure scenarios.
- The number of Resource Groups for supporting native PCIe features has increased to four from two to reduce the impact of firmware updates and failures.
- OSA-Express6S adds TCP checksum on large send

For more information, see Chapter 9, “Reliability, availability, and serviceability” on page 345.

1.5 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 comes with driver level 32 and HMC application Version 2.14.0.

For more information, see Chapter 11, “Hardware Management Console and Support Elements” on page 389.

1.6 Operating systems

The IBM z14 server 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. Use of various features might require the latest releases. For more information, see Chapter 7, “Operating System Support” on page 227.

1.6.1 Supported operating systems

The following operating systems with required maintenance applied are supported for z14 servers:

- z/OS”
  - Version 2 Release 3 (Previewed and expected to be made available September, 2017)
  - Version 2 Release 2
  - Version 2 Release 1
  - Version 1 Release 13 (compatibility support only, with extended support agreement)
- z/VM:
  - Version 6 Release 4
  - Version 6 Release 3
- z/VSE
  - Version 6 Release 2 (Previewed and expected availability date is fourth quarter, 2017)
  - Version 6 Release 1
  - Version 5 Release 2
  - Version 5 Release 1
- z/TPF Version 1 Release 1
- Linux on z Systems distributions[20]:
  - SUSE Linux Enterprise Server (SLES): SLES 12 and SLES 11.
  - Red Hat Enterprise Linux (RHEL): RHEL 7 and RHEL 6.
  - Ubuntu 16.04 LTS (or higher)
- KVM for IBM Z (KVM for IBM z Systems®) Release 1.1 and later
  - KVM hypervisor which is offered with the following Linux distributions: SLES-12 SP2 or higher, and Ubuntu 16.04 LTS or higher.

[20] Customers should monitor for new distribution releases supported

For supported Linux on z Systems distribution levels, see the following website:

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

For detailed information about the features and functions supported on z14 by operating systems, see Chapter 7, “Operating System Support” on page 227.

z/VM Support

With the PTF for APAR VM65942, planned to be available August 25, 2017, z/VM V6.4 provides support for z14. Functions enabling guests to exploit functions supported by z/VM on z14 include, but not limited to:

- z/Architecture support
- ESA/390-compatibility mode for guests
- New hardware facilities:
  - Miscellaneous-Instruction-Extensions Facility 2
  - Vector Enhancements Facility 1
  - Vector Packed Decimal Facility
  - Message-Security-Assist Extension 6, 7, and 8
- Up to 2 TB per z/VM LPAR
- Improved memory management support
- Support for IBM DPM, enabling z/VM 6.4 to be configured through new DPM interface rather than traditional PR/SM
- Guests exploitation support to use Large Page (1 MB pages)
- Guest exploitation support to use SIMD (Single Instruction Multiple Data)
- Guest exploitation support for the Instruction Execution Protection Facility
- Guest exploitation support for pause-less garbage collection
- Encrypted paging support
- Crypto Clear Key ECC operations
- Dynamic I/O support
- Guest exploitation support for RoCE Express2 and RoCE Express - SMC-R
- Guest exploitation for SMC-D

For detailed information about the features and functions supported on z14 by operating systems, see Chapter 7, “Operating System Support” on page 227.

z/OS Support

z/OS takes advantage of many of the new functions and features of z14. Listed below is a subset of z14 functions z/OS can exploit depending on version and release:

- Up to 170 processors per LPAR or up to 128 physical processors per LPAR in SMT mode (SMT for zIIP)
- Up to 16 TB of real memory per LPAR (dependent on z/OS version)
- Two-way simultaneous multithreading (SMT) optimization and support of SAPs (SAP SMT enabled by default) in addition to zIIP engines
- XL C/C++ **`ARCH(12)`** and **`TUNE(12)`** exploitation
- Exploitation of faster CPACF

  - Pervasive Encryption
    - Coupling Facility Encryption
    - Dataset and network encryption
- HiperDispatch Enhancements
- z14 Hardware Instrumentation Services (HIS)
- Entropy-Encoding Compression Enhancements
- Guarded Storage Facility (GSF)
- Instruction Execution Protection (IEP)
- IBM Virtual Flash Memory (VFM)
- Improved memory management in Real Storage Manager (RSM)
- CF Exploitation of VFM
- CFCC Level 22:
  - List Notification Enhancements
  - CF Processor Scalability
  - CF Request Diagnostics
  - Encryption Support
- Coupling Express Long Reach (CE LR) CHPID type CL5
- zHyperLink Express
- FICON Express16S+
  - OSA Express6S
- RoCE-Express2
- Cryptography:
  - Crypto Express6S
    - Next Generation Coprocessor support
    - Support for Coprocessor in PCI-HSM Compliance Mode
    - Architected for up to 85 domains
  - TKE 9.0 workstation

For detailed information about the features and functions supported on z14 by operating systems, see Chapter 7, “Operating System Support” on page 227.

1.6.2 IBM Compilers

The following IBM compilers for Z servers can use z14 servers:

- 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 z Systems

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

To fully use the capabilities of z14 servers, you must compile it 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 refer to “7.5.4, “z/OS XL C/C++ considerations” on page 291.

2

Central processor complex hardware components

This chapter introduces the z14 central processor complex (CPC) hardware components. It also describes the significant features and functions with their characteristics and options. The objective is to explain the z14 hardware building blocks and how these components interconnect from a physical point of view. This information is useful for planning purposes and can help in defining configurations that fit your requirements.

This chapter includes the following sections:

- Frames and drawers
- CPC drawer
- Single chip modules
- Memory
- Reliability, availability, and serviceability
- Connectivity
- Model configurations
- Power and cooling
- Summary of z14 structure

© Copyright IBM Corp. 2017. All rights reserved.

2.1 Frames and drawers

IBM Z frames are enclosures that are built to Electronic Industries Alliance (EIA) standards. The z14 server has two 42U EIA frames. The two frames, A and Z, are bolted together. They have positions for up to four processor drawers and up to five Peripheral Component Interconnect Express (PCIe) I/O drawers. The other components included in the frames are explained in the following sections.

The z14 server and its predecessor, the z13, have the option of ordering the infrastructure to support the top exit of fiber optic cables (FICON, OSA, 12x InfiniBand, 1x InfiniBand, ICA, zHyperLink Express, Coupling Express LR and RoCE) and copper cables for the 1000BASE-T Ethernet features. On the z14 server, the top exit capability is designed to provide an option for overhead power cabling.

The z14 server can be delivered as an air-cooled system or as a water-cooled system. A picture of an air-cooled system with the optional top exit I/O and power feature is shown in Figure 2-1, with the maximum of four CPC drawers and five PCIe I/O drawers.

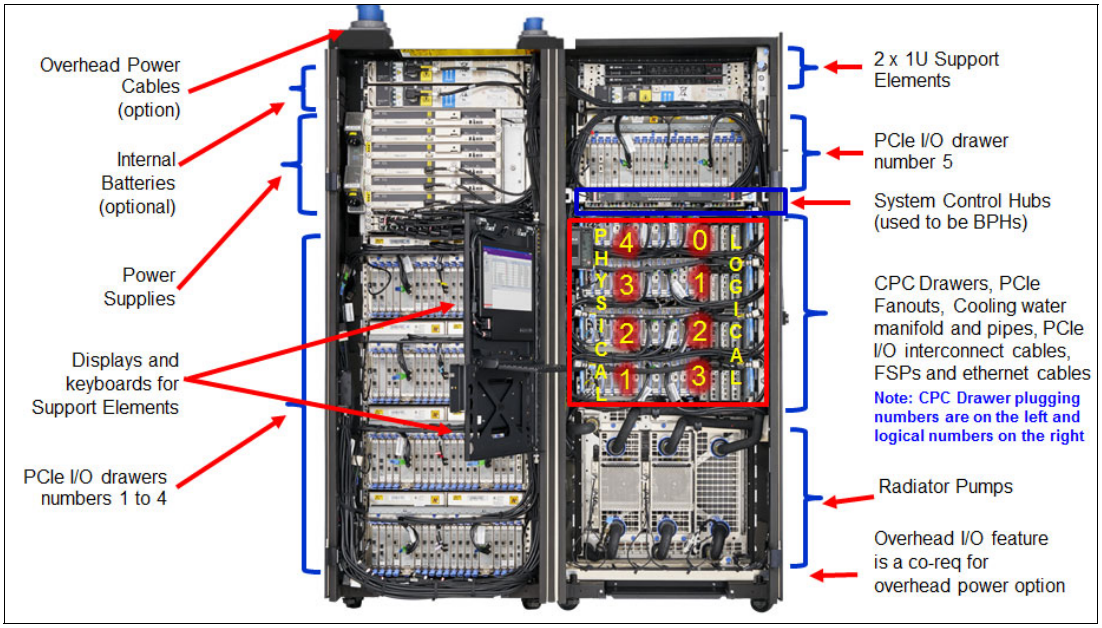


Figure 2-1 z14 internal front view of an air-cooled CPC (models M04 or M05)

A picture of a water-cooled system without the optional top exit I/O and power feature is shown in Figure 2-2, with the maximum of four CPC drawers and five PCIe I/O drawers.

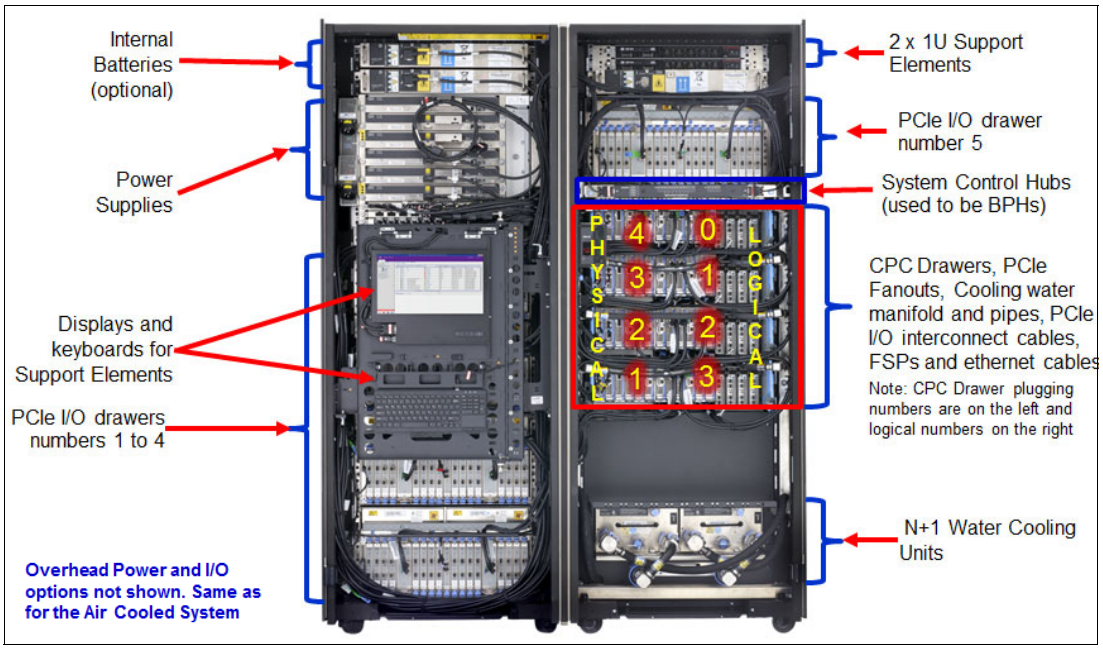


Figure 2-2 z14 internal front view of a water-cooled CPC (models M04 or M05)

2.1.1 The A Frame

As shown in Figure 2-1 on page 34 and Figure 2-2, the A frame has the following major components (from top to bottom of the frame):

- Two Support Element (SE) 1U servers are installed at the top of the A frame.
- Two optional integrated battery features (IBFs), which provide the function of a local uninterrupted power source. The IBF enhances the robustness of the power design, increasing power line disturbance immunity.
- One PCIe I/O drawer. The presence of this PCIe I/O drawer depends on the I/O configuration of the z14 server.
- Two System Control Hubs (SCHs). The SCHs are the replacement of the Bulk Power Hubs that were used in previous IBM Z prior to z13 servers.
- Up to four CPC drawers. The number of the CPC drawers depends on the model number of the z14 server. At least one CPC drawer must be installed, and the additional slots are filled up from the bottom to the top.
- The cooling units for the closed loop water-cooling system of the CPC drawers in the bottom of the A Frame differ for air-cooled and water-cooled systems:
  a. For an air-cooled z14 server (Figure 2-1 on page 34), two pumps and two blowers (N+1 redundant design) are installed.
  b. For a water-cooled z14 server (Figure 2-2), two Water Conditioning Units (WCUs) are installed. The WCUs are connected to a customer-operated (data center) chilled water supply.

2.1.2 Z Frame

As shown in Figure 2-1 on page 34 and Figure 2-2 on page 35, the Z frame has the following major components (from top to bottom of the frame):

- Two or four optional IBFs. The number of batteries depends on the number of bulk power regulators that are installed. IBFs are always installed in pairs.
- Bulk power regulators (BPRs). The number of BPRs varies depending on the configuration of the z14 server. For more information about the required number of BPRs, see 2.8.1, “Power and cooling” on page 75.
- The Keyboard and Display tray (newer design for z14) which is in front of the I/O drawer slots, contains the keyboards and the displays that are connected to the Support Elements.
- Up to four PCIe I/O drawers which install from the top position (location Z22B) to bottom location Z01B).
- An optional overhead power cable feature is shown in Figure 2-1 on page 34. When this feature is ordered, it is present on the Z frame. The top I/O exit cabling feature must also be installed in this case.

2.1.3 z14 Cover (door) design

Figure 2-3 on page 36 shows the standard cover set FC #0160 which includes the rear vectored covers and the new optional thin cover set FC #0161 for space constrained installations.

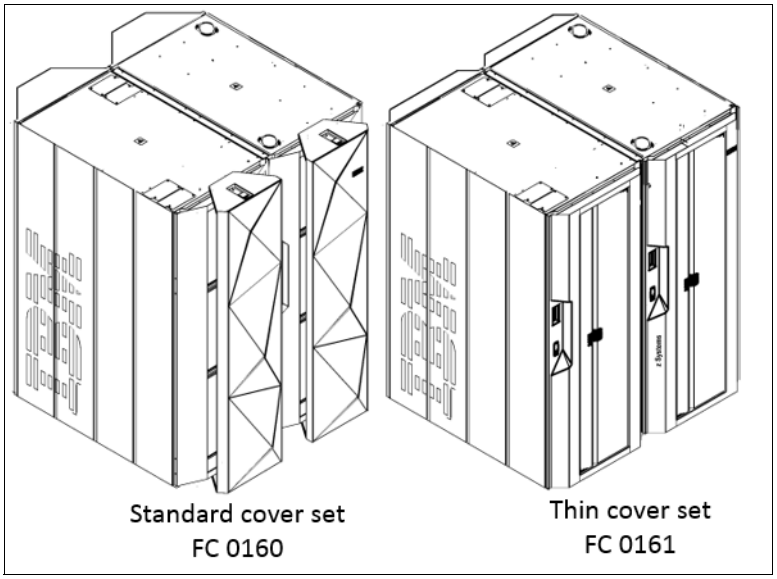


Figure 2-3 Cover sets available with z14

2.1.4 Top exit I/O cabling

Like the z13 and zEC12, the z14 server supports the Top Exit I/O Cabling feature (FC 7942). This feature routes all coupling links and all I/O cables, including a 1000BASE-T Ethernet cable from I/O drawers or PCIe I/O drawers through four frame extensions, out the top of the frame.

In the z14 server, the Top Exit I/O Cabling feature (FC 7942) is available for both radiator-cooled (air-cooled) models and water-cooled models. For more information, see *IBM z14 Installation Manual for Physical Planning,* GC28-6965 and 10.3, “Physical planning” on page 380.

2.1.5 PCIe I/O drawer

Each PCIe I/O drawer (up to 5) has 38 slots each to support the PCIe I/O infrastructure with a bandwidth of 16 GBps.

- 32 I/O cards spread over 4 I/O domains
- Four PCIe switch cards that provide connectivity to the PCIe fanouts installed in the processor drawer. Each domain has a dedicated support partition (4 per system) to manage the native PCIe card types installed in that domain.
- Two Flexible Support Processor (FSP) cards used to control the drawer function
- Redundant N+1 blowers and power supplies

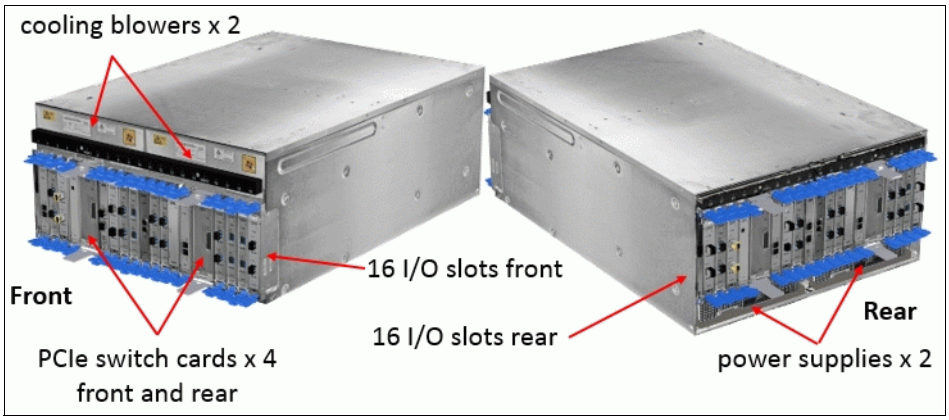


Figure 2-4 PCIe I/O drawer front and rear view

PCIe I/O infrastructure

The PCIe I/O infrastructure uses the PCIe fanout installed in the processor (CPC) drawer(s) to connect to the PCIe I/O drawer, which can contain the following features:

- New features FICON Expresss16+ two port card, long wavelength (LX) or short wavelength (SX), containing two physical channel IDs (PCHIDs)
- FICON Express features (only for a carry-forward MES):
  - FICON Express16S two port card, long wavelength (LX) or short wavelength (SX), containing two PCHIDs
  - FICON Express8S two port card, long wavelength (LX) or short wavelength (SX), containing two PCHIDs
- Open System Adapter (OSA)-Express6S new features:
  - OSA-Express6S 10 Gb Ethernet one port card, Long Reach (LR) or Short Reach (SR), containing one physical channel ID PCHID
  - OSA-Express6S Gb Ethernet two port card, LX or SX, containing one PCHID

  - OSA-Express6S 1000BASE-T Ethernet, two port card, RJ-45, containing one PCHID
- Open System Adapter (OSA)-Express5S and 4S features (only for a carry-forward MES):
  - OSA-Express5S 10 Gb Ethernet (one port card, LR or SR, and one PCHID)
  - OSA-Express5S Gb Ethernet (two port card, LX or SX, and one PCHID)
  - OSA-Express5S and 4S 1000BASE-T Ethernet (two port card, RJ-45, and one CHPID)
- New feature Crypto Express6S. Each 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.
- Crypto Express5S feature (only for a carry-forward MES). Each 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.
- New feature zHyperLink Express is a two port adapter card directly connected to DASD storage designed to increase the scalability of System z transaction processing and lower I/O latency. This feature is installed in the I/O drawer.
- New feature Coupling Express LR (CE LR) is a 2 port feature used for long distance coupling connectivity and utilizes a new coupling channel type: CL5. It utilizes 10GbE RoCE technology and is designed to drive distances up to 10km and support a link data rate of 10 Gigabits per second (Gbps). This feature is installed in the I/O drawer and each port supports up to 4 CHPIDs per PCHID and supports point-to-point connectivity only.
- New feature 10 GbE 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 2-port card and support for up to 63 virtual functions per port / 126 VFs per PCHID (feature).
- 10 GbE RoCE Express - Remote Direct Memory Access (RDMA) over Converged Ethernet (only for a carry-forward MES). It is a 2-port card, and supports for up to 31 virtual functions per adapter.
- zEnterprise Data Compression (zEDC) Express. The zEnterprise Data Compression Express feature occupies one I/O slot, but it does not have a CHPID type. Up to 15 partitions can share the feature concurrently.

2.2 CPC drawer

The z14 server continues the design of z13 by packaging processors in drawers. A z14 CPC drawer contains up to seven single chip modules (SCMs), memory, symmetric multiprocessor (SMP) connectivity, and connectors to support PCIe I/O drawers (through PCIe fanout hubs), and coupling links to other CPCs. The CPC drawers are in the A frame. The z14 server can have up to four CPC drawers installed (the minimum is one drawer). A CPC drawer and its components are shown in Figure 2-5.

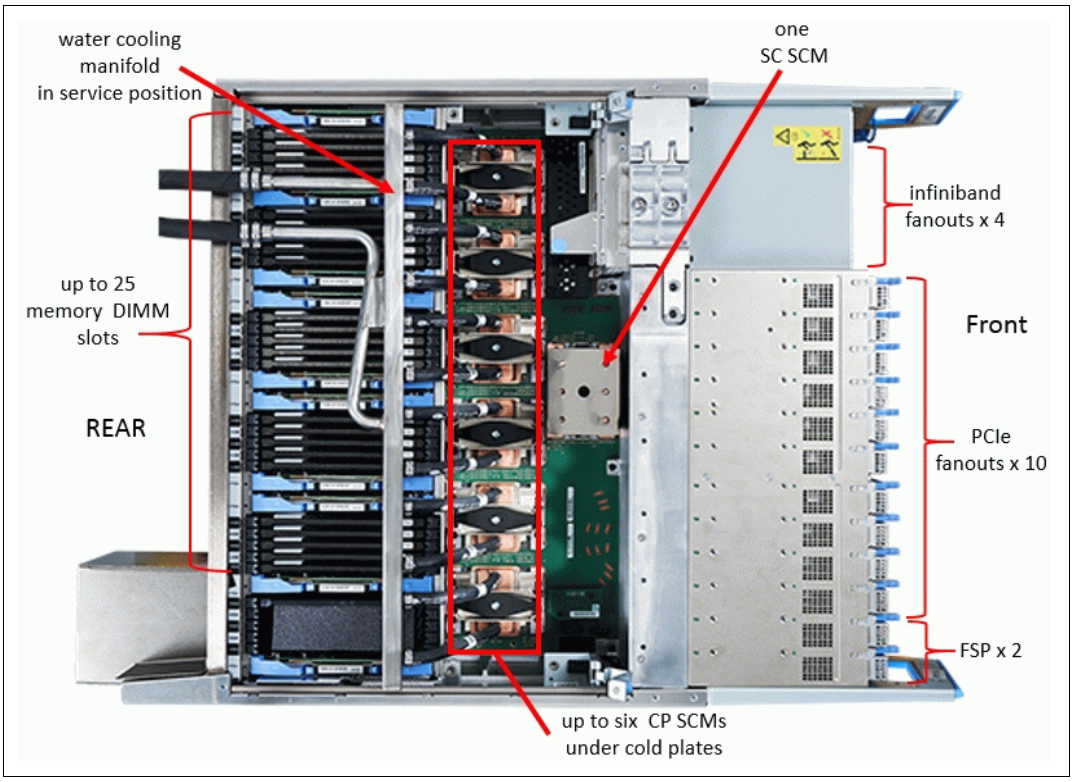


Figure 2-5 CPC drawer components (top view) of a model M05

The z14 CPC drawer is slightly packaged different than z13 and contains the following components:

- Two Drawer Sizes with each containing two logical CP clusters
  - 5 PU SCMs + 1 SC SCM / drawer (41 PUs) – for models M01, M02, M03, M04
  - 6 PU SCMs + 1 SC SCM / drawer (49 PUs) – for model M05
- PU SCM 14nm SOI technology, 17 layers of metal, core running at 5.2GHz
  - 10 PUs / SCM, (7,8,9,10 active core PNs)
- One System Controller (SC) SCM, with a 672 MB L4 cache.
- Five DDR3 dual in-line module (DIMM) slots per memory controller, for a total of up to 25 DIMMs per drawer.
- DIMMs plugged in to 15, 20 or 25 DIMM slots, providing 640 - 40960 GB of physical memory (includes RAIM) and 512 - 32576 GB of addressable memory in a 4 drawer system.
- 10 PCIe Gen3 x16 fanouts (16 GBps bandwidth) per CPC Drawer
  - PCIe Gen3 I/O fanout for PCIe I/O Drawer

  - ICA SR, PCIe fanout
- Four GX++ slots for IFB fanouts (6 GBps bandwidth)
  - HCA3-O, HCA3-O LR
- Two flexible service processor (FSP) cards for system control.
- Two DC converter assemblies (DCAs) that provide power to the CPC drawer. Loss of one DCA leaves enough power to satisfy the drawer's power requirements (n+1 redundancy). The DCAs can be concurrently removed and replaced (one at a time).
- Water-cooling manifold for PU chips.
- Three SMP connectors for inter-drawer connectivity.

Figure 2-6 shows the front view of a CPC drawer, with fanouts slots and connector for water cooling, and the rear view of drawer, with the DIMM slots and DCA connector.

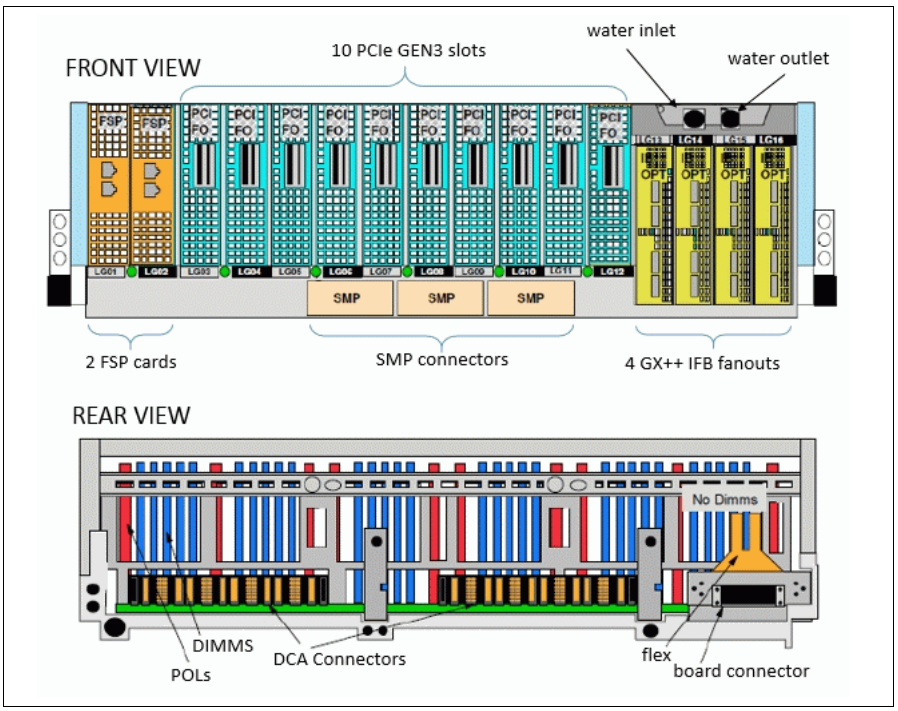


Figure 2-6 Front and rear view of the CPC drawer

Figure 2-7 on page 41 shows the front view of a fully populated processor drawer. Redundant FSP adapters (2) will always be installed, and PCIe I/O fanouts are plugged in specific slots for best performance and availability. SMP interconnect cables are present when multiple drawers are present. Also present are the ICA SR and Infiniband fanouts for coupling connectivity.

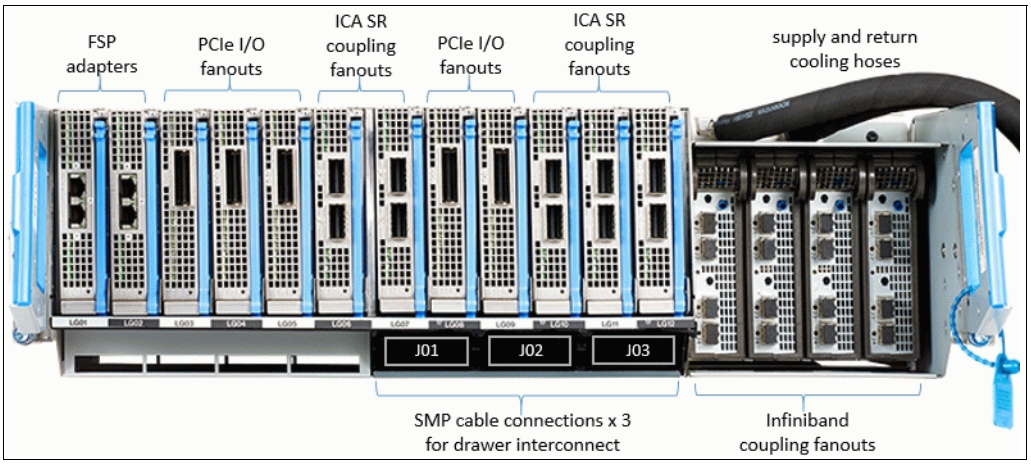


Figure 2-7 Fully populated processor drawer

Figure 2-8 shows the CPC drawer logical structure, component connections (including the CP SCMs), and the storage control (SC) SCMs.

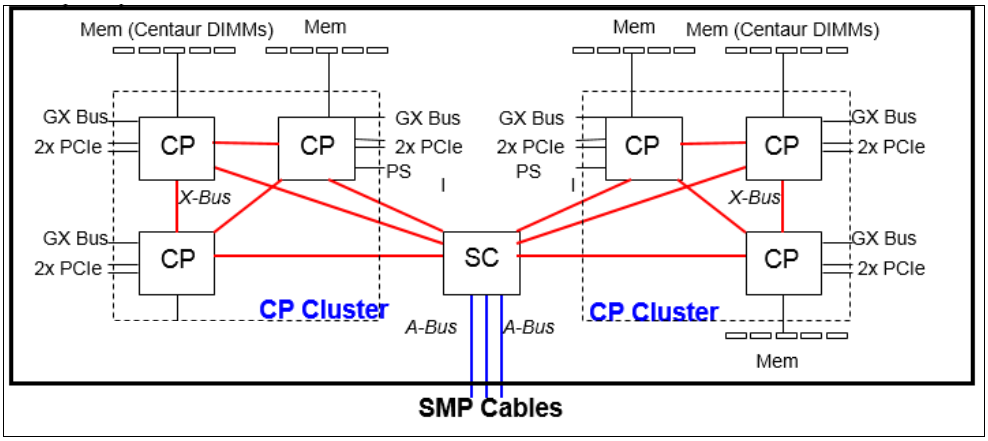


Figure 2-8 CPC drawer logical structure

Memory is connected to the SCMs through memory control units (MCUs). Five MCUs can be placed in a drawer, one per PU SCM, that provide the interface to the controller on memory DIMM. A memory controller drives five DIMM slots.

The buses are organized as follows:

- 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 chip and PUs chips to each other, in the same node.
- The A-bus provides interconnects between SC chips in different drawers using SMP cables.
- Processor support interfaces (PSIs) are used to communicate with FSP cards for system control.

2.2.1 CPC drawer interconnect topology

Figure 2-9 shows the point-to-point topology for CPC drawers and nodes communication. Each CPC drawer communicates directly to all of the other processors drawers in the CPC through two links.

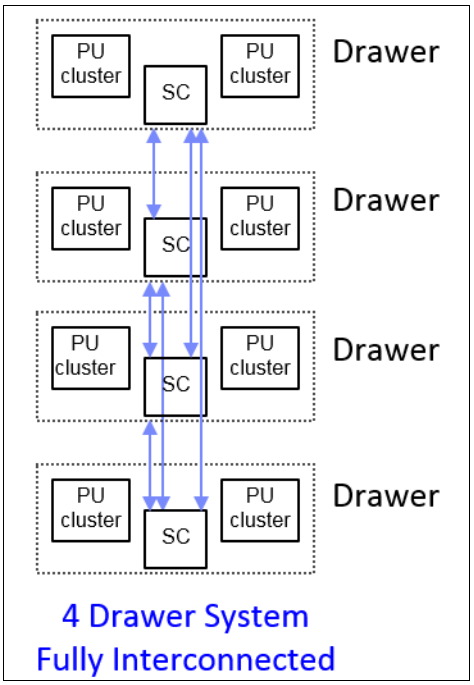


Figure 2-9 Drawer to drawer communication

The CPC drawers are in the Frame A and are populated from bottom to top.

Table 2-1 indicates the order of CPC drawer installation and position in the Frame A.

Table 2-1 CPC drawer installation order and position in Frame A

CPC drawer	CPC drawer 1	CPC drawer 2	CPC drawer 3	CPC drawer 4
Installation order	First	Second	Third	Fourth
Position in Frame A	A15A	A19A	A23A	A27A

CPC drawer installation is concurrent, except for the upgrade to the model M05, which can only be obtained from manufacturing with the model M05 processor drawers already installed. Concurrent drawer repair requires a minimum of two drawers.

2.2.2 Oscillator

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

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

configuration allows time-of-day (TOD) synchronization in a heterogeneous platform environment.

The accuracy of an STP-only CTN is improved by using an NTP server with the PPS output signal as the External Time Source (ETS). NTP server devices with PPS output are available from several vendors that offer network timing solutions. A cable connection from the PPS port on the OSC to the PPS output of the NTP server is required when z14 is using STP and is configured in an STP-only CTN using NTP with PPS as the external time source. The z14 server cannot participate in a mixed CTN. It can participate only in an STP-only CTN.

STP tracks the highly stable and accurate PPS signal from the NTP server and maintains an accuracy of 10 µs to the ETS, as measured at the PPS input of the z14 server.

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

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

Figure 2-10 shows the location of the two OSC cards with BNC connectors for PPS on the CPC, which is beside the drawer 2 and drawer 3 locations.

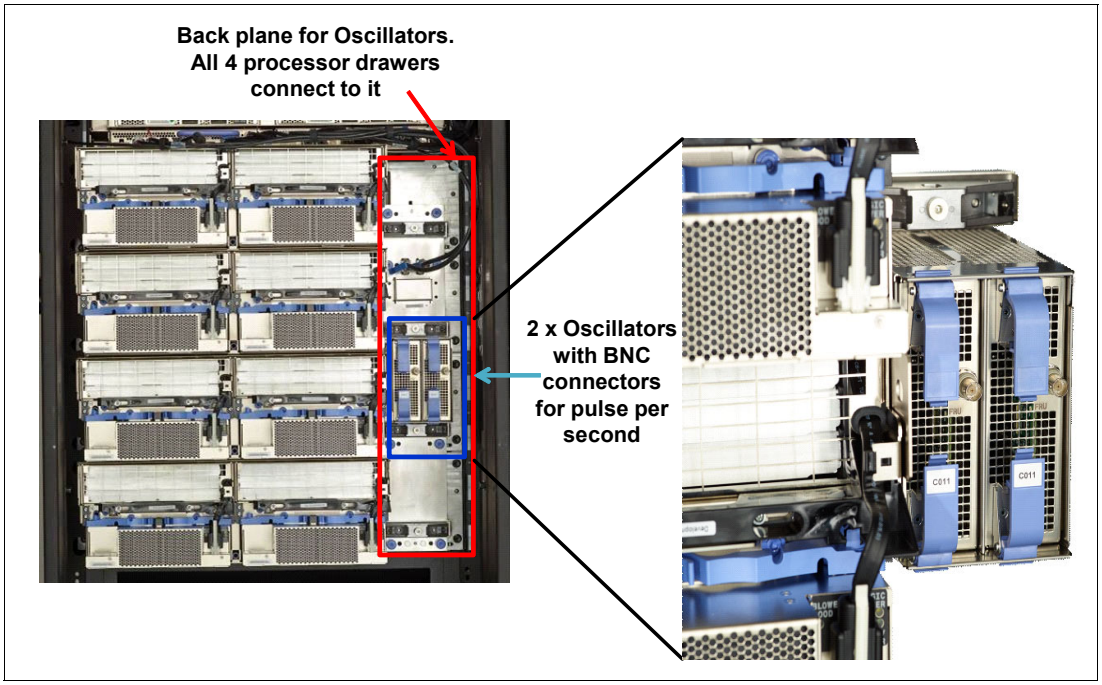


Figure 2-10 Oscillators cards

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

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

2.2.3 System control

The various system elements are managed through the Flexible Support 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 system control hubs (SCH1 and SCH2). The FSPs communicate with the SEs and provide a subsystem interface (SSI) for controlling components.

Figure 2-11 is a conceptual overview of the system control design.

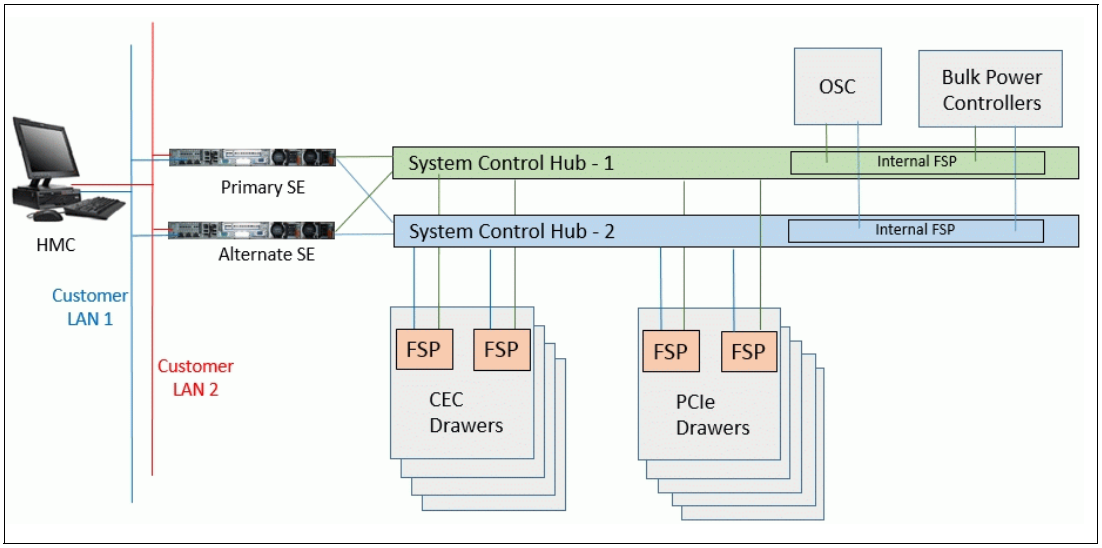


Figure 2-11 Conceptual overview of system control elements

Note: The maximum configuration has four CPC drawers and five PCIe / I/O drawers for z14. The diagram in Figure 2-11 references the various supported FSP connections.

A typical FSP operation is to control a power supply. An SE sends a command to the FSP to start the power supply. The FSP (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.

2.2.4 CPC drawer power

Each CPC drawer gets its power from two DCAs. The DCAs provide the required power for the drawer in an *n*+1 configuration. Loss of one DCA leaves enough power to meet power requirements for the entire drawer. The DCAs can be concurrently serviced, and are accessed from the rear of the frame A.

Figure 2-12 shows the location of DCAs on the backplane of the A frame.

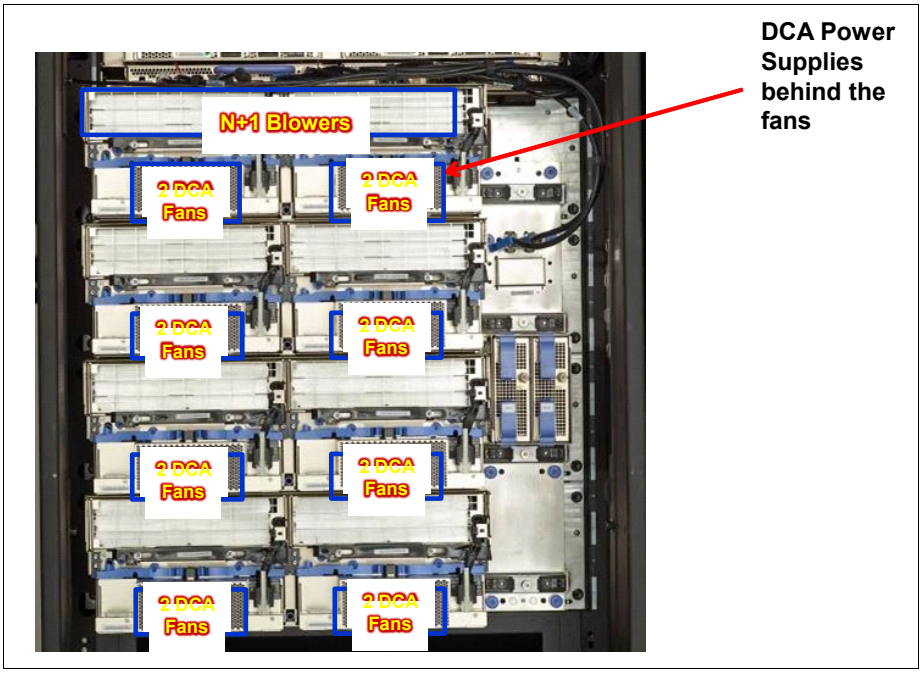


Figure 2-12 DCA power supplies

2.3 Single chip modules

The single chip module (SCM) is a multi-layer metal substrate module that holds either one PU chip or a System Controller (SC) chip. Both CP and SC chip size is 696 mm^2 (25.3 mm x 27.5 mm). Each CPC drawer has six (model M01-M04)) or seven SCMs (model M05), five or six PU SCMs (6.1 billion transistors each), and one SC SCM (9.7 billion transistors).

The two types of SCMs (PU and SC) are shown in Figure 2-13.

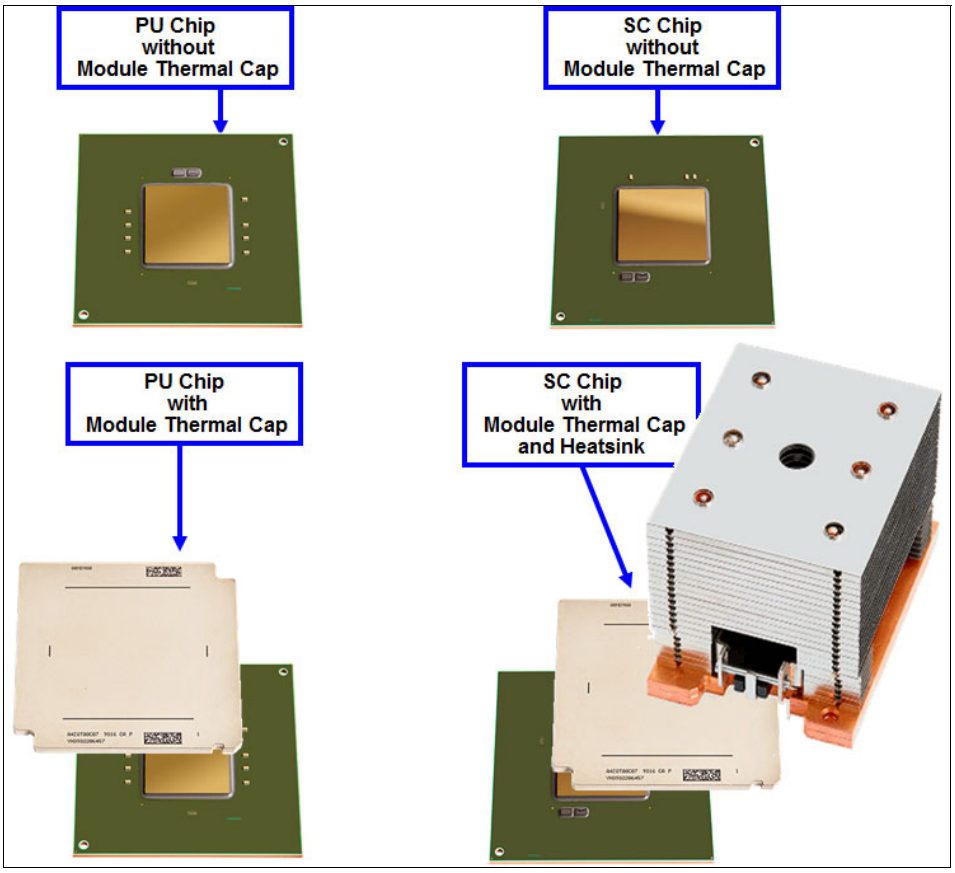


Figure 2-13 Single chip modules (PU SCM and SC SCM)

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

The SCMs are plugged into a socket that is part of the CPC drawer packaging. The interconnectivity between the CPC drawers is accomplished through SMP connectors and cables. There are three inter-drawer connections on each CPC drawer. This configuration allows a multidrawer system to act as a symmetric multiprocessor (SMP) system.

Figure 2-14 depicts the logical cluster structure diagram for a model M05, showing the 6 PU SCMs, the SC SCM and their connections. For models M01-M04, the PU SCM to the far left in Logical Cluster 1 is not present. The number of cores on each of the 5 or 6 populated PU SCMs can range from 7-10 cores.

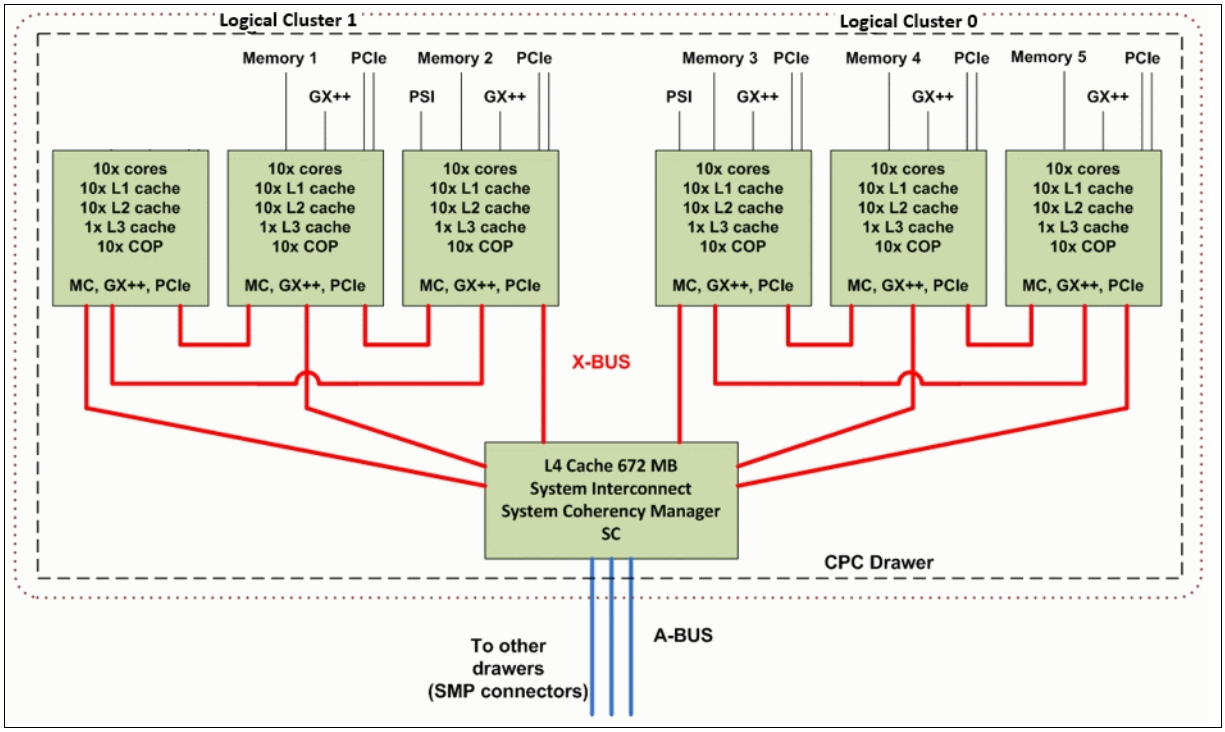


Figure 2-14 Logical drawer structure

2.3.1 Processor unit chip

The z14 PU chip (installed as a PU SCM) is an evolution of the z13 core design. It uses CMOS 14nm SOI technology, out-of-order instruction processing, pipeline enhancements, dynamic improved simultaneous multithreading (SMT), enhanced single-instruction multiple-data (SIMD), and redesigned, larger on-chip caches.

Each PU chip has up to ten active cores that run at 5.2 GHz, which means that the cycle time is 0.192 ns. The PU chips come in four versions: seven, eight, nine or ten active cores. For models M01, M02, M03, and M04, the processor units in each drawer are implemented with 41 active cores per drawer. This configuration means that model M01 has 41, model M02 has 82, mode, M03 has 123, and model M04 has 164 active cores.

Model M05 has 49 active cores per drawer. This configuration provides 196 active cores on model M05. A schematic representation of the PU chip is shown in Figure 2-15.

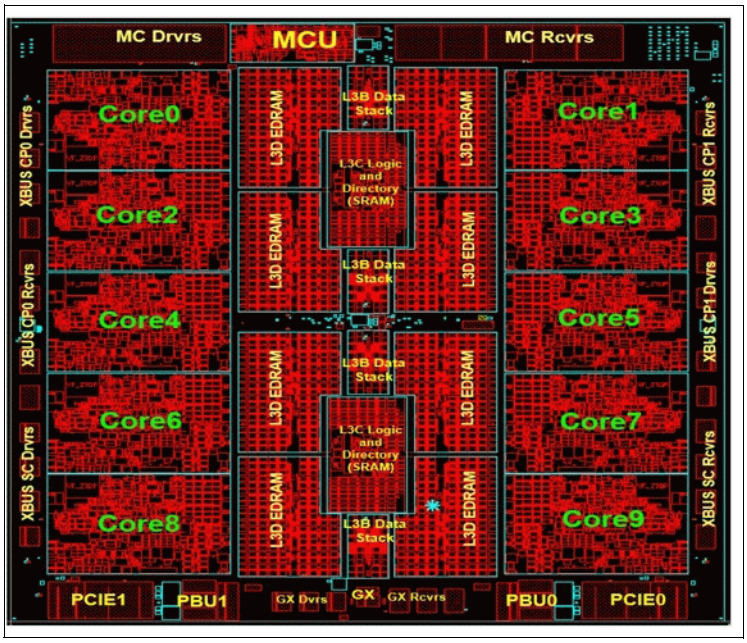


Figure 2-15 PU SCM floor plan

The PU chip contains the following enhancements:

- Cache Improvements:
  - New power efficient logical directory design
  - 33% larger L1 I$ (128K), private
  - 2x larger L2 D$ (4MB), private
  - 2x larger L3 cache with symbol ECC, shared
- New Translation/TLB2 design
  - 4 concurrent translations
  - Reduced latency
  - Lookup integrated into L2 access pipe
  - 2x CRSTE growth
  - 1.5X PTE growth
  - New 64 entry 2gig TLB2
- Pipeline Optimizations
  - Improved instruction delivery
  - Faster branch wake-up
  - Reduced execution latency
  - Improved OSC avoidance
  - Optimized 2nd generation SMT2
- Better Branch Prediction
  - 33% Larger BTB1 & BTB2
  - New Perceptron Predictor
  - New Simple Call Return Stack

2.3.2 Processor unit (core)

Each processor unit (Figure 2-16), or core, is a superscalar and out-of-order processor that has ten execution units:

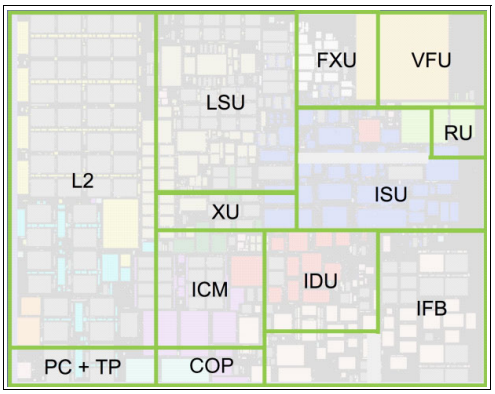


Figure 2-16 Core layout

- 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
- Instruction fetch and branch (IFB) (prediction) and Instruction cache & merge (ICM). These two sub units (IFB and ICM) 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
- Fixed-point unit (FXU): The FXU handles fixed-point arithmetic.
- 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 both floating-point and decimal fixed-point operations and fixed-point division operations.
- Vector execution unit (VXU)
- Level 2 cache (L2), Instructions and data (L2I/L2D)

2.3.3 PU characterization

In each CPC drawer, PUs can be characterized for client use. The characterized PUs can be used in general to run supported operating systems, such as z/OS, z/VM, and Linux on

z Systems. They also can run specific workloads, such as Java, XML services, IPSec, and some DB2 workloads, or functions such as Coupling Facility Control Code (CFCC). For more information about PU characterization, see 3.5, “Processor unit functions” on page 105.

The maximum number of characterized PUs depends on the z14 model. Some PUs are characterized by the system as standard system assist processors (SAPs) to run the I/O processing. By default, there are at least two spare PUs per system that are used to assume the function of a failed PU. The remaining installed PUs can be characterized for client use. A z14 model nomenclature includes a number that represents the maximum number of PUs that can be characterized for client use, as shown in Table 2-2.

Table 2-2 Number of PUs per z14 model

Model	CPs	IFLs	Unassigned IFLs	zIIP	ICFs	IFPs	Std SAPs	Add'l SAPs	Spares
M01	0-33	0-33	0-32	0-22	0-33	1	5	0-4	2
M02	0-69	0-69	0-68	0-46	0-69	1	10	0-8	2
M03	0-105	0-105	0-104	0-70	0-105	1	15	0-12	2
M04	0-141	0-141	0-140	0-94	0-141	1	20	0-16	2
M05	0-170	0-170	0-169	0-112	0-170	1	23	0-16	2

2.3.4 System Controller (SC) chip

The SC chip uses the CMOS 14nm SOI technology, with 17 layers of metal. It measures 25.3 x 27.5 mm, and has 9.7 billion transistors. Each node of the CPC drawer has one SC chip. Figure 2-17 shows a schematic representation of the SC chip.

- X-Bus (CP-CP and CP-SC) – significant changes to allow SC to fit more X-Bus connections
- A Bus (SC-SC off drawer) – minor changes to reflect protocol improvements and new system topology
- 672 MB shared eDRAM L4 Cache
- L4 Directory is built with eDRAM
- New L4 Cache Management
  - Ratio of L3 to L4 cache capacity is increasing
  - New on-drawer Cluster-to-Cluster (topology change) management

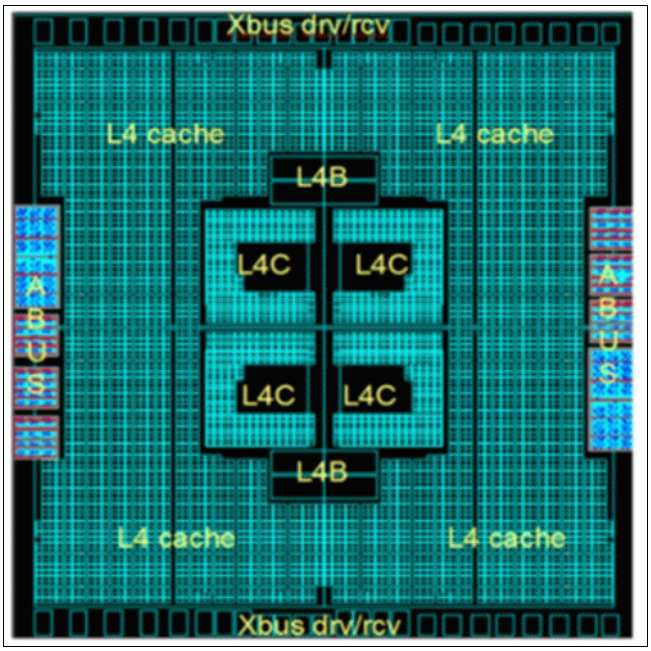


Figure 2-17 SC chip layout

2.3.5 Cache level structure

The cache structure comparison between z13 and z14 is shown in Figure 2-18.

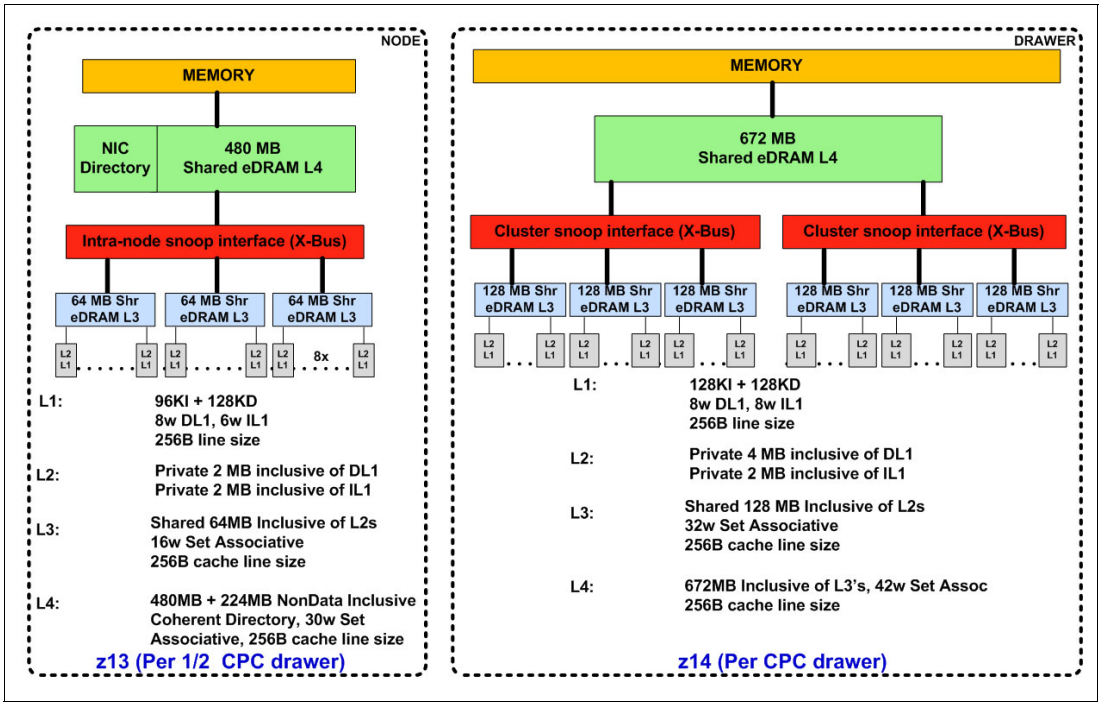


Figure 2-18 Cache structure comparison: z14 vs. z13

2.4 Memory

The maximum physical memory size is directly related to the number of CPC drawers in the system. Each CPC drawer can contain up to 8,000 GB of physical memory, for a total of 32,676 GB (32 TB) of installed (physical) memory per system.

Table 2-3 shows the maximum and minimum memory sizes that you can order for each z13 model.

Table 2-3 z14 memory sizes

Model	Number of CPC drawers	Customer memory (GB)
M01	1	256 - 8000
M02	2	256 - 16192
M03	3	256 - 24384
M04	4	256 - 32576
M05	4	256 - 32576

The minimum physical installed memory is 512 GB per CPC drawer. The minimum initial amount of memory that can be ordered is 256 GB for all z14 models. The maximum customer memory size is based on the physical installed memory minus the RAIM and minus the hardware system area (HSA) memory, which has a fixed amount of 192 GB.

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

Table 2-4 Memory granularity

Granularity (GB)	Customer memory (GB)
64	256 - 576
128	704 - 1600
256	1856 - 3904
512	4416 - 32576

With the z14, the memory granularity varies from 64 GB (for customer memory sizes 256 - 576 GB) up to 512 GB (for CPCs having 4416 - 32576 GB of customer memory).

2.4.1 Memory subsystem topology

The z14 memory subsystem uses high speed, differential-ended communications memory channels to link a host memory to the main memory storage devices.

Figure 2-19 shows an overview of the CPC drawer memory topology of a z14 server.

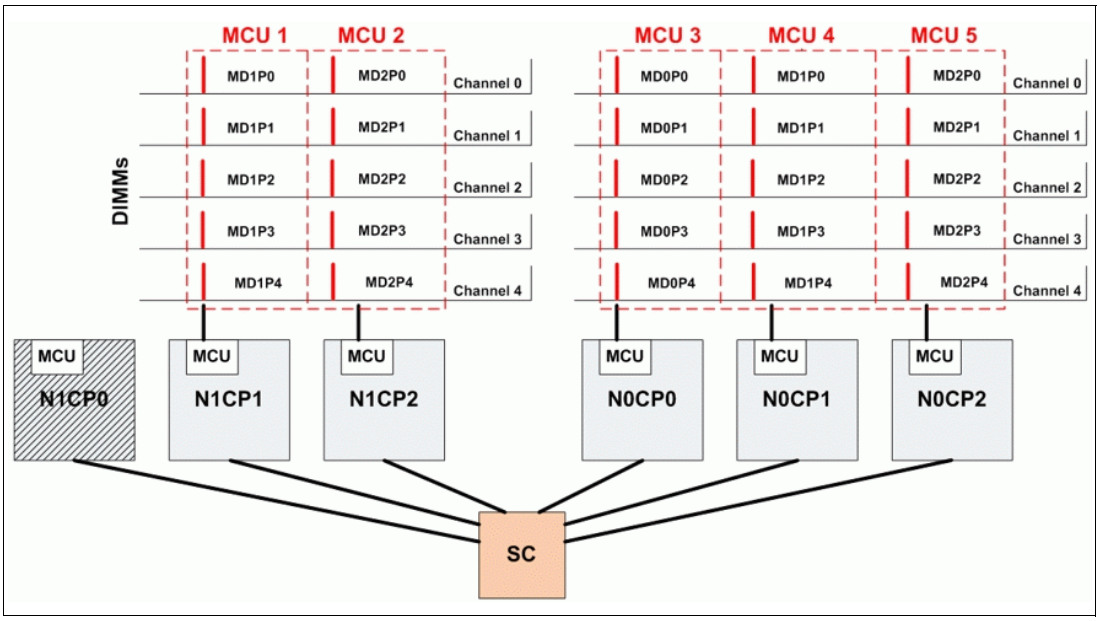


Figure 2-19 CPC drawer memory topology

Each CPC drawer has 15, 20 or 25 DIMMs. DIMMs are connected to each PU chip through the memory control units (MCU). Each PU SCM has one MCU, which uses five channels, one for each DIMM, and one for RAIM implementation, in a 4 +1 design. Each CPC drawer can have three, four or five populated MCUs.

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

2.4.2 Redundant array of independent memory

The z14 server uses the RAIM technology. 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 2-20.

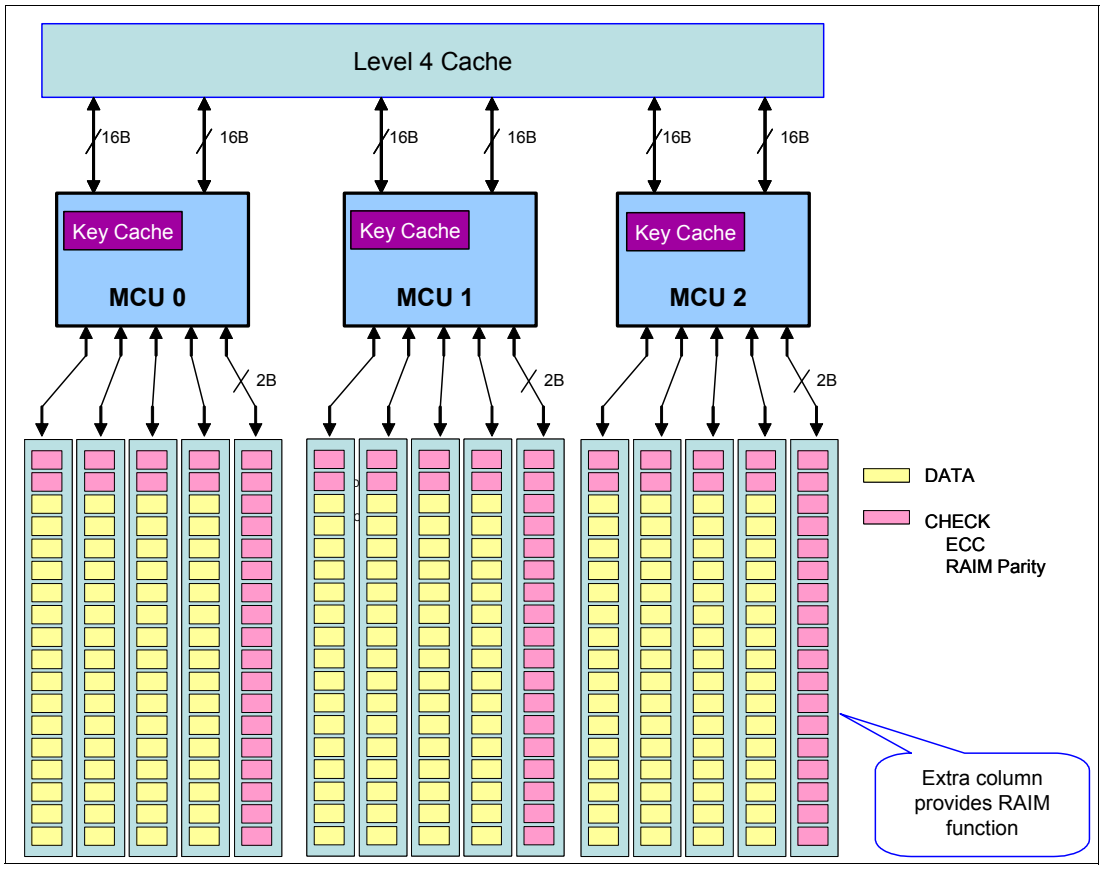


Figure 2-20 RAIM configuration per logical cluster

The fifth channel in each MCU enables memory to be implemented as a Redundant Array of Independent Memory (RAIM). This technology has 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. So, RAIM takes 20% of DIMM capacity. (There is no non-RAIM option.)

The RAIM design provides the following layers of memory recovery:

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

2.4.3 Memory configurations

Memory sizes in each CPC drawer do not have to be similar. Different CPC drawers can contain different amounts of memory. Table 2-5 shows the 15 drawer configurations supporting memory. Total physical memory includes RAIM (20%) and HSA (192 GB).

Table 2-5 Physically installed memory drawer configurations

	Phys. mem.	32GB #DIMMs	64GB #DIMMs	128GB #DIMMs	256GB #DIMMs	512GB #DIMMs	-RAIM GB	-HSA GB	HDW Dialed increments
1	640	10	5	0	0	0	512	320	
2	960	0	15	0	0	0	768	576	256
3	1280	0	10	5	0	0	1024	832	256
4	1600	0	5	10	0	0	1280	1088	256
5	1920	0	0	15	0	0	1536	1344	256
6	2240	0	5	15	0	0	1792	1600	256
7	2560	0	0	20	0	0	2048	1856	256
8	3200	0	0	25	0	0	2560	2368	512
9	3840	0	0	20	5	0	3072	2880	512
10	4480	0	0	15	10	0	3584	3392	512
11	5120	0	0	10	15	0	4096	3904	512
12	6400	0	0	0	25	0	5120	4928	1024
13	7680	0	0	0	20	5	6144	5952	1024
14	8960	0	0	0	15	10	7168	6976	1024
15	10240	0	0	0	10	15	8192	8000	1024

- A CPC drawer always contains a minimum of 10 32 GB DIMMs and 5 64 GB DIMMS (640 GB) as seen in drawer configuration number 1 in Table 2-5.
- A CPC drawer can have more memory that is installed than enabled. The amount of memory that can be enabled by the client is the total physically installed memory minus the RAIM amount (20%) and minus the 192 GB of HSA memory.
- A CPC drawer can have available unused memory, which can be ordered as a memory upgrade and enabled by LIC without DIMM changes.
- DIMM changes require a disruptive power-on reset (POR) on z14 model M01. DIMM changes are always done concurrently on z14 models with multiple CPC drawers that use Enhanced Drawer Availability (EDA).

DIMM plugging for the configurations in each CPC drawer do not have to be similar. Each memory 5 slot DIMM bank must have the same DIMM size, however a drawer can have a mix of DIMM banks. See Table 2-6 for the memory population by DIMM bank for the 15 configurations shown in Table 2-5).

Table 2-6 Memory population by DIMM Bank (one CPC drawer)

	MD01-MD05	MD06-MD10	MD11-MD15	MD16-MD20	MD21-MD25	MD26-MD30	Total -RAIM
1			64	32	32		512
2			64	64	64		768
3			10	64	64		1024
4			128	128	64		1280
5			128	128	128		1536
6		64	128	128	128		1792
7		128	128	128	128		2048
8		128	128	128	128	128	2560
9		256	128	128	128	128	3072
10		256	128	128	256	128	3584
11		256	256	128	256	128	4096
12		256	256	256	256	256	5120
13		512	256	256	256	256	6144
14		512	256	256	512	256	7168
15		512	512	256	512	256	8192

- The support element *View Hardware Configuration* task can be used to determine the size and quantity of the memory plugged in each drawer. Figure 2-21 show an example of the location and description of memory modules installed.

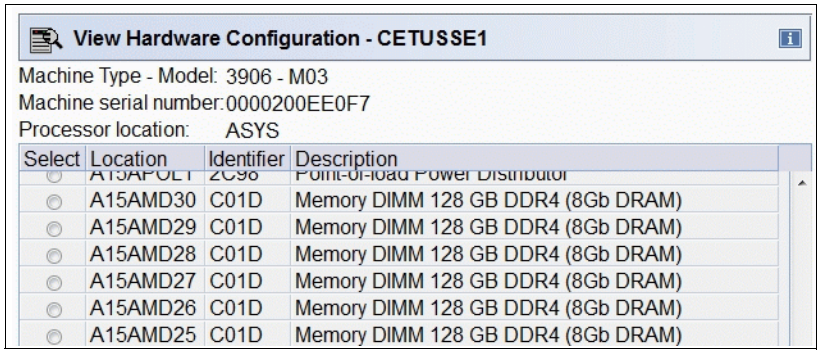


Figure 2-21 View Hardware Configuration task on the Support Element

Figure 2-22 on page 57 displays the CPC drawer and DIMM locations for a z14

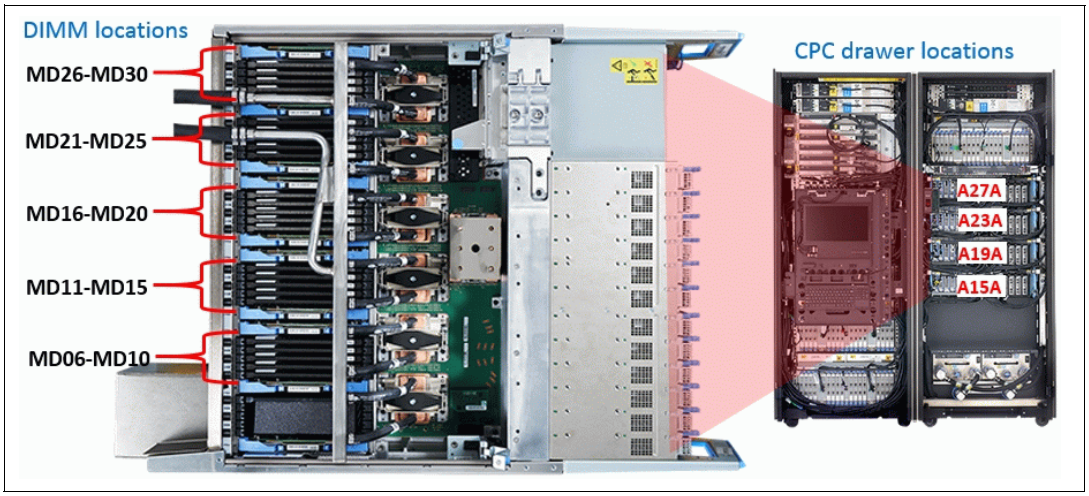


Figure 2-22 CPC drawer and DIMM locations

Table 2-7 contains the physical memory DIMM plugging configurations by feature code from manufacturing when the system is ordered. The drawer column(s) for the specific model will contain the memory configuration number for the specific drawer. Use available unused memory that can be enabled by LIC when required.

- If additional storage is ordered via other feature codes such as Virtual Flash Memory, Flexible Memory and Preplanned memory, the additional storage is installed and plugged as necessary.

Table 2-7 System memory configurations

			M01		M02			M03				M04 / M05				
FC	Inc	Memory Inc GB	DRW1	Dial Max	DRW1	DRW2	Dial Max	DRW1	DRW2	DRW3	Dial Max	DRW1	DRW2	DRW3	DRW4	Dial Max
1660	64	256	1	320	1	1	832	1	1	1	1344	1	1	1	1	1856
1661		320	1	320	1	1	832	1	1	1	1344	1	1	1	1	1856
1662		384	2	576	1	1	832	1	1	1	1344	1	1	1	1	1856
1663		448	2	576	1	1	832	1	1	1	1344	1	1	1	1	1856
1664		512	2	576	1	1	832	1	1	1	1344	1	1	1	1	1856
1665		576	2	576	1	1	832	1	1	1	1344	1	1	1	1	1856
1666		704	3	832	1	1	832	1	1	1	1344	1	1	1	1	1856
1667		832	3	832	1	1	832	1	1	1	1344	1	1	1	1	1856
1668		960	4	1088	1	2	1088	1	1	1	1344	1	1	1	1	1856
1669		1088	4	1088	1	2	1088	1	1	1	1344	1	1	1	1	1856
1670		1216	5	1344	1	2	1344	1	1	1	1344	1	1	1	1	1856

FC	Inc	Memory Inc GB	M01		M02			M03				M04 / M05				
			DRW1	Dial Max	DRW1	DRW2	Dial Max	DRW1	DRW2	DRW3	Dial Max	DRW1	DRW2	DRW3	DRW4	Dial Max
1671		1344	5	1344	2	2	1344	1	1	1	1344	1	1	1	1	1856
1672		1472	6	1600	2	3	1600	1	1	2	1600	1	1	1	1	1856
1673		1600	6	1600	2	3	1600	1	1	2	1600	1	1	1	1	1856
1674		1856	7	1856	3	3	1856	2	1	2	1856	1	1	1	1	1856
1675		2112	8	2368	3	4	2112	2	2	2	2112	1	1	1	2	2112
1676		2368	8	2368	4	4	2368	2	2	3	2368	2	1	1	2	2368
1677		2624	9	2880	4	5	2624	3	2	3	2624	2	1	2	2	2624
1678		2880	9	2880	5	5	2880	3	3	3	2880	2	2	2	2	2880
1679		3136	10	3392	5	6	3136	3	3	4	3136	2	2	2	3	3136
1680		3392	10	3392	6	6	3392	4	3	4	3392	3	2	2	3	3392
1681		3648	11	3904	6	7	3648	4	4	4	3648	3	2	3	3	3648
1682		3904	11	3904	7	7	3904	4	4	5	3904	3	3	3	3	3904
1683		4416	12	4928	7	8	4416	5	5	5	4416	4	3	3	4	4416
1684		4928	12	4928	8	8	4928	6	5	6	4928	4	4	4	4	4928
1685		5440	13	5952	8	9	5440	6	6	7	5440	5	4	4	5	5440
1686		5952	13	5952	9	9	5952	7	7	7	5952	5	5	5	5	5952
1687		6464	14	6976	9	10	6464	7	7	8	6464	6	5	5	6	6464
1688		6976	14	6976	10	10	6976	8	7	8	6976	6	6	6	6	6976
1689		7488	15	8000	10	11	7488	8	8	8	7488	7	6	6	7	7488
1690		8000	15	8000	11	11	8000	8	8	9	8000	7	7	7	7	8000
1691		8512			11	12	9024	9	8	9	8512	7	7	7	8	8512
1692		9024			11	12	9024	9	9	9	9024	8	7	7	8	9024
1693		9536			12	12	10048	9	9	10	9536	8	7	8	8	9536
1694		10048			12	12	10048	10	9	10	10048	8	8	8	8	10048
1695		10560			12	13	11072	10	10	10	10560	8	8	8	9	10560
1696		11072			12	13	11072	10	10	11	11072	9	8	8	9	11072
1697		11584			13	13	12096	11	10	11	11584	9	8	9	9	11584
1698		12096			13	13	12096	11	11	11	12096	9	9	9	9	12096
1699		12608			13	14	13120	11	11	12	13120	9	9	9	10	12608
1700		13120			13	14	13120	11	11	12	13120	10	9	9	10	13120
1701		13632			14	14	14144	12	11	12	14144	10	9	10	10	13632
1702		14144			14	14	14144	12	11	12	14144	10	10	10	10	14144
1703		14656			14	15	15168	12	12	12	15168	10	10	10	11	14656
1704		15168			14	15	15168	12	12	12	15168	11	10	10	11	15168
1705		15680			15	15	16192	12	12	13	16192	11	10	11	11	15680

FC	Inc	Memory Inc GB	M01		M02			M03				M04 / M05				
			DRW1	Dial Max	DRW1	DRW2	Dial Max	DRW1	DRW2	DRW3	Dial Max	DRW1	DRW2	DRW3	DRW4	Dial Max
1706		16192			15	15	16192	12	12	13	16192	11	11	11	11	16192
1707		16704						13	12	13	17216	11	11	11	12	17216
1708		17216						13	12	13	17216	11	11	11	12	17216
1709		17728						13	13	13	18240	12	11	11	12	18240
1710		18240						13	13	13	18240	12	11	11	12	18240
1711		18752						13	13	14	19264	12	11	12	12	19264
1712		19264						13	13	14	19264	12	11	12	12	19264
1713		19776						14	13	14	20288	12	12	12	12	20288
1714		20288						14	13	14	20288	12	12	12	12	20288
1715		20800						14	14	14	21312	12	12	12	13	21312
1716		21312						14	14	14	21312	13	12	12	13	21312
1717		21824						14	14	15	22336	13	12	12	13	22336
1718		22336						14	14	15	22336	13	12	12	13	22336
1719		22848						15	14	15	23360	13	12	13	13	23360
1720		23360						15	14	15	23360	13	12	13	13	23360
1721		23872						15	15	15	24384	13	13	13	13	24384
1722		24384						15	15	15	24384	13	13	13	13	24384
1723		24896										13	13	13	14	25408
1724		25408										13	13	13	14	25408
1725		25920										14	13	13	14	26432
1726		26432										14	13	13	14	26432
1727		26944										14	13	14	14	27456
1728		27456										14	13	14	14	27456
1729		27968										14	14	14	14	28480
1730		28480										14	14	14	14	28480
1731		28992										14	14	14	15	29504
1732		29504										14	14	14	15	29504
1733		30016										15	14	14	15	30528
1734		30528										15	14	14	15	30528
1735		31040										15	14	15	15	31552
1736		31552										15	14	15	15	31552
1737		32064										15	15	15	15	32576
1738		32576										15	15	15	15	32576

2.4.4 Memory upgrades

Memory upgrades can be ordered and enabled by LIC, by upgrading the DIMM cards, by adding new DIMM cards, or by adding a CPC drawer.

For a model upgrade that results in the addition of a CPC drawer, the minimum memory increment is added to the system. Each CPC drawer has a minimum physical memory size of 320 GB.

During a model upgrade, adding a CPC drawer is a concurrent operation. Adding physical memory to the added drawer is also concurrent. If all or part of the added memory is enabled for use, it might become available to an active LPAR if the partition has reserved storage that is defined. For more information, see 3.7.3, “Reserved storage” on page 129. Alternatively, the added memory can be used by an already-defined LPAR that is activated after the memory addition.

Note: Memory downgrades are always disruptive. Model downgrades (removal of a CPC drawer) are not supported.

2.4.5 Drawer replacement and memory

With EDA, which is supported for z14 servers, sufficient resources must be available to accommodate resources that are rendered unavailable when a CPC drawer is removed for upgrade or repair. For more information, see 2.6.2, “Enhanced drawer availability” on page 67.

Most of the time, the removal of a CPC drawer results in the removal of active memory. With the flexible memory option, evacuating the affected memory and reallocating its use elsewhere in the system are possible. For more information, see 2.4.7, “Flexible Memory Option” on page 61. This process requires more available memory to compensate for the memory that is lost with the removal of the drawer.

2.4.6 Virtual Flash Memory

IBM Virtual Flash Memory (VFM) is the replacement for the Flash Express features (#0402,#0403) which were available on the IBM zEC12 and IBM z13. No application changes are required to change from IBM Flash Express to VFM. 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[1]. 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.

VFM can also be used in coupling facility images to provide extended capacity and availability for workloads making use of Websphere MQ Shared Queues structures. Using 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 therefore help organizations meet their most demanding service level agreements and compete more effectively. VFM is designed to be easy to configure, and to provide rapid time to value.

[1] z/OS V1.13 has additional requirements. See the Software Requirements section.

2.4.7 Flexible Memory Option

With the Flexible Memory Option, more physical memory is supplied to support the activation of the actual purchased memory entitlement in a single CPC drawer that is out of service during activation (POR), or in a scheduled concurrent drawer upgrade (memory add) or drawer maintenance (n+1 repair) with the use of enhanced drawer availability.

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

Flexible memory is available only on the M02, M03, M04, and M05 models. Table 2-8 shows the flexible memory sizes that are available for the z14.

- Standard Memory – Provides minimum physical memory required to hold customer purchase memory plus 192 GB HSA.
- Flexible Memory – Provides additional physical memory needed to support activation base customer memory and HSA on a multiple CPC drawer z14 with one drawer out of service.

Table 2-8 z14 (customer usable) memory sizes

Model	Standard memory (GB)	Flexible memory (GB)
M01	320 - 8000	N/A
M02	320 - 16192	320 - 8000
M03	320 - 24384	320 - 16192
M04	320 - 32576	320 - 24384
M05	320 - 32576	320 - 24384

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

2.4.8 Pre-planned memory

Pre-planned memory provides the capability for concurrent and permanent memory upgrades by changing the Licensed Internal Code Configuration Control (LICCC) without using EDA. It differs from the flexible memory option. The flexible memory option is meant to anticipate nondisruptive CPC drawer replacement. Therefore, the usage of flexible memory is temporary, in contrast with plan-ahead memory.

When you are preparing for a future memory upgrade, memory can be pre-plugged, based on a target capacity. The pre-planned memory is available for all z14 models, and can be ordered with flexible memory on a multiple drawer z14 model. The pre-plugged memory can be made available through a LICCC update.

You can order this LICCC through these channels:

- The IBM Resource Link® (a login is required):

  http://www.ibm.com/servers/resourcelink/
- Your IBM representative

The installation and activation of any pre-planned memory requires the purchase of the required feature codes (FCs), as shown in Table 2-9.

Table 2-9 Feature codes for plan-ahead memory

Memory	z14 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 1990 - 32 GB ► FC 1991 - 64 GB
Virtual Flash Memory (VFM) Preplanned 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 1898/1938[a] ► FC 1893/1939[a] ► FC 1894/1940[a] These are 32 GB, 64 GB and 256 GB memory capacity Increments respectively

a. (Main memory <1TB) / (main memory >1TB).

The payment for plan-ahead memory is a two-phase process. One charge takes place when the plan-ahead memory is ordered, and another charge takes place when the prepaid memory is activated for actual usage. For the exact terms and conditions, contact your IBM representative.

Installation of pre-planned memory is done by ordering FC 1990 (32 GB) or FC 1991 (64 GB). The ordered amount of plan-ahead memory is charged at a reduced price compared to the normal price for memory. One FC 1990 is needed for each 32 GB of usable memory (40 GB RAIM), or one FC 1991 is needed for each 64 GB of usable memory (80 GB RAIM).

If VFM is present, it is included in the Flexible memory calculations. Normal plan ahead memory increments will be consumed first with the normal 'feature conversion' action. When normal plan ahead features are totally consumed, the VFM 1.5TB plan ahead increment will be deleted and normal 32GB memory increments will be added. NO feature conversions will be allowed from the 1.5TB VFM plan ahead increment to regular memory, plan ahead memory or VFM memory.

Activation of installed pre-planned memory is achieved by ordering FC 1989/1938, FC 1893/1939, or FC 1894/1940 which causes the other portion of the previously contracted charge price to be invoiced.

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

2.5 Reliability, availability, and serviceability

IBM Z servers continue to deliver enterprise class RAS with IBM z14 servers. The main philosophy behind RAS is about preventing or tolerating (masking) outages. It is also about providing the necessary instrumentation (in hardware, LIC/microcode, and software) to capture (collect) the relevant failure information to help identify an issue without requiring a reproduction of the event. These outages can be planned or unplanned. Planned and

unplanned outages can include the following situations (examples are not related to the RAS features of IBM Z servers):

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

The IBM Z hardware has decades of intense engineering behind it, which has resulted in a robust and reliable platform. The hardware has many RAS features that are built into it. For more information see also Chapter 9, “Reliability, availability, and serviceability” on page 345.

2.5.1 RAS in the CPC memory subsystem

Patented error correction technology in the memory subsystem continues to provide the most robust error correction from IBM to date. Two full DRAM failures per rank can be spared and a third full DRAM failure can be corrected. DIMM level failures, including components such as the memory controller application-specific integrated circuit (ASIC), the power regulators, the clocks, and the system board can be corrected. Memory channel failures, such as signal lines, control lines, and drivers/receivers on the MCM, can be corrected. Upstream and downstream data signals can be spared by using two spare wires on both the upstream and downstream paths. One of these signals can be used to spare a clock signal line (one upstream and one downstream). The following improvements were also added in the z14 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.

2.5.2 General z14 RAS features

The z14 server has the following RAS features:

- The z14 server provides a true N+1 cooling function for the radiator-cooled (air-cooled) model and a true N+1 (fully redundant) cooling function for the water-cooled model.
- The power supplies for the z14 server are also based on the N+1 design. The second power supply can maintain operations and avoid an unplanned outage of the system.
- The IBM Z processors have improved chip packaging (encapsulated chip connectors) and use SER hardened latches throughout the design.
- There is N+2 point-of-load (POL) power conversion. This redundancy protects the processor from the loss of voltage because of POL failures.
- There is N+2 redundancy on the environmental sensors (ambient temperature, relative humidity, air density[2], and corrosion).
- Enhanced bus structure using integrated time domain reflectometry (TDR) technology.
- There are Peripheral Component Interconnect Express (PCIe) service enhancements:
  - Mandatory end-to-end cyclic redundancy check (ECRC)
  - Customer operation code separate from maintenance code
  - Native PCIe firmware stack running on the integrated firmware processor (IFP) to manage isolation and recovery

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

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

For more information, see Chapter 9, “Reliability, availability, and serviceability” on page 345.

2.6 Connectivity

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

Figure 2-23 shows the location of the fanouts for a four CPC drawer system. There are 10 PCIe fanout slots and four IFB fanout slots per CPC drawer. Each CPC drawer has two FSPs for system control. LGXX is the location code.

Figure 2-23 Location of the PCIe and IFB fanouts

Up to 10 PCIe fanouts (LG03 - LG12) and four IFB fanouts (LG13 - LG16) can be installed in each CPC drawer.

A fanout can be repaired concurrently with the use of redundant I/O interconnect. For more information, see 2.6.1, “Redundant I/O interconnect (RII)” on page 65.

Five 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 two or z14 or z13 servers.
- Host Channel Adapter (HCA3-O (12xIFB)): This optical fanout provides 12x InfiniBand coupling link connectivity up to 150 meters (492 ft.) distance to z14, z13s, z13, zEC12, zBC12, z196, and z114 servers.
- Host Channel Adapter (HCA3-O LR (1xIFB)): This optical long range fanout provides 1x InfiniBand coupling link connectivity up to a 10 km (6.2 miles) unrepeated (or 100 km (62 miles) when extended by using IBM Z qualified dense wavelength division multiplexing (DWDM) equipment) distance to z14, z13s, z13, zEC12, zBC12, z196, and z114 servers.

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

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

2.6.1 Redundant I/O interconnect (RII)

Redundancy is provided for PCIe I/O interconnects.

The PCIe I/O drawer supports up to 32 PCIe features. They are organized in four hardware domains per drawer, as shown in Figure 2-24.

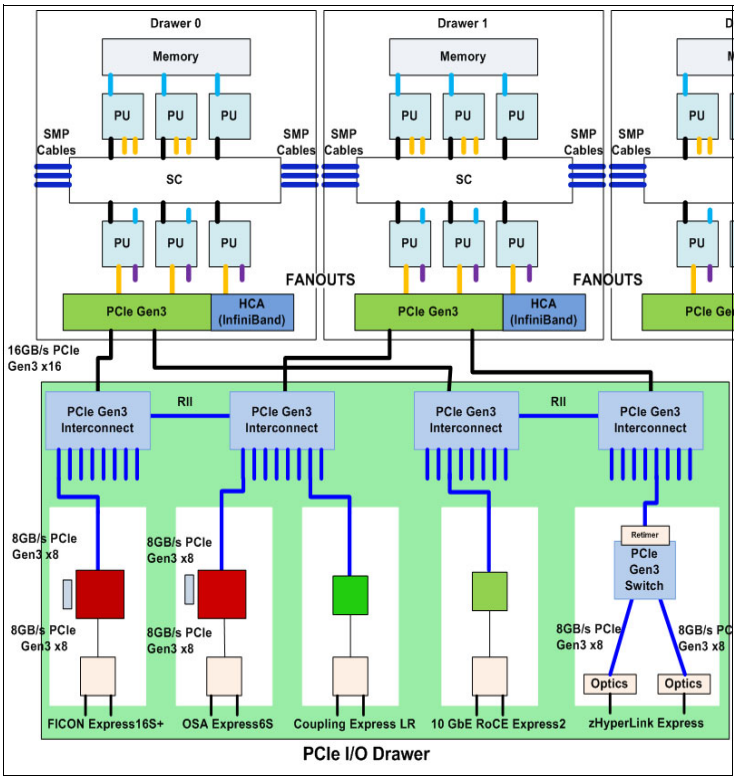


Figure 2-24 Redundant I/O interconnect for PCIe I/O drawer

Each domain is driven through a PCIe Gen3 Interconnect switch adapter. The two PCIe switch cards provide a backup path for each other through the passive connection in the PCIe I/O drawer backplane. During a PCIe fanout or cable failure, all 16 PCIe features in the two domains can be driven through a single PCIe switch card (Figure 2-25).

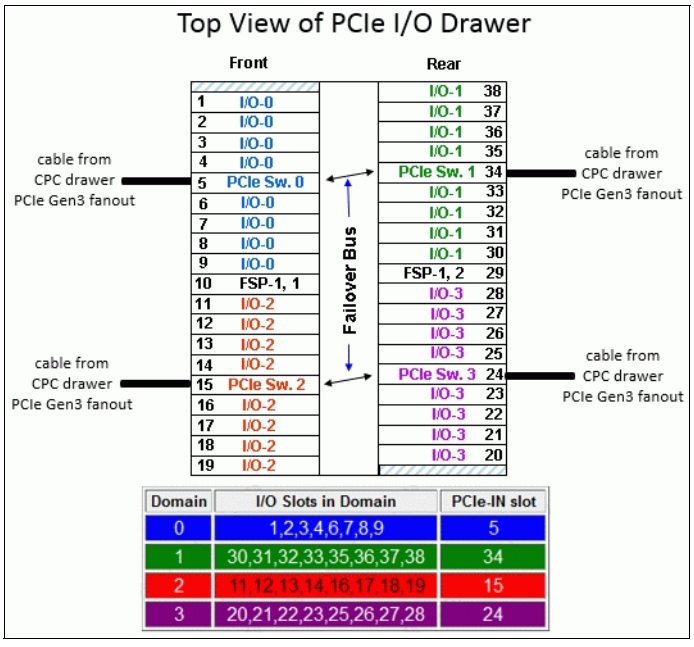


Figure 2-25 Redundant I/O Interconnect

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

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

2.6.2 Enhanced drawer availability

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

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

Removal of a CPC drawer also removes the CPC drawer connectivity to the I/O drawers, PCIe I/O drawers, and coupling links. The effect of the removal of the CPC drawer on the system is limited by the use of redundant I/O interconnect. For more information, see 2.6.1, “Redundant I/O interconnect (RII)” on page 65. However, all PSIFB or ICA-SR links on the removed CPC drawer must be configured offline.

If the enhanced drawer availability and flexible memory options are *not* used when a CPC drawer needs to be replaced, the memory in the failing drawer is also removed. This process might be necessary during an upgrade or a repair action. Until the removed CPC drawer is replaced, a power-on reset of the system with the remaining CPC drawer(s) is supported. The CPC drawer can then be replaced and added back into the configuration concurrently.

2.6.3 CPC drawer upgrade

All fanouts that are used for I/O and coupling links are rebalanced concurrently as part of a CPC drawer addition to support better RAS characteristics.

2.7 Model configurations

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

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

Capacity marked CP A processor that is purchased for future use as a CP is marked as available capacity. It is offline and not available for use until an upgrade for the CP is installed. It does not affect software licenses or maintenance charges.

IFL	The Integrated Facility for Linux (IFL) is a processor that is purchased and activated for use by z/VM for Linux guests and Linux on z Systems 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 restricted to DB2 type applications.
Additional SAP	An optional processor that is purchased and activated for use as an SAP.

A minimum of one PU that is characterized as a CP, IFL, or ICF is required per system. The maximum number of CPs, IFLs, and ICFs is 170. The maximum number of zIIPs is always up to twice the number of PUs that are characterized as CPs.

Not all PUs on a model must be characterized.

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

- SAP to be used by the channel subsystem. The number of predefined SAPs depends on the z14 model.
- One IFP, which is used in the support of designated features, such as zEDC, RoCE, RoCE-2, zHyperLink Express, ISM (Internal Shared Memory SMC-D), and RCE
- Two spare PUs, which can transparently assume any characterization during a permanent failure of another PU.

The z14 model nomenclature is based on the number of PUs that are available for client use in each configuration. The models are summarized in Table 2-10.

Table 2-10 z14 processor configurations

Model	Drawers / PUs	CPs	IFLs uIFL	ICFs	zIIPs	Add. SAPs	Std. SAPs	Spares	IFP
M01	1/33	0 - 33	0 - 33 0 - 32	0 - 33	0 - 22	0 - 4	5	2	1
M02	2/69	0 - 69	0 - 69 0 - 68	0 - 69	0 - 46	0 - 8	10	2	1
M03	3/105	0 - 105	0 - 105 0 - 104	0 - 105	0 - 70	0 - 12	15	2	1
M04	4/141	0 - 141	0 - 141 0 - 140	0 - 141	0 - 94	0 - 16	20	2	1
M05	4/170	0 - 170	0 - 170 0 - 169	0 - 170	0 - 112	0 - 16	23	2	1

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

is used for charging monthly license charge (MLC) software, or when charged on a per-processor basis.

2.7.1 Upgrades

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

Spare PUs are used to replace defective PUs. There are always two spare PUs on a z14 server. In the rare event of a PU failure, a spare PU is activated concurrently and transparently, and assigned the characteristics of the failing PU.

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

Although upgrades from one z14 model to another z14 model are concurrent, meaning that one or more CPC drawers can be added, there is one exception. Upgrades from any z14 server (model M01, M02, M03, or M04) to a model M05 is not supported. M05 model is factory only.

Table 2-11 shows the possible upgrades within the z14 configuration range.

Table 2-11 z14 to z14 upgrade paths

To 3906 From 3906	Model M01	Model M02	Model M03	Model M04	Model M05
Model M01	-	Yes	Yes	Yes	No
Model M02	-	-	Yes	Yes	No
Model M03	-	-	-	Yes	No
Model M04	-	-	-	-	No

You can also upgrade a IBM zEnterprise zEC12 (2827) or a IBM z13 (2964) to a z14 server and preserve the CPC serial number (S/N). The I/O cards can also be carried forward (with certain restrictions) to the z14 server.

Important: Upgrades from z Enterprise EC12 (zEC12) and IBM z13 are always disruptive.

Upgrade paths from any z Enterprise EC12 (zEC12) to any z14 server are supported as listed in Table 2-12.

Table 2-12 IBM zEnterprise zEC12 to z14 upgrade paths

To 3906 From 2827	Model M01	Model M02	Model M03	Model M04	Model M05
Model H20	Yes	Yes	Yes	Yes	No
Model H43	Yes	Yes	Yes	Yes	No
Model H66	Yes	Yes	Yes	Yes	No
Model H89	Yes	Yes	Yes	Yes	No
Model HA1	Yes	Yes	Yes	Yes	No

Upgrades from any IBM z13 to any z14 server are supported as listed in Table 2-13.

Table 2-13 z13 to z14 upgrade paths

To 3906 From 2964	Model M01	Model M02	Model M03	Model M04	Model M05
Model N30	Yes	Yes	Yes	Yes	No
Model N63	Yes	Yes	Yes	Yes	No
Model N96	Yes	Yes	Yes	Yes	No
Model NC9	Yes	Yes	Yes	Yes	No
Model NE1	Yes	Yes	Yes	Yes	No

2.7.2 Concurrent PU conversions

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

Most conversions are nondisruptive. In exceptional cases, the conversion might be disruptive, for example, when a model z14 with 30 CPs is converted to an all IFL system. In addition, an LPAR might be disrupted when PUs must be freed before they can be converted. Conversion information is summarized in Table 2-14.

Table 2-14 Concurrent PU conversions

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

2.7.3 Model capacity identifier

To recognize how many PUs are characterized as CPs, the Store System Information (STSI) instruction returns a Model Capacity Identifier (MCI). The MCI determines the number and speed of characterized CPs. Characterization of a PU as an IFL, an ICF, or a zIIP is not reflected in the output of the STSI instruction because characterization has no effect on software charging. For more information about STSI output, see “Processor identification” on page 339.

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

- For full-capacity engines, model capacity identifiers 701 - 7H0 are used. They express capacity settings for 1 - 170 characterized CPs.
- Three model capacity identifier ranges offer a unique level of granular capacity at the low end. They are available when no more than 33 CPs are characterized. These three

subcapacity settings are applied to up to 33 CPs, which combined offer 90 more capacity settings. For more information, see “Granular capacity” on page 71.

Granular capacity

The z14 server offers 99 capacity settings at the low end of the processor. Only 33 CPs can have granular capacity. When subcapacity settings are used, other PUs beyond 33 can be characterized only as specialty engines.

The three defined ranges of subcapacity settings have model capacity identifiers numbered from 401- 433, 501 - 533, and 601 - 633.

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

List of model capacity identifiers

Table 2-15 shows that regardless of the number of CPC drawers, a configuration with one characterized CP is possible. For example, model 7H0 might have only one PU characterized as a CP.

Table 2-15 Model capacity identifiers

z14	Model capacity identifier
Model M01	701 - 733, 601 - 633, 501 - 533, and 401 - 433
Model M02	701 - 769, 601 - 633, 501 - 533, and 401 - 433
Model M03	701 - 7A5, 601 - 633, 501 - 533, and 401 - 433
Model M04	701 - 7E5, 601 - 633, 501 - 533, and 401 - 433
Model M05	701 - 7H0, 601 - 633, 501 - 533, and 401 - 433

Important: On z14 servers, model capacity identifier 400 is used ICFs only configurations, and 400 or 401 for IFLs only model

2.7.4 Model capacity identifier and MSU value

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

http://www.ibm.com/systems/z/resources/swprice/reference/exhibits/hardware.html

2.7.5 Capacity Backup

Capacity Backup (CBU) delivers temporary backup capacity in addition to the capacity that an installation might already 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
- Optional SAPs

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

When CBU for CP capacity is acquired by switching from one capacity setting to another, no more 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 disaster strikes, you decide how many of each of the contracted CBUs of any type to activate. The CBU rights are registered in one or more records in the CPC. Up to eight records can be active, which can contain various CBU activation variations that apply to the installation.

You can test the CBU. The number of CBU test activations that you can run for no additional fee in each CBU record is now determined by the number of years that are purchased with the CBU record. For example, a three-year CBU record has three test activations, as compared to a 1-year CBU record that has one test activation. You can increase the number of tests up to a maximum of 15 for each CBU record. The real activation of CBU lasts up to 90 days with a grace period of two days to prevent sudden deactivation when the 90-day period expires. The contract duration can be set from 1 - 5 years.

