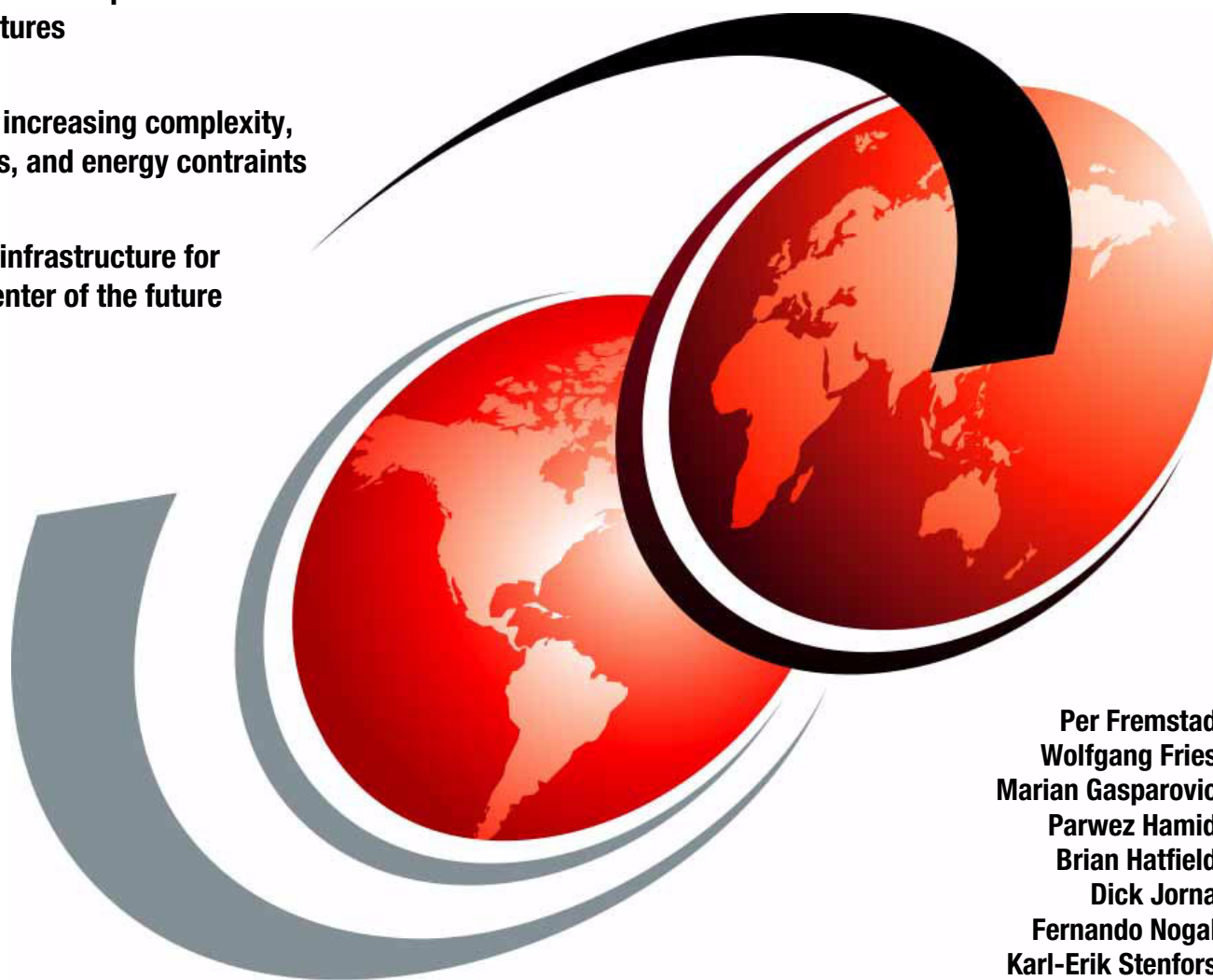


IBM System z10 Enterprise Class Technical Guide

Describes the Enterprise Class server and related features

Addresses increasing complexity, rising costs, and energy constraints

Discusses infrastructure for the data center of the future



Per Fremstad
Wolfgang Fries
Marian Gasparovic
Parwez Hamid
Brian Hatfield
Dick Jorna
Fernando Nogal
Karl-Erik Stenfors

Redbooks



International Technical Support Organization

IBM System z10 Enterprise Class Technical Guide

November 2009

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

Third Edition (November 2009)

This edition applies to the IBM System z10 Enterprise Class server, as described in IBM United States Hardware Announcement 108-794, dated October 21, 2008.

© Copyright International Business Machines Corporation 2008, 2009. All rights reserved.

Note to U.S. Government Users Restricted Rights -- Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Contents

Notices	xi
Trademarks	xii
Preface	xv
The team who wrote this book	xv
Become a published author	xvii
Comments welcome	xvii
Chapter 1. Introducing the System z10 Enterprise Class	1
1.1 Wanted: an infrastructure (r)evolution	3
1.1.1 Simplified	4
1.1.2 Shared	4
1.1.3 Dynamic	5
1.1.4 z10 at the core of a dynamic infrastructure	6
1.1.5 Storage is part of the System z10 stack	6
1.2 System z10 EC highlights	7
1.3 System z10 EC Models	9
1.3.1 Model upgrade paths	10
1.3.2 Concurrent processing unit conversions	10
1.4 System functions and features	10
1.4.1 Processor	12
1.4.2 Memory subsystem	13
1.4.3 Central processor complex cage	13
1.4.4 I/O connectivity	14
1.4.5 I/O subsystems	14
1.4.6 Cryptography	16
1.4.7 Parallel Sysplex support	17
1.4.8 Reliability, availability, and serviceability	18
1.5 Performance	19
1.6 Operating systems and software	20
Chapter 2. Hardware components	23
2.1 Frames and cages	24
2.1.1 Frame A	24
2.1.2 Frame Z	26
2.1.3 I/O cages	26
2.2 Book concept	27
2.2.1 Book power	28
2.2.2 Cooling	28
2.3 Multi-Chip Module	30
2.4 Processing units and storage control chips	31
2.4.1 PU chip	31
2.4.2 Processing unit (core)	32
2.4.3 SC chip	34
2.5 Memory	35
2.5.1 Memory RAS	38
2.5.2 Memory upgrades	38
2.5.3 Book replacement and memory	39
2.5.4 Flexible memory option	39

2.5.5 Plan-ahead memory	39
2.6 Connectivity.	41
2.6.1 Redundant I/O interconnect	44
2.6.2 Enhanced book availability	44
2.6.3 Book upgrade	45
2.7 Model configurations	45
2.7.1 Upgrades	47
2.7.2 PU characterization.	48
2.7.3 Concurrent PU conversions	48
2.7.4 Model capacity identifier	49
2.7.5 Model capacity identifier and MSU values	50
2.7.6 Capacity Backup	51
2.7.7 On/Off Capacity on Demand and CPs	53
2.8 Summary of z10 EC structure	54
Chapter 3. System design	57
3.1 Design highlights.	58
3.2 Book design	59
3.2.1 Book interconnect topology.	64
3.2.2 System control	65
3.3 Processing unit	66
3.3.1 Superscalar processor	67
3.3.2 Compression unit on a chip	67
3.3.3 CP Assist for Cryptographic Function	68
3.3.4 Decimal floating point accelerator.	69
3.3.5 Processor error detection and recovery	69
3.3.6 Branch prediction	70
3.3.7 Wild branch.	70
3.3.8 IEEE floating point	71
3.3.9 Translation look-aside buffer.	71
3.3.10 Instruction fetching, decode, and grouping	71
3.3.11 Extended translation facility	72
3.3.12 Instruction set extensions	72
3.4 Processing unit functions	72
3.4.1 Central processors	74
3.4.2 Integrated Facility for Linux.	75
3.4.3 Internal Coupling Facilities	76
3.4.4 System z10 Application Assist Processors.	77
3.4.5 System z10 Integrated Information Processor	81
3.4.6 zAAP on zIIP capability.	83
3.4.7 System assist processors	83
3.4.8 Reserved processors	84
3.4.9 Processor unit characterization.	85
3.4.10 Transparent CP, IFL, ICF, zAAP, zIIP, and SAP sparing	85
3.4.11 Dynamic SAP sparing and reassignment	86
3.4.12 Increased flexibility with z/VM-mode partitions	86
3.5 Memory design	87
3.5.1 Central storage	89
3.5.2 Expanded storage.	89
3.5.3 Hardware system area	90
3.6 Logical partitioning	90
3.6.1 Storage operations	96
3.6.2 Reserved storage	99

3.6.3	Logical partition storage granularity	100
3.6.4	LPAR dynamic storage reconfiguration.	100
3.7	Intelligent Resource Director.	100
3.8	Clustering technology	102
Chapter 4. I/O system structure		105
4.1	Introduction	106
4.1.1	InfiniBand advantages	106
4.1.2	Data, signalling, and link rates	107
4.2	I/O system overview	107
4.2.1	Characteristics	108
4.2.2	Summary of supported I/O features	108
4.3	I/O cages.	109
4.4	Fanouts	111
4.4.1	HCA2-C fanout	112
4.4.2	HCA2-O fanout	112
4.4.3	HCA2-O LR fanout	113
4.4.4	MBA fanout	114
4.4.5	Fanout considerations.	115
4.4.6	Fanout summary	119
4.5	I/O feature cards	120
4.5.1	I/O feature card types	120
4.5.2	PCHID report	121
4.6	Connectivity.	123
4.6.1	I/O feature support and configuration rules.	124
4.6.2	ESCON channels	127
4.6.3	FICON channels	128
4.6.4	OSA-Express3	136
4.6.5	OSA-Express2	139
4.6.6	Open Systems Adapter selected functions.	141
4.6.7	HiperSockets.	145
4.7	Parallel Sysplex connectivity.	146
4.7.1	Coupling links	148
4.7.2	External time reference.	154
4.7.3	Cryptographic feature	154
Chapter 5. Channel subsystem		157
5.1	Channel subsystem.	158
5.1.1	CSS elements	159
5.1.2	Multiple CSSs concept	160
5.1.3	Multiple CSSs structure.	160
5.1.4	Logical partition name and identification.	161
5.1.5	Physical channel ID	162
5.1.6	Multiple subchannel sets.	163
5.1.7	Multiple CSS construct	166
5.1.8	Adapter ID.	166
5.1.9	Channel spanning	167
5.1.10	Summary of CSS-related numbers.	169

5.2 I/O Configuration management	169
5.3 System-initiated CHPID reconfiguration	170
5.4 Multipath initial program load	170
Chapter 6. Cryptography	171
6.1 Cryptographic synchronous functions	172
6.2 Cryptographic asynchronous functions	173
6.2.1 Secure key functions.	174
6.2.2 Other key functions	175
6.2.3 Cryptographic feature codes	176
6.3 CP Assist for Cryptographic Function	177
6.4 Crypto Express2	178
6.4.1 Crypto Express2 coprocessor	179
6.4.2 Crypto Express2 accelerator	180
6.4.3 Configuration rules	181
6.5 Crypto Express3	182
6.6 TKE workstation feature	184
6.7 Cryptographic functions comparison	186
6.8 Software support	187
Chapter 7. Software support	189
7.1 Operating systems summary	190
7.2 Support by operating system	190
7.2.1 z/OS	190
7.2.2 z/VM	191
7.2.3 z/VSE	191
7.2.4 Linux on System z	191
7.2.5 TPF and z/TPF	192
7.2.6 z10 EC functions support summary	192
7.3 Support by function	200
7.3.1 Single system image	201
7.3.2 zAAP on zIIP capability	202
7.3.3 Maximum main storage size	203
7.3.4 Large page support	203
7.3.5 Guest support for execute-extensions facility	204
7.3.6 Hardware decimal floating point	204
7.3.7 Up to 60 logical partitions	205
7.3.8 Separate LPAR management of PUs	205
7.3.9 Dynamic LPAR memory upgrade	205
7.3.10 Capacity Provisioning Manager	206
7.3.11 Dynamic PU exploitation	206
7.3.12 HiperDispatch	206
7.3.13 The 63.75 K subchannels	207
7.3.14 Multiple subchannel sets	207
7.3.15 MIDAW facility	208
7.3.16 Enhanced CPACF	208
7.3.17 HiperSockets multiple write facility	208
7.3.18 HiperSockets IPv6	208
7.3.19 HiperSockets Layer 2 support	209
7.3.20 High performance FICON for System z10	209
7.3.21 FCP provides increased performance	210
7.3.22 Request node identification data	210
7.3.23 FICON link incident reporting	211

7.3.24	N_Port ID virtualization	211
7.3.25	VLAN management enhancements	211
7.3.26	OSA-Express3 10 Gigabit Ethernet LR and SR	212
7.3.27	OSA-Express3 Gigabit Ethernet LX and SX	212
7.3.28	OSA-Express3 1000BASE-T Ethernet	213
7.3.29	GARP VLAN Registration Protocol	214
7.3.30	OSA-Express3 and OSA-Express2 OSN support	214
7.3.31	OSA-Express2 1000BASE-T Ethernet	215
7.3.32	OSA-Express2 10 Gigabit Ethernet LR	215
7.3.33	Program directed re-IPL	216
7.3.34	Coupling over InfiniBand	216
7.3.35	Dynamic I/O support for InfiniBand CHPIDs	216
7.4	Cryptographic support	217
7.4.1	CP Assist for Cryptographic Function	217
7.4.2	Crypto Express3 and Crypto Express2	218
7.4.3	Web deliverables	218
7.4.4	z/OS ICSF FMIDs	218
7.4.5	ICSF migration considerations	221
7.5	z/OS migration considerations	221
7.5.1	General recommendations	221
7.5.2	HCD	221
7.5.3	InfiniBand coupling links	221
7.5.4	Large page support	222
7.5.5	HiperDispatch	222
7.5.6	Capacity Provisioning Manager	222
7.5.7	Decimal floating point and z/OS XL C/C++ considerations	223
7.6	Coupling facility and CFCC considerations	223
7.7	MIDAW facility	224
7.7.1	MIDAW technical description	225
7.7.2	Extended format data sets	227
7.7.3	Performance benefits	228
7.8	IOCP	228
7.9	Worldwide portname (WWPN) prediction tool	228
7.10	ICKDSF	229
7.11	Software licensing considerations	229
7.11.1	Workload License Charge	230
7.11.2	System z New Application License Charge	231
7.11.3	Select Application License Charge	231
7.11.4	Midrange Workload License Charge	232
7.11.5	System z International Licensing Agreement	232
7.12	References	232
Chapter 8.	System upgrades	233
8.1	Upgrade types	234
8.1.1	Terminology related to CoD for System z10 servers	235
8.1.2	Permanent upgrades	237
8.1.3	Temporary upgrades	238
8.2	Concurrent upgrades	239
8.2.1	Model upgrades	240
8.2.2	Customer Initiated Upgrade facility	241
8.2.3	Summary of concurrent upgrade functions	245
8.3	MES upgrades	246
8.3.1	MES upgrade for processors	247

8.3.2	MES upgrade for memory	249
8.3.3	MES upgrades for I/O	250
8.3.4	Plan-ahead concurrent conditioning	251
8.4	Permanent upgrade through the CIU facility	251
8.4.1	Ordering	253
8.4.2	Retrieval and activation	254
8.5	On/Off Capacity on Demand	255
8.5.1	Overview	255
8.5.2	Ordering	256
8.5.3	On/Off CoD testing	260
8.5.4	Activation and deactivation	261
8.5.5	Termination	262
8.5.6	z/OS capacity provisioning	263
8.6	Capacity for Planned Event	266
8.7	Capacity Backup	268
8.7.1	Ordering	269
8.7.2	CBU activation and deactivation	270
8.7.3	Automatic CBU enablement for GDPS	272
8.8	Nondisruptive upgrades	273
8.9	Summary of Capacity on Demand offerings	278
Chapter 9. RAS	279
9.1	z10 Availability characteristics	280
9.2	z10 RAS functions	281
9.3	z10 Enhanced book availability	283
9.3.1	Planning considerations	284
9.3.2	Enhanced book availability processing	286
9.4	z10 Enhanced driver maintenance	292
Chapter 10. Environmental requirements	295
10.1	z10 Power and cooling	296
10.1.1	Power consumption	296
10.1.2	Internal Battery Feature	297
10.1.3	Emergency power-off	297
10.1.4	Cooling requirements	298
10.2	z10 Physical specifications	298
10.2.1	Weights	298
10.2.2	Dimensions	299
10.3	Power estimation tool	300
Chapter 11. Hardware Management Console	303
11.1	HMC and SE introduction	304
11.2	HMC and SE connectivity	304
11.3	Remote Support Facility	308
11.4	HMC remote operations	308
11.5	z10 EC HMC and SE key capabilities	309
11.5.1	CPC management	310
11.5.2	LPAR management	310
11.5.3	Operating system communication	311
11.5.4	SE access	311
11.5.5	Monitoring	311
11.5.6	HMC Console Messenger	312
11.5.7	Capacity on Demand support	313
11.5.8	Server Time Protocol support	314

11.5.9 NTP client/server support on HMC	314
11.5.10 System Input/Output Configuration Analyzer on the SE/HMC	315
11.5.11 Network Analysis Tool for SE Communication	316
11.5.12 Automated operations.	316
11.5.13 Cryptographic support.	316
11.5.14 z/VM virtual machine management.	316
11.5.15 Installation support for z/VM using the HMC.	317
Related publications	319
IBM Redbooks publications	319
Other publications	319
Online resources	320
How to get Redbooks publications.	321
Help from IBM	321
Index	323

Notices

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

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information about 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 give you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785 U.S.A.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

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

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

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

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

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


COPYRIGHT LICENSE:

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

Trademarks

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

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

CICS®	HiperSockets™	Sysplex Timer®
Cool Blue™	IBM Systems Director Active Energy	System p®
DB2 Connect™	Manager™	System Storage™
DB2®	IBM®	System x®
Distributed Relational Database	IMS™	System z10™
Architecture™	Language Environment®	System z9®
Domino®	Lotus®	System z®
DRDA®	MQSeries®	VM/ESA®
DS8000®	Parallel Sysplex®	WebSphere®
Dynamic Infrastructure®	PR/SM™	z/Architecture®
ECKD™	Processor Resource/Systems	z/OS®
ESCON®	Manager™	z/VM®
eServer™	RACF®	z/VSE™
FICON®	Redbooks®	z9®
GDPS®	Redbooks (logo)  ®	zSeries®
Geographically Dispersed Parallel	Resource Link™	
Sysplex™	S/390®	

The following terms are trademarks of other companies:

AMD, the AMD Arrow logo, and combinations thereof, are trademarks of Advanced Micro Devices, Inc.

InfiniBand, and the InfiniBand design marks are trademarks and/or service marks of the InfiniBand Trade Association.

Ambassador, and the LSI logo are trademarks or registered trademarks of LSI Corporation.

Novell, SUSE, the Novell logo, and the N logo are registered trademarks of Novell, Inc. in the United States and other countries.

Oracle, JD Edwards, PeopleSoft, Siebel, and TopLink are registered trademarks of Oracle Corporation and/or its affiliates.

Red Hat, and the Shadowman logo are trademarks or registered trademarks of Red Hat, Inc. in the U.S. and other countries.

SAP, and SAP logos are trademarks or registered trademarks of SAP AG in Germany and in several other countries.

Java, and all Java-based trademarks are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

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

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

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

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

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

Preface

This IBM® Redbooks® publication discusses the IBM System z10™ Enterprise Class, which offers a continuation of IBM scalable mainframe servers. Based on z/Architecture®, the IBM System z10 Enterprise Class (z10 EC) server provides major extensions by:

- ▶ Increasing the maximum number of processor units
- ▶ Providing fixed HSA where all devices, channel subsystems, and multiple subchannel sets are defined, thus better supporting dynamic changes
- ▶ Providing a base for major server consolidation by further removing memory, processor, and channel constraints
- ▶ Increasing the flexibility of capacity upgrades

This book provides an overview of the z10 EC and its functions, features, and associated software support. Greater detail is offered in areas relevant to technical planning.

The changes to this edition are based on the System z® hardware announcement, dated October 20, 2009.

This book is intended for systems engineers, consultants, planners, and anyone wanting to understand the System z10 Enterprise Class functions and plan for their usage. It is not intended as an introduction to mainframes. Readers are expected to be generally familiar with existing IBM System z technology and terminology.

The team who wrote this book

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

Per Fremstad is an IBM Certified Senior IT Specialist from the IBM Systems and Technology Group in IBM Norway. He has worked for IBM since 1982 and has extensive experience with mainframes and z/OS®. Per also works extensively with Linux® on System z and z/VM®. During the past 25 years he has worked in various roles within IBM and with a large number of customers. He frequently teaches about z/OS and z/Architecture subjects, and has been actively teaching at Oslo University College for the last 5 years. Per holds a BSc from the University of Oslo, Norway.

Wolfgang Fries is a Senior Consultant in the System z Support Center in Germany. He spent several years in the European support center in Montpellier, France, to provide international HW support for System z servers. He has 31 years of experience in supporting large System z customers. His area of expertise includes System z servers and connectivity.

Marian Gasparovic is an IT Specialist working for the IBM Server and Technology Group in IBM Slovakia. He worked as an Administrator for z/OS at Business Partner for 6 years. He joined IBM in 2004 as a Storage Specialist. Currently, he holds dual roles: one role is Field Technical Sales Support for System z in the CEMAAS region as a member of a team that handles new workloads; another role is for ITSO in Poughkeepsie, NY.

Parwez Hamid is a Executive IT Consultant working for the IBM Server and Technology Group. During the past 36 years he has worked in various IT roles within IBM. Since 1988 he has worked with a large number of IBM mainframe customers and spent much of his time

introducing new technology. Currently, he provides pre-sales technical support for the IBM System z product portfolio and is the lead System z technical specialist for UK and Ireland. Parwez co-authors a number of ITSO Redbooks and prepares technical material for the world-wide announcement of System z Servers. Parwez works closely with System z product development in Poughkeepsie and provides input and feedback for 'future' product plans. Additionally, Parwez is a member of the IBM IT Specialist profession certification board in the UK and is also a Technical Staff member of the IBM UK Technical Council which is made of senior technical specialist representing all IBM Client, Consulting, Services and Product groups. Parwez teaches and presents at numerous IBM user group and IBM internal conferences.

Brian Hatfield is a Certified Consulting Learning Specialist working for the IBM Systems and Technology Group in Atlanta, Georgia. He has over 30 years of experience in the IBM mainframe environment, starting his career as a Large System Customer Engineer in Southern California. He has been in education for the past 16 years and currently develops and delivers technical training for the System z environment.

Dick Jorna is an Executive IT Specialist working for IBM Server and Technology Group in the Netherlands. During the past 39 years he has worked in various roles within IBM and with a large number of mainframe customers. He currently provides pre-sales System z technical consultancy in support of large and small System z customers. In addition, he acts as a System z Product Manager in the Netherlands and is responsible for all activities related to System z.

Fernando Nogal is an IBM Certified Consulting IT Specialist working as an STG Technical Consultant for the Spain, Portugal, Greece, Israel, and Turkey IMT. He specializes in on-demand infrastructures and architectures. In his 26 years with IBM he has held a variety of technical positions, mainly providing support for mainframe customers. Previously, he was on assignment to the Europe Middle East and Africa (EMEA) zSeries® Technical Support group, working full time on complex solutions for e-business on zSeries. His job included, and still does, presenting and consulting in architectures and infrastructures, and providing strategic guidance to System z customers regarding the establishment and enablement of e-business technologies on System z, including the z/OS, z/VM, and Linux environments. He is a zChampion and a core member of the System z Business Leaders Council. An accomplished writer, he has authored and co-authored 16 Redbooks and several technical papers. Other activities include chairing a Virtual Team of IBMers interested in e-business on System z and serving as a University Ambassador. He travels extensively on direct customer engagements and as a speaker at IBM and customer events, and trade shows.

Karl-Erik Stenfors is a Senior IT Specialist in the Product and Solutions Support Centre (PSSC) in Montpellier, France. He has more than 40 years of experience in the large systems field, as a Systems Programmer, as a consultant with IBM customers, and, since 1986, with IBM. His areas of expertise include IBM System z hardware and operating systems, including z/VM, z/OS and Linux. He teaches at numerous IBM user group and IBM internal conferences. He is currently working with the System z lab in Poughkeepsie, providing customer requirement input to create an IBM System vision for the future-the zChampion workgroups.

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

Connie Beuselinck, William Clark, Cathy Cronin, Darelle Gent, Michael Gerhart, Gary King, Jeff Kubala, Scott Langenthal, Kenneth Oakes, Patrick Rausch, Charlie Shapley, Charles Webb, Frank Wisnewski
IBM Poughkeepsie

Les Geer, Reed Mullen, Brian Valentine
IBM Endicott

Harv Emery, Greg Hutchison
IBM Gaithersburg

Hans Wijngaard
Field Technical Sales Support, IBM Netherlands

Octavian Lascu
Global Technology Services, IBM Romania

Franck Injey, Bill White
International Technical Support Organization, IBM Poughkeepsie

Become a published author

Join us for a two- to six-week residency program! Help write a book dealing with specific products or solutions, while getting hands-on experience with leading-edge technologies. You will have the opportunity to team with IBM technical professionals, Business Partners, and Clients.

Your efforts will help increase product acceptance and customer satisfaction. As a bonus, you will develop a network of contacts in IBM development labs, and increase your productivity and marketability.

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 in one of the following ways:

- ▶ Use the online **Contact us** review Redbooks form found at:

ibm.com/redbooks

- ▶ Send your comments in an e-mail 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



Introducing the System z10 Enterprise Class

The IBM System z10 Enterprise Class (z10 EC) server represents both a revolution and an evolution of mainframe technology. With the newly designed z10 Enterprise quad-core chip, the fastest in the industry at 4.4 GHz, the z10 EC server can be configured up to a 64-way and 1.5 TB of memory, and offers new connectivity options while adopting advanced technologies such as InfiniBand.

The z10 EC breaks away from past designs, while continuing to enhance the traditional mainframe qualities, delivering in a single footprint unprecedented performance and capacity growth. The z10 EC is a well-balanced general-purpose server that is equally at ease with compute-intensive workloads as it is with I/O-intensive workloads.

The System z server design continues to follow the fundamental principle of being able to simultaneously support a large number of heterogeneous workloads while providing the highest qualities of service. But the workloads in themselves have changed a lot, and the design must adapt to this change.

The last couple of decades have witnessed an explosion in applications, architectures, and platforms. A lot of experimentation occurred in the marketplace. With the generalized availability of the internet and the appearance of commodity hardware and software, several patterns have emerged that have gained center stage.

Multi-tier application architectures and their deployment on heterogeneous infrastructures are common today. When these applications are mission critical, however, a great amount of effort must be done to ensure that the infrastructure provides the required qualities of service, and careful engineering of the application's several tiers is required to provide the robustness, scaling, consistent response, and other characteristics demanded by the users and lines of business.

Providing the required service level in a distributed environment implies acquiring and installing extra equipment and software to ensure availability and security, and additional manpower to configure, administer, troubleshoot, and tune such a complex set of separate and diverse environments. Often, by the end of the distributed equipment's life cycle its residual value is null, requiring new acquisitions and software licences, re-certification, and so

on, taking us back to square one. In today's resource constrained environments there must be another way.

The z10 EC offers an extensive software portfolio that spans from IBM WebSphere®, full support for SOA, Web services, J2EE, Linux, and open standards, to the more traditional batch and transactional environments such as CICS® and IMS™. For instance, considering just the Linux on System z environment, more than 3,000 applications are offered by over 400 independent software vendors (ISVs). The z10 EC is a platform of choice for the integration of the new generations of applications with existing applications and data.

The z10 EC expands the subcapacity settings offer with three different subcapacity levels for the first 12 processors, giving a total of 100 distinct capacity settings in the system, and providing for a range of 1:140 in processing power. The z10 EC delivers scalability and granularity to meet the needs of medium-sized enterprises, while also satisfying the requirements of large enterprises having large-scale, mission-critical transaction and data processing requirements. The z10 EC continues to offer all the specialty engines available with System z9®.

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

1.1 Wanted: an infrastructure (r)evolution

Exploitation of information technology (IT) by enterprises continues to grow and the demands placed upon it are increasingly complex. The world is not stopping. In fact, business pace is accelerating. The pervasiveness of the Internet fuels ever-increasing utilization modes and users. And the most rapidly growing type of user is not people, but devices. All sorts of services are being offered and new business models are being implemented. The demands placed on the network and computing resources will reach a breaking point unless something changes.

Awareness that the very foundation of IT infrastructures is not up to the job is growing. Most existing infrastructures are too complex, too inefficient, and too inflexible. How then can those infrastructures evolve and what must they become in order to avoid the breaking point? And, while they are evolving, the need to improve service delivery, manage the escalating complexity, and maintain a secure enterprise continues to be felt. To compound it, there is a daily pressure to cost-effectively run the business while supporting growth and innovation. Aligning IT with the goals of the business is an absolute top priority.

In the IBM vision of the future, transformation of the IT delivery model is strongly based on new levels of efficiency and service excellence for businesses, driven by and from the data center.

To achieve success in the transformation of their IT model and truly maximize the benefits of this new approach, organizations must develop and follow a plan for their transformation or journey towards that goal. IBM has developed a roadmap to help enterprises build such a plan. The roadmap lets IT free itself from operational complexity and reallocate scarce resources to drive business innovation. The roadmap follows a model based on an infrastructure supporting a highly dynamic, efficient, and shared environment. This is a new view of the data center. It allows IT to better manage costs, improve operational performance and resiliency, and more quickly respond to business needs.

By implementing this evolved infrastructure, organizations can better position themselves to adopt and integrate new technologies, such as Web 2.0 and cloud computing, and deliver dynamic and seamless access to IT services and resources.

Clouds, as seen from their users' side, offer services through the network. User requirements are in the functionality but also in the availability, ease of access, and security areas, so much so that organizations may decide to adopt private clouds while also exploiting public or hybrid clouds. From the service provider viewpoint, guaranteeing availability and security, along with repeatable and predictable response times, requires a very flexible IT infrastructure and advanced resource management.

IBM calls this evolved environment a Dynamic Infrastructure® and the IBM System z10 is at its core. Due to its advanced characteristics, the mainframe already provides many of the qualities of service and functions required, as will be discussed next.

Through its own transformation and engagements with thousands of enterprise clients, IBM has identified three stages of adoption along the way:

- ▶ Simplified
- ▶ Shared
- ▶ Dynamic

These are described in this section.

1.1.1 Simplified

In this stage, to drive new levels of economics in the data center, operational issues are addressed through consolidation, virtualization, energy offerings and service management. Most enterprises start their journey here.

The z10 EC supports advanced server consolidation and offers the best virtualization in the industry. The Processor Resource/Systems Manager™ (PR/SM™) function, responsible for hardware virtualization of the server, provides up to 60 logical partitions (LPARs). PR/SM technology has received Common Criteria EAL5¹ security certification for the System z10 EC. Each logical partition is as secure as a standalone server.

The z10 EC also offers software virtualization, through z/VM. z/VM's extreme virtualization capabilities, which have been perfected since its introduction in 1967, enable virtualization of thousands of distributed servers on a single z10 EC server. IBM is conducting a very large internal consolidation project, which aims to consolidate approximately 3,900 distributed servers into approximately 30 mainframes, using z/VM and Linux on System z. The project expects to achieve reductions of over 80% in the use of space and energy. So far, expectations are being fulfilled. Similar results have been publicly presented by various clients, and these reductions directly translate into significant monetary savings.

Consider also the potential gains in software licensing. The pricing model for many distributed software products is linked to the number of processors or processor cores. Consolidating under z/VM and exploiting the specialized Integrated Facility for Linux (IFL) processors can achieve a large reduction in the number of used cores.

In addition to server consolidation and image reduction by vertical growth under z/VM, z/OS provides a highly sophisticated environment for application integration and co-residence with data, especially for the mission-critical applications.

Most upgrades are concurrent to the hardware. As will be described later, the z10 EC reaches new availability levels by eliminating several preplanning requirements and other disruptive operations.

Further simplification is possible by exploiting the z10 EC HiperSockets™² and z/VM's Virtual Switch functions. These may be used, at no additional cost, to replace physical routers, switches and their cables, while eliminating security exposures and simplifying configuration and administration tasks. In some real simplification cases cables have been reduced by 97%.

IT operational simplification benefits also from the intrinsic autonomic characteristics of the z10 EC, the consolidation and reduction of the number of system images, the management best practices and products developed and available for the mainframe, in particular for the z/OS environment.

1.1.2 Shared

By shifting the focus from operational management to service management, this stage creates a shared IT infrastructure that can be provisioned and scaled rapidly and efficiently. Organizations can create virtualized resource pools for server platforms, storage systems, networks and applications, delivering IT capabilities to end users in a more flexible way.

¹ Evaluation Assurance Level with specific Target of Evaluation, Certificate for System z10 EC published October 29th 2008.

² For a description of HiperSockets see "HiperSockets" on page 15. The z/VM Virtual Switch is a z/VM system function that uses memory to emulate switching hardware.

An important point is that the z10 *stack* consists of much more than just a server. This is because of the total systems view that guides System z development. The *z-stack* is built around services, systems management, software, and storage. It delivers a complete range of policy-driven functions, pioneered and most advanced in the z/OS environment, including:

- ▶ Access management to authenticate and authorize who can access specific business services and associated IT resources
- ▶ Utilization management to drive maximum use of the system. Unlike other classes of servers, z10 is designed to run at 100% of utilization 100% of the time, based on the varied demands of its users.
- ▶ Just-in-time capacity to deliver additional processing power and capacity when needed
- ▶ Virtualization security to enable clients to allocate resources on demand without fear of security risks
- ▶ Enterprise-wide operational management and automation, leading to a more autonomic environment

In addition to the hardware-enabled resource sharing, other uses of virtualization include:

- ▶ Isolating production, test, training, and development environments
- ▶ Supporting back-level applications
- ▶ Enabling parallel migration to new system or application levels, and providing easy back-out capabilities

The resource sharing abilities of the z/VM operating system can drive additional savings by:

- ▶ Allowing dormant servers that do not use resources to be activated when required. This can help reduce hardware, software, and maintenance costs.
- ▶ Pooling resources such as processor, I/O facilities, and disk space. Virtual servers can be dynamically provisioned out of these pools, and, when their useful life ends, the resources are returned to the pools and recycled, with the utmost security.
- ▶ Offering very fast virtual server provisioning. A complete server can be deployed and ready for use in just a few minutes, using resources from the pool and image cloning.
- ▶ Eliminating the need to re-certify servers for specific purposes. Environments are certified to the virtual server. This has to be done only once, even if the server requires scaling up, because the underlying hardware and architecture does not change. Significant reductions in time and manpower can be achieved.
- ▶ Use virtualized resources to test hardware configurations without incurring the cost of buying the actual hardware, and providing the flexibility to easily optimize these configurations.

1.1.3 Dynamic

At this stage, organizations achieve alignment with business goals and can respond dynamically as business needs arise. Opposite from the “break/fix” mentality gripping many data centers, this new environment creates an infrastructure that is economical, integrated, agile and responsive, having harvested new technologies to support the new types of business enterprises. Social networks, highly integrated Web 2.0 applications and cloud computing deliver a rich environment and real-time information, as needed.

System z is the premier server offering from IBM, and the result of sustained and continuous investment and development policies. Commitment to IBM Systems design means that z10 EC brings all this innovation while helping customers leverage their current investment in the mainframe, as well as helping to improve the economics of IT.

The System z10 EC continues the evolution of the mainframe, building upon the z/Architecture definitions. The System z10 EC extends and integrates key platform characteristics: dynamic and flexible partitioning, resource management for mixed and unpredictable workload environments, availability, scalability, clustering, and security and systems management with emerging e-business on demand application technologies, such as WebSphere, Java™, and Linux.

All of these technologies and improvements come into play when the z10 EC is at the heart of the service-oriented architecture (SOA) solutions for an enterprise. In particular, the high availability, security, and scalability requirements of an Enterprise Service Bus (ESB) make its deployment on a mainframe environment highly advisable.

1.1.4 z10 at the core of a dynamic infrastructure

A dynamic infrastructure is able to rapidly respond to sudden requirements, even unforeseen ones. It is resilient, highly automated, optimized, and efficient and offers a catalog of services while granularly metering and billing those services.

The z10 EC enhances the availability and flexibility of just-in-time deployment of additional resources, known as Capacity on Demand (CoD). With the proper contracts, up to eight temporary capacity offerings can be installed on the server. Additional capacity resources can be dynamically activated, either fully or in part, by using granular activation controls directly from the management console, without the having to interact with IBM Support.

IBM has further enhanced and extended the z10 EC leadership with improved access to data and the network. The following list indicates several of many enhancements:

- ▶ Tighter security with CPACF protected key and longer personal account numbers for stronger protection of data
- ▶ Enhancements for improved performance connecting to the network
- ▶ Increased flexibility in defining your options to handle backup requirements
- ▶ Enhanced time accuracy to an external time source

A fast-growing number of enterprises are reaching the limits of available physical space and electrical power at their data centers. The extreme virtualization capabilities of the System z10 Enterprise Class enable the creation of dense and simplified infrastructures that are highly secure and can lower operational costs.

In summary, System z10 characteristics and qualities of service offer an excellent match to the requirements of a dynamic infrastructure, and this is why it is claimed to be at the core of such an infrastructure. System z10 is the most powerful tool available to reduce cost, energy, and complexity in enterprise data centers.

1.1.5 Storage is part of the System z10 stack

Recent advances in IBM System Storage™ disk technology provide clients with the opportunity to take advantage of IBM disk offerings' increased function and value, especially in the area of secure data encryption. Those offerings include updated business continuity features that make the most of the new mainframe's power.

Also for the System z10, the IBM System Storage Virtual Tape solution delivers improved tape processing while supporting business continuity and security through innovative enhancements.

Most topics mentioned in this chapter are discussed in greater detail later in this book. In this chapter, we introduce components of the system design. In subsequent chapters, we focus on specific features and functions that are relevant to technical planning.

1.2 System z10 EC highlights

The z10 EC provides a record level of capacity over the previous System z servers, achieved by both increasing the performance of the individual processor units and increasing the number of processor units (PUs) per server. The increased performance and the total system capacity available, along with possible energy savings, offer the opportunity to continue to consolidate diverse applications on a single platform and turn it into real financial savings. New features help to ensure that System z10 EC is an innovative, security-rich platform that can help maximize resource exploitation and utilization, and can help provide the ability to integrate applications and data across the enterprise IT infrastructure.

IBM continues its technology leadership with the z10 EC. The server is built using IBM modular multibook design that supports one to four books per server. The book contains a Multi-Chip Module (MCM), which hosts the newly designed CMOS 11S processor units, storage control chips, and high-z connectors for I/O. This approach enables many of the high-availability and nondisruptive operations capabilities that differentiate it from other servers. In addition, a new system I/O bus takes advantage of the InfiniBand technology, which is also exploited in coupling links. Figure 1-1 shows an external view of the z10 EC.



Figure 1-1 System z10 Enterprise Class

The Parallel Sysplex® cluster takes the commercial strengths of the z/OS platform to improved levels of system management, competitive price/performance, scalable growth, and continuous availability.

The z10 EC has five model offerings ranging from one to 64 configurable processor units (PUs). The first four models (E12, E26, E40, and E56) have 17 PUs per book, and the high capacity model (the E64) has one 17 PU book and three 20 PU books. Model E64 is estimated to provide up to 70% more total system capacity than the z9 EC Model S54, with up to three times the available memory. This comparison is based on the Large Systems Performance Reference (LSPR) mixed workload average.

Flexibility in customizing traditional capacity to meet individual needs has led to the introduction on the z9 EC of subcapacity CPs. The z10 EC has increased the number of subcapacity CPs available in a server to twelve. When the capacity backup (CBU) function is invoked, the number of total subcapacity processors cannot exceed twelve.

Depending on the model, the z10 EC can support from a minimum of 16 GB to a maximum of 1520 GB of memory, with up to 384 GB per book. In addition, a fixed amount of 16 GB is reserved for HSA (Hardware System Area) and is not part of customer-purchased memory.

There are up to 48 high-performance fanouts for data communications between the server and the peripheral environment. The multiple channel subsystems (CSS) architecture allows up to four CSSs, each with 256 channels. I/O constraint relief, using multiple subchannel sets (MSS), allows access to a greater number of logical volumes.

Processor Resource/System Manager (PR/SM) manages all the installed and enabled resources (processors and memory) as a single large SMP system. It enables the configuration and operation of up to 60 logical partitions, which have processors, memory, and I/O resources assigned from the installed books. PR/SM dispatching has been redesigned to work together with the z/OS dispatcher in a function called HiperDispatch. HiperDispatch provides work alignment to logical processors, and alignment of logical processors to physical processors. This alignment optimizes z/OS work dispatching and increases throughput.

The z10 EC continues the mainframe reliability, availability, and serviceability (RAS) tradition of reducing all sources of outages by continuous focus by IBM on keeping the system running. It is a design objective to provide higher availability with a focus on reducing planned and unplanned outages. With a properly configured z10 EC, further reduction of outages can be attained through improved nondisruptive replace, repair, and upgrade functions for memory, books, and I/O adapters, as well as extending nondisruptive capability to download Licensed Internal Code (LIC) updates.

Enhancements include removing preplanning requirements with the new fixed 16 GB HSA. Customers will no longer need to worry about using their purchased memory when defining their I/O configurations with reserved capacity or new I/O features. Maximums can be configured and IPLed so that insertion at a later time can be dynamic and not require a power on reset of the server.

Capacity on Demand

On demand enhancements enable customers to have more flexibility in managing and administering their temporary capacity requirements. System z10 has a new architectural approach for temporary offerings that has the potential to change the thinking about on demand capacity. Within System z10, one or more flexible configuration definitions can be available to solve multiple temporary situations and multiple capacity configurations can be active simultaneously.

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

1.3 System z10 EC Models

The System z10 EC has a machine type of 2097. Five models are offered: E12, E26, E40, E56, and E64. The last two digits of each model indicate the maximum number of PUs available for purchase. A PU is the generic term for the z/Architecture processor on the Multi-Chip Module (MCM) that can be characterized as any of the following items:

- ▶ Central processor (CP).
- ▶ Internal coupling facility (ICF) to be used by the Coupling Facility Control Code (CFCC).
- ▶ Integrated Facility for Linux (IFL)
- ▶ Additional system assist processor (SAP) to be used by the channel subsystem.
- ▶ System z10 Application Assist Processor (zAAP). One CP must be installed with or prior to installation of any zAAPs.
- ▶ System z10 Integrated Information Processor (zIIP). One CP must be installed with or prior to any zIIPs being installed.

In the five-model structure, only 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 may not exceed the total number available for that model.

The multibook system design provides an opportunity to concurrently increase the capacity of the system in three ways:

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

I/O features or channel types supported are:

- ▶ ESCON® (ESCON is Enterprise Systems Connection)
- ▶ FICON® Express8 (FICON Fibre Channel connection)
- ▶ FICON Express4, FICON Express2, and FICON Express (only when carried forward from a previous System z server)
- ▶ OSA-Express3 and OSA-Express2
- ▶ Crypto Express2 (only when carried forward from a previous System z server)
- ▶ Crypto Express3
- ▶ Coupling Links - peer mode only (ICB-4 and ISC-3)
- ▶ The Parallel Sysplex InfiniBand coupling link (PSIFB)

1.3.1 Model upgrade paths

Any z10 EC can be upgraded to a z10 EC hardware model. Upgrade of models E12, E26, E40, and E56 to E64 is disruptive. When you upgrade to a Model E64, the first book is retained. Any z990 or z9 EC model may be upgraded to any z10 EC model. A z10 Business Class (z10 BC) may be upgraded to a z10 EC model E12. Figure 1-2 presents a diagram of the upgrade path.

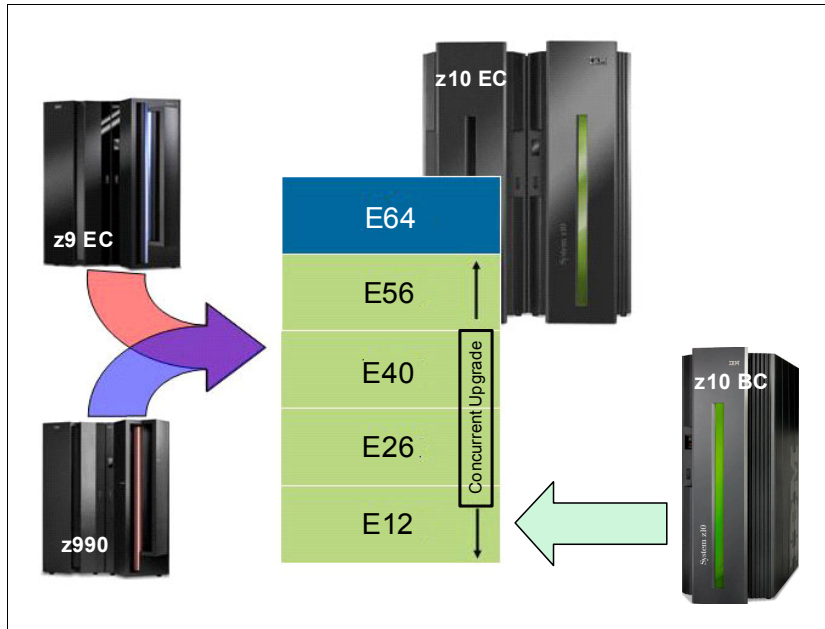


Figure 1-2 System z upgrades

1.3.2 Concurrent processing unit conversions

The z10 EC supports concurrent conversion between different PU types, providing flexibility to meet changing business environments. CPs, IFLs, zAAPs, zIIPs, ICFs, or optional SAPs may be converted to CPs, IFLs, zAAPs, zIIPs, ICFs, or optional SAPs.

1.4 System functions and features

The z10 EC is a two-frame server. The frames contain the key components. The server comprises the following:

- ▶ The CEC cage with up to four books
- ▶ One to three I/O cages
- ▶ Power supplies
- ▶ An optional internal battery feature (IBF)
- ▶ Modular cooling units
- ▶ Support Elements

Functions and features include:

- ▶ High speed 4.4 GHz quad-core processor
- ▶ Single processor core sparing
- ▶ Large memory of up to 1.5 TB (A maximum of 1TB is configurable to a logical partition)

- ▶ Large page (1 MB)
- ▶ Availability enhancements
- ▶ 16 GB fixed HSA supporting:
 - Maximum configuration of 60 LPARs, 4 LCSSs and 2 MSSs
 - Dynamic add/remove of a new logical partition (LPAR) to new or existing logical channel subsystem (LCSS)
 - Dynamic addition and removal Crypto Express2 and Crypto Express3 features
 - Dynamic I/O enabled as a default
 - Add/change number of logical CPs, IFLs, ICFs, zAAPs, and zIIPs processors per partition
 - Dynamic LPAR PU assignment optimization CPs, ICFs, IFLs, zAAPs, and zIIPs
- ▶ Redundant I/O interconnect
- ▶ Hot swap ICB-4 and HCA
- ▶ Redundant 100 Mb Ethernet Service Network with VLAN
- ▶ Enhanced security features (in CPACF, Crypto Express2, Crypto Express3, and TKE)
- ▶ InfiniBand coupling links for local and remote (up to 10 km or 6.2 miles unrepeated) connections
- ▶ Coupling Facility Control Code Level 16 to help provide faster service time for coupling facility (CF) Duplexing and improvements to scheduling to help improve CPU utilization
- ▶ Support for Server Time Protocol (STP) feature
- ▶ Increased flexibility for Capacity on Demand just-in-time offerings with ability for more temporary offerings installed on the central processor complex (CPC) and ways to acquire capacity backup

Design highlights

The z10 EC provides:

- ▶ Uniprocessor performance improvement, which can be up to 60% more than the z9 EC, based on LSPR mixed workload average
- ▶ Up to 70% more total system capacity than the z9 EC, with up to 64 processor units compared with a maximum of 54 PUs on z9 EC. This comparison is of the z10 EC 64-way and the z9 EC S54 and is based on LSPR mixed workload average running z/OS V1 R8
- ▶ Up to 384 GB memory per book
- ▶ Increased bandwidth between memory and I/O
- ▶ Up to 16 InfiniBand DDR (Dual Data Rate) Interconnect connections per book
- ▶ Reduction in the impact of planned and unplanned server outages:
 - Enhanced book availability
 - Redundant I/O interconnect
 - Enhanced driver maintenance
 - Concurrent Host Channel Adapter (HCA-O and HCA-C) fanout and memory bus adapter (MBA) fanout card hot-plug
- ▶ Multiple subchannel sets (MSS), which are designed to allow improved device connectivity for Parallel Access Volumes (PAVs)

- ▶ More capacity over native FICON channels for programs that process data sets, which exploit striping and compression (such as DB2®, VSAM, PDSE, HFS, and zFS) by reducing channel, director, and control unit overhead when using the Modified Indirect Data Address Word (MIDAW) facility
- ▶ Improved access to data with High Performance FICON for System z (zHPF) on FICON Express8, FICON Express4, and FICON Express2 channels
- ▶ 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
- ▶ Increased number of open exchanges (concurrent I/O operations) that can be active simultaneously for FICON channels
- ▶ Up to 336 FICON Express8 channels
- ▶ Two additional OSA-Express3 features

1.4.1 Processor

A minimum of one CP, IFL, or ICF must be purchased for each model. One zAAP or zIIP or both can be purchased for each CP purchased.

Processor features

The z10 EC book has a new Multi-Chip Module with five new IBM z10™ Enterprise chips. The chip has new quad-core design, with either three or four active cores, and operates at 4.4 GHz. Depending on the MCM version (17 PU or 20 PU), from 17 to 77 PUs are available, on one to four books.

This new MCM provides a significant increase in system scalability and an additional opportunity for server consolidation. All books are interconnected with very high-speed internal communications links, in a fully connected star topology through the L2 cache, which allows the system to be operated and controlled by the PR/SM facility as a memory-coherent symmetric multiprocessor (SMP).

The PU configuration is made up of two spare PUs per server and a variable number of system assist processors (SAPs), which scale with the number of books installed in the server, such as three SAPs with one book installed and up to eleven when four books are installed. The remaining PUs can be characterized as central processors (CPs), Integrated Facility for Linux (IFL) processors, System z10 Application Assist Processors (zAAPs), System z10 Integrated Information Processors (zIIPs), internal coupling facility (ICF) processors, or additional SAPs.

The PU chip includes data compression and cryptographic functions, such as the CP Assist for Cryptographic Function (CPACF). Hardware data compression can play a significant role in improving performance and saving costs over doing compression in software. Standard clear key cryptographic processors right on the processor translate to high-speed cryptography for protecting data in storage, integrated as part of the PU.

Each core on the PU has its own hardware decimal floating point unit. Much of today's commercial computing is decimal floating point, so on-core hardware decimal floating point meets the requirements of business and user applications, and provides improved performance, precision, and function.

Increased flexibility with z/VM-mode partitions

System z10 EC provides for the definition of a z/VM-mode logical partition (LPAR) containing a mix of processor types including CPs and specialty processors such as IFLs, zIIPs, zAAPs, and ICFs.

z/VM V5R4 and above support this capability that increases flexibility and simplifies systems management. In a single LPAR, z/VM can manage guests that exploit Linux on System z on IFLs, z/VSE™ and z/OS on CPs, execute designated z/OS workloads, such as parts of DB2 DRDA® processing and XML, on zIIPs, and provide an economical Java execution environment under z/OS on zAAPs.

1.4.2 Memory subsystem

A buffered DIMM has been developed for the z10 EC. For this purpose IBM has developed a chip that controls communication with the PU and redrives address and control from DIMM to DIMM. The DIMM uses DDR2 DRAM chips of 1 Gb and 2 Gb in size to provide DIMM capacities of 4 GB and 8 GB, respectively.

Memory topology

Memory topology provides:

- ▶ Maximum of 1.5 TB of physical memory (with a maximum of 1 TB configurable to a single logical partition)
- ▶ One memory port for each CP chip; up to four memory ports per node
- ▶ Asymmetrical memory size and DRAM technology across nodes
- ▶ Key storage
- ▶ Storage protection key array kept in physical memory
- ▶ Storage protection (memory) key is also kept in every L1.5 and L2 cache directory entry
- ▶ Large, fixed-size HSA eliminates having to plan for HSA

1.4.3 Central processor complex cage

This section highlights new characteristics in the central processor complex (CPC).

MCM technology

The z10 EC is built on a proven superscalar microprocessor architecture. On each book, there is one MCM. The MCM has five PU chips and two SC chips. The PU chip has up to four cores, which can be characterized as CPs, IFLs, ICFs, zIIPs, zAAPs, or SAPs. Two MCM sizes are offered, which are 17 or 20 cores.

Because of increased frequency (4.4 GHz versus 1.7 GHz), the chips on the MCM generate more heat than the z9 EC chips. The MCMs on the z10 EC therefore will continue to be modular refrigeration unit (MRU) cooled with air backup.

Host channel adapter fanout hot-plug

A host channel adapter fanout provides the path for data between memory and the I/O cards using InfiniBand (IFB) cables. The HCA fanout is hot-pluggable.

In the event of an outage, an HCA fanout can be concurrently repaired without loss of access to its associated I/O cards, using redundant I/O interconnect.

Up to eight HCA fanouts are available per book.

1.4.4 I/O connectivity

The z/10 EC offers several new and improved features and exploits new technologies such as InfiniBand. In this section we briefly review the most relevant I/O capabilities.

InfiniBand

In 1999, two competing input/output (I/O) standards called Future I/O (developed by Compaq, IBM, and Hewlett-Packard) and Next Generation I/O (developed by Intel®, Microsoft®, and Sun) merged into a unified I/O standard called InfiniBand. InfiniBand is an industry-standard specification that defines an input/output architecture used to interconnect servers, communications infrastructure equipment, storage, and embedded systems. InfiniBand is a true fabric architecture that leverages switched, point-to-point channels with current supported data transfers of up to 120 Gbps, both in chassis backplane applications as well as through external copper and optical fiber connections.

InfiniBand is a pervasive, low-latency, high-bandwidth interconnect that requires low processing overhead and is ideal to carry multiple traffic types (clustering, communications, storage, management) over a single connection. As a mature and field-proven technology, InfiniBand is used in thousands of data centers, high-performance compute clusters, and embedded applications that scale from two nodes up to a single cluster that interconnects thousands of nodes.

The z10 EC takes advantage of InfiniBand to implement:

- ▶ A new I/O bus, which includes the InfiniBand Double Data Rate (IB-DDR) infrastructure. This replaces the self-timed interconnect features found in prior System z servers.
- ▶ Parallel Sysplex coupling over InfiniBand (PSIFB). This link has a bandwidth of 6 GBps between two z10 servers and 3 GBps between System z10 and System z9 servers.
- ▶ Server Time Protocol.

1.4.5 I/O subsystems

The I/O subsystem direction is evolutionary, drawing on developments from z990 and z9 EC. The I/O subsystem is supported by a new I/O bus, and includes the InfiniBand Double Data Rate (IB-DDR) infrastructure (replacing self-timed interconnect found in the prior System z servers). This new infrastructure is designed to reduce overhead and latency, and provide increased throughput. The I/O expansion network uses the InfiniBand Link Layer (IB-2, Double Data Rate).

The z10 EC has Host Channel Adapter (HCA) fanouts residing on the front of the book. The z10 EC generation of the I/O platform is intended to provide significant performance improvement over the current I/O platform. It will be the primary platform to support future high-bandwidth requirements for FICON/Fibre Channel, Open Systems Adapters, and Crypto.

I/O cage

The z10 EC has a minimum of one CEC cage and one I/O cage in the A frame. The Z frame can accommodate two additional I/O cages, for a total of three for the entire system. One I/O cage can accommodate the following card types:

- ▶ Up to eight Crypto Express2 or Crypto Express3 features
- ▶ Up to 28 FICON Express8, FICON Express4, FICON Express2, or FICON Express
- ▶ Up to 24 OSA-Express2 and OSA-Express3
- ▶ Up to 28 ESCON
- ▶ Up to 12 ISC-M (48 ISC-3 links)

It is possible to populate the 28 I/O slots of each I/O cage with any mix of the previously mentioned cards.

ESCON channels

The high-density ESCON feature (FC 2323) has 16 ports, of which 15 can be activated. One port is always reserved as a spare in the event of a failure of one of the other ports.

FICON channels

Up to 336 FICON Express8, FICON Express4, or FICON Express2 channels and up to 120 FICON Express channels are supported:

- ▶ The FICON Express8 features support a link data rate of 2, 4, or 8 Gbps.
- ▶ The FICON Express4 features support a link data rate of 1, 2, or 4 Gbps.
- ▶ The FICON Express2 features support a link data rate of 1 or 2 Gbps.

The z10 EC supports FCP channels, switches, and FCP/SCSI devices with full fabric connectivity under Linux on System z.

Open Systems Adapter

The z10 EC can have up to 24 features of the Open Systems Adapter (OSA) family, for a maximum of 96 ports of LAN connectivity.

You can choose any combination of the supported OSA-Express2 or OSA-Express3 Ethernet features. The OSA-Express Token Ring is not supported.

OSA-Express3 feature highlights

The z10 EC has five OSA-Express3 features. When compared to similar OSA-Express2 features, which they replace, OSA-Express3 features provide the following important benefits:

- ▶ Doubling the density of ports
- ▶ For TCP/IP traffic, reduced latency and improved throughput for standard and jumbo frames.

Performance enhancements are the result of the data router function present in all OSA-Express3 features. What previously was performed in firmware, the OSA-Express3 now performs in hardware. Additional logic in the IBM ASIC handles packet construction, inspection, and routing, thereby allowing packets to flow between host memory and the LAN at line speed without firmware intervention.

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

For more information about the OSA-Express3 features refer to 4.6.4, “OSA-Express3” on page 136.

HiperSockets

The HiperSockets function, also known as internal queued direct input/output (iQDIO, or internal QDIO), is an integrated function of the System z10 that provides users with attachments to up to 16 high-speed virtual LANs with minimal system and network overhead.

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, eliminating attachment costs while improving availability and performance.

HiperSockets eliminates having to use I/O subsystem operations and to traverse an external network connection to communicate between logical partitions in the same System z10 server. HiperSockets offers significant value in server consolidation by connecting many virtual servers, and can be used instead of certain coupling link configurations in a Parallel Sysplex.

1.4.6 Cryptography

Integrated cryptographic features provide leading cryptographic performance and functionality. Reliability, availability, and serviceability (RAS) support is unmatched in the industry and the cryptographic solution has received the highest standardized security certification. The crypto cards are supported with additional capabilities to add or move crypto processors to logical partitions without pre-planning.

CP Assist for Cryptographic Function

The z10 EC uses the Cryptographic Assist Architecture. The CP Assist for Cryptographic Function (CPACF) offers the full complement of the Advanced Encryption Standard (AES) algorithm and Secure Hash Algorithm (SHA). Support for CPACF is also available by using the Integrated Cryptographic Service Facility (ICSF). ICSF is a component of z/OS, and can transparently use the available cryptographic functions, CPACF, Crypto Express2, or Crypto Express3, to balance the workload and help address the bandwidth requirements of your applications.

The enhancements to CPACF are exclusive to the System z10 and supported by z/OS, z/VM, z/VSE, and Linux on System z.

Configurable Crypto Express features

The Crypto Express features has two PCI-X adapters, which can each be configured as a coprocessor or an accelerator:

- ▶ Crypto Express Coprocessor is for secure key encrypted transactions (default).
- ▶ Crypto Express Accelerator is for Secure Sockets Layer (SSL) acceleration.

A recently added function includes support for secure key AES and 13-digit through 19-digit Personal Account Numbers, often used by credit card companies security code computations.

Because the features are implemented in Licensed Internal Code, current Crypto Express2 features carried forward from z990 and z9 can take advantage of configuration options on z10 EC.

The configurable Crypto Express2 and Crypto Express3 features are supported by z/OS, z/VM, z/VSE, Linux on System z, and (as an accelerator only) by z/TPF.

TKE workstation and continued support for Smart Card Reader

The Trusted Key Entry (TKE) workstation (FC 0839³) and the TKE 5.3 LIC (FC 0854) or TKE 6.0 LIC (FC0858) are optional features on the System z10 EC. The TKE workstation offers security-rich local and remote key management, providing authorized persons a method of operational and master key entry, identification, exchange, separation, and update. Recent enhancements include support for the AES encryption algorithm, audit logging, and an infrastructure for payment card industry data security standard (PCIDSS).

³ A next-generation TKE workstation (FC0840) is planned to start shipping to customers in the European Union and Switzerland beginning January 1, 2010.

Support for an optional Smart Card Reader attached to the TKE workstation allows for 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 is protected by a user-defined personal identification number (PIN).

1.4.7 Parallel Sysplex support

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

Coupling links support

Coupling connectivity in support of Parallel Sysplex environments is improved with the Parallel Sysplex InfiniBand (PSIFB) link. Parallel Sysplex connectivity now supports:

- ▶ Internal Coupling Channels (ICs) operating at memory speed
- ▶ Integrated Cluster Bus-4 (ICB-4), operating at 2 GBps and supported by a 10-meter copper cable provided as a feature (maximum of 7 meters distance, in practice). The ICB-4 uses a dedicated self-timed interconnect (STI) for communication.
- ▶ InterSystem Channel-3 (ISC-3) operating at 2 Gbps and supporting an unrepeated link data rate of 2 Gbps over 9 μ m single mode fiber optic cabling with an LC Duplex connector.
- ▶ InfiniBand (IB) short range (SR) coupling links offer up to 6 GBps of bandwidth between z10 EC servers and up to 3 GBps of bandwidth between System z10 and System z9 for a distance up to 150 m (492 feet).
- ▶ InfiniBand long reach (LR) up to 5 Gbps connection bandwidth between System z10 servers for a distance up to 10 km (6.2 miles).

InfiniBand coupling links can be used to carry Server Time Protocol (STP) messages.

Coupling Facility Control Code Level 16

Coupling Facility Control Code (CFCC) Level 16 is available for the IBM System z10 EC. Enhancements include:

- ▶ CF Duplexing enhancements

Prior to CFCC Level 16, System-Managed CF Structure Duplexing required two protocol enhancements to occur synchronous to CF processing of the duplexed structure request. CFCC Level 16 allows one of these signals to be asynchronous to CF processing, which allows faster service time, with more benefits as the distances between coupling facilities are further apart, such as in a multi-site Parallel Sysplex.

- ▶ List Notification improvements

Prior to CFCC Level 16, when a list changed state from empty to non-empty, it would notify its connectors. The first one to respond would read the new message, but when the others would read, they found nothing, paying the cost for the *false scheduling*. CFCC Level 16 can help improve processor utilization for IMS Shared Queue and WebSphere MQ Shared Queue environments by the coupling facility by notifying only one connector in a round-robin fashion. If the shared queue is read within a fixed period of time, the other connectors do not have to be notified, thereby saving the cost of the false scheduling. If the list is not read within the time limit, then the other connectors are notified as today.

External time reference facility

Two external time reference (ETR) cards are shipped as a standard feature with the server. They provide a dual-path interface to the IBM Sysplex Timers, which can be used for timing synchronization between systems in a sysplex environment. The ETR facility allows

continued operation even if a single ETR card fails. This redundant design also allows concurrent maintenance.

Each card also has a coaxial connector to link to the pulse per second (PPS) signal.

Server Time Protocol facility

Server Time Protocol (STP) is a server-wide facility that is implemented in the Licensed Internal Code of System z servers and coupling facilities. STP presents a single view of time to PR/SM and provides the capability for multiple servers and coupling facilities to maintain time synchronization with each other. Any System z servers or CFs may be enabled for STP by installing the STP feature. Each server and CF that are planned to be configured in a coordinated timing network (CTN) must be STP-enabled.

The STP feature is designed to be the supported method for maintaining time synchronization between System z servers and coupling facilities. The STP design uses the CTN concept, which is a collection of servers and coupling facilities that are time-synchronized to a time value called *coordinated server time*.

Network Time Protocol (NTP) client support has been added to the STP code on the System z10 and on System z9. With this functionality the System z10 and the System z9 can be configured to use an NTP server as an external time source (ETS).

This implementation answers the need for a single time source across the heterogeneous platforms in the enterprise, allowing an NTP server to become the single time source for the System z10 and the System z9, as well as other servers that have NTP clients (UNIX®, NT, and so on). NTP can only be used for an STP-only CTN where no server can have an active connection to a Sysplex Timer®.

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

Improved security can be obtained by providing NTP server support on the Hardware Management Console (HMC), as the HMC is normally attached to the private dedicated LAN for System z maintenance and support.

1.4.8 Reliability, availability, and serviceability

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

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

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

1.5 Performance

This section briefly discusses the results of the performance tests that can be found on the LSPR Web site:

<http://www.ibm.com/servers/eserver/zseries/lspr/>

Workload performance variation

As with previous servers with the same multibook structure, the z10 EC is likely to have workload variability. This variability can be observed in several ways. The range of performance ratings across the individual LSPR workloads is likely to have a large spread. There is also a performance variation of individual logical partitions because the affect of fluctuating resource requirements of other partitions can be more pronounced, which is a result of the book structure of IBM System z10 Enterprise Class. You can see the affect of this increased variability as increased deviations of workloads from single-number metric-based factors, such as MIPS, MSUs, and CPU time charge-back algorithms.

LSPR workload suite

The LSPR workloads, updated for z9 EC and again for z10 EC, are considered to reasonably reflect current and growth workloads of the customer. The set continues to contain:

- ▶ Traditional online transaction processing workload (OLTP-T, formerly known as IMS)
- ▶ Web-enabled online transaction processing workload (OLTP-W, also known as Web/CICS/DB2)
- ▶ Java-based online stock trading application (WASDB, previously referred to as Trade2-EJB)
- ▶ Batch processing, represented by the commercial batch with long-running jobs (CB-L CBW2)
- ▶ Java batch workload (ODE-B, replacing the CB-J workload)

The System z10 EC LSPR provides performance ratios for individual workloads and the *default mixed workload*, which is composed of equal amounts of four of the workloads (OLTP-T, OLTP-W, WASDB, and CB-L). LSPR rates all z/Architecture processors running in LPAR mode and 64-bit mode. The single-number metrics, MIPS, MSUs, and SRM constants are based on a combination of the default mixed workload ratios, typical multiple LPAR configurations, and expected early-program migration scenarios.

The LSPR has two tables:

- ▶ Single image z/OS from 1-way to 32-way
- ▶ Typical logical partition configuration from 1-way to 64-way, based on customer profiles. This logical partition configuration table is used to establish single-number metrics.

In addition to these z/OS workloads used for setting the single-number metrics, the LSPR also contains information pertaining to Linux and z/VM environments. These environments, updated for z10 EC, tend to fall within the range of the z/OS workloads and are expected to continue in that range.

Capacity ratio estimates

With a modular book design, System z10 EC model E64 can provide up to 1.7 times more total system capacity than the z9 EC Model S54, and has up to three times the available memory⁴. The performance of the z10 EC (Machine Type 2097) Model 701 is 1.62 times that of the z9 EC (2094) Model 701 (in an LSPR mixed workload).

The LSPR contains the internal throughput rate ratios (ITRRs) for the z10 EC and the previous generation processor families, based upon measurements and projections that use standard IBM benchmarks in a controlled environment. The actual throughput that any user experiences can vary depending on considerations, such as the amount of multiprogramming in the user's job stream, the I/O configuration, and the workload processed. Therefore, no assurance can be given that an individual user can achieve throughput improvements equivalent to the performance ratios stated.

Consult the Large System Performance Reference (LSPR) when you consider performance on the z10 EC. The range of performance ratings across the individual LSPR workloads is likely to have a large spread. More performance variation of individual logical partitions exists because the impact of fluctuating resource requirements of other partitions can be more pronounced with the increased numbers of partitions and additional PUs available.

For detailed performance information, see the LSPR Web site:

<http://www.ibm.com/servers/eserver/zseries/lspr/>

The MSU ratings are available from:

<http://www.ibm.com/servers/eserver/zseries/library/swpriceinfo>

1.6 Operating systems and software

The z10 EC is supported by a large set of software, including ISV applications. This section lists only the supported operating systems. Exploitation of some features might require the latest releases. Further information is contained in Chapter 7, “Software support” on page 189.

System z10 EC supports any of the following operating systems:

- ▶ z/OS Version 1 Release 7 with IBM Lifecycle Extension and z/OS Version 1 Release 8 with IBM Lifecycle Extension. Note that z/OS.e is not supported.
- ▶ z/OS Version 1 Release 9 and later.
- ▶ z/VM Version 5 Release 3 and later.
- ▶ z/VSE Version 4 Release and later.
- ▶ TPF Version 4 Release 1 and z/TPF Version 1 Release 1.
- ▶ Linux on System z distributions:
 - Novell SUSE: SLES⁵ 9, SLES 10, and SLES 11
 - Red Hat: RHEL⁶ 4 and RHEL 5

Note: Regular service support for z/OS V1 R8 ended in September 2009. However, by ordering the IBM Lifecycle Extension for z/OS V1.8 product, fee-based corrective service can be obtained for up to two years after withdrawal of service. Similarly, by ordering the IBM Lifecycle Extension for z/OS V1.7 product, customers can obtain support up to September 2010.

⁴ This is a comparison of the z10 EC 64-way and the z9 EC 54-way and is based on the LSPR mixed workload average.

⁵ SLES is the abbreviation for Novell SUSE Linux Enterprise Server.

⁶ RHEL is the abbreviation for Red Hat Enterprise Linux.

Finally, a large software portfolio is available to the IBM System z10 Enterprise Class, including an extensive collection of middleware and ISV products that implement the most recent proven technologies.

With support for IBM WebSphere software, full support for SOA, Web services, J2EE, Linux, and Open Standards, the System z10 is intended to be a platform of choice for integration of a new generation of applications with existing applications and data.



Hardware components

This chapter introduces IBM System z10 Enterprise Class hardware components along with significant features and functions with their characteristics and options. Our objective is to explain the IBM System z10 Enterprise Class hardware building blocks, and how these components interconnect from a physical point of view. This information can be useful for planning purposes and helps to define configurations that best fit the requirements.

This chapter discusses the following topics:

- ▶ 2.1, “Frames and cages” on page 24
- ▶ 2.2, “Book concept” on page 27
- ▶ 2.3, “Multi-Chip Module” on page 30
- ▶ 2.4, “Processing units and storage control chips” on page 31
- ▶ 2.5, “Memory” on page 35
- ▶ 2.6, “Connectivity” on page 41
- ▶ 2.7, “Model configurations” on page 45
- ▶ 2.8, “Summary of z10 EC structure” on page 54

2.1 Frames and cages

The frames are enclosures built to Electronic Industry Association (EIA) standards. The server always has two frames that are composed of two 42U EIA frames, shown in Figure 2-1. The two frames, A and Z, are bolted together and have two cage positions (top and bottom):

- ▶ Frame A has the CEC cage at the top and I/O cage 1 at the bottom.
- ▶ Frame Z can be one of the following configurations:
 - Without an I/O cage
 - With one I/O cage, I/O cage 2, at the bottom
 - With two I/O cages, I/O cage 2 at the bottom and I/O cage 3 on top

All books, including the DCAs on the books, and the cooling components are located in the CEC cage in the top half of the A frame. Figure 2-1 shows the front view of both frame A (with four books installed) and frame Z.

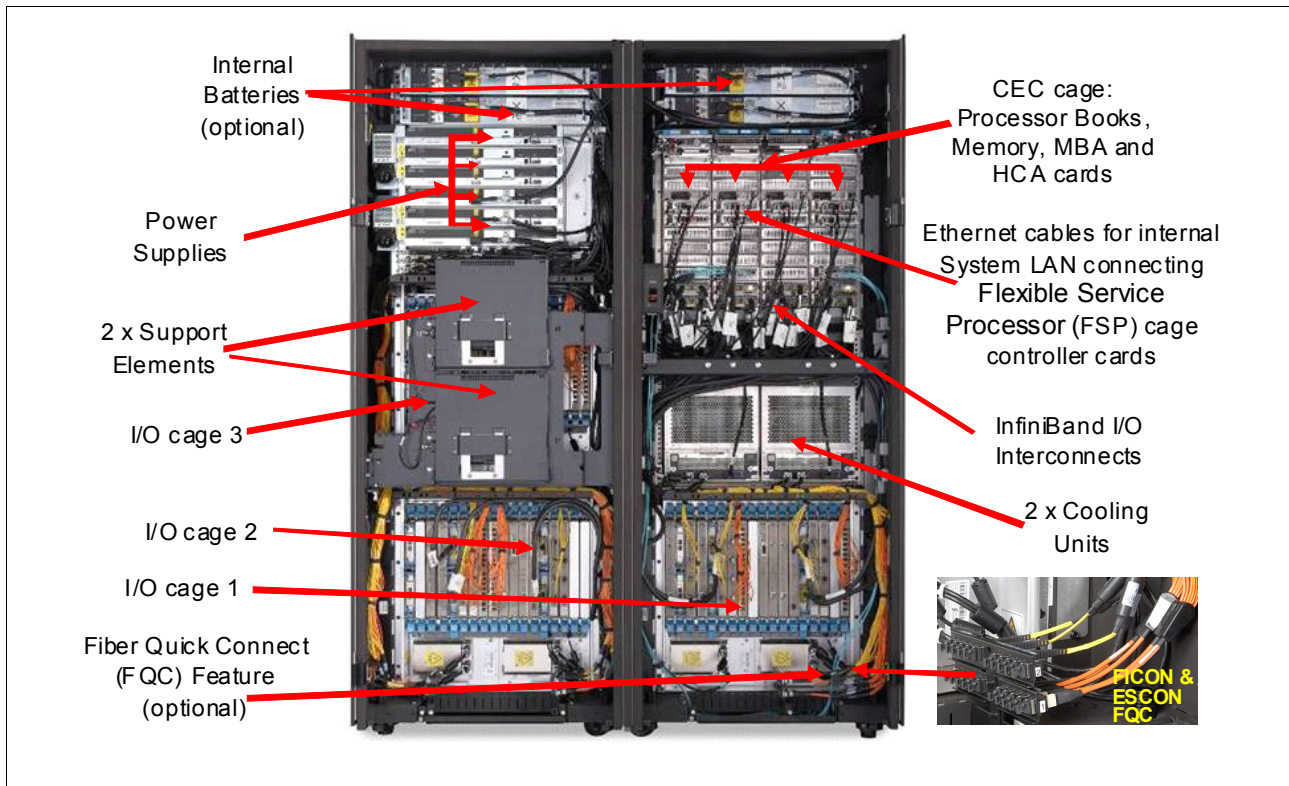


Figure 2-1 CEC cage and I/O cage locations

2.1.1 Frame A

As shown in Figure 2-1, the main components in frame A are:

- ▶ Two optional Internal Battery Features (IBFs), which provide the function of a local uninterrupted power source

The IBF further enhances the robustness of the power design, increasing Power Line Disturbance immunity. It provides battery power to preserve processor data in case of a loss of power on all four AC feeds from the utility company. The IBF can hold power briefly over a *brownout*, or for orderly shutdown in case of a longer outage. The IBF provides up

to 10 minutes of full power, depending on the I/O configuration. Table 2-1 shows the IBF hold-up times for configurations with one, two, or three I/O cages.

Table 2-1 IBF estimated power time

Model	I/O configuration		
	One I/O cage	Two I/O cages	Three I/O cages
E12	9 minutes	10 minutes	10 minutes
E26	9 minutes	6 minutes	6 minutes
E40	6 minutes	4.5 minutes	4.5 minutes
E56	4.5 minutes	3.5 minutes	3.5 minutes
E64	4.5 minutes	3.5 minutes	3.5 minutes

The batteries are installed in pairs. Two to six battery units can be installed. The number is determined based on the z10 EC model and power requirements.

- ▶ One or two modular refrigeration units (MRUs), which are air-cooled by their own internal cooling fans
- ▶ CEC cage (see Figure 2-2), which contains up to four books, each with two insulated refrigeration lines to an MRU

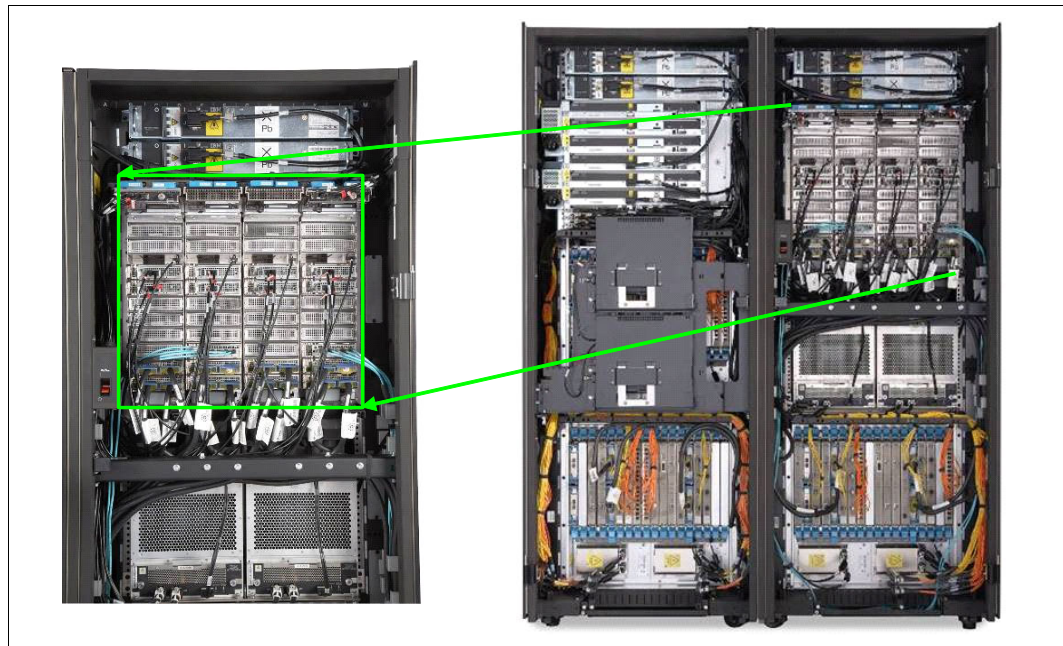


Figure 2-2 The CEC cage

- ▶ I/O cages, which can house all supported types of channel cards
An I/O cage has 28 I/O card slots for installation of ESCON channels, FICON Express8 channels, OSA-Express2, OSA- Express3, Crypto Express2, and Crypto Express3 features. Up to three I/O cages are supported.
- ▶ Air-moving devices (AMD), which provide N+1 redundant cooling for the fanouts, memory, and DCAs.

2.1.2 Frame Z

As shown in Figure 2-1 on page 24, the main components in the frame Z are:

- ▶ Two optional Internal Battery Features (IBFs)
- ▶ Bulk Power Assemblies (BPAs)
- ▶ I/O cage 2 (bottom) and I/O cage 3 (top). Note that both I/O cages are the same as the cage in frame A, and can house all supported types of channel cards.

Frame Z can hold only the bottom cage (I/O cage 2), or both the bottom and top I/O cages (I/O cage 2 and I/O cage 3).

- ▶ The Support Element (SE) tray, located in front of I/O cage 2, contains the two SEs.

2.1.3 I/O cages

There are up to eight dual port fanouts per book for data transfer, each port with bidirectional bandwidth of 6 GBps. The HCA2 and ICB-4 fanouts each drive two ports. Up to 16 InfiniBand (IB) fanout connections provide an aggregated bandwidth of up to 96 GBps per book.

The HCA2-C fanout connects to I/O cages that can contain a variety of channel, Coupling Link, OSA-Express, and Cryptographic feature cards:

- ▶ ESCON channels (16 port cards, 15 usable ports, and one spare)
- ▶ FICON channels (FICON or FCP modes)
 - FICON Express channels (two port cards); carried forward during an upgrade only
 - FICON Express2 channels (four port cards); carried forward during an upgrade only
 - FICON Express4 channels (four port cards); carried forward during an upgrade only
 - FICON Express8 channels (four port cards)
- ▶ ISC-3 links (up to four coupling links, two links per daughter card). Two daughter cards (ISC-D) plug into one mother card (ISC-M).
- ▶ OSA-Express channels:
 - OSA-Express3 10 Gb Ethernet Long Reach and Short Reach (two ports per feature, LR and SR)
 - OSA-Express3 Gb Ethernet (four port cards, LX and SX)
 - OSA-Express3 1000BASE-T Ethernet (four port cards)
 - OSA-Express2 10 Gb Ethernet LR (one port card; carry forward in an upgrade only)
 - OSA-Express2 Gb Ethernet (two port cards, SX, LX, until no longer available)
 - OSA-Express2 1000BASE-T Ethernet (two port cards, until no longer available)
- ▶ Crypto Express2, with two PCI-X adapters per feature. A PCI-X adapter can be configured as a cryptographic coprocessor for secure key operations or as an accelerator for clear key operations.
- ▶ Crypto Express3, with two PCI Express adapters per feature. A PCI Express adapter can be configured as a cryptographic coprocessor for secure key operations or as an accelerator for clear key operations.

ICB-4 channels do not require a slot in the I/O cage and attach directly to the memory bus adapter (MBA) fanout of the server with a bandwidth of 2 GBps.

InfiniBand coupling to a coupling facility is achieved directly from the HCA2-O fanout to the coupling facility with a bandwidth of 6 GBps, or 3 GBps when to a coupling facility on a z9 EC or z9 BC.

The HCA2-O LR fanout supports long distance coupling links for up to 10 km (6.2 miles) or 100 km (62.15 miles) when extended by using DWDM equipment. Supported bandwidths are 5 Gbps (1x IB DDR) and 2.5 Gbps (1x IB SDR), depending on the DWDM equipment used. HCA2-O LR is only available on System z10 (EC and BC).

2.2 Book concept

The central processor complex (CPC) uses a packaging concept based on books. A book contains processor units (PUs), memory, and connectors to I/O cages and other servers. Books are located in the CPC cage in frame A. The z10 EC has at least one book, but may have up to four books installed. A book and its components are shown in Figure 2-3.

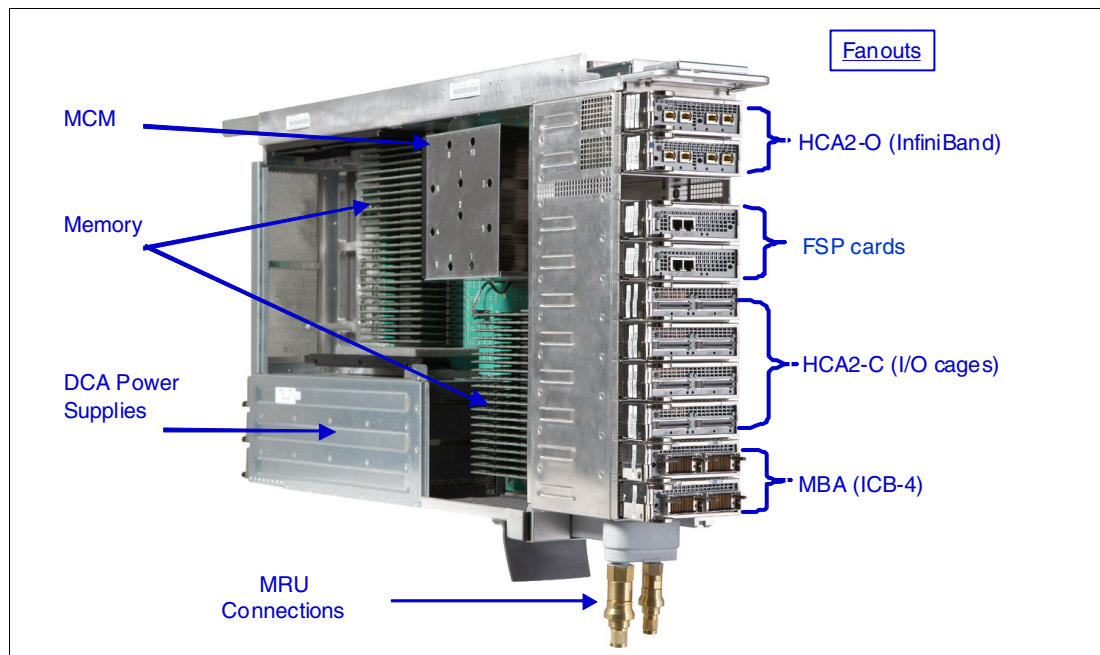


Figure 2-3 Book structure and components

Each book contains:

- ▶ Either 17 or 20 processing units (PUs), depending on the model. The PUs reside on microprocessor chips located on a Multi-Chip Module (MCM).
- ▶ Memory DIMMs plugged into 48 available slots, providing 64 GB to 384 GB of physical memory. The minimum memory in a book is 64 GB, installed in 16 DIMMs of 4 GB each.
- ▶ A combination of up to eight 2-port memory bus adapter (MBA) fanouts, two-port Host Channel Adapter fanouts (HCA2-O, which is optical; HCA2-O LR, which is optical long reach; or HCA2-C, which is copper). These support up to 16 connections to the I/O cages, external coupling links, or ICB-4s.
- ▶ Three Distributed Converter Assemblies (DCAs) that provide power to the book (for 2+1 redundancy).

Up to four books can reside in the CEC cage. Books slide into a mid-plane card that supports up to four books and is located in the top of the CEC cage. The mid-plane card is also the location of two oscillator cards and two external time reference (ETR) cards. The oscillator cards act as a primary and a backup. If the primary oscillator fails, the backup card detects the failure and continues to provide the clock signal so that no outage occurs as a result of oscillator failure. The ETR cards provide two connections to an external time source (Sysplex Timer) and two connections to a Pulse Per Second (PPS) source.

The location of books is indicated in the following list. Table 2-2 indicates the order of book installation and position in cage.

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

Table 2-2 Book installation order and position in cage

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

Book installation from one to four books is concurrent. Concurrent book replacement requires a minimum of two books.

Note: The CEC cage slot locations are important in the sense that in the physical channel ID (PCHID) report, resulting from the IBM configurator tool, locations 01, 06, 10, and 15 are used to indicate whether book features like fanouts and AID assignments relate to the first, second, third, or fourth book in the CEC cage.

2.2.1 Book power

Each book gets its power from three distributed converter assemblies (DCAs) that reside in the book. The DCAs provide the required power for the book. Loss of a DCA leaves enough book power to satisfy its power requirements. The DCAs can be concurrently maintained and are accessed from the rear of the frame.

2.2.2 Cooling

IBM System z10 Enterprise Class is an air-cooled system assisted by refrigeration. Refrigeration is provided by a closed-loop liquid cooling subsystem. The entire cooling subsystem has a modular construction. Besides the refrigeration unit, an air-cooling backup system is in place.

Subsystems

Cooling components and functions are found throughout the cages, and are made up of two subsystems:

- ▶ The modular refrigeration units (MRU)
 - One (or two) MRUs (MRU0 and MRU1), located in the front of the frame A below the books, provide refrigeration to the content of the books together with two large blower assemblies at the rear of the CEC cage, one for each MRU. The assemblies, which are the motor scroll assembly (MSA) and the motor drive assembly (MDA), are connected to the bulk power adapter (BPA) that regulates cooling by increasing the blower speed in combination with an air-moving assembly located in the top part of the CEC cage.
 - A one-book system has MRU0 installed. MRU1 is installed when you upgrade to a two-book system, providing all refrigeration requirements for a four-book system. Concurrent repair of an MRU is possible by taking advantage of the hybrid cooling implementation described in the next section.

- ▶ The motor drive assembly (MDA)

MDAs found throughout the frames provide air cooling where required. They are located at the bottom front of each cage, and in between the CEC cage and I/O cage, in combination with the MSAs.

Hybrid cooling system

IBM System z10 Enterprise Class has a hybrid cooling system that is designed to lower power consumption. Normal cooling is provided by one or two MRUs connected to the evaporator heat sinks mounted on all MCMs in all books.

Refrigeration cooling is the primary cooling source that is backed up by an air-cooling system. If one of the MRUs fails, backup blowers are switched on to compensate for the lost refrigeration capability with additional air cooling. At the same time, the oscillator card is set to a slower cycle time, slowing the system down by up to 17% of its maximum capacity, to allow the degraded cooling capacity to maintain the proper temperature range. Running at a slower clock speed, the MCMs produce less heat. The slowdown process is done in steps, based on the temperature of the books.

Figure 2-4 shows the refrigeration scope of MRU0 and MRU1.

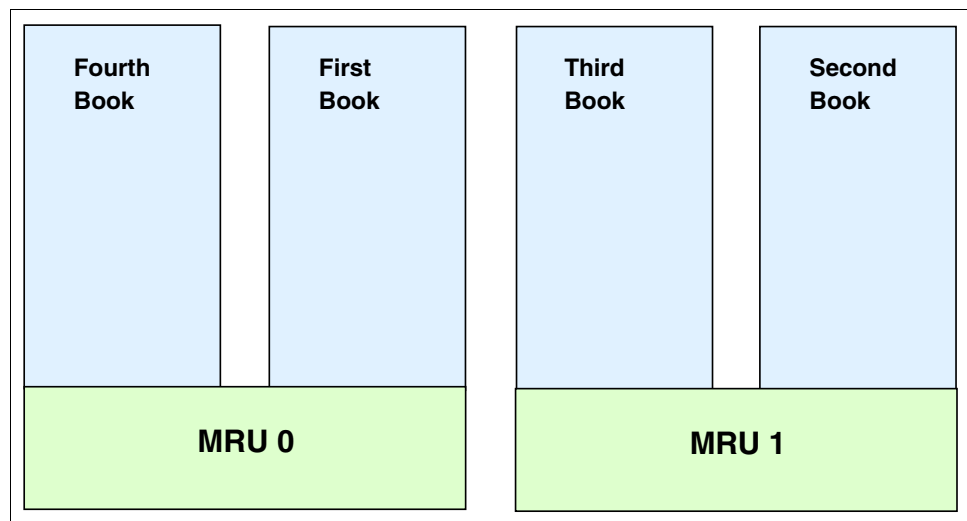


Figure 2-4 MRU scope

2.3 Multi-Chip Module

The Multi-Chip Module (MCM) is a 103-layer glass ceramic substrate (size is 96 x 96 mm) containing seven chip sites and 7356 LGA (land grid array) connections. There are five processor chips and two storage control (SC) chips. Figure 2-5 illustrates the chip locations. The total number of transistors on all chips on the MCM is more than 8 billion.

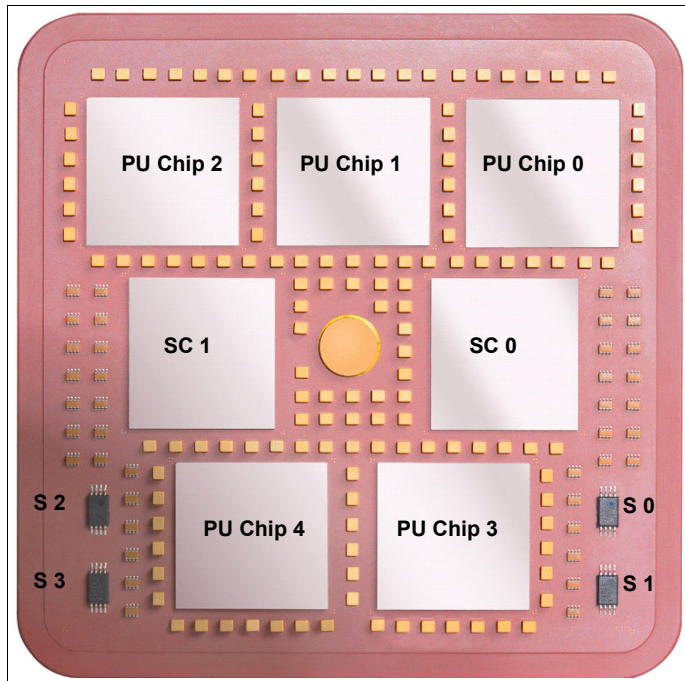


Figure 2-5 z10 EC Multi-Chip Module

The MCM plugs into a card that is part of the book packaging, as shown in Figure 2-6. The book itself is plugged into the mid-plane board to provide interconnectivity between the books, so that a multibook system appears as a symmetric multiprocessor (SMP).

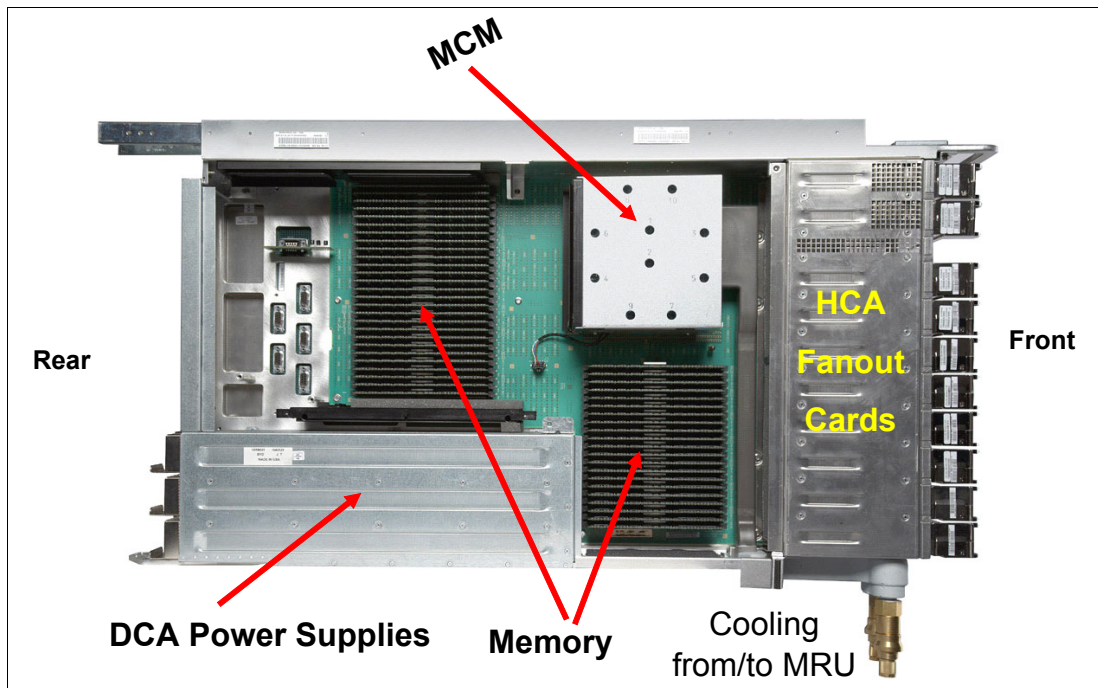


Figure 2-6 Book components

2.4 Processing units and storage control chips

Both processing unit (PU) and storage control (SC) chips on the MCM use CMOS 11s chip technology. CMOS 11s is state-of-the-art microprocessor technology based on ten-layer copper interconnections and silicon-on insulator technologies. The chip lithography line width is $0.065 \mu\text{m}$ (65 nm). On the MCM, four Serial Electrically Erasable Programmable ROM (SEEPRM) chips, identified as S0, S1, S2, and S3 in Figure 2-5 on page 30, are rewritable memory chips that hold data without power, use the same technology, and are used for retaining product data for the MCM and relevant engineering information.

2.4.1 PU chip

Each PU chip is a four-core (quad-core) chip. There are five PU chips on each MCM. The five PU chips come in two versions. For models E12, E26, E40, and E56, the processor units on the MCM in each book are implemented with a mix of three active cores and four active cores per chip (3 x 3 cores active, plus 2 x 4 cores active) resulting in 17 active cores per MCM. All MCMs in all models that we have discussed have 17 active cores. This means that Model E12 has 17, Model E26 has 34, Model E40 has 51, and Model E56 has 68 active PUs.

For the Model E64, the PUs on the MCM in the first book are implemented with 17 active cores (3 x 3 cores active, plus 2 x 4 cores active), plus three MCMs with 20 active cores (5 x 4 cores active). This means that there are 77 active PUs.

A schematic representation of the PU chip is shown in Figure 2-7. The four PUs (cores) are shown in each corner and include the L1 and L1.5 caches plus all microprocessor functions. In the figure, each of the two coprocessors (COP) is shared by two of the four cores. The coprocessors are accelerators for data compression and cryptographic functions.

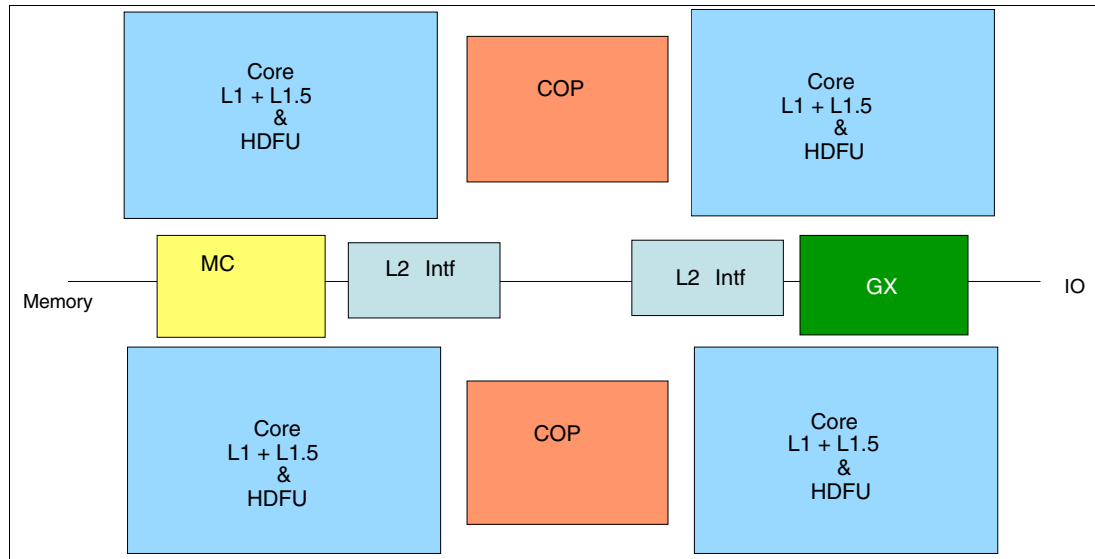


Figure 2-7 PU chip

The L2 cache interface (L2 Intf) is shared by all four cores. MC indicates the memory controller (MC) function controls access to memory. GX indicates the I/O bus controller that controls the interface to the host channel adapters accessing the I/O. The chip controls traffic between the microprocessors (cores), memory, I/O, and the L2 cache on the SC chips.

2.4.2 Processing unit (core)

The following functional areas are on each core (their locations on the core are shown in Figure 2-8 on page 33).

- ▶ Instruction fetch unit (IFU)

The IFU contains the instruction cache, branch prediction logic, instruction fetching controls and buffers. Its relative size is the result of the elaborate branch prediction design, which is further described in 3.3.1, “Superscalar processor” on page 67.
- ▶ Instruction decode unit (IDU)

The IDU is fed from the IFU buffers and is responsible for parsing and decoding of all z/Architecture operation codes.
- ▶ Load-store unit (LSU)

The LSU contains the data cache and is responsible for handling all types of operand accesses of all lengths, modes and formats as defined in the z/Architecture.
- ▶ Translation unit (XU)

The XU has a large translation look-aside buffer (TLB) and the Dynamic Address Translation (DAT) function that handles the dynamic translation of logical to physical addresses.
- ▶ Fixed-point unit (FXU)

The FXU handles fixed point arithmetic.

- ▶ Binary floating-point unit (BFU)

The BFU handles all binary and hexadecimal floating-point, and fixed-point multiplication and division operations.
- ▶ Decimal floating-point unit (DFU)

The DFU executes both floating- point and fixed-point decimal operations.
- ▶ Recovery unit (RU)

The RU keeps a copy of the complete state of the system, including all registers, collects hardware fault signals, and manages the hardware recovery actions.

Each core on the chip runs at a cycle time of 0.227 nanoseconds (4.4 GHz). Each quad-core PU chip measures 21.97 x 21.17 mm, contains 6 km of wire, features 1188 signal and 8765 I/O connections, and has close to one billion (994 million) transistors.

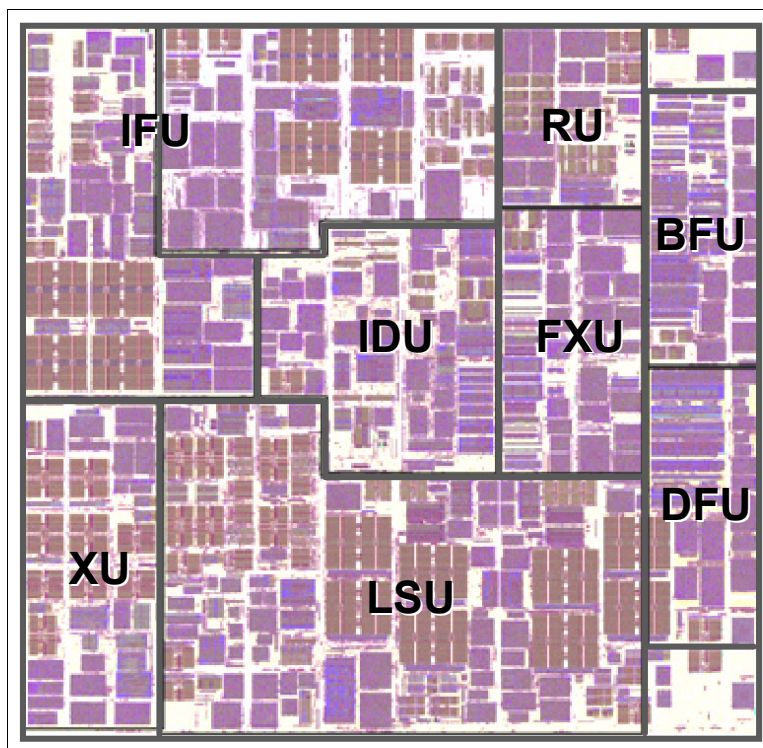


Figure 2-8 PU (core) layout

In each MCM, 12 to 16 available PUs may be characterized for customer use. Up to three SAPs may reside in an MCM, depending on the model and the book in which they reside. Throughout the system, two spare PUs (cores) are available that may be allocated on any MCM in the system. Up to two spare PUs (cores) may be allocated on an MCM.

The following list summarizes the PU distribution in each of the models, listed in detail for each MCM in Table 2-3.

- ▶ Model E12 has one MCM with 17 PUs, of which 12 can be characterized. The five remaining PUs are three system assist processors and two spares.
- ▶ Model E26 has two MCMs with 17 PUs in each MCM for a total of 34 PUs, of which 26 can be characterized. The eight remaining PUs are six system assist processors, three in each book, and two spares, one in each book.
- ▶ Model E40 has three MCMs with 17 PUs in each MCM for a total of 51 PUs, of which 40 can be characterized. The eleven remaining PUs are nine system assist processors, three in each book, and two spares, one in the first book and one in the second book.
- ▶ Model E56 has four MCMs with 17 PUs in each MCM for a total of 68 PUs, of which 56 can be characterized. The 12 remaining PUs are 10 system assist processors, three in the first, second, and third books, and one in the fourth book, plus two spares in the fourth book.
- ▶ Model E64 has four MCMs with 17 PUs in the first MCM, 20 PUs in each of the other three for a total of 77 PUs, of which 64 can be characterized. In Model E64, each MCM has 16 PUs available for customer use. The 13 remaining PUs are 11 system assist processors, which includes one in the first book, three in the second and third books, and four in the fourth book, and one spare in the second and third books.

Table 2-3 Model summary

Models	First book				Second book				Third book				Fourth book			
	Available PUs	SAPs	Spare	MCM size	Available PUs	SAPs	Spare	MCM size	Available PUs	SAPs	Spare	MCM size	Available PUs	SAPs	Spare	MCM size
E12	12	3	2	17	-	-	-	-	-	-	-	-	-	-	-	-
E26	13	3	1	17	13	3	1	17	-	-	-	-	-	-	-	-
E40	13	3	1	17	13	3	1	17	14	3	0	17	-	-	-	-
E56	14	3	0	17	14	3	0	17	14	3	0	17	14	1	2	17
E64	16	1	0	17	16	3	1	20	16	3	1	20	16	4	0	20

Each PU has a 192 KB on-chip Level 1 cache (L1) that is split into a 64 KB L1 cache for instructions (I-cache) and a 128 KB L1 cache for data (D-cache). A second level on chip cache, the L1.5 cache, has a size of 3 MB per PU. The two levels of on-chip cache structure are necessary for optimizing performance so that cache is tuned to the high-frequency properties of each of the microprocessors (cores).

2.4.3 SC chip

The MCM has two SC chips. The L1 and L1.5 PU caches communicate with the L2 caches (SC chips) by five bidirectional 16-byte data buses. The bus/clock ratio of 1.5:1 between the L2 cache and the PU is controlled by the storage controller on the SC chip.

The SC chip also acts as an L2 cache cross-point switch for L2-to-L2 traffic to up to three remote MCMs or books by three bidirectional 16-byte data buses with a 3:1 bus/clock ratio. The SC chip measures 21.11 x 21.71 mm and has 1.6 billion transistors. The L2 SRAM cache

size on the SC chip measures 24 MB, resulting in a combined L2 cache size of 48 (2 x 24) MB per book. The clock function is distributed among both SC chips, and the wire length of the chip is to 3 km.

Figure 2-9, a schematic representation of the SC chip, is shown with the various elements of the SC chip. Most of the space is taken by the SRAM L2 cache. L2C indicates the controller function of the chip for point-to-point inter-book communication. Directory and addressing function locations are shown also.

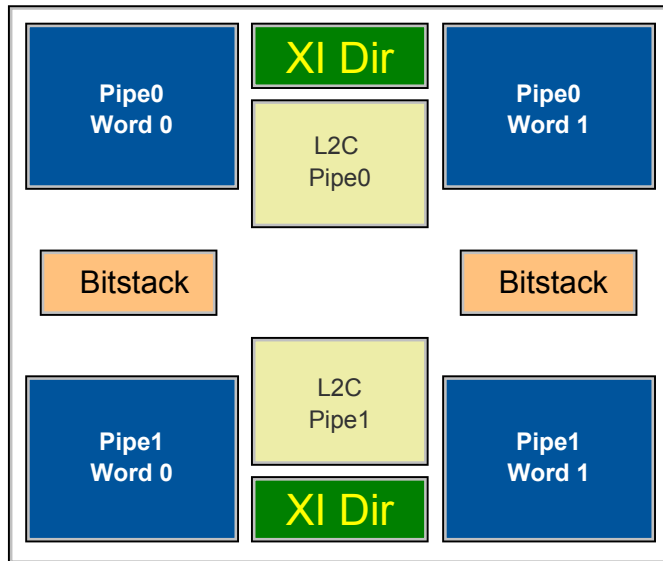


Figure 2-9 SC chip

2.5 Memory

Maximum physical memory size is directly related to the number of books in the system. Each book may contain up to 384 GB of physical memory, for a total of 1536 GB of installed memory. You may purchase up to 1520 GB of physical memory (4 books x 384 GB minus 16 GB reserved for HSA). The 16 GB HSA memory is managed separately from customer memory.

Memory in a book is organized in two logical pairs:

- ▶ Logical pair 0: Memory control unit 0 (MCU 0) and MCU 1 each control three groups of four DIMMs.
- ▶ Logical pair 1: MCU 2 and MCU 3 each control three groups of four DIMMs.

The DIMM size is controlled by a logical MCU pair (4 GB or 8 GB), and DRAM technology must be the same (for example, you cannot mix 1 Gb and 2 Gb DRAMs). In addition, the total memory capacity for each logical MCU pair must be the same. The logical view is shown in Figure 2-10.

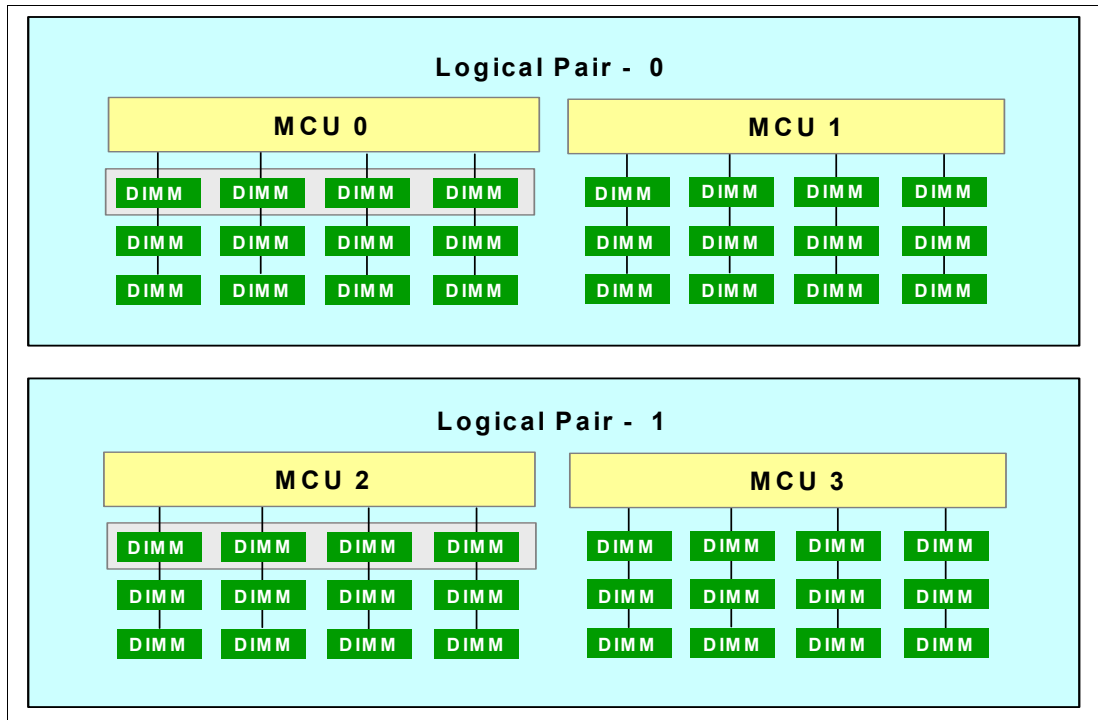


Figure 2-10 Memory logical pairs and MCUs

Memory sizes in each book do not have to be similar. Different books can contain different amounts of memory. However, the total memory capacity for each MCU within a logical pair must be the same. The minimum initial amount of memory that can be ordered is 16 GB for all models.

The IBM System z10 Enterprise Class uses 4 GB and 8 GB DIMMs. Physical memory sizes up to 192 GB can be achieved by using 4 GB DIMMs. Above that amount, 8 GB DIMMs are installed.

Figure 2-11 on page 37 shows how the 48 DIMM slots are organized in a book. There are two banks, one with 27 slots (MD1 - MD27) and one with 21 slots (MD 28 - MD48). The location of the DIMM slots of each of the four MCUs are identified. The first row (a) of each of the three groups of four DIMMs per MCU is the master DIMM.

The master DIMM is a buffered DIMM that redrives address, control, and data from DIMM to DIMM. In Figure 2-11, the master and subordinate DIMMs are shown in blue (dark) color.

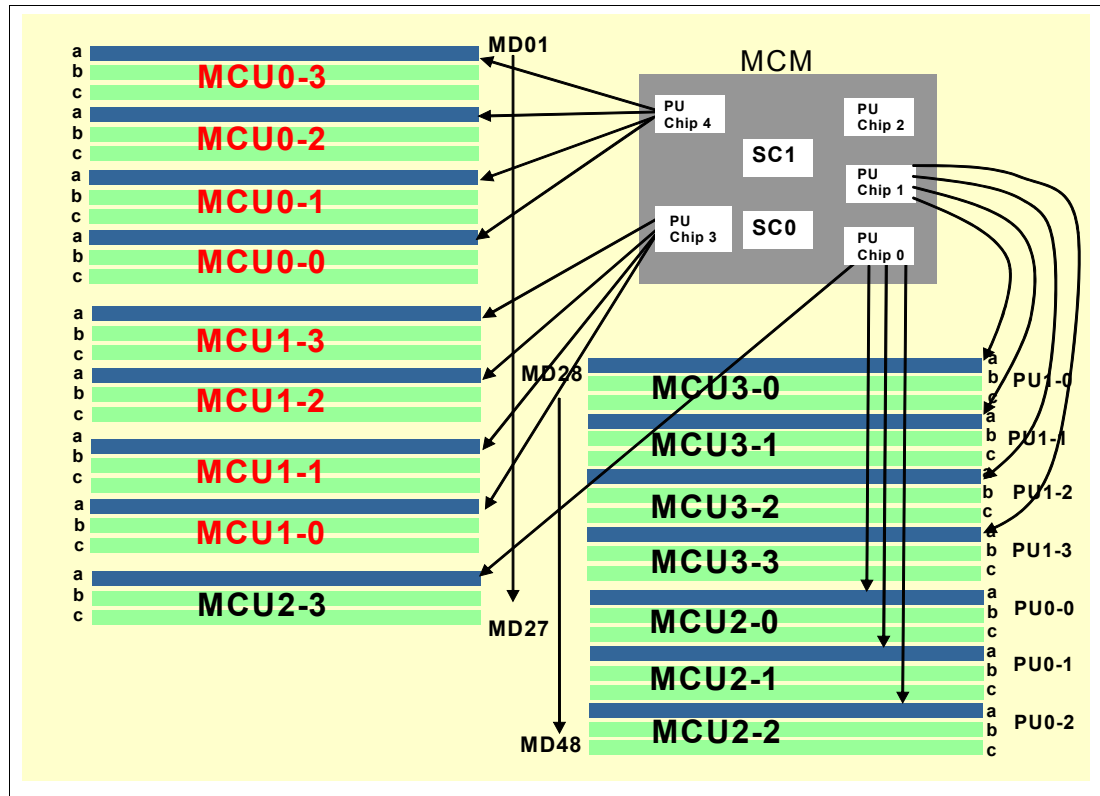


Figure 2-11 Physical DIMM slot layout on book

Four of the five PU chips on the MCM use their memory I/O capabilities to connect to the master DIMMs of an MCU. PU chip 0 connects to the four master DIMMs of MCU 2, PU chip 1 to MCU3, PU chip 3 to MCU 1, and PU chip 4 to MCU 0. PU chip 2 does not connect to memory.

Memory can be purchased in increments of 16 GB up to a total size of 256 GB. From 256 GB, the increment size doubles to 32 GB until 512 GB. From 512 GB to 944 GB, the increment is 48 GB, Beyond 512 GB, up to 1520 GB, an increment of 64 GB is used. Memory may be ordered as follows:

- ▶ A one-book system (model E12) may contain from 64 GB up to 384 GB of physical memory. Memory may be ordered in 16 GB or 32 GB increments up to 352 GB.
- ▶ A two-book system (model E26) may contain from 128 GB up to 768 GB of physical memory. Memory may be ordered in 16 GB, 32 GB, or 48 GB increments up to 752 GB.

- ▶ A three-book system (model E40) may contain from 192 GB up to a maximum of 1152 GB of physical memory. Memory may be ordered in 16 GB, 32 GB, 48 GB, or 64 GB increments up to 1136 GB.
- ▶ A four-book system (model E56 or model S64) may contain from 288 GB up to a maximum of 1536 GB of physical memory. Memory may be ordered in 16 GB, 32 GB, 48 GB, or 64 GB increments up to 1520 GB.

Note: The maximum amount of memory that can be ordered for each of the models *is not* equal to the maximum supported amount of physical memory. This is because 16 GB of physical memory is set aside for the hardware system area (HSA).

Physically, memory is organized as follows:

- ▶ A book always contains a minimum of 16 DIMMs of 4 GB each (64 GB).
- ▶ A book may have more memory installed than enabled. The unused amount of memory can be enabled by a Licensed Internal Code (LIC) code load when required.
- ▶ Memory upgrades are satisfied from already-installed unused memory capacity until this is exhausted. When no more unused memory is available from the installed memory cards (DIMMs), one of the following additions must occur:
 - Memory cards have to be upgraded to a higher capacity.
 - An additional book with additional memory is necessary.
 - Memory cards (DIMMs) must be added.

Note: The amount of memory available for use is the sum of all enabled physical memory on all memory DIMMs in all books.

When activated, a logical partition can use memory resources located in any book. No matter in which book the memory resides, a logical partition has access to that memory for up to a maximum one TB. Despite the book structure, the z10 EC is still a symmetric multiprocessor (SMP).

A memory upgrade is concurrent when it requires no change of the physical memory cards. A memory card change is disruptive when no use is made of Enhanced Book Availability. See 2.6.2, “Enhanced book availability” on page 44.

2.5.1 Memory RAS

Error detection and recovery in the memory subsystem hardware is implemented by error correction code (ECC) and is protected by Chipkill technology. Chipkill is an advanced error correction mechanism that corrects multi-bit memory errors. If a chip fails or exceeds a bit error threshold, the affected chip is taken out of commission and replaced by a spare chip. The connection between the memory DIMMs and the memory controller is protected by ECC. This ECC provides failure protection for virtually every type of packet transfer failure that can be corrected spontaneously; the data portion of the packet-transfers can benefit because the transfers are now protected.

2.5.2 Memory upgrades

For a model upgrade that results in the addition of a book, the minimum memory increment is added to the system. As previously mentioned, the minimum physical memory size in a book is 64 GB. During a model upgrade, the addition of a book is a concurrent operation. The addition of the physical memory that is in the added book is also concurrent. If all or part of

the additional memory is enabled for installation use (if it has been purchased), it becomes available to an active logical partition if this partition has reserved storage defined. For more information, see 3.6.2, “Reserved storage” on page 99. Alternately, additional memory may be used by an already-defined logical partition that is activated after the memory addition.

2.5.3 Book replacement and memory

With enhanced book availability as supported for z10 EC (see 2.6.2, “Enhanced book availability” on page 44), sufficient resources must be available to accommodate resources that are lost when a book is removed for upgrade or repair. Most of the time, removal of a book results in removal of active memory. With the flexible memory option (see 2.5.4, “Flexible memory option” on page 39), evacuating the affected memory and reallocating its use elsewhere in the system is possible. This requires additional available memory to compensate for the memory lost with the removal of the book.

2.5.4 Flexible memory option

With the flexible memory option, sufficient inactive memory resources are made available for use when replacing a book. When ordering memory, you may request additional flexible memory. For example, on a Model E26, when 352 GB is ordered and flexible memory is requested, the result is that one book’s worth of memory (384 GB) is set aside as flexible memory. The order results in the installation of 768 GB of physical memory. Of this amount, 384 GB is set aside for flexible memory, 16 GB for HSA, and 16 GB as a result of the rounding for the increment size of 32 GB, resulting in 352 GB for use (as ordered) on this Model E26. The equation is:

$$768 - 384 - 16 - 16 = 352$$

For a four-book system, such as Model E56, an order of 512 GB results in the installation of 736 GB of physical storage. Of this amount, 192 GB is set aside for flexible memory, 16 GB for HSA, and 16 GB as a result of the rounding for the increment size of 32 GB. This results in 512 GB for use (as ordered) on the Model E56. The equation is:

$$736 - 192 - 16 - 16 = 512$$

Both examples show that sufficient memory is set aside so that the maximum content of memory in a book to be removed can be moved to the excess memory in the remaining books. Flexible memory can be purchased but cannot be used for normal everyday use. For that reason, a different purchase price for the flexible memory is offered to increase the overall availability of the system. Flexible memory granularity follows the same increment scheme as standard memory:

- ▶ Increments from 16 GB to 384 GB are 16 GB. For flexible memory, the same increment is used.
- ▶ Increments from 384 GB to 752 GB are 32 GB. For flexible memory, the same increment is used.
- ▶ Increments from 752 GB to 1520 GB are 64 GB. For flexible memory, the same increment is used from 752 GB to 1136 GB.

2.5.5 Plan-ahead memory

Plan-ahead memory provides the ability to plan for nondisruptive permanent memory upgrades. It differs from the flexible memory option. The flexible memory option is meant to

anticipate nondisruptive bank replacement. The usage of flexible memory is therefore temporary, in contrast with plan-ahead memory.

When preparing in advance for a future memory upgrade, note that memory can be pre-plugged, based on a target capacity. The pre-plugged memory can be made available through a LIC Configuration Code (LICCC) update. You may order this LICCC through

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

The installation and activation of any pre-planned memory requires the purchase of the required feature codes (FC), described in table Table 2-4.

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.

Table 2-4 Feature codes for plan-ahead memory

Memory	z10 EC feature code
Pre-planned memory Charged when physical memory is installed. Used for tracking the quantity of physical increments of plan-ahead memory capacity.	FC1996
Pre-planned memory activation Charged when plan-ahead memory is enabled. Used for tracking the quantity of increments of plan-ahead memory that is being activated.	FC1997

Installation of pre-planned memory is done by ordering FC1996. The ordered amount of plan-ahead memory is charged with a reduced price compared to the normal price for memory.

Activation of installed pre-planned memory is achieved by ordering FC1997 that causes the the other portion of the previously charged price to be invoiced.

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

2.6 Connectivity

Connections to I/O cages, coupling facilities, and ICB-4 links are driven from the memory bus adapters and host channel adapter fanouts that are located on the front of the book.

Figure 2-12 shows the location of the fanouts and connectors for a two-book system. In the figure, ETR is external time reference card; OSC is oscillator card; FSP is flexible support processor; and LG is location code for logic card. See *System z10 Enterprise Class Service Guide*, GC28-6866.

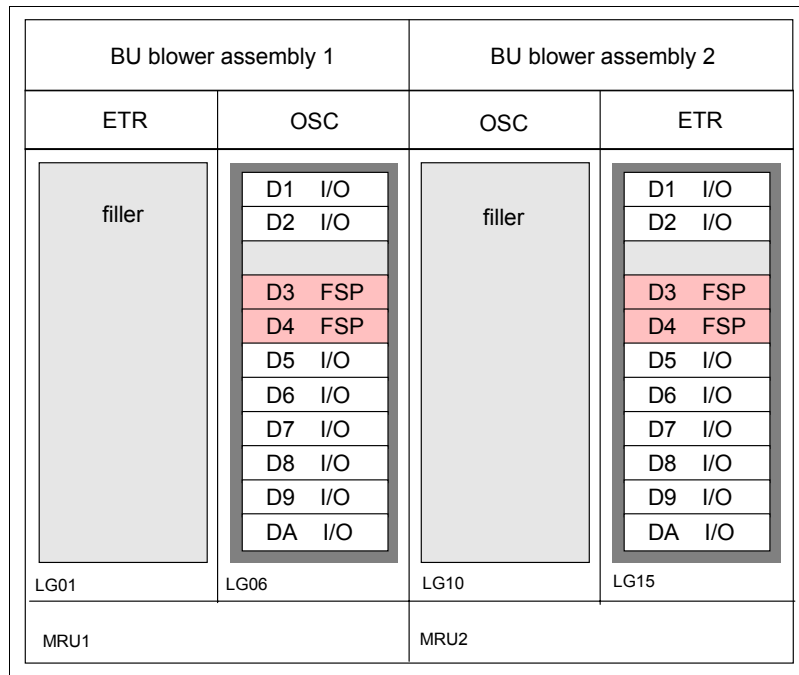


Figure 2-12 Location of the host channel adapter fanouts

Each book has up to eight fanouts (numbered D1, D2, and D5 through DA), each driving two InfiniBand connector cables, resulting in up to 16 physical connections per book.

Figure 2-12 shows a two-book system without fanouts in all D1 and D2 positions. A fanout can be repaired concurrently with the use of redundant I/O interconnect. See 2.6.1, “Redundant I/O interconnect” on page 44.

The four types of two-port fanouts are:

- ▶ Host Channel Adapter2-C (HCA2-C) provides copper connections for InfiniBand I/O interconnect to all I/O, ISC-3, and Crypto Express cards in I/O cages.
- ▶ Host Channel Adapter2-O (HCA2-O) provides optical connections for InfiniBand I/O interconnect for coupling links (PSIFB). The HCA2-O provides a point-to-point connection over a distance of up to 150 m (492 feet), using four 12x MPO fiber connectors and OM3 fiber optic cables (50/125 μm).

System z10 to System z10 connections use 12-lane InfiniBand Double Data Rate (12 x IB-DDR) link at 6 GBps. If the connection is from System z10 to a System z9, 12-lane InfiniBand Single Data Rate (12 x IB-SDR) at 3 GBps is used.

- ▶ The HCA2-O LR fanout supports PSIFB Long Reach (PSIFB LR) coupling links for distances of up to 10 km and up to 100 km when repeated through a DWDM. This fanout is supported on System z10 only.

PSIFB LR coupling links operate at up to 5.0 Gbps (1x IB-DDR) between two z10 servers, or automatically scales down to 2.5 Gbps (1x IB-SDR) depending on the capability of the attached equipment.

Note: The InfiniBand link data rates (6 GBps, 3 GBps, 5 GBps, or 2.5 GBps) do not represent the actual performance of the link. The actual performance depends on several factors, such as latency, cable lengths, and the type of workload. Although the link data rates are higher than that of ICB-4, or ISC-3, higher service times can result because the actual throughput might be less than with ICB-4 or ISC-3 links.

- ▶ The MBA fanout provides up to two ICB-4 links at a rate of 2 GBps to z10, z9, z990, and z890.

Figure 2-13 presents a front view of a processor book that is fully populated with fanout cards.

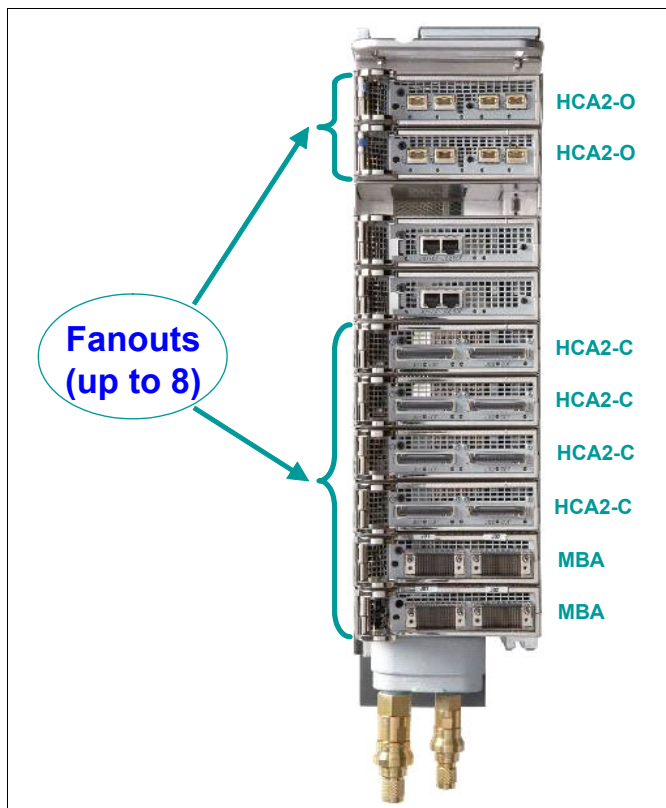


Figure 2-13 Fanout cards

Note the following information about models and fanout positions:

- ▶ On a model E12, all fanout positions may be populated, for up to 16 I/O connections of any type. On a model E26, all fanout positions may also be populated, for up to 32 I/O connections.
- ▶ On a model E40, all fanout positions may be populated only on the first book. Positions D1 and D2 must remain free of fanouts on both the second and third books, for up to 40 I/O connections of any type.
- ▶ On models E56 and E64, all D1 and D2 positions must remain free of any fanout. This results in up to 48 I/O connections of any type.

Figure 2-14 shows the InfiniBand connectors used for each of the three HCA types and the MBA connector. From left to right, HCA types are HCA2-O LR, HCA2-O, HCA2-C.

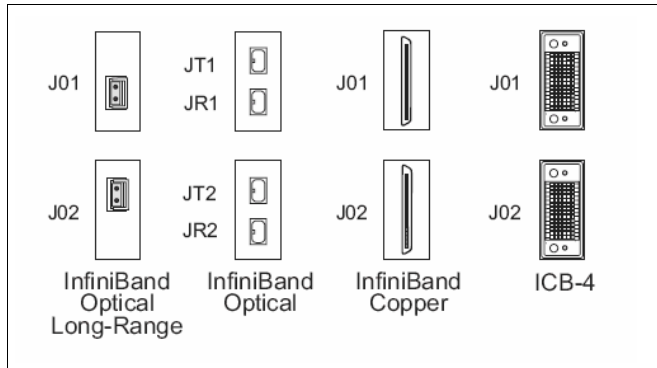


Figure 2-14 InfiniBand (HCA2) and MBA (ICB-4) connectors

Up to two InfiniBand connector cables can be connected to a fanout. When configuring for availability, channels, links, and OSAs should be balanced across books. In a system configured for maximum availability, alternate paths maintain access to critical I/O devices, such as disks, networks, and so on.

Enhanced book availability allows a single book in a multibook server to be concurrently removed and reinstalled for an upgrade or a repair. Removing a book means that the connectivity to the I/O devices connected to that book is lost. To prevent connectivity loss, the redundant I/O interconnect feature allows you to maintain connection to critical devices, except for ICB-4s and Parallel Sysplex InfiniBand coupling (PSIFB), when a book is removed.

In the configuration report, fanouts are identified by their locations in the CPC drawer. Fanout locations are numbered from D3 through D8. The jacks are numbered J01 and J02 for each HCA2-C, HCA2-O LR, or ICB-4 fanout port. Jack numbering for HCA2-O fanout ports for transmit and receive jacks is JT1 and JR, and JT2 and JR2 (Figure 2-14).

2.6.1 Redundant I/O interconnect

Redundant I/O interconnect is accomplished by the facilities of the InfiniBand I/O connections to the InfiniBand Multiplexer (IFB-MP) card. Each IFB-MP card is connected to a jack located in the InfiniBand fanout of a book. IFB-MP cards are half-high cards and are interconnected with cards called STI-A8 and STI-A4, allowing redundant I/O interconnect in case the connection coming from a book ceases to function, as happens when, for example, a book is removed. A conceptual view of how redundant I/O interconnect is accomplished is shown in Figure 2-15.

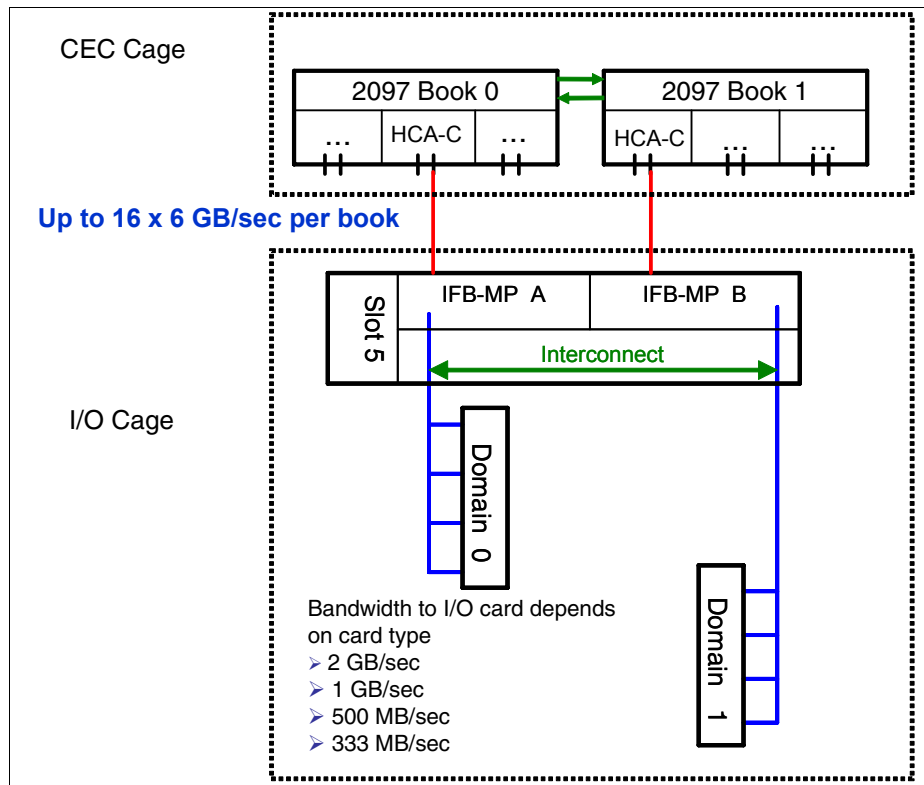


Figure 2-15 Redundant I/O interconnect

Normally, the HCA2-C fanout in the first book connects to the IFB-MP (A) card and services domain 0 (I/O cage slots 01, 03, 06, and 08). In the same fashion, the HCA2-C fanout of the second book connects to the IFB-MP (B) card and services domain 1 (I/O cage slots 02, 04, 07, and 09). If the second book is removed, or the connections from the second book to the cage are removed, connectivity to domain 1 is maintained by guiding the I/O to domain 1 through the interconnect between IFB-MP (A) and IFB-MP (B).

In configuration reports, books are identified by their location in the CEC cage. HCA2-C fanouts are numbered from D1, D2, and D5 to DA. The jacks are numbered J01 and J02 for each HCA2-C fanout port. Jack numbering for HCA2-O fanout ports is JT1, JR1, and JT2 JR2 for transmit and receive jacks, respectively.

2.6.2 Enhanced book availability

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

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

Removal of a book also removes the book connectivity to the I/O cages. The impact of the removal of the book on the system is limited by the use of redundant I/O interconnect, which is described in 2.6.1, “Redundant I/O interconnect” on page 44. However, all PSIFBs and ICBs on the removed book must be configured offline.

If the enhanced book availability and flexible memory options are *not* used when a book must be replaced (for example because of an upgrade or a repair action), the memory in the failing book is removed also. Until the removed book is replaced, a power-on reset of the server with the remaining books is supported.

2.6.3 Book upgrade

All fanouts used for I/O and HCA fanouts used for PSIFB are concurrently rebalanced as part of the book addition. However, to use fanouts for rebalancing ICBs, order the STI rebalance feature (FC 2400). Although the rebalance feature is disruptive, it is highly recommended when you upgrade from model E12 to a larger model. When upgrading from a multiple book to a larger model, we recommend performing a thorough evaluation. See “STI rebalance (FC2400)” on page 119.

2.7 Model configurations

The z10 EC model nomenclature is based on the number of PUs available for customer use in each configuration. The following five models are available:

- | | |
|------------------|--|
| Model E12 | 12 PUs are available for characterization as CPs, IFLs, and ICFs; up to six zAAPs and zIIPs; or up to three additional SAPs. |
| Model E26 | 26 PUs are available for characterization as CPs and IFLs; up to 16 ICFs; up to 13 zAAPs and zIIPs; or up to 7 additional SAPs. |
| Model E40 | 40 PUs are available for characterization as CPs and IFLs; up to 16 ICFs; up to 20 zAAPs and zIIPs; or up to 11 additional SAPs. |
| Model E56 | 56 PUs are available for characterization as CPs and IFLs; up to 16 ICFs; up to 28 zAAPs and zIIPs; or up to 18 additional SAPs. |
| Model E64 | 64 PUs are available for characterization as CPs and IFLs; up to 16 ICFs; up to 32 zAAPs and zIIPs; or up to 21 additional SAPs. |

The models are summarized in Table 2-5. In the table, Opt indicates optional.

Table 2-5 System z10 configurations

Model	Books	PUs per MCM	Active PUs			zAAPs	zIIPs	Opt. SAPs	Base SAPs	Spares
			CPs	IFLs/uIFL	ICFs					
E12	1	17	0–12	0–12	0–12	0–6	0–6	0–3	3	2
E26	2	17	0–26	0–26	0–16	0–13	0–13	0–7	6	2
E40	3	17	0–40	0–40	0–16	0–20	0–20	0–11	9	2
E56	4	17	0–56	0–56	0–16	0–28	0–28	0–18	10	2
E64	4	17/20	0–64	0–64	0–16	0–32	0–32	0–21	11	2

When a z10 EC order is configured, PUs are characterized according to their intended usage. They can be ordered as any of the following items:

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

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

IFL The Integrated Facility for Linux is a processor that is purchased and activated for use by the z/VM for Linux guests and Linux on System z operating systems.

Unassigned IFL A processor 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 purchased and activated for use by the Coupling Facility Control Code.

zAAP A System z10 Application Assist Processor (zAAP) purchased and activated to run Java code under control of z/OS JVM¹ or z/OS XML System Services.

zIIP A System z10 Integrated Information Processor (zIIP) purchased and activated to run eligible workloads such as DB2 DRDA or z/OS¹ Communication Server IPsec.

Additional SAP An optional processor that is purchased and activated for use as a system assist processor (SAP).

A capacity marker identifies that a certain number of CPs have been purchased. This number of purchased CPs is higher than or equal to the number of CPs actively used. The capacity marker marks the availability of purchased but unused capacity 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 MSU value that is used for charging MLC software, or when charged on a per-processor basis.

Unassigned IFLs are those that are purchased for the intention to be used as future IFLs, and usually have this unassigned status for charging software and maintenance. Unassigned IFLs do not count in establishing the charges for either z/VM or Linux.

¹ z/VM V5 R3 supports zAAP and zIIP processors for guest exploitation.

This charging method prevents request for price quotation (RPQ) handling in case a temporary downgrade is required. When the capacity need arises, the marked CPs and unassigned IFLs can be assigned nondisruptively.

2.7.1 Upgrades

Concurrent CP, IFL, ICF, zAAP, zIIP, or SAP upgrades are done within a z10 EC. Concurrent upgrades require available PUs. Spare PUs are used to replace defective PUs. The number of spare PUs depends on machine model. Concurrent processor upgrades require that additional be PUs installed (at a prior time) but not activated.

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

Although upgrades from one z10 EC model to another z10 EC model are concurrent, meaning that one or more books can be added, there is one exception. Upgrades from any z10 EC (model E12, E26, E40, and E56) to a model E64 is disruptive because this upgrade requires the addition or replacement of three books. Table 2-6 shows the possible upgrades within the z10 EC configuration range.

Table 2-6 z10 EC to z10 EC upgrade paths

To 2097 From 2097	Model E12	Model E26	Model E40	Model E56	Model E64 ^a
Model E12	-	Yes	Yes	Yes	Yes
Model E26	-	-	Yes	Yes	Yes
Model E40	-	-	-	Yes	Yes
Model E56	-	-	-	-	Yes

a. Disruptive upgrade

You may also upgrade a System z9 or z990 to a System z10 EC, preserving the server serial number (S/N). The I/O cards are also moved up (with certain restrictions).

Note: Upgrades from System z9 and z990 are disruptive.

Upgrade paths from any z9 EC to any z10 EC are supported as listed in Table 2-7.

Table 2-7 z9 EC to z10 EC upgrade paths

To 2097 From 2094	Model E12	Model E26	Model E40	Model E56	Model E64
Model S08	Yes	Yes	Yes	Yes	Yes
Model S18	Yes	Yes	Yes	Yes	Yes
Model S28	Yes	Yes	Yes	Yes	Yes
Model S38	Yes	Yes	Yes	Yes	Yes
Model S54	Yes	Yes	Yes	Yes	Yes

Upgrades from any z990 to any z10 EC are supported as listed in Table 2-8 on page 48.

Table 2-8 z990 to z10 EC upgrade paths

To 2097 From 2084	Model E12	Model E26	Model E40	Model E56	Model E64
Model A08	Yes	Yes	Yes	Yes	Yes
Model B16	Yes	Yes	Yes	Yes	Yes
Model C24	Yes	Yes	Yes	Yes	Yes
Model D32	Yes	Yes	Yes	Yes	Yes

A z10 BC can be upgraded to a z10 EC model E12.

2.7.2 PU characterization

A minimum of one PU characterized as a CP, IFL, or ICF is required per system. The maximum number of CPs is 64, the maximum number of IFLs is 64, and the maximum number of ICFs is 16. The maximum number of zAAPs is 32, but requires an equal or greater number of characterized CPs. The maximum number of zIIPs is also 32 and requires an equal or greater number of characterized CPs. The sum of all zAAPs and zIIPs cannot be larger than two times the number of characterized CPs.

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

2.7.3 Concurrent PU conversions

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

Most conversions are not disruptive. In exceptional cases, the conversion can be disruptive, for example, when a model E12 with 12 CPs is converted to an all IFL system. In addition, a logical partition might be disrupted when PUs must be freed before they can be converted. Conversion information is summarized in Table 2-9.

Table 2-9 Concurrent PU conversions

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

2.7.4 Model capacity identifier

To recognize how many PUs are characterized as CPs, the store system information (STSI) instruction returns a value that can be seen as a model capacity identifier (MCI), which determines the number and speed of characterized CPs. Characterization of a PU as an IFL, an ICF, a zAAP, or a zIIP is not reflected in the output of the STSI instruction, because these have no effect on software charging. More information about the STSI output is in “Processor identification” on page 274.

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

- ▶ For full-capacity engines, model capacity identifiers 701 to 764 are used. They express the 64 possible capacity settings from one to 64 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 twelve CPs are characterized. These three subcapacity settings applied to up to twelve CPs offer 36 additional capacity settings. See “Granular capacity” on page 49.

Granular capacity

The z10 EC offers 36 capacity settings at the low end of the processor. Only 12 CPs can have granular capacity. When subcapacity settings are used, other PUs, beyond 12, can only be characterized as specialty engines.

The three defined ranges of subcapacity settings have model capacity identifiers numbered from 401 to 412, 501 to 512, and 601 to 612.

Note: Within a z10 EC, all CPs have the same capacity identifier. IFLs and specialty engines operate at full speed.

List of model capacity identifiers

Table 2-10 shows that regardless of the number of books, a configuration with one characterized CP is possible. For example, model E64 may have only one PU characterized as a CP.

Table 2-10 Model capacity identifiers

z10 EC	Model capacity identifier
Model E12	701–712, 601–612, 501–512, 401–412
Model E26	701–726, 601–612, 501–512, 401–412
Model E40	701–740, 601–612, 501–512, 401–412
Model E56	701–756, 601–612, 501–512, 401–412
Model E64	701–764, 601–612, 501–512, 401–412

Note: Model capacity identifier 700 is used for IFL or ICF only configurations.

2.7.5 Model capacity identifier and MSU values

All model capacity identifiers have a related MSU value (millions of service units) that is used to determine the software license charge for MLC software. Table 2-11 and Table 2-12 on page 51 show MSU values for each model capacity identifier.

Table 2-11 Model capacity identifier and MSU values

Model capacity identifier	MSU	Model capacity identifier	MSU	Model capacity identifier	MSU
701	115	723	1690	745	2886
702	215	724	1748	746	2934
703	312	725	1805	747	2981
704	401	726	1865	748	3028
705	488	727	1922	749	3075
706	571	728	1979	750	3120
707	651	729	2037	751	3166
708	729	730	2092	752	3214
709	804	731	2146	753	3262
710	875	732	2200	754	3305
711	944	733	2257	755	3352
712	1011	734	2309	756	3395
713	1076	735	2366	757	3438
714	1139	736	2422	758	3480
715	1202	737	2478	759	3525
716	1264	738	2530	760	3570
717	1329	739	2585	761	3611
718	1390	740	2636	762	3652
719	1451	741	2687	763	3695
720	1512	742	2740	764	3739
721	1571	743	2789	-	-
722	1631	744	2838	-	-

Table 2-12 Model capacity identifier and MSU values for subcapacity models

Model capacity identifier	MSU	Model capacity identifier	MSU	Model capacity identifier	MSU
401	27	501	58	601	79
402	51	502	110	602	149
403	75	503	160	603	215
404	97	504	207	604	277
405	118	505	252	605	339
406	139	506	296	606	398
407	160	507	340	607	455
408	180	508	382	608	511
409	199	509	422	609	565
410	218	510	462	610	617
411	237	511	500	611	668
412	255	512	537	612	717

2.7.6 Capacity Backup

Capacity Backup (CBU) delivers temporary backup capacity in addition to what an installation might have already installed in numbers of assigned CPs, IFLs, ICFs, zAAPs, zIIPs, and optional SAPs. The six CBU types are:

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

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

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

CBU and granular capacity

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

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

You may test the CBU. Each CBU record has an allowance of five tests of 10 days each, for the contract duration. You may increase the number of tests up to a maximum of 15 for each CBU record. The real activation of CBU lasts up to 90 days with a grace period of two days to prevent sudden deactivation when the 90-day period expires. The contract duration can be set from one to five years.

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

- ▶ Number of CP CBUs allowed to be activated
- ▶ Number of IFL CBUs allowed to be activated
- ▶ Number of ICF CBUs allowed to be activated
- ▶ Number of zAAP CBUs allowed to be activated
- ▶ Number of zIIP CBUs allowed to be activated
- ▶ Number of SAP CBUs allowed to be activated
- ▶ Number of additional CBU tests 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, 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 Selecons:
 NEW00001 - CBU - CP(1) - Years(5) - Tests(0)
 Expiration(09/10/2012)

Resulting feature numbers in configuration:

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

In Example 2-2, a second CBU record is added to the same configuration for two CP CBUs, two IFL CBUs, two zAAP CBUs, and two zIIP CBUs, with five additional tests and a 5-year CBU contract. The result is now a total number of 10 years of CBU ordered, which is the standard five years in the first record and an additional five years in the second record. Two CBU records from which to choose are in the system. Five additional CBU tests have been requested, and because there is a total of five years contracted for a total of 3 CP CBUs, two IFL CBUs, two zAAPs, and two zIIP CBUs, they are shown as 15, 10, 10, and 10 CBU years for their respective types.

Example 2-2 Second CBU record and resulting configuration features

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

Resulting cumulative feature numbers in configuration:

6817	Total CBU Years Ordered	10
6818	CBU Records Ordered	2
6819	5 Additional CBU Tests	1
6820	Single CBU CP-Year	15
6822	Single CBU IFL-Year	10
6826	Single CBU zAAP-Year	10
6828	Single CBU zIIP-Year	10

CBU for CP rules

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

- ▶ The total CBU CP capacity features are equal to the number of added CPs plus the number of permanent CPs changing capacity level. For example, if 2 CBU CPs are added to the current model 503, and the capacity level does not change, the 503 becomes 505:

$$(503 + 2 = 505)$$

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

$$(3 + 3 = 6)$$

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

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

CBU for specialty engines

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

Table 2-13 Capacity BackUp matrix

Model	Total PUs available	CBU CPs min-max	CBU IFLs min-max	CBU ICFs min-max	CBU zAAPs min-max	CBU zIIPs min-max	CBU SAPs min-max
Model E12	12	0-12	0-12	0-12	0-6	0-6	0-3
Model E26	26	0-26	0-26	0-16	0-13	0-13	0-7
Model E40	40	0-40	0-40	0-16	0-20	0-20	0-11
Model E56	56	0-56	0-56	0-16	0-28	0-28	0-18
Model E64	64	0-64	0-64	0-16	0-32	0-32	0-21

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

2.7.7 On/Off Capacity on Demand and CPs

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

On/Off CoD and granular capacity

When temporary capacity requested by On/Off CoD for CPs matches the model capacity identifier range of the permanent CP feature, the total number of active CP equals the sum of the number of permanent CPs plus the number of temporary CPs ordered. For example,

when a model capacity identifier 504 has two CP5s added temporarily, it becomes a model capacity identifier 506.

When the addition of temporary capacity requested by On/Off CoD for CPs results in a cross-over from one capacity identifier range to another, the total number of CPs active when the temporary CPs are activated is equal to the number of temporary CPs ordered. For example, when a server with model capacity identifier 504 specifies six CP6 temporary CPs through On/Off CoD, the result is a server with model capacity identifier 606. A cross-over does not necessarily mean that the CP count for the additional temporary capacity will increase. The same 504 could temporarily be upgraded to a server with model capacity identifier 704. In this case, the number of CPs does not increase, but additional temporary capacity is achieved.

On/Off CoD guidelines

When you request temporary capacity, consider the following guidelines

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

Table 8-3 on page 258 shows possible On/Off CoD CP upgrades for granular capacity models. For more information about temporary capacity increases, see Chapter 8, “System upgrades” on page 233.

2.8 Summary of z10 EC structure

Table 2-14 summarizes all aspects of the System z10 EC structure.

Table 2-14 System structure summary

Description	Model E12	Model E26	Model E40	Model E56	Model E64
Number of MCMs	1	2	3	4	4
Total number of PUs	17	34	51	68	77
Maximum number of characterized PUs	12	26	40	56	64
Number of CPs	0–12	0–26	0–40	0–56	0–64
Number of IFLs	0–12	0–26	0–40	0–56	0–64
Number of ICFs	0–12	0–16	0–16	0–16	0–16
Number of zAAPs	0–6	0–13	0–20	0–28	0–32
Number of zIIPs	0–6	0–13	0–20	0–28	0–32
Standard SAPs	3	6	9	10	11
Additional SAPs	0–3	0–7	0–11	0–18	0–21
Standard spare PUs	2	2	2	2	2
Enabled memory sizes	16–352 GB	16–752 GB	16–1136 GB	16–1520 GB	16–1520 GB

Description	Model E12	Model E26	Model E40	Model E56	Model E64
L1 cache per PU	64-l/128-D KB	64-l/128-D KB	64-l/128-D KB	64-l/128-D KB	64-l/128-D KB
L1.5 cache per PU	3 MB	3 MB	3 MB	3 MB	3 MB
L2 cache	48 MB	96 MB	144 MB	192 MB	192 MB
Cycle time (ns)	0.227	0.227	0.227	0.227	0.2273
Clock frequency	4.4 GHz	4.4 GHz	4.4 GHz	4.4 GHz	4.4 GHz
Maximum number of fanout ports	16	32	40	48	48
I/O interface per IB cable	6 GBps	6 GBps	6 GBps	6 GBps	6 GBps
Maximum I/O cages	3	3	3	3	3
Number of support elements	2	2	2	2	2
External power	3 phase	3 phase	3 phase	3 phase	3 phase
Internal Battery Feature	Optional	Optional	Optional	Optional	Optional



System design

The objective of this chapter is to explain how the IBM System z10 Enterprise Class (z10 EC) is designed. This information can be used to understand the functions that make the z10 EC a server that suits a broad mix loads for large enterprises.

This chapter discusses the following topics:

- ▶ 3.1, “Design highlights” on page 58
- ▶ 3.2, “Book design” on page 59
- ▶ 3.3, “Processing unit” on page 66
- ▶ 3.4, “Processing unit functions” on page 72
- ▶ 3.5, “Memory design” on page 87
- ▶ 3.6, “Logical partitioning” on page 90
- ▶ 3.7, “Intelligent Resource Director” on page 100
- ▶ 3.8, “Clustering technology” on page 102

The design of the z10 EC symmetric multiprocessor (SMP) is the next step in an evolutionary trajectory stemming from the introduction of CMOS technology back in 1994. Over time, and for the z10 EC once again, the design has been adapted to the changing requirements dictated by the shift towards e-business applications that customers are becoming more and more dependent on.

The z10 EC offers very high levels of serviceability, availability, reliability, resilience, and security, and fits in the IBM strategy in which mainframes play a central role in realizing an intelligent, energy efficient, integrated infrastructure. The z10 EC is designed in such a way that not only the server is considered important for the infrastructure, but also everything around it in terms of operating systems, middleware, storage, security, and network technologies supporting open standards, all to help customers achieve their business goals.

The modular book design aims to reduce planned and unplanned outages by offering concurrent repair, replace, and upgrade functions for processors, memory, and I/O. The z10 EC with its ultra-high frequency, superscalar processor design, and flexible configuration options is the next implementation to address the ever-changing IT environment.

3.1 Design highlights

The physical packaging of the z10 EC is comparable to the packaging used for z990 and z9 EC systems. Its modular book design creates the opportunity to address the ever-increasing costs related to building systems with ever-increasing capacities. The modular book design is flexible and expandable and might contain even larger capacities in the future.

The main objectives of the z10 EC system design, which are discussed in this and subsequent chapters, are as follows:

- ▶ Offer a *flexible infrastructure* to concurrently accommodate a wide range of operating systems and applications, from the traditional systems (for example z/OS and z/VM) to the world of Linux and e-business.
- ▶ Offer state-of-the-art *integration* capability for server consolidation, offering virtualization techniques, such as:
 - Logical partitioning, which allows 60 independent logical servers
 - z/VM, which can virtualize hundreds to thousands of servers as independently running virtual machines
 - HiperSockets, which implement virtual LANs between logical partitions within a serverThis allows for a logical and virtual server coexistence and maximizes system utilization and efficiency, by sharing hardware resources.
- ▶ Offer *high performance* to achieve the outstanding response times required by new workload-type applications, based on high frequency, superscalar processor technology, architecture, and high bandwidth channels, which offer second-to-none data rate connectivity.
- ▶ Offer the *high capacity* and *scalability* required by the most demanding applications, both from single-system and clustered-systems points of view.
- ▶ Offer the capability of *concurrent upgrades* for processors, memory, and I/O connectivity, avoiding server outages in planned situations.
- ▶ Implement a system with *high availability* and *reliability*, from the redundancy of critical elements and sparing components of a single system, to the clustering technology of the Parallel Sysplex environment.
- ▶ Have broad internal and external *connectivity* offerings, supporting open standards such as Gigabit Ethernet (GbE), and Fibre Channel Protocol (FCP) for Small Computer System Interface (SCSI).
- ▶ Provide the highest level of *security* in which every two PUs share a CP Assist for Cryptographic Function (CPACF). Optional Crypto Express features with Cryptographic Coprocessors and Cryptographic Accelerators for Secure Sockets Layer (SSL) transactions of e-business applications can be added.
- ▶ Be *self-managing* and *self-optimizing*, adjusting itself on workload changes to achieve the best system throughput, through the Intelligent Resource Director or the Workload Manager functions, assisted by HiperDispatch.
- ▶ Have a *balanced system* design, providing large data rate bandwidths for high performance connectivity along with processor and system capacity.

The following sections describe the z10 EC system structure, showing a logical representation of the data flow from PUs, L2 cache, memory cards, and a variety of I/O interconnect capabilities, which connect to the I/O cages and provide direct coupling links.

3.2 Book design

A book contains a Multi-Chip Module (MCM) with five quad-core microprocessor chips. Of these chips, 12, 13, 14, or 16 are available as processor units for customer use depending on the model and the book that the MCM is in; 48 DIMM slots, and up to 16 I/O ports are organized in up to eight fanouts. Additionally, each book has its own power supplies (DCAs, known as distributed converter assemblies). MCM components are shown in Figure 3-1.