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

- Number of CP CBUs allowed to be activated
- Number of IFL CBUs allowed to be activated
- Number of ICF CBUs allowed to be activated
- Number of zIIP CBUs allowed to be activated
- Number of SAP CBUs allowed to be activated
- Number of additional CBU tests that are allowed for this CBU record
- Number of total CBU years ordered (duration of the contract)
- Expiration date of the CBU contract

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

Example 2-1 Simple CBU record and related configuration features

```
On Demand Capacity Selections:
NEW00001 - CBU - CP(1) - Years(5) - Tests(5)

Resulting feature numbers in configuration:

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

In Example 2-2, a second CBU record is added to the configuration for two CP CBUs, two IFL CBUs, and two zIIP CBUs, with five more tests and a 5-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 have been 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(10) - Years(5)

Resulting cumulative feature numbers in configuration:

6805  5 Additional CBU Tests                    5
6817  Total CBU Years Ordered                  10
6818  CBU Records Ordered                       2
6820  Single CBU CP-Year                       15
6822  Single CBU IFL-Year                      10
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 503, and the capacity level does not change, the 503 becomes 505:

  `(503 + 2 = 505)`

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

  `(3 + 3 = 6)`

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

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

CBU for specialty engines

Specialty engines (ICFs, IFLs, and zIIPs) run at full capacity for all capacity settings. This fact also applies to CBU for specialty engines. Table 2-16 shows the minimum and maximum (min-max) numbers of all types of CBUs that can be activated on each of the models. The CBU record can contain larger numbers of CBUs than can fit in the current model.

Table 2-16 Capacity Backup matrix

Model	Total PUs available	CBU CPs min - max	CBU IFLs min - max	CBU ICFs min - max	CBU zIIPs min - max	CBU SAPs min - max
Model M01	33	0-170	0-170	0-170	0 - 114	0 - 16
Model M02	69	0-170	0-170	0-170	0 - 114	0 - 16
Model M03	105	0-170	0-170	0-170	0 - 114	0 - 16
Model M04	141	0-170	0-170	0-170	0 - 114	0 - 16
Model M05	170	0-170	0-170	0-170	0 - 114	0 - 16

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

2.7.6 On/Off Capacity on Demand and CPs

On/Off Capacity on Demand (CoD) provides temporary capacity for all types of characterized PUs. Relative to granular capacity, On/Off CoD for CPs is treated similarly to the way 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 504 has two CP5s added temporarily, it becomes a model capacity identifier 506.

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 CPC with model capacity identifier 504 specifies six CP6 temporary CPs through On/Off CoD, the result is a CPC with model capacity identifier 606. A cross-over does not necessarily mean that the CP count for the additional temporary capacity increases. The same 504 might temporarily be upgraded to a CPC with model capacity identifier 704. In this case, the number of CPs does not increase, but more temporary capacity is achieved.

On/Off CoD guidelines

When you request temporary capacity, consider the following guidelines:

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

For more information about temporary capacity increases, see Chapter 8, “System upgrades” on page 299.

2.8 Power and cooling

The z14 power and cooling system is a continuation of IBM z13 with the addition of improvements in availability, temperature tolerance, and vectored (directional) air output. The air-cooled z14 server now has a radiator unit (RU) N+1 design for the pumps and blowers. With the new rear cover design for vectored air output, you can choose whether the output air goes up or down. The water-cooled system is still an option for the z14 server. The Top Exit Power feature is available for the z14 server. Combined with the Top Exit I/O Cabling feature, it gives you more options when you are planning your computer room cabling. For more information about the z14 Top Exit features, see 10.3, “Physical planning” on page 380.

2.8.1 Power and cooling

The system operates with two fully redundant power supplies. One is in the front side of the Z frame, and the other is in the rear side of the Z frame. Each power supply has either one or two power cords. The number of power cords that are required depends on the system configuration. The total loss of one power supply has no impact on system operation.

Systems that specify two power cords can be brought up with one power cord and continue to run.

The larger systems that have a minimum of four BPR pairs that are installed must have four power cords installed. Systems that specify four power cords can be started with two power cords on the *same side* with sufficient power to keep the system running.

Power cords attach to either a three-phase, 50/60 Hz, 200 - 480 V AC power source, or a 380 - 520 V DC power source.

A Balanced Power Plan Ahead feature is available for future growth, helping to ensure adequate and balanced power for all possible configurations. With this feature, system downtime for upgrading a server is eliminated by including the maximum power requirements in terms of BPRs and power cords to your installation.

For ancillary equipment, such as the Hardware Management Console, its display, and its switch, extra single phase outlets are required.

The power requirements depend on the cooling facility that is installed, and on the number of CPC drawers and I/O units that are installed. For more information about the requirements that are related to the number of installed I/O units, see 10.1.2, “Power requirements and consumption” on page 372.

2.8.2 High Voltage Direct Current power feature

The High Voltage Direct Current power feature is an option for z14 servers. It allows z14 servers to directly use the high voltage (HV) DC distribution as shown in Figure 2-26. A direct high voltage DC data center power design can improve data center energy efficiency by removing the need for a DC to AC inversion step.

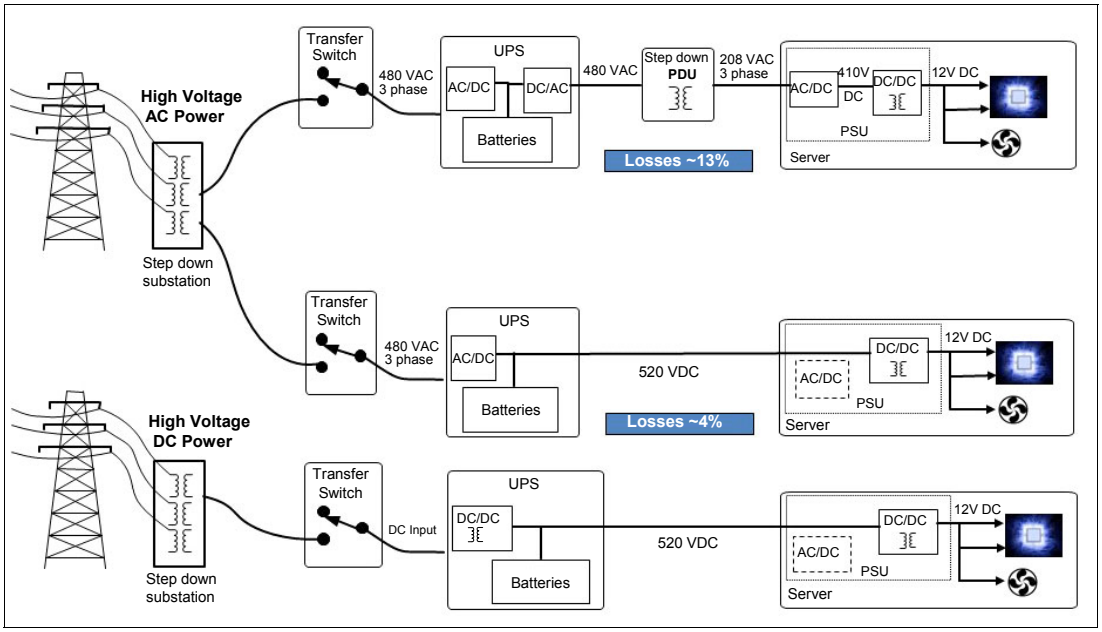


Figure 2-26 AC versus DC distribution

The z14 bulk power supplies have been modified to support HV DC, so the only difference in the shipped hardware to implement this option is the DC power cords. Because HV DC is a new technology, there are multiple proposed standards. z14 servers support both ground referenced and dual polarity HV DC supplies, such as +/-190 V, +/-260 V, and +380 V. Beyond the data center uninterruptible power supply and power distribution energy savings, a z14 server that runs on HV DC power draws 1 - 3% less input power. HV DC does not change the number of power cords that a system requires.

2.8.3 Internal Battery Feature

The Internal Battery Feature (IBF) is an optional feature on the z14 server. It is shown in Figure 2-1 on page 34 for air-cooled (radiator) models and in Figure 2-2 on page 35 for water-cooled models. The IBF provides a local uninterrupted power source.

The IBF further enhances the robustness of the power design, increasing power line disturbance immunity. It provides battery power to preserve processor data during a loss of power on all power feeds from the computer room. The IBF can hold power briefly during a brownout, or for orderly shutdown if there is a longer outage. For information about the hold times, which depend on the I/O configuration and amount of CPC drawers, see 10.1.4, “Internal Battery Feature (IBF)” on page 378.

2.8.4 Power estimation tool

The power estimation tool for the z14 server allows you to 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 Estimation Tools**. Specify the quantity for the features that are installed in your system. This tool estimates the power consumption for the specified configuration. The tool does not verify that the specified configuration can be physically built.

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

2.8.5 Cooling

The PU SCMs are cooled by a cold plate that is connected to the internal water-cooling loop. The SC SCMs are air-cooled. In an air-cooled system, the radiator unit dissipates the heat from the internal water loop with air. The radiator unit provides improved availability with N+1 pumps and blowers. The WCUs are fully redundant in an N+1 arrangement.

Air-cooled models

In z14 servers, the CPC drawer, SC SCMs, PCIe I/O drawers, I/O drawers, and power enclosures are all cooled by forced air with blowers that are controlled by the Move Device Assembly (MDA).

The PU SCMs in the CPC drawers are cooled by water. The internal closed water loop takes heat away from PU SCMs by circulating water between the radiator heat exchanger and the cold plate that is mounted on the PU SCMs. For more information, see 2.8.6, “Radiator Unit” on page 78.

Although the PU SCMs are cooled by water, the heat is exhausted into the room from the radiator heat exchanger by forced air with blowers. At the system level, z14 servers are still air-cooled systems.

Water-cooled models

z14 servers are available as water-cooled systems. With WCU technology, z14 servers can transfer most of the heat that they generate into the building’s chilled water, which effectively reduces the heat output to the computer room.

Unlike the radiator in air-cooled models, a WCU has two water loops: An internal closed water loop and an external (chilled) water loop. The external water loop connects to the client-supplied building’s chilled water. The internal water loop circulates between the WCU heat exchanger and the PU SCMs cold plates. The loop takes heat away from the PU SCMs and transfers it to the external water loop in the WCU’s heat exchanger. For more information, see 2.8.7, “Water-cooling unit” on page 80.

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

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

2.8.6 Radiator Unit

The Radiator Unit provides cooling to PU SCMs with closed loop water. No connection to an external chilled water supply is required. For the z14 server, the internal circulating water is conditioned water that is added to the radiator unit during system installation with the Fill and Drain Tool (FC 3380). The Fill and Drain Tool (FDT) is included with new z14 servers, but if you already have an FDT from a IBM z13 (FC 3380) or a zEC12 (FC 3378) in the data center, you can order an upgrade kit (FC 3379) to have the same equipment as in the FC 3380, and it can be used for the z14, IBM z13, and zEC12 servers. The FDT is used to provide the internal water at the installation and for maintenance, and to remove it at discontinuance.

Figure 2-27 shows the FDT.

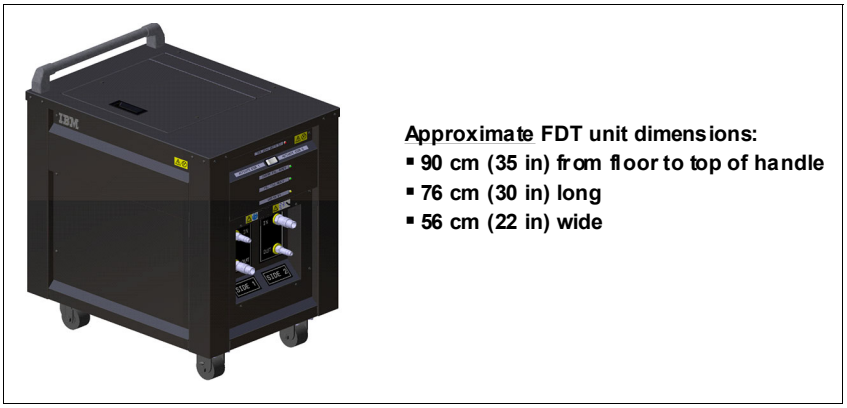


Figure 2-27 Fill and Drain Tool (FDT)

The radiator unit connects to all four CPC drawers (if installed) and cools all PU SCMs simultaneously. The cooling capability is a redundant N+2 design, so a single working pump and blower can support the entire load. The replacement of one pump or blower can be done concurrently and has no performance impact.

As Figure 2-28 shows, the water pumps, manifold assembly, radiator assembly (which includes the heat exchanger), and blowers are the main components of the z14 RU.

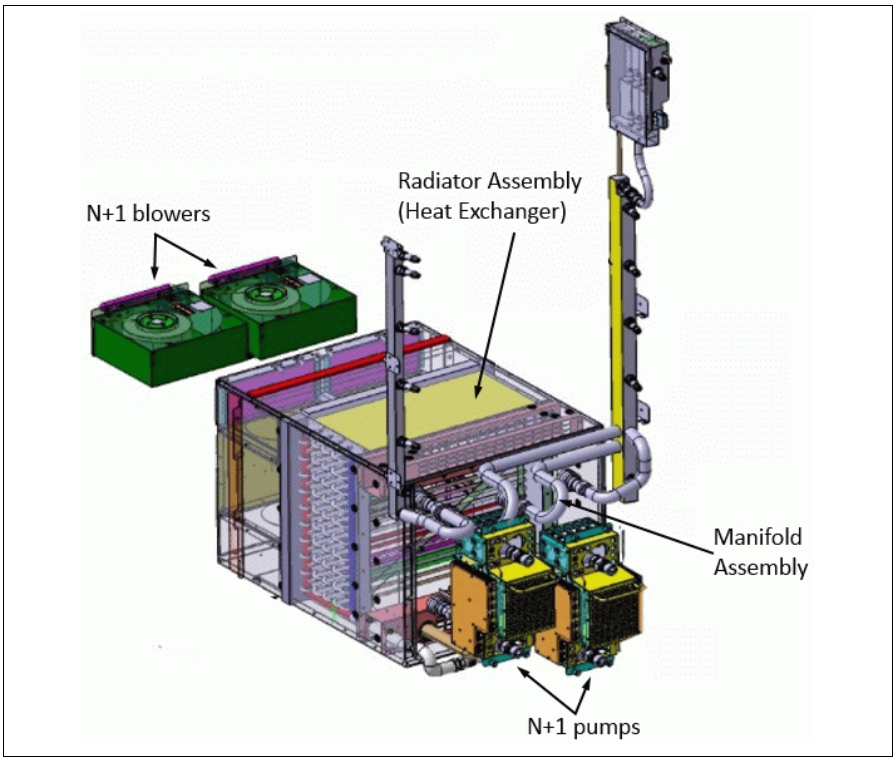


Figure 2-28 Radiator unit

Figure 2-29 shows the closed water loop in the radiator unit. The warm water exiting from the PU SCMs cold plates enters pumps through a common manifold, and is pumped through a heat exchanger where heat is extracted by the air flowing across the heat exchanger fins. The cooled water is then recirculated back into the PU SCMs cold plates.

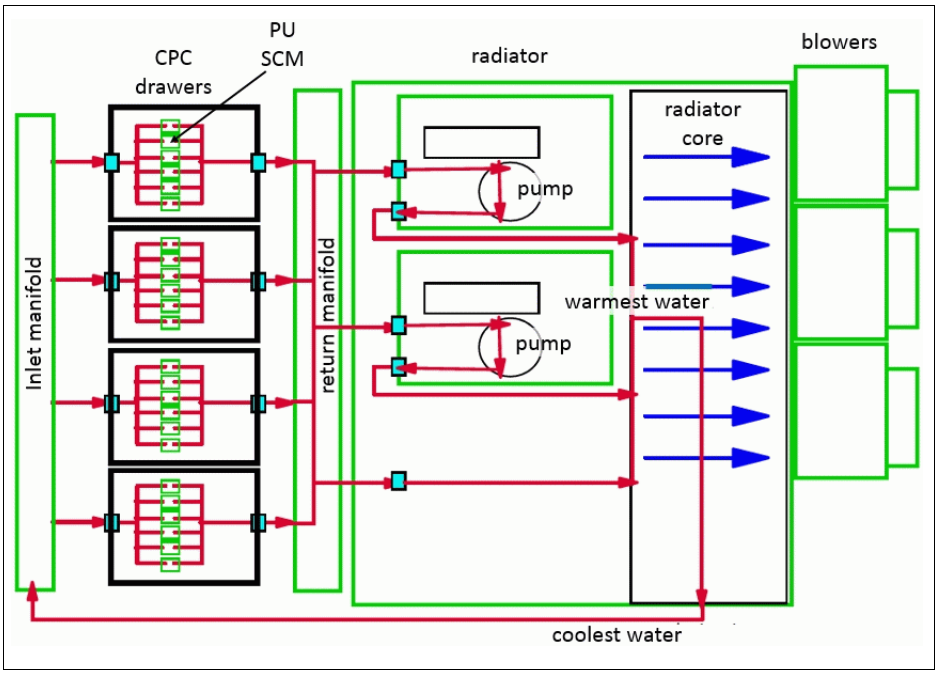


Figure 2-29 Radiator cooling system

2.8.7 Water-cooling unit

z14 servers continue to provide the ability to cool systems with building-provide chilled water by employing the WCU technology. The PU SCMs in the CPC drawers are cooled by internal closed loop water. The internal closed loop water exchanges heat with building-provided chilled water in the WCU heat exchanger. The source of the building's chilled water is provided by the client.

Figure 2-30 shows a WCU.

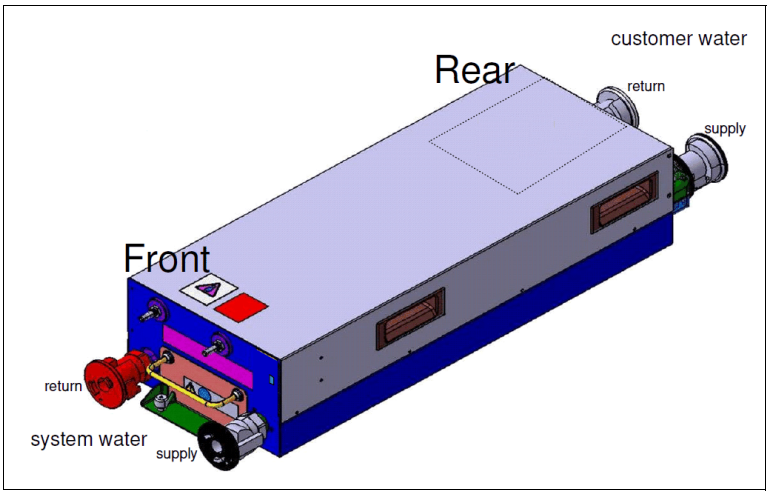


Figure 2-30 Water-cooling unit

The water in the closed loop within the system exchanges heat with the continuous supply of building-provided chilled water. This water circulates between the PU SCMs cold plates and a heat exchanger within the WCU. Heat from the PU SCMs is transferred to the cold plates,

where it is in turn transferred to the circulating system water (closed loop). The system water then dissipates its heat to the building-provided chilled water within the WCU's heat exchanger. The PU SCMs are cooled efficiently in this manner.

A simplified high-level diagram that illustrates this principle is shown in Figure 2-31.

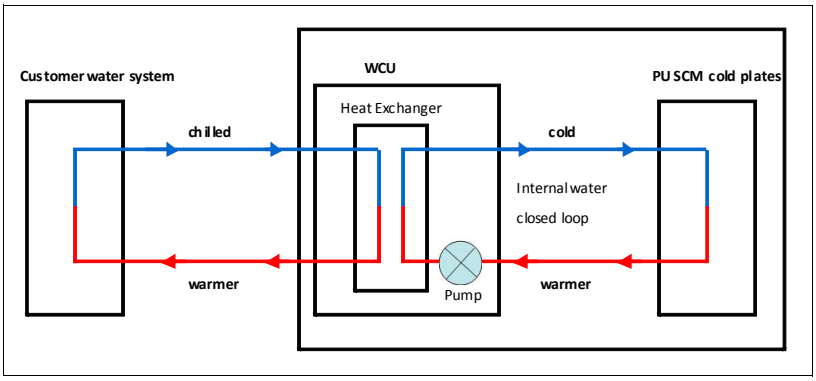


Figure 2-31 WCU water loop

z14 servers operate with two fully redundant WCUs. These water-cooling units have each their own facility feed and return water connections. If water is interrupted to one of the units, the other unit picks up the entire load, and the server continues to operate without interruption. You must provide independent redundant water loops to the water-cooling units to obtain full redundancy.

The internal circulating water is conditioned water that is added to the radiator unit during system installation with the Fill and Drain Tool (FC 3380). The FDT is included with new z14 servers, but if you already have an FDT from a zEC12 (FC 3378) in the data center, you can order an upgrade kit (FC 3379) to have the same equipment as in the FC 3380, and it can be used for the zEC12, z13, and z14 servers. The FDT is used to provide the internal water at the installation and for maintenance, and to remove it at discontinuance. Figure 2-27 on page 78 shows the FDT.

Exhaust Air Heat Exchanger

In z14 servers, all water-cooled models have two Exhaust Air Heat Exchanger units that are installed on the rear of the A and Z frames, as shown in Figure 2-32. These units remove heat from the internal system water loop, and internal air exits the server into the hot air exhaust aisle.

Figure 2-32 Water-cooled model: Rear view

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

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

Considerations before you order

The water-cooling option is preferable because it can substantially lower the total power consumption of z14 servers and the total cost of ownership for the CPC. This savings is greater for the larger models of z14 servers, as shown in Table 2-17.

The water-cooling option cannot be installed in the field. Therefore, you must carefully consider the present and future computer room and CPC configuration options before you decide what cooling option to order. For more information, see 10.1.3, “Cooling requirements” on page 375.

Table 2-17 Heat to water based on temperature

Temperature	Three CPC drawer typical configuration	Four CPC drawer typical configuration	Four CPC drawer maximum power configuration
Water-cooled system power in normal room/hot room			
	12.9 kW/14.1 kW	17.4 kW/19.0 kW	24.7 kW/26.3 kW
Inlet air temperature	**Heat to water and as % of total system heat load**		

Temperature	Three CPC drawer typical configuration	Four CPC drawer typical configuration	Four CPC drawer maximum power configuration
18°C (64°F)	7.3 kW (57%)	9.8 kW (56%)	12.6 kW (51%)
23°C (73°F)	9.5 kW (74%)	12.6 kW (72%)	15.6 kW (63%)
27°C (80.6°F)	11.5 kW (89%)	14.8 kW (85%)	18.0 kW (73%)
32°C (89.6°F) (hot room)	14.8 kW (105%)	18.2 kW (96%)	21.6 kW (82%)

2.9 Summary of z14 structure

Table 2-18 summarizes all aspects of the z14 structure.

Table 2-18 System structure summary

Description	Model M01	Model M02	Model M03	Model M04	Model M05
Number of CPC drawers	1	2	3	4	4
Number of SCMs	6	12	18	24	28
Total number of PUs	41	82	123	164	196
Maximum number of characterized PUs	33	69	105	141	170
Number of CPs	0 - 33	0 - 69	0 - 105	0 - 141	0 - 170
Number of IFLs	0 - 33	0 - 69	0 - 105	0 - 141	0 - 170
Number of ICFs	0 - 33	0 - 69	0 - 105	0 - 141	0 - 170
Number of zIIPs	0 - 22	0 - 46	0 - 70	0 - 94	0 - 112
Standard SAPs	5	10	15	20	23
Additional SAPs	0 - 4	0 - 8	0 - 12	0 - 16	0 - 16
Number of IFP	1	1	1	1	1
Standard spare PUs	2	2	2	2	2
Enabled memory sizes GB	320 - 8000	320 - 16192	320 - 24384	320 - 32576	320 - 32576
Flexible memory sizes GB	N/A	320 - 8000 GB	320 - 16192	320 - 24384	320 - 24384
L1 cache per PU (I/D)	128/128 KB	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)	2/4 MB (I/D)
L3 shared cache per PU chip	128 MB	128 MB	128 MB	128 MB	128 MB
L4 shared cache per node	672 MB	672 MB	672 MB	672 MB	672 MB
Cycle time (ns)	0.192	0.192	0.192	0.192	0.192
Clock frequency	5.2 GHz	5.2 GHz	5.2 GHz	5.2 GHz	5.2 GHz

Description	Model M01	Model M02	Model M03	Model M04	Model M05
Maximum number of PCIe fanouts	10	20	30	40	40
Maximum number of IFB fanouts	4	8	12	16	16
I/O interface per IFB cable	6 GBps	6 GBps	6 GBps	6 GBps	6 GBps
I/O interface per PCIe cable	16 GBps	16 GBps	16 GBps	16 GBps	16 GBps
Number of support elements	2	2	2	2	2
External AC power	3 phase	3 phase	3 phase	3 phase	3 phase
Optional external DC	520 V/380 V	520 V/380 V	520 V/380 V	520 V/380 V	520 V/380 V
Internal Battery Feature	Optional	Optional	Optional	Optional	Optional

3

Central processor complex system design

This chapter explains how the IBM z14 processor unit is designed. This information can be used to understand the functions that make the z14 server a system that accommodates a broad mix of workloads for large enterprises.

This chapter includes the following sections:

- Overview
- Design highlights
- CPC drawer design
- Processor unit design
- Processor unit functions
- Memory design
- Logical partitioning
- Intelligent Resource Director (IRD)
- Clustering technology
- Virtual Flash Memory

© Copyright IBM Corp. 2017. All rights reserved.

3.1 Overview

The z14 symmetric multiprocessor (SMP) system is the next step in an evolutionary trajectory that began with the introduction of the IBM System/360 in1964. Over time, the design was adapted to the changing requirements dictated by the shift toward new types of applications that clients depend on.

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

The modular CPC drawer design aims to reduce, or in some cases even eliminate, planned and unplanned outages. The design does so by offering concurrent repair, replace, and upgrade functions for processors, memory, and I/O. For more information about the z14 reliability, availability, and serviceability (RAS) features, see Chapter 9, “Reliability, availability, and serviceability” on page 345.

z14 servers have ultra-high frequency, large high-speed buffers (caches) and memory, superscalar processor design, out-of-order core execution, simultaneous multithreading (SMT), single-instruction multiple-data (SIMD), and flexible configuration options. It is the next implementation of IBM Z servers to address the ever-changing IT environment.

3.2 Design highlights

The physical packaging of z14 servers is a continuation and an evolution of the z13 systems. Its modular CPC drawer and single chip module (SCM) design address the augmenting costs that are related to building systems with ever-increasing capacities. The modular CPC drawer design is flexible and expandable, offering unprecedented capacity to meet consolidation needs, and might contain even larger capacities in the future.

z14 servers continue the line of mainframe processors that are compatible with an earlier version. Evolution® brings several processor design enhancements, such as: 10 cores per CP chip, pipeline optimization, improved Simultaneous Multi-Threading - SMT and Single Instruction Multiple Data - SIMD, better branch prediction, improved co-processor functionality, and so on. It uses 24-bit, 31-bit, and 64-bit addressing modes, multiple arithmetic formats, and multiple address spaces for robust interprocess security.

The z14 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, which range from the traditional systems (for example, z/OS and z/VM) to the world of Linux, cloud, analytics, and mobile computing.
- Offer state-of-the-art *integration* capability for server consolidation by using virtualization capabilities in a highly *secure environment*:
  - Logical partitioning, which allows 85 independent logical servers.
  - z/VM, which can virtualize hundreds to thousands of servers as independently running virtual machines (guests).

  - HiperSockets, which implement virtual LANs between logical partitions (LPARs) within the system.
  - Efficient data transfer using 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 running on one partition (LPAR) cannot access another application on a different partition, providing essentially the same security as an air-gapped system.

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

- Offer *high performance computing* to achieve the outstanding response times that are required by new workload-type applications. This performance is achieved by high frequency, enhanced superscalar processor technology, out-of-order core execution, large high-speed buffers (cache) and memory, an architecture with multiple complex instructions, and high-bandwidth channels.
- Offer the *high capacity* and *scalability* that are required by the most demanding applications, both from the single-system and clustered-systems points of view.
- Offer the capability of *concurrent upgrades* for processors, memory, and I/O connectivity, 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.
- Have internal and external *connectivity* offerings, supporting open standards, such as Gigabit Ethernet (GbE) and Fibre Channel Protocol (FCP).
- Provide leading *cryptographic* performance. Every processor unit (PU) has a dedicated and optimized CP Assist for Cryptographic Function (CPACF). Optional Crypto Express features with cryptographic coprocessors provide the highest standardized security certification.[1] These optional features can also be configured as Cryptographic Accelerators to enhance the performance of Secure Sockets Layer/Transport Layer Security (SSL/TLS) transactions.
- Be *self-managing* and *self-optimizing*, adjusting itself when the workload changes to achieve the best system throughput. This process can be done through the Intelligent Resource Director or the Workload Manager functions, which are assisted by HiperDispatch.
- Have a *balanced system* design with pervasive encryption, providing large data rate bandwidths for high performance connectivity along with processor and system capacity, while protecting every single byte entering and exiting the z14.

The remaining sections describe the z14 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 system can have up to four CPC drawers in a full configuration, up to 170 PUs that can be characterized for customer use, and up to 32TB of customer usable memory capacity.

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

Each CPC drawer is logically divided in two clusters to improve the processor and memory affinity and availability.

There are two types of drawer configurations for z14:

- Five CP SCMs populated - Modes M01, M02, M03, M04
- Six CP SCMs populated for Model M05.

For serviceability, the CP SCM spare is supplied as an Universal field Spare.

The z14 has up to 20 memory controller units (MCUs) (five MCUs per CPC drawer). 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 made by IBM has eliminated that limitation. In addition, eDRAM offers higher density, less power utilization, fewer soft errors, and better performance. Concurrent maintenance allows dynamic central processing complex (CPAC) drawer add and repair.[2]

z14 servers use CMOS Silicon-on-Insulator (SOI) 14 nm chip technology, with advanced low latency pipeline design, creating high-speed yet power-efficient circuit designs. The PU SCM has a dense packaging, allowing closed water loop cooling. The heat exchange from the closed loop is either air-cooled by a radiator unit (RU) or, optionally, water-cooled by a water-cooling unit (WCU). The water-cooling option can lower the total power consumption of the system. This benefit is significant for larger configurations. For more information, see 2.8.1, “Power and cooling” on page 75.

3.3.1 Cache levels and memory structure

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

Figure 3-1 shows the z14 cache levels and memory hierarchy.

[2] For configurations with two or more CPC drawers installed

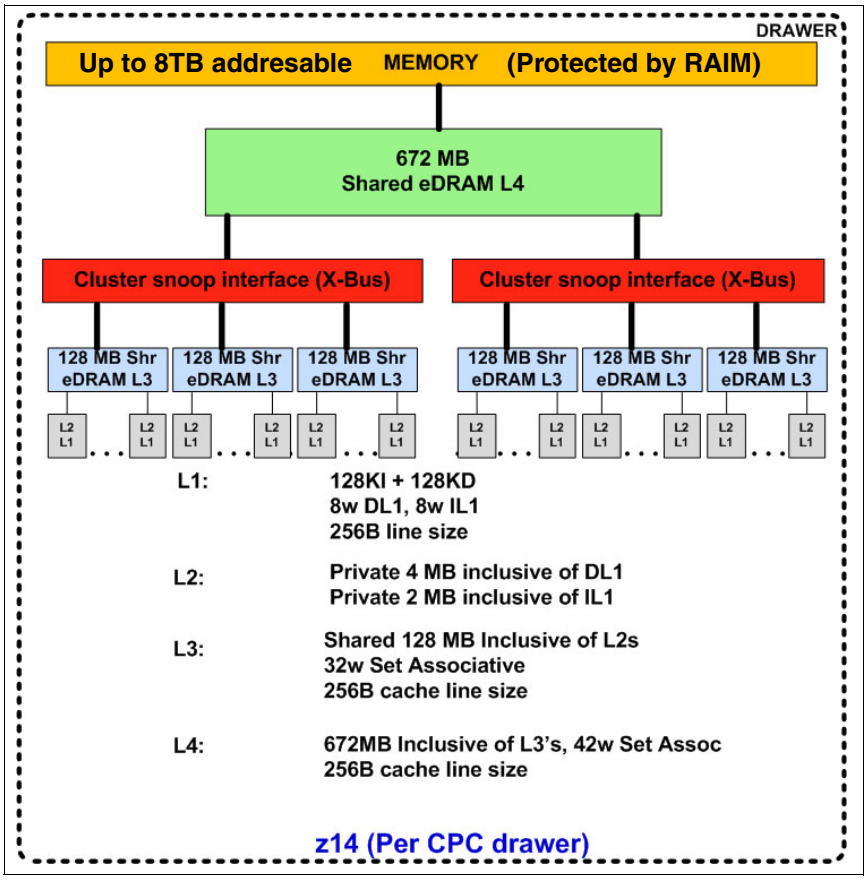


Figure 3-1 z14 cache levels and memory hierarchy

While L1, L2, and L3 caches are implemented on the CP SCM, the fourth cache level (L4) is implemented within the system controller (SC) SCM. There is one L4 cache in each CPC drawer, shared by all CP SCMs. The cache structure of the z14 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 LRU (Least Recent Used) Demote handling
- L3 cache also uses eDRAM and is shared by all ten cores within the PU chip. Each CPC drawer has five (M01 through M04) or six L3 caches (M05). A four-CPC drawer system Model M05 therefore has 24 of them, resulting in 3072 MB (24 x 128MB) 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 on the CPC drawer. Each L4 cache has 672 MB inclusive of L3's, 42w Set Associative and 256B cache line size. A four-CPC drawer system has 2688 MB (4 x 672 MB) of shared L4 cache. In most real wold situations there are a fair amount of cache lines that exist in multiple L3's underneath a given L4. The L4 doesn't 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, 672MB of inclusive L4 can easily cover 768MB of underlying L3 caches.
- Main storage has up to 8 TB addressable memory per CPC drawer, using 25 DIMMs. A four-CPC drawer system can have up to 32 TB of main storage.

Considerations

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

Also, the distance to remote caches as seen from the microprocessor becomes a significant factor. An example is an L4 cache that is not on the microprocessor (and might not even be in the same CPC drawer). Although the L4 cache is rather large, several cycles are needed to travel the distance to the cache. Figure 3-2 shows the node-cache topology of z14 servers.

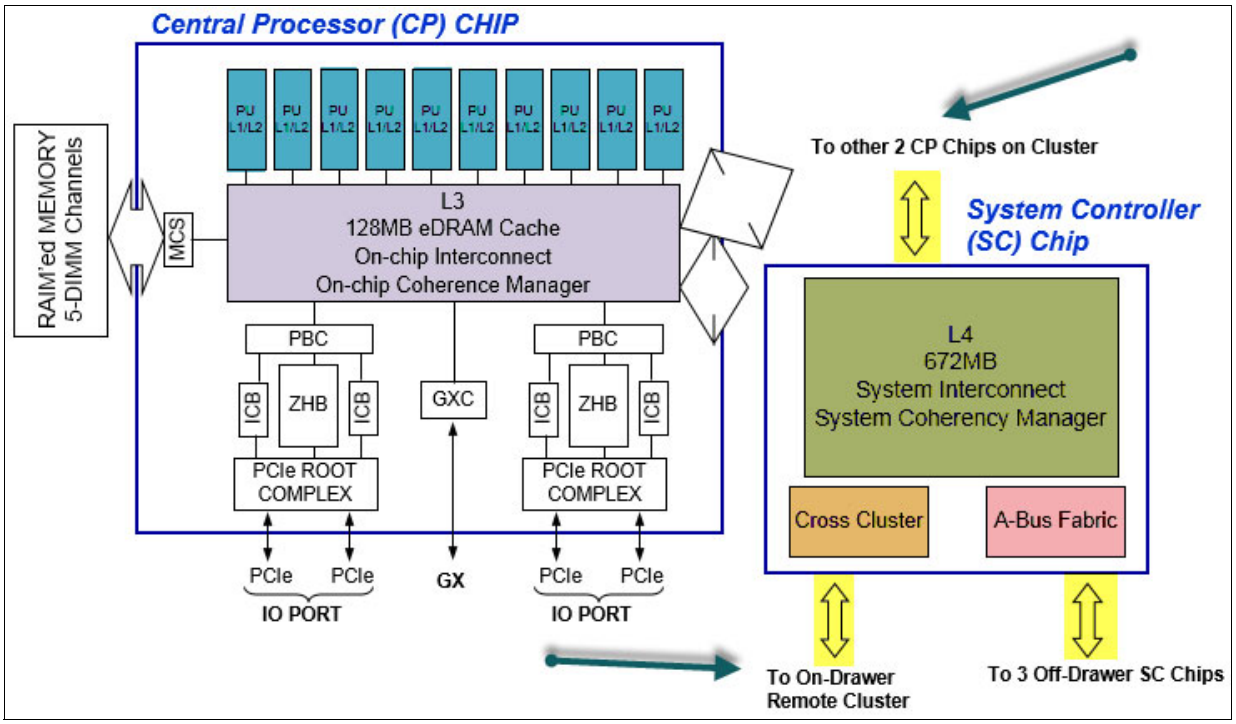


Figure 3-2 z14 cache topology

Although large caches mean increased access latency, the new technology of CMOS 14S0 (14 nm chip lithography) and the lower cycle time allows z14 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 there is a cache miss 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 cache design, some lines of the L3 cache are not included in the L4 cache.

Another approach is available for avoiding L4 cache access delays (latency). The L4 cache straddles up to four CPC drawers. This configuration means that relatively long distances exist between the higher-level caches in the processors and the L4 cache content. To overcome the delays that are inherent in the SMP CPC drawer design and save cycles to access the remote L4 content, keep instructions and data as close to the processors as possible. This can be managed by directing as much work of a particular LPAR workload to the processors in the same CPC drawer as the L4 cache. This configuration is achieved by having the IBM Processor Resource/Systems Manager (PR/SM) scheduler and the z/OS WLM and dispatcher work together. Have them keep as much work as possible within the

boundaries of as few processors and L4 cache space (which is best within a CPC drawer boundary) without affecting throughput and response times.

Figure 3-3 compares the cache structures of z14 servers with the previous generation of IBM Z servers, z13.

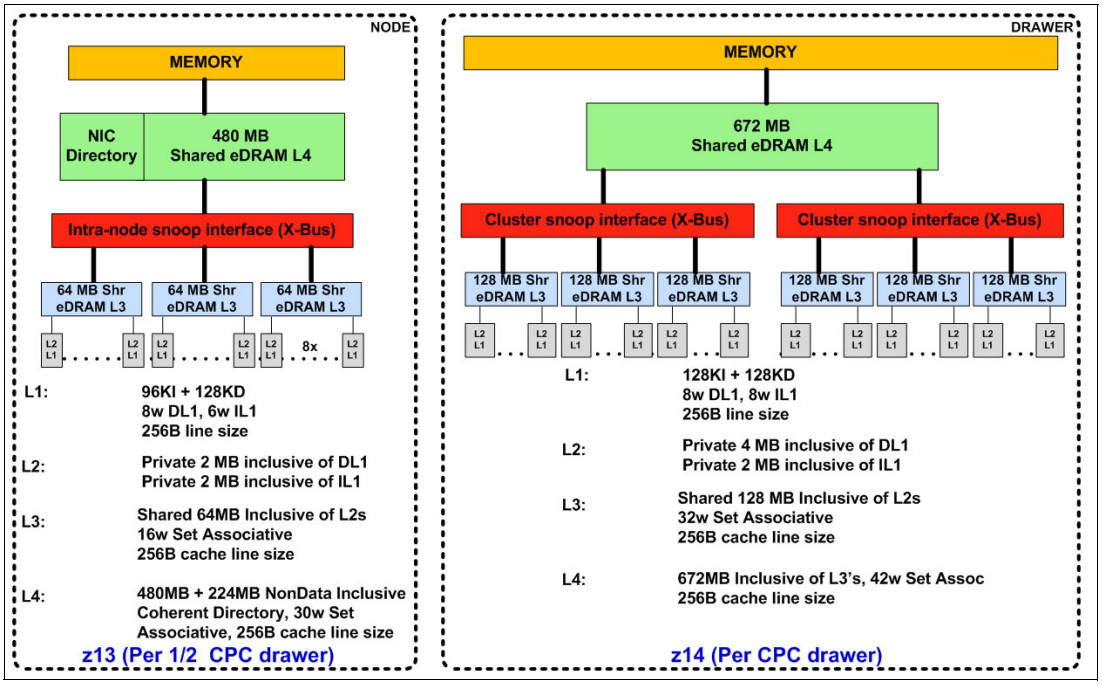


Figure 3-3 z14 and z13 cache level comparison

Compared to z13, the z14 cache design has larger L1, L2, and L3 cache sizes. In z14 servers, there is more affinity between the memory of a partition, the L4 cache in the SC, accessed by the two logical clusters in the same CPC drawer, and the cores in the PU. The access time of the private cache usually occurs in one cycle. The z14 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, thus keeping 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 cache topology, which has 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, node, and drawer to improve the reuse of shared caches by processors of the same partition. It can use dynamic memory relocation (DMR) to move a running partition's memory to different physical memory to improve the affinity and reduce the distance between the memory of a partition and the processors of the partition. For more information about HiperDispatch, see 3.7, "Logical partitioning" on page 120.

3.3.2 CPC drawer interconnect topology

CPC drawers are interconnected in a point-to-point topology at SC level, allowing a CPC drawer to communicate with every CPC drawer. Data transfer does not always have to go through another CPC drawer (L4 cache) to address the requested data or control information.

Figure 3-4 shows the z14 inter-CPC drawer communication structure.

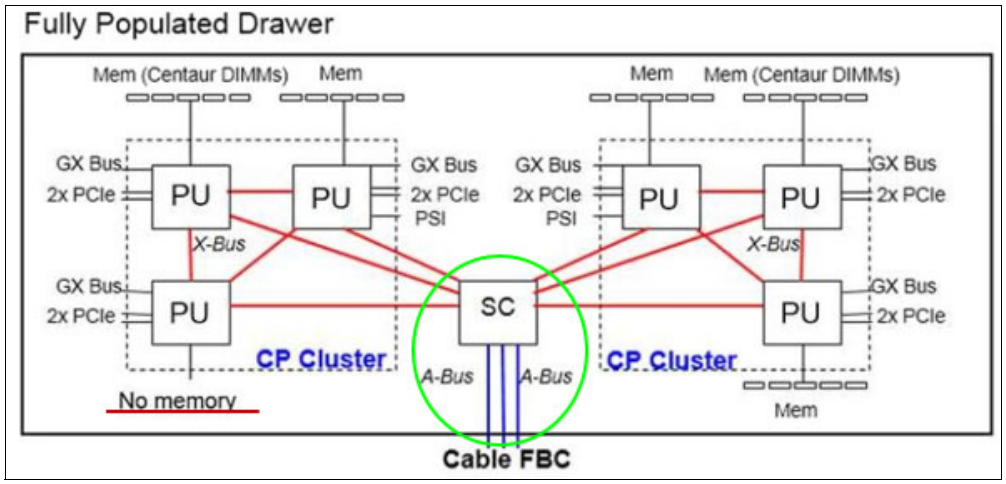


Figure 3-4 z14 CPC drawer communication topology

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

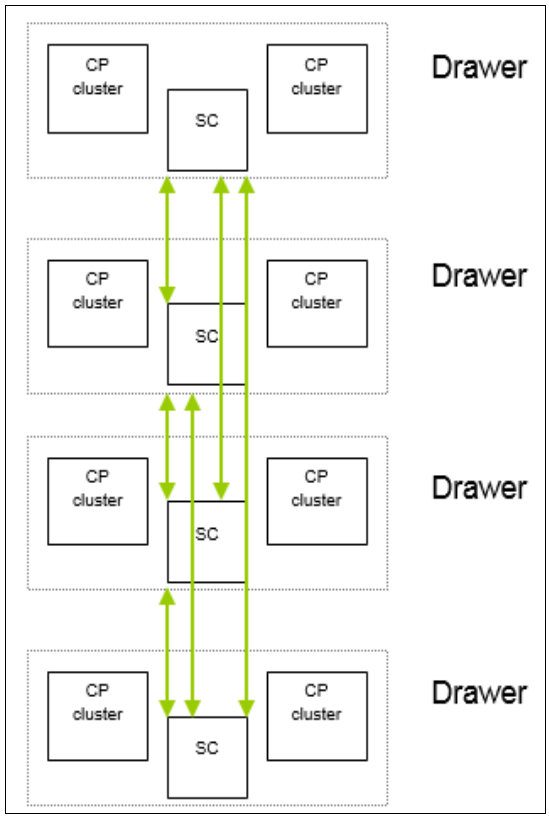


Figure 3-5 Point-to-point topology with four CPC drawers

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

3.4 Processor unit design

Processor cycle time is especially important for processor-intensive applications. Current systems design is driven by processor cycle time, although improved cycle time does not automatically mean that the performance characteristics of the system improve. The System z10® EC introduced a dramatic PU cycle time improvement. Its succeeding generations, the z196 and the zEC12, reduced the cycle time even further, reaching 0.192 ns (5.2 GHz) and 0.178 ns (5.5 GHz) respectively. While chip lithography has driven higher on chip core and cache density, the thermal design and cache sizes have added some challenges to chip frequency evolution.

Through innovative processor design - pipeline and cache management redesigns - the IBM Z processor performance has continued to evolve. With the introduction of out-of-order execution, ever improving branch prediction mechanism, simultaneous multi-threading, the processing performance has been enhanced despite processor frequency variations (z13 core runs at 5.0 GHz)

z14 core furnace is 5.2GHz, which allows the increased number of processors that share larger caches to have quick access times and improved capacity and performance. Although the cycle time of the z14 processor was only slightly reduced (4%) compared to z13, the processor performance was increased even further (z14 uni-processor PCI up 8% compared to z13) through improved processor design, such as pipeline, 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 433.

z13 servers introduced architectural extensions with instructions designed to allow reduced processor quiesce effects, reduced cache misses, reduced pipeline disruption, and increased parallelism with instructions that process several operands in a single instruction (SIMD). The processor architecture has been further developed for z14 and includes the following features:

- Optimized second generation SMT (Simultaneous Multi-Threading)
- 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 enhanced Instruction Set Architecture (ISA) includes a set of instructions that are added to improve compiled code efficiency. These instructions optimize PUs to meet the demands of a wide variety of business and analytics workload types without compromising the performance characteristics of traditional workloads.Simultaneous multithreading

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

SMT is supported only for Integrated Facility for Linux (IFL) and IBM Z Integrated Information Processor (zIIP) speciality engines on $z14^3$ 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 Systems) 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 227.

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.

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

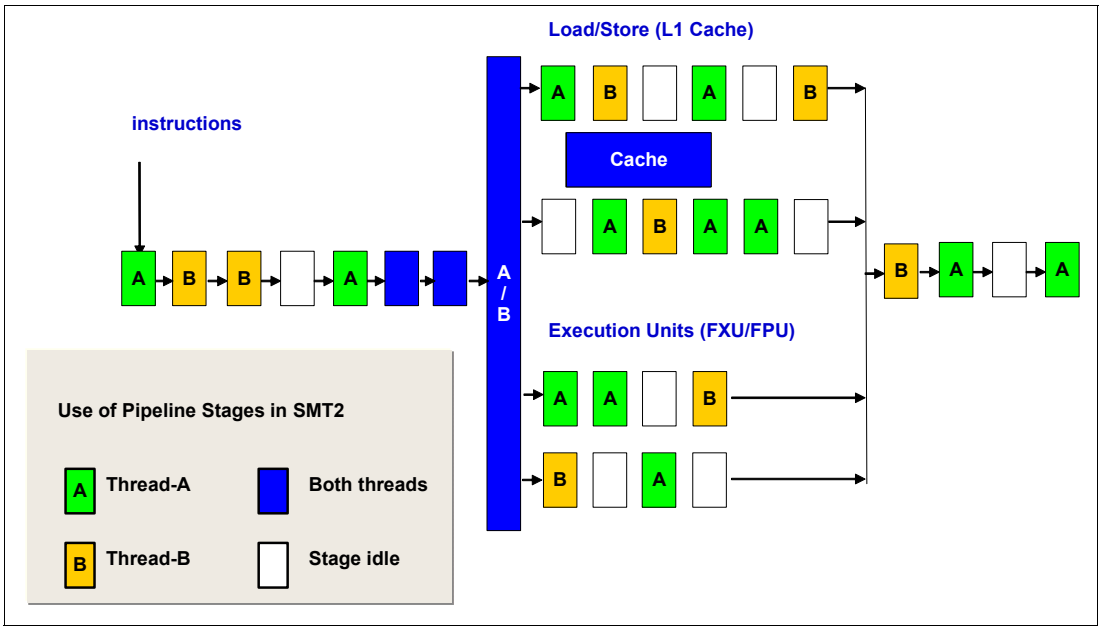


Figure 3-6 Two threads running simultaneously on the same processor core

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

Although SMT increases the processing capacity, the performance in some cases might be superior if 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 has introduced SMT as default for SAPs (not user controllable).

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

3.4.1 Single-instruction multiple-data

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

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

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

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

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

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

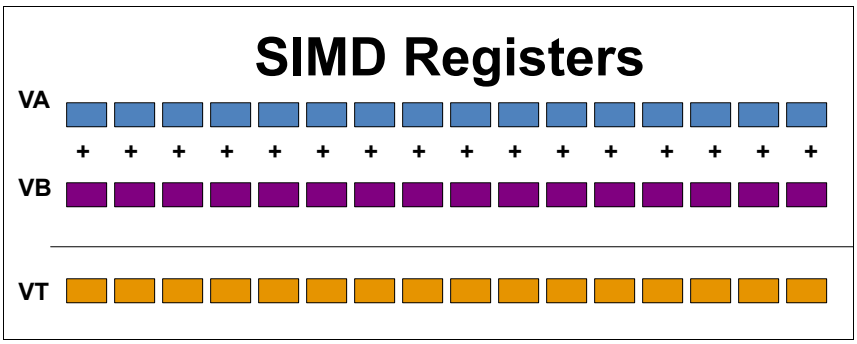


Figure 3-7 Schematic representation of add SIMD instruction with 16 elements in each vector

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

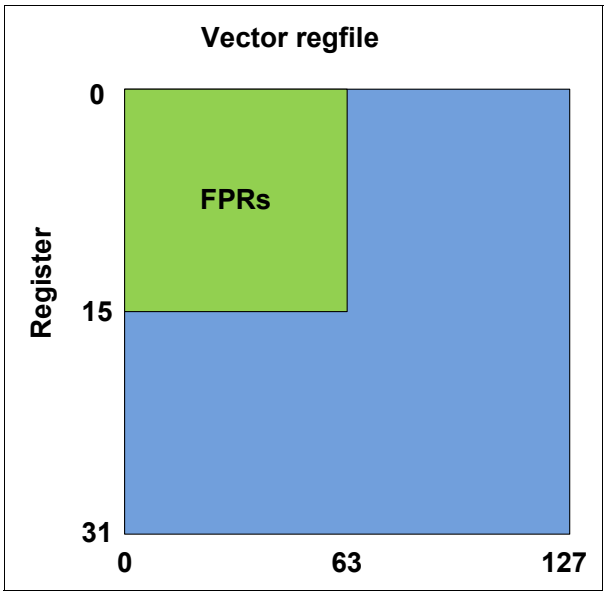


Figure 3-8 Floating point registers overlaid by vector registers

Here are some examples of SIMD instructions:

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

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

z14 SIMD enhancements include:

- Doubled vector double precision BFP[4] throughput (2x 64b)
- Added vector single precision BFP (4x 32b)
- Added vector quad precision BFP (128b)
- Added additional binary FMA[5] operations to speed up code
- Vector SP/DP/QP[6] compare/min/max with programming language support
- Enhanced to Storage-to-Storage BCD[7]
- Vector load/store rightmost with length

3.4.2 Out-of-order execution

z14 servers have an Out-of-Order core, much like the z13, zEC12, and z196. Out-of-Order yields significant performance benefits for compute-intensive applications. It does so by reordering instruction execution, allowing later (younger) instructions to be run ahead of a stalled instruction, and reordering storage accesses and parallel storage accesses. Out-of-Order maintains good performance growth for traditional applications. Out-of-order execution can improve performance in the following ways:

[4] BFP - Binary Floating Point operations
[5] FMA - Fixed Multiply Add operations
[6] Single Precision/ Double Precision/ Quad Precision
[7] BCD Binary Coded Decimal

- 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 processor has pipeline enhancements that benefit Out-of-Order execution. The IBM Z processor design has advanced micro-architectural innovations that provide these 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 chip, at 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 OSC[8] (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, ending up with the same results 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 core is shown in Figure 3-9.

[8] OSC - Operand Store Compare

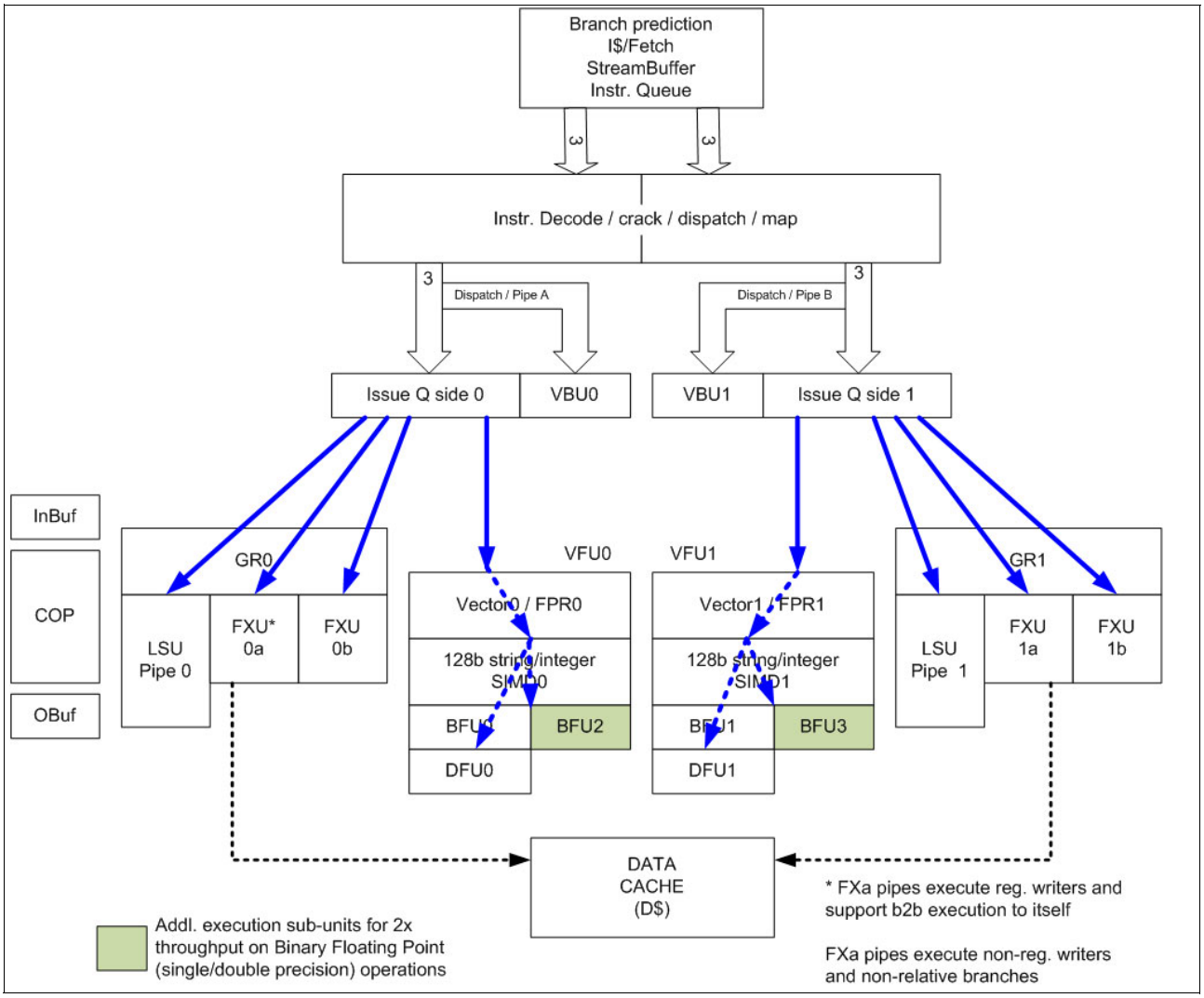


Figure 3-9 z14 PU core logical diagram

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

The z14 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 ten instructions/operations can be initiated to run per clock cycle (<0.192 ns). The execution of the instructions can occur out of program order, and memory address generation and memory accesses can also occur out of program order. Each core has special circuitry to display execution and memory accesses in order to the software.

The z14 superscalar PU core can have up to 10 instructions/operations running per cycle. This technology results in shorter workload run time.

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 3-9 on page 98. The branch target buffer (BTB) runs ahead of instruction cache pre-fetches to prevent branch misses in an early stage. Furthermore, a branch history table (BHT) in combination with a pattern history table (PHT) and the use of tagged multi-target prediction technology branch prediction offer a high branch prediction success rate.

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

3.4.3 Superscalar processor

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

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

On z14 servers, up to 6 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 has made significant strides in avoiding address generation interlock (AGI) situations. Instructions that require information from memory locations can suffer multi-cycle delays to get the needed memory content. Because high-frequency processors wait “faster” (spend processor cycles more quickly while idle), the cost of getting the information might become prohibitive.

3.4.4 Compression and cryptography accelerators on a chip

This section introduces the compression and cryptography feature (CPACF) enhancements for z14.

Coprocessor units

There is one coprocessor unit 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 of clear key operations.

Here are some of the characteristics of the coprocessors:

- Each core has an independent compression and cryptographic engine.
- The coprocessor has been redesigned from scratch 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.

Figure 3-10 shows the location of the coprocessor on the chip.

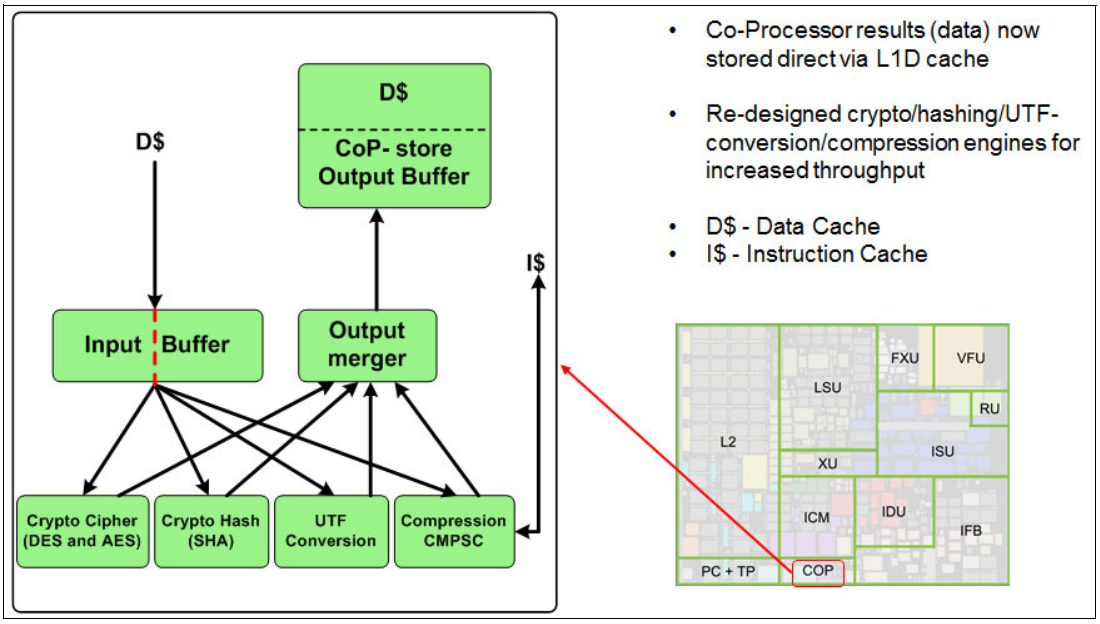


Figure 3-10 Compression and cryptography accelerators on a core in the chip

Compression enhancements

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

CP Assist for Cryptographic Function

CPACF accelerates the encrypting and decrypting of SSL/TLS transactions, virtual private network (VPN)-encrypted data transfers, and data-storing applications that do not require 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

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

For more information about cryptographic functions on z14 servers, see Chapter 6, “Cryptographic features” on page 195.

3.4.5 Decimal floating point accelerator

The 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 has made floating point arithmetic complex and error-prone for programmers who use it for applications in which the data is typically decimal.

Hardware decimal floating point computational instructions provide the following features:

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

Benefits of the DFP accelerator

The DFP accelerator offers the following benefits:

- Avoids rounding issues, such as those that happen with binary-to-decimal conversions.
- Controls existing binary-coded decimal (BCD) operations better.
- Follows the standardization of the dominant decimal data and decimal operations in commercial computing, supporting 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 take advantage of the z/Architecture DFP instructions.

z14 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 servers have new decimal floating point packed conversion facility support with the following benefits:

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

Software support

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

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

3.4.6 IEEE floating point

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

The z14 core implements two additional execution sub-units for 2x throughput on Binary Floating Point (single/double precision) operations (see Figure 3-9 on page 98).

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.

3.4.7 Processor error detection and recovery

The PU uses a process called *transient recovery* as an error recovery mechanism. When an error is detected, the instruction unit tries the instruction again and attempts to recover the error. If the second attempt is unsuccessful (that is, a permanent fault exists), a relocation process is started that restores the full capacity by moving work to another PU. Relocation under hardware control is possible because the R-unit has the full designed state in its buffer. The principle is shown in Figure 3-11.

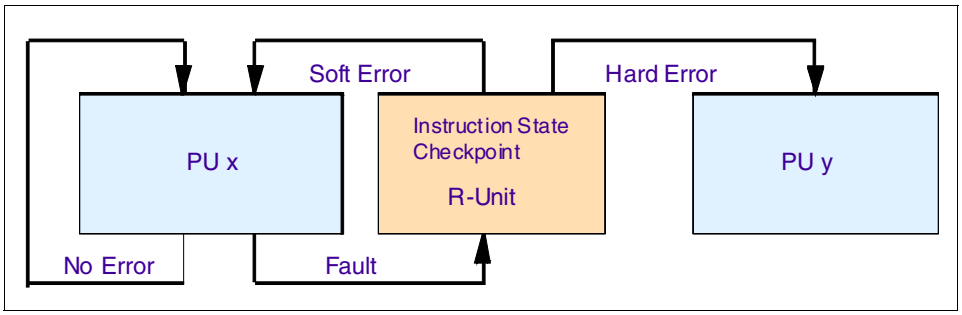


Figure 3-11 PU error detection and recovery

3.4.8 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 has been continuously improved.

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

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

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

The z14 branch prediction enhancements over z13 include:

- Branch prediction search pipeline extended from 5 to 6 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): 1K rows X 6 sets -> 2K rows X 4 sets
  - Level 2 Branch Target Buffer (BTB2): 16K rows X 6 sets -> 32K rows X 4 sets
- Better power efficiency: several structures have been re-designed to maintain their accuracy while consuming less power through smart access algorithms.
- New static IA(0:31) regions expanded from 4 to 8. To conserve space, prediction structures do not store full target addresses. They instead exploit the locality and limited ranges of “4gig regions” of virtual instruction addresses => IA(0:31).

3.4.9 Wild branch

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

With the wild branch hardware facility, the last address from which a successful branch instruction was run is kept. z/OS uses this information with debugging aids, such as the **SLIP** command, to determine where a wild branch came from. It 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 from.

3.4.10 Translation lookaside buffer

The 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 have recently increased significantly as a result of the introduction of 64-bit addressing, a smaller working set is represented by the TLB. To increase the working set representation in the TLB without enlarging the TLB, large (1 MB) page and giant page (2 GB) support is available and can be used when appropriate. For more information, see “Large page support” on page 118.

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

z14 TLB enhancements

IBM z14 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) while no longer needs to access TLB for L1 cache hit. TLB2 is accessed in parallel to L2, saving 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.

3.4.11 Instruction fetching, decoding, and grouping

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

Instruction fetching

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

Instruction decoding

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

Instruction grouping

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

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

3.4.12 Extended Translation Facility

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

3.4.13 Instruction set extensions

The processor supports many instructions to support functions:

- Hexadecimal floating point instructions for various unnormalized multiply and multiply-add instructions.
- Immediate instructions, including various add, compare, OR, exclusive-OR, subtract, load, and insert formats. Use of these instructions improves performance.
- Load instructions for handling unsigned halfwords, such as those used for Unicode.

- Cryptographic instructions, which are known as the MSA, offer the full complement of the AES, SHA-1, SHA-2, and DES algorithms. They also include functions for random number generation.
- Extended Translate Facility-3 instructions, enhanced to conform with the current Unicode 4.0 standard.
- Assist instructions that help eliminate hypervisor processor usage.
- SIMD instructions, which allow the parallel processing of multiple elements in a single instruction.

3.4.14 Transactional Execution

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

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

3.4.15 Runtime Instrumentation

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

3.5 Processor unit functions

This section describes the PU functions.

3.5.1 Overview

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

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

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

This design brings outstanding flexibility to z14 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 227.

Concurrent upgrades

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

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

If the PU chips in the installed CPC drawers have no available remaining PUs, an upgrade results in a model upgrade and the installation of an extra CPC drawer. However, there is a limit of four CPC drawers. CPC drawer installation is nondisruptive, but takes more time than a simple LIC upgrade.

For more information about Capacity on Demand, see Chapter 8, “System upgrades” on page 299.

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 servers have two spare PUs. PUs that are not characterized on a CPC configuration can also be used as extra spare PUs. For more information about PU sparing, see 3.5.10, “Sparing rules” on page 116.

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 107.

All assigned PUs are grouped in the PU pool. These PUs are dispatched to online logical PUs. As an example, consider a z14 server with 10 CPs, two IFLs, five zIIPs, and one ICF. This system has a PU pool of 18 PUs, called the *pool width*. Subdivision defines these pools:

- A CP pool of 10 CPs
- An ICF pool of one ICF
- An IFL pool of two IFLs
- A zIIP pool of five 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 the addition of PUs during a CBU
- Following a capacity on-demand upgrade through On/Off CoD or CIU

PUs are removed from their pools when a concurrent downgrade takes place as the result of the removal of a CBU. They are also removed through 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 is also the case when an LPAR logically configures a PU 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 170 PUs can be characterized as CPs, depending on the configuration.

The z14 server can be initialized either in LPAR (PR/SM) mode or in Dynamic Partition Manger (DPM) mode. For more information, see Appendix E, “IBM Dynamic Partition Manager” on page 483. CPs are defined as either dedicated or shared. Reserved CPs can be defined to an LPAR to allow for nondisruptive image upgrades. If the operating system in the LPAR supports the logical processor add function, reserved processors are no longer needed. Regardless of the installed model, an LPAR can have up to 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 HMC workplace. Any z/Architecture operating systems, CFCCs, and IBM zAware can run on CPs that are assigned from the CP pool.

The z14 server recognizes four distinct capacity settings for CPs. Full-capacity CPs are identified as CP7. In addition to full-capacity CPs, three subcapacity settings (CP6, CP5, and CP4), each for up to 33 CPs, are offered.

These capacity settings appear in hardware descriptions:

- CP7 Feature Code #1932
- CP6 Feature Code #1931
- CP5 Feature Code #1930
- CP4 Feature Code #1929

Granular capacity adds 90 subcapacity settings to the 170 capacity settings that are available with full capacity CPs (CP7). Each of the 90 subcapacity settings applies only to up to 33 CPs, independently of the model installed.

Information about CPs in the remainder of this chapter applies to all CP capacity settings, unless indicated otherwise. For more information about granular capacity, see 2.7, “Model configurations” on page 67.

3.5.3 Integrated Facility for Linux

An IFL is a PU that can be used to run Linux, Linux guests on z/VM operating systems, and Secure Service Container (SSC). Up to 170 PUs can be characterized as IFLs, depending on the configuration. IFLs can be dedicated to a Linux, a z/VM, or an IBM zAware LPAR, or can be shared by multiple Linux guests, z/VM LPARs, or SSC running on the same z14 server.

Only z/VM, Linux on z Systems operating systems, SSC, and designated software products can run on IFLs. IFLs are orderable by using FC #1933.

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 Hardware Management Console (HMC) workplace.

IFLs 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 IFLs.

Unassigned IFLs

An IFL that is purchased but not activated is registered as an unassigned IFL (FC 1937). When the system is later upgraded with an additional IFL, the system recognizes that an IFL already was purchased and is present.

3.5.4 Internal Coupling Facility

An Internal Coupling Facility (ICF) is a PU that is used to run the CFCC for Parallel Sysplex environments. Within the sum of all unassigned PUs in up to four CPC drawers, up to 170 ICFs can be characterized, depending on the model. However, the maximum number of ICFs that can be defined on a coupling facility LPAR is limited to 16. ICFs are orderable by using FC 1934.

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

Coupling Thin Interrupts

With the introduction of Driver 15F (zEC12 and 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 already 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, existing “poll for work” logic in both CFCC and z/OS can be used largely as is to locate and process the work. The new interrupt simply expedites the redispatching of the partition.

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

Coupling facility processor combinations

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

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

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

In Figure 3-12, the CPC on the left 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.

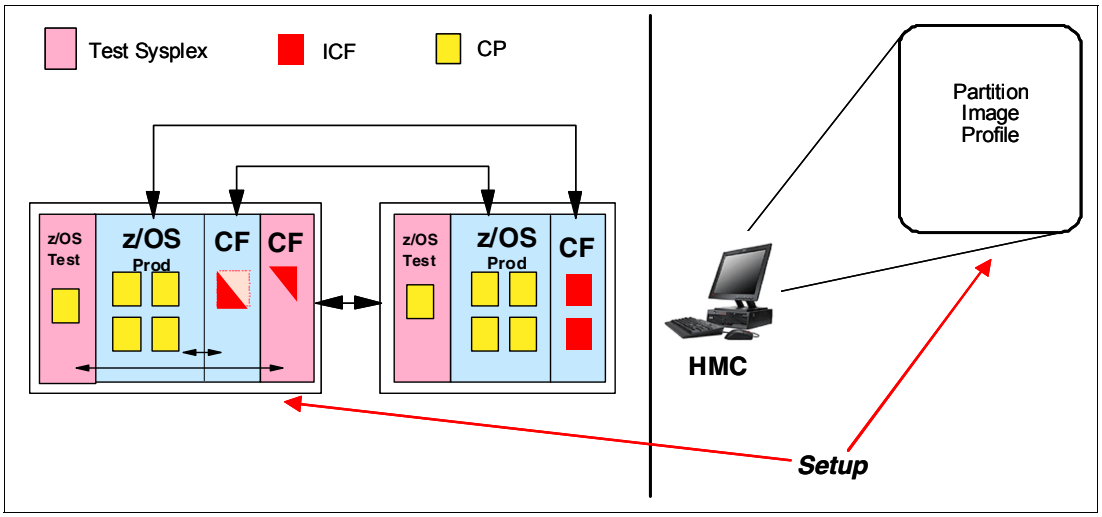


Figure 3-12 ICF options - shared ICFs

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

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

Dynamic coupling facility dispatching

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

CFCC Level 19 introduced Coupling Thin Interrupts and the new DYNDISP specification. It allows more environments with multiple CF images to coexist in a server, and to share CF engines with reasonable performance. For more information, see 3.9.3, “Dynamic CF dispatching” on page 135.

Coupling Facility Processor Scalability

Coupling Facility work management and dispatcher have changed to allow improved efficiency as additional 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 may be forced to split the CF workload across more CF images, but this brings additional 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 existing 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):

- Non-prioritized (FIFO-based) work queues; This avoids overhead of maintaining ordered queues in the CF
- Streamlined system-managed duplexing protocol; This avoids costly latching deadlocks that can occur between primary/secondary structure
- “Functionally specialized” ICF processors that operate for CF images with dedicated processors defined under certain conditions
  - 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 lots of inter-processor contention previously associated with CF dispatching

3.5.5 IBM Z Integrated Information Processor

A zIIP[9] reduces the standard processor (CP) capacity requirements for z/OS Java, XML system services applications, and a portion of work of z/OS Communications Server and DB2 UDB for z/OS Version 8 or later, freeing up capacity for other workload requirements.

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

z14 is the second generation of IBM Z processors to support SMT. z14 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 has been enhanced for z14 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

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

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 redispatches the JVM task on an available CP. After this process occurs, normal processing is resumed.

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

Figure 3-13 shows the logical flow of Java code running on a z14 server that has a zIIP available. When JVM starts the execution of a Java program, it passes control to the z/OS dispatcher that verifies the availability of a zIIP.

The availability is treated in the following manner:

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

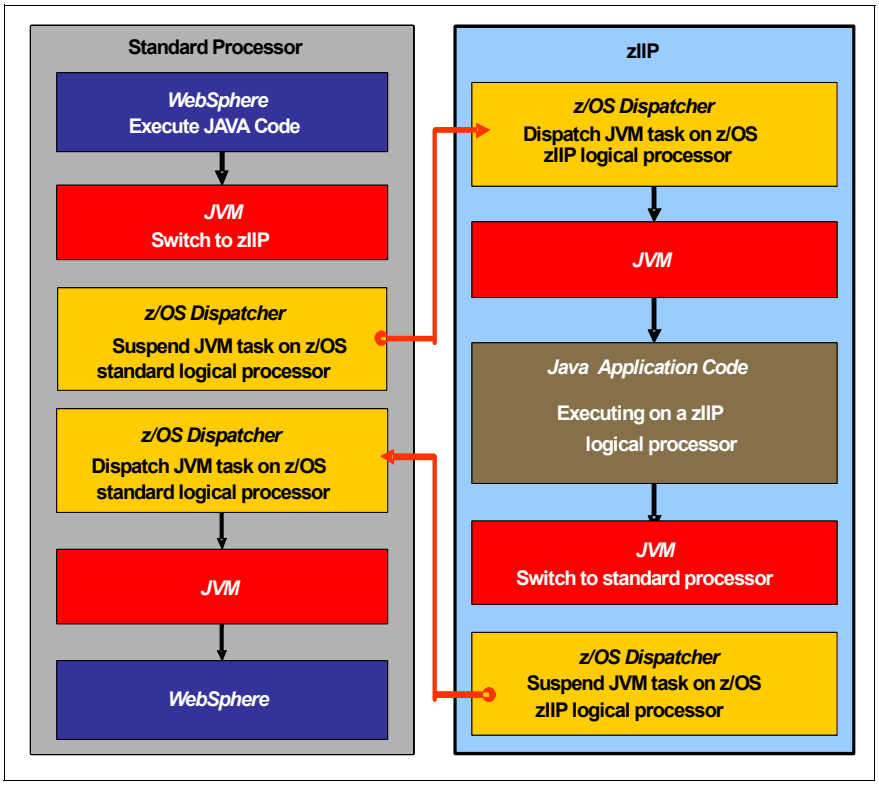


Figure 3-13 Logical flow of Java code execution on a zIIP

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

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

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

- Option 1: Java dispatching by priority (IIPHONORPRIORITY=YES): This is the default option, and specifies that CPs must not automatically consider zIIP-eligible work for dispatching on them. The zIIP-eligible work is dispatched on the zIIP engines until 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, and allows it to spill over onto the CPs only when the zIIPs are overcommitted.

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

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

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

zIIPs provide the following benefits:

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

The following DB2 UDB for z/OS V8 or later workloads are eligible to run in 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, mostly used in business intelligence (BI) work. A *star schema* is a relational database schema for representing multidimensional data. It stores data in a central fact table and is surrounded by more dimension tables that hold information about each perspective of the data. A star schema query, for example, joins various dimensions of a star schema data set.
- DB2 utilities that are used for index maintenance, such as LOAD, REORG, and REBUILD. Indexes allow quick access to table rows, but over time, as data in large databases is manipulated, the databases become less efficient and must be maintained.

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

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

For more information about zIIP and eligible workloads, see the IBM zIIP website:

http://www.ibm.com/systems/z/hardware/features/ziip/index.html

zIIP installation information

One CP must be installed with or before any zIIP is installed. In z14, the zIIP to CP ratio is 2:1, which means that up to 112 zIIPs on a model M05 can be characterized. Table 3-1 shows the allowed number of zIIPs for each model.

Table 3-1 Number of zIIPs per model

Model	M01	M02	M03	M04	M05
Maximum zIIPs	0 - 22	0 - 46	0 - 70	0 - 94	0 - 112

zIIPs are orderable by using FC 1936. Up to two zIIPs can be ordered for each CP or marked CP configured in the system. If the installed CPC drawer has no remaining unassigned PUs, the assignment of the next zIIP might require the installation of an additional CPC drawer.

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

The number of permanent zIIPs plus temporary zIIPs cannot exceed twice the number of purchased CPs plus temporary CPs. Also, the number of temporary zIIPs cannot exceed the number of permanent zIIPs.

zIIPs and logical partition definitions

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

Logical partition: In a logical partition, as many zIIPs as are available can be defined together with at least one CP.

3.5.6 System assist processors

A system assist processor (SAP) is a PU that runs the channel subsystem LIC to control I/O operations. All SAPs run I/O operations for all LPARs. z14 is the first IBM Z server to enable SMT for SAPs. All models have standard SAPs configured. The number of standard SAPs depends on the z14 model, as shown in Table 3-2.

Table 3-2 SAPs per model

Model	M01	M02	M03	M04	M05
Standard SAPs	5	10	15	20	23

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. Assignment of more SAPs can increase the capability of the channel subsystem to run I/O operations. In z14 systems, the number of SAPs can be greater than the number of CPs. However, additional SAPs plus standard SAPs cannot exceed 128.

Optional additional orderable SAPs

The option to order more SAPs is available on all models (FC #1935). These additional SAPs increase the capacity of the channel subsystem to run I/O operation, which is suggested for TPF environments. The maximum number of optional additional orderable SAPs depends on the configuration and the number of available uncharacterized PUs. The number of SAPs is listed in Table 3-3.

Table 3-3 Optional SAPs per model

Model	M01	M02	M03	M04	M05
Optional SAPs	0 - 4	0 - 8	0 - 12	0 - 16	0 - 16

3.5.7 Reserved processors

Reserved processors are defined by the PR/SM to allow for a nondisruptive capacity upgrade. Reserved processors are like spare logical processors, and can be shared or dedicated. Reserved CPs can be defined to an LPAR dynamically to allow for nondisruptive image upgrades.

Reserved processors can be dynamically configured online by an operating system that supports this function, if enough unassigned PUs are available to satisfy 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 configure online, nondisruptively, more logical processors after additional 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 already 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 120.

3.5.8 Integrated firmware processor

An IFP is allocated from the pool of PUs and is available for the whole system. Unlike other characterized PUs, the IFP is standard and not defined by the client. It is a single PU dedicated solely to supporting the *native* Peripheral Component Interconnect Express (PCIe) features (10GbE Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE) Express, RoCE Express2, zEnterprise Data Compression (zEDC) Express, IBM zHyperLink Express, and Coupling Express Long Reach) and is initialized at POR. The IFP supports Resource Group (RG) LIC to provide native PCIe I/O feature management and virtualization functions. For more information, see Appendix C, “Native Peripheral Component Interconnect Express (PCIe)” on page 451.

3.5.9 Processor unit assignment

The processor unit assignment of characterized PUs is done at POR time, when the system is initialized. The initial assignment rules keep PUs of the same characterization type grouped as much as possible in relation to PU chips and CPC drawer boundaries to optimize shared cache usage.

The PU assignment is based on CPC drawer plug ordering. The CPC drawers are populated from the bottom upward. This process defines the low-order and the high-order CPC drawers:

- CPC drawer 1: Plug order 1 (low-order CPC drawer)
- CPC drawer 2: Plug order 2
- CPC drawer 3: Plug order 3
- CPC drawer 4: Plug order 4 (high-order CPC drawer)

The assignment rules follow this order:

- Spare: CPC drawers 1 and 2 are assigned one spare each on the high PU chip. In the model M01, the two spares are assigned to CPC drawer 1.
- IFP: One IFP is assigned to CPC drawer 1.
- SAPs: Spread across CPC drawers and high PU chips. Each CPC drawer has six standard SAPs. Start with the highest PU chip high core, then the next highest PU chip high core. This process prevents all the SAPs from being assigned on one PU chip.

- IFLs and ICFs: Assign IFLs and ICFs to cores on chips in higher CPC drawers working downward.
- CPs and zIIPs: Assign CPs and zIIPs to cores on chips in lower CPC drawers working upward.

These rules are intended to isolate, as much as possible, on different CPC drawers and even on different PU chips, processors that are used by different operating systems. This configuration ensures that different operating systems do not use the same shared caches. For example, CPs and zIIPs are all used by z/OS, and can benefit by using the same shared caches. However, IFLs are used by z/VM and Linux, and ICFs are used by CFCC. Therefore, for performance reasons, the assignment rules prevent them from sharing L3 and L4 caches with z/OS processors.

This initial PU assignment, which is done at POR, can be dynamically rearranged by an LPAR by swapping an active core to a core in a different PU chip in a different CPC drawer or node to improve system performance. For more information, see “LPAR dynamic PU reassignment” on page 126.

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

3.5.10 Sparing rules

On a z14 system, two PUs are reserved as spares. The reserved spares are available to replace any two characterized PUs, whether they are CP, IFL, ICF, zIIP, SAP, or IFP.

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

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

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

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

Application preservation

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

Dynamic SAP and IFP sparing and reassignment

Dynamic recovery is provided if there is a failure of the SAP or IFP. If the SAP or IFP fails, and if a spare PU is available, the spare PU is dynamically assigned as a new SAP or IFP. If no spare PU is available, and more than one CP is characterized, a characterized CP is reassigned as an SAP or IFP. In either case, client intervention is not required. This capability eliminates an unplanned outage and allows a service action to be deferred to a more convenient time.

3.6 Memory design

This section describes various considerations of the z14 memory design.

3.6.1 Overview

The z14 memory design also provides flexibility, high availability, and upgrades:

- Concurrent memory upgrades (if the physically installed capacity is not yet reached): z14 servers can have more physically installed memory than the initial available capacity. Memory upgrades within the physically installed capacity can be done concurrently by LIC, and no hardware changes are required. However, memory upgrades *cannot* be done through CBU or On/Off CoD.
- Concurrent memory upgrades (if the physically installed capacity is reached): Physical memory upgrades require a processor cage to be removed and reinstalled after replacing the memory cards in the processor cage. Except for a model M01, the combination of enhanced drawer availability and the flexible memory option allows you to concurrently add memory to the system. For more information, see 2.4.5, “Drawer replacement and memory” on page 60 and 2.4.7, “Flexible Memory Option” on page 61.

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

Memory allocation

When system is activated by using a POR, PR/SM determines the total installed memory and the customer enabled memory. Later in the process, during LPAR activation, PR/SM assigns and allocates each partition memory according to their image profile.

PR/SM has control over all physical memory, and can so make physical memory available to the configuration when a CPC drawer is added.

In prior IBM Z processors, memory allocation was striped across the available CPC drawers because there was relatively fast connectivity between the drawers, and splitting the work between all of the memory controllers allowed a smooth performance variability.

With the z14, z13 and z13s memory design, the memory allocation algorithm has changed. PR/SM tries to allocate memory into a single CPC drawer. Basically, the PR/SM memory and logical processor resources allocation goal is to place all partition resources on a single CPC drawer, if possible. The resources, like memory and logical processors, are assigned to the logical partitions at the time of their activation. Later on, when all partitions are activated, PR/SM can move memory between CPC drawers to benefit the performance of each LPAR, without OS knowledge. This process was done on the previous families of IBM Z servers only for PUs that use PR/SM dynamic PU reallocation.

With z14 servers, this process happens each time the configuration changes, such as in these circumstances:

- Activating an LPAR
- Deactivating an LPAR
- Changing the LPARs processing weights
- Upgrade the system through a temporary or permanent record
- Downgrade the system through a temporary record

PR/SM schedules a global reoptimization of the resources in use. It does so by looking at all the partitions that are active and prioritizing them based on their processing entitlement and weights, creating a high and low priority rank. Then the resources, such as logical processors and memory, can be moved from one CPC drawer to another to address the priority ranks that were just created.

When partitions are activated, PR/SM tries to find a home assignment CPC drawer, home assignment node, and home assignment chip for the logical processors that are defined to them. The PR/SM goal is to allocate all the partition logical processors and memory to a single CPC drawer (the home drawer for that partition). If all logical processors can be assigned to a home drawer and the partition defined memory is greater than what is available in that drawer, the exceeding memory amount is allocated on another CPC drawer. If all the logical processors cannot fit in one CPC drawer, the remaining logical processors spill to another CPC drawer. When that overlap happens, PR/SM stripes the memory, if possible, across the CPC drawers where the logical processors are assigned.

The process of reallocating memory is based on the *memory copy/reassign* function, which is used to allow enhanced drawer availability (EDA) and concurrent drawer replacement (CDR)[10]. This process has been enhanced for z13 and z13s to provide more efficiency and speed to the process without affecting system performance. Furthermore, z14 implements a faster dynamic memory reallocation mechanism, especially useful during service operations (EDA, CDR). PR/SM controls the reassignment of the content of a specific physical memory array in one CPC drawer to a physical memory array in another CPC drawer. To do that, PR/SM uses all the available physical memory in the system. This memory includes the memory not in use by the system, that is available but not purchased by the client, and the planned memory options, if installed.

Because of the memory allocation algorithm, systems that undergo a number of miscellaneous equipment specification (MES) upgrades for memory can have different memory mixes and quantities in all processor drawers of the system. If the memory fails, it is technically feasible to run a POR of the system with the remaining working memory resources. After the POR completes, the memory distribution across the processor drawers will be different, as will be the total amount of available memory.

Large page support

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

The TLB exists to reduce the amount of time that is required to translate a virtual address to a real address. This translation is done by dynamic address translation (DAT) when it must find the correct page for the correct address space. Each TLB entry represents one page. Like other buffers or caches, lines are discarded from the TLB on a least recently used (LRU) basis. The worst-case translation time occurs when a TLB miss occurs and both the segment table (which is needed to find the page table) and the page table (which is needed to find the entry for the particular page in question) are not in cache. This case involves two complete real memory access delays plus the address translation delay. The duration of a processor cycle is much shorter than the duration of a memory cycle, so a TLB miss is relatively costly.

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

[10] In previous IBM Z generations (prior to z13), these were known as enhanced book availability (EBA) and concurrent book repair (CBR).

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 both elapsed time and processor usage. These improvements are because DAT and memory operations are part of processor busy time even though the processor waits for memory operations to complete without processing anything else in the meantime.

To overcome the processor usage that is associated with creating a 1 MB page, a process must run for some time. It must maintain frequent memory access to keep the pertinent addresses in the TLB.

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

Long-running work with low memory-access frequency is less likely to maintain its entries in the TLB. However, when it does run, a smaller number of address translations is required to resolve all the memory it needs. Therefore, a long-running process can benefit even without frequent memory access. Weigh the benefits of whether something in this category must use large pages as a result of the system-level costs of tying up real storage. There is a balance between the performance of a process that uses large pages, and the performance of the remaining work on the system.

On z14 servers, 1 MB large pages become pageable if Virtual Flash Memory[11] is available and enabled. They are available only for 64-bit virtual private storage, such as virtual memory above 2 GB.

It is easy to assume that increasing the TLB size is a feasible option to deal with TLB-miss situations. However, this process is not as straightforward as it seems. As the size of the TLB increases, so does the processor usage that is involved in managing the TLB'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 these functions:

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

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

[11] Virtual Flash Memory has replaced IBM zFlash Express for z14. There is no carry forward of zFlash Express

3.6.3 Hardware system area

The HSA is a non-addressable storage area that contains system LIC and configuration-dependent control blocks. On z14 servers, the HSA has a fixed size of 192 GB and is not part of the purchased memory that you order and install.

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

- Six channel subsystems (CSSs)
- Fifteen LPARs in the five CSSs and 10 LPARs for the sixth CSS for a total of 85 LPARs
- Subchannel set 0 with 63.75-K devices in each CSS
- Subchannel set 1 with 64-K devices in each CSS
- Subchannel set 2 with 64-K devices in each CSS
- Subchannel set 3 with 64-K devices in each CSS

The HSA has sufficient reserved space to allow for dynamic I/O reconfiguration changes to the maximum capability of the processor.

3.6.4 Virtual Flash Memory

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

3.7 Logical partitioning

This section addresses logical partitioning features.

3.7.1 Overview

Logical partitioning is a function implemented by the PR/SM on z14. z14 runs either in LPAR-, or in Dynamic Partition Manager (DPM) mode. DPM provides a dynamic GUI for managing PR/SM. Therefore, all system aspects are controlled by PR/SM functions.

PR/SM is aware of the processor drawer structure on z14 servers. However, LPARs do not have this awareness. LPARs have resources that are allocated to them from various physical resources. From a systems standpoint, LPARs have no control over these physical resources, but the PR/SM functions do.

PR/SM manages and optimizes allocation and the dispatching of work on the physical topology. Most physical topology that was previously handled by the operating systems is the responsibility of PR/SM.

As shown in 3.5.9, “Processor unit assignment” on page 115, the initial PU assignment is done during POR by using rules to optimize cache usage. This is the “physical” step, where CPs, zIIPs, IFLs, ICFs, and SAPs are allocated on the processor drawers.

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

Memory allocation has changed from the previous IBM Z servers. IBM System z9® memory used to be spread across all books. This optimization was done by using a round-robin

algorithm with a number of increments per book to match the number of memory controllers (MCs) per book. This memory allocation design is driven by performance results, which minimizes variability for most workloads.

With z14 servers, memory allocation has changed from the model that was used for the z9. Partition memory is now allocated in a per processor drawer basis and striped across processor clusters. For more information, see “Memory allocation” on page 117.

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

Processor drawers and node level assignments are more important because they optimize L4 cache usage. Therefore, logical processors from a specific LPAR are packed into a processor drawer as much as possible.

PR/SM optimizes chip assignments within the assigned processor drawers (or drawers) to maximize L3 cache efficiency. Logical processors from an LPAR are dispatched on physical processors on the same PU chip as much as possible. The number of processors per chip (10) matches the number of z/OS processor affinity queues that is used by HiperDispatch, achieving optimal cache usage within an affinity node.

PR/SM also tries to redispatch a logical processor on the same physical processor to optimize private cache (L1 and L2) usage.

HiperDispatch

PR/SM and z/OS work in tandem to 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 redispatching units of work to the same processor group, keeping processes running near their cached instructions and data, and minimizing transfers of data ownership among processors and processor drawers.

The nested topology is returned to z/OS by the Store System Information (STSI) instruction. HiperDispatch uses the information to concentrate logical processors around shared caches (L3 at PU chip level, and L4 at drawer level), and dynamically optimizes the assignment of logical processors and units of work.

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

Tip: z/VM V6.3 and later also support HiperDispatch.

Logical partitions

PR/SM enables z14 servers to be initialized for a logically partitioned operation, supporting up to 85 LPARs. Each LPAR can run its own operating system image in any image mode, independently from the other LPARs.

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

Each LPAR has the same resources as a real CPC:

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

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

  The weight and number of online logical processors of an LPAR can be dynamically managed by the LPAR CPU Management function of the Intelligent Resource Director (IRD). These can be used to achieve the defined goals of this specific partition and of the overall system. The provisioning architecture of z14 servers, described in Chapter 8, “System upgrades” on page 299, adds another dimension to the dynamic management of LPARs.

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

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

  For the z/OS Workload License Charges (WLC) pricing metric, and metrics that are based on it, such as Advanced Workload License Charges (AWLC), an LPAR *defined capacity* can be set. This defined capacity enables the soft capping function. Workload charging introduces the capability to pay software license fees based on the processor utilization of the LPAR on which the product is running, rather than on the total capacity of the system:

  – In support of WLC, the user can specify a defined capacity in millions of service units (MSUs) per hour. The defined capacity sets the capacity of an individual LPAR when soft capping is selected.

    The defined capacity value is specified on the Options tab in the Customize Image Profiles window.

  – WLM keeps a 4-hour rolling average of the processor usage of the LPAR. When the 4-hour average processor consumption exceeds the defined capacity limit, WLM dynamically activates LPAR capping (soft capping). When the rolling 4-hour average returns below the defined capacity, the soft cap is removed.

  For more information about WLM, see *System Programmer's Guide to: Workload Manager*, SG24-6472. For a review of software licensing, see 7.8, “Software licensing” on page 294.

  Weight settings: When defined capacity is used to define an uncapped LPAR’s capacity, carefully consider the weight settings of that LPAR. If the weight is much smaller than the defined capacity, PR/SM uses a discontinuous cap pattern to achieve the defined capacity setting. This configuration means PR/SM alternates between capping the LPAR at the MSU value corresponding to the relative weight settings, and no capping at all. It is best to avoid this scenario, and try to establish a defined capacity that is equal or close to the relative weight.

- Memory: Memory (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, enabling nondisruptive memory addition to and removal from an LPAR, by using the LPAR dynamic storage reconfiguration (z/OS and z/VM). For more information, see 3.7.5, “LPAR dynamic storage reconfiguration” on page 130.

- Channels: Channels can be shared between LPARs by including the partition name in the partition list of a channel-path identifier (CHPID). I/O configurations are defined by the IOCP or the HCD with the CHPID mapping tool (CMT). The CMT is an optional tool that is used to map CHPIDs onto physical channel IDs (PCHIDs). PCHIDs represent the physical location of a port on a card in an I/O cage, I/O drawer, or PCIe I/O drawer.

  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*, SB10-7163. HCD is available on the z/OS and z/VM operating systems. Consult the appropriate 2964DEVICE Preventive Service Planning (PSP) buckets before implementation.

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

Modes of operation

Table 3-4 shows the nodes of operation, summarizing all available mode combinations, including their operating modes and processor types, operating systems, and addressing modes. Only the currently supported versions of operating systems are considered.

Table 3-4 z14 modes of operation

Image mode	PU type	Operating system	Addressing mode
General[a] (formerly ESA/390 mode)	CP *and* zIIP	z/OS z/VM	64-bit
	CP	z/VSE Linux on z Systems z/TPF	64-bit
Coupling facility	ICF or CP	CFCC	64-bit
Linux only	IFL *or* CP	Linux on z Systems (64-bit)	64-bit
		z/VM	
		Linux on z Systems (31-bit)	31-bit
z/VM	CP, IFL, zIIP, or ICF	z/VM	64-bit
SSC[b]	IFL or CP	z/VSE Network Appliance[c]	64 bit

a. General mode uses 64 bit z/Architecture
b. Secure Service Container
c. More appliances to be announced and supported in the future

The 64-bit z/Architecture mode has no special operating mode because the architecture mode is not an attribute of the definable images operating mode. The 64-bit operating systems are in 31-bit mode at IPL and change to 64-bit mode during their initialization. The operating system is responsible for taking advantage of the addressing capabilities that are provided by the architectural mode.

For information about operating system support, see Chapter 7, “Operating System Support” on page 227

Logically partitioned mode

If the z14 server runs in LPAR mode, each of the 85 LPARs can be defined to operate in one of the following image modes:

- General (formerly ESA/390) mode to run the following systems:
  - A z/Architecture operating system, on dedicated or shared CPs
  - ESA/390 operating systems
  - A Linux on z Systems operating system, on dedicated or shared CPs
  - z/OS, on any of the following processor units:
    - Dedicated or shared CPs
    - Dedicated CPs *and* dedicated zIIPs
    - Shared CPs *and* shared zIIPs

    zIIP usage: zIIPs can be defined to an ESA/390 mode or z/VM mode image, as shown in Table 3-4 on page 123. However, zIIPs are used only by z/OS. Other operating systems cannot use zIIPs even if they are defined to the LPAR. z/VM V5R4[a] and later provide support for real and virtual zIIPs to guest z/OS systems.

    a. z/VM V5R4 is not supported on IBM z14

- General mode is used also to run the z/TPF operating system, on dedicated or shared CPs
- Coupling facility mode, by loading the CFCC code into the LPAR that is defined as one of these types:
  - Dedicated or shared CPs
  - Dedicated or shared ICFs
- Linux only mode to run these systems:
  - A Linux on z Systems operating system, on either of these types:
    - Dedicated or shared IFLs
    - Dedicated or shared CPs
  - A z/VM operating system, on either of these types:
    - Dedicated or shared IFLs
    - Dedicated or shared CPs
- z/VM mode to run z/VM on dedicated or shared CPs or IFLs, plus zIIPs and ICFs
- Secure Service Container (SSC) mode LPAR can run on:
  - Dedicated or shared CPs
  - Dedicated or shared IFLs

Table 3-5 shows all LPAR modes, required characterized PUs, operating systems, and the PU characterizations that can be configured to an LPAR image. The available combinations of dedicated (DED) and shared (SHR) processors are also shown. For all combinations, an LPAR also can have reserved processors that are defined, allowing nondisruptive LPAR upgrades.

Table 3-5 LPAR mode and PU usage

LPAR mode	PU type	Operating systems	PUs usage
General (formerly ESA/390)	CPs	z/Architecture operating systems ESA/390 operating systems Linux on z Systems	CPs DED or CPs SHR
	CPs *and* zIIPs	z/OS z/VM (V6R3 and later for guest exploitation)	CPs DED *and* (*or*) zIIPs DED *or* CPs SHR *and* (*or*) zIIPs SHR
General	CPs	z/TPF	CPs DED or CPs SHR
Coupling facility	ICFs *or* CPs	CFCC	ICFs DED *or* ICFs SHR, *or* CPs DED *or* CPs SHR
Linux only	IFLs *or* CPs	Linux on z Systems z/VM	IFLs DED *or* IFLs SHR, *or* CPs DED *or* CPs SHR
z/VM	CPs, IFLs, zIIPs, or ICFs	z/VM (V6R2 and later)	All PUs must be SHR or DED
SSC[a]	IFLs, *or* CPs	IBM zAware[b] z/VSE Network Appliance[c]	IFLs DED *or* IFLs SHR, *or* CPs DED *or* CPs SHR

a. Secure Service Container
b. Encapsulated as a firmware appliance
c. Planned

Dynamically adding or deleting a logical partition name

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

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

Adding logical processors to a logical partition

Logical processors can be concurrently added to an LPAR by defining them as reserved in the image profile and later configuring them online to the operating system by using the appropriate console commands. Logical processors 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 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 Express5S features to an LPAR on the crypto page in the image profile by defining the Cryptographic Candidate List, and the Usage and Control Domain indexes, in the partition profile. By using the Change LPAR Cryptographic Controls task, you can add crypto adapters dynamically to an LPAR without an outage of the LPAR. Also, dynamic deletion or moving of these features does not require pre-planning. Support is provided in z/OS, z/VM, z/VSE, Secure Service Container (based on appliance requirements), and Linux on z Systems.

LPAR dynamic PU reassignment

The system configuration is enhanced to optimize the PU-to-CPC drawer assignment of physical processors dynamically. The initial assignment of client-usable physical processors to physical processor drawers can change dynamically to better suit the actual LPAR configurations that are in use. For more information, see 3.5.9, “Processor unit assignment” on page 115. 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 processor drawers.

LPAR dynamic PU reassignment can swap client processors of different types between processor drawers. For example, reassignment can swap an IFL on processor drawer 1 with a CP on processor drawer 2. Swaps can also occur between PU chips within a processor drawer or a node and can include spare PUs. The goals are to pack the LPAR on fewer processor drawers and also on fewer PU chips, based on the z14 processor drawers’ topology. The effect of this process is evident in dedicated and shared LPARs that use HiperDispatch.

LPAR dynamic PU reassignment is transparent to operating systems.

LPAR group capacity limit (LPAR group absolute capping)

The group capacity limit feature allows the definition of a group of LPARs on a z14 system, and limits the combined capacity usage by those LPARs. This process allows the system to manage the group so that the group capacity limits in MSUs per hour are not exceeded. To take advantage of this feature, you must be running z/OS V1.10 or later in the all LPARs in the group.

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

LPAR absolute capping

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

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

Absolute capping provides an optional maximum capacity setting for logical partitions that is specified in 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 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 mode of operation that provides a simplified approach to create and manage virtualized environments, reducing the barriers of its adoption for new and existing customers.

The implementation provides built-in integrated capabilities that allow advanced virtualization management on IBM Z servers. With DPM, you can use your existing 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, using a firmware managed environment, that coordinate the physical system resources shared by the partitions. The partitions' resources include processors, memory, network, storage, crypto, and accelerators.

DPM provides a new mode of operation for IBM Z servers that provide these 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 an additional hypervisor for IBM Z servers. DPM uses the existing PR/SM hypervisor infrastructure and on top of it, provides an intelligent interface that allows customers to define, use, and operate the platform virtualization without prior IBM Z experience or skills. For more information about DPM, see Appendix E, "IBM Dynamic Partition Manager" on page 483.

3.7.2 Storage operations

In z14 servers, memory can be assigned as main storage only (expanded storage is not supported) supporting up to 85 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 V6R3 support 1 TB while 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 130.

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

Table 3-6 shows the z14 storage allocation and usage possibilities, depending on the image mode.

Table 3-6 Main storage definition and usage possibilities

Image mode	Architecture mode (addressability)	Maximum main storage	
		Architecture	z14 definition
General (formerly ESA/390)	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[a]	z/Architecture (64-bit)	16 EB	2 TB
SSC[b]	z/Architecture (64-bit)	16 EB	2 TB

a. z/VM mode is supported by z/VM V6R2 and later.
b. 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, allowing for virtual addresses up to 16 exabytes (16 EB). The 64-bit architecture theoretically allows a maximum of 16 EB to be used as main storage. However, the current main storage[12] limit for LPARs is 16 TB of main storage. The operating system that runs in z/Architecture mode must be able to support the real storage. Currently, z/OS supports up to 4 TB[13] of real storage (z/OS V1R10 and later releases).

Coupling facility mode

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

- CFCC Level 22, available o z14 (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
- CFCC Level 19, available for the zEC12 with Driver Level 15F

For more information, see 3.9.1, “Coupling Facility Control Code” on page 133. Expanded storage cannot be defined for a coupling facility image. Only IBM CFCC can run in coupling facility mode.

[12] Expanded storage is NOT supported on z14
[13] 1 TB if an I/O drawer is installed in the z13 system (carry forward only). z14 does not support I/O drawer.

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 Systems 64-bit distributions (SUSE SLES 10 and later, and Red Hat RHEL 5 and later) use 64-bit addressing and operate in z/Architecture mode. z/VM also uses 64-bit addressing and operates in z/Architecture mode.

z/VM mode

In z/VM mode, certain types of processor units can be defined within one LPAR. This increases flexibility and simplifies systems management by allowing z/VM to run the following tasks in the same z/VM LPAR:

- Manage guests to operate Linux on z Systems on IFLs
- Operate z/VSE and z/OS on CPs
- Offload z/OS system software processor usage, such as DB2 workloads on zIIPs
- Provide an economical Java execution environment under z/OS on zIIPs

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 definition limit for LPARs is 10 TB of storage.

Currently, the z/VSE Network Appliance which is available on z14 and z13s servers runs in a SSC LPAR.

3.7.3 Reserved storage

Reserved storage can be optionally defined to an LPAR, allowing a nondisruptive image memory upgrade for this partition. Reserved storage can be defined to both central and expanded storage, and to any image mode except coupling facility mode.

An LPAR must define an amount of main storage and, optionally (if not a coupling facility image), an amount of expanded storage. Both main storage and expanded storage can have two storage sizes defined:

- The initial value is the storage size that is allocated to the partition when it is activated.
- The reserved value is an additional storage capacity beyond its initial storage size that an LPAR can acquire dynamically. The reserved storage sizes that are defined to an LPAR do not have to be available when the partition is activated. They are predefined storage sizes to allow a storage increase, from an LPAR point of view.

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

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

The additional storage capacity for an LPAR upgrade can come from these sources:

- Any unused available storage
- Another partition that has released storage
- A memory upgrade

A concurrent LPAR storage upgrade uses DSR. z/OS uses the reconfigurable storage unit (RSU) definition to add or remove storage units in a nondisruptive way.

z/VM V5R4 and later releases support the dynamic addition of memory to a running LPAR by using reserved storage. It also virtualizes this support to its guests. Removal of storage from the z/VM LPAR is disruptive. Removal of memory from a z/VM guest is not disruptive to the z/VM LPAR.

SLES 11 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 either initial or reserved main storage, as shown in Table 3-7.

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

Logical partition: Largest main storage amount	**Logical partition: 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
8 TB < main storage amount <= 16 TB	32 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 16 TB of main storage. For z/VM V6R3, the limit is 1 TB, whereas for z/VM V6R4 the limit is 2 TB.

3.7.5 LPAR dynamic storage reconfiguration

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

With dynamic storage reconfiguration, the unused storage does not have to be continuous.

When an operating system running on an LPAR assigns a storage increment to its configuration, PR/SM determines whether any free storage increments are available. PR/SM then dynamically brings the storage online.

PR/SM dynamically takes offline a storage increment and makes it available to other partitions when an operating system running on an LPAR releases a storage increment.

3.8 Intelligent Resource Director (IRD)

IRD is a z14 and IBM Z capability that is used only by z/OS. IRD is a function that optimizes processor and channel resource utilization across LPARs within a single 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 or Z servers running in LPAR mode, and in the same Parallel Sysplex, into an *LPAR cluster*. This configuration allows WLM to manage resources, both processor and I/O, not just in one single image, but across the entire cluster of system images.

Figure 3-14 shows an LPAR cluster. It contains three z/OS images, and one Linux image that is managed by the cluster. Included as part of the entire Parallel Sysplex is another z/OS image and a coupling facility image. In this example, the scope over which IRD has control is the defined LPAR cluster.

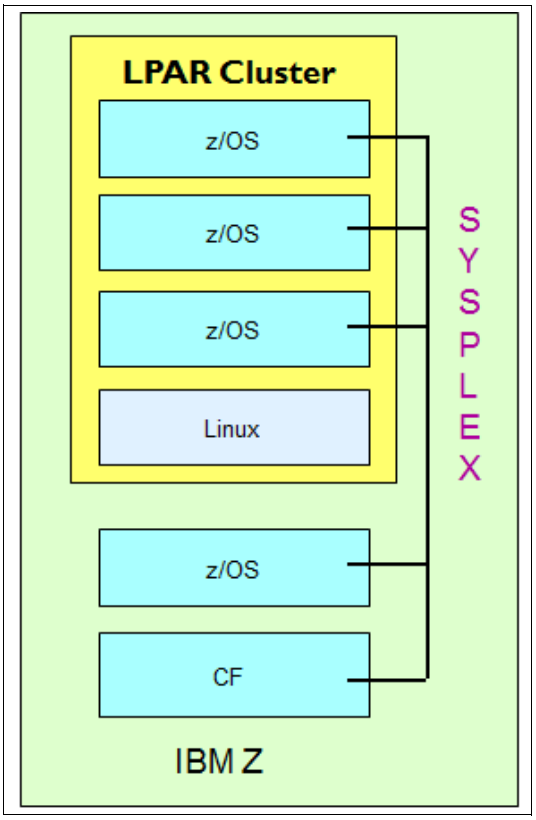


Figure 3-14 IRD LPAR cluster example

IRD has the following characteristics:

- IRD processor management: WLM dynamically adjusts the number of logical processors within an LPAR and the processor weight based on the WLM policy. The ability to move the processor weights across an LPAR cluster provides processing power where it is most needed, based on WLM goal mode policy.

  The processor management function is automatically deactivated when HiperDispatch is active. However, the LPAR weight management function remains active with IRD with HiperDispatch. For more information about HiperDispatch, see 3.7, “Logical partitioning” on page 120.

  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 servers support DCM within a channel subsystem.
- Channel subsystem priority queuing: This function on z14 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 servers. Figure 3-15 illustrates the components of a Parallel Sysplex as implemented within the Z architecture. The figure is intended only as an example. It shows one of many possible Parallel Sysplex configurations.

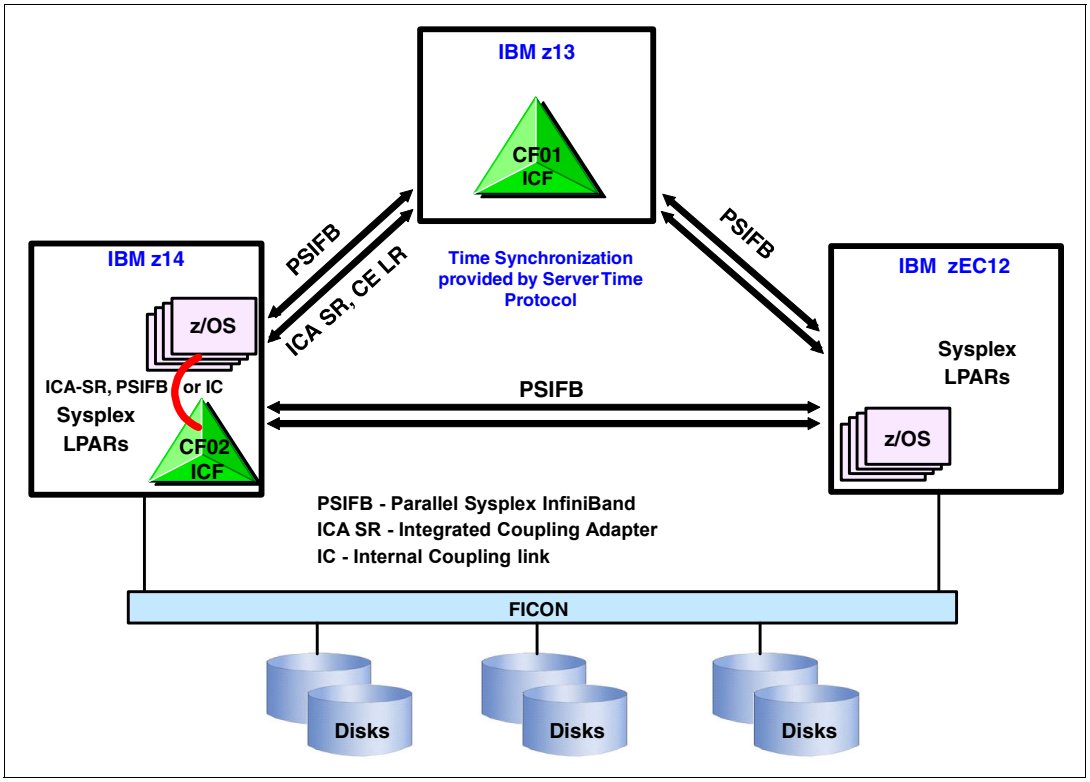


Figure 3-15 Sysplex hardware overview

Figure 3-15 shows a z14 system that contains multiple z/OS sysplex partitions. It contains an internal coupling facility (CF02), a z13 system that contains a stand-alone CF (CF01), and a zEC12 that contains multiple z/OS sysplex partitions. STP over coupling links provides time synchronization to all systems. The appropriate CF link technology (Integrated Coupling Adapter (ICA SR), Coupling Express Long reach (CE LR), 1x InfiniBand (IFB), or 12x IFB) 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 176.

Parallel Sysplex technology is an enabling technology, allowing 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 these ways:

- Continuous (application) availability: Changes can be introduced, such as software upgrades, one image at a time, while the remaining images continue to process work. For more information, see *Parallel Sysplex Application Considerations*, SG24-6523.
- High capacity: 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, and provides techniques for achieving simplified systems management consistent with this requirement.
- Resource sharing: A number of base (z/OS) components use the coupling facility shared storage. This configuration enables sharing of physical resources with significant improvements in cost, performance, and simplified systems management.
- Single system image: The collection of system images in the Parallel Sysplex is displayed as a single entity to the operator, the user, and the database administrator. A single system image ensures reduced complexity from both operational and definition perspectives.
- N-2 support: Multiple hardware generations (normally three) are supported in the same Parallel Sysplex. This configuration provides for a gradual evolution of the systems in the Parallel Sysplex without having to change all of them simultaneously. Similarly, software support for multiple releases or versions is supported.

Through state-of-the-art cluster technology, the power of multiple images can be harnessed to work in concert on common workloads. The 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 running the CFCC can be on z14, z13, z13s, zEC12 and zBC12 systems. For more information about CFCC requirements for supported systems, see , “Coupling facility and CFCC considerations” on page 259.

Consideration: z14, z13, z13s, zEC12, zBC12 servers cannot coexist in the same sysplex with System z196 and previous systems. The introduction of z14 servers into existing installations might require additional planning.

CFCC Level 22

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

- CF Enhancements:

  - Coupling Facility structure encryption
    Coupling Facility Structure encryption is completely transparent to CF-exploiting middleware and applications, while CF exploiters are neither aware of, nor involved in, the encryption. All data and adjunct data flowing 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 is not encrypted, including locks and lock structures.
    z/OS generates the required structure-related encryption keys and does much of the key management automatically via CFRM using secure, protected keys (never clear keys).Secure keys maintained in CFRM couple dataset.
- Coupling Facility Asynchronous Duplexing for Lock Structures
  - New asynchronous duplexing protocol for lock structures.
    - z/OS sends command only to primary CF
    - 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 only for lock structures
    - 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
    - Reduced 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
- Coupling Facility Processor Scalability
  - Coupling Facility work management and dispatcher changes to allow improved efficiency as additional 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 lots of inter-processor contention previously 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, thin interrupt), which are useful when investigating functional performance problems

z14 systems with CFCC Level 22 require z/OS V1R12 with PTFs or later, and z/VM V6R3 or later for guest virtual coupling.

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

z14 servers (CFCC level 22) can coexist in a sysplex with CFCC levels 19, 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 coupling facility LPAR. If the LPAR running the CFCC has only dedicated processors (CPs or ICFs), using all processor capacity (cycles) is not a problem. However, this configuration can be an issue if the LPAR that is running the CFCC also has shared processors. Therefore, enable dynamic dispatching on the CF LPAR.

With CFCC Level 19 and Coupling Thin Interrupts, shared-processor CF can provide more consistent CF service time and acceptable usage in a broader range of configurations. For more information, see 3.9.3, “Dynamic CF dispatching” on page 135.

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

CF structure sizing changes are expected when going from CFCC Level 17 (or earlier) to CFCC Level 20 or later. Review the CF structure size by using the CFSizer tool, which is available at this website:

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

For latest recommended levels, see the current exception letter that is published on Resource Link at the following website:

https://www.ibm.com/servers/resourcelink/lib03020.nsf/pages/exceptionLetters?OpenDocument

3.9.2 Coupling Thin Interrupts

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

- The interrupt causes a shared logical processor coupling facility partition to be dispatched by PR/SM, if it is not already dispatched, allowing the request or signal to be processed in a more timely manner. The coupling facility gives up control when work is exhausted or when PR/SM takes the physical processor away from the logical processor.
- The use of Coupling Thin Interrupts is controlled by the new DYNDISP specification.

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

3.9.3 Dynamic CF dispatching

Dynamic CF dispatching provides the following functions on a coupling facility:

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

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

This capability allows ICF engines to be shared by several CF images. In this environment, it provides faster and far more consistent CF service times. It can also provide performance that is reasonably close to dedicated-engine CF performance if the CF engines are not Coupling Facility Control Code thin interrupts. The introduction of thin interrupts allows a CF to run by using a shared processor while maintaining good performance. The shared engine is allowed to be undispatched when there is no more work, as in the past. The new thin interrupt now gets the shared processor that is dispatched when a command or duplexing signal is presented to the shared engine.

This function saves processor cycles and is an excellent option to be used by a production backup CF or a testing environment CF. This function is activated by the CFCC command **DYNDISP ON**.

The CPs can run z/OS operating system images and CF images. For software charging reasons, generally use only ICF processors to run coupling facility images.

Figure 3-16 shows dynamic CF dispatching.

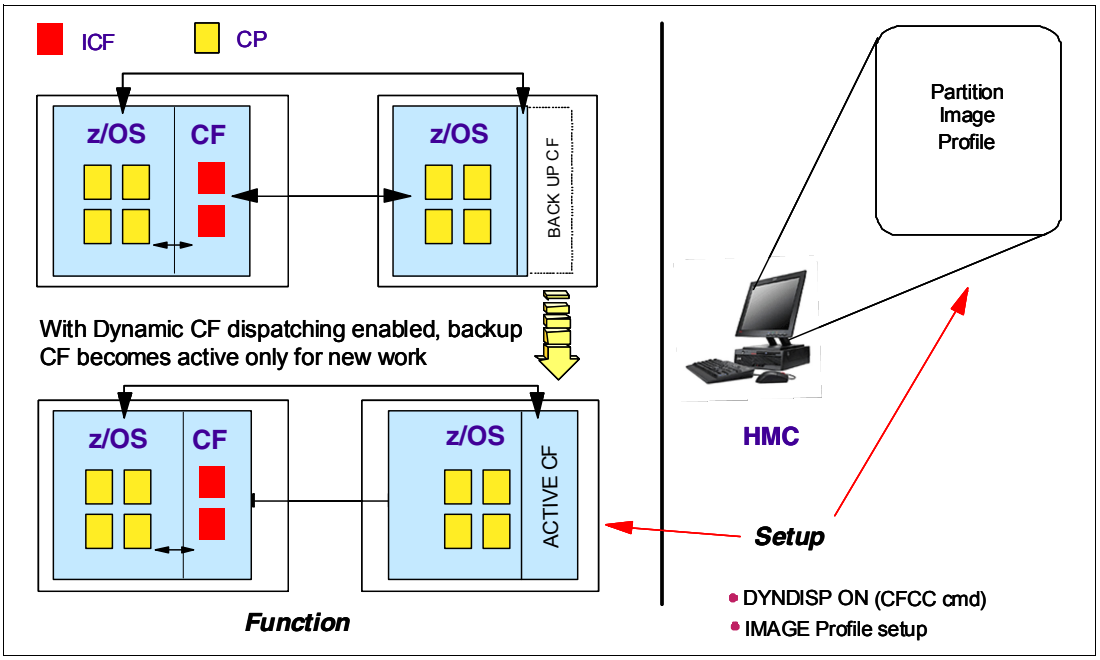


Figure 3-16 Dynamic CF dispatching (shared CPs or shared ICF PUs)

For more information about CF configurations, see *Coupling Facility Configuration Options*, GF22-5042.

3.10 Virtual Flash Memory

Flash Express is not supported on z14. This feature has been replaced by IBM Z Virtual Flash Memory (zVFM), FC #0604.

3.10.1 What is IBM Z Virtual Flash Memory

VFM is an IBM solution to replace existing external zFlash Express feature with support based on main memory.

The “storage class memory” currently provided by Flash Express adapters is replaced with memory allocated from main memory: Virtual Flash Memory (VFM).

VFM is designed to help improve availability and handling of paging workload spikes when running z/OS V2.1, V2.2, or V2.3. 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.

VFM can also be used in coupling facility images to provide extended capacity and availability for workloads making use of Websphere MQ Shared Queues structures. Using 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 Virtual Flash Memory Feature

A VFM feature (FC #0604) is 1.5 TB on z14 (that is approximately same size as a current Flash Express pair of adapters). The maximum number of VFM features is 4 per z14.

Simplification in its management is of great value since there is no hardware adapter to manage. It has no hardware repair and verify. It has a better performance since no “I/O” to attached adapter takes place. And because this is part of memory, it is protected by RAIM and ECC.

3.10.3 Virtual Flash Memory administration

The allocation and definition information of VFM for all partitions is viewed through the “Storage Information” panel under the “Operational Customization” panel.

It will be relocated during CDR in a manner identical to that previously used for “expanded storage”. VFM is much simpler to manage (HMC task) and there is no hardware repair and verify (no cables, no adapters). Also, since this is part of internal memory, VFM is protected by RAIM and ECC and can potentially provide better performance since there since no I/O to an attached adapter takes place.

Notes:

Use cases and exploitation for Flash have not been changed (e.g. z/OS paging, CF shared queue overflow), they just transparently benefit from the changes in the hardware implementation.

There is no option for VFM plan ahead. The only option is to always include zVFM plan ahead when Flexible Memory option is selected.

4

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 servers.

This chapter includes the following sections:

- Introduction to I/O infrastructure
- I/O system overview
- PCIe I/O drawer
- PCIe I/O drawer
- PCIe I/O drawer offerings
- Fanouts
- I/O features (cards)
- Connectivity
- Parallel Sysplex connectivity
- Cryptographic functions
- Integrated firmware processor
- zEDC Express

© Copyright IBM Corp. 2017. All rights reserved.

4.1 Introduction to I/O infrastructure

The IBM z14 servers support PCIe I/O drawers only, no I/O cage or I/O drawer anymore.

4.1.1 I/O infrastructure

IBM extends the use of industry standards on the IBM Z servers 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 consists of the following components:

- PCIe fanouts that support 16 GBps I/O bus interconnection for CPC drawer connectivity to the PCIe I/O drawers
- The 7U, 32-slot, and 4-domain PCIe I/O drawer for PCIe I/O features

The z14 PCIe I/O infrastructure provides these 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 doubles the number of I/O ports compared to an I/O drawer (z13 or earlier only). Up to 64 channels (32 PCIe I/O features) are supported in the PCIe I/O drawer.
- Granularity for the storage area network (SAN) and the local area network (LAN):
  - The FICON Express16S+ has 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 have two ports each (LAN connectivity)
  - 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),
  - 10 GbE Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE) Express2 (new for z14)
  - Coupling Express Long Reach (CE LR) (new for z14 and z13, z13s)
  - zEnterprise Data Compression (zEDC) Express
  - 10 GbE Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE) Express (carry forward only)
  - Crypto Express6s (new for 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 has 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) become 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 one lane (x1), x2, x4, x8, x12, x16, and x32.

The data transmission rate of a PCIe link is determined by the link width (numbers of lanes), the signaling rate of each lane, and the signal encoding rule. The signaling rate of a PCIe Generation 3 lane is 8 gigatransfers/s (GT/s), which means that nearly 8 gigabits are transmitted per second (Gbps).

A PCIe Gen3 x16 link has the following data transmission rates:

- Data transmission rate per lane: 8 Gb/s * 128/130 bit (encoding) = 7.87 Gb/s=984.6 MB/s
- Data transmission rate per link: 984.6 MB/s * 16 (lanes) = 15.75 GB/s

Considering the PCIe link is full-duplex mode, the data throughput rate of a PCIe Gen3 x16 link is 31.5 GBps (15.75 GB/s in both directions).

Link performance: The link speeds do not represent the actual performance of the link. The actual performance depends on many factors that include latency through the adapters, cable lengths, and the type of workload.

PCIe Gen3 x16 links are used in IBM z14 servers for driving the PCIe I/O drawers, and coupling links for CPC to CPC communications.

Note: Unless specified otherwise, when PCIe is mentioned in remaining sections of this chapter, it refers to PCIe Generation 3.

4.2 I/O system overview

This section lists the z14 I/O characteristics and a summary of supported features.

4.2.1 Characteristics

The z14 I/O subsystem is designed to provide great flexibility, high availability, and excellent performance characteristics:

- High bandwidth:

  Link performance: The link speeds do not represent the actual performance of the link. The actual performance depends on many factors that include latency through the adapters, cable lengths, and the type of workload.

  - IBM z14 servers use PCIe as an internal interconnect protocol to drive PCIe I/O drawers and CPC to CPC connections. The I/O bus infrastructure data rate increases up to 160 GBps per drawer (10 PCIe Gen3 Fanout slots). For more information about coupling link connectivity, see 4.7.4, “Parallel Sysplex connectivity” on page 175.
- Connectivity options:
  - IBM z14 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 compare to FICON).
  - For CPC to CPC connections, IBM z14 servers use Integrated Coupling Adapter (ICA SR), Coupling Express (CE LR), or Parallel Sysplex InfiniBand (IFB).
  - The 10GbE RoCE Express2 (and the 10GbE RoCE Express) feature provides 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 IBM z14 servers if unused I/O slot positions are available.
- Concurrent PCIe I/O drawer upgrade:
  - Additional PCIe I/O drawers can be installed concurrently if free frame slots for the PCIe I/O drawers are available.
- Dynamic I/O configuration:
  - Dynamic I/O configuration supports the dynamic addition, removal, or modification of the channel path, control units, and I/O devices without a planned outage.
- Pluggable optics:
  - The FICON Express16S+, FICON Express16S and FICON Express8S, OSA Express6S, OSA Express5S, 10GbE RoCE Express2 and 10GbE RoCE Express features have 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.
  - For zHyperLink Express, it use the cable with MTP connector, and the cable goes to a CXP connection to the adapter. The CXP optics are provided with the adapter.
- Concurrent I/O card maintenance:
  - Every I/O card that is plugged in an I/O drawer or PCIe I/O drawer supports concurrent card replacement during a repair action.

4.2.2 Summary of supported I/O features

The following I/O features are supported:

- Up to 320 FICON Express16S+ channels (up to 160 on M01)
- Up to 320 FICON Express16S channels (up to 160 on M01)
- Up to 320 FICON Express8S channels (up to 160 on M01)
- Up to 96 OSA-Express6S ports
- Up to 96 OSA-Express5S ports
- Up to 16 zEDC Express features
- Up to 8 10GbE RoCE Express2 features
- Up to 8 10GbE RoCE Express features
- Up to 16 zHyperLink Express features
- Up to 40 ICA SR features with up to 80 coupling links
- Up to 32 CE LR features with up to 64 coupling links
- Up to 16 InfiniBand fanouts (features) with one of these options:
  - Up to 32 12x InfiniBand coupling links (HCA3-O fanouts, two ports per feature)
  - Up to 64 1x InfiniBand coupling links (HCA3-O LR fanouts, four ports per feature)

[1] OSA-Express4S, 5S and 6S 1000BASE-T features do not have optics (copper only - RJ45 connectors)

Notes:

- The maximum number of coupling CHPIDs on a IBM z14 server is 256 - a combination of (not all combinations are possible; subject to I/O configuration options):
  - up to 80 ICA SR ports
  - up to 64 CE LR ports
  - up to 32 HCA3-O 12x IFB ports
  - up to 64 HCA3-O LR 1x IFB ports
- IBM Virtual Flash Memory replaces IBM zFlash Express feature on IBM z14 servers.
- 10GbE RoCE Express2 (new) can be mixed with 10 GbE RoCE Express (carry forward only), but maximum number of RoCE and RoCE2 features is 8.

4.3 PCIe I/O drawer

The PCIe I/O drawers (Figure 4-1) are attached to the CPC drawer through a PCIe bus and use PCIe as the infrastructure bus within the drawer. The PCIe I/O bus infrastructure data rate is up to 16 GBps.

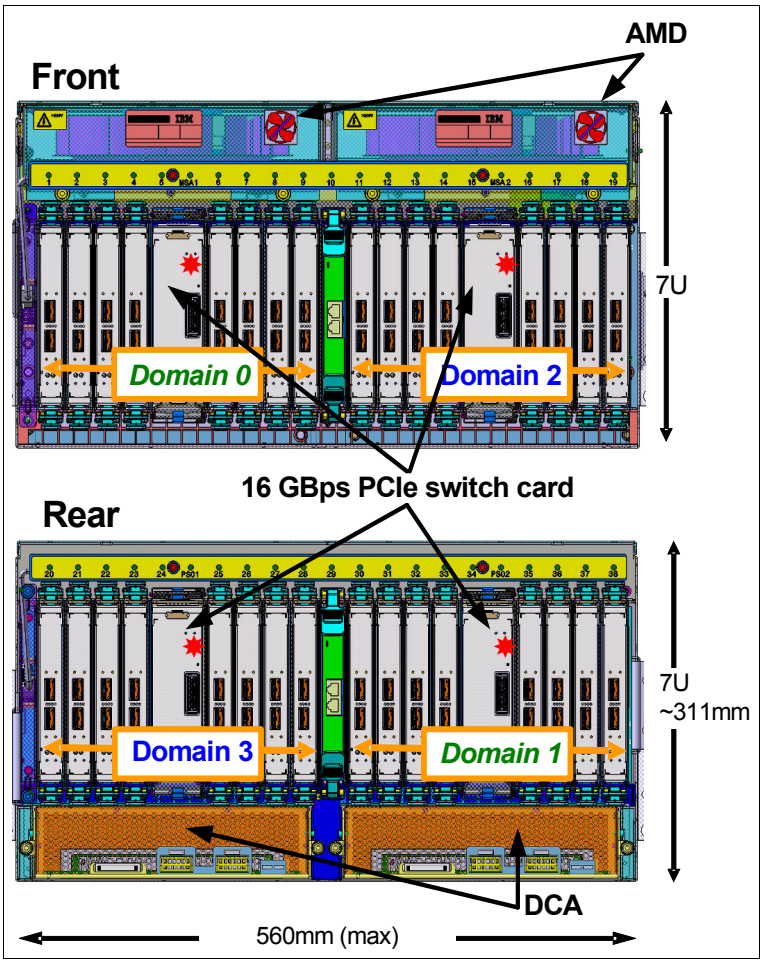


Figure 4-1 PCIe I/O drawer

PCIe switch application-specific integrated circuits (ASICs) are used to fan out the host bus from the processor drawers to the individual I/O features. Maximum 32 PCIe I/O features (up to 64 channels) per drawer are supported.

The PCIe I/O drawer is a two-sided drawer (I/O cards on both sides - front and back) that is 7U high. The drawer contains the 32 I/O slots for PCIe features, four switch cards (two in front, two in the back), two DCAs to provide redundant power, and two air-moving devices (AMDs) for redundant cooling (Figure 4-1 on page 143).

The PCIe I/O drawer slots numbers are shown in Figure 4-2.

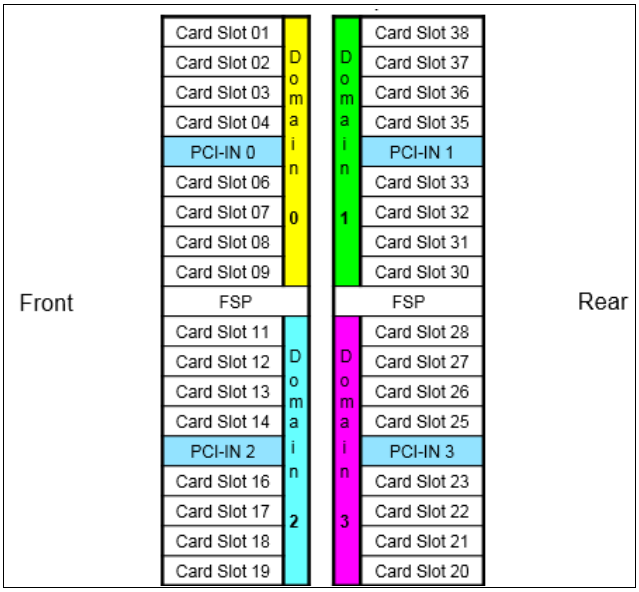


Figure 4-2 PCIe I/O drawer slots numbers

The I/O structure in a z14 CPC is shown in Figure 4-3. The PCIe switch card provides the fanout from the high-speed x16 PCIe host bus to eight individual card slots. The PCIe switch card is connected to the drawers through a single x16 PCIe 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 enable you to increase I/O capability as needed without having to plan ahead.

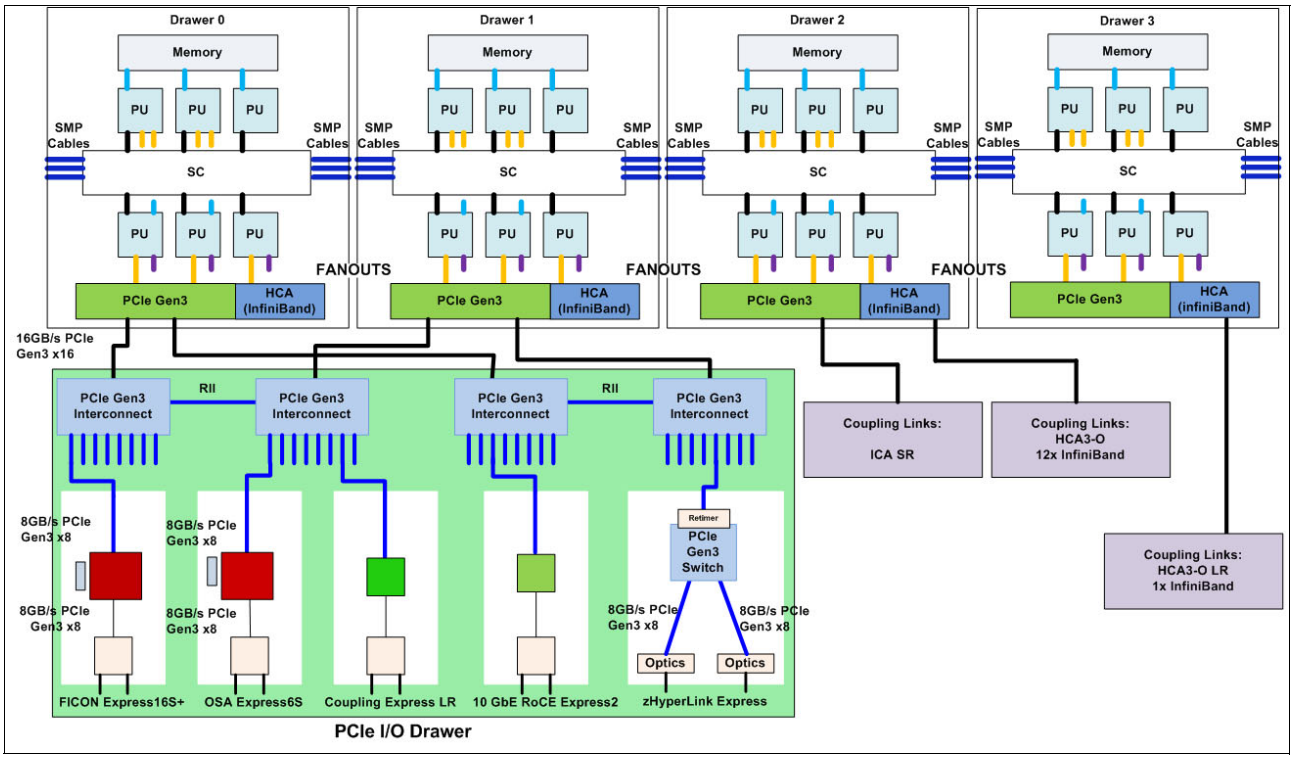


Figure 4-3 z14 I/O connectivity

The PCIe I/O slots are organized into four hardware 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.

A switch card in the front is connected to a switch card in the rear through the PCIe I/O drawer board (through the Redundant I/O Interconnect - RII). In addition, switch cards in same PCIe I/O drawer are connected to PCIe fanouts across nodes and CPC drawers for higher availability.

The RII design provides a failover capability during a PCIe fanout card failure or CPC drawer upgrade. All four domains in one of these PCIe I/O drawers can be activated with four fanouts. The flexible service processors (FSPs) are used for system control.

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

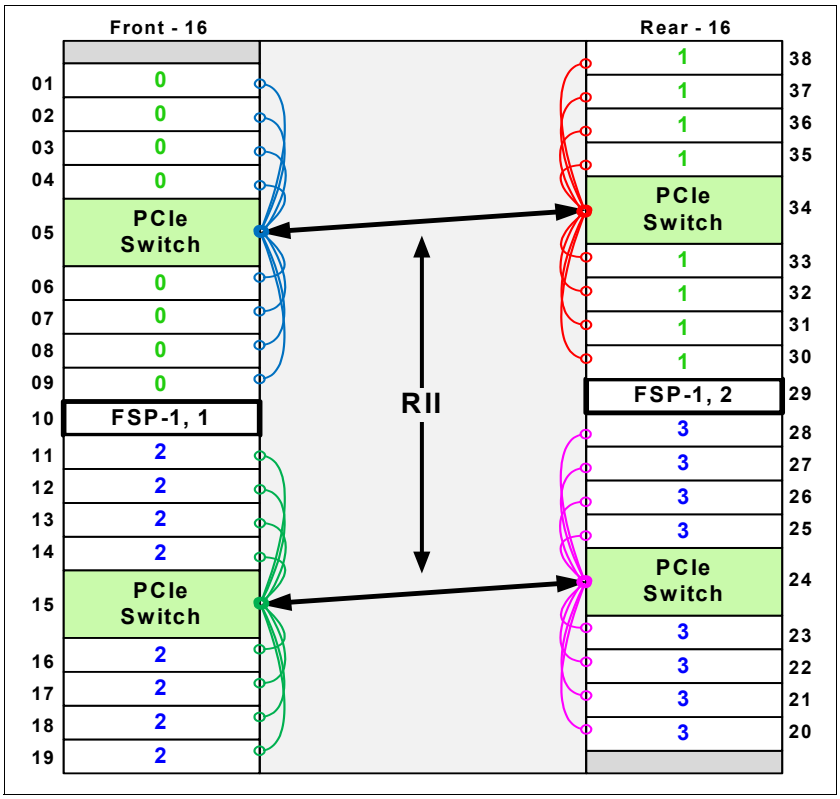


Figure 4-4 PCIe I/O drawer with 32 PCIe slots and four I/O domains

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

Table 4-1 I/O domains of PCIe I/O drawer

Domain	I/O slot in domain
0	01, 02, 03, 04, 06, 07, 08, and 09
1	30, 31, 32, 33, 35, 36, 37, and 38
2	11, 12, 13, 14, 16, 17, 18, and 19
3	20, 21, 22, 23, 25, 26, 27, and 28

4.4 PCIe I/O drawer offerings

A maximum of five PCIe I/O drawers can be installed that support up to 160 PCIe I/O features.

For an upgrade to IBM z14 servers, only the following PCIe I/O features carried forward:

- FICON Express16S
- FICON Express8S
- OSA-Express5S (all 5S features)
- OSA-Express4S 1000BaseT (only)
- 10GbE RoCE Express

- Crypto Express5S
- zEDC Express
- Coupling Express Long Reach (CE LR)

Consideration: On a *new build* IBM z14 server, only PCIe I/O drawers are supported. No carry-forward of I/O drawers or associated features are supported on upgrades to a IBM z14 server.

A new build IBM z14 server supports the following PCIe I/O feature hosted in the PCIe I/O drawers:

- FICON Express16S+
- OSA-Express6S
- 10GbE RoCE Express2
- Crypto Express6S
- zEDC Express
- Coupling Express Long Reach (CE LR)
- zHyperLink Express

Note: Model-upgrades allowed to IBM z14 from z13 or zEC12, downgrades are not allowed from IBM z14. Capacity upgrades or downgrades are allowed in an upgrade to IBM z14 from z13 or zEC12.

For frame roll MES from zEC12 and z13 to IBM z14, new frames are shipped, new PCIe I/O drawers will be supplied with the MES for zEC12 to replace the I/O drawers.

4.5 Fanouts

The z14 server uses fanout cards to connect the I/O subsystem to the CPC drawer. The fanout cards also provide the ICA SR and InfiniBand coupling links for Parallel Sysplex. All fanout cards support concurrent add, delete, and move.

The internal z14 I/O infrastructure consists of:

- The PCIe Generation3 fanout cards: one port card (feature) and connects to a PCIe I/O drawer supporting an eight-slot I/O domain. This card is always installed in pairs to support I/O domain redundant connectivity.
- InfiniBand HCA3-C fanout card, HCA2-O fanout card is NOT supported on IBM z14 servers.

IBM z14 will be the last z Systems and IBM Z server to support HCA3-O and HCA3-O LR adapters[a]: z14 will be last z Systems and IBM Z server to support HCA3-O fanout for 12x IFB (#0171) and HCA3-O LR fanout for 1x IFB (#0170).

a. IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM's sole discretion.

The PCIe and InfiniBand fanouts are installed in the front of each CPC drawer. Each CPC drawer has 10 PCIe Gen3 fanout slots and four InfiniBand fanout slots. The PCIe fanout slots are named LG03 - LG12, left to right. The InfiniBand fanout slots are in the right of the CPC drawer and are named LG13 - LG16, left to right. Slots LG01 and LG02 are used for Flexible Support Processor (FSP).

Five types of fanout cards are supported by IBM z14 servers. Each slot can hold one of the following five fanouts:

- PCIe Gen3 fanout card: This copper fanout provides connectivity to the PCIe switch card in the PCIe I/O drawer.
- Integrated Coupling Adapter (ICA SR): This adapter provides coupling connectivity between z14, z13 and z13s servers, up to 150-meter (492 ft) distance, 8 GB/s link rate.
- Host Channel Adapter (HCA3-O (12xIFB)): This optical fanout provides 12x InfiniBand coupling link connectivity up to 150-meter (492 ft) distance to a IBM z14, z13, z13s, zEC12, zBC12 servers.
- Host Channel Adapter (HCA3-O LR (1xIFB)): This optical long range fanout provides 1x InfiniBand coupling link connectivity to IBM z14, z13, z13s, zEC12, zBC12 servers. HCA3-O LR supports up to 10 km (6.2 miles) unrepeated distance or 100 km (62 miles) when using IBM Z qualified dense wavelength division multiplexing (DWDM) equipment.

The PCIe Gen3 fanout card comes with one port, the HCA3-O LR (1xIFB) fanout comes with four ports, and other fanouts come with two ports.

Note:

- HCA2-O fanout card carry-forward is no longer supported on IBM z14 servers.

Figure 4-7 on page 179 illustrates the following PCIe and IFB connections from the CPC drawer:

- A PCIe I/O drawer (PCIe Gen3)
- Another z server connected through InfiniBand (either 12x or 1x HCA3-O)
- Another z server connected through a dedicated PCIe ICA SR

Figure 4-3 on page 145 shows an I/O connection scheme that is not tied to a particular CPC drawer. In a real configuration I/O connectivity is mixed across multiple CPC drawers, if available, for I/O connection redundancy.

4.5.1 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 ft) to 4.0 meters (13.1 ft) is used for connection to the PCIe switch card in the PCIe I/O drawer. PCIe fanout cards are always plugged in pairs and provide redundancy for I/O domains within the PCIe I/O drawer.

The pairs of PCIe fanout cards of a z14 are named as LG03 to LG12 from left to right. All z14 models (except for model M01) split the PCIe fanout pairs across different processor drawers for redundancy purposes.

PCIe fanout: The PCIe fanout is used exclusively for I/O and cannot be shared for any other purpose.

4.5.2 Integrated Coupling Adapter (FC #0172)

The IBM ICA SR, introduced with IBM z13, 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 150m 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 and z13/z13s CPCs.

The ICA SR fanout is housed in the PCIe I/O fanout slot on the z14 CPC drawer, which supports 10 PCIe I/O slots. Up to 10 ICA SR fanouts and up to 20 ICA SR ports are supported on a z14 CPC drawer, enabling greater connectivity for short distance coupling on a single processor node compared to previous generations. The maximum number of ICA SR fanout features is 20 per system on IBM z14 servers.

The ICA SR can be used for coupling connectivity between z14 and z13/z13s servers. It does not support connectivity to zEC12, zBC12 servers, and it cannot be connected to HCA3-O or HCA3-O LR coupling fanouts.

The ICA SR fanout requires cabling that different from the 12x IFB cables. For distances up to 100 m, OM3 fiber optic, while for distances up to 150 m, OM4 fiber optic cables can be used. For more information, see *Planning for Fiber Optic Links*, GA23-1407, and *IBM z14 Installation Manual for Physical Planning,* GC28-6965.

4.5.3 HCA3-O (12x IFB) fanout (FC #0171)

The HCA3-O fanout for 12x InfiniBand provides an optical interface that is used for coupling links. The two ports on the fanout are dedicated to coupling links that connect to z14, z13, z13s CPCs. Up to 16 HCA3-O (12x IFB) fanouts are supported, and provide up to 32 ports for coupling links.

The fiber optic cables are industry standard OM3 (2000 MHz-km) 50-µm multimode optical cables with multifiber push-on (MPO) connectors. The maximum cable length is 150 meters (492 ft). Each port (link) has 12 pairs of fibers: 12 fibers for transmitting, and 12 fibers for receiving. The HCA3-O (12xIFB) fanout supports a link data rate of 6 GBps.

Important: The HCA3-O fanout has two ports (1 and 2). Each port has one connector for transmitting (TX) and one connector for receiving (RX). Ensure that you use the correct cables. An example is shown in Figure 4-5.

For more information, see *Planning for Fiber Optic Links*, GA23-1407, and *IBM z14 Installation Manual for Physical Planning,* GC28-6965.

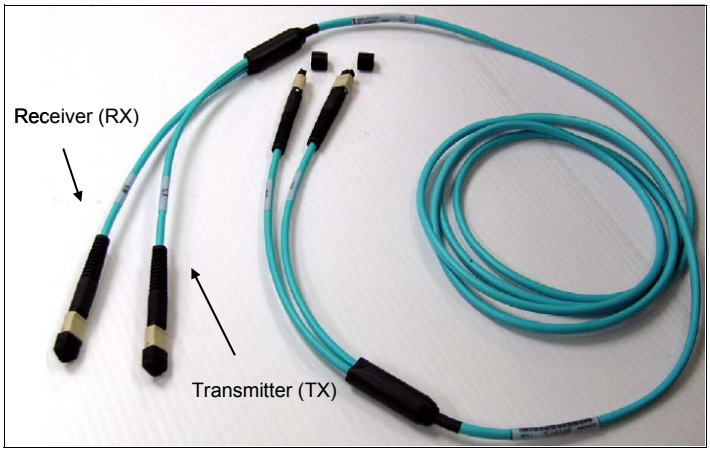


Figure 4-5 OM3 50/125 µm multimode fiber cable with MPO connectors

A fanout has two ports for optical link connections, and supports up to 16 CHPIDs across both ports. These CHPIDs are defined as channel type CIB in the input/output configuration data set (IOCDS). The coupling links can be defined as shared between images within a channel subsystem (CSS). They also can be spanned across multiple CSSs in a CPC.

Each HCA3-O (12x IFB) fanout has an assigned Adapter ID (AID) number. This number must be used for definitions in IOCDS to create a relationship between the physical fanout location and the CHPID number. For more information about AID numbering, see “Adapter ID number assignment” on page 151.

For more information about how the AID is used and referenced in the HCD, see *Implementing and Managing InfiniBand Coupling Links on System z* SG24-7539.

When STP is enabled, IFB coupling links can be defined as timing-only links to other z14, z13, z13s, zEC12 and zBC12 CPCs.

12x IFB and 12x IFB3 protocols

These protocols are supported by the HCA3-O for 12x IFB feature:

- 12x IFB3 protocol: When HCA3-O (12xIFB) fanouts are communicating with HCA3-O (12x IFB) fanouts and are defined with four or fewer CHPIDs per port, the 12x IFB3 protocol is used.
- 12x IFB protocol: If more than four CHPIDs are defined per HCA3-O (12xIFB) port, or HCA3-O (12x IFB) features are communicating with HCA2-O (12x IFB) features on zEC12 and zBC12 CPCs, links run with the 12x IFB protocol.

The HCA3-O feature that supports 12x InfiniBand coupling links is designed to deliver improved service times. When no more than four CHPIDs are defined per HCA3-O (12xIFB) port, the 12x IFB3 protocol is used. When you use the 12x IFB3 protocol, synchronous service times are up to 40% faster than when you use the 12x IFB protocol.

4.5.4 HCA3-O LR (1x IFB) fanout (FC #0170)

The HCA3-O LR fanout for 1x InfiniBand provides an optical interface that is used for coupling links. The four ports on the fanout are dedicated to coupling links to connect to z14, z13, z13s servers. Up to 16 HCA3-O LR (1xIFB) fanouts are supported on IBM z14 servers (up to 64 1x IFB3 ports for coupling).

The HCA3-O LR fanout supports InfiniBand 1x optical links that offer long-distance coupling links. The cable has one lane that contains two fibers. One fiber is used for transmitting, and the other fiber is used for receiving data.

Each connection supports a link rate of up to 5 Gbps if connected to a z14, z13, z13s server. HCA3-O LR supports also a link rate of 2.5 Gbps when connected to IBM Z qualified DWDM equipment. The link rate is auto-negotiated to the highest common rate.

The fiber optic cables are 9-µm SM optical cables that terminate with an LC Duplex connector. With direct connection, the supported unrepeated distance[2] is up to 10 km (6.2 miles), and up to 100 km (62 miles) with IBM Z qualified DWDM equipment.

A fanout has four ports for optical link connections, and supports up to 16 CHPIDs across all four ports. These CHPIDs are defined as channel type CIB in the IOCDS. The coupling links can be defined as shared between images within a channel subsystem, and also can be spanned across multiple channel subsystems in a server.

Each HCA3-O LR (1xIFB) fanout can be used for link definitions to another server, or a link from one port to a port in another fanout on the same server.

The source and target operating system image, CF image, and the CHPIDs that are used on both ports in both servers are defined in IOCDS.

Each HCA3-O LR (1xIFB) fanout has an assigned AID number. This number must be used for definitions in IOCDS to create a relationship between the physical fanout location and the CHPID number. For more information about AID numbering, see “Adapter ID number assignment” on page 151.

When STP is enabled, HCA3-O LR coupling links can be defined as timing-only links to other z14, z13, z13s, zEC12, and zBC12 CPCs.

4.5.5 Fanout considerations

Fanout slots in each CPC drawer can be used to plug different fanouts. One drawer can hold up to 10 PCIe fanouts and four InfiniBand fanout cards.

Adapter ID number assignment

PCIe and IFB fanouts and ports are identified by an AID that is initially dependent on their physical locations, which is unlike channels that are installed in a PCIe I/O drawer or I/O drawer. Those channels are identified by a physical channel ID (PCHID) number that is related to their physical location. This AID must be used to assign a CHPID to the fanout in the IOCDS definition. The CHPID assignment is done by associating the CHPID to an AID port.

Table 4-2 AID number assignment

Drawer	Location	Fanout slot	AIDs
First	A15A	LG03-LG12 (PCIe)	2E-37
		LG13-LG16 (IFB)	0C-0F
Second	A19A	LG03-LG12 (PCIe)	24-2D
		LG13-LG16 (IFB)	08-0B

[2] On special request, see http://www.ibm.com/systems/z/advantages/pso/cflinks.html

Drawer	Location	Fanout slot	AIDs
Third	A23A	LG03-LG12 (PCIe)	1A-23
		LG13-LG16 (IFB)	04-07
Fourth	A27A	LG03-LG12 (PCIe)	10-19
		LG13-LG16 (IFB)	00-03

Fanout slots

The fanout slots are numbered LG03 - LG16 left to right, as shown in Table 4-2 on page 151. All fanout locations and their AIDs for all four drawers are shown for reference only. Slots LG01 and LG02 never have a fanout installed because they are dedicated for FSPs.

Important: The AID numbers in Table 4-2 are valid only for a new build system or if new processor drawers are added. If a fanout is moved, the AID follows the fanout to its new physical location.

The AID assignment is listed in the PCHID REPORT that is provided for each new server or for an MES upgrade on existing servers. Example 4-1 shows part of a PCHID REPORT for a model M03. In this example, one fanout card is installed in the first drawer (location A15A slot LG14), then it is assigned as AID 0D. Another fanout card is installed in the drawer (location A19A slot LG14), it is assigned as AID 09.

Example 4-1 AID assignment in PCHID REPORT

```
CHPIDSTART
  19567745                      PCHID REPORT                      Jul 14,2017
  Machine: 3906-M03  00000NEW1
  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
  Source           Drwr  Slot  F/C    PCHID/Ports or AID             Comment
  A23/LG14         A23A  LG14  0170   AID=05

  A23/LG16         A23A  LG16  0171   AID=07

  A19/LG14         A19A  LG14  0171   AID=09

  A19/LG15         A19A  LG15  0171   AID=0A

  A19/LG16         A19A  LG16  0170   AID=0B

  A15/LG13         A15A  LG13  0171   AID=0C

  A15/LG14         A15A  LG14  0170   AID=0D

  A15/LG16         A15A  LG16  0170   AID=0F

  A23/LG04         A23A  LG04  0172   AID=1B

  A23/LG11         A23A  LG11  0172   AID=22

  A19/LG04         A19A  LG04  0172   AID=25

  A15/LG11         A15A  LG11  0172   AID=36
```

Fanout features that are supported by the z14 server are shown in Table 4-3, which provides 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
HCA3-O (12xIFB)	0171	Coupling link	50-µm MM OM3 (2000 MHz-km)	MPO	150 m (492 ft)	6 GBps[b]
HCA3-O LR (1xIFB)	0170	Coupling link	9-µm SM	LC Duplex	10 km (6.2 miles) 100 km[c] (62 miles)	5.0 Gbps
ICA SR	0172	Coupling link	OM4	MTP	150 m (492 ft)	8 Gbps
			OM3	MTP	100 m (328 ft)	8 Gbps

a. The link data rates do not represent the actual performance of the link. The actual performance depends on many factors that include latency through the adapters, cable lengths, and the type of workload.
b. When using the 12x IFB3 protocol, synchronous service times are 40% faster than the 12x IFB protocol.
c. Up to 100 km (62 miles) with repeaters (IBM Z qualified DWDM).

4.6 I/O features (cards)

I/O features (cards) have ports[3] to connect the z14 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.

4.6.1 I/O feature card ordering information

Table 4-4 lists the I/O features that are supported by z14 servers and the ordering information for them.

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-Express6S 10 GbE LR	0424	Y	N/A

[3] Certain I/O features do not have external ports, such as Crypto Express and zEDC

Channel feature	Feature code	New build	Carry-forward
OSA-Express6S 10 GbE 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 10 GbE LR	0415	N	Y
OSA-Express5S 10 GbE 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 1000BASE-T Ethernet	0408	N	Y
Integrated Coupling Adapter (ICA SR)	0172	Y	Y
Coupling Express LR	0433	Y	Y
HCA3-O (12xIFB)	0171	Y	Y
HCA3-O LR (1xIFB)	0170	Y	Y
Crypto Express6S	0893	Y	N/A
Crypto Express5S	0890	N	Y
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: IBM z14 servers do not support the ISC-3, HCA2-O (12x), or HCA2-O LR (1x) features and cannot participate in a Mixed Coordinated Timing Network (CTN).

4.6.2 Physical channel (PCHID) report

A PCHID reflects the physical location of a channel-type interface. A PCHID number is based on these factors:

- The I/O drawer, and PCIe I/O drawer location
- The channel feature slot number
- The port number of the channel feature

A CHPID does not directly correspond to a hardware channel port, but it is assigned to a PCHID in the hardware configuration definition (HCD) or IOCP.

A PCHID REPORT is created for each new build server and for upgrades on existing servers. The report lists all I/O features that are installed, the physical slot location, and the assigned PCHID. Example 4-2 shows a portion of a sample PCHID REPORT. For more information about the AID numbering rules for InfiniBand coupling links, see Table 4-2 on page 151.

Example 4-2 PCHID REPORT

```
CHPIDSTART
  12519541                      PCHID REPORT                      May 05,2017
  Machine: 3906-M05  NEW1
  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
  Source          Drwr  Slot  F/C    PCHID/Ports or AID             Comment
  A23/LG16        A23A  LG16  0171   AID=07
  A19/LG14        A19A  LG14  0171   AID=09
  A15/LG14        A15A  LG14  0170   AID=0D
  A27/LG05        A27A  LG05  0172   AID=12
  A15/LG09        A15A  LG09  0172   AID=34
  A19/LG07/J01    Z22B  01    0428   100/D1 101/D2
  A19/LG07/J01    Z22B  02    0420   104                    RG1
  A19/LG07/J01    Z22B  03    0412   108/D1D2               RG1
  A19/LG07/J01    Z22B  04    0431   10C/D1D2               RG1
  A19/LG07/J01    Z22B  06    0433   110/D1D2               RG1
  A19/LG07/J01    Z22B  07    0893   114/P00
  A19/LG07/J01    Z22B  08    0424   118/D1
  A19/LG07/J01    Z22B  09    0425   11C/D1
  A27/LG07/J01    Z22B  11    0424   120/D1
  A27/LG07/J01    Z22B  12    0424   124/D1
  A27/LG07/J01    Z22B  13    0427   128/D1 129/D2
  A27/LG07/J01    Z22B  14    0428   12C/D1 12D/D2
  A27/LG07/J01    Z22B  16    0433   130/D1D2               RG3
  A27/LG07/J01    Z22B  17    0431   134/D1D2               RG3
  A27/LG07/J01    Z22B  18    0412   138/D1D2               RG3
  A27/LG07/J01    Z22B  19    0420   13C                    RG3
  A15/LG03/J01    Z22B  20    0420   140                    RG4
  A15/LG03/J01    Z22B  21    0412   144/D1D2               RG4
  A15/LG03/J01    Z22B  22    0431   148/D1D2               RG4
  A15/LG03/J01    Z22B  23    0433   14C/D1D2               RG4
  A15/LG03/J01    Z22B  25    0428   150/D1 151/D2
  A15/LG03/J01    Z22B  26    0427   154/D1 155/D2
  A15/LG03/J01    Z22B  27    0893   158/P00
  A15/LG03/J01    Z22B  28    0425   15C/D1
  A23/LG03/J01    Z22B  30    0425   160/D1
  A23/LG03/J01    Z22B  31    0424   164/D1
  A23/LG03/J01    Z22B  32    0427   168/D1 169/D2
  A23/LG03/J01    Z22B  33    0428   16C/D1 16D/D2
  A23/LG03/J01    Z22B  35    0433   170/D1D2               RG2
  A23/LG03/J01    Z22B  36    0433   174/D1D2               RG2
  A23/LG03/J01    Z22B  37    0412   178/D1D2               RG2
  A23/LG03/J01    Z22B  38    0412   17C/D1D2               RG2
  Legend:
  Source  Book Slot/Fanout Slot/Jack
  A15A    CEC Drawer 1 in A frame
  A27A    CEC Drawer 4 in A frame
  A23A    CEC Drawer 3 in A frame
  A19A    CEC Drawer 2 in A frame
  Z22B    PCIe Drawer 1 in Z frame
  Z15B    PCIe Drawer 2 in Z frame
  Z08B    PCIe Drawer 3 in Z frame
  0428    16GB FICON Express16S+ SX 2 Ports
  RG1     Resource Group 1
  0420    zEDC Express
```

```
0412    10GbE  RoCE Express
0431    zHyperLink Express
0433    Coupling Express LR
0893    Crypto Express6S
0424    OSA Express6S 10 GbE LR 1 Ports
0425    OSA Express6S 10 GbE SR 1 Ports
0427    16GB FICON Express16S+ LX 2 Ports
RG3     Resource Group 3
RG2     Resource Group 2
RG4     Resource Group 4
0171    HCA3 O PSIFB 12x 2 Links
0170    HCA3 O LR PSIFB 1x 4 Links
0172    ICA SR 2 Links
```

The following list explains the content of the sample PCHID REPORT:

- Feature code 0170 (HCA3-O LR (1xIFB)) is installed in CPC drawer 1 (location A15A, slot LG14), and has AID 0D assigned.
- Feature code 0172 (Integrated Coupling Adapter (ICA SR) is installed in CPC drawer 4 (location A27A, slot LG05), and has AID 12 assigned.
- Feature code 0424 (OSA-Express6S 10 GbE LR) is installed in PCIe I/O drawer 1 (location Z22B, slot 11), and has PCHID 120 assigned.
- Feature code 0427 (FICON Express16S+ long wavelength (LX) 10 km (6.2 miles)) is installed in PCIe I/O drawer 2 (location Z22B, slot 26), and has PCHIDs 154 and 155 assigned.
- Feature code 0431 (zHyperLink Express) is installed in PCIe I/O drawer 2 (location Z22B, slot 04), and has PCHID10C assigned. PCHID 240 is shared by ports D1 and D2.

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, RG4).

For more information about resource groups, see Appendix C, “Native Peripheral Component Interconnect Express (PCIe)” on page 451.

The preassigned PCHID number of each I/O port relates directly to its physical location (jack location in a specific slot).

4.7 Connectivity

I/O channels are part of the CSS. They provide connectivity for data exchange between servers, or between servers and external control units (CUs) and devices, or between networks.

- Connectivity to external I/O subsystems, disks for example, this part will be introduce as “Storage connectivity” on page 159.
- Communication to LANs will be introduced as “Network Connectivity” on page 166.
- Communication between servers is implemented by using CE LR, ICA SR, or coupling over InfiniBand, or channel-to-channel (CTC) connections. This part will be introduced as “Parallel Sysplex connectivity” on page 175.

4.7.1 I/O feature support and configuration rules

Table 4-5 lists the I/O features that are supported. The table shows the number of ports per card, port increments, the maximum number of feature cards, and the maximum number of channels for each feature type. Also, the CHPID definitions that are used in the IOCDS are listed.

Table 4-5 z14 supported I/O features

I/O feature	Number of		Maximum number of		PCHID	CHPID definition
	Ports per card	Port increments	Ports	I/O slots		
FICON Express16S+ LX/SX	2	2	320	160	Yes	FC, FCP[a]
FICON Express16S LX/SX	2	2	320	160	Yes	FC, FCP
FICON Express8S LX/SX	2	2	320	160	Yes	FC, FCP
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 1000BASE-T	2	2	96	48	Yes	OSC, OSD, OSE, OSM
10GbE RoCE Express2	2	2	16	8	Yes	N/A[b]
10GbE RoCE Express	2	2	16	8	Yes	N/A[b]
Coupling Express LR	2	2	64	32	Yes	CL5
Integrated Coupling Adapter (ICA SR)	2	2	40	20	N/A	CS5
HCA3-O for 12x IFB and 12x IFB3	2	2	32	16	N/A	CIB
HCA3-O LR for 1x IFB	4	4	64	16	N/A	CIB
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 using Function IDs (FIDs)

At least one I/O feature (FICON) or one coupling link feature (ICA SR or HCA3-O) must be present in the minimum configuration.

The following features can be shared and spanned:

- FICON channels that are defined as FC or FCP
- OSA-Express features that are defined as OSC, OSD, OSE, OSM, or OSX
- Coupling links that are defined as CS5, CIB, 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 have a CHPID type. However, LPARs in all CSSs have access to the features. Each Crypto Express adapter can be defined to up to 85 LPARs.
- Each RoCE Express/Express2 feature occupies one I/O slot but does not have a CHPID type. However, LPARs in all CSSs have access to the feature. The 10GbE RoCE Express can be defined to up to 31 LPARs per PCHID, the 10GbE RoCE Express2 supports up to 126 LPARs per PCHID.
- Each zEDC Express feature occupies one I/O slot but does not have a CHPID type. However, LPARs in all CSSs have access to the feature. The zEDC feature can be defined to up to 15 LPARs.
- Each zHyperLink Express feature occupies one I/O slot but does not have a CHPID type. However, LPARs in all CSSs have access to the feature. The zHyperLink Express adapter works as native PCIe adapter, 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 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, a SAN, a LAN, or the end-to-end enterprise.

Cables: All fiber optic cables, cable planning, labeling, and installation are client responsibilities for new z14 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 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 having to extend the maintenance window.

Table 4-6 lists the required connector and cable type for each I/O feature on IBM z14 servers.

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
0424	OSA-Express6S 10 GbE 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
0408	OSA-Express4S 1000BASE-T	RJ-45	Category 5 UTP
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 (4.7 GHz-km)[b]
0171	HCA3-O (12xIFB)	MPO	50 µm MM OM3 (2 GHz-km)
0170	HCA3-O LR (1xIFB)	LC Duplex	9 µm SM
0431	zHyperLink Express	MPO	50 µm MM OM4 (4.7 GHz-km)

a. UTP is unshielded twisted pair. Consider using category 6 UTP for 1000 Mbps connections.
b. Or 50 µm MM OM3 (2 GHz-km), but OM4 is highly recommended.
MM = Multi-Mode
SM = Single-Mode

4.7.2 Storage connectivity

Connectivity to external I/O subsystems, disks for example, is provided by FICON channels and zHyperLink[4].

[4] zHyperLink feature operates together with a FICON channel

FICON channels

z14 supports the following FICON features:

- FICON Express16S+
- FICON Express16S
- 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)

They 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 has two ports, each supporting an LC Duplex connector, with one PCHID and one CHPID 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.

IBM z14 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:

- CHPID type FC: FICON, zHPF, and FCTC. All of these protocols are supported simultaneously.
- CHPID type FCP: Fibre Channel Protocol that supports attachment to SCSI devices directly or through Fibre Channel switches or directors.

FICON channels (CHPID type FC or FCP) can be shared among LPARs and can be defined as spanned. All ports on a FICON feature must be of the same type, either LX or SX. The features are connected to a FICON capable control unit, either point-to-point or switched point-to-point, through a Fibre Channel switch.

FICON 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 two types of FICON Express16S+ optical transceivers that are supported are the LX and the SX (no mix on same card):

- FICON Express16S+ 10 km 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+ 10 km LX feature uses an optical transceiver that supports an unrepeated distance of 10 km (6.2 miles) by using 9 µm single-mode fiber.

Each port of the FICON Express16S SX feature uses an optical transceiver that supports to up to 125 m (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 two types of FICON Express16S optical transceivers that are supported are the LX and the SX:

- FICON Express16S 10 km LX feature, FC #0418, with two ports per feature, supporting LC Duplex connectors
- FICON Express16S SX feature, FC #0419, with two ports per feature, supporting LC Duplex connectors

Each port of the FICON Express16S 10 km LX feature uses an optical transceiver that supports an unrepeated distance of 10 km (6.2 miles) by using 9 µm single-mode fiber.

Each port of the FICON Express16S SX feature uses an optical transceiver that supports to up to 125 m (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 two types of FICON Express8S optical transceivers that are supported are the LX and the SX:

- FICON Express8S 10 km LX feature, FC #0409, with two ports per feature, supporting LC Duplex connectors
- FICON Express8S SX feature, FC #0410, with two ports per feature, supporting LC Duplex connectors

Each port of the FICON Express8S 10 km LX feature uses an optical transceiver that supports an unrepeated distance of 10 km (6.2 miles) by using 9 µm single-mode fiber.

Each port of the FICON Express8S SX feature uses an optical transceiver that supports up to 150 m (492 feet) of distance depending on the fiber used.

FICON enhancements

Together with the FICON Express16S+, IBM z14 servers provide enhancements for FICON in both functional and performance aspects.

Forward Error Correction

Forward Error Correction (FEC) is a technique used for reducing data errors when transmitting over unreliable or noisy communication channels (improving signal to noise ratio). By adding redundancy error-correction code (ECC) to the transmitted information, the receiver can detect and correct a number of errors without requiring retransmission. This process has improved signal reliability and bandwidth utilization by reducing retransmissions due to bit errors, especially for connections across long distance, like an inter-switch link (ISL) in a GDPS Metro Mirror environment.

The FICON Express16S+ and FICON Express16S are designed to support FEC coding on top of its 64b/66b data encoding for 16Gbps connections. This design is capable of correcting up to 11 bit errors per 2112 bits transmitted. Thus, 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. They do this while retaining the same reliability and robustness that FICON channels have traditionally been known for.

With the IBM DS8870 or above, IBM z14 (as well as z13/z13s) servers are capable of extending 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, 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 provided by the FICON director manufacturers, for example, Brocade's exchange-based routing (EBR) and Cisco's originator exchange ID (OxID)[5] 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 based on first come first served, when a port starts a fabric login (FLOGI) to a destination domain. The ISL is round robin selected for assignment. Thus, I/O flow from same incoming port to same destination domain will always be assigned the same ISL route, regardless of the destination port of each I/O. This setup can result in some ISLs overloaded while some are underutilized. The ISL routing table is changed every time a Z server undergoes a power-on-reset (POR), so the ISL assignment is somewhat unpredictable.

Device-based routing (DBR) assigns the ISL routes statically based on a hash of the source and destination port. That I/O flow from same incoming port to same destination is assigned

[5] Check with the switch provider for their support statement.

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 using a static SAN routing policy, the FICON director has limited capability to assign ISL routes based on workload, which may result in unbalanced use of ISLs (some may be overloaded, while others are underutilized).

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, IBM z14 servers have advantages for performance and management in configurations with ISL and cascaded FICON directors:

- Support sharing of ISLs between FICON and FCP (PPRC or distributed)
- I/O traffic is better balanced between all available ISLs
- Improve utilization of FICON director and ISL
- Easier to manage with a predicable and repeatable I/O performance

FICON dynamic routing can be enabled by definition of dynamic routing capable switches and control units in HCD. Also, z/OS has implemented a health check function for FICON dynamic routing.

Improved zHPF I/O execution at distance

By introducing the concept of pre-deposit writes, zHPF reduces the number of round trips of standard FCP I/Os to a single round trip. Originally this benefit is limited to writes that are less than 64 KB. zHPF on IBM z14, z13s and z13 servers has been enhanced to allow all large write operations (> 64 KB) at distances up to 100 km to be run in a single round trip to the control unit. This 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) has been added to the Fibre Channel T11 standard to allow Z servers to obtain extra diagnostic data from the SFP optics located throughout the SAN fabric.

IBM z14 and z13 servers now can read this additional diagnostic data for all the ports accessed in the I/O configuration and make the data available to an LPAR. For z/OS LPARs using FICON channels, z/OS displays the data with a new message and display command. For Linux on z Systems, z/VM, z/VSE, and KVM LPARs using FCP channels, this diagnostic data is available in a new window in the SAN Explorer tool.

N_Port ID Virtualization enhancement

N_Port ID Virtualization (NPIV) allows multiple system images (in LPARs or z/VM guests) to use a single FCP channel as though each were the sole user of the channel. This feature, first introduced with z9 EC, can be used with earlier FICON features that have been carried forward from earlier servers.

Using the FICON Express16S (or above) as an FCP channel with NPIV enabled, the maximum numbers of these aspects for one FCP physical channel are doubled:

1. Maximum number of NPIV hosts defined: Increased from 32 to 64
2. Maximum number of remote N_Ports communicated: Increased from 512 to 1024
3. Maximum number of addressable LUNs: Increased from 4096 to 8192

4. Concurrent I/O operations: Increased from 764 to 1528

For operating system information of NPIV supporting, see , “N_Port ID Virtualization” on page 273.

Export / import physical port WWPNs for FCP Channels

IBM Z servers automatically assign worldwide port names (WWPNs) to the physical ports of an FCP channel based on the PCHID. This WWPN assignment changes when an FCP channel is moved to a different physical slot position. IBM z14, z13 and z13s servers will allow for the modification of these default assignments, allowing FCP channels to keep previously assigned WWPNs, even after being moved to a different slot position. This capability can eliminate the need for reconfiguration of the SAN in many situations, and is especially helpful during a system upgrade (FC #0099 - WWPN Persistence).

Note: For more information about the FICON enhancement of IBM z14 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 only supported a single ISL (a single hop) in a cascaded FICON SAN environment. The IBM z14, 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 detail information about the FICON multi-hop, see:

http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP102704

FICON feature summary

Table 4-7 shows the FICON feature codes, cable type, maximum unrepeated distance, and the link data rate on a IBM z14 server. 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

Channel feature	Feature codes	Bit rate	Cable type	Maximum unrepeated distance[a] (MHz -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 v11 and v12 with z/OS 2.1 with patches.

The zHyperLink Express feature (FC #0431) provides a low latency direct connection between z14 CPC 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 CPC directly to I/O Bay of the DS8880, using PCIe Gen3 x 8 physical link (up to 150m distance). A new transport protocol is defined for reading and writing ECKD™ data records, as documented in the zHyperLink interface specification.

On z14, zHyperLink Express card is a new PCIe adapter which installed in the PCIe I/O drawer. HCD definition support has been added for new PCIe function type with PORT attributes.

Requirements of zHyperLink

The zHyperLink Express feature is available on z14 servers, and requires:

- 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 CPC (up to 32 links).

The zHyperLink Express is managed 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 gives a maximum of 254 VFs per adapter.

- zHyperLink connector on DS8880 I/O Bay
  For DS8880 firmware R8.3 above, the I/O Bay planar is updated to support the zHyperLink interface. This includes the update of the PEX 8732 switch to PEX8733 that includes a DMA engine for the zHyperLink transfers, and also 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 150m.

4.7.3 Network Connectivity

Communication for LANs is provided by the OSA-Express6S, OSA-Express5S, OSA-Express4S (1000BaseT only), 10GbE RoCE Express2 and 10GbE RoCE Express features.

OSA-Express6S

The OSA-Express6S feature is installed in the PCIe I/O drawer. The following OSA-Express6S features can be installed on z14 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

Table 4-8 lists the supported OSA-Express6S features.

Table 4-8 OSA-Express6S features

I/O feature	Feature code	Number of ports per feature	Port increment	Maximum number of ports	Maximum number of features	CHPID type
OSA-Express6S 10 GbE LR	0424	1	1	48	48	OSD, OSX
OSA-Express6S 10 GbE SR	0425	1	1	48	48	OSD, OSX
OSA-Express6S GbE LX	0422	2	2	96	48	OSD
OSA-Express6S GbE SX	0423	2	2	96	48	OSD

I/O feature	Feature code	Number of ports per feature	Port increment	Maximum number of ports	Maximum number of features	CHPID type
OSA-Express6S 1000BASE-T	0426	2	2	96	48	OSC, OSD, OSE, OSM

OSA-Express6S 10 Gigabit Ethernet LR (FC #0424)

The OSA-Express6S 10 Gigabit Ethernet (GbE) Long Reach (LR) feature has one PCIe adapter and one port per feature. The port supports CHPID types OSD and OSX. 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), but ordering of any zBX features has been 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 has 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 μm single-mode fiber optic cable that terminates 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 has 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 can be 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 terminates with an LC Duplex connector is required for connecting each port on this feature to the selected device.

OSA-Express5S Gigabit Ethernet LX (FC #0422)

The OSA-Express6S GbE LX feature has 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 terminates with an LC Duplex connector is required for connecting each port on this feature to the selected device. If multimode fiber optic cables are being reused, a pair of Mode Conditioning Patch cables is required, with one cable for each end of the link.

OSA-Express6S Gigabit Ethernet SX (FC #0423)

The OSA-Express6S GbE SX feature has 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 SX feature supports the use of an LC Duplex connector. Ensure that the attaching or downstream device has an SX transceiver. The sending and receiving transceivers must be the same (SX to SX).

A multi-mode fiber optic cable that terminates with an LC Duplex connector is required for connecting each port on this feature to the selected device.

OSA-Express5S 1000BASE-T Ethernet feature (FC #0426)

Feature code 0426 occupies one slot in the PCIe I/O drawer. It has two ports that connect to a 1000 Mbps (1 Gbps) or 100 Mbps Ethernet LAN. Each port has a 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 m (328 ft). The SFP allows a concurrent repair or replace action.

OSA-Express6S 1000BASE-T adapters[a]**:** OSA-Express6S 1000BASE-T adapters (#0426) will be the last generation of OSA 1000BASE-T adapters to support connections operating at 100 Mb/second link speed. Future OSA-Express 1000BASE-T adapter generations will support operation only at 1000 Mb/second (1Gb/s) link speed.

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, 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 you are not using auto-negotiate, the OSA-Express port attempts to join the LAN at the specified speed and duplex mode. If this specified speed and duplex mode do not match the speed and duplex mode of the signal on the cable, the OSA-Express port does not connect.

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

Table 4-8 lists the OSA-Express5S features.

Table 4-9 OSA-Express5S features for z14

I/O feature	Feature code	Number of ports per feature	Port increment	Maximum number of ports	Maximum number of features	CHPID type
OSA-Express5S 10 GbE LR	0415	1	1	48	48	OSD, OSX
OSA-Express5S 10 GbE SR	0416	1	1	48	48	OSD, OSX
OSA-Express5S GbE LX	0413	2	2	96	48	OSD
OSA-Express5S GbE SX	0414	2	2	96	48	OSD
OSA-Express5S 1000BASE-T	0417	2	2	96	48	OSC, OSD, OSE, OSM

OSA-Express5S 10 Gigabit Ethernet LR (FC #0415)

The OSA-Express5S 10 Gigabit Ethernet (GbE) Long Reach (LR) feature has one PCIe adapter and one port per feature. The port supports CHPID types OSD and OSX. The 10 GbE feature is designed to support attachment to a single-mode fiber 10 Gbps Ethernet LAN or Ethernet switch that is capable of 10 Gbps. The port can be defined as a spanned channel and can be shared among LPARs within and across logical channel subsystems.

The OSA-Express5S 10 GbE LR feature supports the use of an industry standard small form factor LC Duplex connector. Ensure that the attaching or downstream device has an LR transceiver. The transceivers at both ends must be the same (LR to LR).

The OSA-Express5S 10 GbE LR feature does not support auto-negotiation to any other speed and runs in full duplex mode only.

A 9 µm single-mode fiber optic cable that terminates with an LC Duplex connector is required for connecting this feature to the selected device.

OSA-Express5S 10 Gigabit Ethernet SR (FC #0416)

The OSA-Express5S 10 GbE Short Reach (SR) feature has one PCIe adapter and one port per feature. The port supports CHPID types OSD and OSX. When defined as CHPID type OSX, the 10 GbE port provides connectivity and access control to the IEDN from 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 can be shared among LPARs within and across logical channel subsystems.

The OSA-Express5S 10 GbE SR feature supports the use of an industry standard small form factor LC Duplex connector. Ensure that the attaching or downstream device has an SR transceiver. The sending and receiving transceivers must be the same (SR to SR).

The OSA-Express5S 10 GbE SR feature does not support auto-negotiation to any other speed and runs in full duplex mode only.

A 50 or a 62.5 μm multimode fiber optic cable that terminates with an LC Duplex connector is required for connecting each port on this feature to the selected device.

OSA-Express5S Gigabit Ethernet LX (FC #0413)

The OSA-Express5S GbE LX feature has one PCIe adapter and two ports. The two ports share a channel path identifier (CHPID type OSD exclusively). The ports support attachment to a 1 Gbps Ethernet LAN. Each port can be defined as a spanned channel and can be shared among LPARs and across logical channel subsystems.

The OSA-Express5S GbE LX feature supports the use of an LC Duplex connector. Ensure that the attaching or downstream device has an LX transceiver. The sending and receiving transceivers must be the same (LX to LX).

A 9 μm single-mode fiber optic cable that terminates with an LC Duplex connector is required for connecting each port on this feature to the selected device. If multimode fiber optic cables are being reused, a pair of Mode Conditioning Patch cables is required, with one cable for each end of the link.

OSA-Express5S Gigabit Ethernet SX (FC #0414)

The OSA-Express5S GbE SX feature has one PCIe adapter and two ports. The two ports share a channel path identifier (CHPID type OSD exclusively). The ports support attachment to a 1 Gbps Ethernet LAN. Each port can be defined as a spanned channel and can be shared among LPARs and across logical channel subsystems.

The OSA-Express5S GbE SX feature supports the use of an LC Duplex connector. Ensure that the attaching or downstream device has an SX transceiver. The sending and receiving transceivers must be the same (SX to SX).

A multi-mode fiber optic cable that terminates with an LC Duplex connector is required for connecting each port on this feature to the selected device.

OSA-Express5S 1000BASE-T Ethernet feature (FC #0417)

Feature code 0417 occupies one slot in the PCIe I/O drawer. It has two ports that connect to a 1000 Mbps (1 Gbps) or 100 Mbps Ethernet LAN. Each port has a 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 m (328 ft). The SFP allows a concurrent repair or replace action.

The OSA-Express5S 1000BASE-T Ethernet feature supports auto-negotiation when attached to an Ethernet router or switch. If you allow the LAN speed and duplex mode to default to

auto-negotiation, the OSA-Express port and the attached router or switch auto-negotiate the LAN speed and duplex mode settings between them. They then connect at the highest common performance speed and duplex mode of interoperation. If the attached Ethernet router or switch does not support auto-negotiation, the OSA-Express port examines the signal that it is receiving and connects at the speed and duplex mode of the device at the other end of the cable.

The OSA-Express5S 1000BASE-T Ethernet feature can be configured as CHPID type OSC, OSD, OSE, or OSM. Non-QDIO operation mode requires CHPID type OSE.

The following settings are supported on the OSA-Express5S 1000BASE-T Ethernet feature port:

- Auto-negotiate
- 100 Mbps half-duplex or full-duplex
- 1000 Mbps full-duplex

If you are not using auto-negotiate, the OSA-Express port 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 addresses 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.

Table 4-10 lists the characteristics of the OSA-Express4S features which supported on IBM z14.

Table 4-10 OSA-Express4S features on IBM z14 servers

I/O feature	Feature code	Number of ports per feature	Port increment	Maximum number of ports (CHPIDs)	Maximum number of features	CHPID type
OSA-Express4S 1000BASE-T	0408	2	2	96	48	OSC, OSD, OSE, OSM

OSA-Express4S 1000BASE-T Ethernet feature (FC #0408)

The OSA-Express4S 1000BASE-T Ethernet feature occupies one slot in the PCIe I/O drawer. It has two ports that connect to a 1000 Mbps (1 Gbps), 100 Mbps, or 10 Mbps Ethernet LAN. Each port has an RJ-45 receptacle for cabling to an Ethernet switch. The RJ-45 receptacle must be attached by using an EIA/TIA Category 5 or Category 6 UTP cable with a maximum length of 100 meters (328 ft).

The OSA-Express4S 1000BASE-T Ethernet feature supports auto-negotiation when attached to an Ethernet router or switch. If you allow the LAN speed and duplex mode to default to auto-negotiation, the OSA-Express port and the attached router or switch auto-negotiate the LAN speed and duplex mode settings between them. They connect at the highest common performance speed and duplex mode of interoperation. If the attached Ethernet router or switch does not support auto-negotiation, the OSA-Express port examines the signal that it is receiving. It then connects at the speed and duplex mode of the device at the other end of the cable.

The OSA-Express4S 1000BASE-T Ethernet feature can be configured as CHPID type OSC, OSD, OSE, or OSM. Non-QDIO operation mode requires CHPID type OSE.

The following settings are supported on the OSA-Express4 1000BASE-T Ethernet feature port:

- Auto-negotiate
- 10 Mbps half-duplex or full-duplex
- 100 Mbps half-duplex or full-duplex
- 1000 Mbps full-duplex

If you are not using auto-negotiate, the OSA-Express port attempts to join the LAN at the specified speed and duplex mode. If these settings do not match the speed and duplex mode of the signal on the cable, the OSA-Express port does not connect.

10 GbE RoCE Express2

RoCE Express2 (FC #0412) is installed in the PCIe I/O drawer and is supported on IBM z14. 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 IBM z14 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 terminates with an LC Duplex connector. Both point-to-point connections and switched connections with an enterprise-class 10 GbE switch are supported.

Switch configuration for RoCE 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 unrepeated distance, point-to-point, is 300 meters (984 ft). A client-supplied cable is required. Three types of cables can be used for connecting the port to the selected 10 GbE switch or to the 10GbE RoCE Express2 feature on the attached server:

- OM3 50 micron multimode fiber optic cable that is rated at 2000 MHz-km that terminates with an LC Duplex connector (supports 300 meters (984 ft))
- OM2 50 micron multimode fiber optic cable that is rated at 500 MHz-km that terminates with an LC Duplex connector (supports 82 meters (269 ft))
- OM1 62.5 micron multimode fiber optic cable that is rated at 200 MHz-km that terminates with an LC Duplex connector (supports 33 meters (108 ft))

For more information about the management and definition of the 10GbE RoCE2, see Appendix D., “Shared Memory Communications” on page 457 and Appendix C, “Native Peripheral Component Interconnect Express (PCIe)” on page 451.

10 GbE RoCE Express

RoCE Express (FC #0411) is installed in the PCIe I/O drawer and is supported z13, zEC12 and zBC12, can be carried forward to IBM z14 on an upgrade. 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 only be dedicated to an LPAR, 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 and z13. The 10GbE RoCE Express feature uses SR optics, and supports the use of a multimode fiber optic cable that terminates with an LC Duplex connector. Both point-to-point connections and switched connections with an enterprise-class 10 GbE switch are supported.

Switch configuration for RoCE: If the IBM 10GbE RoCE Express features are connected to 10 GbE switches, the switches must meet the following requirements:

- Global Pause function enabled
- Priority flow control (PFC) disabled
- No firewalls, no routing, and no IEDN

The maximum supported unrepeated distance, point-to-point, is 300 meters (984 ft). A client-supplied cable is required. Three types of cables can be used for connecting the port to the selected 10 GbE switch or to the 10GbE RoCE Express feature on the attached server:

- OM3 50 micron multimode fiber optic cable that is rated at 2000 MHz-km that terminates with an LC Duplex connector (supports 300 meters (984 ft))
- OM2 50 micron multimode fiber optic cable that is rated at 500 MHz-km that terminates with an LC Duplex connector (supports 82 meters (269 ft))
- OM1 62.5 micron multimode fiber optic cable that is rated at 200 MHz-km that terminates with an LC Duplex connector (supports 33 meters (108 ft))

For more information about the management and definition of the 10GbE RoCE, see Appendix D., “Shared Memory Communications” on page 457 and Appendix C, “Native Peripheral Component Interconnect Express (PCIe)” on page 451.

Shared Memory Communications (SMC) functions

The Shared Memory Communication capabilities of the z14 help optimize the communications between applications for server-to-server (SMC-R) or LPAR-to-LPAR (SMC-D) connectivity.

SMC-R provides application transparent exploitation of the RoCE-Express feature reducing the network overhead and latency of data transfers, effectively offering the benefits of optimized network performance across processors.

SMC-D was exploited 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 providing 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 Shared Memory Communications solution, providing synergy with SMC-R.

Both SMC-R and SMC-D use shared memory architectural concepts, eliminating the TCP/IP processing in the data path, yet preserving TCP/IP Qualities of Service for connection management purposes.

Internal Shared Memory

Internal Shared Memory (ISM) is a function supported by IBM z14, z13 and z13s machines. It is the firmware that provides connectivity using shared memory access between multiple operating system images within the same CPC. ISM creates virtual adapters with shared memory 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 using HiperSockets can have network latency and CPU reduction benefits and performance improvement, by using the SMC-D over ISM.

IBM z14 and z13 servers support up to 32 ISM VCHIDs per CPC, and each VCHID supports up to 255 VFs, with a total maximum of 8K VFs.

For more information about the SMC-D and ISM, see Appendix D, “Shared Memory Communications” on page 457.

HiperSockets

The HiperSockets function of IBM z14 servers provides up to 32 high-speed virtual LAN attachments.

HiperSockets IOCP definitions on IBM z14**:** A parameter has been added for HiperSockets IOCP definitions on IBM z14 and z13 servers. Therefore, the IBM z14 IOCP definitions need to be migrated to support the HiperSockets definitions (CHPID type IQD).

On IBM z14 and z13 servers, the CHPID statement of HiperSockets devices requires the keyword VCHID. VCHID specifies the virtual channel identification number associated with the channel path. Valid range is 7E0 - 7FF.

VCHID is not valid on Z servers before z13.

For more information, see the *IBM Z Input/Output Configuration Program User's Guide for ICP IOCP,* SB10- 7163.

HiperSockets can be customized to accommodate varying traffic sizes. Because HiperSockets does not use an external network, it can free up system and network resources. This advantage can help eliminate attachment costs, and improve availability and performance.

HiperSockets eliminates having to use I/O subsystem operations and having to traverse an external network connection to communicate between LPARs in the same z14 server. HiperSockets offers significant value in server consolidation when connecting many virtual servers. It can be used instead of certain coupling link configurations in a Parallel Sysplex.

HiperSockets internal networks support two transport modes:

- Layer 2 (link layer)
- Layer 3 (network or IP layer)

Traffic can be IPv4 or IPv6, or non-IP, such as AppleTalk, DECnet, IPX, NetBIOS, or SNA.

HiperSockets devices are protocol-independent and Layer 3 independent. Each HiperSockets device (Layer 2 and Layer 3 mode) has its own 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 on IBM z14 and z13 servers is supported by Linux on z Systems, and by z/VM for Linux guest use.

IBM z14 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, during high volume situations, data can be temporarily held until the receiver has buffers that are available in its inbound queue. The HiperSockets Completion Queue function requires the following applications at a minimum[6]:

- z/OS V1.13
- Linux on z Systems distributions:
  - Red Hat Enterprise Linux (RHEL) 6.2
  - SUSE Linux Enterprise Server (SLES) 11 SP2
- z/VSE V5.1.1
- z/VM V6.2[7] with maintenance

In z/VM V6.2, the virtual switch function is enhanced to transparently bridge a guest virtual machine network connection on a HiperSockets LAN segment. This bridge allows a single HiperSockets guest virtual machine network connection to communicate directly with these systems:

- Other guest virtual machines on the virtual switch
- External network hosts through the virtual switch OSA UPLINK port

4.7.4 Parallel Sysplex connectivity

Coupling links are required in a Parallel Sysplex configuration to provide connectivity from the z/OS images to the coupling facility. A properly configured Parallel Sysplex provides a highly reliable, redundant, and robust IBM Z technology solution to achieve near-continuous

[6] Minimum OS support for z14 may differ, see Chapter 7, “Operating System Support” on page 227
[7] z/VM V6.2 is not supported on z14. z/VM V6.3 or later is needed

availability. A Parallel Sysplex is composed of one or more z/OS operating system images that are coupled through one or more coupling facilities.

Coupling links

The type of coupling link that is used to connect a coupling facility (CF) to an operating system LPAR is important. The link performance has a significant effect on response times and coupling processor usage. For configurations that cover large distances, the time that is spent on the link can be the largest part of the response time.

These links are available to connect an operating system LPAR to a coupling facility:

- Integrated Coupling Adapter (ICA SR) for short distance connectivity, defined as CHPID type CS5. The ICA SR can be used only for coupling connectivity between z14, z13 and z13s servers. It does not support connectivity to zEC12, zBC12 servers, and it cannot be connected to HCA3-O or HCA3-O LR coupling fanouts. The ICA SR supports distances up to 150 m and a link data rate of 8 GBps. OM3 fiber optic cable is used for distances up to 100 m, and OM4 for distances up to 150 m. ICA SR supports four CHPIDs per port and seven subchannels (devices) per CHPID. ICA SR supports transmission of Server Time Protocol (STP) messages.
- Parallel Sysplex using IFB 12x connects z14, z13, z13s, zEC12, and zBC12 servers. 12x InfiniBand coupling links are fiber optic connections that support a maximum distance of up to 150 meters (492 ft). IFB coupling links are defined as CHPID type CIB. IFB supports transmission of STP messages.
- Parallel Sysplex using InfiniBand 1x Long Reach (IFB LR) connects z14, z13, z13s, zEC12, and zBC12. 1x InfiniBand coupling links are fiber optic connections that support a maximum unrepeated distance of up to 10 km (6.2 miles), and up to 100 km (62 miles) with a IBM Z qualified DWDM. IFB LR coupling links are defined as CHPID type CIB. IFB LR supports transmission of STP messages.
- Coupling Express Long Reach: Coupling Express LR (FC #0433) is recommended for Long Distance Coupling IBM z14/z13/z13s to z13 and above. It support a maximum unrepeated distance to 10kM (6.2 miles) and up to 100kM with a qualified DWDM. CE LR coupling links are defined as CHPID CL5. CE LR utilizes same 9µm single mode fiber cable as 1x IFB.

Table 4-11 shows the coupling link options.

Table 4-11 Coupling link options supported on z14

Type	Description	Use for connecting	Link rate[a]	Distance	z14 Maximum # of ports
IFB	12x InfiniBand (HCA3-O)[b]	z14/z13/z13s/zEC12/zBC12 to z14/z13/z13s/zEC12/zBC12	6 GBps	150 meters (492 feet)	32
IFB LR	1x IFB (HCA3-O LR)	z14/z13/z13s/zEC12/zBC12 to z14/z13/z13s/zEC12/zBC12	5.0 Gbps	10 km unrepeated (6.2 miles) 100 km repeated (62 miles)	64
CE LR	Coupling Express LR	z14/z13/z13s to z14/z13/z13s	10 Gbps	10 km unrepeated (6.2 miles) 100 km repeated (62 miles)	64

Type	Description	Use for connecting	Link rate[a]	Distance	z14 Maximum # of ports
ICA SR	Integrated Coupling Adapter	z14/z13/z13s to z14/z13/z13s	8 GBps	150 meters (492 feet)	80
IC	Integrated Coupling Adapter	Internal communication	Internal speeds	N/A	40

a. The link data rates 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.
b. 12x IFB3 protocol supports a maximum of four CHPIDs and connects to the other HCA3-O port. When connected to a HCA2-O port, 12x IFB protocol is used. The protocol is auto-configured when conditions are met for IFB3. For more information, see 4.5.3, “HCA3-O (12x IFB) fanout (FC #0171)” on page 149.

The maximum number of combined external coupling links (active CE LR, ICA SR links and IFB LR) is 144 per IBM z14 server. And IBM z14 servers support up to 256 coupling CHPIDs per CPC. Figure 4-6 shows z14 coupling link support summary for z14.

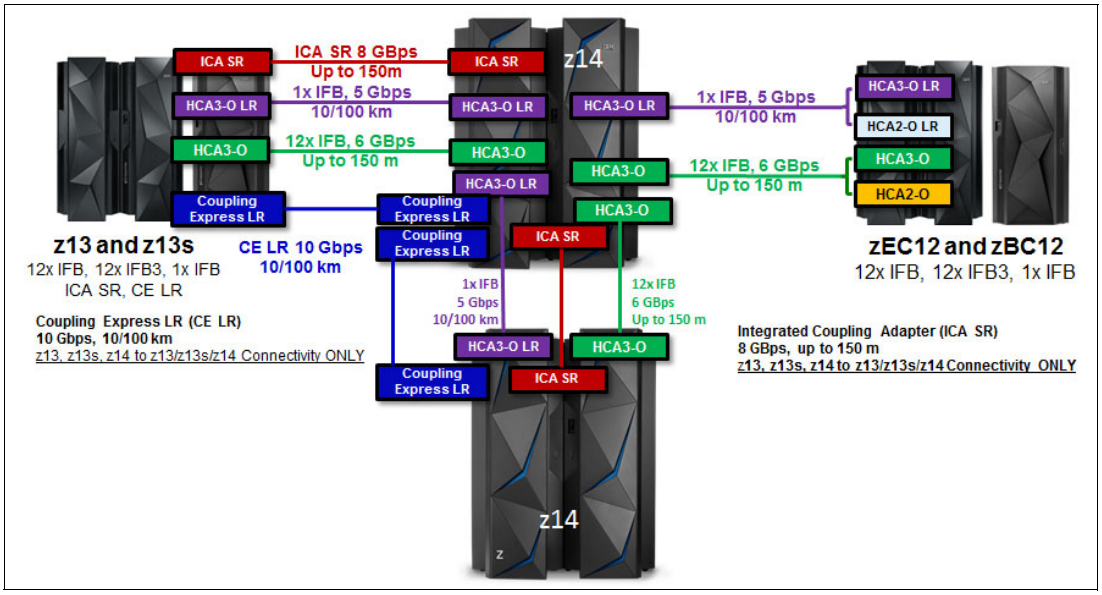


Figure 4-6 z14 Parallel Sysplex coupling connectivity

When defining IFB coupling links (CHPID type CIB), HCD defaults to seven subchannels. 32 subchannels are supported only on HCA2-O LR (1xIFB) and HCA3-O LR (1xIFB) on zEC12 and later when both sides of the connection use IFB protocol.

In a Parallel Sysplex configuration, z/OS and coupling facility images can run on the same or on separate servers. There must be at least one CF that is connected to all z/OS images, even though other CFs can be connected only to selected z/OS images. Two coupling facility images are required for system-managed CF structure duplexing. In this case, each z/OS image must be connected to both duplexed CFs.

To eliminate any single points of failure in a Parallel Sysplex configuration, have at least the following components:

- Two coupling links between the z/OS and coupling facility images.
- Two coupling facility images not running on the same server.

- One stand-alone coupling facility. If using system-managed CF structure duplexing or running with *resource sharing* only, a stand-alone coupling facility is not mandatory.

Coupling link features

IBM z14 server supports the following coupling link feature:

- HCA3-O fanout for 12x InfiniBand, FC #0171
- HCA3-O LR fanout for 1x InfiniBand, FC #0170
- ICA SR fanout, FC #0172
- CE LR adapter, FC #0433

Extended distance support

For 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 servers. 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 (implemented in Licensed Internal Code (LIC)), and so does not require any hardware or cabling.

An IC link is a fast coupling link that uses memory-to-memory data transfers. IC links do not have PCHID numbers, but do require CHPIDs.

IC links require an ICP channel path definition at the z/OS and the CF end of a channel connection to operate in peer mode. They are always defined and connected in pairs. The IC link operates in peer mode, and its existence is defined in HCD/IOCP.

IC links have the following attributes:

- Operate in peer mode (channel type ICP) on IBM Z servers.
- Provide the fastest connectivity, which is faster than any external link alternatives.
- Result in better coupling efficiency than with external links, effectively reducing the server cost that is associated with Parallel Sysplex technology.
- Can be used in test or production configurations, and reduce the cost of moving into Parallel Sysplex technology while also enhancing performance and reliability.
- Can be defined as spanned channels across multiple CSSs.
- Are available for no extra fee (no feature code). Employing ICFs with IC channels results in considerable cost savings when you are configuring a cluster.

IC links are enabled by defining channel type ICP. A maximum of 32 IC channels can be defined on a Z server.

Migration considerations

Upgrading from previous generations of IBM Z servers in an existing Parallel Sysplex to IBM z14 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 CPC drawer, as compared to the number of available fanout slots of the processor books of the previous generation Z servers like zEC12.

IBM z14 servers do not support ISC-3 links, HCA2-o and HCA2-O (LR).

HCA3-O link compatibility: HCA3-O (LR) links *can* connect to HCA2-O (LR) on zEC12, and zBC12.

z196, z114 are NOT supported in same Parallel Sysplex or STP CTN with IBM z14.

The ICA SR fanout provides short-distance connectivity to another IBM z14/z13s/z13 server. For more information, see Table 4-11 on page 176.

The CE LR adapter provides long-distance connectivity to another IBM z14/z13s/z13 server. For more information, see “Coupling links” on page 176

The z14 server fanout slots in the CPC drawer provide coupling links connectivity through the ICA SR and IFB fanout cards. In addition to coupling links for Parallel Sysplex, the fanout cards that the fanout slots provide connectivity for the PCIe I/O drawer (PCIe fanout). Up to 10 PCIe and four IFB fanout cards can be installed in each CPC drawer, as shown in Figure 4-7.

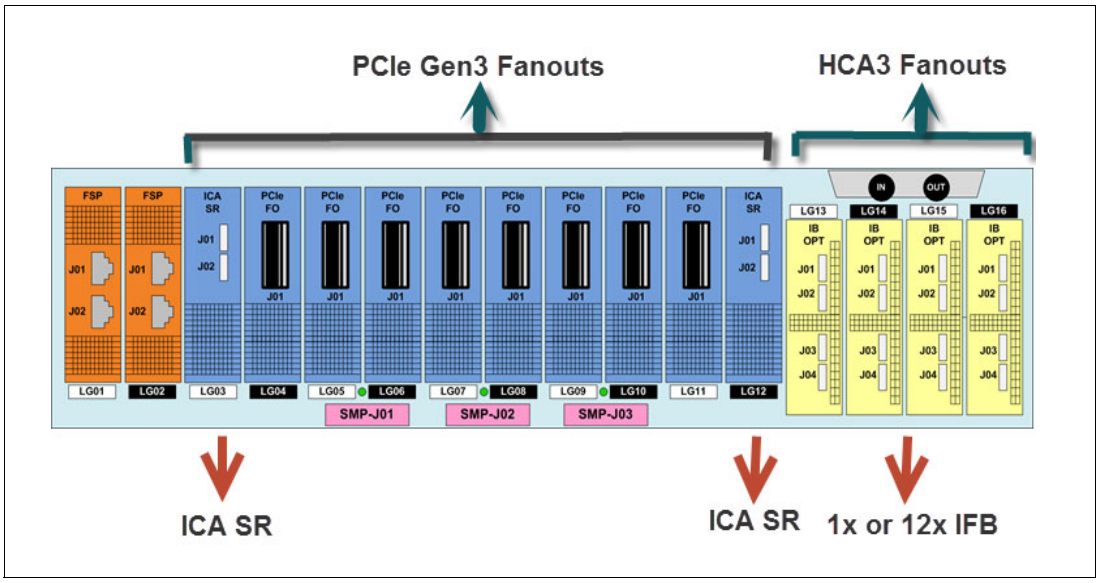


Figure 4-7 CPC drawer front view showing the coupling links

Previous generation IBM Z servers, in particular z196 and zEC12, use processor books, which provide connectivity for up to eight InfiniBand fanouts per book.

As an example of a possible migration case, assume a one-book zEC12 used as stand-alone coupling facility with all eight fanouts that are used for IFB connectivity. In a 1:1 link migration scenario, this server cannot be upgraded to a one z14 CPC drawer because the IBM z14 server cannot accommodate more than four InfiniBand fanouts in a single CPC drawer. For more information, see 4.5.5, “Fanout considerations” on page 151.

In this case, a second CPC drawer is needed to fulfill all IFB connectivity, as shown in Figure 4-8.

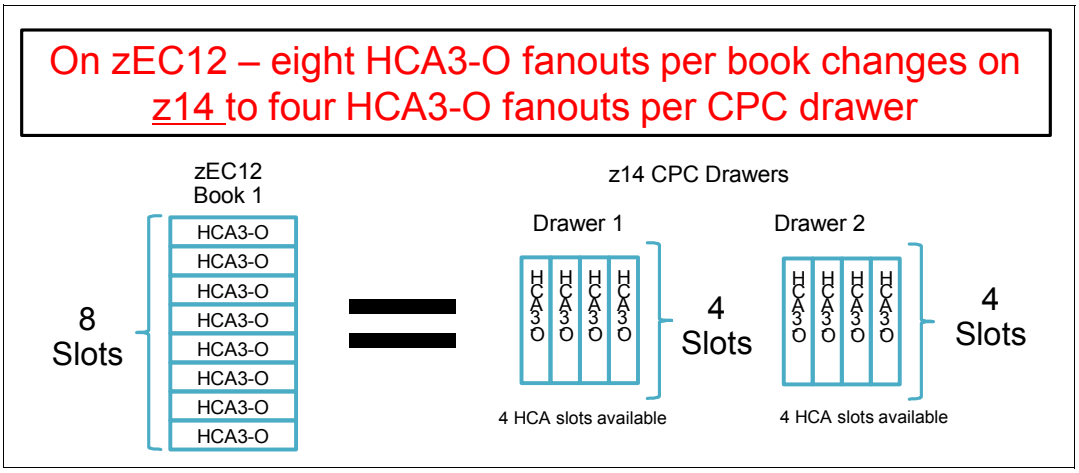


Figure 4-8 HCA3-O Fanouts: z14 versus zEC12 servers

It is beyond the scope of this book to describe all possible migration scenarios. Always involve subject matter experts (SMEs) to help you to develop your migration strategy.

The following list of considerations can help you assess possible migration scenarios. The objective of this list is to enable migration to IBM z14 servers, supporting legacy coupling where essential, and adopting ICA SR where possible, to avoid the need for more CPC drawers and other possible migration issues.

- The IBM zEnterprise EC12 and BC12 are the last generation of Z servers to support ISC-3, 12x HCA2-O, 1x HCA2-O LR. They also are the last Z servers that can be part of a Mixed Coordinated Timing Network (CTN).
- Consider Long Distance Coupling requirements first:
  - HCA3-O 1x or CE LR are the long-distance coupling link that are available on IBM z14 servers.
  - ICA SR or HCA3-O 12x should be used for short distance coupling requirements.
- ISC-3 Migration (IBM z14/z13 servers do not support ISC-3):
  - Evaluate current ISC-3 usage (long- and short-distance, coupling data, or timing only) to determine how to fulfill ISC-3 requirements with the links that are available on IBM z14/z13 servers.
  - You can migrate from ISC-3 to CE LR, ICA SR, 12x InfiniBand, or 1x InfiniBand on IBM z14/z13 servers.
  - 1:1 Mapping of ISC-3 to Coupling over InfiniBand. On previous servers, the HCA2-C fanouts enable ISC-3 coupling in the I/O Drawer. Two HCA2-C fanouts can be replaced by two 1x fanouts (eight 1x links) or two 12x fanouts (four 12x links).
  - ISC-3 supports one CHPID/link. Consolidate ISC-3 CHPIDs into CE LR, ICA SR or IFB, and use multiple CHPIDs per link.
- Evaluate configurations for opportunities to eliminate or consolidate InfiniBand links:
  - Eliminate any redundant links. Two physical links between CPCs is the minimum requirement from a reliability, availability, and serviceability (RAS) perspective.
  - Share logical CHPIDs on a physical IFB link connection to reduce the usage of IFB links in z14 servers (even multiple sysplexes can share a single link).
  - Coupling Link Analysis: Capacity Planning tools and services can help.

- For z14/z13 to z14/z13, use CE LR and ICA SR as much as possible, and use existing InfiniBand links for z13/zEC12/zBC12 to z13/zEC12/zBC12 connectivity.
- Install all the ICA SR links that are required to fulfill future short distance coupling requirements:
  - When upgrading a CPC to a IBM z14 server, configure the IBM z14 server with all the ICA SR coupling links that eventually will be needed (that is, avoid loose piece MES with ICA SR links) in your Parallel Sysplex configuration.
  - Consider a plan-ahead configuration. In a multi-CPC configuration, a final approach can be to establish z14 to z14 connections by using mainly ICA SR connectivity. Even if the links are not used immediately, especially within the zEC12/zBC12 migration, most coupling connectivity uses InfiniBand. However, after the IBM z14 server is migrated, coupling connectivity can be moved to ICA SR links.
- Upgrade CPCs with fewer coupling constraints first:
  - Consider upgrading CPCs to have sufficient IFB links in the target z14 configuration first (for example, multi-CPC drawer CPCs).
  - Test the new CE LR or ICA SR link on the least constrained CPC (for example, z/OS Host CPCs that have the lowest number of links). For the CPCs that are involved in this test that do not have a CF, you might need to add a CF LPAR and ICF engine to one of the CPCs.
  - When migrating CPCs with more coupling links to a IBM z14 server, begin by using enough CE LR or ICA SR links in place of Coupling over InfiniBand (for example, half ICA SR, half 12x) to maintain the CPC footprint (that is, avoid extra CPC drawers).
- Consider replacing two servers that are close to each other at the same time. Assess the risk, and, if acceptable, immediately replace 12x IFB with ICA SR links in z14 peers, and replace 1x IFB with CE LR links, thus freeing InfiniBand links for connecting to old machine like zEC12/zBC12.
- Consider temporary performance reduction when consolidating more than four CHPIDs per IFB link. After the peer system has been migrated to a z14 or z13 server, the InfiniBand links can be migrated to CE LR or ICA SR links.

Coupling links and Server Time Protocol

All external coupling links can be used to pass time synchronization signals by using STP. STP is a message-based protocol in which timing messages are passed over data links between servers. The same coupling links can be used to exchange time and coupling facility messages in a Parallel Sysplex.

Using the coupling links to exchange STP messages has the following advantages:

- By using the same links to exchange STP messages and coupling facility messages in a Parallel Sysplex, STP can scale with distance. Servers exchanging messages over short distances, such as IFB or ICA SR links, can meet more stringent synchronization requirements than servers that exchange messages over long IFB LR links (distances up to 100 km (62 miles)). This advantage is an enhancement over the IBM Sysplex Timer implementation, which does not scale with distance.
- Coupling links also provide the connectivity that is necessary in a Parallel Sysplex. Therefore, there is a potential benefit 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 server does not support attachment to the IBM Sysplex Timer. A IBM z14 server cannot be added into a Mixed CTN. It can participate only in an STP-only CTN.

STP enhancements on z14

Important: For more information about configuring an STP CTN with three or more servers, see the white paper that is found at the following website:

`http://www.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP101833`

If the guidelines are not followed, it might result in all the servers in the CTN becoming unsynchronized. This condition results in a sysplex-wide outage.

Enhancements for STP on z14 are:

- Additional Stratum Level

Previously the limit was Stratum level 3 before z14. The extra stratum allows CPCs to operate as part of CTN stratum level 4. That can avoid the additional complexity and expense of system reconfiguration.

Warning: This additional stratum level should only be used as a temporary state during reconfiguration. Customer should not run with machines at stratum level 4 for extended periods of time because of the lower quality of the time synchronization.

- Graphical display of a Coordinated Timing Network (CTN)

This Graphical display improved user interface to STP controls. This kind of visual display of the CTN status, provided a clearer view of CTNs, can avoid the outage like: Customer intentionally took down the CTS, but not realizing the BTS was already down.

For more information about STP configuration, see these books:

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

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 on a small backplane mounted in the rear of the frame of z14 servers, it is 2x Oscillators with BNC connectors for pulse per second. With connection to all the CEC drawers, to provide redundancy for continued operation and concurrent maintenance when a single oscillator card fails. Each oscillator card has a Bayonet Neill-Concelman (BNC) connector for PPS connection support, attaching to two different ETSs. Two PPS connections from two different ETSs are preferable for redundancy.

The time accuracy of an STP-only CTN is improved by adding an ETS device with the PPS output signal. STP tracks the highly stable accurate PPS signal from ETSs. It maintains accuracy of 10 µs as measured at the PPS input of the z14 server. If STP uses an NTP server without PPS, a time accuracy of 100 ms to the ETS is maintained. ETSs with PPS output are available from various vendors that offer network timing solutions.

4.8 Cryptographic functions

Cryptographic functions are provided by the CP Assist for Cryptographic Function (CPACF) and the PCI Express cryptographic adapters. IBM z14 servers support the Crypto Express6S feature.

4.8.1 CPACF functions (FC #3863)

FC #3863[8] is required to enable CPACF functions.

4.8.2 Crypto Express6S feature (FC #0893)

Crypto Express6S is a new feature on IBM z14 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, is designed to conform 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).

CCA compliance levels:

- Non-compliant (default)
- PCI-HSM 2016
- PCI-HSM 2016 (migration, key tokens in the process of migrating to compliant)

EP11 compliance levels: (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 has no CHPID assigned. However, it has one PCHID.

4.8.3 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 Common Cryptographic Architecture (CCA) coprocessor, as a Secure IBM Enterprise Public Key Cryptography Standards (PKCS) #11 (EP11) coprocessor, or as an accelerator.

Each Crypto Express5S feature occupies one I/O slot in the PCIe I/O drawer, and has no CHPID assigned. However, it has one PCHID.

[8] Subject to export regulations.

4.9 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. These new features are installed in the PCIe I/O drawer:

- zEDC Express
- 10GbE RoCE Express2
- 10GbE RoCE Express
- zHyperLink Express

All native PCIe features should be ordered in pairs for redundancy. According to their physical location in the PCIe I/O drawer, the features are assigned to one of the four resource groups (RGs) that are running on the IFP, providing management functions and virtualization functions. If two features of same type are installed, one will always be 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 has a failure.

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 (PCIe)” on page 451.

4.10 zEDC Express

zEDC Express is an optional feature (FC #0420) that is available on IBM z14, z13, 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 are designed to help improve cross-platform data exchange, reduce CPU consumption, and save disk space.

The feature installs exclusively on the PCIe I/O drawer. Up to sixteen features can be installed on the system. There is one PCIe adapter/compression coprocessor 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 493 and Appendix C, “Native Peripheral Component Interconnect Express (PCIe)” on page 451

5

Central processor complex channel subsystem

This chapter addresses the concepts of the z14 channel subsystem, including multiple channel subsystems and multiple subchannel sets. It also describes the technology, terminology, and implementation aspects of the channel subsystem.

This chapter includes the following sections:

- Channel subsystem
- I/O configuration management
- Channel subsystem summary

© Copyright IBM Corp. 2017. All rights reserved.

5.1 Channel subsystem

Channel subsystem (CSS) is a collective name of facilities that Z servers used 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 proceeded 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 unites. For more information about the channel subsystem, see 5.1.1, “Multiple logical channel subsystems (LCSSs)”.

The design of IBM Z servers 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 set (MSS) since z9.

An overview of the channel subsystem for z14 servers is shown in Figure 5-1. z14 servers are designed to support up to six logical channel subsystems, each with four subchannel sets and up to 256 channels.

z14					
LCSS 0	LCSS 1	LCSS 2	LCSS 3	LCSS 4	LCSS 5
Up to 15 Logical Partitions	**Up to 15 Logical Partitions**	**Up to 15 Logical Partitions**	**Up to 15 Logical Partitions**	**Up to 15 Logical Partitions**	**Up to 10 Logical Partitions**
Subchannel Sets: **SS 0 – 63.75 k** **SS 1 – 64 k** **SS 2 – 64 k** **SS 3 – 64 k**	**Subchannel Sets:** **SS 0 – 63.75 k** **SS 1 – 64 k** **SS 2 – 64 k** **SS 3 – 64 k**	**Subchannel Sets:** **SS 0 – 63.75 k** **SS 1 – 64 k** **SS 2 – 64 k** **SS 3 – 64 k**	**Subchannel Sets:** **SS 0 – 63.75 k** **SS 1 – 64 k** **SS 2 – 64 k** **SS 3 – 64 k**	**Subchannel Sets:** **SS 0 – 63.75 k** **SS 1 – 64 k** **SS 2 – 64 k** **SS 3 – 64 k**	**Subchannel Sets:** **SS 0 – 63.75 k** **SS 1 – 64 k** **SS 2 – 64 k** **SS 3 – 64 k**
Up to 256 Channels	Up to 256 Channels	Up to 256 Channels	Up to 256 Channels	Up to 256 Channels	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 servers, the HSA is pre-allocated in memory with a fixed size of 192 GB, which is in addition to the customer purchased memory. This fixed size memory for HSA eliminates the requirement for additional planning of the initial I/O configuration and pre-planning for future I/O expansions.

CPC drawer **repair:** The HSA can be moved from one CPC drawer to a different one in an enhanced availability configuration as part of a concurrent CPC drawer repair (CDR) action.

These objects are always reserved in the z14 HSA during POR, whether they are defined in the IOCDS for use or not:

- Six CSSs
- Fifteen LPARs in each CSS0 to CSS4

- Ten LPARs in CSS5
- 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
- Subchannel set 3 with 64 K minus one device in each CSS

5.1.1 Multiple logical channel subsystems (LCSSs)

In the Z architecture, a *single channel subsystem* can have up to 256 channel paths defined, which use to limit the total numbers of I/O connectivities on older Z servers to 256.

The introducing of *multiple LCSSs* enabled an IBM Z server 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 transparent to the operating systems and applications. The multiple image facility (MIF) in the structure enables resource sharing across logical partitions within a single LCSS or across the LCSSs.

The multiple LCSS structure has extended 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 word *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 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, and 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 63.75K subchannels (with 0.25K 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 188, extra subchannel sets are available to increase the device addressability of a channel subsystem.

Channel paths

A *channel path* provides a connection between the channel subsystem and control unites that 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, like FICON, or an internal logical one, like HiperSockets.

Each channel path in a channel subsystem has a unique 2-digit hexadecimal identifier that is known as a channel-path identifier (CHPID), which ranges from 00 to FF. Therefore, a total of 256 CHPIDs are supported by a CSS, and a maximum of 1536 CHPIDs are available on a z14 server with six logical channel subsystems.

By assigning a CHPID to a physical port of an I/O feature card 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 has a unique physical channel identifier (PCHID) according to the physical location of this I/O feature card, and the sequence of this port on the card.

In addition, a port on a fanout adapter has a unique adapter identifier (AID), according to the physical location of this fanout card, and the sequence of this port on the card.

A CHPID is assigned to a physical port by defining the corresponding PCHID or AID in the I/O configuration definitions.

Control units

A *control unit* provides the logical capabilities that are necessary to operate and control an I/O device. It adapts the characteristics of each device so that it can respond to the standard form of control that is provided by the CSS.

A control unit can be housed separately, or can be physically and logically integrated with the I/O device, the channel subsystem, or within the Z server itself.

I/O devices

An *I/O device* provides external storage, a means of communication between data-processing systems, or a means of communication between a system and its environment. In the simplest case, an I/O device is attached to one control unit and is accessible through one or more channel paths 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, an operating system for example, running in the LPAR.

Each subchannel has a unique 4-digit hexadecimal number, ranging from 0x0000 to 0xFFFF. Thus, a single subchannel set can address and access up to 64K I/O devices.

MSS was introduced in z9 to extend the maximum number of addressable I/O devices for a channel subsystem.

Just like the z13 server, the z14 support four subchannel sets for each logical channel subsystem. It is capable of accessing a maximum of 255.74K devices for a logical channel subsystem, and thus a logical partition and the programs 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, ranging from 0x0000 to 0xFFFF, assigned to a subchannel within a subchannel set of a channel subsystem. Subchannels in each subchannel set are always assigned subchannel numbers within a single range of contiguous numbers. The lowest-numbered subchannel is subchannel 0, and the highest-numbered subchannel has a subchannel number equal to one less than the maximum numbers of subchannels supported by the subchannel set. Thus, a subchannel number is always unique within a subchannel set of a channel subsystem, and depends on the sequence of assigning.

With the subchannel numbers, a program running on an LPAR, an operating system for example, can specify all I/O functions relative to a specific I/O device by designating a subchannel assigned to the I/O devices.

Normally, subchannel numbers are only used in communication between the programs and the channel subsystem.

Subchannel set identifier

While introducing the MSS, the channel subsystem is extended to assign a value ranging from 0 to 3 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, ranging from 0x0000 to 0xFFFF, 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. Thus a subchannel in a subchannel set of a channel subsystem has a device number together with s subchannel number for designating an I/O operation.

The device number provide a means to identify a device, independent of any limitations imposed by the system model, the configuration, or channel-path protocols.

A device number also can be used to designate an I/O function to a specific I/O device. Because it is an arbitrary number, it can easily be fit into any configuration management and operating management scenarios. For example, that a system administrator can set all the z/OS systems in an environment to device number 1000 for their system RES volumes.

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 defined by user. That an I/O device can always be identified by an SSID in conjunction with a subchannel number or a device number. For example, a device with device number AB00 of subchannel set 1, can be designated as 1AB00.

Normally, the subchannel number is used by the programs to communicate with the channel subsystem and I/O device, whereas the device number is used by a system programmer, operator, and administrator.

Device in subchannel set 0 and additional subchannel sets

An LCSS always has the first subchannel set (SSID 0), which can have up to 63.75K subchannels, with 256 subchannels reserved by the channel subsystem. That user can always define their I/O devices in this subchannel set for general use.

For the additional subchannel sets enabled by the MSS facility, each has 65535 subchannels (64K minus one) for specific types of devices. These additional subchannel sets are referred as *alternate subchannel sets* in z/OS. Also, a device that is defined in an alternate subchannel set is considered a *special device*, which normally has a special device type in the I/O configuration.

Currently, a z14 server running z/OS defines these types of devices in an additional subchannel set, with proper APAP or PTF installed:

- Alias devices of the parallel access volumes (PAV).
- Secondary devices of GDPS Metro Mirror Copy Service (formerly Peer-to-Peer Remote Copy (PPRC)).

- FlashCopy SOURCE and TARGET devices with program temporary fix (PTF) OA46900.
- DB2 data backup volumes with PTF OA24142.

Using an additional subchannel set for these special devices helps reduce the numbers of devices in the subchannel set 0, which increases the growth capability for accessing more devices.

IPL from an alternate subchannel set

z14 servers support initial program load (IPL) from alternate subchannel sets in addition to subchannel set 0. Devices that are used early during IPL processing now can be accessed by using subchannel set 1, subchannel set 2, or subchannel set3 on a z14 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 alternate subchannel set to be used for IPL, an I/O definition file (IODF), and stand-alone memory dump volumes when needed.

IPL from an alternate subchannel set is supported by z/OS V1.13 or later, and Version 1.12 with PTFs.

The display ios,config command

The z/OS **display ios,config(all)** command that is shown in Figure 5-2 includes information about the MSSs.

```
D IOS,CONFIG(ALL)
IOS506I 11.32.19 I/O CONFIG DATA 340
ACTIVE IODF DATA SET = SYS6.IODF39
CONFIGURATION ID = LO6RMVS1       EDT ID = 01
TOKEN:  PROCESSOR DATE     TIME     DESCRIPTION
 SOURCE: SCZP501  14-10-31 08:51:47 SYS6     IODF39
ACTIVE CSS:  0    SUBCHANNEL SETS CONFIGURED: 0, 1, 2, 3
CHANNEL MEASUREMENT BLOCK FACILITY IS ACTIVE
LOCAL SYSTEM NAME (LSYSTEM): SCZP501
HARDWARE SYSTEM AREA AVAILABLE FOR CONFIGURATION CHANGES
PHYSICAL CONTROL UNITS               8099
CSS  0 - LOGICAL CONTROL UNITS       3996
 SS  0   SUBCHANNELS                54689
 SS  1   SUBCHANNELS                58862
 SS  2   SUBCHANNELS                65535
 SS  3   SUBCHANNELS                65535
CSS  1 - LOGICAL CONTROL UNITS       4088
 SS  0   SUBCHANNELS                65280
 SS  1   SUBCHANNELS                65535
 SS  2   SUBCHANNELS                65535
 SS  3   SUBCHANNELS                65535
CSS  2 - LOGICAL CONTROL UNITS       4088
 SS  0   SUBCHANNELS                65280
 SS  1   SUBCHANNELS                65535
 SS  2   SUBCHANNELS                65535
 SS  3   SUBCHANNELS                65535
CSS  3 - LOGICAL CONTROL UNITS       4088
 SS  0   SUBCHANNELS                65280
 SS  1   SUBCHANNELS                65535
 SS  2   SUBCHANNELS                65535
 SS  3   SUBCHANNELS                65535
CSS  4 - LOGICAL CONTROL UNITS       4088
 SS  0   SUBCHANNELS                65280
 SS  1   SUBCHANNELS                65535
 SS  2   SUBCHANNELS                65535
 SS  3   SUBCHANNELS                65535
CSS  5 - LOGICAL CONTROL UNITS       4088
 SS  0   SUBCHANNELS                65280
 SS  1   SUBCHANNELS                65535
 SS  2   SUBCHANNELS                65535
 SS  3   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.

While a shared channel path can be shared by LPARs within a same LCSS, a spanned channel path can be shared by LPARs within and across LCSSs.

By assigning the same CHPID from different LCSSs to the same channel path, a PCHID for example, the channel path can be accessed by any LPARs from these LCSSs at the same time. The CHPID is spanned across those LCSSs. Using spanned channels paths decreases the number of channels needed in an installation of Z servers.

A sample of channel paths defined as dedicated, shared, and spanned is shown in Figure 5-3.

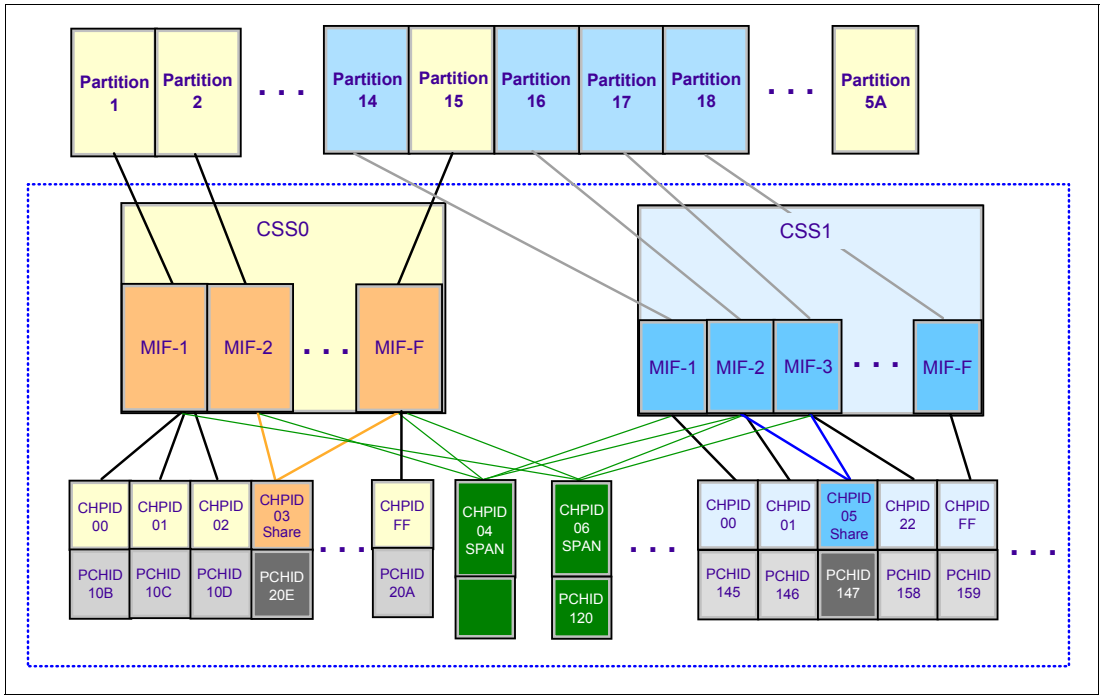


Figure 5-3 IBM Z CSS: Channel subsystems with channel spanning

In the sample, the three different definitions of a channel path are shown:

- CHPID FF, 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, assigned to PCHID 20E is shared access for partition 2, and 15 of LCSS0. The same applies to CHPID 05 of LCSS1.
- CHPID 06, assigned to PCHID 120 is spanned access for partition 1, 15 of LCSS0, and partition 16, 17 of LCSS1. 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 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, supporting the implementation of these dedicated, shared, and spanned paths.

An example of definition of these LPAR related identifications is shown in Figure 5-4.

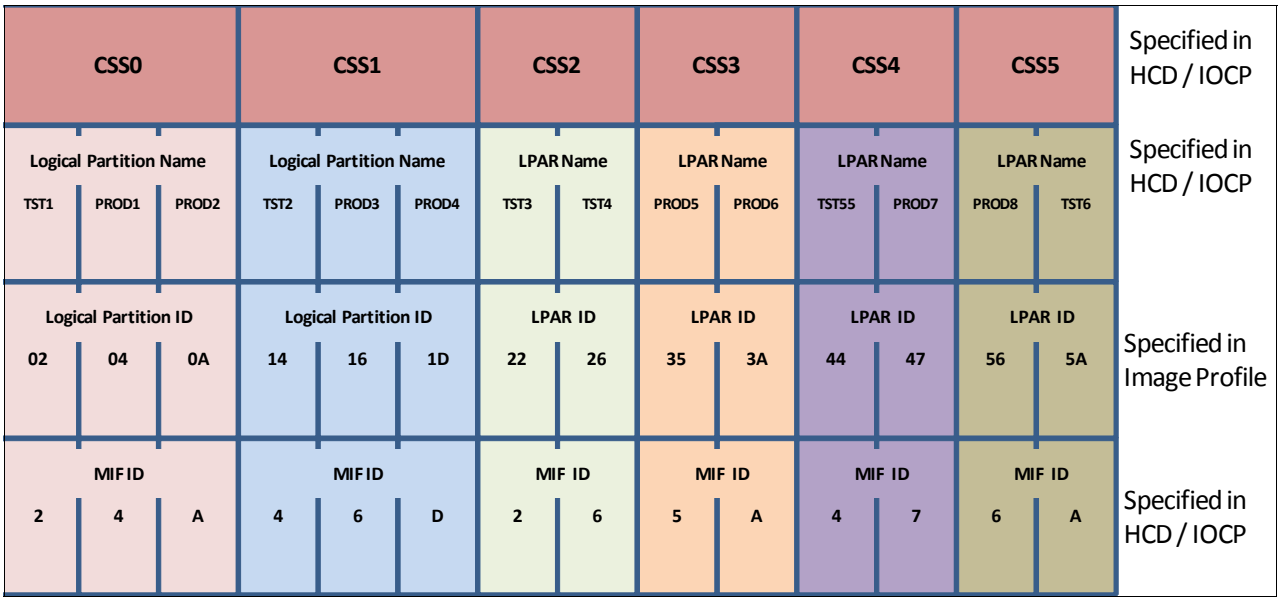


Figure 5-4 CSS, LPAR, and identifier example

LPAR name

The LPAR name is defined as partition name parameter in the **RESOURCE** statement of an I/O configuration. The LPAR name must be unique across the server.

MIF image ID

The MIF image ID is defined as a parameter for each LPAR in the **RESOURCE** statement of an I/O configuration. It ranges from 1 to F, and must be unique within an LCSS. 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 server supports a maximum of six LCSSs, with a total of 85 LPARs that can be defined. Each LCSS of a z14 server can support the following numbers of LPARs:

- LCSS0 to LCSS4 support 15 LPARs each, and the MIF image ID ranges from 1 to F
- LCSS5 supports 10 LPARs, and the MIF image IDs range from 1 to 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 from 00 to 7F. The LPAR ID must be unique across the server. Although it is arbitrarily defined by the user, normally an LPAR ID will be 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 defined in that 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 new configurations or upgrades of an existing 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 what the target configuration will look like.

- Hardware configuration definition (HCD): HCD supplies an interactive dialog to generate the IODF, and later the IOCDS. Generally, use HCD or Hardware Configuration Manager (HCM) to generate the I/O configuration rather than writing 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 both dynamic hardware and software I/O configuration changes.

  Note: Certain functions might require specific levels of an operating system, PTFs, or both.

- Consult the appropriate 3906DEVICE, 2964DEVICE, 2827DEVICE, 2828DEVICE, 2817DEVICE, and 2818DEVICE Preventive Service Planning (PSP) buckets before implementation.
- Hardware Configuration Manager (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 both the physical and the logical aspects of a mainframe's hardware configuration.
- CHPID Mapping Tool (CMT): The CMT helps to map CHPIDs onto PCHIDs based on a 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 has built-in mechanisms to generate a mapping according to customized I/O performance groups. Additional enhancements are implemented in CMT to support z14 servers. The CMT is available for download from the IBM Resource Link website:

  http://www.ibm.com/servers/resourcelink

5.3 Channel subsystem summary

z14 servers support the channel subsystem features of multiple LCSS, MSS and the channel spanning described in this chapter. Table 5-1 lists an overview of the channel subsystem capabilities of z14 servers.

Table 5-1 z13 CSS overview

Maximum number of CSSs	6
Maximum number of LPARs per CSS	CSS0 - CSS4: 15 CSS5: 10
Maximum number of LPARs per system	85
Maximum number of subchannel sets per CSS	4
Maximum number of subchannels per CSS	255.74 K SS0: 65280 SS1 - SS3: 65535
Maximum number of CHPIDs per CSS	256

6

Cryptographic features

This chapter describes the hardware cryptographic functions that are available on IBM z14 servers. 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 is designed for delivering a transparent and consumable approach that enables extensive (pervasive) encryption of data in flight and at rest, with the ultimate goal of substantially simplifying data security and reducing the costs associated with protecting data while achieving compliance mandates.

This chapter provides a short introduction in the principles of cryptography. It then looks at the implementation of cryptography in the hardware and software architecture of IBM Z servers, followed by a more detailed description of the features that IBM z14 servers offer. The chapter is closed by a summary of the cryptographic features and the software required.

This chapter includes the following sections:

- Cryptography enhancements on IBM z14 servers
- Cryptography on IBM z14 servers
- CP Assist for Cryptographic Functions
- Crypto Express6S
- TKE workstation
- Cryptographic functions comparison
- Cryptographic Operating system support

© Copyright IBM Corp. 2017. All rights reserved.

6.1 Cryptography enhancements on IBM z14 servers

IBM z14 introduces the new PCI Crypto Express6S feature, together with an improved CPACF Coprocessor, 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) have been 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 strength 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 further in this chapter.

IBM z14 servers include both standard cryptographic hardware and optional cryptographic features for flexibility and growth capability. IBM has a long history of providing hardware cryptographic solutions. This history stretches from the development of the Data Encryption Standard (DES) in the 1970s to the Crypto Express tamper-sensing and tamper-responding programmable features. Crypto Express is designed to meet the US Government's highest security rating, which is, FIPS 140-2 Level 4[1], and several other security ratings like the Common Criteria for Information Technology Security Evaluation, the Payment Card Industry (PCI) Hardware Security Module (HSM) criteria and the criteria for 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 servers offer.

6.2 Short introduction to cryptography

From the early beginning of human history, when two people were communicating with each other, there has always been the demand to keep certain messages secret, so that a third person is not able to 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 have been used to achieve these objectives, with more or less success. Many procedures and algorithms for encrypting and decrypting data have been developed, increasingly complicated and time consuming.

[1] FIPS 140-2 Security Requirements for Cryptographic Modules

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 which are based on the prime factorization of very large numbers.

Modern cryptography involves the following purposes for protecting information:

- Data protection: The protection of data usually is the main concept associated with cryptography. Only authorized persons should be able to read the message or to get information about it. Data is encrypted by using a known algorithm and secret keys such that the intended party can de-scramble the data, but an interloper cannot. This idea is also referred to as *confidentiality*.
- Authentication (identity validation): This is the process of deciding 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 has been received is identical to what has been sent. It must proven that the data is complete and has not been altered during the moment it has been transmitted (by the sender) and the moment it has been 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 has 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 itself. The goal is to keep the system secure, manageable, and productive.

The basic method data protection is to encrypt and decrypt it, while for authentication, integrity and non-repudiation, hash algorithms, message authentication codes (MACs), digital signatures, and certificates are used.

When encrypting a message, the sender transforms the clear text into a secret text. Doing so requires two main elements: An *algorithm* and a *key*.

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, and that decryption of a ciphered message only results in the original clear message when the correct key is used. Therefore, the receiver of a ciphered message must know which algorithm and which key must be used to decrypt the message.

6.2.2 Kerckhoffs' principle

In modern cryptography, the algorithm is published and known to everyone, whereas the keys are kept secret. This 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. So, the secure management of keys is the primal task of modern cryptographic systems.

The reasons for adhering to Kerckhoffs' Principle are obvious:

- It is much more difficult to keep an algorithm secret than a key.
- It is harder to exchange a compromised algorithm than to exchange a compromised key.
- Secret algorithms can be reconstructed by reverse engineering of software or hardware implementations.
- Errors in public algorithms can generally be found more easily, when many experts look into it.
- In history, most secret encryption methods have proved to be weak and inadequate.
- When using a secret encryption method, it is possible that a back door has been built in.
- If an algorithm is public, many experts can form an opinion about it, and the method can be more thoroughly investigated for potential weaknesses and vulnerabilities.

6.2.3 Keys

The keys that are used for the cryptographic algorithms usually are sequences of numbers and characters, but can also be any other sequence of bits. The length of a key influences the security (strength) of the cryptographic method. The longer the used key, the harder it is to compromise a cryptographic algorithm. For example, algorithm the DES (symmetric key) algorithm uses keys with a length of 56 bits, Triple-DES (TDES) uses keys with a length of 112 bits, the 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.

As previously mentioned, in modern cryptography keys must be kept secret. Depending on the effort that is made to protect the key, keys are classified into three 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, so under certain circumstances the key can be exposed to someone accessing this memory area. This is a risk that must be considered when using clear keys. However, there are many applications where this risk can be accepted. For example, the transaction security for the (widely used) encryption methods Secure Sockets Layer (SSL) and Transport Layer Security (TLS) is based on clear keys.
- The value of a *protected key* is only stored in clear in memory areas that cannot be read by applications or users. The key value does not exist outside of the physical hardware, although the hardware might not be tamper-resistant. The principle of protected keys is unique to IBM Z servers and is explained in more detail in 6.4.2, "CPACF protected key" on page 205.
- For a *secure key*, the key value does not exist in clear outside of a special hardware device, called a hardware security module (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 that of clear keys, as shown in Figure 6-1.

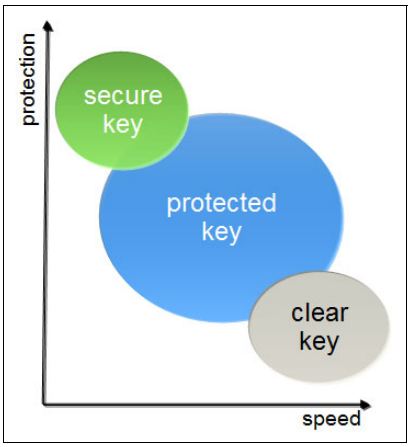


Figure 6-1 Three levels of protection with three levels of speed

6.2.4 Algorithms

The 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 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 huge amounts of data, even if they are not run on specialized hardware. The disadvantage is that the key must be known by both sender and receiver of the messages, 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 the already mentioned DES, TDES, and AES.

- *Asymmetric algorithms* use two distinct but related keys, the *public key* and the *private key*. As the names imply, the private key must be kept secret, 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 only be undone by the other key. For instance, data encrypted by the public key can only be decrypted by the associated private key, and vice versa. Unlike symmetric algorithms, which use distinct functions for encryption and decryption, there is only one function in asymmetric algorithms. Depending on the values passed to this function, it either encrypts or decrypts the data. Asymmetric algorithms are also known as *public key algorithms*.

  Asymmetric algorithms use complex calculations and are relatively slow (about 100 - 1000 times slower than symmetric algorithms), and therefore 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 the already mentioned RSA, and also the elliptic curve algorithms.

- Additionally there are *one-way algorithms* which are, strictly speaking, not cryptographic functions at all. They do not use keys, and they can only scramble data. They cannot de-scramble it. These algorithms are used extensively within cryptographic procedures for digital signing and tend to be developed and governed 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 servers

In principle, cryptographic algorithms can run on processor hardware. But these workloads are compute intensive, and the handling of secure keys also requires special hardware protection. Therefore, IBM Z servers offer several cryptographic hardware features, specialized to meet the requirements for cryptographic workload. Figure 6-2 shows the cryptographic hardware that is supported on IBM z14 servers. All these features are described in more detail further in this chapter.

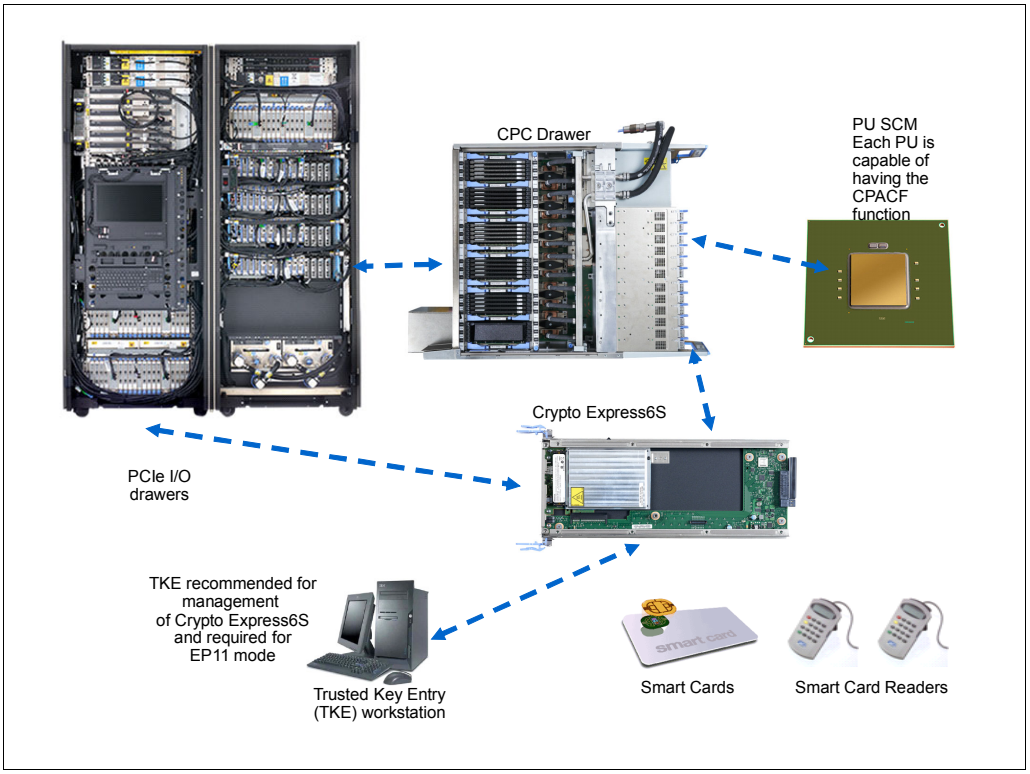


Figure 6-2 Cryptographic hardware supported in IBM z14 servers

Implemented in every processor unit (PU or core) \ in a central processor complex (CPC) is a cryptographic coprocessor that can be used for cryptographic algorithms using clear keys or protected keys. For more information, see 6.4, “CP Assist for Cryptographic Functions” on page 203.

The Crypto Express6S card is a Hardware Security Module (HSM) that is placed in the PCIe I/O drawer of z14 servers. It also supports cryptographic algorithms using secret keys. This feature is described in more detail in 6.5, “Crypto Express6S” on page 207.

Finally, for entering keys in a secure way into the Crypto Express6S HSM, a Trusted Key Entry (TKE) workstation is required, usually also equipped with smart card readers. For more information, see 6.6, “TKE workstation” on page 218.

Table 6-1 lists the feature codes and describes the purpose of the cryptographic hardware features available for IBM z14 servers.

Table 6-1 Cryptographic features for IBM z14 servers

Feature code	Description
3863	CP Assist for Cryptographic Function (CPACF) enablement: This feature is a prerequisite to use CPACF (except for SHA-1, SHA-224, SHA-256, SHA-384, and SHA-512) and the Crypto Express6S and Crypto Express5S feature.
0893	Crypto Express6S card: A maximum of 16 features can be ordered (minimum of two features). This is an optional feature, each feature of which contains one PCI Express cryptographic adapter (adjunct processor). This feature is supported only in z14.
0890	Crypto Express5S card: This feature is available as a carry forward MES from z13. The maximum supported number of Crypto Express5S and Crypto Express6S is 16 in total. This is an optional feature, each feature of which contains one PCI Express cryptographic adapter (adjunct processor). This feature is supported only in z14, z13, and z13s servers.
0086	Trusted Key Entry (TKE) tower workstation: A TKE provides basic key management (key identification, exchange, separation, update, and backup) and security administration. It is optional for running a Crypto Express5S card in CCA mode and required for running it in EP11 mode. The TKE workstation has one Ethernet port, and supports connectivity to an Ethernet local area network (LAN) operating at 10, 100, or 1000 Mbps. Up to 10 features per z14 server can be ordered.
0085	Trusted Key Entry (TKE) rack-mounted workstation: The rack-mounted version of the TKE, which needs a customer provided standard 19-inch rack. It comes with a 1u TKE unit and a 1u console tray (screen, keyboard, and pointing device). When using smart card readers, an additional customer provided tray is needed. Up to 10 features per z13 server can be ordered.
0879	TKE 9.0 Licensed Internal Code (LIC): 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.
0891	TKE Smart Card Reader: Access to information in the smart card is protected by a PIN. One feature code includes two smart card readers, two cables to connect to the TKE workstation, and 20 smart cards.
0892	TKE additional 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 features in a z14, TKE workstation with the TKE 9.0 LIC is required. For more information, see 6.6, “TKE workstation” on page 218.

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 servers, the application must use an application programming interface (API) provided by the operating system. In z/OS, the Integrated Cryptographic Service Facility (ICSF) provides the APIs and is managing the access to the cryptographic devices, as shown in Figure 6-3.