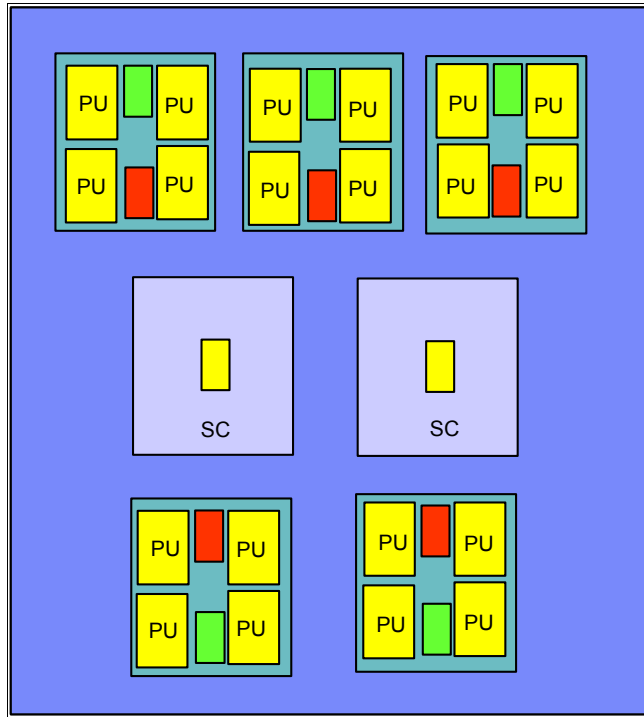


Figure 3-1 z10 EC MCM

Each microprocessor (PU) has its own 192 KB cache Level 1 (L1), split into 128 KB for data (D-cache) and 64 KB for instructions (I-cache). The L1 cache is designed as a store-through cache, meaning that altered data is also stored to the next level of memory. See Figure 3-2.

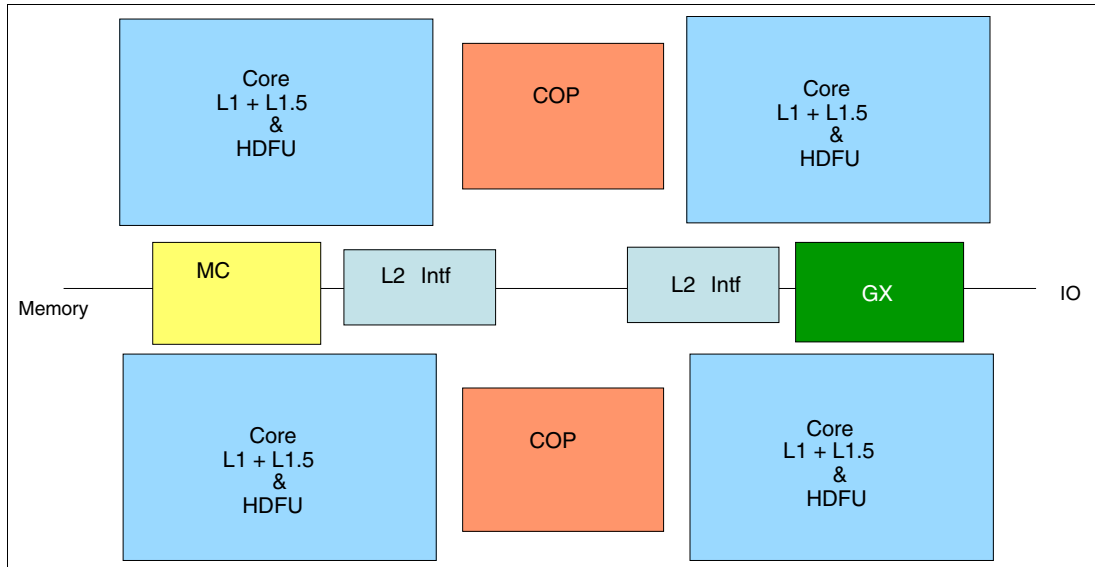


Figure 3-2 PU chip

The next level of memory is the L1.5 cache that is also on each PU and is 3 MB in size. It is a store-through cache. The Level 1.5 cache is needed because in servers with reduced cycle times such as the z10 EC, the distance or latency between the processor and the shared cache (L2) is getting bigger measured in number of cycles needed go to the cache and get the data. The increase in latency is compensated by the insertion of an intermediate level cache reducing the traffic to and from the L2 cache. Memory levels are presented in Figure 3-3.

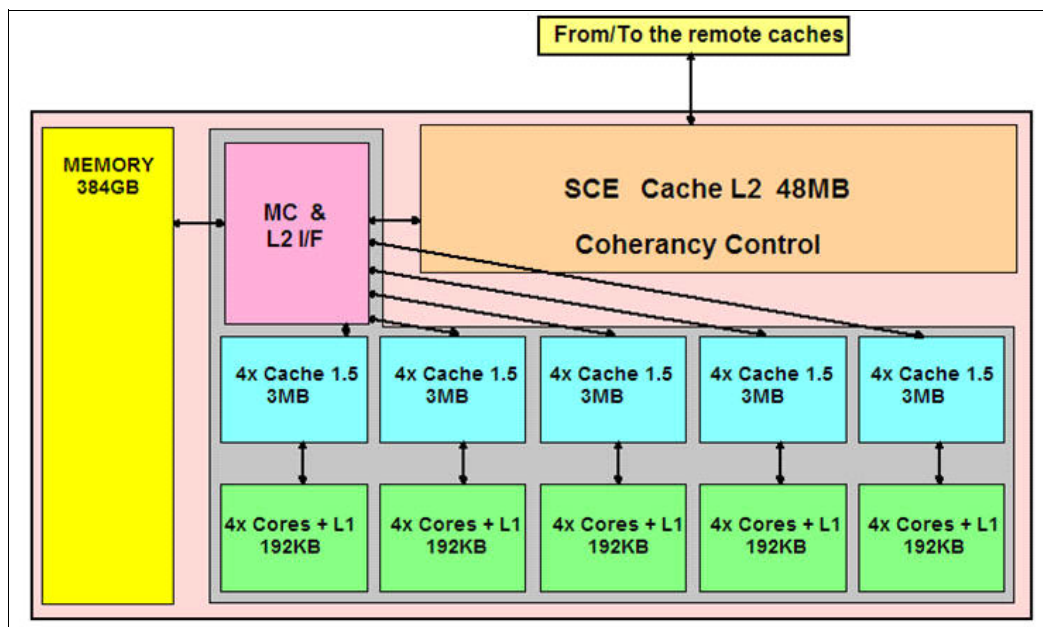


Figure 3-3 L1, L1.5, L2 and L3 cache levels

All models use CMOS 11S technology. The microprocessors are running at 4.4 GHz (0.227 ns cycle time).

The MCM also contains two storage control (SC) chips, each with a Level 2 cache of 24 MB for a total of 48 MB. The SC acts as a coherency manager and is responsible for coherent traffic between the L2 caches in a multiple book system and between the L2 cache and the local microprocessor caches. It optimizes cache traffic and does not look for cache hits in other books when it knows that all resources of a given logical partition are available in the same book. Each L2 cache has a direct path to each of the other L2 caches in remote books on one side and each microprocessor in the MCM on the other side through point-to-point (any-to-any) connections.

Each memory DIMM has a capacity of 4 GB or 8 GB, easily allowing you to install up to 384 GB of physical memory (L3) per book. A four-book z10 EC can have up to 1.5 TB of physical memory. Of the installed amount of physical memory, 16 GB is set aside for the hardware system area (HSA). The 16 GB HSA does not belong to the memory you purchased, meaning that you may purchase only up to 352 GB in a one-book system (352 GB because the upgrade granularity beginning at 256 GB equals 32 GB, and 16 GB is set aside for the HSA).

The L2 cache is the aggregate of all cache space on the SC chips, resulting in a 48 MB L2 cache per book. The SC chip controls the access and storing of data in between the system memory (L3) and the on-chip caches. The L2 cache is shared by all PUs within a book and shared across books, providing the communication between L2 caches across books in systems with more than one book installed. The L2 has a store-in buffer design.

Access to main memory (L3) is controlled from four of the five microprocessor chips by their memory control (MC) function, shown in Figure 3-3 on page 60. Storage access is interleaved between the DIMMs, which tends to equalize storage activity across them.

I/O traffic is handled by up to 16 ports in up to eight fanouts of three different types (numbered D1, D2, and D5 through DA) in front of the book.

The logical book structure is shown in Figure 3-4.

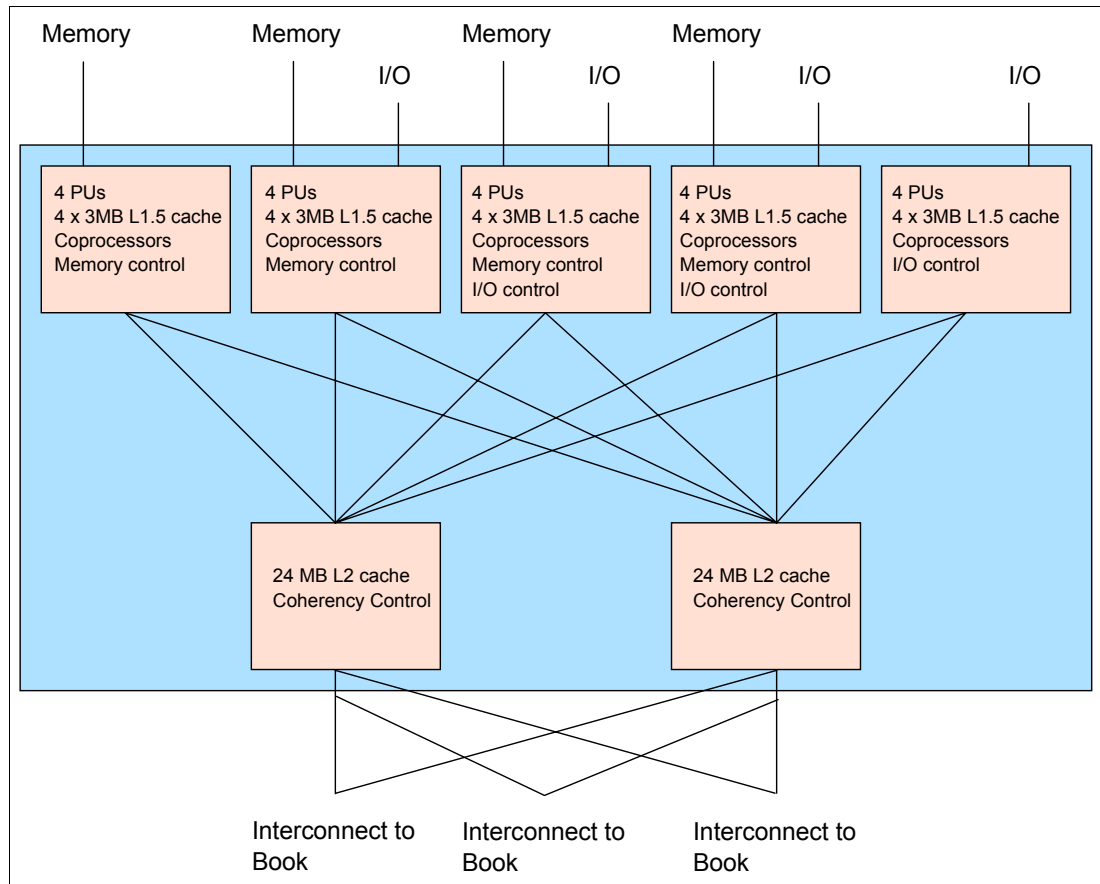


Figure 3-4 20-PU logical book structure

Up to 16 connections per book are for transferring data. Each of these connections have a bidirectional bandwidth of up to 6 GBps. A one-book system can have up to 16 physical connections, a two-book system 32, a three-book system 40, and a four-book system 48. This leads to the support of a maximum aggregated data rate of 288 GBps per system.

The four I/O interconnect types used for all I/O types are:

- ▶ I/O interconnect based on a two-port (6 GBps each) Host Channel Adapter 2 - Copper (HCA2-C) fanout supports an IFB-MP card in an I/O cage to connect to:
 - ESCON
 - ESCON channels (16 port cards)
 - FICON
 - FICON-Express (two port cards): carried forward on upgrade only for FCV
 - FICON Express2 (four port cards): carried forward on upgrade only
 - FICON Express4 (four port cards): carried forward on upgrade only
 - FICON Express4 (four port cards)
 - OSA
 - OSA-Express3 10 Gb Ethernet (LR and SR)
 - OSA-Express3 Gb Ethernet (LX and SX offer four port cards)

- OSA-Express3 1000BASE-T Ethernet (four port cards)
 - OSA-Express2 10 Gb Ethernet LR: carried forward during an upgrade only
 - OSA-Express2 Gb Ethernet (LX and SX): until no longer available and carried forward during an upgrade
 - OSA-Express2 1000BASE-T Ethernet: until no longer available and carried forward during an upgrade
- Crypto
- Crypto Express2 (carried forward on upgrade only) with one (FC 0870) or two (FC 0863) PCI-X adapters per feature
A PCI-X adapter can be configured as a cryptographic coprocessor for secure key operations or as an accelerator for clear key operations.
 - Crypto Express3 with one (FC0871) or two (FC0864) PCI Express adapters per feature
A PCI Express adapter can be configured as a cryptographic coprocessor for secure key operations or as an accelerator for clear key operations.
- ISC
- ISC-3 links, up to four Coupling Links with two links per daughter card (ISC-D)
Two daughter cards plug into one mother card (ISC-M).
- ▶ Coupling for up to 150 meters is based on a two-port (6.0 GBps each) Host Channel Adapter 2 - Optical (HCA2-O) fanout
HCA2-O fanouts support PSIFB coupling links for up to 16 CHPIDs, from System z10 to System z10 or from System z10 to System z9.
 - ▶ Coupling for up to 10 km (or up to 100 km with extenders) is based on a two-port Host Channel Adapter 2 - Optical LR (HCA2-O LR) fanout
HCA2-O LR fanouts support PSIFB coupling links for up to 16 CHPIDs, from System z10 to System z10.

Note: Only four CHPIDs per port are recommended for both HCA2-O and HCA2-O LR.

- ▶ I/O interconnect based on a two-port memory bus adapter fanout is used for ICB-4, directly attaching to a System z10, System z9, z990, or z890. The ICB-4 runs at 2.0 GBps for up to 7 meters.

Note: The ICB-4 feature cannot be ordered on the z10 E64 model.

System z10 servers will be the last server family to support the ICB-4 feature.

For details about I/O connectivity and each channel type see Chapter 4, “I/O system structure” on page 105.

Dual external time reference

Two external time reference (ETR) cards are already installed and shipped with the server and provide a dual-path interface to IBM Sysplex Timers, which may be used for timing synchronization between systems in a sysplex environment. This redundancy allows continued operation even if a single ETR card fails. This redundant design also allows concurrent maintenance. The two connectors to external timers are located above the books and are on the mid-plane to which the books are connected.

Support exists for a Simple Network Time Protocol (SNTP) client on the Support Element. When STP is used, the time of an STP-only Coordinated Timing Network (CTN) can be synchronized with the time provided by a Network Time Protocol (NTP) server, allowing a heterogeneous platform environment to have the same time source.

The time accuracy of an STP-only CTN is improved by adding an NTP server with the pulse per second output signal (PPS) as the External Time Signal (ETS) device. ETS is available from several vendors that offer network timing solutions.

STP tracks the highly stable accurate PPS signal from the NTP server and maintains an accuracy of 10 μ s as measured at the PPS input of the System z server. In comparison, the IBM Sysplex Timer could maintain an accuracy of 100 μ s when attached to an external time source.

If STP uses a dial-out time service or an NTP server without PPS a time accuracy of 100 ms to the ETS is maintained.

Note: Server Time Protocol is available as FC 1021. STP is implemented in the Licensed Internal Code (LIC) of the z10 BC and is designed for multiple servers to maintain time synchronization with each other. See the following publications for more information:

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

Oscillator

The z10 EC has two oscillator cards (a primary and a backup). Although not part of the book design, they are found above the books, connected to the same mid-plane to which the books are connected. If the primary fails, the secondary detects the failure, takes over transparently, and continues to provide the clock signal to the server.

3.2.1 Book interconnect topology

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

Inter-book communication takes place at the L2 cache level. The L2 cache is implemented on two storage control (SC) cache chips in each MCM. Each SC chip holds 24 MB of SRAM cache, resulting in a 48 MB L2 cache per book. The L2 cache is shared by all PUs in the book and has a store-in buffer design. The SC function regulates coherent book-to-book traffic.

The point-to-point topology between the books maintains interbook communication at the L2 cache level. Each book is able to communicate with every other book in the configuration.

Figure 3-5 shows a simplified topology for a four-book system.

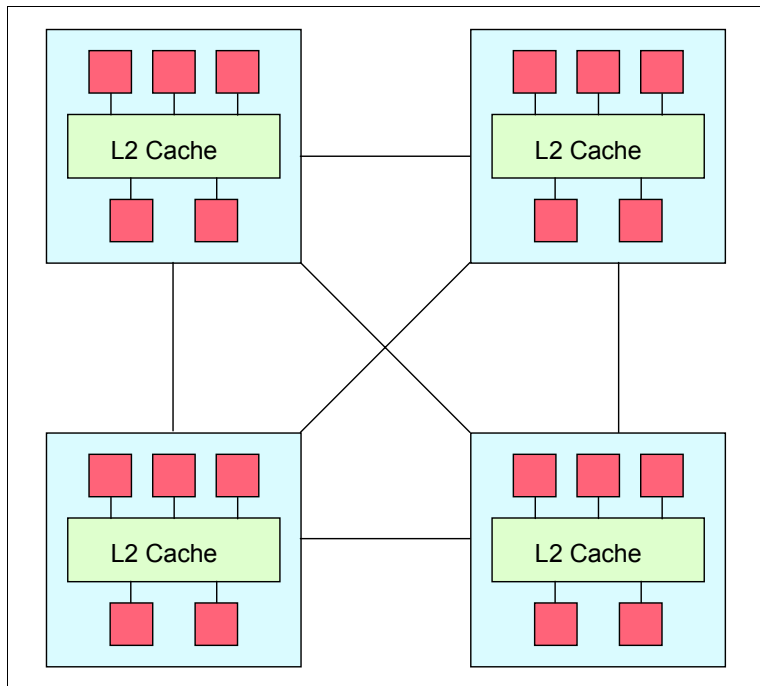


Figure 3-5 Point-to-point topology for book-to-book communication

A memory-coherent controller (L2 Cache Controller -L2C- on the SC chip) optimizes the traffic between the L2 caches.

3.2.2 System control

Various system elements use *flexible service processors* (FSPs). An FSP is based on the IBM Power PC microprocessor. It connects to an internal Ethernet LAN to communicate with the Support Elements (SEs) and provides a subsystem interface (SSI) for controlling components. Figure 3-6 is a conceptual overview of the system control design.

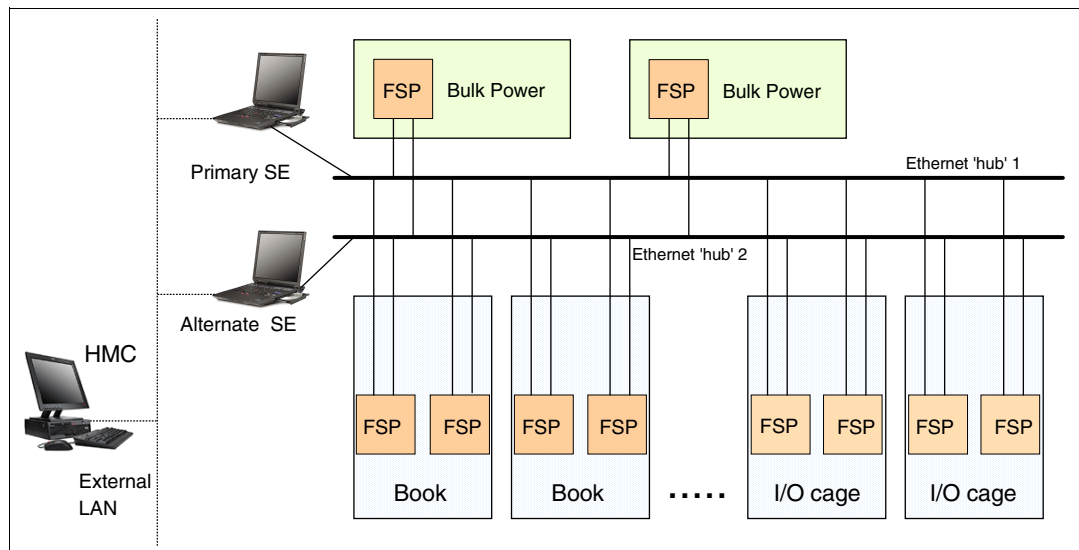


Figure 3-6 Conceptual overview of system control elements

A typical FSP operation is to control a power supply. An SE sends a command to the FSP to bring up 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 report this status to the SE.

Most system elements are duplexed (for redundancy), and each element has an FSP. Two internal Ethernet LANs and two SEs for redundancy, and crossover capability between the LANs are available so that both SEs can operate on both LANs.

The SEs, in turn, are connected to one or two (external) LANs (Ethernet only), and the Hardware Management Consoles (HMCs) are connected to these external LANs. One or more HMCs can exist. In a production environment, the server is normally managed from the HMCs. If necessary, the system can be managed from either SE. Several or all HMCs can be disconnected without affecting system operation.

3.3 Processing unit

Today's systems design is driven by processor cycle time, though this does not automatically mean that the performance characteristics of the system improve. One of the first things to realize is that 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. Cache sizes must be limited because larger distances must be traveled to reach long cache lines.

This phenomenon of shrinking cache sizes can be seen in the design of the z10 EC, where the instruction and data caches (L1) have been managed back in size to accommodate the reduced cycle times that limit the distance that can be traveled in one cycle, potentially causing increased latency. Also, the distance to remote caches as seen from the microprocessor becomes a significant factor. An example of this is the L2 cache that is not on the microprocessor (and might not even be in the book). One way to solve the problem is by the introduction of additional cache levels in combination with denser packaging.

Although the L2 cache is rather large, the reduced cycle time has the effect that more cycles are needed to travel the same distance. In order to overcome this and avoid potential latency, the z10 EC introduces an intermediate local non-shared cache level on each microprocessor (the L1.5 cache) to reduce traffic to and from the shared L2 cache. Only when there is a cache miss in both L1 and L1.5 is a request sent to L2. L2 is the coherence manager, meaning that all memory fetches must be in the L2 cache before that data can be used by the processor.

As seen in Figure 3-4 on page 62, memory fetches go through the processor when transferred to L2 but bypass any processor function. Instruction fetches are fetched into the I-cache. If the instruction is not in L2, it is fetched from memory and installed in the I-cache, L1.5, and in L2 caches.

Another approach is available for avoiding L2 cache access delays (latency) as much as possible. The L2 cache straddles up to four books. This means relatively large distances exist between the higher-level caches in the processors and the L2 cache content. To overcome the delays that are inherent to the book design and to save cycles to access the *remote* L2 content, it is beneficial to keep instructions and data as close to the processors as possible by directing as much work of a given workload (that is, a logical partition) on the processors located in the same book as the L2 cache. This is achieved by having the PR/SM scheduler and the z/OS dispatcher work together to keep as much work as possible within the boundaries of as few processors and L2 cache space (which is best within a book boundary)

as can be achieved without affecting throughput and response times. Preventing PR/SM and the dispatcher from scheduling and dispatching a workload on any processor available, and keeping the workload in as small a portion of the server as possible, contributes to overcoming latency in a high-frequency processor design such as the z10 EC. The cooperation between z/OS and PR/SM has been bundled in a function called HiperDispatch. More information about HiperDispatch is in 3.6, “Logical partitioning” on page 90.

Each processing unit is optimized to meet the demands of a wide variety of business workload types without compromising the performance characteristics of traditional workloads. The PUs in the z10 EC have a superscalar design.

3.3.1 Superscalar processor

A scalar processor is a processor that is based on a single-issue architecture, which means that only a single instruction is executed at a time. A superscalar processor allows concurrent execution of instructions by adding additional resources onto the microprocessor to achieve more parallelism by creating multiple pipelines, each working on its own set of instructions.

A superscalar processor is based on a multi-issue architecture. In such a processor, where multiple instructions can be executed at each cycle, a higher level of complexity is reached, 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.

For more information about pipeline environment, see the IEEE Computer Society article *IBM z10: The Next-Generation Mainframe Microprocessor* by Charles F. Webb, March 2008, available for members at IEEE Micro:

<http://www.computer.org/micro/>

Example of branch prediction

If the branch prediction logic of the microprocessor makes the wrong prediction, removing all instructions in the parallel pipelines also might be necessary. Obviously, the cost of the wrong branch prediction is more costly in a high-frequency processor design, as we discussed previously. Therefore, the branch prediction techniques used are very important to prevent as many wrong branches as possible. For this reason, a variety of history-based branch prediction mechanisms are used. In addition, the z10 EC has a two-level compressed branch target buffer (BTB). The 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 an extremely high branch prediction success rate.

Challenges of creating a superscalar processor

Many challenges exist in creating an efficient superscalar processor. The superscalar design of the PU has made big strides in avoiding address generation interlock (AGI) situations. Instructions requiring information from memory locations can suffer multi-cycle delays to get the desired memory content, and because high-frequency processors wait faster, the cost of getting the information might become prohibitive.

3.3.2 Compression unit on a chip

Each two microprocessor cores on the quad-core chip share a compression unit, providing the hardware compression function. The compression unit is integrated with the CP Assist for Cryptographic Function (CPACF), benefiting from combining (or sharing) the use of buffers

and interfaces. Two sets of two microprocessors on the quad-core chip share the compression unit function.

3.3.3 CP Assist for Cryptographic Function

The CP Assist for Cryptographic Function (CPACF) accelerates the encrypting and decrypting of SSL transactions and VPN-encrypted data transfers. The assist function uses a special instruction set for symmetrical clear key cryptographic encryption and encryption operations. Six special instructions are used with the cryptographic assist function. For more information about the instructions (and micro-programming), see the IBM Resource Link Web site, which requires registration:

<http://www.ibm.com/servers/resourceLink/>

Two sets of two microprocessors on the quad-core chip share the CPACF. The CPACF is integrated with the compression unit, benefiting from combining (sharing) the use of buffers and interfaces (see Figure 3-7). The assist provides high-performance hardware encrypting and decrypting support for clear key operations.

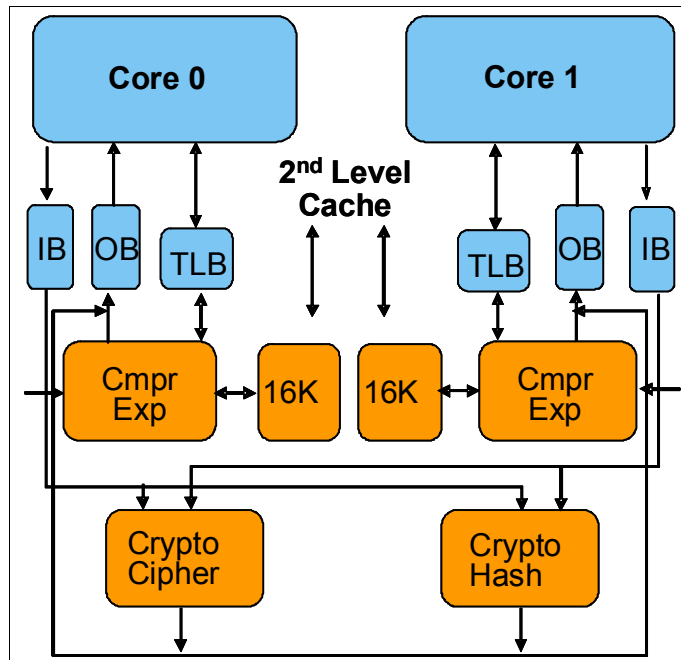


Figure 3-7 Compression and cryptographic coprocessor

CPACF offers a set of symmetric cryptographic functions for high encrypting and decrypting performance of clear key operations for SSL, VPN, and data-storing applications that do not require FIPS 140-2 level 4 security. The cryptographic architecture includes support for:

- ▶ Data Encryption Standard (DES) data encryption and decrypting
- ▶ Triple Data Encryption Standard (triple DES) data encrypting and decrypting
- ▶ Advanced Encryption Standard (AES) for 128-bit, 192-bit, and 256-bit keys
- ▶ Pseudorandom number generator (PRNG)
- ▶ MAC message authorization
- ▶ Secure Hash Algorithm (SHA-1) hashing
- ▶ Secure Hash Algorithm (SHA-2) hashing (SHA-256, SHA-384, and SHA-512)

3.3.4 Decimal floating point accelerator

The decimal floating point (DFP) accelerator function is present on each of the microprocessors (cores) on the quad-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 typically done in decimal arithmetic has involved frequent necessary data conversions and approximation to represent decimal numbers. This has made floating point arithmetic complex and error-prone for programmers using it for applications in which the data is typically decimal data.

Hardware decimal-floating-point computational instructions provide data formats of 4, 8, and 16 bytes, an encoded decimal (base 10) representation for data, instructions for performing decimal floating point computations, and an instruction that performs data conversions to and from the decimal floating point representation.

Benefits of DFP accelerator

The DFP accelerator offers the following benefits:

- ▶ Avoids rounding issues such as those happening with binary-to-decimal conversions.
- ▶ Has better functionality over existing binary coded decimal (BCD) operations.
- ▶ Follows the standardization of the dominant decimal data and decimal operations in commercial computing supporting industry standardization (IEEE 745R) of decimal floating point operations. Instructions are added in support of the Draft Standard for Floating-Point Arithmetic, which is intended to supersede the ANSI/IEEE Std 754-1985.

Software support

Decimal floating point is supported in several programming languages, including:

- ▶ Release 4 and 5 of High Level Assembler
- ▶ C/C++ (requires z/OS 1.9 with program temporary fixes, PTFs, for full support)
- ▶ Enterprise PL/I Release 3.7 and Debug Tool Release 8.1
- ▶ Java Applications using the BigDecimal Class Library
- ▶ SQL support as in DB2 Version 9

Support for decimal floating point data types is provided in SQL as of DB2 Version 9.

Tip: For details, check the 2097DEVICE Preventive Service Planning (PSP) bucket, available through your IBM Software Support representative.

3.3.5 Processor error detection and recovery

The PU uses something called transient recovery as an error recovery mechanism. When an error is detected, the instruction unit retries the instruction and attempts to recover the error. If the retry is not successful (that is, a permanent fault exists), a relocation process is started

that restores the full capacity by moving work to another PU. Relocation under hardware control is possible because the R-unit has the full architected state in its buffer. The principle is shown in Figure 3-8.

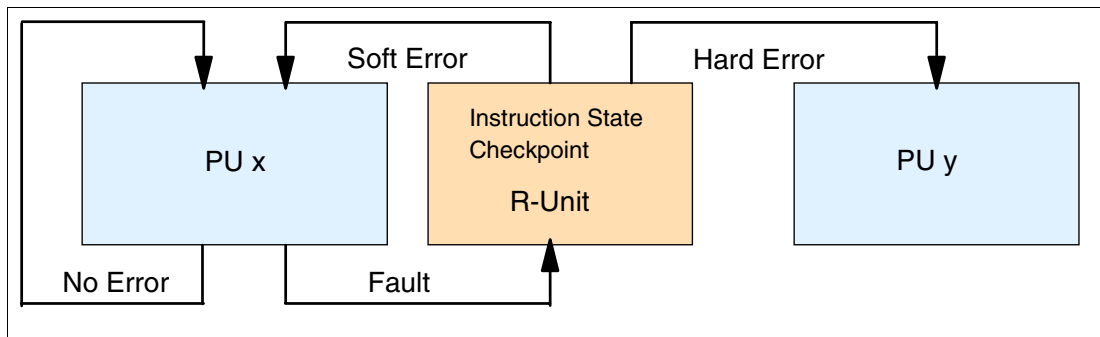


Figure 3-8 PU error detection and recovery

3.3.6 Branch prediction

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

The branch history table (BHT) implementation on processors has a large performance improvement effect, but is not sufficient for the z10 EC. 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. Besides the BHT, the z10 EC uses a variety of techniques to improve the prediction of the correct branch to be executed. The techniques include:

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

The success rate of branch prediction contributes significantly to the superscalar aspects of the z10 EC. 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.

3.3.7 Wild branch

When a bad pointer is used or when code overlays a data area containing a pointer to code, a random branch is the result, causing a 0C1 or 0C4 abend. Random branches are very 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 executed is kept. z/OS uses this information in conjunction with debugging aids, such as the SLIP command, to determine where a wild branch came from and might collect data from that storage location. This approach decreases the many debugging steps necessary when looking for where the branch came from.

3.3.8 IEEE floating point

Over 130 binary and hexadecimal floating-point instructions are present in z10 EC. They incorporate IEEE Standards into the platform.

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

3.3.9 Translation look-aside buffer

The translation look-aside buffer (TLB) in the instruction and data L1 caches use a secondary TLB to enhance performance. In addition, a translator unit is added to translate misses in the secondary TLB.

The size of the TLB is kept as small as possible because of its low access time requirements and hardware space limitations. Because memory sizes have recently increased significantly, as a result of the introduction of 64-bit addressing, a smaller working set is represented by the TLB. To increase the working set representation in the TLB without enlarging the TLB, large page support is introduced and can be used when appropriate. See “Large page support” on page 88.

3.3.10 Instruction fetching, decode, and grouping

The superscalar design of the microprocessor allows for the decoding of up to two instructions per cycle and the execution of three instructions per cycle. Execution takes place in sequence, but 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 awaiting decoding.

Instruction decoding

The processor can decode one or two instructions per cycle. The result of the decoding process is queued and subsequently used to form a group.

Instruction grouping

From the instruction queue, one simple branch instruction and up to two general instructions can be issued every cycle. The instructions are taken from the instruction queue and grouped together. The instructions are assembled according to instruction grouping rules. A complete description of the rules is beyond the scope of this book.

The compiler and JVM are responsible for selecting instructions that best fit with the superscalar microprocessor and abide by the grouping rules to create code that best exploits the superscalar implementation.

3.3.11 Extended translation facility

Instructions have been added to the z/Architecture instruction set in support of the extended translation facility. They are used in data conversion operations for data encoded in Unicode, 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.

3.3.12 Instruction set extensions

The processor supports a large number of instructions to support functions, including:

- ▶ 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 half words (such as those used for Unicode).
- ▶ Cryptographic instructions, extended with AES, SHA-256, and functions for random number generation
- ▶ Extended Translate Facility-3 instructions, enhanced to conform with the current Unicode 4.0 standard
- ▶ Assist instructions, help eliminate hypervisor overhead

3.4 Processing unit functions

A key component of the z10 EC is the processing unit (PU), discussed in 3.3, “Processing unit” on page 66. The PU is the microprocessor where instructions are executed and the related data resides. The instructions and the data are stored in the PU’s high-speed buffer, called the Level 1 cache. Each PU has its own Level 1 cache, split into 128 KB for data and 64 KB for instructions.

The L1 cache is designed as a store-through cache, which means that altered data is synchronously stored into the next level, the L1.5 cache, that holds 3 MB on each PU, where altered data is synchronously passed through to the next level of cache, the L2 cache.

All PUs are physically identical. When the system is initialized, PUs can be characterized to specific functions: CP, IFL, ICF, zAAP, zIIP, or SAP. The function assigned to a PU is set by the Licensed Internal Code, which is loaded when the system is initialized (at power-on reset) and the PU is *characterized*. Only characterized PUs have a designated function. Non-characterized PUs are considered spares.

Note: All PUs assigned to a logical partition are either shared or dedicated.

This design brings outstanding flexibility to the z10 EC server, because any PU can assume any available characterization. This also plays an essential role in system availability, because PU characterization can be done dynamically, with no server outage, allowing the actions discussed in the following sections.

Concurrent upgrades

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

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

For more information about Capacity on Demand see Chapter 8, “System upgrades” on page 233.

PU sparing

In the rare event of a PU failure, the failed PU’s characterization is dynamically and transparently reassigned to a spare PU. More information about PU sparing is provided in “Sparing rules” on page 86.

A minimum of one PU per z10 EC must be ordered as one of the following items:

- ▶ Central processor (CP)
- ▶ Integrated Facility for Linux (IFL)
- ▶ Internal coupling facility (ICF)

The number of CPs, IFLs, ICFs, zAAPs, zIIPs, or SAPs assigned to particular models depends on the configuration. Two spare PUs can reside in any of the MCMs. For example, a model E12 has two spare PUs in its MCM, and a model E26 has one spare PU in the MCM of its first book, and one in the MCM of the second book. For details about spare PU location, see Table 2-3 on page 34. Non-characterized PUs act as spares. The number of these additional spare PUs depends on the number of books in the configuration and how many PUs are non-characterized.

PU pools

PUs defined as CPs, IFLs, ICFs, zIIPs, and zAAPs are grouped together in their own pools, from where they can be managed separately. This significantly simplifies capacity planning and management for logical partitions. The separation also has an effect on weight management because CP, zAAP, and zIIP weights can be managed separately. For more information, see “PU weighting” on page 74.

All assigned PUs are grouped together in the PU pool. These PUs are dispatched to online logical PUs. As an example, consider a z10 EC with 10 CPs, three zAAPs, two IFLs, two zIIPs, and one ICF. The system has a PU pool of 18 PUs, called the *pool width*. Subdivision of the PU pool defines:

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

PUs are placed in the pools according to the following occurrences:

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

Also, when a dedicated logical partition is deactivated or logically unconfigures a logical PU, its PUs are returned to the proper pool.

PUs are removed from their pools when a concurrent downgrade takes place as the result of removal of a CBU, and through On/Off CoD and conversion of a PU. Also, when a dedicated logical partition is activated, its PUs are taken from the proper pools, as is the case when a logical partition logically configures a PU on, if the width of the pool allows.

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

For a logical partition, logical PUs are dispatched from the supporting pool only. This means that logical CPs are dispatched from the CP pool, logical zAAPs are dispatched from the zAAP pool, logical zIIPs from the zIIP pool, logical IFLs from the IFL pool, and the logical ICFs from the ICF pool.

PU weighting

Because zAAPs, zIIPs, IFLs, and ICFs have their own pools from where they are dispatched, they can be given their own weights. zAAPs and zIIPs are not assigned a weight based on their own weight specifications rather than a weight that is based on the logical partition CP weight.

For more information about PU pools and processing weights, see *IBM System z10 Enterprise Class Processor Resource/Systems Manager Planning Guide*, SB10-7153.

3.4.1 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, Linux) and the Coupling Facility Control Code (CFCC).

The z10 EC can only be initialized in LPAR mode. CPs are defined as either dedicated or shared. Reserved CPs can be defined to a logical partition to allow for nondisruptive *image* upgrades. If the operating system in the logical partition supports the *logical processor add* function, reserved processors are no longer needed. Logical processor add is supported by z/VM 5.3 with PTFs and z/OS V1.10.

Regardless of the installed model, a logical partition can have up to 56 logical CPs defined. (This is the sum of active and reserved logical CPs.) On model E64, up to 64 initial and reserved logical CPs can be defined. We recommend defining no more CPs than the operating system supports.

All PUs characterized as CPs within a configuration are grouped into the CP pool. The CP pool can be seen on the hardware management console workplace. Any z/Architecture operating systems and CFCCs can run on CPs that are assigned from the CP pool.

Within the limit of all non-characterized PUs available in the installed configuration, CPs can be concurrently assigned to an existing configuration through On-line Permanent Upgrade, On/Off Capacity on Demand (On/Off CoD), Capacity Backup (CBU), or Capacity for Planned Event (CPE). For more information about all forms of concurrent addition of CP resources see Chapter 8, "System upgrades" on page 233.

If the MCMs in the installed books have no available remaining PUs, the assignment of the next CP results in requiring a model upgrade and the installation of an additional book. Book

installation is nondisruptive, but can take more time than a simple Licensed Internal Code upgrade.

Granular capacity

The z10 EC recognizes four distinct capacity settings for CPs. Full-capacity CPs are identified as CP7. Up to 64 CPs can be configured. In addition to full-capacity CPs, three subcapacity settings (CP6, CP5, and CP4), each for up to 12 CPs, are offered. The four capacity settings appear in hardware descriptions, as follows:

- ▶ CP7 feature code 6810
- ▶ CP6 feature code 6809
- ▶ CP5 feature code 6808
- ▶ CP4 feature code 6807

Granular capacity adds 36 subcapacity settings to the 64 capacity settings that are available with full capacity CPs (CP7). Each of the 36 subcapacity settings applies only to up to 12 CPs, independently of the model installed.

Information about CPs in the remainder of this chapter applies to all CP capacity settings, CP7, CP6, CP5, and CP4, unless indicated otherwise. See 2.7, “Model configurations” on page 45, for more details about granular capacity.

Capacity marker

A capacity marker indicates the presence of purchased capacity. For example, a model E12 can be configured with five full capacity CP7s, of which one has been purchased but is not used. The capacity level is identified by FC 6810, and model capacity marker 705 (FC 7142) indicates the purchased capacity.

The same applies to the subcapacity models. For example, when a model E12 is configured with four subcapacity CP5s, of which one CP has been purchased but is not used. The capacity level is identified with FC 6808, and capacity marker 504 (FC 7116) indicates the purchased capacity.

Note: Capacity settings smaller than the full-capacity setting (CP6, CP5, and CP4) only apply to up to 12 PUs characterized as CPs. Specialty engines such as IFLs, ICFs, zAAPs, and zIIPs always run at full capacity.

3.4.2 Integrated Facility for Linux

An Integrated Facility for Linux (IFL) is a PU that can be used to run Linux or Linux guests on z/VM operating systems. Up to 64 PUs may be characterized as IFLs, depending on the configuration. IFLs can be dedicated to a Linux or a z/VM logical partition, or can be shared by multiple Linux guests or z/VM logical partitions running on the same z10 EC. Only z/VM and Linux on System z operating systems and designated software products can run on IFLs.

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

IFLs do not change the model capacity identifier of the z10 EC. Software product license charges based on the model capacity identifier are not affected by the addition of IFLs.

Adding an IFL

Within the limit of all non-characterized PUs available in the installed configuration, IFLs can be concurrently added to an existing configuration through On-line Permanent Upgrade,

On/Off Capacity on Demand (On/Off CoD), or Capacity for Planned Event (CPE). An IFL CBU might have been purchased to provide IFL backup capacity for lost IFLs elsewhere. If the installed books have no unassigned PUs remaining, the assignment of the next IFL might require the installation of an additional book. For more information see Chapter 8, “System upgrades” on page 233.

Unassigned IFLs

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

3.4.3 Internal Coupling Facilities

An Internal Coupling Facility (ICF) is a PU used to run the Coupling Facility Control Code for Parallel Sysplex environments. Within the capacity of the sum of all unassigned PUs in up to four books, up to 16 ICFs can be characterized, depending on the model. At least a model E26 is necessary to characterize 16 ICFs (model E12 supports only up to 12 ICFs).

Only Coupling Facility Control Code can run on ICF processors. ICFs do not change the model capacity identifier of the z10 EC. Software product license charges based on the model capacity identifier are not affected by the addition of ICFs.

All ICF processors within a configuration are grouped into the ICF pool. The ICF pool can be seen on the hardware management console workplace.

The ICF processors can only be used by coupling facility logical partitions. ICFs are either dedicated or shared. ICF processors can be dedicated to a CF logical partition, or shared by multiple CF logical partitions running in the same server. However, having a logical partition with dedicated *and* shared ICFs at the same time is *not* possible. With Dynamic ICF expansion, a coupling facility image can also use dedicated ICFs and shared CPs.

Thus, a coupling facility image can have one of the following combinations defined in the image profile:

- ▶ Dedicated ICFs
- ▶ Shared ICFs
- ▶ Dedicated CPs
- ▶ Shared CPs
- ▶ Dedicated ICFs *and* shared CPs

Shared ICFs add flexibility. However, running only with shared coupling facility PUs (either ICFs or CPs) is not a recommended production configuration. We recommend that a production CF operates by using ICF dedicated processors.

In Figure 3-9, the server on the left has two environments defined (production and test), each having one z/OS and one coupling facility image. The coupling facility images are sharing the same ICF processor.

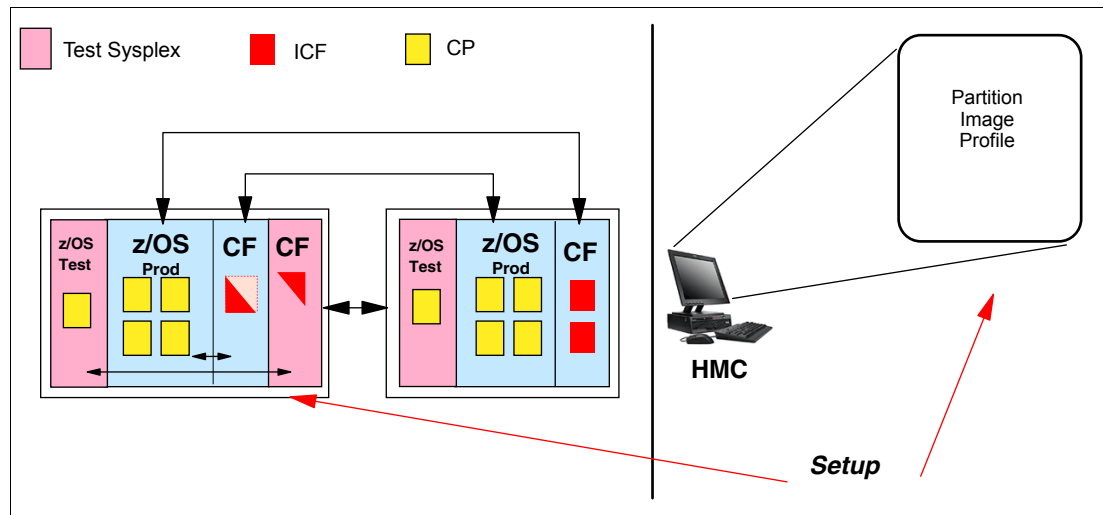


Figure 3-9 ICF options; shared ICFs

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

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

ICFs can be concurrently assigned to an existing configuration through capacity on demand. If the installed books have no remaining non-characterized PUs, the assignment of the next ICF might require the installation of an additional book. An ICF CBU might have been purchased to provide ICF backup capacity for ICFs lost elsewhere. For information about CoD, On/Off CoD, CBU, and CPE see Chapter 8, “System upgrades” on page 233.

Dynamic coupling facility dispatching

The dynamic coupling facility dispatching function has a dispatching algorithm that lets you define a backup coupling facility in a logical partition on the system. When this logical partition is in backup mode, it uses very little processor resources. When the backup CF becomes active, only the resource necessary to provide coupling is allocated.

3.4.4 System z10 Application Assist Processors

A System z10 Application Assist Processor (zAAP) reduces the standard processor (CP) capacity requirements for z/OS Java or XML System Services applications, freeing up capacity for other workload requirements. zAAPs do not increase the MSU value of the processor and therefore do not affect the software license fees.

Note: z/VM V5 R3 and later support zAAP for guest exploitation.

The zAAP is a PU that is used for running z/OS Java or z/OS XML System Services workloads. IBM SDK for z/OS Java 2 Technology Edition (the Java Virtual Machine), in

cooperation with z/OS dispatcher, directs JVM processing from CPs to zAAPs. Also, z/OS XML parsing performed in TCB mode is eligible to be executed on the zAAP processors.

zAAP benefits include:

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

One CP must be installed with or prior to installing a zAAP. The number of zAAPs in a server cannot exceed the number of purchased CPs. Within the capacity of the sum of all purchased PUs in up to four books, up to 32 zAAPs can be characterized. This is on a model E64. Table 3-1 shows the allowed number of zAAPs for each model.

Table 3-1 Number of zAAPs per model

Model	E12	E26	E40	E56	E64
zAAPs	0–6	0–13	0–20	0–28	0–32

Within the limit of all non-characterized PUs available in the installed configuration, zAAPs can be concurrently added to an existing configuration through Capacity on Demand. A zAAP CBU might have been purchased to provide zAAP backup capacity for lost zAAPs elsewhere.

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

For more information about On/Off CoD and CPE see Chapter 8, “System upgrades” on page 233. If the installed books have no remaining unassigned PUs, the assignment of the next zAAP might require the installation of an additional book.

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

zAAPs are orderable by feature code (FC 6814). Up to one zAAP can be ordered for each CP or marked CP configured in the server.

zAAPs and logical partition definitions

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

Restriction: A server cannot have more zAAPs than CPs, as stated before. In a logical partition, as many zAAPs as are available can be defined together with at least one CP.

How zAAPs work

zAAPs are designed for z/OS Java code execution. When Java code must be executed (for example, under control of WebSphere), the z/OS Java Virtual Machine (JVM) calls the function of the zAAP. The z/OS dispatcher then suspends the JVM task on the CP it is running

on and dispatches it on an available zAAP. After the Java application code execution is finished, z/OS redispaches the JVM task on an available CP, after which normal processing is resumed. This process reduces the CP time needed to run Java WebSphere applications, freeing capacity for other workloads.

Figure 3-10 shows the logical flow of Java code running on a z10 EC that has a zAAP available. When JVM starts execution of a Java program, it passes control to the z/OS dispatcher that will verify the availability of a zAAP:

- ▶ If a zAAP is available (not busy), the dispatcher suspends the JVM task on the CP, and assign the Java task to the zAAP. When the task returns control to the JVM, it passes control back to the dispatcher that reassigns the JVM code execution to a CP.
- ▶ If no zAAP is available (all busy) at that time, the z/OS dispatcher may allow a Java task to run on a standard CP, depending on the option used in the OPT statement in the IEAOPTxx member of SYS1.PARMLIB.

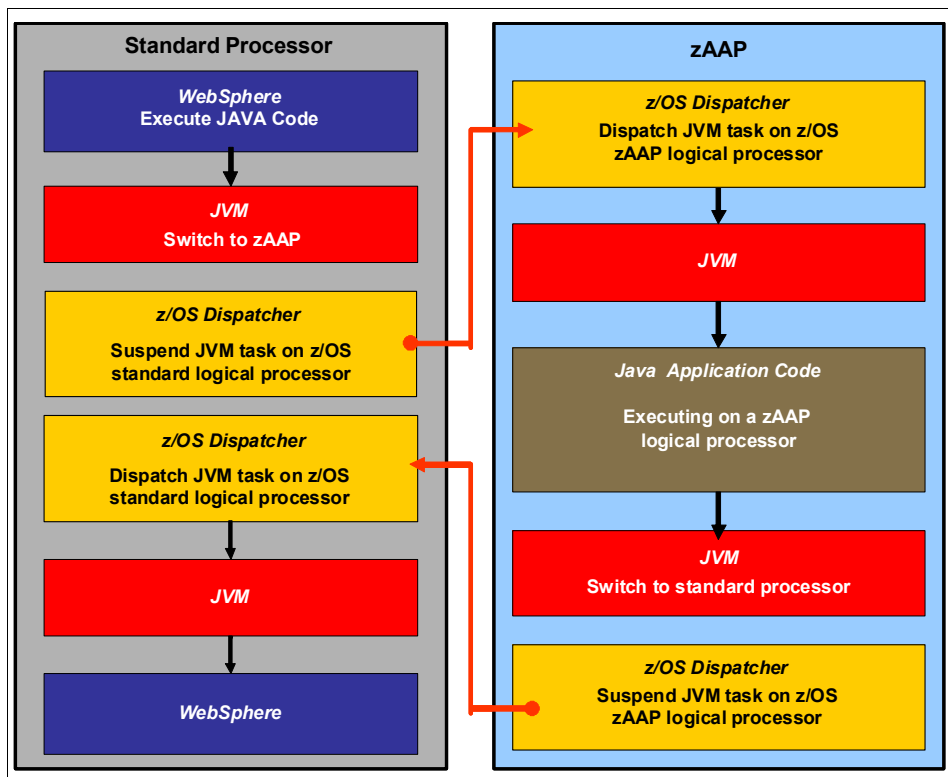


Figure 3-10 Logical flow of Java code execution on a zAAP

Software support

zAAPs do not change the model capacity identifier of the z10 EC. IBM software product license charges based on the model capacity identifier are not affected by the addition of zAAPs. On a z10 EC, z/OS Version 1 Release 7 is the minimum level for supporting zAAPs, together with IBM SDK for z/OS Java 2 Technology Edition V1.4.1.

Exploiters of zAAPs include:

- ▶ Any Java application that is using the current IBM SDK.
- ▶ WebSphere Application Server V5R1 and later, and products based on it, such as WebSphere Portal, WebSphere Enterprise Service Bus (WebSphere ESB), WebSphere Business Integration (WBI) for z/OS and so on.

- ▶ CICS/TS V2R3 and later.
- ▶ DB2 UDB for z/OS Version 8 and later.
- ▶ IMS Version 8 and later.
- ▶ All z/OS XML System Services validation and parsing that execute in TCB mode, which might be eligible for zAAP processing. This eligibility requires z/OS V1R9 and later. For z/OS 1R10 (with appropriate maintenance), middleware and applications requesting z/OS XML System Services can have z/OS XML System Services processing execute on the zAAP.

So that DB2 V9 can exploit a zAAP, the following prerequisites are required:

- ▶ DB2 V9 for z/OS in new function mode
- ▶ The C API for z/OS XML System Services, available with z/OS V1R9 with rollback APARs to z/OS V1R7, and z/OS V1R8
- ▶ One of the following items:
 - z/OS V1R9 has native support.
 - z/OS V1R8 requires an APAR for zAAP support.
 - z/OS V1R7 requires an APAR for zAAP support and an APAR for rollback of z/OS XML System Services.

For more information, see the IBM zAAP Web site:

<http://www-03.ibm.com/systems/z/advantages/zaap/index.html>

The functioning of a zAAP is transparent to all Java programming on JVM V1.4.1 and later.

A zAAP executes only JVM code. JVM is the only authorized user of a zAAP in association with some parts of system code, such as the z/OS dispatcher and supervisor services. A zAAP is not able to process I/O or clock comparator interruptions and does not support operator controls such as IPL.

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

Three execution options for Java code execution are available. These options are user specified in IEAOPTxx and can be dynamically altered by the SET OPT command. The current options that are supported for z/OS V1R8 are:

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

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

This option has the effect of running as much zAAP-eligible work on zAAPs as possible and only allowing it to spill over onto the CPs when the zAAPs are overcommitted.
- ▶ Option 2: Java dispatching by priority (IFAHONORPRIORITY=NO)

zAAP-eligible work executes on zAAPs only while at least one zAAP engine is online. zAAP-eligible work is not normally dispatched on a CP, even if the zAAPs are overcommitted and CPs are unused. The exception to this is that zAAP-eligible work sometimes run on a CP to resolve resource conflicts, and other reasons.

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

- ▶ Option 3: Java discretionary crossover (IFACROSSOVER=YES or NO)

As of z/OS V1R8 (and the IBM zIIP Support for z/OS V1R7 Web deliverable), the IFACROSSOVER parameter is no longer honored.

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

3.4.5 System z10 Integrated Information Processor

A System z10 Integrated Information Processor (zIIP) enables eligible workloads to work with z/OS and have a portion of the workload's enclave service request block (SRB) work directed to the zIIP. The zIIPs do not increase the MSU value of the processor and therefore do not affect the software license fee.

Note: z/VM V5R3 and later support the zIIP for guest exploitation.

z/OS Communication Server and DB2 UDB for z/OS Version 8 (and later) exploit the zIIP by indicating to z/OS which portions of the work are eligible to be routed to a zIIP.

Types of eligible DB2 UDB for z/OS V8 (and later) workloads executing in SRB mode include:

- ▶ Query processing of network-connected applications that access the DB2 database over a TCP/IP connection using Distributed Relational Database Architecture™ (DRDA).

DRDA enables relational data to be distributed among multiple platforms. It is native to DB2 for z/OS, thus reducing the need for additional gateway products that can affect performance and availability. The application uses the DRDA requester or server to access a remote database. (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 additional dimension tables holding information about each perspective of the data. A star schema query, for example, joins several dimensions of a star schema data set.

- ▶ DB2 utilities that are used for index maintenance, such as LOAD, REORG, and REBUILD

Indices allow quick access to table rows, but over time as data in large databases is manipulated, they become less efficient and have to be maintained.

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

z/OS Communications Server exploits the zIIP for eligible Internet Protocol Security (IPSec) network encryption workloads. This requires z/OS V1R8 with PTFs or z/OS V1R9. 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 performing the encryption processing, the zIIP also handles cryptographic validation of message integrity and IPSec header processing.

z/OS Global Mirror (zGM), formerly known as Extended Remote Copy (XRC), exploits the zIIP too. Most z/OS DFSMS SDM (System Data Mover) processing associated with zGM is eligible to run on the zIIP. This requires z/OS V1R10, z/OS V1R9, or z/OS V1R8 with PTFs.

The first IBM exploiter of z/OS XML System Services is DB2 V9. With regard to DB2 V9 prior to 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 when processing is used as part of any workload running in enclave SRB mode.

For more information, see the IBM zIIP Web site:

<http://www-03.ibm.com/systems/z/advantages/zIIP/about.html>

zIIP installation information

One CP must be installed with or prior to any zIIP being installed. The number of zIIPs in a server cannot exceed the number of CPs and unassigned CPs in that server. Within the capacity of the sum of all unassigned PUs in up to four books, up to 32 zIIPs on a model E64 can be characterized. Table 3-2 shows the maximum number of zIIPs per model.

Table 3-2 Maximum number of zIIPs per model

Model	E12	E26	E40	E56	E64
Maximum zIIPs	6	13	20	28	32

zIIPs are orderable by feature code (FC 6815). Up to one zIIP can be ordered for each CP or marked CP configured in the server. If the installed books have no remaining unassigned PUs, the assignment of the next zIIP may require the installation of an additional book.

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

Within the limit of all non-characterized PUs available in the installed configuration, zIIPs can be added concurrently to an existing configuration through Capacity on Demand. zIIP capacity can be purchased to provide zIIP backup capacity.

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

For more information about capacity on demand see Chapter 8, “System upgrades” on page 233.

zIIPs and logical partition definitions

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

Restriction: A server cannot have more zIIPs than CPs. However, in a logical partition, as many zIIPs as are available can be defined together with at least one CP.

3.4.6 zAAP on zIIP capability

As described previously, zAAPs and zIIPs support different types of workloads. However, there are installations that do not have enough eligible workloads to justify buying a zAAP or a zAAP and a zIIP. IBM is now making available the possibility of combining zAAP and zIIP workloads on zIIP processors, provided that no zAAPs are installed on the server. This may provide the following benefits:

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

This capability does not eliminate the need to have one or more CPs for every zIIP processor in the server. Support is provided by z/OS. See 7.3.2, “zAAP on zIIP capability” on page 202.

When zAAPs are present¹ this capability is not available, as it is neither intended as a replacement for zAAPs, which continue to be available, nor as an overflow possibility for zAAPs. IBM does not recommend converting zAAPs to zIIPs in order to take advantage of the zAAP to zIIP capability:

- ▶ Having both zAAPs and zIIPs maximizes the system potential for new workloads.
- ▶ zAAPs have been available for over five years and there may exist applications or middleware with zAAP-specific code dependencies. For example, the code may use the number of installed zAAP engines to optimize multithreading performance.

We recommend planning and testing before eliminating all zAAPs, as there may be application code dependencies that may affect performance.

3.4.7 System assist processors

A system assist processor (SAP) is a PU that runs the channel subsystem Licensed Internal Code to control I/O operations.

All SAPs perform I/O operations for all logical partitions. All models have standard SAPs configured. Model E12 has three SAPs, model E26 has six SAPs, model E40 has nine SAPs, and models E54 and E64 have 10 and 11 SAPs, respectively, as the standard configuration.

SAP configuration

A standard SAP configuration provides a very well-balanced system for most environments. However, there are application environments with very high I/O rates (typically some TPF environments). In this case, optional additional SAPs can be ordered. Assignment of additional SAPs can increase the capability of the channel subsystem to perform I/O operations. In z10 EC servers, the number of SAPs can be greater than the number of CPs.

¹ The zAAP on zIIP capability is available to z/OS when running as a guest of z/VM on machines with zAAPs installed, provided that no zAAPs are defined to the z/VM LPAR. This would allow, for instance, testing this capability to estimate usage before committing to production.

Optional additional orderable SAPs

An option available on all models is additional orderable SAPs. These additional SAPs increase the capacity of the channel subsystem to perform I/O operations, usually suggested for Transaction Processing Facility (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 are listed in Table 3-3.

Table 3-3 Optional SAPs per model

Model	E12	E26	E40	E56	E64
Optional SAPs	0–3	0–7	0–11	0–18	0–21

Optionally assignable SAPs

Assigned CPs may be optionally reassigned as SAPs instead of CPs by using the reset profile on the Hardware Management Console (HMC). This reassignment increases the capacity of the channel subsystem to perform I/O operations, usually for some specific workloads or I/O-intensive testing environments.

If you intend to activate a modified server configuration with a modified SAP configuration, a reduction in the number of CPs available reduces the number of logical processors that can be activated. Activation of a logical partition can fail if the number of logical processors that you attempt to activate exceeds the number of CPs available. To avoid a logical partition activation failure, verify that the number of logical processors assigned to a logical partition does not exceed the number of CPs available.

3.4.8 Reserved processors

Reserved processors are defined by the Processor Resource/System Manager (PR/SM) to allow for a nondisruptive *capacity* upgrade. Reserved processors are like spare *logical* processors. They can be shared or dedicated.

Reserved CPs should be defined to a logical partition to allow for nondisruptive *image* upgrades. If the operating system in the logical partition supports the logical processor add function, reserved processors are no longer needed.

Notes:

- ▶ z/OS V1 R10 supports logical processor add.
- ▶ z/OS V1R8 and z/OS V1R7 support up to 32 processors.
- ▶ z/OS V1R9 supports up to 64 processors including CPs, zAAPs, and zIIPs.
- ▶ z/VM V5R3 supports up to 32 processors.
- ▶ z/VM V5R3 with PTFs supports logical processor add.

Reserved processors can be dynamically configured online by an operating system that supports this function, if enough unassigned PUs are available to satisfy this request. The PR/SM rules regarding logical processor activation remain unchanged.

Reserved processors provide the capability to define to a logical partition more logical processors than the number of available CPs, IFLs, ICFs, zAAPs, and zIIPs in the configuration. This makes it possible to configure online, nondisruptively, more logical processors after additional CPs, IFLs, ICFs, zAAPs, and zIIPs have been made available concurrently with one of the Capacity on Demand options.

On model E56 and lower, a logical partition can have up to 56 logical CPs defined, which is the sum of initial and reserved logical CPs. A partition can specify a total of 64 logical processors of any type (CPs, zAAPs, zIIPs, IFLs) if the number of logical CPs is not larger than 56. On model E64, the sum of initial and reserved logical CPs defined to a partition can be 64. The maximum number of logical processors of all types (CPs, zAAPs, zIIPs, IFLs) still cannot exceed 64.

When no reserved processors are defined to a logical partition, an addition of a processor to that logical partition is disruptive, requiring the following tasks:

1. Partition deactivation
2. A logical processor definition change
3. Partition activation

The maximum number of reserved processors that can be defined to a logical partition depends on the number of logical processors that are already defined.

Do not define more active and reserved processors than the operating system for the logical partition can support. For more information about logical processors and reserved processors definition see 3.6, “Logical partitioning” on page 90.

3.4.9 Processor unit characterization

Processor unit characterization is done at power-on reset time when the server is initialized. The z10 EC is always initialized in LPAR mode, and it is the PR/SM hypervisor that has responsibility for the PU assignment.

Additional SAPs are characterized first, then CPs, followed by IFLs, ICFs, zAAPs, and zIIPs. For performance reasons, CPs for a logical partition are grouped together as much as possible. Having all CPs grouped in as few books as possible limits memory and cache interference to a minimum.

When an additional book is added concurrently after power-on reset and new logical partitions are activated, or processor capacity for active partitions is dynamically expanded, the additional PU capacity may be assigned from the new book. It is only after the next power-on reset that the processing unit allocation rules take into consideration the newly installed book.

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

Characterized PUs, whether CPs, IFLs, ICFs, zAAPs, zIIPs, or SAPs, are transparently spared, following distinct rules.

The z10 EC has two spare PUs that can be used throughout the system. Depending on the model, sparing of CP, IFL, ICF, zAAP, zIIP, and SAP is completely transparent and does not require an operating system or operator intervention.

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

Application preservation

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

3.4.11 Dynamic SAP sparing and reassignment

Dynamic recovery is provided in case of failure of the system assist processor (SAP). If the SAP fails, and if a spare PU is available, the spare PU is dynamically assigned as a new SAP. If no spare PU is available, and more than one CP is characterized, a characterized CP is reassigned as an SAP. In either case, customer intervention is not required. This capability eliminates an unplanned outage and permits a service action to be deferred to a more convenient time.

Sparing rules

Two PUs are reserved as spares. The reserved spares are available to replace two PUs. The spare PUs can be used for sparing any characterization, whether it is a CP, IFL, ICF, zAAP, zIIP, or SAP. On a model E12, two spares are located in the one book present. In multibook systems, the two spares are distributed across the books. For the location of the spares in a multibook system see Table 2-3 on page 34.

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

Sparing rules are as follows:

- ▶ When a PU failure occurs on a chip that has four active cores, the two standard spare PUs are used to recover the failing PU and the parent PU that shares function (for example, the compression unit and CPACF) with the failing PU, even though only one of the PUs has failed.
- ▶ When a PU failure occurs on a chip that has three active cores, one standard spare PU is used to replace the PU that does not share any function with another PU.
- ▶ When no spares are left, non-characterized PUs are used for sparing, following the previous two rules.

The system does not issue a call to the Remote Support Facility (RSF) in any of the above circumstances. When non-characterized PUs are used for sparing and might be required to satisfy an On/Off CoD request, an RSF call occurs to request a book repair.

3.4.12 Increased flexibility with z/VM-mode partitions

System z10 EC provides a capability for the definition of a z/VM-mode logical partition (LPAR) that contains a mix of processor types including CPs and specialty processors, such as IFLs, zIIPs, zAAPs, and ICFs.

z/VM V5R4 and later support this capability, which increases flexibility and simplifies systems management. In a single LPAR, z/VM can:

- ▶ Manage guests that exploit Linux on System z on IFLs, z/VSE and z/OS on CPs.
- ▶ Execute designated z/OS workloads, such as parts of DB2 DRDA processing and XML, on zIIPs.
- ▶ Provide an economical Java execution environment under z/OS on zAAPs.

3.5 Memory design

For PUs and the I/O subsystem designs, the memory design equally provides flexibility and high availability, allowing:

- ▶ Concurrent memory upgrades (if the physically installed capacity is not yet reached)
The z10 EC may have more physically installed memory than the initial available capacity. Memory upgrades within the physically installed capacity can be done concurrently by the Licensed Internal Code, and no hardware changes are required. Concurrent memory upgrades can be done through Capacity on Demand. Note that memory upgrades *cannot* be done through Capacity BackUp (CBU) or On/Off CoD. For more information see Table 8-2 on page 245.
- ▶ Concurrent memory upgrades (if the physically installed capacity is reached)
Physical memory upgrades require a book to be removed and re-installed after having replaced the memory cards in the book. Except for a model E12, the combination of enhanced book availability and the flexible memory option allow you to concurrently add memory to the system. For more information see 2.5.3, “Book replacement and memory” on page 39, and 2.5.4, “Flexible memory option” on page 39.
- ▶ Partial memory restart
In the rare event of a memory card failure, a partial-memory restart enables the system to be restarted with only part of the original memory. In a one-book system, the memory DIMMs that make up logical pair 0 or logical pair 1 (depending on where the failure resides) are deactivated, after which the system can be restarted with the memory in the remaining logical pair cards.

In a multibook system, all physical memory in the book containing the failing memory is taken offline, which allows you to bring up the system with the remaining physical memory in the other books. In this way, processing can be resumed until a replacement memory card is installed.

The memory DIMMs use the latest fast 1 Gb synchronous DRAMs. Memory access is interleaved to equalize memory activity across the DIMMs. Memory DIMMs have 4 GB or 8 GB of capacity. DIMMs installed in a book do not necessarily have the same capacity (as long as the DRAM sizes are the same). Books may contain different memory sizes.

The total capacity installed may have more usable memory than required for a configuration, and Licensed Internal Code Configuration Control (LICCC) determines how much memory is used from each card. The sum of the LICCC provided memory from each card is the amount available for use in the system.

Memory allocation

Memory assignment or allocation is done at power-on reset (POR) when the system is initialized. PR/SM is responsible for the memory assignments.

PR/SM has knowledge of the amount of purchased memory and how it relates to the available physical memory in each of the installed books. PR/SM has control over all physical memory and therefore is able to make physical memory available to the configuration when a book is nondisruptively added. PR/SM also controls the reassignment of the content of a specific physical memory array in one book to a memory array in another book. This is known as the memory copy/reassign function, which is used to reallocate the memory content from the memory in a book to another memory location. It is used when enhanced book availability is applied to concurrently remove and re-install a book in case of an upgrade or repair action.

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

Large page support

By default, page frames are allocated with a 4 KB size. The z10 EC supports a large page size of 1 MB. The first z/OS release that supports large pages is z/OS V1R9. Linux on System z support for large pages is available in SLES 10 SP2 and RHEL 5.2.

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

It is very desirable to have one's addresses in the TLB. With 4 K pages, holding all the addresses for 1 MB of storage takes 256 TLB lines. When using 1 MB pages, it takes only 1 TLB line. This means that large page size exploiters have a much smaller TLB footprint.

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

Exploiters of large pages are better represented in the TLB and are expected to see performance improvement in both elapsed time and CPU time. This is because DAT and memory operations are part of CPU busy time even though the CPU waits for memory operations to complete without processing anything else in the meantime.

Overhead is associated with creating a 1 MB page. To overcome that overhead, a process has to run for a period of time and maintain frequent memory access to keep the pertinent addresses in the TLB.

Very short-running work does not overcome the overhead; short processes with small working sets are expected to provide 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. So, a very long-running process can benefit somewhat even without frequent memory access. You should weigh the benefits of whether something in this category should use large pages as a result of the system-level costs of tying up real storage.

There is a balance between the performance of a process using large pages, and the performance of the remaining work on the system.

Large pages are treated as fixed pages. They are only available for 64-bit virtual private storage such as virtual memory located above 2 GB. Decide on the use of large pages based on knowledge of memory usage and page address translation overhead for a specific workload.

One would be inclined to think, that increasing the TLB size is a feasible option to deal with TLB-miss situations. However, this is not as straightforward as it seems. As the size of the TLB increases, so does the overhead involved in managing the TLB's contents. Correct sizing of the TLB is subject to very complex statistical modelling in order to find the optimal trade-off between size and performance.

3.5.1 Central storage

Central storage (CS) consists of main storage, addressable by programs, and storage not directly addressable by programs. Non-addressable storage includes the hardware system area (HSA). Central storage provides:

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

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

3.5.2 Expanded storage

Expanded storage (ES) can optionally be defined on z10 EC servers. Expanded storage is physically a section of processor storage. It is controlled by the operating system and transfers 4 KB pages to and from central storage.

Except for z/VM, z/Architecture operating systems do *not* use expanded storage. Because they operate in 64-bit addressing mode, they can have all the required storage capacity allocated as central storage. z/VM is an exception because, even when operating in 64-bit mode, it can have guest virtual machines running in 31-bit addressing mode, which can use expanded storage. In addition, z/VM exploits expanded storage for its own operations.

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

The z10 EC only runs in LPAR mode. Storage is placed into a single storage pool called LPAR Single storage pool, which can be dynamically converted to expanded storage and back to central storage as needed when partitions are activated or de-activated.

LPAR single storage pool

In LPAR mode, storage is not split into central storage and expanded storage at power-on reset. Rather, the storage is placed into a single central storage pool that is dynamically assigned to expanded storage and back to central storage, as needed.

On the Hardware Management Console, the Storage Assignment tab of a reset profile shows the *customer storage*, which is the total installed storage minus the 16 GB hardware system area. Logical partitions are still defined to have central storage and, optionally, expanded storage.

Activation of logical partitions and dynamic storage reconfiguration cause the storage to be assigned to the type needed (central or expanded), and does not require a power-on reset.

3.5.3 Hardware system area

The hardware system area (HSA) is a non-addressable storage area that contains server Licensed Internal Code and configuration-dependent control blocks. The HSA has a fixed size of 16 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 HCD/IOCP always reserves:

- ▶ Four channel subsystems (CSSs)
- ▶ Fifteen logical partitions in each CSS for a total of 60 logical partitions
- ▶ Subchannel set 0 with 63.75 K devices in each CSS
- ▶ Subchannel set 1 with 64 K devices in each CSS

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

3.6 Logical partitioning

Logical partitioning is a function implemented by the Processor Resource/Systems Manager (PR/SM) on all z10 EC servers. The z10 EC runs only in LPAR mode. This means that all system aspects are controlled by PR/SM functions.

PR/SM is aware of the book structure on the z10 EC. Logical partitions, however, do not have this awareness. Logical partitions have resources allocated to them from a variety of physical resources. From a systems standpoint, logical partitions 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.

PR/SM always attempts to allocate all real storage for a logical partition within one book, and attempts to dispatch a logical PU on a physical PU in a book that also has the central storage for that logical partition. If this is not possible, a PU in an adjacent book is chosen.

In general, PR/SM tries to minimize the number of books required to allocate the resources of a given logical partition. In addition, PR/SM always tries to redispach a logical PU on the same physical PU to assure that as much as possible of the L1 cache content can be reused.

On the z10 EC, support for affinity is more advanced. PR/SM and z/OS now work in tandem to more efficiently use processor resources. HiperDispatch is a function that combines the dispatcher actions and the knowledge that PR/SM has about the topology of the server. For that purpose, the z/OS dispatcher manages multiple queues with an average number of four CPs per queue and uses these queues to assign work to as few logical processors as are needed for a given logical partition workload. So, even if the logical partition is defined with a large number of logical processors, HiperDispatch optimizes this number of processors

nearest to the required capacity. The optimal number of processors to be used are kept within a book boundary where possible, preventing L2 cache misses that would have occurred when the dispatcher dispatched work, where a processor might be available.

PR/SM enables z10 EC servers to be initialized for a logically partitioned operation, supporting up to 60 logical partitions. Each logical partition can run its own operating system image in any image mode, independent from the other logical partitions.

A logical partition can be added, removed, activated, or deactivated at any time. Changing the number of logical partitions is not disruptive and does not require power-on reset (POR). Several facilities might not be available to all operating systems, because the facilities might have software corequisites.

Each logical partition has the same resources as a real CPC. They are processors, memory, and channels:

- ▶ Processors

Called *logical processors*, they can be defined as CPs, IFLs, ICFs, zAAPs, or zIIPs. They can be dedicated to a logical partition or shared among logical partitions. When shared, a processor weight can be defined to provide the required level of processor resources to a logical partition. Also, the capping option can be turned on, which prevents a logical partition from acquiring more than its defined weight, limiting its processor consumption.

Logical partitions for z/OS can have CP, zAAP, and zIIP logical processors. All three logical processor types can be defined as either all dedicated or all shared. The zAAP and zIIP support is available in z/OS.

Figure 3-11 shows the logical processor assignment window of the Customize Image Profiles in the Hardware Management Console. The panel allows the definition of:

- Dedicated or shared logical processors, including CPs, zAAPs, and zIIPs
- Initial weight, capping option, enable workload manager option, and minimum and maximum processing weight for shared CPs, zAAPs, and zIIPs
- Optional group profile name to which the logical partition is assigned
- Number of initial and optional reserved CPs, zAAPs, and zIIPs
- Sum of initial and reserved logical processors in a logical partition (limited to 64). z/OS V1R7 supports 32, z/OS V1R8 supports 54, and z/OS V1R9 supports 64 logical processors in a logical partition. The limit applies to the sum of CP, zAAP, and zIIP logical processors. z/VM V5R3 and later support 32 processors.

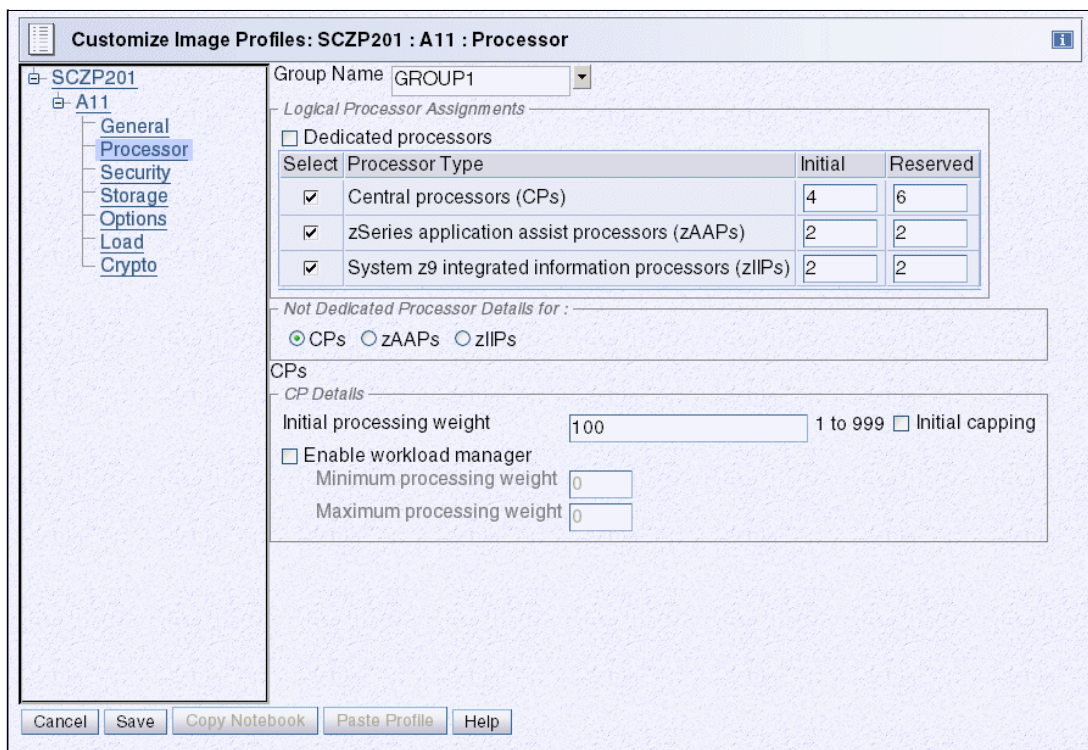


Figure 3-11 Customize Image Profiles: Processor page

The weight and the number of online logical processors of a logical partition can be dynamically managed by the LPAR CPU Management function of the Intelligent Resource Director to achieve the defined goals of this specific partition and of the overall system. The provisioning architecture of the z10 EC, described in Chapter 8, “System upgrades” on page 233, adds another dimension to dynamic management of logical partitions.

For z/OS Workload License Charge (WLC), a logical partition *defined capacity* can be set, enabling the soft capping function. Workload charging introduces the capability to pay software license fees based on the size of the logical partition on which the product is running, rather than on the total capacity of the server, as follows:

- 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 logical partition when soft capping is selected.

The defined capacity value is specified on the Options tab on the Customize Image Profiles panel.

- WLM keeps a 4-hour rolling average of the CPU usage of the logical partition, and when the 4-hour average CPU 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 regarding WLM, see *System Programmer's Guide to: Workload Manager*, SG24-6472.

Note: When defined capacity is used to define an uncapped logical partition's capacity, looking carefully at the weight settings of that logical partition is important. If the weight is much smaller than the defined capacity, PR/SM will use a discontinuous cap pattern to achieve the defined capacity setting. This means PR/SM will alternate between capping the LPAR at the MSU value corresponding to the relative weight settings, and no capping at all. It is recommended to avoid this case, and try to establish a defined capacity which is equal or close to the relative weight.

► **Memory**

Memory, either central storage or expanded storage, must be dedicated to a logical partition. The defined storage must be available during the logical partition activation. Otherwise, the activation fails.

Reserved storage can be defined to a logical partition, enabling nondisruptive memory addition to and removal from a logical partition, using the LPAR dynamic storage reconfiguration (z/OS and z/VM V5 R4). For more information see 3.6.4, "LPAR dynamic storage reconfiguration" on page 100.

► **Channels**

Channels can be shared between logical partitions by including the partition name in the partition list of a Channel Path ID (CHPID). I/O configurations are defined by the input/output configuration program (IOCP) or the Hardware Configuration Dialog (HCD) in conjunction with the CHPID mapping tool (CMT). The CMT is an optional, but strongly recommended, tool used to map CHPIDs onto physical channel IDs (PCHIDs) that represent the physical location of a port on a card in an I/O cage.

IOCP is available on the z/OS, z/VM, VM/ESA®, and z/VSE operating systems, and as a stand-alone program on the hardware console. HCD is available on z/OS and z/VM operating systems.

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

Modes of operation

Table 3-4 shows the modes of operation, summarizing all available mode combinations: operating modes and their processor types, operating systems, and addressing modes.

Table 3-4 z10 EC modes of operation

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

Image mode	PU type	Operating system	Addressing mode
ESA/390 TPF mode	CP <i>only</i>	TPF	31-bit
	CP <i>only</i>	z/TPF	64-bit
Coupling facility mode	ICF or CP, or both	CFCC	64-bit
Linux-only mode	IFL <i>or</i> CP	Linux on System z (64-bit)	64-bit
		z/VM	
		Linux on System z (31-bit)	31-bit
z/VM-mode	CP, IFL, zIIP, zAAP, ICF	z/VM	64-bit

The 64-bit z/Architecture mode has no special operating mode because the architecture mode is not an attribute of the definable images operating mode. The 64-bit operating systems are IPLed in 31-bit mode and, optionally, can change to 64-bit mode during their initialization. The operating system is responsible for taking advantage of the addressing capabilities provided by the architectural mode.

For information about operating system support see Chapter 7, “Software support” on page 189.

Logically partitioned mode

The z10 EC only runs in LPAR mode. Each of the 60 logical partitions can be defined to operate in one of the following image modes:

- ▶ ESA/390 mode, to run:
 - A z/Architecture operating system, on dedicated *or* shared CPs
 - An ESA/390 operating system, on dedicated *or* shared CPs
 - A Linux operating system, on dedicated *or* shared CPs
 - z/OS, on any of the following processors:
 - Dedicated *or* shared CPs
 - Dedicated CPs *and* dedicated zAAPs *or* zIIPs
 - Shared CPs *and* shared zAAPs *or* zIIPs

Note: zAAPs and zIIPs can be defined to an ESA/390 mode or z/VM-mode image (Table 3-4 on page 93). However, zAAPs and zIIPs are supported only by z/OS. Other operating systems cannot use zAAPs or zIIPs, even if they are defined to the logical partition. z/VM V5R3 and later can provide zAAPs or zIIPs to a guest z/OS.

- ▶ ESA/390 TPF mode, to run TPF or z/TPF operating system, on dedicated *or* shared CPs
- ▶ Coupling facility mode, by loading the CFCC code into the logical partition defined as:
 - Dedicated *or* shared CPs
 - Dedicated *or* shared ICFs

- ▶ Linux-only mode, to run:
 - A Linux operating system, on either:
 - Dedicated *or* shared IFLs
 - Dedicated *or* shared CPs
 - A z/VM operating system, on either:
 - Dedicated *or* shared IFLs
 - Dedicated *or* shared CPs
- ▶ z/VM-mode to run z/VM on dedicated *or* shared CPs or IFLs, plus zAAPs, zIIPs, and ICFs.

Table 3-5 shows all LPAR modes, required characterized PUs, operating systems, and the PU characterizations that can be configured to a logical partition image. The available combinations of dedicated (DED) and shared (SHR) processors are also shown. For all combinations, a logical partition can also have reserved processors defined, allowing nondisruptive logical partition upgrades.

Table 3-5 LPAR mode and PU usage

LPAR mode	PU type	Operating systems	PUs usage
ESA/390	CPs	z/Architecture operating systems ESA/390 operating systems Linux on System z	CPs DED <i>or</i> CPs SHR
	CPs <i>and</i> zAAPs <i>or</i> zIIPs	z/OS z/VM (V5R3 and later for guest exploitation)	CPs DED <i>and</i> zAAPs DED, <i>and</i> (<i>or</i>) zIIPs DED <i>or</i> CPs SHR <i>and</i> zAAPs SHR <i>or</i> zIIPs SHR
ESA/390 TPF	CPs	TPF z/TPF	CPs DED <i>or</i> CPs SHR
Coupling facility	ICFs <i>or</i> CPs	CFCC	ICFs DED <i>or</i> ICFs SHR, <i>or</i> CPs DED <i>or</i> CPs SHR
Linux only	IFLs <i>or</i> CPs	Linux on System z z/VM	IFLs DED <i>or</i> IFLs SHR, <i>or</i> CPs DED <i>or</i> CPs SHR
z/VM-mode	CPs, IFLs, zAAPs, zIIPs, ICFs	z/VM	All PUs must be SHR or DED

Dynamic add or delete of a logical partition name

Dynamic add or delete of a logical partition name is the ability to add or delete logical partitions and their associated I/O resources to or from the configuration without a power-on reset.

The extra channel subsystem and MIF image ID pairs (CSSID/MIFID) can later be assigned to a logical partition for use (or later removed) through dynamic I/O commands using the Hardware Configuration Definition (HCD). At the same time, required channels have to be defined for the new logical partition.

Attention: Cryptographic coprocessors are not tied to partition numbers or MIF IDs. They are set up with AP numbers and domain indices. These are assigned to a partition profile of a given name. The customer assigns these *lanes* to the partitions and continues to have the responsibility to clear them out when their users change.

Add Crypto feature to a logical partition

You can preplan the addition of Crypto Express features to a logical partition on the Crypto page in the image profile by defining the Cryptographic Candidate List, Cryptographic Online List and usage, and Control Domain Indices in advance of installation. By using the Change LPAR Cryptographic Controls task, adding Crypto dynamically to a logical partition without an outage of the logical partition is possible. Also, dynamic deletion or moving of these features no longer requires pre-planning. Support is provided in z/OS, z/VM, z/VSE, and Linux on System z.

LPAR group capacity limit

The group capacity limit feature allows the definition of a capacity limit for a group of logical partitions on z10 EC servers. This feature allows a capacity limit to be defined for each logical partition running z/OS, and to define a group of logical partitions on a server. This allows the system to manage the group in such a way that the sum of the LPAR group capacity limits in MSUs per hour will not be exceeded. To take advantage of this, you must be running z/OS V1.8 or later and all logical partitions in the group have to be z/OS V1.8 and later.

PR/SM and WLM work together to enforce the capacity defined for the group and enforce the capacity optionally defined for each individual logical partition.

LPAR dynamic PU reassignment

System configuration has been enhanced to optimize the CPU-to-book allocation of physical processors dynamically. The initial allocation of customer usable physical processors to physical books can change dynamically to better suit the actual logical partition configurations that are in use. Swapping of specialty engines and general processors with each other, with spare PUs, or with both, can occur as the system attempts to compact logical partition configurations into physical configurations that span the least number of books. The effect of this is evident in dedicated and shared partitions that use HiperDispatch.

LPAR dynamic PU reassignment is available only to System z10 and is transparent to operating systems.

3.6.1 Storage operations

In z10 EC servers, memory can be assigned as a combination of central storage and expanded storage, supporting up to 60 logical partitions. Expanded storage is only used by the z/VM operating system.

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

Central storage can be dynamically assigned to expanded storage and back to central storage as needed without a power-on reset (POR). For details see “LPAR single storage pool” on page 89.

Memory *cannot* be shared between system images. It is possible to dynamically reallocate storage resources for z/Architecture logical partitions running operating systems that support dynamic storage reconfiguration (DSR). This is supported by z/OS and z/VM V5R4. z/VM in turn virtualizes this support to its guests. For details see 3.6.4, “LPAR dynamic storage reconfiguration” on page 100.

Operating systems running under z/VM can exploit 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 z10 EC storage *allocation* and *usage* possibilities, depending on the image mode.

Table 3-6 Storage definition and usage possibilities

Image mode	Architecture mode (addressability)	Maximum central storage		Expanded storage	
		Architecture	z10 EC definition	z10 EC definable	Operating system usage ^a
ESA/390	z/Architecture (64-bit)	16 EB	1 TB	Yes	Yes
	ESA/390 (31-bit)	2 GB	128 GB	Yes	Yes
z/VM ^b	z/Architecture (64-bit)	16 EB	256 GB	Yes	Yes
ESA/390 TPF	ESA/390 (31-bit)	2 GB	2 GB	Yes	No
Coupling facility	CFCC (64-bit)	1.5 TB	1 TB	No	No
Linux only	z/Architecture (64-bit)	16 EB	256 GB	Yes	<i>Only by z/VM</i>
	ESA/390 (31-bit)	2 GB	2 GB	Yes	<i>Only by z/VM</i>

a. z/VM supports the use of expanded storage.

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

ESA/390 mode

In ESA/390 mode, storage addressing can be 31 or 64 bits, depending on the operating system architecture *and* the operating system configuration.

An ESA/390 mode image is always initiated in 31-bit addressing mode. During its initialization, a z/Architecture operating system can change it to 64-bit addressing mode and operate in the z/Architecture mode.

Some z/Architecture operating systems, such as z/OS, *always* change the 31-bit addressing mode and operate in 64-bit mode. Other z/Architecture operating systems, such as z/VM, can be configured to change to 64-bit mode or to stay in 31-bit mode and operate in the ESA/390 architecture mode.

The modes are:

► z/Architecture mode

In z/Architecture mode, storage addressing is 64-bit, allowing for virtual addresses up to 16 exabytes (16 EB). The 64-bit architecture theoretically allows a maximum of 16 EB to be used as central storage. However, the current central storage limit for logical partitions is 1 TB of central storage. The operating system that runs in z/Architecture mode has to be able to support the real storage. Currently, z/OS for example, supports up to 4 TB of real storage (z/OS V1.8 and higher releases).

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

► **ESA/390 architecture mode**

In ESA/390 architecture mode, storage addressing is 31-bit, allowing for virtual addresses up to 2 GB. A maximum of 2 GB can be used for central storage. Because the processor storage can be configured as central and expanded storage, memory above 2 GB may be configured as expanded storage. In addition, this mode permits the use of either 24-bit or 31-bit addressing, under program control.

Because an ESA/390 mode image can be defined with up to 128 GB of central storage, the central storage above 2 GB is *not* used, but remains configured to this image.

Note: Either a z/Architecture mode or an ESA/390 architecture mode operating system can run in an ESA/390 image on a z10 EC. Any ESA/390 image can be defined with more than 2 GB of central storage *and* can have expanded storage. These options allow you to configure more storage resources than the operating system is capable of addressing.

z/VM-mode

In z/VM-mode, several types of System z10 processors can be defined within one LPAR. This increases flexibility and simplifies systems management by allowing z/VM to perform the following tasks all in the same z/VM LPAR:

- Manage guests to operate Linux on System z on IFLs
- Operate z/VSE and z/OS on CPs
- Offload z/OS system software overhead, such as DB2 workloads on zIIPs
- Provide an economical Java execution environment under z/OS on zAAPs

This support is exclusive for the z10 and is supported by z/VM V5R4 and later.

ESA/390 TPF mode

In ESA/390 TPF mode, storage addressing follows the ESA/390 architecture mode; the TPF/ESA operating system runs in the 31-bit addressing mode.

Coupling facility mode

In coupling facility mode, storage addressing is 64-bit for a coupling facility image running CFCC Level 12 or later, allowing for an addressing range up to 16 EB. However, the current z10 EC definition limit for logical partitions is 1 TB of storage.

CFCC Level 15 and CFCC Level 16 are available for the z10 EC. CFCC Level 16 provides important functions:

- CF Duplexing enhancements
- List notification improvements

For details see “Coupling Facility Control Code” on page 103.

Expanded storage cannot be defined for a coupling facility image. Only IBM Coupling Facility Control Code can run in coupling facility mode. See the *System z10 Enterprise Class Processor Resource/Systems Manager Planning Guide*, SB10-7153.

Linux-only mode

In Linux-only mode, storage addressing can be 31-bit or 64-bit, depending on the operating system architecture *and* the operating system configuration, in exactly the same way as in ESA/390 mode.

Only Linux and z/VM operating systems can run in Linux-only mode, as follows:

- ▶ Linux on System z 64-bit distributions (SLES 9, SLES 10, SLES 11, RHEL 4, RHEL 5) use 64-bit addressing and operate in the z/Architecture mode.
- ▶ Linux on System z 31-bit distributions (SLES 9, RHEL 4) use 31-bit addressing and operate in the ESA/390 Architecture mode.
- ▶ z/VM uses 64-bit addressing and operates in the z/Architecture mode.

3.6.2 Reserved storage

Reserved storage can optionally be defined to a logical partition, 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 the coupling facility mode.

A logical partition must define an amount of central storage and, optionally (if not a coupling facility image), an amount of expanded storage. Both central and expanded storages can have two storage sizes defined:

- ▶ The initial value is the storage size allocated to the partition when it is activated.
- ▶ The reserved value is an additional storage capacity beyond its initial storage size that a logical partition can acquire dynamically. The reserved storage sizes defined to a logical partition do not have to be available when the partition is activated. They are simply predefined storage sizes to allow a storage increase, from a logical partition point of view.

Without the reserved storage definition, a logical partition storage upgrade is disruptive, requiring:

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

The additional storage capacity to a logical partition upgrade can come from:

- ▶ Any unused available storage
- ▶ Another partition that has released some storage
- ▶ A concurrent memory upgrade

A concurrent logical partition storage upgrade uses dynamic storage reconfiguration (DSR). z/OS uses the reconfigurable storage unit (RSU) definition to add or remove storage units in a nondisruptive way. z/VM V5R4 supports the dynamic addition of memory to a running LPAR by using reserved storage, and also virtualizes this support to its guests. Removal of storage from the guests or z/VM is disruptive.

3.6.3 Logical partition storage granularity

Granularity of central storage for a logical partition depends on the largest central storage amount defined for either initial or reserved central storage, as shown in Table 3-7.

Table 3-7 Logical partition main storage granularity

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

The granularity applies across all central storage defined, both initial and reserved. For example, for a logical partition with an initial storage amount of 30 GB and a reserved storage amount of 48 GB, the central storage granularity of both initial and reserved central storage is 256 MB.

Expanded storage granularity is fixed at 256 MB.

Logical partition storage granularity information is required for logical partition image setup and for z/OS Reconfigurable Storage Units definition. Logical partitions are limited to a maximum size of 1 TB of central storage. For z/VM V5R3 and later the limitation is 256 GB.

3.6.4 LPAR dynamic storage reconfiguration

Dynamic storage reconfiguration on z10 EC servers allows an operating system running in a logical partition to add (nondisruptively) its reserved storage amount to its configuration, if any unused storage exists. This unused storage can be obtained when another logical partition releases some 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 in a logical partition assigns a storage increment to its configuration, Processor Resource/Systems Manager (PR/SM) determines whether any free storage increments are available and 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 in a logical partition releases a storage increment.

LPAR dynamic storage reconfiguration is described in *System z10 Enterprise Class Processor Resource/Systems Manager Planning Guide*, SB10-7153.

3.7 Intelligent Resource Director

Intelligent Resource Director (IRD) is only available on System z running z/OS. IRD is a function that optimizes processor CPU and channel resource utilization across logical partitions within a single System z server.

IRD is a feature that extends the concept of goal-oriented resource management by allowing grouping system images that are resident on the same System z running in LPAR mode, and

in the same Parallel Sysplex, into an *LPAR cluster*. This gives Workload Manager the ability to manage resources, both processor and I/O, not just in one single image, but across the entire cluster of system images.

Figure 3-12 shows an LPAR cluster. It contains three z/OS images, and one Linux image managed by the cluster. Note that included as part of the entire Parallel Sysplex is another z/OS image, and a coupling facility image. In this example, the scope that IRD has control over is the defined LPAR cluster.

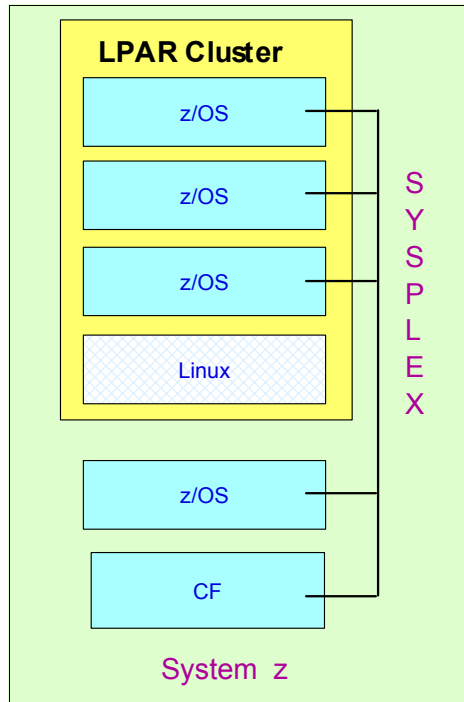


Figure 3-12 IRD LPAR cluster example

IRD addresses three separate but mutually supportive functions:

► LPAR CPU management

WLM dynamically adjusts the number of logical processors within a logical partition and the processor weight based on the WLM policy. The ability to move the CPU weights across an LPAR cluster provides processing power to where it is most needed, based on WLM goal mode policy.

HiperDispatch was introduced in 3.1, “Design highlights” on page 58, in 3.3, “Processing unit” on page 66, and in 3.6, “Logical partitioning” on page 90.

HiperDispatch manages the number of logical CPs in use. It adjusts the number of logical processors within a logical partition in order to achieve the optimal balance between CP resources and the requirements of the workload in the logical partition in cooperation with the weight management part of LPAR CPU management.

HiperDispatch also adjusts the number of logical processors. The goal is to map the logical processor to as few physical processors as possible. Doing this efficiently uses the CP resources by attempting to stay within the local cache structure, making efficient use of the advantages of the high-frequency microprocessors and improving throughput and response times.

- ▶ Dynamic channel path management (DCM)

DCM moves ESCON and FICON channel bandwidth between disk control units to address current processing needs. The z10 EC supports DCM within a channel subsystem.
- ▶ Channel subsystem priority queuing

This function on the System z allows the priority queuing of I/O requests in the channel subsystem and the specification of relative priority among logical partitions. WLM in goal mode sets the priority for a logical partition and coordinates this activity among clustered logical partitions.

For information about implementing LPAR CPU management under IRD, see *z/OS Intelligent Resource Director*, SG24-5952.

3.8 Clustering technology

Parallel Sysplex continues to be the clustering technology used with IBM System z10 Enterprise Class. Figure 3-13 illustrates the components of a Parallel Sysplex as implemented within the System z architecture. The figure is intended only as an example. It shows one of many possible Parallel Sysplex configurations. Many other possibilities exist.

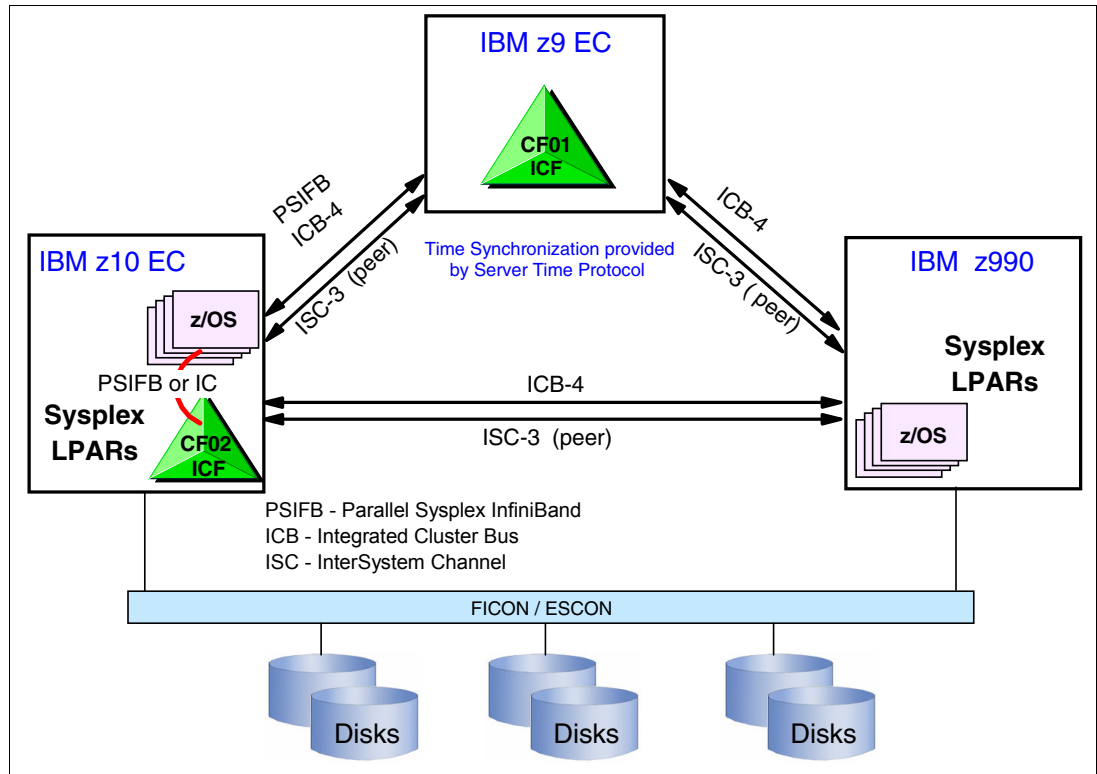


Figure 3-13 Sysplex hardware overview

Figure 3-13 shows a z10 EC containing multiple z/OS sysplex partitions and an internal coupling facility (CF02), a z9 EC containing a stand-alone ICF (CF01), and a z990 containing multiple z/OS sysplex partitions. STP over coupling links provides time synchronization to all servers. CF link technology (PSIFB, ICB-4, ISC-3) selection depends on sever configuration. Link technologies are described in 4.7.1, “Coupling links” on page 148.

Parallel Sysplex technology is an enabling technology, allowing highly reliable, redundant, and robust System z technology to achieve near-continuous availability. A Parallel Sysplex comprises one or more (z/OS) operating system images coupled through one or more coupling facilities. The images can be combined together to form clusters. A properly configured Parallel Sysplex cluster maximizes availability, as follows:

- ▶ 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 details, see *Parallel Sysplex Application Considerations*, SG24-6523.
- ▶ High capacity: Scales can be from 2 to 32 images.
- ▶ Dynamic workload balancing: Viewed as a single logical resource, work can be directed to any similar operating system image in a Parallel Sysplex cluster having available capacity.
- ▶ Systems management: Architecture provides the infrastructure to satisfy customer 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 exploit coupling facility shared storage. This exploitation 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 appears as a single entity to the operator, the user, the database administrator, and so on. A single system image ensures reduced complexity from both operational and definition perspectives.

Through state-of-the-art cluster technology, the power of multiple images can be harnessed to work in concert on common workloads. The System z Parallel Sysplex cluster takes the commercial strengths of the platform to improved levels of system management, competitive price for performance, scalable growth, and continuous availability.

Coupling Facility Control Code

Coupling Facility Control Code (CFCC) Level 16 is made available on the z10 EC.

Up to this level of CFCC, System Managed CF Structure Duplexing required two protocol exchanges to occur synchronously to CF processing of the duplex structure request. With CFCC Level 16, one of these signals can now be asynchronous. This results in faster service times especially if both coupling facilities are further apart.

CFCC Level 16 also has better list notification processing. Today, when a list changes its state from empty to non-empty, all its connectors are notified. The first connector notified reads the new message but subsequent readers find nothing. CFCC Level 16 approaches this differently to improve processor utilization. It only notifies one connector in a round-robin fashion, and if the shared queue (such as in IMS Shared Queue and WebSphere MQ Shared Queue) is read within a fixed period of time, the other connectors do not need to be notified. If the list is not read again within the time limit the other connectors are informed.

The CF Control Code, the *CF Operating System*, is implemented using the *active wait* technique. This technique means that the CF Control Code is always running (processing or searching for service) and never enters a wait state. This also means that the CF Control Code uses all the processor capacity (cycles) available for the coupling facility logical partition. If the LPAR running the CF Control Code has only dedicated processors (CPs or ICFs), then using all processor capacity (cycles) is not a problem. However, this can be an issue if the LPAR that is running the CF Control Code also has shared processors. Therefore, the recommendation is to enable dynamic dispatching on the CF LPAR.

Dynamic CF dispatching

Dynamic CF dispatching provides the following function on a coupling facility:

1. If there is no work to do, CF enters a wait state (by time).
2. After an elapsed time, CF wakes up to see whether there is any new work to do (requests in the CF Receiver buffer).
3. If there is no work, CF sleeps again for a longer period of time.
4. If there is new work, CF enters into the normal active wait until there is no more work, starting the process all over again.

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, using only ICF processors to run coupling facility images is better.

Figure 3-14 shows the dynamic CF dispatching.

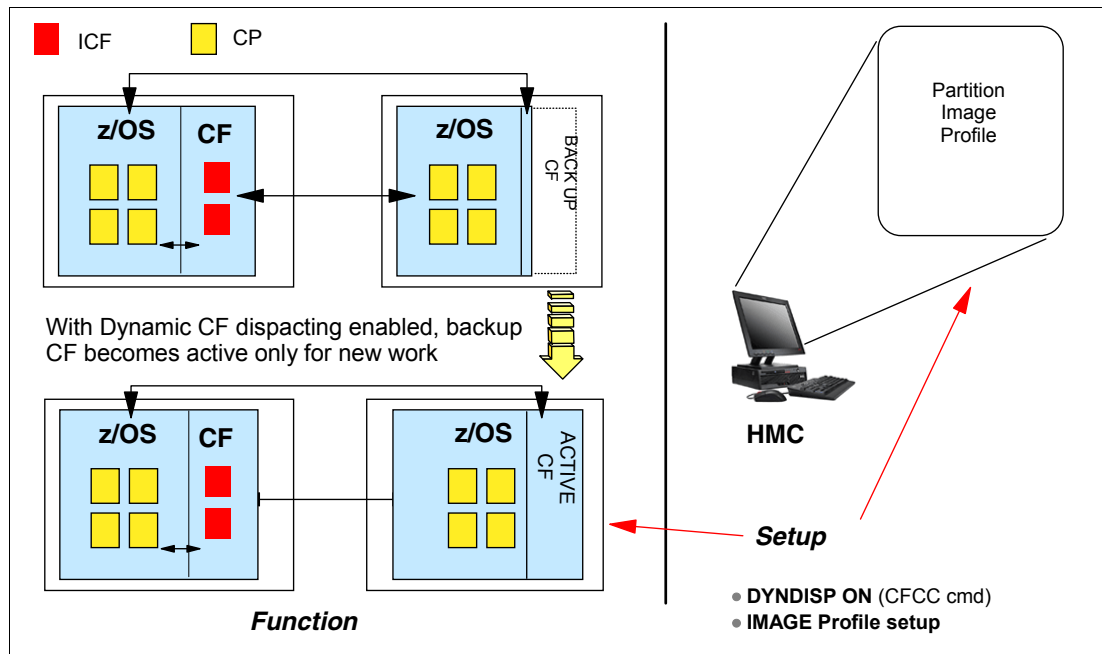


Figure 3-14 Dynamic CF dispatching (shared CPs or shared ICF PUs)

For additional details regarding CF configurations, see *Coupling Facility Configuration Options*, GF22-5042, also available from the Parallel Sysplex Web site:

<http://www.ibm.com/systems/z/advantages/ps0/index.html>



I/O system structure

This chapter describes the I/O system structure and the connectivity options available on System z10 Enterprise Class (z10 EC).

This chapter discusses the following topics:

- ▶ 4.1, “Introduction” on page 106
- ▶ 4.2, “I/O system overview” on page 107
- ▶ 4.3, “I/O cages” on page 109
- ▶ 4.4, “Fanouts” on page 111
- ▶ 4.5, “I/O feature cards” on page 120
- ▶ 4.6, “Connectivity” on page 123
- ▶ 4.7, “Parallel Sysplex connectivity” on page 146

4.1 Introduction

The z10 EC uses InfiniBand as a new interconnect protocol for various connectivity types. Before describing the InfiniBand implementation on the z10 EC, we provide a short general introduction to InfiniBand.

Note: Not all properties and functions offered by InfiniBand are implemented on the z10 EC. Only a subset is used to fulfill the interconnect requirements that have, up to now, been defined for z10 EC.

4.1.1 InfiniBand advantages

InfiniBand addresses the challenges that IT infrastructures face as more demand is placed on the interconnect with ever-increasing requirements for computing and storage resources. InfiniBand has the following advantages:

- ▶ Superior performance

InfiniBand provides superior bandwidth and latency performance. It supports 20 Gbps node-to-node and 60 Gbps switch-to-switch connections. Additionally, InfiniBand has a defined road map to 120 Gbps—the fastest support specification of any industry-standard interconnect. See also:

http://www.infinibandta.org/itinfo/IB_roadmap

- ▶ Reduced complexity

InfiniBand allows for the consolidation of multiple I/Os on a single cable or backplane interconnect, which is critical for blade servers, data center computers, storage clusters, and embedded systems. InfiniBand also consolidates the transmission of clustering, communications, storage and management data types over a single connection. The consolidation of I/O onto a unified InfiniBand fabric significantly lowers the overall power and infrastructure required for server and storage clusters. Other interconnect technologies are less suited for unified fabrics because their fundamental architectures are not designed to support multiple traffic types.

- ▶ Highest interconnect efficiency

InfiniBand is developed to provide efficient scalability of multiple systems. InfiniBand provides communication processing functions in hardware—relieving the CPU of this task—and enables the full resource utilization of each node added to the cluster. In addition, InfiniBand incorporates Remote Direct Memory Access (RDMA), which is an optimized data transfer protocol that further enables the server processor to focus on application processing. RDMA contributes to optimal application processing performance in server and storage clustered environments.

- ▶ Reliable and stable connections

InfiniBand provides reliable end-to-end data connections and defines this capability to be implemented in hardware. In addition, InfiniBand facilitates the deployment of virtualization solutions, which allow multiple applications to run on the same interconnect with dedicated application partitions. As a result, multiple applications run concurrently over stable connections, thereby minimizing downtime. InfiniBand fabrics are typically constructed with multiple levels of redundancy so that if a link goes down, the fault is limited to the link and an additional link can automatically take over to ensure that connectivity continues throughout the fabric. Creating multiple paths through the fabric results in intra-fabric redundancy and further contributes to the overall fabric reliability.

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

Similar to Fibre Channel, PCI Express, Serial ATA, and many other contemporary interconnects, InfiniBand is a point-to-point, bidirectional serial link intended for the connection of processors with high-speed peripherals, such as disks. InfiniBand supports several signalling rates and, as with PCI Express, links can be bonded together for additional bandwidth.

The serial connection's signalling rate is 2.5 Gbps on one lane in each direction (SDR)¹, per physical connection. InfiniBand also supports double (DDR) and quad speeds (QDR), for 5 Gbps or 10 Gbps, respectively.

4.1.2 Data, signalling, and link rates

Links use 8b/10b encoding (every ten bits sent carries eight bits of data), so that the useful data transmission rate is four-fifths of the signalling rate (signalling rate equals raw bit rate). Thus, single, double, and quad rates carry 2, 4, or 8 Gbps of useful data, respectively.

Links can be aggregated in units of 4 or 12, indicated as 4x or 12x. A quad-rate 12x (12x QDR) link therefore carries 120 Gbps raw or 96 Gbps of payload (useful) data. Larger systems with 12x links are typically used for cluster and supercomputer interconnects, as implemented on the z10 EC, and for inter-switch connections.

Table 4-1 lists the effective theoretical InfiniBand data throughput in different configurations.

Table 4-1 Effective data rates of aggregated links

Number of links	Single (SDR)	Double (DDR)	Quad (QDR)
1X	2 Gbps	4 Gbps	8 Gbps
4X	8 Gbps	16 Gbps	32 Gbps
12X	24 Gbps	48 Gbps	96 Gbps

Throughout this chapter the following terminology is used:

Data rate The data transfer rate is expressed in bytes; one byte equals eight bits.

Signalling rate The raw bit rate is expressed in bits.

Link rate The rate is equal to the signalling rate expressed in bits.

For details and the standard for InfiniBand, see the InfiniBand Web site:

<http://www.infinibandta.org>

4.2 I/O system overview

This section lists characteristics and a summary of features that are supported.

¹ SDR is Single Data Rate, DDR is Dual Data Rate, QDR is Quad Data Rate

4.2.1 Characteristics

The z10 EC I/O system design provides great flexibility, high availability, and excellent performance characteristics, as follows:

- ▶ High bandwidth

The z10 EC uses InfiniBand as the internal interconnect protocol to drive ESCON and FICON channels, OSA ports, and ISC-3 coupling links. As a connection protocol, InfiniBand supports InfiniBand coupling (PSIFB²) with a link rate of up to 6 GBps.

- ▶ Wide connectivity

The z10 EC can be connected to an extensive range of interfaces such as Gigabit Ethernet (GbE), FICON/Fibre Channel, ESCON, and coupling links.

- ▶ Concurrent I/O upgrade

You may concurrently add I/O cards to the server if an unused I/O slot position is available. Additional I/O cages can be installed in advance to provide greater capacity for concurrent upgrades.

- ▶ Dynamic I/O configuration

Dynamic I/O configuration supports the dynamic addition, removal, or modification of channel path, control units, and I/O devices without a planned outage.

- ▶ ESCON port sparing

One unused port on the 16-port I/O card is dedicated for sparing in the event of a port failure on that card. Other unused ports are available for growth of ESCON channels without requiring new hardware. Unused ports can be enabled through Licensed Internal Code (LIC).

- ▶ Pluggable optics

The FICON Express8 and FICON Express4 features have Small Form Factor Pluggable (SFP) optics to permit each channel to be individually serviced in the event of a fiber optic module failure. The traffic on the other channels on the same feature can continue to flow if a channel requires servicing.

- ▶ Concurrent I/O card maintenance

Each I/O card plugged in an I/O cage supports concurrent card replacement in case of a repair action.

4.2.2 Summary of supported I/O features

The following I/O features are supported:

- ▶ Up to 1024 ESCON channels (up to 960 on the model E12)
- ▶ Up to 120 FICON Express channels (when carried forward on upgrade only)
- ▶ Up to 336 FICON Express2 channels (when carried forward on upgrade only)
- ▶ Up to 336 FICON Express4 channels (when carried forward on upgrade only)
- ▶ Up to 336 FICON Express8 channels
- ▶ Up to 24 OSA-Express3 features
- ▶ Up to 24 OSA-Express2 features
- ▶ Up to 48 ISC-3 coupling links
- ▶ Up to 16 ICB-4 coupling links
- ▶ Up to 32 InfiniBand coupling links
- ▶ Two external time reference (ETR) connections
- ▶ Two pulse-per-second (PPS) connections

² Parallel Sysplex InfiniBand

Note: The maximum number of coupling links combined (IC, ICB-4, ISC-3, and PSIFB coupling links) cannot exceed 64 for each server.

4.3 I/O cages

The z10 EC has two frames. The A frame holds the CEC cage on top, and one I/O cage at the bottom. The Z frame holds two optional I/O cages that might be necessary for accommodating the I/O configuration requirements.

Each cage supports up to seven I/O domains (named A to G) for a total of 28 I/O card slots. Each I/O domain supports four I/O card slots, as shown in Figure 4-1.

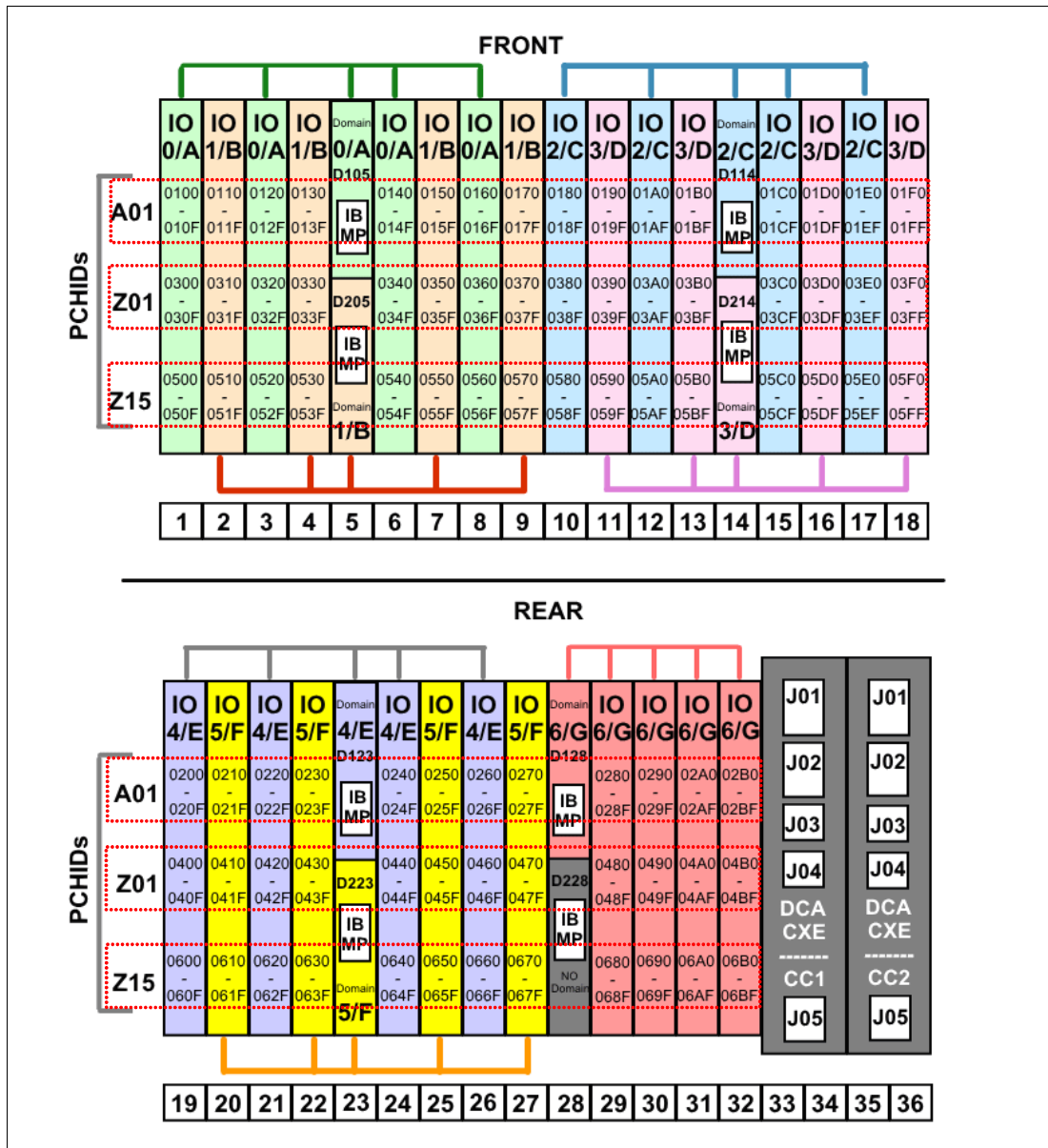


Figure 4-1 I/O cage

Each I/O domain uses an IFB-MP card (IB MP in Figure 4-1 on page 109) in the I/O cage and a copper cable connected to an Host Channel Adapter (HCA) fanout in the CPC cage. A maximum of seven I/O domains are available in each cage. An eighth IFB-MP card is installed to provide an alternate path to I/O cards in slots 29, 30, 31, and 32 in case of a repair action.

Domain number 6 (G) is not used until all other domains are full in all other I/O cages. If more than 24 or 48 I/O cards are required, a new I/O cage must be installed. Only when more than 72 I/O cards are required will domain G be populated.

Figure 4-2 illustrates the I/O structure in a z10 EC. An InfiniBand (IFB) cable connects the HCA2-C fanout to an IFB-MP card in the I/O cage. The passive connection between two IFB-MP cards allows for redundant I/O interconnection. The IFB cable between an HCA2-C fanout in a book, and each IFB-MP card in the I/O cage supports a 6 GBps bandwidth.

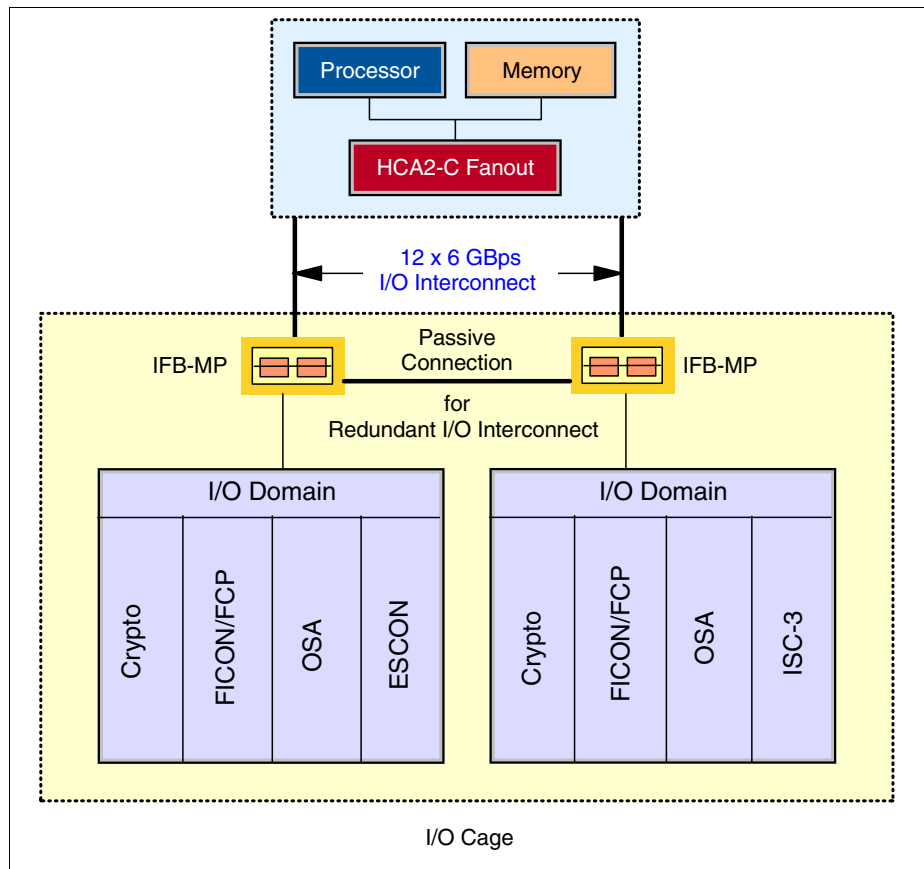


Figure 4-2 z10 EC and z9 EC I/O structure

Note: Installing an additional I/O cage is disruptive. The plan-ahead process allows you to avoid an outage by including additional I/O cages at initial order time.

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

A fully populated system with three I/O cages installed has a total of 84 available I/O card slots.

Table 4-2 lists the I/O domains and their related I/O slots (see also Figure 4-1 on page 109).

Table 4-2 I/O domains

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

The configuration process selects which I/O slots are used for I/O cards and provides the required number of I/O cages, HCA2-C fanout cards, IFB-MP cards, and IFB cables, either for a new build server or a server upgrade.

If you order the Power Sequence Control (PSC) feature, the PSC24V card is always plugged in to slot 29 of domain G. Installing a PSC24V card is always disruptive.

4.4 Fanouts

InfiniBand offers a point-to-point bidirectional serial, high-bandwidth, low-latency link that is used for the connection of processors. Its use is introduced for the connection to other systems in a Parallel Sysplex, and for the internal connection to I/O cages in which the cards for the connection to peripheral devices and networks reside. The InfiniBand fanouts are located in the front of each book.

Each book has eight fanout slots. They are named D1 to DA, top to bottom; slots D3 and D4 are not used for fanouts. Each fanout has two ports to connect an ICB or IFB cable, depending on the type of fanout. There are three types of Host Channel Adapters (HCAs). One uses a copper cable (HCA2-C) to connect to an I/O cage; the other two use optical connections (HCA2-O, HCA2-O LR). Each slot holds one of the following four fanouts:

- ▶ Host Channel Adapter (HCA2-C): This copper fanout provides connectivity to the IFB-MP card in the I/O cage.
- ▶ Host Channel Adapter (HCA2-O): This optical fanout provides coupling link connectivity up to 150 meters (492 feet) distance to other z10 or z9 servers.
- ▶ Host Channel Adapter (HCA2-O LR): This optical long range fanout provides coupling link connectivity up to 10 km (6.2 miles) unrepeated distance to other z10 servers.
- ▶ Memory bus adapter (MBA): This fanout is used for copper cable ICB-4 links only.

Figure 4-3 illustrates the IFB connection from the CEC cage to an I/O cage, the Integrated Cluster Bus (ICB-4), and coupling over InfiniBand (PSIFB).

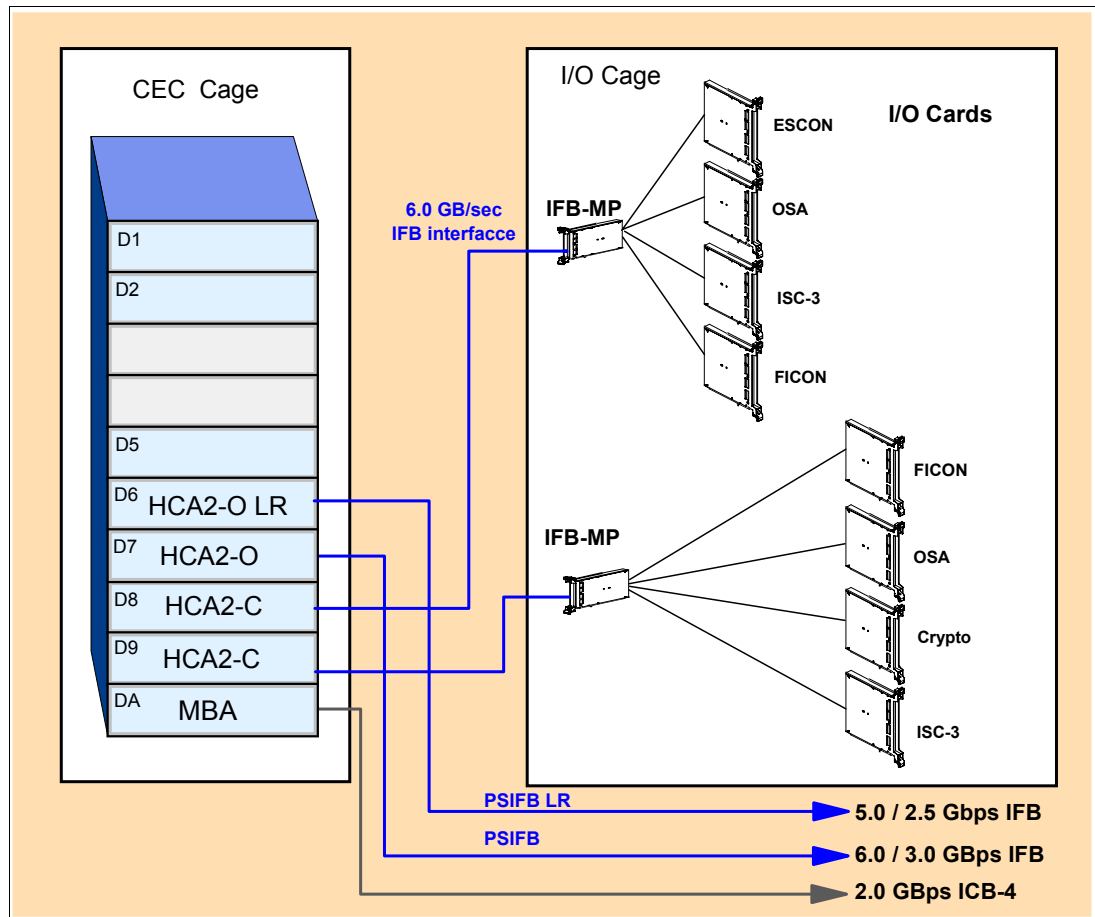


Figure 4-3 PSIFB, MBA, and IFB I/O cage interface connections

4.4.1 HCA2-C fanout

The HCA2-C fanout is used to connect to an I/O cage using a copper cable. The two ports on the fanout are dedicated to I/O. The bandwidth of each port on the HCA2-C fanout supports a link rate of up to 6 GBps.

A copper cable of 1.5 to 3.5 meters long is used for connection to the IFB-MP card in the I/O cage. If the maximum of three fully populated I/O cages is installed, 12 HCA2-C fanouts (24 ports) are required.

Note: The HCA2-C fanout is used exclusively for I/O and cannot be shared for any other purpose.

4.4.2 HCA2-O fanout

The HCA2-O fanout provides an optical interface used for coupling links. The two ports on the fanout are dedicated to coupling links to connect to System z10 or System z9 servers, or to connect to a coupling port in the same server by using a fiber cable. Each fanout has an optical transmitter and receiver module and allows dual simplex operation. Up to 16 HCA2-O

fanouts are supported and provide up to 32 ports for coupling links. The combined maximum of all PSIFB and ICB-4 links is 32.

The HCA2-O fanout supports InfiniBand double data rate (12x IB-DDR) and InfiniBand single data rate (12x IB-SDR) optical links that offer longer distance, configuration flexibility, and high bandwidth for enhanced performance of coupling links. There are 12 lanes (two fibers per lane) in the cable, which is 24 fibers used in parallel for data transfer.

The fiber cables are industry standard OM3 (2000 MHz-km) 50 µm multimode optical cables with Multi-Fiber Push-On (MPO) connectors. The maximum cable length is 150 meters (492 feet).

Each fiber supports a link rate of 6 GBps (12x IB-DDR) if connected to a z10 EC server or 3 GBps (12x IB-SDR) when connected to a System z9 server. The link rate is auto-negotiated to the highest common rate.

Note: Ports on the HCA2-O fanout are exclusively used for coupling links and cannot be used or shared for other purpose.

A fanout has two ports for optical link connections and supports up to 16 CHPIDs across both ports. These CHPIDs are defined in IOCDS as coupling links.

Note: The recommendation is to define only four CHPIDs for each port.

Each HCA2-O fanout used for coupling links has an assigned adapter ID (AID) number that must be used for definitions in IOCDS to create a relationship between the physical fanout location and the CHPID number. For details about AID numbering, see “Adapter ID number assignment” on page 118.

For detailed information about how the AID is used and referenced in HCD, see *Getting Started with InfiniBand on System z10 and System z9*, SG24-7539.

4.4.3 HCA2-O LR fanout

The HCA2-O LR fanout provides an optical interface used for coupling links. The two ports on the fanout are dedicated to coupling links to connect to other z10 servers. Up to 16 HCA2-O LR fanouts are supported and provide 32 ports for coupling link. The combined maximum of all PSIFB and ICB-4 links is 32.

The HCA-O LR fanout supports InfiniBand double data rate (1x IB-DDR) and InfiniBand single data rate (1x IB-SDR) optical links that offer longer distance of coupling links. The cable has one lane containing two fibers; one fiber is used for transmitting and one fiber used for receiving data.

Each fiber supports a link rate of 5 Gbps (1x IB-DDR) if connected to a System z10 sever or to a repeater (DWDM³) supporting IB-DDR, and a data link rate of 2.5 Gbps (1x IB-SDR) when connected to a repeater (DWDM) that supports IB-SDR. The link rate is auto-negotiated to the highest common rate.

Note: Ports on the HCA2-O LR fanout are used exclusively for coupling links and cannot be used or shared for other purpose.

³ dense wavelength division multiplexing

The fiber cables are 9 μm single mode (SM) optical cables terminated with an LC Duplex connector. The maximum unrepeated distance is 10 km (6.2 miles) and up to 100 km (62 miles) with repeaters (DWDM).

A fanout has two ports for optical link connections and supports up to 16 CHPIDs across both ports. These CHPIDs are defined in IOCDS as coupling links and require a fiber cable to connect to other z10 servers or the same server.

Note: It is recommended that you define only four CHPIDs per port.

Each HCA2-O LR fanout used for coupling links has an assigned adapter ID (AID) number that must be used for definitions in IOCDS to create a relationship between the physical fanout location and the CHPID number. See “Adapter ID number assignment” on page 118 for details about AID numbering.

4.4.4 MBA fanout

The MBA fanout provides coupling links (ICB-4) to either System z10 servers or System z9, z990, and z890 servers. This construction allows the use of the z10 EC and earlier servers in the same Parallel Sysplex.

MBA fanouts are only for ICB-4 coupling links and cannot be used for any other purpose. Up to eight MBA fanouts, providing up to 16 links, are supported. The combined maximum of all PSIFB and ICB-4 links is 32.

Important: When upgrading to a z10 EC from a System z9 or z990 with ICB-4 coupling links, new ICB copper cables are required because connector types used in the z10 servers are different from the ones used for z9 and z990.

Note: The ICB-4 feature cannot be ordered on a model E64 server.

The physical channel ID (PCHID) numbers for ICB-4 coupling links are assigned by the physical location of the MBA fanout in a book. Table 4-3 lists the PCHID numbers assigned to each port on the MBA fanout in each book.

Table 4-3 MBA fanout PCHID assignment

Fanout location	Fourth book	First book	Third book	Second book
D1	000/001	010/011	020/021	030/031
D2	002/003	012/013	022/023	032/033
D3	-	-	-	-
D4	-	-	-	-
D5	004/005	014/015	024/025	034/035
D6	006/007	016/017B	026/027	036/037
D7	008/009	018/019	028/029	038/039
D8	00A/00B	01A/01B	02A/02B	03A/03B

Fanout location	Fourth book	First book	Third book	Second book
D9	00C/00D	01C/01D	02C/02D	03C/03D
DA	00E/00F	01E/01F	02E/02F	03E/03F

4.4.5 Fanout considerations

Because fanout slots in each book can be used to plug different fanouts, where each fanout is designed for a special purpose, some restrictions might apply to the number of available channels located in the I/O cage.

A fully populated server has three I/O cages. Each cage requires eight connections to support all 28 slots for I/O cards. This is a total of 24 connections required for all three cages, which is equivalent to 12 HCA2-C fanouts (24 ports) dedicated to I/O links. For I/O cage details, see Figure 4-1 on page 109.

If fewer than 12 HCA-C fanouts are available, the number of supported I/O cards and the number of CHPIDs available in I/O cages can decrease. The number of HCA2-C fanouts for cage connections depends on the number of HCA2-O LR, HCA2-O and MBA fanouts used for coupling links, and vice versa. Also, the fanouts for I/O are always plugged in pairs.

Depending on the model, the number of fanouts varies. The following sections show the relationship between number of fanouts used for coupling links and the remaining available I/O domains and CHPIDs for each model.

The plugging rules for fanouts for each model are illustrated in Figure 4-4.

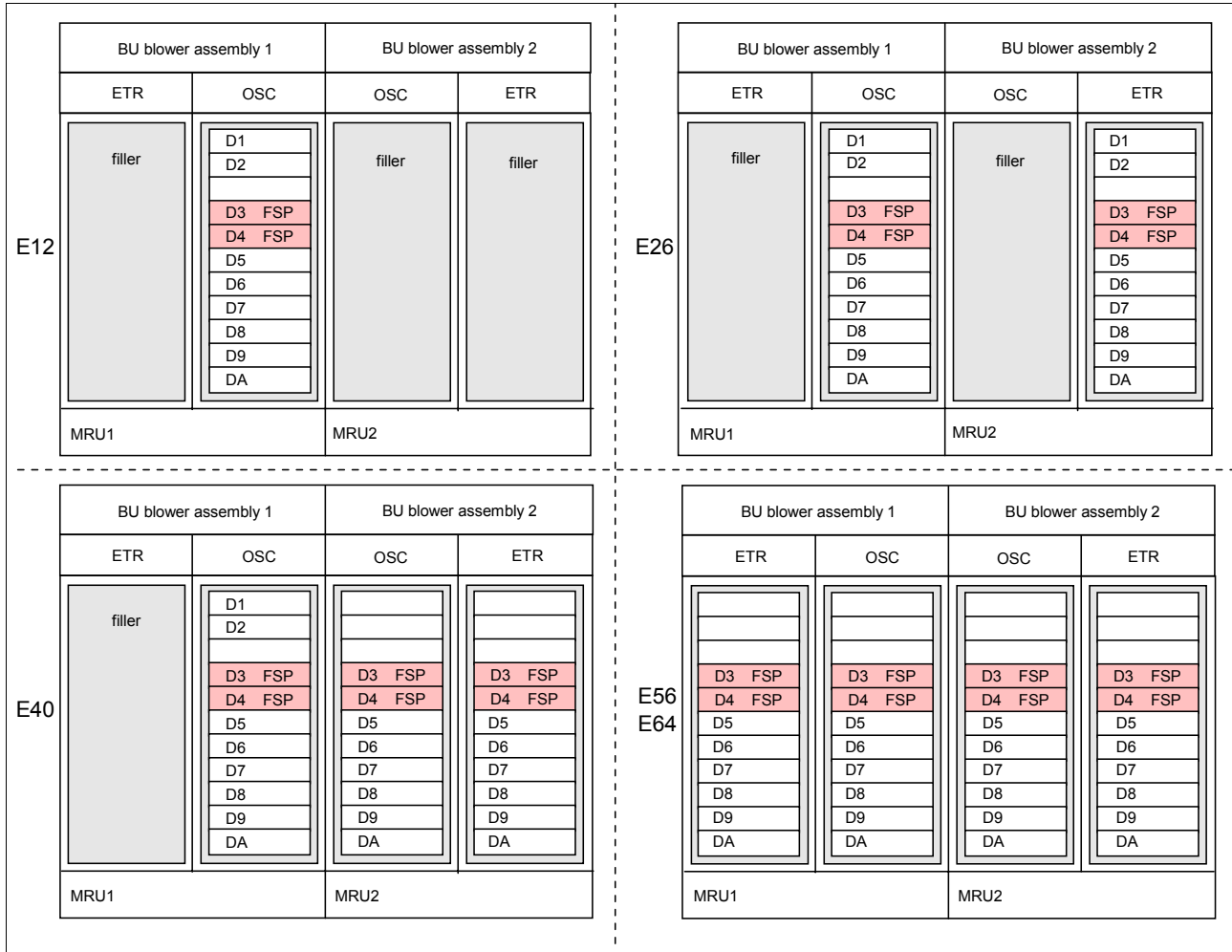


Figure 4-4 Fanout plugging rules

Fanouts in a model E12 (one book)

Model E12 has one book installed, supporting eight fanouts. The maximum number of (ESCON) channels supported is 960. This number is decreased by each fanout used for a coupling link. For example, if four fanouts of any type designated for coupling links are installed, the maximum number of I/O domains is eight, supporting up to 480 (ESCON) CHPIDs in the I/O cage, as shown in Table 4-4.

A maximum of eight HCA2-O LR, HCA2-O, and MBA fanouts used for coupling links is supported.

Table 4-4 Available CHPIDs in I/O cage (one book)

Maximum of eight fanouts									
Number of HCA2-O LR, HCA2-O, and MBA fanouts	0	1	2	3	4	5	6	7	8
Available I/O domains	16	12	12	8	8	4	4	0	0
Maximum number of CHPIDS in I/O cage	960	720	720	480	480	240	240	0	0

Fanouts in a model E26 (two books)

Model E26 has two books installed, supporting 16 fanouts. The maximum number of (ESCON) channels supported is 1024 using 69 features (across 18 domains). This number can decrease if fewer than 10 fanouts for I/O connectivity remain. See Table 4-5 for details.

A maximum of 16 HCA2-O LR, HCA2-O, and MBA fanouts, used for coupling links, is supported.

Table 4-5 Available CHPIDs in I/O cage (two books)

Maximum of 16 fanouts													
Number of HCA2-O LR, HCA2-O, and MBA fanouts	0 - 4	5	6	7	8	9	10	11	12	13	14	15	16
Available I/O domains	21	19	19	16	16	12	12	8	8	4	4	0	0
Maximum number of CHPIDs in I/O cage	1024	1024	1024	960	960	720	720	480	480	240	240	0	0

Fanouts in model E40 (three books)

Model E40 has three books, supporting 20 fanouts. The maximum number of (ESCON) channels supported is 1024 using 69 features (across 18 domains). This number can decrease if fewer than 10 fanouts for I/O connectivity remain. See Table 4-6 for details. A maximum of 16 HCA2-O LR, HCA2-O, and MBA fanouts, used for coupling links, is supported.

Table 4-6 Available CHPIDs in I/O cage (three books)

Maximum of 20 fanouts										
Number of HCA2-O LR, HCA2-O and MBA fanouts	0 - 8	9	10	11	12	13	14	15	16	
Available I/O domains	21	19	19	16	16	12	12	8	8	
Maximum number of CHPIDs in I/O cage	1024	1024	1024	960	960	720	720	480	480	

Fanouts in models E56 and E64 (four books)

Models E56 and E64 have four books, supporting 24 fanouts. The maximum number of channels supported is 1024 using 69 features (across 18 domains). This number can decrease if fewer than 10 fanouts used for I/O connectivity remain. See Table 4-7 for details.

A maximum of 16 HCA2-O LR, HCA2-O, and MBA fanouts, used for coupling links, is supported.

Table 4-7 Available CHPIDs in I/O cage (four book)

Maximum of 24 fanouts					
Number of HCA2-O LR, HCA2-O and MBA ^a fanouts	0 - 12	13	14	15	16
Available I/O domains	21	19	19	16	16
Maximum number of CHPIDs in I/O cage	1024	1024	1024	960	960

a. MBA fanouts are not supported on model E64 server

Adapter ID number assignment

Unlike channels installed in an I/O cage, which are identified by a PCHID number related to their physical location, PSIFB fanouts and ports are identified by an adapter ID (AID), initially dependent on their physical locations. 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 4-8 illustrates the AID assignment for each fanout slot relative to the book location on a new build system.

Table 4-8 AID number assignment

Book	Slot	Fanout slot	AIDs
First	6	D1, D2, D5-DA	08, 09, 0A-0F
Second	15	D1, D2, D5-DA	18, 19, 1A-1F
Third	10	D1, D2, D5-DA	10, 11, 12-17
Fourth	1	D1, D2, D5-DA	00, 01, 02-07

The fanout slots are numbered D1 to DA top to bottom, as shown in Table 4-9. All fanout locations and their AIDs for all four books are shown in the table for reference only. Fanouts in locations D1 and D2 are not available on all models. Slots D3 and D4 will never have a fanout installed (dedicated for FSPs).

Note: Slots D1 and D2 are not used in a 4-book server, and only partially in a 3-book server.

Table 4-9 Fanout AID numbers

Fanout location	Fourth book	First book	Third book	Second book
D1	00	08	10	18
D2	01	09	11	19
D3	-	-	-	-
D4	-	-	-	-
D5	02	0A	12	1A
D6	03	0B	13	1B
D7	04	0C	14	1C
D8	05	0D	15	1D
D9	06	0E	16	1E
DA	07	0F	17	1F

Important Note: The AID numbers in Table 4-9 are valid only for a new build server or for new books added. If a fanout is moved, the AID follows the fanout to its new physical location.

The AID assigned to a fanout is found in the PCHID REPORT provided for each new server or for MES upgrade on existing servers.

Example 4-1 shows part of a report, named PCHID REPORT, for a model E26. In this example, one fanout is installed in the first book (location 06) and one fanout is installed in the second book (location 15), both in location D6. The assigned AID for the fanout in the first book is 0B; the AID assigned to the fanout in the second book is 1B.

Example 4-1 AID assignment in PCHID report

CHPIDSTART						
19756694		PCHID REPORT			Jul 16,2007	
Machine: 2097-E26 SNxxxxxxx						

Source	Cage	Slot	F/C	PCHID/Ports or AID	Comment	
06/D6	A25B	D606	0163	AID=0B		
15/D6	A25B	D615	0163	AID=1B		

STI rebalance (FC2400)

In a newly built z10 EC with multiple books installed, the fanouts are balanced across the available books following the plugging rules for new build servers. If books are added by an MES upgrade, the fanout cards used for I/O and PSIFB are rebalanced across all books.

MBA fanouts used for ICB-4 are not rebalanced. STI rebalance (FC 2400) can be ordered to rebalance ICB-4 MBA fanouts across all books. Installation of this feature is disruptive.

For fanout plugging rules on E12, E26, E40, E56, and E64, see Figure 4-4 on page 116.

4.4.6 Fanout summary

Fanout features supported by the z10 EC server are shown in Table 4-10. The table provides the feature type and code, total maximum (Max.) number of features and ports, and information about the link supported by the fanout feature.

Table 4-10 Fanout summary

Fanout feature	Feature code	Max. features	Max. ports	Use	Cable type	Connector type	Max. distance	Link data rate
HCA2-C	0162	12	24	Connect to I/O cage	Copper	n/a	3.5 m	6 GBps
HCA2-O	0163	16 ^a	32	Coupling link	50 µm MM OM3 (2000 MHz-km)	MPO	150 m	6 GBps ^b
HCA2-O LR	0168	16 ^a	32	Coupling link	9 µm SM	LC Duplex	10 km ^c	5.0 Gbps 2.5 Gbps ^d
MBA	0164	8 ^a	16	Coupling link	Copper	n/a	10 m	2 GBps

a. A maximum of 16 combined of these features (FC 0163, 0168, and 0164) is supported

b. 3 GBps link data rate if connected to a System z9 server

c. Up to 100 km with repeaters (DWDM)

d. Autonegotiated, depending on DWDM equipment

4.5 I/O feature cards

I/O cards have ports to connect the z10 EC to external devices, networks, or other servers. I/O cards are plugged into the I/O cage 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.

4.5.1 I/O feature card types

The I/O features listed in Table 4-11 on page 120 can be ordered for newly built servers.

Table 4-11 I/O feature codes

Card type	Feature code
ESCON (16-port)	2323
FICON Express8 LX (10 km)	3325
FICON Express8 SX	3326
OSA-Express2 1000BASE-T	3366
OSA-Express3 10 GbE LR	3370
OSA-Express3 10 GbE SR	3371
OSA-Express3 GbE LX	3362
OSA-Express3 GbE SX	3363
OSA-Express3 1000BASE-T	3367
ISC-3	0217 (ISC-M) 0218 (ISC-D)
ISC-3 up to 20 km	RPQ 8P2197 (ISC-D)
Crypto Express3	0864

Table 4-12 lists I/O features that are available only if carried over during an upgrade.

Table 4-12 I/O feature codes

Card type	Feature code
FICON Express LX	2319
FICON Express SX	2320
FICON Express2 LX	3319
FICON Express2 SX	3320
FICON Express4 LX (4 km)	3324
FICON Express4 LX (10 km)	3321
FICON Express4 SX	3322
OSA-Express2 10 GbE LR	3368
OSA-Express2 GbE LX	3364

Card type	Feature code
OSA-Express2 GbE SX	3365
Crypto Express2	0863

4.5.2 PCHID report

A physical channel ID (PCHID) number is assigned to each I/O card port and the Crypto Express2 and Crypto Express3 card plugged in the I/O cage. Each enabled port has a PCHID number assigned, depending on the physical I/O slot location of where the card is plugged in, and on the physical port on the card.

A PCHID report is created for each new build server and for upgrades on existing servers. The report lists all I/O features installed, the physical slot location, and the assigned PCHID.

Example 4-2 shows a portion of a sample PCHID report.

The AID numbering rules for InfiniBand coupling links are described in “Adapter ID number assignment” on page 118.

The PCHID number assignment rules for MBA fanout ports are described in 4.4.4, “MBA fanout” on page 114.

Example 4-2 PCHID report

```

CHPIDSTART
19756694                PCHID REPORT                Jul 16,2007
Machine: 2097-E26  SN1
-----
Source          Cage Slot F/C  PCHID/Ports or AID          Comment
06/D6           A25B D606 0163  AID=0B
15/D6           A25B D615 0163  AID=1B
06/DA           A25B DA06 3393  01E/J01 01F/J02
15/DA           A25B DA15 3393  03E/J01 03F/J02
15/D9/J01       A01B 01   0864 100/P00 101/P01
06/D9/J01       A01B D102 0218  110/J00 111/J01
06/D9/J01       A01B D202 0218  118/J00 119/J01
15/D9/J01       A01B 03   3365  120/J00 121/J01
06/D9/J01       A01B 04   3366  130/J00 131/J01
06/D9/J01       A01B 07   3324  150/D1 151/D2 152/D3 153/D4
06/D8/J01       A01B 12   3319  1A0/J00 1A1/J01 1A2/J02 1A3/J03

```

```

06/D8/J01      A01B  17   2323  1E0/J00 1E1/J01 1E2/J02 1E3/J03
                1E4/J04 1E5/J05 1E6/J06 1E7/J07
                1E8/J08 1E9/J09 1EA/J10 1EB/J11
                1EC/J12 1ED/J13

15/D7/J01      Z01B  06   2319  340/J00 341/J01

```

The following list explains the content of the sample PCHID REPORT:

- ▶ Feature code 0864 (Crypto Express3) is installed in cage A01B slot 1 and has PCHID 100 and 101 assigned.
- ▶ Feature code 0218 (ISC-3) is installed in cage A01B slot 2 and has PCHID 110 and 111 assigned to the two ports on the upper daughter card, and PCHID 118 and 119 to the two ports on the lower daughter card.
- ▶ Feature code 3365 (OSA-Express2 GbE SX) is installed in cage A01B slot 3 and has PCHID 120 and 121 assigned.
- ▶ Feature code 3366 (OSA-Express2 1000BASE-T-EN) is installed in cage A01B slot 4 and has PCHID 130 and 131 assigned.
- ▶ Feature code 3324 (FICON Express4 LX 4 km) is installed in cage A01B slot 7 and has PCHID 150, 151, 152, and 153 assigned.
- ▶ Feature code 3319 (FICON Express2 LX) is installed in cage A01B slot 12 and has PCHID 1A0, 1A1, 1A2, and 1A3 assigned.
- ▶ Feature code 2323 (ESCON 16-port) is installed in cage A01B slot 17 and has PHCHIDs 1E0 to 1ED for the 14 ports enabled on that adaptor card.
- ▶ Feature code 2319 (FICON Express LX) is installed in cage Z01B slot 6 and has PCHID 340 and 341 assigned.

The pre-assigned PCHID number of each I/O port relates directly to its physical location (jack location in a specific slot). For PCHID numbers and their locations, see Table 4-13.

Table 4-13 PCHID numbers and locations

I/O cage slot ^a	PCHID numbers ^b		
	First I/O cage frame A bottom	Second I/O cage frame Z bottom	Third I/O cage frame Z top
1	100-10F	300-30F	500-50F
2	110-11F	310-31F	510-51F
3	120-12F	320-32F	520-52F
4	130-13F	330-33F	530-53F
6	140-14F	340-34F	540-54F
7	150-15F	350-35F	550-55F
8	160-16F	360-36F	560-56F
9	170-17F	370-37F	570-57F
10	180-18F	380-38F	580-58F
11	190-19F	390-39F	590-59F
12	1A0-1AF	3A0-3AF	5A0-5AF

I/O cage slot ^a	PCHID numbers ^b		
	First I/O cage frame A bottom	Second I/O cage frame Z bottom	Third I/O cage frame Z top
13	1B0-1BF	3B0-3BF	5B0-5BF
15	1C0-1CF	3C0-3CF	5C0-5CF
16	1D0-1DF	3D0-3DF	5D0-5DF
17	1E0-1EF	3E0-3EF	5E0-5EF
18	1F0-1FF	3F0-3FF	5F0-5FF
19	200-20F	400-40F	600-60F
20	210-21F	410-41F	610-61F
21	220-22F	420-42F	620-62F
22	230-23F	430-43F	630-63F
24	240-24F	440-44F	640-64F
25	250-25F	450-45F	650-65F
26	260-26F	460-46F	660-66F
27	270-27F	470-47F	670-67F
29	280-28F	480-48F	680-68F
30	290-29F	490-49F	690-69F
31	2A0-2AF	4A0-4AF	6A0-6AF
32	2B0-2BF	4B0-4BF	6B0-6BF

a. Slots 5, 14, 23, and 28 are reserved for IFB-MP cards.

b. The PCHID number range from 000 to 03F is reserved for ICB-4 links.

4.6 Connectivity

I/O channels are part of the channel subsystem (CSS). They provide connectivity for data exchange between servers, or between servers and external control units (CU) and devices, or networks.

Communication between servers is implemented by using intersystem channels (ISC-3), Integrated Cluster Bus (ICB-4), coupling using InfiniBand (IFB), or channel-to-channel connections (CTC).

Communication to local area networks (LANs) is provided by the OSA-Express2 and OSA-Express3 features.

Connectivity to I/O subsystems to exchange data is provided by ESCON and FICON channels.

4.6.1 I/O feature support and configuration rules

Table 4-14 lists the I/O features supported. The table shows the feature code numbers, number of ports per card, port increments, and the maximum number of feature cards and the maximum of channels for each feature type. Also, the CHPID definition used in the IOCDs are listed.

Table 4-14 Supported I/O features

I/O feature	Feature codes	Number of		Max. number of		PCHID	CHPID definition
		Ports per card	Port increments	Ports	I/O slots		
ESCON ^a	2323 ^b 2324 ^b	16 (1 spare)	4 (LICCC)	1024	69	Yes	CNC, CVC, CTC, CBY
FICON Express LX/SX ^c	2319/2320	2	2	120	60	Yes	FC, FCP, FCV
FICON Express2 LX/SX ^c	3319/3320	4	4	336	84	Yes	FC, FCP
FICON Express4 LX/SX	3324/3321/3322	4	4	336	84	Yes	FC, FCP
FICON Express8 LX/SX	3325/3326	4	4	336	84	Yes	FC, FCP
OSA- Express2 GbE LX/SX	3364/3365	2	2	48	24 ^d	Yes	OSD, OSN
OSA- Express2 10 GbE LR ^c	3368	1	1	24	24 ^d	Yes	OSD
OSA- Express2 1000BASE-T	3366	2	2	48	24 ^d	Yes	OSE, OSD, OSC, OSN
OSA- Express3 10 GbE LR/SR	3370/3371	2	2	48	24 ^d	Yes	OSD
OSA-Express3 GbE LX/SX	3362/3363	4	4	96	24 ^d	Yes	OSD, OSN
OSA-Express3 1000BASE-T	3367	4	4	96	24 ^d	Yes	OSE, OSD, OSC, OSN
ISC-3 2 Gbps (10 km) ^e	0217,0218,0219	2 / ISC-D	1	48	12	Yes	CFP
ISC-3 1 Gbps (20 km) ^e	RPQ 8P2197	2 / ISC-D	2	48	12	Yes	CFP
ICB-4 ^e	3393	2	1	16 ^f	-	Yes	CBP
InfiniBand coupling (IFB) ^e	0163	2	2	32 ^f	-	No	CIB
InfiniBand coupling (IFB LR) ^e	0168	2	2	32 ^f	-	No	CIB

a. The maximum number of ESCON ports on a model E12 is 960.

b. Feature code 2323 is the ESCON 16-port card; feature code 2324 is for the amount of ESCON ports ordered in increments of four. Each ESCON card has 15 usable ports and one spare port.

c. This feature is only available if carried over on an upgrade.