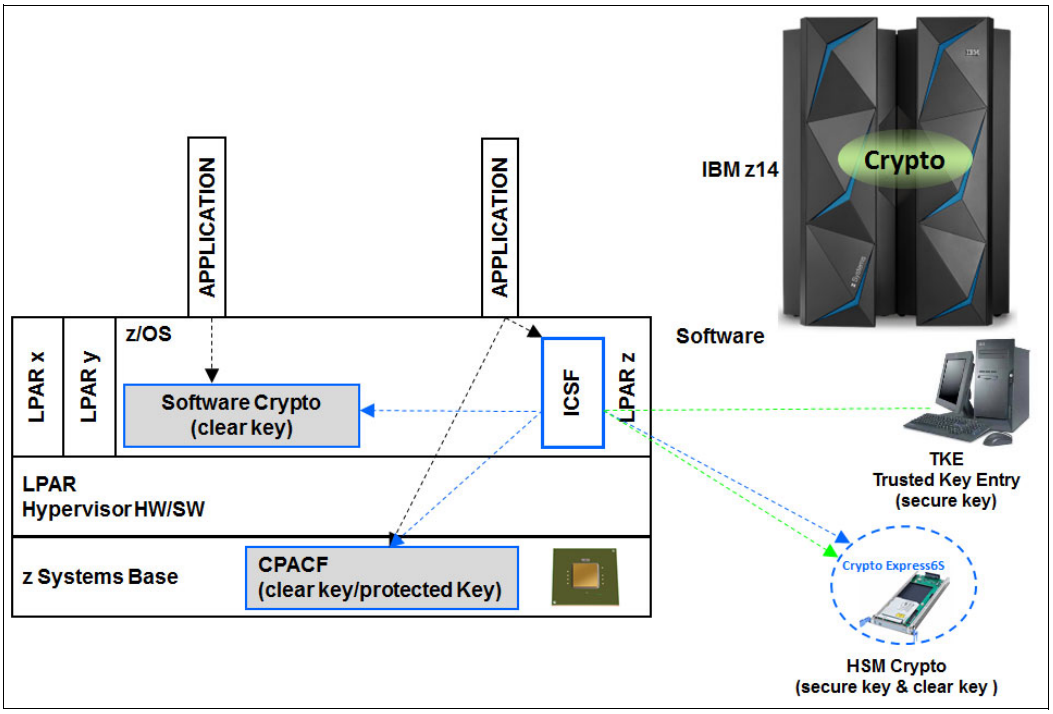


Figure 6-3 z13 cryptographic support in z/OS

ICSF is a software component of z/OS. ICSF works with the hardware cryptographic features and the Security Server (IBM Resource Access Control Facility (RACF®) element) to provide secure, high-speed cryptographic services in the z/OS environment. ICSF provides the APIs by which applications request the cryptographic services, as well from the CPACF and the Crypto Express6S feature. ICSF transparently routes application requests for cryptographic services to one of the integrated cryptographic engines, either CPACF or a Crypto Express6S card, depending on performance or requested cryptographic function. ICSF is also the means by which the secure Crypto Express6S features are loaded with master key values, allowing the hardware features to be used by applications. The cryptographic hardware that is installed in z14 servers determines the cryptographic features and services available to the applications.

The exploiters 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 being generated for a cryptographic hardware feature.

6.4 CP Assist for Cryptographic Functions

As already mentioned, attached to every PU on an SCM in a CPC of a z14 server are two independent engines, one for compression and one for cryptographic purposes, as shown in Figure 6-4. This cryptographic coprocessor, called the CPACF, is not qualified as an HSM, and therefore 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.

Central Processor Assist for Cryptographic Function (CPACF), now supports pervasive encryption. Simple policy controls allow business to enable encryption to protect data in mission-critical databases without needing to stop the database or recreate database objects. Database administrators can leverage z/OS Dataset Encryption, z/OS Coupling Facility Encryption, z/VM encrypted hypervisor paging, and z/TPF transparent database encryption, which exploit performance enhancements in the hardware.

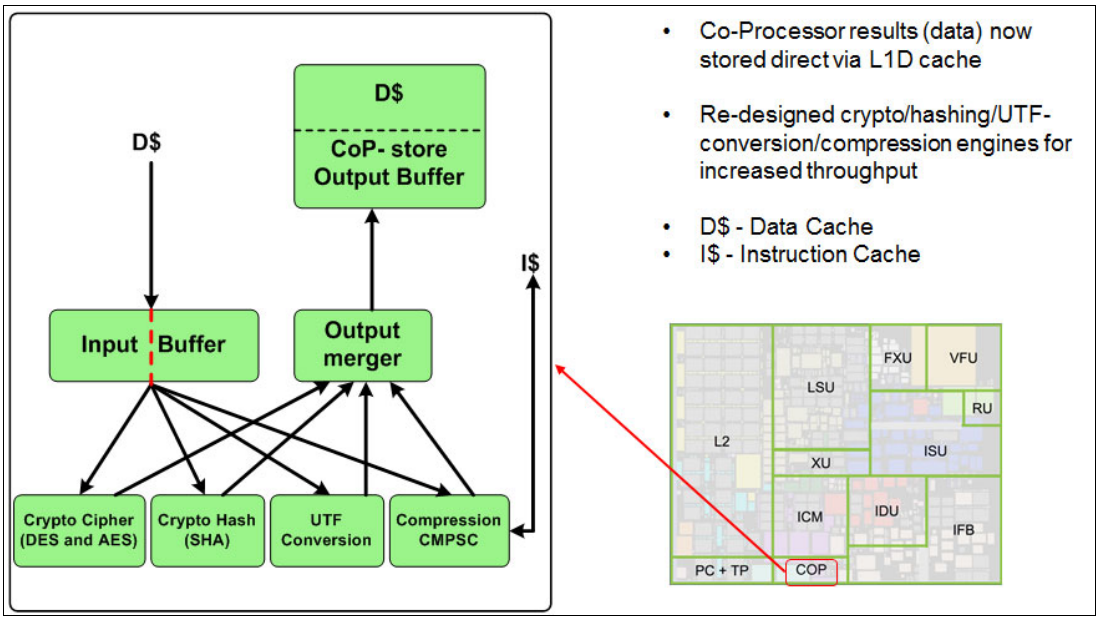


Figure 6-4 The cryptographic coprocessor CPACF

The CPACF offers a set of symmetric cryptographic functions that enhance the encryption and decryption performance of clear key operations. These functions are for SSL, virtual private network (VPN), and data-storing applications that do not require FIPS 140-2 Level 4 security.

CPACF is designed to facilitate the privacy of cryptographic key material when used for data encryption through key wrapping implementation. It ensures that key material is not visible to applications or operating systems during encryption operations. For more information, see 6.4.2, “CPACF protected key” on page 205

The CPACF feature provides hardware acceleration for DES, Triple-DES, AES-128, AES-192, AES-256 (all for clear and protected keys), SHA-1, SHA-256, SHA-384, SHA-512, PRNG, DRNG and TRNG cryptographic services. It provides high-performance hardware encryption, decryption, hashing, and random number generation support. The following instructions support the cryptographic assist function:

KMAC	Compute Message Authentic Code
KM	Cipher Message
KMC	Cipher Message with Chaining
KMF	Cipher Message with CFB
KMCTR	Cipher Message with Counter
KMO	Cipher Message with OFB
KIMD	Compute Intermediate Message Digest
KLMD	Compute Last Message Digest
PCKMO	Provide Cryptographic Key Management Operation

These functions are provided as problem-state z/Architecture instructions that are directly available to application programs. These instructions are known as Message-Security Assist (MSA). When enabled, the CPACF runs at processor speed for every CP, IFL, and zIIP. For more information about MSA instructions, see *z/Architecture Principles of Operation*, SA22-7832.

The CPACF must be explicitly enabled by using an enablement feature (feature code 3863), which is available for no additional 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 running on Linux on z Systems can also start CPACF synchronous functions.

The CPACF coprocessor in z14 servers is redesigned for improved performance compared to the z13, depending on the function that is being used. These tools that might benefit from the throughput improvements are:

- DB2/IMS encryption tool
- DB2 built-in encryption
- z/OS Communication Server: IPsec/IKE/AT-TLS
- z/OS System SSL
- z/OS Network Authentication Service (Kerberos)
- DFDSS Volume encryption
- z/OS Java SDK
- z/OS Encryption Facility
- Linux on z Systems: Kernel, openSSL, openCryptoki, and GSKIT

The z14 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 servers offer the following synchronous functions:

- Data encryption and decryption algorithms for data privacy and confidentially:
  - 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 and new for z14 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, enabled by a single HW instruction

For the SHA hashing algorithms and the random number generation algorithms, only clear keys are used. For the symmetric encryption/decryption DES and AES algorithms as well as clear keys, protected keys can also be used. Protected keys require a Crypto Express6S or a Crypto Express5S card running in CCA mode. For more information, see 6.5.2, “Crypto Express6S as a CCA coprocessor” on page 210.

The hashing algorithms SHA-1, SHA-2 and SHA-3 support for SHA-224, SHA-256, SHA-384, and SHA-512, are included enabled on all servers and do not require the CPACF enablement feature. For all other algorithms, the no-charge CPACF enablement feature (FC 3863) is required.

The CPACF functions are supported by z/OS, z/VM, z/VSE, zTPF, and Linux on z Systems.

6.4.2 CPACF protected key

z14 servers support the protected key implementation. Since PCIXCC[2] deployment, secure keys are processed on the PCI-X and PCIe cards. This process requires an asynchronous operation to move the data and keys from the general-purpose central processor (CP) to the crypto cards. Clear keys process faster than secure keys because the process is done synchronously on the CPACF. Protected keys blend the security of Crypto 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.

[2] IBM 4764 PCI-X cryptographic coprocessor

Wrapping keys are generated during the clear reset each time an LPAR is activated or reset. There is no customizable option available at Support Element (SE) or Hardware Management Console (HMC) that permits or avoids the wrapping key generation. Figure 6-5 shows this function flow.

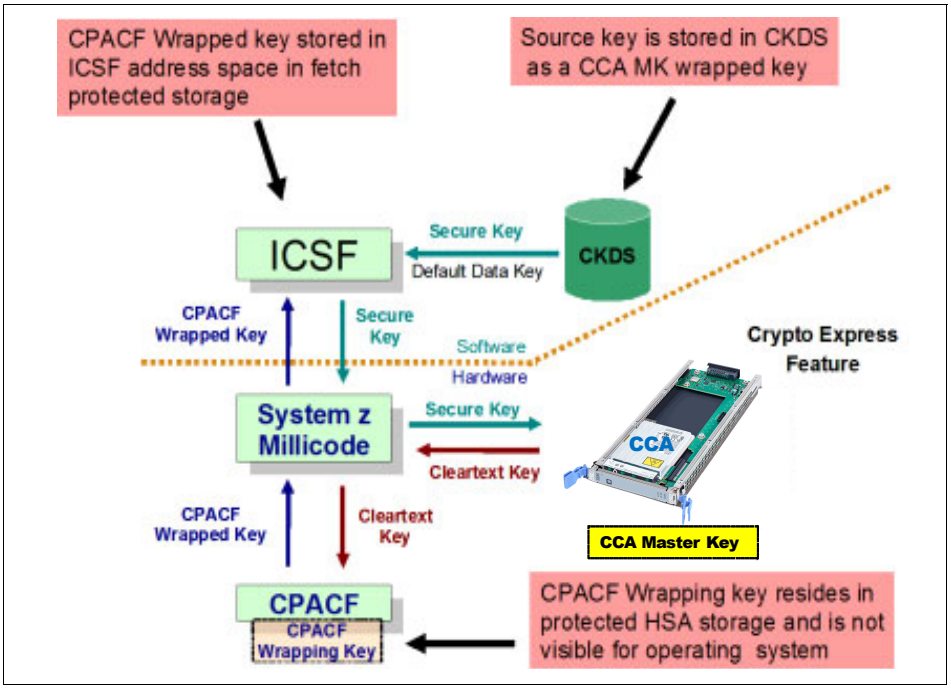


Figure 6-5 CPACF key wrapping

If a Crypto 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 using 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 described in Figure 6-5 considers a secure key as being the source of a protected key.

The source key in this case is already stored in the ICSF Cryptographic Key Data Set (CKDS) as a secure key, which has been encrypted under the master key. This secure key is sent to CEX6C or CEX5C to be deciphered, and sent to the CPACF in clear text. At the CPACF, the key is wrapped under the LPAR wrapping key, and is then returned to ICSF. After the key is wrapped, ICSF can keep the protected value in memory. It then passes it to the CPACF, where the key is unwrapped for each encryption/decryption operation.

The protected key is designed to provide substantial throughput improvements for a large volume of data encryption and low latency for encryption of small blocks of data. A high performance secure key solution, also known as a protected key solution, requires the ICSF HCR7770 as a minimum release.

6.5 Crypto Express6S

The Crypto Express6S feature (FC #0893) is an optional feature that is exclusive to z14 servers. Each feature has one PCIe cryptographic adapter. The Crypto Express6S (CEX6S) feature occupies one I/O slot in a z14 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 can be in one of these 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 210.
- Secure IBM Enterprise PKCS #11 (EP11) coprocessor (CEX6P) implements an industry-standardized set of services that adhere to the PKCS #11 specification V2.20 and more recent amendments. It was designed for extended FIPS and Common Criteria evaluations to meet public sector requirements. This new cryptographic coprocessor mode introduced the PKCS #11 secure key function. For more information, see 6.5.3, “Crypto Express6S as an EP11 coprocessor” on page 214.

  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 215.

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 card secrets. The exception is when you are switching from Secure CCA to accelerator, and vice versa.

The Crypto Express6S feature has been released for enhanced cryptographic performance. Clients who migrated to variable-length AES key tokens cannot take advantage of faster encryption speeds 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 allows for faster AES encryption speeds, when using variable-length tokens, while maintaining strong levels of security

The Crypto Express6S feature does not have external ports and does not use optical fiber or other cables. It does not use channel path identifiers (CHPIDs), but requires one slot in the PCIe I/O drawer and one physical channel ID (PCHID) for each PCIe cryptographic adapter. Removal of the feature or card *zeroizes* its content. Access to the PCIe cryptographic adapter is controlled through the setup in the image profiles on the SE.

Adapter: Although PCIe cryptographic adapters have no CHPID type and are not identified as external channels, all logical partitions (LPARs) in all channel subsystems have access to the adapter. In z14 servers, there are up to 85 LPARs per adapter. Having access to the adapter requires a setup in the image profile for each partition. The adapter must be in the candidate list.

Each z14 server 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 carried forward from a z13 vie an MES. Table 6-2 shows configuration information for Crypto Express6S.

Table 6-2 Crypto Express6S features

Feature	Quantity
Minimum number of orderable features for each server[a]	2
Order increment above two features	1
Maximum number of features for each server (CEX6S and CEX5S in total)	16
Number of PCIe cryptographic adapters for each feature (coprocessor or accelerator)	1
Number of cryptographic domains for each PCIe adapter[b]	85

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 CPs 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:

- 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 85). For example, you might define a configuration for backup situations. However, only one of the LPARs can be active at a time.

Fore more information, see 6.5.5, “Management of Crypto Express6S” on page 215.

6.5.1 Cryptographic asynchronous functions

The optional PCIe cryptographic coprocessors Crypto Express6S provides asynchronous cryptographic functions to z14 servers. Over 300 Cryptographic algorithms and modes are supported, including:

- 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 have been replaced by the AES as the official US standard, but it is still used in the industry, together 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, 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-3[3] (224, 256, 384, 512), HMAC: The Secure Hash Algorithm (SHA-1 and the enhanced SHA-2 or SHA-3 for different block sizes) as well as the older message-digest (MD5) algorithm and the advanced keyed-hash method authentication code (HMAC) are used for verifying both the data integrity and the authentication of a message.
- Visa Format Preserving Encryption (VFPE): A method of encryption where the resulting cipher text has the same form as the input clear text, developed for use with credit cards.
- RSA (512, 1024, 2048, 4096): RSA was been published in 1977, and was named with the initial letters of the surnames of its authors Ron Rivest, Adi Shamir, and Leonard Adleman. It is 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, 521 Prime/NIST): ECC is a family of asymmetric cryptographic algorithms based on the algebraic structure of elliptic curves. ECC can be used for encryption, pseudo-random number generation, and digital certificates. The Elliptic Curve Digital Signature Algorithm (ECDSA) Prime/NIST method is used for ECC digital signatures, which are recommended for government use by NIST.
- ECDSA (160, 192, 224, 256, 320, 384, 512 BrainPool): ECC Brainpool is a workgroup of companies and institutions that collaborate on developing ECC algorithms. The ECDSA algorithms that are recommended by this group are supported.
- ECDH (192, 224, 256, 384, 521 Prime/NIST): Elliptic Curve Diffie-Hellman (ECDH) is an asymmetric protocol used for key agreement between two parties using ECC-based private keys. The recommendations by NIST are supported.
- ECDH (160, 192, 224, 256, 320, 384, 512 BrainPool): ECDH according to the Brainpool recommendations.
- Montgomery Modular Math Engine: The Montgomery Modular Math Engine is a method for fast modular multiplication. Many crypto systems like RSA and Diffie-Hellman key Exchange can use this method.
- RNG (Random Number Generator): The generation of random numbers for cryptographic key generation is supported.(PRNG, DRNG, and TRNG)
- PNG (Prime Number Generator): The generation of prime numbers is also supported.

[3] SHA-3 was standardized by NIST in 2015. SHA-2 is still acceptable and there is no indication that SHA-2 is vulnerable or that SHA-3 is more or less vulnerable than SHA-2.

- 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 them can also run with a clear key on the CPACF. Many standards are only supported when the Crypto Express6S card is running in CCA mode. Others are supported only when the card is running in EP11 mode. The three modes for the Crypto Express6S card are further described in the following sections. A summary which algorithm is supported in which mode is shown in 6.7, “Cryptographic functions comparison” on page 222.

6.5.2 Crypto Express6S as a CCA coprocessor

A Crypto Express6S card running in CCA mode supports IBM CCA. CCA is both an architecture and a set of APIs. It provides cryptographic algorithms and 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 these 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

UDX allows the user to add customized operations to a cryptographic coprocessor. User-Defined Extensions to the CCA support customized operations that run within the Crypto Express features when defined as a coprocessor.

UDX is supported under a special contract through an IBM or approved third-party service offering. The Crypto Cards website directs your request to an IBM Global Services location for your geographic location. A special contract is negotiated between IBM Global Services and you 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 servers is tied to specific versions of the CCA card 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 servers, up to four UDX files can be imported. These files can be imported only from a DVD. The UDX configuration window is updated to include a **Reset to IBM Default** button.

Consideration: CCA has a new code level starting with z13 servers, and the UDX clients require a new UDX.

On z14 servers, the crypto Express6S card is delivered with CCA Level 5.2 firmware. A new set of cryptographic functions and callable services are provided by the IBM CCA LIC to enhance the functions that secure financial transactions and keys:

- Greater than 16 domains support up to 85 LPARs on z14 servers.
- Payment Card Industry (PCI) PIN Transaction Security (PTS) Hardware Security Module (HSM) Certification exclusive to CEX6S and z14
- Visa Format Preserving Encryption (VFPE) support, introduced to z13 or z13s servers.
- AES PIN support for the German banking industry
- PKA Translate UDX function into CCA
- Verb Algorithm Currency

Greater than 16 domains support

z14 servers have support for up to 85 LPARs. The IBM Z crypto architecture was designed to support 16 domains, which matched the LPAR maximum at the time. Before z13 servers, in customer environments where the number of LPARs was larger than 16, crypto workload separation could be complex. These customers had to map a large set of LPARs to a small set of crypto domains.

Now in z14 servers, with the adjunct processor (AP) extended addressing (APXA) facility that is installed, the IBM Z crypto architecture can support up to 256 domains in an AP. As such, the Crypto Express cards are enhanced to handle 256 domains, and the IBM Z firmware provides up to 85 domains to customers (to match the current LPAR maximum). Customers have the flexibility of mapping individual LPARs to unique crypto domains or continuing to share crypto domains across LPARs.

Here are the requirements to support 85 domains:

- Hardware requirements:
  - z14 servers and Crypto Express6S (or Crypto Express5S)
- Operating systems requirements:
  - z/OS all functions require ICSF WD17 (HCR77C1) for exploitation unless otherwise noted. WD17 supports z/OS V2R1, V2R2, and V2R3.
  - z/VM V6.3 and Version 6.4 with PTFs for guest use

Payment Card Industry (PCI) - Hardware Security Module (HSM) Certification

PCI standards are developed to help ensure security in the payment card industry. PCI itself defines their standards as a set of security standards designed to ensure that all companies that accept, process, store, or transmit credit card information maintain a secure environment.

Compliance with the PCI-HSM standard has a great deal of value for customers, particularly those who are in the banking and finance industry. There are two fundamental reasons that this certification is important to clients.

1. Compliance is increasingly becoming mandatory.
2. The requirements in PCI-HSM make the system more secure.

Let's look at each of these in a little more detail.

Industry requirements for PCI-HSM compliance

The PCI organization itself has no power to require compliance with its standards. Compliance with PCI standards is enforced by the payment card brands: 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 have the ability to impose requirements on you if you want to process their cards. One set of requirements they have been increasingly enforcing are 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 PCI-DSS, the PCI Data Security Standard. Some of the other standards were written or required later, and PCI-HSM is one of the last ones to be developed. In addition, the standards themselves have been 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 themselves 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 will eventually have to comply with all of them.

Improved security through use of PCI-HSM

The main reason PCI-HSM was developed was to improve security in payment card systems. It imposes requirements in key management, HSM API functions, and device physical security, and controls during manufacturing and delivery, device administration, and a number of 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 have to be updated because they used some of the things that are now prohibited. While this 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 both the institution and its clients.

The requirements for using PCI-HSM are:

- Hardware requirements:

z14 servers and Crypto Express6S

- Operating systems requirements:
  - z/OS - ICSF WD17 (HCR77C1) for exploitation unless otherwise noted. WD17 supports z/OS V2R1, V2R2, and V2R3
  - z/VM V6.3 and Version 6.4 with PTFs for guest use

Visa Format Preserving Encryption

VFPE refers to a method of encryption where the resulting cipher text has the same form as the input clear text. The form of the text can vary according to use and application. One of the classic examples is a 16-digit credit card number. After using VFPE to encrypt a credit card number, the resulting cipher text is another 16-digit number. This helps legacy databases contain encrypted data of sensitive fields without having to restructure the database or applications.

VFPE allows customers to add encryption to their applications in such a way that the encrypted data can flow through their systems without requiring a massive redesign of their application. In the previous example, if the credit card number is VFPE-encrypted at the point of entry, the cipher text still behaves like a credit card number. It can flow through business

logic until it meets a back-end transaction server that can VFPE-decrypt it to get the original credit card number to process the transaction.

Here are the FPE requirements:

- Hardware requirements:
  - z14 servers and Crypto Express 6S (or Crypto Express5S with CCA V5.2 firmware).
- Operating systems requirements:
  - z/OS - all functions require ICSF WD17 (HCR77C1) for exploitation 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 V6.3 and Version 6.4 with PTFs for guest use

AES PIN support for the German banking industry

The German banking industry organization, DK, has defined a new set of PIN processing functions to be used on the internal systems of banks and their servers. CCA is designed to support the functions that are essential to those parts of the German banking industry that are governed by DK requirements. The functions include key management support for new AES key types, AES key derivation support, and several DK-specific PIN and administrative functions.

This support includes PIN method APIs, PIN administration APIs, new key management verbs, and new access control points support that is needed for DK-defined functions.

Here are the requirements for AES PIN support:

- Hardware requirements:
  - z13s servers and Crypto Express5S with CCA V5.2 firmware
  - z13 servers and Crypto Express5S with CCA V5.0 or later firmware
  - zEC12 or zBC12 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) for exploitation 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

PKA Translate UDX function into CCA

UDX is custom code that allows the client to add unique operations/extensions to the CCA firmware. Certain UDX functions are integrated into the base CCA code over time to accomplish the following tasks:

- Removes headaches and pain points that are associated with UDX management and currency.
- Popular UDX functions are made available to a wider audience to encourage adoption.

UDX is integrated into the base CCA code to support translating an external RSA 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 either CBC or ECB mode. In addition, AES CMAC support is delivered.

Here are the requirements for this function:

- Hardware requirements:
  - z14 servers and Crypto Express6S (or Crypto Express5S with CCA V5.2 firmware)
  - z13s servers and Crypto Express5S with CCA V5.2 firmware
  - z13 servers and Crypto Express5S with CCA V5.0 or later firmware
- Operating systems requirements:
  - z/OS - ICSF WD17 (HCR77C1) for exploitation 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

Note: Older IBM Z servers and operating systems are also supported, but out of the scope of this redbook.

Verb Algorithm Currency

Verb Algorithm Currency is a collection of CCA verb enhancements related to customer requirements, with the intent of maintaining currency with cryptographic algorithms and standards. It is intended for customers who want to maintain the 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

Here are the requirements for this function:

- Hardware requirements:
  - z14 servers and Crypto Express6S (or Crypto Express5S with CCA V5.2 firmware)
  - z13s or z13 servers and Crypto Express5S with CCA V5.2 firmware
- Software requirements:
  - z/OS - ICSF WD17 (HCR77C1) for exploitation unless otherwise noted. WD17 supports z/OS V2R1, V2R2, and V2R3.
  - z/VM 6.3, and 6.4 with PTFs for guest exploitation

6.5.3 Crypto Express6S as an EP11 coprocessor

A Crypto Express6S card that is configured in Secure IBM Enterprise PKCS #11 (EP11) coprocessor mode provides PKCS #11 secure key support for public sector requirements. Before EP11, the ICSF PKCS #11 implementation supported only clear keys. In EP11, keys can now be generated and securely wrapped under the EP11 Master Key. The secure keys never leave the secure coprocessor boundary unencrypted.

The secure IBM Enterprise PKCS #11 (EP11) coprocessor runs the following tasks:

- Encrypt and decrypt (AES, DES, TDES, and RSA)
- Sign and verify (DSA, RSA, and ECDSA)
- Generate keys and key pairs (DES, AES, DSA, ECC, and RSA)
- HMAC (SHA1, SHA2 or SHA3 [SHA224, SHA256, SHA384, and SHA512])
- Digest (SHA1, SHS2 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.

6.5.4 Crypto Express6S as an accelerator

A Crypto Express6S card running in accelerator mode supports only RSA clear key and SSL Acceleration. A request is processed fully in hardware. The Crypto Express accelerator is a coprocessor that is reconfigured by the installation process so that it uses only a subset of the coprocessor functions at a higher speed. Reconfiguration is disruptive to coprocessor and accelerator operations. The coprocessor or accelerator must be deactivated before you begin the reconfiguration.

FIPS 140-2 certification is not relevant to the accelerator because it operates with clear keys only. The function extension capability through UDX is not available to the accelerator.

The functions that remain available when the Crypto 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 bits to 4,096 bits, in the Modulus-Exponent (ME) and Chinese Remainder Theorem (CRT) formats.

6.5.5 Management of Crypto Express6S

With zEC12 and older servers, each cryptographic coprocessor has 16 physical sets of registers or queue registers. Starting with z13 servers, this number is raised to 85. This increase corresponds to the maximum number of LPARs running on a z14 server, which is also 85. Each of these 85 sets belongs to a domain as follows:

- A cryptographic domain index, in the range of 0 - 84, is allocated to a logical partition in its image profile. The same domain must also be allocated to the ICSF instance running in the logical partition that uses the Options data set.
- Each ICSF instance accesses only the Master Keys or queue registers corresponding to the domain number specified in the logical partition image profile at the SE and in its Options data set. Each ICSF instance sees a logical cryptographic coprocessor that consists of the physical cryptographic engine and the unique set of registers (the domain) allocated to this logical partition

The installation of CP Assist for Cryptographic Functions (CPACF) DES/TDES enablement, Feature Code #3863, is required to use the Crypto Express5S feature.

Each Crypto Express5S feature contains one PCI-X adapter. The adapter can be in the following configurations:

- IBM Enterprise Common Cryptographic Architecture (CCA) Coprocessor (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 EP11coprocessor requires a Trusted Key Entry (TKE) tower workstation (FC #0086) or a TKE rack-mounted workstation (FC #0085) with TKE 9.0 (FC #0879) LIC.

The Crypto Express5S 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 card, 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, either from the HMC or from the SE. Modify the cryptographic initial definition from the Crypto option in the image profile, as shown in Figure 6-6.

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.

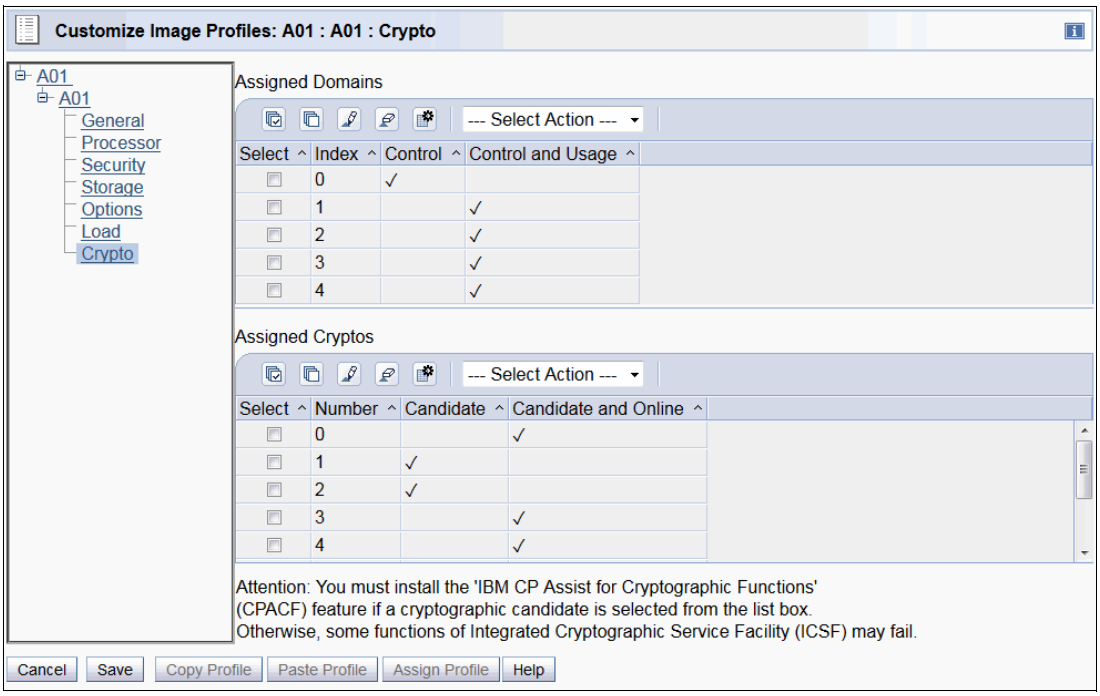


Figure 6-6 Customize Image Profiles: Crypto

The cryptographic resource definitions have the following meanings:

Control Domain Identifies the cryptographic coprocessor domains that can be administered from this logical partition if it is being set

	up as the TCP/IP host for the TKE. If you are setting up the host TCP/IP in this logical partition to communicate with the TKE, the partition is used as a path to other domains' Master Keys. Indicate all the control domains that you want to access (including this partition's own control domain) from this partition.
Control and Usage Domain	Identifies the cryptographic coprocessor domains that are assigned to the partition for all cryptographic coprocessors that are configured on the partition. The usage domains cannot be removed if they are online. The numbers that are selected must match the domain numbers 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 of which CSS they are defined to. 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 are eligible to 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. Select the CPCs, and click **View LPAR Cryptographic Controls** in the CPC Operational Customization window. The resulting window displays the definition of Usage and Control domain indexes, and PCI Cryptographic candidate and online lists, as shown in Figure 6-7. Information is provided only for active logical partitions.

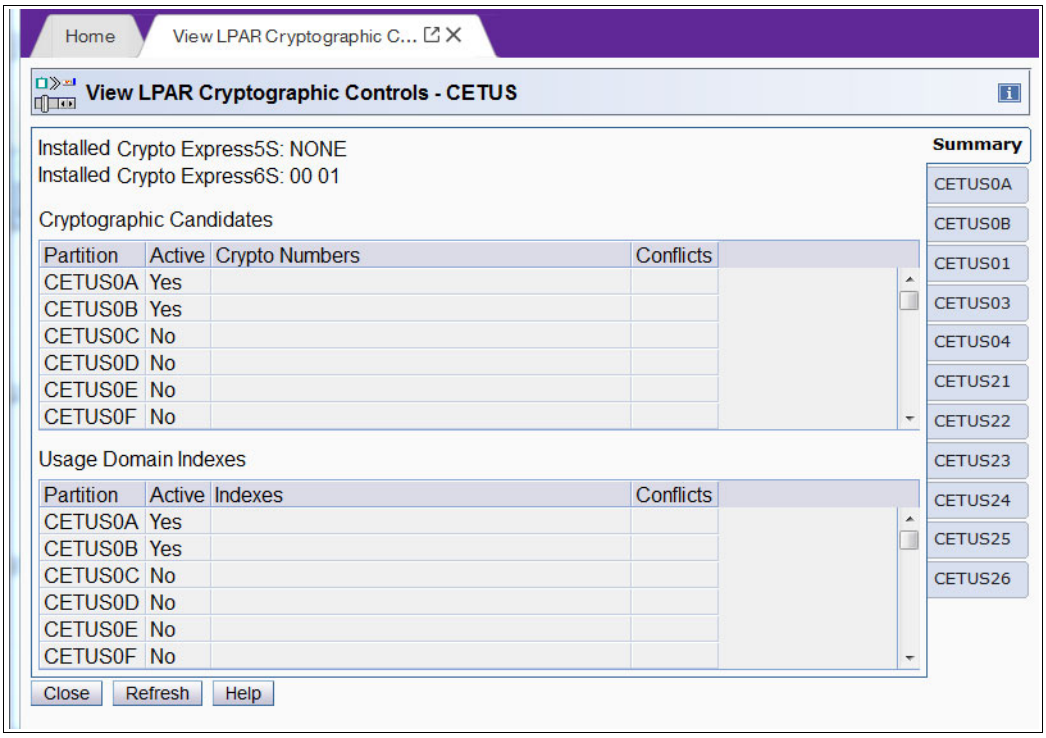


Figure 6-7 SE: View LPAR Cryptographic Controls

Operational changes can be made by using the Change LPAR Cryptographic Controls task from the SE, which reflects the cryptographic definitions in the image profile for the partition. With this function, the cryptographic feature can be added and removed dynamically, without stopping a running operating system.

For more information about the management of Crypto Express6S cards, see the corresponding chapter in *IBM z14 Configuration Setup,* SG24-8460.

6.6 TKE workstation

The TKE workstation is an optional feature that offers key management functions. A TKE tower workstation feature (FC #0085) or a TKE rack-mounted workstation feature (FC #0086) is required for z14 servers to manage the Crypto Express6S feature, or the Crypto Express5S feature, respectively.

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 using 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 server 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 server through a TCP/IP connection. The TKE workstation is available with Ethernet LAN connectivity only. Up to 10 TKE workstations can be ordered. TKE FCs #0085 and #0086 can be used to control the Crypto Express6S or Crypto Express5S cards on z14 servers. They also can be used to control the Crypto Express5S on z13 and z13s servers, and the Crypto cards on older still supported servers.

This are the enhancements for the new TKE 9.0 LIC:

Key material copy to alternate zone

TKE 9.0 allows you to copy key material from smart cards in one TKE zone to smart cards in another zone. You might have old 1024 bit strength TKE zones, and may wish to move/copy the key material in those zones into a new, stronger TKE zone. To use this new feature you create new TKE and/or EP11 smart cards on your TKE 9.0 system. You enroll the new TKE and/or EP11 smart cards in an alternate zone. This allows you to copy smart card content from a smart card enrolled in the alternate 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

The TKE application now has the ability for administrators to create key parts without opening a host. This 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

There is a new TKE Audit Log application available for the Privileged Mode Access ID of AUDITOR. This application provides a new, easy-to-use interface to view the TKE workstation security audit records from the TKE workstation.

Heartbeat audit record

TKE workstations cuts an audit record when the TKE boots or when no audit events have occurred during a client-configured duration of time. The record shows the serial number of the TKE local crypto adapter and indicates if the local crypto adapter has been changed since the last check.

Performance improvements for domain groups

With CCA version 5.3, depending on the size of a domain group, you may experience performance improvements when doing a Load, Set, or Clear operation 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 issued to the module will drop 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 will now have the ability to manage a set of parent X.509 certificates for validating operating X.509 certificates used by applications 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 thefollowing modes:

- Normal Mode
- Imprint Mode
- Compliant Mode

Imprint and compliant mode were added to indirectly and directly meet 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, there is the ability to set the host crypto module's clock. 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 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 which prevent roles from being able to both issue and cosign a command. Refer to the TKE User's Guide for detailed information on how to manage imprint and compliant mode domains.

Setup PCI Environment Wizard

To simplify the management of a compliant domain, the TKE provides a setup wizard that will create a minimum set of forced dual control roles and authorities needed to manage a compliant domain. Refer to the TKE User's Guide for detailed information on how to manage imprint and compliant mode domains.

Tip: For more information about handling a TKE, see the series of instructive video clips at:

`https://www.youtube.com/watch?v=Y8T9rSd-qrQ`

6.6.1 Logical partition, TKE host, and TKE target

If one or more LPARs are configured to use Crypto Express6S 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 either a TKE host or a TKE target. An LPAR with a TCP/IP connection to the TKE is referred to as the *TKE host*. All other partitions are *TKE targets*.

The cryptographic controls that are set for an LPAR through the SE determine whether the workstation is a TKE host or a TKE target.

6.6.2 Optional smart card reader

An optional smart card reader (FC #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 only be used with smart cards that have applets that were loaded from a TKE 8.1 or later. This cards are FIPS certified.
Smart card readers from feature code 0885 or 0891 can carried forward. Existing 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 additional smart cards can be ordered for backup. The additional smart card feature code is FC #0892. When one feature code is ordered, 10 smart cards are shipped. The order increment is 1 - 99 (10 - 990 blank smart cards).

If smart cards with applets not supported by the new smart card reader will be 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.

6.6.3 TKE hardware support and migration information

The new TKE 9.0 LIC (FC #0879) is shipped with any new z14 server. If a new TKE 9.0 is purchased, two versions are available:

- TKE 9.0 tower workstation (FC #0086)
- TKE 9.0 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 the best option for you.

The TKE 9.0 LIC requires the 4768 crypto adapter. The TKE 8.0 and TKE 8.1 workstations can be upgraded to the TKE 9.0 tower workstation with purchasing a 4768 crypto adapter.

The Omnikey Cardman 3821 smart card readers can be carried forward to any TKE 9.0 workstation. Smart cards 45D3398, 74Y0551, and 00JA710 can be used on TKE 9.0

When performing a MES upgrade to a TKE 9.0 installation, the following steps must be completed.

Upgrade from TKE 7.3, TKE 8.0, or TKE 8.1 to TKE 9.0:

1. Save Upgrade Data on 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. Perform the Frame Roll installation to apply Save Upgrade Data (client data) to the TKE 9.1 system
5. Run the TKE Workstation Setup wizard.

For more information about TKE hardware support, see Table 6-3. For some functionality requirements needs to consider. E.g. characterization of an Crypto Express card in EP 11 mode always require 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
	HW Feature Code	0814	0842	0847	0847 or 0097	0085 or 0086
	LICC	0850	0872	0877	0878	0879
	Smart Card Reader	0885	0885	0891	0891	0895
	Smart Card	0884	0884	0892	0892	0892
Server supported	z14			Yes	Yes	Yes
	z13			Yes	Yes	Yes
	z13s			Yes	Yes	Yes
	zEC12	Yes	Yes	Yes	Yes	Yes
	zBC12		Yes	Yes	Yes	Yes
Manage Host Crypto Module	CEC3C (CCA)	Yes	Yes	Yes	Yes	Yes
	CEX4C (CCA)	Yes	Yes	Yes	Yes	Yes
	CEX4P (EP11)	Yes	Yes	Yes	Yes	Yes
	CEX5C (CCA)			Yes	Yes	Yes
	CEX5P (EP11)			Yes	Yes	Yes
	CEX6C (CCA)			Yes	Yes	Yes
	CEX6P (EP11)			Yes	Yes	Yes

Attention: The TKE is unaware of the CPC type where the host crypto module is installed. That is, the TKE does not care whether a Crypto Express is running on a zEC12, zBC12, z13, 13s, or z14 server. Therefore, the LIC can support any CPC where the coprocessor is supported, but the TKE LIC must support the specific crypto module.

6.7 Cryptographic functions comparison

Table 6-4 lists the functions or attributes on z14 servers for the two cryptographic hardware features. In the table, “X” indicates that the function or attribute is supported.

Table 6-4 Cryptographic functions on z14 servers

Functions or attributes	CPACF	CEX6C	CEX6P	CEX6A
Supports z/OS applications using ICSF	X	X	X	X
Supports Linux on z Systems CCA applications	X	X	-	X
Encryption and decryption using secret-key algorithm	-	X	X	-
Provides the highest SSL/TLS handshake performance	-	-	-	X
Supports SSL/TLS functions	X	X	-	X

Functions or attributes	CPACF	CEX6C	CEX6P	CEX6A
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	-
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 performing 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

Functions or attributes	CPACF	CEX6C	CEX6P	CEX6A
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	-
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	-	-

a. To make the addition of the Crypto Express features nondisruptive, the logical partition must be predefined with the appropriate PCI Express cryptographic adapter number. This number must be selected from its candidate list in the partition image profile.
b. This feature is not required for Linux if only RSA clear key operations are used. DES or triple DES encryption requires CPACF to be enabled.
c. This feature is physically present but is not used when configured as an accelerator (clear key only).

6.8 Cryptographic Operating system support

The following section gives an overview of the Operating Systems requirements in relation to cryptographic elements.

Crypto Express6S (#0893) Toleration

Treats Crypto Express6S cryptographic coprocessors and accelerators as Crypto Express5 coprocessors and accelerators, requires at a minimum:

- z/OS V2.3 with PTFs.
- z/OS V2.2 with PTFs.
- z/OS V2.1 with PTFs.
- z/VM V6.4 with PTFs for guest exploitation.
- z/VM V6.3 with PTFs for guest exploitation.
- z/VSE V6.2 with PTFs.
- z/VSE V6.1 with PTFs.
- z/VSE V5.2 with PTFs.

- Linux on z Systems: IBM is working with its Linux distribution partners to provide support via maintenance or future distribution releases for:
  - SLES 12 and SLES 11.
  - HEL 7 and RHEL 6.
  - Ubuntu 16.04 LTS (or higher).
- KVM hypervisor which is offered with the following Linux distributions: SLES-12 SP2 or higher, and Ubuntu 16.04 LTS or higher. For minimal and recommended distribution levels refer to the following website:

  https://www-03.ibm.com/systems/z/os/linux/resources/testedplatforms.html

Crypto Express6S (#0893) support of Visa Format Preserving Encryption

This functionality requires at a minimum:

- 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 V6.4 with PTFs for guest exploitation.
- z/VM V6.3 with PTFs for guest exploitation.
- KVM hypervisor which is offered with the following Linux distributions: SLES-12 SP2 or higher, and Ubuntu 16.04 LTS or higher. For minimal and recommended distribution levels refer to the following website:

  https://www-03.ibm.com/systems/z/os/linux/resources/testedplatforms.html

Crypto Express6S (#0893) support of greater than 16 Domains

The support of more than 16 Domains requires at a minimum:

- 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 V6.4 with PTFs for guest exploitation.
- z/VM V6.3 with PTFs for guest exploitation.
- z/VSE V6.2 with PTFs.
- z/VSE V6.1 with PTFs.
- z/VSE V5.2 with PTFs.
- Linux on z Systems: IBM is working with its Linux distribution partners to provide support via maintenance or future distribution releases for:
  - SLES 12 and SLES 11.
  - RHEL 7 and RHEL 6.
  - Ubuntu 16.04 LTS (or higher).

The software support levels for cryptographic functions are listed in Chapter 7, “Operating System Support” on page 227.

7

Operating System Support

This chapter lists the minimum operating system requirements and support considerations for the IBM z14 servers and their features. It addresses z/OS, z/VM, z/VSE, z/TPF, Linux on z Systems, and KVM for IBM Z. Because this information is subject to change, see the Preventive Service Planning (PSP) bucket 3906DEVICE for the most current information.

Support of z14 functions depends on the operating system, its version, and release.

This chapter includes the following sections:

- Operating systems summary
- Support by operating system
- z14 features and function support summary
- Support by features and functions
  - LPAR Configuration and Management
  - Base CPC Features and Functions
  - Coupling and clustering Features and Functions
  - Storage Connectivity Related Features and Functions
  - Networking Features and Functions
  - Cryptography Features and Functions Support
  - Special-purpose Features and Functions
- z/OS migration considerations
- z/VM migration considerations
- z/VSE migration considerations
- Software licensing
- References

© Copyright IBM Corp. 2017. All rights reserved.

7.1 Operating systems summary

Table 7-1 lists the minimum operating system levels that are required on z14 servers.

End-of-service operating systems: Operating system levels that are no longer in service are not covered in this publication. These older levels might provide support for some features.

<< Simplified prior table (7.1) with the drop of ESA/390 support and added reference to z/Architecture mode box>>

Table 7-1 z14 minimum operating systems requirements

Operating systems[a]	Supported Version and Release on z14[b]
z/OS	V1R13[c]
z/VM	V6R3[d]
z/VSE	V5R1[e]
z/TPF	V1R1
Linux on z Systems	See Table 7-2 on page 231
KVM for IBM Z	V1R1M2[f]
KVM Hypervisor[g]	Offered with the following Linux distributions SLES-12 SP2 or higher, and Ubuntu 16.04 LTS or higher.

a. Only z/Architecture mode is supported. See the box titled “z/Architecture mode”.
b. Service is required. See the box titled “Features”.
c. 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.
d. As announced on February 3, 2015, end of service date for z/VM V6R3 is December 31, 2017.
e. As announced on May 23, 2014, end of service date for z/VSE V5R1 is June 30, 2016.
f. End of Service date is announced as March 31, 2018
g. For minimal and recommended distribution levels refer to the Linux on z Systems website.

z/Architecture mode: As announced on January 14, 2015 with Announcement letter 115-001, beginning with IBM z14, all IBM Z servers will only support operating systems running in z/Architecture mode. This applies to operating systems running native on PR/SM as well as operating systems running as second-level guests. IBM operating systems that run in ESA/390 mode are either no longer in service or only currently available with extended service contracts, and they are not usable on systems beginning with IBM z14. However, IBM z14 does provide ESA/390-compatibility mode, an environment supporting a subset of DAT-off ESA/390 applications in a hybrid architectural mode.

All 24-bit and 31-bit problem state application programs originally written to run on the ESA/390 architecture will be unaffected by this change.

Features: Usage 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, z/VM, z/VSE, and z/TPF subsets of the 3906DEVICE PSP buckets. The PSP buckets are continuously updated, and contain the latest information about maintenance.

- Hardware and software buckets contain installation information, hardware and software service levels, service guidelines, and cross-product dependencies.
- For Linux on z Systems distributions and KVM hypervisors, consult the distributor's support information.
- For KVM for IBM Z, see *Getting Started with KVM for IBM z Systems*, SG24-8332.

7.2 Support by operating system

z14 servers introduce several new functions. This section addresses support of those functions by the current operating systems. Also included are some of the functions that were introduced in previous IBM Z servers and carried forward or enhanced in z14 servers. Features and functions that are available on previous servers but no longer supported by z14 servers have been removed.

See 7.3, "z14 features and function support summary" on page 232 for supported functions based on operating systems. 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 1 is the earliest in-service release that supports z14 servers. Although service support for z/OS Version 1 Release 13 ended in September of 2016, a fee-based extension for defect support (for up to three years) can be obtained by ordering IBM Software Support Services - Service Extension[1] for z/OS 1.13.

z14 capabilities differ depending on the z/OS release. Toleration support is provided on z/OS V1R13. Exploitation support is provided only on z/OS V2R1 and later.

For a list of supported functions and their minimum required support levels, see 7.3, "z14 features and function support summary" on page 232.

7.2.2 z/VM

z/VM V6R3 and z/VM V6R4 provide support that will enable guests to exploit function supported by z/VM on IBM z14, which includes:

- 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-Express6S OSD CHPIDs, FICON Express16S+ FC and FCP CHPIDs, zHyperLink Express, RoCE Express2 features, and Regional Crypto Enablement (RCE).
- Improved memory management

Fulfillment of Statements of Directions:

- *Removal of support for Expanded Storage (XSTORE)*: z/VM V6.3 is the last z/VM release that supported Expanded Storage (XSTORE) for either host or guest use. z14 servers do not support Expanded Storage (XSTORE).
- *Product Delivery of z/VM on DVD/Electronic only*: z/VM V6.3 is the last release of z/VM that is available on tape. Subsequent releases are available only on DVD or electronically.

For a list of supported functions and their minimum required support levels, see 7.3, “z14 features and function support summary” on page 232.

Statements of Directions[a]:

- *Stabilization of z/VM V6.3 support:* IBM z14 is planned to be the last z Systems and IBM Z high-end server and z13s is planned to be the last mid-range z Systems server supported by z/VM V6.3 and the last z Systems and IBM Z servers that will be supported when z/VM V6.3 is running as a guest (second level). z/VM V6.3 will continue to be supported until December 31, 2017, as announced in announcement letter # 915-025.
- *Future z/VM release guest support:* z/VM V6.4 will be the last release supported as a guest of z/VM V6.2 or older releases.
- *Disk-only support for z/VM dumps*: z/VM V6.4 will be 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 support is provided by z/VSE V5R1 and later, with the following considerations:

- z/VSE runs in z/Architecture mode only.
- z/VSE supports 64-bit real and virtual addressing
- z/VSE V5 has an architectural level set (ALS) that requires IBM System z9 or later.
- z/VSE V6 has an ALS that requires IBM System z10 or later.

For a list of supported functions and their minimum required support levels, see 7.3, “z14 features and function support summary” on page 232.

7.2.4 z/TPF

z14 support is provided by z/TPF V1R1 with PTFs. For a list of supported functions and their minimum required support levels, see 7.3, “z14 features and function support summary” on page 232.

7.2.5 Linux on z Systems

Generally, a new machine is transparent to Linux on z Systems. For z14, toleration support is required for:

- IPL in “z Architecture” mode
- Crypto Express6S cards
- RoCE Express cards
- 8-byte LPAR offset

Table 7-2 shows the service levels of SUSE, Red Hat and Ubuntu releases supported at the time this document was created.

Table 7-2 Linux on z Systems distributions

Linux on z Systems distribution[a]	Supported Version and Release on z14[b]
SUSE Linux Enterprise Server	12 SP2
SUSE Linux Enterprise Server	11 SP4
Red Hat RHEL	7.3
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 the latest information about supported Linux distributions on IBM Z servers, see this website:

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

IBM is working with Linux distribution Business Partners to provide further use of selected z14 functions in future Linux on z Systems 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 servers.
- Update any Linux distribution to the latest service level before migrating to z14 servers.
- Adjust the capacity of any z/VM and Linux on z Systems LPAR guests, and z/VM guests, in terms of the number of IFLs and CPs, real or virtual, according to the PU capacity of the z14 servers.

7.2.6 KVM for IBM Z

KVM for IBM Z (KVM for IBM z Systems) is an open source virtualization alternative for IBM Z servers built on Linux and KVM. KVM for IBM Z delivers a Linux-familiar administrator experience that can enable simplified virtualization management and operation.

KVM is a strategic open source component for the IBM Z and IBM LinuxONE platforms. With Announcement Letter 917-04 dated March 7, 2017; IBM is changing how KVM is delivered. KVM will now be offered through our Linux distribution partners for IBM Z and LinuxONE to help simplify the delivery and installation. Linux and KVM will be provided from a single

source, and with KVM being included in the Linux distribution, it should make ordering and installing KVM easier.

KVM hypervisor is now offered with the following Linux distributions:

- SLES-12 SP2 or higher, and
- Ubuntu 16.04 LTS or higher.

For minimal and recommended distribution levels refer to the IBM Z website.

KVM for IBM Z Version 1.1.2 will be the last release delivered by IBM[2]. However, IBM will continue to provide the Linux and KVM enablement for IBM Z and LinuxONE to the open source community, enabling our Linux distribution partners to deliver KVM and Linux together for existing and new IBM Z and IBM LinuxONE hardware.

This change provides below advantages:

- Simplified delivery of open source infrastructure
- Linux and KVM now provided from a single source
- Easier obtaining and installing of KVM
- Distributions include open source cloud infrastructure components
- Simplified enabling of other technologies, such as container management and cloud management.

The last day to order KVM for IBM Z from IBM is August 28, 2017 and the End of Service date is March 31, 2018.

7.3 z14 features and function support summary

The following tables summarize the z14 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 233
- Table 7-4, “Supported base CPC functions for z/VSE, z/TPF and Linux on z Systems” on page 234
- Table 7-5, “Supported coupling and clustering functions for z/OS and z/VM” on page 235
- Table 7-6, “Supported storage connectivity functions for z/OS and z/VM” on page 236
- Table 7-7, “Supported storage connectivity functions for z/VSE, z/TPF and Linux on z Systems” on page 237
- Table 7-8, “Supported network connectivity functions for z/OS and z/VM” on page 239
- Table 7-9, “Supported network connectivity functions for z/VSE, z/TPF and Linux on z Systems” on page 241
- Table 7-10, “Supported cryptography functions for z/OS and z/VM” on page 243
- Table 7-11, “Supported cryptography functions for z/VSE, z/TPF and Linux on z Systems” on page 244
- Table 7-12, “Supported Special-purpose features for z/OS and z/VM” on page 245

Information about Linux on z Systems refers exclusively to the appropriate distributions of SUSE, Red Hat and Ubuntu.

All tables use the following conventions:

Y The function is supported.

[2] https://www-03.ibm.com/systems/z/solutions/virtualization/kvm/

N The function is not supported.
- The function is not applicable to that specific operating system.

Please note that these 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. Therefore, for the current information, see the PSP bucket for 3906DEVICE.

7.3.1 Supported CPC functions

Table 7-3 lists the supported Base CPC Functions or z/OS and z/VM

Table 7-3 Supported Base CPC Functions or z/OS and z/VM

Function[a]	z/OS V2R2	z/OS V2R1	z/OS V1R13	z/VM V6R4	z/VM V6R3
z14 servers	Y	Y	Y	Y	Y
Maximum processor unit (PUs) per system image	170[b]	170[b]	100	64[c]	64[c]
Maximum main storage size[d]	4 TB	4 TB	1 TB	2 TB	1 TB
85 LPARs	Y	Y	Y	Y	Y
Separate LPAR management of PUs	Y	Y	Y	Y	Y
Dynamic PU add	Y	Y	Y	Y	Y
Dynamic LPAR memory upgrade	Y	Y	Y	Y	Y
LPAR group absolute capping	Y	Y	Y	Y	Y
Capacity Provisioning Manager	Y	Y	Y	N	N
Program-directed re-IPL	-	-	-	Y	Y
HiperDispatch	Y	Y	Y	Y	Y
IBM Z Integrated Information Processors (zIIPs)	Y	Y	Y	Y	Y
Transactional Execution	Y	Y	Y	Y[ef]	N
Java Exploitation of Transactional Execution	Y	Y	Y	Y[f]	N
Simultaneous multithreading (SMT)	Y	Y	N	Y[g]	Y
Single Instruction Multiple Data (SIMD)	Y	Y	N	Y[fh]	Y[f]
Hardware decimal floating point[i]	Y	Y	Y	Y	Y
2 GB large page support	Y	Y	Y	N	N
Large page (1 MB) support	Y	Y	Y	Y[f]	N
Out-of-order execution	Y	Y	Y	Y	Y
CPUMF (CPU measurement facility) for z14	Y	Y	N	Y	Y
Enhanced flexibility for Capacity on Demand (CoD)	Y	Y	Y	Y	Y
IBM Virtual Flash Memory (VFM)	Y	Y	Y[j]	N	N
1 MB pageable large pages[k]	Y	Y	Y	N	N
Guarded Storage Facility (GSF)	Y	N	N	Y[f]	N

Function[a]	z/OS V2R2	z/OS V2R1	z/OS V1R13	z/VM V6R4	z/VM V6R3
Instruction Execution Protection (IEP)	Y	N	N	Y[f]	N
Co-processor Compression Enhancements	Y	Y	N	N	N

a. PTFs may be required for toleration support or exploitation of z14 features and functions.
b. 170-way without multithreading. 128-way with multithreading enabled.
c. 64-way without multithreading and 32-way with multithreading enabled.
d. 32 TB of real storage is supported per server.
e. Guests are informed that TX facility is available for use
f. Guest exploitation support
g. Dynamic SMT with z14
h. Guests are informed that SIMD is available for use
i. Packed decimal conversion support.
j. A web deliverable is required, available at http://www-03.ibm.com/systems/z/os/zos/downloads/
k. With IBM Virtual Flash Memory, for middleware exploitation

Table 7-4 lists the supported base CPC functions for z/VSE, z/TPF and Linux on z Systems.

Table 7-4 Supported base CPC functions for z/VSE, z/TPF and Linux on z Systems

Function[a]	z/VSE V6R1	z/VSE V5R2	z/VSE V5R1	z/TPF V1R1	Linux on z Systems[b]
z14 servers	Y	Y	Y	Y	Y
Maximum processor unit (PUs) per system image	10	10	10	86	170[c]
Maximum main storage size[d]	32 GB	32 GB	32 GB	4 TB	16 TB[e]
85 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[f]	Y[f]	Y[f]	N	Y
HiperDispatch	N	N	N	N[g]	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)	N	N	N	N	Y
Hardware decimal floating point[h]	N	N	N	N	Y
2 GB large page support	N	N	N	Y	Y
Large page (1 MB) support	Y[i]	Y[i]	Y[i]	Y	Y
Out-of-order execution	Y	Y	Y	Y	Y

Function[a]	z/VSE V6R1	z/VSE V5R2	z/VSE V5R1	z/TPF V1R1	Linux on z Systems[b]
CPUMF (CPU measurement facility) for z14	N	N	N	Y	N[j]
Enhanced flexibility for CoD	N	N	N	N[g]	N
IBM Virtual Flash Memory (VFM)	N	N	N	N	Y
1 MB pageable large pages[k]	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 may be required for toleration support or exploitation of z14 features and functions.
b. Support statement varies based on Linux on z Systems distribution and release
c. For SLES12/RHEL7/Ubuntu 16.10, Linux supports 256 cores without SMT and 128 cores with SMT (=256 threads)
d. 32 TB of real storage is supported per server.
e. Linux on z Systems releases can support up to 64 TB memory
f. on SCSI disks
g. Availability expected in fourth quarter of 2017
h. Packed decimal conversion support.
i. Supported for data spaces
j. IBM is working with its Linux distribution Business Partners to provide this feature
k. With IBM Virtual Flash Memory, for middleware exploitation

7.3.2 Coupling and clustering

Table 7-5 lists the supported coupling and clustering functions for z/OS and z/VM.

Table 7-5 Supported coupling and clustering functions for z/OS and z/VM

Function[a]	z/OS V2R2	z/OS V2R1	z/OS V1R13	z/VM V6R4	z/VM V6R3
Server Time Protocol (STP)	Y	Y	Y	-	-
CFCC Level 22[b]	Y	Y	Y	Y[c]	Y[c]
CFCC Level 22 Coupling Thin Interrupts	Y	Y	Y	N	N
CFCC Level 22 Large Memory support	Y	Y	Y	N	N
CFCC Level 22 Support for 256 Coupling CHPIDs per CPC	Y	Y	Y	Y[c]	Y[c]
CFCC Level 22 Coupling Facility Processor Scalability	Y	Y	N	Y[c]	Y[c]
CFCC Level 22 List Notification Enhancements	Y	N	N	Y[c]	Y[c]
CFCC Level 22 Encryption Support[d]	N[e]	N[e]	N	Y[c]	Y[c]
CFCC Level 22 Exploitation of VFM (Virtual Flash Memory)	Y	Y	Y	N	N
RMF coupling channel reporting	Y	Y	Y	N	N
Coupling over InfiniBand CHPID type CIB	Y	Y	Y	N	N

Function[a]	z/OS V2R2	z/OS V2R1	z/OS V1R13	z/VM V6R4	z/VM V6R3
InfiniBand coupling links 12x at a distance of 150 m (492 ft.)	Y	Y	Y	N	N
InfiniBand coupling links 1x at an unrepeated distance of 10 km (6.2 miles)	Y	Y	Y	N	N
Integrated Coupling Adapter (ICA-SR) links CHPID CS5	Y	Y	Y	Y[f]	Y[f]
Coupling Express LR (CE LR) CHPID CL5	Y	Y	N	Y[f]	Y[f]
z/VM Dynamic I/O support for InfiniBand CHPIDs	-	-	-	Y[f]	Y[f]
z/VM Dynamic I/O support for ICA CHPIDs	-	-	-	Y[f]	Y[f]
Asynchronous CF Duplexing for lock structures	Y	N	N	Y[c]	N

a. PTFs may be required for toleration support or exploitation of z14 features and functions.
b. CFCC Level 22 with Driver 32.
c. Virtual guest coupling
d. Will be supported with z/OS 2.3
e. Toleration support (“locking out” down level systems who cannot use encrypted structure) will be provided for z/OS 2.2 and z/OS 2.1
f. To define, modify, and delete CHPID type CS5 when z/VM is the controlling LPAR for dynamic I/O.

In addition to operating system support listed in Table 7-5 on page 235, the following functions are supported on:

- Server Time Protocol is supported also on z/TPF V1R1 and Linux on z Systems.
- CFCC Level 22 is supported for z/TPF V1R1

Storage connectivity

Table 7-6 lists the supported storage connectivity functions for z/OS and z/VM.

Table 7-6 Supported storage connectivity functions for z/OS and z/VM

Function[a]	z/OS V2R2	z/OS V2R1	z/OS V1R13	z/VM V6R4	z/VM V6R3
zHyperLink Express	Y	Y	N	N	N
The 63.75-K subchannels	Y	Y	Y	Y	Y
Six logical channel subsystems (LCSSs)	Y	Y	Y	Y	Y
Four subchannel set per LCSS	Y	Y	Y	Y[b]	Y[b]
Health Check for FICON Dynamic routing	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
FICON Express16S+ for support of zHPF (z Systems High Performance FICON)	Y	Y	Y	Y[d]	Y[d]
FICON Express16S for support of zHPF	Y	Y	Y	Y[d]	Y[d]
FICON Express8S for support of zHPF	Y	Y	Y	Y[d]	Y[d]

Function[a]	z/OS V2R2	z/OS V2R1	z/OS V1R13	z/VM V6R4	z/VM V6R3
MIDAW (Modified Indirect Data Address Word)	Y	Y	Y	Y[d]	Y[d]
zDAC (z/OS Discovery and Auto-Configuration)	Y	Y	Y	N	N
FICON Express16S+ when using FICON or CTC (channel-to-channel)	Y	Y	Y	Y	Y
FICON Express16S when using FICON or CTC	Y	Y	Y	Y	Y
FICON Express8S when using FICON or CTC	Y	Y	Y	Y	Y
Global resource serialization (GRS) FICON CTC toleration	Y	Y	Y	N	N
IPL from an alternative subchannel set	Y	Y	Y	N	N
32 K subchannels for the FICON Express16S+	Y	Y	Y	Y	Y
32 K subchannels for the FICON Express16S	Y	Y	Y	Y	Y
Request node identification data	Y	Y	Y	N	N
FICON link incident reporting	Y	Y	Y	N	Y
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[d]	Y[d]
FICON Express16S support of hardware data router	-	-	-	Y[d]	Y[d]
FICON Express8S support of hardware data router	-	-	-	Y[d]	Y[d]
FICON Express16S+ T10-DIF support	-	-	-	Y[d]	Y[d]
FICON Express16S T10-DIF support	-	-	-	Y[d]	Y[d]
FICON Express8S T10-DIF support	-	-	-	Y[d]	Y[d]
Increased performance for the FCP protocol	-	-	-	Y	Y
N_Port ID Virtualization (NPIV)	-	-	-	Y	Y
Worldwide port name tool	-	-	-	Y	Y

a. PTFs may be required for toleration support or exploitation of z14 features and functions.
b. For specific Geographically Dispersed Parallel Sysplex (GDPS) usage only
c. Transparent to operating systems
d. For guest use

Table 7-7 lists the supported storage connectivity functions for z/VSE, z/TPF and Linux on z Systems.

Table 7-7 Supported storage connectivity functions for z/VSE, z/TPF and Linux on z Systems

Function[a]	z/VSE V6R1	z/VSE V5R2	z/VSE V5R1	z/TPF V1R1	Linux on z Systems[b]
zHyperLink Express	-	-	-	-	-

Function[a]	z/VSE V6R1	z/VSE V5R2	z/VSE V5R1	z/TPF V1R1	Linux on z Systems[b]
The 63.75-K subchannels	N	N	N	N	Y
Six logical channel subsystems (LCSSs)	Y	Y	Y	N	Y
Four 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]	N	N	N	Y	Y
FICON Express16S for support of zHPF	N	N	N	Y	Y
FICON Express8S for support of zHPF	N	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
FICON Express16S when using FICON or CTC	Y	Y	Y	Y	Y
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

Function[a]	z/VSE V6R1	z/VSE V5R2	z/VSE V5R1	z/TPF V1R1	Linux on z Systems[b]
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 may be required for toleration support or exploitation of z14 features and functions.
b. Support statement varies based on Linux on z Systems distribution and release
c. Transparent to operating systems
d. Will be supported on z/VSE V6.2 with PTFs

7.3.3 Network connectivity

Table 7-8 lists the supported network connectivity functions for z/OS and z/VM.

Table 7-8 Supported network connectivity functions for z/OS and z/VM

Function[a]	z/OS V2R2	z/OS V2R1	z/OS V1R13	z/VM V6R4	z/VM V6R3
Checksum offload for IPV6 packets	Y	Y	Y	Y[b]	Y[b]
Checksum offload for LPAR-to-LPAR traffic with IPv4 and IPv6	Y	Y	Y	Y[b]	Y[b]
Querying and displaying an OSA configuration	Y	Y	Y	N	N
QDIO data connection isolation for z/VM	-	-	-	Y	Y
QDIO interface isolation for z/OS	Y	Y	Y	-	-
QDIO OLM (Optimized Latency Mode)	Y	Y	Y	-	-
Adapter interruptions for QDIO	N	N	N	Y	Y
QDIO Diagnostic Synchronization	Y	Y	Y	N	N
IWQ (Inbound Workload Queuing) for OSA	Y	Y	Y	Y[b]	Y[b]
VLAN management enhancements	Y	Y	Y	Y[c]	Y[c]
GARP VLAN Registration Protocol	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[b]	Y[b]
z/VM Dynamic I/O Support for OSA-Express6S OSD CHPIDs	-	-	-	Y	Y
OSA Dynamic LAN idle	Y	Y	Y	N	N
OSA Layer 3 virtual MAC for z/OS environments	Y	Y	Y	-	-
Network Traffic Analyzer	Y	Y	Y	N	N

Function[a]	z/OS V2R2	z/OS V2R1	z/OS V1R13	z/VM V6R4	z/VM V6R3
Hipersockets					
HiperSockets[d]	Y	Y	Y	Y	Y
32 HiperSockets	Y	Y	Y	Y	Y
HiperSockets Completion Queue	Y	Y	Y	Y	Y
HiperSockets Virtual Switch Bridge	-	-	-	Y	Y
HiperSockets Multiple Write Facility	Y	Y	Y	N	N
HiperSockets support of IPV6	Y	Y	Y	Y	Y
HiperSockets Layer 2 support	Y	Y	Y	Y	Y
HiperSockets Network Traffic Analyzer for Linux on z Systems	-	-	-	-	-
SMC-D and SMC-R					
SMC-D[e] over ISM (Internal Shared Memory)	Y	N	N	Y[b]	Y[b]
10GbE RoCE[f] Express	Y	Y	Y	Y[b]	Y[b]
10GbE RoCE Express2 for SMC-R	Y	Y	N	Y[b]	Y[b]
10GbE RoCE Express2 for Ethernet communications[g] including Single Root I/O Virtualization (SR-IOV)	N	N	N	Y[b]	Y[b]
z/VM Dynamic I/O support for RoCE Express2	-	-	-	Y	Y
Shared RoCE environment	Y	Y	N	Y	Y
Open Systems Adapter (OSA)[h]					
OSA-Express6S 1000BASE-T Ethernet CHPID type OSC	Y	Y	Y	Y	Y
OSA-Express5S 1000BASE-T Ethernet CHPID type OSC	Y	Y	Y	Y	Y
OSA-Express4S 1000BASE-T Ethernet CHPID type OSC	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-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-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

Function[a]	z/OS V2R2	z/OS V2R1	z/OS V1R13	z/VM V6R4	z/VM V6R3
OSA-Express4S 1000BASE-T Ethernet CHPID type OSD	Y	Y	Y	Y	Y
OSA-Express6S 1000BASE-T Ethernet CHPID type OSE	Y	Y	Y	Y	Y
OSA-Express5S 1000BASE-T Ethernet CHPID type OSE	Y	Y	Y	Y	Y
OSA-Express4S 1000BASE-T Ethernet CHPID type OSE	Y	Y	Y	Y	Y

a. PTFs may be required for toleration support or exploitation of z14 features and functions.
b. For guest use/exploitation
c. Support of guests is transparent to z/VM if the device is directly connected to the guest (pass through).
d. On z14, the CHPID statement of HiperSockets devices requires the keyword VCHID. Therefore, the z14 IOCP definitions need to be migrated to support the HiperSockets definitions (CHPID type IQD). VCHID specifies the virtual channel identification number associated with the channel path. Valid range is 7E0 - 7FF. VCHID is not valid on z Systems 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, OSM

Table 7-9 lists the supported network connectivity functions for z/VSE, z/TPF and Linux on z Systems.

Table 7-9 Supported network connectivity functions for z/VSE, z/TPF and Linux on z Systems

Function[a]	z/VSE V6R1	z/VSE V5R2	z/VSE V5R1	z/TPF V1R1	Linux on z Systems[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	-	-	-	-	-
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

Function[a]	z/VSE V6R1	z/VSE V5R2	z/VSE V5R1	z/TPF V1R1	Linux on z Systems[b]
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	Y	Y	Y	N	Y
HiperSockets Network Traffic Analyzer for Linux on z Systems	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]
10GbE RoCE Express2 for SMC-R	N	N	N	N	N
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]					
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-Express4S 1000BASE-T Ethernet CHPID type OSC	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

Function[a]	z/VSE V6R1	z/VSE V5R2	z/VSE V5R1	z/TPF V1R1	Linux on z Systems[b]
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-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-Express4S 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
OSA-Express4S 1000BASE-T Ethernet CHPID type OSE	Y	Y	Y	N	N

a. PTFs may be required for toleration support or exploitation of z14 features and functions.
b. Support statement varies based on Linux on z Systems distribution and release
c. By using VLANs
d. On z14, the CHPID statement of HiperSockets devices requires the keyword VCHID. Therefore, the z14 IOCP definitions need to be migrated to support the HiperSockets definitions (CHPID type IQD). VCHID specifies the virtual channel identification number associated with the channel path. Valid range is 7E0 - 7FF. VCHID is not valid on z Systems 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 exploit 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, OSM

7.3.4 Cryptographic functions

Table 7-10 on page 243 lists the supported cryptography functions for z/OS and z/VM.

Table 7-10 Supported cryptography functions for z/OS and z/VM

Function[a]	z/OS V2R2	z/OS V2R1	z/OS V1R13	z/VM V6R4	z/VM V6R3
CP Assist for Cryptographic Function (CPACF)	Y	Y	Y	Y[b]	Y[b]
CPACF greater than 16 Domain Support	Y	Y	Y	Y[b]	Y[b]
CPACF AES-128, AES-192, and AES-256	Y	Y	Y	Y[b]	Y[b]
CPACF SHA-1, SHA-224, SHA-256, SHA-384, and SHA-512	Y	Y	Y	Y[b]	Y[b]
CPACF protected key	Y	Y	Y	Y[b]	Y[b]
Crypto Express6S	Y[c]	Y[c]	Y[c]	Y[b]	Y[b]

Function[a]	z/OS V2R2	z/OS V2R1	z/OS V1R13	z/VM V6R4	z/VM V6R3
Crypto Express6S Support for Visa Format Preserving Encryption	Y[c]	Y[c]	Y[c]	Y[b]	Y[b]
Crypto Express6S Support for Coprocessor in PCI-HSM Compliance Mode[d]	Y[c]	Y[c]	N	Y[b]	Y[b]
Crypto Express6S spouting up to 85 domains	Y[c]	Y[c]	Y	Y[b]	Y[b]
Crypto Express5S	Y	Y	Y	Y[b]	Y[b]
Crypto Express5S spouting up to 85 domains	Y	Y	Y	Y[b]	Y[b]
Elliptic Curve Cryptography (ECC)	Y	Y	Y	Y[b]	Y[b]
Secure IBM Enterprise PKCS #11 (EP11) coprocessor mode	Y	Y	Y	Y[b]	Y[b]
RCE (Regional Crypto Enablement)	Y	N	N	Y[b]	Y[b]
z/VM Dynamic I/O Support for RCE	-	-	-	Y	Y
z/OS Data Set Encryption	Y	N	N	-	-
z/VM Encrypted paging support	-	-	-	Y	N
RMF Support for Crypto Express6	Y	Y	N	-	-
z/OS encryption readiness technology (zERT)	Y	Y	N	-	-
z/TPF transparent database encryption	-	-	-	-	-

a. PTFs may be required for toleration support or exploitation of z14 features and functions.
b. For guest use/exploitation
c. A web deliverable is required, available at http://www-03.ibm.com/systems/z/os/zos/downloads/
d. Requires TKE 9.0

Table 7-11 lists the supported cryptography functions for z/VSE, z/TPF and Linux on z Systems.

Table 7-11 Supported cryptography functions for z/VSE, z/TPF and Linux on z Systems

Function[a]	z/VSE V6R1	z/VSE V5R2	z/VSE V5R1	z/TPF V1R1	Linux on z Systems[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

Function[a]	z/VSE V6R1	z/VSE V5R2	z/VSE V5R1	z/TPF V1R1	Linux on z Systems[b]
Crypto Express6S spouting up to 85 domains	Y	Y	Y	N	Y
Crypto Express5S	Y	Y	Y	Y	Y
Crypto Express5S spouting up to 85 domains	Y	Y	Y	N	Y
Elliptic Curve Cryptography (ECC)	N	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 may be required for toleration support or exploitation of z14 features and functions.
b. Support statement varies based on Linux on z Systems 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

Table 7-12 lists the supported Special-purpose features for z/OS and z/VM.

Table 7-12 Supported Special-purpose features for z/OS and z/VM

Function[a]	z/OS V2R2	z/OS V2R1	z/OS V1R13	z/VM V6R4	z/VM V6R3
zEDC[b] Express	Y	Y	N	Y[c]	Y[c]

a. PTFs may be required for toleration support or exploitation of z14 features and functions.
b. zEnterprise Data Compression.
c. For guest exploitation

7.4 Support by features and functions

This section addresses operating system support by function. Only the currently in-support releases are covered.

Tables in this section use the following convention:

N/A	Not applicable
NA	Not available

7.4.1 LPAR Configuration and Management

A single system image can control several processor units, such as CPs, zIIPs, or IFLs.

This section covers the following sub-topics:

- "Maximum number of PUs per system image" on page 246
- "Maximum main storage size" on page 247
- "Up to 85 LPARs" on page 247
- "Separate LPAR management of PUs" on page 247
- "Dynamic PU add" on page 247
- "Dynamic LPAR memory upgrade" on page 248
- "LPAR group absolute capping" on page 248
- "Capacity Provisioning Manager" on page 248
- "Program-directed re-IPL" on page 248
- "IOCP" on page 249

Maximum number of PUs per system image

Table 7-13 lists the maximum number of PUs supported by each operating system image and by special-purpose LPARs.

Table 7-13 Maximum number of PUs per system image

Operating system	Maximum number of PUs per system image
z/OS V2R2	170[a b]
z/OS V2R1	170[a b]
z/OS V1R13	100[b]
z/VM V6R4	64[c]
z/VM V6R3	64[c]
z/VSE V5R1 and later	z/VSE Turbo Dispatcher can use up to 4 CPs, and tolerates up to 10-way LPARs
z/TPF V1R1	86 CPs
CFCC Level 22	16 CPs or ICFs CPs and ICFs cannot be mixed.
Linux on z Systems	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: 256 CPs or IFLs
KVM for IBM Z	170
KVM Hypervisor	Offered with the following Linux distributions: SUSE Linux Enterprise Server 12: 256 CPs or IFLs. Ubuntu 16.04 LTS: 256 CPs or IFLs
Secure Service Container	80
GDPS Virtual Appliance	80

a. 170-way without multithreading. 128-way with multithreading.
b. Total characterizable PUs including zIIPs and CPs.
c. 64-way without multithreading and 32-way with multithreading enabled.

Maximum main storage size

Table 7-14 lists the maximum amount of main storage that is supported by current operating systems. A maximum of 16 TB of main storage can be defined for an LPAR on a z14 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/VM V6R3 supports 1 TB
z/VSE	z/VSE V5R1 and later support 32 GB.
z/TPF	z/TPF supports 4 TB.
CFCC	Level 22 supports up to 3 TB.
Secure Service Containers	Supports up to 3 TB.
Linux on z Systems (64-bit)	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 supports 10 TB**

a. A LPAR on z14 supports up to 16TB of memory
b. z14 servers support 32 TB user configurable memory per server.

Up to 85 LPARs

This feature, first made available on z13 servers, 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 six channel subsystems to reach the 85 LPARs limit.

See Table 7-3 on page 233 and Table 7-4 on page 234 for list of supported operating systems.

Remember: A virtual appliance 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 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.

See Table 7-3 on page 233 and Table 7-4 on page 234 for list of supported operating systems.

Dynamic PU add

Planning of an LPAR configuration allows defining reserved PUs that can be brought online when extra capacity is needed. Operating system support is required to use this capability without an IPL, that is, nondisruptively. This support has been in z/OS for a long time.

The dynamic PU add function enhances this support by allowing you to define and change dynamically the number and type of reserved PUs in an LPAR profile, removing any planning

requirements. The new resources are immediately made available to the operating system and, in the case of z/VM, to its guests.

See Table 7-3 on page 233 and Table 7-4 on page 234 for list of supported operating systems.

Dynamic LPAR memory upgrade

An LPAR can be defined with both an initial and a reserved amount of memory. At activation time, the initial amount is made available to the partition and the reserved amount can be added later, partially or totally. Those two memory zones do not have to be contiguous in real memory, but appear as logically contiguous to the operating system that runs in the LPAR.

z/OS can take advantage of this support and nondisruptively acquire and release memory from the reserved area. z/VM V6R2 and later can acquire memory nondisruptively, and immediately make it available to guests. z/VM virtualizes this support to its guests, which 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 Systems also supports both acquiring and releasing memory nondisruptively. This feature is enabled for SUSE Linux Enterprise Server 11 and RHEL 6 and later releases.

LPAR group absolute capping

On z13 servers, PR/SM is enhanced to support an option to limit the amount of physical processor capacity that is consumed by an individual LPAR when a PU that is defined as a 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.

See Table 7-3 on page 233 and Table 7-4 on page 234 for list of supported operating systems.

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 337. The new process is inherently more flexible, and can be automated. This capability can result in easier, faster, and more reliable management of the processing capacity.

The Capacity Provisioning Manager, a function that is first available with z/OS V1R9, interfaces with z/OS Workload Manager (WLM) and implements capacity provisioning policies. Several implementation options are available, from an analysis mode that issues only guidelines, to an autonomic mode that provides fully automated operations.

Replacing manual monitoring with autonomic management or supporting manual operation with guidelines can help ensure that sufficient processing power is available with the least possible delay. See Table 7-3 on page 233 for list of supported operating systems.

Program-directed re-IPL

First available on System z9, program directed re-IPL allows an operating system on a z14 to IPL again without operator intervention. This function is supported for both SCSI and IBM extended count key data (IBM ECKD) devices.

See Table 7-3 on page 233 and Table 7-4 on page 234 for list of supported operating systems.

IOCP

All IBM Z servers require a description of their I/O configuration. This description is stored in input/output configuration data set (IOCDS) files. The input/output configuration program (IOCP) allows the creation of the IOCDS file from a source file that is known as the input/output configuration source (IOCS).

The IOCS file contains definitions of LPARs, channel subsystems as well as detailed information for each channel and path assignment, each control unit, and each device in the configuration.

IOCP for z14 provides support for the following:

- z14 Base machine definition
- New PCI function adapter for zHyperLink (HYL)
- New PCI function adapter for RoCE Express2 (CX4)
- New IOCP Keyword MIXTYPE required for prior FICON[3] cards

IOCP required level for z14 **servers:** The required level of IOCP for the z14 is IOCP 5.4.0 or later with PTFs. For more information, see the following manuals:

- *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.

7.4.2 Base CPC Features and Functions

This section covers the following sub-topics:

- "HiperDispatch" on page 249
- "zIIP support" on page 251
- "Transactional Execution" on page 252
- "Simultaneous multithreading" on page 252
- "Single-instruction multiple-data" on page 255
- "Hardware decimal floating point" on page 256
- "Out-of-order Execution" on page 256
- "CPU Measurement Facility" on page 256
- "Large page support" on page 257
- "Virtual Flash Memory" on page 257
- "Guarded Storage Facility (GSF)" on page 257
- "Instruction Execution Protection (IEP)" on page 257
- "Co-processor Compression (CMPSC) Enhancements" on page 258

HiperDispatch

The **HIPERDISPATCH=YES/NO** parameter in the IEAOPTxx member of SYS1.PARMLIB and on the **SET OPT=xx** command controls whether HiperDispatch is enabled or disabled for a z/OS image. It can be changed dynamically, without an IPL or any outage.

The default is that HiperDispatch is disabled on all releases, from z/OS V1R10 (requires PTFs for zIIP support) through z/OS V1R12.

Beginning with z/OS V1R13, when running on a z14, z13, z13s, zEC12, or zBC12 server, the IEAOPTxx keyword **HIPERDISPATCH** defaults to `YES`. If `HIPERDISPATCH=NO` is specified, the specification is accepted as it was on previous z/OS releases.

[3] FICON Express16S+ does not allow a mixture of CHPID types on new cards

The use of SMT on z14 servers requires that HiperDispatch is enabled on the operating system. See for “Simultaneous multithreading” on page 252 more information.

Additionally, with z/OS V1R12 or later, any LPAR running with more than 64 logical processors is required to operate in HiperDispatch Management Mode.

The following rules control this environment:

- If an LPAR is defined at IPL with more than 64 logical processors, the LPAR automatically operates in HiperDispatch Management Mode, regardless of the `HIPERDISPATCH=` specification.
- If more logical processors are added to an LPAR that has 64 or fewer logical processors and the additional logical processors raise the number of logical processors to more than 64, the LPAR automatically operates in HiperDispatch Management Mode regardless of the `HIPERDISPATCH=YES/NO` specification. That is, even if the LPAR has the `HIPERDISPATCH=NO` specification, that LPAR is converted to operate in HiperDispatch Management Mode.
- An LPAR with more than 64 logical processors running in HiperDispatch Management Mode cannot be reverted to run in non-HiperDispatch Management Mode.

HiperDispatch on z14 servers uses a new chip and CPC drawer configuration to improve the access cache performance. Beginning with z/OS V2R1, HiperDispatch changed to use the new node cache structure of z14 servers. The base support is provided by PTFs identified by IBM.device.server.z14-3906.requiredservice.

The PR/SM in the System z9 EC to zEC12 servers stripes the memory across all books in the system to take advantage of the fast book interconnection and spread memory controller work. The PR/SM on z14 servers seeks to assign all memory in one CPC drawer striped across the clusters of that drawer in order to take advantage of the lower latency memory access in a drawer.

The PR/SM in the System z9 EC to zEC12 servers attempts to assign all logical processors to one book, packed into PU chips of that book in cooperation with operating system HiperDispatch optimize shared cache usage. The PR/SM on z14 servers seeks to assign all logical processors of a partition to one CPC drawer, packed into PU chips of that CPC drawer in cooperation with operating system HiperDispatch optimize shared cache usage.

The PR/SM automatically keeps partition’s memory and logical processors on the same CPC drawer. This arrangement looks simple for a partition, but it is a complex optimization for multiple logical partitions because some must be split among processors drawers.

To use HiperDispatch effectively, WLM goal adjustment might be required. Review the WLM policies and goals, and update them as necessary. 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 servers.

z/VM V6R4

z/VM V6R4 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 system. The supported processor limit has been increased to 64, whereas with SMT, it remains at 32, supporting up to 64 threads running simultaneously.

CPU polarization support in Linux on z Systems

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 Systems, see:

http://www.ibm.com/support/knowledgecenter/linuxonibm/com.ibm.linux.z.lgdd/lgdd_t_cpu_pol.html

z/TPF

z/TPF on z14 can utilize additional processors immediately without reactivating the LPAR or IPL'ing the z/TPF system

In installation prior to z14, z/TPF workload is evenly distributed across all available processors, even in low utilization situations. This causes cache and core contention with other LPARs. When z/TPF is running in a shared processor configuration, the achieved MIPS will be higher when z/TPF is utilizing a minimum set of processors.

In low utilization periods, z/TPF will now minimize the processor footprint, by compressing TPF workload onto a minimal set of I-streams (engines), thus reducing the impact on other LPARs, allowing the entire CPC to operate more efficiently.

As a consequence, z/OS and z/VM will experience less contention from the z/TPF system when the z/TPF system is operating at periods of low demand.

See Table 7-3 on page 233 and Table 7-4 on page 234 for list of supported operating systems.

zIIP support

zIIPs do not change the model capacity identifier of z14 servers. IBM software product license charges based on the model capacity identifier are not affected by the addition of zIIPs. On a z14 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 these applications:

- DB2 V8 and later for z/OS data serving, for applications that use data Distributed Relational Database Architecture (DRDA) over TCP/IP, such as data serving, data warehousing, and selected utilities.
- z/OS XML services.
- z/OS CIM Server.
- z/OS Communications Server for network encryption (Internet Protocol Security (IPSec)) and for large messages that are sent by HiperSockets.
- IBM GBS Scalable Architecture for Financial Reporting.
- IBM z/OS Global Mirror (formerly XRC) and System Data Mover.
- IBM OMEGAMON® XE on z/OS, OMEGAMON XE on DB2 Performance Expert, and DB2 Performance Monitor.
- Any Java application that is using the current IBM SDK.
- WebSphere Application Server V5R1 and later, and products that are based on it, such as WebSphere Portal, WebSphere Enterprise Service Bus (WebSphere ESB), and WebSphere Business Integration (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. See Table 7-3 on page 233 for list of supported operating systems.

On z14 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, z13, z13s, and zBC12 servers.

This feature enables software to indicate to the hardware the beginning and end of a group of instructions that need to be treated in an atomic way. Either all of their results happen or none happens, in true transactional style. The execution is optimistic. The hardware provides a memory area to record the original contents of affected registers and memory as the instruction's execution takes place. If the transactional execution group is canceled or must be rolled back, the hardware transactional memory is used to reset the values. Software can implement a fallback capability.

This capability enables more efficient software by providing a way to avoid locks (lock elision). This advantage is of special importance for speculative code generation and highly parallelized applications.

TX is used by IBM Java virtual machine (JVM), but potentially can be used by other software. See Table 7-3 on page 233 and Table 7-4 on page 234 for list of supported operating systems.

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, increasing the overall throughput of the system.

The z14 can run up two threads simultaneously in the same processor, dynamically sharing 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.

SMT[4] is supported for zIIPs and IFLs.

- For zIIPs and IFLs, SMT must be enabled on z/OS, z/VM, or Linux on z Systems 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.

[4] On z14 SMT is also enabled (not user configurable) by default for SAPs

See Table 7-3 on page 233 and Table 7-4 on page 234 for list of supported operating systems.

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 have run previously. PR/SM attempts to dispatch a vertical low logical processor in the same physical processor. If that 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 point of view of an application, SMT is transparent and no changes are required in the application for it to run in an SMT environment, as shown in Figure 7-1.

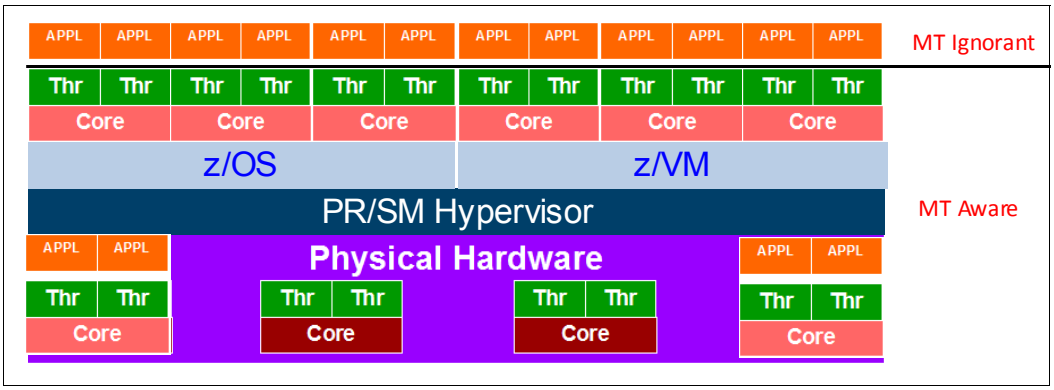


Figure 7-1 Simultaneous multithreading

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, where a core can have one or more threads. The number of threads is limited by z14 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 running in z14, HiperDispatch is forced to run as enabled, and you cannot disable HiperDispatch. The **PROCVIEW** statement cannot be changed dynamically, so you must run an IPL after changing it to make the new setting effective.

The **MT_ZIIP_MODE** parameter in the IEAOPTxx controls zIIP SMT mode. It can be 1 (the default), where only one thread can be running in a core, or 2, where up two threads can be running in a core. If **PROCVIEW CPU** is specified, the **MT_ZIIP_MODE** is always 1. Otherwise, the use of SMT to dispatch two threads in a single zIIP logical processor (MT_ZIIP_MODE=2) can be changed dynamically by using the **SET OPT=xx** setting in the IEAOPTxx parmlib. Changing the MT mode for all cores can take some time to complete.

The activation of SMT mode also requires that the HMC Customize/Delete Activation Profiles task “Do not end the time slice if a partition enters a wait state” must not be selected. This is the recommended default setting.

PROCVIEW CORE requires DISPLAY M=CORE and CONFIG CORE to display the core states and configure an entire core.

With the introduction of Multi-Threading support for SAPs in z14, the maximum number of logical SAPs can reach up to 88. RMF is updated to support this change by implementing page break support in the I/O Queuing Activity report generated by the RMF Post processor.

z/VM V6R4

The use of SMT in z/VM is enabled via **MULTITHREADING** statement in the system configuration file. Multithreading will be 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, it is now possible to change the number of active threads per core 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, 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 Systems

The upstream kernel 4.0 has SMT functionality developed by the Linux on z Systems development team. SMT is supported on LPAR only (not as a second level guest). For more information see:

https://www.ibm.com/developerworks/linux/linux390/kernel-4.0.html

Listed below are the minimum releases of Linux on z Systems distributions natively supporting SMT:

- Red Hat RHEL 7.2
- SUSE Linux Enterprise Server12 SP1
- Ubuntu 16.04 LTS
- KVM for IBM Z V1.1.1

- KVM hypervisor which is offered with the following Linux distributions: SLES-12 SP2 or higher, and Ubuntu 16.04 LTS or higher.

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 is equipped with new set of instructions to improve the performance of complex mathematical models and analytic workloads through vector processing and new complex instructions, which can process a lot of 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 has 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.

See Table 7-3 on page 233 and Table 7-4 on page 234 for list of supported operating systems. Operating System support includes the following items[5]:

- Enablement of vector registers.
- Use of vector registers using XL C/C++ ARCH(11) and TUNE(11).
- A math library with an optimized and tuned math function (Mathematical Acceleration Subsystem (MASS)) that can be used in place of some of the C standard math functions. It has a SIMD vectorized and non-vectorized version.
- A specialized math library, which is known as Automatically Tuned Linear Algebra Software (ATLAS), 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 SMID 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 have built-in functions for SIMD:

- IBM Java
- XL C/C++
- Enterprise COBOL
- Enterprise PL/I

Code must be developed to take advantage of the SIMD functions, and applications with SIMD 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.

[5] Features listed here may not be available on all operating systems listed in the tables.

The MASS and standard math library have 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 servers inherited the decimal floating point accelerator feature that was introduced with z10 EC

z14 has a new decimal architecture with Vector Enhancements Facility and Vector Packed Decimal Facility for Data Access Accelerator. Vector Packed Decimal Facility introduces set of instructions that perform operation on decimal type using vector registers in order to improve performance.

See Table 7-3 on page 233 and Table 7-4 on page 234 for list of supported operating systems. For more information, see 7.5.4, “z/OS XL C/C++ considerations” on page 291.

Out-of-order Execution

OOO yields significant performance benefits for compute-intensive applications by reordering instruction execution, allowing later (younger) instructions to be run ahead of a stalled instruction, and reordering storage accesses and parallel storage accesses. OOO maintains good performance growth for traditional applications.

See Table 7-3 on page 233 and Table 7-4 on page 234 for list of supported operating systems. For more information, see “3.4.2, “Out-of-order execution” on page 96.

CPU Measurement Facility

Also known as Hardware Instrumentation Services (HIS), CPU Measurement Facility (CPUMF) was initially introduced with z10 EC in order 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. See Table 7-3 on page 233 for list of supported operating systems.

Refer to Load-Program-Parameter and the CPU-Measurement Facilities for detailed describe this function:

http://www.ibm.com/support/docview.wss?uid=isg26fcd1cc32246f4c8852574ce0044734a&aid=1

Review below IBM techdoc for an overview of the CPU Measurement Facility: what it is, how to enable it, why you should use it.

http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/TC000066

For more information, see “12.2, “LSPR workload suite” on page 431.

Large page support

In addition to the existing 1-MB large pages, 4-KB pages, and page frames, z14 servers support pageable 1-MB large pages, large pages that are 2 GB, and large page frames. See Table 7-3 on page 233 and Table 7-4 on page 234 for list of supported operating systems.

Virtual Flash Memory

IBM Virtual Flash Memory (VFM) is the replacement for the Flash Express features (#0402, #0403) which were available on the IBM zEC12 and IBM z13. No application changes are required to change from IBM Flash Express to VFM for it implements EADM Architecture using HSA-like memory instead of Flash card pairs.

IBM Virtual Flash Memory (#0604) offers up to 6.0 TB of memory in 1.5 TB 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.

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

See Table 7-3 on page 233 and Table 7-4 on page 234 for list of supported operating systems.

Guarded Storage Facility (GSF)

Also known as less-pausing garbage collection, Guarded Storage Facility is a new architecture introduced with z14 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 gets 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 un-dispatch and re-dispatch. Enabling a less-pausing approach improves Java garbage collection. Function is provided on z14 running z/OS 2.2 and later with APAR OA51643 installed. **`MACHMIG`** statement in **`LOADxx`** of SYS1.PARMLIB provides ability to disable the function.

z/VM

With the PTF for APAR VM65987, planned to be available December 15, 2017, z/VM V6.4 will provide support for guest exploitation of the z14 guarded storage facility. This facility is designed to improve the performance of garbage-collection processing by various languages, in particular Java.

See Table 7-3 on page 233 and Table 7-4 on page 234 for list of supported operating systems.

Instruction Execution Protection (IEP)

Instruction Execution Protection is a new hardware function introduced with z14 that enables software like LE (Language Environment) to mark certain memory regions (E.g. 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 explicit software requests to obtain memory areas as non-executable it is possible to protect areas of memory from unauthorized execution. 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 a Protection Exception will occur.

z/OS

In order to exploit IEP, RSM (Real Storage Manager) is enhanced to request non-executable memory allocation. Use new keyword **`EXECUTABLE=YES|NO`** on either **`STORAGE OBTAIN`** or **`IARV64`** to indicate whether memory to be used contains executable code. RTM (Recovery Termination Manager) writes LOGREC record of any program-check that results from IEP.

IEP support is for z/OS 2.2 and later running on z14 with APARs OA51030 and OA51643 installed.

z/VM

Guest exploitation support for the Instruction Execution Protection Facility is provided with APAR VM65986.

See Table 7-3 on page 233 and Table 7-4 on page 234 for list of supported operating systems.

Co-processor Compression (CMPSC) Enhancements

Since z196, IBM Z processors have on-chip co-processors with built-in capabilities like compression, expansion and translation. Each follow-on server generation improved on-chip co-processors functionality and performance and z14 is taking this even further with following enhancements:

Entropy Encoding for CMPSC with Huffman Coding

z14 enables increased compression ratio (using Huffman coding) for the on-chip compression coprocessor resulting in fewer CPU cycles to enable further compression of data. This improves memory, transfer, and disk efficiency. Software-based expansion of Huffman-encoded text will be used when hardware support is not present. Support is provided by z/OS 2.1 and later running on z14 with OA49967 installed.

Order Preserving Compression

Order Preserving Compression allows comparisons against compressed data. It helps to achieve further disk and memory savings by being able to compress search trees, index or sort files that have not been practical to compress previously.

Combining Order Preserving Compression (individual compression of index entries) with Entropy Encoding (to compress non-index data) may result up to 35% data reduction[6] for DB2 considering that significant portion of DB2 disk space is in indices for fast access to data.

See Table 7-3 on page 233 and Table 7-4 on page 234 for list of supported operating systems.

7.4.3 Coupling and clustering Features and Functions

This section covers the following sub-topics:

[6] Results based on modeling, not actual measurements

- "Coupling facility and CFCC considerations" on page 259
- "CFCC Level 22" on page 259
- "Coupling links support" on page 260
- "Virtual Flash Memory exploitation by CFCC" on page 261
- "CFCC Coupling Thin Interrupts" on page 262
- "Asynchronous CF Duplexing for lock structures" on page 262
- "z/VM Dynamic I/O support for InfiniBand and ICA CHPIDs" on page 262

Coupling facility and CFCC considerations

Coupling facility connectivity to a z14 is supported on the z13, zEC12, zBC12, or another z14. The LPAR running the CFCC can be on any of the previously listed supported systems. See Table 7-15 for CFCC Levels supported on Z servers.

Table 7-15 IBM Z CFCC code-levels

IBM Z server	Code level
z14	CFCC Level 22
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 z196 and previous systems is not supported, introducing z14 into existing installations requires extra planning. Consider the level of CFCC. For more information, see "Migration considerations" on page 178.

CFCC Level 22

CFCC Level 22 is delivered on z14 servers with driver level 32. CFCC Level 22 introduces the following enhancements:

- *Coupling Express Long Range (CE LR):* A new link type introduced with z14 for long distance coupling connectivity.
- *Coupling Facility (CF) Processor Scalability:* Coupling Facility work management and dispatching changes for IBM z14 allow improved efficiency and scalability for coupling facility images.
  - First, ordered work queues have been eliminated from the CF in favor of first-in / first-out queues, avoiding the overhead of maintaining ordered queues.
  - Second, protocols for system-managed duplexing have been simplified to avoid the potential for latching deadlocks between duplexed structures.
  - Third, the CF image is now able to use its processors to perform specific work management functions when the number of processors in the CF image exceeds a threshold. Together, these changes will improve the processor scalability and throughput for a CF image.
- *CF List Notification Enhancements:* There are significant enhancements 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, reducing notification overhead.

  - Second, there is now an option for delivering “aggressive” notifications, which can drive a notification when new elements are added to a queue, providing 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, allowing exploiters to re-drive 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 exploiters, and yields more consistent, timely notifications.

- *Coupling Link Constraint Relief*: IBM z14 provides additional physical and logical coupling link connectivity compared to z13.
  - The maximum number of physical ICA SR coupling links (ports) is increased from 40 per CPC to 80 per CPC. These higher limits on z14 support concurrent use of Infiniband coupling, ICA SR, and CE LR links, for coupling link technology migration purposes.
  - Maximum number of coupling CHPIDs (of all types) is 256 per CPC (same as z13)
- *CF Encryption:* z/OS 2.3 provides support for end-to-end encryption for both 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 CF images are not required, but are recommended in order to simplify some sysplex recovery and reconciliation scenarios involving encrypted CF structures. (Note that the CF image itself never decrypts, nor encrypts, any data). IBM z14 z/OS images are not required, but are recommended for the improved AES CBC encrypt/decrypt performance that z14 provides.

See Table 7-5 on page 235 for list of supported operating systems.

For more information about CFCC code levels, see the Parallel Sysplex website at:

http://www.ibm.com/systems/z/pso/cftable.html

For the latest CFCC code levels, see the current exception letter that is published on Resource Link at the following website:

https://www.ibm.com/servers/resourcelink/lib03020.nsf/pages/exceptionLetters?OpenDocument

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, found at the following website:

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

Sizer Utility, an authorized z/OS program download, is useful when you are upgrading a CF. It is available at the following web page:

http://www.ibm.com/systems/support/z/cfsizer/altsize.html

Before the migration, install the compatibility/coexistence PTFs. A planned outage is required when you upgrade the CF or CF LPAR to CFCC Level 22.

Coupling links support

Integrated Coupling Adapter (ICA) Short Reach, Coupling Express Long Reach (CE LR) and InfiniBand (IFB) coupling link options provide high-speed connectivity at both short distances and longer distances over fiber optic interconnections. Several areas of this book address CE LR, ICA SR and InfiniBand characteristics and support. For more information, see 4.7.4, “Parallel Sysplex connectivity” on page 175.

Integrated Coupling Adapter

PCIe Gen3 fanout (also known as Integrated Coupling Adapter Short Range (ICA SR) supports a maximum distance of 150 meters (492 feet) and defined as CHPID type CS5 in IOCP. See Table 7-5 on page 235 for list of supported operating systems.

Coupling Express Long Reach

The CE LR link provides point-to-point coupling connectivity at distances of 10 km unrepeated and defined as CHPID type CL5 in IOCP. See Table 7-5 on page 235 for list of supported operating systems.

InfiniBand coupling links

HCA3-O[7] (12xIFB) fanout supports InfiniBand coupling links 12x at a maximum distance of 150 meters (492 feet) and defined as CHPID type CIB in IOCP. See Table 7-5 on page 235 for list of supported operating systems.

InfiniBand coupling links at an unrepeated distance of 10 km (6.2 miles)

HCA3-O LR[7] (1xIFB) fanout supports InfiniBand coupling links 1x at an unrepeated distance of 10 KM (6.2 miles). See Table 7-5 on page 235 for list of supported operating systems.

IBM z14 will be the last z Systems and IBM Z server to support HCA3-O and HCA3-O LR adapters[a]**:** z14 will be last z Systems and IBM Z server to support HCA3-O fanout for 12x IFB (#0171) and HCA3-O LR fanout for 1x IFB (#0170). As announced previously, z13s is the last mid-range z Systems server to support these adapters. Enterprises should begin migrating from HCA3-O and HCA3-O LR adapters to ICA SR and/or Coupling Express Long Reach (CE LR) adapters on z14, 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 requiring a DWDM, enterprises will need to determine their desired DWDM vendor's plan to qualify the planned replacement long-range coupling links**:**

- IBM Z enterprises should plan to migrate off of 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 requiring a DWDM, enterprises will need to determine their desired DWDM vendor's plan to qualify the CE LR. Reference Hardware Announcement 117-031, dated March 2017.

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.

Virtual Flash Memory exploitation by CFCC

VFM can be used in coupling facility images to provide extended capacity and availability for workloads making use of Websphere MQ Shared Queues structures. Using 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.

[7] HCA2-O is not supported on z14.

CFCC Coupling Thin Interrupts

The Coupling Thin Interrupts enhancement is delivered with CFCC 19. It improves the performance of a Coupling Facility partition and improves the dispatching of z/OS LPARs awaiting the arrival of returned asynchronous CF requests, when used in a shared engine environment. See Table 7-5 on page 235 for list of supported operating systems, and “Coupling Thin Interrupts” on page 108.

Asynchronous CF Duplexing for lock structures

Asynchronous CF Duplexing enhancement is a general-purpose interface for any CF Lock structure exploiters. It enabling secondary structure updates to be performed asynchronously with respect to 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:

- 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 V12 with PTFs for APAR PI66689
- IRLM V2.3 with PTFs for APAR PI68378

See Table 7-5 on page 235 for list of supported operating systems.

z/VM Dynamic I/O support for InfiniBand and ICA 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 both the server and z/VM without shutting down the system.

This function refers exclusively to the z/VM dynamic I/O support of InfiniBand and ICA coupling links. Support is available for the CIB and CS5 CHPID type in the z/VM dynamic commands, including the **`change channel path`** dynamic I/O command. Specifying and changing the system name when entering and leaving configuration mode are also supported. z/VM does not use InfiniBand or ICA, and does not support the use of InfiniBand or ICA coupling links by guests. See Table 7-5 on page 235 for list of supported operating systems.

7.4.4 Storage Connectivity Related Features and Functions

This section covers the following sub-topics:

- “zHyperlink Express” on page 263
- “FICON Express16S+” on page 263
- “FICON Express16S” on page 263
- “FICON Express8S” on page 264
- “Extended distance FICON” on page 264
- “zHPF (High Performance FICON)” on page 264
- “MIDAW (Modified Indirect Data Address Word facility)” on page 266
- “ICKDSF” on page 270
- “z/OS Discovery and Auto-Configuration” on page 270
- “Platform and name server registration in FICON channel” on page 270
- “The 63.75-K subchannels” on page 271
- “Multiple subchannel sets” on page 271
- “Fourth subchannel set” on page 271
- “IPL from an alternative subchannel set” on page 272
- “32 K subchannels for the FICON Express16S+ & FICON Express16S” on page 272
- “Request node identification data” on page 272

- “FICON link incident reporting” on page 272
- “Health Check for FICON Dynamic routing” on page 272
- “Global resource serialization (GRS) FICON CTC toleration” on page 272
- “Increased performance for the FCP protocol:” on page 273
- “T10-DIF support” on page 273
- “N_Port ID Virtualization” on page 273
- “Worldwide port name tool” on page 273

zHyperlink Express

z14 introduces IBM zHyperLink Express as a brand new IBM Z input/output (I/O) channel link technology since FICON. zHyperLink Express is designed to help bring data close 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 150m), 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 resides in the DS8880 cache. This eliminates the un-dispatch of the running request, the queuing delays to resume the request, and the PU cache disruption.

Statement of Direction[a]:

IBM intends to deliver VSAM exploitation of z14 and DS8880 zHyperLink Express.

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.

See Table 7-6 on page 236 and Table 7-7 on page 237 for list of supported operating systems.

FICON Express16S+

FICON Express16S+ supports a link data rate of 16 gigabits per second (Gbps) and autonegotiation to 4 or 8 Gbps for synergy with existing switches, directors, and storage devices. With support for native FICON, High Performance FICON for Z (zHPF), and Fibre Channel Protocol (FCP), the IBM z14 server enables you to position your SAN for even higher performance -- helping 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 will work with your existing fiber optic cabling environment, both single mode and multimode optical cables. The FICON Express16S+ feature running at end-to-end 16 Gbps link speeds will provide reduced latency for large read/write operations and increased bandwidth compared to the FICON Express8S feature.

See Table 7-6 on page 236 and Table 7-7 on page 237 for list of supported operating systems.

FICON Express16S

FICON Express16S supports a link data rate of 16 gigabits per second (Gbps) and autonegotiation to 4 or 8 Gbps for synergy with existing switches, directors, and storage devices. With support for native FICON, zHPF, and FCP, the z14 server enables SAN for even higher performance, helping to prepare for an end-to-end 16 Gbps infrastructure to meet the increased bandwidth demands of your applications.

The new features for the multimode and single mode fiber optic cabling environments reduce latency for large read/write operations and increase bandwidth compared to the FICON Express8S features.

See Table 7-6 on page 236 and Table 7-7 on page 237 for list of supported operating systems.

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 have the same type).

Statement of Direction[a]:

IBM z14 *will be the last z Systems and IBM Z high-end server to support FICON Express8S:* z14 will be last z Systems and IBM Z high-end server to support FICON Express8S (#0409 and #0410) channels. Enterprises should begin migrating from FICON Express8S channels to FICON Express16S+ channels (#0427 and #0428). FICON Express8S will not be 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 contrasted to FICON Express8, FICON Express4, and FICON Express2, which have a firmware-only zHPF implementation.

See Table 7-6 on page 236 and Table 7-7 on page 237 for list of supported operating systems.

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.

See Table 7-6 on page 236 and Table 7-7 on page 237 for list of supported operating systems.

zHPF (High Performance FICON)

zHPF was first provided on System z10, and is a FICON architecture for protocol simplification and efficiency. It reduces the number of information units (IUs) processed. Enhancements have been made to the z/Architecture and the FICON interface architecture to provide optimizations for online transaction processing (OLTP) workloads.

zHPF is available on z14, z13, z13s, zEC12, and zBC12 servers. The FICON Express16S+, FICON Express16S, and FICON Express8S (CHPID type FC) concurrently support both the existing 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. Additionally, the changes to the architectures provide end-to-end system enhancements to improve reliability, availability, and serviceability (RAS).

zHPF is compatible with these standards:

- Fibre Channel Framing and Signaling standard (FC-FS)
- Fibre Channel Switch Fabric and Switch Control Requirements (FC-SW)
- Fibre Channel Single-Byte-4 (FC-SB-4) standards

The zHPF channel programs can be used, for example, by the z/OS OLTP I/O workloads, DB2, VSAM, the partitioned data set extended (PDSE), and the z/OS file system (zFS).

At the zHPF announcement, zHPF supported the transfer of small blocks of fixed size data (4 K) from a single track. This capability has been extended, first to 64 KB, and then to multitrack operations. The 64 KB data transfer limit on multitrack operations was removed by z196. This improvement allows the channel to fully use the bandwidth of FICON channels, resulting in higher throughputs and lower response times.

The multitrack operations extension applies exclusively to the FICON Express16S+, FICON Express16S, and FICON Express8S, on the z14, 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 (> 64 KB) at distances up to 100 km to be run in a single round trip to the control unit. This process does not elongate the I/O service time for these write operations at extended distances. This enhancement to zHPF removes a key inhibitor for clients adopting zHPF over extended distances, especially when using the IBM HyperSwap capability of z/OS.

From the z/OS point of view, the existing FICON architecture is called *command mode* and the zHPF architecture is called *transport mode*. During link initialization, the channel node and the control unit node indicate whether they support zHPF.

Requirement: All FICON 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, there is a parameter in the IECIOSxx member of SYS1.PARMLIB (`ZHPF=YES` or `NO`) and in the **`SETIOS`** system command to control whether zHPF is enabled or disabled. The default is `ZHPF=NO`.

Support is also added for the **`D IOS,ZHPF`** system command to indicate whether zHPF is enabled, disabled, or not supported on the server.

Similar to the existing FICON channel architecture, the application or access method provides the channel program (CCWs). The way that zHPF (transport mode) manages channel program operations is different from the CCW operation for the existing FICON architecture (command mode). While in command mode, each CCW is sent to the control unit for

execution. In transport mode, multiple channel commands are packaged together and sent over the link to the control unit in a single control block. Fewer processors are used compared to the existing FICON architecture. Certain complex CCW chains are not supported by zHPF.

See Table 7-6 on page 236 and Table 7-7 on page 237 for list of supported operating systems.

For more information about FICON channel performance, see the performance technical papers on the IBM Z I/O connectivity website at:

http://www.ibm.com/systems/z/hardware/connectivity/ficon_performance.html

MIDAW (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 exploitation that is designed to improve FICON performance. This facility was first made available on System z9 servers, and is used by the Media Manager in z/OS.

The MIDAW facility provides a more efficient CCW/IDAW structure for certain categories of data-chaining I/O operations.

MIDAW can improve FICON performance for extended format data sets. Non-extended data sets can also benefit from MIDAW.

MIDAW can improve channel utilization, and can improve I/O response time. It reduces FICON channel connect time, director ports, and control unit processor usage.

IBM laboratory tests indicate that applications that use EF data sets, such as DB2, or long chains of small blocks can gain significant performance benefits by using the MIDAW facility.

MIDAW is supported on FICON channels that are configured as CHPID type FC. See Table 7-6 on page 236 and Table 7-7 on page 237 for list of supported operating systems.

MIDAW technical description

An IDAW is used to specify data addresses for I/O operations in a virtual environment.[8] The existing IDAW design allows the first IDAW in a list to point to any address within a page. Subsequent IDAWs in the same list must point to the first byte in a page. Also, IDAWs (except the first and last IDAW) in a list must deal with complete 2 K or 4 K units of data.

Figure 7-2 shows a single CCW that controls the transfer of data that spans non-contiguous 4 K frames in main storage. When the IDAW flag is set, the data address in the CCW points to a list of words (IDAWs). Each IDAW contains an address that designates a data area within real storage.

[8] There are exceptions to this statement, and many details are skipped in this description. This section assumes that you can merge this brief description with an existing understanding of I/O operations in a virtual memory environment.

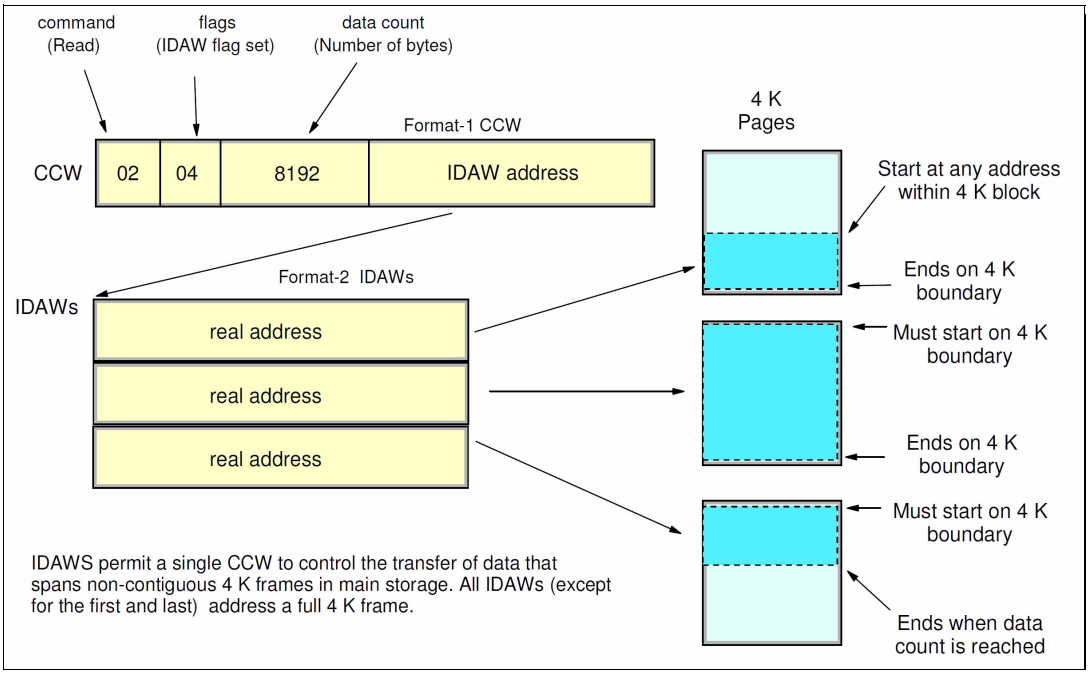


Figure 7-2 IDAW usage

The number of required IDAWs for a CCW is determined by these factors:

- The IDAW format as specified in the operation request block (ORB)
- The count field of the CCW
- The data address in the initial IDAW

For example, three IDAWS are required when these events occur:

- The ORB specifies format-2 IDAWs with 4 KB blocks.
- The CCW count field specifies 8 KB.
- The first IDAW designates a location in the middle of a 4 KB block.

CCWs with data chaining can be used to process I/O data blocks that have a more complex internal structure, in which portions of the data block are directed into separate buffer areas. This process is sometimes known as scatter-read or scatter-write. However, as technology evolves and link speed increases, data chaining techniques 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.

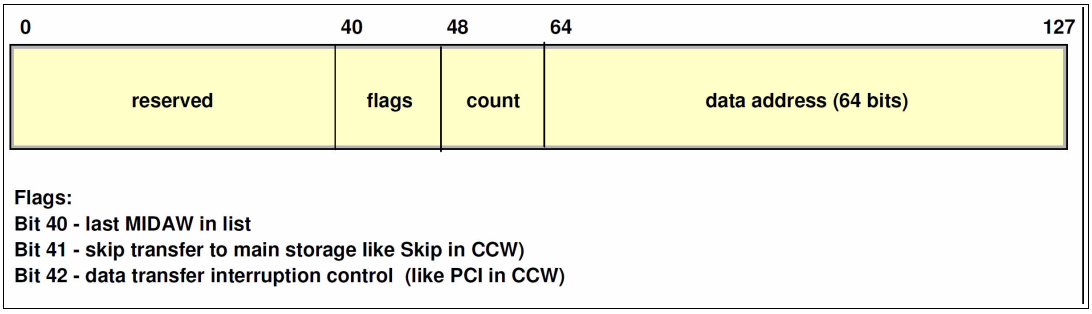


Figure 7-3 MIDAW format

An example of MIDAW usage is shown in Figure 7-4.

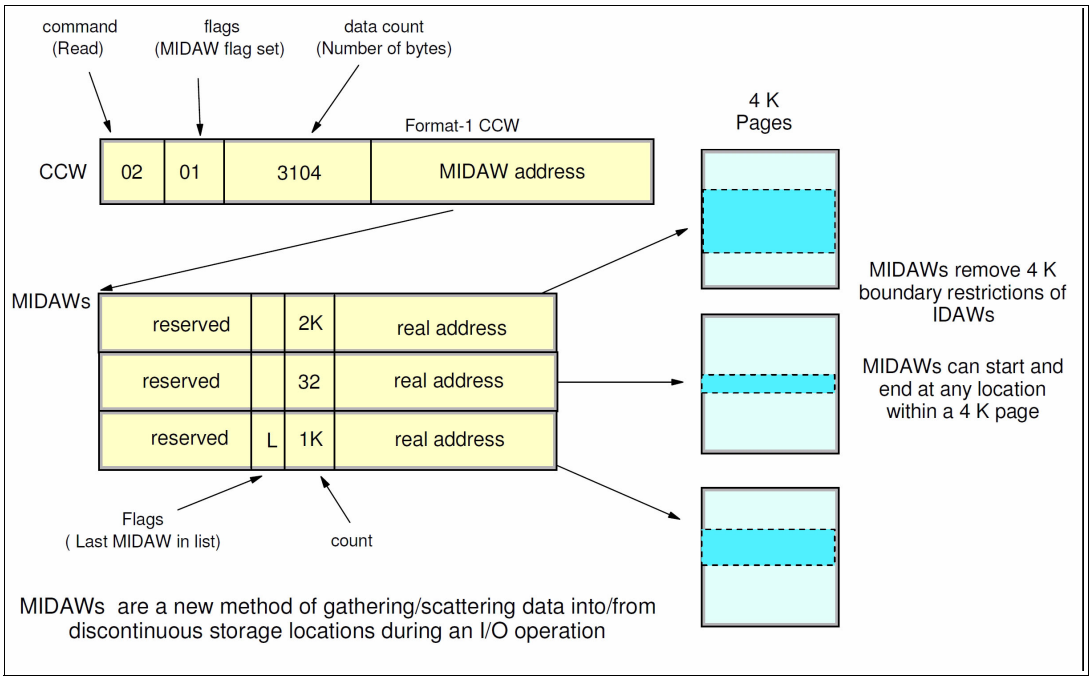


Figure 7-4 MIDAW usage

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 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 here.

Both 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 in the following paragraphs. Therefore, for example, if the DCB BLKSIZE or VSAM CI size is equal to 8192, the actual block on storage consists of 8224 bytes. The control unit itself does not distinguish between suffixes and user data. The suffix is transparent to the access method and database.

In addition to reliability, EF data sets enable three other functions:

- DFSMS striping
- Access method compression
- Extended addressability (EA)

EA is useful for creating large DB2 partitions (larger than 4 GB). Striping can be used to increase sequential throughput, or to spread random I/Os across multiple logical volumes. DFSMS striping is useful for using multiple channels in parallel for one data set. The DB2 logs are often striped to optimize the performance of DB2 sequential inserts.

Processing an I/O operation to an EF data set normally requires at least two CCWs with data chaining. One CCW is used for the 32-byte suffix of the EF data set. With MIDAW, the additional CCW for the EF data set suffix is eliminated.

MIDAWs benefit both EF and non-EF data sets. For example, to read twelve 4 K records from a non-EF data set on a 3390 track, Media Manager chains 12 CCWs together by using data chaining. To read twelve 4 K records from an EF data set, 24 CCWs are chained (two CCWs per 4 K record). Using Media Manager track-level command operations and MIDAWs, an entire track can be transferred by using a single CCW.

Performance benefits

z/OS Media Manager has I/O channel program support for implementing EF data sets, and automatically uses MIDAWs when appropriate. Most disk I/Os in the system are generated by using Media Manager.

Users of the Executing Fixed Channel Programs in Real Storage (EXCPVR) instruction can construct channel programs that contain MIDAWs. However, doing so requires that they construct an IOBE with the IOBEMIDA bit set. Users of the EXCP instruction cannot construct channel programs that contain MIDAWs.

The MIDAW facility removes the 4 K boundary restrictions of IDAWs and, 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 therefore 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 website contains material about FICON channel performance:

  http://www.ibm.com/systems/z/connectivity/

- *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 processor.

ICKDSF supports a modified format of the CPU information field that contains a two-digit LPAR identifier. ICKDSF uses the CPU information field instead of CCW reserve/release for concurrent media maintenance. It prevents multiple systems from running ICKDSF on the same volume, and at the same time allows user applications to run while ICKDSF is processing. To prevent data corruption, ICKDSF must be able to determine all sharing systems that can potentially run ICKDSF. Therefore, this support is required for z14.

Remember: The need for ICKDSF Release 17 applies even to systems that are not part of the same sysplex, or are running an operating system other than z/OS, such as z/VM.

z/OS Discovery and Auto-Configuration

z/OS Discovery and Auto Configuration (zDAC) is designed to automatically run a number of I/O configuration definition tasks for new and changed disk and tape controllers that are connected to a switch or director, when attached to a FICON channel.

The zDAC function is integrated into the existing hardware configuration definition (HCD). Clients can define a policy that can includes 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 systems that run z/OS, and reduces complexity and setup time.

zDAC applies to all FICON features that are supported on z14 when configured as CHPID type FC. See Table 7-6 on page 236 and Table 7-7 on page 237 for list of supported operating systems.

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 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 63.75-K subchannels

Servers before z9 EC reserved 1024 subchannels for internal system use, out of a maximum of 64K subchannels. Starting with z9 EC, the number of reserved subchannels was reduced to 256, increasing the number of subchannels that are available. Reserved subchannels exist only in subchannel set 0. One subchannel is reserved in each of subchannel sets 1, 2, and 3.

The informal name, *63.75-K subchannels*, represents 65280 subchannels, as shown in the following equation:

```
63 x 1024 + 0.75 x 1024 = 65280
```

The above equation is applicable for subchannel set 0. For subchannel sets 1, 2 and 3, the available subchannels are derived by using the following equation:

```
(64 X 1024) -1=65535
```

See Table 7-6 on page 236 and Table 7-7 on page 237 for list of supported operating systems.

Multiple subchannel sets

Multiple subchannel sets (MSS), first introduced in z9 EC, provides a mechanism for addressing more than 63.75-K I/O devices and aliases for FICON (CHPID types FC) on the z14, z13, z13s, zEC12, and zBC12. z196 introduced the third subchannel set (SS2). With z13, one more subchannel set (SS3) has been introduced, which expands the alias addressing by 64-K more I/O devices.

z/VM V6R3 MSS support for mirrored direct access storage device (DASD) provides a subset of host support for the MSS facility to allow using an alternative subchannel set for Peer-to-Peer Remote Copy (PPRC) secondary volumes.

See Table 7-6 on page 236 and Table 7-7 on page 237 for list of supported operating systems. For more details about channel subsystem, refer to Chapter 5, “Central processor complex channel subsystem” on page 185.

Fourth subchannel set

With z13, a *fourth subchannel set* (SS3) was introduced. Together with the second subchannel set (SS1) and third subchannel set (SS2), SS3 can be used for disk alias devices of both primary and secondary devices, and as Metro Mirror secondary devices. This set helps facilitate storage growth and complements other functions, such as extended address volume (EAV) and Hyper Parallel Access Volumes (HyperPAV).

See Table 7-6 on page 236 and Table 7-7 on page 237 for list of supported operating systems.

IPL from an alternative subchannel set

z14 supports IPL from subchannel set 1 (SS1), subchannel set 2 (SS2), or subchannel set 3 (SS3), in addition to subchannel set 0.

See Table 7-6 on page 236 and Table 7-7 on page 237 for list of supported operating systems. For more information, refer to “IPL from an alternate subchannel set” on page 190.

32 K subchannels for the FICON Express16S+ & FICON Express16S

To help facilitate growth and continue to enable server consolidation, the z14 supports up to 32 K 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 63.75 K for subchannel set 0 and 64 K (64 X 1024)-1 for subchannel sets 1, 2, and 3.

This support is exclusive to the z14, 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.

See Table 7-6 on page 236 and Table 7-7 on page 237 for list of supported operating systems.

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. See Table 7-6 on page 236 for list of supported operating systems.

FICON link incident reporting

FICON link incident reporting allows an operating system image (without operator intervention) to register link incident reports. See Table 7-6 on page 236 for list of supported operating systems.

Health Check for FICON Dynamic routing

Starting with z13, the channel microcode has been changed to support FICON dynamic routing. There is no change required in z/OS to support dynamic routing but, there can be I/O errors if the FICON switches are configured for dynamic routing despite the missing support in either the processor and/or storage controllers. Therefore, a health check is being provided that interrogates the switch to determine if dynamic routing is enabled in the switch fabric.

There is no action is required on z/OS to enable the health check for it will automatically be enabled at IPL and react to changes that might cause problems. The health check can be disabled via modify commands, PARMLIB or via SDSF.

See Table 7-6 on page 236 for list of supported operating systems. For more information about FICON Dynamic Routing (FIDR), refer to “Central processor complex I/O system structure” on page 139.

Global resource serialization (GRS) FICON CTC toleration

For some configurations depending on ESCON CTC definitions, GRS FICON CTC toleration provided with APAR OA38230 has become essential especially after removal of ESCON channel support from IBM Z starting with zEC12.

See Table 7-6 on page 236 for list of supported operating systems.

Increased performance for the FCP protocol:

The FCP LIC is modified to help provide increase I/O operations per second for both small and large block sizes, and to support 16-Gbps link speeds.

For more information about FCP channel performance, see the performance technical papers on the IBM Z I/O connectivity website at:

http://www.ibm.com/systems/z/hardware/connectivity/fcp_performance.html

The FCP protocol is supported by z/VM, z/VSE, and Linux on z Systems. See Table 7-6 on page 236 and Table 7-7 on page 237 for list of supported operating systems.

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 consisting of CRC (Cyclic Redundancy Checking), LBA (Logical Block Address), and host application tags to each sector of FB data on a logical volume.

IBM Z support applies to FCP channels only. See Table 7-6 on page 236 and Table 7-7 on page 237 for list of supported operating systems.

N_Port ID Virtualization

N_Port ID Virtualization (NPIV) allows multiple system images (in LPARs or z/VM guests) to use a single FCP channel as though each were the sole user of the channel. This feature, first introduced with z9 EC, can be used with supported FICON features on z14 servers. See Table 7-6 on page 236 and Table 7-7 on page 237 for list of supported operating systems.

Worldwide port name tool

Part of the z14 system installation is the pre-planning of the SAN environment. IBM has a stand-alone tool to assist with this planning before the installation.

The capabilities of the WWPN are extended to calculate and show WWPNs for both virtual and physical ports ahead of system installation.

The tool assigns WWPNs to each virtual FCP channel/port by using the same WWPN assignment algorithms that a system uses when assigning WWPNs for channels using NPIV. Therefore, the SAN can be set up in advance, allowing operations to proceed much faster after the server is installed. In addition, the SAN configuration can be retained instead of altered by assigning the WWPN to physical FCP ports when a FICON feature is replaced.

The WWPN tool takes a `.csv` file that contains the FCP-specific I/O device definitions and creates the WWPN assignments that are required to set up the SAN. A binary configuration file that can be imported later by the system is also created. The `.csv` file can either be created manually, or exported from the HCD/HCM. See Table 7-6 on page 236 and Table 7-7 on page 237 for list of supported operating systems.

The WWPN tool is applicable to all FICON channels that are defined as CHPID type FCP (for communication with SCSI devices) on z14. It is available for download at the Resource Link at the following website:

http://www.ibm.com/servers/resourcelink/

Note: An optional feature can be ordered for WWPN persistency before shipment to keep the same I/O serial number on the new CPC. Current information must be provided during the ordering process.

7.4.5 Networking Features and Functions

This section covers the following sub-topics:

- “10GbE RoCE Express2” on page 274
- “10GbE RoCE Express” on page 275
- “Shared Memory Communication - Direct Memory Access (SMC-D)” on page 275
- “HiperSockets Completion Queue” on page 275
- “HiperSockets Virtual Switch Bridge” on page 275
- “HiperSockets Multiple Write Facility” on page 276
- “HiperSockets support of IPV6” on page 276
- “HiperSockets Layer 2 support” on page 276
- “HiperSockets network traffic analyzer for Linux on z Systems” on page 276
- “OSA-Express6S 10-Gigabit Ethernet LR and SR” on page 277
- “OSA-Express5S 10-Gigabit Ethernet LR and SR” on page 277
- “OSA-Express6S Gigabit Ethernet LX and SX” on page 277
- “OSA-Express5S Gigabit Ethernet LX and SX” on page 277
- “OSA-Express6S 1000BASE-T Ethernet” on page 277
- “OSA-Express5S 1000BASE-T Ethernet” on page 278
- “OSA-Express4S 1000BASE-T Ethernet” on page 278
- “OSA-Express NCP support (OSN)” on page 279
- “OSA-ICC (Integrated Console Controller)” on page 279
- “Checksum offload for in QDIO mode (CHPID type OSD)” on page 279
- “Querying and displaying an OSA configuration” on page 280
- “QDIO data connection isolation for z/VM” on page 280
- “QDIO interface isolation for z/OS” on page 280
- “QDIO optimized latency mode” on page 280
- “QDIO Diagnostic Synchronization” on page 281
- “Adapter interruptions for QDIO” on page 281
- “Inbound workload queuing (IWQ) for OSA” on page 281
- “VLAN management enhancements” on page 282
- “GARP VLAN Registration Protocol (GVRP)” on page 282
- “Link aggregation support for z/VM” on page 282
- “Multi-VSwitch Link Aggregation” on page 283
- “Large send for IPv6 packets” on page 283
- “OSA Dynamic LAN idle” on page 283
- “OSA Layer 3 virtual MAC for z/OS environments” on page 284
- “Network Traffic Analyzer” on page 284

10GbE RoCE Express2

z14 introduces the next generation of RoCE technology with the IBM 10GbE RoCE Express2 providing 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, in conjunction with an OSA feature, enables shared memory communications between two CPCs using a shared switch.

RoCE Express2 provides increased virtualization (sharing capability) by supporting 127 Virtual Functions (VFs) per physical port for a total of 254 VFs per PCHID allowing 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 transparent to both upper layers of the CS (the SMC-R and TCP/IP stack) and application software (exploiting 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 will also be exploited by Linux on z Systems 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 239 and Table 7-9 on page 241 for list of supported operating systems.

10GbE RoCE Express

z14 servers support carrying forward the 10GbE RoCE Express feature which provides support to the second port on the adapter and sharing the ports to up 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 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 server or for server-to-server communication in a multiple CPC environment.

See Table 7-8 on page 239 and Table 7-9 on page 241 for list of supported operating systems. For more information, see Appendix D, “Shared Memory Communications” on page 457.

Shared Memory Communication - Direct Memory Access (SMC-D)

The SMC-D feature, first introduced with z13 servers, maintains the socket-API transparency aspect of SMC-R so that applications that use TCPI/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 preserving TCP/IP Qualities of Service for connection management purposes.

See Table 7-8 on page 239 and Table 7-9 on page 241 for list of supported operating systems. For more information, see Appendix D, “Shared Memory Communications” on page 457.

HiperSockets Completion Queue

The HiperSockets Completion Queue function is implemented on z14, 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. See Table 7-8 on page 239 and Table 7-9 on page 241 for list of supported operating systems.

HiperSockets Virtual Switch Bridge

The HiperSockets Virtual Switch Bridge is implemented on z14, 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:

- Other guest virtual machines on the virtual switch
- External network hosts through the virtual switch OSA UPLINK port

See Table 7-8 on page 239 and Table 7-9 on page 241 for list of supported operating systems.

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 174. See Table 7-8 on page 239 for list of supported operating systems.

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. See Table 7-8 on page 239 and Table 7-9 on page 241 for list of supported operating systems.

HiperSockets Layer 2 support

For flexible and efficient data transfer for IP and non-IP workloads, the HiperSockets internal networks on z14 can support two transport modes. These modes are Layer 2 (Link Layer) and the current Layer 3 (Network or IP Layer). Traffic can be Internet Protocol (IP) Version 4 or Version 6 (IPv4, IPv6) or non-IP (AppleTalk, DECnet, IPX, NetBIOS, or SNA).

HiperSockets devices are protocol-independent and Layer 3-independent. Each HiperSockets device has its own Layer 2 Media Access Control (MAC) address. This MAC address allows the use of applications that depend on the existence of Layer 2 addresses, such as Dynamic Host Configuration Protocol (DHCP) servers and firewalls.

Layer 2 support can help facilitate server consolidation. Complexity can be reduced, network configuration is simplified and intuitive, and LAN administrators can configure and maintain the mainframe environment the same way as they do a non-mainframe environment.

See Table 7-8 on page 239 and Table 7-9 on page 241 for list of supported operating systems.

HiperSockets network traffic analyzer for Linux on z Systems

HiperSockets network traffic analyzer (HS NTA), introduced with IBM System z10, provides support for tracing Layer2 and Layer3 HiperSockets network traffic in Linux on z Systems. This support allows Linux on z Systems to control the trace for the internal virtual LAN to capture the records into host memory and storage (file systems).

Linux on z Systems tools can be used to format, edit, and process the trace records for analysis by system programmers and network administrators.

OSA-Express6S 10-Gigabit Ethernet LR and SR

z14 introduces an Ethernet technology refresh with OSA-Express6S 10-Gigabit Ethernet features to be installed in the PCIe I/O drawer, 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.

See Table 7-8 on page 239 and Table 7-9 on page 241 for list of supported operating systems.

OSA-Express5S 10-Gigabit Ethernet LR and SR

The OSA-Express5S 10-Gigabit Ethernet feature, introduced with the zEC12 and zBC12, is installed exclusively in the PCIe I/O drawer. Each feature has one port, which is defined as CHPID type OSD supporting the queued direct input/output (QDIO) architecture for high-speed TCP/IP communication.

See Table 7-8 on page 239 and Table 7-9 on page 241 for list of supported operating systems.

OSA-Express6S Gigabit Ethernet LX and SX

z14 introduces an Ethernet technology refresh with OSA-Express6S Gigabit Ethernet features to be installed in the PCIe I/O drawer, 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.

See Table 7-8 on page 239 and Table 7-9 on page 241 for list of supported operating systems.

OSA-Express5S Gigabit Ethernet LX and SX

The OSA-Express5S Gigabit Ethernet feature is installed exclusively in the PCIe I/O drawer. Each feature has one PCIe adapter and two ports. The two ports share a channel path identifier (CHPID type OSD exclusively). Each port supports attachment to a 1 Gigabit per second (Gbps) Ethernet local area network (LAN). The ports can be defined as a spanned channel, and can be shared among LPARs and across logical channel subsystems.

Note: Operating system support is required to recognize and use the second port on the OSA-Express5S Gigabit Ethernet feature.

See Table 7-8 on page 239 and Table 7-9 on page 241 for list of supported operating systems.

OSA-Express6S 1000BASE-T Ethernet

z14 introduces an Ethernet technology refresh with OSA-Express6S 1000BASE-T Ethernet features to be installed in the PCIe I/O drawer, 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.

See Table 7-8 on page 239 and Table 7-9 on page 241 for list of supported operating systems.

Statement of Direction[a]:

- *OSA-Express6S 1000BASE-T adapters (#0426)* will be the last generation of OSA 1000BASE-T adapters to support connections operating at 100 Mb/second link speed. Future OSA-Express 1000BASE-T adapter generations will support operation only at 1000 Mb/second (1Gb/s) 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 has 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.

See Table 7-8 on page 239 and Table 7-9 on page 241 for list of supported operating systems.

OSA-Express4S 1000BASE-T Ethernet

The OSA-Express4S 1000BASE-T Ethernet feature is installed exclusively in the PCIe I/O drawer. Each feature has one PCIe adapter and two ports. The two ports share a CHPID, which is defined as OSC, OSD, 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 and OSN
- 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-Express4S 1000BASE-T Ethernet feature.

See Table 7-8 on page 239 and Table 7-9 on page 241 for list of supported operating systems.

OSA-Express NCP support (OSN)

Fulfillment of Statement of Directions:

- *Removal of support for configuring OSN CHPID types*: The IBM z13 and z13s has been the last z Systems and IBM Z servers to support configuring OSN CHPID types. IBM z14 servers do not support CHPID Type = OSN.
  OSN CHPIDs were used to communicate between an operating system instance running in one logical partition and the IBM Communication Controller for Linux on z Systems (CCL) product in another logical partition on the same CPC. See announcement letter #914-227 dated 12/02/2014 for details regarding withdrawal from marketing for the CCL product.

OSA-ICC (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 has multiple LPAR support, both as shared or spanned channels.

With the OSA-ICC function, 3270 emulation for console session connections is integrated in the z14 through a port on the OSA-Express6S 1000BASE-T, OSA-Express5S 1000BASE-T or OSA-Express4S 1000BASE-T features.

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.

In order to improve security of console operations and to provide a secure, validated connectivity; staring with z13 GA2 (Driver level 27), OSA-ICC supports TLS/SSL (Transport Layer Security/Secure Sockets Layer) with Certificate Authentication.

Note: OSA-ICC supports up to 48 *secure* sessions per CHPID (the overall maximum of 120 connections is unchanged).

See Table 7-8 on page 239 and Table 7-9 on page 241 for list of supported operating systems.

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.

See Table 7-8 on page 239 and Table 7-9 on page 241 for list of supported operating systems.

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 a TCP/IP operator command called **`display OSAINFO`**.

Using **`display OSAINFO`** allows the operator to monitor and verify the current OSA configuration and helps improve the overall management, serviceability, and usability of OSA-Express cards.

The **`display OSAINFO`** command is exclusive to z/OS, and applies to CHPID type OSD. See Table 7-8 on page 239 for list of supported operating systems.

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. See Table 7-8 on page 239 and Table 7-9 on page 241 for list of supported operating systems.

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. See Table 7-8 on page 239 and Table 7-9 on page 241 for list of supported operating systems.

QDIO optimized latency mode

QDIO optimized latency mode (OLM) can help improve performance for applications that have a critical requirement to minimize response times for inbound and outbound data.

OLM optimizes the interrupt processing in the following manner:

- For inbound processing, the TCP/IP stack looks more frequently for available data to process. This process ensures that any new data is read from the OSA-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.

See Table 7-8 on page 239 for list of supported operating systems.

QDIO Diagnostic Synchronization

QDIO Diagnostic Synchronization enables system programmers and network administrators to coordinate and simultaneously capture both software and hardware traces. It allows z/OS to signal OSA-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 when in QDIO mode (CHPID type OSD). See Table 7-8 on page 239 for list of supported operating systems.

Adapter interruptions for QDIO

Linux on z Systems and z/VM work together to provide performance improvements by using extensions to the QDIO architecture. Adapter interruptions, first added to z/Architecture with HiperSockets, provide an efficient, high-performance technique for I/O interruptions to reduce path lengths and processor usage. These reductions are in both the host operating system and the adapter (supported OSA-Express cards when using CHPID type OSD).

In extending the use of adapter interruptions to OSD (QDIO) channels, the processor utilization to handle a traditional I/O interruption is reduced. This benefits OSA-Express TCP/IP support in z/VM, z/VSE, and Linux on z Systems. See Table 7-8 on page 239 and Table 7-9 on page 241 for list of supported operating systems.

Inbound workload queuing (IWQ) 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 has unique service and processing requirements. The IWQ function allows z/OS to preassign the appropriate processing resources for each input queue. This approach allows multiple concurrent z/OS processing threads to process each unique input queue (workload), avoiding traditional resource contention. 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-Express4S[9] features defined as CHPID type OSD. See Figure 7-5 below depicting OSA inbound workload queuing (IWQ).

[9] Only OSA-Express4S 1000BASE-T cards are supported on z14 as carry forward.

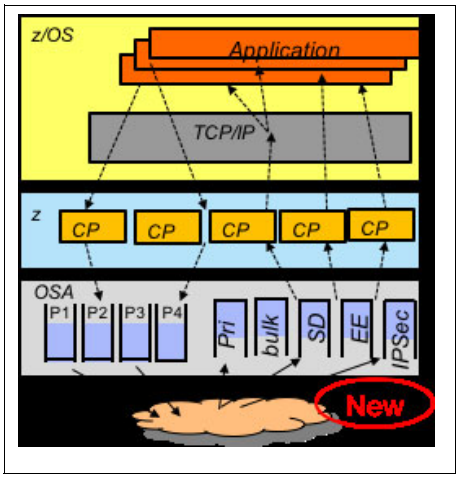


Figure 7-5 OSA inbound workload queuing (IWQ)

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.

See Table 7-8 on page 239 and Table 7-9 on page 241 for list of supported operating systems.

VLAN management enhancements

VLAN management enhancements are valid for supported OSA-Express features on z14 defines as CHPID type OSD. See Table 7-8 on page 239 and Table 7-9 on page 241 for list of supported operating systems.

GARP VLAN Registration Protocol (GVRP)

All OSA-Express6S, OSA-Express5S and OSA-Express4S[9] features support VLAN prioritization, a component of the IEEE 802.1 standard. GARP VLAN Registration Protocol (GVRP) support allows an OSA-Express port to register or unregister its VLAN IDs with a GVRP-capable switch and dynamically update its table as the VLANs change. This process simplifies the network administration and management of VLANs because manually entering VLAN IDs at the switch is no longer necessary. See Table 7-8 on page 239 and Table 7-9 on page 241 for list of supported operating systems.

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-Express6S, OSA-Express5S or OSA-Express4S[9] 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). See Table 7-8 on page 239 and Table 7-9 on page 241 for list of supported operating systems.

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. See Table 7-8 on page 239 and Table 7-9 on page 241 for list of supported operating systems.

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-Express6S, OSA-Express5S and OSA-Express4S[9] features (CHPID type OSD) on z14, 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 added TCP checksum on large send which reduces the cost (CPU time) of error detection for large send.

See Table 7-8 on page 239 and Table 7-9 on page 241 for list of supported operating systems.

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 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, avoiding thrashing and frequent updates to the OAT. Based on the TCP/IP settings, OSA holds the packets before presenting them to the host. A dynamic setting is designed to avoid or minimize host interrupts.

OSA Dynamic LAN idle is supported by the OSA-Express6S, OSA-Express5S, and OSA-Express4S[9] features on z14 when in QDIO mode (CHPID type OSD). Table 7-8 on page 239 for list of supported operating systems.

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-Express6S, OSA-Express5S and OSA-Express4S[9] features on z14 when in QDIO mode (CHPID type OSD). See Table 7-8 on page 239 for list of supported operating systems.

Network Traffic Analyzer

The z14 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-Express6S, OSA-Express5S and OSA-Express4S[9] features on z14 when in QDIO mode (CHPID type OSD). See Table 7-8 on page 239 for list of supported operating systems.

7.4.6 Cryptography Features and Functions Support

IBM z14 provides two 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 listed in the following sections. Obtain and review the current PSP buckets to ensure that the latest support levels are known and included as part of the implementation plan.

This section covers the following sub-topics:

- “CP Assist for Cryptographic Function” on page 284
- “Crypto Express6S” on page 285
- “Crypto Express5S (carry forward on z14)” on page 285
- “Regional Crypto Enablement (RCE)” on page 285
- “Web deliverables” on page 285
- “z/OS Integrated Cryptographic Service Facility” on page 286
- “RMF Support for Crypto Express6” on page 286
- “z/OS Data Set Encryption” on page 287
- “z/VM encrypted hypervisor paging (Encrypted paging support)” on page 287
- “z/TPF transparent database encryption” on page 288

CP Assist for Cryptographic Function

Central Processor Assist for Cryptographic Function (CPACF), standard on every z14 core, now supports pervasive encryption. Simple policy controls allow business to enable encryption to protect data in mission-critical databases without needing to stop the database or recreate database objects. Database administrators can leverage z/OS Dataset Encryption, z/OS Coupling Facility Encryption, z/VM encrypted hypervisor paging, and z/TPF transparent database encryption, which exploit performance enhancements in the hardware.

In z14, CPACF supports the following features:

- The Advanced Encryption Standard (AES, symmetric encryption)
- The Data Encryption Standard (DES, symmetric encryption)
- The 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 203.

See Table 7-10 on page 243 and Table 7-11 on page 244 for list of supported operating systems.

Crypto Express6S

Introduced with z14, Crypto Express6S complies with below 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 not only by operating system and release but also by the way the card is configured as a coprocessor or an accelerator. For more information, see 6.5, “Crypto Express6S” on page 207. See Table 7-10 on page 243 and Table 7-11 on page 244 for list of supported operating systems.

Crypto Express5S (carry forward on z14)

Support of Crypto Express5S functions varies not only by operating system and release but also by the way the card is configured as a coprocessor or an accelerator. See Table 7-10 on page 243 and Table 7-11 on page 244 for list of supported operating systems.

Regional Crypto Enablement (RCE)

Starting with z13 GA2, IBM has enabled geo-specific cryptographic support that is supplied by IBM approved vendors. China is the first geography to exploit 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 support will reserve the I/O slot(s) for the IBM approved vendor-supplied cryptographic card(s). Clients will need to directly contact the IBM approved vendor for purchasing information.

RCE is a framework to enable the integration of IBM certified 3rd party cryptographic hardware for regional or industry encryption requirements and supports the use of cryptography algorithms and equipment from selected providers in conjunction with IBM Z in specific countries. Support for exploiting international algorithms (AES, DES, RSA, ECC) with regional crypto devices (supporting regional algorithms like SMx) is added to the ICSF PKCS#11 services.

See Table 7-10 on page 243 and Table 7-11 on page 244 for list of supported operating systems.

Web deliverables

For web-deliverable code on z/OS, see the z/OS downloads website:

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

For Linux on z Systems, support is delivered through IBM and the distribution partners. For more information, see Linux on z Systems on the IBM developerWorks® website:

http://www.ibm.com/developerworks/linux/linux390/

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 HCR77C1 - Cryptographic Support for z/OS V2R1 - z/OS V2R3

ICSF Web Deliverable HCR77C1 provides support for;

- Usage and administration of Crypto Express6S feature
  - May be configured as an accelerator (CEX6A), a CCA coprocessor (CEX6C), or an EP-11 coprocessor (CEX6P)
- Coprocessor in PCI-HSM Compliance Mode
  - Enablement will require TKE 9.0
- z14 CPACF support covered in “CP Assist for Cryptographic Function” on page 284

Listed below are the software enhancements in ICSF Web Deliverable HCR77C1 when running on z14 server:

- *Crypto Usage Statistics:* When enabled, ICSF aggregates statistics related to crypto workloads and logs to an SMF record.
- *Panel-based CKDS Administration:* ICSF has added an ISPF, panel-driven interface which allows interactive administration (View, Create, Modify, Delete) of CKDS keys.
- *CICS End User Auditing:* When enabled, ICSF retrieves the CICS end user identity and include it as a log string in the SAF resource check. The end user identity is not checked for access to the resource. It is simply included in the resource check (SMF Type 80) records logged for any of the ICSF SAF classes protecting crypto keys and services (CSFKEYS, XCSFKEY, CRYPTOZ, CSFSERV)

For a list of ICSF versions and FMID cross-references, see the Technical Documents website:

http://www.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/TD103782

For PTFs that allow prior levels of ICSF to coexist with the Cryptographic Support for z/OS 2.1 - z/OS V2R3 (HCR77C1) web deliverable, check below FIXCAT:

```
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 exploit performance data for the new Crypto Express6 (CEX6) card. RMF supports all valid card configurations on z14 and provides CEX6 crypto activity data in:

- SMF type 70 subtype 2 records
- RMF Postprocessor Crypto Activity Report

See Table 7-10 on page 243 for list of supported operating systems.

z/OS Data Set Encryption

Aligned with IBM Z Pervasive Encryption initiative, IBM announced[10] plans to deliver application transparent, policy-controlled dataset encryption in IBM z/OS.

Policy driven z/OS Data Set Encryption enables users to;

- 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 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 exploit z/OS Data Set Encryption.

With z/OS Data Set Encryption DFSMS enhances data security with support for data set level encryption via DFSMS access methods. This function is designed to give users the ability to encrypt their data sets without changing their application programs. DFSMS users have the ability to identify which data sets require encryption via 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/restore, migration/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 used. Considering the significant enhancements introduced with z14, the encryption mode of XTS will be 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. See Table 7-10 on page 243 for list of supported operating systems.

Crypto Analytics Tool for Z

The IBM CAT is an analytics solution that collects data on your z/OS cryptographic infrastructure, present reports and can analyze if there are vulnerabilities. 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. Using 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, planned to be available December 15, 2017, z/VM V6.4 will provide support for encrypted paging, in support of the z14 pervasive encryption philosophy of encrypting all data in flight and at rest. Ciphering will occur as data moves between active memory and a paging volume owned by z/VM. Included in the support is the ability to dynamically control whether a running z/VM system is encrypting this data. This support will protect guest paging data from administrators and/or users with access to volumes. Enabled with AES encryption, z/VM Encrypted Paging comes with very low overhead via CPACF.

See Table 7-10 on page 243 for list of supported operating systems.

[10] Statement of Direction in Software Announcement Letter 216-392, dated October 4, 2016. Available at https://www.ibm.com/common/ssi/rep_ca/2/897/ENUS216-392/ENUS216-392.PDF

z/TPF transparent database encryption

Shipped in August 2016, z/TPF at-rest Data Encryption provides following features and benefits:

- Automatic encryption of at-rest data using AES CBC (128 or 256)
- No application changes required
- Database level encryption 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 an existing DB from un-encrypted to encrypted state or change the encryption key/algorithm for a given DB while transactions are flowing (no DB downtime)

Pervasive Encryption for Linux on z Systems

Pervasive encryption for Linux on z Systems combines the full power of Linux ecosystem with z14 capabilities via the support of the following features:

- Kernel Crypto - z14 CPACF
- LUKS dm-crypt Protected-Key CPACF
- libica & openssl - z14 CPACF and acceleration of RSA handshakes using SIMD
- Secure Service Container - High security virtual appliance deployment infrastructure

Protection of data at-rest

By leveraging the integration of industry-unique hardware accelerated CPACF encryption into the standard Linux components, users can achieve optimized encryption transparently in order to prevent raw key material from being visible to OS and applications.

Protection of data in-flight

Due to the potential costs and overheads, most of the organizations avoid using host-based network encryption today. By exploiting enhanced CPACF & SIMD on z14, TLS & IPSec can leverage hardware performance gains while enjoying transparent enablement. Reduced cost of encryption enables broad use of network encryption.

7.4.7 Special-purpose Features and Functions

This section covers the zEnterprise Data Compression Express.

zEnterprise Data Compression Express

The growth of data that must be captured, transferred, and stored for a long time is unrelenting. Software-implemented compression algorithms are costly in terms of processor resources, and storage costs are not negligible either.

zEnterprise Data Compression (zEDC) Express is an optional feature that is available on z14, z13, z13s, zEC12, and zBC12 servers that addresses those requirements by providing hardware-based acceleration for data compression and decompression. zEDC provides data compression with lower CPU consumption than the compression technology that previously was available on Z 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. Software decompression is slow and uses considerable processor resources, so it is not recommended for production environments.

zEDC supports QSAM/BSAM (non-VSAM) data set compression via any of the following ways:

- Data class level: Two new values, zEDC Required (ZR) and zEDC Preferred (ZP), can be set with the **COMPACTION** option in the data class.
- System Level: Two new values, zEDC Required (ZEDC_R) and zEDC Preferred (ZEDC_P), can be specified with the **COMPRESS** parameter found in the IGDSMSXX member of the SYS1.PARMLIB data set

Data class takes precedence over system level. See Table 7-12 on page 245 for list of supported operating systems.

For more information about zEDC Express, see Appendix F, “IBM zEnterprise Data Compression Express” on page 493.

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 needed depends on z/OS release. Although z14 servers *do not* require any ‘*functional*’ software, it is recommended to install all z14 service prior to upgrading to the new server. Refer to Table 7-16 on page 289 listing support matrix for z/OS releases and the Z servers supporting them.

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	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[c]	09/2021[c]
V2R2		X	X	X	X	X	09/2020[c]	09/2023[c]
V2R3[d]				X	X	X	09/2022[c]	09/2025[c]

a. Server has been withdrawn from marketing
b. 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. Planned. All statements regarding IBM's plans, directions, and intent are subject to change or withdrawal without notice.
d. Planned to be Generally Available in September 2017

7.5.1 General guidelines

The IBM z14 introduces the latest IBM Z technology. Although support is provided by z/OS starting with z/OS V1R13, the capabilities and use of z14 depends on the z/OS release. Also, web deliverables[11] 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.

[11] E.g. Exploitation of Crypto Express6S requires the Cryptographic Support for z/OS V2R1 - z/OS V2R3 web deliverable.

- Migrate to an STP-only network before introducing a z14 into a sysplex.
- 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 and plan accordingly
- Determine the changes in IOCP, HCD and HCM to support defining z14 configuration and the new features and functions it introduces.
- Ensure that none of the new z/Architecture Machine Instructions (mnemonics) introduced with z14 are colliding with the names of Assembler macro instructions you use[12].
- Check the use of Use of **MACHMIG** Statements in **LOADxx** PARMLIB

7.5.2 Hardware PSP Buckets and Fix Categories

Support for z14 is provided via a combination of web deliverables and PTFs which are documented in PSP Bucket: **Upgrade = 3906DEVICE, Subset = 3906/ZOS**

Base support[13] has fixes that are required to run z/OS on the IBM z14 server. They are identified by: **IBM.Device.Server.nnn-3906.RequiredService**

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.nnn-3906.Exploitation**

Recommended service is identified by: **IBM.Device.Server.nnn-3906.RecommendedService**

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 exploit the High Performance FICON function:
IBM.Function.zHighPerformanceFICON

PTFs that allow prior 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 SMP/E **REPORT MISSINGFIX** command to determine whether any FIXCAT APARs exist that are applicable and have not yet been installed, and whether any SYSMODs are available to satisfy the missing FIXCAT APARs. Visit below URL for more information about IBM Fix Category Values and Descriptions:

http://www.ibm.com/systems/z/os/zos/features/smpe/fix-category.html

7.5.3 Coupling links

z14 servers only support active participation in the same Parallel Sysplex with z13, z13s, IBM zEC12 and zBC12. Configurations with z/OS on one of these servers can add a z14 Server to their Sysplex for either a z/OS or a Coupling Facility image. Configurations with a Coupling Facility on one of these servers can add a z14 Server to their Sysplex for either a z/OS or a Coupling Facility image. z14 does not support participating in a Parallel Sysplex with System z196 and earlier systems.

[12] Check IBM technote available at http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/PRS5289
[13] nnn = Server Name, z14

Each system can use, or not use, internal coupling links, InfiniBand 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. See 4.7.4, “Parallel Sysplex connectivity” on page 175 for supported coupling link technologies on z14 and consult the *Coupling Facility Configuration Options* white paper, which is available at the following website:

http://www.ibm.com/systems/z/advantages/pso/whitepaper.html

7.5.4 z/OS XL C/C++ considerations

z/OS V2R1 with PTFs or higher is required to use the latest level (12) of the following two 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.
- **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.

Listed below are new functions that provide performance improvements for applications by exploiting new z14 instructions.

- Vector Programming Enhancements
- New z14 hardware instruction support
- Packed Decimal support using vector registers
- Auto-SIMD enhancements to make use of new data types

To enable the use of new functions, Specify **ARCH(12)** and **VECTOR** for compilation. Please note that the binaries produced by the compiler on z14 can only be executed on z14 and above, as it will make use of the vector facility on z14 for new functions. Using older versions of the compiler on z14 would not enable new functions.

For more information about the **ARCHITECTURE**, **TUNE** and **VECTOR** compiler options, see the *z/OS V2R2.0 XL C/C++ User’s Guide,* SC09-4767 manual.

Important: Use the previous Z **ARCHITECTURE** or **TUNE** options for C/C++ programs if the same applications run on both the z14 and on previous IBM Z servers. However, if C/C++ applications run only on z14 servers, use the latest **ARCHITECTURE** and **TUNE** options to ensure that the best performance possible is delivered through the latest instruction set additions.

For more information, see *Migration from z/OS V2R1 to z/OS V2R2*, GA32-0889.

7.5.5 z/OS V2.3 Preview

IBM has announced preview of IBM z/OS Version 2 Release 3 - Engine for digital transformation via Announcement letter 217-085 on February 21, 2017. Focusing on three critical areas, Security, Simplification, and Cloud, z/OS V2.3 is planning to provide a simple, transparent, and consumable approach to enable extensive encryption of data, simplify the overall management of the z/OS ecosystem to increase productivity, and provide a simple, consumable approach for self-service provisioning and rapid delivery of software as a

service, while enabling for the API economy. Planned availability date for z/OS 2.3 is September 29, 2017.

Below are key points to consider prior to migrating z/OS 2.3 to IBM z14:

- IBM z/OS V2.3 with IBM z14 will require a minimum of 8GB of memory: When running as a z/VM guest or on an IBM System z Personal Development Tool, a minimum of 2GB will be required for z/OS V2.3. If the minimum is not met, a warning WTOR will be issued at IPL. Continuing with less than the minimum memory could impact 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 8GB.
- Dynamic splitting and merging of Coordinated Timing Network (CTN) is available with z14.
- HyperSwap support is enhanced to allow RESERVE processing. When a system executes a request to swap to secondary devices managed by HyperSwap, z/OS will detect when RESERVEs are held and will ensure that the devices being swapped also hold the RESERVE. This enhancement is provided with collaboration from z/OS, GDPS HyperSwap and CSM HyperSwap.
- 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 using I/O processors (SAPs) on next generation processors. This will be designed to eliminate the CPU cost for this operation and help improve performance of RSM first reference page fault processing, and of system services such as IARV64 and STORAGE OBTAIN.
- z/OS V2.3 exploits a DS8000 interface for z/OS environments, as announced in Hardware Announcement 116-044, dated May 31, 2016, that will release any extents on a thinly provisioned extent space-efficient (ESE) volume that only contains empty space. Thin provisioning with ESE volumes in z/OS environments is designed to simplify storage configurations and management by standardizing device sizes without wasting space, and to reduce replication and tiering overheads by avoiding movement of unallocated extents. Additionally, spare storage capacity across sysplexes or between storage groups potentially can be shared. A new DFSMSdss space release command, SPACEREL, will exploit the DS8000 interface, and a new RACF FACILITY class profile will be supported to protect the new SPACEREL command. This function is available on z/OS V2.1 and V2.2 with PTFs for OA50675.
- PDS/E member names processing is enhanced to preserve mixed case names and provide name inversion with a new keyword provided in the NFS Server. The new "strictmaplower" option may be set as site attribute, on mount and file command level and has a priority over maplower/nomaplower setting.
- 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 and provide 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.

7.6 z/VM migration considerations

IBM z14 supports z/VM 6.3 and z/VM 6.4.

z/VM V6.3 cannot be installed directly on an IBM z14 server. Instead, z/VM V6.3 must be installed on a prior Z server and the PTFs for APARs VM65942, VM65921, and VM65922 must be applied prior to moving the image to z14. After applying required service and before IPLing z/VM V6.3 on an IBM z14 server, an updated SAPL must be installed by using the updated SALIPL MODULE provided by the PTF for APAR VM65856. Otherwise, z/VM V6.3

will not successfully IPL on a z14 server. In addition, after applying the PTFs for APARs VM65921 and VM65922, the stand-alone dump program must be re-installed using the SDINST utility.

z/VM V6.4 can be installed directly on a z14 server with an image obtained from IBM after August 25, 2017. The PTF for APAR VM65942 must be applied immediately after installing z/VM V6.4 and prior to doing any configuration of the new z/VM system.

See Table 7-17 below for z/VM Release Status Summary.

Table 7-17 z/VM Release Status Summary

z/VM Level	General Availability	End of Marketing	End of Service	Minimum Processor Level	Maximum Processor Level
6.4	November, 2016	Not announced	Not announced	z196 & z114	-
6.3	July, 2013	November, 2016	December, 2017[a]	z10	-
6.2	December, 2011	July, 2013	July, 2017[b]	z10	z13
5.4	September, 2008	March, 2012	December, 2017[c]	zSeries 800 & zSeries 900	zEC12

a. Announced February 3, 2015
b. Announced February 2, 2016
c. Announced August 2, 2016

7.6.1 ESA/390-compatibility mode for guests:

IBM z14 no longer supports the full ESA/390 architectural mode. However, IBM z14 does provide ESA/390-compatibility mode, an environment supporting a subset of DAT-off ESA/390 applications in a hybrid architectural mode. z/VM will provide the support necessary for DAT-off guests to run in this new compatibility mode, which allows guests such as CMS, GCS, and those that start in ESA/390 mode briefly before switching to z/Architecture mode to continue to run on IBM z14.

The available PTF for APAR VM65976 provides infrastructure support for ESA/390 compatibility mode within z/VM V6.2, V6.3, and V6.4, and must be installed on all members of an SSI cluster before any z/VM V6.3 or z/VM V6.4 member of the cluster is run on an IBM z14 server.

7.6.2 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 servers.

7.7 z/VSE migration considerations

As covered in “z/VSE” on page 230, IBM z14 supports z/VSE 6.1, z/VSE 5.2 and z/VSE 5.1[14].

[14] z/VSE 5.1 is end of support since June 2016. It can be IPL’ed on z14 after applying APAR DY47654 (PTF UD54170).

z/VSE 6.2 which was previewed with the announcement letter 217-091 on April, 2017; will also be supported on z14. The focus of z/VSE V6.2 will be online transaction processing, security, and connectivity to improve integration of z/VSE in a heterogeneous environment. Planned availability date for z/VSE 6.2 is 4th quarter of 2017.

Below are general guidelines to consider when migrating z/VSE environment to z14 servers:

- Collect reference information before migration
  - Baseline reflecting the current status, e.g. performance data, CPU utilization of reference workload, I/O activity, elapsed times, etc.
  - This information is not only required to size z14 but also the only way to compare workload characteristics after migration
  - See “z/VSE Release and Hardware Upgrade document for more details
- Apply required maintenance for z14
  - Review PSP (Preventive Service Planning) bucket 3906DEVICE for z14
  - Apply required PTFs for IBM and ISV (Independent Software Vendor) products
- Note that z14 supports only z/Architecture mode

7.8 Software licensing

The IBM z14 software portfolio includes operating system software (that is, z/OS, z/VM, z/VSE, and z/TPF) and middleware that runs on these operating systems. The portfolio also includes middleware for Linux on z Systems environments.

For the z14, two metric groups for software licensing are available from IBM, depending on the software product:

- Monthly license charge (MLC)

  MLC pricing metrics have a recurring charge that applies each month. In addition to the permission to use the product, the charge includes access to IBM product support during the support period. MLC 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 additional charge.

Software Licensing References

For more information about software licensing, see the following websites:

- 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, limiting the 4-hour rolling average partition utilization to the defined capacity value. Soft capping controls the maximum 4-hour rolling average usage (the last 4-hour average value at every 5-minute interval), but does not control the maximum instantaneous partition use.

You can also use an LPAR group capacity limit, which sets soft capping by PR/SM for a group of logical partitions 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, allowing 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 details about how to combine logical partition utilization, see *z/OS Planning for Sub-Capacity Pricing,* SA23-2301-01.

Key MLC Metrics and Offerings

MLC metrics include various offerings. Below is a non-exhaustive list of metrics and pricing schemes. Please note that offerings are usually tied to or made available only on certain Z servers. Always refer to IBM Z Software Pricing Reference Guide for most up to date information about available software licensing options available for IBM Z generations.

- 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 (S/390® Usage Pricing)

One of the recent changes in software licensing for z/OS and z/VSE has been MVM (Multi-Version Measurement) replacing Single Version Charging (SVC), Migration Pricing Option (MPO) and the IPLA Migration Grace Period.

Multi-Version Measurement (MVM) for z/OS and z/VSE removes time limits for running multiple eligible versions of a software program. Clients may run different versions of a program simultaneously for an unlimited duration during a program version upgrade. Clients may 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, providing more flexible control over their program upgrade cycles. Refer to ***Software Announcement 217-093, dated February 14, 2017*** for more information.

Technology Transition Offerings with z14

Complementing the announcement of the z14 server IBM has introduced following TTOs (Technology Transition Offerings):

- Technology Update Pricing for the IBM z14.
- 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 extends the software price/performance provided by AWLC and CMLC for z14 servers. The new and revised Transition Charges for Sysplexes or Multiplexes offerings provide a transition to Technology Update Pricing for the IBM z14 for customers who have not yet fully migrated to z14 servers. This ensures that aggregation benefits are maintained and also phases in the benefits of Technology Update Pricing for the IBM z14 pricing as customers migrate.

When a z14 server is in an actively coupled Parallel Sysplex or a Loosely Coupled Complex, you may choose either 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 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 additional information about software pricing for the z14 server, refer to Software Announcement 217-nnn, dated July 17, 2017, Technology Transition Offerings for the IBM z14 offer price-performance advantages.

When a z14 server is running z/VSE, you may 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, refer to:

http://ibm.com/systems/z/resources/swprice/

7.9 References

For current planning information, see the landing pages for each of the following operating systems:

- z/OS
- z/VM
- z/VSE
- z/TPF
- Linux on z Systems
- KVM for IBM Z

8

System upgrades

This chapter provides an overview of z14 upgrade capabilities and procedures, with an emphasis on capacity on demand (CoD) offerings. The upgrade offerings to the z14 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 detailed tutorials about all aspects of system upgrades, go to the IBM Resource Link[1] website. Click **Resource Link** → **Client Initiated Upgrade Information**, and then select **Education**. Select your particular product from the list of available systems:

https://www.ibm.com/servers/resourcelink/hom03010.nsf/pages/CIUInformation?OpenDocument

The growth capabilities that are provided by the z14 servers include the following benefits:

- Enabling exploitation 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 impacts

This chapter includes the following sections:

- Upgrade types
- Concurrent upgrades
- Miscellaneous equipment specification upgrades
- Permanent upgrade through the CIU facility
- On/Off Capacity on Demand
- Capacity for Planned Event
- Capacity Backup
- Nondisruptive upgrades
- Summary of Capacity on Demand offerings

[1] http://www.ibm.com/servers/resourcelink/. Registration is required to access Resource Link.

© Copyright IBM Corp. 2017. All rights reserved.

8.1 Upgrade types

The types of upgrades for a z14 server are summarized in this section.

8.1.1 Overview of upgrade types

Upgrades can be categorized as described in the following 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, more I/O cards, or more processor capacity. However, only a short-term upgrade might be necessary to handle a peak workload, or to temporarily replace a system that is down during a disaster or data center maintenance. z14 servers offer the following solutions for such situations:

- Permanent:
  - Miscellaneous equipment specification (MES):

    The MES upgrade order is always performed by IBM personnel. The result can be either real hardware or installation of Licensed Internal Code Configuration Control (LICCC) to the system. In both cases, installation is performed by IBM personnel.
  - Customer Initiated Upgrade (CIU):

    Using the CIU facility for a system requires that the online CoD buying feature (FC 9900) is installed on the system. The CIU facility supports only LICCC upgrades.
- Temporary:

  All temporary upgrades are LICCC-based. The one billable capacity offering is On/Off Capacity on Demand (On/Off CoD). The two replacement capacity offerings available are Capacity Backup (CBU) and Capacity for Planned Event (CPE).

For more information, see 8.1.2, “Terminology that is related to CoD for z14 systems” on page 301.

Tip: An MES provides system upgrades that can result in more enabled processors, a different central processor (CP) capacity level, and in additional processor drawers, memory, PCIe I/O drawers, and I/O features (physical upgrade). Additional 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 impact on the system and application availability, upgrades can be classified in the following manner:

- Concurrent

  In general, *concurrency* addresses the continuity of operations of the hardware part of an upgrade. For example, whether a system (hardware) is required to be turned off during the upgrade. For more information, see 8.2, “Concurrent upgrades” on page 304.
- Non-concurrent

  This type of upgrade requires turning off the hardware that is being upgraded. Examples include model upgrades from any z14 model to the z14 M05 model, and certain physical memory capacity upgrades.

- Disruptive

  An upgrade is considered *disruptive* when resources modified or added to an operating system image require that the operating system be 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 337.

8.1.2 Terminology that is related to CoD for z14 systems

Table 8-1 lists the most frequently used terms that are related to CoD for z14 systems.

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, either 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 Capacity on Demand.
Capacity levels	Can be full capacity or subcapacity. For the z14 system, capacity levels for the CP engine are 7, 6, 5, and 4: ► 1 - 99 in decimal and A0 - H0, where A0 represents 100 and H0 represents 170, for capacity level 7*nn*. ► 1 - 33 for capacity levels 6*yy* and 5*yy*. ► 0 - 33 for capacity levels 4*xx*. An all Integrated Facility for Linux (IFL) or an all integrated catalog facility (ICF) system has a capacity level of 400.
Capacity setting	Derived from the capacity level and the number of processors. For the z14 system, the capacity levels are 7*nn*, 6*yy*, 5*yy*, and 4*xx*, where *xx*, *yy*, or *nn* indicates the number of active CPs. The number of processors can have the following ranges: ► 1 - 99 in decimal and A0 - H0, where A0 represents 100 and H0 represents 170, for capacity level 7*nn*. ► 1 - 33 for capacity levels 6*yy* and 5*yy*. ► 0 - 33 for capacity levels 4*xx*. An all IFL or an all ICF system has a capacity level of 400.
Customer Initiated Upgrade (CIU)	A web-based facility where you can request processor and memory upgrades by using the IBM Resource Link and the system’s Remote Support Facility (RSF) connection.
Capacity on Demand (CoD)	The ability of a computing system to increase or decrease its performance capacity as needed to meet fluctuations in demand.

Term	Description
Capacity Provisioning Manager (CPM)	As a component of z/OS Capacity Provisioning, CPM monitors business-critical workloads that are running on z/OS on z14 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.
Full capacity CP feature	For z14 servers, feature (CP7) provides full capacity. Capacity settings 7nn are full capacity settings.
High-water mark	Capacity that is purchased and owned by the client.
Installed record	The LICCC record is downloaded, staged to the Support Element (SE), and is installed on the 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 system, including all replacement and billable capacity. For z14 servers, the model capacity identifier is in the form of 7*nn*, 6*yy*, 5*yy*, or 4*xx*, where *xx*, *yy*, or *nn* indicates the number of active CPs: ► 1 - 99 in decimal and A0 - H0, where A0 represents 100 and H0 represents 170, for capacity level 7*nn*. ► *yy* can have a range of 01 - 33. ► *xx* can have a range of 00 - 33. An all IFL or an all ICF system has a capacity level of 400.
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, z13 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.
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 *entitlement record* contains the encrypted representation of the upgrade configuration with the associated time limit conditions.
Replacement capacity	A temporary capacity that is used for situations in which processing capacity in other parts of the enterprise is lost. This loss can be a planned event or an unexpected disaster. The two replacement offerings available are Capacity for Planned Events and Capacity Backup.
Resource Link	The IBM Resource Link is a technical support website that provides a comprehensive set of tools and resources. It is available at the IBM Systems technical support website: http://www.ibm.com/servers/resourcelink/

Term	Description
Secondary approval	An option, which 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 capacity upgrade, either temporary or permanent, is retrieved and loaded on the SE disk.
Subcapacity	For z14 servers, CP features (CP4, CP5, and CP6) provide reduced capacity relative to the full capacity CP feature (CP7).
Temporary capacity	An optional capacity that is added to the current system capacity for a limited amount of time. It can be capacity that is owned or not owned by the client.
Vital product data (VPD)	Information that uniquely defines system, hardware, software, and microcode elements of a processing system.

8.1.3 Permanent upgrades

Permanent upgrades can be obtained by using these processes:

- Ordered through an IBM marketing representative
- Initiated by the client with the CIU on the IBM Resource Link

Tip: The use of the CIU facility for a system requires that the online CoD buying feature (FC 9900) is installed on the system. The CIU facility itself is enabled through the permanent upgrade authorization feature code (FC 9898).

Permanent upgrades that are ordered through an IBM representative

Through a permanent upgrade, you can accomplish these tasks:

- Add processor drawers
- 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)

Considerations: Most of the MESs can be concurrently applied without disrupting the existing workload. For more information, see 8.2, “Concurrent upgrades” on page 304. However, certain MES changes are disruptive, such as model upgrades from any z14 model to the z14 M05 model.

Memory upgrades that require dual inline memory module (DIMM) changes can be made nondisruptively if there are multiple CPC drawers and the flexible memory option is used.

Permanent upgrades initiated through CIU on the IBM Resource Link

Ordering a permanent upgrade by using the CIU application through Resource Link allows you to add capacity to fit within your existing hardware:

- Add model capacity
- Add specialty engines

- Add memory
- Activate unassigned model capacity or IFLs
- Deactivate activated model capacity or IFLs

8.1.4 Temporary upgrades

z14 servers offer three types of temporary upgrades:

- On/Off Capacity on Demand (On/Off CoD):

  This offering allows you to temporarily add more capacity or specialty engines to cover seasonal activities, period-end requirements, peaks in workload, or application testing. This temporary upgrade can be ordered only by using the CIU application through Resource Link.

- Capacity Backup (CBU):

  This offering allows you to replace model capacity or specialty engines in a backup system that is used in an unforeseen loss of system capacity because of a disaster.

- Capacity for Planned Event (CPE):

  This offering allows you to replace model capacity or specialty engines because of a relocation of workload during system migrations or a data center move.

CBU or CPE temporary upgrades can be ordered by using the CIU application through Resource Link or by calling your IBM marketing representative.

Temporary upgrade capacity changes can be billable or a replacement.

Billable capacity

To handle a peak workload, you can activate up to double the purchased capacity of any processor unit (PU) type temporarily. You are charged 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 offerings are the two replacement capacity offerings:

- Capacity Backup
- Capacity for Planned Event

8.2 Concurrent upgrades

Concurrent upgrades on z14 servers can provide more capacity with no system outage. In most cases, with prior planning and operating system support, a concurrent upgrade can be nondisruptive to the operating system.

The concurrent capacity growth capabilities that are provided by z14 servers include, but are not limited to, these 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.

The subcapacity models allow more configuration granularity within the family. The added granularity is available for models that are configured with up to 33 CPs, and provides 99 extra capacity settings. Subcapacity models provide for CP capacity increase in two dimensions that can be used together to deliver configuration granularity. The first dimension is adding CPs to the configuration. The second is changing the capacity setting of the CPs currently installed to a higher model capacity identifier.

z14 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. There are two ways to accomplish this addition:

- With planning ahead for the future need of extra processors. In the LPAR's profile, reserved processors can be specified. When the extra processors are installed, the number of active processors for that LPAR can be increased without the need for a partition reactivation and 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 servers support dynamic processor addition in the same way that the z13, zEC12, z196, and z10 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 Model upgrades

z14 servers have a machine type and model, and model capacity identifiers:

- The machine type and model are 3906-M*vv*.

  The *vv* can be 01, 02, 03, 04, or 05. The model number indicates how processor drawers are installed. The model M01 can have maximum 33 PUs for client characterization, M02 can have maximum 69, M03 can have maximum A5 which stands for 105, where A equals 10 in decimal plus 5), M04 can have E5 which stands for 145 and model M05 can have H0 which stands for 170, where H equals 17). Model M01 has one processor drawer that is installed, model M02 contains two processor drawers, model M03 contains three processor drawers, and models M04 and M05 contain four processor drawers.

- The model capacity identifiers are 4*xx*, 5*yy*, 6*yy*, or 7*nn*.

  The *xx* is a range of 00 - 33[2], *yy* is a range of 01 - 33, and *nn* is a range of 01 - 99, A0 - H0, where A0 represents the decimal number 100, combining the hexadecimal A with decimal 0 and H0 represents the decimal number 170, obtained by continuing the hexadecimal counting to F that equals 15, G would equal 16 and H would equal 17 and adding the decimal digit 0 to make 170. A z14 server with 170 client usable processors is a z14 7H0. The model capacity identifier describes how many CPs are characterized (*xx*, *yy*, or *nn*) and the capacity setting (4, 5, 6, or 7) of the CPs.

[2] The z13 zero CP MCI is 400. This setting applies to an all-IFL or all-ICF system.

A hardware configuration upgrade always requires more physical hardware (processor drawers, PCIe I/O drawers, or both[3]). A system upgrade can change either, or both, of the system model and the MCI.

Consider the following model upgrade information:

- LICCC upgrade:
  - Does not change the system model 3906-M0v because more processor drawers are not added
  - Can change the model capacity identifier, the capacity setting, or both
- Hardware installation upgrade:
  - Can change the system model 3906-M0v if one or more processor drawers are added
  - Can change the model capacity identifier, the capacity setting, or both

The system model and the model capacity identifier can be concurrently changed. Concurrent upgrades can be performed for both permanent and temporary upgrades.

Tip: A model upgrade can be concurrent by using concurrent drawer add (CDA), except for upgrades to Model M05 which are factory ship only.

Licensed Internal Code upgrades (MES ordered)

The LICCC provides for system upgrades without hardware changes by activation of additional (previously installed) unused capacity. Concurrent upgrades through LICCC can be performed for these resources:

- Processors (CPs, ICFs, IBM z Integrated Information Processors (zIIPs), IFLs, and SAPs) if unused PUs are available on the installed processor drawers, or if the model capacity identifier for the CPs can be increased.
- Memory, when unused capacity is available on the installed memory cards. Plan-ahead memory and the flexible memory option are available to give you better control over future memory upgrades. For more information, see 2.4.7, “Flexible Memory Option” on page 61, and 2.4.8, “Pre-planned memory” on page 61.

Concurrent hardware installation upgrades (MES ordered)

Configuration upgrades can be concurrent when installing the following resources:

- Processor drawers (which contain processors, memory, and fanouts). Up to three processor drawers can be added concurrently on the model z14 M01.
- PCIe fanouts.
- I/O cards, when slots are still available on the installed PCIe I/O drawers.
- PCIe I/O drawers.

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 servers support concurrent conversion between all PU types, which includes SAPs, to provide flexibility to meet changing business requirements.

3

Important: The LICCC-based PU conversions require that at least one PU, either CP, ICF, or IFL, remains unchanged. Otherwise, the conversion is disruptive. The PU conversion generates 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 have support to configure processors offline or online so that the PU conversion can be done nondisruptively.

Considerations: Client planning and operator action are required to use concurrent PU conversion. Consider the following information about PU conversion:

- It is disruptive if *all* current PUs are converted to different types.
- It might require individual LPAR outages if dedicated PUs are converted.

Unassigned CP capacity is recorded by a model capacity identifier. CP feature conversions change (increase or decrease) the model capacity identifier.

8.2.2 Customer Initiated Upgrade facility

The 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 servers at any time, generally it is added when ordering a z14 server. The CIU facility itself is controlled through the permanent upgrade authorization feature code, FC 9898.

After you place an order through the CIU facility, you receive a notice that the order is ready for download. You can then download and apply the upgrade by using functions that are available through the Hardware Management Console (HMC), along with the RSF. After all the prerequisites are met, the entire process, from ordering to activation of the upgrade, is performed by the client.

After download, the actual upgrade process is fully automated and does not require any onsite presence of IBM SSRs.

CIU prerequisites

The CIU facility supports LICCC upgrades only. It does not support I/O upgrades. All additional capacity that is required for an upgrade must be previously installed. Additional 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:

`http://www.ibm.com/servers/resourcelink/`

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 and start the CIU application to upgrade a system for processors or memory. Requesting a client order approval to conform to your operational policies is possible. You can allow the definition of more IDs to be authorized to access the CIU. Additional IDs can be authorized to enter or approve CIU orders, or only view existing orders.

Permanent upgrades

Permanent upgrades can be ordered by using the CIU facility. Through the CIU facility, you can generate online permanent upgrade orders to concurrently add processors (CPs, ICFs, zIIPs, IFLs, and SAPs) and memory, or change the model capacity identifier. You can do so up to the limits of the installed processor drawers on an existing system.

Temporary upgrades

The base model z14 server describes permanent and dormant capacity using the capacity marker and the number of PU features installed on the system. Up to eight temporary offerings can be present. Each offering has its own policies and controls, and each can be activated or deactivated independently in any sequence and combination. Although multiple offerings can be active at any time, if enough resources are available to fulfill the offering specifications, only one On/Off CoD offering can be active at any time.

Temporary upgrades are represented in the system by a *record*. All temporary upgrade records are resident 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.

Figure 8-1 shows a representation of the provisioning architecture.

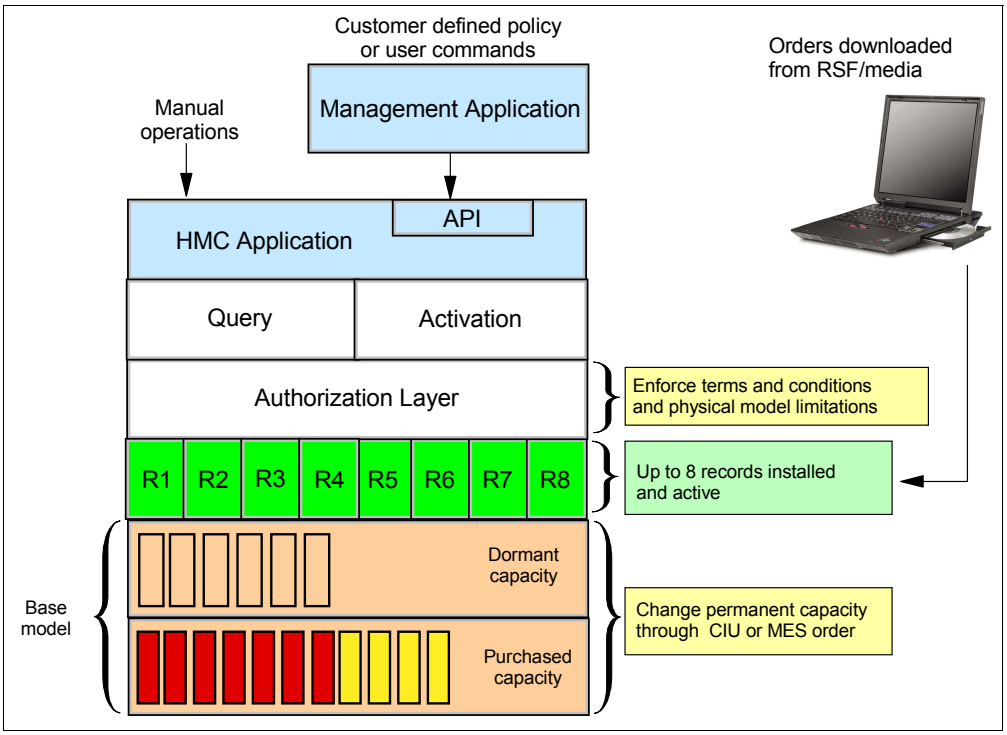


Figure 8-1 The provisioning architecture

The authorization layer enables administrative control over the temporary offerings. The activation and deactivation can be driven either manually or under control of an application through a documented application programming interface (API).

By using the API approach, you can customize, at activation time, the resources 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, for CPE, the ordered configuration is the only possible activation.

In addition, this approach enables you to update and replenish temporary upgrades, even in situations where the upgrades are already active. Likewise, depending on the configuration, permanent upgrades can be performed while temporary upgrades are active. Figure 8-2 shows examples of the activation sequence of multiple temporary upgrades.

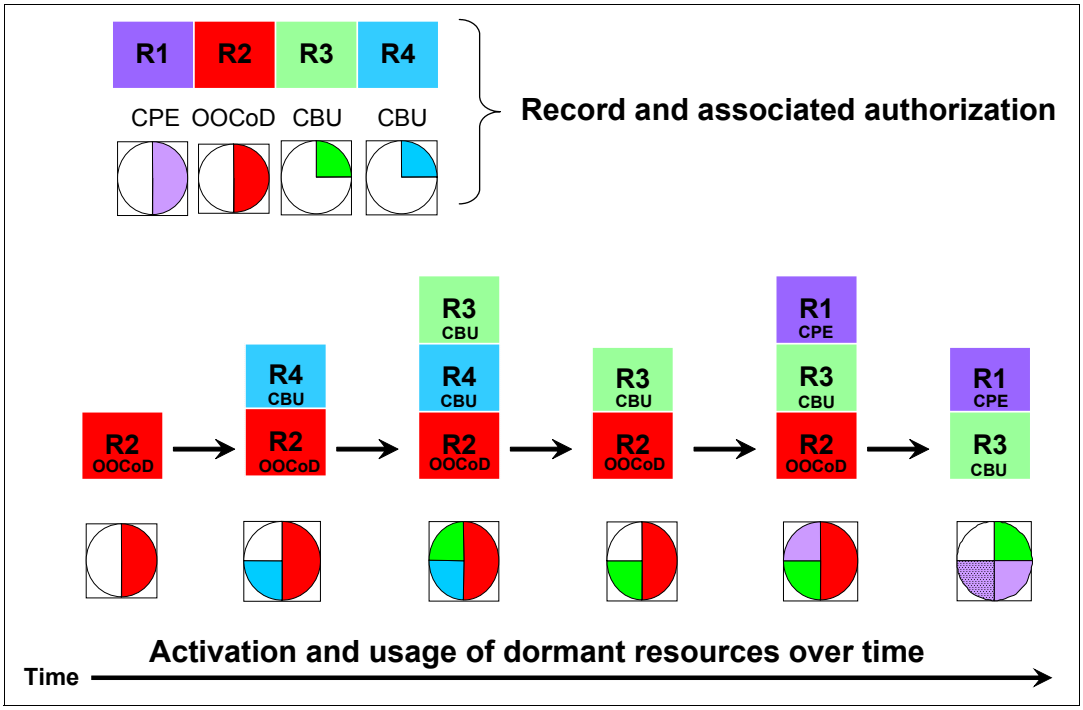


Figure 8-2 Example of temporary upgrade activation sequence

If R2, R3, and R1 are active at the same time, only parts of R1 can be activated because not enough resources are available to fulfill all of R1. When R2 is deactivated, the remaining parts of R1 can be activated as shown.

Temporary capacity can be billable as On/Off CoD, or replacement capacity as CBU or CPE:

- On/Off CoD is a function that enables *concurrent* and *temporary* capacity growth of the system.

  On/Off CoD can be used for client peak workload requirements, for any length of time, and has a daily hardware and maintenance charge. The software charges can vary according to the license agreement for the individual products. See your IBM Software Group representative for exact details.

  On/Off CoD can concurrently add processors (CPs, ICFs, 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 an existing system, and is restricted to twice the currently installed capacity. On/Off CoD requires a contractual agreement between you and IBM.

  You decide whether to either pre-pay or post-pay On/Off CoD. Capacity tokens inside the records are used to control activation time and resources.

- CBU is a concurrent and temporary activation of more CPs, ICFs, zIIPs, IFLs, and SAPs, an increase of the model capacity identifier, or both.

  CBU *cannot* be used for peak workload management in any form. As stated, On/Off CoD is the correct way to do that. A CBU activation can last up to 90 days when a disaster or recovery situation occurs. CBU features are optional, and require unused capacity to be

available on installed processor drawers of the backup system. They can be available as unused PUs, an increase in the model capacity identifier, or both.

A CBU contract must be in place before the special code that enables this capability can be loaded on the system. The standard CBU contract provides for five 10-day tests[4] (the *CBU test activation*) and one 90-day activation over a five-year period. Contact your IBM representative for details.

You can run production workload on a CBU upgrade during a CBU test. At least an *equivalent amount* of production capacity must be shut down during the CBU test. If you already have existing CBU contracts, you also must sign an Amendment (US form #Z125-8145) with IBM to allow you to run production workload on a CBU upgrade during your CBU tests.

- CPE is a concurrent and temporary activation of 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 either as unused PUs, as a possibility to increase the model capacity identifier on a subcapacity system, or as both. A CPE contract must be in place before the special code that enables this capability can be loaded on the system. The standard CPE contract provides for one 3-day planned activation at a specific date. Contact your IBM representative for details.

8.2.3 Summary of concurrent upgrade functions

Table 8-2 summarizes the possible concurrent upgrades combinations.

Table 8-2 Concurrent upgrade summary

Type	Name	Upgrade	Process
Permanent	MES	CPs, ICFs, zIIPs, IFLs, SAPs, processor drawer, memory, and I/Os	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

[4] z14 servers provides more improvements in the CBU activation windows. These windows have been improved to prevent inadvertent CBU activation.

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 ports. For subcapacity models, MES upgrades allow the concurrent adjustment of both the number of processors and the capacity level. The MES upgrade can be performed by using LICCC only, by installing more processor drawers, by adding PCIe I/O drawers, by adding I/O[5] features, or by a combination:

- MES upgrades for processors are done by any of the following methods:
  - LICCC assigning and activating unassigned PUs up to the limit of the installed processor drawers.
  - LICCC to adjust the number and types of PUs, to change the capacity setting, or both.
  - Installing more processor drawers, and LICCC assigning and activating unassigned PUs on the installed processor drawers.
- MES upgrades for memory are done by one of the following methods:
  - Using LICCC to activate more memory capacity up to the limit of the memory cards on the currently installed processor drawers. Plan-ahead and flexible memory features enable you to have better control over future memory upgrades. For more information about the memory features, see these descriptions:
    - 2.4.8, “Pre-planned memory” on page 61
    - 2.4.7, “Flexible Memory Option” on page 61
  - Installing more processor drawers and using LICCC to activate more memory capacity on installed processor drawers.
  - Using the CPC Enhanced Drawer Availability (EDA), where possible, on multi-drawer systems to add or change the memory cards.
- MES upgrades for I/O[5] are done by installing more I/O[5] features and supporting infrastructure, if required, on PCIe drawers that are already installed, or installing more PCIe 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.

To better use the MES upgrade function, carefully plan the initial configuration to allow a concurrent upgrade to a target configuration. The availability of PCIe I/O drawers improves the flexibility to perform unplanned I/O configuration changes concurrently.

The Store System Information (STSI) instruction gives more useful and detailed information about the base configuration and temporary upgrades. You can more easily resolve billing situations where independent software vendor (ISV) products are in use.

The model and model capacity identifiers that are returned by the STSI instruction are updated to coincide with the upgrade. For more information, see “Store System Information instruction” on page 340.

Upgrades: The MES provides the physical upgrade, resulting in more enabled processors, different capacity settings for the CPs, and more memory, I/O ports, I/O adapters, and I/O drawers. Extra planning tasks are required for nondisruptive logical upgrades. For more information, see “Guidelines to avoid disruptive upgrades” on page 342.

[5] Other adapter types, such as zHyperlink, Coupling Express LR, zEDC, and Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE), also can be added to the PCIe I/O drawers through an MES.

8.3.1 MES upgrade for processors

An MES upgrade for processors can concurrently add CPs, ICFs, zIIPs, IFLs, and SAPs to a z14 server by assigning available PUs on the processor drawers through LICCC. Depending on the quantity of the additional processors in the upgrade, more processor drawers might be required, and can be concurrently installed before the LICCC is enabled. With the subcapacity models, more capacity can be provided by adding CPs, by changing the capacity identifier on the current CPs, or by doing both.

Limits: The sum of CPs, inactive CPs, ICFs, 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.

Figure 8-3 is an example of an MES upgrade for processors, showing two upgrade steps.

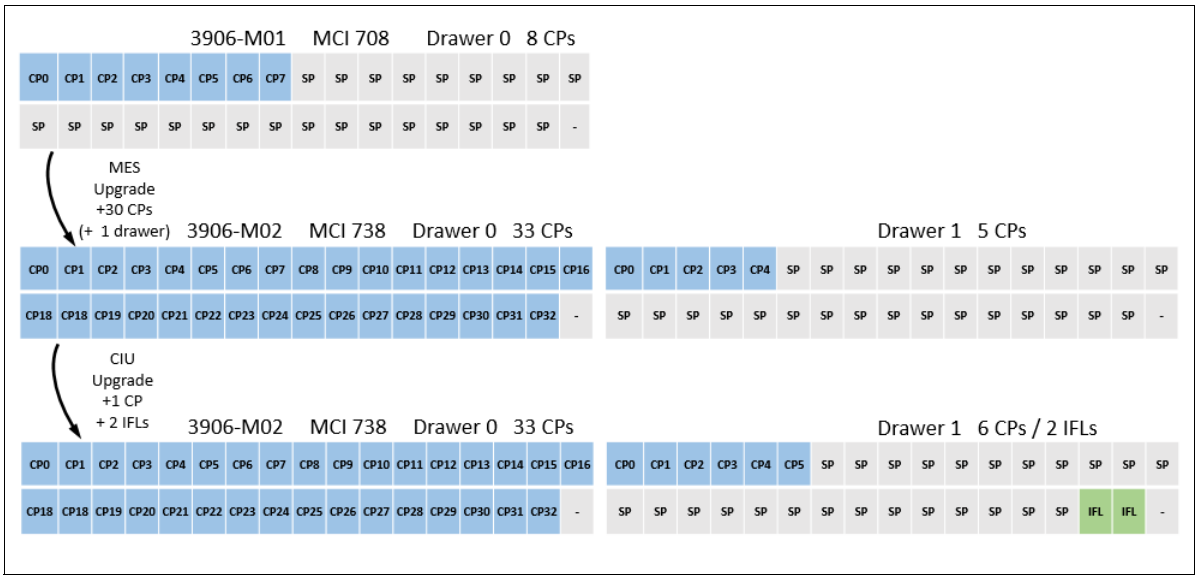


Figure 8-3 MES for processor example

A model M01 (one processor drawer), model capacity identifier 708 (eight CPs), is concurrently upgraded to a model M02 (two processor drawers), with MCI 738 (38 CPs). The model upgrade requires adding a processor drawer and assigning and activating 38 PUs as CPs. Then, model M02, MCI 738, is concurrently upgraded to a capacity identifier 739 (39 CPs) with two IFLs. This process is done by assigning and activating three more unassigned PUs (one as CP and two as IFLs). If needed, more LPARs can be created concurrently to use the newly added processors.

The example in Figure 8-3 on page 313 was used to show how the addition of PUs as CPs and IFLs, and the addition of a processor drawer, works. In reality, the addition of a processor drawer to a z14 Model M01 upgrades the machine model to M02. In addition, one of the two spare PUs on CPC drawer 0 is moved over to CPC drawer 1 to have one spare PU on each CPC drawer. After the second CPC drawer addition, CPC drawer 0 has 33 configurable PUs and CPC drawer 1 has 33 configurable PUs, which allow 66 PUs to be characterized on the new M02 model.

Consideration: Up to 170 logical processors, including reserved processors, can be defined to an LPAR. However, do not define more processors to an LPAR than the target operating system supports.

Table 8-3 describes the number of processors that are supported by various z/OS and z/VM releases.

Table 8-3 Number of processors that are supported by the operating system

Operating system	Number of processors that are supported
z/OS V1R13 with PTFs	100
z/OS V2R1	170 PUs per z/OS LPAR in non-SMT mode and 128 PUs per z/OS LPAR in simultaneous multithreading (SMT) mode. For both, the PU total is the sum of CPs and zIIPs
z/OS V2R2	170 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/OS V2R3	170 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 V6R3	64[a]
z/VM V6R4	64[a]
z/VSE	z/VSE Turbo Dispatcher can use up to 4 CPs, and tolerates up to 10-way LPARs
z/TPF	86 CPs
Linux on z Systems -170 CPs	SLES12/RHEL7/Ubuntu 16.10, Linux supports 256 cores without SMT and 128 cores with SMT (256 threads)

a. 32 in SMT mode

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 294.

8.3.2 MES upgrades for memory

MES upgrades for memory can concurrently add more memory in the following ways:

- Enabling, through LICCC, more capacity up to the limit of the currently installed DIMM memory cards
- Concurrently installing more CPC drawers and LICCC-enabling memory capacity on the new CPC drawers.

The Preplanned Memory Feature is available to allow better control over future memory upgrades. For more information about plan-ahead memory features, see 2.4.7, “Flexible Memory Option” on page 61 and 2.4.8, “Pre-planned memory” on page 61.

If the z14 server is a multiple processor drawer configuration, you can use the EDA feature to remove a processor drawer and add DIMM memory cards. It can also be used to upgrade the already installed memory cards to a larger capacity size. You can then use LICCC to enable the additional memory. 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.

Concurrency: Upgrades requiring DIMM changes can be concurrent by using the EDA feature. Planning is required to see whether this is a viable option for your configuration. Using the flexible memory option and the Preplanned Memory Feature (FC 1996 for the 16-GB increment, or FC 1990 for the 32-GB increment) ensures that EBA can work with the least disruption.

The one-processor drawer model M01 has a minimum of 320 GB physical installed memory. The client addressable storage in this case is 256 GB. If you require more, an extra memory upgrade can install up to 8 TB of memory. It does so by changing the existing DIMM sizes and adding more DIMMs in all available slots in the processor drawer. You can also add memory by *concurrently* adding a second processor drawer with sufficient memory into the configuration and then using LICCC to enable that memory.

An LPAR can dynamically take advantage of a memory upgrade if reserved storage is defined to that LPAR. The reserved storage is defined to the LPAR as part of the image profile. Reserved memory can be configured online to the LPAR by using the LPAR dynamic storage reconfiguration (DSR) function. DSR allows a z/OS operating system image and z/VM partitions to add reserved storage to their configuration if any unused storage exists.

The nondisruptive addition of storage to a z/OS and z/VM partition requires that pertinent operating system parameters have been prepared. If reserved storage is not defined to the LPAR, the LPAR must be deactivated, the image profile changed, and the LPAR reactivated. This process allows the additional storage resources to be available to the operating system image.

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 already installed PCIe I/O drawer.
- The installed PCIe I/O drawer providing 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.

For more information about I/O drawers and PCIe I/O drawers, see 4.2, “I/O system overview” on page 141.

Table 8-4 gives an overview of the number of I/O drawers and PCIe I/O drawers that can be present in a z14 server.

Table 8-4 I/O drawers and PCIe drawer summary

Description	New build	Carry-forward	MES add
PCIe I/O drawer	0 - 5	0 - 5[a]	0 - 5

a. In an upgrade from zEC12 new PCIe I/O Drawers will be delivered

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.

To better use the MES for I/O capability, carefully plan the initial configuration to allow concurrent upgrades up to the target configuration. If original I/O features are removed from the I/O drawer, the configurator does not physically remove the drawer unless the I/O frame slots are required to install a new PCIe I/O drawer.

If a PCIe I/O drawer is added to an existing z14 server and original features must be physically moved to another PCIe I/O drawer, original card moves are disruptive.

z/VSE, z/TPF, Linux on z Systems, and CFCC do *not* provide dynamic I/O configuration support. The installation of the new hardware is performed concurrently, but defining the new hardware to these operating systems requires an IPL.

Tip: z14 servers have a hardware system area (HSA) of 192 GB. z13 servers have a 96 GB HSA. 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 and its contents can be viewed under the Manage window, as shown in Figure 8-4. A staged record can be removed without installing it. An FoD record can only be installed completely. There is no selective feature or partial record installation. The features that are installed are merged with the CPC LICCC after activation.

An FoD record can be installed only one time. If it is removed, a new FoD record is needed to install it again. A remove action cannot be undone.

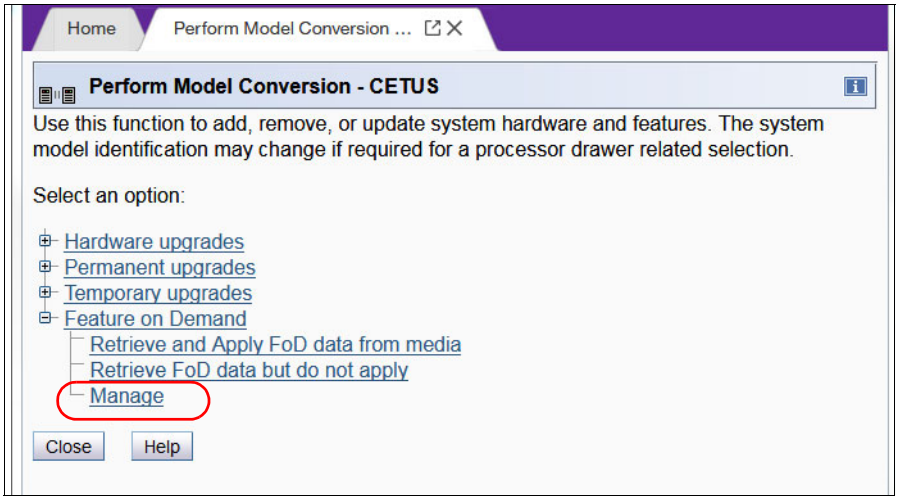


Figure 8-4 Features on Demand window for zBX blade feature HWMs

8.3.5 Summary of plan-ahead features

A number of plan-ahead features exist for z14 servers. The following list provides an overview of those features:

- Flexible memory
  Flexible memory has no feature code (FC) associated with it. The purpose of flexible memory is to enable enhanced processor drawer availability. If a processor drawer must

be serviced, the flexible memory is activated to accommodate the storage of the CPC drawer that is taken offline. After the repair action, the memory is taken offline again and is made unavailable for use.

- Pre-planned memory
  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 is enabled by using an LICCC order when more memory capacity is needed. FC 1990 provides 32 GB of pre-planned memory, and FC 1991 provides 64 GB of pre-planned memory, while 1999 provides 64 GB of VFM preplanned memory. Se table Table 8-5 for feature codes required to activate previously installed pre-planned memory,

Table 8-5 Feature codes for preplanned memory activation

FC	Capacity increment (GB)	Notes
1894	256	
1893	64	
1898	32	
1938	256	>= 1 TB
1939	64	>= 1 TB
1940	32	>= 1 TB

- Balanced Power Plan Ahead
  Balanced Power Plan Ahead is designed to anticipate future upgrade power needs on z14 servers. When more processor drawers are added to the system, the power consumption also rises. If necessary, one or more bulk power regulators (BPRs) must be added. This process increases the time that is needed for the upgrade. When ordering this feature, regardless of the configuration, all six BPR pairs are installed and activated. Balanced Power Plan Ahead has FC 3003.
- Line Cord plan ahead
  This option allows you to plan ahead for the second set of power cords. It is normally not configured until the addition of extra BPRs requires them. A plan-ahead option allows you to plan for a lengthy outage that is caused by installing circuit breakers or power feeds, or the routing of cables under the floor. The Line Cord plan-ahead option is FC 2000.

Tip: Accurate planning and the definition of the target configuration allows you to maximize the value of these plan-ahead features.

8.4 Permanent upgrade through the CIU facility

By using the CIU facility (through the IBM Resource Link on the web), you can start a permanent upgrade for CPs, ICFs, zIIPs, IFLs, SAPs, or memory. When performed through the CIU facility, you add the resources without having IBM personnel present at your location. You can also unassign previously purchased CPs and IFL processors through the CIU facility.

Adding permanent upgrades to a system through the CIU facility requires that the permanent upgrade enablement feature (FC 9898) is installed on the system. A permanent upgrade might change the system model capacity identifier (4*xx*, 5*yy*, 6*yy*, or 7*nn*) if more CPs are requested, or if the capacity identifier is changed as part of the permanent upgrade. However,

it cannot change the system model. If necessary, more LPARs can be created concurrently to use the newly added processors.

Consideration: A permanent upgrade of processors can provide a physical concurrent upgrade, resulting in more enabled processors 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 342.

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 294.

Figure 8-5 illustrates the CIU facility process on IBM Resource Link.

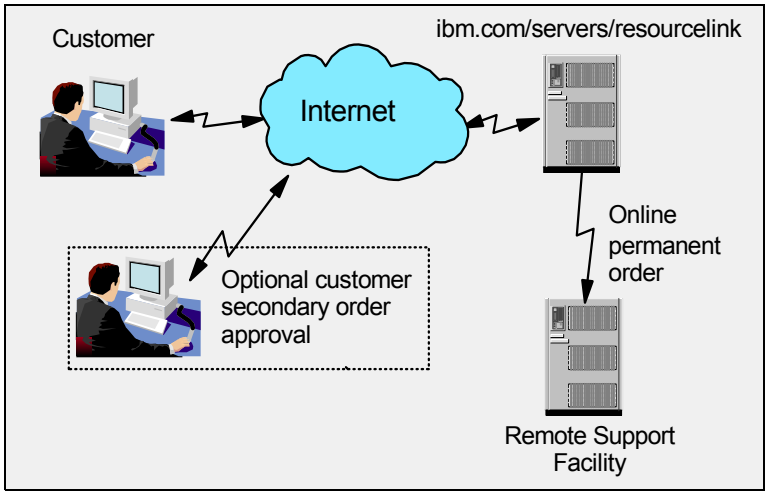


Figure 8-5 Permanent upgrade order example

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 details that are associated with the user ID are displayed.
3. Select the system to receive the upgrade. The current configuration (PU allocation and memory) is shown for the selected system.
4. Select **Order Permanent Upgrade**. The Resource Link limits the options to those that are valid or possible for the selected configuration (system).
5. After the target configuration is verified by the system, accept or cancel the order. An order is created and verified against the pre-established agreement.
6. Accept or reject the price that is quoted. A secondary order approval is optional. Upon confirmation, the order is processed. The LICCC for the upgrade will be available within hours.

Figure 8-6 illustrates the process for a permanent upgrade. When the LICCC is passed to the Remote Support Facility, you are notified through an email that the upgrade is ready to be downloaded.

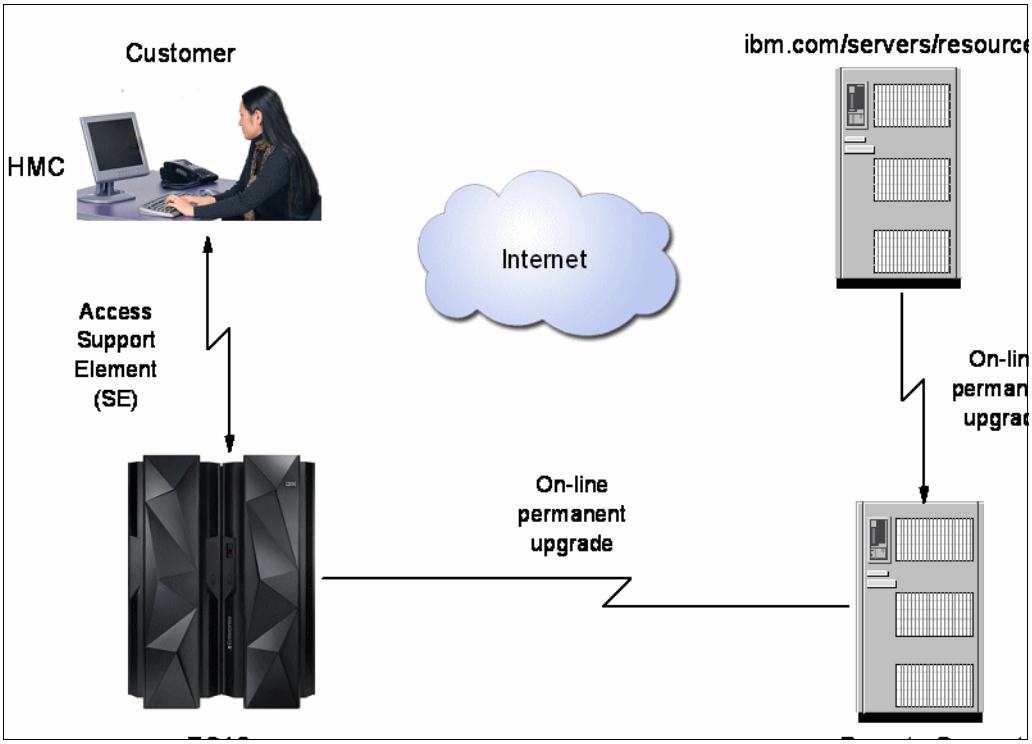


Figure 8-6 CIU-eligible order activation example

The two major components in the process are *ordering* and *retrieval* (along with activation).

8.4.1 Ordering

Resource Link provides the interface that enables you to order a concurrent upgrade for a 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 a tutorial, see this website:

https://www.ibm.com/servers/resourcelink/hom03010.nsf/pages/CIUInformation?OpenDocument

Figure 8-7 shows the initial view of the Machine profile on Resource Link.

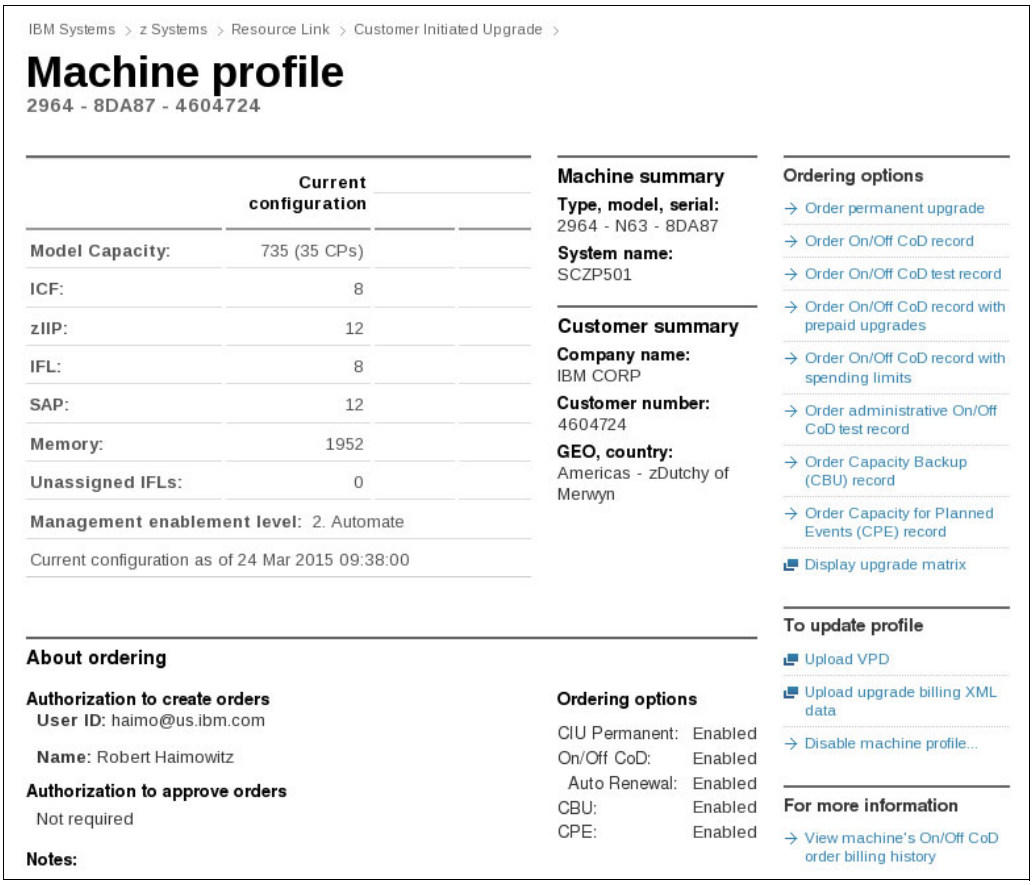


Figure 8-7 Machine profile window

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 web page.

Resource Link retrieves and stores relevant data that is associated with the processor configuration, such as the number of CPs and installed memory cards. It allows you to select only those upgrade options that are deemed valid by the order process. It allows upgrades only within the bounds of the currently installed hardware.

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.

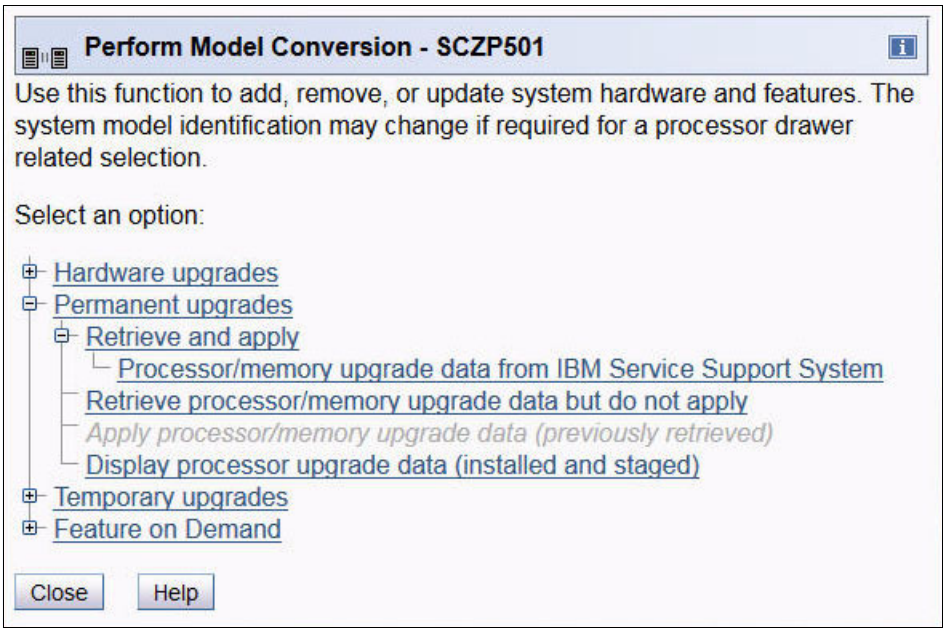


Figure 8-8 z14 Perform Model Conversion window

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.

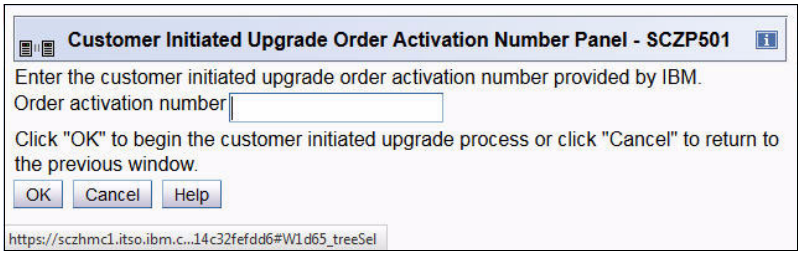


Figure 8-9 Customer Initiated Upgrade Order Activation Number window

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 has its own tokens, and each On/Off CoD record has separate token pools for each capacity type. During the ordering sessions on Resource Link, select how many tokens of each type to create for an offering record. Each engine type must have tokens for that engine type to be activated. Capacity that has no tokens cannot be activated.

When resources from an On/Off CoD offering record that contains capacity tokens are activated, a *billing window* is started. A billing window is always 24 hours. Billing takes place at the end of each billing window. The resources that are billed are the highest resource usage inside each billing window for each capacity type. An activation period is one or more complete billing windows. The activation period is the time from the first activation of resources in a record until the end of the billing window in which the last resource in a record is deactivated. At the end of each billing window, the tokens are decremented by the highest usage of each resource during the billing window. If any resource in a record does not have enough tokens to cover usage for the next billing window, the entire record is deactivated.

On/Off CoD requires that the Online CoD Buying feature (FC 9900) is installed on the system that 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 resources eligible 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 processor drawers 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 (4*xx*, 5*yy*, 6*yy*, or 7*nn*).

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, as described in z/OS MVS™ Capacity Provisioning User's Guide, SA33-8299

Using the capacity provisioning function requires these prerequisites:

- TCP/IP connectivity to observed systems.
- RMF Distributed Data Server must be active.
- CIM server must be active.

- Security and CIM customization.
- Capacity Provisioning Manager customization.

In addition, the Capacity Provisioning Control Center must be downloaded from the host and installed on a PC server. This application is used only to define policies. It is not required for regular operation.

Customization of the capacity provisioning function is required on the following systems:

- Observed z/OS systems: These are the systems in one or multiple sysplexes that are to be monitored. For more information about the capacity provisioning domain, see 8.8, “Nondisruptive upgrades” on page 337.
- Runtime systems: These are the systems where the Capacity Provisioning Manager is running, or to which the server can fail over after server or system failures.

8.5.3 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
- Up to 5 SAPs for model M01, 10 for an M02, 15 for an M03, and 20 for an M04 and 23 for an M05

On/Off CoD can provide CP temporary capacity in two ways:

- By increasing the number of CPs.
- For subcapacity models, capacity can be added by increasing the number of CPs, changing the capacity setting of the CPs, or both. The capacity setting for all CPs must be the same. If the On/Off CoD is adding CP resources that have a capacity setting different from the installed CPs, the base capacity settings are changed to match.

  On/Off CoD has the following limits associated with its use:

  – The number of CPs cannot be reduced.

  – The target configuration capacity is limited to these amounts:

    • Twice the currently installed capacity, expressed in MSUs for CPs.

    • Twice the number of installed IFLs, ICFs, and zIIPs. The number of SAPs that can be activated depends on the model. For more information, see 8.2.1, “Model upgrades” on page 305.

On/Off CoD can be ordered as prepaid or postpaid:

- A prepaid On/Off CoD offering record contains resource descriptions, MSUs, a number of speciality engines, and tokens that describe the total capacity that can be used. For CP capacity, the token contains MSU-days. For speciality engines, the token contains speciality engine-days.
- When resources on a prepaid offering are activated, they must have enough capacity tokens to allow the activation for an entire billing window, which is 24 hours. The resources remain active until you deactivate them or until one resource consumes all of its capacity tokens. When that happens, all activated resources from the record are deactivated.
- A postpaid On/Off CoD offering record contains resource descriptions, MSUs, speciality engines, and can contain capacity tokens that denote MSU-days and speciality engine-days.
- When resources in a postpaid offering record without capacity tokens are activated, those resources remain active until they are deactivated, or until the offering record expires. The record usually expires 180 days after its installation.
- When resources in a postpaid offering record with capacity tokens are activated, those resources must have enough capacity tokens to allow the activation for an entire billing window (24 hours). The resources remain active until they are deactivated, until all of the resource tokens are consumed, or until the record expires. The record usually expires 180 days after its installation. If one capacity token type is consumed, resources from the entire record are deactivated.

As an example, for a z14 server with capacity identifier 502 (two CPs), a capacity upgrade through On/Off CoD can be delivered in these ways:

- The first option is to add CPs of the same capacity setting. With this option, the model capacity identifier can be changed to a 503, adding one more CP to make it a 3-way. It can also be changed to a 504, which adds two CPs, making it a 4-way.
- The second option is to change to a different capacity level of the current CPs and change the model capacity identifier to a 602 or to a 702. The capacity level of the CPs is increased, but no additional CPs are added. The 502 also can be temporarily upgraded to a 603, increasing the capacity level and adding another processor. The capacity setting 430 does not have an upgrade path through On/Off CoD.

Use the Large System Performance Reference (LSPR) information to evaluate the capacity requirements according to your workload type. LSPR data for current IBM processors is available at this website:

https://www.ibm.com/servers/resourcelink/lib03060.nsf/pages/lsprindex

The On/Off CoD hardware capacity is charged on a 24-hour basis. There is a grace period at the end of the On/Off CoD day. This grace period allows up to an hour after the 24-hour billing period to either change the On/Off CoD configuration for the next 24-hour billing period or deactivate the current On/Off CoD configuration. The times when the capacity is activated and deactivated are maintained in the z14 server and sent back to the support systems.

If On/Off capacity is already active, more On/Off capacity can be added without having to return the system to its original capacity. If the capacity is increased multiple times within a 24-hour period, the charges apply to the highest amount of capacity active in that period. If more capacity is added from an already active record that contains capacity tokens, the systems checks whether the resource has enough capacity to be active for an entire billing window (24 hours). If that criteria is not met, no additional resources are activated from the record.

If necessary, more LPARs can be activated concurrently to use the newly added processor resources.

Consideration: On/Off CoD provides a concurrent hardware upgrade, resulting in more enabled processors 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 342.

To participate in this offering, you must have accepted contractual terms for purchasing capacity through the Resource Link, established a profile, and installed an On/Off CoD enablement feature on the system. Later, you can concurrently install temporary capacity up to the limits in On/Off CoD and use it for up to 180 days. Monitoring occurs through the system call-home facility, and an invoice is generated if the capacity is enabled during the calendar month. You are billed for the use of temporary capacity until the system is returned to the original configuration. If the On/Off CoD support is no longer needed, remove the enablement code.

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 order sequence. If a staged order is installed out of sequence, and later an order is staged that has a higher price, 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, giving 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, producing a new LICCC record for that specific order. Alternatively, you can use an *auto-renewal* feature to eliminate the need for a manual replenishment of the On/Off CoD order. This feature is implemented in Resource Link, and you must also select this feature in the machine profile, as shown in Figure 8-10.

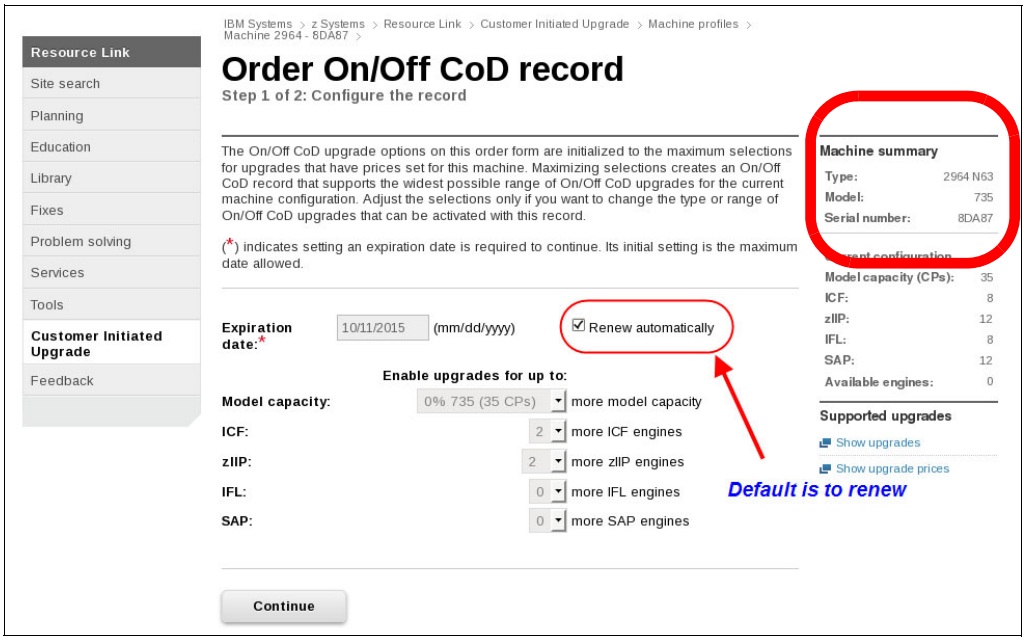


Figure 8-10 Order On/Off CoD record window

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, commencing upon the activation of any capacity resource that is contained in the On/Off CoD record. Activation levels of capacity can change during the 24-hour test period. The On/Off CoD test automatically ends at the end of the 24-hour period.

In addition, you can perform administrative testing. No additional capacity is added to the system, but you can test all the procedures and automation for the management of the On/Off CoD facility.

Figure 8-11 is an example of an On/Off CoD order on the Resource Link web page.

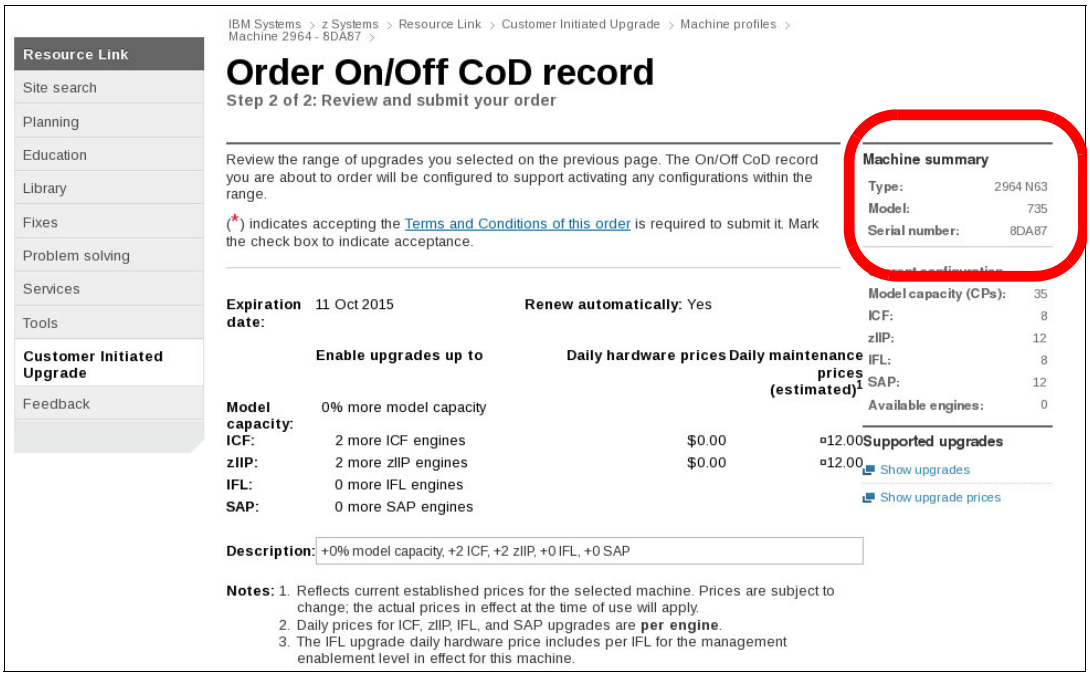


Figure 8-11 On/Off CoD order example

The example order in Figure 8-11 is an On/Off CoD order for 0% more CP capacity (system is already at capacity level 7), 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 installed processor drawers. The maximum number of SAPs is determined by the model number and the number of available PUs on the already installed processor drawers.

To finalize the order, you must accept Terms and Conditions for the order, as shown in Figure 8-12.

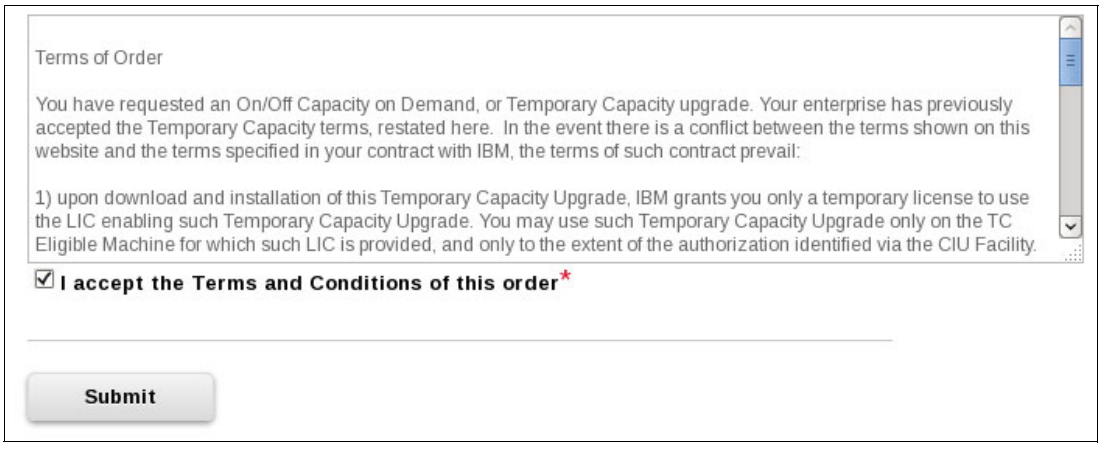


Figure 8-12 CIU order Terms and Conditions

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 when the capacity is needed, either manually or through automation.

If the On/Off CoD offering record does not contain resource tokens, you must deactivate the temporary capacity manually. Deactivation is accomplished from the SE, and is nondisruptive. Depending on how the additional capacity was added to the LPARs, you might be required to perform tasks at the LPAR level to remove 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 one time. If a different On/Off CoD order is required or a permanent upgrade is needed, it can be downloaded and activated without having to restore the system to its original purchased capacity.

In support of automation, an API is provided that allows the activation of the On/Off CoD records. The activation is performed from the HMC, and requires specifying the order number. With this API, automation code can be used to send an activation command along with the order number to the HMC to enable the order.

8.5.6 Termination

A client is contractually obligated to terminate the On/Off CoD right-to-use feature when a transfer in asset ownership occurs. A client also can choose to terminate the On/Off CoD right-to-use feature without transferring ownership. Application of FC 9898 terminates the right to use the On/Off CoD. This feature cannot be ordered if a temporary session is already active. Similarly, the CIU enablement feature cannot be removed if a temporary session is active. Any time that 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 involving physical hardware are supported while an On/Off CoD upgrade is active on a particular z14 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 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 340.

8.5.7 z/OS capacity provisioning

The z14 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.

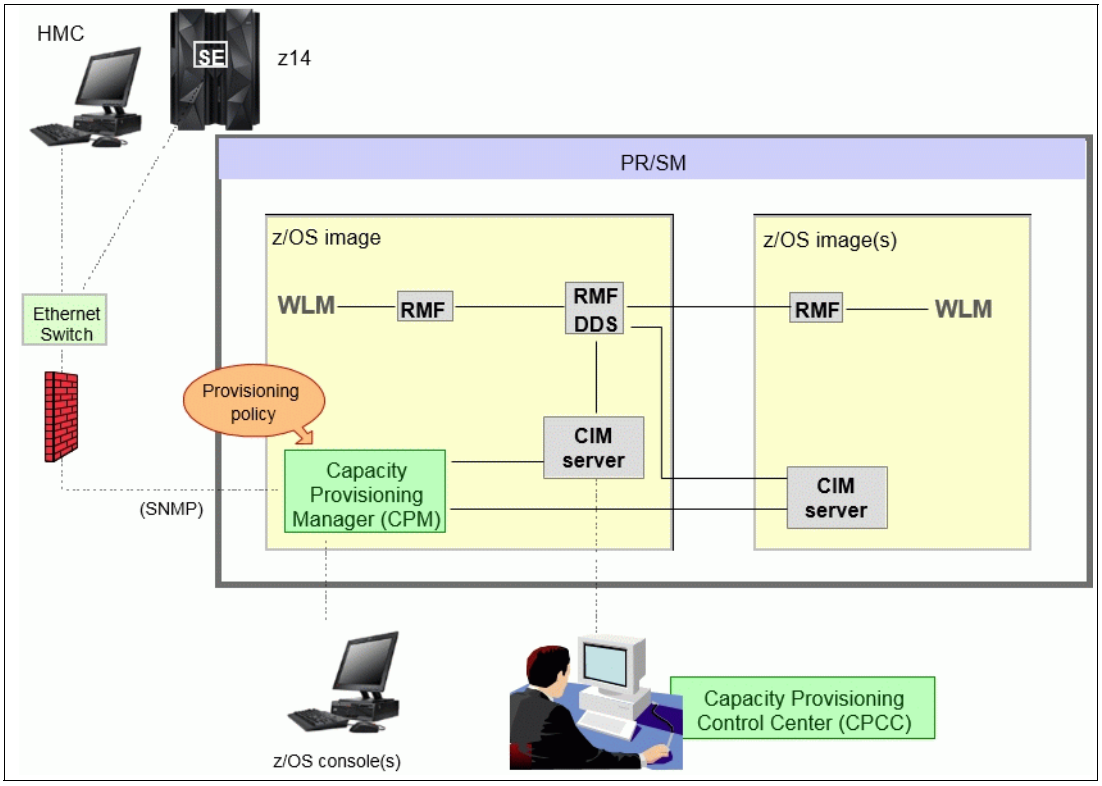


Figure 8-13 The capacity provisioning infrastructure

The z/OS WLM manages the workload by goals and business importance on each z/OS system. WLM metrics are available through existing interfaces, and are reported through IBM Resource Measurement Facility™ (RMF) Monitor III, with one RMF gatherer for each z/OS system.

Sysplex-wide data aggregation and propagation occur in the RMF Distributed Data Server (DDS). The RMF Common Information Model (CIM) providers and associated CIM models publish the RMF Monitor III data.

The CPM, a function inside z/OS, 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 has visibility of the resources in the individual offering records and the capacity tokens. When CPM activates resources, a check is run to determine whether enough capacity tokens remain for the specified resource to be activated for at least 24 hours. If insufficient tokens remain, no resource from the On/Off CoD record is activated.

If a capacity token is consumed during an activation that is driven by the CPM, the corresponding On/Off CoD record is deactivated prematurely by the system. This process occurs even if the CPM has activated this record, or parts of it. However, you do receive warning messages if capacity tokens are getting close to being fully consumed. You receive the messages five days before a capacity token is fully consumed. The five days are based on the assumption that the consumption is constant for the five days. You must put operational procedures in place to handle these situations. You can either deactivate the record manually, allow it happen automatically, or replenish the specified capacity token by using the Resource Link application.

The Capacity Provisioning Control Center (CPCC), which is on a workstation, provides an interface to administer capacity provisioning policies. The CPCC is not required for regular CPM operation. The CPCC will over time be moved into the z/OS Management Facility (z/OSMF). Parts of the CPCC are included in z/OSMF V1R13.

Capacity Provisioning Domain

The 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.

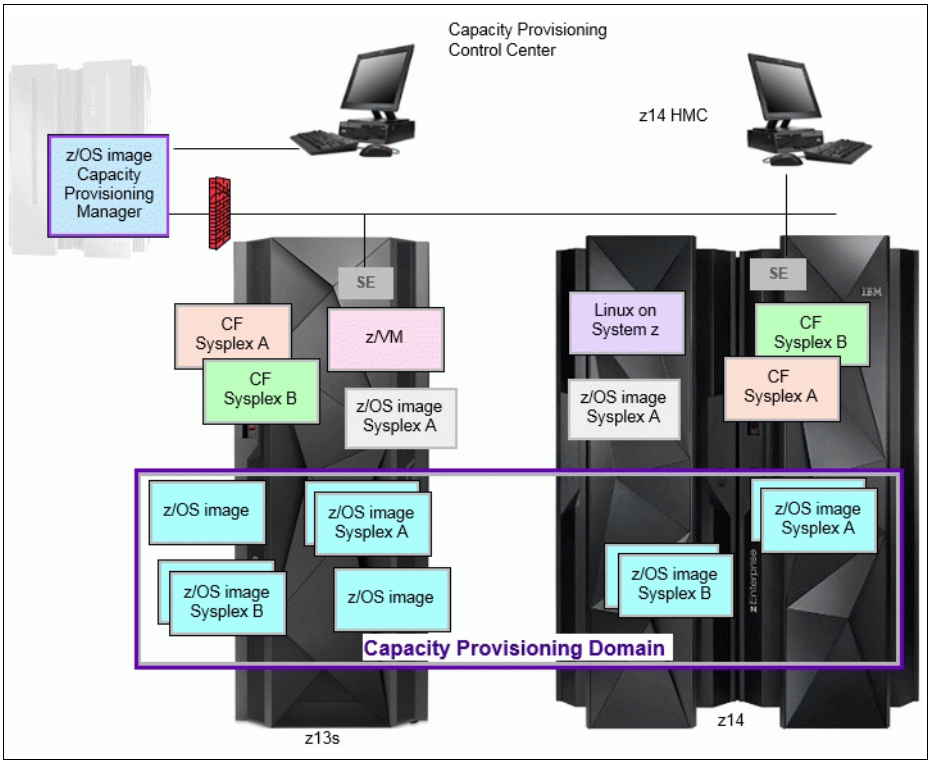


Figure 8-14 The Capacity Provisioning Domain

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. Sysplexes and CPCs do not have to be contained in a domain, but must not belong to more than one domain. Each domain has one active capacity provisioning policy. The CPCC is the CPM user interface component. Administrators work through this interface to define the domain configuration and provisioning policies. The CPCC is installed on a Microsoft Windows workstation.

CPM operates in four modes, allowing four different levels of automation:

- Manual mode:

  Use this command-driven mode when no CPM policy is active.

- Analysis mode:

  - In analysis mode, CPM processes capacity-provisioning policies and informs the operator when a provisioning or deprovisioning action is required according to policy criteria.
  - In analysis mode, the operator determines whether to ignore the information or to manually upgrade or downgrade the system by using the HMC, the SE, or available CPM commands.

- Confirmation mode:

  In this mode, CPM processes capacity provisioning policies and interrogates the installed temporary offering records. Every action that is proposed by the CPM must be confirmed by the operator.

- Autonomic mode:

  This mode is similar to the confirmation mode, but no operator confirmation is required.

A number of reports are available in all modes that contain information about workload and provisioning status, and the rationale for provisioning guidelines. User interfaces are provided through the z/OS console and the CPCC application.

The provisioning policy defines the circumstances under which more capacity can be provisioned (when, which, and how). The criteria have three 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 these 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, expressed in MSUs.

  The number of zIIPs must be one specification per CPC that is part of the CPD. They are specified in MSUs.

  The maximum provisioning scope is the maximum additional capacity that can be activated for all the rules in the CPD.

The provisioning rule is, in the specified time interval, that if the specified workload is behind its objective, up to the defined additional capacity can be activated.

The rules and conditions are named and stored in the Capacity Provisioning Policy.

For more information about z/OS Capacity Provisioning functions, see *z/OS MVS Capacity Provisioning User's Guide*, SA33-8299.

Planning considerations for using automatic provisioning

Although only one On/Off CoD offering can be active at any one time, several On/Off CoD offerings can be present on the system. Changing from one to another requires stopping the active one before the inactive one can be activated. This operation decreases the current capacity during the change.

The provisioning management routines can interrogate the installed offerings, their content, and the status of the content of the offering. To avoid the decrease in capacity, create only one On/Off CoD offering on the system by specifying the maximum allowable capacity. The CPM can then, when an activation is needed, activate a subset of the contents of the offering sufficient to satisfy the demand. If more capacity is needed later, the Provisioning Manager can activate more capacity up to the maximum allowed increase.

Having an unlimited number of offering records pre-staged on the SE hard disk is possible. Changing the content of the offerings, if necessary, is also possible.

Remember: The CPM has control over capacity tokens for the On/Off CoD records. In a situation where a capacity token is consumed, the system deactivates the corresponding offering record. Therefore, you must prepare routines for catching the warning messages about capacity tokens being consumed, and have administrative procedures in place for such a situation. The messages from the system begin five days before a capacity token is fully consumed. To avoid capacity records being deactivated in this situation, replenish the necessary capacity tokens before they are consumed.

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 these feature codes:

- FC #6833: Capacity for Planned Event enablement
- FC #0116: 1 CPE Capacity Unit
- FC #0117: 100 CPE Capacity Unit
- FC #0118: 10000 CPE Capacity Unit
- FC #0119: 1 CPE Capacity Unit - IFL
- FC #0120: 100 CPE Capacity Unit - IFL
- FC #0121: 1 CPE Capacity Unit - ICF
- FC #0122: 100 CPE Capacity Unit - ICF
- FC #0125: 1 CPE Capacity Unit - zIIP
- FC #0126: 100 CPE Capacity Unit - zIIP
- FC #0127: 1 CPE Capacity Unit - SAP
- FC #0128: 100 CPE Capacity Unit - SAP

The feature codes are calculated automatically when the CPE offering is configured. Whether using the eConfig tool or the Resource Link, a target configuration must be ordered. The configuration consists of a model identifier, a number of speciality engines, or both. Based on the target configuration, a number of feature codes from the list are calculated automatically, and a CPE offering record is constructed.

CPE is intended to replace capacity that is lost within the enterprise because of a planned event, such as a facility upgrade or system relocation. *CPE is intended for short duration events that last a maximum of three days*. Each CPE record, after it is activated, gives you access to dormant PUs on the system for which you have a contract, as described by the feature codes. Processor units can be configured in any combination of CP or specialty engine types (zIIP, 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 processors that can be activated by CPE come from the available unassigned PUs on any installed processor drawer. CPE features can be added to an existing z14 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 additional PUs is predetermined by the number and type of feature codes that are configured, as described by the feature codes. The number of PUs that can be activated is limited by the unused capacity that is available on the system:

- A model M03 with 26 CPs, and no IFLs or ICFs has 79 unassigned PUs available.
- A model M04 with 38 CPs, one IFL, and one ICF has 101 unassigned PUs available.

When the planned event is over, the system must be returned to its original configuration. You can deactivate the CPE features at any time before the expiration date.

A CPE contract must be in place before the special code that enables this capability can be installed on the system. CPE features can be added to an existing z14 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 these 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 an amount of 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, and *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 be able to use the CBU process, a CBU enablement feature (FC 9910) must be ordered and installed. You must order the quantity and type of PU that you require using these feature codes:

- FC 6805: Additional CBU test activations
- FC 6817: Total CBU years ordered
- FC 6818: CBU records ordered
- FC 6820: Single CBU CP-year
- FC 6821: 25 CBU CP-year
- FC 6822: Single CBU IFL-year
- FC 6823: 25 CBU IFL-year
- FC 6824: Single CBU ICF-year
- FC 6825: 25 CBU ICF-year
- FC 6828: Single CBU zIIP-year
- FC 6829: 25 CBU zIIP-year
- FC 6830: Single CBU SAP-year

[6] All new CBU contract documents contain new CBU test terms to allow execution of production workload during CBU test. Existing CBU clients must run the IBM client Agreement Amendment for IBM Z Capacity Backup Upgrade Tests (US form #Z125-8145).

- 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 additional CBU year added to the CBU entitlement record.

FC 6805 allows for ordering more tests in increments of one. The total number of tests that is allowed is 15 for each FC 6818.

The processors that can be activated by CBU come from the available unassigned PUs on any installed processor drawer. The maximum number of CBU features that can be *ordered* is 101. The number of features that can be *activated* is limited by the number of unused PUs on the system:

- A model M01 with Capacity Model Identifier 410 can activate up to 40 CBU features: 20 to change the capacity setting of the existing CPs, and 20 to activate unused PUs.
- A model M02 with 15 CPs, four IFLs, and one ICF has 43 unused PUs available. It can *activate* up to 43 CBU features.

However, the ordering system allows for over-configuration in the order itself. You can *order* up to 170 CBU features regardless of the current configuration. However, at *activation*, only the capacity that is already 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 done. On the full-capacity models, the CBU features indicate the amount of extra capacity needed. If the amount of necessary CBU capacity is equal to four CPs, the CBU configuration is four CBU CPs.

The subcapacity models have multiple capacity settings of 4*xx*, 5*yy*, or 6*yy*. The standard models have the capacity setting 7*nn*. The number of CBU CPs must be equal to or greater than the number of CPs in the base configuration. All the CPs in the CBU configuration must have the same capacity setting. For example, if the base configuration is a 2-way 402, providing a CBU configuration of a 4-way of the same capacity setting requires two CBU feature codes. If the required CBU capacity changes the capacity setting of the CPs, going from model capacity identifier 402 to a CBU configuration of a 4-way 504 requires four CBU feature codes with a capacity setting of 5*yy*.

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 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 (4*xx*, 5yy, or 6*yy*), but does not change the system model. The CBU feature cannot *decrease* the capacity setting.

A CBU contract must be in place before the special code that enables this capability can be installed on the system. CBU features can be added to an existing z14 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 additional capacity greater than the system's original, *permanent* configuration. At activation time, determine the capacity that you require for that situation. You can decide to activate only a subset of the capacity that is specified in the CBU contract.

The base system configuration must have sufficient memory and channels to accommodate the potential requirements of the large CBU target system. Ensure that all required functions and resources are available on the backup systems. These include CF LEVELs for coupling facility partitions, memory, and cryptographic functions, 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 processors, 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 342.

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/deactivated concurrently from the HMC by using the API. On the SE, CBU is activated either by using the Perform Model Conversion task or through the API. The API enables task automation.

CBU activation

CBU is activated from the SE, by using the 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 server can have more capacity, more active PUs, or both. The additional resources go into the resource pools and are available to the LPARs. If the LPARs must increase their share of the resources, the LPAR weight can be changed or the number of logical processors can be concurrently increased by configuring reserved processors online. The operating system must be able to concurrently configure more processors online. If necessary, more LPARs can be created to use the newly added capacity.

CBU deactivation

To deactivate the CBU, the additional resources must be released from the LPARs by the operating systems. In some cases, this process is a matter of varying the resources offline. In other cases, it can mean shutting down operating systems or deactivating LPARs. After the resources are released, the same facility on the 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, by using the same SE Perform a Model Conversion window and selecting the **Temporary upgrades** option. The HMC windows have been 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 capacity backup operation is 12 CBU features that are installed on a backup model M02 with model capacity identifier 708. When a production model M01 with model capacity identifier 708 has an unplanned outage, the backup system can be temporarily upgraded from model capacity identifier 708 - 720. This process allows the capacity to take over the workload from the failed production system.

Furthermore, you can configure systems to back up each other. For example, if you use two models of M01 model capacity identifier 705 for the production environment, each can have five or more features installed. If one system suffers an outage, the other one uses a temporary upgrade to recover the approximate original total capacity.

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:

- Runs the analysis that is required to determine the scope of the failure. This process minimizes operator intervention and the potential for errors.
- Automates authentication and activation of the reserved CPs.
- Automatically restarts the critical applications after reserved CP activation.
- Reduces the outage time to restart critical workloads from several hours to minutes.

The GDPS service is for z/OS only, or for z/OS in combination with Linux on z Systems.

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 servers allow *concurrent* upgrades, which mean that dynamically adding more 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 do not have 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 be able to concurrently configure more capacity online. z/OS operating systems have this capability. z/VM can concurrently configure new processors and I/O devices online, and memory can be dynamically added to z/VM partitions.

If the concurrent upgrade is intended to satisfy the need for more operating system images, more LPARs can be created *concurrently* on the z14 system. These LPARs include all resources that are needed. These additional LPARs can be activated concurrently.

These enhanced configuration options are available through the separate HSA, which was introduced on the zEnterprise 196.

Linux operating systems, in general, cannot add more resources concurrently. However, Linux, and other types of virtual machines that run under z/VM, can benefit from the z/VM capability to nondisruptively configure more resources online (processors and I/O).

With z/VM, Linux guests can manipulate their logical processors by using the Linux CPU hotplug daemon. The daemon can start and stop logical processors that are based on the Linux *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 here:

- Processors
- Memory
- I/O
- Cryptographic adapters
- Special features

Processors

CPs, ICFs, zIIPs, IFLs, and SAPs can be added concurrently to a z14 server if unassigned PUs are available on any installed processor drawer. The number of zIIPs cannot exceed twice the number of CPs plus unassigned CPs. More processor drawers can also be installed concurrently, allowing further processor upgrades.

If necessary, more LPARs can be created concurrently to use the newly added processors.

The Coupling Facility Control Code (CFCC) can also configure more processors online to coupling facility LPARs by using the CFCC image operations window.

Memory

Memory can be added concurrently up to the physical installed memory limit. Additional processor drawers can also be installed concurrently, allowing further memory upgrades by LICCC, enabling memory capacity on the new processor drawers.

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 Systems 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), allowing 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 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 server can be upgraded concurrently from one model to another, either temporarily or permanently.

Enabling and using the additional processor capacity is transparent to most applications. However, certain programs depend on processor model-related information, such as ISV products. Consider the effect on the software that is running on a z14 server when you perform any of these configuration upgrades.

Processor identification

Two instructions are used to obtain processor information:

- Store System Information (STSI) instruction

  STSI reports the processor model and model capacity identifier for the base configuration, and for any additional configuration changes through temporary upgrade actions. It fully supports the concurrent upgrade functions, and is the preferred way to request processor information.

- Store CPU ID (STIDP) instruction

  STIDP is provided for compatibility with an earlier version.

Store System Information instruction

Figure 8-15 shows the relevant output from the STSI instruction. The STSI instruction returns the model capacity identifier for the permanent configuration and the model capacity identifier for any temporary capacity. This data is key to the functioning of CoD offerings.

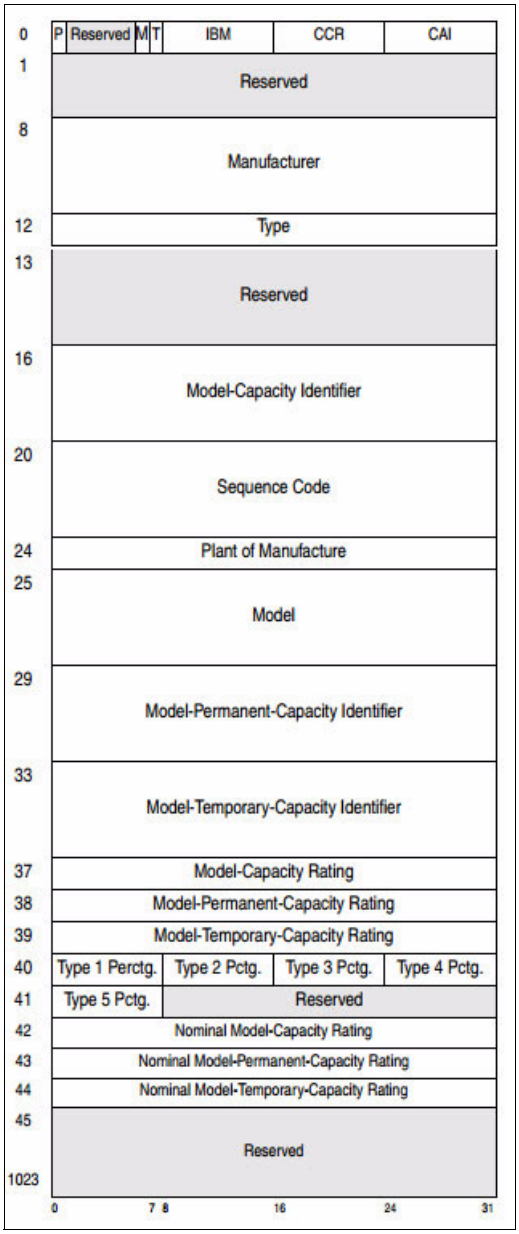


Figure 8-15 STSI output on a z14 server

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 (STIDP) instruction

The STIDP instruction provides information about the processor type, serial number, and LPAR identifier, as shown in Table 8-6. The LPAR identifier field is a full byte to support more than 15 LPARs.

Table 8-6 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'2964'	x'8000'[c]

a. The version code for z14 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 returned in bits 8 - 15 is a two-digit value.

When issued from an operating system that is running as a guest under z/VM, the result depends on whether the **SET CPUID** command was used:

- Without the use of the **SET CPUID** command, bits 0 - 7 are set to FF by z/VM. However, the remaining bits are unchanged, which means that they are exactly as they were without running as a z/VM guest.
- If the **SET CPUID** command is issued, bits 0 - 7 are set to FF by z/VM and bits 8 - 31 are set to the value that is entered in the **SET CPUID** command. Bits 32 - 63 are the same as they were without running as a z/VM guest.

Table 8-7 lists the possible output that is returned to the issuing program for an operating system that runs as a guest under z/VM.

Table 8-7 z/VM guest 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
Without **SET CPUID** command	x'FF'	LPAR ID	4-digit number that is derived from the CPC serial number	x'3906'	x'8000'
With **SET CPUID** command	x'FF'	6-digit number as entered by the command **SET CPUID** = *nnnnnn*		x'3906'	x'8000'

Planning for nondisruptive upgrades

Online permanent upgrades, On/Off CoD, CBU, and CPE can be used to upgrade concurrently a z14 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 in advance. Planning ahead is a key factor for nondisruptive upgrades.

The following list describes the main reasons for disruptive upgrades. However, by carefully planning and reviewing “Guidelines to avoid disruptive upgrades” on page 342, you can minimize the need for these outages.

- LPAR memory upgrades when reserved storage was not previously defined are disruptive to image upgrades. z/OS and z/VM support this function.
- Installation of an I/O cage is disruptive.
- An I/O upgrade when the operating system cannot use the dynamic I/O configuration function is disruptive to that partition. Linux, z/VSE, z/TPF, and CFCC do not support dynamic I/O configuration.

Guidelines to avoid disruptive upgrades

Based on reasons for disruptive upgrades (“Planning for nondisruptive upgrades” on page 342), here are guidelines for avoiding or at least minimizing these situations, increasing the chances for nondisruptive upgrades:

- 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 be able 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 V1R11, z/OS V1R12, and z/OS V1R13 with PTFs support up to 100 processors. z/OS V2R1 supports 170 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 be able to 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 V2.2 and V2.1, and 16 TB for z/OS V2.3. z/VM 6.4 supports 2 TB memory partitions.
- Consider the flexible and plan-ahead memory options.

  Use a convenient entry point for memory capacity, and select memory options that allow future upgrades within the memory cards that are installed on the CPC drawers. For more information about the offerings, see these sections:

  – 2.4.7, “Flexible Memory Option” on page 61
  – 2.4.8, “Pre-planned memory” on page 61

Considerations when installing additional CPC drawers

During an upgrade, more processor drawers can be installed concurrently. Depending on the number of additional processor drawers in the upgrade and your I/O configuration, a fanout rebalancing might be needed for availability reasons.

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 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 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 consumption (and therefore financial exposure) can be controlled by using capacity tokens in the On/Off CoD offering records.

The CPM is an example of an application that uses the CoD APIs to provision On/Off CoD capacity based on the requirements of the workload. The CPM cannot control other offerings.

For more information about any of the topics in this chapter, see the *IBM Z Capacity on Demand User's Guide*, SC28-6943.

9

Reliability, availability, and serviceability

The z14 RAS design is driven, from the Quality perspective, by a set of high level program RAS objectives. The IBM Z platform continues to drive toward Continuous Reliable Operation (CRO) at the single footprint level. The key objectives, in order of priority, are to ensure data integrity, computational integrity, reduce or eliminate unscheduled outages, reduce scheduled outages, reduce planned outages and reduce 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 sections:

- The RAS strategy
- Technology change
- Structure change
- Reducing complexity
- Reducing touches
- z14 availability characteristics
- z14 RAS functions
- z14 enhanced drawer availability
- z14 Enhanced Driver Maintenance
- RAS capability for the Hardware Management Console and Support Element

© Copyright IBM Corp. 2017. All rights reserved.

9.1 The RAS strategy

The RAS strategy is to manage change by learning from previous generations and investing in new RAS function to eliminate or minimize all sources of outages. Enhancements to z13 RAS designs are implemented on the z14 system through the introduction of new technology, structure, and requirements. Continuous improvements in RAS are associated with new features and functions to ensure that Z servers deliver exceptional value to clients.

Overriding RAS requirements/principles as depicted in Figure 9-1:

- Inclusion of existing (or equivalent) RAS characteristics from previous generations.
- Learning 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, both hardware and firmware, that provide IBM Z and Customer valued differentiation.

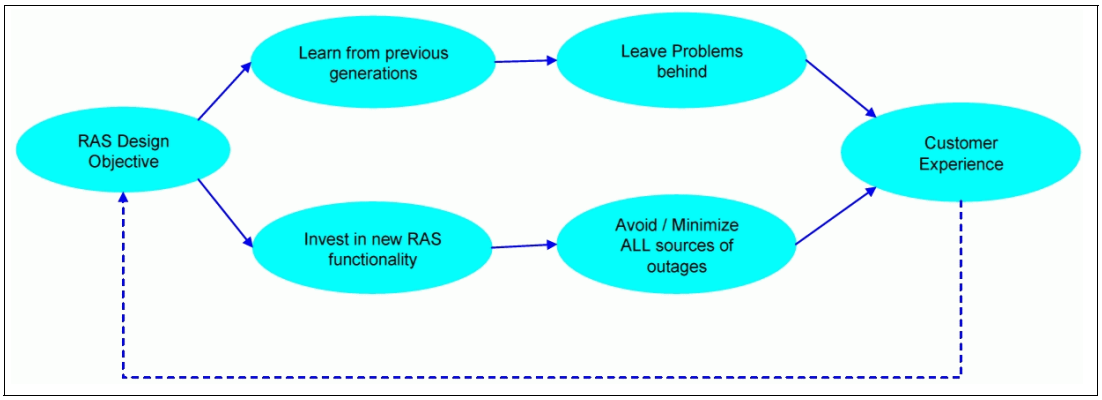


Figure 9-1

9.2 Technology change

The IBM z14 processor builds upon the Reliability, Availability and Serviceability (RAS) of the z13 family with the following RAS improvements:

- Improved SCM (Single Chip Module) cold plate and the layer of metal between the SCM and the cold plate that enhances the thermal conductivity and improves the reliability
- Symbol ECC on L3 data cache for better availability. This is now another layer of ECC protection which previously was with the main memory RAIM and L4 symbol ECC
- L3 ability to monitor (dynamically) fenced macros
- Ability to spare the PU core upon non-L2 cache/DIR Core Array delete.
- Improved error thresholding on PU cores 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

- L3 wordline span reduced (less impact)
- L1 and L1+, L2 protected by PU sparing
- PU Core mandatory address checking
- Redundant parity on error in RU bit to protect WL

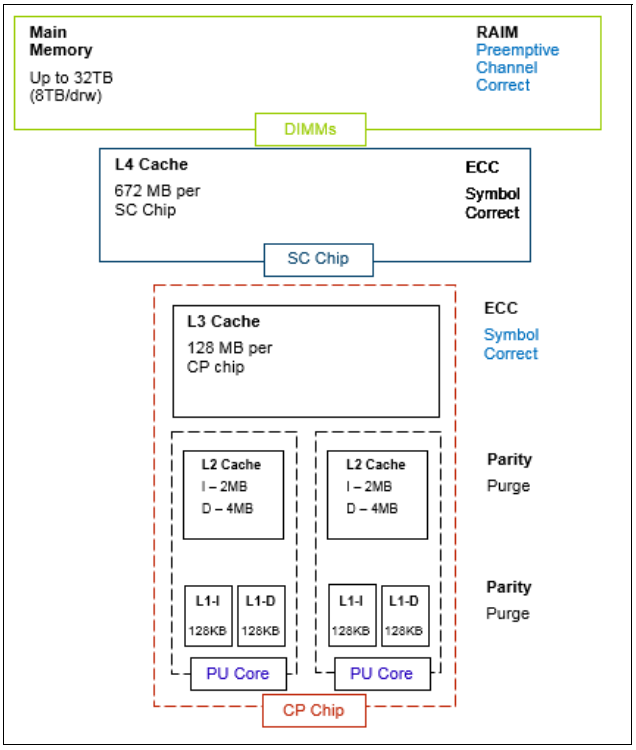


Figure 9-2 Memory and Cache Structure

SC chip and Main Memory:

- Reduced chip count (single storage controller (SC) chip) improves reliability with less components involved
- Reduced SMP cables (fewer SC chips) improves reliability and also results in faster repair actions and hardware upgrade times due to fewer components to quiesce, test and resume.
- Pre-emptive memory channel marking
  - Analysis of uncorrectable errors considers pattern of prior correctable errors
  - More robust uncorrectable error handling
  - Simplified repair action
- Improved resilience in KUE[1] repair capability
- Virtual Flash Memory (Flash Express replacement) solution is moved to DIMM
  - Solution is moved to more robust storage RAIM protected (same function that main memory uses)
  - New solution includes Concurrent Drawer Repair (CDR) and concurrent drawer add (CDA)

[1] Key in storage error uncorrected: Indicates that the hardware could not repair a storage key that was in error.

I/O and Service:

- OSA-express6S add 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, support partitions, for managing native PCIe IO
  - Now four partitions, was two on previous systems
  - Reduced impact for MCL updates
  - Better availability
- Faster Dynamic Memory Relocation engine
  - Enables faster reallocation of memory (2x faster) used for LPAR activations, Concurrent Drawer Repair and Concurrent upgrade
  - Provides faster, more robust service actions
- Dynamic TDR (Time Domain Reflectometry) which used to be static
  - Hardware facility used to isolate failures on wires provides better FRU isolation and improved service actions
- Universal spare for CP SCMs and SC SCMs and processor drawer
- Same Lift tool and FDT (Fill and Drain tool) used on z13

9.3 Structure change

With z14 servers, the processing infrastructure within the CPC is designed using drawer technology. Cables are keyed to ensure that correct lengths are plugged, plug detection ensures correct location, and custom latches ensure retention. Further improvements to the fabric bus include symmetric multiprocessing (SMP) cables that connect the drawers together. To improve field-replaceable unit (FRU) isolation, time domain reflectometry (TDR) techniques are applied to the SMP cables, between chips (CP-CP, and CP-SC), and between the CP chips and dual inline memory modules (DIMMs).

Enhancements to thermal RAS were introduced as well, such as a field-replaceable water manifold for CP cooling. Two redundant oscillator cards are attached to the central processor complex (CPC) drawers through a rigid backplane. The following z13 characteristics are continued with z14 servers:

- Processing infrastructure is designed by using drawer technology.
- Keyed cables and plugging detection.
- SMP cables that are used for fabric bus connections. There are fewer SMP cables for z14 than z13 used.
- Water manifold is a FRU.
- Master-master redundant oscillator design in the main memory.
- Processor and nest chips are separate FRUs.
- Point of load cards are separate FRUs.
- Oscillator cards plug straight in (no blind mating of connector)
- Built in time domain reflectometer for FRU isolation in interface errors

9.4 Reducing complexity

z14 servers continue the z13 enhancements that reduced system RAS complexity. Specifically, in the memory subsystem design, simplifications were made in RAIM recovery. Memory DIMMs are no longer 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. Further 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 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 has been 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, and SC-SMP fabric
- N+1 radiator pumps, controllers, blowers, and sensors
- SCH N+1 (with N+1 SCH power supplies)
- 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

9.6 z14 availability characteristics

The following functions include availability characteristics on z14 servers:

- Enhanced drawer availability (EDA)

  EDA is a *procedure* under which a CPC drawer in a multidrawer system can be removed and reinstalled during an upgrade or repair action with no impact on the workload.

- Concurrent memory upgrade or replacement

  Memory can be upgraded concurrently by using Licensed Internal Code Configuration Control (LICCC) if physical memory is available on the drawers. If the physical memory cards must be changed in a multidrawer configuration, requiring the drawer to be removed, the EDA function can be useful. It requires the availability of more memory resources on other drawers or reducing the need for memory resources during this action. To help ensure that the appropriate level of memory is available in a multiple-drawer configuration, select the flexible memory option. This option provides more resources to use EDA when repairing a drawer or memory on a drawer. They also are available when upgrading memory where larger memory cards might be required.

Memory can be upgraded concurrently by using LICCC if physical memory is available. The plan-ahead memory function that is available with z14 servers allows you to plan for nondisruptive memory upgrades by having the system pre-plugged based on a target configuration. You can enable the pre-plugged memory by placing an order through LICCC.

- Enhanced driver maintenance (EDM)

  One of the greatest contributors to downtime during planned outages is LIC driver updates that are performed in support of new features and functions. z14 servers are designed to support the concurrent activation of a selected new driver level.

- Plan Ahead for Balanced Power (FC 3003)

  The *Plan Ahead for Balanced Power* feature allows you to order the maximum number of bulk power regulators (BPRs) on any server configuration. This feature helps to ensure that your configuration is in a balanced power environment if you intend to add CPC drawers and I/O drawers to your server in the future. Regardless of your configuration, all six BPR pairs are shipped, installed, and activated.

  Note: When this feature is ordered, a corequisite feature, the Plan Ahead for Line Cords feature (FC 2000), is automatically selected.

- Concurrent fanout addition or replacement

  A PCIe or InfiniBand (IFB) fanout card provides the path for data between memory and I/O through PCIe or IFB cables. With z14 servers, a hot-pluggable and concurrently upgradeable fanout card is available. Up to 10 PCIe and up to four IFB fanout cards per CPC drawer are available for z14 servers. A z14 Model M04 (or M05) holds four CPC drawers and can have 40 PCIe fanout slots and 16 IFB fanout slots. Internal I/O paths from the CPC drawer fanout ports to either a PCIe drawer or an I/O drawer are spread across multiple CPC drawers (for models M02, M03, M04, and M05) and across different nodes within a single CPC drawer model M01. During an outage, a fanout card that is used for I/O can be repaired concurrently while redundant I/O interconnect ensures that no I/O connectivity is lost.

- Redundant I/O interconnect

  Redundant I/O interconnect helps maintain critical connections to devices. z14 servers allow a single drawer, in a multidrawer system, to be removed and reinstalled concurrently during an upgrade or repair. Connectivity to the system I/O resources is maintained through a second path from a different drawer.

- Dynamic oscillator switch-over

  z14 servers have two oscillator cards, a primary and a backup. During a primary card failure, the backup card is designed to detect transparently the failure, switch over, and provide the clock signal to the system.

- Processor unit (PU) sparing

  z14 servers have two spare PUs 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 sparing for failed processors is supported. There are two spare PUs per system and sparing is supported across the drawers in the unlikely event that the drawer with the failure does not have spares available.

- Application preservation

Application preservation is used when a CP fails and no spares are left. The state of the failing CP is passed to another active CP, where the operating system uses it to successfully resume the task, in most cases without client intervention.

- Cooling improvements

  The z14 air-cooled configuration comes with a newly designed radiator cooling system. The radiator pumps, blowers, controls, and sensors are N+1 redundant. In normal operation, one active pump supports the system. A second pump is turned on and the original pump is turned off periodically, which improves reliability of the pumps. The replacement of pumps or blowers is concurrent with no performance impact.

  A water-cooling system also is an option in z14 servers, with water-cooling unit (WCU) technology. Two redundant WCUs run with two independent chilled water feeds. One WCU and one water feed can support the entire system load. The water-cooled configuration is backed up by the rear door heat exchangers in the rare event of a problem with the chilled water facilities of the customer.

  As with z13, the design of the rear door covers addresses past data center issues regarding airflow challenges. The covers can be installed for a vectored down or up orientation, fulfilling the requirements of the modern data center. For z14 there is an option to order the system with thin front and rear door covers if required.

- FICON Express16S+ with Forward Error Correction (FEC)

  FICON Express16S+ features continues to provide a new standard for transmission of 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 running at 16 Gbps can take advantage of 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, while retaining the same reliability and robustness for which FICON channels have traditionally been 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 should see fewer I/O errors, reducing the potential impact 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), use link data returned from Read Diagnostic Parameters to differentiate between errors because of failures in the optics versus failures because of dirty or faulty links. Key metrics can be displayed on the operator console. Figure 9-3 displays the results of a display matrix command with the LINKINFO=FIRST parameter which will collect information from each device in the path from the channel to the I/O device. The output displays:

  – 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

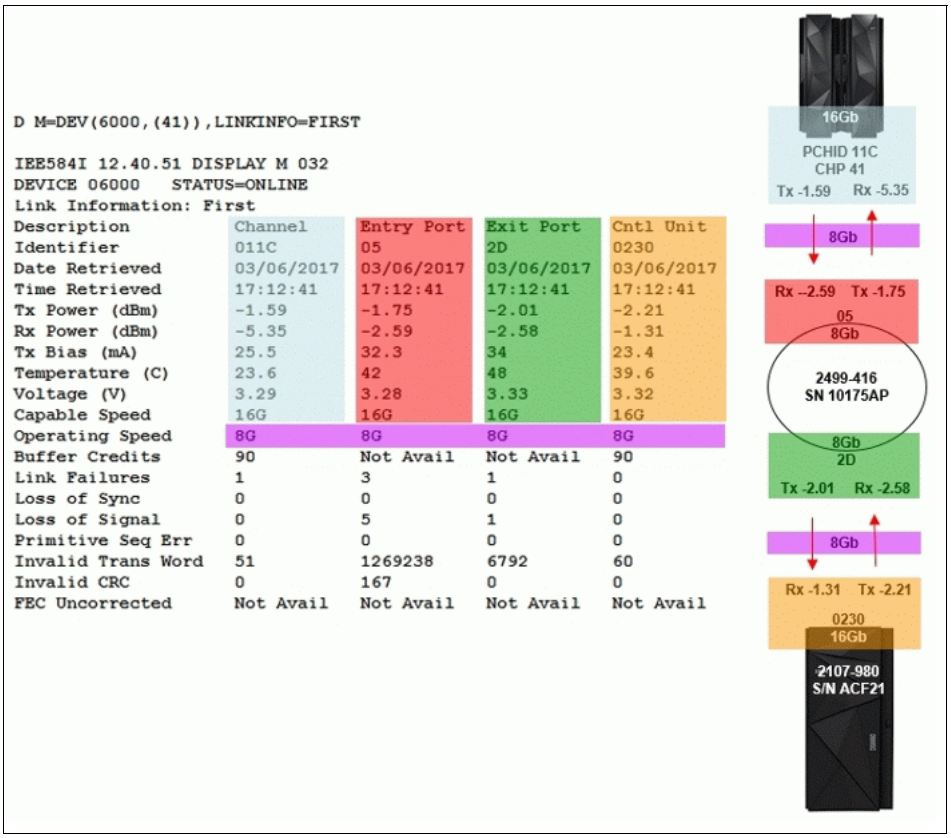


Figure 9-3 Read Diagnostic Parameters function

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 usage of storage area network (SAN) dynamic routing policies in the fabric. With the z14 server, 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 based on the Fibre Channel Exchange ID. Each I/O operation has a unique exchange ID. FIDR is designed to support static SAN routing policies, and dynamic routing policies.

  FICON Dynamic Routing can help clients reduce costs by providing these abilities:

  – 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 assigning 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 RAS functions

Hardware RAS function improvements focus on addressing all sources of outages. Sources of outages have these classifications:

Unscheduled This outage occurs because of an unrecoverable malfunction in a hardware component of the system.

Scheduled This outage is caused by changes or updates that must be done to the system in a timely fashion. A scheduled outage can be caused by a disruptive patch that must be installed, or other changes that must be made to the system.

Planned This outage is caused by changes or updates that must be done to the system. A planned outage can be caused by a capacity upgrade or a driver upgrade. A planned outage is usually requested by the client, and often requires pre-planning. The z14 design phase focuses on enhancing planning to simplify or eliminate planned outages.

The difference between scheduled outages and planned outages is, perhaps, not obvious. The general consensus is that scheduled outages take place 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. This chapter does not distinguish between scheduled and planned outages.

Preventing unscheduled, scheduled, and planned outages has been addressed by the IBM System z system design for many years.

z14 servers introduce a fixed size HSA of 192 GB. This size helps eliminate pre-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.

In addition, by addressing the elimination of planned outages, the following tasks 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 307.

[2] Some pre-planning considerations might exist. For more information, see Chapter 8, “System upgrades” on page 299.

9.7.1 Scheduled outages

Concurrent hardware upgrades, concurrent parts replacement, concurrent driver upgrades, and concurrent firmware fixes that are available with z14 servers all address the elimination of scheduled outages. Furthermore, the following indicators and functions that address scheduled outages are included:

- Double memory data bus lane sparing.

  This feature reduces the number of repair actions for memory.
- Single memory clock sparing.
- Double DRAM chipkill tolerance.
- Field repair of the cache fabric bus.
- Processor drawer power distribution *N*+2 design.

  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. If you can use EDA, the replacement of POL cards is concurrent for the whole Z server.
- Redundant N+1 System Control Hubs (SCHs).

  The bulk power hub (BPH) in former Z servers has been repacked into a new part, the SCH. The SCH contains an Ethernet hub, a flexible support processor (FSP), power supplies for the SEs and displays, and blowers. There are two SCHs in the A frame for redundancy.
- Redundant (*N*+2) humidity sensors.
- Redundant (*N*+2) altimeter sensors.
- Redundant (*N*+2) ambient temperature 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 time to concurrently repair the storage module if the z14 is a multidrawer system. The process to repair the storage module is to isolate or “fence off” the drawer, remove the drawer, replace the failing storage module, and then add the drawer. The flexible memory option might be necessary to maintain sufficient capacity while repairing the storage module.
- Single processor core checkstop and sparing.

  This indicator shows that a processor core has malfunctioned and is *spared*. IBM determines what to do based on the system and the history of that system.
- Point-to-point fabric for symmetric multiprocessing (SMP).

  Having fewer components that can fail is an advantage. In a multidrawer system, all of the drawers are connected by point-to-point connections. A drawer can always be added concurrently.
- Air-cooled system: radiator with redundant (*N*+1) pumps.

  z14 servers implement true *N*+1 redundancy on pumps and blowers for the radiator. The radiator cooling system can support one and two CPC drawers simultaneously with a redundant design that consists of two pumps and two blowers. One active pump and blower can support the entire system load. If you have three or four CPC drawers, one additional blower is there, and one active pump and two blowers can support the entire system load. The replacement of a pump or blower causes no performance impact.

- Water-cooled system: N+1 Water-Cooling Units (WCUs).

  A water-cooling system is an option in z14 servers, with WCU technology. Two redundant WCUs run with two independent chilled water feeds. One WCU and one water feed can support the entire system load. The water-cooled configuration is backed up by the rear door heat exchangers in the rare event of a problem with the chilled water facilities of the customer.

- Hot-swap IFB hub cards.

  When correctly configured for redundancy, hot swapping (replacing) the IFB (HCA2-O (12xIFB) or HCA3-O (12xIFB)) hub cards is possible. This process avoids any interruption when you must replace these types of cards.

- Redundant 1 Gbps Ethernet service network with virtual LAN (VLAN).

  The service network in the system gives the machine code the capability to monitor each internal function in the system. This process helps to identify problems, maintain the redundancy, and concurrently replace a part. Through the implementation of the VLAN to the redundant internal Ethernet service network, these advantages are improved, making the service network easier to handle and more flexible.

- The PCIe I/O drawer is available for z14 servers. It and all of the PCIe I/O drawer-supported features can be installed concurrently.

- Memory interface logic to maintain channel synchronization when one channel goes into replay. z14 servers can isolate recovery to just the failing channel.

- Out of band access to DIMM (for background maintenance functions).

  Out of band access (using an I2C interface) allows maintenance (such as logging) without disrupting customer memory accesses.

- Lane shadowing function to each lane that periodically is taken offline (for recalibration).

  The (logical) spare bit lane is rotated through the (physical) lanes. This allows the lane to be tested and recalibrated transparently to customer operations.

- Automatic lane recalibration on offline lanes on the main memory interface. Hardware support for transparent recalibration.

- Automatic dynamic lane sparing based on pre-programmed CRC thresholds on the main memory interface. Hardware support to detect a defective lane and spare it out.

- Improved DIMM exerciser for testing memory during IML.

- PCIe redrive hub cards plug straight in (no blind mating of connector). Simplified plugging that is more reliable.

- ICA (short distance) coupling cards plug straight in (no blind mating of connector). Simplified plugging that is more reliable.

- Coupling Express LR (CE LR) coupling cards plug into the PCIe I/O Drawer allowing more connections with a faster bandwidth.

- Hardware driven dynamic lane sparing on fabric (SMP) buses. More bit lane sparing.

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 Z servers 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 addition of one memory channel that is dedicated for RAS. The parity of the four data DIMMs is stored in the DIMMs that are attached to the fifth memory channel. Any failure in a memory component can be detected and corrected dynamically. z14 servers inherited this memory architecture.

  The memory system on z14 servers 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 server further improved this chip marking technology. There is graduated DRAM marking, and there are channel marking and scrubbing calls for replacement on the third DRAM failure. For more information about the memory system on z14 servers, see 2.4, “Memory” on page 52.

- Improved thermal, altitude, and condensation management
- Soft-switch firmware

  z14 servers are equipped with the capabilities of soft-switching firmware. Enhanced logic in this function ensures that every affected circuit is powered off during the soft-switching of firmware components. For example, when you are upgrading the microcode of a FICON feature, enhancements are implemented to avoid any unwanted side effects that were detected on previous systems.

- Server Time Protocol (STP) recovery enhancement

  When HCA3-O (12xIFB) or HCA3-O Long Reach (LR) (1xIFB) or PCIe based integrated communication adapter (ICA) Short Reach (SR) links are used, an unambiguous “going away signal” is sent when the server on which the coupling link is running is about to enter a failed (check stopped) state.

  When the “going away signal” that is sent by the Current Time Server (CTS) in an STP-only Coordinated Timing Network (CTN) is received by the Backup Time Server (BTS), the BTS can safely take over as the CTS without relying on the previous Offline Signal (OLS) in a two-server CTN, or as the Arbiter in a CTN with three or more servers.

  Enhanced Console Assisted Recovery (ECAR) 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 389.

  Coupling Express LR does not support the “going away signal”, however ECAR can be used to assist with recovery in these configurations.

- 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 enhanced drawer availability

EDA is a procedure where a drawer in a multidrawer system can be removed and reinstalled during an upgrade or repair action. This procedure has no impact on the running workload.

The EDA procedure and careful planning help ensure that all the resources are still available to run critical applications in an (*n*-1) drawer configuration. This process allows you to avoid planned outages. Consider the flexible memory option to provide more memory resources when you are replacing a drawer. For more information about flexible memory, see 2.4.7, “Flexible Memory Option” on page 61.

To minimize the effect on current workloads, ensure that sufficient inactive physical resources exist on the remaining drawers to complete a drawer removal. Also, consider deactivating non-critical system images, such as test or development LPARs. After you stop these non-critical LPARs and free their resources, you might find sufficient inactive resources to contain critical workloads while completing a drawer replacement.

9.8.1 EDA planning considerations

To use the EDA function, configure enough physical memory and engines so that the loss of a single drawer does not result in any degradation to critical workloads during the following occurrences:

- A degraded restart in the rare event of a drawer failure
- A drawer replacement for repair or a physical memory upgrade

The following configurations especially enable the use of the EDA function. These z14 models need enough spare capacity so that they can cover the resources of a fenced or isolated drawer. This configuration imposes limits on the number of the client-owned PUs that can be activated when one drawer within a model is fenced:

- A maximum of 69 client PUs are configured on the M02.
- A maximum of 105 client PUs are configured on the M03.
- A maximum of 141 client PUs are configured on the M04.
- A maximum of 170 client PUs are configured on the M05.
- No special feature codes are required for PU and model configuration.
- For all z14 models, there are five SAPs in every drawer (model M05 has 23 total).
- The flexible memory option delivers physical memory so that 100% of the purchased memory increment can be activated even when one drawer is fenced.

The system configuration must have sufficient dormant resources on the remaining drawers in the system for the *evacuation* of the drawer that is to be replaced or upgraded. Dormant resources include the following possibilities:

- Unused PUs or memory that are not enabled by LICCC
- Inactive resources that are enabled by LICCC (memory that is not being used by any activated LPARs)
- Memory that is purchased with the flexible memory option
- Additional drawers

The I/O connectivity must also support drawer removal. Most of the paths to the I/O have redundant I/O interconnect support in the I/O infrastructure (drawers) that enable connections through multiple fanout cards.

If sufficient resources are not present on the remaining drawers, certain non-critical LPARs might have to be deactivated. One or more CPs, specialty engines, or storage might have to be configured offline to reach the required level of available resources. Plan to address these possibilities to help reduce operational errors.

Exception: Single-drawer systems cannot use the EDA procedure.

Include the planning as part of the initial installation and any follow-on upgrade that modifies the operating environment. A client can use the Resource Link machine information report to determine the number of drawers, active PUs, memory configuration, and channel layout.

If the z14 server is installed, click **Prepare for Enhanced Drawer Availability** in the Perform Model Conversion window of the EDA process on the Hardware Management Console (HMC). This task helps you determine the resources that are required to support the removal of a drawer with acceptable degradation to the operating system images.

The EDA process determines which resources, including memory, PUs, and I/O paths, are free to allow for the removal of a drawer. You can run this preparation on each drawer to determine which resource changes are necessary. Use the results as input in the planning stage to help identify critical resources.

With this planning information, you can examine the LPAR configuration and workload priorities to determine how resources might be reduced and still allow the drawer to be concurrently removed.

Include the following tasks in the planning process:

- Review of the z14 configuration to determine the following values:
  - Number of drawers that are installed and the number of PUs enabled. Note the following information:
    - Use the Resource Link machine information or the HMC to determine the model, number, and types of PUs (CPs, IFLs, ICFs, and zIIPs).
    - Determine the amount of memory, both physically installed and LICCC-enabled.
    - Work with your IBM Service Support Representative (IBM SSR) to determine the memory card size in each drawer. The memory card sizes and the number of cards that are installed for each drawer can be viewed from the SE under the CPC configuration task list. Use the View Hardware Configuration option.
  - Channel layouts, ICA, and host channel adapter (HCA) to channel connections.

    Use the Resource Link machine information to review the channel configuration, including the HCA paths. This process is a normal part of the I/O connectivity planning. The alternative paths must be separated as far into the system as possible.
- Review the system image configurations to determine the resources for each image.
- Determine the importance and relative priority of each LPAR.
- Identify the LPAR or workloads and the actions to be taken:
  - Deactivate the entire LPAR.
  - Configure PUs.
  - Reconfigure memory, which might require the use of reconfigurable storage unit (RSU) values.
  - Vary off the channels.

- Review the channel layout and determine whether any changes are necessary to address single paths.
- Develop a plan to address the requirements.

When you perform the review, document the resources that can be made available if the EDA is used. The resources on the drawers are allocated during a POR of the system and can change after that process. Perform a review when changes are made to z14 servers, such as adding drawers, CPs, memory, or channels. Also, perform a review when workloads are added or removed, or if the HiperDispatch feature was enabled and disabled since the last time you performed a POR.

9.8.2 Enhanced drawer availability processing

To use the EDA, first ensure that the following conditions are satisfied:

- Free the used processors (PUs) on the drawer that will be removed.
- Free the used memory on the drawer.
- For all I/O domains that are connected to the drawer, ensure that alternative paths exist. Otherwise, place the I/O paths offline.

For the EDA process, this is the preparation phase. It is started from the SE, either directly or on the HMC by using the **Single object operation** option on the Perform Model Conversion window from the CPC configuration task list, as shown in Figure 9-4 on page 361.

Processor availability

Processor resource availability for reallocation or deactivation is affected by the type and quantity of the resources in use:

- Total number of PUs that are enabled through LICCC
- PU definitions in the profiles that can be dedicated and dedicated reserved or shared
- Active LPARs with dedicated resources at the time of the drawer repair or replacement

To maximize the PU availability option, ensure that sufficient inactive physical resources are on the remaining drawers to complete a drawer removal.

Memory availability

Memory resource availability for reallocation or deactivation depends on these factors:

- Physically installed memory
- Image profile memory allocations
- Amount of memory that is enabled through LICCC
- Flexible memory option
- Virtual Flash Memory if enabled and configured

For more information, see 2.6.2, “Enhanced drawer availability” on page 67.

Fanout card to I/O connectivity requirements

The optimum approach is to maintain maximum I/O connectivity during drawer removal. The redundant I/O interconnect (RII) function provides for redundant HCA connectivity to all installed I/O domains in the PCIe I/O drawers, and I/O drawers.

Preparing for enhanced drawer availability

The Prepare Concurrent Drawer replacement option validates that enough dormant resources exist for this operation. If enough resources are not available on the remaining drawers to complete the EDA process, the process identifies those resources. It then guides

you through a series of steps to select and free up those resources. The preparation process does not complete until all processors, memory, and I/O conditions are successfully resolved.

Preparation: The preparation step does not reallocate any resources. It is only used to record client choices and produce a configuration file on the SE that is used to run the concurrent drawer replacement operation.

The preparation step can be done in advance. However, if any changes to the configuration occur between the preparation and the physical removal of the drawer, you must rerun the preparation phase.

The process can be run multiple times because it does not move any resources. To view the results of the last preparation operation, click **Display Previous Prepare Enhanced Drawer Availability Results** from the Perform Model Conversion window in the SE.

The preparation step can be run without performing a drawer replacement. You can use it to dynamically adjust the operational configuration for drawer repair or replacement before IBM SSR activity. Figure 9-4 shows the Perform Model Conversion window where you click **Prepare for Enhanced Drawer Availability**.

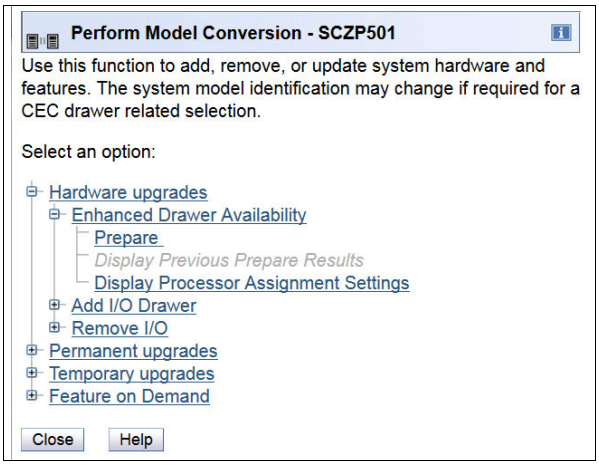


Figure 9-4 Clicking Prepare for Enhanced Drawer Availability

After you click **Prepare for Enhanced Drawer Availability**, the Enhanced Drawer Availability window opens. Select the drawer that is to be repaired or upgraded, then select **OK**, as shown in Figure 9-5. Only one target drawer can be selected at a time.

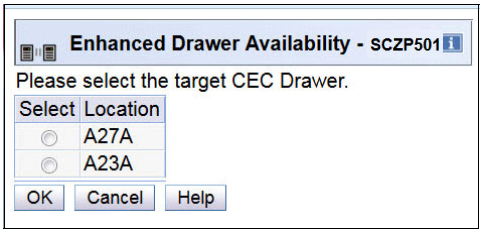


Figure 9-5 Selecting the target drawer

The system verifies the resources that are required for the removal, determines the required actions, and presents the results for review. Depending on the configuration, the task can take from a few seconds to several minutes.

The preparation step determines the readiness of the system for the removal of the targeted drawer. The configured processors and the memory in the selected drawer are evaluated against unused resources available across the remaining drawers. The system also analyzes I/O connections that are associated with the removal of the targeted drawer for any single path I/O connectivity.

If insufficient resources are available, the system identifies the conflicts so that you can free other resources.

Three states can result from the preparation step:

- The system is ready to run the EDA for the targeted drawer with the original configuration.
- The system is not ready to run the EDA because of conditions that are indicated by the preparation step.
- The system is ready to run the EDA for the targeted drawer. However, to continue with the process, processors are reassigned from the original configuration. Review the results of this reassignment relative to your operation and business requirements. The reassignments can be changed on the final window that is presented. However, before making changes or approving reassignments, ensure that the changes are reviewed and approved by the correct level of support based on your organization's business requirements.

Preparation tabs

The results of the preparation are presented for review in a tabbed format. Each tab indicates conditions that prevent the EDA option from being run. Tabs are for processors, memory, and various single path I/O conditions. The following tab selections are available:

- Processors
- Memory
- Single I/O
- Single Domain I/O
- Single Alternate Path I/O

Only the tabs that have conditions that prevent the drawer from being removed are displayed. Each tab indicates the specific conditions and possible options to correct them.

The preparation identifies, for example, single I/O paths that are associated with the removal of the selected drawer. These paths must be varied offline to perform the drawer removal. After you address the condition, rerun the preparation step to ensure that all the required conditions are met.

Preparing the system to perform enhanced drawer availability

During the preparation, the system determines the CP configuration that is required to remove the drawer. Figure 9-6 shows the results and provides the option to change the assignment on non-dedicated processors.

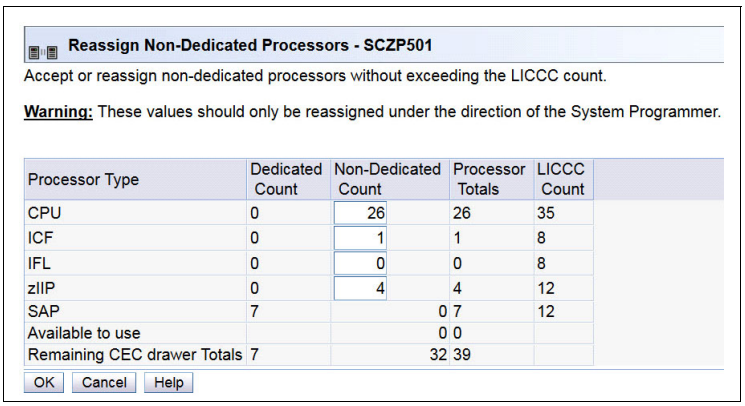


Figure 9-6 Reassign Non-Dedicated Processors results

Important: Consider the results of these changes relative to the operational environment. Understand the potential impact of making such operational changes. Changes to the PU assignment, although technically correct, can result in constraints for critical system images. In certain cases, the solution might be to defer the reassignments to another time that has less impact on the production system images.

After you review the reassignment results and make any necessary adjustments, click **OK**.

The final results of the reassignment, which include the changes that are made as a result of the review, are displayed as shown in Figure 9-7. These results are the assignments when the drawer removal phase of the EDA is completed.

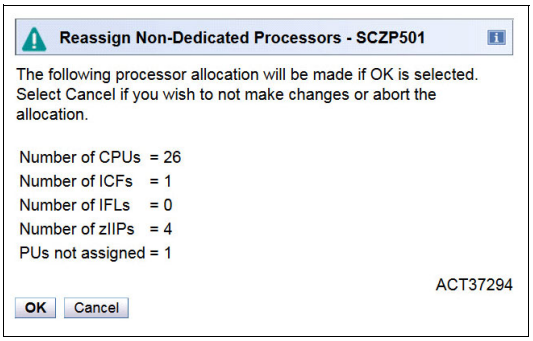


Figure 9-7 Reassign Non-Dedicated Processors, message ACT37294

Summary of the drawer removal process steps

To remove a drawer, the following resources must be moved to the remaining active drawers:

- PUs: Enough PUs must be available on the remaining active drawers, including all types of PUs that can be characterized (CPs, IFLs, ICFs, zIIPs, the IFP[3], and SAPs).
- Memory: Enough installed memory must be available on the remaining active drawers.
- I/O connectivity: Alternative paths to other drawers must be available on the remaining active drawers, or the I/O path must be taken offline.

By understanding both the system configuration and the LPAR allocation for memory, PUs, and I/O, you can make the best decision about how to free the necessary resources to allow for drawer removal.

To replace concurrently a drawer, complete these steps:

1. Run the preparation task to determine the necessary resources.
2. Review the results.
3. Determine the actions to perform to meet the required conditions for EDA.
4. When you are ready for the drawer removal, free the resources that are indicated in the preparation steps.
5. Rerun the step in Figure 9-4 on page 361 (the Prepare for Enhanced Drawer Availability task) to ensure that the required conditions are all satisfied.
6. Upon successful completion, the system is ready for the removal of the drawer.

The preparation process can be run multiple times to ensure that all conditions are met. It does not reallocate any resources. It only produces a report. The resources are not reallocated until the Perform Drawer Removal process is started.

Rules during EDA

During EDA, the following processor, memory, and single I/O rules are enforced:

- Processor rules

  All processors in any remaining drawers are available to be used during EDA. This requirement includes the two spare PUs or any available PU that is non-LICCC.

  The EDA process also allows conversion of one PU type to another PU type. One example is converting a zIIP to a CP during the EDA function. The preparation for the concurrent drawer replacement task indicates whether any SAPs must be moved to the remaining drawers.

- Memory rules

  All physical memory that is installed in the system, including flexible memory, is available during the EDA function. Any physical installed memory, whether purchased or not, is available to be used by the EDA function.

- Single I/O rules

  Alternative paths to other drawers must be available, or the I/O path must be taken offline.

Review the results. The result of the preparation task is a list of resources that must be made available before the drawer replacement can take place.

[3] If any native PCIe features are installed on the system.

Freeing any resources

At this stage, create a plan to free these resources. The following list shows the resources and actions that are necessary to free them:

- To free any PUs:
  - Vary off the CPs by using the Perform a Model Conversion window, which reduces the number of CPs in the shared CP pool.
  - Deactivate the LPARs.
- To free memory:
  - Deactivate an LPAR.
  - Vary offline a portion of the reserved (online) memory. For example, in z/OS, run the following command:

    ```
    CONFIG_STOR(E=1),<OFFLINE/ONLINE>
    ```

    This command enables a storage element to be taken offline. The size of the storage element depends on the RSU value. In z/OS, the following command configures offline smaller amounts of storage than the amount that was set for the storage element:

    ```
    CONFIG_STOR(nnM),<OFFLINE/ONLINE>
    ```
  - A combination of both LPAR deactivation and varying memory offline.

Reserved storage: If you plan to use the EDA function with z/OS LPARs, set up reserved storage and an RSU value. Use the RSU value to specify the number of storage units that are to be kept free of long-term fixed storage allocations. This configuration allows for storage elements to be varied offline.

9.9 z14 Enhanced Driver Maintenance

EDM is one more step toward reducing both the necessity for and the duration of a scheduled outage. One of the components to planned outages is LIC Driver updates that are run in support of new features and functions.

When correctly configured, z14 servers support concurrently activating a selected new LIC Driver level. Concurrent activation of the selected new LIC Driver level is supported only at specific released sync points. Concurrently activating a selected new LIC Driver level anywhere in the maintenance stream is not possible. Certain LIC updates do not allow a concurrent update/upgrade.

Consider the following key points of EDM:

- The HMC can query whether a system is ready for a concurrent driver upgrade.
- Previous firmware updates, which require an initial machine load (IML) of the z14 system 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.
- z14 servers introduce Concurrent Driver Upgrade (CDU) cloning support to other CPCs for CDU preinstallation and activation.

- Concurrent crossover from Driver level *N* to Driver level *N*+1, 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 initiated. 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.
- 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 installation of 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

9.9.1 Resource Group and native PCIe MCLs

Microcode fixes, referred to as individual MCLs or packaged in Bundles, will occasionally 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, there will be times the maintenance updates will require the Resource Group or the effected native PCIe to be toggled offline and online to implement the updates.

Native PCIe cards for z14

native PCIe adapter type	Feature Code	Resource required to be offline
10GbE RoCE Express	0412	FID(s) / PCHID
ZEDC Express	0420	FID(s) / PCHID
zHyperlink Express	0431	
Coupling Express LR	0433	CHPID(s) / PCHID

- Updates to the Resource Group require all native PCIe adapters installed in that RG to be offline. The requirement is depicted in the Manage PCIe
- Updates to the native PCIe adapter requires the adapter to be offline. If the adapter is not defined, the MCL session will automatically install the maintenance related to the adapter.

To help alleviate the amount of adapters impacted for Resource Group code updates, z14 increased the amount of Resource Groups from two with previous systems to four. The re

PCIe native adapters are configured with FIDs (Function IDs) and may need to be configured off line when changes to code are necessary

Note: Other adapter types such as FICON Express, OSA Express, and Crypto Express installed in the PCIe I/O drawer are not effected as they are not managed by the Resource Groups

Figure 9-8 below displays the front, rear and top view of the PCIe I/O drawer and the Resource Group assignment by card slot. All PCIe I/O drawers installed in the system have the same Resource Group assignment.

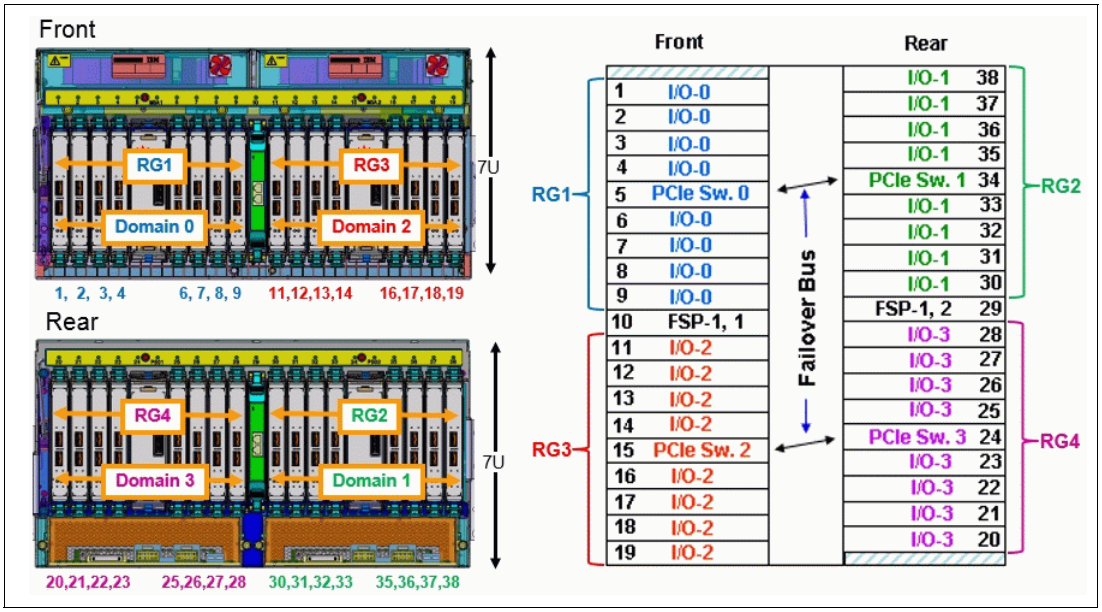


Figure 9-8 Resource Group slot assignment

The adapter locations and PCHIDs for the four Resource Groups

Table 9-1 Resource Group affinity to native PCIe adapter locations

Resource Group	Adapter Locations	PCHIDs	
RG1 Front left	Z22BLG01-04,06-09	100-11F	
	Z15BLG01-04,06-09	180-19F	
	Z08BLG01-04,06-09	200-21F	
	Z01BLG01-04,06-09	280-29F	
	A32BLG01-04,06-09	300-31F	
RG2 Rear right	Z22BLG30-33,35-38	160-17F	
	Z15BLG30-33,35-38	1E0-1FF	
	Z08BLG30-33,35-38	260-27F	
	Z01BLG30-33,35-38	2E0-2FF	
	A32BLG30-33,35-38	360-37F	

Resource Group	Adapter Locations	PCHIDs	
RG3 Front right	Z22BLG11-14,16-19	120-13F	
	Z15BLG11-14,16-19	1A0-1AF	
	Z08BLG11-14,16-19	220-22F	
	Z01BLG11-14,16-19	2A0-2BF	
	A32BLG11-14,16-19	320-33F	
RG4 Rear left	Z22BLG20-23,25-28	140-15F	
	Z15BLG20-23,25-28	1C0-1DF	
	Z08BLG20-23,25-28	240-25F	
	Z01BLG20-23,25-28	2C0-2DF	
	A32BLG20-23,25-28	340-35F	

z14 servers introduce the support to concurrently[4] activate an MCL on an OSA-ICC channel to improve the availability and simplification of the firmware maintenance. The OSD channels already have this capability.

9.10 RAS capability for the Hardware Management Console and Support Element

The HMC and the SE have the following RAS capabilities:

- Back up from HMC and SE

  For the customers who do not have a FTP server defined for backups, new with z14, the HMC can be configured as a FTP server.

  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 alternate SE HDD. In addition, you can save the backup to a defined FTP server.

  For more information, see 11.2.5, “New backup options for HMCs and primary SEs” on page 395.

- 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 406.

- 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 412.

- Support Element (SE)

[4] For up-to-date information, see *Driver 27 Service Exception Letter*, GC28-6957.

z14 servers are provided with two 1U trusted servers inside the IBM Z server A frame. One is always the primary SE and the other is the alternate SE. The primary SE is the active one. The alternate acts as the backup. Once per day, information is mirrored. The SE servers have N+1 redundant power supplies.

For more information, see 11.2.4, “New Support Elements” on page 394.

- Hardware Management Console (HMC) in an ensemble

  The serviceability function for the components of an ensemble is delivered through the traditional HMC/SE constructs, as for earlier Z severs. 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. Therefore, each ensemble must be equipped with two HMCs: A primary and an alternate. The primary HMC performs all HMC activities (including Unified Resource Manager activities). The alternate is only the backup. The alternate cannot be used for tasks or activities.

  Failover: The primary HMC and its alternate must be connected to the same LAN segment. This configuration allows the alternate HMC to take over the IP address of the primary HMC during failover processing.

- Alternate HMC preload function

  The Manage Alternate HMC task allows you to reload internal code onto the alternate HMC to minimize HMC downtime during an upgrade to a new driver level. After the new driver is installed on the alternate HMC, it can be made active by running an HMC switchover.

10

Environmental requirements

This chapter addresses the environmental requirements for IBM z14 servers. It lists the dimensions, weights, power, and cooling requirements that are needed to plan for the installation of an z14 server.

Select from these options for the physical installation of the server:

- Air or water 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 frames
- A high-voltage DC power supply or the usual AC power supply

For more information about physical planning see *IBM z14 Installation Manual for Physical Planning,* GC28-6965.

This chapter includes the following sections:

- Power and cooling
- Physical specifications
- Physical planning
- Energy management

© Copyright IBM Corp. 2017. All rights reserved.

10.1 Power and cooling

Thez14 server is always a two-frame system. Frames are shipped separately and then are bolted together during the installation procedure. z14 servers support installation on a raised floor or non-raised floor. It is worth to note that the water-cooling feature can be installed only on a raised floor as water hoses are attached to the server from underneath the raised floor. Standard exit for power and I/O cables is also located on the bottom of the server frames unless the tope exit features are installed:

- Top Exit I/O Cabling feature code (FC 7942)
- Top Exit cord DC (FC 8948)
- Top Exit cord Low Voltage, 3 phase (FC 8949)
- Top Exit cord HiV, 3 phase (FC 8951)

These options allow I/O cables and power cables to exit through the top of the server into overhead cabling rails.

10.1.1 Rear cover design for vectored air output

The rear door covers were designed to address data center airflow challenges. You can change the cover (door) fin orientation down or up, as shown in Figure 10-1. This allows you to direct the hot air that is exhausted by z14 servers to avoid adversely affecting cooling of other systems on the hot aisle.

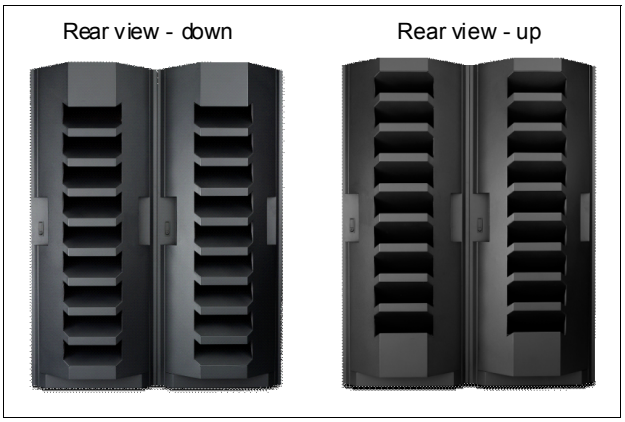


Figure 10-1 Rear cover vectored air output

The new rear doors are all the same part. In the installation planning meeting, you can decide in which orientation the IBM Service Support Representative (IBM SSR) should install the covers. For more information about the vectored orientation, see the *IBM z14 Installation Manual for Physical Planning,* GC28-6965 or ask your IBM SSR.

10.1.2 Power requirements and consumption

This section describes the power requirements and consumption for z14 servers.

Power requirements

The system operates with two fully redundant power supplies. One is in the front, and the other is in the rear of the Z frame. Each power supply has either one or two power cords. The number of power cords required depends on the system configuration. The total loss of one power supply has no impact on system operation.

Systems with two power cords (one in the front and one in the rear) can be started with one power cord and continue to run.

The larger systems with a minimum of four bulk power regulator (BPR) pairs must have four power cords installed. A system with four power cords can be started with two power cords on the *same power supply* with sufficient power to keep the system running.

Power cords can be attached to either 3-phase, 50/60 Hz, 200 - 480 V AC power, or 380 - 520V DC power.

The High-Voltage Direct Current (HVDC) feature is an option for z14 servers. It allows to directly use the high voltage (HV) DC distribution. A direct HVDC data center power design improves data center energy efficiency by removing the need for a DC to AC inversion step.

The z14 bulk power supplies have been modified to support HVDC, so the only difference in the included hardware to implement this option is the DC power cords. HVDC is a quite fresh technology with many standards. z14 server supports two of them - ground referenced and dual polarity HVDC supplies, such as +/-190 V, +/-260 V, and +380 V. HVDC brings many advantages. Beyond the data center uninterruptible power supply and power distribution energy savings, a z14 server that runs on HVDC power draws 1 - 3% less input power. HVDC does not change the number of power cords that a system requires.

For additional equipment, such as the Hardware Management Console (HMC), its display, and Ethernet switch, extra single-phase outlets are required.

The power requirements depend on the cooling facility installed, the number of central processor complex (CPC) drawers, and the number of I/O units.

If you initially need only one power cord pair but you plan to use a second pair in the future, you can order the Line Cord Plan Ahead feature (FC 2000). This feature gives you four power cords at the initial configuration. Additionally, Balanced Power Plan Ahead feature (FC 3003) results in initial configuration with four power cords and twelve BPRs. If the z14 server is configured with the Internal Battery Feature (IBF), Balanced Power Plan Ahead automatically supplies the maximum number of batteries, six IBFs, with the system.

Table 10-1 lists the BPR requirements depending on the number of CPC drawers and I/O units. A second pair of power cords is installed if the number of BPR pairs is four or higher.

Table 10-1 Number of BPRs that are required per side

CPC Drawers	I/O units[a]					
	0	1	2	3	4	5
1	2[b]	2[b]	2[b]	2[b]	3[b]	3[b]
2	2[b]	3[b]	3[b]	3[b]	3[b]	4[c]
3	3[c]	3[c]	4[c]	4[c]	4[c]	5[c]
4	4[c]	4[c]	5[c]	5[c]	5[c]	6[c]

a. I/O units = the number of I/O drawers or PCIe I/O drawers.
b. Single-line power cord pair.
c. Two-line power cord pair.

Table 10-2 shows the number of power cords installed on one power supply, which depends on the number of I/O units and the number of CPC drawers.

Table 10-2 Number of power cords that are installed per power supply

CPC Drawers	Number of I/O units[a]					
	0	1	2	3	4	5
1	1	1	1	1	1	1
2	1	1	1	1	1	2
3	1	1	2	2	2	2
4	2	2	2	2	2	2

a. I/O units = the number of I/O drawers or PCIe I/O drawers.

Power consumption

This section lists the maximum power consumption for the air-cooled and water-cooled models.

Consideration: Power consumption is lower when in a normal ambient temperature room, and for configurations with lesser number of I/O slots, smaller amount of memory, and with fewer processors. Power consumption is also slightly lower for DC input voltage. The numbers below assume that batteries are present and charging.

Power estimation for any configuration, power source, and room condition can be obtained by using the power estimation tool at IBM Resource Link (authentication required), found at the following website:

https://www.ibm.com/servers/resourcelink/

On the Resource Link page, click **Tools** → **Power and weight estimation**.

Table 10-3 lists the absolute maximum power consumption for the air-cooled models in a warm room (power will be lower for DC input voltage).

Table 10-3 Power consumption for air-cooled models

CPC drawers	I/O units[a]					
	0	1	2	3	4	5
1	6.5	8.4	10.2	11.3	12.4[b]	13.5[c]
2	11.1	13.0	14.9	16.7	18.6	20.3
3	15.6	17.5	19.4	21.2	23.1	24.8
4	20.4	22.3	24.2	25.9	27.9	29.6
5	20.6	22.5	24.4	26.1	28.1	29.8

a. I/O units = the number of I/O drawers or PCIe I/O drawers.
b. Three PCIe I/O drawers and one or two I/O drawers.
c. Assumes the maximum supported configuration (maximum I/O adapters are installed).

Table 10-4 lists the absolute maximum power consumption for the water-cooled models in a warm room (power will be lower for DC input voltage).

Table 10-4 Power consumption for water-cooled models<< wait for input from frank bosco >>

Model (number of CPC drawers)	Number of I/O units[a]					
	0	1	2	3	4	5
1	6.2	7.9	9.8	10.8	11.9[b]	13.0[c]
2	10.3	12.2	14.1	15.9	17.8	19.5
3	14.7	16.5	18.4	20.1	22.1	24.0
4	19.0	20.8	22.8	24.7	26.4	28.3
5	19.3	21.1	23.1	25.0	26.7	28.6

a. I/O units = the number of I/O drawers or PCIe I/O drawers.
b. Three PCIe I/O drawers and one or two I/O drawers.
c. Assumes a maximum supported configuration (maximum I/O adapters are installed).

10.1.3 Cooling requirements

The z14 cooling system comes with two options: air-cooling or water-cooling. On top of that, single chip modules (SCMs) are always cooled with an internal water loop, no matter which z14 cooling option is chosen. The liquid in internal water system can be cooled using either a radiator (for air-cooling option) or customer-supplied chilled water supply (for water-cooling option). I/O drawers, PCIe I/O drawers, power enclosures, and CPC drawers are cooled by chilled air with blowers.

z14 servers have a recommended (long-term) ambient temperature range from 18°C (64.4°F) to 27°C (80.6°F). The minimum allowed ambient temperature is 15°C (59°F) and the maximum allowed temperature is 32°C (89.6°F).

For more information about the environmental specifications, see *IBM z14 Installation Manual for Physical Planning,* GC28-6965.

Radiator cooling system requirements

The radiator cooling system requires chilled air to fulfill the air-cooling requirements. Normal air exhausts from the front to the rear of the frames. The chilled air is provided through perforated floor panels in front of the system.

Figure 10-2 does not represent any particular server system type, and is intended only to show hot and cold airflow and the arrangement of server aisles.

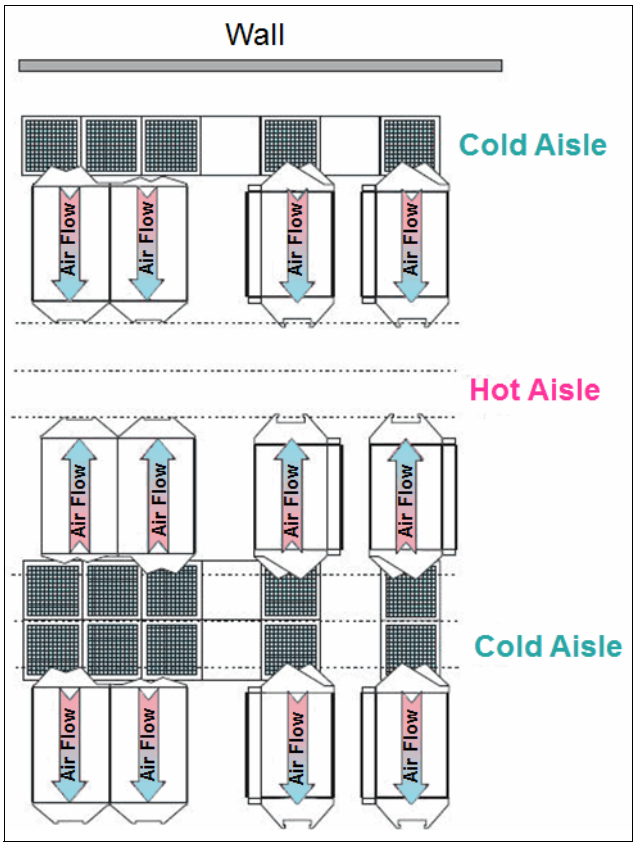


Figure 10-2 Hot and cold aisles

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.

With the standard z14 rear covers (FC #0160), the exiting airflow direction can be customized, which provides you more flexibility in placing z14 servers in your data center. Optional Thin doors (non acoustic, non-vectored) are available also (FC #0161).

For more information about the requirements for air-cooling options, see *IBM z14 Installation Manual for Physical Planning,* GC28-6965.

Water-cooling system requirements

The water-cooling system requires a chilled customer building water supply to be supplied to the z14 water-cooling units (WCUs). A z14 server requires four connections to the facility water: Two feeds and two returns.

Raised floor: The minimum raised floor height for a water-cooled system is 22.86 cm (8.6 in.).

These connections are made by using hoses that are fixed to the facility plumbing and are routed up through the front tailgate of the system. They terminate with quick connect couplings.

Before you install z14 servers with water-cooled option, your facility must meet following requirements:

- Total water hardness must not exceed 200 mg/L of calcium carbonate.
- The pH must be 7 - 9.
- Turbidity must be less than 10 Nephelometric Turbidity Units (NTUs).
- Bacteria must be less than 1000 colony-forming units (CFUs)/ml.
- The water must be as free of particulate matter as feasible.
- The allowable system inlet water temperature range is 6°C to 20°C (43°F to 68°F) using standard building chilled water. A special water system is typically not required.
- The required flow rate to the frame is 3.7 - 79.4 lpm (1 - 21 gpm), depending on the inlet water temperature and the number of processor drawers in the z13 server. Colder inlet water temperatures require less flow than warmer water temperatures. Fewer processor drawers require less flow than a maximum populated z13 server.
- The minimum water pressure that is required across the IBM hose ends is 0.34 - 2.32 BAR (5 - 33.7 psi), depending on the minimum flow required.
- The maximum water pressure that is supplied at the IBM hose connections to the client's water supply cannot exceed 6.89 BAR (100 psi).

For more information about the requirements for water-cooling options, see *IBM z14 Installation Manual for Physical Planning,* GC28-6965*,* and Figure 10-3.

Supply hoses

The z14 water-cooled system includes 4.2 m (13.7 ft) water hoses. Figure 10-3 shows the WCU water supply connections.

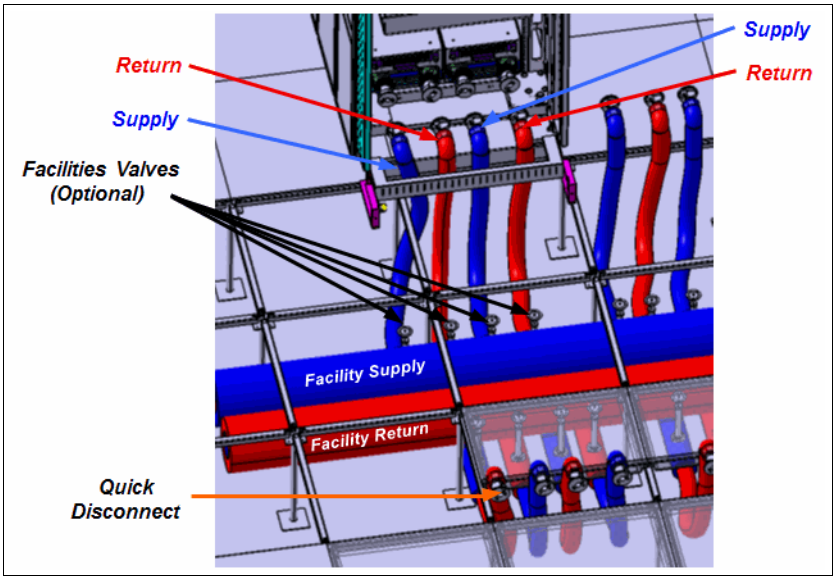


Figure 10-3 WCU water supply connections

The client's ends of the hoses are left open, allowing you to cut the hose to the length that you need. An insulation clamp is provided to secure the insulation and protective sleeving after you cut the hose to the correct length and install it onto your plumbing.

Use shut-off valves in front of the hoses. This configuration allows for the removal of the hoses for a service procedure or relocation. Valves are not included in the order. A stainless steel fitting is available for ordering. The fitting is barbed on one side and has a 2.54 cm (1 in.) male national pipe thread (NPT). For a complete list of tools for the water supply connections, see the *IBM z14 Installation Manual for Physical Planning,* GC28-6965.

10.1.4 Internal Battery Feature (IBF)

In case of power shutdown the optional IBF provides sustained system operations for a relatively short time, allowing a proper z14 server's shutdown. In addition, an external uninterrupted power supply system can be connected, allowing for longer periods of sustained operation.

The IBF can provide emergency power for the estimated time that is listed in Table 10-5. The number of IFBs depends on the number of BPRs. For the number of BPRs that are installed in relation to I/O units and the number of CPC drawers, see Table 10-5. They are installed in pairs. You can have two, four, or six batteries (odd numbers are not allowed).

Table 10-5 Internal Battery Feature holdup times

CPC drawers	Number of I/O units[a]					
	0	1	2	3	4	5
1	19.9 min	13.7 min	10.3 min	8.9 min	13.9 min	12.4 min
2	8.8 min	12.5 min	10.5 min	9.0 min	7.9 min	7.1 min
2	9.6 min	8.3 min	7.4 min	6.6 min	6.1 min	5.0 min
4	6.7 min	6.1 min	5.0 min	4.5 min	4.0 min	3.7 min

a. I/O units = the number of I/O drawers or PCIe I/O drawers.

Consideration: The system holdup times in Table 10-5 assume that both sides are functional and have fresh batteries under normal room ambient conditions.

Holdup times are greater for configurations that do not have every I/O slot plugged, the maximum installed memory, and are not using the maximum processors.

These holdup times are estimates. Your particular battery holdup time for any specific circumstance might be different.

Holdup times vary depending on the number of BPRs that are installed. As the number of BPRs increases, the holdup time also increases until the maximum number of BPRs is reached. After six BPRs (three per side) are installed, no additional batteries are added, so the time decreases from that point.

Holdup times for actual configurations are provided in the power estimation tool at IBM Resource Link (authentication required) found at the following website:

`https://www.ibm.com/servers/resourcelink/`

On the Resource Link page, click **Tools** → **Machine information**, select your Z system, and click **Power Estimation Tool**.

10.1.5 Emergency power-off switch

There is an emergency power-off switch on the front of the A frame. When activated, the switch immediately disconnects utility *and battery power* from the server. This event results in a loss of all volatile data in the server.

If the server is connected to a room's emergency power-off switch, and the IBF is installed, the batteries take over if the switch is engaged.

To avoid the takeover, connect the room emergency power-off switch to the server power-off switch. Then, when the room emergency power-off switch is engaged, all power is disconnected from the power cords and the Internal Battery Features. However, all volatile data in the server is lost.

10.2 Physical specifications

This section describes the weights and dimensions of z14 server. Base overall z14 dimensions are identical with z13 server. The exception is when z14 comes with thin cover option that will shorten the system depth by approximately 15.5" and reduce the system weight by 108 lbs.

z14 server can be installed on a raised or on a non-raised floor. For more information about weight distribution and floor loading tables, see the *IBM z14 Installation Manual for Physical Planning,* GC28-6965. This data is used with the maximum frame weight, frame width, and frame depth to calculate the floor loading.

Table 10-6 indicates the maximum system dimension and weights for the M04/M05 model. The weight ranges are based on configuration models with five PCIe I/O drawers, IBFs, and with the top exit cable features.

Table 10-6 System dimensions and weights

Maximum	**A and Z frames with IBF (FC 3212)**	**A and Z frames with IBFs (FC 3212) and Top Exit Cabling Features (FC 7942 and FC 7901)**
Radiator-cooled servers		
Weight kg (lbs)	2499 (5508)	2566.7 (5657))
Width mm (in.)	1568 (61.7)	1847 (72.7)
Depth mm (in.)	1869 (73.6)	1806 (71.1)
Height mm (in.)	2015 (79.3)	2154 (84.8)
Height reduction mm (in.)	1803 (71.0)	1803 (71.0)
Water-cooled servers		
Weight kg (lbs)	2564 (5653)	2631.7 (5800)
Width mm (in)	1568 (61.7)	1847 (72.7)
Depth mm (in)	1971 (77.7)	1908 (75.1)
Height mm (in)	2015 (79.3)	2154 (84.8)
Height reduction mm (in)	1809 (71.2)	1809 (71.2)

Maximum	A and Z frames with IBF (FC 3212)	A and Z frames with IBFs (FC 3212) and Top Exit Cabling Features (FC 7942 and FC 7901)
Notes: ► Weight is based on the maximum system configuration. ► Weight does not include covers. Covers add 67.7 kg (150 lbs) to each frame. Width, depth, and height are also indicated without covers. ► Be certain that the raised floor on which you are going to install the server can support the weight.		

The power and weight estimation tool for Z servers on Resource Link covers the estimated weight for your designated configuration. It is available on IBM Resource Link (authentication required) at the following website:

http://www.ibm.com/servers/resourcelink

On the Resource Link page, click **Tools** → **Power and weight estimation**.

10.3 Physical planning

This section describes the floor mounting options, and power and I/O cabling options. For more information, see the *IBM z14 Installation Manual for Physical Planning,* GC28-6965.

10.3.1 Raised floor or non-raised floor

z14 servers can be installed on a raised or non-raised floor. The water-cooled models require a raised floor.

Raised floor

If the z14 server is installed in a raised floor environment, both air-cooled and water-cooled models are supported. You can select top exit features to route I/O cables and power cables from the top frame of the z14 server. The following top exit options are available for z14 servers:

- Top Exit I/O Cabling feature code (FC 7942)
- Top Exit Line Cord for DC (FC 8948)
- 3-phase, Low Voltage Top Exit Line Cord (FC 8949)
- 3-phase, High Voltage Top Exit Line Cord (FC 8951)

Figure 10-4 shows the top exit feature options of z14 servers in a raised floor environment.

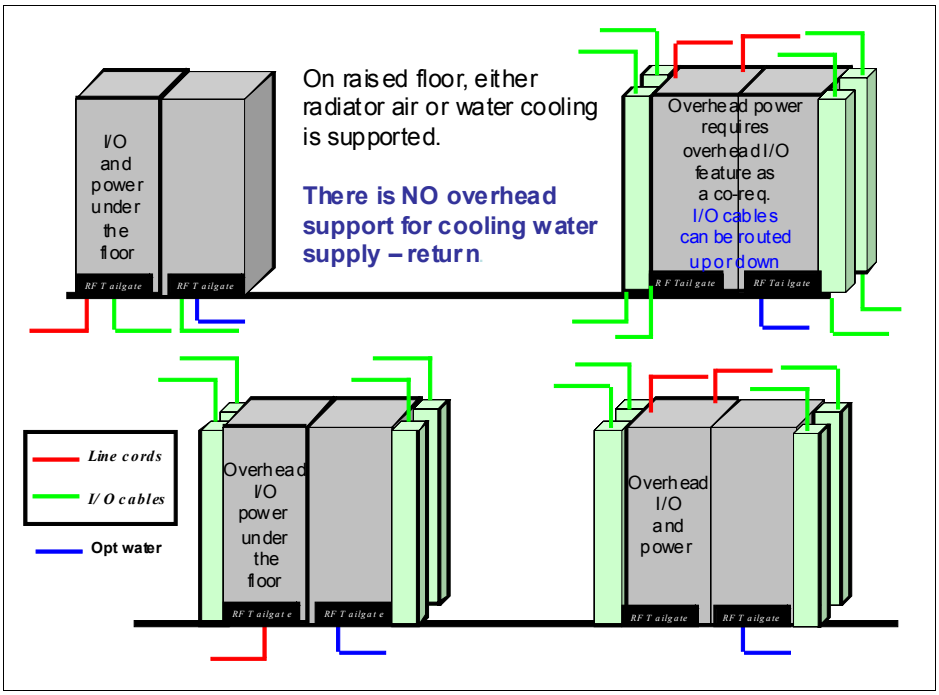


Figure 10-4 Raised floor options

Note: There is no top exit feature support for water hoses, which must go through the system from underneath the raised floor.

Non-raised floor

If you install the z14 server in a non-raised floor environment, you can select only radiator-cooled models. The Non-Raised Floor Support feature code (FC 7998) is required. The Top Exit I/O Cabling feature code (FC 7942) and one of three types of Top Exit Line Cords (FC 8948, FC 8949, FC 8951) also must be ordered. All cables must exit from the top frame of the z14 server, as shown in Figure 10-5.

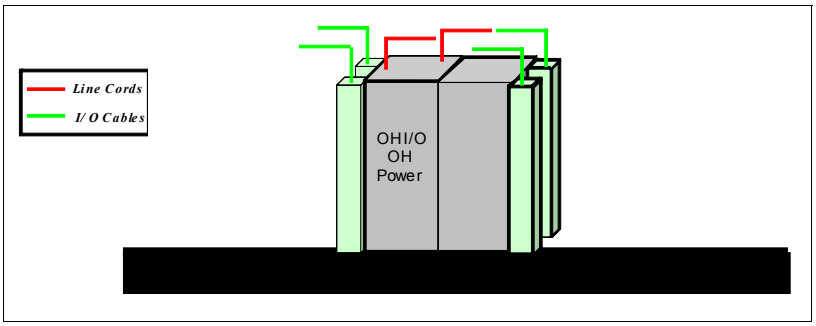


Figure 10-5 Non-raised floor options

10.3.2 Top Exit Power option

Top Exit Power option is supported with or without Top Exit I/O cabling option. You can route your power cables from the top of the frame, once one of following features is chosen:

- Top Exit Line Cord for DC (FC 8948)
- 3-phase, Low Voltage Top Exit Line Cord (FC 8949)
- 3-phase, High Voltage Top Exit Line Cord (FC 8951)

Two types of power cords are offered in this feature:

- Cut cords

  These cords are 4.3 m (14 ft) long from the exit point of the frame, with an attached mount bracket that you can use to fix power cords at the top of frame, as shown in Figure 10-5.

- Plugged cords

  For z14 servers, the fittings are 60A and require much more force to plug in successfully. For the 60A plugs, the “power cord” is a short connection from the power enclosure to the top of the frame. The plug is rigidly fixed to the frame. The client drop must come down to the frame to meet the system input plug.

Figure 10-6 shows the difference between cut cords and plugged cords.

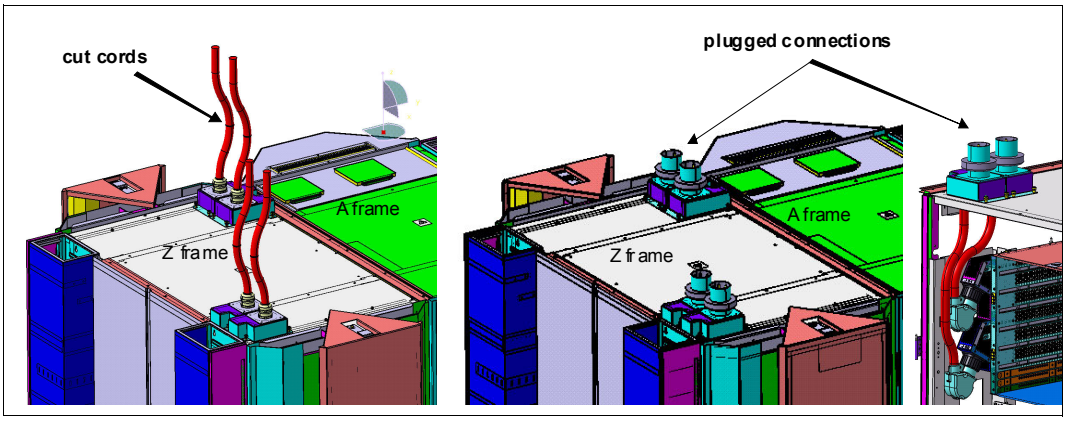


Figure 10-6 Top Exit Power feature

10.3.3 Top Exit I/O Cabling feature

Like the z13, z14 servers support the Top Exit I/O Cabling feature (FC 7942). This feature routes all coupling links and all I/O cables, including 1000BASE-T Ethernet cable from I/O drawers or PCIe I/O drawers, through four more frame extensions out the top of the frame.

Figure 10-7 shows the frame extensions (*chimneys*) installed on frame A or on frame Z when the Top Exit I/O Cabling feature (FC 7942) is ordered.

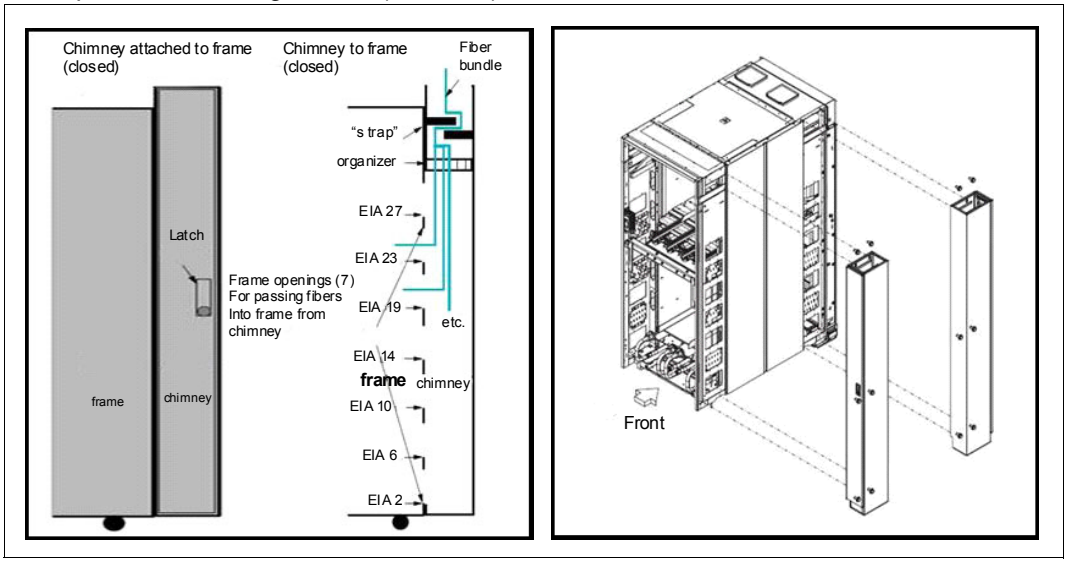


Figure 10-7 Top Exit I/O Cabling feature

The Top Exit I/O Cabling feature adds 15 cm (6 in.) to the width of each frame and about 95 lbs (43 kg) to the weight.

For z13 servers, the Top Exit I/O Cabling feature (FC 7942) is available for both radiator-cooled models and water-cooled models.

10.3.4 Weight distribution plate

The *weight distribution plate* is designed to distribute the weight of a frame onto two floor panels in a raised-floor installation. As listed in Table 10-6 on page 379, the weight of a frame can be substantial. A concentrated load on a caster or leveling foot can be half of the total frame weight. In a multiple system installation, one floor panel can have two casters from two adjacent systems on it, potentially inducing a highly concentrated load on a single floor panel. The weight distribution plate distributes the weight over two floor panels. The weight distribution kit is ordered and delivered by using FC 9970.

Always consult the floor tile manufacturer to determine the load rating of the tile and pedestal structure. Additional panel support might be required to improve the structural integrity because cable cutouts reduce the floor tile rating.

10.3.5 Bolt-down kit for raised floor

A bolt-down kit for raised floor environments can be ordered for the z14 servers. The kit provides hardware to enhance the ruggedness of the frames and to tie down the frames to a concrete floor beneath a raised floor of 15 cm - 91 cm (5.9 in. - 35.8 in.). The kit is offered in the following configurations:

- The Bolt-Down kit for an air-cooled system (FC 8003) is designed for air-cooled server.It provides frame stabilization and bolt-down hardware for securing the frames to a concrete floor beneath the raised floor.

- The Bolt-Down kit for a water-cooled system (FC 8004) is designed for water-cooled server. It provides frame stabilization and bolt-down hardware for securing the frames to a concrete floor beneath the raised floor.

Those kits help secure the frames and their contents from damage when exposed to shocks and vibrations, such as those generated by a seismic event. The frame tie-downs are intended for securing a frame that weighs up to 1632 kg (3600 lbs).

Seismic (Earthquake) resistance

In earthquake areas, the z14 server equipped with appropriate earthquake equipment are certified to meet requirements ICC ES AC156 with the test parameters in Table 10-7. Tests were run for two options of S_{ds} parameter (S_{ds} = 2.5g and S_{ds} = 2.0g).

Table 10-7

Test criteria	S_{DS} (g)	z/h	Horizontal		Vertical	
			A_{FLX-H}	A_{RIG-H}	A_{FLX-V}	A_{RIG-V}
ICC-ES AC156	2.5	1.0	4.00	3.00	1.68	0.68
ICC-ES AC156	2.0	1.0	3.20	2.40	1.34	0.54

Notes:

S_{DS}	Design spectral response acceleration at short period
z/h	Quotient of z (height of the structure with respect to attachment point) and h (average building/structure roof height with respect to the base elevation)
A_{FLX-H}	Horizontal spectral acceleration calculated for flexible components
A_{RIG-H}	Horizontal spectral acceleration calculated for rigid components
A_{FLX-V}	Vertical spectral acceleration calculated for flexible components
A_{RIG-V}	Vertical spectral acceleration calculated for rigid components

The above S_{ds} parameter 2.5g represents the high magnitude covering most of densely populated area in California. As an example S_{ds} values for Los Angeles (1.29 g), San Francisco (2.00 g), Santa Barbara (2.00 g), San Diego (1.60 g).

z14 structure consists of a frame of rack, drawers with central processor units, I/O equipment, memory, and other electronic equipment. The primary function of the frame is to protect critical electronic equipment in two modes. The first mode is during shipping shock and vibration, which provides excitation primary in the vertical direction. The second mode is protecting the equipment during seismic events where horizontal vibration can be significant. For more information see *IBM z14 Installation Manual for Physical Planning,* GC28-6965.

10.3.6 Nonraised floor frame tie-down kit

The nonraised floor frame tie-down kit (FC 8005) is designed for air-cooled server. It helps to secure the frames and its contents from damage when exposed to vibrations and shocks, such as those in a seismic event.

10.3.7 Service clearance areas

z14 servers require specific service clearance to ensure the fastest possible repair in the unlikely event that a part needs to 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 z14 Installation Manual for Physical Planning,* GC28-6965.

10.4 Energy management

This section addresses 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 10-8.

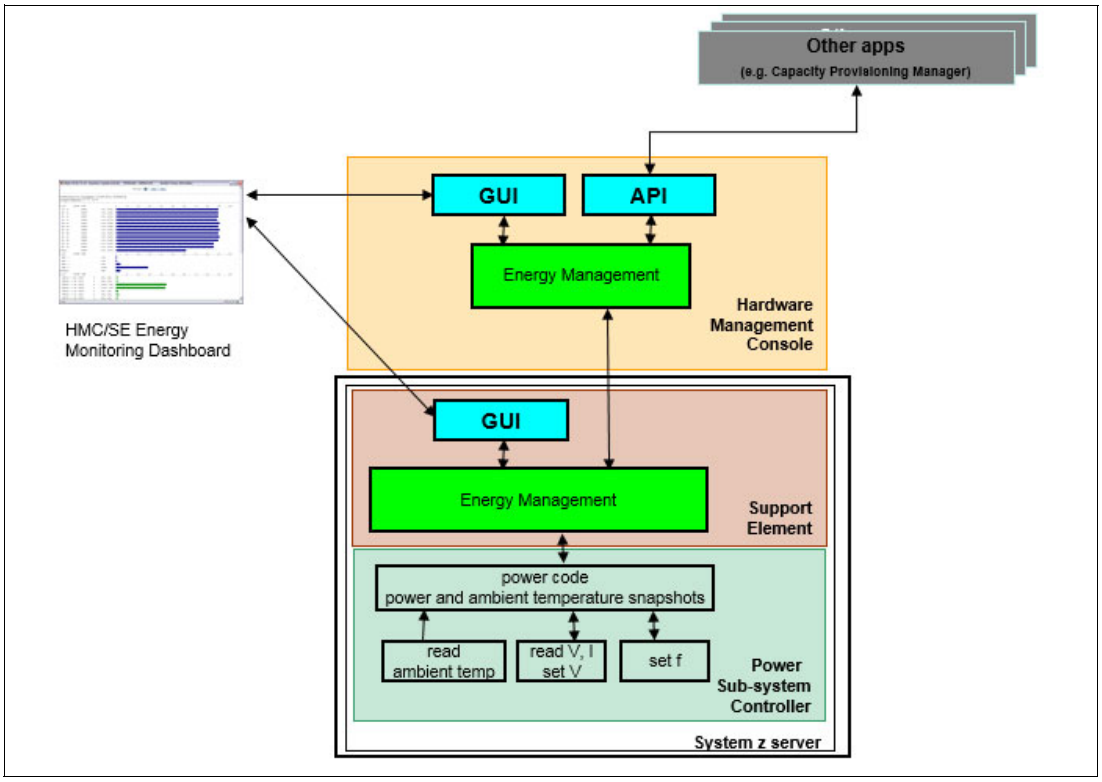


Figure 10-8 z14 energy management

The hardware components in the z14 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.

A few tools are available to plan and monitor the energy consumption of z14 servers:

- The power estimation tool on Resource Link
- The Energy Management task for maximum potential power on HMC and SE
- The Monitors Dashboard and Environmental Efficiency Statistics tasks on HMC and SE

10.4.1 Environmental monitoring

This section introduces energy monitoring HMC and SE tasks.

Monitor task group

The Monitor task group on the HMC and SE includes monitoring-related tasks for z14 servers, as shown in Figure 10-9.

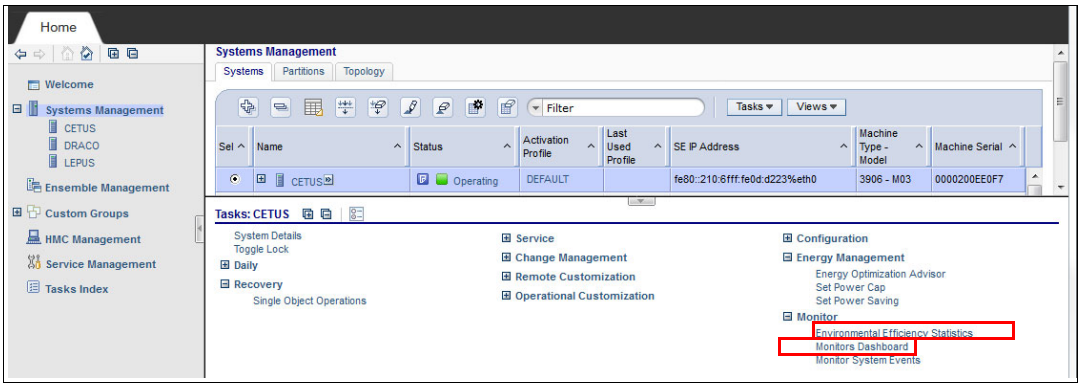


Figure 10-9 HMC Monitor task group

The Monitors Dashboard task

In z14 servers, the Monitors Dashboard task in the Monitor task group provides a tree-based view of resources. Multiple graphical views exist for displaying data, including history charts. 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.

Figure 10-10 shows an example of the Monitors Dashboard task.

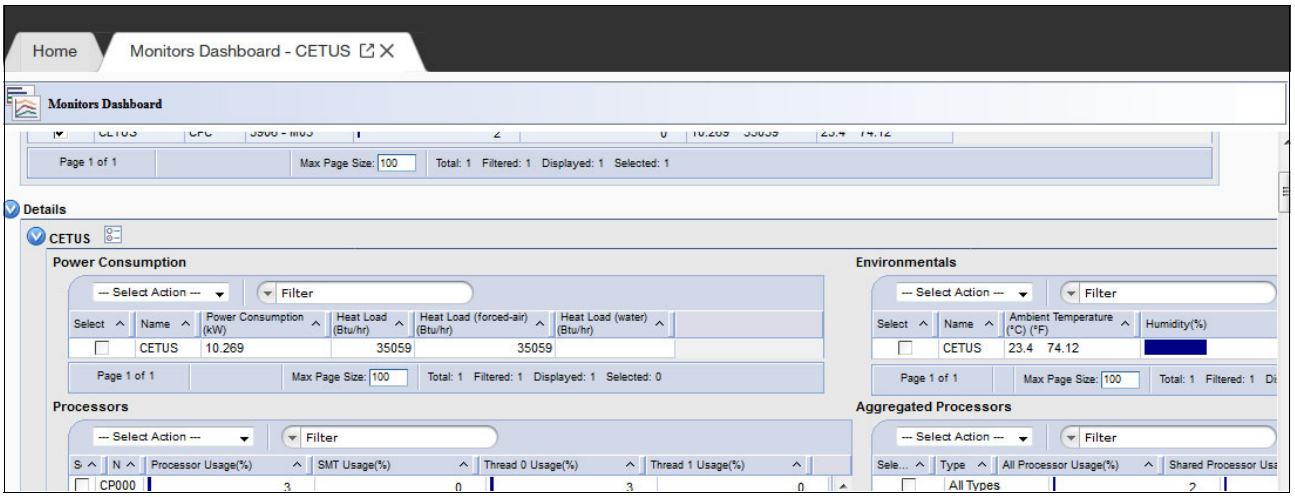


Figure 10-10 Monitors Dashboard task

Environmental Efficiency Statistics task

The Environmental Efficiency Statistics task (Figure 10-11) 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.

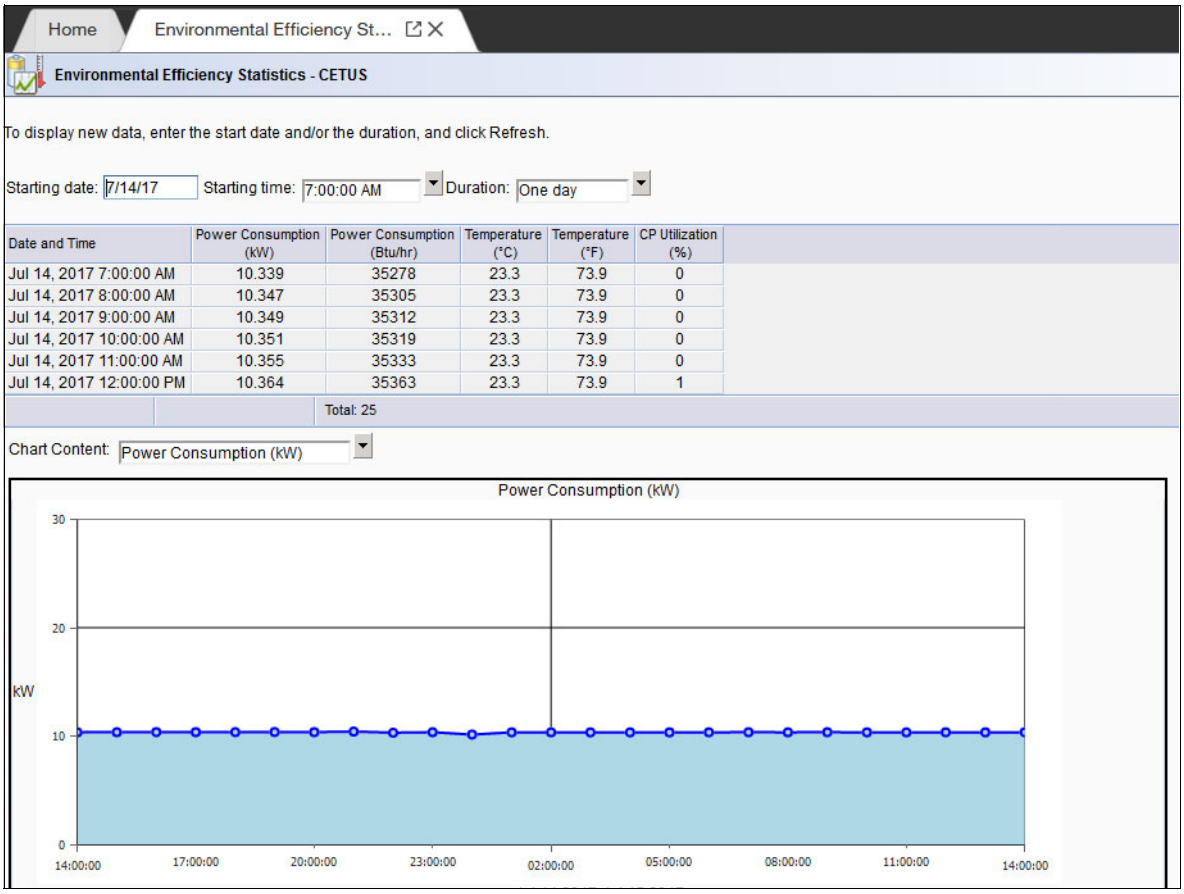


Figure 10-11 Environmental Efficiency Statistics

11

Hardware Management Console and Support Elements

The Hardware Management Console supports many functions and tasks to extend the management capabilities of IBM z14 servers. When tasks are performed on the HMC, the commands are sent to one or more Support Elements (SEs), which then issue commands to their central processor complexes (CPCs).

This chapter includes the following sections:

- Introduction to the HMC and SE
- HMC and SE changes and new features
- HMC and SE connectivity
- Remote Support Facility
- HMC and SE key capabilities

© Copyright IBM Corp. 2017. All rights reserved.

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 server. 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 A frame that are supplied together with the z14 server. One is the primary SE and the other is the alternate SE. The primary SE is the active one. The alternate SE acts as the backup. Like the HMCs, the SEs are closed systems, and no other applications can be installed on them.

When tasks are performed at the HMC, the commands are routed to the active SE of the CPC. 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 409.

With Driver 27 (Version 2.13.0), the IBM Dynamic Partition Manager (DPM) was introduced for Linux only CPCs with Fibre Channel Protocol (FCP) attached storage. 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:

http://www.ibm.com/support/knowledgecenter

After you get to the IBM Knowledge Center, click on 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 406.

11.2 HMC and SE changes and new features

The initial release that comes with z14 servers is HMC application Version 2.14.0. Use the “What’s New” task to explore the new features that are available for each release. For a complete list of HMC and SE functions, use the HMC and SE (Version 2.14.0) console help system or go to the IBM Knowledge Center at the following website:

http://www.ibm.com/support/knowledgecenter

After you get to the IBM Knowledge Center, click **IBM Z,** and then click **z14**.

11.2.1 Driver Level 32 HMC and SE changes and new features

The z14 HMC and SE with Driver Level 32 has several 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 11.2.2, “Enhanced Computing and z14 HMC” on page 392
- Starting with version 2.13.1 HMC Tasks no longer have 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 alternative method with remote browsing on Support Element is still available.
- Starting with z14, all FTP operations originating from the SE will be proxied through a managing HMC. This now allows the FTP SE originated operations to follow our security recommendation. In addition, all HMC/SE tasks that support FTP will provide three options of FTP: FTP, FTPS and SFTP. For more details see 11.3.1, “Network planning for the HMC and SE” on page 401.
- Secure console-to-console communication was established for z14 HMC consoles putting new security standards even on internal communication. For more details see 11.3.1, “Network planning for the HMC and SE” on page 401.
- There are functional enhancements in SNMP/BCPii API interfaces such as queries to Virtual Flash Memory or queries to Secure Service Container. Security of BCPii interface was also enhanced. You can disable s BCPii sending or receiving capability for each partition. Cross-partition authority setting on SE remains the same.
- Remote Browser IP Address Limiting function was implemented due to security reasons. It allows to specify valid remote browser IP or valid mask for group of IP addresses. Global setting to enable/disable remote access is still available.
- Multi-factor authentication was implemented for z14 HMC/SE/TKE. The feature enables you to login with higher security level using two factors. First factor is traditional login and password and second factor is a passcode sent on your smartphone. For more information see 11.3.6, “HMC Multi-factor authentication” on page 406.
- The HMC Global OSA/SF will now provide a global view of all OSA PCHIDs and the monitoring and diagnostic information previously available in the Query Host command For more information see 11.3.4, “OSA Support Facility changes” on page 404.
- Compliance mode for CCA PCI-HSM and EP11 and other certificates is now displayed on Support Element. Setup and administration tasks are done on TKE. For more information see 11.5.12, “Cryptographic support” on page 424.
- Enhancements and changes have been made for the Server Time Protocol (STP), which are described in 11.5.7, “Server Time Protocol support” on page 418.
- A new Mobile application interface will be 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 408
- 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 2.14.0*
  - *IBM Z Hardware Management Console Operations Guide for Ensembles Version 2.14.0*
  - *IBM Z Support Element Operations Guide Version 2.14.0*

Alternatively, see the HMC and SE (Version 2.14.0) console help system or go to the IBM Knowledge Center at the following website:

http://www.ibm.com/support/knowledgecenter

After you get to the IBM Knowledge Center, click **IBM Z**, and then click **z14**.

- 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 2.14.0*
  - *IBM Z Hardware Management Console Operations Guide for Ensemble Version 2.14.0*
  - *IBM Z Support Element Operations Guide Version 2.14.0*

  Alternatively, see the HMC and SE console help system or go to the IBM Knowledge Center at the following website:

  http://www.ibm.com/support/knowledgecenter

  After you get to the IBM Knowledge Center, click **IBM Z**, and then click **z14**.

11.2.2 Enhanced Computing and z14 HMC

Enhanced Computing is a set of security features that have been implemented with z14 HMC Console to improve its resistance to security attacks. z14 HMC consoles, TKE devices, and Support Elements have firmware compliant with NIST Computer Security Standard 800-147, which translates to the following:

- Current IBM Z Firmware is protected during delivery using Digital Signatures.
- BIOS Secure boot function is used by SE/HMC/TKE.
- Signature/hash verification of SE/HMC IBM Z Firmware
- Non-stop monitoring and checking the integrity of files
- Code measurements is stored in TPM (Trusted Platform Module) on SE/HMC
- Provides security logs for internal analysis
- Trusted third Party Validation (IBM Resource Link via zRSF data)
- Analyzes periodic call home measurement data
- Initiates challenge/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)

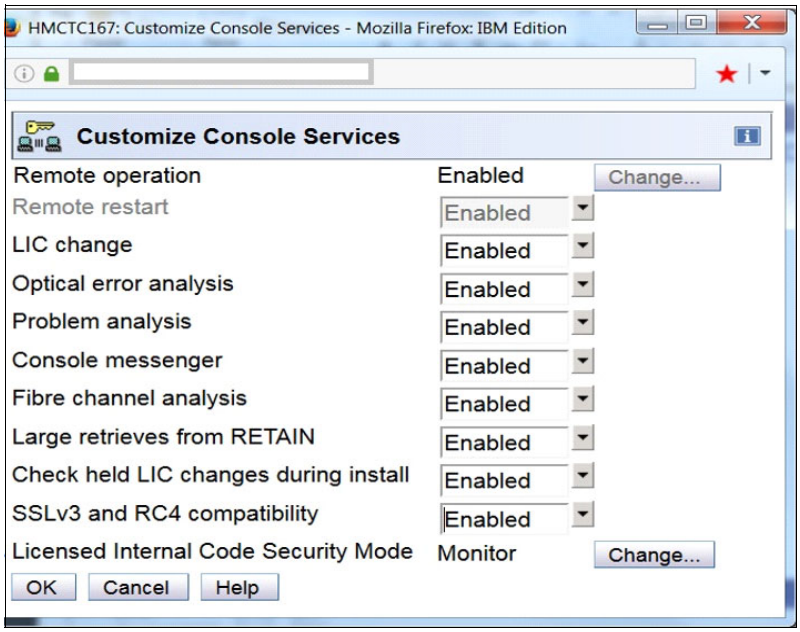


Figure 11-1 Enhanced computing

Firmware tamper detection

z14 will also offer an enhancement on the Support Element that provides notification if tampering (see Figure 11-1) with booting of firmware on the server (CPC) is detected. This enhancement is designed to meet 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 Support Element will issue a customer alert via a warning or a lock of the Support Element, depending on the configuration.

11.2.3 Rack-mounted HMC

Feature code FC #0083 provides a rack-mounted HMC.

The HMC is a 1U IBM server and comes with an IBM 1U standard tray that contains a monitor and a keyboard. The system unit and tray must be mounted in the rack in two adjacent 1U locations in the “ergonomic zone” between 21U and 26U in a standard 19” rack.

The customer must provide the rack. Three C13 power receptacles are required: Two for the system unit and one for the display and keyboard, as shown in Figure 11-2.

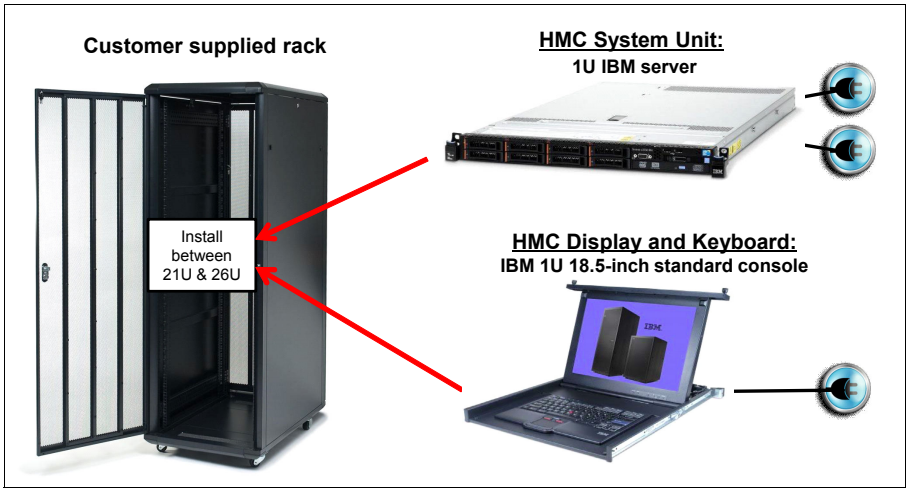


Figure 11-2 Rack-mounted HMC

11.2.4 New Support Elements

The SEs are no longer two notebook computers in one z14 server. They are now two servers that are installed at the top of the A frame. They are managed by the keyboards, pointing devices, and displays that are mounted in the front and rear of the tray of the Z frame (where the SE notebooks were in previous IBM Z servers), as shown in Figure 11-3. The SEs have internal USB attached smart card readers to support Flash Express and Feature on Demand (FoD).

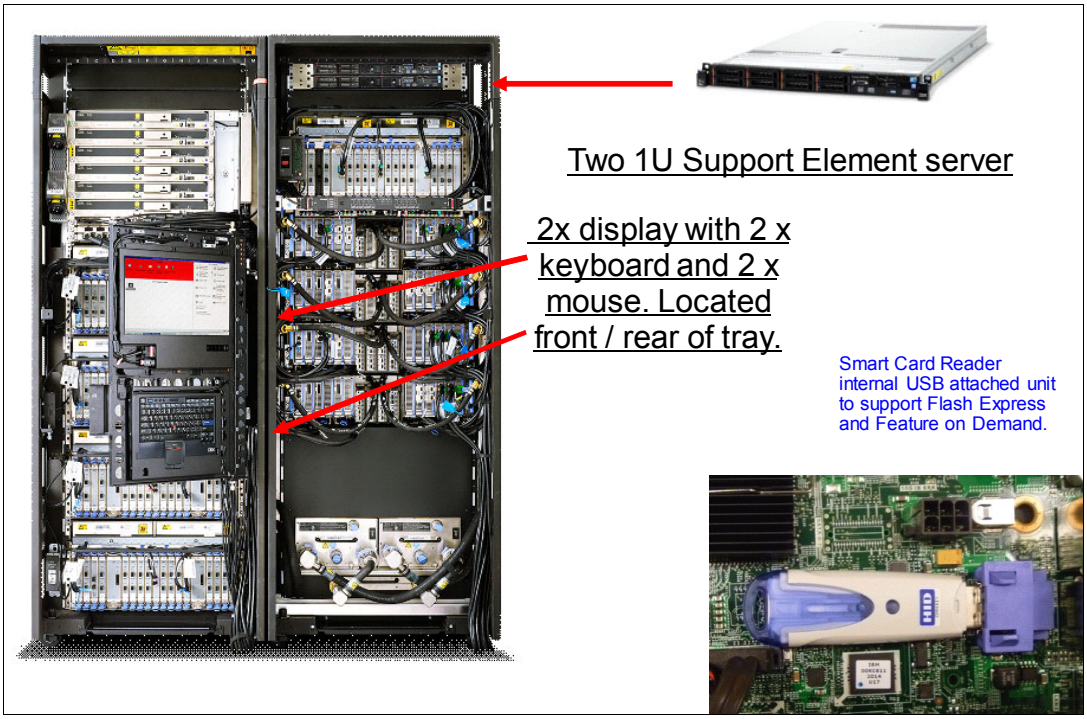


Figure 11-3 SEs location

11.2.5 New backup options for HMCs and primary SEs

This section provides a short description of the new backup options that are available for HMC Version 2.14.0.

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 you have three FTP options: FTP, FTPS or SFTP. For more information see 11.3.1, “Network planning for the HMC and SE” on page 401.

Note: If you do a backup to an FTP server for a z14 server, ensure that you have set up a connection to the FTP server by using the Configure Backup Setting task. If you have not set up a connection to the FTP server, a message appears that prompts you to configure it.

The FTP server must be supplied by the customer. You can enable a secure FTP connection to your server.

Figure 11-4 shows the information that is required to configure your backup FTP server.

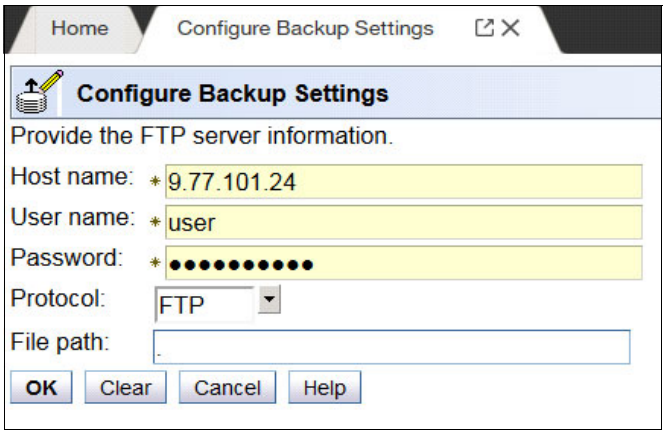


Figure 11-4 Configure Backup Settings

Note: Backup FTP site is a static setting for an HMC. If an alternate FTP site is needed to perform a backup, then this process should be done from another HMC.

Backup of HMCs

A backup of the HMC can be performed to the following media

- A USB flash memory drive (UFD)
- An FTP server
- A UFD and FTP server

Figure 11-5 shows the destination options of the Backup Critical Console Data task.

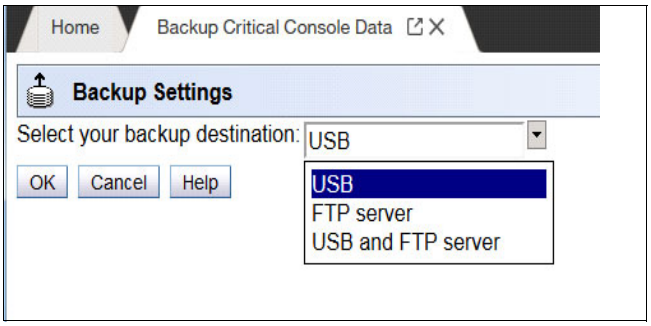


Figure 11-5 Backup Critical Console Data destinations

Optional 32 GB UFD FC #0848

With z13 server the 32 GB UFD is available for backups. An 8 GB UFD is included by default with the system. SE and HMC backup files have been getting larger with the later IBM Z servers.

Backup of primary SEs

The backup for the primary SE of a z14 can be made to the following media:

- The primary SE HDD and alternate SE HDD
- The primary SE HDD and alternate SE HDD and FTP server

It is no longer possible to do the primary SE backup to an UFD of a z14. Table 11-1 shows the SE Backup options for external media.

Table 11-1 SE Backup options

System Type	UFD Media	FTP Server
z14	No	Yes
z13 / z13s	No	Yes
zBX 004	No	Yes
zEC12 / zBC12	Yes	No
z196 / z114	Yes	No
z10EC / z10BC	Yes	No
z9EC / /z9BC	Yes	No

Figure 11-5 on page 396 shows examples of the different destination options of the SE Backup Critical Data for different CPC machine types.

For more information, see the HMC and SE console help system or go to the IBM Knowledge Center at the following link:

`http://www.ibm.com/support/knowledgecenter`

After you get to the IBM Knowledge Center, click **IBM Z**, and then click **z13 or z14**.

Scheduled operations for the backup of HMCs and SEs

The Scheduled Operation task with the new backup options for HMC is changed, as shown in Figure 11-6.

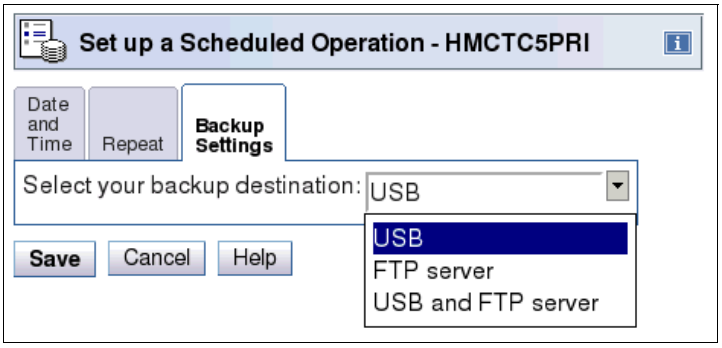


Figure 11-6 Scheduled Operation for HMC backup

The Scheduled Operation task with the new backup options for the SEs is changed, as shown in Figure 11-7.

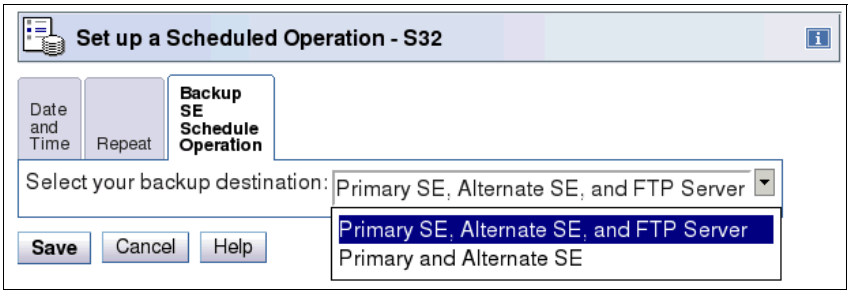


Figure 11-7 Scheduled Operation for SEs backup

11.2.6 SE driver support with the HMC driver

The driver of the HMC and SE is equivalent to a specific HMC and SE version, as shown in these examples:

- Driver 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

An HMC with Version 2.14.0 can support different IBM Z types. Some functions that are available on Version 2.14.0 and later are supported only when the HMC is connected to a z13 server with Driver 27.

Table 11-2 shows a summary of the SE drivers and versions that are supported by the new z14 HMC Version 2.14.0 (Driver 32) and earlier versions/

Table 11-2 Summary of SE drivers

IBM Z family name	Machine type	SE driver	SE version	Ensemble node potential
z14	3906	32	2.14.0	Yes
z13s	2965	27	2.13.1	Yes[a]
z13	2964	22, 27	2.13.0, 2.13.1	Yes[a]
zBX Node	2458 Model 004	22	2.13.0	Required
zBC12	2828	15	2.12.1	Yes
zEC12	2827	15	2.12.1	Yes
z114	2818	93	2.11.1	Yes
z196	2817	93	2.11.1	Yes
z10 BC	2098	79	2.10.2	No
z10 EC	2097	79	2.10.2	No
z9 BC	2096	67	2.9.2	No
z9 EC	2094	67	2.9.2	No

a. A CPC in DPM mode cannot be a member of a ensemble, but the CPC can still be managed by the ensemble HMC.

Note: 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 have these older systems, consider managing these systems by using separate HMCs running older drivers.

11.2.7 HMC feature codes

HMCs older than FC #0091 are not supported for z14 servers at Driver 32.

FC #0082 M/T 2461-TW2

FC #0082 a tower model HMC working with z14, z13 and z13s servers.

FC #0083 M/T 2461-SE1

FC #0083 is the new rack-mounted HMC working with z14 13 and z13s servers.

Previous HMCs that can be carried forward (The carry forward HMCs do not provide Enhanced feature):

- Tower FC #0092
- Tower FC #0095
- 1U Rack FC #0094
- 1U Rack FC #0096

11.2.8 User Interface

Starting with HMC Version 2.14.0 only *Tree Style User Interface* (default) is available. The *Classic Style User Interface* has been discontinued.

11.2.9 Customize Product Engineering Access - Best practice

There may be times when the Hardware Management Console or the Support Element needs to be accessed in a support role to perform problem determination.

The figure below is the task to set the authorization for IBM Product Engineering access to the console. 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.

The task (Figure 11-8) is only available to users with ACSADMIN authority:

- Customers should ensure they have redundant admin users for each console
- Customers should document contact information and/or 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/SE access is needed
- By default the options are disabled

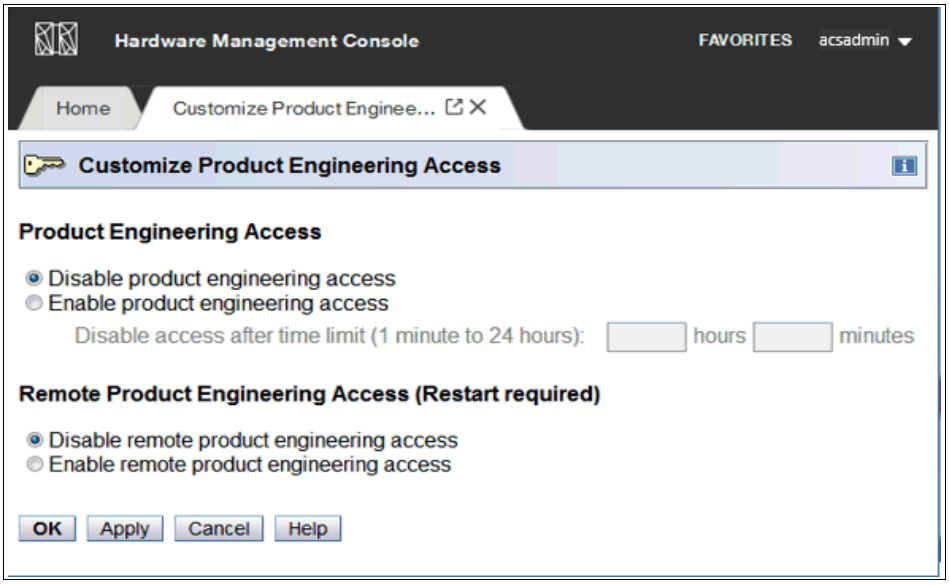


Figure 11-8 Customize Product Engineering access

11.3 HMC and SE connectivity

The HMC has two Ethernet adapters, which are supported by HMC Driver 22 or later for connectivity to up to two different Ethernet LANs.

The SEs on z14 servers are connected to the System Control Hubs (SCH) to control the internal network. In previous IBM Z servers, 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 only possible through an Ethernet switch connected to the J03 or J04 port on the SEs. Other IBM Z servers and HMCs also can be connected to the switch. To provide redundancy, install two Ethernet switches.

Only the switch (and not the HMC directly) can be connected to the SEs.

Figure 11-9 shows the connectivity between HMCs and the SEs.

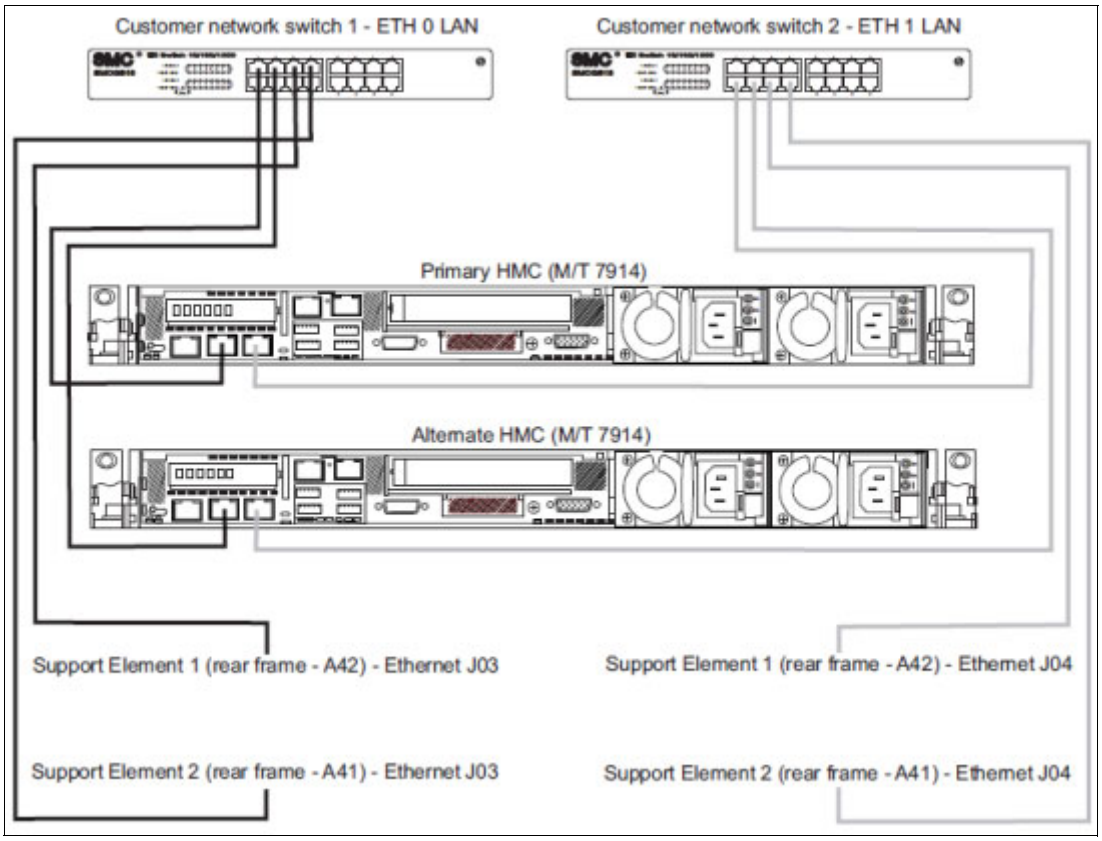


Figure 11-9 HMC and SE connectivity

The LAN ports for the SEs installed in the CPC are shown in Figure 11-10.

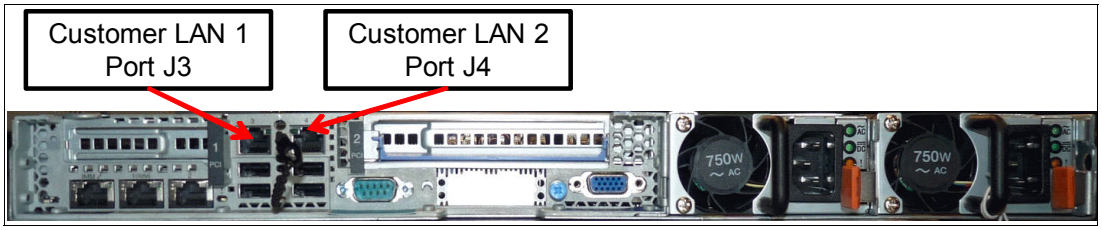


Figure 11-10 SE Physical connection

Various methods are available for setting up the network. It is your responsibility to plan and design the HMC and SE connectivity. Select the method based on your connectivity and security requirements.

Security: Configuration of network components, such as routers or firewall rules, is beyond the scope of this book. Whenever the networks are interconnected, security exposures can exist. For more information about HMC security, see *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 (v.2.14.0) console help system or go to the IBM Knowledge Center at the following link:

`http://www.ibm.com/support/knowledgecenter`

After you get to the IBM Knowledge Center, click **IBM Z**, and then click **z14**.

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. These functions, which depend on the HMC connectivity, are available to the HMC:

- Lightweight Directory Access Protocol (LDAP) support, which can be used for HMC user authentication
- Network Time Protocol (NTP) client/server support
- RSF through broadband
- HMC access through a remote web browser
- Enablement of the SNMP and CIM APIs to support automation or management applications, such as IBM System Director Active Energy Manager (AEM)

These examples are shown in Figure 11-11.

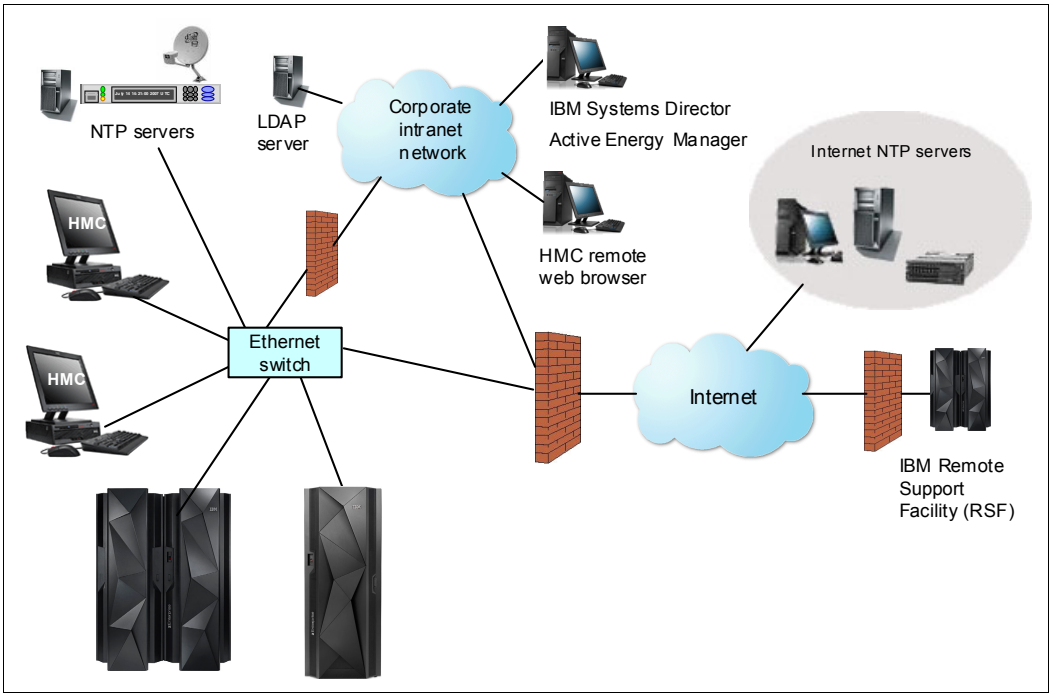


Figure 11-11 HMC connectivity

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 Socket Layer cryptographic protocol (SSL) and needs certificates to authenticate servers. SFTP is based on Secure Shell cryptographic protocol (SSH) and needs 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. There is a number of tasks that previously used only removable media and with z14 HMC console they can use FTP connections. Figure 11-12 shows the recommended network topology for HMC, SE and FTP sever.

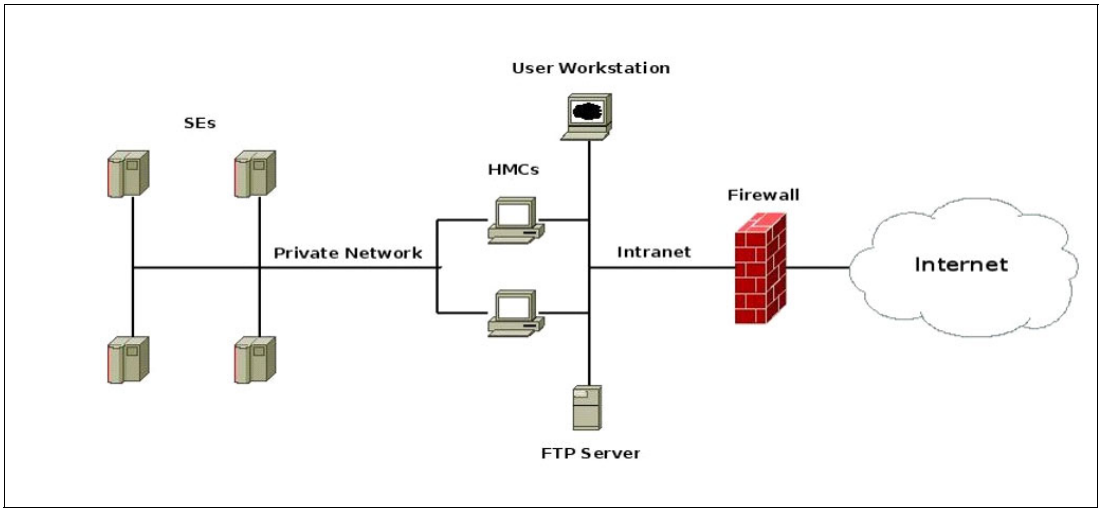


Figure 11-12 Recommended Network Topology for HMC, SE and FTP server

FTP server requirements:

- Must support ‘Passive’ data connections
- A server configuration allowing the client to connect on an ephemeral port

FTPS server requirements:

- Must operate in ‘Explicit’ mode
- Allows a server to offer both secure and unsecured connections
- Must support ‘Passive’ data connections
- Must support secure data connections

SFTP server requirements

- Must support password based authentication

Figure 11-13 on page 403 shows the FTP server choices for HMC.

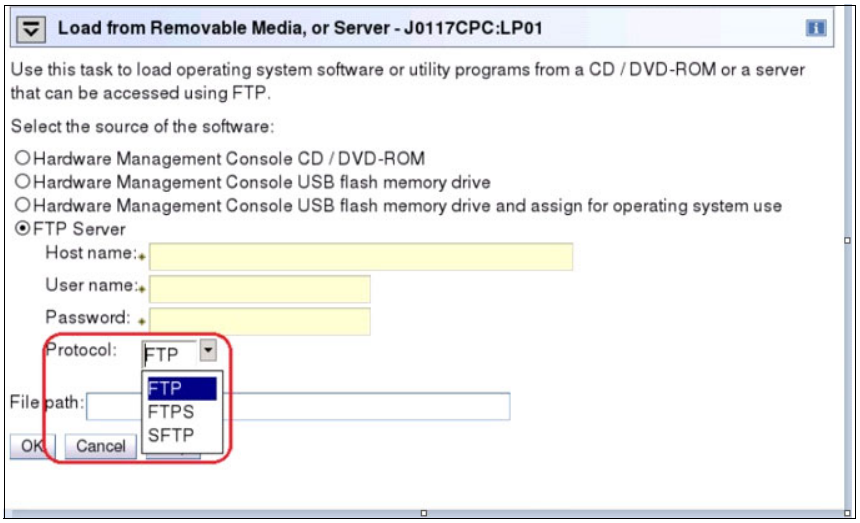


Figure 11-13 New FTP protocols dropdown list

FTP through HMC

It is highly recommended to keep IBM Z servers, HMC consoles, and Support Elements on an isolated network. This approach prevents Support Elements making FTP connections with outside networks. This is valid for all FTP protocols supported: FTP, FTPS, and SFTP. With z14 HMC, all FTP connections originating from Support Elements will be taken to HMC consoles. Secure FTP server credentials will need to be imported to one or more managing HMC consoles. Once HMC console will finish FTP operations, the HMC console performs the FTP operation on the SE's behalf and returns the results. The IBM Z platform will need to be managed by at least one HMC to allow FTP operations to work.

Secure Console-to-Console Communication

Before z14 appeared 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, and this fact potentially can be source of security breach.

Therefore z14 HMC consoles abandon anonymous cipher suite and implement an industry standard based password driven cryptography system. The existing Domain Security Settings will be used to provide authentication as well as high quality encryption. Because of these changes, we now recommend that customers use unique Domain Security settings to provide maximum security. The new system will provide 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 shops. If communication with older systems is needed, the previous password limits must be followed (6 to 8 characters, only uppercase and number characters allowed).

For more information HMC networks, see the following resources:

- The HMC and SE (Version 2.14.0) console help system, or go to the IBM Knowledge Center at the following link:

  `http://www.ibm.com/support/knowledgecenter`

  After you get to the IBM Knowledge Center, click **IBM Z**, and then click **z14**.
- *IBM z14 Installation Manual for Physical Planning*, GC28-6965

11.3.2 Hardware prerequisite changes

The following HMC changes are important for z14 servers:

- IBM does not provide Ethernet switches with the system.
- RSF is broadband-only.

Ethernet switches

Ethernet switches for HMC and SE connectivity are customer provided. Existing supported switches can still be used.

Ethernet switches/hubs typically have these characteristics:

- Sixteen auto-negotiation ports
- 100/1000 Mbps data rate
- Full or half duplex operation
- Auto medium dependent interface crossover (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 HMC. Broadband is needed for hardware problem reporting and service. For more information, see 11.4, “Remote Support Facility” on page 406.

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 a SE is unnecessary if the SE communicates only with HMCs on the same subnet. The HMC and SE can use IPv6 link-local addresses to communicate with each other.

IPv6 link-local addresses have the following characteristics:

- Every IPv6 network interface is assigned a link-local IP address.
- A link-local address is used only on a single link (subnet) and is never routed.
- Two IPv6-capable hosts on a subnet can communicate by using link-local addresses, without having any other IP addresses assigned.

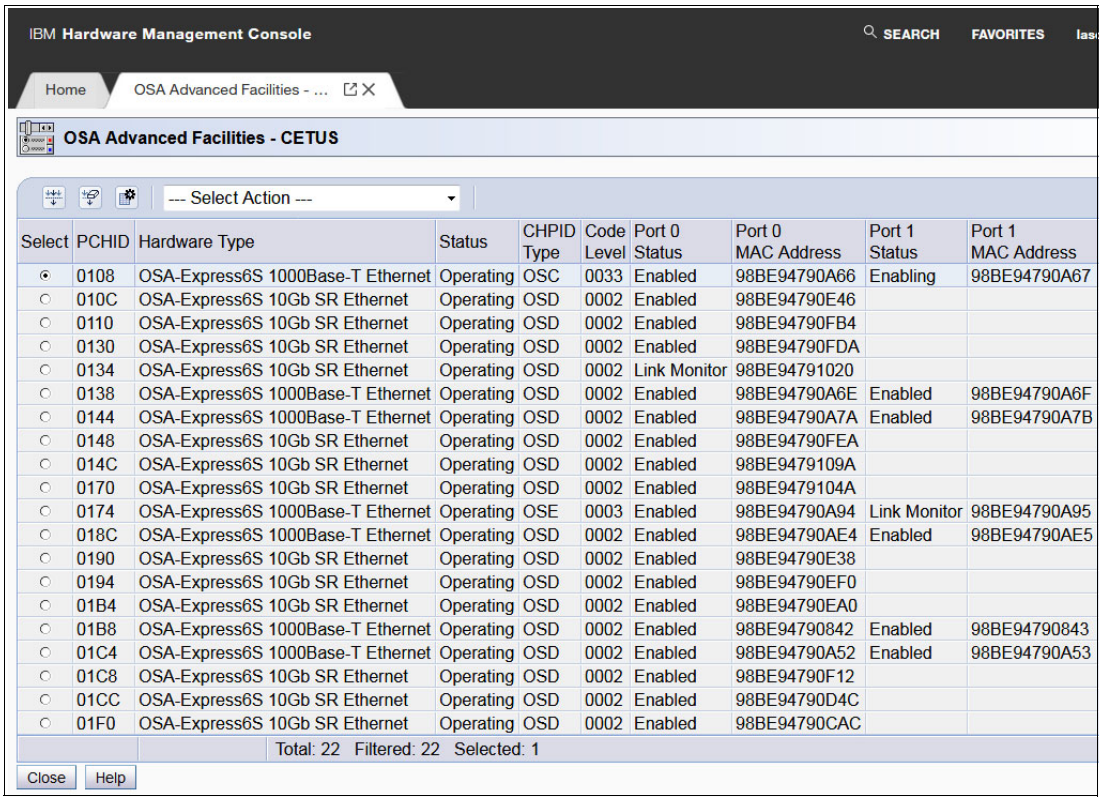
11.3.4 OSA Support Facility changes

Since OSA/SF was moved from z/OS to HMC/SE environment, it was noted that there is no longer an easy way to obtain a global view of all OSA PCHIDs and the monitoring and diagnostic information previously available in the Query Host command.

To address this issue following changes were made:

- If a CPC is targeted, the initial panel will provide a global view of all OSA PCHIDs.
- The user will be able to navigate to various OSA Advanced Facilities subtasks from the initial panel. So there will be a less cumbersome way of getting to them.
- 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 will export the data for all OSA PCHIDs.

- Initial panel is changed to display status information of all OSA PCHID (Figure 11-14)



Select	PCHID	Hardware Type	Status	CHPID Type	Code Level	Port 0 Status	Port 0 MAC Address	Port 1 Status	Port 1 MAC Address
◉	0108	OSA-Express6S 1000Base-T Ethernet	Operating	OSC	0033	Enabled	98BE94790A66	Enabling	98BE94790A67
○	010C	OSA-Express6S 10Gb SR Ethernet	Operating	OSD	0002	Enabled	98BE94790E46		
○	0110	OSA-Express6S 10Gb SR Ethernet	Operating	OSD	0002	Enabled	98BE94790FB4		
○	0130	OSA-Express6S 10Gb SR Ethernet	Operating	OSD	0002	Enabled	98BE94790FDA		
○	0134	OSA-Express6S 10Gb SR Ethernet	Operating	OSD	0002	Link Monitor	98BE94791020		
○	0138	OSA-Express6S 1000Base-T Ethernet	Operating	OSD	0002	Enabled	98BE94790A6E	Enabled	98BE94790A6F
○	0144	OSA-Express6S 1000Base-T Ethernet	Operating	OSD	0002	Enabled	98BE94790A7A	Enabled	98BE94790A7B
○	0148	OSA-Express6S 10Gb SR Ethernet	Operating	OSD	0002	Enabled	98BE94790FEA		
○	014C	OSA-Express6S 10Gb SR Ethernet	Operating	OSD	0002	Enabled	98BE9479109A		
○	0170	OSA-Express6S 10Gb SR Ethernet	Operating	OSD	0002	Enabled	98BE9479104A		
○	0174	OSA-Express6S 1000Base-T Ethernet	Operating	OSE	0003	Enabled	98BE94790A94	Link Monitor	98BE94790A95
○	018C	OSA-Express6S 1000Base-T Ethernet	Operating	OSD	0002	Enabled	98BE94790AE4	Enabled	98BE94790AE5
○	0190	OSA-Express6S 10Gb SR Ethernet	Operating	OSD	0002	Enabled	98BE94790E38		
○	0194	OSA-Express6S 10Gb SR Ethernet	Operating	OSD	0002	Enabled	98BE94790EF0		
○	01B4	OSA-Express6S 10Gb SR Ethernet	Operating	OSD	0002	Enabled	98BE94790EA0		
○	01B8	OSA-Express6S 1000Base-T Ethernet	Operating	OSD	0002	Enabled	98BE94790842	Enabled	98BE94790843
○	01C4	OSA-Express6S 1000Base-T Ethernet	Operating	OSD	0002	Enabled	98BE94790A52	Enabled	98BE94790A53
○	01C8	OSA-Express6S 10Gb SR Ethernet	Operating	OSD	0002	Enabled	98BE94790F12		
○	01CC	OSA-Express6S 10Gb SR Ethernet	Operating	OSD	0002	Enabled	98BE94790D4C		
○	01F0	OSA-Express6S 10Gb SR Ethernet	Operating	OSD	0002	Enabled	98BE94790CAC		

Figure 11-14 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 using the MAC address as part of the address to ensure uniqueness.

An SE can have the following IP addresses:

- Statically assigned IPv4 or statically assigned IPv6
- Auto-configured IPv6 as link-local or router-advertised

IP addresses on the SE cannot be dynamically assigned through DHCP to ensure repeatable address assignments. Privacy extensions are not used.

The HMC uses IPv4 and IPv6 multicasting[1] 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 has 16 bytes. It is written as eight 16-bit hexadecimal blocks that are separated by colons, as shown in the following example:

```
2001:0db8:0000:0000:0202:b3ff:fe1e:8329
```

Because many IPv6 addresses are not fully qualified, shorthand notation can be used. In shorthand notation, the leading zeros can be omitted, and a series of consecutive zeros can be replaced with a double colon. The address in the previous example also can be written in the following manner:

```
2001:db8::202:b3ff:fe1e:8329
```

For remote operations that use a web browser, if an IPv6 address is assigned to the HMC, navigate to it by specifying that address. The address must be surrounded with square brackets in the browser's address field:

```
https://[fdab:1b89:fc07:1:201:6cff:fe72:ba7c]
```

Using link-local addresses must be supported by your browser.

11.3.6 HMC Multi-factor authentication

Multi-factor authentication is an optional feature, configured on per-user, per-template basis. It enhances security of the z14 systems by requiring not only what you know, which is first factor, but also what you have, which means that only person owning a specific phone number can log in. Multi-factor authentication first factor is login and password, while second factor is time-based one-time password sent to your smartphone. This password is defined in RFC 6238 standard and uses a cryptographic hash function combining secret key with current time to generate user-friendly one-time password. The secret key is generated by HMC/SE/TKE while user is doing first factor logon. The secret key is only known to HMC/SE/TKE and its user - smartphone. For that reason it needs to 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 has been generated. Algorithm within HMC responsible for MFA code generation is changing the code every 30 seconds. However to make things less tough HMC/SE console accepts current, previous and next MFA codes. It is also important to have HMC/SE/TKE and smartphone clocks synced. If not, MFA logon attempt will not succeed. Another important fact here is that time zone differences are irrelevant as MFA code algorithm uses UTC.

11.4 Remote Support Facility

The HMC RSF provides important communication to a centralized IBM support network for hardware problem reporting and service. The following types of communication are provided:

- Problem reporting and repair data
- Microcode Change Level (MCL) delivery
- Hardware inventory data, which is also known as vital product data (VPD)
- On-demand enablement

Consideration: RSF through a modem *is not supported* on the z14 HMC. Broadband connectivity is needed for hardware problem reporting and service.

[1] For a customer-supplied switch, multicast must be enabled at the switch level.

11.4.1 Security characteristics

The following security characteristics are in effect:

- RSF requests always are initiated from the HMC to IBM. An inbound connection is never initiated from the IBM Service Support System.
- All data that is transferred between the HMC and the IBM Service Support System is encrypted with high-grade Secure Sockets Layer (SSL)/Transport Layer Security (TLS) encryption.
- When starting the SSL/TLS-encrypted connection, the HMC validates the trusted host with the digital signature that is issued for the IBM Service Support System.
- Data that is sent to the IBM Service Support System consists of hardware problems and configuration data.

Additional resource: 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*, SC28-6927.

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 will work.

At the time of writing, RSF still uses the “traditional” RETAIN connection. You must add access to the new Enhanced IBM Service Support System to your current RSF infrastructure (proxy, firewall, and so on).

To have the best availability and redundancy and to be prepared for the future, the HMC must have access to the Internet to IBM through RSF in the following manner. Transmission to the enhanced IBM Support System requires a domain name server (DNS). The DNS must be configured on the HMC if you are not using a proxy for RSF. If you are using a proxy for RSF, the proxy must provide the DNS.

The following host names and IP addresses are used and your network infrastructure must allow the HMC to have access to the following host names:

- Host names:
  - www-945.ibm.com on port 443
  - esupport.ibm.com on port 443

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 existing IP addresses are no longer supported.

11.4.3 HMC and SE remote operations

You can use these methods to perform remote manual operations on the HMC:

- Using a remote HMC

  A remote HMC is a physical HMC that is on a different subnet from the SE. This configuration prevents the SE from being automatically discovered with IP multicast. A remote HMC requires TCP/IP connectivity to each SE to be managed. Therefore, any existing customer-installed firewalls between the remote HMC and its managed objects must permit communications between the HMC and the SE. For service and support, the remote HMC also requires connectivity to IBM, or to another HMC with connectivity to IBM through RSF. For more information, see 11.4, “Remote Support Facility” on page 406.

- Using a web browser to connect to an HMC

  The z14 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.

  - 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**... from the **Customize Console Services** window. as shown in Figure 11-15.

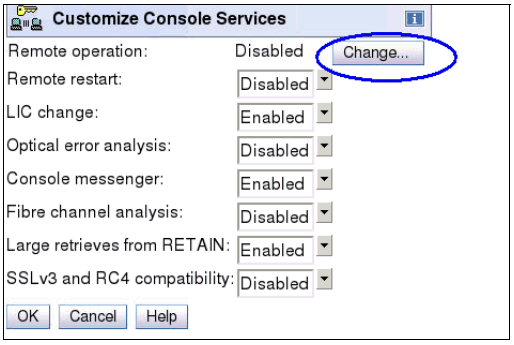


Figure 11-15 Customize HMC remote operation

Microsoft Internet Explorer, Mozilla Firefox, and Goggle Chrome were tested as remote browsers. For detailed web browser requirements, see the HMC and SE console help system or go to the IBM Knowledge Center at the following link:

http://www.ibm.com/support/knowledgecenter

After you get to the IBM Knowledge Center, click **IBM Z**, and then click **z14**.

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 one on the SE. For more information, see the HMC and SE console help system or go to the IBM Knowledge Center at the following link:

http://www.ibm.com/support/knowledgecenter

After you get to the IBM Knowledge Center, click **IBM Z**, and then click **z14**.

HMC Mobile Interface

The new mobile application interface will allow HMC users to securely monitor and manage systems from anywhere. iOS and Android HMC applications will be available to provide system and partition views, the ability to monitor status and Hardware and Operating System Messages, and the ability to receive mobile push notifications from the HMC using the existing zRSF (IBM Z Remote Support Facility) connection. A full set of granular security controls are provided from HMC console, to user, to monitor only, and finally app password including multi-factor authentication. This Mobile interface is optional and by default disabled.

11.5 HMC and SE key capabilities

The HMC and SE have many capabilities. This section covers the key areas. For a complete list of capabilities, see the HMC and SE (Version 2.14.0) console help system or go to the IBM Knowledge Center at the following link:

http://www.ibm.com/support/knowledgecenter

After you get to the IBM Knowledge Center, click **IBM Z**, and then click **z14**.

11.5.1 Central processor complex management

The HMC is the primary place for CPC control. For example, the input/output configuration data set (IOCDS) contains definitions of LPARs, channel subsystems, control units, and devices, and their accessibility from LPARs. IOCDS can be created and put into production from the HMC.

The HMC is used to start the power-on reset (POR) of the server. During the POR, processor units (PUs) are characterized and placed into their respective pools, memory is put into a single storage pool, and the IOCDS is loaded and initialized into the hardware system area (HSA).

The Hardware messages task displays hardware-related messages at the CPC level, LPAR level, or SE level. It also displays hardware messages that relate to the HMC itself.

11.5.2 Logical partition management

Use the HMC to define LPAR properties, such as the number of processors of each type, how many are reserved, and how much memory is assigned to it. These parameters are defined in LPAR profiles, and are stored on the SE.

Because Processor Resource/Systems Manager (PR/SM) must manage LPAR access to processors and the initial weights of each partition, weights are used to prioritize partition access to processors.

You can use the Load task on the HMC to perform an IPL of an operating system. This task causes a program to be read from a designated device, and starts that program. You can perform the IPL of the operating system from storage, the HMC DVD-RAM drive, the USB flash memory drive (UFD), or a File Transfer Protocol (FTP) server.

When an LPAR is active and an operating system is running in it, you can use the HMC to dynamically change certain LPAR parameters. The HMC provides an interface to change partition weights, add logical processors to partitions, and add memory.

LPAR weights can 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 servers can export the Change LPAR Controls table data to a comma-separated value (.csv)-formatted file. This support is available to a user when connected to the HMC remotely by a web browser.

Partition capping values can be scheduled and are specified on the Change LPAR Controls scheduled operation support. Viewing details about an existing Change LPAR Controls scheduled operation is available on the SE.

Absolute physical HW LPAR capacity setting

Driver 15 introduced the capability to define, in the image profile for shared processors, the absolute processor capacity that the image is allowed to use (independent of the image weight or other cappings).

To indicate that the LPAR can use the undedicated processors absolute capping, select **Absolute capping** on the Image Profile Processor settings to specify an absolute number of processors to cap the LPAR's activity at. The absolute capping value can either be "None" or a value for the number of processors (0.01 - 255.0).

LPAR group absolute capping

This is the next step in partition capping options available on z14 and z13/z13s servers at Driver level 27 and above. Follow on to LPAR absolute capping, LPAR group absolute capping uses a similar methodology to enforce:

- Customer licensing
- Non-z/OS partitions where group soft capping is not an option
- z/OS partitions where ISV does not support software capping

A group name, processor capping value, and partition membership are specified at HW console:

- Set an absolute capacity cap by CPU type on a group of LPARs

- Allows each of the partitions to consume capacity up to their individual limits as long as the group's aggregate consumption does not exceed the group absolute capacity limit
- Includes updated SysEvent QVS support (used by vendors who implement software pricing)
- Only shared partitions managed in these groups
- Can specify caps for one or more processor types in the group
- Specified in absolute processor capacity (for example, 2.5 processors)
- Use Change LPAR Group Controls (same windows that are used for software group defined capacity) as shown in Figure 11-16 (snapshot on a z13 server).

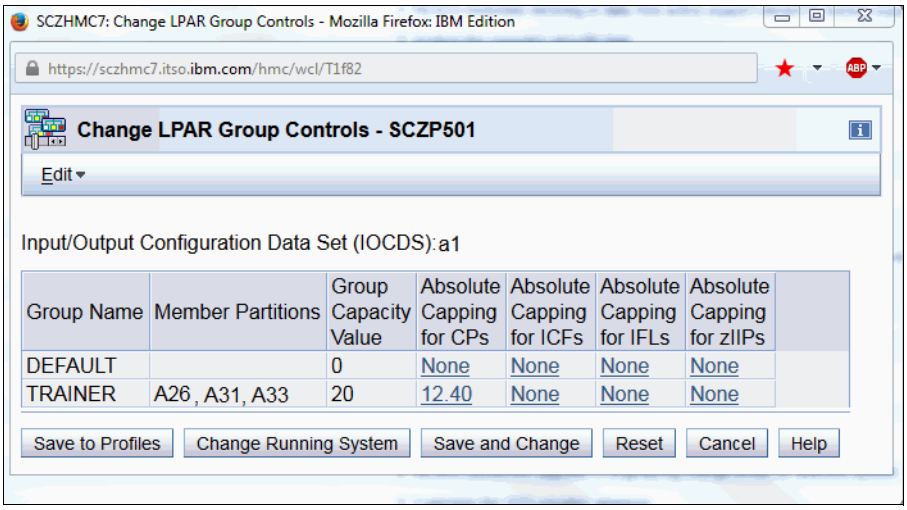


Figure 11-16 Change LPAR Group Controls - Group absolute capping

Absolute capping is specified as an absolute number of processors to cap the group's activity to. The value is specified to hundredths of a processor (for example, 4.56 processors) worth of capacity.

The value is not tied to the LICCC. Any value from 0.01 to 255.00 can be specified. This configuration makes the profiles more portable and means that you will not have issues in the future when profiles are migrated to new machines.

Although the absolute cap can be specified to hundredths of a processor, the exact amount might not be that precise. The same factors that influence the “machine capacity” also influence the precision with which the absolute capping actually works.

11.5.3 Operating system communication

The Operating System Messages task displays messages from an LPAR. You also can enter operating system commands and interact with the system. This task is especially valuable for entering Coupling Facility Control Code (CFCC) commands.

The HMC also provides integrated 3270 and ASCII consoles. These consoles allow an operating system to be accessed without requiring other network or network devices, such as TCP/IP or control units.

Updates to x3270 support

The Configure 3270 Emulators task on the HMC and TKE consoles was enhanced with Driver 15 to verify the authenticity of the certificate that is returned by the 3270 server when a secure and encrypted SSL connection is established to an IBM host. This 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 11-17.

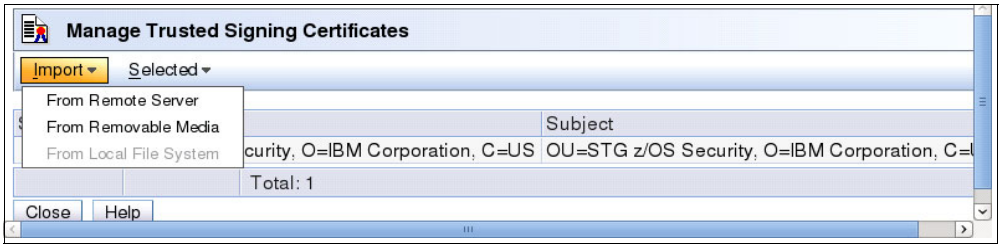


Figure 11-17 Manage Trusted Signing Certificates

If the connection between the console and the IBM host can be trusted at the time of importing the certificate, the import from the remote server option can be used, as shown in Figure 11-18. Otherwise, import the certificate by using removable media.

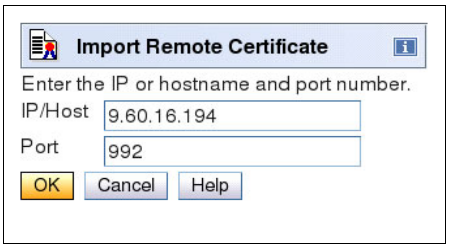


Figure 11-18 Import Remote Certificate example

A secure Telnet connection is established by adding the prefix `L:` to the IP address:port of the IBM host, as shown in Figure 11-19.

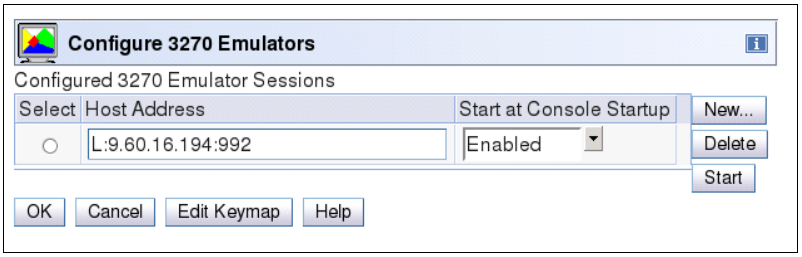


Figure 11-19 Configure 3270 Emulators

11.5.4 HMC and SE microcode

The microcode for the HMC, SE, and CPC is included in the driver/version. The HMC provides the management of the driver upgrade through Enhanced Driver Maintenance

(EDM). EDM also provides the installation of the latest functions and 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 9.9, “z14 Enhanced Driver Maintenance” on page 365.

Microcode Change Level

Regular installation of MCLs is key for reliability, availability, and serviceability (RAS), optimal performance, and new functions:

- Install MCLs on a quarterly basis at a minimum.
- Review hiper MCLs continuously to decide whether to wait for the next scheduled fix application session or to schedule one earlier if the risk assessment warrants.

Tip: The IBM Resource Link[a] provides access to the system information for your IBM Z server according to the system availability data that is sent on a scheduled basis. It provides more information about the MCL status of your z14 servers. To access the Resource Link, to the following website:

`http://www.ibm.com/servers/resourcelink`

After you reach the Resource Link, click **Tools** → **Machine Information**, choose your IBM Z server, and click **EC/MCL**.

a. Registration is required to access the IBM Resource Link.

Microcode terms

The microcode has these characteristics:

- The driver contains engineering change (EC) streams.
- Each EC stream covers the code for a specific component of z14 servers. It has a specific name and an ascending number.
- The EC stream name and a specific number are one MCL.
- MCLs from the same EC stream must be installed in sequence.
- MCLs can have installation dependencies on other MCLs.
- Combined MCLs from one or more EC streams are in one bundle.
- An MCL contains one or more Microcode Fixes (MCFs).

Figure 11-20 shows how the driver, bundle, EC stream, MCL, and MCFs interact with each other.

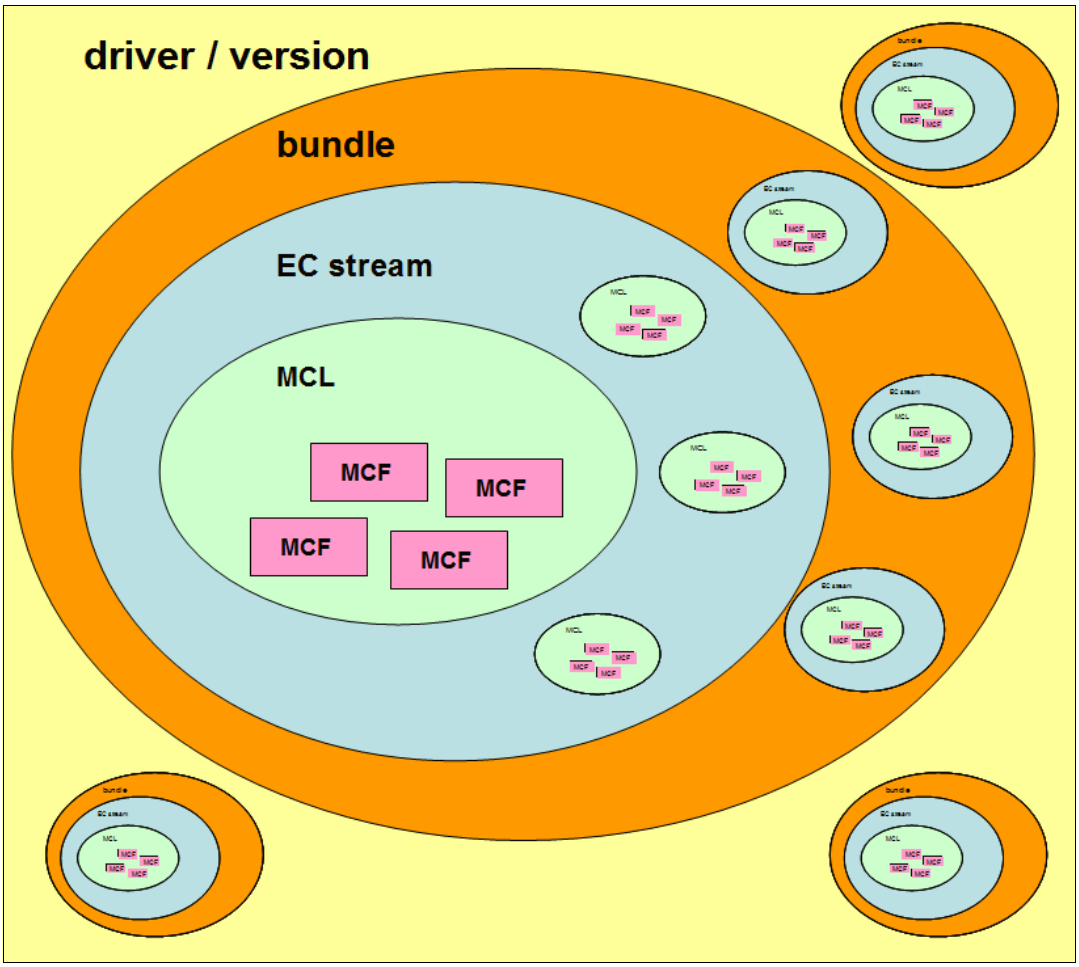


Figure 11-20 Microcode terms and interaction

Microcode installation by MCL bundle target

A *bundle* is a set of MCLs grouped during testing and released as a group on the same date. You can install an MCL to a specific target bundle level. The System Information window is enhanced to show a summary bundle level for the activated level, as shown in Figure 11-21.

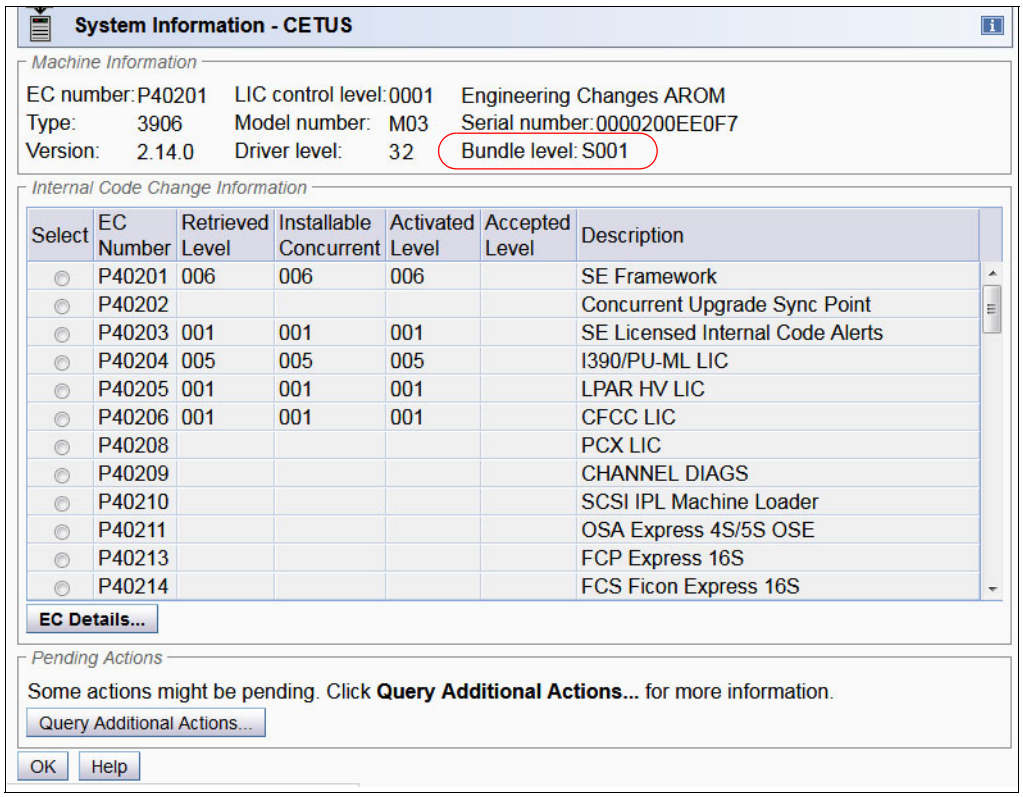


Figure 11-21 System Information: Bundle level

OSC Concurrent Patch

Concurrent patch for OSC channels is now supported

11.5.5 Monitoring

This section addresses monitoring considerations.

Monitor task group

The Monitor task group on the HMC and SE includes monitoring-related tasks for z14 servers, as shown in Figure 11-22.

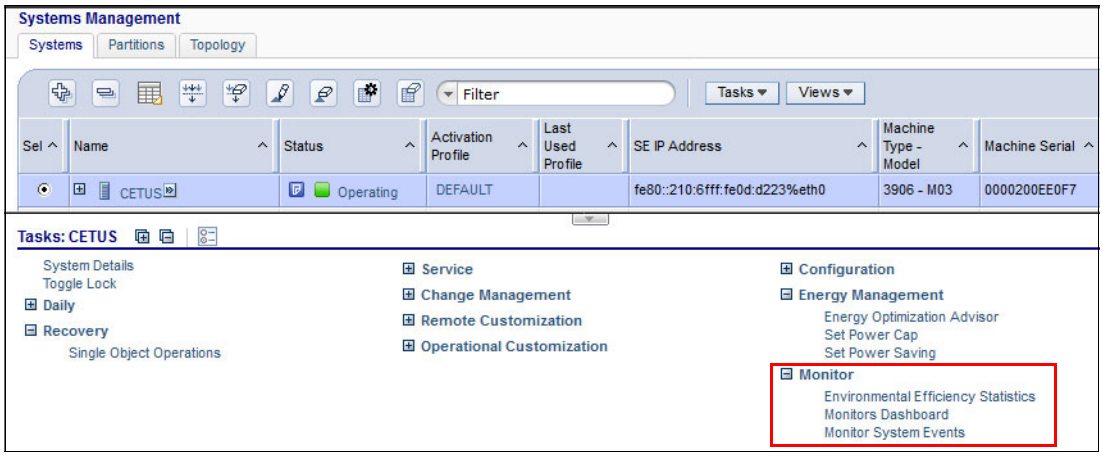


Figure 11-22 HMC Monitor Task Group

The Monitors Dashboard task

The Monitors Dashboard task supersedes the System Activity Display (SAD). In the z14 server, the Monitors Dashboard task in the Monitor task group provides a tree-based view of resources.

Multiple graphical views exist for displaying data, including history charts. The Open Activity task, which is known as SAD (System Activity Display), monitors processor and channel usage. It produces data that includes power monitoring information, power consumption, and the air input temperature for the server.

Figure 11-23 shows an example of the Monitors Dashboard task.

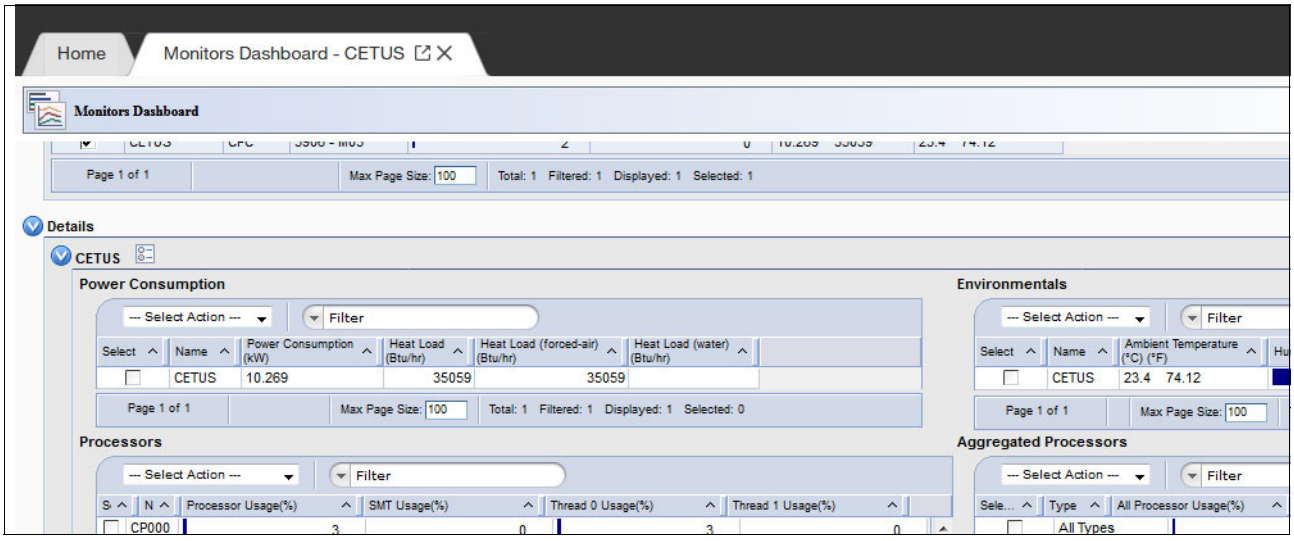


Figure 11-23 Monitors Dashboard task

You can display details for (see Figure 11-24 on page 417):

- Power consumption
- Environmentals

- Aggregated processors
- Processors (with SMT information)
- System Assist Processors
- Logical Partitions
- Channels
- Adapters (Crypto utilization percentage is displayed according to the physical channel ID (PCHID) number.)

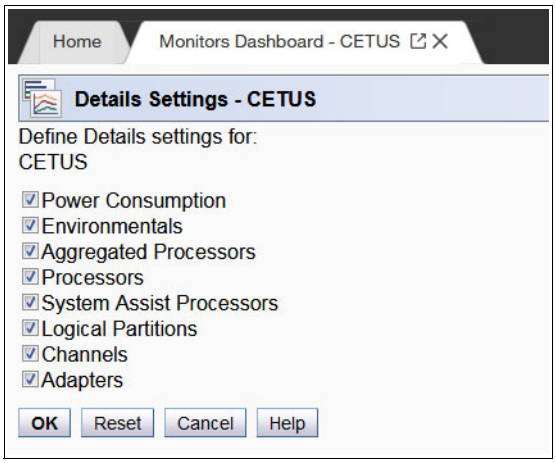


Figure 11-24 Monitors dashboard Detailed settings

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 Licensed Internal Code (LIC) configuration code (LICCC) records to help you manage them. It also shows a history of recorded activities.

The HMC for IBM z14 servers has these CoD capabilities:

- SNMP API support:
  - API interfaces for granular activation and deactivation
  - API interfaces for enhanced CoD query information
  - API event notification for any CoD change activity on the system
  - CoD API interfaces, such as On/Off CoD and Capacity BackUp (CBU)
- SE window features (accessed through HMC 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 the details of installed and staged permanent records
- More details for the Attention state by providing seven more flags

HMC and SE are a part of the z/OS Capacity Provisioning environment. The Capacity Provisioning Manager (CPM) communicates with the HMC through 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 these 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 STP functions, the role of the HMC is extended to provide the user interface for managing the Coordinated Timing Network (CTN):

- z14 servers 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 server. 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 server cannot be in the same CTN with a z196 system or earlier systems and cannot become member of an mixed CTN.
- An STP-only CTN can be managed by using different HMCs. However, the HMC must be at the same driver level (or later) than any SE that is to be managed. Furthermore, all SEs to be managed must be known (defined) to that HMC.

In a STP-only CTN, the HMC can be used to perform the following tasks:

- Initialize or modify the CTN ID.
- Initialize the time, manually or by contacting an NTP server.
- Initialize the time zone offset, Daylight Saving Time offset, and leap second offset.
- Assign the roles of preferred, backup, and current time servers, and arbiter.
- Adjust time by up to plus or minus 60 seconds.
- Schedule changes to the offsets listed. STP can automatically schedule Daylight Saving Time, based on the selected time zone.
- Monitor the status of the CTN.
- Monitor the status of the coupling links that are initialized for STP message exchanges.
- For diagnostic purposes, the PPS port state on a z14 server can be displayed and fenced ports can be reset individually.

STP changes and enhancements

The STP related functions had changes dramatically and gained new, intuitive GUI. Administrator of IBM Z is guided through a system time management workflow reducing need to refer to external documentation. The in-line definition of technical terms eliminates need to look up documentation to find out definitions. Detailed instructions and guidelines are provided within task workflow. New tasks give 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 enable catching more errors.

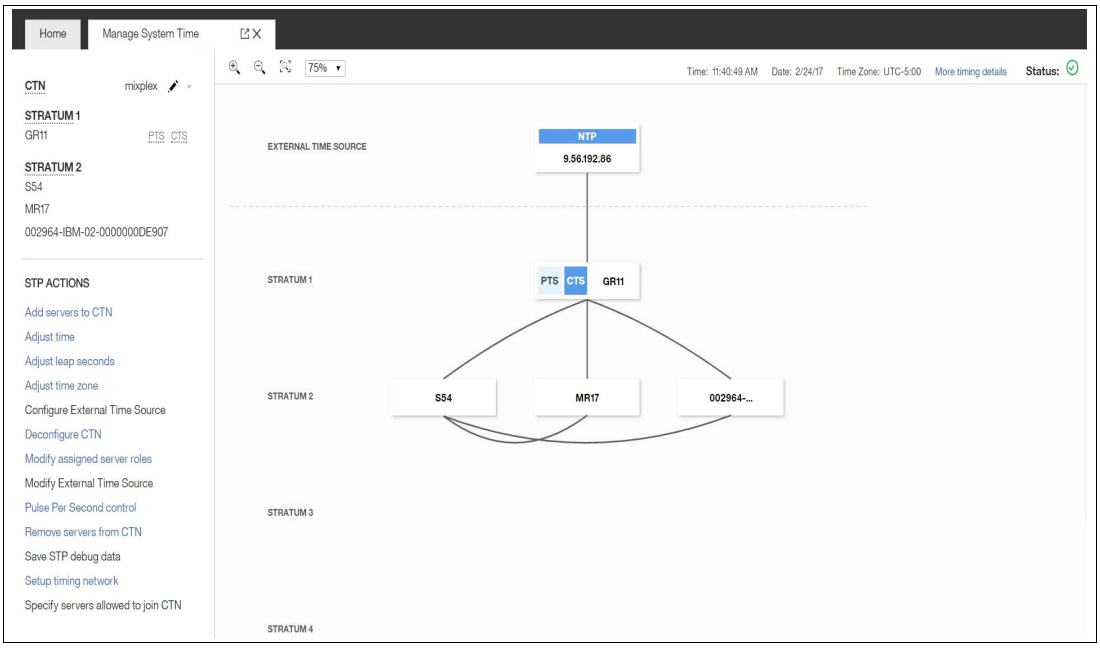


Figure 11-25 CTN topology visible on HMC Manage System Time window

Attention: Figure 11-25 shows a Schedule leap second offset change to 26 seconds that is scheduled for 12/31/2014. This is *not* a real leap second that is released by the International Earth Rotation and Reference System Services. It was temporarily set *only* to show the panel appearance.

Enhanced Console Assisted Recovery

Enhanced Console Assisted Recovery (ECAR) speeds up the process of BTS takeover by using these 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 for these reasons:

- The console path goes from a 2-way path to an one-way path
- There is almost no lag time 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 only available on z14 and z13/z13s servers on Driver 27 and later. In a mixed environment with previous generation machines, you should define a z14, z13 or z13s server as the PTS and CTS.

For more planning and setup information, see the following publications:

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

11.5.8 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 accuracy for the STP-only CTN
- Clients who use a common time reference across heterogeneous systems

The NTP server becomes the single time source, the ETS, for STP and other servers that are not IBM Z servers (such as AIX®, Microsoft Windows, and others) that have NTP clients.

The HMC can act as an NTP server. With this support, the z14 server can get time from the HMC without accessing a LAN other than the HMC/SE network. When the HMC is used as an NTP server, it can be configured to get the NTP source from the Internet. For this type of configuration, a LAN that is separate from the HMC/SE LAN can be used.

HMC NTP broadband authentication support

HMC NTP authentication can be used since HMC Driver 15. The SE NTP support is unchanged. To use this option on the SE, configure the HMC with this option as an NTP server for the SE.

Authentication support with a proxy

Some client configurations use a proxy for external access outside the corporate data center. NTP requests are User Datagram Protocol (UDP) socket packets and cannot pass through the proxy. The proxy must be configured as an NTP server to get to target servers on the web. Authentication can be set up on the client's proxy to communicate with the target time sources.

Authentication support with a firewall

If you use a firewall, HMC NTP requests can pass through it. Use HMC authentication to ensure untampered time stamps.

NTP symmetric key and autokey authentication

With symmetric key and autokey authentication, the highest level of NTP security is available. HMC Level 2.12.0 and later provide windows that accept and generate key information to be configured into the HMC NTP configuration. They can also issue NTP commands.

The HMC offers symmetric key and autokey authentication and NTP commands:

- Symmetric key (NTP V3-V4) authentication

  Symmetric key authentication is described in RFC 1305, which was made available in NTP Version 3. Symmetric key encryption uses the same key for both encryption and decryption. Users exchanging data keep this key to themselves. Messages encrypted with

a secret key can be 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

11.5.9 Security and user ID management

This section addresses security and user ID management considerations.

HMC and SE security audit improvements

With the Audit and Log Management task, audit reports can be generated, viewed, saved, and offloaded. The Customize Scheduled Operations task allows you to schedule audit report generation, saving, and offloading. The Monitor System Events task allows Security Logs to send email notifications by using the same type of filters and rules that are used for both hardware and operating system messages.

With z14 servers, you can offload the following HMC and SE log files for customer audit:

- Console event log
- Console service history
- Tasks performed log
- Security logs
- System log

Full log offload and delta log offload (since the last offload request) are provided. Offloading to removable media and to remote locations by FTP is available. The offloading can be manually started by the new Audit and Log Management task or scheduled by the Customize Scheduled Operations task. The data can be offloaded in the HTML and XML formats.

HMC user ID templates and LDAP user authentication

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 the capability to change the roles for a default user ID, create your own version by copying an existing default user ID.

View-only user IDs and view-only access for HMC and SE

With HMC and SE user ID support, users can be created that have "view-only" access to selected tasks. Support for "view-only" user IDs is available for the following purposes:

- Hardware messages
- Operating system messages
- Customize or delete activation profiles
- Advanced facilities
- Configure on and off

HMC and SE secure FTP support

You can use a secure FTP connection from a HMC/SE FTP client to a customer FTP server location. This configuration is implemented by using the Secure Shell (SSH) File Transfer Protocol, which is an extension of SSH. You can use the Manage SSH Keys console action, which is available to both the HMC and SE, to import public keys that are associated with a host address.

The Secure FTP infrastructure allows HMC and SE applications to query whether a public key is associated with a host address and to use the Secure FTP interface with the appropriate public key for a host. Tasks that use FTP now provide a selection for the secure host connection.

When selected, the task verifies that a public key is associated with the specified host name. If none is provided, a message window is displayed that points to the Manage SSH Keys task to input a public key. The following tasks provide this support:

- Import/Export IOCDS
- Advanced Facilities FTP IBM Content Collector Load
- Audit and Log Management (Scheduled Operations only)
- FCP Configuration Import/Export
- OSA view Port Parameter Export
- OSA Integrated Console Configuration Import/Export

11.5.10 System Input/Output Configuration Analyzer on the SE and HMC

The System Input/Output Configuration Analyzer task supports the system I/O configuration function.

The information that is needed to manage a system's I/O configuration must be obtained from many separate sources. The System Input/Output Configuration Analyzer task enables the system hardware administrator to access, from one location, the information from those sources. Managing I/O configurations then becomes easier, particularly across multiple servers.

The System Input/Output Configuration Analyzer task runs the following functions:

- Analyzes the current active IOCDS on the SE.
- Extracts information about the defined channel, partitions, link addresses, and control units.
- Requests the channels' node ID information. The Fibre Channel connection (FICON) channels support remote node ID information, which is also collected.

The System Input/Output Configuration Analyzer is a view-only tool. It does not offer any options other than viewing. With the tool, data is formatted and displayed in five different views. The tool provides various sort options, and data can be exported to a UFD for later viewing.

The following five views are available:

- PCHID Control Unit View shows PCHIDs, channel subsystems (CSS), CHPIDs, and their control units.
- PCHID Partition View shows PCHIDS, CSS, CHPIDs, and the partitions in which they exist.
- Control Unit View shows the control units, their PCHIDs, and their link addresses in each CSS.
- Link Load View shows the Link address and the PCHIDs that use it.
- Node ID View shows the Node ID data under the PCHIDs.

11.5.11 Automated operations

As an alternative to manual operations, an application can interact with the HMC and SE through an API. The interface allows a program to monitor and control the hardware components of the system in the same way you can. The HMC APIs provide monitoring and control functions through SNMP and the CIM. These APIs can get and set a managed object's attributes, issue commands, receive asynchronous notifications, and generate SNMP traps.

The HMC supports the CIM as an extra systems management API. The focus is on attribute query and operational management functions for IBM Z servers, such as CPCs, images, and activation profiles. z13 servers contain a number of enhancements to the CIM systems management API. The function is similar to that provided by the SNMP API.

For more information about APIs, see *IBM Z Application Programming Interfaces*, SB10-7164.

11.5.12 Cryptographic support

This section lists the cryptographic management and control functions that are available in the HMC and the SE.

Cryptographic hardware

z14 servers include both standard cryptographic hardware and optional cryptographic features for flexibility and growth capability.

The HMC/SE interface provides the following capabilities:

- Defining the cryptographic controls
- Dynamically adding a Crypto feature to a partition for the first time
- Dynamically adding a Crypto feature to a partition that already uses Crypto
- Dynamically removing a Crypto feature from a partition

The Crypto Express6S, 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)

Figure 11-26 shows an example of the Cryptographic Configuration window.

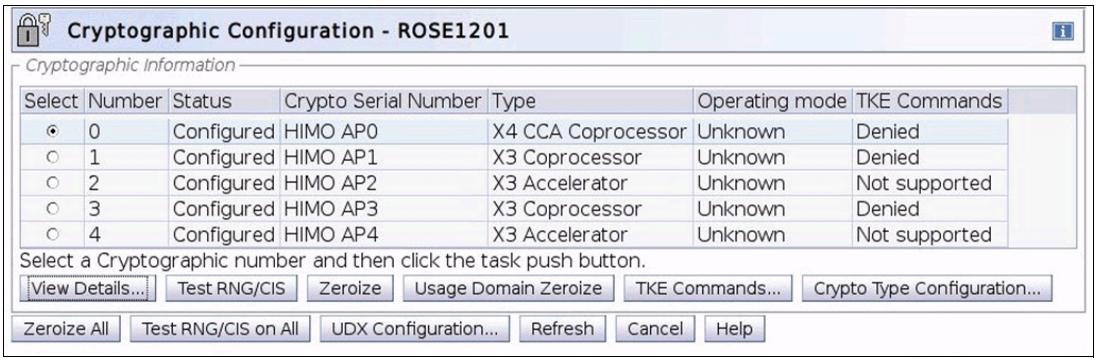


Figure 11-26 Cryptographic Configuration window

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 z13 Configuration Setup*, SG24-8260.

Digitally signed firmware

Critical issues with firmware upgrades are security and data integrity. Procedures are in place to use a process to digitally sign the firmware update files that are sent to the HMC, the SE, and the TKE. Using a hash algorithm, a message digest is generated that is then encrypted with a private key to produce a digital signature.

This operation ensures that any changes that are made to the data are detected during the upgrade process by verifying the digital signature. It helps ensure that no malware can be installed on IBM Z products during firmware updates. It enables the z14 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

CCA compliance levels for Crypto Express6S that are possible to see on Support Element are listed below:

- CCA: Non-compliant (default)
- CCA: PCI-HSM 2016
- CCA: PCI-HSM 2016

EP11 compliance levels (Crypto Express5S and Crypto Express6S) are listed below:

- FIPS 2009 (default
- FIPS 2011
- BSI 2009
- BSI 2011

11.5.13 Installation support for z/VM using the HMC

Starting with z/VM V5R4 and z10, Linux on z Systems can be installed in a z/VM virtual machine from HMC workstation media. This Linux on z Systems installation can use the existing communication path between the HMC and the SE. No external network or additional network setup is necessary for the installation.

11.5.14 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 a function called “Enable Dynamic Partition Manager”, under the SE CPC Configuration menu.

Figure 11-27 shows enabling DPM with the SE interface.

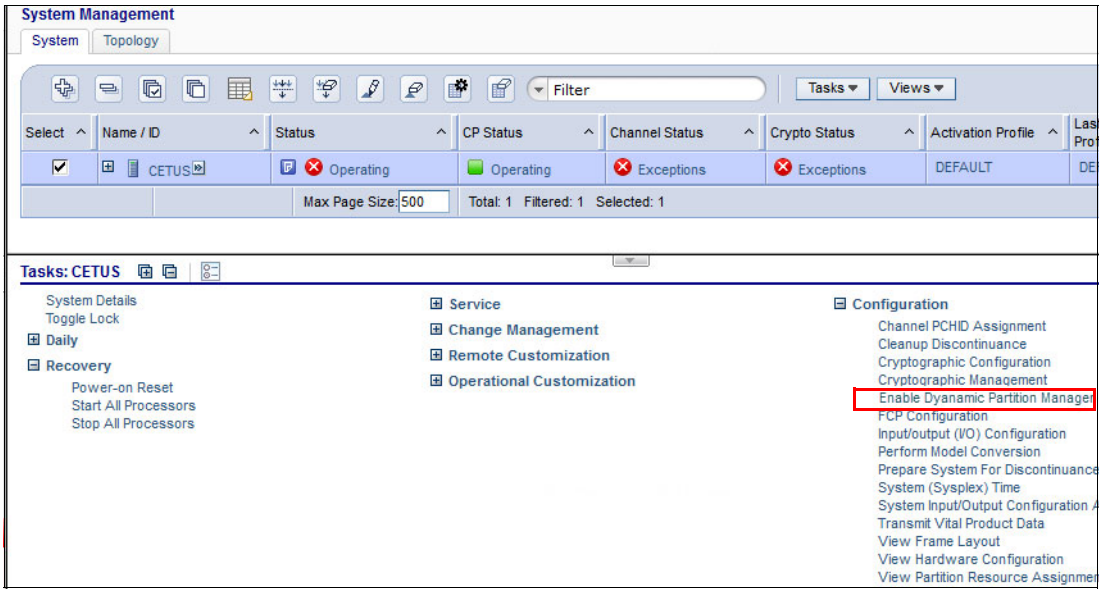


Figure 11-27 Enabling DPM (SE interface)

After the CPC is restarted and you log on to the HMC that has this CPC defined, the HMC shows another welcome window that is shown in Figure 11-28.

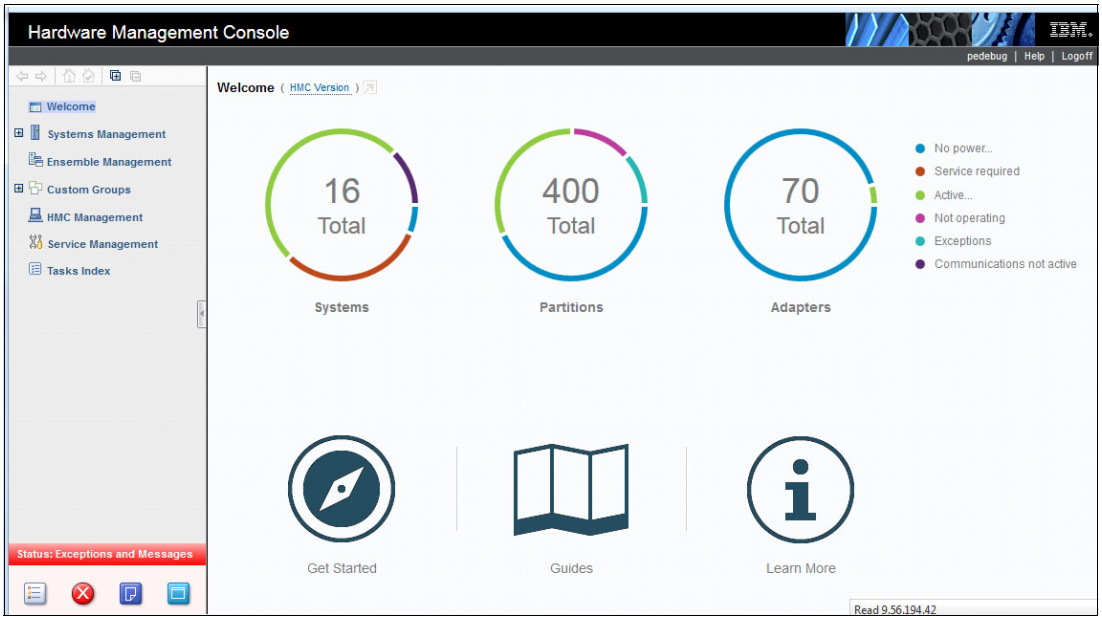


Figure 11-28 HMC welcome window

New LPARs can be added by selecting **Get Started**. For more information see IBM Knowledge Center:

http://www.ibm.com/support/knowledgecenter

After you get to the IBM Knowledge Center, click on the search engine window and enter '**dpm**'.

12

Performance

This chapter describes the performance considerations for IBM z14 servers.

This chapter includes the following sections:

- IBM z14 performance characteristics
- LSPR workload suite
- Fundamental components of workload performance
- Relative Nest Intensity
- LSPR workload categories based on relative nest intensity
- Relating production workloads to LSPR workloads
- Workload performance variation

© Copyright IBM Corp. 2017. All rights reserved.

12.1 IBM z14 performance characteristics

z14 Model M05 is designed to offer up to 35% more capacity and **3.2** times the amount of memory than the z13 Model HA1 system.

Uniprocessor performance has also increased. A z14 Model 701 offers, on average, performance improvements of more than 8% over the z13 Model 701. Figure 12-1 shows the estimated capacity ratios for z13, zEC12, z196, and z10 EC.

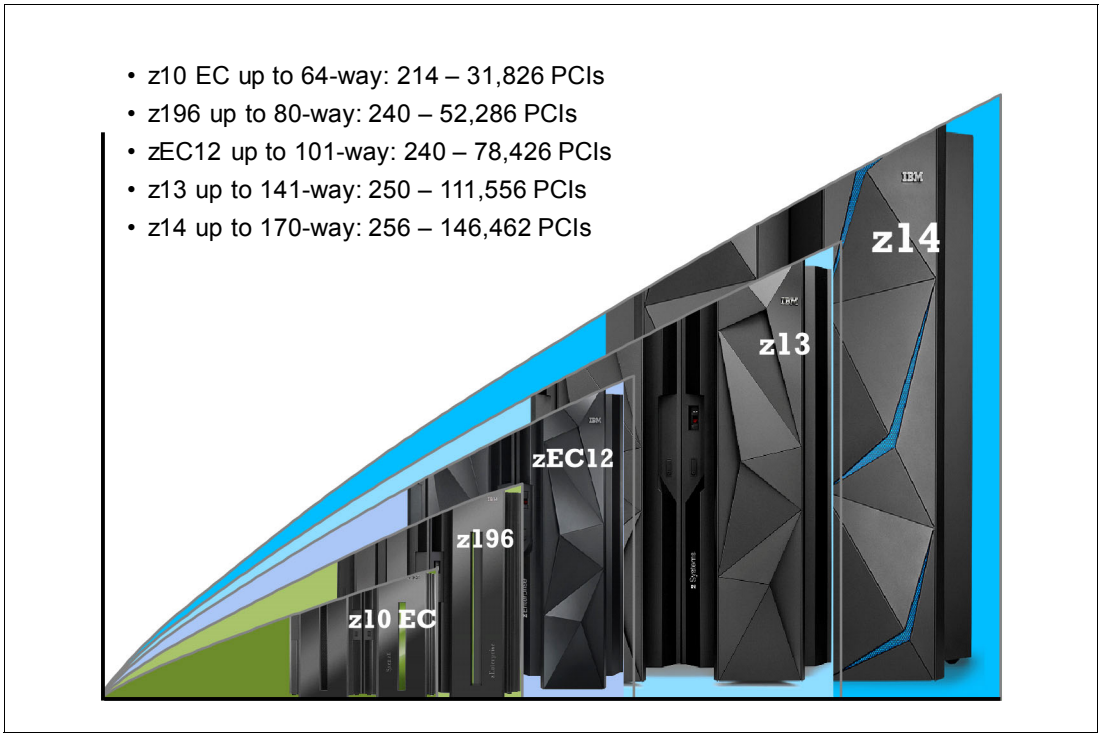


Figure 12-1 IBM Z server generations capacity comparison

The Large System Performance Reference (LSPR) provides capacity ratios among various processor families based on a variety of mesasured workloads. It has become common practice to assign a capacity scaling value to processors as a high-level approximation of their capacities. For z/OS V2R2 studies, the capacity scaling factor commonly associated with the reference processor is set to a 2094-701 with a PCI (Processor Capacity Index) value of 593. This is unchanged since z/OS V1R11 LSPR and using 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.

Performance data for z14 servers were obtained with z/OS V2R2 (running DB2 for z/OS V11, CICS TS V5R3, IMS V14, Enterprise COBOL V6R1 and Websphere Application Server for z/OS V8.5.5.9). All IBM Z server generations are measured in the same environment with the same workloads at high utilizations.

On average, z14 servers can deliver up to 35% more performance in a 170-way configuration than an z13 141-way. However, the observed performance increase varies depending on the workload type.

Consult the LSPR when you consider performance on the z14. The range of performance ratings across the individual LSPR workloads is likely to have 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 436.

For detailed performance information, see the LSPR website:

https://www.ibm.com/servers/resourcelink/lib03060.nsf/pages/lsprindex

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

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

12.2 LSPR workload suite

Historically, LSPR capacity tables, including pure workloads and mixes, have been identified with application names or a *software* characteristic. Examples are 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 based on the underlying hardware sensitivities, in particular, the processor access to caches and memory. This is known as *nest activity intensity*. Using this data, LSPR introduces three workload capacity categories that replace all prior primitives and mixes.

LSPR contains the internal throughput rate ratios (ITRRs) for the z14 and the previous generation processor families. These ratios are based on measurements and projections that use standard IBM benchmarks in a controlled environment. The throughput that any user experiences can vary depending on the amount of multiprogramming in the user’s job stream, the I/O configuration, and the workload processed. Therefore, no assurance can be given that an individual user can achieve throughput improvements equivalent to the performance ratios that are stated.

12.3 Fundamental components of workload performance

Workload performance is sensitive to three major factors:

- Instruction path length
- Instruction complexity
- Memory hierarchy and memory nest

This section examines each of these three factors.

12.3.1 Instruction path length

A transaction or job runs a set of instructions to complete its task. These instructions are composed of various paths through the operating system, subsystems, and application. The total count of instructions that are run across these software components is referred to as the *transaction* or *job path length*. The path length varies for each transaction or job, and depends on the complexity of the tasks that must be run. For a particular transaction or job,

[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

the application path length tends to stay the same, assuming that the transaction or job is asked to run the same task each time.

However, the path length that is associated with the operating system or subsystem can vary based on a number of factors:

- Competition with other tasks in the system for shared resources. As the total number of tasks grows, more instructions are needed to manage the resources.
- The 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*. Many 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]

As workloads are moved between microprocessors with various designs, performance varies. However, when on a processor, this component tends to be similar across all models of that processor.

12.3.3 Memory hierarchy and memory nest

The *memory hierarchy* of a processor generally refers to the caches, data buses, and memory arrays that stage the instructions and data that must be run on the microprocessor to complete a transaction or job.

Many design choices affect this component:

- Cache size
- Latencies (sensitive to distance from the microprocessor)
- Number of levels, the Modified, Exclusive, Shared, Invalid (MESI) protocol, controllers, switches, the number and bandwidth of data buses, and 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 anIBM Z processor refers to the shared caches and memory along with the data buses that interconnect them.

[5] Only available for IFL, zIIP, and SAP processors

Figure 12-2 shows a memory nest in a fully populated z14 CPC drawer.

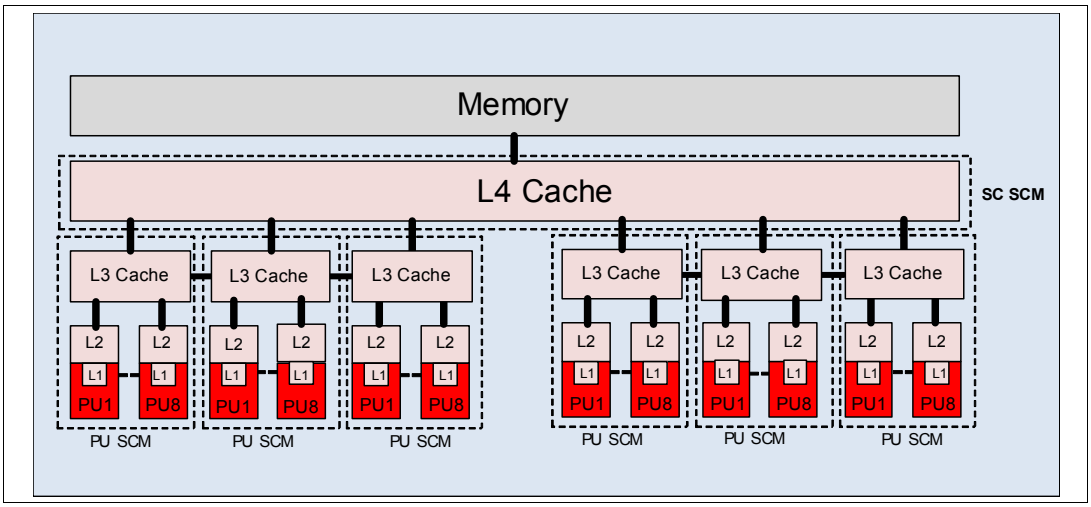


Figure 12-2 Memory hierarchy in a fully populated z14 CPC drawer

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. Additionally, 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, these factors:

- Locality of reference
- I/O rate
- Competition from other applications and LPARs

12.4 Relative Nest Intensity

The most performance-sensitive area of the memory hierarchy is the activity to the memory nest. This is the distribution of activity to the shared caches and memory. The term Relative Nest Intensity (RNI) indicates the level of activity to this part of the memory hierarchy. Using data from CPU MF, the RNI of the workload running in an LPAR can be calculated. The higher the RNI, the deeper into the memory hierarchy the processor must go to retrieve the instructions and data for that workload.

RNI reflects the distribution and latency of sourcing data from shared caches and memory, as shown in Figure 12-3.

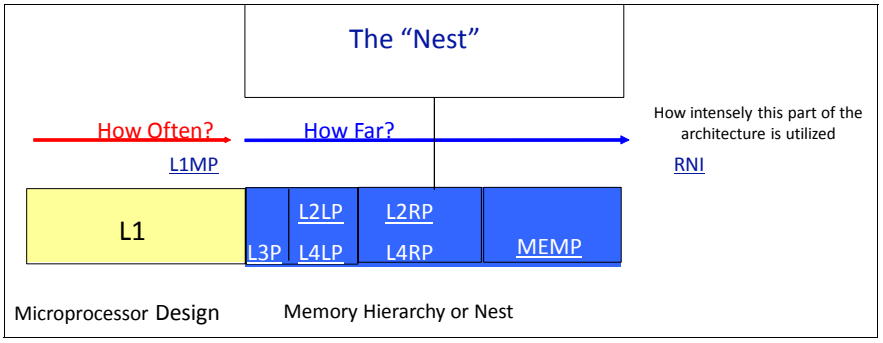


Figure 12-3 Relative Nest Intensity

Many factors influence the performance of a workload. However, usually what these factors are influencing is the RNI of the workload. The interaction of all these factors results in a net RNI for the workload, which in turn directly relates to the performance of the workload.

These factors are tendencies, 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 listed along with their RNI tendency in Figure 12-4.

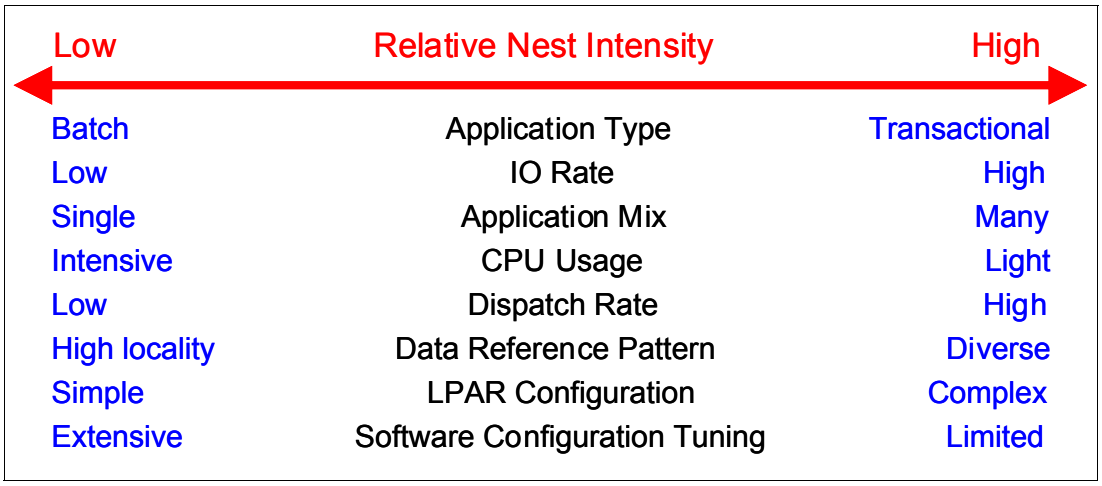


Figure 12-4 The traditional factors that were used to categorize workloads

Little can be done to affect most of these factors. An application type is whatever is necessary to do the job. The data reference pattern and processor usage tend to be inherent to the nature of the application. The LPAR configuration and application mix are mostly a function of what must be supported on a system. The I/O rate can be influenced somewhat through buffer pool tuning.

However, one factor, *software configuration tuning*, is often overlooked but can have a direct effect on RNI. This term refers to the number of address spaces (such as CICS application-owning regions (AORs) or batch initiators) that are needed to support a workload.

This factor has 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. Re-tuning 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.

Three 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, a number of techniques were used to match production workloads to LSPR workloads:

- Application name (a client running CICS can use the CICS LSPR workload)
- Application type (create a mix of the LSPR online and batch workloads)
- I/O rate (the low I/O rates used a mix of low I/O rate LSPR workloads)

The IBM Processor Capacity Reference for IBM Z (zPCR) tool supports the following workload categories:

- Low
- Low-Average

- Average
- Average-high
- High

The IBM zPCR tool reflects the latest IBM LSPR measurements. It is available at no extra charge at:

http://www.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/PRS1381

However, as addressed in 12.5, “LSPR workload categories based on relative nest intensity” on page 435, 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 435.

The AVERAGE RNI LSPR workload is intended to match most client workloads. When no other data is available, use it 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 approaching the size of atoms which stands as a fundamental physical barrier, industry is no longer able to double the processor chip’s performance every two years - which is known as the Moore’s Law[6].

A holistic performance approach is required when the performance gains due to frequency are reduced like today. This is why, hardware and software synergy becomes an absolute requirement.

Starting with z13, IPC (Instructions Per Cycle) improvements in core and cache became the driving factor for performance gains. As these microarchitectural features increase - contibuting to instruction parallellism - overall workload performance variability also increases because not all workloads react the same way to these enhancements.

Because of the nature of the z14 multi-CPC drawer system and resource management across those drawers, performance variability from application to application is expected.

Also, the memory and cache designs affect various workloads in many ways. All workloads are improved, with cache-intensive loads benefiting the most. For instance, having more PUs per CPC drawer, each with higher capacity than z13, more workload can fit on a z14 CPC drawer, which may result in better performance.

[6] Refer to http://www.mooreslaw.org

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.

Nevertheless, the workload variability for moving from z13 to z14 is expected to be stable. Workloads migrating from zEC12 and prior generations to z14 can expect to see similar results with slightly less variability than the typical z13 experience.

Experience demonstrates that IBM Z servers can be run at up to 100% utilization levels, sustained. However, most clients prefer to leave a bit of room and run at 90% or slightly under. For any capacity comparison exercise, using a single metric, such as MIPS or MSU, is not a valid method. When deciding the number of processors and the uniprocessor capacity, remember both the workload characteristics and LPAR configuration. For these reasons, when you plan capacity, using zPCR and involving IBM technical support are recommended.

12.7.1 Main performance improvement drivers with z14 servers

z14 servers are designed to deliver new levels of performance and capacity for large-scale consolidation and growth. The following attributes and design points of z14 servers contribute to overall performance and throughput improvements as compared to the z13.

The *z/Architecture implementation* has the following enhancements:

- Guarded Storage Facility, enhancement for garbage collected languages
- Vector Packed Decimal Facility, enhancement on packed decimal operations
- Vector Enhancements Facility, several vector enhancements including adding support for single precision floating point and VMSL for cryptographic computations.
- Order Preserving Compression and Entropy Encoding for CMPSC, allowing comparisons against compressed data. Entropy Encoding increases compression ratio
- Miscellaneous New General Instructions, enhancement on 64-bit halfword operations and new multiply instructions
- Removal of Runtime Instrumentation External Interruption, avoiding RI Buffer Full interrupt
- Semaphore Assist Facility, Enhanced NIAI (next instruction access intent) codepoints
- (MSA 6, 7, 8), adding SHA-3 hash, true random number generation, and AES-GCM mode

The z14 *microprocessor* has the following design enhancements:

- 14nm FINFET SOI technology with IBM embedded dynamic static random access memory (eDRAM) technology
- Up to ten active processor cores per chip
- Clock frequency at 5.2 GHz
- A new translation/TLB2 design with four hardware-implemented translation engines reducing latency when compared with one pico-coded engine on z13
- Branch prediction improvements:
  - 33% BTB1-and-2 growth (Branch Target Buffer)
  - new perceptron predictor
  - simple call-return stack
- Pipeline optimization
  - improved instruction delivery

  - faster branch wakeup
  - reduced execution latency
  - improved OSC prediction
- Second generation of Simultaneous multithreading (SMT)
  - including SAPs in addition to 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 96K to 128K per Core (33%)
  - L2 D-Cache increased from 2MB to 4MB per Core (2x)
  - L3 Cache (shared) increased from 64MB to 128MB per CP SCM(2x)
  - New L4 Cache design with 672MB (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 additional operations: SHA-3, SHAKE hashing algorithms and True Random-number generation (TRNG)
  - Improved GCM (Galois Counter Mode) performance
  - Entropy-Encoding Compression Enhancement with Huffman encoding
  - Order-Preserving compression
- Adjusted Hiperdispatch to exploit new chip configuration

The z14 design has the following enhancements as compared with the z13:

- Increased total number of PUs that are available on the system, from 168 to 196, and number of characterizable cores, from 141 to 170
- Hardware system area (HSA) increased from 96GB to 192GB
- 32TB of addressable memory (configurable to LPARs) with up to 16 TB of memory per LPAR
- PR/SM enhancements:
  - Improved memory affinity
  - optimized LPAR placement algorithms
- SMT enablement for system assist processor (SAP) processors
- New Coupling Facility Control Code (CFCC) with improved performance and following enhancements:
  - Coupling Facility (CF) Processor Scalability
  - CF List Notification Enhancements
  - CF Request Diagnostics

  - Coupling Link Constraint Relief
  - CF Encryption

The following new features are available on z14 servers:

- Coupling Express Long Reach
- zHyperlink Express
- FICON Express16S+
- 10GbE RoCE Express2
- Crypto Express6S with up to 85 domains

A

IBM Secure Service Container Infrastructure

This chapter introduces the IBM Secure Service Container[1] (SSC) framework. The SSC framework is available on IBM z14 (Driver Level 32), IBM z13 and z13s (Driver level 27) servers. This appendix briefly describes the reason why IBM has created the SSC framework and also a description about how the SSC environment is intended to be used.

This appendix includes the following sections:

- What is SSC?
- Why use SSC?
- IBM Z servers and SSC

[1] Former name: IBM z Appliance Container Infrastructure (zACI)

© Copyright IBM Corp. 2017. All rights reserved.

A.1 What is SSC?

An appliance is an application (software) that provides a specified function or set of functions (service) and is packaged and deployed with a specific (trimmed) operating system in either a virtual machine or a dedicated COTS (commodity of the shelf) hardware (physical server), and requires little to no intervention from a system administrator (SW 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 a system administrators or other privileged users) as well as simple deployment and maintenance.

In the current IT deployments, various components serving the business processes (databases, middleware, applications) require specialized management functions (access management, enterprise directories, secure key management, backup and restore - to name a few). The development requirements of the management functions do not follow the dynamic of the actual business functions.

In addition, due to the diversity of the platforms on which the business applications run, if the management functions are deployed alongside the mainstream business applications, each time maintenance and/or upgrade of the platform is required, the management function must also be maintained (updated, tested or even certified), increasing complexity and associated IT spending,

As such, these management functions can be deployed using an appliance model where the functions provided is available and accessible through standardized methods.

A large diversity of appliances is available from various suppliers, each appliance having:

- Separate administration (and deployment process)
- Different HW configuration requirements
- Different performance profile and management requirements
- Different security characteristics (that require alignment with enterprise requirements)

IBM Z Appliance

An IBM Z Appliance is an integration of operating system, middleware and software components that work autonomously and provide core services and infrastructure focusing on consumability and security. An appliance is deployed as bootable system image that contains all of the layers necessary to provide a specific set of services or functions. IBM Z Appliance are implemented as:

- Firmware appliance
- Software appliance

IBM Secure Service Container

IBM has developed the Secure Service Container (SSC) framework which provides the base infrastructure to create and deploy an IBM Z 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 workload deployed from being accessed by a system administrator or an external attacker.

SSC LPAR

For IBM Z servers, the SSC Partition is an LPAR type running an IBM Z Appliance based on Secure Service Container framework.

Multiple virtual appliances integrated into IBM Secure Service Container can be deployed on IBM z14 (z13 and z13s also). These virtual appliances have:

- Common administration (deployment)
- Common HW configuration
- Commonly managed performance profiles
- Common security characteristics (aligned with enterprise requirements)

At the time of writing, the following appliances are available from IBM:

- z/VSE Network Appliance
- IBM Z Advanced Workload Analysis Reporter (IBM zAware), now deployed as software appliance and integrated with IBM Operations Analytics for Z.

More appliances are expected to appear in the future. Appliances can be implemented as firmware or software, depending on the environment in which the appliance runs and the function it must provide.

The SSC framework is available on IBM z14, z13, and z13s servers.

A.2 Why SSC?

The Secure Service Container framework simplifies the process a team has to apply to create an appliance, and enforces a common set of behaviors for operations that all appliances have to perform.

Secure Service Container framework provides a set of utilities 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 standalone appliance that is easily installable onto the IBM Z platform.

The SSC framework enables the decision to release a product as software or firmware 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.3 IBM Z servers and SSC

Appliances based on the SSC framework share following features and characteristics:

- Encapsulated Operating Systems
- Services provided via 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 either as firmware or software

The deployment model for an appliance is shown in Figure A-1 on page 444.

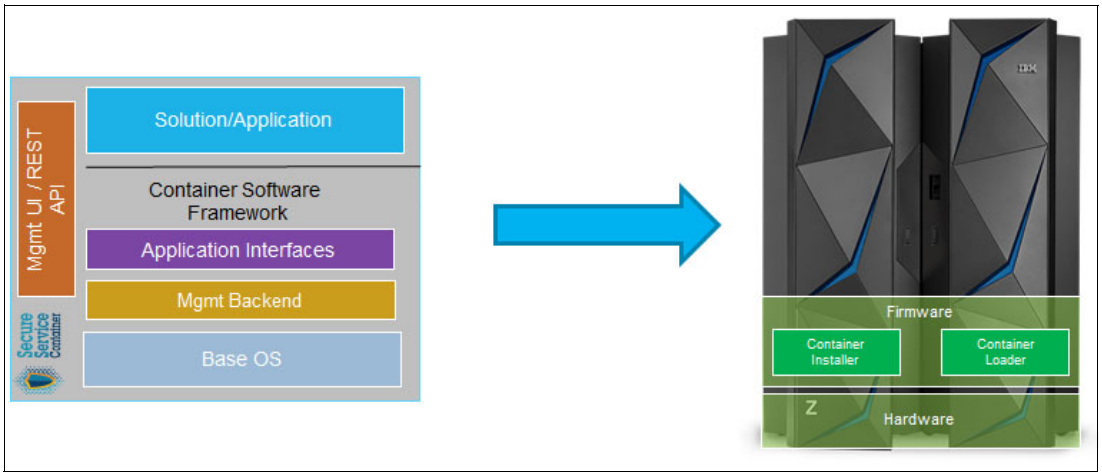


Figure A-1 Appliance deployment in a SSC LPAR on IBM Z

Secure Service Container provides a highly secure context (Figure A-2) for deploying appliances that:

- Allows no system admin access
  - Once the appliance image is built, OS access (ssh) is not possible
  - Only Remote APIs available
  - Memory access of system admin is disabled
- Data storage uses encrypted disk
- Debug data (dumps) encrypted
- Strong isolation between container instances
- High assurance isolation

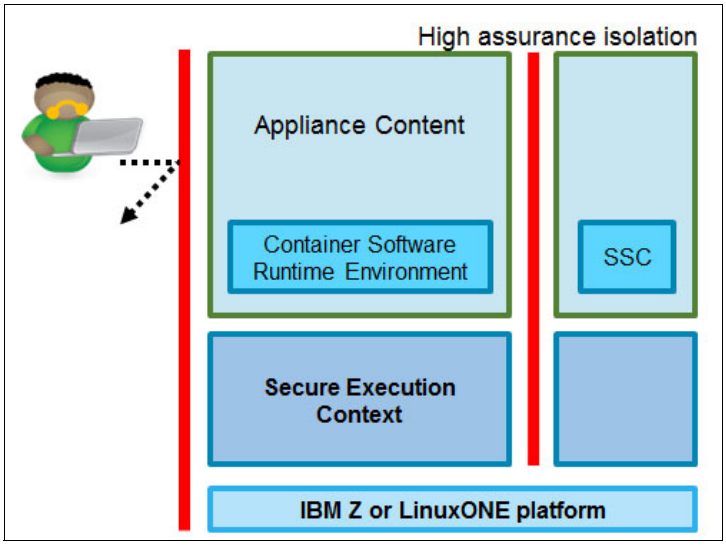


Figure A-2 SSC protection

The deployment of appliance follows these steps:

- Buying the software appliance
- Downloading the appliance image
- Creating and activating the SSC LPAR.
- deploying the appliance using the appliance installer
- configuring and using appliance through REST API or web UI.

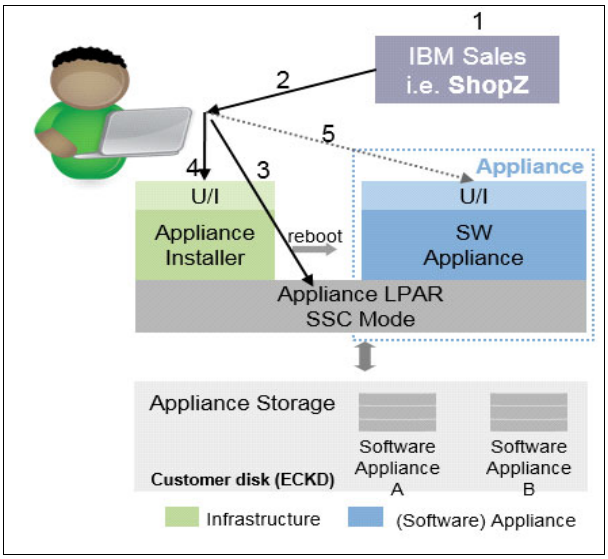


Figure A-3 SSC deployment in 5 steps

The Secure Service Container Framework provides following appliance management controls for appliance administrators:

- View messages and events
- Network management
- User management
- Disk management
- View Appliance status
- Export/import data
- Apply service, updates
- Software license support

At the time of this writing, the SSC Software Framework provides support for:

- 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, virsh packages
- Embedded OS upgrades
- Support smart card machine unique key handling

B

Channel options

This appendix describes all channel attributes, the required cable types, the maximum unrepeated distance, and the bit rate for z14 servers. The fatures 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 12xIFB connection has one row of 12 fibers. The MPO connector of the zHyperLink, and the ICA connection has 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.

Table B-1 lists the attributes of the channel options that are supported on z14 servers.

Table B-1 z14 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,	100m	New build
			OM4	150 m	
FICON Express16S+ 10KM LX	0427	4, 8, or 16	SM 9 µm	10 km (6.2 miles)	New build
FICON Express16S+ SX	0428	4, 8, or 16	OM2, OM3, OM4	See Table B-2 on page 449.	New build
FICON Express16S 10KM LX	0418	4, 8, or 16	SM 9 µm	10 km (6.2 miles)	Carry forward
FICON Express16S SX	0419	4, 8, or 16	OM2, OM3, OM4	See Table B-2 on page 449.	Carry forward

© Copyright IBM Corp. 2017. All rights reserved.

Channel feature	Feature codes	Bit rate[a] in Gbps (or stated)	Cable type	Maximum unrepeated 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 449.	Carry forward
Open Systems Adapter (OSA) and Remote Direct Memory over Converged Ethernet (RoCE)					
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-Expess6S 10 GbE SR	0425	10	MM 62.5 µm MM 50 µm	33 m (200) 82 m (500) 300 m (2000)	New build
OSA-Expess5S 10 GbE SR	0416				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-Express6S GbE SX	0423	1.25	MM 62.5 µm MM 50 µm	275 m (200) 550 m (500)	New build
OSA-Express5S GbE SX	0414				Carry forward
OSA-Express6S 1000BASE-T	0426	100 or 1000 Mbps	Cat 5, Cat 6 unshielded twisted pair (UTP)	100m	New build
OSA-Express5S 1000BASE-T	0417				Carry forward
OSA-Express4S 1000BASE-T	0408	10, 100, or 1000 Mbps			
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 (PCIe-O SR)	0172	8 GBps	OM4	150 m	New build or Carry forward
			OM3	100 m	
HCA3-O (12x IFB)	0171	6 GBps	OM3	150 m	New build or Carry forward
HCA3-O LR (1x IFB)	0170	2.5 or 5 Gbps	SM 9 µm	10 km (6.2 miles)	
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 that include latency through the adapters, cable lengths, and the type of workload.
b. Where applicable, the minimum fiber bandwidth distance in MHz-km for multi-mode fiber optic links is included in parentheses.
c. 600 meters maximum when sharing the same switch across 2 RoCE Express features

Table B-2 shows the maximum unrepeated distances for FICON short wavelength (SX) features.

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)

C

Native Peripheral Component Interconnect Express (PCIe)

This appendix introduces the native Peripheral Component Interconnect Express (PCIe) features management on z13 servers and includes concepts of the integrated firmware processor (IFP) and resource groups (RG).

The following topics are included:

- Design of native PCIe adapter management
- Native PCIe feature plugging rules
- Native PCIe feature definitions

© Copyright IBM Corp. 2017. All rights reserved.

C.1 Design of native PCIe adapter management

The native PCIe adapter is a new category of features introduced since zEC12. IBM z14 supports the following “native” PCIE features:

- 10GbE RoCE Express2 (FC #0412)
- 10GbE RoCE Express (FC #0411) (carry forward only)
- zEDC Express (FC #0420)
- zHyperLink Express (FC #0431)
- Coupling Express Long Reach (CE LR) (FC #0433)

These adapters are installed into a PCIe I/O drawer, together with the existing I/O features and have a physical channel ID (PCHID) assigned according to its physical location.

For all the feature adapters installed in an I/O drawer, management functions in the form of device drivers and diagnostic tools are always implemented to support virtualization of the adapter, service, and maintenance.

Traditionally, these management functions are integrated on the adapter with specific hardware design. For the newly introduced native PCIe adapters, these functions are moved out of the adapter and are now handled by an IFP.

This section covers the following topics:

- Native PCIe adapter
- Integrated firmware processor
- Resource groups
- Management tasks

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 will always download the device drivers and diagnostic tools from the Support Element (SE) and run the management functions on the adapter. In the new design, there is no ASIC chip for management function on the native PCIe feature adapters.

For the RoCE Express, zHyperlink Express 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 still implemented.

C.1.2 Integrated firmware processor

The IFP is a processor unit (PU) exclusively used to manage native PCIe feature adapters that are installed in the PCIe I/O drawer. On previous systems, this processor was not used and was called a *reserved processor*. It is allocated from the system PU pool and is not counted in the PUs available for characterization.

If a native PCIe feature is installed in the system, the system allocates and initializes an IFP during its power-on reset (POR) phase. Although the IFP is allocated to one of the physical PUs, it is not visible to the users. In an error or failover scenario, PU sparing also happens for an IFP, with the same rules as other PUs.

C.1.3 Resource groups

The IFP allocates four resource groups for running the management functions of native PCIe feature adapters. A native PCIe feature adapter is managed by one of the resource groups according to which I/O domain this adapter is located in.

As shown in Figure C-1, each I/O domain in a PCIe I/O drawer of a z14 server is logically attached to one resource group. The native PCIe I/O feature adapters managed by their respective RG for device drivers and diagnostic tools functions.

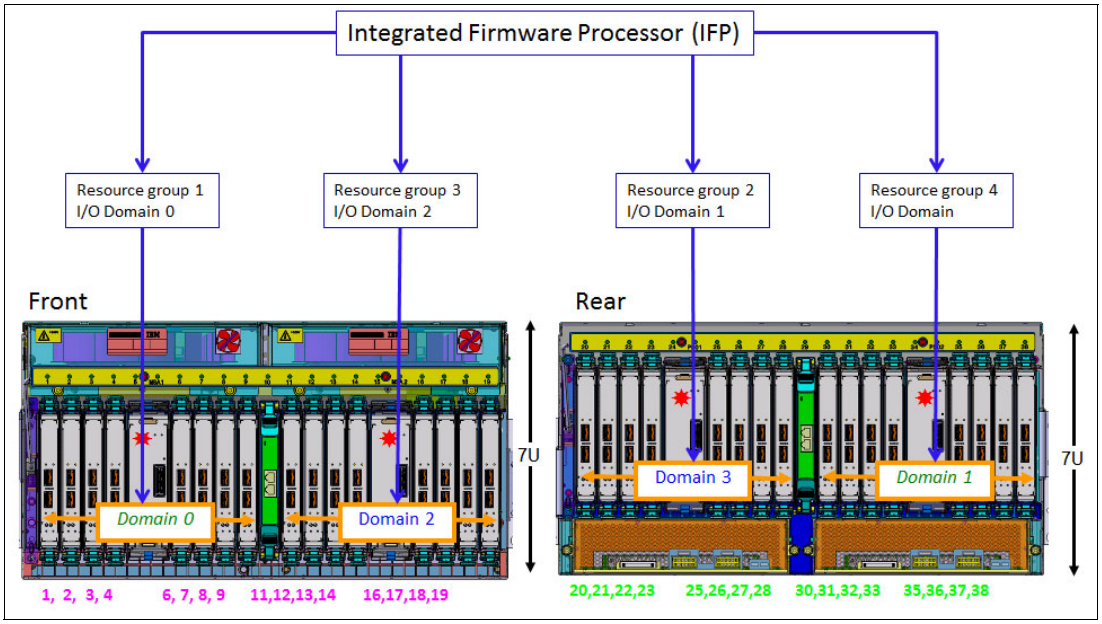


Figure C-1 I/O domains and resource groups managed by the IFP - z14

Up to five PCIe I/O drawers are supported on z14 servers. The same type of native PCIe features are always assigned to different I/O domains in different resource groups, and different PCIe I/O drawers if the configuration has them, to eliminate the possibility of a single point of failure.

Currently, an I/O domain of the PCIe I/O drawer can support a total of two native PCIe feature adapters, together with the existing PCIe feature cards (FICON, OSA, and Crypto).

C.1.4 Management tasks

The IFP and resource groups perform all management tasks on the native PCIe adapters:

- Firmware update of adapters and resource groups
- Error recovery and failure data collection
- Diagnostic and maintenance taskss

Firmware update of adapters and resource groups

Firmware of native PCIe adapters and resource groups are part of the system's microcode and can be updated by a Microcode Change Level (MCL) upgrade. MCL upgrades on adapters or on the code of the resource groups require the specific adapter or all native PCIe adapters managed by the specific resource group (depending on the type of u-code that it applies) to be offline during activation of the MCL. However, to maintain availability, MCLs can only be applied to one resource group at a time. While one resource group is offline, the

second resource group and all adapters in it remain active. An MCL application for a native PCIe adapter or resource group is not possible if an error condition exists within the other resource group.

Error recovery and failure data collection

If an error in one of the resource groups or features that are assigned to one of the resource groups occurs, the IFP manages error recovery and collects error data. The error data is sent by the IFP to the SE, which then provides a message on the SE and the Hardware Management Console (HMC). If an error requires maintenance, a call home to the IBM Support system is initiated by the HMC.

Diagnostic and maintenance tasks

Any maintenance action on a native PCIe feature is managed by the IFP, including testing or replacing a feature card. Before configuring a feature offline, the IFP ensures that the same type of feature is available in the same or the other resource group (if applicable).

C.2 Native PCIe feature plugging rules

The following are the maximum number of native PCIe adapters that can be installed in a z14 server:

- Up to eight 10GbE RoCE Express/Express2 features. Each feature has one adapter configured.
- Up to Eight zEDC Express features. Each feature has one adapter configured.
- Up to 16 zHyperLink Express features
- Up to 32 Copupling Express Long Reach features

Considering availability, install adapters of the same type in slots of different I/O domains, drawers, fanouts, and resource groups . The next sections describe more information about achieving a highly available configuration.

C.3 Native PCIe feature definitions

During the ordering process of the native PCIe adapters, such as the zEDC Express and 10GbE RoCE Express features, features of the same type are evenly spread across two resource groups (resource group 1 and resource group 2) for availability and serviceability.

Notes:

- Although Flash Express features are counted as native PCIe cards for the total number of Native PCIe features, they are not part of any resource group.
- The Flash Express features are not defined by using the IOCDS.

Figure C-2 on page 455 shows a sample PCHID report of a z13 configuration with four zEDC Express features and four 10GbE RoCE Express features. It lists the following information for each adapter:

- PCHID and ports
- The Resource Group that the adapter is attached to (Comment column)
- Physical location (drawer, slot)

In Figure C-2, FC #0420 = zEDC Express, and FC #0412 = 10GbE RoCE Express2.