- d. The maximum number of combined OSA features is 24.
- e. The Order Process does not allow you to order more than 64 links (ICB-4, ISC, and IFB). HCD and IOCP prevent you from defining more than 64 coupling CHPIDs (CBP+CFP+CIB+ICP).
- f. The maximum number of IFB and MBA fanouts is 16 (32 links).

At least one I/O feature (FICON or ESCON) or one coupling link feature (IFB, ICB-4, or ISC-3) must be present in the minimum configuration. A maximum of 256 channels is configurable per channel subsystem and per operating system image.

Spanned and shared channels

The multiple image facility (MIF) allows sharing channels within a channel subsystem, as follows:

- ▶ Shared channels are shared by logical partitions within a channel subsystem (CSS).
- ▶ Spanned channels are shared by logical partitions within and across CSSs.

The following channel *cannot* be shared or spanned: ESCON-to-parallel channel conversion (defined as CVC and CBY).

The following channels can be shared but *cannot* be spanned:

- ▶ ESCON channels defined as CNC or CTC
- ▶ FICON LX channels defined as FCV (conversion mode)

The following channels can be shared and spanned:

- ▶ FICON channels defined as FC or FCP
- ▶ OSA-Express2 and OSA-Express3, defined as OSD, OSE, OSC, or OSN
- ▶ Coupling links defined as CFP, CBP, ICP, or CIB
- ▶ HyperSockets defined as IQD

The Crypto Express2 and Crypto Express3 features do not have a CHPID type, but logical partitions in all CSSs have access to the features. Each adapter on a Crypto Express2 feature can be defined to up to 16 active logical partitions and each adapter on a Crypto Express3 feature can be defined to up to 32 logical partitions.

I/O feature cables and connectors

Note: All fiber optic cables, cable planning, labeling, and installation are customer responsibilities for new z10 EC installations and upgrades. Fiber optic conversion kits and mode conditioning patch (MCP) cables are not orderable as features on z10 EC servers. All other cables have to be sourced separately.

IBM Facilities Cabling Services - fiber transport system offers a total cable solution service to help with cable ordering requirements, and is highly recommended. These services consider the requirements for all of the protocols and media types supported (for example, ESCON, FICON, Coupling Links, and OSA), whether the focus is the data center, the storage area network (SAN), local area network (LAN), or the end-to-end enterprise.

The Enterprise Fiber Cabling Services make use of 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 integrated in the frame of a z10 EC for *quick* connection, which is offered as a feature on z10 EC servers for connection to FICON LX and ESCON 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 4-15 lists the required connector and cable type for each I/O feature and the ETR feature on the z10 EC.

Table 4-15 I/O features connector and cable types

Feature code	Feature name	Connector type	Cable type
0163	InfiniBand coupling (PSIFB)	MPO	50 µm MM ^a OM3 (2000 MHz-km)
0168	InfiniBand coupling (PSIFB LR)	LC Duplex	9 µm SM ^b
0219	ISC-3	LC Duplex	9 µm SM
2324	ESCON	MT-RJ	62.5 µm MM
2319 ^c	FICON Express LX	LC Duplex	9 µm SM
2320 ^c	FICON Express SX	LC Duplex	50, 62.5 µm MM
3319 ^c	FICON Express2 LX	LC Duplex	9 µm SM
3320 ^c	FICON Express2 SX	LC Duplex	50, 62.5 µm MM
3321 ^c	FICON Express4 LX 10 km	LC Duplex	9 µm SM
3322 ^c	FICON Express4 SX	LC Duplex	50, 62.5 µm MM
3324 ^c	FICON Express4 LX 4 km	LC Duplex	9 µm SM
3325	FICON Express8 LX 10 km	LC Duplex	9 µm SM
3326	FICON Express8 SX	LC Duplex	50, 62.5 µm MM
3364 ^c	OSA-Express2 GbE LX	LC Duplex	9 µm SM
3365 ^c	OSA-Express2 GbE SX	LC Duplex	50, 62.5 µm MM
3366	OSA-Express2 1000BASE-T	RJ-45	Category 5 UTP ^d
3368	OSA-Express2 10 GbE LR	SC Duplex	9 µm SM
3370	OSA-Express3 10 GbE LR	LC Duplex	9 µm SM
3371	OSA-Express3 10 GbE SR	LC Duplex	50, 62.5 µm MM
3362	OSA-Express3 GbE LX	LC Duplex	9 µm SM
3363	OSA_Express3 GbE SX	LC Duplex	50, 62.5 µm MM
3367	OSA-Express3 1000BASE-T	RJ-45	Category 5 UTP ^d
ETR	External time reference	MT-RJ	62.5 µm MM

a. MM is multimode fiber.

b. SM is single mode fiber.

c. This feature is available only if carried over on an upgrade.

d. UTP is unshielded twisted pair.

4.6.2 ESCON channels

ESCON channels support the ESCON architecture and directly attach to ESCON-supported I/O devices.

Sixteen-port ESCON feature

The 16-port ESCON feature (FC 2323) occupies one I/O slot in an I/O cage. Each port on the feature uses a 1300 nanometer (nm) light-emitting diode (LED) transceiver, designed to be connected to 62.5 μm multimode fiber optic cables only.

The feature has 16 ports with one PCHID associated with each port, up to a maximum of 15 active ESCON channels per feature. Each feature has a minimum of one spare port to allow for channel-sparing in the event of a failure of one of the other ports.

The 16-port ESCON feature port utilizes a small form factor optical transceiver that supports a fiber optic connector called MT-RJ. The MT-RJ is an industry standard connector that has a much smaller profile compared to the original ESCON Duplex connector. The MT-RJ connector, combined with technology consolidation, allows for the much higher density packaging implemented with the 16-port ESCON feature.

Notes:

- ▶ The 16-port ESCON feature does *not* support a multimode fiber optic cable terminated with an ESCON Duplex connector. However, 62.5 μm multimode ESCON Duplex jumper cables *can* be reused to connect to the 16-port ESCON feature. This is done by installing an MT-RJ/ESCON conversion kit between the 16-port ESCON feature MT-RJ port and the ESCON Duplex jumper cable. This protects the investment in the existing ESCON Duplex cabling infrastructure.
- ▶ Fiber optic conversion kits and mode conditioning patch (MCP) cables are not orderable as features. Fiber optic cables, cable planning, labeling, and installation are all customer responsibilities for new installations and upgrades.
- ▶ IBM Facilities Cabling Services - fiber transport system offers a total cable solution service to help with cable ordering needs, and are highly recommended.

ESCON channel port enablement feature

The 15 active ports on each 16-port ESCON feature are activated in groups of four ports through Licensed Internal Code Control Code (LICCC) by using the ESCON channel port feature (FC 2324).

The first group of four ESCON ports requires two 16-port ESCON features. After the first pair of ESCON cards is fully allocated (by seven ESCON port groups, using 28 ports), single cards are used for additional ESCON ports groups.

Ports are activated equally across all installed 16-port ESCON features for high availability. In most cases, the number of physically installed channels is greater than the number of active channels that are LICCC-enabled. The reason is because the last ESCON port (J15) of every 16-port ESCON channel card is a spare, and because several physically installed channels are typically inactive (LICCC-protected). These inactive channel ports are available to satisfy future channel adds.

Note: It is the intent of IBM for ESCON channels to be phased out. System z10 EC and System z10 BC will be the last servers to support greater than 240 ESCON channels. We recommend that customers review the usage of their installed ESCON channels and where possible migrate to FICON channels.

The PRIZM Protocol Converter Appliance from Optica Technologies Incorporated provides a FICON-to-ESCON conversion function that has been System z qualified. For more information see:

<http://www.opticatech.com>

Note: IBM cannot confirm the accuracy of compatibility, performance, or any other claims by vendors for products that have not been System z qualified. Questions regarding these capabilities and device support should be addressed to the suppliers of those products.

4.6.3 FICON channels

The FICON Express8, FICON Express4, FICON Express2, and FICON Express LX and SX features conform to the Fibre Channel connection (FICON) architecture and directly attach to FICON-supported I/O devices.

FICON channels can be shared among logical partitions and can be defined as spanned. All ports on a FICON feature must be of the same type, either LX or SX.

FICON Express8

This generation of FICON features for the z10 servers is designed to support a link data rate of 8 Gbps with auto negotiation to 2 or 4 Gbps to support existing devices.

The FICON Express8 features are exclusive to System z10 and are designed to deliver increased performance compared with the FICON Express4 features. For more information about FICON channel performance see the technical papers on the System z I/O connectivity Web site at:

http://www-03.ibm.com/systems/z/hardware/connectivity/ficon_performance.html

The two types of FICON Express8 channel transceivers supported on new build servers are the long wavelength (LX) laser version and the short wavelength (SX) LED version:

- ▶ FICON Express8 10km LX feature FC 3325, with four ports per feature, supporting LC Duplex connectors
- ▶ FICON Express8 SX feature FC 3326, with four ports per feature, supporting LC Duplex connectors

All channels on a feature are of the same type, either 10 km 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.

Up to 336 FICON Express8 channels (up to 84 features) can be installed in the z10 EC server. The Model E12 can have up to 256 FICON channels (64 features). The number is limited by the number of available IFB connections (based on HCA2-C fanouts) on the E12 model.

All FICON Express8 features use small form-factor pluggable (SFP) optics that allow for concurrent repair or replacement for each SFP. The data flow on the unaffected channels

on the same feature can continue. A problem with one FICON port no longer requires replacement of a complete feature.

The FICON Express8 10 km LX feature supports an unrepeated distance of 10 km using 9 μm single-mode fiber. The FICON Express8 SX feature supports varying distances depending on the fiber used (50 or 62.5 μm multimode fiber) and the link speed (2 Gbps, 4 Gbps, or 8 Gbps).

FICON Express8 10km LX feature (FC 3325)

The FICON Express8 10 km LX feature occupies one I/O slot in the I/O cage. It has four ports, each supporting an LC Duplex connector, with one PCHID and one CHPID associated with each port. It supports link speeds of 2 Gbps, 4 Gbps, or 8 Gbps up to an unrepeated distance of 10 km (6.2 miles).

Each port supports attachment to the following items:

- ▶ Fibre Channel switches that support 2 Gbps, 4 Gbps, 8 Gbps, and FICON LX Fibre Channels
- ▶ Control units that support 2 Gbps, 4 Gbps, and 8 Gbps FICON LX Fibre Channels
- ▶ FICON channels in Fibre Channel Protocol (FCP) mode

Each port of the FICON Express8 10 km LX feature uses a 1300 nanometer (nm) fiber bandwidth transceiver. The port supports connection to a 9 μm single-mode fiber optic cable terminated with an LC Duplex connector. Use of MCP cables limits the link speed to 1 Gbps and the unrepeated distance to 550 meters (1804 feet).

FICON Express8 SX feature (FC 3326)

The FICON Express8 SX feature occupies one I/O slot in the I/O cage. It has two Peripheral Component Interconnect (PCI) cards. The PCI cards have a higher performing infrastructure, which can improve performance compared with the FICON Express4 LX feature. Each PCI card has two ports supporting an LC Duplex connector, with one CHPID associated with each port, and supports link speeds of 2 Gbps, 4 Gbps, or 8 Gbps.

Each port supports attachment to the following items:

- ▶ Fibre Channel switches that support 2 Gbps, 4 Gbps, 8 Gbps, and FICON SX Fibre Channels
- ▶ Control units that support 2 Gbps, 4 Gbps, and 8 Gbps FICON SX Fibre Channels
- ▶ FICON channels in Fibre Channel Protocol (FCP) mode

Each port of the FICON Express8 SX feature uses an 850 nanometer (nm) fiber bandwidth SX transceiver. The port supports connection to a 62.5 μm or 50 μm multimode fiber optic cable terminated with an LC Duplex connector. Unrepeated distances vary with the use of 50 μm or 62.5 μm fiber optic cable and the data rate.

FICON Express4

The three types of FICON Express4 channel transceivers supported only if carried over during an upgrade are the two long wavelength (LX) laser versions and one short wavelength (SX) LED version:

- ▶ FICON Express4 10km LX feature FC 3321, with four ports per feature, supporting LC Duplex connectors
- ▶ FICON Express4 4km LX feature FC 3324, with four ports per feature, supporting LC Duplex connectors
- ▶ FICON Express4 SX feature FC 3322, with four ports per feature, supporting LC Duplex connectors

All channels on a feature are of the same type, either 10 km LX, 4 km 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.

Up to 336 FICON Express4 channels (up to 84 features) can be installed in the z10 EC server. The Model E12 can have up to 256 FICON channels (64 features). The number is limited by the number of available IFB connections (based on HCA2-C fanouts) on the E12 model.

All FICON Express4 features use small form-factor pluggable (SFP) optics that allow for concurrent repair or replacement for each SFP. The data flow on the unaffected channels on the same feature can continue. A problem with one FICON port no longer requires replacement of a complete feature.

Two FICON Express4 LX features are available. One supports an unrepeated distance of 10 km, and the other an unrepeated distance of 4 km, using 9 µm single-mode fiber. The FICON Express4 SX feature supports varying distances depending on the fiber used (50 or 62.5 µm multimode fiber) and the link speed (1 Gbps, 2 Gbps, or 4 Gbps).

FICON Express4 10km LX feature (FC 3321)

The FICON Express4 10 km LX feature occupies one I/O slot in the I/O cage. It has four ports, each supporting an LC Duplex connector, with one PCHID and one CHPID associated with each port. It supports link speeds of 1 Gbps, 2 Gbps, or 4 Gbps up to an unrepeated distance of 10 km (6.2 miles).

Interoperability of 10 km transceivers with 4 km transceivers is supported, provided the unrepeated distance between the two transceivers does not exceed 4 km (2.5 miles).

Each port supports attachment to the following items:

- ▶ Fibre Channel switches that support 1 Gbps, 2 Gbps, 4 Gbps, and FICON LX Fibre Channels
- ▶ Control units that support 1 Gbps, 2 Gbps, and 4 Gbps FICON LX Fibre Channels
- ▶ FICON channels in Fibre Channel Protocol (FCP) mode

Each port of the FICON Express4 10 km LX feature uses a 1300 nanometer (nm) fiber bandwidth transceiver. The port supports connection to a 9 µm single-mode fiber optic cable terminated with an LC Duplex connector. Use of MCP cables limits the link speed to 1 Gbps and the unrepeated distance to 550 meters (1804 feet).

FICON Express4 4km LX feature (FC 3324)

The FICON Express4 4km LX feature occupies one I/O slot in the I/O cage. It has four ports, each supporting one LC Duplex connector, with one PCHID and one CHPID associated with

each port. It supports link speeds of 1 Gbps, 2 Gbps, or 4 Gbps up to an unrepeated distance of 4 km (2.5 miles). Interoperability of 10 km transceivers with 4 km transceivers is supported, provided that the unrepeated distance between the two transceivers does not exceed 4 km.

Each port supports attachment to the following items:

- ▶ Fibre Channel switches that support 1 Gbps, 2 Gbps, 4 Gbps, and FICON LX Fibre Channels
- ▶ Control units that support 1 Gbps, 2 Gbps, and 4 Gbps FICON LX Fibre Channels
- ▶ FICON channels in Fibre Channel Protocol (FCP) mode

Each port of the FICON Express4 4km LX feature uses a 1300 nm fiber bandwidth transceiver. The port supports connection to a 9 µm single-mode fiber optic cable terminated with an LC Duplex connector. Use of MCP cables limits the link speed to 1 Gbps and the unrepeated distance to 550 meters (1804 feet).

FICON Express4 SX feature (FC 3322)

The FICON Express4 SX feature occupies one I/O slot in the I/O cage. It has two Peripheral Component Interconnect (PCI) cards. The PCI cards have a higher performing infrastructure, which can improve performance compared to the FICON Express2 LX feature. Each PCI card has two ports supporting an LC Duplex connector, with one CHPID associated with each port, and supports link speeds of 1 Gbps, 2 Gbps, or 4 Gbps.

Each port supports attachment to the following items:

- ▶ Fibre Channel switches that support 1 Gbps, 2 Gbps, 4 Gbps, and FICON SX Fibre Channels
- ▶ Control units that support 1 Gbps, 2 Gbps, and 4 Gbps FICON SX Fibre Channels
- ▶ FICON channels in Fibre Channel Protocol (FCP) mode

Each port of the FICON Express4 SX feature uses an 850 nanometer (nm) fiber bandwidth SX transceiver. The port supports connection to a 62.5 µm or 50 µm multimode fiber optic cable terminated with an LC Duplex connector. Unrepeated distances vary with the use of 50 µm or 62.5 µm fiber optic cable and the data rate.

Note: FICON Express4 is the last FICON family able to negotiate link speed down to 1 Gbps.

FICON Express2

The FICON Express2 feature is supported on a z10 EC only if carried over on an upgrade. The two types of FICON Express2 channel transceivers that are supported on z10 EC servers when carried forward on an upgrade are a long wavelength (LX) laser version and a short wavelength (SX) LED version:

- ▶ FICON Express2 LX feature FC 3319, with four ports per feature, supporting LC Duplex connectors
- ▶ FICON Express2 SX feature FC 3320, with four ports per feature, supporting LC Duplex connectors

The features are connected to a FICON-capable control unit, either point-to-point or switched point-to-point, through a Fibre Channel switch.

Up to 336 FICON Express2 channels (84 features) can be installed. The model E12 can have up to 256 FICON channels (64 features). The number is limited by the number of available IFB connections (based on HCA2-C fanouts) on the E12 model.

FICON Express2 LX feature (FC 3319)

The FICON Express2 LX feature occupies one I/O slot in the I/O cage. It has four ports, each supporting an LC Duplex connector, with one PCHID and one CHPID associated with each port. It supports link speeds of 1 Gbps or 2 Gbps.

Each port supports attachment to the following items:

- ▶ Fibre Channel switches that support 1 Gbps and 2 Gbps FICON LX Fibre Channels
- ▶ Control units that support 1 Gbps and 2 Gbps FICON LX Fibre Channels
- ▶ FICON channels in Fibre Channel Protocol (FCP) mode

Each port of the FICON Express2 LX feature uses a 1,300 nm fiber bandwidth transceiver. The port supports connection to a 9 µm single-mode fiber optic cable terminated with an LC Duplex connector.

FICON Express2 SX feature (FC 3320)

The FICON Express2 SX feature occupies one I/O slot in the I/O cage. It has four ports, each supporting an LC Duplex connector, with one PCHID and one CHPID associated with each port. It supports link speeds of 1 Gbps or 2 Gbps.

Each port supports attachment to the following items:

- ▶ Fibre Channel switches that support 1 Gbps and 2 Gbps FICON SX Fibre Channel
- ▶ Control units that support 1 Gbps and 2 Gbps FICON SX Fibre Channels
- ▶ FICON channels in Fibre Channel Protocol (FCP) mode

Each port of the FICON Express SX feature uses an 850 nm fiber bandwidth transceiver. The port supports connection to a 62.5 µm or 50 µm multimode fiber optic cable terminated with an LC Duplex connector.

FICON Express

FICON Express features (FC 2319 and FC 2320) are carried forward to the z10 EC when you upgrade from a System z9 or z990 server.

FICON Express LX feature (FC 2319)

The FICON Express LX feature occupies one I/O slot in the I/O cage. It has two ports, each supporting an LC Duplex connector, with one PCHID and one CHPID associated with each port. It supports link speeds of 1 Gbps or 2 Gbps.

Each port supports attachment to the following items:

- ▶ FICON LX Bridge 1-port feature of an IBM 9032 ESCON Director at *only* 1 Gbps
- ▶ Fibre Channel switches that support 1 Gbps and 2 Gbps FICON LX Fibre Channels
- ▶ Control units that support 1 Gbps and 2 Gbps FICON LX Fibre Channels
- ▶ FICON channels in Fibre Channel Protocol (FCP) mode

Each port of the FICON Express LX feature uses a 1300 nm fiber bandwidth transceiver. The port supports connection to a 9 µm single-mode fiber optic cable terminated with an LC Duplex connector.

Note: FICON Express2 and FICON Express4 features do not support FCV mode. FCV mode is available on z10 EC only if FICON Express LX feature 2319 is carried over on upgrades. It is intended that the z10 EC is the last server to support FICON Express LX feature 2319 and CHPID type FCV.

FICON Express SX feature (FC 2320)

The FICON Express SX feature occupies one I/O slot in the I/O cage. It has two ports, each supporting an LC Duplex connector, with one PCHID and one CHPID associated with each port. It supports link speeds of 1 Gbps or 2 Gbps.

Each port supports attachment to the following items:

- ▶ Fibre Channel switches that support 1 Gbps and 2 Gbps FICON SX Fibre Channels
- ▶ Control units that support 1 Gbps and 2 Gbps FICON SX Fibre Channels
- ▶ FICON channels in Fibre Channel Protocol (FCP) mode

Each port of the FICON Express SX feature uses an 850 nm fiber bandwidth transceiver. The port supports connection to a 62.5 μm or 50 μm multimode fiber optic cable terminated with an LC Duplex connector.

Notes:

- ▶ A multimode (62.5 or 50 μm) fiber optic cable may be used with the FICON Express LX, FICON Express2 LX, and FICON Express4 LX features *only* for 1 Gbps. The use of this multimode cable type requires a mode conditioning patch cable to be used at each end of the fiber optic link or at each optical port in the link. Use of the single mode to multimode MCP cables reduces the supported distance of the 1 Gbps link to an end-to-end maximum of 550 meters.
- ▶ Fiber optic conversion kits and MCP cables are not orderable as features. Fiber optic cables, cable planning, labeling, and installation are all customer responsibilities for new installations and upgrades.
- ▶ IBM Facilities Cabling Services - fiber transport system offers total a cable solution service to help with cable ordering needs, and is highly recommended

High performance FICON for System z (zHPF)

zHPF is an enhancement of the FICON channel architecture and is compatible with:

- ▶ Fibre Channel Physical and Signaling standard (FC-FS)
- ▶ Fibre Channel Switch Fabric and Switch Control Requirements (FC-SW)
- ▶ Fibre Channel Single-Byte-4 (FC-SB-4) standards

Exploiting zHPF by the FICON channel, the z/OS operating system, and the DS8000® control unit (and other subsystems) can reduce the FICON channel overhead. This is achieved by protocol simplification and reducing the number of information units (IUs) processed, resulting in more efficient usage of the fiber link.

The FICON Express8, FICON Express4, and FICON Express2 features, when configured as CHPID type FC, support both the existing FICON architecture and the zHPF architecture.

From the z/OS point of view, the existing FICON architecture is called *command mode* and zHPF architecture is called *transport mode*. During link initialization, the channel node and the control unit node indicate whether they support zHPF.

Note: All FICON channel paths (CHPIDs) defined to the same Logical Control Unit (LCU) must support zHPF. The inclusion of any non-compliant zHPF features in the path group (for instance, FICON Express feature codes 2319 and 2320) will cause the entire path group to support command mode only.

The mode used for an I/O operation depends on the control unit supporting zHPF and settings in the z/OS operating system. For z/OS exploitation there is a parameter in the

IECIOSxx member of SYS1.PARMLIB (ZHPF=YES or NO) and in the SETIOS system command to control whether zHPF is enabled or disabled. The default is ZHPF=NO.

Support is also added for the D IOS,ZHPF system command to indicate whether zHPF is enabled, disabled, or not supported on the server.

Similar to the existing FICON channel architecture, the application or access method provides the channel program (channel command words, CCWs). The way that zHPF (transport mode) manages channel program operations is significantly different from the CCW operation for the existing FICON architecture (command mode). While in command mode, each single CCW is sent to the control unit for execution. In transport mode, multiple channel commands are packaged together and sent over the link to the control unit in a single control block. Less overhead is generated compared to the existing FICON architecture. Certain complex CCW chains are not supported by zHPF.

Platform and name server registration in FICON channel

The FICON Express8, FICON Express4, FICON Express2, and FICON Express features on the System z10 servers support platform and name server registration to the fabric if the FICON feature is defined as CHPID type FC.

Information about the channels 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 System z10 servers:

- ▶ Platform information
- ▶ Channel information
- ▶ World Wide Port Name (WWPN)
- ▶ Port type (N_Port_ID)
- ▶ FC-4 types supported
- ▶ Classes of service supported by the channel

The platform and name server registration service are defined in the Fibre Channel - Generic Services 4 (FC-GS-4) standard.

Extended distance FICON

An enhancement to the industry standard FICON architecture (FC-SB-3) helps avoid degradation of performance at extended distances by implementing a new protocol for *persistent* information unit (IU) pacing. Extended distance FICON is transparent to operating systems and applies to all the FICON Express8, FICON Express4, and FICON Express2 features carrying native FICON traffic (CHPID type FC).

For exploitation, the control unit must support the new IU pacing protocol. IBM System Storage DS8000 series supports extended distance FICON for IBM System z environments. The channel defaults to current pacing values when it operates with control units that cannot exploit extended distance FICON.

Worldwide port name (WWPN) prediction tool

A part of the installation of your IBM System z10 server is the pre-planning of the SAN environment. IBM has made available a standalone tool to assist with this planning prior to the installation.

The tool, known as the worldwide port name (WWPN) prediction tool, assigns WWPNs to each virtual Fibre Channel Protocol (FCP) channel/port using the same WWPN assignment algorithms that a system uses when assigning WWPNs for channels utilizing N_Port Identifier

Virtualization (NPIV). Thus, the SAN can be set up in advance, allowing operations to proceed much faster once the server is installed.

The WWPN prediction tool takes a .csv file containing the FCP-specific I/O device definitions and creates the WWPN assignments that are required to set up the SAN. A binary configuration file that can be imported later by the system is also created. The .csv file can either be created manually or exported from the Hardware Configuration Definition/Hardware Configuration Manager (HCD/HCM).

FICON feature summary

Table 4-16 shows the FICON card feature codes on a z10 EC and their respective specifications, such as connector and cable type, maximum unrepeated distance, and the link data rate.

Table 4-16 FICON Channel specifications

Feature code	Feature name	Connector type	Cable type ^a	Unrepeated max. distance	Link data rate
2319 ^b	FICON Express LX	LC Duplex	SM 9 μm	10 km	1 or 2 Gbps ^c
			with MCP: MM 50 μm or MM 62.5 μm	550 m (1,804 feet)	1 Gbps
2320 ^b	FICON Express SX	LC Duplex	MM 62.5 μm	120 m (394 feet) ^d	1 or 2 Gbps ^b
			MM 50 μm	300 m (984 feet) ^c	
3319 ^b	FICON Express2 LX	LC Duplex	SM 9 μm	10 km	1 or 2 Gbps ^c
			With MCP: MM 50 μm or MM 62.5 μm	550 m (1,804 feet)	1 Gbps
3320 ^b	FICON Express2 SX	LC Duplex	MM 62.5 μm	120 m (394 feet) ^c	1 or 2 Gbps ^c
			MM 50 μm	300 m (984 feet) ^c	
3321	FICON Express4 10KM LX	LC Duplex	SM 9 μm	10 km	1, 2, or 4 Gbps ^c
			SM 9 μm with MCP	550 m (1,804 feet) ^e	
3322	FICON Express4 SX	LC Duplex	MM 62.5 μm	55 m (180 feet) ^f at 160 MHz-km 70 m (230 feet) ^f at 200 MHz-km	1, 2, or 4 Gbps ^c
			MM 50 μm	150 m (492 feet) ^f at 500 MHz-km 270 m (886 feet) ^f at 2,000 MHz-km	
3324	FICON Express4 4KM LX	LC Duplex	SM 9 μm	4 km	1, 2, or 4 Gbps ^c
			SM 9 μm with MCP	550 m (1804 feet) ^e	
3325	FICON Express8 10KM LX	LC Duplex	SM 9 μm	10 km	2, 4, or 8 Gbps ^c

Feature code	Feature name	Connector type	Cable type ^a	Unrepeated max. distance	Link data rate
3326	FICON Express8 SX	LC Duplex	MM 62.5 µm	21 m (69 feet) ^g at 200 MHz-km	2, 4, or 8 Gbps ^c
			MM 50 µm	50 m (164 feet) ^g at 500 MHz-km 150 m (492 feet) ^g at 1500 MHz-km	

- a. MM is multimode; SM is single mode
- b. Feature is only available if carried over on an upgrade
- c. Supports auto-negotiate with neighbor node
- d. Maximum unrepeated distance at 2 Gbps
- e. Maximum unrepeated distance at 1 Gbps
- f. Maximum unrepeated distance at 4 Gbps
- g. Maximum unrepeated distance at 8 Gbps

Note: FICON Express8 features do not support auto-negotiation to a data link rate of 1 Gbps.

4.6.4 OSA-Express3

This section discusses the connectivity options offered by the OSA-Express3 features.

The OSA-Express3 features provide improved performance by reducing latency at the TCP/IP application. Direct access to the memory allows packets to flow directly from the memory to the LAN without firmware intervention in the adapter.

The following OSA-Express3 features can be installed on z10 EC servers:

- ▶ OSA-Express3 10 Gigabit Ethernet (GbE) Long Range (LR), feature code 3370
- ▶ OSA-Express3 10 Gigabit Ethernet (GbE) Short Reach (SR), feature code 3371
- ▶ OSA-Express3 Gigabit Ethernet (GbE) Long wavelength (LX), feature code 3362
- ▶ OSA-Express3 Gigabit Ethernet (GbE) Short wavelength (SX), feature code 3363
- ▶ OSA Express3 1000BASE-T Gigabit Ethernet (GbE), feature code 3367

All OSA-Express3 GbE features are available on newly built servers. Up to 24 OSA-Express3 features are supported on the z10 EC, which is a total of 48 ports when on 2-port OSA-Express3 features and up to 96 ports on 4-port OSA-Express3 features.

Note that the maximum number of OSA-Express2 and OSA-Express3, in combination, is 24 features, system-wide.

Table 4-17 lists the OSA-Express3 features.

Table 4-17 OSA -Express3 feature

I/O feature	Feature code	Number ports per feature	Port increment	Maximum number ports (CHPIDs)	Maximum number features	PCHID	CHPID type
OSA Express3 10 GbE LR	3370	2	2	48	24	Yes	OSD
OSA-Express3 10 GbE SR	3371	2	2	48	24	Yes	OSD

I/O feature	Feature code	Number ports per feature	Port increment	Maximum number ports (CHPIDs)	Maximum number features	PCHID	CHPID type
OSA-Express3 GbE LX	3362	4	4	96 (48)	24	Yes	OSD, OSN
OSA-Express3 GbE SX	3363	4	4	96 (48)	24	Yes	OSD, OSN
OSA-Express3 1000BASE-T	3367	4	4	96 (48)	24	Yes 2 ports	OSC, OSD, OSE, OSN

OSA-Express3 data router

OSA-Express3 features help reduce latency and improve throughput by providing a data router. What was previously done in firmware (packet construction, inspection, and routing) is now performed in hardware. With the data router, there is now direct memory access. Packets flow directly from host memory to the LAN without firmware intervention. OSA-Express3 is also designed to help reduce the round-trip networking time between systems. Up to a 45% reduction in latency at the TCP/IP application layer has been measured.

The OSA-Express3 features are also designed to improve throughput for standard frames (1492 byte) and jumbo frames (8992 byte) to help satisfy bandwidth requirements for applications. Up to a 4x improvement has been measured (compared to OSA-Express2).

These statements are based on OSA-Express3 performance measurements performed in a laboratory environment on a System z10 and do not represent actual field measurements. Results can vary.

OSA-Express3 10 GbE LR (FC 3370)

The OSA-Express3 10 GbE LR feature occupies one slot in an I/O cage and has two ports that connect to a 10 Gbps Ethernet LAN through a 9 µm single mode fiber optic cable terminated with an LC Duplex connector. Each port on the card has a PCHID assigned. The feature supports an unrepeated maximum distance of 10 km.

Compared to the OSA-Express2 10 GbE LR feature, the OSA-Express3 10 GbE LR feature has double port density (two ports for each feature) and improved performance for standard and jumbo frames.

The OSA-Express3 10 GbE LR feature does not support auto-negotiation to any other speed and runs in full-duplex mode only. It supports 64B/66B encoding, whereas GbE supports 8B/10B encoding. Therefore, auto-negotiation to any other speed is not possible.

The OSA-Express3 10 GbE LR feature has two CHPIDs, with each CHPID having one port. The supported CHPID type is OSD (QDIO mode), which is supported by z/OS, z/VM, z/VSE, TPF, and Linux on System z.

OSA-Express3 10 GbE SR (FC 3371)

The OSA-Express3 10 GbE SR feature (FC 3371) occupies one slot in the I/O cage and has two CHPIDs, with each CHPID having one port. The supported CHPID type is OSD (QDIO mode), which is supported by z/OS, z/VM, z/VSE, TPF, and Linux on System z.

External connection to a 10 Gbps Ethernet LAN is done through a 62.5 μm or 50 μm multimode fiber optic cable terminated with an LC Duplex connector. The maximum supported unrepeated distance is 33 meters (108 feet) on a 62.5 μm multimode (200 MHz) fiber optic cable, 82 meters (269 feet) on a 50 μm multi mode (500 MHz) fiber optic cable, and 300 meters (984 feet) on a 50 μm multimode (2000 MHz) fiber optic cable.

The OSA-Express3 10 GbE SR feature does not support auto-negotiation to any other speed and runs in full-duplex mode only. OSA-Express3 10 GbE SR supports 64B/66B encoding, whereas GbE supports 8B/10 encoding, making auto-negotiation to any other speed impossible.

OSA-Express3 GbE LX (FC 3362)

Feature code 3362 is exclusive to the z10 server and occupies one slot in the I/O cage. It has four ports that connect to a 1 Gbps Ethernet LAN through a 9 μm single mode fiber optic cable terminated with an LC Duplex connector, supporting an unrepeated maximum distance of 5 km (3.1 miles). Multimode (62.5 or 50 μm) fiber optic cable can be used with this features.

Note: The use of these multimode cable types requires a mode conditioning patch (MCP) cable at each end of the fiber optic link. Use of the single mode to multimode MCP cables reduces the supported distance of the link to a maximum of 550 meters (1084 feet).

The OSA-Express3 GbE LX feature does not support auto-negotiation to any other speed and runs in full-duplex mode only.

The OSA-Express3 GbE LX feature has two CHPIDs, with each CHPID in OSD mode having two ports for a total of four ports per feature. Exploitation of all four ports requires operating system support. See 7.2, “Support by operating system” on page 190.

OSA-Express3 GbE SX (FC 3363)

Feature code 3363 is exclusive to the z10 server and occupies one slot in the I/O cage. It has four ports that connect to a 1 Gbps Ethernet LAN through a 50 μm or 62.5 μm multimode fiber optic cable terminated with an LC Duplex connector over an unrepeated distance of 550 meters (for 50 μm fiber) or 220 meters (for 62.5 μm fiber).

The OSA-Express2 GbE SX feature does not support auto-negotiation to any other speed and runs in full-duplex mode only.

The OSA-Express3 GbE SX feature has two CHPIDs in OSD mode, with each CHPID having two ports for a total of four ports per feature. Exploitation of all four ports requires operating system support. See section 7.2, “Support by operating system” on page 190.

OSA-Express3 1000BASE-T Ethernet feature (FC 3367)

Feature code 3367 is exclusive to the z10 servers and occupies one slot in the I/O cage. It has four ports that connect to a 1000 Mbps (1 Gbps), 100 Mbps, or 10 Mbps Ethernet LAN. Each port has an RJ-45 receptacle for cabling to an Ethernet switch. The RJ-45 receptacle is required to be attached using EIA/TIA category 5 unshielded twisted pair (UTP) cable with a maximum length of 100 meters (328 feet).

The OSA-Express3 1000BASE-T Ethernet feature supports auto-negotiation when attached to an Ethernet hub, router, or switch. If you allow the LAN speed and duplex mode to default to auto-negotiation, the OSA-Express port and the attached hub, router, or switch auto-negotiate the LAN speed and duplex mode settings between them and connect at the highest common performance speed and duplex mode of interoperation. If the attached Ethernet hub, router, or switch does not support auto-negotiation, the OSA-Express port

examines the signal it is receiving and connects at the speed and duplex mode of the device at the other end of the cable.

The following settings are supported on the OSA-Express3 1000BASE-T Ethernet feature port:

- ▶ Auto-negotiate
- ▶ 10 Mbps half-duplex
- ▶ 10 Mbps full-duplex
- ▶ 100 Mbps half-duplex
- ▶ 100 Mbps full-duplex
- ▶ 1000 Mbps full-duplex

If you are not using auto-negotiate, the OSA-Express port will attempt to join the LAN at the specified speed and duplex mode. If this does not match the speed and duplex mode of the signal on the cable, the OSA-Express port will not connect.

4.6.5 OSA-Express2

This section discusses the connectivity options offered by the OSA-Express2 features.

The following OSA-Express2 features can be installed on z10 EC servers:

- ▶ OSA-Express2 Gigabit Ethernet (GbE) Long Wavelength (LX), feature code 3364
- ▶ OSA-Express2 Gigabit Ethernet (GbE) Short Wavelength (SX), feature code 3365
- ▶ OSA-Express2 1000BASE-T Ethernet, feature code 3366
- ▶ OSA-Express2 Gigabit Ethernet 10 GbE LR, feature code 3368

OSA-Express features installed in previous servers are *not* supported on a z10 EC and cannot be carried forward on an upgrade.

A z10 EC supports up to 24 OSA-Express2 features (48 ports). The maximum number of combined OSA-Express2 and OSA-Express3 features is 24.

Table 4-18 lists the OSA-Express2 features.

Table 4-18 OSA-Express2 features

I/O feature	Feature code	Number of		Max. number of		PCHID	CHPID type
		Ports per feature	Port increments	Ports	I/O slots		
OSA-Express2 GbE LX/SX	3364 3365	2	2	48	24	Yes	OSD, OSN
OSA-Express2 ^a 10 GbE LR	3368	1	1	24	24	Yes	OSD
OSA-Express2 1000BASE-T	3366	2	2	48	24	Yes	OSE, OSD, OSC, OSN

a. This feature is available only if carried over on an upgrade.

OSA-Express2 10 GbE LR (FC 3368)

The OSA-Express2 10 GbE LR feature occupies one slot in an I/O cage and has one port that connects to a 10 Gbps Ethernet LAN through a 9 μm single mode fiber optic cable terminated

with an LC Duplex connector. The feature supports an unrepeated maximum distance of 10 km.

The OSA-Express2 10 GbE LR feature does not support auto-negotiation to any other speed and runs in full-duplex mode only. The OSA-Express 10 GbE LR feature is defined as CHPID type OSD.

CHPID type OSD is supported by z/OS, z/VM, z/VSE, TPF, and Linux on System z.

OSA-Express2 GbE LX (FC 3364)

The OSA-Express2 Gigabit (GbE) Long Wavelength (LX) feature occupies one slot in an I/O cage and has two independent ports, with one PCHID associated with each port.

Each port supports a connection to a 1 Gbps Ethernet LAN through a 9 μ m single-mode fiber optic cable terminated with an LC Duplex connector. This feature uses a long wavelength laser as the optical transceiver.

A multimode (62.5 or 50 μ m) fiber cable may be used with the OSA-Express2 GbE LX feature. The use of these multimode cable types requires a mode conditioning patch (MCP) cable to be used at each end of the fiber link. Use of the single-mode to multimode MCP cables reduces the supported optical distance of the link to a maximum end-to-end distance of 550 meters.

The OSA-Express2 GbE LX feature supports Queued Direct Input/Output (QDIO) and OSN modes only, full-duplex operation, jumbo frames, and checksum offload. It is defined with CHPID types OSD or OSN.

OSA-Express2 GbE SX (FC 3365)

The OSA-Express2 Gigabit (GbE) Short Wavelength (SX) feature occupies one slot in an I/O cage and has two independent ports, with one PCHID associated with each port.

Each port supports a connection to a 1 Gbps Ethernet LAN through a 62.5 μ m or 50 μ m multimode fiber optic cable terminated with an LC Duplex connector. The feature uses a short wavelength laser as the optical transceiver.

The OSA-Express2 GbE SX feature supports Queued Direct Input/Output (QDIO) and OSN mode only, full-duplex operation, jumbo frames, and checksum offload. It is defined with CHPID types OSD or OSN.

OSA-Express2 1000BASE-T Ethernet (FC 3366)

The OSA-Express2 1000BASE-T Ethernet occupies one slot in the I/O cage and has two independent ports, with one PCHID associated with each port.

Each port supports connection to either a 1000BASE-T (1000 Mbps), 100BASE-TX (100 Mbps), or 10BASE-T (10 Mbps) Ethernet LAN. The LAN must conform either to the IEEE 802.3 (ISO/IEC 8802.3) standard or to the DIX V2 specifications.

Each port has an RJ-45 receptacle for cabling to an Ethernet switch that is appropriate for the LAN speed. The RJ-45 receptacle is required to be attached using EIA/TIA category 5 unshielded twisted pair (UTP) cable with a maximum length of 100 m (328 ft).

The OSA-Express2 1000BASE-T Ethernet feature supports auto-negotiation and automatically adjusts to 10 Mbps, 100 Mbps, or 1000 Mbps, depending upon the LAN.

The OSA-Express2 1000BASE-T Ethernet feature supports CHPID types, OSC, OSD, OSE, and OSN.

You may choose any of the following settings for the OSA-Express2 1000BASE-T Ethernet and OSA-Express2 1000BASE-T Ethernet features:

- ▶ Auto-negotiate
- ▶ 10 Mbps half-duplex or full-duplex
- ▶ 100 Mbps half-duplex or full-duplex
- ▶ 1000 Mbps or 1 Gbps full-duplex

LAN speed and duplexing mode default to auto negotiation. The feature port and the attached switch automatically negotiate these settings. If the attached switch does not support auto-negotiation, the port enters the LAN at the default speed of 1000 Mbps and full-duplex mode.

The 1000BASE-T Ethernet feature can be configured as CHPID type OSC, OSD, OSE, or OSN.

Non-QDIO operation mode requires CHPID type OSE. When configured at 1 Gbps, the 1000BASE-T Ethernet feature has the same attributes as the fiber Gigabit Ethernet features:

- ▶ Operates in QDIO mode only (CHPID type OSD)
- ▶ Carries TCP/IP packets only
- ▶ Operates in full-duplex mode only
- ▶ Supports jumbo frames
- ▶ Supports checksum offload

4.6.6 Open Systems Adapter selected functions

This section discusses several OSA functions that are particularly important regarding performance, availability, manageability, or security.

Open System Adapter for NCP

OSA-Express3 GbE, OSA-Express3 1000BASE-T Ethernet, OSA-Express2 GbE, and OSAExpress2 1000BASE-T Ethernet features can provide channel connectivity from an operating system in a z10 EC to IBM Communication Controller for Linux on System z (CCL) with the Open Systems Adapter for NCP (OSN), in support of the Channel Data Link Control (CDLC) protocol. OSN eliminates requiring an external communication medium for communications between the operating system and the CCL image.

With OSN, using an external ESCON channel is unnecessary. Data flow of the logical-partition to the logical-partition is accomplished by the OSA-Express3 or OSA-Express2 feature without ever exiting the card. OSN support allows multiple connections between the same CCL image and the same operating system (such z/OS or TPF). The operating system must reside in the same physical server as the CCL image.

For CCL planning information see *IBM Communication Controller for Linux on System z V1.2.1 Implementation Guide*, SG24-7223.

For the most recent CCL information, see:

<http://www-01.ibm.com/software/network/ccl/>

Integrated Console Controller

The 1000BASE-T Ethernet features also provide the Integrated Console Controller (OSA-ICC) function, which supports TN3270E (RFC 2355) and non-SNA DFT 3270 emulation. The OSA-ICC function uses a definition as CHPID type OSC and console controller, and has multiple logical partitions support, both as shared or spanned channels.

With the OSA-ICC function, 3270 emulation for console session connections is integrated in the z10 EC through a port on the OSA-Express3 or OSA-Express2 1000BASE-T features. This function eliminates the requirement for external console controllers, such as 2074 or 3174, helping to reduce cost and complexity. Each port can support up to 120 console session connections.

OSA-ICC can be configured on a PCHID-by-PCHID basis and is supported at any of the feature settings (10, 100, or 1000 Mbps, half-duplex or full-duplex).

Link aggregation support for z/VM

Link aggregation (IEEE 802.3ad) controlled by the z/VM Virtual Switch (VSWITCH) allows the dedication of an OSA-Express3 or OSA-Express2 port to the z/VM operating system, when the port is participating in an aggregated group configured in Layer 2 mode. Link aggregation (trunking) is designed to allow combining multiple physical OSA-Express3 or OSA-Express2 ports into a single logical link for increased throughput and for nondisruptive failover in the event that a port becomes unavailable. The target links for aggregation must be of the same type.

Link aggregation is exclusive on System z10 and System z9 servers. It is applicable to the OSA-Express, OSA-Express3, and OSA-Express2 features when configured as CHPID type OSD (QDIO). Link aggregation is supported by z/VM V5.3.

QDIO data connection isolation for z/VM

The Queued Direct I/O (QDIO) data connection isolation function provides a higher level of security on System z10 and System z9 servers when sharing the same OSA connection in z/VM environments that use the Virtual Switch (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, and provides a means for creating security zones and preventing network traffic between the zones.

VSWITCH isolation support is provided by APAR VM64281. z/VM 5.3 and z/VM 5.4 support is provided by CP APAR VM64463 and TCP/IP APAR PK67610.

QDIO data connection isolation is supported by all OSA-Express3 and OSA-Express2 features on System z10, and by all OSA-Express2 features on System z9, with an MCL update (refer to 2097DEVICE for details).

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 will discard any packets destined for a z/OS LPAR that is registered in the OAT as isolated.

QDIO interface isolation is supported by Communications Server for z/OS V1R11 and all OSA-Express3 and OSA-Express2 features on System z10.

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 as follows:

- ▶ For inbound processing, the TCP/IP stack looks more frequently for available data to process, ensuring that any new data is read from the OSA-Express3 without requiring additional program controlled interrupts (PCIs).
- ▶ For outbound processing, the OSA-Express3 also looks more frequently for available data to process from the TCP/IP stack, thus not requiring a Signal Adapter (SIGA) instruction to determine whether more data is available.

Checksum offload for IPv4 packets when in QDIO mode

A function referred to as *checksum offload*, supports z/OS and Linux on System z environments. It is offered on the OSA-Express3 GbE, OSA-Express3 100BASE-T Ethernet, OSA-Express2 GbE, and OSA-Express2 1000BASE-T Ethernet features. Checksum offload provides the capability of calculating the Transmission Control Protocol (TCP), User Datagram Protocol (UDP), and Internet Protocol (IP) header checksum. Checksum verifies the accuracy of files. By moving the checksum calculations to a Gigabit or 1000BASE-T Ethernet feature, host CPU cycles are reduced and performance is improved.

When checksum is offloaded, the OSA-Express feature performs the checksum calculations for Internet Protocol Version 4 (IPv4) packets. The checksum offload function applies to packets that go to or come from the LAN. When multiple IP stacks share an OSA-Express, and an IP stack sends a packet to a next hop address owned by another IP stack that is sharing the OSA-Express, the OSA-Express then sends the IP packet directly to the other IP stack without placing it out on the LAN. Checksum offload does not apply to such IP packets.

Checksum offload is supported by the GbE features (FC 3362, FC 3363, FC 3364, and FC 3365) and the 1000BASE-T Ethernet features (FC 3366 and FC 3367) when operating at 1000 Mbps (1 Gbps). Checksum offload is applicable to the QDIO mode only (channel type OSD).

z/OS support for checksum offload is available in all in-service z/OS releases, and in all supported Linux on System z distributions.

Adapter interruptions for QDIO

Linux on System z and z/VM work together to provide performance improvements by exploiting extensions to the Queued Direct I/O (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 overhead in both the host operating system and the adapter (OSA-Express3 and OSA-Express2 when using type OSD CHPID).

In extending the use of adapter interruptions to OSD (QDIO) channels, the programming overhead to process a traditional I/O interruption is reduced. This benefits OSA-Express TCP/IP support in Linux on System z, z/VM and z/VSE.

Adapter interruptions apply to all of the OSA-Express3 and OSA-Express2 features on z10 EC when in QDIO mode (CHPID type OSD).

OSA Dynamic LAN idle

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, which was previously a static setting.

For latency-sensitive applications, the blocking algorithm is modified to be *latency sensitive*. For streaming (throughput-sensitive) applications, the blocking algorithm is adjusted to maximize throughput. In all cases, the TCP/IP stack determines the best setting based on the

current system and environmental conditions (inbound workload volume, processor utilization, traffic patterns, and so on) and can dynamically update the settings. OSA-Express3 and OSA-Express2 features adapt to the changes, avoiding thrashing and frequent updates to the OSA address table (OAT). Based on the TCP/IP settings, OSA holds the packets before presenting them to the host. A dynamic setting is designed to avoid or minimize host interrupts.

OSA Dynamic LAN idle is exclusive to System z10 and System z9 servers, is supported by the OSA-Express2 and OSA-Express3 features (CHPID type OSD), and is exploited by z/OS V1.8 (or higher) with program temporary fixes (PTFs).

VLAN management

To simplify the network administration and management of VLANs, the GARP VLAN Registration Protocol (GVRP) can be used. All OSA-Express3 and OSA-Express2 features support VLAN prioritization, a component of the IEEE 802.1 standard. With this support, manually entering VLAN IDs at the switch is no longer necessary. The OSA-Express3 and OSA-Express2 features, when in QDIO mode (CHPID type OSD), can have GVRP dynamically register VLAN IDs.

OSA Layer 3 Virtual MAC for z/OS environments

To help simplify the infrastructure and to facilitate load balancing when a logical partition is sharing the same OSA Media Access Control (MAC) address with another logical partition, each operating system instance can have its own unique *logical* or *virtual* MAC (VMAC) address. All IP addresses associated with a TCP/IP stack are accessible by using their own VMAC address, instead of sharing the MAC address of an OSA port, which also applies to Layer 3 mode and to an OSA port spanned among channel subsystems.

OSA Layer 3 VMAC is exclusive to System z10 and System z9 servers and is applicable to the OSA-Express3 and OSA-Express2 features when configured as CHPID type OSD (QDIO), and is supported by z/OS V1.8 (or higher).

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 an OSA-Express3 or OSA-Express2 feature (by using a diagnostic assist function) to stop traces and capture the current trace records.

QDIO Diagnostic Synchronization is exclusive to System z10 and System z9 servers on OSA-Express3 and OSA-Express2 features when configured as CHPID type OSD.

z/OS V1.8 (and higher) implements software support for QDIO Diagnostic Synchronization.

Network Traffic Analyzer

With the large volume and complexity of today's network traffic, the System z10 offers systems programmers and network administrators the ability to more easily solve network problems. With the availability of the OSA-Express Network Traffic Analyzer and QDIO Diagnostic Synchronization on the server, you can capture trace and trap data, and forward it to z/OS tools for easier problem determination and resolution.

The Network Traffic Analyzer is exclusive to System z10 and System z9 servers, and OSA-Express3 and OSA-Express2 features when configured as CHPID type OSD. Support is available in z/OS V1.8 or higher.

4.6.7 HiperSockets

The HiperSockets function is an integrated function of System z10 that provides users with attachments to up to sixteen high-speed virtual LANs with minimal system and network overhead. HiperSockets is also known as internal Queued Direct Input/Output (iQDIO) or internal QDIO.

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, which 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 logical partitions in the same System z10 server. HiperSockets offers significant value in server consolidation connecting many virtual servers, and can be used instead of certain coupling link configurations in a Parallel Sysplex.

HiperSockets multiple write facility

The HyperSockets function has been enhanced on System z10 server to support multiple output buffers on a single SIGA write instruction. This operation is beneficial for the streaming of bulk data over a HyperSockets link between two logical partitions.

The receiving partition processes a much larger amount of data per I/O interrupt. This is transparent to the operating system in the receiving partition. HiperSockets Multiple Write Facility with fewer I/O interrupts is designed to reduce CPU utilization of the sending and receiving partitions.

System z10 HiperSockets Layer 2 support

HiperSockets internal networks on System z10 servers 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 and Layer 3-independent. Each HiperSockets device (Layer 2 and Layer 3 mode) has its own MAC address designed to allow the use of applications that depend on the existence of Layer 2 addresses, such as DHCP servers and firewalls. Layer 2 support helps facilitate server consolidation, can reduce complexity, can simplify network configuration, and allows LAN administrators to maintain the mainframe network environment similarly as for non-mainframe environments.

Packet forwarding decisions are based on Layer 2 information instead of Layer 3. The HiperSockets device can perform automatic MAC address generation to create uniqueness within and across logical partitions and servers. The use of Group MAC addresses for multicast is supported as well as 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 logical partition 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 enables the creation of high-performance and high-availability link layer switches between the internal HiperSockets

network and an external Ethernet or to connect to the HiperSockets Layer 2 networks of different servers.

HiperSockets Layer 2 support is exclusive on System z10 EC, supported by Linux on System z, and by z/VM for Linux guest exploitation.

4.7 Parallel Sysplex connectivity

Coupling links are required in a Parallel Sysplex configuration to provide connectivity from the z/OS images to the coupling facility. A properly configured Parallel Sysplex provides a highly reliable, redundant, and robust System z technology solution to achieve near-continuous availability. A Parallel Sysplex comprises one or more z/OS operating system images coupled through one or more coupling facilities.

The type of coupling link that is used to connect a CF to an operating system logical partition is important because of the effect of the link performance on response times and coupling overheads. For configurations covering large distances, the time spent on the link can be the largest part of the response time.

The types of links that are available to connect an operating system logical partition to a coupling facility are:

- ▶ ISC-3

The ISC-3 type is available in peer mode only. ISC-3 links can be used to connect to System z10, System z9, z990, or z890 servers. They are fiber links that support a maximum distance of 10 km, 20 km with RPQ 8P2197, and 100 km with dense wave division multiplexing (DWDM). ISC-3s operate in single mode only. Link bandwidth is 200 MBps for distances up to 10 km, and 100 MBps when RPQ 8P2197 is installed. Each port operates at 2 Gbps. Ports are ordered in increments of one. The maximum number of ISC-3 links per z10 EC is 48. ISC-3 supports transmission of Server Time Protocol (STP) messages.

- ▶ ICB-4

The ICB-4 type connects a System z10 to a System z10, System z9, z990, or z890 server. The maximum distance between the two servers is seven meters (maximum cable length is 10 meters). The link bandwidth is 2 GBps. ICB-4 links can be defined only in peer mode. The maximum number of ICB-4 links is 16 per z10 EC. ICB-4 supports transmission of STP messages.

Note: The System z10 servers will be the last family of servers to support ICB-4s.

- ▶ PSIFB

Parallel Sysplex using Infiniband (PSIFB) connects a System z10 to another System z10 or a System z10 to a z9 EC or z9 BC. PSIFB links are fiber connections that support a maximum distance of up to 150 meters. PSIFB coupling links are defined as CHPID type CIB. The maximum number of PSIFB links is 32 for each z10 EC. PSIFB supports transmission of STP messages.

► PSIFB LR

Parallel Sysplex using InfiniBand connects a System z10 to another z10 server. PSIFB links are fiber connections that support a maximum unrepeated distance of up to 10 km and up to 100 km with dense wave division multiplexing (DWDM). PSIFB LR (Long Reach) coupling links are defined as CHPID type CIB. The maximum number of PSIFB LR links is 32 for each z10 EC server. PSIFB LR supports transmission of STP messages.

► IC

CHPIDs (type ICP) defined for internal coupling can connect a CF to a z/OS logical partition in the same System z10. IC connections require two CHPIDs to be defined, which can only be defined in peer mode. The bandwidth is greater than 2 GBps. A maximum of 32 IC CHPIDs (16 connections) can be defined.

Table 4-19 shows the coupling link options.

Table 4-19 Coupling link options

Type	Description	Use	Link rate	Distance	z10 EC maximum
ISC-3	Fiber connection	System z10 to System z10, System z9, z990, z890	2 Gbps	10 km unrepeated (6.2 miles) 100 km repeated	48
ICB-4	Copper connection	System z10 to System z10, System z9, z990, z890	2 GBps	10 meters (33 feet)	16
PSIFB	12x IB-DDR fiber connection	System z10 to System z10	6 GBps	150 meters (492 feet)	32
	12x IB-SDR fiber connection	System z10 to System z9	3 GBps ^a		
PSIFB LR	1x IB-SDR ^b fiber connection	System z10 to System z10	2.5 Gbps 5.0 Gbps	10 km unrepeated (6.2.miles) 100 km repeated	32
IC	Internal coupling channel	Internal communication	Internal speeds	N/A	32

a. When connected to a System z9 EC or System z9 BC.

b. Double data rate (1x IB-DDR) is supported if connected to a DWDM supporting DDR.

The maximum PSIFB + ICB-4 links is 32. The maximum number of coupling links combined cannot exceed 64 per server (ICB-4, ISC-3 links, PSIFB). There is a maximum of 64 coupling CHPIDs, including CIB, CFP, CBP, and ICP per server.

The z10 EC supports several connectivity options depending on the connected System z server. Figure 4-5 shows z10 EC coupling link support to System z servers.

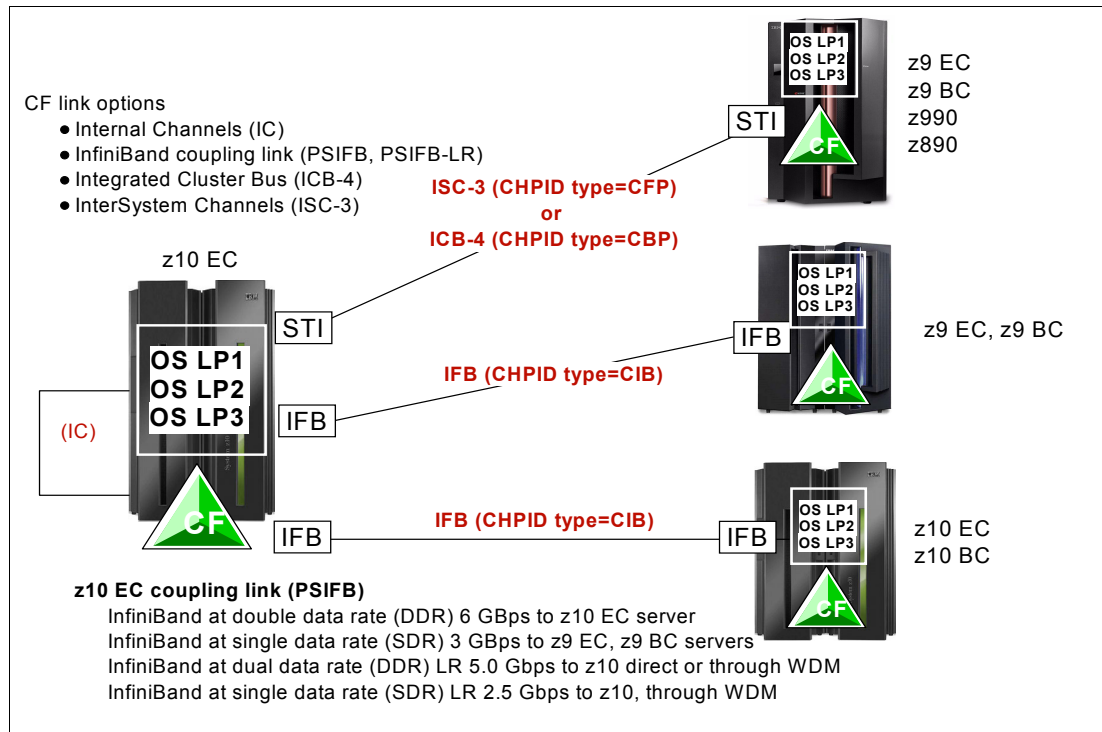


Figure 4-5 z10 EC to System z CF connectivity options

z/OS and coupling facility images may be running on the same or on separate servers. There must be at least one CF connected to all z/OS images, although there can be other CFs that are connected only to selected z/OS images. Two coupling facility images are required for system-managed CF structure duplexing and, 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, there should be at least:

- ▶ 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 you are using system-managed CF structure duplexing or running with *resource sharing* only, then a stand-alone coupling facility is not mandatory.

4.7.1 Coupling links

A coupling link provides connectivity to servers in a Parallel Sysplex environment. Coupling links are also used to transmit messages when Server Time Protocol (STP) is enabled.

Coupling link features

The z10 EC supports four types of coupling link options:

- ▶ Inter-system channel, ISC-3 , FC 0217, FC 0218, and FC 0219
- ▶ Integrated cluster bus, ICB-4, FC 3393
- ▶ Parallel Sysplex using InfiniBand (PSIFB) coupling, FC 0163
- ▶ Parallel Sysplex using InfiniBand Long Reach (PSIFB LR) coupling, FC 0168

The coupling link features available on the z10 EC connect z10 EC servers to the identified System z servers by various link options:

- ▶ ISC-3 at 2 Gbps to System z10, z9 EC, z9 BC, z990 and z890
- ▶ ICB-4 at 2 GBps to System z10, z9 EC, z9 BC, z990 and z890
- ▶ PSIFB at 6 GBps to System z10 or 3 GBps to z9 EC and z9 BC
- ▶ PSIFB LR at 5.0 or 2.5 Gbps to System z10 servers

ISC-3 coupling links

Three feature codes are available to implement ISC-3 coupling links:

- ▶ FC 0217, ISC-3 mother card
- ▶ FC 0218, ISC-3 daughter card
- ▶ FC 0219, ISC-3 port

The ISC mother card (FC 0217) occupies one slot in the I/O cage and supports up to two daughter cards. The ISC daughter card (FC 0218) provides two independent ports with one PCHID associated with each enabled port. The ISC-3 ports are enabled and activated by Licensed Internal Code.

When the quantity of ISC links (FC 0219) is selected, the quantity of ISC-3 port features selected determines the appropriate number of ISC-3 mother and daughter cards to be included in the configuration, up to a maximum of 12 ISC-M cards. Additional ISC-M cards can be ordered, up to the number of ISC-D features or twelve, whichever is smaller.

Each active ISC-3 port in peer mode supports a 2 Gbps (200 MBps) connection through 9 μ m single mode fiber optic cables terminated with an LC Duplex connector. The maximum unrepeated distance for an ISC-3 link is 10 km. With repeaters the maximum distance extends to 100 km. ISC-3 links can be defined as *timing-only links* when STP is enabled. Timing-only links are coupling links that allow two servers to be synchronized using STP messages when a CF does not exist at either end of the link.

RPQ 8P2197 extended distance option

The RPQ 8P2197 daughter card provides two ports that are active and enabled when installed and do not require activation by LIC.

This RPQ allows the ISC-3 link to operate at 1 Gbps (100 MBps) instead of 2 Gbps (200 MBps). This lower speed allows an extended unrepeated distance of 20 km. One RPQ daughter is required on both ends of the link to establish connectivity to other servers. This RPQ supports STP if defined as either a coupling link or timing-only.

ICB-4 link (FC 3393)

The ICB-4 link option uses a 10 m copper cable to connect to other z10 EC, System z9, z990, or z890 servers. The ICB-4 copper cable is plugged directly into an MBA fanout on both sides of the link. One ICB-4 feature is required for each end of the link. The maximum of 16 ICB-4 links is supported. The ICB-4 link operates at 2 GBps.

Different ICB cables are required when connecting a z10 EC to another z10 EC or connecting a z10 EC to a System z9, z990, or z890 server. FC 0229 provides the 10 meter copper cable

to connect to System z9, z990, or z890; FC 0230 provides a cable to connect to a z10 EC server. When you order an ICB-4, you must specify the cable feature code.

The PCHID assigned to the ICB-4 depends on the physical location of the fanout.

When STP is enabled, ICB-4 links can be defined as timing-only links to other System z10, System z9, z990, and z890 servers.

Note: IBM intends for System z10 EC to be the last server to support ICB-4 links. When adding coupling links to a z10 server to a possible future server, for migration purposes we recommend adding PSIFB coupling links.

InfiniBand coupling links (FC 0163)

The Parallel Sysplex using InfiniBand (PSIFB) coupling option uses InfiniBand over an optical interface. FC 0163 supports the optical coupling link.

Each fanout provides two ports. Both ports on the HCA2-O fanout are exclusively used for coupling links and *cannot* be shared for other functions. Up to 16 CHPIDs are supported for each HCA2-O fanout.

The maximum distance between servers connected by the ICB-4 copper cable is 10 meters; the maximum distance for PSIFB coupling links is 150 meters.

The maximum number of FC 0163s on a z10 EC is 16, supporting 32 optical links. The maximum number of links to a System z9 is 16.

The InfiniBand coupling link is defined as channel type CIB in the IOCDS. The coupling links can be defined as shared between images within a channel subsystem and they can be also be spanned across multiple CSSs in a server.

Each fanout for optical links (FC 0163) supports up to 16 CHPIDs across both ports, which 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.

Note: We recommend no more than four CHPIDs per port.

When connected to an external server, the same physical link is used for all 16 CHPIDs. The source and the target operating system or CF image must be defined in the IOCDS.

Figure 4-6 shows an optical link connection between two servers. Port J02 on the fanout in location D2 is connected to port J01 on fanout D1.

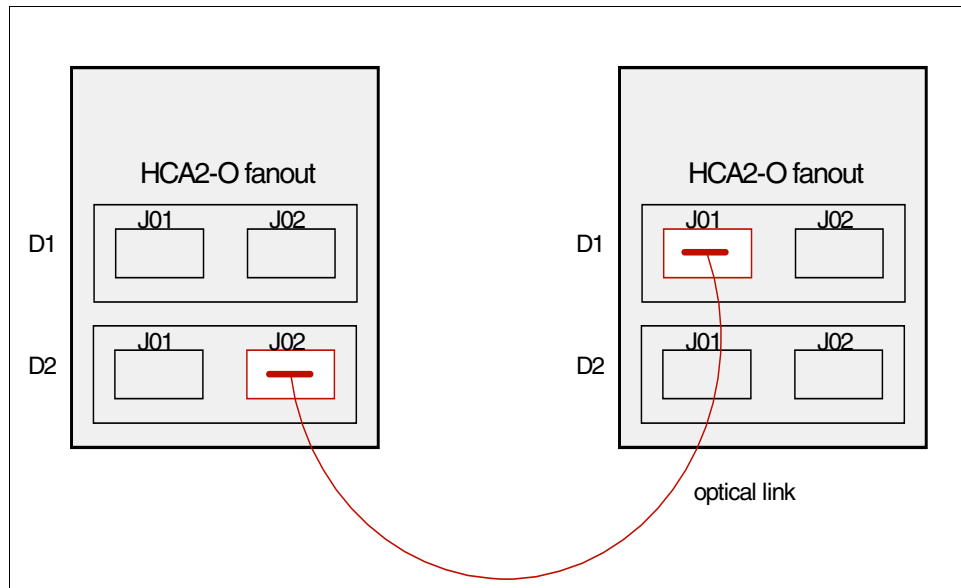


Figure 4-6 InfiniBand optical link

Definitions of the source and target operating system image, CF image, and the CHPIDs used on both ports in both servers, are defined in IOCDS.

Coupling links to System z9 servers require FC 0167 to be ordered for the System z9 server. Up to 16 InfiniBand coupling links are supported in the System z9 to connect to a z10 EC server, and up to 32 InfiniBand coupling links are supported in the System z10 EC to connect to a System z10 or z9 server. There is a combined maximum of 32 PSIFB and ICB-4 links. The link rate for coupling links to System z9 servers is auto-negotiated to the highest maximum rate, which is 3 GBps on the System z9 server.

When STP is enabled, PSIFB coupling links can be defined as timing-only links to other System z10 and System z9 servers.

InfiniBand coupling links LR (FC 0168)

The Parallel Sysplex using InfiniBand Long Reach (PSIFB LR) coupling option uses InfiniBand over an optical interface. FC 0168 supports the optical coupling link.

Each fanout provides two ports. Both ports on the HCA2-O LR fanout are exclusively used for coupling links and *cannot* be shared for other functions. Up to 16 CHPIDs are supported per HCA2-O LR fanout.

Note: We recommend no more than four CHPIDs per port.

The maximum unrepeated distance for PSIFB LR coupling links is 10 km, and up to 100 km if using repeaters.

The maximum number of FC 0168s on a z10 EC is 16, supporting 32 optical links. The PSIFB coupling links are intended to replace the existing ISC-3 coupling links. It uses the same fiber link components as are used for ISC-3 coupling links, which is a 9 μ m single mode (SM) cable terminated with an LC Duplex connector.

The InfiniBand coupling link is defined as channel type CIB in the IOCDs. The coupling links can be defined as shared between images within a channel subsystem and they can be also be spanned across multiple CSSs in a server.

Each fanout for optical links (FC 0168) supports up to 16 CHPIDs across both ports, which 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.

Definitions of the source and target operating system image, CF image, and the CHPIDs used on both ports in both servers, are defined in IOCDs.

When STP is enabled, PSIFB LR coupling links can be defined as timing-only links to other System z10 servers.

The PSIFB LR feature is exclusive to System z10 servers and PSIFB LR coupling link connectivity to other servers is not supported.

Internal coupling links

The z10 EC provides the capability to define a coupling link that does not use any external cables.

IC CHPIDs connect a CF to z/OS logical partitions in the same server. These CHPIDs are available on all System z servers, including the z10 EC. The definition and usage of these CHPIDs are the same as previous servers.

IC CHPIDs are used when an ICF logical partition is on the same server as other system images participating in the sysplex. An IC CHPID is a fast coupling connection, using memory-to-memory data transfers. IC connections do not have PCHID numbers, but do require CHPIDs.

An IC connection requires an ICP channel path definition at the z/OS and the CF end of a channel connection to operate in peer mode. IC connections are always defined and connected in pairs. The IC connection operates in peer mode and its existence is defined in HCD/IOCP.

Coupling link migration considerations

IBM has issued a Statement of Direction (SOD) regarding IBM System z9 Business Class (BC) and Enterprise Class (EC). The z9 EC and z9 BC will be the last servers to support active participation in the same Parallel Sysplex with IBM eServer™ zSeries 900 (z900), IBM eServer zSeries 800 (z800), and older S/390® Parallel Enterprise Server systems.

The following restrictions apply to customers that are running a Parallel Sysplex including one or more z900 or z800 servers, and are installing a z10 EC server:

- ▶ The z10 EC cannot be added to the Parallel Sysplex.
- ▶ Rolling IPLs cannot be performed to introduce the z10 EC.
- ▶ If the sysplex also includes any z990, z890, z9 EC, or z9 BC that is being upgraded, then the z900 and z800 in the sysplex must either be upgraded or removed from the sysplex.
- ▶ When the z900 or z800 is being used as a coupling facility, the coupling facility *must* be moved to a z990 or z890, or later, *before* introducing a z10 EC for a z/OS image or ICF.

The ICB connector is different from those on previous machines, requiring new cables and connectors to be installed on downlevel machines to connect them to z10 EC through ICB.

Note: The InfiniBand link data rates of 6 GBps, 3 GBps, 2.5 Gbps, or 5 Gbps do not represent the performance of the link. The actual performance depends on many factors including latency through the adapters, cable lengths, and the type of workload.

When comparing coupling links data rates, InfiniBand (12x IB-SDR or 12x IB-DDR) can be higher than ICB-4, and InfiniBand (1x IB-SDR or 1x IB-DDR) can be higher than that of ISC-3. However, with InfiniBand, the service times of coupling operations are greater, and the actual throughput can be less than with ICB-4 links or ISC-3 links.

For a more specific explanation of when to continue using the current ICB or ISC-3 technology versus migrating to InfiniBand coupling links, see the *Coupling Facility Configuration Options* white paper, available from:

<http://www.ibm.com/systems/z/advantages/pso/whitepaper.html>

Coupling links and Server Time Protocol

All external coupling links can be used to pass time synchronization signals by using Server Time Protocol (STP). Server Time Protocol is a message-based protocol in which STP 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 PSIFB or ICB-4 links, can meet more stringent synchronization requirements than servers exchanging messages over long ISC-3 links (distances up to 100 km). This advantage is an enhancement over the IBM Sysplex Timer implementation, which does not scale with distance.
- ▶ Coupling links also provide the connectivity necessary in a Parallel Sysplex. Therefore, there is a potential benefit of minimizing the number of cross-site links required in a multi-site Parallel Sysplex.

Between any two servers that are intended to exchange STP messages, we recommend that each server be configured so that at least two coupling links exist for communication between the servers. This configuration prevents the loss of one link, causing the loss of STP communication between the servers. If a server does not have a CF logical partition, timing-only links can be used to provide STP connectivity. STP is the System z technology that is replacing the Sysplex Timer function.

The z10 EC supports attachment to the IBM Sysplex Timer through the external time reference (ETR) feature. Timing networks can be implemented using ETR only, mixed ETR and STP, or STP-only Coordinated Timing Network (CTN) in a Parallel Sysplex configuration.

Note: System z10 servers will be the last family of servers to support the Sysplex Timer.

For Sysplex Timer connectivity and configuration information see *IBM System z Connectivity Handbook*, SG24-5444.

For STP configuration information, see *Server Time Protocol Planning Guide*, SG24-7280, and *Server Time Protocol Implementation Guide*, SG24-7281.

4.7.2 External time reference

The external time reference (ETR) is a standard feature providing two ETR cards plugged in the CEC cage. The ETR cards provide attachment to the Sysplex Timer. Each ETR card should connect to a different Sysplex Timer in an expanded availability configuration. Each feature has a single port supporting an MT-RJ fiber optic connector to provide the capability to attach to a Sysplex Timer Unit. The two ETR cards are supported in two CEC cage card slots on top of the books and provide attachment to a Sysplex Timer.

The Sysplex Timer provides the synchronization for the time-of-day (TOD) clocks of multiple servers, and thereby allows events started by different servers to be properly sequenced in time. When multiple servers update the same database and database reconstruction is necessary, all updates are required to be time stamped in proper sequence.

The ETR Network ID of the attached Sysplex Timer Network must be manually set in the Support Element (SE) at installation. The SE checks that the ETR Network ID being received in the timing signals through the ETR ports matches the ETR Network ID manually set in the server's SE.

The port cards support concurrent maintenance. The ETR card port has a small form factor optical transceiver that supports an MT-RJ connector only.

The ETR card does not support a multimode fiber optic cable terminated with an ESCON Duplex connector as on the Sysplex Timer.

Multimode ESCON Duplex jumper cables (62.5 μm) can be reused to connect to the ETR card. This is done by installing an MT-RJ/ESCON Conversion kit between the ETR card MT-RJ port and the ESCON Duplex jumper cable.

Fiber optic conversion kits and mode conditioning patch (MCP) cables are not orderable as features. Fiber optic cables, cable planning, labeling, and installation are all customer responsibilities for new z10 EC installations and upgrades.

IBM Facilities Cabling Services - fiber transport system offers a total cable solution service to help with cable ordering requirements, and is highly recommended.

4.7.3 Cryptographic feature

Cryptographic functions are provided by CP Assist for Cryptographic Function (CPACF) and the Crypto Express2 or Crypto Express3 features. Feature code (FC) 3863 is required to enable CPACF functions.

Crypto Express2 feature (FC 0863)

Crypto Express2 is an optional feature. It cannot be ordered on a new server, but it can be carried forward from a server that is being upgraded to a z10 EC.

Each Crypto Express2 feature holds two PCI-X cryptographic adapters that can be configured as coprocessors or accelerators. Either of the adapters can be configured by the installation as a coprocessor or accelerator.

Each Crypto Express2 feature occupies one I/O slot in an I/O cage and has no CHPIDs assigned, but uses two PCHIDS.

Cryptographic functions are described in Chapter 6, "Cryptography" on page 171.

Crypto Express3 feature (FC 0864)

Crypto Express3 is an optional feature. On the initial order, the minimum of two features are installed. After the initial configuration, the number of features increase one at a time up to a maximum of eight.

Each Crypto Express3 feature holds two PCI Express cryptographic adapters that can be configured as coprocessors or accelerators. Either of the adapters can be configured by the installation as a coprocessor or accelerator.

Each Crypto Express3 feature occupies one I/O slot in an I/O cage and has no CHPIDs assigned, but uses two PCHIDS.

Cryptographic functions are described in Chapter 6, "Cryptography" on page 171.



Channel subsystem

This chapter describes the concept of multiple channel subsystems. It also discusses the technology, terminology, and implementation aspects of the channel subsystem.