```
 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
Source             Drwr   Slot  F/C*    PCHID/Ports or AID        Comment
  A15/LG07/J01     Z22B   02    0420    104                       RG1

  A27/LG07/J01     Z22B   19    0412    13C/D1D2                  RG3

  A23/LG03/J01     Z22B   38    0412    17C/D1D2                  RG2

  A19/LG07/J01     Z15B   02    0412    184/D1D2                  RG1

  A27/LG03/J01     Z15B   20    0412    1C0/D1D2                  RG4

  A15/LG03/J01     Z15B   38    0420    1FC                       RG2

  A27/LG09/J01     Z08B   19    0420    23C                       RG3

  A27/LG05/J01     Z01B   20    0420    2C0                       RG4
```

Figure C-2 Sample output of AO data or PCHID report

The native PCIe features are not part of the traditional channel subsystem (CSS). They do not have a channel-path identifier (CHPID) assigned, but they have a PCHID 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 in Figure C-3 on page 455 defines zEDC Express and 10GbE RoCE Express2 features 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=104
FUNCTION FID=06,VF=1,PART=((LP14),(LP01)),TYPE=ZEDC,PCHID=1FC

zEDC Express Functions for LPAR LP15, Reconfigurable to LP02:
FUNCTION FID=07,VF=1,PART=((LP15),(LP02)),TYPE=ZEDC,PCHID=23C
FUNCTION FID=08,VF=1,PART=((LP15),(LP02)),TYPE=ZEDC,PCHID=2C0

10GbE RoCE Express Functions for LPAR LP14, Reconfigurable to LP03 or LP04:
FUNCTION FID=9,VF=01,PART=((LP14),(LP03,LP04)),PNETID=(NET1,NET2),  *
         TYPE=ROCE,PCHID=13C
FUNCTION FID=A,VF=1,PART=((LP14),(LP03,LP04)),PNETID=(NET1,NET2),  *
         TYPE=ROCE,PCHID=17C

10GbE RoCE Express Functions for LPAR LP15, Reconfigurable to LP03 or LP04:
FUNCTION FID=B,VF=01,PART=((LP15),(LP03,LP04)),PNETID=(NET1,NET2),  *
         TYPE=ROCE,PCHID=184
FUNCTION FID=C,VF=01,PART=((LP15),(LP03,LP04)),PNETID=(NET1,NET2),  *
         TYPE=ROCE,PCHID=1C0
```

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 between 000 and FFF that you use to assign a PCHID to the FUNCTION to identify the specific hardware feature in the PCIe I/O drawer. Because the FUNCTION is not related to a channel subsystem, all LPARs on a 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.

C.3.2 Virtual function number

If you want several LPARs to be able to use a zEDC Express feature (the 10GbE RoCE Express feature cannot be shared between LPARs), you need to use a Virtual Function (VF) number. A VF number is a number between 1 and *nnn*, where *nnn* is the maximum number of LPARs that the feature supports. The maximum is 15 for the zEDC Express feature and 126 for the RoCE Express2 feature.

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 10GbE RoCE Express features. Each FUNCTION definition supports up to four PNETIDs.

Notes:

- For more information about FUNCTION statement, see *IBM Z Input/Output Configuration Program User's Guide for ICP IOCP*, SB10-7163.
- Definition of10GbE RoCE Express/Express2 feature is required to pair up with an OSD CHPID definition, by the parameter of PNETID. The OSD CHPID definition statement is not listed in the example.

D

Shared Memory Communications

This appendix briefly describes the optional Shared Memory Communications (SMC) function 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.

This appendix includes the following sections:

- Shared Memory Communications overview
- Shared Memory Communication over RDMA
- Shared Memory Communications - Direct Memory Access

© Copyright IBM Corp. 2017. All rights reserved.

D.1 Shared Memory Communications overview

As the volume of data being generated and transmitted by technologies that are driven by cloud, mobile, analytics, and social computing applications grows, pressure increases on business IT organizations to be able to provide fast access to that data across the web, application, and database tiers that comprise most enterprise workloads. Shared Memory Communications helps to access data faster and with less latency as well as reducing CPU resource consumption over traditional TCP/IP for communications.

D.2 Shared Memory Communication over RDMA

IBM z14 delivers a new RoCE feature, 10GbE RoCE Express2 (FC #0412). The 10GbE RoCE Express (FC #0411) feature can be carried forward to z14 (from z13/zEC12). SMC-R (RoCE) highlights

- 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 existing TCP applications to benefit transparently from RDMA for transferring data:
  - SMC-R uses 10GbE RoCE Express/Express2 adapter as the physical transport layer.
  - Initial deployment is limited to z/OS to z/OS communications with a goal to expand exploitation to more operating systems, and possibly appliances and accelerators.
- Single Root I/O Virtualization (SR-IOV) technology provides the capability to share the 10GbE RoCE Express adapter between logical partitions (LPARs):
  - RoCE Express2 supports 63 Virtual Functions (VFs) per physical port for a total of 126 VFs per PCHID.
  - RoCE Express supports 31 Virtual Functions (VFs) per PCHID.
- Maximum number of RoCE Express and RoCE Express2 features supported per z14 is eight (8) (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 transfer of data to or from memory on a remote system with low latency, high throughput, and low CPU utilization.

Traditional Ethernet transports, such as TCP/IP, typically use software-based mechanisms for error detection and recovery, and are based on the underlying Ethernet fabric using a “best-effort” policy. With the traditional policy, the switches typically discard packets in congestion and rely on the upper-level transport for packet retransmission.

RoCE. however, uses hardware-based error detection and recovery mechanisms that are defined by the InfiniBand specification. A RoCE transport performs best when the underlying Ethernet fabric provides a lossless capability, where packets are not routinely dropped. This 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 10GbE RoCE Express features. This capability is enabled, in the 10GbE RoCE Express feature by default.

The two key requirements for RDMA are shown in Figure D-1 on page 459:

- A reliable “lossless” Ethernet network fabric (LAN for layer 2 data center network distance)
- An RDMA network interface card (RNIC)

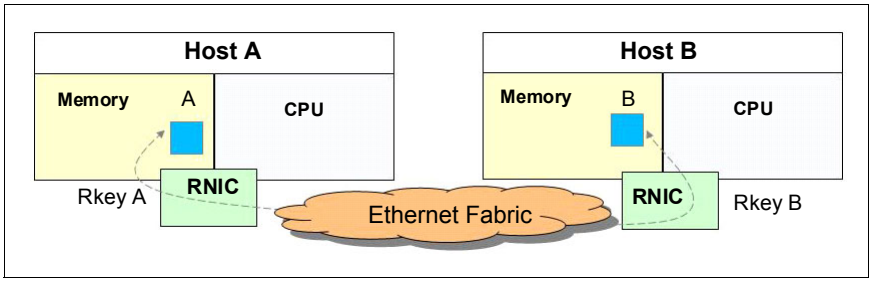


Figure D-1 RDMA technology overview

RDMA technology is now available on Ethernet. RoCE uses an existing Ethernet fabric (switches with Global Pause enabled) and requires advanced Ethernet hardware (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.

SMC-R is a “hybrid” solution as shown in Figure D-2:

- It uses an existing TCP connection to establish the SMC-R connection.
- A TCP option (SMCR) controls switching from TCP to “out of band” SMC-R.
- The SMC-R information is exchanged within the TCP data stream.
- Socket application data is exchanged through RDMA (write operations).
- The TCP connection remains to control the SMC-R connection.
- This model preserves many critical existing operational and network management features of TCP/IP.

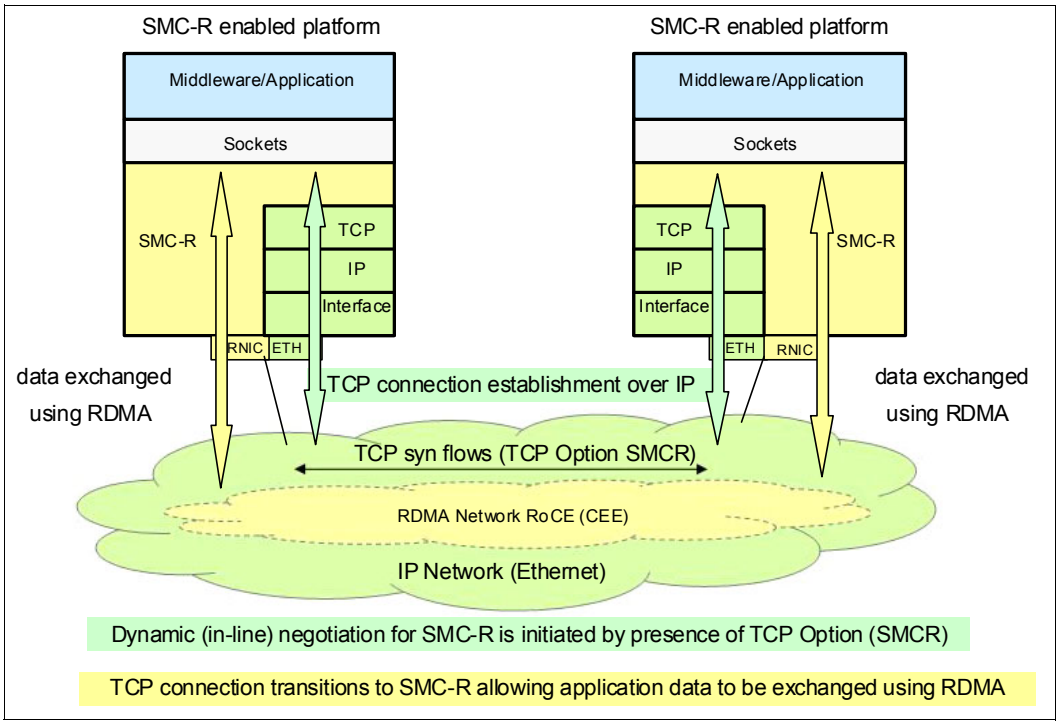


Figure D-2 Dynamic transition from TCP to SMC-R

The hybrid model of SMC-R uses these key existing attributes:

- It follows the standard TCP/IP connection setup.
- The hybrid model switches to RDMA (SMC-R) dynamically.
- The TCP connection remains active (idle) and is used to control the SMC-R connection.
- The hybrid model preserves the following critical operational and network management TCP/IP features:
  - Minimal (or zero) IP topology changes
  - Compatibility with TCP connection-level load balancers
  - Preservation of the existing IP security model, such as IP filters, policies, virtual LANs (VLANs), and Secure Sockets Layer (SSL)
  - Minimal network administration and management changes
- Host application software is not required to change, so all host application workloads can benefit immediately.

D.2.3 Single Root I/O virtualization (SR-IOV)

SR-IOV is a technology that is designed to provide the capability to share the adapter between up to 31 LPARs. SR-IOV is also designed to provide isolation of virtual functions within the PCIe 10GbE RoCE Express adapter. For example, one LPAR cannot cause errors visible to other virtual functions or other LPARs. Each operating system LPAR has its own application queue in its own memory space.

Figure D-3 shows the concept of the Shared RoCE Mode.

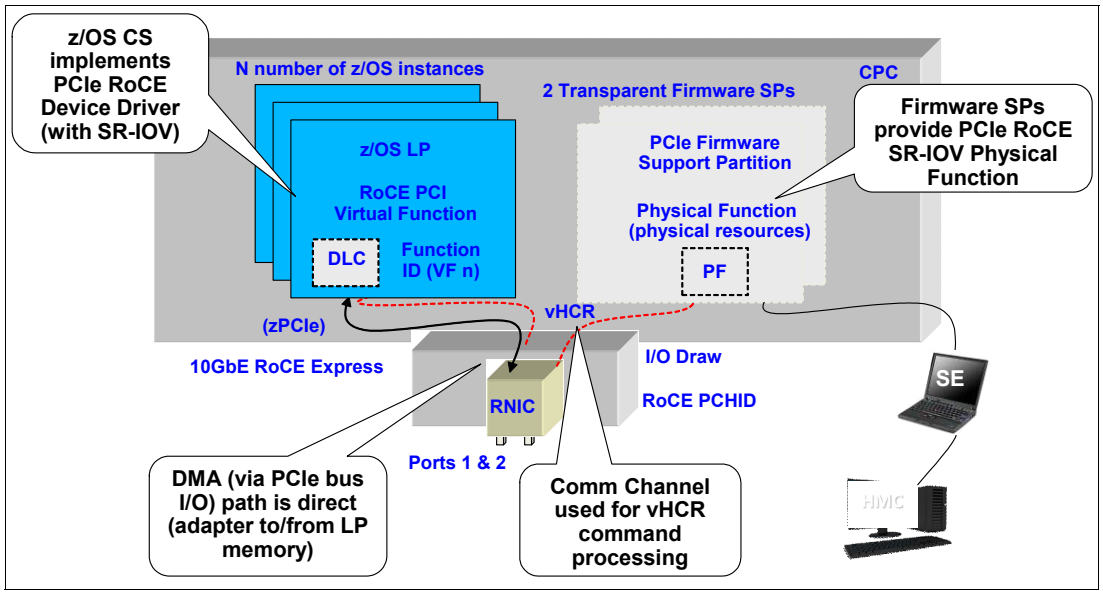


Figure D-3 Shared RoCE mode concepts

The Physical Function Driver communicates to the physical function in the PCIe adapter. The Physical Function Driver has a relatively limited function:

- Manages 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 10 Gigabit Ethernet (10GbE) RoCE Express feature (FC #0411) and 10GbE RoCE Express2 (FC #0412) are RDMA-capable NICs. The integrated firmware processor (IFP) has four resource groups (RGs) that contain firmware for the RoCE Express feature. For more information, see C.1.3, “Resource groups” on page 453.

D.2.5 10GbE RoCE Express/Express2 features

The 10GbE RoCE Express/Express2 features 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. Using the 10GbE RoCE Express/Express2 features also helps to reduce network latency with memory-to-memory transfers that use SMC-R in z/OS V2.1 or later. It is transparent to applications and can be used for LPAR-to-LPAR communications on a single z/OS system or server-to-server communications in a multiple CPC environment.

Table D-1 shows the differences in the number of ports and shared support for different systems.

Table D-1 RoCE number of enabled ports and shared/dedicated environment

System Name	z/OS Supported Ports	Shared mode	Dedicated mode
z14	2	YES[a]	No
z13	2	YES[b]	No
z13s	2	YES	No
zEC12	1	NO	YES
zBC12	1	NO	YES

a. Up to 126 Virtual Functions (VFs) per PCHID for RoCE Express2 (FC #0412)
b. Up to 31 VFs supported per PCHID for RoCE Express (FC #0411)

The 10GbE RoCE Express feature shown in Figure D-4 is installed in the PCIe I/O drawer.

Figure D-4 10GbE RoCE Express

Each feature has one PCIe adapter and two ports. A maximum of eight (8) features can be installed. The 10GbE RoCE Express/Express2 feature uses a short reach (SR) laser as the optical transceiver and supports the use of a multimode fiber optic cable terminated with an LC Duplex connector. Both point-to-point connection (with another 10GbE RoCE Express/Express2 adapter) and switched connection with an enterprise-class 10 GbE switch are supported.

Although SMC-R can be taken advantage of by using direct RoCE Express to RoCE Express connectivity (without any switch), this type of direct physical connectivity forms a single physical point-to-point connection, disallowing any other connectivity with other LPARs such

as other SMC-R peers. Although this is a viable option for test scenarios, it is not practical (nor recommended) for production deployment.

If the IBM 10GbE RoCE Express/Express2 features are connected to 10 GbE switches, the switches must support the following requirements:

- Global Pause function frame (as described in the IEEE 802.3x standard) should be enabled
- Priority flow control (PFC) disabled
- No firewalls, no routing, and no intraensemble data network (IEDN)

The maximum supported unrepeated distance, point-to-point is 300 meters (984.25 ft).

A client-supplied cable is required. Three types of cables can be used:

- OM3 50 micron multimode fiber optic cable rated at 2000 MHz-km terminated with an LC Duplex connector (supports 300 m (984.25 ft))
- OM2 50 micron multimode fiber optic cable rated at 500 MHz-km terminated with an LC Duplex connector (support 82 m (269 ft))
- OM1 62.5 micron multimode fiber optic cable rated at 200 MHz-km terminated with an LC Duplex connector (support 33 m (108.2 ft))

D.2.6 10GbE RoCE Express/Express2 configuration example

Figure D-5 illustrates a sample configuration that allows redundant SMC-R connectivity among LPAR A and C, and LPAR 1, 2 and 3. Each feature can be shared or dedicated to an LPAR. Like the sample configuration, two features per LPAR are advised for redundancy.

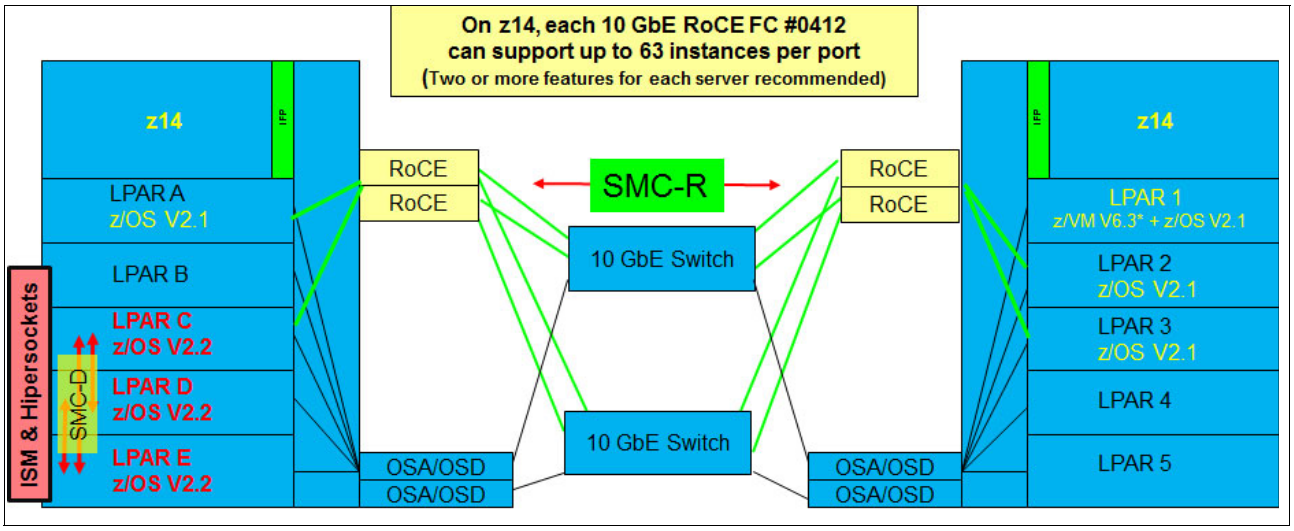


Figure D-5 10GbE RoCE Express sample configuration

The configuration that is shown in Figure D-5 allows redundant SMC-R connectivity among LPAR A, LPAR C, LPAR 1, LPAR 2, and LPAR 3. LPAR to LPAR OSD connections are required to establish the SMC-R communications. 1 GbE OSD connections can be used instead of 10 GbE. OSD connections can flow through the same 10 GbE switches or different switches.

Note: The OSA-Express Adapter and the RoCE Express feature must be associated to each other by having equal PNET IDs (defined in the hardware configuration definition (HCD)).

An OSA-Express feature, defined as channel-path identifier (CHPID) type OSD, is required to establish SMC-R. Figure D-6 shows the interaction of OSD and the RNIC. The OSA feature might be a single or pair of 10 GbE, 1 GbE, or 1000Base-T OSAs. The OSA needs to be connected to another OSA on the system with which the RoCE feature is communicating. In Figure D-5, 1 GbE OSD connections can still be used instead of 10 GbE and OSD connections can flow through the same 10 GbE switches.

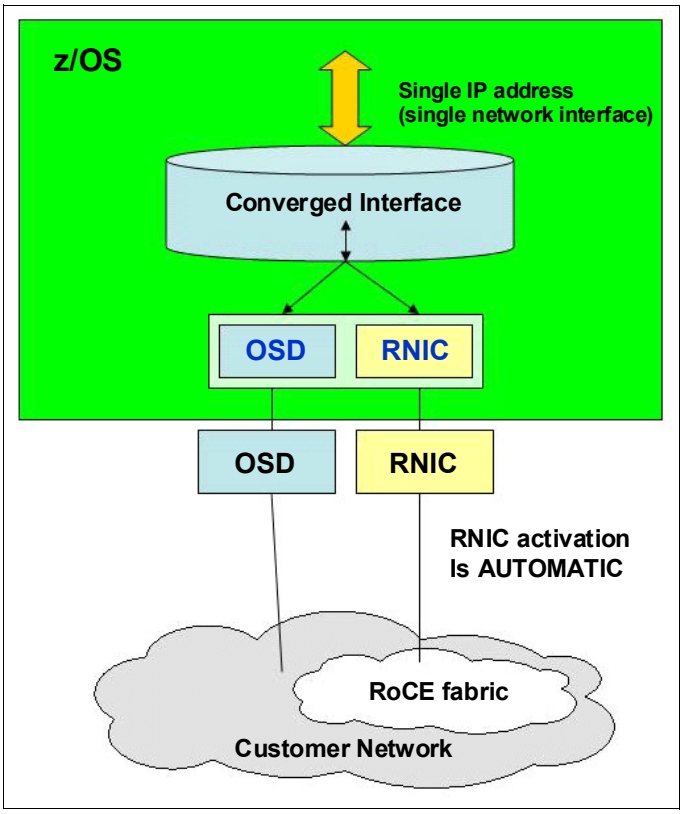


Figure D-6 RNIC and OSD interaction

The following notes refer to Figure D-6:

- The z/OS system administrator must only configure and manage the OSD interface.
- The Communications Server transparently splits and converges network traffic to and from the converged interface.
- Only OSD connectivity must be configured.

With SMC-R, the RNIC interface is dynamically and transparently added and configured.

D.2.7 Hardware configuration definitions

The following HCDs are important.

Function ID

The RoCE feature is identified by a hexadecimal Function Identifier (FID). It has a dedicated limit in the range 00 - FF, while the shared limit is 000 - 0FFF in the HCD or Hardware Management Console (HMC), to create the input/output configuration program (IOCP) input. A FID can only be configured to one LPAR, but it is reconfigurable. The RoCE feature, as installed in a specific PCIe I/O drawer and slot, is to be used for the defined function. The physical installation (drawer and slot) determines the physical channel identifier (PCHID). Only one FID can be defined for dedicated mode.

Virtual Function ID

Virtual Function ID is defined when PCIe hardware is shared between LPARs. Virtual Function ID has a decimal Virtual Function Identifier (VF=) in the range 1 – nn, where nn is the maximum number of partitions that the PCIe feature supports. For example, the RoCE Express2 feature supports up to 126 partitions, and a zEDC Express feature 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 both OSA and RNIC interfaces in the HCD. A PNet ID is defined for each physical port. z/OS Communications Server gets the information during the activation of the interfaces and associates the OSD interfaces with the RNIC interfaces that have matching PNet ID values.

Attention: If you do not configure a PNet ID for the RNIC adapter, activation fails. If you do not configure a PNet ID for the OSA adapter, activation succeeds, but the interface is not eligible to use SMC-R.

Figure D-7 shows the three physically separate networks defined.

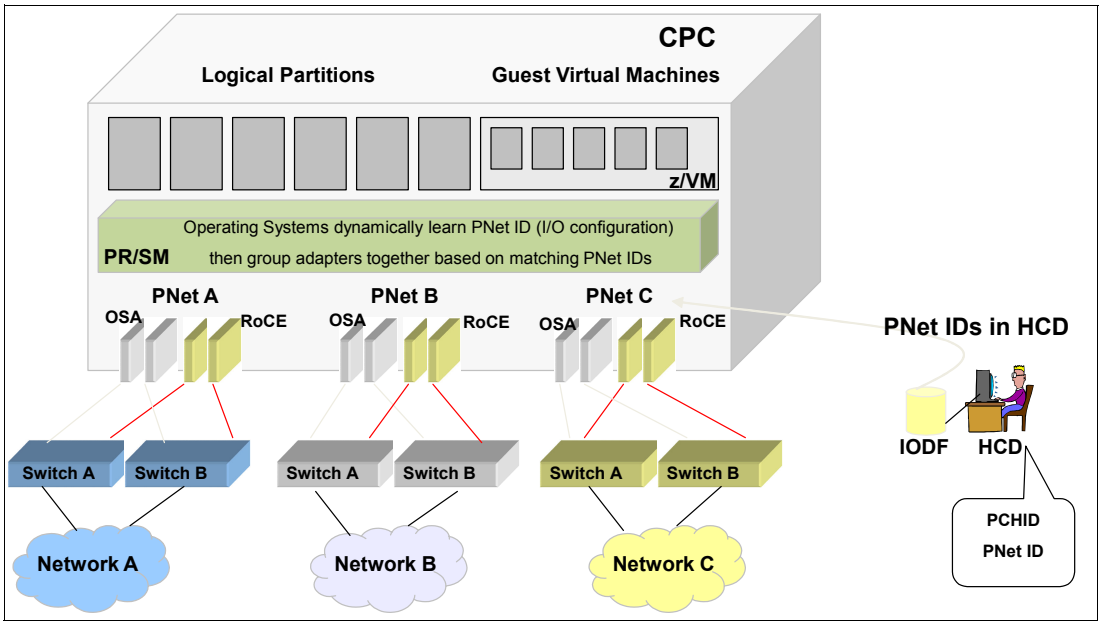


Figure D-7 Physical network ID example

D.2.8 Software exploitation of SMC-R

SMC-R can be implemented on the RoCE and can communicate memory to memory, thus avoiding the CPU resources of TCP/IP by reducing network latency and improving wall clock time. It focuses on “time to value” and widespread performance benefits for all TCP socket-based middleware.

The following advantages are gained as shown in Figure D-8:

- No middleware or application changes (transparent)
- Ease of deployment (no IP topology changes)
- LPAR-to-LPAR communications on a single central processing complex (CPC)
- Server-to-server communications in a multi-CPC environment
- Retained key qualities of service that TCP/IP offers for enterprise class server deployments (high availability, load balancing, and an IP security-based framework)

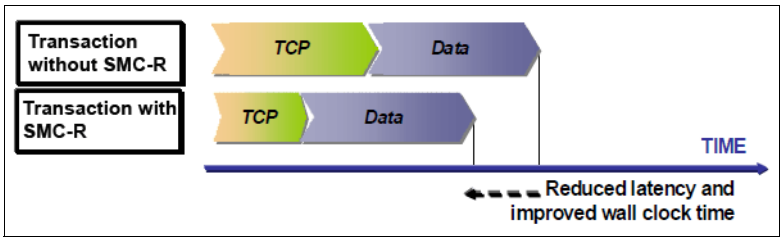


Figure D-8 Reduced latency and improved wall clock time with SMC-R

D.2.9 SMC-R support overview

SMC-R needs both hardware and software support.

Hardware

SMC-R requires the following hardware:

- PCIe-based RoCE Express2
  - z14 servers
  - Dual port 10 GbE adapter
  - Maximum of 8 RoCE Express2 features per CPC
- PCIe-based RoCE Express
  - z14, z13, z13s, zEC12, and zBC12 servers
  - Dual port 10 GbE adapter
  - Maximum of 8 RoCE Express features per CPC
- HCD and input/output configuration data set (IOCDS):
  - PCIe FID, VF (sharing), and RoCE configuration with PNet ID
- Optional: Standard 10 GbE switch (CEE-enabled switch is not required)
- Required queued direct input/output (QDIO) Mode OSA connectivity between z/OS LPARs as shown in Figure D-5 on page 463.
- Adapter MUST be dedicated to an LPAR on a zEC12 or zBC12. And it MUST be shared (or at least in shared mode) to one or more LPARs on a z14, z13 or z13s server.
- SMC-R cannot be used in IEDN.

Software

SMC-R requires the following software:

- z/OS V2R1 (with PTFs) or higher are the only supported operating systems for the SMC-R protocol. You cannot roll back to previous z/OS releases.
- z/OS guests under z/VM 6.3 or later are supported to use 10GbE RoCE features.
- IBM is working with its Linux distribution partners to include support in future Linux on z Systems distribution releases.

IOCP required level for z14 servers: The required level of IOCP for z13 servers is V5 R2 L1 (IOCP 5.4.1) or later with program temporary fixes (PTFs). For more information, see the following manuals:

- *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.

Other RoCE considerations

RoCE has the following additonal considerations:

- RoCE system limits:
  - 128 unique VLANs per PCHID physical port
  - Each VF guarantees a minimum of 2 VLANs max of 16
- z/OS CS consumption of RoCE virtual resources:
  - One VF consumed 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
  - Existing z/OS RoCE users might be required to make a TCP/IP configuration change (that is, existing TCP/IP profile (PFIDs) might be compatible with shared RoCE)
- Changes are required for existing RoCE users for the following cases:
  - z/OS users who use multiple TCP/IP stacks and both stacks currently use the same RoCE feature (single z/OS image sharing a physical card among multiple stacks).
  - z/OS users who need to use both physical RoCE ports from the same z/OS instance (not “best practices”, but is allowed).
  - z/OS users who could not continue using (coordinate) the same PFID values (continue using the existing PFID value that is used in the dedicated environment for a specific z/OS instance) when adding multiple PFIDs and VFs to the same card (for additional shared users).

D.2.10 SMC-R use cases for z/OS to z/OS

SMC-R with RoCE provides high-speed communications and “HiperSockets-like” performance across physical processors. It can help all TCP-based communications across z/OS LPARs that are in different CPCs.

The following list shows several typical communications patterns:

- Optimized Sysplex Distributor intra-sysplex load balancing

- WebSphere Application Server type 4 connections to remote DB2, IMS, and CICS instances
- IBM Cognos® to DB2 connectivity
- CICS to CICS connectivity through Internet Protocol interconnectivity (IPIC)

Optimized Sysplex Distributor intra-sysplex load balancing

Dynamic virtual IP address (VIPA) and Sysplex Distributor support are often deployed for high availability (HA), 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 having to traverse the Sysplex Distributor node for every inbound packet, which is the current model with TCP/IP. In the new model, only connection establishment flows must go through the Sysplex Distributor node.

Sysplex Distributor before RoCE

Figure D-9 shows a traditional Sysplex Distributor.

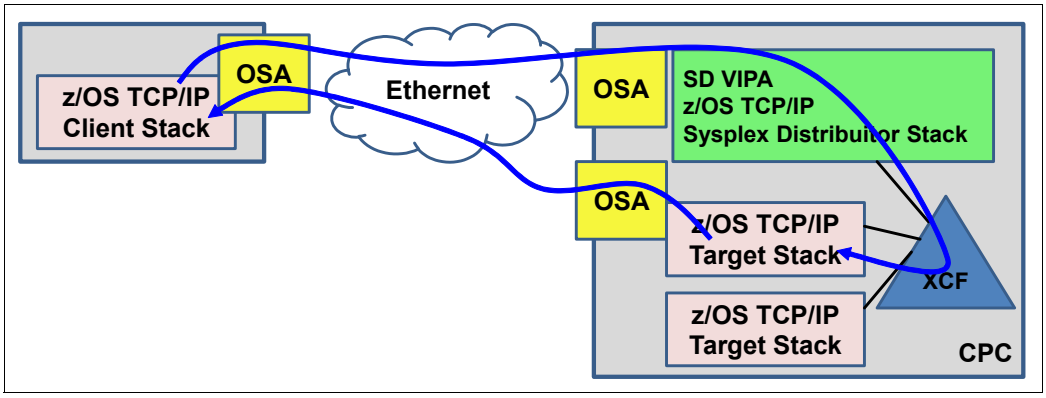


Figure D-9 Sysplex Distributor before RoCE

The traditional Sysplex Distributor has these characteristics:

- All traffic from the client to the target application goes through the Sysplex Distributor TCP/IP stack.
- All traffic from the target application goes directly back to the client by using the TCP/IP routing table on the target TCP/IP stack.

Sysplex Distributor after RoCE

Figure D-10 shows a RoCE Sysplex Distributor:

- The initial connection request goes through the Sysplex Distributor stack.
- The session then flows directly between the client and the target over the RoCE cards.

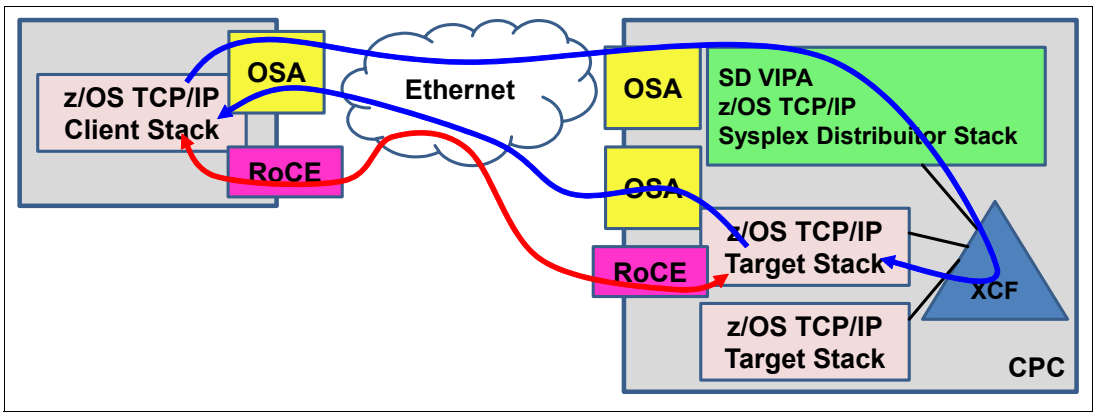


Figure D-10 Sysplex Distributor after RoCE

Note: As with all RoCE Communications, the session end also flows over OSAs.

D.2.11 Enabling SMC-R support in z/OS Communications Server

The following checklist provides a task summary for enabling SMC-R support in z/OS Communications Server. This list assumes that you start with an existing IP configuration for LAN access 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 feature) in a specific LPAR (each stack must define a unique FID value).
- ☐ Start the IPAQENET or IPAQENET6 INTERFACE with CHPIDTYPE OSD:
  - ☐ SMC-R is enabled, by default, for these interface types.
  - ☐ SMC-R is not supported on any other interface types.

  Note: The IPv4 INTERFACE statement (IPAQENET) must also specify an IP subnet mask

- ☐ Repeat in each host (at least two hosts).

Start the TCP/IP traffic and monitor it with Netstat and IBM VTAM® displays.

D.3 Shared Memory Communications - Direct Memory Access

This section describes the new Shared Memory Communication - Direct Memory Access (SMC-D) functions implemented in IBM z14 (Driver level 32), z13 and z13s (Driver Level 27) servers.

Concepts

Co-location of multiple tiers of a workload onto a single IBM Z physical server allows for the exploitation of HiperSockets, 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, so it requires no physical cabling or external network connection to purchase, maintain, or replace. The lack of external components also provides for a secure and low latency network connection because data transfer occurs much like a cross-address-space memory move.

SMC-D maintains the socket-API transparency aspect of SMC-R so that applications that use TCP/IP communications can benefit immediately without requiring any application software or IP topology changes. SMC-D completes the overall Shared Memory Communications solution, providing synergy with SMC-R. Both protocols use shared memory architectural concepts, eliminating TCP/IP processing in the data path, yet preserving TCP/IP Qualities of Service for connection management purposes.

From an operations standpoint, SMC-D is similar to SMC-R. The objective is to provide consistent operations and management tasks for both SMC-D and SMC-R. SMC-D uses a new virtual PCI adapter called Internal Shared Memory (ISM). The ISM Interfaces are associated with IP interfaces (for example HiperSockets or OSA, ISM interfaces do not exist without an IP interface). ISM interfaces are not defined in software. Instead, ISM interfaces are dynamically defined and created, and automatically started and stopped. You do not need to operate (Start or Stop) ISM interfaces. Unlike RoCE, ISM FIDs (PFIDs) are not defined in software. Instead, they are auto-discovered based on their PNet ID.

SMC-R uses RDMA (RoCE), which is based on Queue Pair (QP) technology:

- RC-QPs represent SMC Links (logical point-to-point connection).
- RC-QPs over unique RNICs are logically bound together to form Link Groups (used for HA and load balancing).
- Link Groups (LGs) and Links are provided in many Netstat displays (for operational and various network management tasks).

SMC-D over ISM does not use QPs:

- Links and LGs based on QPs (or other hardware constructs) are not applicable to ISM. So the SMC-D information in the Netstat command displays are related to ISM link information rather than LGs.
- SMC-D protocol (like SMC-R) has a design concept of a “logical point-to-point connection” and therefore preserves the concept of an SMC-D Link (for various reasons that include network administrative purposes).

Note: The SMC-D information in the Netstat command displays is related to ISM link information (not LGs).

D.3.1 Internal Shared Memory technology overview

ISM is a function supported by the z14, z13 and z13s machines. It is the firmware that provides the connectivity for shared memory access between multiple operating systems within the same CPC. It provides the same functionality as SMC-R but without physical adapters like the RoCE card, using instead virtual ISM devices as SMC-R. It is a Hipersocket like function that provides guest-to-guest communications within the same machine. Figure D-11 shows a possible solution using SMC-D only.

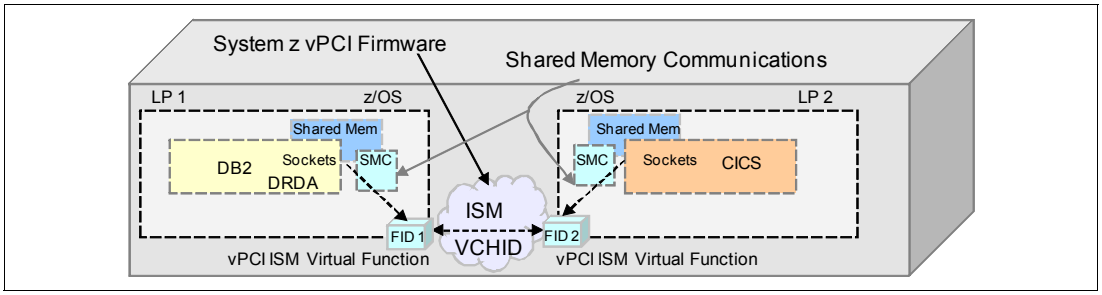


Figure D-11 Connecting two LPARs on the same CPC using SMC-D

Both SMC-D and SMC-R technologies can be used at the same time on the same CPCs.

Figure D-12 shows a fully configured three-tier solution using both SMC-D and SMC-R.

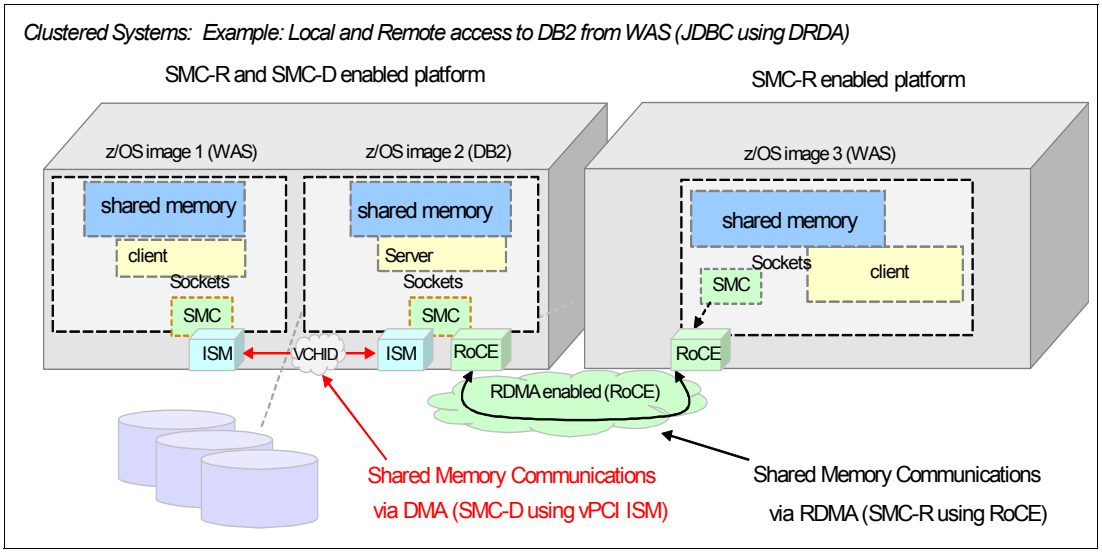


Figure D-12 Clustered systems: Multitier application solution. RDMA, and DMA

D.3.2 SMC-D over Internal Shared Memory

ISM is a virtual channel similar to IQD for HiperSockets. A virtual adapter is created in each OS. Using the SMC protocol, the memory is logically shared. The network is firmware provided. A new device is required to manage that virtual function. SMC is based on a TCP/IP connection and preserves the entire network infrastructure.

SMC-D is a protocol that allows TCP socket applications to transparently use ISM.

SMC-D is a “hybrid” solution as shown in Figure D-13:

- It uses a TCP connection to establish the SMC-D connection.

- The TCP connection can be either through the OSA adapter or IQD HiperSockets
- A TCP option (SMCD) controls switching from TCP to “out of band” SMC-D.
- The SMC-D information is exchanged within the TCP data stream.
- Socket application data is exchanged through ISM (write operations).
- The TCP connection remains to control the SMC-D connection.
- This model preserves many critical existing operational and network management features of TCP/IP.

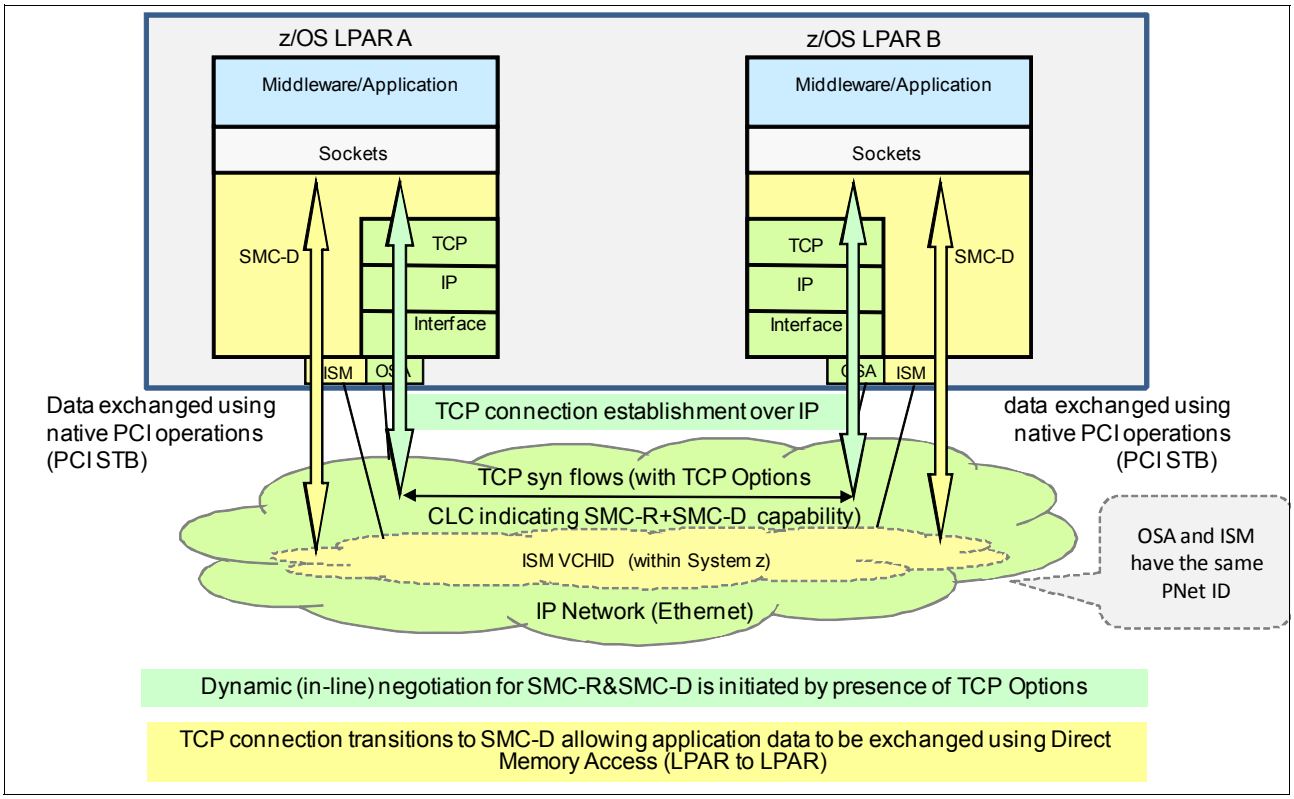


Figure D-13 Dynamic transition from TCP to SMC-D by using two OSA-Express adapters

The hybrid model of SMC-D uses these key existing attributes:

- It follows the standard TCP/IP connection setup.
- The hybrid model switches to ISM (SMC-D) dynamically.
- The TCP connection remains active (idle) and is used to control the SMC-D connection.
- The hybrid model preserves the following critical operational and network management TCP/IP features:
  - Minimal (or zero) IP topology changes
  - Compatibility with TCP connection-level load balancers
  - Preservation of the existing IP security model, such as IP filters, policies, VLANs, and SSL
  - Minimal network administration and management changes
- Host application software is not required to change, so all host application workloads can benefit immediately.

D.3.3 Internal Shared Memory introduction

The IBM z14, IBM z13 (Driver 27) and z13s servers support the ISM virtual PCI function. ISM is a virtual PCI network adapter that enables direct access to shared virtual memory providing a highly optimized network interconnect for 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 existing IBM Z PCIe architecture (that is virtual PCI function / adapter). It introduces a new PCI Function Group and type (ISM virtual PCI). There will be a new virtual adapter.

The system admin, configuration, and operations tasks follow the same process (HCD/IOCDS) as existing PCI functions such as RoCE Express, zEDC Express, and so on. ISM supports dynamic I/O.

ISM Provides adapter virtualization (Virtual Functions) with high scalability:

- It supports up to 32 ISM VCHIDs per CPC (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 8k VFs per CPC), which provide significant scalability.

  Note: There is no concept of a PCI Physical Function to provide virtualization. There is no concept of MACs, or MTU or Frame size.

  - Each ISM VCHID represents a unique and isolated internal network, each having a unique Physical Network ID (PNet IDs are configured in HCD/IOCDS).
- ISM VCHIDs support VLANs, so subdividing a VCHID using virtual LANs is supported.
- ISM provides a Global Identifier (GID) that is internally generated to correspond with each ISM FID.
- ISM is supported by z/VM in pass-through mode (PTF required).

D.3.4 Virtual PCI Function (vPCI Adapter)

Virtual Function ID is defined when PCIe hardware is shared between LPARs. Virtual Function ID has a decimal Virtual Function Identifier (VF=) in the range 1 – nn, where nn is the maximum number of partitions that the PCIe feature supports. For example, the SMC-D ISM feature supports up to 32 partitions, and a zEDC Express feature supports up to 15.

The following basic infrastructure is available:

- zPCI architecture
- RoCE, zEDC, ISM
- zPCI layer in z/OS and Linux for z systems
- vPCI for SD queues

Figure D-14 shows the basic concept vPCI adapters:

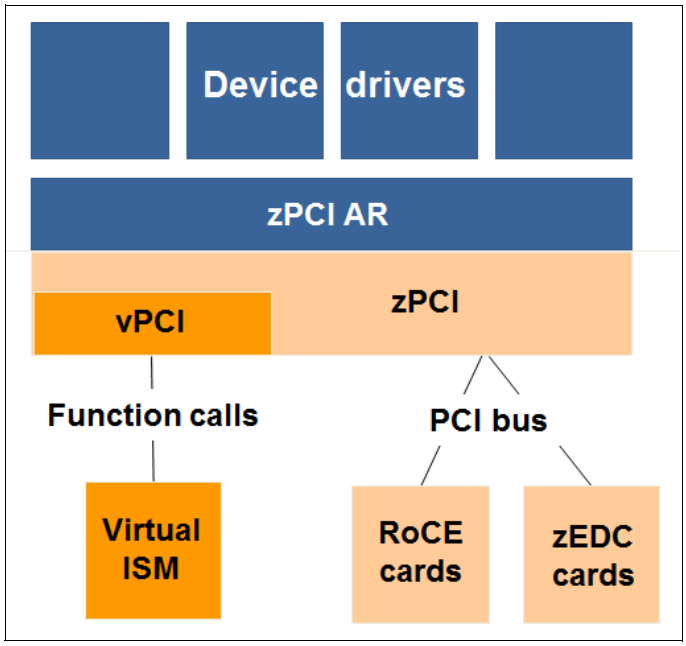


Figure D-14 Concept of vPCI adapter implementation

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 basically unchanged

Figure D-15 shows a SMC-D configuration in which Ethernet provides the connectivity.

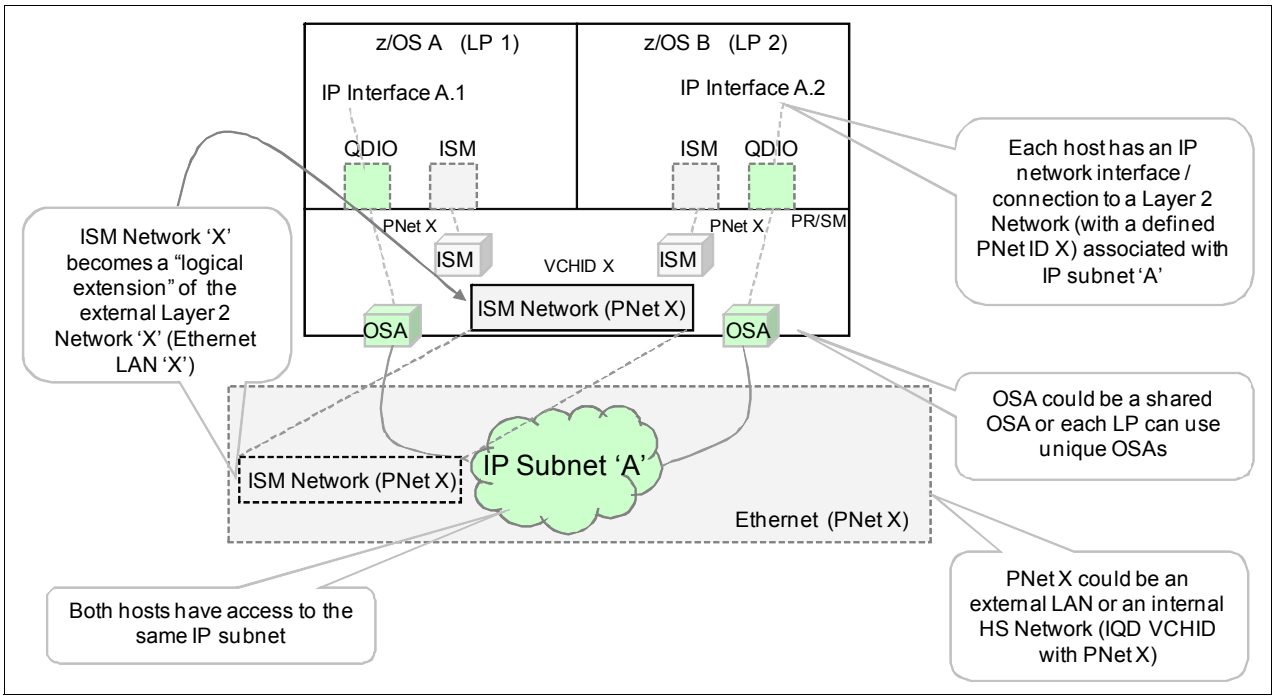


Figure D-15 SMC-D configuration using Ethernet to provide connectivity

Figure D-16 shows a SMC-D configuration in which HiperSockets provide the connectivity.

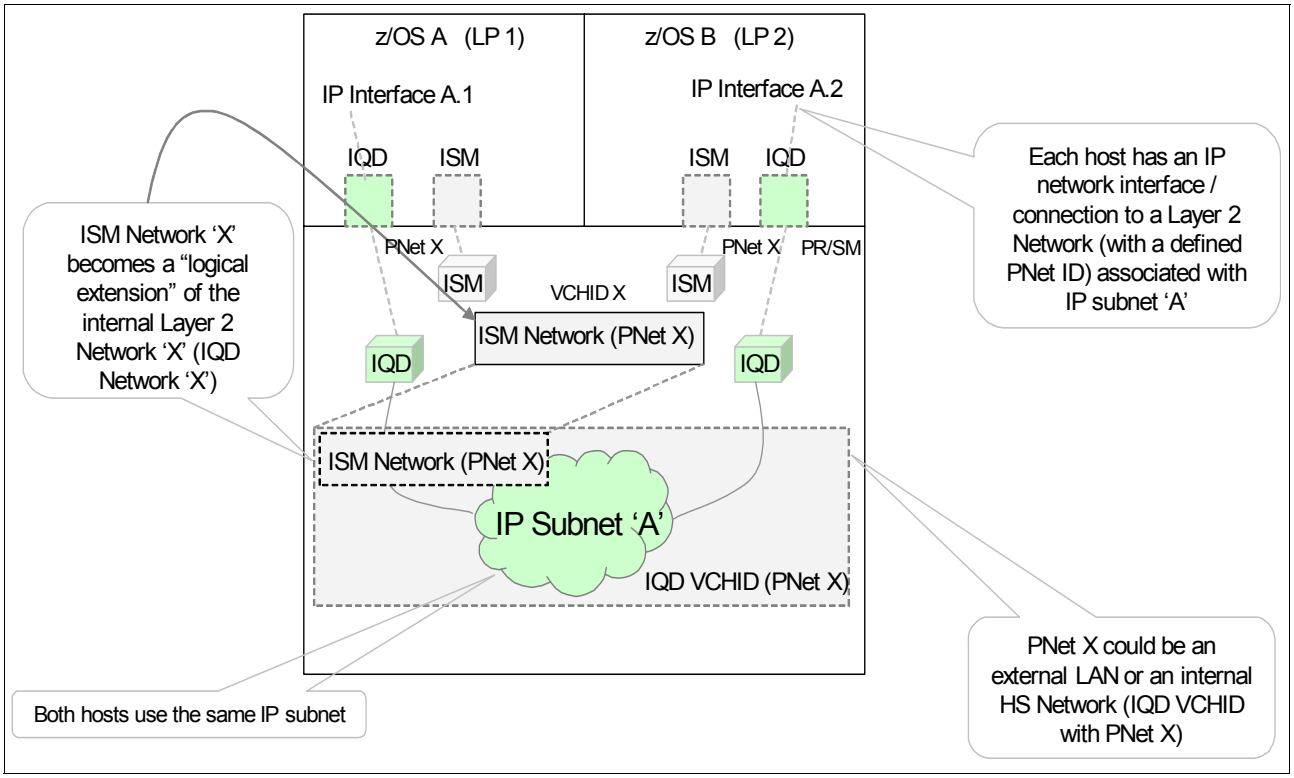


Figure D-16 SMC-D configuration using HiperSockets to provide connectivity

D.3.5 Planning considerations

In the z/OS SMC-D implementation, z/OS uses a single Virtual Function (VF) per ISM PNet. This is true for a single VLAN or for multiple VLANs per PNet. The number of VLANs defined for a given 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 is not affected by the version of the IP (that is, this is true even if both IPv4 and IPv6 are used).

Reasons why z/OS might use additional ISM FIDs:

- IBM supports up to 8 TCP/IP stacks per z/OS LPAR. SMC-D could use up to 8 FIDs or VFs (one per TCP/IP stack).
- IBM supports up to 32 ISM PNet IDs per CEC. Each TCP/IP stack can have access to PNet ID consuming up to 32 FIDs (one VF per PNet ID).

D.3.6 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 using HCD.
4. Enable SMC-D in at least two z/OS instances, which is a single parameter in TCP/IP Global configuration. Both z/OS instances must execute on the same CPC.
5. Review and adjust if / as needed available real memory and fixed memory usage limits (z/OS and CS). SMC requires fixed memory. You might need to review limits and provision additional real memory for z/OS.
6. Review IP topology, VLAN usage considerations, and IP security. For more information, see the SMC-R Security 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, there are numerous updates to these items:
   - Messages (VTAM and TCP stack)
   - Netstat (status, monitoring, and display information)
   - CS diagnostic tools (VIT, Packet trace, CTRACE, and IPCS formatted dumps)

Note: There are no application changes (transparent to Socket applications). There are no required optional operation changes (for example starting or stopping devices).

ISM Functions must be associated with another channel (CHID) of one of these types:

- IQD (a single IQD HiperSockets) channel
- OSD channels

Note: A single ISM PCHID cannot be associated with both (IQD and OSD)

D.3.7 Sample IOCP FUNCTION statements

Example D-1 shows IOCP FUNCTION statements that describe the configuration that defines ISM adapters shared between LPARs on the same CPC as shown in Figure D-17.

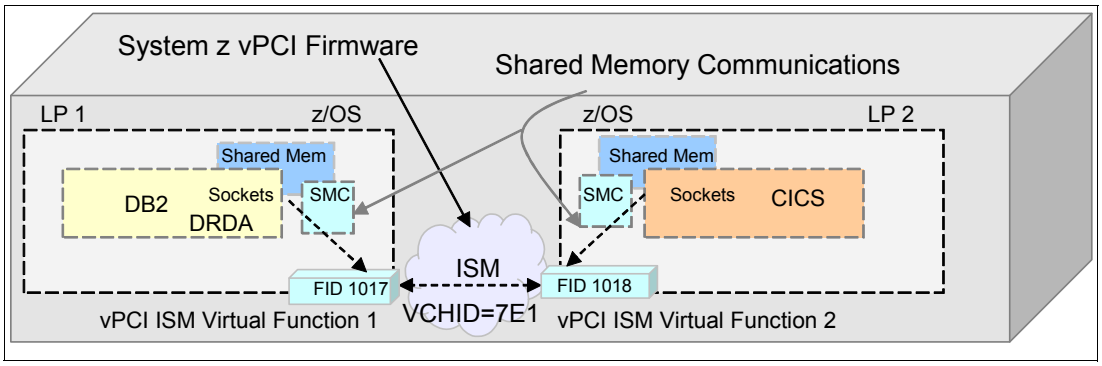


Figure D-17 ISM adapters shared between LPARs

Example D-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
```

Note: On the IOCDS statement, the VCHID is defined as 7E1. In Figure D-17 on page 477, the ISM network “PNET 1” is referenced by the IOCDS VCHID statement. ISM (just like IQD) doesn’t use physical cards or card slots (PCHID), but only logical (firmware) instances that are defined as VCHIDs in IOCDS.

Example D-2 shows a sample IOCP FUNCTION configuration that defines ISM adapters shared between LPSRs and multiple VLANs on the same CPC as shown in Figure D-18.

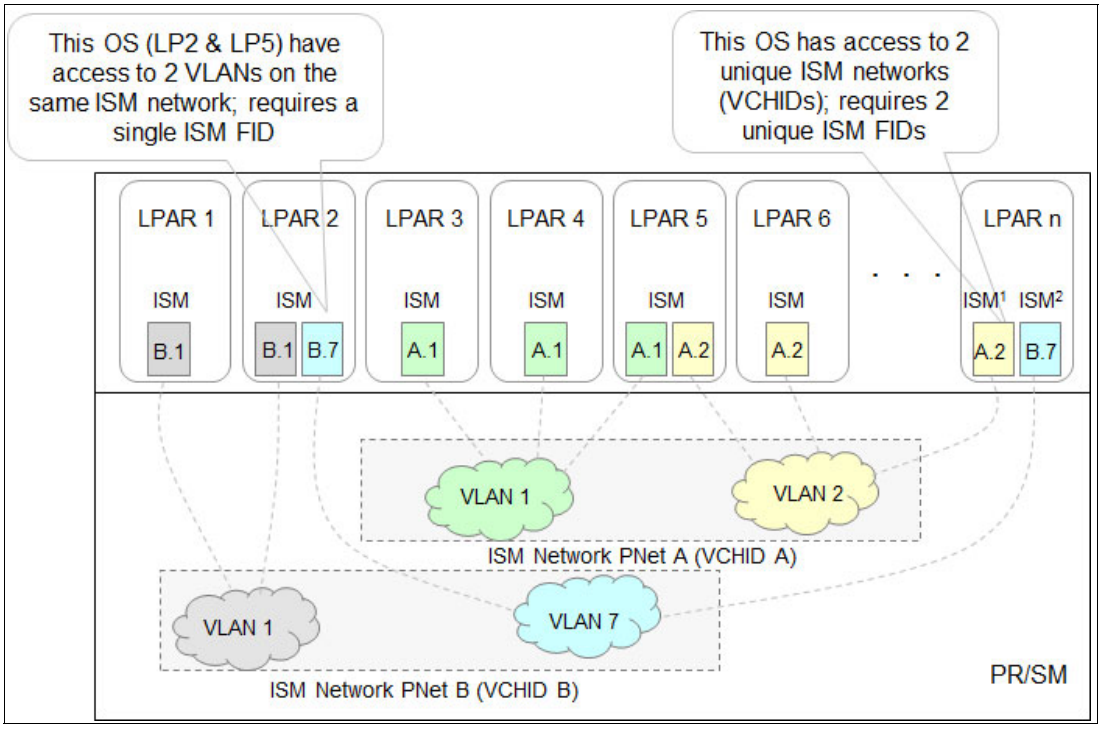


Figure D-18 Multiple LPARs connected through multiple VLANs

Workloads can be logically isolated on separate ISM VCHIDs. Alternatively, workloads can be isolated by exploiting VLANs. The ISM VLAN definitions are inherited from the associated IP network (OSA or HiperSockets).

Example D-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
```

Configuration considerations

The IOCDS (HCD) definitions for ISM PCI VFs are not directly related to the software (SMC-D) exploitation of ISM (that is, the z/OS TCP/IP and SMC-D implementation and usage are not directly related to the I/O definition).

The user defines a list if ISM FIDs (VFs) in IOCDS (HCD), and z/OS dynamically selects an eligible FID based on the required PNet ID. FIDs or VFs are NOT defined in Communications Server for z/OS TCP/IP. Instead, z/OS selects an available FID for a specific PNET. Access to additional VLANs does not require configuration of additional VFs.

Note: Consider over-provisioning the I/O definitions (e.g.consider defining eight FIDs instead of 5.

For native PCI devices, FIDs need to be defined. Each FID in turn will also define a corresponding VF. In terms of operating system administration tasks, the administrator will typically reference FIDs. Usually VFs (and VF numbers) are transparent.

D.3.8 Software exploitation of ISM

ISM enables SMC-D, which provides SMC capability within the CPC (SMC without requiring RoCE hardware/network equipment). Host virtual memory is managed by each OS (similar to SMC-R, logically shared memory) following existing IBM Z PCI I/O translation architecture. Only minor changes required for z/VM guests. An OS can be enabled for both SMC-R and SMC-D. SMC-D is used when both peers are within the same CPC (and ISM PNet and VLAN). After the ISM HCD configuration is complete, SMC-D can be enabled in z/OS with a single TCP/IP parameter (GLOBALCONFIG SMCD). ISM FIDs must be associated with an IP network. The association is accomplished by matching PNet IDs (for example HiperSockets and ISM).

Note: ISM FIDs must be associated with HiperSockets or with an OSA adapter by using a PNet ID. It cannot be associated to both.

D.3.9 SMC-D over ISM prerequisites

SMC-D over ISM has these prerequisites:

- IBM z14, z13s or z13 server (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 only communicate with another z/OS V2.2 instance and peer hosts must be on the same CPC and ISM PNet.
  - SMC-D requires an IP Network with access through OSA or HiperSockets that has a defined PNet ID that matches the ISM PNet ID.
- If running as a z/OS guest under z/VM, z/VM 6.3 with APAR VM65716, including APARs is required for guest access to RoCE (Guest Exploitation only).
- Linux support is planned for a future deliverable.

Table D-2 shows a list of required APARs per z/OS subsystem.

Table D-2 Table with prerequisite APARs for SMC-D enablement

Subsystem	FMID	APAR
IOS	HBB77A0	OA47913
CommServer SNA VTAM	HVT6220	OA48411
CommServer 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 these circumstances:

- Peer hosts are not within the same IP subnet and VLAN
- TCP traffic requires IPSec or the server exploits FRCA.

D.3.10 Enabling SMC-D support in z/OS Communications Server

The new parameter SMCD (see Figure D-19) will be 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. 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 associated with the OSD or HiperSockets. Figure D-19 shows an extract from z/OS Communications Server: IP Configuration Reference.

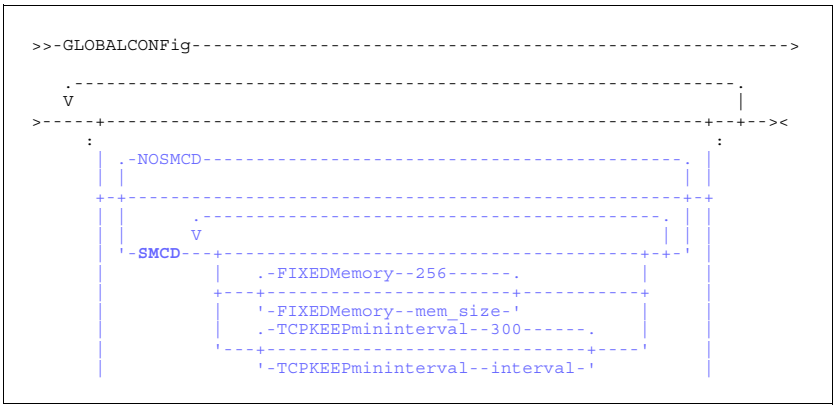


Figure D-19 SMCD parameter in GLOBALCONFIG

D.3.11 SMC-D support overview

SMC-D requires IBM z14, or IBM z13 and IBM z13s servers at driver level 27 or later for support of ISM.

IOCP required level: The required level of IOCP for the z14 server is V5 R4 L1 or later with PTFs. Defining ISM devices on machines other the z14, z13 or z13s servers will not be possible. For more information, see the following manuals:

- *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-2 on page 479) or later is the only supported operating systems for the SMC-D protocol:
  - HCD APAR (OA46010) is required.
  - You cannot roll back to previous z/OS releases.
- z/OS guests under z/VM 6.3 are supported to use SMC-D.
- IBM is working with its Linux distribution partners to include support in future Linux on z Systems distribution releases.

Other ISM considerations

ISM systems have the following limits:

- 32 ISM VCHIDs (in IOCDS / HCD) per CPC. Each IOCDS / HCD VCHID represents a unique internal shared memory network each with a unique Physical Network ID.
- 255 VFs per VCHID (8k VFs per CPC). For example, the maximum number of virtual servers that can communicate over the same ISM VCHID is 255.
- Each ISM VCHID in IOCDS / HCD represents a unique (isolated) internal network, each having a unique Physical Network ID (PNet IDs are configured in HCD/IOCDS).
- ISM VCHIDs support VLANs (can be subdivided into VLANs).
- ISM provides a GID (internally generated) to correspond with each ISM FID.
- MACs (VMACs), MTU, physical ports, and Frame size are all N/A.
- ISM is supported by z/VM (for pass-through guest access to support the new PCI function).

Additional documentation

A configuration example for SMC-D is presented in the Redbook *IBM z/OS V2R2 Communications Server TCP/IP Implementation - Volume 1*, SG24-8360.

E

IBM Dynamic Partition Manager

This appendix contains an introduction to IBM Dynamic Partition Manager (DPM) on IBM Z. It provides a description about how the DPM environment can be set up and managed.

This appendix includes the following sections:

- Introduction to IBM Dynamic Partition Manager
- Reasons to use DPM
- IBM Z servers and DPM
- Setting up the DPM environment

© Copyright IBM Corp. 2017. All rights reserved.

E.1 Introduction to IBM Dynamic Partition Manager

DPM is an IBM Z mode of operation that provides a simplified approach to creating and managing virtualized environments. DPM reduces the barriers to adoption of virtualized environments 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 existing Linux and virtualization skills while getting 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, using a firmware managed environment, that coordinates the physical system resources shared by the partitions[1]. The partitions' resources include processors, memory, network, storage, Crypto, and Accelerators.

DPM provides a new mode of operation for IBM Z servers that provides these 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 additional hypervisor for IBM Z servers. DPM uses the existing 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 are supported:

- Linux virtual servers
- KVM for IBM Z
- z/VM with Linux guests
- Virtual appliances running in Secure Service Container (SSC)

E.2 Reasons to use DPM

DPM mode is targeted at customers (distributed market) with no specific IBM Z skills or knowledge of z/VM who want to implement and use the cloud infrastructure to consolidate and integrate their IT. DPM also eases the management and administration of their Linux environment and workloads.

DPM is of special value for customer segments with these characteristics:

- New IBM Z, or Linux adopters, or distributed driven:
  - Likely not z/VM users
  - Looking for integration into their distributed business models
  - Want ease of migration of distributed environments to IBM Z servers and improve centralized management
- Currently not running on IBM Z servers:
  - No IBM Z skills
  - Want to implement cloud

[1] DPM uses the term "partition", which is the same as logical partition (LPAR).

– Have expectations acquired from another hypervisors' management such as VMware, KVM, and Citrix

E.3 IBM Z servers and DPM

Traditional IBM Z servers are highly virtualized with the goal of maximizing the utilization of compute and I/O (storage and network) resources, and simultaneously lowering the total amount of resources needed for workloads. For decades, virtualization has been embedded in the 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 additional software to run.

PR/SM allows the defining and managing of subsets of the IBM Z resources in LPARs. The LPAR definitions include a number of logical processing units (LPUs), memory, and I/O resources. LPARs can be added, modified, activated, or deactivated in IBM Z platforms by using the traditional Hardware Management Console (HMC) interface.

DPM uses all the capabilities previously mentioned as the foundation for the new user experience. On top of these capabilities, DPM provides an HMC user interface that allows customers to define, implement, and run Linux partitions without requiring deep knowledge of the underlying IBM Z infrastructure management (for example, input/output configuration program (IOCP) or hardware configuration definition (HCD)).

The DPM infrastructure is depicted in Figure E-1.

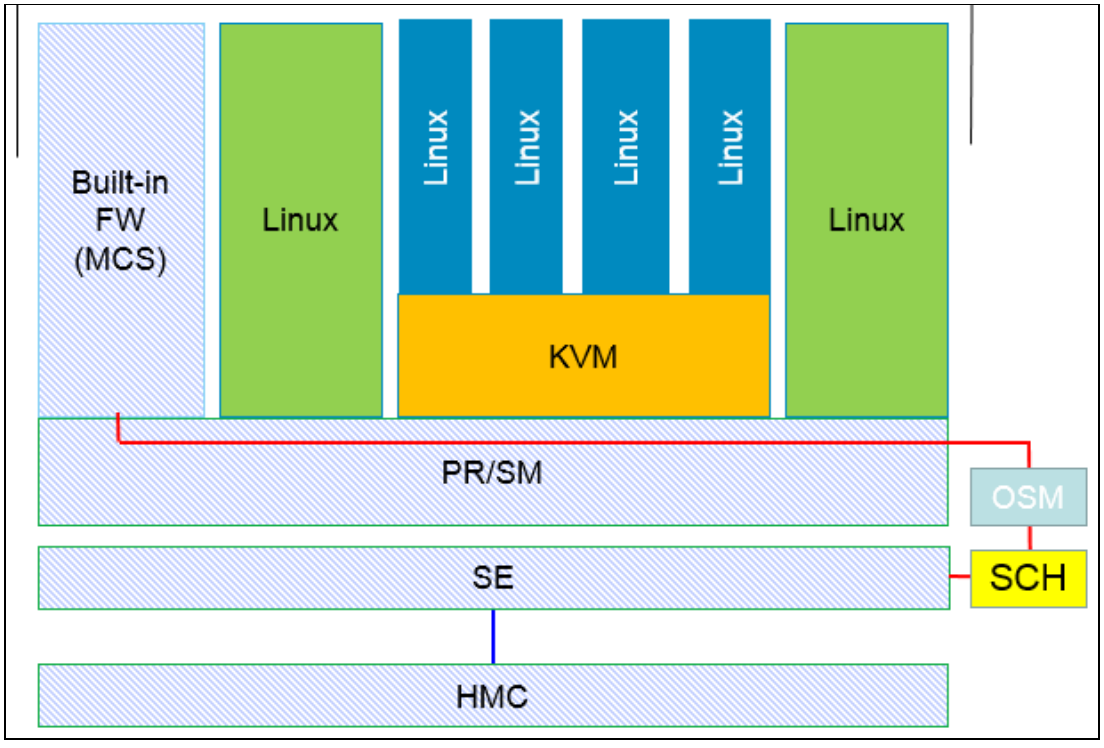


Figure E-1 High-level view of DPM implementation

The firmware partition (similar to the existing PCIe support partitions, also known as master control services (MCS) partition), along with the Support Element (SE), provides services to create and manage the Linux native partitions, or partitions 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 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 IBM Z mode of operation that requires IBM z14, 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 called **Enable Dynamic Partition Manager**, under the **CPC Configuration** menu, in the SE interface, as shown in Figure E-2. The DPM mode of operation 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. After it is selected, a pair of OSA Express 1000BASE-T adapters must be included in the configuration.

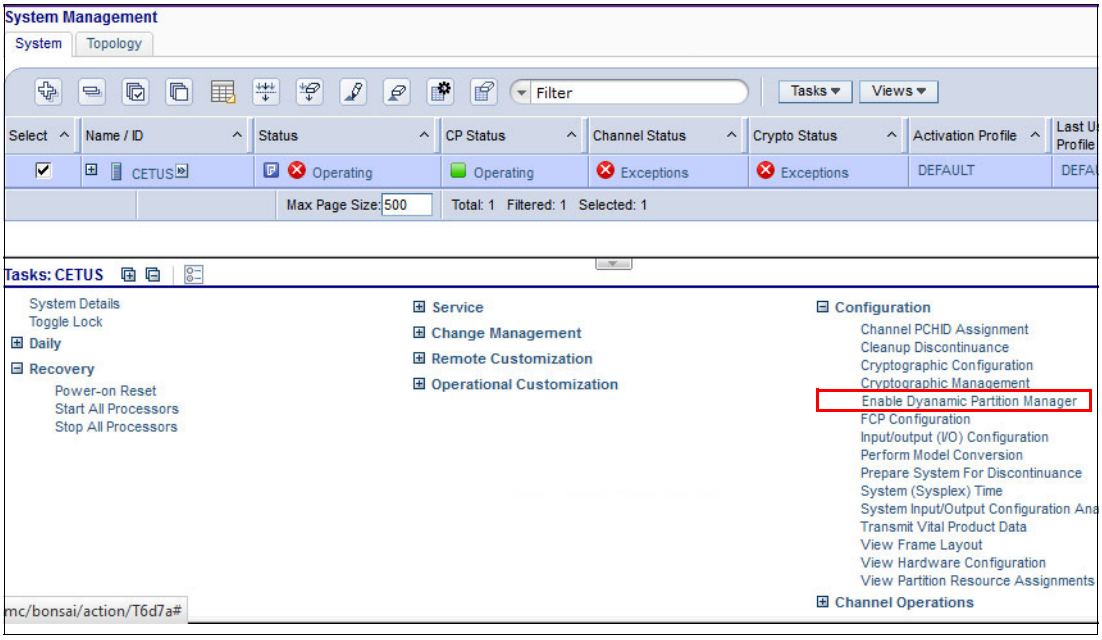


Figure E-2 Enabling DPM mode of operation from the SE CPC configuration options

After the option is selected, a new window is displayed where you enter the two OSA Express 1000BASE-T ports selected and cabled to the System Control Hubs (SCHs) during the Z server installation. This window is shown in Figure E-3.

Note: During the machine installation process, the IBM SSR will connect the two OSA Express4/5/6s 1000BASE-T cables to the SCHs provided ports.

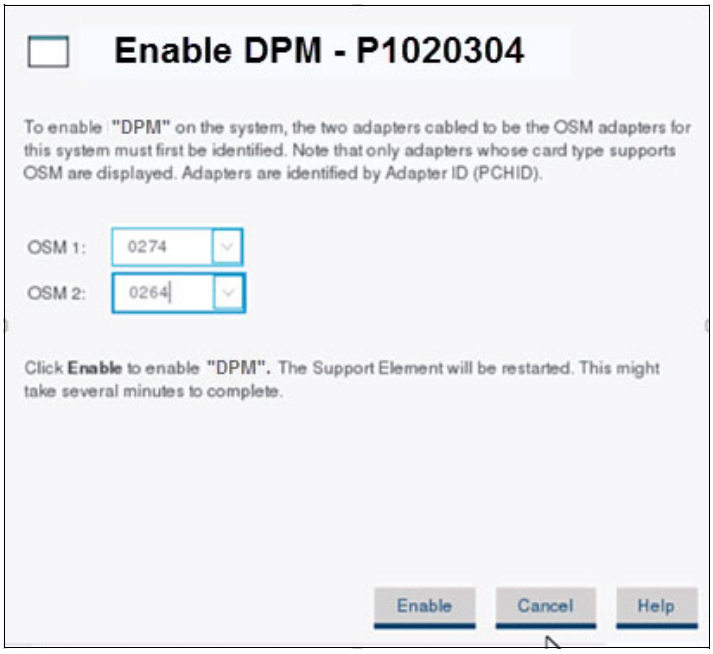


Figure E-3 Entering the OSA ports that will be used by the management network

After entering the OSA adapter port numbers that were previously cabled to the SCHs, click **Enable**. The SE then restarts, and, when finished, the DPM mode becomes active and operational.

Important:

- A CPC in DPM mode cannot be part of an Ensemble managed by Unified Resource Manager. The HMC used to enable the CPC in DPM mode must NOT be an Ensemble HMC (either Primary or Backup Ensemble HMC).
- All definitions that are made for the CPC, if any, before the DPM mode is activated, are saved and can be brought back if 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.

Figure E-4 shows the DPM mode welcome window. The three options at the bottom (**Getting Started**, **Guides**, and **Learn More** have mouse-over functions that either briefly describe their meaning or provide additional functions.

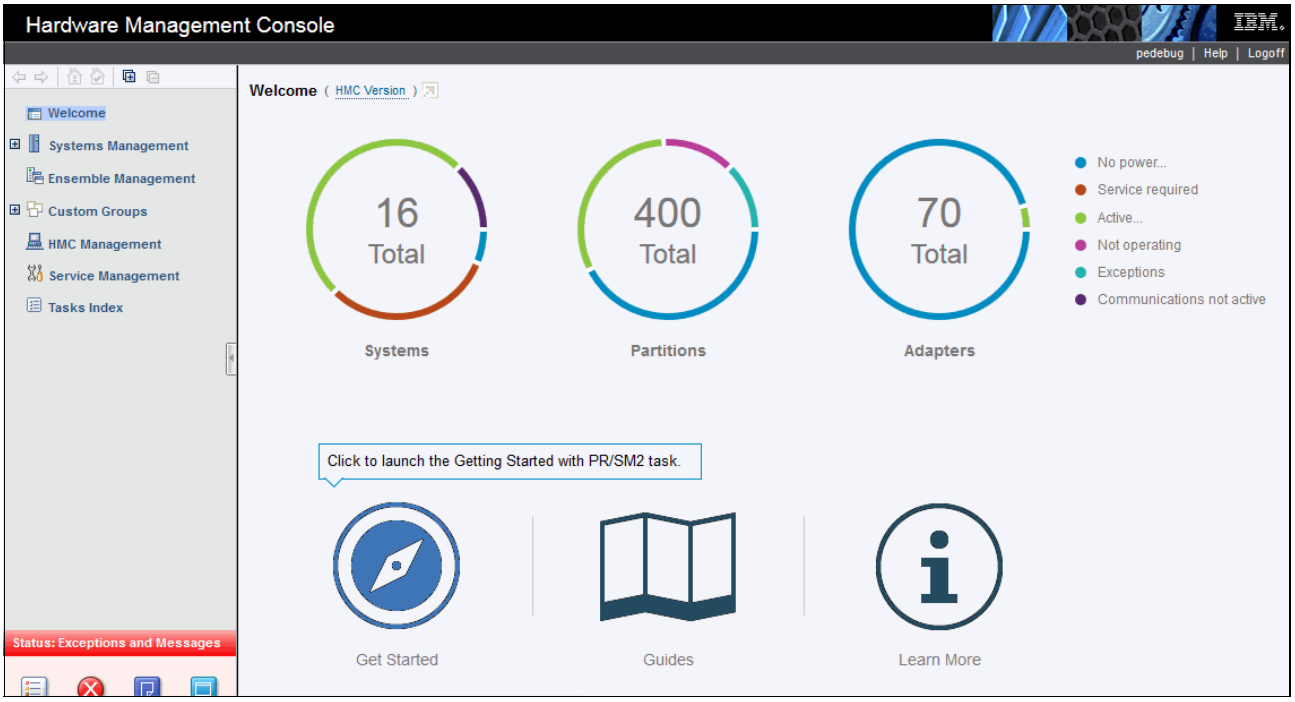


Figure E-4 DPM mode welcome window

The HMC can monitor and control up to 32 IBM Z CPCs. The monitored and controlled CPCs need to be *defined* to the HMC by using the Object Definition task, and adding the CPC object.

The welcome window shown in Figure E-4 only opens when at least one HMC defined CPC is active in DPM mode. Otherwise, the traditional HMC window is presented when you log on to the HMC.

Figure E-5 shows the welcome window from a traditional HMC when none of the defined CPC objects are running in DPM mode.

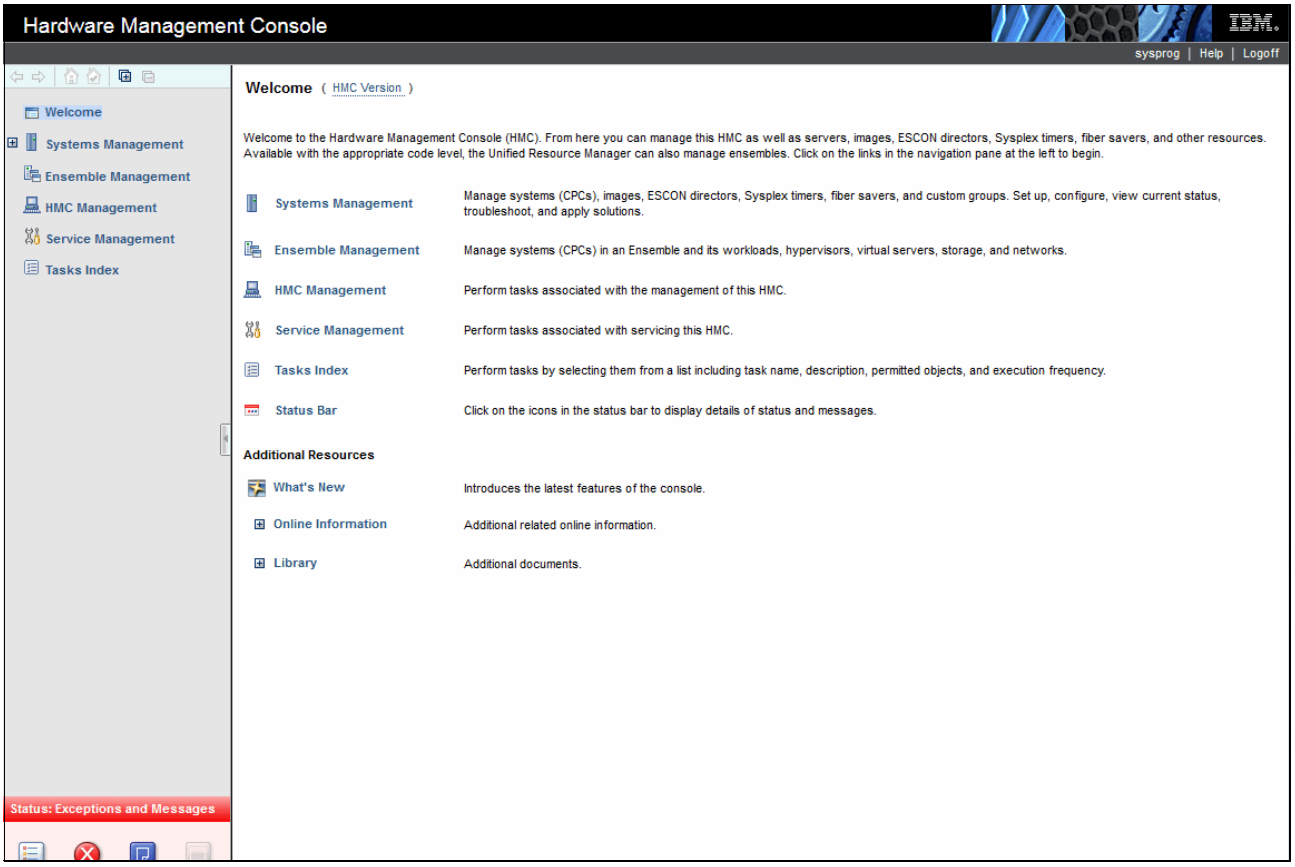


Figure E-5 Traditional HMC Welcome window when no defined CPCs are running in DPM mode

E.4.1 Defining partitions in DPM mode

After the CPC is in DPM mode, the user can choose one of the options provided in the welcome window to learn more about the process of defining partitions, browse the tutorials, or start creating the environment by defining and activating the partitions.

Figure E-6 shows the three options that are presented to the user in the HMC welcome page when at least one CPC is running in DPM mode.

Figure E-6 User Options when the HMC presents the DPM welcome window

The **Guides** option provides tutorials, videos, and information about What's New in DPM. The **Learn More** option covers the application programming interfaces (APIs), and the **Support** option takes the user to the IBM ResourceLink website.

The first option on the left of the window shown on Figure E-6 on page 489, 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.

Figure E-7 on page 491 shows, on the left banner, the HMC create partition wizard steps that 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**: Window where users define partition's network NICs resources.
- **Storage**: Where the storage connectivity host bus adapters (HBAs) are defined.
- **Accelerators:** Partition resources such as zEDC can be added in this window.
- **Cryptos**: Wizard window where users define their cryptographic resources.
- **Boot**: In this wizard 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
  - ISO image
- **Summary:** This wizard 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.

Figure E-7 shows an option at the lower left called **Advanced.** The **Advanced** option allows users to open a window that contains all definitions made for the partition. This window provides additional settings for some of the definitions, such as defining a processor as shared or dedicated, and associating a weight for a partition.

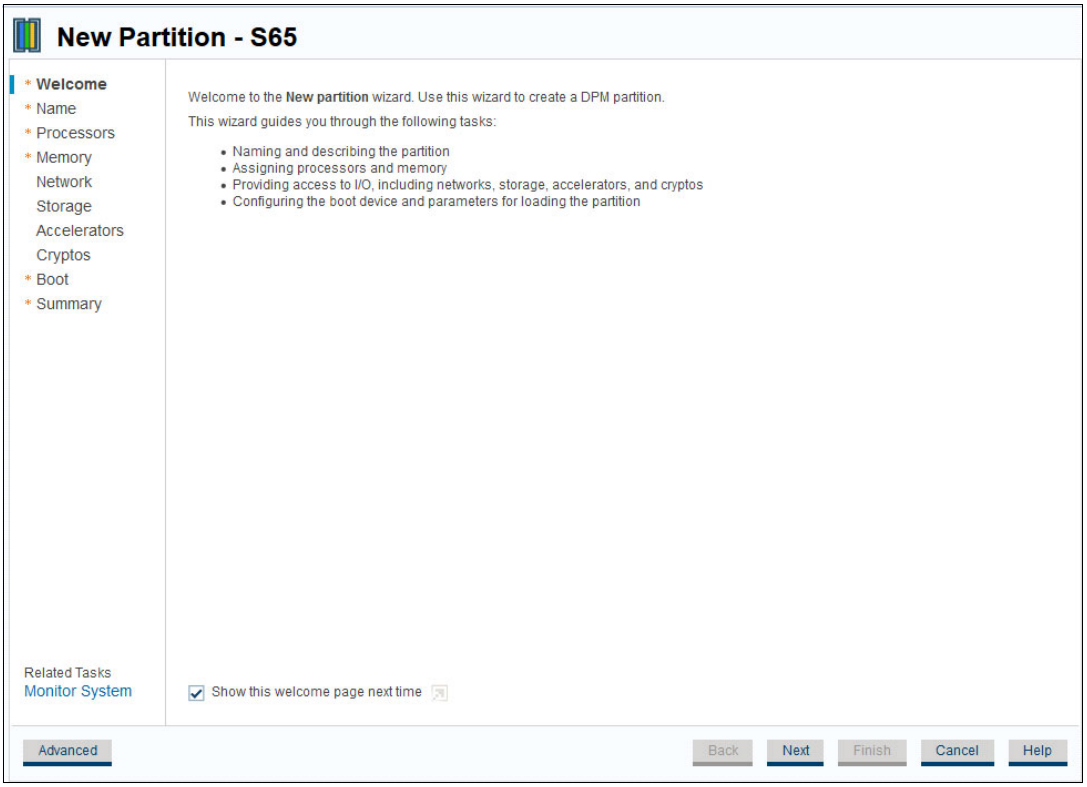


Figure E-7 DPM wizard welcome window options

An important additional 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 utilization in the last 36 hours
- Utilization 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 Systems and KVM on for z/VM as hypervisor, running in a partition.

The new mode, DPM, provides partition lifecycle and dynamic I/O management capabilities using the Hardware Management Console (HMC) for these tasks:

- Create and provision: Creation of new partitions, assignment of processors and memory, configuration of I/O Adapters (Network, FCP Storage, Crypto, and Accelerators)
- Manage the environment: Modification of system resources without disrupting running workloads
- Monitor and troubleshoot the environment: Source identification of system failures, conditions, states, or events that might lead to workload degradation

A CPC can be in either DPM mode or standard PR/SM mode. The mode is enabled before the CPC power-on reset (POR).

DPM mode requires two OSA-Express 1000BASE-T Ethernet features for primary and backup connectivity (OSA-Express4/5/6S 1000BASE-T Ethernet), along with associated cabling (HW for DPM FC0016).

F

IBM zEnterprise Data Compression Express

This appendix briefly describes the optional IBM zEnterprise Data Compression (zEDC) Express feature of the z14, z13, z13s, zEC12, and IBM zBC12 servers.

This appendix includes the following sections:

- Overview
- zEDC Express
- Software support

© Copyright IBM Corp. 2017. All rights reserved.

F.1 Overview

The growth of data that needs to be captured, transferred, and stored for large periods of time is not relenting. In addition, software-implemented compression algorithms can be costly in terms of processor resources and storage costs.

zEDC Express, an optional feature available for 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 previously available on the IBM Z server.

Using the zEDC Express feature with the z/OS V2R1 zEnterprise Data Compression acceleration capability (or later releases) is designed to deliver an integrated solution. 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 exclusively on the Peripheral Component Interconnect Express (PCIe) I/O drawer. Between one and eight features can be installed on the system. There is one PCIe adapter/compression coprocessor per feature, which implements compression as defined by RFC1951 (DEFLATE).

A zEDC Express feature can be shared by up to 15 logical partitions (LPARs) on the same CPC.

Adapter support for zEDC is provided by Resource Group code that runs on the system integrated firmware processor (IFP). The recommended high availability configuration per server is four features. This configuration provides continuous availability during concurrent update. For resilience, the z14 system always has four independent RGs on the system, sharing the IFP. Install a minimum of two zEDC features, for resilience and throughput.

Figure F-1 illustrates the PCIe I/O drawer structure and the relationships among card slots, domains, and resource groups.

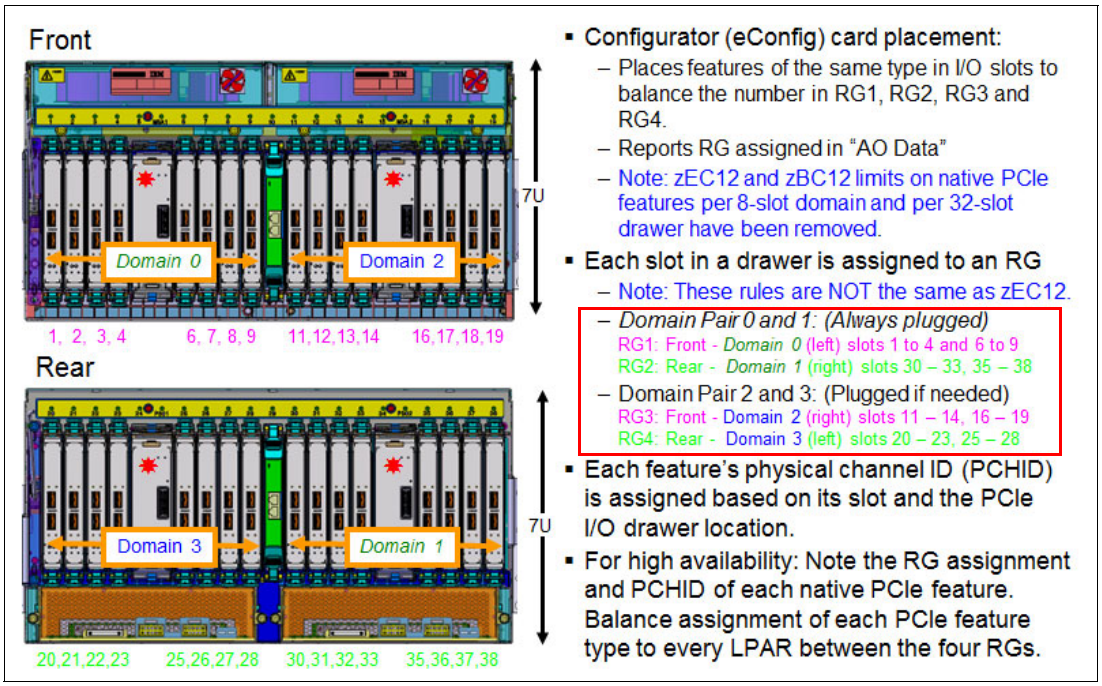


Figure F-1 Relationships among PCIe I/O drawer card slots, I/O domains, and resource groups

F.3 Software support

Support of the zEDC Express function is provided by z/OS V2R1 zEnterprise Data Compression or later for both data compression and decompression. Support for data recovery (decompression) in the case that zEDC is not installed, or installed but not available, is provided through software in z/OS V2R2, V2R1, V1R13, and V1R12 with the appropriate program temporary fixes (PTFs). Software decompression is slow and can use considerable processor resources. Therefore, it is not suggested for production environments.

Refer to the appropriate fixcat for SMP/E to install prerequisite PTFs. The fixcat function is explained here:

`http://www.ibm.com/systems/z/os/zos/features/smpe/fix-category.html`

A specific fix category 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, provide optimized cross-platform exchange of data, and provide higher write rates for SMF data. For more information, see:

`http://www.vm.ibm.com/zvm630/apars.html`

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 (zBNA)

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 alternate processors.

zBNA can be used to analyze client-provided System Management Facilities (SMF) records to identify jobs and data sets that are candidates for zEDC compression, across a specified time window, typically a batch window. zBNA is able to generate lists of data sets by job:

- Those jobs that already perform hardware compression and might be candidates for zEDC
- Those jobs that might be zEDC candidates, but are not in extended format

Therefore, zBNA can help you estimate the use of zEDC features and help determine the number of features needed:

- IBM Employees can obtain zBNA and other CPS tools by using the IBM intranet:

  `http://w3.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/PRS5126`
- IBM Business Partners can obtain zBNA and other CPS tools by using the Internet:

  `https://www.ibm.com/partnerworld/wps/servlet/mem/ContentHandler/tech_PRS5133`
- IBM clients can obtain zBNA and other CPS tools by using the Internet:

  `http://www.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/PRS5132`

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this book.

IBM Redbooks

The following IBM Redbooks publications provide additional information about the topic in this document. Note that some publications referenced in this list might be available in softcopy only.

- *IBM z14 Technical Introduction*, SG24-8450
- *IBM z Systems Connectivity Handbook*, SG24-5444
- *IBM z Systems Functional Matrix*, REDP-5157
- *IBM z13 Configuration Setup*, SG24-8260
- *z Systems Simultaneous Multithreading Revolution*, REDP-5144
- *z/OS Infrastructure Optimization using Large Memory*, REDP-5146

You can search for, view, download or order these documents and other Redbooks, Redpapers, Web Docs, draft and additional materials, at the following website:

ibm.com/redbooks

Other publications

These publications are also relevant as further information sources:

- *Capacity on Demand User's Guide,* SC28-6985
- *Installation Manual for Physical Planning,* GC28-6965
- *PR/SM Planning Guide,* SB10-7169
- *IOCP Users Guide*, SB10-7163

Online resources

These 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

© Copyright IBM Corp. 2017. All rights reserved.

Help from IBM

IBM Support and downloads

ibm.com/support

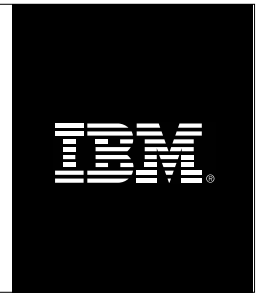
IBM Global Services

ibm.com/services

To determine the spine width of a book, you divide the paper PPI into the number of pages in the book. An example is a 250 page book using Plainfield opaque 50# smooth which has a PPI of 526. Divided 250 by 526 which equals a spine width of .4752". In this case, you would use the .5” spine. Now select the Spine width for the book and hide the others: **Special>Conditional Text>Show/Hide>SpineSize(-->Hide:)>Set** . Move the changed Conditional text settings to all files in your book by opening the book file with the spine.fm still open and **File>Import>Formats** the Conditional Text Settings (ONLY!) to the book files.

SG24-8451-00

ISBN

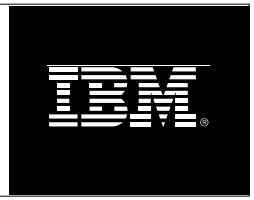


(1.5” spine)
1.5”<-> 1.998”
789 <->1051 pages

IBM z14 Technical Guide

SG24-8451-00

ISBN



(1.0” spine)
0.875”<->1.498”
460 <-> 788 pages

Redbooks

SG24-8451-00

ISBN

(0.5” spine)
0.475”<->0.873”
250 <-> 459 pages

Redbooks

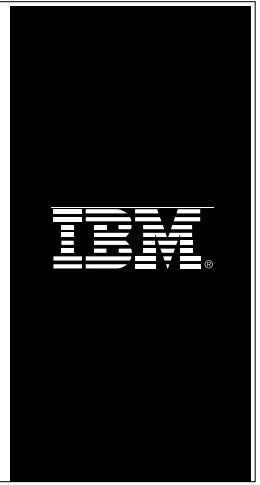
(0.2”spine)
0.17”<->0.473”
90<->249 pages

(0.1”spine)
0.1”<->0.169”
53<->89 pages

To determine the spine width of a book, you divide the paper PPI into the number of pages in the book. An example is a 250 page book using Plainfield opaque 50# smooth which has a PPI of 526. Divided 250 by 526 which equals a spine width of .4752". In this case, you would use the .5” spine. Now select the Spine width for the book and hide the others: **Special>Conditional Text>Show/Hide>SpineSize(-->Hide:)>Set** . Move the changed Conditional text settings to all files in your book by opening the book file with the spine.fm still open and **File>Import>Formats** the Conditional Text Settings (ONLY!) to the book files.

SG24-8451-00

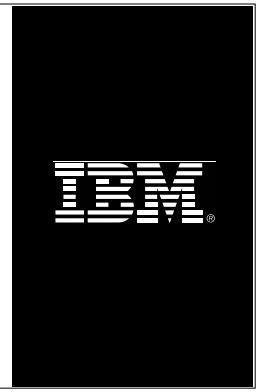
ISBN



(2.5” spine)
2.5”<->nnn.n”
1315<-> nnnn pages

SG24-8451-00

ISBN



(2.0” spine)
2.0” <-> 2.498”
1052 <-> 1314 pages

SG24-8451-00

ISBN

Printed in U.S.A.

Get connected