This chapter discusses the following topics:

- ▶ 5.1, “Channel subsystem” on page 158
- ▶ 5.2, “I/O Configuration management” on page 169
- ▶ 5.3, “System-initiated CHPID reconfiguration” on page 170
- ▶ 5.4, “Multipath initial program load” on page 170

5.1 Channel subsystem

The role of the channel subsystem (CSS) is to control communication of internal and external channels to control units and devices. The configuration definitions of the CSS define the operating environment for the correct execution of all system I/O operations. The CSS provides the server communications to external devices through channel connections. The channels permit transfer of data between main storage and I/O devices or other servers under the control of a channel program. The CSS allows channel I/O operations to continue independently of other operations within the central processors (CPs).

The building blocks that make up a channel subsystem are shown in Figure 5-1.

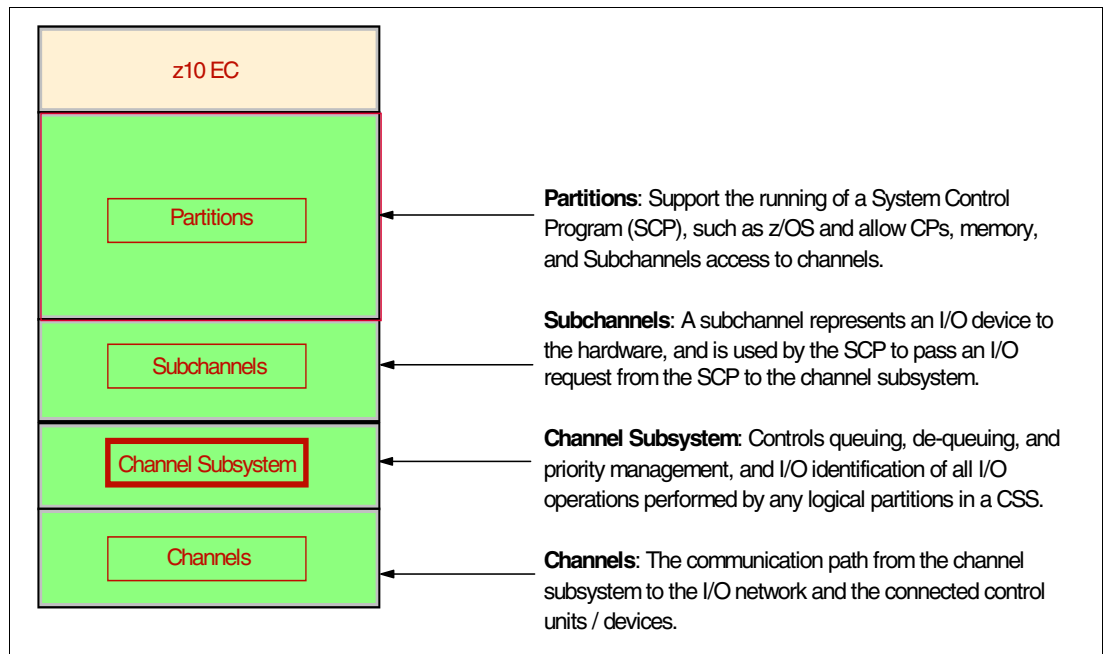


Figure 5-1 Channel subsystem overview

The structure provides up to four channel subsystems (Figure 5-2). Each CSS has from one to 256 CHPIDs, and may be configured with up to 15 logical partitions that relate to that particular channel subsystem. CSSs are numbered from 0 to 3, and are sometimes referred to as the CSS image ID (CSSID 0, 1, 2, and 3).

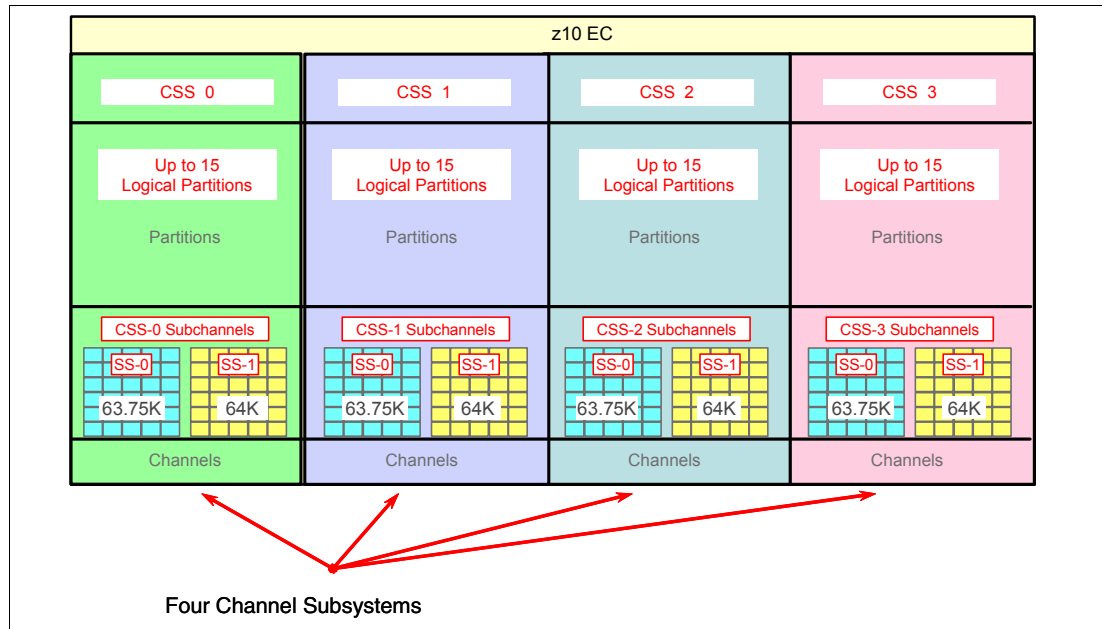


Figure 5-2 Four channel subsystems

The z10 EC provides for up to four channel subsystems, 1024 CHPIDs, and up to 60 logical partitions for the entire system.

The health checker function in z/OS V1.10 introduces a health check in the I/O Supervisor that can help system administrators identify single points of failure in the I/O configuration.

5.1.1 CSS elements

A CSS can have up to 256 channel paths. A channel path is a single interface between a server and one or more control units. Commands and data are sent across a channel path to perform I/O requests. The entities that encompass the CSS are described in this section.

Subchannels

A subchannel provides the logical representation of a device to the program and contains the information required for sustaining a single I/O operation. A subchannel is assigned for each device defined to the logical partition.

Note that multiple subchannel sets (MSS) are available to increase addressability. Two subchannel sets are supported on System z10; subchannel set 0 can have up to 63.75 K subchannels, and subchannel set 1 can have up to 64 K subchannels. See 5.1.6, "Multiple subchannel sets" on page 163 for more information.

Channel path identifier

A channel path identifier (CHPID) is a value assigned to each channel path of the system that uniquely identifies that path. A total of 256 CHPIDs are supported by the CSS.

The channel subsystem communicates with I/O devices by means of channel paths between the channel subsystem and control units. On System z a CHPID number is assigned to a physical location (slot/port) by the user through HCD or IOCP.

Control units

A control unit provides the logical capabilities necessary to operate and control an I/O device and adapts the characteristics of each device so that it can respond to the standard form of control provided by the CSS. A control unit may be housed separately, or it may be physically and logically integrated with the I/O device, the channel subsystem, or within the 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 channel path.

5.1.2 Multiple CSSs concept

The multiple channel subsystems concept provides the ability to define more than 256 CHPIDs in System z servers. The z10 EC supports four CSSs. The design of System z servers offers considerable processing power, memory sizes, and I/O connectivity. In support of the larger I/O capability, the CSS concept has been scaled up correspondingly to provide relief for the number of supported logical partitions, channels, and devices available to the server.

Each CSS may have from 1 to 256 channels and be configured with 1 to 15 logical partitions. Therefore, four CSSs support a maximum of 60 logical partitions. CSSs are numbered from 0 to 3 and are sometimes referred to as the CSS image ID (CSSID 0, 1, 2 or 3).

5.1.3 Multiple CSSs structure

The structure of the multiple CSSs provides channel connectivity to the defined logical partitions in a manner that is transparent to subsystems and application programs.

The System z servers provide the ability to define more than 256 CHPIDs in the system through the multiple CSSs. CSS defines CHPIDS, control units, subchannels, and so on, enabling the definition of a balanced configuration for the processor and I/O capabilities.

For ease of management, we strongly recommend that the Hardware Configuration Definitions (HCDs) be used to build and control the I/O configuration definitions. HCD support for multiple channel subsystems is available with z/VM and z/OS. HCD provides the capability to make both dynamic hardware and software I/O configuration changes.

No logical partitions can exist without at least one defined CSS. Logical partitions are defined to a CSS, not to a server. A logical partition is associated with one CSS only. CHPID numbers are unique within a CSS and range from 00 to FF. However, the same CHPID number can be reused within any other CSS.

All channel subsystem images (CSS images) are defined within a single I/O configuration data set (IOCDS). The IOCDS is loaded and initialized into the hardware system area during power-on reset.

The HSA is pre-allocated in memory with a size of 16 GB. This eliminates planning for HSA and pre-planning for HSA expansion, because HCD/IOCP always reserves the following items by the IOCDS process:

- ▶ Four CSSs
- ▶ Fifteen LPARs in each CSS
- ▶ Subchannel set 0 with 63.75 K devices in each CSS
- ▶ Subchannel set 1 with 64 K devices in each CSS

All these are designed to be activated and used with dynamic I/O changes.

Figure 5-3 shows a logical view of the relationships. Note that each CSS supports up to 15 logical partitions. System-wide, a total of up to 60 logical partitions are supported.

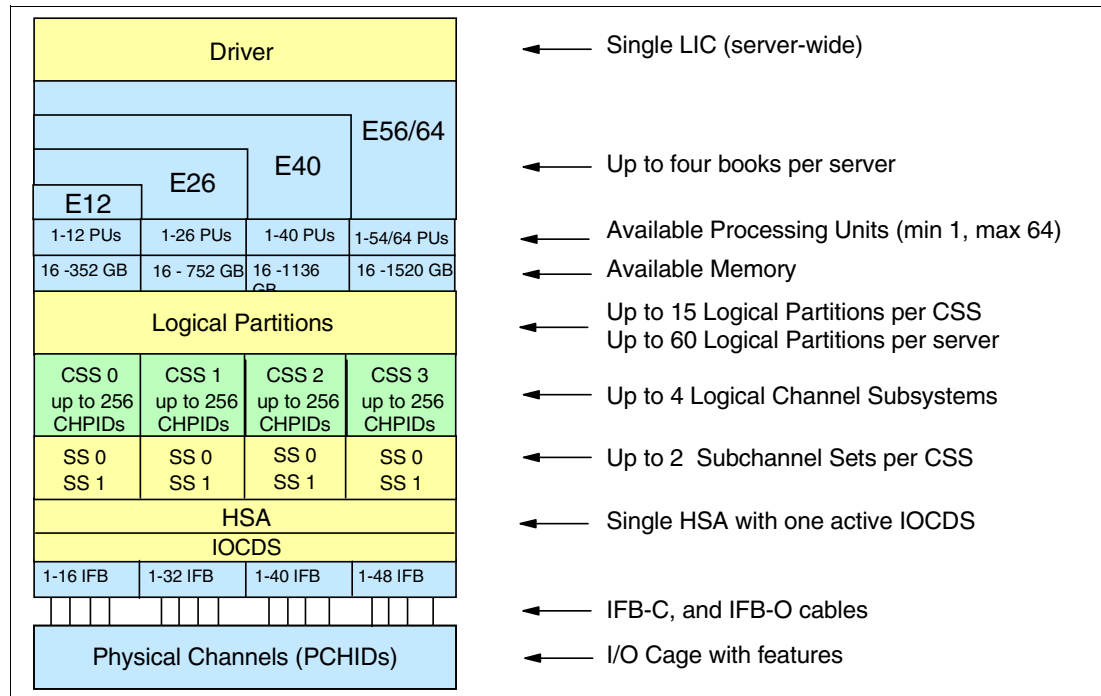


Figure 5-3 Logical view of z10 EC models, CSSs, IOCDS, and HSA

Note: The HSA can be moved from one book to a different book in an enhanced availability configuration as part of a concurrent book repair action.

The channel definitions of a CSS are not bound to a single book. A CSS can define resources that are physically connected to any InfiniBand cable of any book in a multibook CPC.

5.1.4 Logical partition name and identification

A logical partition is identified through its name, its identifier, and its multiple image facility (MIF) image ID (MIF ID).

The logical partition name is user defined through HCD or the IOCP and is the partition name in the RESOURCE statement in the configuration definitions. Each name must be unique across the CPC.

The logical partition identifier is a number in the range of 00 - 3F assigned by the user on the image profile through the Support Element (SE) or the Hardware Management Console (HMC). It is unique across the CPC and may also be referred to as the user logical partition ID (UPID).

The MIF ID is a number that is defined through the Hardware Configuration Dialog (HCD) or directly through the IOCP. It is specified in the RESOURCE statement in the configuration definitions. It is in the range of 1 - F and is unique within a CSS. However, because of multiple CSSs, the MIF ID is not unique within the CPC.

The multiple image facility enables resource sharing across logical partitions within a single CSS or across the multiple CSSs. When a channel resource is shared across logical partitions in multiple CSSs, this is known as spanning. Multiple CSSs may specify the same MIF image ID. However, the combination CSSID.MIFID is unique across the CPC.

Summary of identifiers

Figure 5-4 summarizes the identifiers and how they are defined.

CSS0			CSS1			CSS2	CSS3		Specified in HCD / IOCP
Logical	Partition	Name	Logical	Partition	Name	Log Part Name	Logical Partition Name		Specified in HCD / IOCP
TST1	PROD1	PROD2	TST2	PROD3	PROD4	TST3	TST4	PROD5	
Logical Partition ID			Logical Partition ID			Log Part ID	Logical Partition ID		Specified in HMC Image Profile
02	04	0A	14	16	1D	22	35	3A	
MIF ID 2	MIF ID 4	MIF ID A	MIF ID 4	MIF ID 6	MIF ID D	MIF ID 2	MIF ID 5	MIF ID A	Specified in HCD / IOCP

Figure 5-4 CSS, logical partition, and identifiers example

We recommend establishing a naming convention for the logical partition identifiers. As shown in Figure 5-4, you could use the CSS number concatenated to the MIF ID, which means that logical partition ID 3A is in CSS 3 with MIF IDA. This fits within the allowed range of logical partition IDs and conveys useful information to the user.

Dynamic addition or deletion of a logical partition name

All undefined logical partitions are reserved partitions. They are automatically predefined in the HSA with a name placeholder and a MIF ID.

5.1.5 Physical channel ID

A physical channel ID (PCHID) reflects the physical identifier of a channel-type interface. A PCHID number is based on the I/O cage location, the channel feature slot number, and the port number of the channel feature. A CHPID does not directly correspond to a hardware channel port, and may be arbitrarily assigned. A hardware channel is identified by a PCHID.

Within a single channel subsystem, 256 CHPIDs can be addressed. That gives a maximum of 1,024 CHPIDs when four CSSs are defined. Each CHPID number is associated with a single channel. The physical channel, which uniquely identifies a connector jack on a channel feature, is known by its PCHID number.

PCHIDs identify the physical ports on cards located in I/O cages and follow the numbering scheme shown in Table 5-1.

Table 5-1 PCHIDs numbering scheme

Cage	Front PCHID ##	Rear PCHID ##
I/O cage 1	100-1FF	200-2BF
I/O cage 2	300-3FF	400-4BF
I/O cage 3	500-5FF	600-6BF
CEC cage	000-03F reserved for ICB-4s	

CHPIDs are not pre-assigned. The installation is responsible to assign the CHPID numbers through the use of the CHPID Mapping Tool (CMT) or HCD/IOCP. Assigning CHPIDs means that a CHPID number is associated with a physical channel port location and a CSS. The CHPID number range is still from 00 - FF and must be unique within a CSS. Any non-internal CHPID that is not defined with a PCHID can fail validation when an attempt is made to build a production IODF or an IOCDS.

5.1.6 Multiple subchannel sets

Do not confuse the multiple subchannel set (MSS) functionality with multiple channel subsystems.

In most cases, a subchannel represents an addressable device. For example, a disk control unit with 30 drives uses 30 subchannels (for base addresses), and so forth. An addressable device is associated with a device number and the device number is commonly (but incorrectly) known as the device address.

Subchannel numbers (including their implied path information to a device) are limited to four hexadecimal digits by architecture (0x0000 to 0xFFFF). Four hexadecimal digits provide 64 K addresses, known as a *set*. IBM has reserved 256 subchannels, leaving over 63 K subchannels for general use¹.

Again, addresses, device numbers, and subchannels are often used as synonyms, although this is not technically correct. We may hear that there is *a maximum of 63 K addresses* or *a maximum of 63 K device numbers*.

The processor architecture allows for *sets* of subchannels (addresses), with a current implementation of two sets. Each set provides 64 K addresses. Subchannel set 0, the first set, still reserves 256 subchannels for IBM use. Subchannel set 1 provides a full range of 64 K subchannels. In principle, subchannels in either set could be used for any device-addressing purpose. However, the current implementation in z/OS restricts subchannel set 1 to disk *alias* subchannels. Subchannel set 0 may be used for base addresses and for alias addresses.

¹ The number of reserved subchannels is 256. We abbreviate this to 63 K in this discussion to easily differentiate it from the 64 K subchannels available in subchannel set 1.

Figure 5-5 summarizes the multiple subchannel sets.

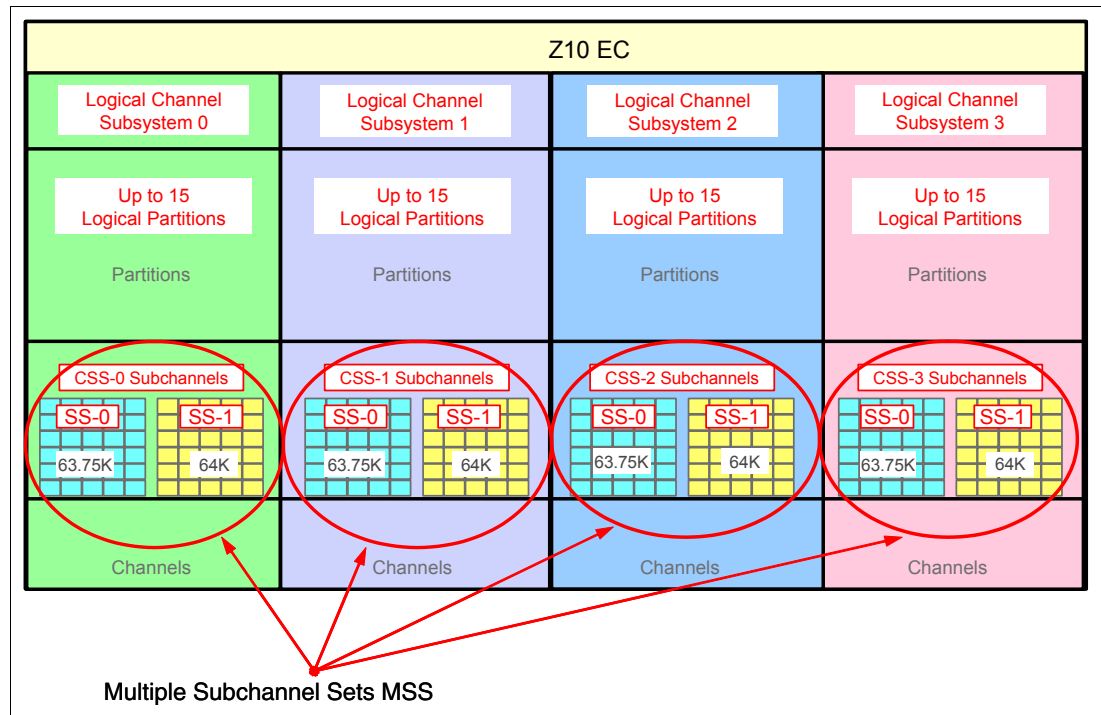


Figure 5-5 Multiple subchannel sets

Correspondence is not required between addresses in the two sets, and is not required between the device numbers used in the two subchannel sets.

The additional subchannel set, in effect, adds an extra high-order digit (either 0 or 1) to existing device numbers. For example, we might think of an address as 08000 (subchannel set 0) or 18000 (subchannel set 1). Add a digit is not done in system code or in messages because of the architectural requirement for four-digit addresses (device numbers or subchannels). However, some messages do contain the subchannel set number, and you can mentally use that as a high-order digit for device numbers. There should be few requirements for this because subchannel set 1 is used only for alias addresses and users, through JCL, messages, or programs that rarely refer directly to an *alias* address.

Moving the alias devices into the second subchannel set creates additional space for device number growth. The appropriate subchannel set number must be included in IOCP definitions or in the HCD definitions that produce the IOCDs. The subchannel set number defaults to zero. Channel sets are exploited by the Peer-to-Peer Remote Copy (PPRC) function by the ability to have the PPRC primary devices defined in channel set 0, while secondary devices can be defined in channel set 1, thus providing more connectivity through channel set 0.

Parallel access volume (PAV) support enables a single System z server to simultaneously process multiple I/O operations to the same logical volume, which can help to significantly reduce device queue delays. Dynamic PAV allows the dynamic assignment of aliases to volumes to be under WLM controls.

With the availability of HyperPAV, the requirement for PAV devices is greatly reduced. HyperPAV allows an alias address to be used to access any base on the same control unit image per I/O base. It also allows different HyperPAV hosts to use one alias to access different bases, which reduces the number of alias addresses required. HyperPAV is designed to enable applications to achieve equal or better performance than possible with the

original PAV feature alone, while also using the same or fewer z/OS resources. HyperPAV is an optional feature on the IBM DS8000 series.

To further reduce the complexity of managing large I/O configurations System z introduces Extended Address Volumes (EAV). EAV is designed to build very large disk volumes using virtualization technology. By being able to extend the disk volume size a customer may potentially need fewer volumes to hold his data, therefore making systems management and data management less complex.

The 63.75 K subchannels

On the z10 EC, 256 subchannels are reserved for IBM use in subchannel set 0. No subchannels are reserved in subchannel set 1. The informal name, *63.75 K subchannel*, represents the following equation:

$$(63 \times 1024) + (0.75 \times 1024) = 65280$$

The display ios,config command

The `display ios,config(a11)` command, shown in Figure 5-6, includes information about the MSSs.

```
D IOS,CONFIG(ALL)
IOS506I 18.21.37 I/O CONFIG DATA 610
ACTIVE IODF DATA SET = SYS6.IODF45
CONFIGURATION ID = TEST2097 EDT ID = 01
TOKEN: PROCESSOR DATE      TIME      DESCRIPTION
SOURCE: SCZP201 08-03-04 09:20:58 SYS6      IODF45
ACTIVE CSS: 0 SUBCHANNEL SETS CONFIGURED: 0, 1
CHANNEL MEASUREMENT BLOCK FACILITY IS ACTIVE
HARDWARE SYSTEM AREA AVAILABLE FOR CONFIGURATION CHANGES
PHYSICAL CONTROL UNITS          8131
CSS 0 - LOGICAL CONTROL UNITS    4037
SS 0 SUBCHANNELS                 62790
SS 1 SUBCHANNELS                 61117
CSS 1 - LOGICAL CONTROL UNITS    4033
SS 0 SUBCHANNELS                 62774
SS 1 SUBCHANNELS                 61117
CSS 2 - LOGICAL CONTROL UNITS    4088
SS 0 SUBCHANNELS                 65280
SS 1 SUBCHANNELS                 65535
CSS 3 - LOGICAL CONTROL UNITS    4088
SS 0 SUBCHANNELS                 65280
SS 1 SUBCHANNELS                 65535
ELIGIBLE DEVICE TABLE LATCH COUNTS
0 OUTSTANDING BINDS ON PRIMARY EDT
```

Figure 5-6 Display ios,config(all) with MSS

5.1.7 Multiple CSS construct

A pictorial view of a z10 EC with multiple CSSs defined is shown in Figure 5-7. In this example, two channel subsystems are defined (CSS0 and CSS1). Each CSS has three logical partitions with their associated MIF image identifiers.

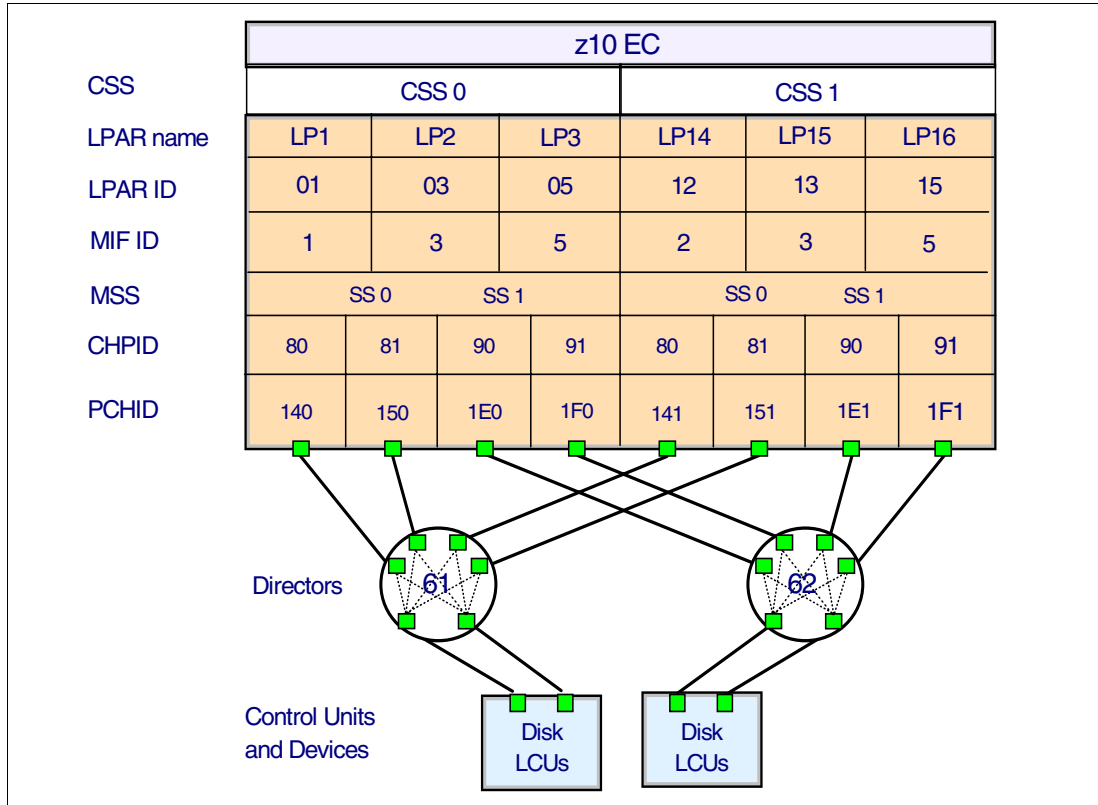


Figure 5-7 z10 EC CSS connectivity

In each CSS, the CHPIDs are shared across all logical partitions. The CHPIDs in each CSS can be mapped to their designated PCHIDs using the CHPID Mapping Tool (CMT) or manually using HCD or IOCP. The output of the CMT is used as input to HCD or the IOCP to establish the CHPID to PCHID assignments.

5.1.8 Adapter ID

The adapter ID (AID) number is used for assigning a CHPID to a port through HCD/IOCP for Parallel Sysplex over InfiniBand (PSIFB) coupling links.

The AID is a number from 00 through 1F. If the fanout is moved to another slot, the AID changes for that specific fanout and it might be necessary to readjust the IOCDs.

The AID is bound to the serial number of the fanout. If the fanout is moved, the AID moves with it. No IOCDs update is required if adapters are moved to a new physical location.

Table 5-22 shows the assigned AID numbers for a new build z10 EC.

Table 5-2 Fanout AID numbers

Fanout location	Fourth book	First book	Third book	Second book
D1	00	08	10	18
D2	01	09	11	19
D3	N/A	N/A	N/A	N/A
D4	N/A	N/A	N/A	N/A
D5	02	0A	12	1A
D6	03	0B	13	1B
D7	04	0C	14	1C
D8	05	0D	15	1D
D9	06	0E	16	1E
DA	07	0F	17	1F

The AIDs are shown in the PCHID report provided by an IBM representative for new build System z10 servers or for upgrades. Part of a PCHID report is shown in Example 5-1.

Example 5-1 AID assignment in a PCHID report

```

CHPIDSTART
19756694                PCHID REPORT
Machine: 2097-E26  SNxxxxxxx
-----
Source          Cage  Slot  F/C   PCHID/Ports or AID          Comment
06/D6          A25B D606 0163  AID=0B
15/D6          A25B D615 0163  AID=1B

```

For more information regarding PSIFB coupling link features, see “InfiniBand coupling links (FC 0163)” on page 150.

5.1.9 Channel spanning

Channel spanning extends the MIF concept of sharing channels across logical partitions to sharing channels across logical partitions *and* channel subsystems.

Spanning is the ability for a physical channel (PCHID) to be mapped to CHPIDs defined in multiple channel subsystems. When defined that way, the channels can be transparently shared by any or all of the configured logical partitions, regardless of the channel subsystem to which the logical partition is configured.

A channel is considered a spanned channel if the same CHPID number in different CSSs is assigned to the same PCHID in IOCP, or is defined as *spanned* in HCD.

In the case of internal channels (for example, IC links and HiperSockets), the same applies, but with no PCHID association. They are defined with the same CHPID number in multiple CSSs.

CHPIDs that span CSSs reduce the total number of channels available. The total is reduced, because no CSS can have more than 256 CHPIDs. For a z10 EC with two CSSs defined, a total of 512 CHPIDs is supported. If all CHPIDs are spanned across the two CSSs, then only 256 channels are supported. For a z10 EC with four CSSs defined, a total of 1024 CHPIDs is supported. If all CHPIDs are spanned across the four CSSs, then only 256 channels can be supported.

Channel spanning is supported for internal links (HiperSockets and Internal Coupling (IC) links) and for certain external links (FICON Express8, FICON Express4, and FICON Express2 channels, OSA-Express2, OSA-Express3, and Coupling Links).

Note: Spanning of ESCON channels and FICON converter (FCV) channels is not supported.

In Figure 5-8, CHPID 04 is spanned to CSS0 and CSS1. Because it is not an external channel link, no PCHID is assigned. CHPID 06 is an external spanned channel and has a PCHID assigned.

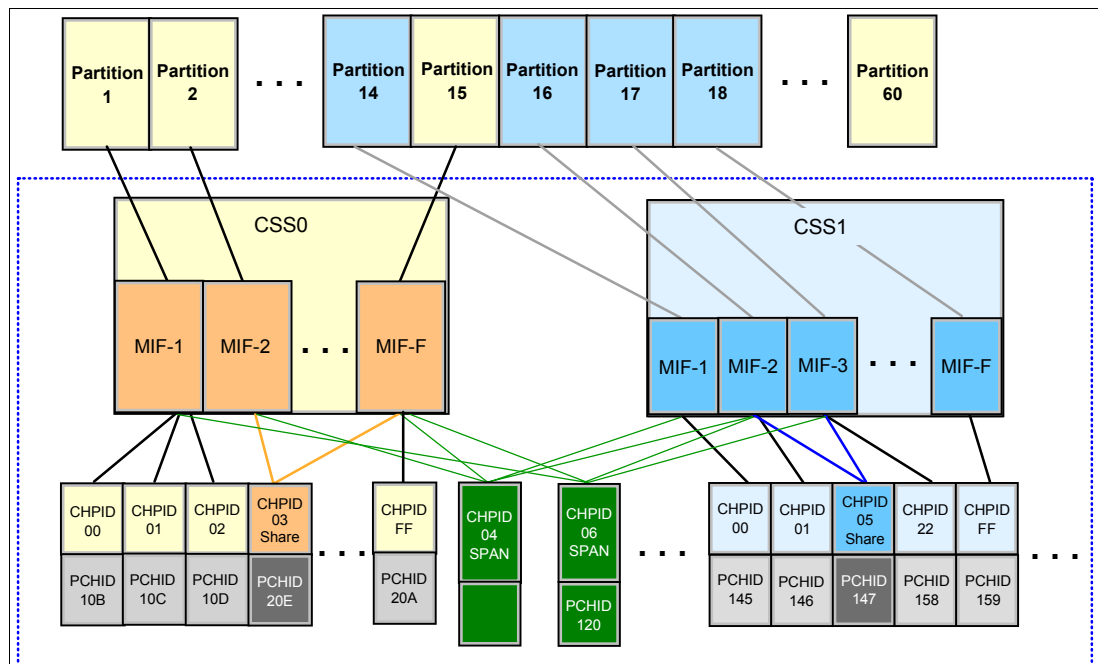


Figure 5-8 z10 EC CSS: two channel subsystems with channel spanning

5.1.10 Summary of CSS-related numbers

Table 5-3 shows CSS-related information in terms of maximum values for devices, subchannels, logical partitions, and CHPIDs.

Table 5-3 z10 EC CSS overview

Setting	z10 EC
Maximum number of CSSs	4
Maximum number of CHPIDs	1024
Maximum number of LPARs supported per CSS	15
Maximum number of LPARs supported per system	60
Maximum number of HSA subchannels	7665 K (127.75 K per partition x 60 partitions)
Maximum number of devices	511 K (4 CSSs x 127.75 K devices)
Maximum number of CHPIDs per CSS	256
Maximum number of CHPIDs per logical partition	256
Maximum number of devices/subchannels per logical partition	127.75 K

5.2 I/O Configuration management

Tools are provided to help maintain and optimize the I/O configuration:

- ▶ IBM Configurator for e-business (eConfig)

The eConfig tool is available to your IBM representative. It is used to configure new configurations or upgrades of an existing configuration, and maintains installed features of those configurations. Reports produced by eConfig are helpful in understanding the changes being made for a system upgrade and what the final configuration will look like.

- ▶ Hardware Configuration Dialog (HCD)

HCD supplies an interactive dialog to generate the I/O definition file (IODF) and subsequently the input/output configuration data set (IOCDS). We strongly recommend that HCD or HCM be used to generate the I/O configuration, as opposed to writing IOCP statements. The validation checking that HCD performs as data is entered helps minimize the risk of errors before the I/O configuration is implemented.

The HCD version shipped in z/OS V1R7 generates an IODF with a Version 5 format. HCD provides a conversion function to upgrade from a V4 to a V5 IODF.

When accessing an IODF in V4 format from a z/OS 1.7 system, the IODF format is converted to an in-storage IODF V5, and message CBDG549 is issued to inform the user that a back-level IODF is being accessed. However, as long as no migration is requested, the V5 IODF format will not be saved and the copy on disk will remain in the V4 IODF format. After migration to a Version 5 IODF, only z/OS V1R7 or later can make updates to this IODF version.

- ▶ CHPID Mapping Tool (CMT)

The CHPID Mapping Tool provides a mechanism to map CHPIDs onto PCHIDs as required. Additional enhancements have been built into the CMT to cater to the requirements of the z10 EC. It provides the best availability recommendations for the

installed features and defined configuration. CMT is a workstation-based tool available for download from IBM Resource Link site:

<http://www.ibm.com/servers/resourceLink>

5.3 System-initiated CHPID reconfiguration

The system-initiated CHPID reconfiguration function is designed to reduce the duration of a repair action and minimize operator interaction when an ESCON or FICON channel, an OSA port, or an ISC-3 link is shared across logical partitions on an z10 EC server. When an I/O card is to be replaced for a repair, it usually has some failed channels and some that are still functioning.

To remove the card, all channels must be configured offline from all logical partitions sharing those channels. Without system-initiated CHPID reconfiguration, this means that the CE must contact the operators of each affected logical partition and have them set the channels offline, and then after the repair, contact them again to configure the channels back online.

With system-initiated CHPID reconfiguration support, the Support Element sends a signal to the IOP that a channel needs to be configured offline. The IOP determines all the logical partitions sharing that channel and sends an alert to the operating systems in those logical partitions. The operating system then configures the channel offline without any operator intervention. This cycle is repeated for each channel on the card. When the card is replaced, the Support Element sends another signal to the IOP for each channel. This time, the IOP alerts the operating system that the channel should be configured back online. This process minimizes operator interaction to configure channels offline and online.

System-initiated CHPID reconfiguration is supported by z/OS.

5.4 Multipath initial program load

Multipath initial program load (IPL) helps increase availability and helps eliminate manual problem determination during IPL execution. This happens by allowing IPL to complete, if possible, using alternate paths when executing an IPL from a device connected through ESCON and FICON channels. If an error occurs, an alternate path is selected.

Multipath IPL is applicable to ESCON channels (CHPID type CNC) and to FICON channels (CHPID type FC). z/OS supports multipath IPL.



Cryptography

This chapter describes the hardware cryptographic functions available on the z10 EC. Similar to the System z9 and earlier generations, the Cryptographic Assist Architecture (CAA), together with the CP Assist for Cryptographic Function (CPACF), offer a balanced use of resources and unmatched scalability.

The z10 EC includes both standard cryptographic hardware and optional cryptographic features for flexibility and growth capability. IBM has a long history of providing hardware cryptographic solutions, from the development of Data Encryption Standard (DES) in the 1970s to have the Crypto Express tamper-resistant features designed to meet the U.S. Government's highest security rating FIPS 140-2 Level 4¹.

The cryptographic functions include the full range of cryptographic operations necessary for e-business, e-commerce, and financial institution applications. Custom cryptographic functions can also be added to the set of functions that the z10 EC offers.

Today, e-business applications increasingly rely on cryptographic techniques to provide the confidentiality and authentication required in this environment. Secure Sockets Layer/Transport Layer Security (SSL/TLS) technology is a key technology for conducting secure e-commerce using Web servers, and it is in use by a rapidly increasing number of e-business applications, demanding new levels of security, performance, and scalability.

This chapter discusses the following topics:

- ▶ 6.1, “Cryptographic synchronous functions” on page 172
- ▶ 6.2, “Cryptographic asynchronous functions” on page 173
- ▶ 6.3, “CP Assist for Cryptographic Function” on page 177
- ▶ 6.4, “Crypto Express2” on page 178
- ▶ 6.5, “Crypto Express3” on page 182
- ▶ 6.6, “TKE workstation feature” on page 184
- ▶ 6.7, “Cryptographic functions comparison” on page 186
- ▶ 6.8, “Software support” on page 187

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

6.1 Cryptographic synchronous functions

Cryptographic synchronous functions are provided by the CP Assist for Cryptographic Function (CPACF).

The z10 EC hardware includes the implementation of algorithms as hardware synchronous operations, which means holding the PU processing of the instruction flow until the operation has completed. The secure key functions are:

- ▶ Data encryption and decryption algorithms
 - Data Encryption Standard (DES), which includes:
 - Double-key DES (double DES)
 - Triple-key DES (triple DES)
 - Advanced Encryption Standard (AES) with secure encrypted 128-bit, 192-bit, and 256-bit keys (secure key AES is exclusive to System z10).
- ▶ Hashing algorithms, such as SHA-1, and SHA-2 support for SHA-224, SHA-256, SHA-384, and SHA-512
- ▶ Message authentication code (MAC)
 - Single-key MAC
 - Double-key MAC
- ▶ Pseudo Random number generation (PRNG)
- ▶ Random number generation long (RNGL) with 8 bytes to 8096 bytes
- ▶ Random number generation (RNG) with up to 4096-bit key RSA support

Note: Keys must be provided in clear form only.

SHA-1, and SHA-2 support for SHA-224, SHA-256, SHA-384, and SHA-512 are shipped enabled on all servers and do not require the CPACF enablement feature. The CPACF functions are supported by z/OS, z/VM, z/VSE, and Linux on System z.

An enhancement to CPACF is designed to facilitate the continued privacy of cryptographic key material when used for data encryption. CPACF, using key wrapping, ensures that key material is not visible to applications or operating systems during encryption operations. Figure 6-1 shows this function.

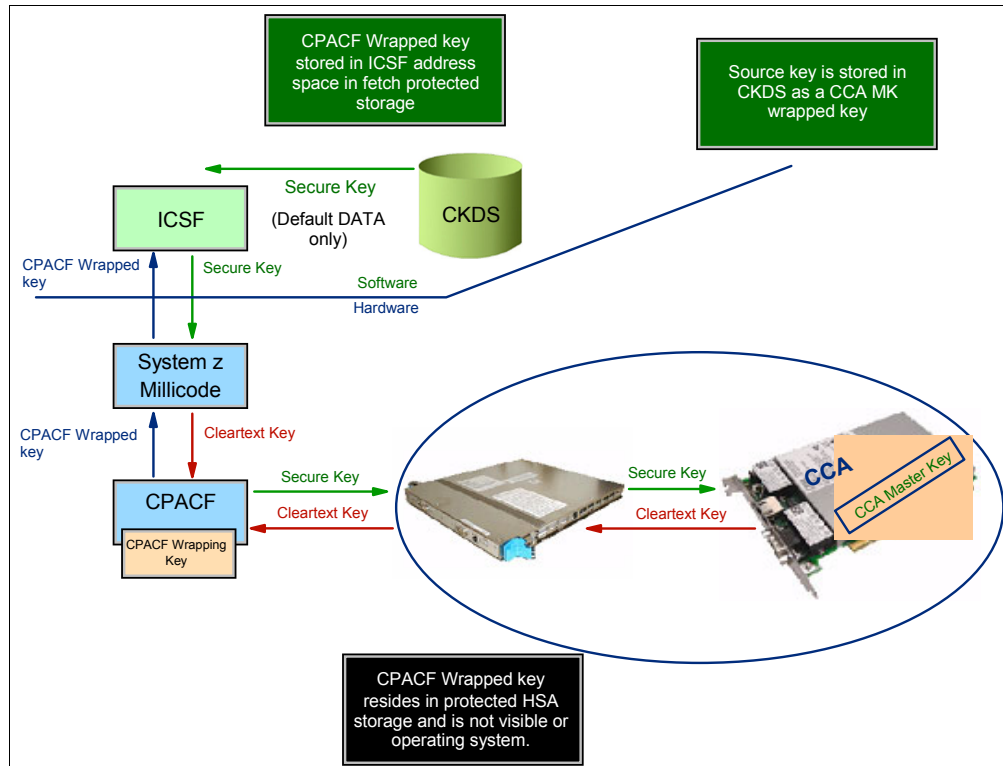


Figure 6-1 CPACF key wrapping

Protected key CPACF is designed to provide substantial throughput improvements for large volume data encryption as well as low latency for encryption of small blocks of data. Further, changes to the information management tool, IBM Encryption Tool for IMS and DB2 Databases, improves performance for protected key applications.

6.2 Cryptographic asynchronous functions

Cryptographic asynchronous functions are provided by the PCI-X and PCI Express cryptographic adapters.

6.2.1 Secure key functions

The following secure key functions are provided as cryptographic asynchronous functions. System internal messages are passed to the cryptographic coprocessors to initiate the operation, then messages are passed back from the coprocessors to signal completion of the operation.

- ▶ Data encryption and decryption algorithms
 - Data Encryption Standard (DES)
 - Double-key DES (double DES)
 - Triple-key DES (triple DES)
- ▶ DES key generation and distribution
- ▶ PIN generation, verification, and translation functions
- ▶ Pseudorandom number generator (PRNG)
- ▶ Public key algorithm (PKA) facility

Supported operating system commands intended for application programs that use PKA include:

- Importing RSA public-private key pairs in clear and encrypted forms
- Rivest-Shamir-Adelman (RSA), which can provide:
 - Key generation, up to 4,096-bit
 - Signature verification, up to 4,096-bit
 - Import and export of DES keys under an RSA key, up to 4,096-bit
- Public key encryption (PKE)

The PKE service is provided for assisting the SSL/TLS handshake. When used with the Mod_Raised_to Power (MRP) function, PKE is also used to offload compute-intensive portions of the Diffie-Hellman protocol onto the cryptographic adapters.

- Public key decryption (PKD)

PKD supports a zero-pad option for clear RSA private keys. PKD is used as an accelerator for raw RSA private operations, such as those required by the SSL/TLS handshake and digital signature generation. The Zero-Pad option is exploited by Linux to allow the use of cryptographic adapters for improved performance of digital signature generation.

- Derived Unique Key Per Transaction (DUKPT)

The service is provided to write applications that implement the DUKPT algorithms as defined by the ANSI X9.24 standard. DUKPT provides additional security for point-of-sale transactions that are standard in the retail industry. DUKPT algorithms are supported on the Crypto Express2 feature coprocessor for triple DES with double-length keys.

- Europay Mastercard VISA (EMV) 2000 standard

Applications may be written to comply with the EMV 2000 standard for financial transactions between heterogeneous hardware and software. Support for EMV 2000 applies only to the Crypto Express2 feature coprocessor of the z9 EC and z10 EC.

The Crypto Express3 card, a PCI Express cryptographic adapter, offers SHA-2 and RSA functions similar to those functions offered in the CPACF. This is in addition to the functions mentioned above.

6.2.2 Other key functions

Other key functions of the Crypto Express features serve to enhance the security of public and private key encryption processing:

- ▶ Remote loading of initial ATM keys

This function provides the ability to remotely load the initial ATM keys. Remote-key loading refers to the process of loading DES keys to ATM from a central administrative site without requiring someone to manually load the DES keys on each machine. The process uses ICSF callable services along with the Crypto Express features to perform the remote load.

ICSF has added two callable services, Trusted Block Create (CSNDTBC) and Remote Key Export (CSNDRKX). CSNDTBC is a callable service that is used to create a trusted block containing a public key and certain processing rules. The rules define the ways and formats in which keys are generated and exported. CSNDRKX is a callable service that uses the trusted block to generate or export DES keys for local use and for distribution to an ATM or other remote device. The PKA Key Import (CSNDPKI), PKA Key Token Change (CSNDKTC), and Digital Signature Verify (CSFNDFV) callable services support remote key loading.

- ▶ Key exchange with non-CCA cryptographic systems

This function allows for the changing of the operational keys between the remote site and the non-CCA system, such as the asynchronous transfer mode (ATM). IBM Common Cryptographic Architecture (CCA) employs control vectors to control usage of cryptographic keys. Non-CCA systems use other mechanisms, or can use keys that have no associated control information. The key exchange functions added to CCA enhance the ability to exchange keys between CCA systems (and systems that do not use control vectors) by allowing the CCA system owner to define permitted types of key import and export while preventing uncontrolled key exchange that can open the system to an increased threat of attack.

- ▶ ISO 16609 CBC Mode T-DES MAC support

In support of ISO 16609:2004, the cryptographic facilities support the requirements for message authentication, using symmetric techniques. The Crypto Express features provide the ISO 16609 CBC Mode T-DES MAC support. This support is accessible through ICSF callable services. ICSF callable services used to invoke the support are MAC Generate (CSNBMGN) and MAC Verify (CSNVMVR).

- ▶ Retained key support (RSA private keys generated and kept stored within the secure hardware boundary)

- ▶ 4753 Network Security Processor migration support

- ▶ User-Defined Extensions (UDX) support, including:

- Activate UDX requests:
 - Establish owner
 - Relinquish owner
 - Emergency Burn of Segment
 - Remote Burn of Segment
- Import UDX File function
- Reset UDX to IBM default function
- Query UDX Level function

UDX allows the user to add customized operations to a cryptographic processor. User-Defined Extensions to the Common Cryptographic Architecture (CCA) support

customized operations that execute within the Crypto Express features. UDX is supported through an IBM or approved third-party service offering.

More information can be found on the IBM CryptoCards Web site:

<http://www.ibm.com/security/cryptocards>

Under a special contract with IBM, Crypto Express feature customers can define and load custom cryptographic functions. The CryptoCards Web site directs your request to an IBM Global Services location appropriate for your geographic location. A special contract is negotiated between you and IBM Global Services. The contract is for development of the UDX by IBM Global Services according to your specifications and an agreed-upon level of the UDX.

The UDX toolkit for System z with the Crypto Express3 feature is made available on the general availability date for the feature. In addition, there will be a migration path for customers with UDX on a previous feature to migrate their code to the Crypto Express3 feature. A UDX migration is no more disruptive than a normal MCL or ICSF release migration.

6.2.3 Cryptographic feature codes

Table 6-1 lists the cryptographic features available.

Table 6-1 Cryptographic features for System z10

Feature code	Description
3863	CPACF DES or triple DES enablement This feature is a prerequisite to use CPACF (except for SHA-1, SHA-224, SHA-256, SHA-384, and SHA-512) and Crypto Express features.
0863	Crypto Express2 feature A maximum of eight features may be ordered. Each feature contains two PCI-X cryptographic adapters.
0864	Crypto Express3 feature A maximum of eight features may be ordered. Each feature contains two PCI Express cryptographic adapters.
0839 ^a	Trusted Key Entry (TKE) workstation This feature is optional. It offers local and remote key management and supports connectivity to an Ethernet LAN at operating speeds of 10, 100, and 1000 Mbps. This workstation may also be used to control z10 BC, z9 EC, z9 BC, z990, and z890 servers. Up to three features per z10 EC may be installed
0859	Trusted Key Entry workstation, only when carried forward. This feature is not orderable on System z10 EC. If it is installed at the time of an upgrade to the System z10 EC, it may be retained. TKE 5.2 or TKE 5.3 LIC must be used to control the z10 EC and z10 BC. TKE 5.0 and 5.1 workstations (FC 0839) may be used to control z9 EC, z9 BC, z990, and z890 servers.
0854	TKE 5.3 Licensed Internal Code (TKE 5.3 LIC) TKE 5.3 LIC can store trusted key parts on DVD-RAM, paper, and smart card. Use of diskettes is limited to read-only. TKE 5.3 LIC controls coprocessors by using a password protected authority signature key pair in a binary file or on a smart card.
0858	TKE 6.0 Licensed Internal Code (TKE 6.0 LIC) The 6.0 LIC can operate on Crypto Express2 features on z9 machines, similar to the LIC 5.3. The 6.0 LIC can operate on both Crypto Express features on the z10 EC.

Feature code	Description
0885	TKE Smart Card Reader Access to information about the smart card is protected by a personal identification number (PIN)
0884	TKE additional smart cards

a. A next-generation TKE workstation (FC0840) is planned to ship to customers starting January 1, 2010.

TKE includes support for the EAS encryption algorithm with 256-bit master keys and key management functions to load or generate master keys to the cryptographic coprocessor.

If the TKE option is chosen for key management of the cryptographic adapters, a TKE workstation with the TKE 5.3 LIC or later is required.

If the TKE workstation is chosen to operate the Crypto Express features and use certain operational enhancements, a TKE workstation with the TKE 6.0 LIC or later is required. See 6.6, “TKE workstation feature” on page 184 for a more detailed description.

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 the customer’s responsibility to understand and adhere to these regulations when moving, selling, or transferring these products.

6.3 CP Assist for Cryptographic Function

The CP Assist for Cryptographic Function (CPACF) offers a set of symmetric cryptographic functions that enhance the encryption and decryption performance of clear key operations for SSL, VPN, and data-storing applications that do not require FIPS 140-2 level 4 security².

CPACF is designed to facilitate the privacy of cryptographic key material when used for data encryption. CPACF, using key wrapping, ensures that key material is not visible to applications or operating systems during encryption operations

The CPACF feature provides hardware acceleration for DES, triple DES, MAC, AES-128, AES-192, AES-256, SHA-1, SHA-224, SHA-256, SHA-384, and SHA-512 cryptographic services. It provides high-performance hardware encryption, decryption, and hashing support.

The following instructions support the cryptographic assist function:

KMAC	Compute Message Authentic Code.
KM	Cipher Message.
KMC	Cipher Message with Chaining.
KIMD	Compute Intermediate Message Digest.
KLMD	Compute Last Message Digest.
PCKMO	Provide Cryptographic Key Management Operation.

The functions are provided as problem-state z/Architecture instructions, directly available to application programs. When enabled, the CPACF runs at processor speed for every CP, IFL, zIIP, and zAAP.

² Federal Information Processing Standard

The cryptographic architecture includes DES, triple DES, MAC message authentication, AES data encryption and decryption, SHA-1, and SHA-2 support for SHA-224, SHA-256, SHA-384, and SHA-512 hashing.

The functions of the CPACF must be explicitly enabled using FC 3863 by the manufacturing process or at the customer site as an MES installation, except for SHA-1, and SHA-2 support for SHA-224, SHA-256, SHA-384, and SHA-512, which are always enabled.

6.4 Crypto Express2

The Crypto Express2 feature has two Peripheral Component Interconnect eXtended (PCI-X) cryptographic adapters. Each PCI-X cryptographic adapter can be configured as:

- ▶ Cryptographic coprocessor
- ▶ Cryptographic accelerator

Reconfiguration of the PCI-X cryptographic adapter between coprocessor and accelerator mode is also supported for Crypto Express2 features carried forward from z990, z890, z9 EC, and z9 BC systems to the z10 EC, as follows:

- ▶ When the PCI-X cryptographic adapter is configured as a coprocessor, the adapter provides functions (plus several additional functions) equivalent to the *PCICC* card on previous systems with a *higher* level of performance. When the PCI-X adapter is configured as a coprocessor, the adapter also provides functions (plus several additional functions) equivalent to the *PCICA* card on previous systems with the *same* level of performance.
- ▶ When the PCI-X cryptographic adapter is configured as an accelerator, it provides PCICA-equivalent functions with an expected throughput of approximately three times the PCICA throughput on previous systems.

A physical layout of the Crypto Express2 feature is shown in Figure 6-2.

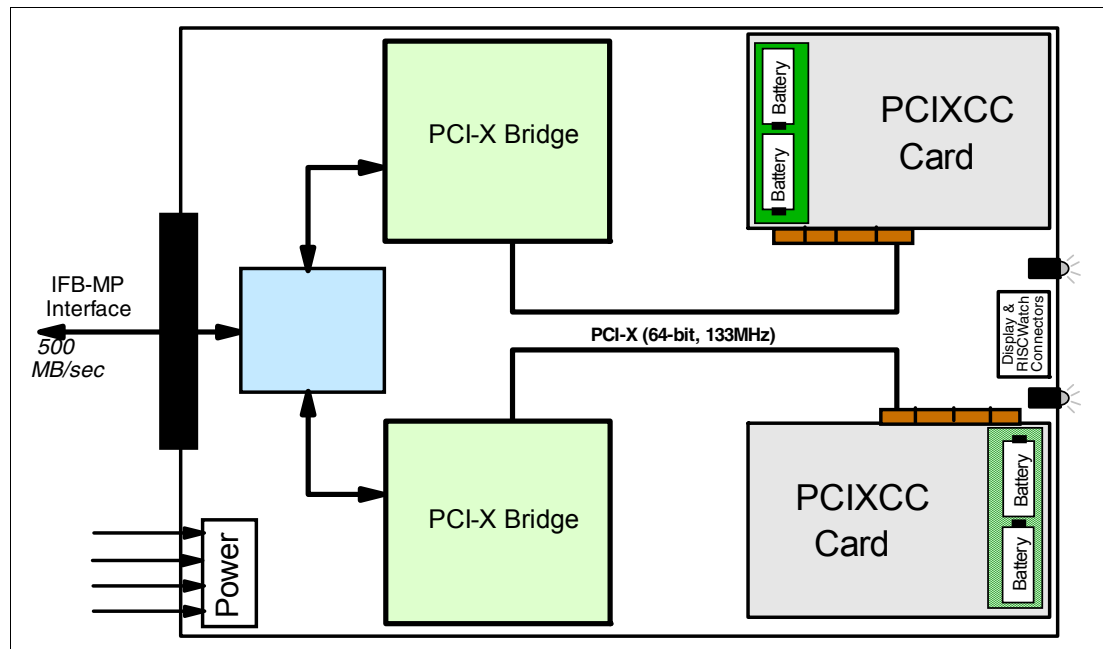


Figure 6-2 Crypto Express2 feature layout

The Crypto Express2 feature does not have external ports and does not use fiber optic or other cables. It does not use CHPIDs, but requires one slot in the I/O cage and one PCHID for each PCI-X cryptographic adapter. The feature is attached to a self-timed interface (STI) and has no other external interfaces. Removal of the feature or card *zeroizes* the content.

The z10 EC supports a maximum of eight Crypto Express2 features, offering a combination of up to 16 coprocessor and accelerators. Access to the PCI-X cryptographic adapter is controlled through the setup in the image profiles on the SE.

Note: Although PCI-X cryptographic adapters have no CHPID type and are not identified as external channels, all logical partitions in all channel subsystems have access to the adapter (up to 16 logical partitions per adapter). Having access to the adapter requires setup in the image profile for the partition. The adapter must be in the candidate list. For details about setting up the image profile see *IBM System z10 Enterprise Class Configuration Setup*, SG24-7571.

6.4.1 Crypto Express2 coprocessor

The Crypto Express2 coprocessor is a PCI-X cryptographic adapter configured as a coprocessor and provides a high-performance cryptographic environment with added functions.

The Crypto Express2 coprocessor provides asynchronous functions only.

The Crypto Express2 feature contains two PCI-X cryptographic adapters. The two adapters are actually two PCI-X CC cryptographic processors that provide the equivalent (plus additional) functions as the PCIXCC feature on the z990 with doubled throughput.

PCI-X cryptographic adapters, when configured as coprocessors, are designed for FIPS 140-2 Level 4 compliance rating for secure cryptographic hardware modules. Unauthorized removal of the adapter or feature *zeroizes* its content.

The Crypto Express2 coprocessor enables the user to:

- ▶ Encrypt and decrypt data by using secret-key algorithms. Triple-length key DES and double-length key DES algorithms are supported.
- ▶ Generate, install, and distribute cryptographic keys securely by using both public and secret-key cryptographic methods.
- ▶ Generate, verify, and translate personal identification numbers (PINs).
- ▶ Generate, verify, and translate 13 through 19-digit personal account numbers (PANs).
- ▶ Ensure the integrity of data by using message authentication codes (MACs), hashing algorithms, and Rivest-Shamir-Adelman (RSA) public key algorithm (PKA) digital signatures.

The Crypto Express2 coprocessor also provides the functions listed for the Crypto Express2 accelerator, however, with a lower performance than the Crypto Express2 accelerator can provide.

Three methods of master key entry are provided by Integrated Cryptographic Service Facility (ICSF) for the Crypto Express2 feature coprocessor:

- ▶ A pass-phrase initialization method, which generates and enters all master keys that are necessary to fully enable the cryptographic system in a minimal number of steps.
- ▶ A simplified master key entry procedure provided through a series of Clear Master Key Entry panels from a TSO terminal.
- ▶ A Trusted Key Entry (TKE) workstation, which is available as an optional feature in enterprises that require enhanced key-entry security.

The security-relevant portion of the cryptographic functions is performed inside the secure physical boundary of a tamper-resistant card. Master keys and other security-relevant information are also maintained inside this secure boundary.

A Crypto Express2 coprocessor operates with the Integrated Cryptographic Service Facility (ICSF) and IBM Resource Access Control Facility (RACF®), or equivalent software products, in a z/OS operating environment to provide data privacy, data integrity, cryptographic key installation and generation, electronic cryptographic key distribution, and personal identification number (PIN) processing.

The Processor Resource/Systems Manager (PR/SM) fully supports the Crypto Express2 feature coprocessor to establish a logically partitioned environment on which multiple logical partitions can use the cryptographic functions. A 128-bit data-protection master key and one 192-bit public key algorithm (PKA) master key are provided for each of 16 cryptographic domains that a coprocessor can serve.

Use the dynamic addition or deletion of a logical partition name to rename a logical partition. Its name can be changed from NAME1 to * (single asterisk) and then changed again from * to NAME2. The logical partition number and MIF ID are retained across the logical partition name change. The master keys in the Crypto Express2 feature coprocessor that were associated with the old logical partition NAME1 are retained. No explicit action is taken against a cryptographic component for this dynamic change.

Note: Cryptographic coprocessors are not tied to logical partition numbers or MIF IDs. They are set up with PCI-X adapter numbers and domain indices that are defined in the partition image profile. The customer can dynamically configure them to a partition and change or clear them when needed.

6.4.2 Crypto Express2 accelerator

The Crypto Express2 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. Note the following information about the reconfiguration:

- ▶ It is done through the Support Element.
- ▶ It is done at the PCI-X cryptographic adapter level. A Crypto Express2 feature can host a coprocessor and an accelerator, two coprocessors, or two accelerators.
- ▶ It works both ways, from coprocessor to accelerator and from accelerator to coprocessor. Master keys in the coprocessor domain can be optionally preserved when it is reconfigured to be an accelerator.
- ▶ It is disruptive to coprocessor and accelerator operations. The coprocessor or accelerator must be deactivated before engaging 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 configured as an accelerator are used for the acceleration of modular arithmetic operations (that is, the RSA cryptographic operations used with the SSL/TLS protocol), as follows:

- ▶ PKA Decrypt (CSNDPKD), with PKCS-1.2 formatting
- ▶ PKA Encrypt (CSNDPKE), with zero-pad formatting
- ▶ Digital Signature Verify

The RSA encryption and decryption functions support key lengths of 512 bit to 4,096 bit, in the Modulus Exponent (ME) and Chinese Remainder Theorem (CRT) formats.

6.4.3 Configuration rules

Each system supports up to eight Crypto Express2 features, which equals up to a maximum of 16 PCI-X cryptographic adapters. In a one-book system up to eight features may be installed and configured.

Table 6-2 summarizes configuration information for Crypto Express2.

Table 6-2 *Crypto Express2 feature*

Minimum number of orderable features for each server ^a	2
Order increment above two features	1
Maximum number of features for each server	8
Number of PCI-X cryptographic adapters for each feature (coprocessor or accelerator)	2
Maximum number of PCI-X adapters for each server	16
Number of cryptographic domains for each PCI-X adapter ^b	16

a. The minimum initial order of Crypto Express2 features is two. After the initial order, additional Crypto Express2 can be ordered one feature at a time up to a maximum of eight.

b. More than one partition, defined to the same CSS or to different CSSs, can use the same domain number when assigned to different PCI-X cryptographic adapters.

The concept of *dedicated processor* does not apply to the PCI-X cryptographic adapter. Whether configured as coprocessor or accelerator, the PCI-X cryptographic adapter is made available to a logical partition as directed by the domain assignment and the candidate list in the logical partition image profile, regardless of the shared or dedicated status given to the CPs in the partition.

When installed non-concurrently, Crypto Express2 features are assigned PCI-X cryptographic adapter numbers sequentially during the power-on reset following the installation. When a Crypto Express2 feature is installed concurrently, the installation can select an out-of-sequence number from the unused range. When a Crypto Express2 feature is removed concurrently, the PCI-X adapter numbers are automatically freed.

The definition of domain indexes and PCI-X cryptographic adapter numbers in the candidate list for each logical partition should be planned ahead to allow for nondisruptive changes, as follows.

- ▶ Operational changes can be made by using the Change LPAR Cryptographic Controls task from the Support Element, which reflects the cryptographic definitions in the image profile for the partition. With this function, adding and removing the cryptographic feature without stopping a running operating system can be done dynamically.
- ▶ The same usage domain index may be defined more than once across multiple logical partitions. However, the PCI-X cryptographic adapter number coupled with the usage domain index specified must be unique across all active logical partitions.

The same PCI-X cryptographic adapter number and usage domain index combination may be defined for more than one logical partition, for example to define a configuration for backup situations. Note that only one of the logical partitions can be active at any one time.

The z10 EC allows for up to 60 logical partitions to be active concurrently. Each PCI-X adapter supports 16 domains, whether it is configured as a Crypto Express2 accelerator or a Crypto Express2 coprocessor. The server configuration must include at least two Crypto Express2 (four PCI-X adapters and 16 domains per PCI-X adapter) when all 60 logical partitions require concurrent access to cryptographic functions. More Crypto Express2 features may be needed to satisfy application performance and availability requirements.

For availability, assignment of multiple PCI-X adapters of the same type (Crypto Express2 accelerator or coprocessor) to one logical partition should be spread across multiple features.

6.5 Crypto Express3

The Crypto Express3 feature (FC 0864) has two PCI Express cryptographic adapters. Each of the PCI Express cryptographic adapters can be configured as a cryptographic coprocessor or a cryptographic accelerator.

The Crypto Express3 feature is the newest state-of-the-art generation cryptographic feature. Like its predecessors it is designed to complement the functions of CPACF. This feature is tamper-sensing and tamper-responding. It provides dual processors operating in parallel supporting cryptographic operations with high reliability.

Figure 6-3 shows the physical layout of the Crypto Express3 feature.

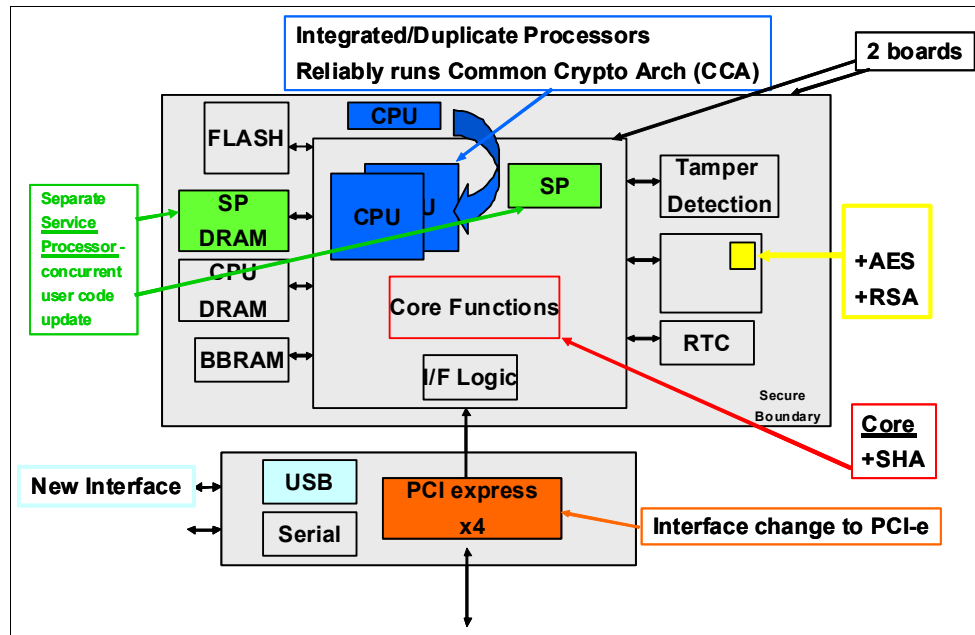


Figure 6-3 Crypto Express3 feature layout

The Crypto Express3 feature has the same configuration options as, and contains all the functions of, the Crypto Express2 feature and introduces a number of additional functions, including:

- ▶ SHA2 functions similar to the SHA2 function in CPACF
- ▶ RSA functions similar to the RSA function in CPACF
- ▶ Dynamic power management designed to keep within the temperature limits of the feature and at the same time maximize RSA performance
- ▶ Up to 32 LPARs in all logical channel subsystems have access to the feature
- ▶ Improved RAS over previous crypto features due to dual processors and the service processor
- ▶ Function update while installed using secure code load
- ▶ When a PCI Express adapter is defined as a coprocessor lock-step checking by the dual processors enhances error detection and fault isolation
- ▶ Dynamic addition and configuration of the Crypto Feature3 to LPARs without an outage
- ▶ Updated cryptographic algorithms used to load the LIC from the TKE
- ▶ Support for smart card applications using Europay, MasterCard, and VISA specifications

It is not possible to mix and match UDX definitions across Crypto Express2 and Crypto Express3 features. Panels on the HMC and SE ensure that UDX files are applied to the appropriate crypto card type.

The UDX toolkit for System z with the Crypto Express3 feature is made available on the general availability date for the feature. In addition, there will be a migration path for customers with UDX on a previous feature to migrate their code to the Crypto Express3 feature. A UDX migration is no more disruptive than a normal MCL or ICSF release migration.

The Crypto Express3 feature is designed to deliver throughput improvements for both symmetric and asymmetric operations.

For less error-prone and easier migration a Crypto Express3 migration wizard is available. The wizard allows the user to collect configuration data from a Crypto Express2 or Crypto Express3 feature configured as a coprocessor and migrate that data to a different Crypto Express coprocessor. The target for this migration must be a coprocessor with equivalent or greater capabilities.

6.6 TKE workstation feature

The TKE workstation is an optional feature that offers key management functions. The TKE workstation, with TKE 5.2 or later Licensed Internal Code (LIC), is required to support cryptographic key management on the z10 EC.

Note: TKE 5.3 LIC is required in support of the AES algorithm and includes master key management functions to load or generate AES master keys to cryptographic coprocessors.

TKE workstation with LIC 5.2 or later can control cryptographic features on z10 EC, z10 BC, z9 EC, z9 BC, z990, and z890 servers.

Note: The TKE workstation supports Ethernet adapters only to connect to a LAN.

A TKE workstation is part of a customized solution for using the Integrated Cryptographic Service Facility for z/OS program product (ICSF for z/OS) to manage cryptographic keys of a z10 EC that has Crypto Express features installed and that is configured for using DES and PKA cryptographic keys.

The TKE workstation provides secure control for the Crypto Express2 and Crypto Express3 coprocessors, including loading of master keys.

The TKE workstation with LIC 6.0 offers a number of usability enhancements:

- ▶ Grouping of up to 16 domains across one or more cryptographic adapters. These adapters may be installed on one or more servers or LPARs. Grouping of domains applies to CryptoExpress3 and Crypto Express2 features.
- ▶ Greater flexibility and efficiency by executing domain-scoped commands on every domain in the group. For example, a TKE user can load master key parts to all domains with one command.
- ▶ Efficiency by executing Crypto Express2 and Crypto Express3 scoped commands on every coprocessor in the group. This allows a substantial reduction of the time required for loading new master keys from a TKE workstation into a Crypto Express3 or Crypto Express2 feature.

Furthermore, the LIC 6.0 strengthens the cryptography for TKE protocol inbound and outbound authentication. TKE uses cryptographic algorithms and protocols in communication with the target cryptographic adapters on the host systems that it administers. Cryptography is first used to verify that each target adapter is a valid IBM cryptographic coprocessor. It then ensures that there are secure messages between the TKE workstation and the target Crypto Express2 or Crypto Express3 feature. The cryptography has been updated to keep pace with industry developments and with recommendations from experts and standards organizations.

The enhancements are in the following areas:

- ▶ TKE Certificate Authorities (CAs) initialized on a TKE workstation with TKE 6.0 LIC can issue certificates with 2048-bit keys. Previous versions of TKE used 1024-bit keys.
- ▶ The transport key used to encrypt sensitive data sent between the TKE workstation and a Crypto Express3 coprocessor has been strengthened from a 192-bit TDES key to a 256-bit AES key.
- ▶ The signature key used by the TKE workstation and the Crypto Express3 coprocessor has been strengthened from 1024-bit to a maximum of 4096-bit strength.
- ▶ Replies sent by a Crypto Express3 coprocessor on the host are signed with a 4096-bit key.

The TKE LIC 6.0 increases the key strength for TKE Certificate Authority smart cards, TKE smart cards, and signature keys stored on smart cards from 1024-bit to 2048-bit strength.

Only smart cards (# 0884) with smart card readers (# 0885) support the creation of TKE Certificate Authority (CA) smart cards, TKE smart cards, or signature keys with the new 2048-bit key strength. Smart cards (#0888) and smart card readers (# 0887) will continue to work with the 1024-bit key strength.

Logical partition, TKE host, and TKE target

If one or more logical partitions are customized for using Crypto Express coprocessors, the TKE workstation can be used to manage DES master keys and PKA master keys for all cryptographic domains of each Crypto Express coprocessor feature assigned to logical partitions defined by the TKE workstation.

Each logical partition in the same system using a domain managed through a TKE workstation connection is either a TKE host or a TKE target. A logical partition with a TCP/IP connection to the TKE is referred to as TKE host. All other partitions are TKE targets.

The cryptographic controls as set for a logical partition through the Support Element determine whether the workstation is a TKE host or TKE target.

Optional smart card reader

Adding an optional smart card reader (FC 0885) to the TKE workstation is possible. The reader supports the use of smart cards that contain an embedded microprocessor and associated memory for data storage that can contain the keys to be loaded into the Crypto Express features. Access to and use of confidential data on the smart card is protected by a user-defined personal identification number (PIN). Up to 99 additional smart cards can be ordered for backup. The smart card feature code is FC 0884. The older features, FC 0887 and FC 0888, will be withdrawn from marketing on November 20, 2009.

6.7 Cryptographic functions comparison

Table 6-3 lists functions or attributes on z10 EC of the three cryptographic hardware features. In the table, X indicates the function or attribute is supported.

Table 6-3 Cryptographic functions on z10 EC

Functions or attributes	CPACF	Crypto Express2 and Crypto Express3 Coprocessor	Crypto Express2 and Crypto Express3 Accelerator
Supports z/OS applications using ICSF	X	X	X
Encryption and decryption using secret-key algorithm	-	X	-
Provides highest SSL handshake performance	-	-	X ^a
Provides highest symmetric (clear key) encryption performance	X	-	-
Provides highest asymmetric (clear key) encryption performance	-	-	X
Provides highest asymmetric (encrypted key) encryption performance	-	X	-
Disruptive process to enable	-	Note ^b	Note ^b
Requires IOCDs definition	-	-	-
Uses CHPID numbers	-	-	-
Uses PCHIDs	-	X ^c	X ^c
Physically embedded on each CP and IFL	X	-	-
Requires CPACF DES or triple DES enablement (FC 3863)	X ^d	X ^d	X ^d
Requires ICSF to be active	-	X	X
Offers user programming function (UDX)	-	X	-
Usable for data privacy: encryption and decryption processing	X	X	-
Usable for data integrity: hashing and message authentication	X	X	-
Usable for financial processes and key management operations	-	X	-
Crypto performance RMF monitoring	-	X	X
Requires system master keys to be loaded	-	X	-
System (master) key storage	-	X	-
Retained key storage	-	X	-
Tamper-resistant hardware packaging	-	X	X ^e
Designed for FIPS 140-2 Level 4 certification	-	X	-

Functions or attributes	CPACF	Crypto Express2 and Crypto Express3 Coprocessor	Crypto Express2 and Crypto Express3 Accelerator
Supports SSL functions	X	X	X
Supports Linux applications doing SSL handshakes	-	-	X
RSA functions	-	X	X
High performance SHA-1 to SHA-512	X	SHA2 for Crypto Express3	SHA2 for Crypto Express3
Clear key DES or triple DES	X	-	-
Advanced Encryption Standard (AES) for 128-bit, 192-bit, and 256-bit keys	X	-	-
Pseudorandom number generator (PRNG)	X	X	-
Clear key RSA	-	-	X
Double length DUKPT support	-	X	-
Europay Mastercard VISA (EMV) support	-	X	-
Public Key Decrypt (PKD) support for Zero-Pad option for clear RSA private keys	-	X	X
Public Key Encrypt (PKE) support for MRP function	-	X	X
Remote loading of initial keys in ATM	-	X	-
Improved key exchange with non CCA system	-	X	-
ISO 16609 CBC mode triple DES MAC support	-	X	-

- a. Requires CPACF DES or triple DES enablement feature code 3863.
- b. To make the addition of the Crypto Express features nondisruptive, the logical partition must be predefined with the appropriate PCI-X or PCI Express cryptographic adapter number selected in its candidate list in the partition image profile.
- c. One PCHID is required for each PCI-X cryptographic adapter.
- d. This is not required for Linux if only RSA clear key operations are used. DES or triple DES encryption requires CPACF to be enabled.
- e. This is physically present but is not used when configured as an accelerator (clear key only).

6.8 Software support

The software support levels are listed in 7.4, “Cryptographic support” on page 217.



Software support

This chapter lists the minimum operating system requirements and support considerations for the z10 EC and its features. It discusses z/OS, z/VM, z/VSE, TPF, z/TPF, and Linux on System z. Because this information is subject to change, see the Preventive Service Planning (PSP) bucket for 2097DEVICE for the most current information,

Support of IBM System z10 Enterprise Class functions is dependent on the operating system, version, and release.

This chapter discusses the following topics:

- ▶ 7.1, “Operating systems summary” on page 190
- ▶ 7.2, “Support by operating system” on page 190
- ▶ 7.3, “Support by function” on page 200
- ▶ 7.4, “Cryptographic support” on page 217
- ▶ 7.5, “z/OS migration considerations” on page 221
- ▶ 7.6, “Coupling facility and CFCC considerations” on page 223
- ▶ 7.7, “MIDAW facility” on page 224
- ▶ 7.8, “IOCP” on page 228
- ▶ 7.10, “ICKDSF” on page 229
- ▶ 7.11, “Software licensing considerations” on page 229
- ▶ 7.12, “References” on page 232

7.1 Operating systems summary

Table 7-1 lists the minimum operating system levels required on the z10 EC. Note that operating system levels that are no longer in service are not covered in this publication. These older levels may provide support for some features.

Table 7-1 z10 EC minimum operating systems requirements

Operating systems	ESA/390 (31-bit mode)	z/Architecture (64-bit mode)	Notes
z/OS V1R7 ^a	No	Yes	Service is required. See the following shaded Note box.
z/VM V5R3	No	Yes ^b	
z/VSE V4	No	Yes	
z/TPF V1R1	Yes	Yes	
TPF V4R1	Yes	No	
Linux on System z	See Table 7-2 on page 191.	See Table 7-2 on page 191.	Novell SUSE SLES 9 Red Hat RHEL 4

a. Regular service support for z/OS V1R7 ended in September 2008. However, by ordering the IBM Lifecycle Extension for z/OS V1.7 product, fee-based corrective service can be obtained for up to two years after withdrawal of service (September 2010). Similarly, the IBM Lifecycle Extension for z/OS V1.8 product provides corrective service up to September 2011.

b. z/VM supports both 31-bit and 64-bit mode guests.

Note: Exploitation of certain features depends on a particular operating system. In all cases, PTFs might be required with the operating system level indicated. Check the z/OS, z/VM, z/VSE, z/TPF, and TPF subsets of the 2097DEVICE Preventive Service Planning (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 recommendations, and cross-product dependencies.

7.2 Support by operating system

System z10 EC introduces several new functions. In this section, we discuss support of those by the current operating systems. Also included are some of the functions previously introduced by z9 EC and z990 and carried forward or enhanced in the z10 EC.

For a list of supported functions and the z/OS and z/VM minimum required support levels, see Table 7-3 on page 193. For z/VSE, Linux on System z, z/TPF, and TPF see Table 7-4 on page 197. The tabular format is intended to help determine, by a quick scan, which functions are supported and the minimum operating system level required.

7.2.1 z/OS

z/OS Version 1 Release 9 is the earliest in-service release supporting the z10 EC. Although service support for z/OS Version 1 Release 8 ended in September of 2009, a fee-based extension for defect support (for up to two years) can be obtained by ordering the IBM

Lifecycle Extension for z/OS V1.8. Similarly, IBM Lifecycle Extension for z/OS V1.7 provides fee-based support for z/OS Version 1 Release 7 until September 2010. Support for z/OS Version 1 Release 6 ended on September 30, 2007. Also note that z/OS.e is not supported on z10 EC and that z/OS.e Version 1 Release 8 was the last release of z/OS.e.

See Table 7-3 on page 193 for a list of supported functions and their minimum required support levels.

7.2.2 z/VM

At general availability:

- ▶ z/VM V5R4 and later provide exploitation support.
- ▶ z/VM V5R3 provides compatibility support only.

See Table 7-3 on page 193 for a list of supported functions and their minimum required support levels.

Notes: We recommend that the capacity of any z/VM logical partitions, and any z/VM guests, in terms of the number of IFLs and CPs, real or virtual, be adjusted to accommodate the PU capacity of the z10 EC.

7.2.3 z/VSE

z/VSE V4:

- ▶ Executes in z/Architecture mode only
- ▶ Exploits 64-bit real memory addressing
- ▶ Does not support 64-bit virtual addressing

See Table 7-4 on page 197 for a list of supported functions and their minimum required support levels.

7.2.4 Linux on System z

Linux on System z distributions are built separately for the 31-bit and 64-bit addressing modes of the z/Architecture. The newer distribution versions are built for 64-bit only. You can run 31-bit applications in the 31-bit emulation layer on a 64-bit Linux on System z distribution. None of the current versions of Linux on System z distributions (SLES 9, SLES 10, SLES 11, RHEL 4, and RHEL 5)¹ require System z10 toleration support. Table 7-2 shows the most recent service levels of the current SUSE and Red Hat releases at the time of writing.

Table 7-2 Current Linux on System z distributions as of October 2009

Linux on System z distribution	ESA/390 (31-bit mode)	z/Architecture (64-bit mode)
Novell SUSE SLES 9 SP4	Yes	Yes
Novell SUSE SLES 10 SP3	No	Yes
Novell SUSE SLES 11	No	Yes
Red Hat RHEL 4.8	Yes	Yes
Red Hat RHEL 5.4	No	Yes

IBM is working with its Linux distribution partners to provide further exploitation of selected z10 EC functions in future Linux on System z distribution releases.

We recommend that:

- ▶ SUSE SLES 11 or Red Hat RHEL 5 be used in any new projects for the z10 EC.
- ▶ Any Linux distributions be updated to their latest service level before migration to z10 EC.
- ▶ The capacity of any z/VM and Linux on System z logical partitions guests, as well as z/VM guests, in terms of the number of IFLs and CPs, real or virtual, be adjusted according to the PU capacity of the z10 EC.

7.2.5 TPF and z/TPF

See Table 7-4 on page 197 for a list of supported functions and their minimum required support levels.

7.2.6 z10 EC functions support summary

In the following tables, although we attempt to note all functions requiring support, the PTF numbers are not given. Therefore, for the most current information, see the Preventive Service Planning (PSP) bucket for 2097DEVICE.

The following two tables summarize the z10 EC functions and their minimum required operating system support levels:

- ▶ Table 7-3 on page 193 is for z/OS and z/VM.
- ▶ Table 7-4 on page 197 is for z/VSE, Linux on System z, z/TPF, and TPF.

Information about Linux on System z refers exclusively to appropriate distributions (either 31-bit or 64-bit) of Novell SUSE SLES 9 and Red Hat RHEL 4.

Both tables use the following conventions:

Y	The function is supported.
N	The function is not supported.
-	The function is not applicable to that specific operating system.

¹ SLES is SUSE Linux Enterprise Server
RHEL is Red Hat Enterprise Linux

Table 7-3 z10 EC functions minimum support requirements summary, part 1

Function	z/OS V1R11	z/OS V1R10	z/OS V1R9	z/OS V1R8	z/OS V1R7	z/VM V6R1	z/VM V5R4	z/VM V5R3
z10 EC	Y	Y	Y	Y	Y	Y	Y	Y ^a
Greater than 54 PUs single system image	Y	Y	Y	N	N	N ^b	N ^b	N ^b
zIIP	Y	Y	Y	Y	Y	Y ^c	Y ^c	Y ^c
zAAP	Y	Y	Y	Y	Y	Y ^c	Y ^c	Y ^c
zAAP on zIIP	Y	Y ^j	Y ^j	N	N	Y ^d	Y ^d	N
Large memory (> 128 GB)	Y	Y	Y	Y	N	Y ^e	Y ^e	Y ^e
Large page support	Y	Y	Y	N	N	N ^f	N ^f	N ^f
Guest support for execute-extensions facility	-	-	-	-	-	Y	Y	Y
Hardware decimal floating point	Y ^g	Y ^g	Y ^g	Y ^g	Y ^g	Y ^c	Y ^c	Y ^c
60 logical partitions	Y	Y	Y	Y	Y	Y	Y	Y
LPAR group capacity limit	Y	Y	Y	Y	N	-	-	-
CPU measurement facility	Y	Y ^j	Y ^j	Y ^j	N	N	N	N
Separate LPAR management of PUs	Y	Y	Y	Y	Y	Y	Y	Y
Dynamic add and delete logical partition name	Y	Y	Y	Y	Y	Y	Y	Y
Capacity provisioning	Y	Y	Y ^j	N	N	N ^f	N ^f	N ^f
Enhanced flexibility for CoD	Y ^g	Y ^g	Y ^g	Y ^g	Y ^g	Y ^g	Y ^g	N ^f
HiperDispatch	Y	Y	Y	Y	Y ^h	N ^f	N ^f	N ^f
63.75 K subchannels	Y	Y	Y	Y	Y	Y	Y	Y
Four logical channel subsystems (LCSS)	Y	Y	Y	Y	Y	Y	Y	Y
Dynamic I/O support for multiple LCSS	Y	Y	Y	Y	Y	Y	Y	Y
Multiple subchannel sets	Y	Y	Y	Y	Y	N ^f	N ^f	N ^f
MIDAW facility	Y	Y	Y	Y	Y	Y ^c	Y ^c	Y ^c
Cryptography								
CPACF protected public key	Y ⁱ	Y ⁱ	Y ⁱ	N	N	N ^f	N ^f	N ^f
CPACF enhancements	Y ⁱ	Y ⁱ	Y ⁱ	Y ⁱ	Y ⁱ	Y ^c	Y ^c	Y ^c
CPACF AES, PRNG, and SHA-256	Y	Y	Y	Y	Y ⁱ	Y ^c	Y ^c	Y ^c
CPACF	Y	Y	Y	Y	Y ⁱ	Y ^c	Y ^c	Y ^c
Personal Account Numbers of 13 to 19 digits	Y ⁱ	Y ⁱ	Y ⁱ	Y ⁱ	Y ⁱ	Y ^c	Y ^c	Y ^c
Crypto Express3	Y ⁱ	Y ⁱ	Y ⁱ	N	N	Y ^c	Y ^c	Y ^c

Function	z/OS V1R11	z/OS V1R10	z/OS V1R9	z/OS V1R8	z/OS V1R7	z/VM V6R1	z/VM V5R4	z/VM V5R3
Crypto Express2	Y	Y	Y	Y	Y ⁱ	Y ^c	Y ^c	Y ^c
Remote key loading for ATMs, ISO 16609 CBC mode triple DES MAC	Y	Y	Y	Y	Y ⁱ	Y ^c	Y ^c	Y ^c
HiperSockets								
HiperSockets multiple write facility	Y	Y	Y ^j	N	N	N ^f	N ^f	N ^f
HiperSockets support of IPV6	Y	Y	Y	Y	Y	Y	Y	Y
HiperSockets Layer 2 Support	N	N	N	N	N	Y ^c	Y ^c	Y ^c
HiperSockets	Y	Y	Y	Y	Y	Y	Y	Y
ESCON (Enterprise Systems CONnection)								
16-port ESCON feature	Y	Y	Y	Y	Y	Y	Y	Y
FICON (Fiber CONnection) and FCP (Fibre Channel Protocol)								
High Performance FICON for System z (zHPF)	Y	Y ^j	Y ^j	Y ^j	N ^f	N ^f	N ^f	N ^f
FCP - increased performance for small block sizes	N	N	N	N	N	Y	Y	Y
Request node identification data	Y	Y	Y	Y	Y	N	N	N
FICON link incident reporting	Y	Y	Y	Y	Y	N	N	N
N_Port ID Virtualization for FICON (NPIV) CHPID type FCP	N	N	N	N	N	Y	Y	Y
FCP point-to-point attachments	N	N	N	N	N	Y	Y	Y
FICON SAN platform & name server registration	Y	Y	Y	Y	Y	Y	Y	Y
FCP SAN management	N	N	N	N	N	N	N	N
SCSI IPL for FCP	N	N	N	N	N	Y	Y	Y
Cascaded FICON Directors CHPID type FC	Y	Y	Y	Y	Y	Y	Y	Y
Cascaded FICON Directors CHPID type FCP	N	N	N	N	N	Y	Y	Y
FICON Express8, FICON Express4 and FICON Express2 support of SCSI disks CHPID type FCP	N	N	N	N	N	Y	Y	Y
FICON Express8	Y ^k	Y ^k	Y ^k	Y ^k	Y ^k	Y ^k	Y ^k	Y ^k
FICON Express4 ^l	Y	Y	Y	Y	Y	Y	Y	Y
FICON Express2 ^l	Y	Y	Y	Y	Y	Y	Y	Y
FICON Express ^l CHPID type FCV	Y	Y	Y	Y	N	Y	Y	Y

Function	z/OS V1R11	z/OS V1R10	z/OS V1R9	z/OS V1R8	z/OS V1R7	z/VM V6R1	z/VM V5R4	z/VM V5R3
FICON Express ¹ CHPID type FC	Y	Y	Y	Y	N	Y	Y	Y
FICON Express ¹ CHPID type FCP	N	N	N	N	N	Y	Y	Y
OSA (Open Systems Adapter)								
VLAN management	Y	Y	Y	Y	Y	Y	Y	Y
VLAN (IEE 802.1q) support	Y	Y	Y	Y	Y	Y	Y	Y
QDIO data connection isolation for z/VM virtualized environments	-	-	-	-	-	Y	Y ^j	Y ^j
OSA Layer 3 Virtual MAC	Y	Y	Y	Y	N	Y ^c	Y ^c	Y ^c
OSA Dynamic LAN idle	Y	Y	Y	Y	N	Y ^c	Y ^c	Y ^c
OSA/SF enhancements for IP, MAC addressing (CHPID=OSD)	Y	Y	Y	Y	Y	Y	Y	Y
QDIO diagnostic synchronization	Y	Y	Y	Y	N	Y ^c	Y ^c	Y ^c
OSA-Express2 Network Traffic Analyzer	Y	Y	Y	Y	N	Y ^c	Y ^c	Y ^c
Broadcast for IPv4 packets	Y	Y	Y	Y	Y	Y	Y	Y
Checksum offload for IPv4 packets	Y	Y	Y	Y	Y	Y ^m	Y ^m	Y ^m
OSA-Express3 10 Gigabit Ethernet LR CHPID type OSD	Y	Y	Y	N	N	Y	Y	Y
OSA-Express3 10 Gigabit Ethernet SR CHPID type OSD	Y	Y	Y	Y	Y	Y	Y	Y
OSA-Express3 Gigabit Ethernet LX (using four ports) CHPID types OSD, OSN	Y	Y	Y ^j	Y ^j	N	Y	Y	Y ^j
OSA-Express3 Gigabit Ethernet LX using two ports. CHPID types OSD, OSN	Y	Y	Y	Y	Y	Y	Y	Y
OSA-Express3 Gigabit Ethernet SX (using four ports) CHPID types OSD, OSN	Y	Y	Y ^j	Y ^j	N	Y	Y	Y ^j
OSA-Express3 Gigabit Ethernet SX (using 2 ports) CHPID types OSD, OSN	Y	Y	Y	Y	Y	Y	Y	Y
OSA-Express3 1000BASE-T (using 1 + 1 port) CHPID type OSC	Y	Y	Y	Y	Y	Y	Y	Y
OSA-Express3 1000BASE-T (using four ports) CHPID type OSD	Y	Y	Y ^j	Y ^j	N	Y	Y	Y ^j
OSA-Express3 1000BASE-T (using two ports) CHPID type OSD	Y	Y	Y	Y	Y	Y	Y	Y

Function	z/OS V1R11	z/OS V1R10	z/OS V1R9	z/OS V1R8	z/OS V1R7	z/VM V6R1	z/VM V5R4	z/VM V5R3
OSA-Express3 1000BASE-T (using two or four ports) CHPID type OSE	Y	Y	Y	Y	Y	Y	Y	Y
OSA-Express3 1000BASE-T CHPID type OSN	Y	Y	Y	Y	Y	Y	Y	Y
OSA-Express2 10 Gigabit Ethernet LR ⁿ CHPID type OSD	Y	Y	Y	Y	Y	Y	Y	Y
OSA-Express2 Gigabit Ethernet LX and SX ⁿ CHPID type OSD	Y	Y	Y	Y	Y	Y	Y	Y
OSA-Express2 Gigabit Ethernet LX and SX ⁿ CHPID type OSN	Y	Y	Y	Y	Y	Y	Y	Y
OSA-Express2 1000BASE-T Ethernet CHPID type OSC	Y	Y	Y	Y	Y	Y	Y	Y
OSA-Express2 1000BASE-T Ethernet CHPID type OSD	Y	Y	Y	Y	Y	Y	Y	Y
OSA-Express2 1000BASE-T Ethernet CHPID type OSE	Y	Y	Y	Y	Y	Y	Y	Y
OSA-Express2 1000BASE-T Ethernet CHPID type OSN	Y	Y	Y	Y	Y	Y	Y	Y
Parallel Sysplex and other								
z/VM integrated systems management	-	-	-	-	-	Y	Y	Y
System-initiated CHPID reconfiguration	Y	Y	Y	Y	Y	-	-	-
Program-directed re-IPL	-	-	-	-	-	Y	Y	Y
Multipath IPL	Y	Y	Y	Y	Y	N	N	N
STP enhancements	Y	Y	Y	Y	Y	-	-	-
Server Time Protocol	Y	Y	Y	Y	Y	-	-	-
Coupling over InfiniBand CHPID type CIB	Y	Y	Y	Y	Y	Y ^o	Y ^o	Y ^o
InfiniBand coupling links (1x IB-SDR or 1xIB DDR) at an unrepeated distance of 10 km	Y	Y	Y ^j	Y ^j	N	Y ^o	Y ^o	Y ^o
Dynamic I/O support for InfiniBand CHPIDs	-	-	-	-	-	Y	Y	Y
CFCC Level 16	Y	Y	Y	Y	Y	Y ^c	Y ^c	Y ^c

- a. Support is for compatibility only. z/VM and guests are supported at the System z9 functionality level. There is no exploitation of new hardware unless otherwise noted.
- b. A maximum of 32 PUs per system image is supported. Guests can be defined with up to 64 virtual PUs. z/VM V5R3 and V5R4 support up to 32 PUs.
- c. Support is for guest use only.
- d. Available for z/OS on virtual machines without virtual zAAPs defined when the z/VM LPAR does not have zAAPs defined.

- e. 256 GB of central memory are supported by z/VM V5R3 and later. z/VM V5R3 and later are designed to support more than 1 TB of virtual memory in use for guests.
- f. Not available to guests.
- g. Support varies by operating system and by version and release.
- h. This requires support for zIIP.
- i. FMIDs are shipped in a Web Deliverable.
- j. PTFs are required.
- k. Support varies with operating system and level. See “FICON Express8” on page 210“FICON Express8” on page 210 for details.
- l. FICON Express4 10KM LX, 4KM LX, and SX features are withdrawn from marketing. All FICON Express2 and FICON features are withdrawn from marketing.
- m. Supported for dedicated devices only.
- n. Withdrawn from marketing
- o. Support is for dynamic I/O configuration only.

Table 7-4 z10 EC functions minimum support requirements summary, part 2

Function	z/VSE V4R2	z/VSE V4R1	Linux on System z	z/TPF V1R1	TPF V4R1
z10 EC	Y	Y	Y	Y	Y
Greater than 54 PUs single system image	N	N	Y	Y	N
zIIP	-	-	-	-	-
zAAP	-	-	-	-	-
zAAP on zIIP	-	-	-	-	-
Large memory (> 128 GB)	N	N	Y	Y	N
Large page support	N	N	Y	N	N
Guest support for Execute-extensions facility	-	-	-	-	-
Hardware decimal floating point ^a	N	N	Y ^b	N	N
60 logical partitions	Y	Y	Y	Y	Y
CPU measurement facility	N	N	N	N	N
LPAR group capacity limit	-	-	-	-	-
Separate LPAR management of PUs	Y	Y	Y	Y	Y
Dynamic add/delete logical partition name	N	N	Y	N	N
Capacity provisioning	-	-	-	N	N
Enhanced flexibility for CoD	-	-	-	N	N
HiperDispatch	N	N	N	N	N
63.75 K subchannels	N	N	Y	N	N
Four logical channel subsystems	Y	Y	Y	N	N
Dynamic I/O support for multiple LCSS	N	N	Y	N	N
Multiple subchannel sets	N	N	Y	N	N
MIDAW facility	N	N	N	N	N
Cryptography					
CPACF protected public key	N	N	Y	N	N

Function	z/VSE V4R2	z/VSE V4R1	Linux on System z	z/TPF V1R1	TPF V4R1
CPACF enhancements	Y	Y	N	N	N
CPACF AES, PRNG, and SHA-256	Y	Y	Y	N	N
CPACF	Y	Y	Y	Y	Y
Personal Account Numbers of 13 to 19 digits	N	N	-		
Crypto Express3	Y ^c	N	Y ^d	Y	N
Crypto Express2	Y	Y	Y	Y	N
Remote key loading for ATMs, ISO 16609 CBC mode triple DES MAC	N	N	-	N	N
HiperSockets					
HiperSockets multiple write facility	N	N	N	N	N
HiperSockets support of IPV6	N	N	Y	N	N
HiperSockets Layer 2 Support	N	N	Y	N	N
HiperSockets	Y	Y	Y	N	N
ESCON (Enterprise System CONnection)					
16-port ESCON feature	Y	Y	Y	Y	Y
Fiber CONnection (FICON) and Fibre Channel Protocol (FCP)					
High Performance FICON for System z (zHPF)	N	N	N	N	N
FCP - increased performance for small block sizes	Y	Y	Y	N	N
Request node identification data	-	-	-	-	-
FICON link incident reporting	N	N	N	N	N
N_Port ID Virtualization for FICON (NPIV) CHPID type FCP	Y	Y	Y	N	N
FCP point-to-point attachments	Y	Y	Y	N	N
FICON SAN platform and name registration	Y	Y	Y	Y	Y
FCP SAN management	N	N	Y	N	N
SCSI IPL for FCP	Y	Y	Y	N	N
Cascaded FICON Directors CHPID type FC	Y	Y	Y	Y	Y
Cascaded FICON Directors CHPID type FCP	Y	Y	Y	N	N
FICON Express8, FICON Express4, and FICON Express2 support of SCSI disks CHPID type FCP	Y	Y	Y	N	N
FICON Express8	Y ^e	Y ^e	Y ^e	Y ^e	Y ^e
FICON Express4 ^f	Y	Y	Y	Y	Y

Function	z/VSE V4R2	z/VSE V4R1	Linux on System z	z/TPF V1R1	TPF V4R1
FICON Express2 ^f	Y	Y	Y	Y	Y
FICON Express ^f CHPID type FCV	Y	Y	N	N	N
FICON Express ^f CHPID type FC	Y	Y	Y	N	N
FICON Express ^f CHPID type FCP	Y	Y	Y	N	N
Open Systems Adapter (OSA)					
VLAN management	N	N	N	N	N
VLAN (IEE 802.1q) support	N	N	Y	N	N
QDIO data connection isolation for z/VM virtualized environments	-	-	-	-	-
OSA Layer 3 Virtual MAC	N	N	N	N	N
OSA Dynamic LAN idle	N	N	N	N	N
OSA/SF enhancements for IP, MAC addressing (CHPID=OSD)	N	N	N	N	N
OSA-Express2 QDIO Diagnostic Synchronization	N	N	N	N	N
OSA-Express2 Network Traffic Analyzer	N	N	N	N	N
Broadcast for IPv4 packets	N	N	Y	N	N
Checksum offload for IPv4 packets	N	N	Y	N	N
OSA-Express3 10 Gigabit Ethernet LR CHPID type OSD	Y	Y	Y	Y	Y
OSA-Express3 10 Gigabit Ethernet SR CHPID type OSD	Y	Y	Y	Y	Y
OSA-Express3 Gigabit Ethernet LX 4-ports CHPID types OSD	Y	Y	Y	Y	N
OSA-Express3 Gigabit Ethernet SX 4-ports CHPID types OSD	Y	Y	Y	Y	N
OSA-Express3 1000BASE-T CHPID type OSC	Y	Y	-	N	N
OSA-Express3 1000BASE-T 4-ports CHPID type OSD	Y	Y	Y	Y	N
OSA-Express3 1000BASE-T 4-ports CHPID type OSE	Y	Y	N	N	N
OSA-Express3 1000BASE-T Ethernet CHPID type OSN	Y	Y	Y	Y	Y
OSA-Express2 10 Gigabit Ethernet LR ⁹ CHPID type OSD	Y	Y	Y	Y	Y

Function	z/VSE V4R2	z/VSE V4R1	Linux on System z	z/TPF V1R1	TPF V4R1
OSA-Express2 Gigabit Ethernet LX and SX ^g CHPID type OSD	Y	Y	Y	Y	Y
OSA-Express2 Gigabit Ethernet LX and SX ^g CHPID type OSN	Y	Y	Y	Y	Y
OSA-Express2 1000BASE-T Ethernet CHPID type OSC	Y	Y	N	N	N
OSA-Express2 1000BASE-T Ethernet CHPID type OSD	Y	Y	Y	Y	Y
OSA-Express2 1000BASE-T Ethernet CHPID type OSE	Y	Y	N	N	N
OSA-Express2 1000BASE-T Ethernet CHPID type OSN	Y	Y	Y	Y	Y
Parallel Sysplex and other					
z/VM integrated systems management	-	-	-	-	-
System-initiated CHPID reconfiguration	-	-	Y	-	-
Program-directed re-IPL	Y ^h	Y ^h	Y	-	-
Multipath IPL	-	-	-	-	-
STP enhancements	-	-	-	-	-
Server Time Protocol	-	-	-	-	-
Coupling over InfiniBand CHPID type CIB	-	-	-	Y	Y
InfiniBand coupling links (1x IB-SDR or IB-DDR) at unrepeated distance of 10 km	-	-	-	-	-
Dynamic I/O support for InfiniBand CHPIDs	-	-	-	-	-
CFCC Level 16	-	-	-	Y	Y

- a. Support varies with operating system and level.
- b. Supported by Novell SUSE SLES 11.
- c. Service is required.
- d. Toleration support only. Requires SLES 10 SP3 or RHEL 5.4.
- e. See “FICON Express8” on page 210 for details.
- f. FICON Express4 10KM LX, 4KM LX, and SX features are withdrawn from marketing. All FICON Express2 and FICON features are withdrawn from marketing.
- g. Withdrawn from marketing.
- h. This is for FCP-SCSI disks.

7.3 Support by function

In this section, we discuss operating system support by function.

7.3.1 Single system image

A single system image can control several processor units such as CPs, zIIPs, zAAPs, or IFLs, as appropriate.

Maximum number of PUs

Table 7-5 shows the maximum number of PUs supported for each operating system image.

Table 7-5 Single system image software support

Operating system	Maximum number of (CPs+zIIPs+zAAPs) ^a or IFLs per system image
z/OS V1R11	64
z/OS V1R10	64
z/OS V1R9	64
z/OS V1R8	32
z/OS V1R7	32 ^b
z/VM V6R1	32 ^{c,d}
z/VM V5R4	32 ^{c,d}
z/VM V5R3	32 ^c
z/VSE V4	z/VSE Turbo Dispatcher can exploit up to 4 CPs and tolerates up to 10-way LPARs
Linux on System z	Novell SUSE SLES 9: 64 CPs or IFLs
	Novell SUSE SLES 10: 64 CPs or IFLs
	Novell SUSE SLES 11: 64 CPs or IFLs
	Red Hat RHEL 4: 8 CPs or IFLs
	Red Hat RHEL 5: 64 CPs or IFLs
z/TPF V1R1	64 CPs
TPF V4R1	16 CPs

a. The number of purchased zAAPs and the number of purchased zIIPs each cannot exceed the number of purchased CPs. A logical partition can be defined with any number of the available zAAPs and zIIPs. The total refers to the sum of these PU characterizations.

b. z/OS V1R7 requires IBM zIIP support for z/OS V1R7 Web deliverable to be installed to enable HiperDispatch.

c. z/VM guests can be configured with up to 64 virtual PUs.

d. The z/VM-mode LPAR supports CPs, zAAPs, zIIPs, IFLs and ICFs.

The z/VM-mode logical partition

System z10 supports a logical partition (LPAR) mode, named z/VM-mode, which is exclusive for running z/VM. The z/VM-mode requires z/VM V5R4 or later and allows z/VM to utilize a wider variety of specialty processors in a single LPAR. For instance, in a z/VM-mode LPAR, z/VM can manage Linux on System z guests running on IFL processors while also managing z/VSE and z/OS on central processors (CPs), and allowing z/OS to fully exploit IBM System z10 Integrated Information Processors (zIIPs) and IBM System z10 Application Assist Processors (zAAPs).

7.3.2 zAAP on zIIP capability

This new capability, exclusive to System z10 and System z9 servers under defined circumstances, enables workloads eligible to run on Application Assist Processors (zAAPs) to run on Integrated Information Processors (zIIP). It is intended as a means to optimize the investment on existing zIIPs and not as a replacement for zAAPs. The rule of at least one CP installed per zAAP and zIIP installed still applies.

Exploitation of this capability is by z/OS only, and is only available in these situations:

- ▶ When there are no zAAPs installed on the server.
- ▶ When z/OS is running as a guest of z/VM V5R4 or later and there are no zAAPs defined to the z/VM LPAR. The server may have zAAPs installed. Because z/VM can dispatch both virtual zAAPs and virtual zIIPs on real CPs², the z/VM partition does not require any real zIIPs defined to it, although we recommend using real zIIPs due to software licensing reasons.

Table 7-6 summarizes this support.

Table 7-6 Availability of zAAP on zIIP support

		z/OS is running on an LPAR ^a	z/OS is running as a z/VM guest		
			z/VM LPAR has zAAPs defined	No zAAPs defined to z/VM LPAR	
				Virtual zAAPs defined for z/OS guest	No virtual zAAPs for z/OS guest ^b
zAAPs installed on the server	Yes	No	No	No	Yes
	No	Yes	Not valid	No	Yes

a. zIIPs must be defined to the z/OS LPAR.

b. Virtual zIIPs must be defined to the z/OS virtual machine.

Support is available on z/OS V1R11 and this capability is enabled by default (ZAAPZIIP=YES). To disable it, specify NO for the ZAAPZIIP parameter in the IEASYSxx PARMLIB member.

On z/OS V1R10 and z/OS V1R9 support is provided by PTF for APAR OA27495 and the default setting in the IEASYSxx PARMLIB member is ZAAPZIIP=NO. Enabling or disabling this capability is disruptive. After changing the parameter, z/OS must be re-IPLed for the new setting to take effect.

² The z/VM system administrator can use the SET CPUAFFINITY command to influence the dispatching of virtual specialty engines on CPs or real specialty engines.

7.3.3 Maximum main storage size

Table 7-7 on page 203 lists the maximum amount of main storage supported by current operating systems. Expanded storage, although part of the z/Architecture, is currently exploited only by z/VM. A maximum of 1 TB of main storage can be defined for a logical partition.

Table 7-7 Maximum memory supported by operating system

Operating system	Maximum supported main storage
z/OS	z/OS V1R11 supports 4 TB and up to 1.5 TB per server ^a z/OS V1R10 supports 4 TB and up to 1.5 TB per server ^a z/OS V1R9 supports 4 TB and up to 1.5 TB per server ^a z/OS V1R8 supports 4 TB and up to 1.5 TB per server ^a z/OS V1R7 supports 128 GB
z/VM	z/VM V6R1 supports 256 GB z/VM V5R4 supports 256 GB z/VM V5R3 supports 256 GB
Linux on System z (64-bit)	Novell SUSE SLES 11 supports 4 TB Novell SUSE SLES 10 supports 4 TB Novell SUSE SLES 9 supports 4 TB Red Hat RHEL 5 supports 64 GB Red Hat RHEL 4 supports 64 GB
z/VSE	z/VSE V4R2 supports 32 GB z/VSE V4R1 supports 8 GB
TPF and z/TPF	z/TPF supports 4 TB ^a TPF runs in ESA/390 mode and supports 2 GB

a. System z10 EC restricts the LPAR memory size to 1 TB.

7.3.4 Large page support

In addition to the existing 4 KB pages and page frames, z10 EC supports large pages and large page frames that are 1 MB in size, as described in “Large page support” on page 88. Table 7-8 lists large page support requirements.

Table 7-8 Minimum support requirements for large page

Operating system	Support requirements
z/OS	z/OS V1R9
z/VM	Not supported; not available to guests
Linux on System z	Novell SUSE SLES 10 SP2 Red Hat RHEL 5.2

7.3.5 Guest support for execute-extensions facility

The execute-extensions facility contains several new machine instructions. Support is required in z/VM so that guests can exploit this facility. Table 7-9 lists the minimum support requirements.

Table 7-9 Minimum support requirements for execute-extensions facility

Operating system	Support requirements
z/VM	z/VM V5R4: support is included in the base z/VM V5R3: PTFs required

7.3.6 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 the z9 EC. However, the z10 EC has a new decimal floating point accelerator feature, described in 3.3.4, “Decimal floating point accelerator” on page 69. Therefore, in Table 7-10 we list the operating system support for decimal floating point. See also 7.5.7, “Decimal floating point and z/OS XL C/C++ considerations” on page 223.

Table 7-10 Minimum support requirements for decimal floating point

Operating system	Support requirements
z/OS	z/OS V1R9: Support includes XL, C/C++, HLASM, Language Environment®, DBX, and CDA RTLE. z/OS V1R8: Support includes HL ASM, Language Environment, DBX, and CDA RTLE. z/OS V1R7: Support is for the High Level Assembler (HLASM) only.
z/VM	z/VM V5R3: Support is for guest use.
Linux on System z	Novell SUSE SLES 11.

7.3.7 Up to 60 logical partitions

This feature, first made available in the z9 EC, allows the system to be configured with up to 60 logical partitions. Because channel subsystems can be shared by up to 15 logical partitions, configuring four channel subsystems to reach 60 logical partitions is necessary. Table 7-11 lists the minimum operating system levels for supporting 60 logical partitions.

Table 7-11 Minimum support requirements for 60 logical partitions

Operating system	Support requirements
z/OS	z/OS V1R7
z/VM	z/VM V5R3
z/VSE	z/VSE V4R1
Linux on System z	Novell SUSE SLES 9 Red Hat RHEL 4
TPF and z/TPF	TPF V4R1 and z/TPF V1R1

7.3.8 Separate LPAR management of PUs

The z10 EC uses separate PU pools for each optional PU type. The separate management of PU types enhances and simplifies capacity planning and management of the configured logical partitions and their associated processor resources. Table 7-12 on page 205 lists the support requirements for separate LPAR management of PU pools.

Table 7-12 Minimum support requirements for separate LPAR management of PUs

Operating system	Support requirements
z/OS	z/OS V1R7
z/VM	z/VM V5R3
z/VSE	z/VSE V4R1
Linux on System z	Novell SUSE SLES 9, Red Hat RHEL 4
TPF and z/TPF	TPF V4R1 and z/TPF V1R1

7.3.9 Dynamic LPAR memory upgrade

A logical partition 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 running in the LPAR.

Until now only z/OS was able to take advantage of this support by nondisruptively acquiring and releasing memory from the reserved area. z/VM V5R4 and higher are able to acquire memory nondisruptively, and immediately make it available to guests. z/VM virtualizes this support to its guests, which now can also increase their memory nondisruptively, if the operating system they are running supports it. Releasing memory from z/VM is a disruptive operation to z/VM. Releasing memory from the guest support depends on the operating system.

7.3.10 Capacity Provisioning Manager

The provisioning architecture, described in 8.8, “Nondisruptive upgrades” on page 273, enables you to better control the configuration and activation of the On/Of Capacity on Demand. 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 z/OS V1R9 function, interfaces with z/OS Workload Manager (WLM) and implements capacity provisioning policies. Several implementation options are available from an analysis mode, that only issues recommendations to an autonomic mode providing fully automated operations.

Replacing manual monitoring with autonomic management or supporting manual operation with recommendations can help ensure that sufficient processing power will be available with the least possible delay. Support requirements are listed on Table 7-13.

Table 7-13 Minimum support requirements for capacity provisioning

Operating system	Support requirements
z/OS	z/OS V1R9
z/VM	Not supported; not available to guests

7.3.11 Dynamic PU exploitation

z/OS has long been able to define reserved PUs to an LPAR for the purpose of non-disruptively bringing online the additional computing resources when needed.

z/OS V1R10 and z/VM V5R4 offer a similar, but enhanced, capability because no pre-planning is required. The ability to dynamically define and change the number and type of reserved PUs in an LPAR profile can be used for that purpose. The new resources are immediately made available to the operating systems and, in the z/VM case, to its guests.

7.3.12 HiperDispatch

HiperDispatch, which is exclusive to System z10, represents a cooperative effort between the z/OS operating system and the z10 EC hardware. It improves efficiencies in both the hardware and the software in the following ways:

- ▶ Work may be dispatched across fewer logical processors, therefore reducing the multiprocessor (MP) effects and lowering the interference among multiple partitions.
- ▶ Specific z/OS tasks may be dispatched to a small subset of logical processors that Processor Resource/Systems Manager (PR/SM) will tie to the same physical processors, thus improving the hardware cache reuse and locality of reference characteristics such as reducing the rate of cross-book communication.

For more information, see 3.6, “Logical partitioning” on page 90. Table 7-14 lists HiperDispatch support requirements.

Table 7-14 Minimum support requirements for HiperDispatch

Operating system	Support requirements
z/OS	z/OS V1R7 and later with PTFs (z/OS V1R7 requires IBM zIIP support for z/OS V1R7 Web deliverable)
z/VM	Not supported; not available to guests

7.3.13 The 63.75 K subchannels

Servers prior to the z9 EC reserved 1024 subchannels for internal system use out of the maximum of 64 K subchannels. Starting with the z9 EC, the number of reserved subchannels has been reduced to 256, thus increasing the number of subchannels available. Reserved subchannels exist only in subchannel set 0. No subchannels are reserved in subchannel set 1. The informal name, *63.75 K subchannels*, represents 65280 subchannels, as shown in the following equation:

$$63 \times 1024 + 0.75 \times 1024 = 65280$$

Table 7-15 lists the minimum operating system level required on z10 EC.

Table 7-15 Minimum support requirements for 63.75 K subchannels

Operating system	Support requirements
z/OS	z/OS V1R7
z/VM	z/VM V5R3
Linux on System z	Novell SUSE SLES 9 Red Hat RHEL 4

7.3.14 Multiple subchannel sets

Multiple subchannel sets, first introduced in z9 EC, provide a mechanism for addressing more than 63.75 K I/O devices and aliases for ESCON (CHPID type CNC) and FICON (CHPID types FCV and FC) on the z9 EC and z10 EC.

Multiple subchannel sets are not supported for z/OS running as a guest of z/VM.

Table 7-16 lists the minimum operating systems level required on the z10 EC.

Table 7-16 Minimum software requirement for MSS

Operating system	Support requirements
z/OS	z/OS V1R7
Linux on System z	Novell SUSE SLES 10 Red Hat RHEL 5

7.3.15 MIDAW facility

The modified indirect data address word (MIDAW) facility improves FICON performance. The MIDAW facility provides a more efficient CCW/IDAW structure for certain categories of data-chaining I/O operations.

Support for the MIDAW facility when running z/OS as a guest of z/VM requires z/VM V5R3 or higher. See 7.7, “MIDAW facility” on page 224.

Table 7-17 lists the minimum support requirements for MIDAW.

Table 7-17 Minimum support requirements for MIDAW

Operating system	Support requirements
z/OS	z/OS V1R7
z/VM	z/VM V5R3 for guest exploitation

7.3.16 Enhanced CPACF

Cryptographic functions are described in 7.4, “Cryptographic support” on page 217.

7.3.17 HiperSockets multiple write facility

This capability allows the streaming of bulk data over a HiperSockets link between two logical partitions. The key advantage of this enhancement is that it allows the receiving logical partition to process a much larger amount of data per I/O interrupt. Support for this function is required by the sending operating system. See 4.6.7, “HiperSockets” on page 145.

Table 7-18 Minimum support requirements for HiperSockets multiple write

Operating system	Support requirements
z/OS	z/OS V1R9 with PTFs

7.3.18 HiperSockets IPv6

IPv6 is expected to be a key element in future networking. The IPv6 support for HiperSockets permits compatible implementations between external networks and internal HiperSocket networks.

Table 7-19 lists the minimum support requirements for HiperSockets IPv6 (CHPID type IQD).

Table 7-19 Minimum support requirements for HiperSockets IPv6 (CHPID type IQD)

Operating system	Support requirements
z/OS	z/OS V1R7
z/VM	z/VM V5R3
Linux on System z	Novell SUSE SLES 10 SP2 Red Hat RHEL 5.2

7.3.19 HiperSockets Layer 2 support

For flexible and efficient data transfer for IP and non-IP workloads, the HiperSockets internal networks on System z10 EC can support two transport modes, which 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, which 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 as they do a non-mainframe environment.

Table 7-20 show the requirements for HiperSockets Layer 2 support.

Table 7-20 Minimum support requirements for HiperSockets Layer 2

Operating system	Support requirements
z/VM	z/VM V5R3 for guest exploitation
Linux on System z	Novell SUSE SLES 10 SP2 Red Hat RHEL 5.2

7.3.20 High performance FICON for System z10

High performance FICON for System z10 (zHPF) is a FICON architecture for protocol simplification and efficiency, reducing the number of information units (IUs) processed. Enhancements have been made to the z/Architecture and the FICON interface architecture to provide optimizations for on line transaction processing (OLTP) workloads.

When exploited by the FICON channel, the z/OS operating system, and the control unit (new levels of Licensed Internal Code are required) the FICON channel overhead can be reduced and performance can be improved. Additionally, the changes to the architectures provide end-to-end system enhancements to improve reliability, availability, and serviceability (RAS). Table 7-21 on page 209 lists the minimum support requirements for zHPF.

Table 7-21 Minimum support requirements for zHPF

Operating system	Support requirements
z/OS	z/OS V1R8 with PTFs
z/VM	Not supported; not available to guests
Linux	IBM is working with its Linux distribution partners so that exploitation of appropriate z10 BC functions be provided in future Linux on System z distribution releases.

The zHPF channel programs can be exploited by z/OS OLTP I/O workloads; DB2, VSAM, PDSE and zFS transfer small blocks of fixed size data (4 K). zHPF implementation, along with matching support by the DS8000 series, provides support for I/Os that transfer less than a single track of data as well as multitrack operations.

For more information about FICON channel performance, see the performance technical papers on the System z I/O connectivity Web site at:

http://www-03.ibm.com/systems/z/hardware/connectivity/ficon_performance.html

The zHPF is exclusive to System z10. The FICON Express8, FICON Express4³ and FICON Express2 features (CHPID type FC) concurrently support both the existing FICON protocol and the zHPF protocol in the server Licensed Internal Code.

FICON Express8

FICON Express8 is the newest generation of FICON features. They provide a link rate of 8 Gbps, with autonegotiation to 4 or 2 Gbps, for compatibility with previous devices and investment protection. Both 10KM LX and SX connections are offered (in a given feature all connections must have the same type).

With FICON Express 8 customers may be able to consolidate existing FICON, FICON Express2 and FICON Express4 channels, while maintaining and enhancing performance.

Table 7-22 lists the minimum support requirements for FICON Express8.

Table 7-22 Minimum support requirements for FICON Express8

Operating system	z/OS	z/VM	z/VSE	Linux on System z	z/TPF	TPF
Native FICON and Channel-to-Channel (CTC) CHPID type FC	V1R7	V5R3	V4R1	SUSE SLES 9 RHEL 4	V1R1	V4R1 PUT 16
zHPF single track operations CHPID type FC	V1R7 ^a	NA	NA	NA	NA	NA
zHPF multitrack operations CHPID type FC	V1R9 ^a	NA	NA	NA	NA	NA
Support of SCSI devices CHPID type FCP	NA	V5R3	V4R1	SUSE SLES 9 RHEL 4	NA	NA

a. PTFs required

7.3.21 FCP provides increased performance

The Fibre Channel Protocol (FCP) Licensed Internal Code has been modified to help provide increased I/O operations per second for both small and large block sizes and to support 8 Gbps link speeds.

For more information about FCP channel performance, see the performance technical papers on the System z I/O connectivity Web site at:

http://www-03.ibm.com/systems/z/hardware/connectivity/fcp_performance.html

7.3.22 Request node identification data

Request node identification data (RNID) for native FICON CHPID type FC allows isolation of cabling-detected errors on the z9 EC and z10 EC.

³ FICON Express4 10KM LX, 4KM LX and SX features are withdrawn from marketing. All FICON Express2 and FICON features are withdrawn from marketing.

Table 7-23 lists the minimum support requirements for RNID.

Table 7-23 Minimum support requirements for RNID

Operating system	Support requirements
z/OS	z/OS V1R7

7.3.23 FICON link incident reporting

FICON link incident reporting allows an operating system image (without operator intervention) to register for link incident reports. Table 7-24 lists the minimum support requirements for this function.

Table 7-24 Minimum support requirements for link incident reporting

Operating system	Support requirements
z/OS	z/OS V1R7

7.3.24 N_Port ID virtualization

N_Port ID virtualization (NPIV) provides a way to allow multiple system images (in logical partitions or z/VM guests) to use a single FCP channel as though each were the sole user of the channel. This feature, first introduced with z9 EC, can be used with earlier FICON features that have been carried forward from earlier servers.

Table 7-25 lists the minimum support requirements for NPIV.

Table 7-25 Minimum support requirements for NPIV

Operating system	Support requirements
z/VM	z/VM V5R3 provides support for guest operating systems and VM users to obtain virtual port numbers. Installation from DVD to SCSI disks is supported when NPIV is enabled.
z/VSE	z/VSE V4R1.
Linux on System z	Novell SUSE SLES 9 SP3. Red Hat RHEL 5.

7.3.25 VLAN management enhancements

Table 7-26 lists minimum support requirements for VLAN management enhancements for the OSA-Express2 and OSA-Express features (CHPID type OSD).

Table 7-26 Minimum support requirements for VLAN management enhancements

Operating system	Support requirements
z/OS	z/OS V1R7
z/VM	z/VM V5R3. Support of guests is transparent to z/VM if the device is directly connected to the guest (pass through).

7.3.26 OSA-Express3 10 Gigabit Ethernet LR and SR

The OSA-Express3 10 Gigabit Ethernet features offer two ports, defined as CHPID type OSD, supporting the queued direct input/output (QDIO) architecture for high-speed TCP/IP communication.

Table 7-27 lists the minimum support requirements for OSA-Express3 10 Gigabit Ethernet LR and SR features.

Table 7-27 Minimum support requirements for OSA-Express3 10 Gigabit Ethernet LR and SR

Operating system	Support requirements
z/OS	z/OS V1R7
z/VM	z/VM V5R3
z/VSE	z/VSE V4R1; service required
TPF and z/TPF	z/TPF V1R1 TPF V4R1 PUT 13; service required
Linux on System z	Novell SUSE SLES 9 Red Hat RHEL 4

7.3.27 OSA-Express3 Gigabit Ethernet LX and SX

The OSA-Express3 Gigabit Ethernet features offer two cards with two PCI Express adapters each. Each PCI Express adapter controls two ports, giving a total of four ports per feature. Each adapter has its own CHPID, defined as either OSD or OSN, supporting the queued direct input/output (QDIO) architecture for high-speed TCP/IP communication. Thus, a single feature can support both CHPID types, with two ports for each type.

Operating system support is required in order to recognize and use the second port on each PCI Express adapter. Minimum support requirements for OSA-Express3 Gigabit Ethernet LX and SX features are listed in Table 7-28 (four ports) and Table 7-29 on page 213 (two ports).

Table 7-28 Minimum support requirements for OSA-Express3 Gigabit Ethernet LX and SX, four ports

Operating system	Support requirements when using four ports
z/OS	z/OS V1R8; service required
z/VM	z/VM V5R3; service required
z/VSE	z/VSE V4R1; service required
z/TPF	z/TPF V1R1; service required (not supported by TPF V4R1)
Linux on System z	Novell SUSE SLES 10 SP2 Red Hat RHEL 5.2

Table 7-29 Minimum support requirements for OSA-Express3 Gigabit Ethernet LX and SX, two ports

Operating system	Support requirements when using two ports
z/OS	z/OS V1R7
z/VM	z/VM V5R3
z/VSE	z/VSE V4R1; service required
TPF and z/TPF	z/TPF V1R1 TPF V4R1 PUT 13; service required
Linux on System z	Novell SUSE SLES 9 Red Hat RHEL 4

7.3.28 OSA-Express3 1000BASE-T Ethernet

The OSA-Express3 1000BASE-T Ethernet features offer two cards with two PCI Express adapters each. Each PCI Express adapter controls two ports, giving a total of four ports for each feature. Each adapter has its own CHPID, defined as one of OSC, OSD, OSE or OSN. A single feature can support two CHPID types, with two ports for each type.

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

Operating system support is required in order to recognize and use the second port on each PCI Express adapter. Minimum support requirements for OSA-Express3 1000BASE-T Ethernet feature are listed in Table 7-30 (four ports) and Table 7-31 on page 214.

Table 7-30 Minimum support requirements for OSA-Express3 1000BASE-T Ethernet, four ports

Operating system	Support requirements when using four ports ^{a,b}
z/OS	OSD: z/OS V1R8; service required OSE: z/OS V1R7 OSN ^b : z/OS V1R7
z/VM	OSD: z/VM V5R3; service required OSE: z/VM V5R3 OSN ^b : z/VM V5R3
z/VSE	OSD: z/VSE V4R1; service required OSE: z/VSE V4R1 OSN ^b : z/VSE V4R1; service required
z/TPF	OSD and OSN ^b : z/TPF V1R1; service required
TPF	Not supported
Linux on System z	OSD: <ul style="list-style-type: none"> ▶ Novell SUSE SLES 10 SP2 ▶ Red Hat RHEL 5.2 OSN: <ul style="list-style-type: none"> ▶ Novell SUSE SLES 9 SP3 ▶ Red Hat RHEL 4.3,

a. Applies to CHPID types OSC, OSD, OSE and OSN. For support, see Table 7-31 on page 214.

b. Although CHPID type OSN does not use any ports (because all communication is LPAR to LPAR), it is listed here for completeness.

Table 7-31 Minimum support requirements for OSA-Express3 1000BASE-T Ethernet, two ports

Operating system	Support requirements when using two ports
z/OS	OSD, OSN, and OSE; z/OS V1R7
z/VM	OSD, OSN and OSE: z/VM V5R3
z/VSE	z/VSE V4R1
z/TPF	OSD, OSN, and OSC: z/TPF V1R1
TPF	OSD and OSC: TPF V4R1 PUT 13; service required
Linux on System z	OSD: <ul style="list-style-type: none"> ▶ Novell SUSE SLES 10 ▶ Red Hat RHEL 4 OSN: <ul style="list-style-type: none"> ▶ Novell SUSE SLES 9 SP3 ▶ Red Hat RHEL 4.3

7.3.29 GARP VLAN Registration Protocol

GARP⁴ VLAN Registration Protocol (GVRP) support allows an OSA-Express3 or OSA-Express2 port to register or unregister its VLAN IDs with a GVRP-capable switch and dynamically update its table as the VLANs change. Minimum support requirements are listed in Table 7-32.

Table 7-32 Minimum support requirements for GVRP

Operating system	Support requirements
z/OS	z/OS V1R7
z/VM	z/VM V5R3

7.3.30 OSA-Express3 and OSA-Express2 OSN support

Channel Data Link Control (CDLC), when used with the Communication Controller for Linux, emulates selected functions of IBM 3745/NCP operations. The port used with the OSN support appears as an ESCON channel to the operating system. This support can be used with OSA-Express3 GbE and 1000BASE-T, and OSA-Express2 GbE⁵ and 1000BASE-T features.

Table 7-33 lists the minimum support requirements for OSN.

Table 7-33 Minimum support requirements for OSA-Express3 and OSA-Express2 OSN

Operating system	OSA-Express3 and OSA-Express2 OSN
z/OS	z/OS V1R7
z/VM	z/VM V5R3

⁴ Generic Attribute Registration Protocol

⁵ OSA Express2 GbE is withdrawn from marketing.

Operating system	OSA-Express3 and OSA-Express2 OSN
z/VSE	z/VSE V4R1
Linux on System z	Novell SUSE SLES 9 SP3 Red Hat RHEL 4.3
TPF and z/TPF	TPF V4R1 and z/TPF V1R1

7.3.31 OSA-Express2 1000BASE-T Ethernet

The OSA-Express2 1000BASE-T Ethernet adapter can be configured in:

- ▶ QDIO mode, with CHPID type OSD or OSN
- ▶ Non-QDIO mode, with CHPID type OSE
- ▶ Local 3270 emulation mode with CHPID type OSC

Table 7-34 lists the support for OSA-Express2 1000BASE-T.

Table 7-34 Minimum support requirements for OSA-Express2 1000BASE-T

Operating system	CHPID type OSC	CHPID type OSD	CHPID type OSE
z/OS V1R7	Supported	Supported	Supported
z/VM V5R3	Supported	Supported	Supported
z/VSE V4R1	Supported	Supported	Supported
z/TPF V1R1	Supported	Supported	Not supported
TPF V4R1	Supported	PUT 13 plus PTFs	Not supported
Linux on System z	Not supported	Supported	Not supported

7.3.32 OSA-Express2 10 Gigabit Ethernet LR

Table 7-35 lists the minimum support requirements for OSA-Express2 10 Gigabit (CHPID type OSD).

Table 7-35 Minimum support requirements for OSA-Express2 10 Gigabit (CHPID type OSD)

Operating system	Support requirements
z/OS	z/OS V1R7
z/VM	z/VM V5R3
z/VSE	z/VSE V4R1
TPF and z/TPF	TPF V4R1 and z/TPF V1R1
Linux on System z	Novell SUSE SLES 9 Red Hat RHEL 4

7.3.33 Program directed re-IPL

Program directed re-IPL allows an operating system on a z9 EC or z10 EC to re-IPL without operator intervention. This function is supported for both SCSI and ECKD™ devices. Table 7-36 lists the minimum support requirements for program directed re-IPL.

Table 7-36 Minimum support requirements for program directed re-IPL

Operating system	Support requirements
z/VM	z/VM V5R3
Linux on System z	Novell SUSE SLES 9 SP3 Red Hat RHEL 4.5
z/VSE	V4R1 on SCSI disks

7.3.34 Coupling over InfiniBand

InfiniBand technology can potentially provide high-speed interconnection at short distances, longer distance fiber optic interconnection, and interconnection between partitions on the same system without external cabling. Several areas of this book discuss InfiniBand characteristics and support. For example, see 4.7, “Parallel Sysplex connectivity” on page 146.

InfiniBand coupling links

Table 7-37 lists the minimum support requirements for coupling links over InfiniBand.

Table 7-37 Minimum support requirements for coupling links over InfiniBand

Operating system	Support requirements
z/OS	z/OS V1R7
z/VM	z/VM V5R3 (dynamic I/O support for InfiniBand CHPIDs only; coupling over InfiniBand is not supported for guest use)
TPF and z/TPF	TPF V4R1 compatibility support only

InfiniBand coupling links at an unrepeated distance of 10 km

Support for HCA2-O LR fanout supporting InfiniBand coupling links (1x IB-SDR or 1x IB-DDR) at an unrepeated distance of 10 KM (6.2 miles) is listed on Table 7-38.

Table 7-38 Minimum support requirements for coupling links over InfiniBand at 10 km

Operating system	Support requirements
z/OS	z/OS V1R8; service required
z/VM	z/VM V5R3 (dynamic I/O support for InfiniBand CHPIDs only; coupling over InfiniBand is not supported for guest use)

7.3.35 Dynamic I/O support for InfiniBand CHPIDs

This function refers exclusively to the z/VM dynamic I/O support of InfiniBand coupling links. Support is available for the CIB CHPID type in the z/VM dynamic commands, including the `change channel path` dynamic I/O command. Specifying and changing the system name

when entering and leaving configuration mode is also supported. z/VM does not use InfiniBand and does not support the use of InfiniBand coupling links by guests.

Table 7-39 lists the minimum support requirements of dynamic I/O support for InfiniBand CHPIDs.

Table 7-39 Minimum support requirements for dynamic I/O support for InfiniBand CHPIDs

Operating system	Support requirements
z/VM	z/VM V5R3

7.4 Cryptographic support

z10 EC provides two major groups of cryptographic functions:

- ▶ Synchronous cryptographic functions, provided by the CP Assist for Cryptographic Function (CPACF)
- ▶ Asynchronous cryptographic functions, provided by the Crypto Express2 and Crypto Express3 features

The minimum software support levels are listed in the following sections. Obtain and review the most recent Preventive Service Planning (PSP) buckets to ensure that the latest support levels are known and included as part of the implementation plan.

7.4.1 CP Assist for Cryptographic Function

In z10 EC, the CP Assist for Cryptographic Function (CPACF) is extended to support the full standard for Advanced Encryption Standard (AES, symmetric encryption) and secure hash algorithm (SHA, hashing). For a full description, see 6.3, “CP Assist for Cryptographic Function” on page 177. Support for this function is provided through a Web deliverable. Table 7-40 lists the support requirements for enhanced CPACF.

Table 7-40 Support requirements for enhanced CPACF

Operating system	Support requirements
z/OS ^a	z/OS V1R7 and later: The function varies by release. Protected public key requires z/OS V1R9 and higher plus PTFs.
z/VM	z/VM V5R3 and higher: Supported for guest use. Protected public key not supported.
z/VSE	z/VSE V4R1 and later, and IBM TCP/IP for VSE/ESA V1R5 with PTFs
Linux on System z	Novell SUSE SLES 9 SP3, SLES 10 and SLES 11 Red Hat RHEL 4.3 and RHEL 5 The z10 EC CPACF enhancements can be used with: <ul style="list-style-type: none"> ▶ Novell SUSE SLES 10 SP2 and SLES 11 ▶ Red Hat RHEL 5.2
TPF and z/TPF	TPF V4R1 and z/TPF V1R1

a. CPACF is also exploited by several IBM Software product offerings for z/OS, such as IBM WebSphere Application Server for z/OS.

7.4.2 Crypto Express3 and Crypto Express2

Support of Crypto Express3 and Crypto Express2 functions varies by operating system and release. Table 7-41 lists the minimum software requirements for the Crypto Express3 and Crypto Express2 features when configured as a coprocessor or an accelerator. For a full description, see 6.4, “Crypto Express2” on page 178, and 6.5, “Crypto Express3” on page 182.

Table 7-41 *Crypto Express2 and Crypto Express3 support on z10 EC*

Operating system	Crypto Express3	Crypto Express2
z/OS	V1R11: Web deliverable V1R10: Web deliverable V1R9: Web deliverable V1R8: Not supported V1R7: Not supported	V1R11: Included in base V1R10: Included in base V1R9: Included in base V1R8: Included in base V1R7: Web deliverable
z/VM	V5R3: Service required; supported for guest use only	V5R3; supported for guest use only
z/VSE	V4R2 with IBM TCP/IP for VSE/ESA V1R5. Service required	V4R1 with IBM TCP/IP for VSE/ESA V1R5. Service required
Linux on System z	Note ^a Novell SUSE SLES 11 Novell SUSE SLES 10 SP3 Red Hat RHEL 5.4	Novell SUSE SLES 11 Novell SUSE SLES 10 Novell SUSE SLES 9 SP3 Red Hat RHEL 5.1 Red Hat RHEL 4.4
TPF V4R1	Not supported	Not supported
z/TPF V1R1	Service required (accelerator mode only)	Service required (accelerator mode only)

a. Support for Crypto Express3 is provided at the same functional level as for Crypto Express2

7.4.3 Web deliverables

For Web-delivered code on z/OS, see the z/OS downloads :

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

For Linux on System z, support is delivered through IBM and distribution partners. For more information see Linux on System z on the developerWorks Web site:

<http://www.ibm.com/developerworks/linux/linux390/>

7.4.4 z/OS ICSF FMIDs

Integrated Cryptographic Service Facility (ICSF) is a component of z/OS, and is designed to transparently use the available cryptographic functions, whether CPACF, Crypto Express2, or Crypto Express3 to balance the workload and help address the bandwidth requirements of the applications.

For a list of ICSF versions and FMID cross-references, see the Technical Documents page:

<http://www.ibm.com/support/techdocs/atmastr.nsf/WebIndex/TD103782>

Table 7-42 on page 219 lists the ICSF FMIDs and Web-delivered code for z/OS V1R7 through V1R10.

Table 7-42 z/OS ICSF FMIDs

z/OS	ICSF FMID ^a	Web deliverable name	Supported function
V1R7	HCR7731	Enhancements for cryptographic support for z/OS V1R6 and V1R7 (Web deliverable)	<ul style="list-style-type: none"> ▶ PCI-X Adapter Coprocessor and Accelerator ▶ CPACF enhancements ▶ Remote Key Loading ▶ ISO 16609 CBC Mode TDES MAC
	HCR7750	Enhancements for Cryptographic support for z/OS V1R7 through z/OS V1R9 (Web deliverable)	<ul style="list-style-type: none"> ▶ Cryptographic exploitation ▶ 4096-bit RSA keys ▶ CPACF support for SHA-384 and 512 ▶ Reduced support for retained private key in ICSF
V1R7 and V1R8	HCR7731	Enhancements for Cryptographic support for z/OS V1R6 and V1R7 (included in base)	<ul style="list-style-type: none"> ▶ PCI-X Adapter Coprocessor and Accelerator ▶ CPACF enhancements ▶ Remote Key Loading and ISO 16609 CBC Mode triple DES MAC
	HCR7750	Enhancements for Cryptographic Support for z/OS V1R7 through z/OS V1R9 (Web deliverable)	<ul style="list-style-type: none"> ▶ Cryptographic exploitation z10 BC ▶ 4096-bit RSA keys ▶ CPACF support for SHA-384 and 512 ▶ Reduced support for retained private key in ICSF
	HCR7751	Cryptographic Support for z/OS V1R8-V1R10 and z/OS.e V1R8 (Web deliverable)	<ul style="list-style-type: none"> ▶ Secure key AES ▶ 13 through 19-digit personal account number data ▶ Crypto Query service ▶ Enhanced SAF checking
V1R9	HCR7740	Cryptographic support for z/OS V1R7 through z/OS V1R9 (included in base)	<ul style="list-style-type: none"> ▶ Cryptographic toleration z10 BC
	HCR7750	Enhancements for Cryptographic support for z/OS V1R7 through z/OS V1R9 (Web deliverable)	<ul style="list-style-type: none"> ▶ Cryptographic exploitation z10 BC ▶ 4096-bit RSA keys ▶ CPACF support for SHA-384 and 512 ▶ Reduced support for retained private key in ICSF
	HCR7751	Cryptographic Support for z/OS V1R8 through V1R10 and z/OS.e V1R8 (Web deliverable)	<ul style="list-style-type: none"> ▶ Secure key AES ▶ 13 through 19-digit personal account number data ▶ Crypto Query service ▶ Enhanced SAF checking
	HCR7770	Cryptographic support for z/OS V1R9 through V1R11 (Web deliverable)	<ul style="list-style-type: none"> ▶ Protected key for CPACF ▶ Crypto Express3 and Crypto Express3-1P

z/OS	ICSF FMID ^a	Web deliverable name	Supported function
V1R10	HCR7750	Enhancements for Cryptographic support for z/OS V1R7 through z/OS V1R9 (included in base)	<ul style="list-style-type: none"> ▶ Cryptographic exploitation z10 BC ▶ 4096-bit RSA keys ▶ CPACF support for SHA-384 and 512 ▶ Reduced support for retained private key in ICSF
	HCR7751	Cryptographic Support for z/OS V1R8 through V1R11 and z/OS.e V1R8 (Web deliverable)	<ul style="list-style-type: none"> ▶ Secure key AES ▶ 13 through 19-digit personal account number data ▶ New Crypto Query service ▶ Enhanced SAF checking
	HCR7770	Cryptographic support for z/OS V1R9 through V1R11 (Web deliverable)	<ul style="list-style-type: none"> ▶ Protected key for CPACF ▶ Crypto Express3 and Crypto Express3-1P
V1R11	HCR7751	Cryptographic Support for z/OS V1R8 through V1R11 and z/OS.e V1R8 (included in base)	<ul style="list-style-type: none"> ▶ Secure key AES ▶ 13 through 19-digit personal account number data ▶ New Crypto Query service ▶ Enhanced SAF checking
	HCR7770	Cryptographic support for z/OS V1R9 through V1R11 (Web deliverable)	<ul style="list-style-type: none"> ▶ Protected key for CPACF ▶ Crypto Express3 and Crypto Express3-1P

a. PTF information is located in z10 EC PSP bucket: upgrade 2097DEVICE, subset 2097/ZOS.

Note the following FMID information:

- ▶ FMID HCR7730 is available as a Web download for z/OS V1R7
- ▶ FMID HCR7731 is available as a Web download for z/OS V1R8 in support of the PCI-X cryptographic coprocessor and accelerator functions, and the CPACF AES, PRNG, and SHA support.
- ▶ FMID HCR7740 is integrated in the base of z/OS V1R9, so no download is necessary.
- ▶ FMID HCR7750 must be downloaded and installed for support of the SHA-384 and SHA-512 function on z/OS V1R7, V1R8, and V1R9.
- ▶ FMID HCR7751, which is available for z/OS V1R8 and later, supports functions such as Secure Key AES, Crypto Query Service, enhanced IPv6 support, and enhanced SAF Checking and Personal Account Numbers with 13-19 digits.
- ▶ FMID HCR7770, with a planned availability of November 2009, supports Crypto Express3, Crypto Express3-1P and CPACF protected key on z/OS V1R9 and later.

7.4.5 ICSF migration considerations

Consider the following points about the Web-delivered ICSF code:

- ▶ Increased size of the PKDS file is required in order to allow 4096-bit RSA keys to be stored.

If you use the PKDS for asymmetric keys, copy your PKDS to a larger VSAM data set before using the new version of ICSF. The ICSF options file must be updated with the name of the new data set. ICSF can then be started.

A toleration PTF must be installed on any system that is sharing the PKDS with a system running HCR7750 ICSF. The PTF allows the PKDS to be larger and prevents any service from accessing 4096-bit keys stored in a HCR7750 PKDS.

- ▶ Support is reduced for retained private keys.

Applications that make use of the retained private key capability for key management are no longer able to store the private key in the cryptographic coprocessor card. The applications will continue to be able to list the retained keys and to delete them from the cryptographic coprocessor cards.

7.5 z/OS migration considerations

With the exception of base processor support, z/OS software changes do not require the new z10 EC functions. Equally, the new functions do not require functional software. The approach has been, where applicable, to let z/OS automatically decide to enable a function based on the presence or absence of the required hardware and software.

7.5.1 General recommendations

The IBM System z10 Enterprise Class introduces the latest System z technology. Although support is provided by z/OS starting with z/OS V1R7, exploitation of z10 EC is dependent on the z/OS release. The z/OS.e is *not* supported on z10 EC.

In general, we have the following recommendations:

- ▶ Do not migrate software releases and hardware at the same time.
- ▶ Keep members of sysplex at same software level, except during brief migration periods.
- ▶ Review z10 EC restrictions and migration considerations prior to creating an upgrade plan.

7.5.2 HCD

When using the hardware configuration definition (HCD) on z/OS V1R6 to create a definition for z10 EC, *all* subchannel sets must be defined or the VALIDATE task can fail. On z/OS V1R7, HCD or the Hardware Configuration Manager (HCM) assist in the definitions.

7.5.3 InfiniBand coupling links

Each system can use, or not use, InfiniBand coupling links independently of what other systems are doing, and do so in conjunction with other link types.

InfiniBand coupling connectivity can only be obtained with other systems that also support InfiniBand coupling.

7.5.4 Large page support

The large page support function is not be enabled without the software support. If large page is not specified, page frames are allocated at the current size of 4 K.

In z/OS V1R9 and later, the amount of memory to be reserved for large page support is defined by using parameter LFAREA in the IEASYSxx member of SYS1.PARMLIB, as follows:

```
LFAREA=xx%|xxxxxxM|xxxxxxG
```

The parameter indicates the amount of storage, in percentage, megabytes, or gigabytes. The value cannot be changed dynamically.

7.5.5 HiperDispatch

The HIPERDISPATCH=YES/NO parameter in the IEAOPTxx member of SYS1.PARMLIB and on the SET OPT=xx command can control 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 V1R7 (requires PTFs for zIIP support) through z/OS V1R10.

To effectively exploit HiperDispatch, the Workload Manager (WLM) goal adjustment might be required. We recommend that you review WLM policies and goals, and update them as necessary. You may want to run with the new policies and HiperDispatch on for a period, turn it off and use the older WLM policies while analyzing the results of using HiperDispatch, re-adjust the new policies and repeat the cycle, as needed. In order to change WLM policies, turning HiperDispatch off then on is not necessary.

A health check is provided to verify whether HiperDispatch is enabled on a system image that is running on z10 EC.

7.5.6 Capacity Provisioning Manager

Installation of the capacity provision function on z/OS requires:

- ▶ 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 because V1R7)
- ▶ Performing capacity provisioning customization as described in the publication *z/OS MVS Capacity Provisioning User's Guide*, SA33-8299

Exploitation of the capacity provisioning function requires:

- ▶ 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 has to be downloaded from the host and installed on a PC server. This application is only used to define policies. It is not required for regular operation.

Customization of the capacity provisioning function is required on the following systems:

- ▶ Observed z/OS systems. These are the systems in one or multiple sysplexes that are to be monitored. For a description of the capacity provisioning domain, see 8.8, “Nondisruptive upgrades” on page 273.
- ▶ 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.

7.5.7 Decimal floating point and z/OS XL C/C++ considerations

The following two C/C++ compiler options require z/OS V1R9:

- ▶ The ARCHITECTURE option, which selects the minimum level of machine architecture on which the program will run. Note that certain features provided by the compiler require a minimum architecture level. ARCH(8) exploits instructions available on the z10 EC.
- ▶ The TUNE option, which allows optimization of the application for a specific machine architecture, within the constraints imposed by the ARCHITECTURE option. The TUNE level must not be lower than the setting in the ARCHITECTURE option.

For more information about the ARCHITECTURE and TUNE compiler options, see the *z/OS V1R9.0 XL C/C++ User's Guide*, SC09-4767.

Note: A C/C++ program compiled with the ARCHITECTURE or TUNE options can run only on z10 EC servers, or an operation exception will result. This is a consideration for programs that may have to run on different level servers during development, test, production, and during fallback or DR.

7.6 Coupling facility and CFCC considerations

Coupling facility connectivity to a z10 EC is supported on the z10 BC, z9 EC, z9 BC, z990, z890, or another z10 EC. The logical partition running the Coupling Facility Control Code (CFCC) can reside on any of the supported servers previously listed. See Table 7-43 on page 224 for Coupling Facility Control Code requirements for supported servers.

Note: Because coupling link connectivity to z800 and z900 is *not* supported, this could affect the introduction of z10 EC into existing installations, and require additional planning. Also consider the level of CFCC. For more information, see “Coupling link migration considerations” on page 152.

The initial support of the CFCC on the z10 EC was level 15. CFCC level 16 is available and is exclusive to z10. CFCC level 16 offers the following enhancements:

► CF Duplexing enhancements

Prior to CFCC level 16, System-Managed CF Structure Duplexing required two protocol enhancements to occur synchronously to CF processing of the duplexed structure request. CFCC level 16 allows one of these signals to be asynchronous to CF processing. This enables faster service time, with more benefits because the distances between coupling facilities are further apart, such as in a multiple site Parallel Sysplex.

► List notification improvements

Prior to CFCC level 16, when a list changed state from empty to non-empty, it notified its connectors. The first one to respond would read the new message, but when the others read, they would find nothing, paying the cost for the *false scheduling*.

CFCC level 16 can help improve CPU utilization for IMS Shared Queue and WebSphere MQ Shared Queue environments. The coupling facility only notifies one connector in a round-robin fashion. If the shared queue is read within a fixed period of time, the other connectors do not have to be notified, saving the cost of the false scheduling. If a list is not read within the time limit, then the other connectors are notified as they are prior to CFCC level 16.

Although no significant increase in storage requirements is expected when moving to CFCC level 16, we strongly recommend using the CFSizer Tool, located on the Web at:

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

System z10 servers with CFCC level 16 require z/OS V1R7 or later, and z/VM V5R2 or later for guest virtual coupling.

A planned outage is required when migrating the CF or the CF LPAR to CFCC level 16.

Table 7-43 System z CFCC code level considerations

z10 EC	CFCC level 15 or later
z9 EC or z9 BC	CFCC level 14 or later
z990 or z890	CFCC level 13 or later

The current CFCC level for all System z9 servers is CFCC level 15. To support migration from one CFCC level to the next, different levels of CFCC can be run concurrently while the coupling facility logical partitions are running on different servers (CF logical partitions running on the same server share the same CFCC level).

For additional details about CFCC code levels, see the Parallel Sysplex Web site:

<http://www.ibm.com/systems/z/psocftable.html>

7.7 MIDAW facility

The modified indirect data address word (MIDAW) facility is a system architecture and software exploitation designed to improve FICON performance. This facility is available only on System z9 and System z10 servers and is exploited by the media manager in z/OS.

The MIDAW facility provides a more efficient CCW/IDAW structure for certain categories of data-chaining I/O operations:

- ▶ MIDAW can significantly improve FICON performance for extended format data sets. Non-extended data sets can also benefit from MIDAW.
- ▶ MIDAW can improve channel utilization and can significantly improve I/O response time. It reduces FICON channel connect time, director ports, and control unit overhead.

IBM laboratory tests indicate that applications using 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 ESCON channels configured as CHPID type CNS and on FICON channels configured as CHPID types FCV and FC.

7.7.1 MIDAW technical description

An indirect address word (IDAW) is used to specify data addresses for I/O operations in a virtual environment.⁶ 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 7-1 on page 225 shows a single channel command word (CCW) to control the transfer of data that spans non-contiguous 4 K frames in main storage. When the IDAW flag is set, the data address in the CCW points to a list of words (IDAWs), each of which contains an address designating a data area within real storage.

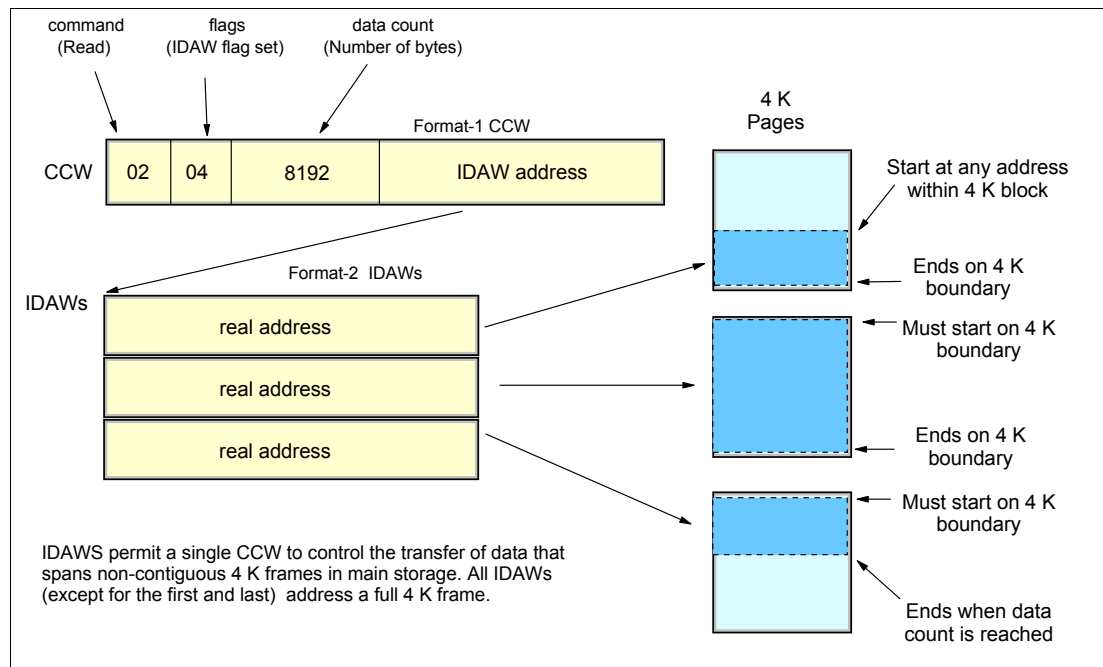


Figure 7-1 IDAW usage

⁶ There are exceptions to this statement and we skip a number of details in the following description. We assume that the reader can merge this brief description with an existing understanding of I/O operations in a virtual memory environment.

The number of IDAWs required for a CCW is determined by the IDAW format as specified in the operation request block (ORB), by the count field of the CCW, and by the data address in the initial IDAW. For example, three IDAWs are required when the following three events occur:

1. The ORB specifies format-2 IDAWs with 4 KB blocks.
2. The CCW count field specifies 8 KB.
3. The first IDAW designates a location in the middle of a 4 KB block.

CCWs with *data chaining* may 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 is sometimes known as scatter-read or scatter-write). However, as technology evolves and link speed increases, data chaining techniques are becoming less efficient in modern I/O environments for reasons involving switch fabrics, control unit processing and exchanges, and others.

The MIDAW facility is a method of gathering and scattering data from and into discontinuous storage locations during an I/O operation. The modified IDAW (MIDAW) format is shown in Figure 7-2. It is 16 bytes long and is aligned on a quadword.

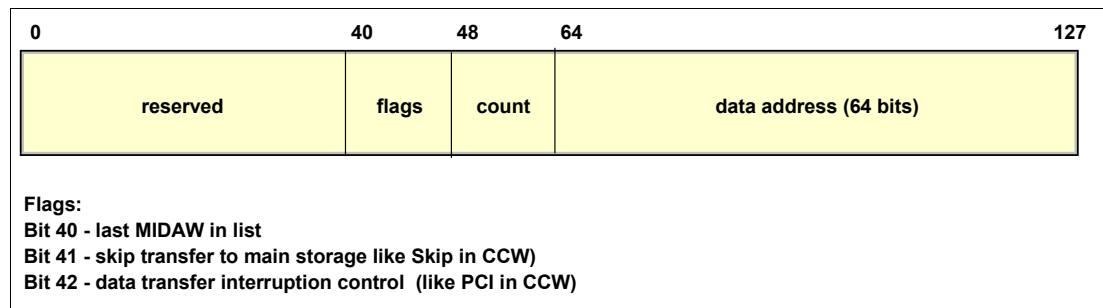


Figure 7-2 MIDAW format

An example of MIDAW usage is shown in Figure 7-3.

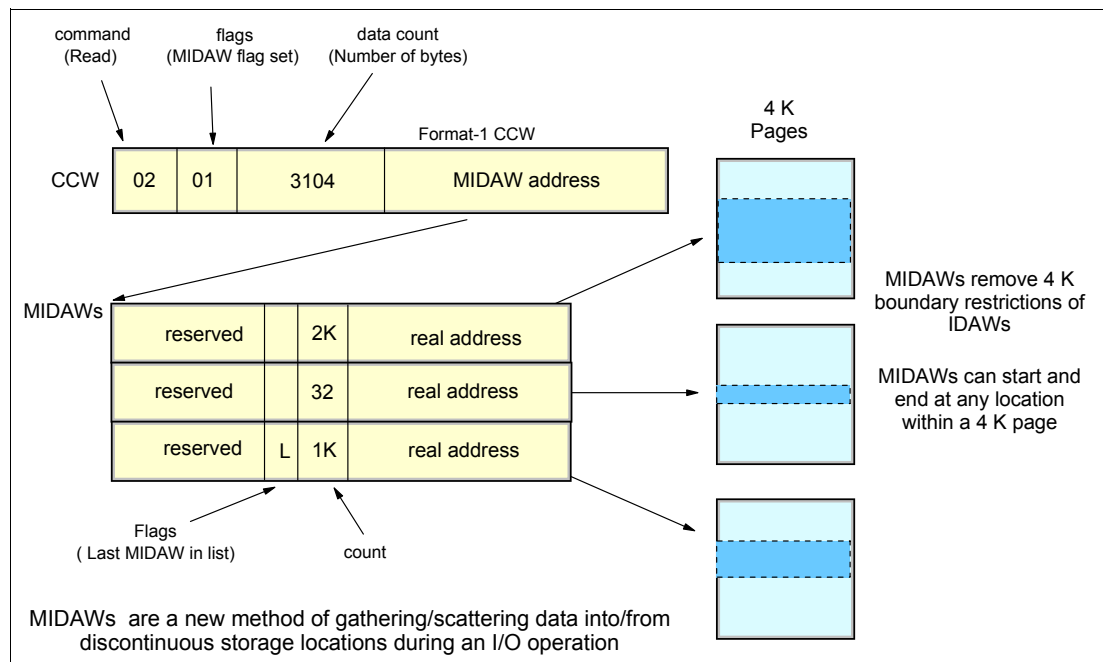


Figure 7-3 MIDAW usage

The use of MIDAWs is indicated by the MIDAW bit in the CCW. If this bit is set, then the *skip flag* cannot be set in the CCW. The skip flag in the MIDAW may be used instead. The data count in the CCW should equal the sum of the data counts in the MIDAWs. The CCW operation ends when the CCW count goes to zero or the last MIDAW (with the *last* flag) ends. The combination of the address and count in a MIDAW cannot cross a page boundary. This means that the largest possible count is 4 K. The maximum data count of all the MIDAWs in a list cannot exceed 64 K, which is the maximum count of the associated CCW.

The scatter-read or scatter-write effect of the MIDAWs makes it possible to efficiently send small control blocks embedded in a disk record to separate buffers from those used for larger data areas within the record. MIDAW operations are on a single I/O block, in the manner of data chaining. Do not confuse this operation with CCW *command* chaining.

7.7.2 Extended format data sets

z/OS extended format data sets use internal structures (usually not visible to the application program) that require scatter-read (or scatter-write) operation. This means that CCW data chaining is required and this produces less than optimal I/O performance. Because the most significant performance benefit of MIDAWs is achieved with extended format (EF) data sets, a brief review of the EF data sets is included here.

Both Virtual Storage Access Method (VSAM) and non-VSAM (DSORG=PS) can be defined as extended format data sets. In the case of non-VSAM data sets, a 32-byte suffix is appended to the end of every physical record (that is, block) on disk. VSAM appends the suffix to the end of every control interval (CI), which normally corresponds to a physical record (a 32 K CI is split into two records to be able to span tracks.) This suffix is used to improve data reliability and facilitates other functions described in the following paragraphs. Thus, for example, if the DCB BLKSIZE or VSAM CI size is equal to 8192, the actual block on DASD consists of 8224 bytes. The control unit itself does not distinguish between suffixes and user data. The suffix is transparent to the access method or database.

In addition to reliability, EF data sets enable three other functions:

- ▶ DFSMS striping
- ▶ Access method compression
- ▶ Extended addressability (EA)

EA is especially 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 especially useful for utilizing multiple channels in parallel for one data set. The DB2 logs are often striped to optimize the performance of DB2 sequential inserts.

To process an I/O operation to an EF data set would normally require at least two CCWs with data chaining. One CCW would be used for the 32-byte suffix of the EF data set. With MIDAW, the additional CCW for the EF data set suffix can be 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 would chain 12 CCWs together using data chaining. To read twelve 4 K records from an EF data set, 24 CCWs would be chained (two CCWs per 4 K record). Using Media Manager track-level command operations and MIDAWs, an entire track can be transferred using a single CCW.

7.7.3 Performance benefits

z/OS Media Manager has the I/O channel programs support for implementing Extended Format data sets and it automatically exploits MIDAWs when appropriate. Today, most disk I/Os in the system are generated using media manager.

Users of the Executing Fixed Channel Programs in Real Storage (EXCPVR) instruction *may* construct channel programs containing MIDAWs provided that they construct an IOBE with the IOBEMIDA bit set. Users of EXCP instruction *may not* construct channel programs containing MIDAWs

The MIDAW facility removes the 4 K boundary restrictions of IDAWs and, in the case of EF data sets, reduces the number of CCWs. Decreasing the number of CCWs helps to reduce the FICON channel processor utilization. Media Manager and MIDAWs do not cause the bits to move any faster across the FICON link, but they do reduce the number of frames and sequences flowing across the link, thus using the channel resources more efficiently.

Use of the MIDAW facility with FICON Express4, operating at 4 Gbps, compared to use of IDAWs with FICON Express2, operating at 2 Gbps, showed an improvement in throughput for all reads on DB2 table scan tests with EF data sets.

The performance of a specific workload can vary according to the conditions and hardware configuration of the environment. IBM laboratory tests found that DB2 gains significant performance benefits by using the MIDAW facility in the following areas:

- ▶ Table scans
- ▶ Logging
- ▶ Utilities
- ▶ Using 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 additional information relating to FICON and MIDAW, consult the following resources:

- ▶ The I/O Connectivity Web site contains the material about FICON channel performance:
<http://www.ibm.com/systems/z/connectivity/>
- ▶ The following publication:
DS8000 Performance Monitoring and Tuning, SG24-7146

7.8 IOCP

The required level of I/O configuration program (IOCP) for z10 EC is V2R1L0 (IOCP 2.1.0) or later.

7.9 Worldwide portname (WWPN) prediction tool

A part of the installation of your IBM System z10 server is the preplanning of the Storage Area Network (SAN) environment. IBM has made available a stand alone tool to assist with this planning prior to the installation.

The tool, known as the worldwide port name (WWPN) prediction tool, assigns WWPNs to each virtual Fibre Channel Protocol (FCP) channel/port using the same WWPN assignment algorithms a system uses when assigning WWPNs for channels utilizing N_Port Identifier Virtualization (NPIV). Thus, the SAN can be set up in advance, allowing operations to proceed much faster once the server is installed.

The WWPN prediction tool takes a .csv file containing the FCP-specific I/O device definitions and creates the WWPN assignments which are required to set up the SAN. A binary configuration file that can be imported later by the system is also created. The .csv file can either be created manually, or exported from the Hardware Configuration Definition/Hardware Configuration Manager (HCD/HCM).

The WWPN prediction tool on System z10 (CHPID type FCP) requires at a minimum:

- ▶ z/OS V1R8, V1R9, V1R10 and V1R11 with PTFs
- ▶ z/VM V5R3, V5R4 and V6R1 with PTFs

The WWPN prediction tool is available for download at Resource Link and is applicable to all FICON channels defined as CHPID type FCP (for communication with SCSI devices) on System z10.

<http://www.ibm.com/servers/resourceLink/>

7.10 ICKDSF

Device Support Facilities, ICKDSF, Release 17 is required on all systems that share disk subsystems with a z10 EC processor.

ICKDSF supports a modified format of the CPU information field, which contains a two-digit logical partition 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 any possible data corruption, ICKDSF must be able to determine all sharing systems that can potentially run ICKDSF. Therefore, this support is required for z10 EC.

Important: The need for ICKDSF Release 17 applies even to systems that are not part of the same sysplex, or that are running an operating system other than z/OS, such as z/VM.

7.11 Software licensing considerations

The IBM System z10 mainframe software portfolio includes operating system software (z/OS, z/VM, z/VSE, and z/TPF) and middleware that runs on these operating systems. It also includes middleware for Linux on System z environments. Two major metrics for software licensing are available from IBM, depending on the software product:

- ▶ Monthly License Charge (MLC)
- ▶ International Program License Agreement (IPLA)

The MLC pricing metrics have a recurring charge that applies each month. In addition to the right to use the product, the charge includes access to IBM product support during the support period. MLC metrics have several offerings that are applicable to the System z10 EC:

- ▶ Workload License Charge (WLC)
- ▶ System z New Application License Charge (zNALC)
- ▶ Parallel Sysplex License Charge (PSLC)
- ▶ Midrange Workload License Charge (MWLC)

IPLA metrics have an single, up-front, charge for an entitlement to use the product. Optionally, a separate annual charge called *subscription and support* entitles customers to receive future releases and versions at no additional charge, and also allows access to IBM product support during the support period.

For details, consult the *IBM System z Software Pricing Reference Guide*, G326-0594:

http://www.ibm.com/servers/eserver/zseries/library/refguides/sw_pricing.html

7.11.1 Workload License Charge

Workload License Charge (WLC) requires z/OS or z/TPF operating systems in 64-bit mode. Any mix of z/OS, z/VM, Linux on System z, VM/ESA, z/VSE, TPF, and z/TPF images is allowed.

The two WLC license types are:

- ▶ Flat WLC (FWLC)

Software products licensed under FWLC are charged at the same flat rate, no matter what capacity (MSUs) the server is.

- ▶ Variable WLC (VWLC)

Products such as z/OS, DB2, IMS, CICS, MQSeries®, and Lotus® Domino® can be charged in two different ways:

- Full-capacity is when the server's total number of MSUs is used for charging. Full-capacity is applicable when the server is not eligible for subcapacity.
- Subcapacity is when software charges are based on the logical partition's utilization where the product is running.

WLC subcapacity allows software charges based on logical partition utilizations instead of the server's total number of MSUs. Subcapacity removes the dependency between software charges and server (hardware) installed capacity.

Subcapacity is based on the logical partition's rolling 4-hour average utilization. It is *not* based on the utilization of each product⁷, but on the utilization of the logical partition or partitions where it runs. The VWLC licensed products running on a logical partition are charged by the maximum value of this partition's rolling 4-hour average utilization within a month.

The logical partition's rolling 4-hour average utilization can be limited by a *defined capacity* definition on the partition's image profiles. The defined capacity definition activates the *soft capping* function of PR/SM, avoiding 4-hour average partition utilizations above the defined capacity value. Soft capping controls the maximum rolling 4-hour average utilization (the last 4-hour average value at every five minutes interval), but does *not* control the maximum instantaneous partition utilization.

⁷ With the exception of products licensed using the Select Application License Charge (SALC) pricing metric.

Even by using the soft-capping option, the partition's utilization can reach its maximum share based on the number of logical processors and weights in the image profile. Only the rolling 4-hour average utilization is tracked, allowing utilization peaks above the defined capacity value.

As with the Parallel Sysplex License Charge (PSLC) software license charge type, the aggregation of servers' capacities within the same Parallel Sysplex is also possible in WLC, following the same prerequisites.

Entry Workload License Charge (EWLC) is not offered for IBM System z10 Enterprise Class.

For further information about WLC and details about how to combine logical partitions utilization, see *z/OS Planning for Workload License Charges*, SA22-7506.

7.11.2 System z New Application License Charge

System z New Application License Charge (zNALC) offers a reduced price for the z/OS operating system on logical partitions running a qualified new workload application such as Java language business applications running under WebSphere Application Server for z/OS, Domino, SAP, PeopleSoft, and Siebel.

z/OS with zNALC provides a strategic pricing model available on the full range of System z servers for simplified application planning and deployment. zNALC allows for aggregation across a qualified Parallel Sysplex, which can provide a lower cost for incremental growth across new workloads that span a Parallel Sysplex.

For additional information see the zNALC Web site:

<http://www.ibm.com/servers/eserver/zseries/swprice/zna1c.html>

7.11.3 Select Application License Charge

Select Application License Charge (SALC) applies only to WebSphere MQ for System z. It allows a WLC customer to license MQ under product utilization rather than the subcapacity pricing provided under WLC.

WebSphere MQ is typically a low-usage product that runs pervasively throughout the customer environment. Customers who run WebSphere MQ at a very low usage can benefit from SALC. Alternatively, one can still choose to license WebSphere MQ under WLC.

A reporting function, which IBM provides in the operating system IBM Software Usage Report Program, is used to calculate the daily MSU number. The rules to determine the billable SALC MSUs for WebSphere MQ use the following algorithm:

1. Determine the highest daily usage of a program⁸ family, which is the highest of 24 hourly measurements recorded each day.
2. Determine the monthly usage of a program family, which is the fourth highest daily measurement recorded for a month.
3. Use the highest monthly usage determined for the next billing period.

For additional information about SALC, see the MWLC Web site:

<http://www.ibm.com/servers/eserver/zseries/swprice/other.html>

⁸ The term *program* refers to all active versions of MQ.

7.11.4 Midrange Workload License Charge

Midrange Workload License Charge (MWLC) applies to z/VSE V4 when it is running on IBM System z10 and IBM System z9 servers. The exceptions are the z10 BC and z9 BC servers at capacity setting A01 to which zSeries Entry License Charge (zELC) applies. Similar to Workload License Charge, MWLC can be implemented in full-capacity or subcapacity mode. MWLC applies to z/VSE V4 and several IBM middleware products for z/VSE. All other z/VSE programs continue to be priced as before.

The z/VSE pricing metric is independent of the pricing metric for other systems, for instance, z/OS, that might be running on the same server. When z/VSE is running as a guest of z/VM, z/VM V5R3 or later is required.

The Subcapacity Report Tool (SCRT) is used to report utilization. One SCRT report is required for each server.

For additional information see the MWLC Web site:

<http://www.ibm.com/servers/eserver/zseries/swprice/mwlc.html>

7.11.5 System z International Licensing Agreement

On the mainframe, the following types of products are generally in the IPLA category:

- ▶ Data Management Tools
- ▶ CICS Tools
- ▶ Application Development Tools
- ▶ Certain WebSphere for System z products
- ▶ System z Linux middleware products
- ▶ z/VM Versions 5 and 6

For additional information, see the System z IPLA Web site:

<http://www.ibm.com/servers/eserver/zseries/swprice/zipla/>

7.12 References

For the most current planning information, see the support Web site for each of the following operating systems:

- ▶ z/OS
<http://www.ibm.com/systems/support/z/zos/>
- ▶ z/VM
<http://www.ibm.com/systems/support/z/zvm/>
- ▶ z/TPF
<http://www.ibm.com/software/http/tpf/pages/maint.htm>
- ▶ z/VSE
<http://www.ibm.com/servers/eserver/zseries/zvse/support/preventive.html>
- ▶ Linux on System z
<http://www.ibm.com/systems/z/os/linux/>



System upgrades

This chapter provides an overview of z10 EC upgrade capabilities and procedures, with an emphasis on Capacity on Demand offerings.

The upgrade offerings to the IBM System z10 EC servers have been developed from previous IBM System z servers. In response to customer demands and changes in market requirements, a number of features have been added. The changes and additions are designed to provide increased customer control over the capacity upgrade offerings with decreased administrative work and with enhanced flexibility. The provisioning environment gives the customer an unprecedented flexibility and a finer control over cost and value.

Given today's business environment, the benefits of the growth capabilities provided by the z10 EC are plentiful, and include, but are not limited to:

- ▶ Enabling exploitation of new business opportunities
- ▶ Supporting the growth of dynamic, on-demand environments
- ▶ Managing the risk of volatile, high-growth, and high-volume applications
- ▶ Supporting 24x365 application availability
- ▶ Enabling capacity growth during lock down periods
- ▶ Enabling planned-downtime changes without availability impacts

This chapter discusses the following topics:

- ▶ 8.1, "Upgrade types" on page 234
- ▶ 8.2, "Concurrent upgrades" on page 239
- ▶ 8.3, "MES upgrades" on page 246
- ▶ 8.4, "Permanent upgrade through the CIU facility" on page 251
- ▶ 8.5, "On/Off Capacity on Demand" on page 255
- ▶ 8.6, "Capacity for Planned Event" on page 266
- ▶ 8.7, "Capacity Backup" on page 268
- ▶ 8.8, "Nondisruptive upgrades" on page 273
- ▶ 8.9, "Summary of Capacity on Demand offerings" on page 278

For more information, see the following publications:

- ▶ *IBM System z10 Enterprise Class Capacity On Demand*, SG24-7504
- ▶ *System z10Capacity on Demand User's Guide*, SC28-6871

8.1 Upgrade types

Types of upgrades for a System z10 Enterprise Class are summarized in this section.

Permanent and temporary upgrades

In different situations, different types of upgrades are needed. After some time, depending on your growing workload, you might require more memory, additional I/O cards, or process more capacity. However, in certain situations, only a short-term upgrade is necessary to handle a peak workload, or to temporarily replace a server that is down during a disaster or data center maintenance. The z10 EC offers the following solutions for such situations:

► Permanent

- Miscellaneous equipment specification (MES)

The MES upgrade order is always performed by IBM personnel. The result can be either real hardware added to the server or installation of LIC configuration control (LICCC) to the server. In both cases, installation is performed by IBM personnel.

- Customer Initiated Upgrade (CIU)

Using the CIU facility for a given server requires that the online CoD buying feature (FC 9900) is installed on the server. The CIU facility supports LICCC upgrades only.

► 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 descriptions see 8.1.1, “Terminology related to CoD for System z10 servers” on page 235.

Note: The MES provides system upgrade that can result in more enabled processors, different CP capacity level, but also in additional books, memory, I/O drawers, and I/O cards (physical upgrade). Additional planning tasks are required for nondisruptive logical upgrades. MES is *ordered* through your IBM representative and *delivered* by IBM service personnel.

Concurrent and nondisruptive upgrades

Depending on the impact on system and application availability, upgrades can be classified as:

► Concurrent

In general, concurrency addresses the continuity of operations of the hardware part of an upgrade, for instance, whether a server (as a box) is required to be switched off during the upgrade. For details see 8.2, “Concurrent upgrades” on page 239.

► Non-concurrent

This type of upgrade requires the stopping the system (HW). Examples of such upgrades include model upgrade from any E12, E26, E40, E56 models to E64 model, certain physical memory capacity upgrades and adding I/O cages

► Disruptive

An upgrade is disruptive when resources added to an operating system image require that the operating system be recycled to configure the newly added resources.

► Nondisruptive

Nondisruptive upgrades do not require the running software or operating system to be restarted for the upgrade to take an effect. Thus, even concurrent upgrades can be disruptive to those operating systems or programs that do not support the upgrades while at the same time being nondisruptive to others. For details see 8.8, “Nondisruptive upgrades” on page 273.

8.1.1 Terminology related to CoD for System z10 servers

Table 8-1 briefly describes the most frequently used terms related to Capacity on Demand for System z10 servers.

Table 8-1 CoD terminology

Term	Description
Activated capacity	Capacity that is purchased and activated. Purchased capacity can be greater than activated capacity.
Billable capacity	Capacity that helps handle workload peaks, either expected or unexpected. The one billable offering available is On/Off Capacity on Demand.
Book	A physical package that contains memory, a Multi-Chip Module (MCM), and the memory bus adapters (MBAs). A book plugs into one of four slots in the central processor complex (CPC) cage of the z10 EC.
Capacity	Hardware resources (processor and memory) able to process workload can be added to the system through various capacity offerings.
Capacity Backup (CBU)	A function that allows the use of spare capacity in a CPC to replace capacity from another CPC within an enterprise, for a limited time. Typically, CBU is used when another CPC of the enterprise has failed or is unavailable because of a disaster event. The CPC using CBU replaces the missing CPC's capacity.
Capacity for planned event (CPE)	Used when temporary replacement capacity is needed for a short term event. CPE activate processor capacity temporarily to facilitate moving machines between data centers, upgrades, and other routine management tasks. CPE is an offering of Capacity on Demand.
Capacity levels	Can be full capacity or subcapacity. For the z10 EC server, capacity levels for the CP engine are 7, 6, 5, and 4: <ul style="list-style-type: none"> ► Full capacity CP engine is indicated by 7. ► Subcapacity CP engines are indicated by 6, 5, and 4.
Capacity setting	Derived from the capacity level and the number of processors. For the z10 EC server, the capacity levels are 7nn, 6xx, 5xx, 4xx, where xx or nn indicates the number of active CPs. The number of processors can have a range of: <ul style="list-style-type: none"> ► 0–64 for capacity level 7nn ► 1–12 for capacity levels 6xx, 5xx, 4xx
Concurrent book add (CBA)	Concurrently adds book hardware, including processors, physical memory, and I/O connectivity
Capacity Backup (CBU)	Provides reserved emergency backup processor capacity for unplanned situations when a loss of capacity occurs in another part of the enterprise

Term	Description
Central processor complex (CPC)	A physical collection of hardware that consists of main storage, one or more central processors, timers, and channels
Customer Initiated Upgrade (CIU)	A Web-based facility where you may request processor and memory upgrades by using the IBM Resource Link and the system's remote support facility (RSF) connection
Capacity on Demand (CoD)	The ability of a computing system to increase or decrease its performance capacity as needed to meet fluctuations in demand
Capacity Provisioning Manager (CPM)	As a component of z/OS Capacity Provisioning, CPM monitors business-critical workloads that are running on z/OS systems on IBM System z10 Enterprise Class servers.
Customer profile	This information resides on Resource Link and contains customer and machine information. A customer profile may contain information about more than one machine.
Enhanced book availability	In a multibook configuration, the ability to have a book concurrently removed from the server and reinstalled during an upgrade or repair action
Full capacity CP feature	For z10 EC feature (CP7), provides full capacity. Capacity settings 7xx are full capacity settings.
High water mark	Capacity purchased and owned by the customer
Installed record	The LICCC record has been downloaded, staged to the SE, and is now installed on the CPC. A maximum of eight different records can be concurrently installed and active.
Licensed Internal Code (LIC)	LIC is microcode, basic I/O system code, utility programs, device drivers, diagnostics, and any other code delivered with an IBM machine for the purpose of enabling the machine's specified functions.
LIC Configuration Control (LICCC)	Configuration control by the LIC to provides for server upgrade without hardware changes by enabling the activation of additional previously installed capacity
Multi-Chip Module (MCM)	An electronic package where multiple integrated circuits (semiconductor dies) and other modules are packaged on a common substrate to be mounted on a PCB (printed circuit board) as a single unit.
Model capacity identifier (MCI)	Shows the current active capacity on the server, including all replacement and billable capacity. For the z10 EC, the model capacity identifier is in the form of 7nn, 6xx, 5xx, or 4xx, where xx or nn indicates the number of active CPs. <ul style="list-style-type: none"> ▶ nn can have a range of 00 - 64. ▶ xx can have a range of 01-12. For the z10 BC the model capacity identifier is in the form of Axx - Zxx where xx indicates the number of active processors and can have a range of 01 - 05.
Model Permanent Capacity Identifier (MPCI)	Keeps information about capacity settings active before any temporary capacity was activated
Model Temporary Capacity Identifier (MTCI)	Reflects the permanent capacity with billable capacity only, without replacement capacity. If no billable temporary capacity is active, Model Temporary Capacity Identifier equals Model Permanent Capacity Identifier.
On/Off Capacity on Demand (CoD)	Represents a function that allows a spare capacity in a CPC to be made available to increase the total capacity of a CPC. For example, On/Off CoD may be used to acquire additional capacity for the purpose of handling a workload peak.

Term	Description
Permanent capacity	The capacity that a customer purchases and activates. This amount might be less capacity than the total capacity purchased.
Permanent upgrade	LIC licensed by IBM to enable the activation of applicable computing resources, such as processors or memory, for a specific CIU-eligible machine on a permanent basis
Purchased capacity	Capacity delivered to and owned by the customer. It can be higher than permanent capacity.
Permanent/Temporary entitlement record	The internal representation of a temporary (TER) or permanent (PER) capacity upgrade processed by the CIU facility. An entitlement record contains the encrypted representation of the upgrade configuration with the associated time limit conditions.
Replacement capacity	A temporary capacity used for situations in which processing capacity in other parts of the enterprise is lost during either a planned event or an unexpected disaster. The two replacement offerings available are, Capacity for Planned Events and Capacity Backup.
Resource Link	IBM Resource Link is a technical support Web site included in the comprehensive set of tools and resources available from the IBM Systems technical support site: http://www.ibm.com/servers/resourcecelink/
Secondary approval	An option, selected by the customer, that a second approver control each Capacity on Demand order. When a secondary approval is required, the request is sent for approval or cancellation to the Resource Link secondary user ID.
Staged record	The point when a record representing a capacity upgrade, either temporary or permanent, has been retrieved and loaded on the Support Element (SE) disk.
Subcapacity	For the z10 EC, CP features (CP4, CP5, and CP6) provide reduced capacity relative to the full capacity CP feature (CP7). For the z10 BC, CP features (CPA - CPY) provide reduced capacity relative to the full capacity CP feature (CPZ).
Temporary capacity	An optional capacity that is added to the current server capacity for a limited amount of time. It can be capacity that is owned or not owned by the customer.
Vital product data (VPD)	Information that uniquely defines system, hardware, software, and microcode elements of a processing system
Miscellaneous equipment specification (MES)	An upgrade process initiated through IBM representative and installed by IBM personnel

8.1.2 Permanent upgrades

Permanent upgrades can be:

- ▶ Ordered through an IBM sales representative
- ▶ Initiated by the customer with the Customer Initiated Upgrade (CIU) on IBM Resource Link

Note: The use of the CIU facility for a given server requires that the online CoD buying feature (FC 9900) is installed on the server. The CIU facility itself is enabled through FC 9898.

Permanent upgrades ordered through an IBM representative

Through a permanent upgrade you can:

- ▶ Add processor books.
- ▶ Add I/O cages 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 (re-characterization).

Attention: Most of the MES can be *concurrently* applied, without disrupting the existing workload (see 8.2, “Concurrent upgrades” on page 239, for details). However, certain MES changes are disruptive (for example, upgrade of models E12, E26, E40, and E56 to E64, or adding I/O cages).

Memory upgrades that require DIMM changes can be made nondisruptive if the flexible memory option is ordered.

Permanent upgrades initiated through CIU on 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, as follows:

- ▶ Add model capacity
- ▶ Add specialty engines
- ▶ Add memory
- ▶ Activate unassigned model capacity or IFLs
- ▶ Deactivate activated model capacity or IFLs

8.1.3 Temporary upgrades

System z10 EC offers three types of temporary upgrades:

- ▶ On/Off Capacity on Demand (On/Off CoD)
This offering allows you to temporarily add additional capacity or specialty engines due to seasonal activities, period-end requirements, peaks in workload, or application testing. This temporary upgrade can only be ordered using the CIU application through Resource Link.
- ▶ Capacity Backup (CBU)
This offering allows you to replace model capacity or specialty engines to a backup server in the event of an unforeseen loss of server capacity because of an emergency.
- ▶ Capacity for Planned Event (CPE)
This offering allows you to replace model capacity or specialty engines due to a relocation of workload during system migrations or a data center move.

CBU or CPE temporary upgrades can be ordered by using the CIU application through Resource Link or by calling your IBM sales representative.

Temporary upgrades capacity changes can be billable or replacement.

Billable capacity

To handle a peak workload, processors can be rented temporarily on a daily basis. You may activate up to double the purchased capacity of any PU type.

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 authorized limit.

The two replacement capacity offerings are:

- ▶ Capacity Backup
- ▶ Capacity for Planned Event

8.2 Concurrent upgrades

Concurrent upgrades on the IBM System z10 Enterprise Class can provide additional capacity with no server outage. In most cases, with prior planning and operating system support, a concurrent upgrade can also be nondisruptive to the operating system.

Given today's business environment, the benefits of the concurrent capacity growth capabilities provided by the z10 EC are plentiful, and include, but are not limited to:

- ▶ Enabling exploitation of new business opportunities
- ▶ Supporting the growth of e-business environments
- ▶ Managing the risk of volatile, high-growth, and high-volume applications
- ▶ Supporting 24x365 application availability
- ▶ Enabling capacity growth during *lock down* 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 additional configuration granularity within the family. The added granularity is available for models configured with up to 12 CPs and provides 36 additional capacity settings. Subcapacity models provide for CP capacity increase in two dimensions that can be used together to deliver configuration granularity. The first dimension is by adding CPs to the configuration, the second is by changing the capacity setting of the CPs currently installed to a higher model capacity identifier.

The z10 EC introduces a function that allows the concurrent addition of processors to a running logical partition. As a result, you can have a flexible infrastructure, in which you may add capacity without pre-planning. This function is supported by z/VM V5R3 (after you install the fixes to APAR VM64249, VM64323, and VM64389), and by z/VM V5R4. Planning ahead is required for z/OS logical partitions, as was the case before. To be able to add processors to a running z/OS, reserved processors must be specified in the logical partition's profile.

Another function concerns the system assist processor (SAP). When additional SAPs are concurrently added to the configuration, the SAP-to-channel affinity is dynamically re-mapped on all SAPs on the server to rebalance the I/O configuration.

8.2.1 Model upgrades

The z10 EC has a machine type and model, and model capacity identifiers:

- ▶ Machine type and model is 2097-Evv.

The *vv* can be 12, 26, 40, 56, or 64. The model number indicates how many PUs (*vv*) are available for customer characterization. Model E12 has one book installed, model E26 contains two books, model E40 contains three books, and models E56 and E64 contain four books.

- ▶ Model capacity identifiers are 4xx, 5xx, 6xx, or 7yy.

The *xx* is a range of 01 - 12 and *yy* is a range of 00 - 64. The model capacity identifier describes how many CPs are characterized (*xx* or *yy*) and the capacity setting (4, 5, 6, or 7) of the CPs.

A hardware configuration upgrade always requires additional physical hardware (books, cages, or both). A server upgrade can change either, or both, the server model and the model capacity identifier (MCI).

Note the following model upgrade information:

- ▶ LICCC upgrade
 - Does not change the server model 2097-Evv, because additional books are not added
 - Can change the model capacity identifier, the capacity setting, or both
- ▶ Hardware installation upgrade
 - Can change the server model 2097-Evv, if additional books are included
 - Can change the model capacity identifier, the capacity setting, or both

The server model and the model capacity identifier can be concurrently changed. Concurrent upgrades can be accomplished for both *permanent* and *temporary* upgrades.

Note: A model upgrade can be concurrent by using concurrent book add (CBA), except for upgrades to Model E64.

Licensed Internal Code upgrades (MES ordered)

The LIC Configuration Control (LICCC) provides for server upgrade without hardware changes by activation of additional (previously installed) unused capacity. Concurrent upgrades through LICCC can be done for:

- ▶ Processors (CPs, ICFs, zAAPs, zIIPs, IFLs, and SAPs) if unused PUs are available on the installed books or if the model capacity identifier for the CPs can be increased.
- ▶ Memory, when unused capacity is available on the installed memory cards. Plan-ahead memory and the flexible memory option are available for customers to gain better control over future memory upgrades. See 2.5.4, “Flexible memory option” on page 39, and 2.5.5, “Plan-ahead memory” on page 39 for more details.
- ▶ I/O card ports (ESCON channels and ISC-3 links), when there are available ports on the installed I/O cards.

Concurrent hardware installation upgrades (MES ordered)

Configuration upgrades can be concurrent when installing additional:

- ▶ Books (which contain processors, memory, MBAs, and HCA2s), when book slots are available in the CEC cage
- ▶ MBA fanouts for ICB4s
- ▶ HCA2 fanouts
- ▶ InfiniBand-Multiplexer (IFB-MP) cards
- ▶ I/O cards, when slots are still available on the installed I/O cages. I/O cages *cannot* be installed concurrently.

The concurrent I/O upgrade capability can be better exploited if a future target configuration is considered during the initial configuration. Using the plan-ahead concept, the required number of I/O cages for concurrent upgrades, up to the target configuration, can be included in the initial configuration.

Concurrent PU conversions (MES ordered)

The z10 EC supports concurrent conversion between all PU types, any-to-any PUs including SAPs, to provide flexibility to meet changing business requirements.

Note: The LICCC-based PU conversions require that at least one PU, either CP, ICF, or IFL, remains unchanged. Otherwise, the conversion is disruptive. The PU conversion generates a new LICCC that can be installed concurrently in two steps:

1. The assigned PU is removed from the configuration.
2. The newly available PU is activated as the new PU type.

Logical partitions might also have to *free* the PUs to be converted, and the operating systems must have support for configure offline or online so that performing the PU conversion can be done nondisruptively.

Note: Customer planning and operator action are required to exploit 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 logical partition outage if dedicated PUs are converted.

Unassigned CP capacity is recorded by a model capacity identifier. CP feature conversions change (increase or decrease) the model capacity identifier.

8.2.2 Customer Initiated Upgrade facility

The Customer Initiated Upgrade (CIU) facility is an IBM online system through which you may order, download, and install permanent and temporary upgrades for a System z server. Access to and use of the CIU facility requires a contract between the customer and IBM, through which the terms and conditions for use of the CIU facility are accepted. The use of the CIU facility for a given server requires that the online CoD buying feature code (FC 9900) is installed on the server. The CIU facility itself is controlled through FC 9898.

After you place an order through the CIU facility, you receive notice that the order is ready to download. You may then download and apply the upgrade by using functions available through the HMC, along with the remote support facility. After all the prerequisites are met, the entire process, from ordering to activation of the upgrade, is performed by the customer.

After the downloading, the actual upgrade process is fully automated and does not require any on-site presence of IBM service personnel.

CIU prerequisites

The CIU facility supports LICCC upgrades only. It does not support I/O upgrades. All additional capacity required for an upgrade must be previously installed. Additional books or I/O cards cannot be installed as part of an order placed through the CIU facility. The sum of CPs, unassigned CPs, ICFs, zAAPs, zIIPs, IFLs, and unassigned IFLs cannot exceed the PU count of the installed books. The total number of zAAPs or zIIPs cannot each exceed the number of purchased CPs.

CIU registration and agreed contract for CIU

To use the CIU facility, you must be registered and the system must be set up. After completing the CIU registration, access the CIU application through the IBM Resource Link Web site:

<http://www.ibm.com/servers/resourceLink/>

As part of the setup, you provide one resource link ID for configuring and placing CIU orders and, if required, a second ID as an approver. The IDs are then set up for access to the CIU support. The CIU facility is beneficial by allowing 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 Web site and invoke the CIU application to upgrade a server for processors, or memory. Requesting a customer order approval to conform to customer operation policies is possible. As previously mentioned, customers may allow the definition of additional 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 may generate online permanent upgrade orders to concurrently add processors (CPs, ICFs, zAAPs, zIIPs, IFLs, and SAPs) and memory, or change the model capacity identifier, up to the limits of the installed books on an existing server.

Temporary upgrades

The base model z10 EC describes permanent and dormant capacity (Figure 8-1) using the capacity marker and the number of PU features installed on the server. Up to eight temporary offerings can be present. Each offering has its own policies and controls and each can be activated or deactivated independently in any sequence and combination. Although multiple offerings can be active at any time, if enough resources are available to fulfill the offering specifications, only one On/Off CoD offering can be active at any time.

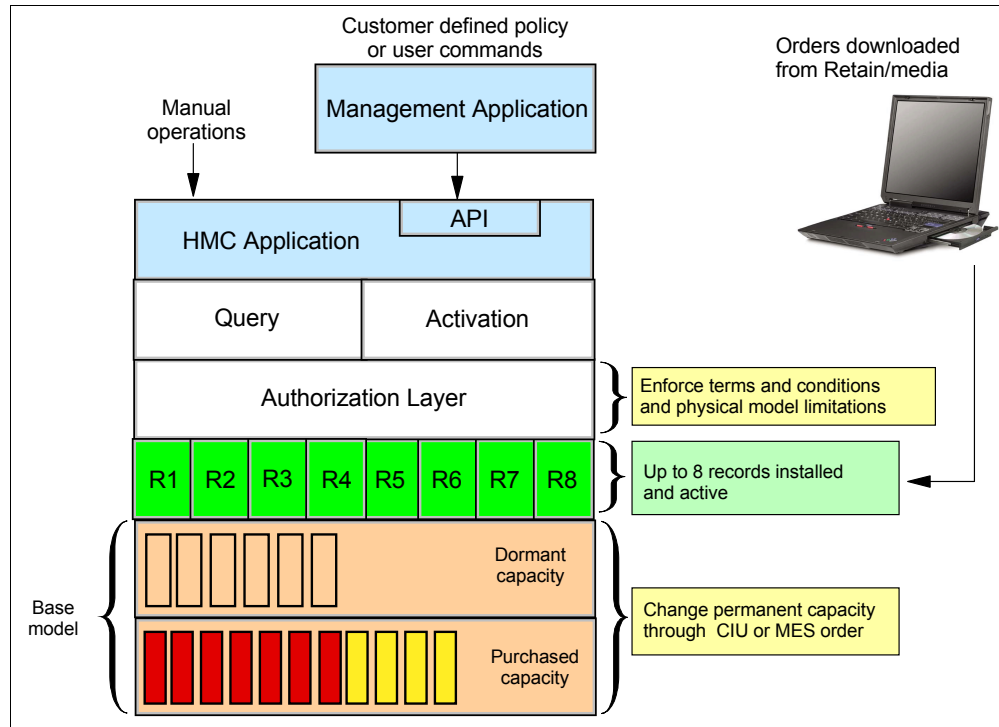


Figure 8-1 The provisioning architecture

Temporary upgrades are represented in the server by a *record*. All temporary upgrade records, downloaded from the remote support facility (RSF) or installed from portable media, are resident on the Service Element (SE) hard drive. At the time of activation, requiring a remote connection to IBM is no longer necessary. You may control everything locally. Figure 8-1 shows a representation of the provisioning architecture.

The authorization layer enables administrative control over the temporary offerings.

The activation and deactivation can be driven either manually or under control of an application through a documented application program interface (API).

By using the API approach, you may customize, at activation time, the resources necessary for responding to the current situation, up to the maximum specified at the time of order. If the situation changes, the you can add more or remove resources without having to go back to the base configuration. This eliminates the need for temporary upgrade specification for all possible scenarios. However, for CPE the ordered configuration is the only possible activation.

In addition, this approach enables you to update and replenish temporary upgrades, even in situations where the upgrades are already active. Likewise, depending on the configuration, permanent upgrades can be performed while temporary upgrades are active. Figure 8-2 shows examples of activation sequences of multiple temporary upgrades.

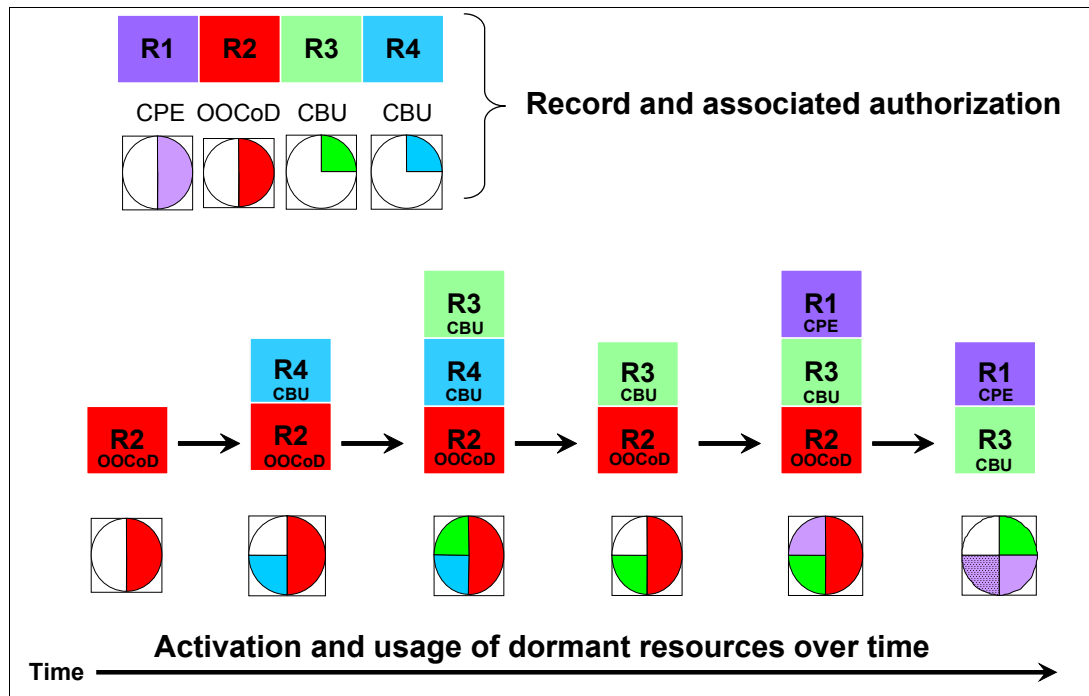


Figure 8-2 Example of temporary upgrade activation sequence

In the case of the R2, R3, and R1 being active at the same time, only parts of R1 can be activated, because not enough resources are available to fulfill all of R1. When R2 is then deactivated, the remaining parts of R1 may be activated as shown.

Temporary capacity can be billable as On/Off Capacity on Demand (On/Off CoD), or replacement as Capacity Backup (CBU) or CPE:

- On/Off CoD is a function that enables *concurrent* and *temporary* capacity growth of the server.

On/Off CoD *can* be used for customer peak workload requirements, for any length of time, and has a daily hardware and maintenance charge. The software charges can vary according to the license agreement for the individual products. See your IBM Software Group representative for exact details.

On/Off CoD can concurrently add processors (CPs, ICFs, zAAPs, zIIPs, IFLs, and SAPs), increase the model capacity identifier, or both, up to the limit of the installed books of an existing server, and is restricted to twice the currently installed capacity. On/Off CoD requires a contract agreement between the customer and IBM.

You decide whether to pre-pay or post-pay On/Off CoD. Capacity tokens inside the records are used to control activation time and resources.

- CBU is a *concurrent* and *temporary* activation of additional CPs, ICFs, zAAPs, zIIPs, IFLs, and SAPs, an increase of the model capacity identifier, or both.

CBU cannot be used for peak load management of customer workload or for CPE. A CBU activation can last up to 90 days when a disaster or recovery situation occurs.

CBU features are optional and require unused capacity to be available on installed books of the backup server, either as unused PUs or as a possibility to increase 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 server. The standard CBU contract provides for five 10-day tests and one 90-day disaster activation over a five-year period. Contact your IBM Representative for details.

- CPE is a *concurrent* and *temporary* activation of additional CPs, ICFs, zAAPs, zIIPs, IFLs, and SAPs or an increase of the model capacity identifier, or both.

The CPE offering is used to replace temporary lost capacity within a customer’s enterprise for planned downtime events, for example, with data center changes. CPE cannot be used for peak load management of customer workload or for a disaster situation.

The CPE feature requires unused capacity to be available on installed books of the backup server, either as unused PUs or as a possibility to increase the model capacity identifier on a subcapacity server, or both. A CPE contract must be in place before the special code that enables this capability can be loaded on the server. The standard CPE contract provides for one three-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, zAAPs, zIIPs, IFLs, SAPs, book, memory, and I/Os	Installed by IBM service personnel
	Online permanent upgrade	CPs, ICFs, zAAPs, zIIPs, IFLs, SAPs, and memory	Performed through CIU facility
Temporary	On/Off CoD	CPs, ICFs, zAAPs, zIIPs, IFLs, and SAPs	Performed through OOCOD facility
	CBU	CPs, ICFs, zAAPs, zIIPs, IFLs, and SAPs	Performed through CBU facility
	CPE	CPs, ICFs, zAAPs, zIIPs, IFLs, and SAPs	Performed through CPE facility

8.3 MES upgrades

Miscellaneous equipment specification (MES) upgrades enable *concurrent* and *permanent* capacity growth. MES upgrades allow the concurrent adding of processors (CPs, ICFs, zAAPs, zIIPs, IFLs, and SAPs), memory capacity, and I/O ports. Regarding subcapacity models, MES upgrades allow the concurrent adjustment of both the number of processors and the capacity level. The MES upgrade can be done using Licensed Internal Code Configuration Control (LICCC) only, by installing additional books, adding I/O cards, or a combination:

- ▶ MES upgrades for processors are done by any of the following methods:
 - LICCC assigning and activating unassigned PUs up to the limit of the installed books
 - LICCC to adjust the number and types of PUs or to change the capacity setting, or both
 - Installing additional books and LICCC assigning and activating unassigned PUs on installed books
- ▶ MES upgrades for memory are done by either of the following methods:
 - Using LICCC to activate additional memory capacity up to the limit of the memory cards on the currently installed books. Plan-ahead and flexible memory features enable you to have better control over future memory upgrades. For details about the memory features, see:
 - 2.5.5, “Plan-ahead memory” on page 39
 - 2.5.4, “Flexible memory option” on page 39
 - Installing additional books and using LICCC to activate additional memory capacity on installed books
 - Using the enhanced book availability (EBA), where possible, on multibook systems to add or change the memory cards
- ▶ MES upgrades for I/O are done by either of the following methods:
 - Using LICCC to activate additional ports on already installed ESCON and ISC-3 cards
 - Installing additional I/O cards and supporting infrastructure if required on I/O cages that are already installed

Important: If the STI rebalance feature (FC 2400) is selected at server upgrade configuration time, it will change the physical channel ID (PCHID) number of ICB-4 links, requiring a corresponding update on the server I/O definition through HCD or HCM.

An MES upgrade requires IBM service personnel for the installation. In most cases, the time required for installing the LICCC and completing the upgrade is short.

To better exploit the MES upgrade function, carefully plan the initial configuration to allow a concurrent upgrade to a target configuration.

By planning ahead, it is possible to enable nondisruptive capacity and I/O growth with no system power down and no associated PORs or IPLs. This enhancement is made possible by having a separate hardware system area (HSA).

In response to customer demands, the store system information (STSI) instruction has been changed to give more useful and detailed information about the base configuration and about temporary upgrades. The change enables you to more easily resolve billing situations where Independent Software Vendor (ISV) products are in use.

The model and model capacity identifier returned by the STSI instruction are updated to coincide with the upgrade. See “Store system information (STSI) instruction” on page 275 for more details.

Note: The MES provides the *physical* upgrade, resulting in more enabled processors, different capacity settings for the CPs, additional memory, and I/O ports. Additional planning tasks are required for *nondisruptive* logical upgrades (see “Recommendations to avoid disruptive upgrades” on page 277).

8.3.1 MES upgrade for processors

An MES upgrade for processors can *concurrently* add CPs, ICFs, zAAPs, zIIPs, IFLs, and SAPs to a z10 EC by assigning available PUs that reside on the books, through LICCC. Depending on the quantity of the additional processors in the upgrade, additional books might be required and can be concurrently installed before the LICCC is enabled. With the subcapacity models, additional capacity can be provided by adding CPs, by changing the capacity identifier on the current CPs, or by doing both.

Note: The sum of CPs, inactive CPs, ICFs, zAAPs, zIIPs, IFLs, unassigned IFLs, and SAPs cannot exceed the maximum limit of PUs available for customer use. The number of zAAPs or zIIPs cannot exceed the number of purchased CPs.

Example of MES upgrade

Figure 8-3 is an example of an MES upgrade for processors, showing two upgrade steps.

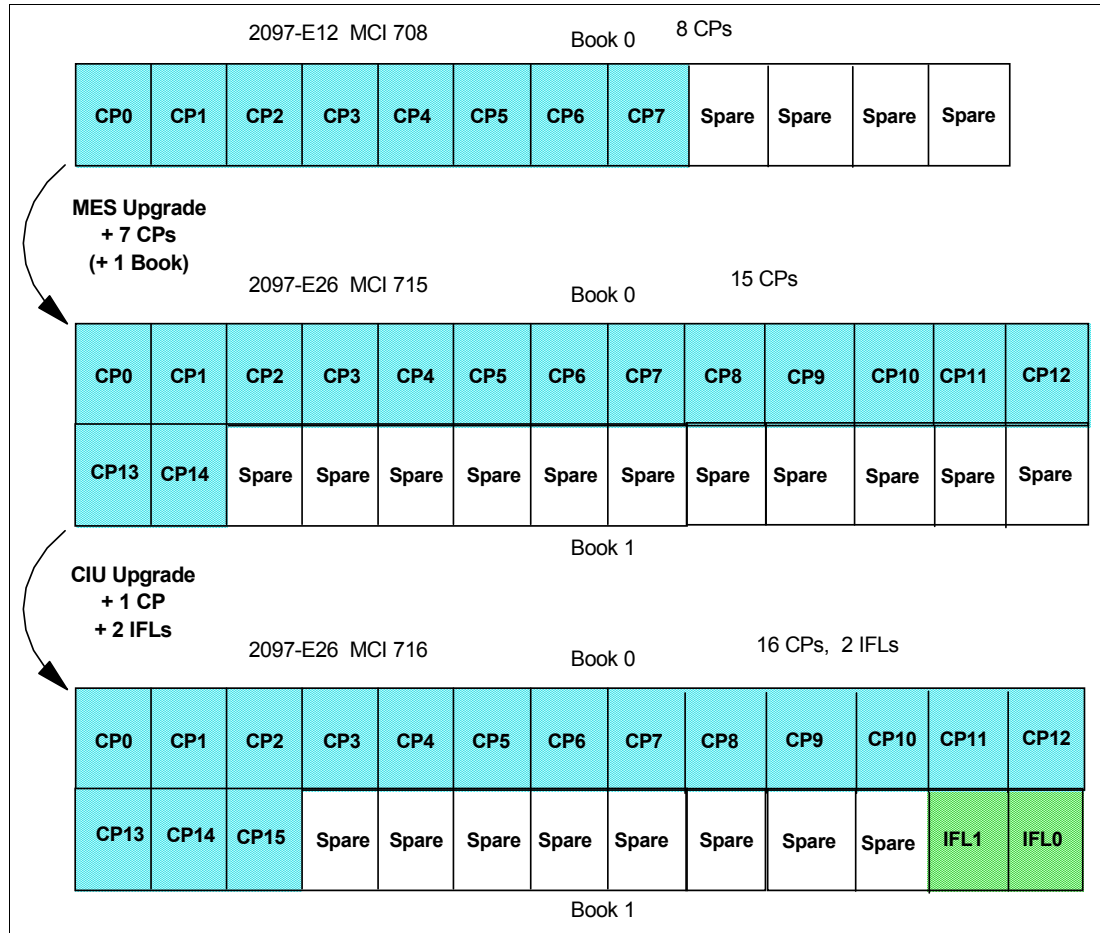


Figure 8-3 MES for processor example

A model E12 (one book), model capacity identifier 708 (eight CPs), is concurrently upgraded to a model E26 (two books), with model capacity identifier (MCI) 715 (which is 15 CPs). The model upgrade requires adding a book and assigning and activating seven PUs as CPs. Then, model E26, capacity identifier 715, is concurrently upgraded to a capacity identifier 716 (which is 16 CPs) with two IFLs by assigning and activating three more unassigned PUs (one as CP and two as IFLs). If needed, additional logical partitions can be created concurrently to use the newly added processors.

Note: Up to 64 logical processors, including reserved processors, can be defined to a logical partition. You should not define more processors to a logical partition than the target operating system supports:

- ▶ z/OS V1R7 supports up to 32 processors.
- ▶ z/OS V1R8 supports up to 54 processors.
- ▶ z/OS V1R9 supports up to 64 as a combination of CPs, zAAPs, and zIIPs.
- ▶ z/VM V5R3 and z/VM V5R4 support up to 32 processors of any type.

Software charges, based on the total capacity of the server on which the software is installed, are adjusted to the new capacity after the MES upgrade.

Software products that use Workload License Charge (WLC) might not be affected by the server upgrade, because their charges are based on partition utilization and not based on the server total capacity. For more information about WLC see 7.11.1, “Workload License Charge” on page 230.

8.3.2 MES upgrade for memory

MES upgrade for memory can *concurrently* add more memory by:

- ▶ Enabling, through LICCC, additional capacity up to the limit of the current installed memory cards
- ▶ Concurrently installing additional books and LICCC-enabling memory capacity on the new books.

Plan-ahead memory features are available to allow better control over future memory upgrades. See 2.5.4, “Flexible memory option” on page 39, and 2.5.5, “Plan-ahead memory” on page 39, for details about plan-ahead memory features.

If the z10 EC is a multiple-book configuration, using the enhanced book availability (EBA) feature to remove a book and add memory cards or to upgrade the already-installed memory cards to a larger size and then using LICCC to enable the additional memory is possible. With proper planning, additional memory can be added non-disruptively to z/OS partitions and z/VM V5R4 partitions. If necessary, new logical partitions can be created non-disruptively to use the newly added memory.

Note: Upgrades requiring DIMM changes can be concurrent by using the enhanced book availability feature. Planning is required to see whether this is a viable option in your configuration. The use of the flexible memory option (FC 1996) and the plan-ahead memory features (FC1991 and FC1992) is the safest way to ensure that EBA can work with the least disruption.

The one-book model has, as a minimum, sixteen 4 GB DIMMs, resulting in 64 GB of installed memory in total. With a fixed HSA size of 16 GB, which leaves up to 48 GB for customer use. If you have this configuration and have purchased 32 GB initially, a concurrent upgrade to 48 GB is possible through LICCC. If you require more than that, a *non-concurrent* upgrade can install up to 352 GB of memory for customer use, by changing the existing DIMM sizes and adding additional DIMMs in all available slots in the book. Another possibility is to add memory by *concurrently* adding a second book with sufficient memory into the configuration and then using LICCC to enable that memory.

A logical partition can dynamically take advantage of a memory upgrade if reserved storage has been defined to that logical partition. The reserved storage is defined to the logical partition as part of the image profile. Reserved memory can be configured online to the logical partition by using the LPAR dynamic storage reconfiguration (DSR) function. DSR allows a z/OS operating system image, and z/VM V5R4 (and higher) 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 V5R4 partition necessitates that pertinent operating system parameters have been prepared. If reserved storage has not been defined to the logical partition, the logical partition must be deactivated, the image profile changed, and the logical partition reactivated to allow 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 ports by either of the following methods:

- ▶ Enabling additional ports on the already installed I/O cards through LICCC
 - LICCC-only upgrades can be done for ESCON channels and ISC-3 links, activating ports on the existing 16-port ESCON or ISC-3 daughter (ISC-D) cards.
- ▶ Installing additional I/O cards on an already installed I/O cage's slots
 - The installed I/O cages must provide the number of I/O slots required by the target configuration.

Note: I/O cages *cannot* be installed concurrently.

Figure 8-4 shows a z10 EC that has 16 ESCON channels available on two 16-port ESCON channel cards installed in an I/O cage. Each channel card has eight ports enabled. In this example, eight additional ESCON channels are concurrently added to the configuration by enabling, through LICCC, four unused ports on each ESCON channel card.

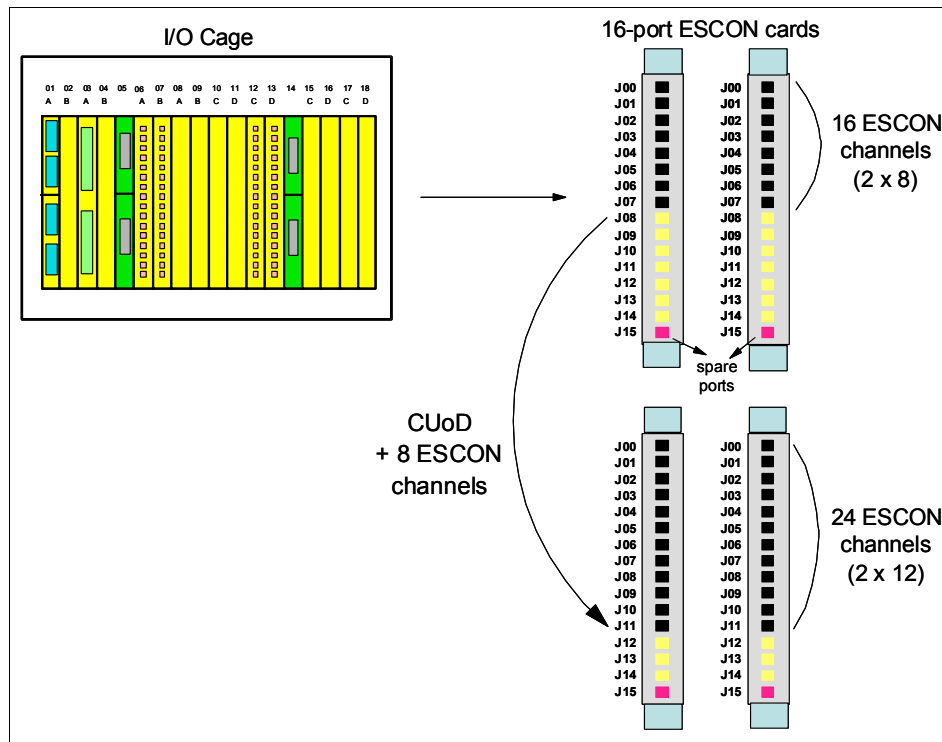


Figure 8-4 MES for I/O LICCC upgrade example

The additional channels installed concurrently to the hardware can also be concurrently defined in HSA and to an operating system by using the dynamic I/O configuration function. Dynamic I/O configuration can be used by z/OS or z/VM operating systems.

z/VSE, TPF, z/TPF, Linux on System z, and CFCC do *not* provide dynamic I/O configuration support. The installation of the new hardware is performed concurrently, but defining the new hardware to these operating systems requires an IPL.

To better exploit the MES for I/O capability, an initial configuration should be carefully planned to allow concurrent upgrades up to the target configuration. The plan-ahead concurrent

conditioning process can include, in the initial configuration, the shipment of additional I/O cages required for future I/O upgrades.

8.3.4 Plan-ahead concurrent conditioning

Concurrent Conditioning (FC 1999) and Control for Plan-Ahead (FC 1995) features, together with the input of a future target configuration, allow upgrades to exploit the order process configurator for concurrent I/O upgrades at a future time. If the initial configuration of a z10 EC can be installed with two power line-cords, order a plan-ahead feature for additional line-cords (FC 2000) if the future configuration will require additional power cords.

The plan-ahead feature identifies the content of the target configuration, which cannot be concurrently installed, thereby avoiding any down time associated with feature installation. As a result, Concurrent Conditioning may include, in the initial order, additional I/O cages to support the future I/O requirements.

Plan-ahead memory features enable you to install memory for future use:

- ▶ FC 1991 specifies memory to be installed but not used.
- ▶ FC 1992 is used to activate previously installed plan-ahead memory and can activate all the pre-installed memory or subsets of it.

Accurate planning and definition of the target configuration is vital to maximize the value of these features.

8.4 Permanent upgrade through the CIU facility

By using the CIU facility (through the IBM Resource Link on the Web), you may initiate a permanent upgrade for CPs, ICFs, zAAPs, zIIPs, IFLs, SAPs, or memory. When performed through the CIU facility, you add the resources; IBM personnel do not have to be present at the customer location. You may also unassign previously purchased CPs and IFLs processors through the CIU facility.

The capability to add permanent upgrades to a given server through the CIU facility requires that the permanent upgrade enablement feature (FC 9898) be installed on the server. A permanent upgrade might change the server model capacity identifier 4xx, 5xx, 6xx, or 7xx if additional CPs are requested or the capacity identifier is changed as part of the permanent upgrade, but it cannot change the server model, for example, 2097-Evv. If necessary, additional logical partitions can be created concurrently to use the newly added processors.

Note: A permanent upgrade of processors can provide a physical concurrent upgrade, resulting in more enabled processors available to a server configuration. Thus, additional planning and tasks are required for *nondisruptive* logical upgrades. See “Recommendations to avoid disruptive upgrades” on page 277 for more information.

Maintenance charges are automatically adjusted as a result of a permanent upgrade.

Software charges based on the total capacity of the server on which the software is installed are adjusted to the new capacity in place after the permanent upgrade is installed. Software products that use Workload License Charge (WLC) might not be affected by the server upgrade, because their charges are based on a logical partition utilization and not based on the server total capacity. See 7.11.1, “Workload License Charge” on page 230, for more information about WLC.

Figure 8-5 illustrates the CIU facility process on IBM Resource Link.

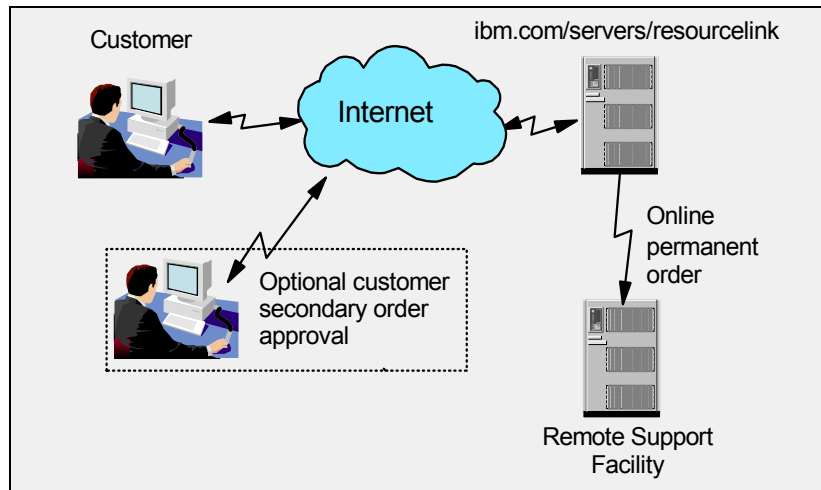


Figure 8-5 Permanent upgrade order example

The following sample sequence on IBM Resource Link initiates an order:

1. Sign on to Resource Link.
2. Select the **Customer Initiated Upgrade** option from the main Resource Link page. Customer and server details associated with the user ID are listed.
3. Select the server that will receive the upgrade. The current configuration (PU allocation and memory) is shown for the selected server.
4. Select **Order Permanent Upgrade** function. Resource Link limits options to those that are valid or possible for this configuration.
5. After the target configuration is verified by the system, accept or cancel the order. An order is created and verified against the pre-established agreement.
6. Accept or reject the price that is quoted. A secondary order approval is optional. Upon confirmation, the order is processed. The LICCC for the upgrade should 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 e-mail that the upgrade is ready to be downloaded.

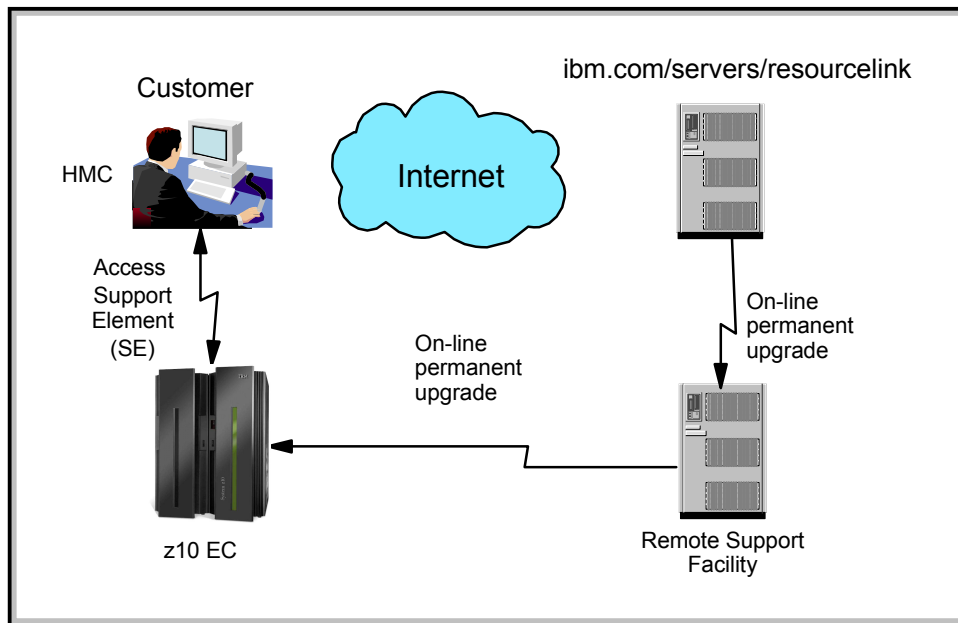


Figure 8-6 CIU-eligible order activation example

The two major components in the process are *ordering* and *retrieval* (and activation).

8.4.1 Ordering

Resource Link provides the interface that enables you to order a concurrent upgrade for a server. You may create, cancel, view the order, and view the history of orders that were placed through this interface. Configuration rules enforce only valid configurations being generated within the limits of the individual server. Warning messages are issued if you select invalid upgrade options. The process allows only one permanent CIU-eligible order for each server to be placed at a time.

Figure 8-7 shows the initial view of the machine profile on Resource Link.

Machine profile
2097 - RED03 - 1234567

	Current configuration	Ordered configuration
Model Capacity:	703 (3 CPs)	704 (4 CPs)
ICF:	4	4
zAAP:	2	2
zIIP:	2	2
IFL:	2	2
SAP:	6	6
Memory:	112	112
Unassigned IFLs:	0	0

Current configuration as of 20 Apr 2007 15:57:31

Machine summary
Type, model, serial: 2097 - E26 - RED03
System name: Not found
Model capacity downgraded from: 706 (6 CPs)

Customer summary
Company name: IBM
Customer number: 1234567
GEO, country: Americas - US

Ordering options
 → Order permanent upgrade
 → Order On/Off CoD record
 → Order On/Off CoD test record
 → Order Capacity Backup (CBU) record
 → Order Capacity for Planned Events (CPE) record
 → Display upgrade matrix

About ordering
Authorization to create orders
 User ID: redbook
 Name: Red Book
Permanent: Enabled
On/Off CoD: Enabled
CBU: Enabled
CPE: Enabled
Authorization to approve orders
 Not required
Notes:
 • A pre-negotiated price agreement exists for this machine.
 • A permanent upgrade order is currently being processed for this machine. You will not be able to create another permanent upgrade order until the billing has been completed for the current order.
 • On/Off CoD Test: 0 staged out of 1 remaining

Capacity on Demand records
 Record number - type - install state

Figure 8-7 CIU order example

The number of CPs, ICFs, zAAPs, zIIPs, IFLs, SAPs, memory size, CBU features, unassigned CPs, and unassigned IFLs on the current configuration are displayed on the left side of the Web page. On the right side are the corresponding updated values of the ordered configuration. This example requests an upgrade from three CPs (model capacity identifier 703) to four CPs (model capacity identifier 704).

Resource Link retrieves and stores relevant data 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 e-mail that contains an activation number. You may then retrieve the order by using the Perform Model Conversion task from the Support Element (SE), or through Single Object Operation to the SE from an HMC.

In the Perform Model Conversion panel, select the **Permanent upgrades** option to start the process. See Figure 8-8.

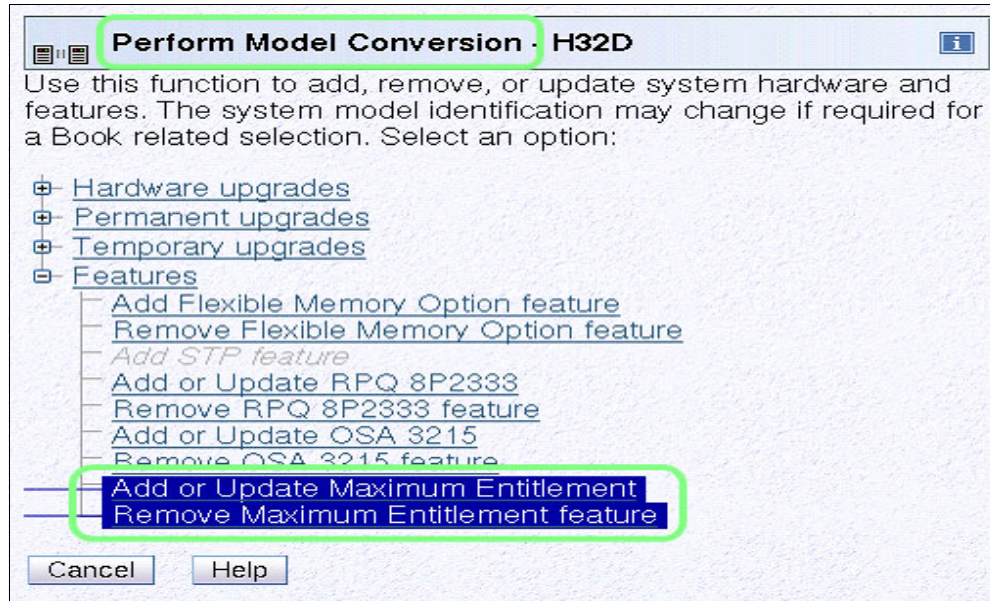


Figure 8-8 z10 EC Perform Model Conversion panel

The panel provides several possible options. If you select the **Retrieve and apply** data option, you are prompted to enter the order activation number to initiate the permanent upgrade. See Figure 8-9.

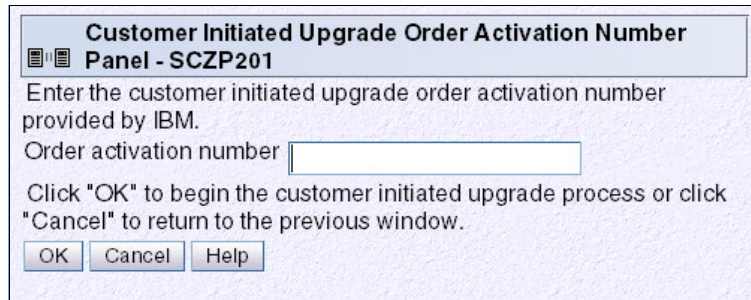


Figure 8-9 Customer Initiated Upgrade Order Activation Number Panel

8.5 On/Off Capacity on Demand

On/Off Capacity on Demand (On/Off CoD) allows you to temporarily enable PUs and unassigned IFLs available within the current model, or to change capacity settings for CPs to help meet your peak workload requirements.

8.5.1 Overview

The capacity for CPs is expressed in MSUs. 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 are represented as follows:

- ▶ For CP capacity, each token represents the amount of CP capacity that will result in one MSU of software cost for one day (an *MSU-day token*).
- ▶ For speciality engines, each token is equivalent to one speciality engine capacity for one day (an *engine-day token*).

Tokens are by capacity type, MSUs for CP capacity, and number of engines for speciality engines. Each speciality engine type has its own tokens, and each On/Off CoD record has separate token pools for each capacity type. During the ordering sessions on Resource Link, you decide how many tokens of each type should be created in an offering record. Each engine type must have tokens for that engine type to be activated. Capacity that has no tokens cannot be activated.

When resources from an On/Off CoD offering record containing capacity tokens are activated, a *billing window* is started. A billing window is always 24 hours in length. Billing takes place at the end of each billing window. The resources billed are the highest resource usage inside each billing window for each capacity type. An activation period is one or more complete billing windows, and represents the time from the first activation of resources in a record until the end of the billing window in which the last resource in a record is deactivated. At the end of each billing window, the tokens are decremented by the highest usage of each resource during the billing window. If any resource in a record does not have enough tokens to cover usage for the next billing window, the entire record will be deactivated.

On/Off CoD requires that the Online CoD Buying feature (FC 9900) be installed on the server that is to be upgraded.

The resources eligible for temporary use are CPs, ICFs, zAAPs, zIIPs, IFLs, and SAPs. Temporary addition of memory and I/O ports is not supported. Unassigned PUs that are on the installed books can be temporarily and concurrently activated as CPs, ICFs, zAAPs, zIIPs, IFLs, SAPs through LICCC, up to twice the currently installed CP capacity and up to twice the number of ICFs, zAAPs, zIIPs, or IFLs. This means that an On/Off CoD upgrade cannot change the server Model 2097-Evv. The addition of new books is not supported. However, activation of an On/Off CoD upgrade can increase the model capacity identifier 4xx, 5xx, 6xx, or 7xx.

8.5.2 Ordering

Concurrently installing temporary capacity by ordering On/Off CoD is possible, as follows:

- ▶ CP features equal to the MSU capacity of installed CPs
- ▶ IFL features up to the number of installed IFLs
- ▶ ICF features up to the number of installed ICFs
- ▶ zAAP features up to the number of installed zAAPs
- ▶ zIIP features up to the number of installed zIIPs
- ▶ SAPs up to three for model E12, seven for an E26, eleven for an E40, eighteen for an E56, and twenty-one for an E64.

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 or by changing the capacity setting of the CPs, or both. The capacity setting for all CPs must be

the same. If the On/Off CoD is adding CP resources that have a capacity setting different from the installed CPs, then 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:
 - Twice the currently installed capacity, expressed in MSUs for CPs
 - Twice the number of installed IFLs, ICFs, zAAPs, and zIIPs. The number of SAPs that can be activated depends on the model described in 8.2.1, “Model upgrades” on page 240.

Table 8-3 on page 258 shows the valid On/Off CoD configurations for CPs on the subcapacity models.

On/Off CoD can be ordered as prepaid or postpaid:

- ▶ A prepaid On/Off CoD offering record contains resource descriptions, MSUs, 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 has consumed 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 may contain capacity tokens describing MSU-days and speciality engine-days.
- ▶ When resources in a postpaid offering record without capacity tokens are activated, those resources remain active until they are deactivated, or until the offering record expires, which is usually 180 days after its installation.
- ▶ When resources in a postpaid offering record with capacity tokens are activated, those resources must have enough capacity tokens to allow the activation for an entire billing window (24 hours). The resources remain active until they are deactivated or until one of the resource tokens are consumed, or until the record expires, usually 180 days after its installation. If one capacity token type is consumed, resources from the entire record are deactivated.

As an example, for a z10 EC with capacity identifier 502, two ways to deliver a capacity upgrade through On/Off CoD exist:

- ▶ The first option is to add CPs of the same capacity setting. With this option, the model capacity identifier could be changed to a 503, which would add one additional CP (making a 3-way) or to a 504, which would add two additional CPs (making 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 could also be temporarily upgraded to a 603 as indicated in the table, thus increasing the capacity level and adding another processor. The 412 does not have an upgrade path through On/Off CoD.

We recommend that you use the Large Systems Performance Reference (LSPR) information to evaluate the capacity requirements according to your workload type. LSPR data for current IBM processors is available at:

<http://www.ibm.com/servers/eserver/zseries/lspr/>

Table 8-3 Valid On/Off CoD upgrade examples

Capacity identifier	On/Off CoD CP4	On/Off CoD CP5	On/Off CoD CP6	On/Off CoD CP7
401	402	-	-	-
402	403, 404	-	-	-
403	404, 405, 406	-	-	-
404	405 - 408	-	-	-
405	406 - 410	-	-	-
406	407 - 412	-	-	-
407	408 - 412	-	-	-
408	409 - 412	-	-	-
409	410 - 412	-	-	-
410	411, 412	-	-	-
411	412	-	-	-
412	-	-	-	-
501	-	502	601	701
502	-	503, 504	602, 603	702
503	-	504, 505, 506	603, 604	703
504	-	505 - 508	604 - 606	704
505	-	506 - 511	605 - 607	705
506	-	507 - 512	606 - 609	706
507	-	508 - 512	607 - 611	707
508	-	509 - 512	608 - 612	708
509	-	510 - 512	609 - 612	709
510	-	511 - 512	610 - 612	710
511	-	512	611, 612	711
512	-	-	612	712
601	-	-	602	701
602	-	-	603, 604	702
603	-	-	604, 605, 606	703, 704
604	-	-	605 - 608	704, 705
605	-	-	606 - 611	705 - 707

Capacity identifier	On/Off CoD CP4	On/Off CoD CP5	On/Off CoD CP6	On/Off CoD CP7
606	-	-	607 - 612	706 - 708
607	-	-	608 - 612	707 - 710
608	-	-	609 - 612	708 - 712
609	-	-	610 - 612	709 - 713
610	-	-	611, 612	710 - 715
611	-	-	612	711 - 717
612	-	-	-	712 - 718
701	-	-	-	702
702	-	-	-	703, 704
703	-	-	-	704, 705, 706
704	-	-	-	705 - 708
705	-	-	-	706 - 711
706	-	-	-	707 - 714
707	-	-	-	708 - 716
708	-	-	-	709 - 719
709	-	-	-	710 - 721
710	-	-	-	711 - 724
711	-	-	-	712 - 726
712	-	-	-	713 - 728

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 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 z10 EC and sent back to the support systems.

If On/Off capacity is already active, additional On/Off capacity can be added without having to return the server to its original capacity. If the capacity is increased multiple times within a 24-hour period, the charges apply to the highest amount of capacity active in the period. If additional capacity is added from an already active record containing capacity tokens, a check is made to control that the resource in question has enough capacity to be active for an entire billing window (24 hours). If that criteria is not met, no additional resources will be activated from the record.

If necessary, additional logical partitions can be activated concurrently to use the newly added processor resources.

Note: On/Off CoD provides a concurrent *hardware* upgrade, resulting in more enabled processors available to a server configuration. Additional planning tasks are required for *nondisruptive* upgrades. See “Recommendations to avoid disruptive upgrades” on page 277.

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 server. Subsequently, you may concurrently install temporary capacity up to the limits in On/Off CoD and use it for up to 180 days. Monitoring occurs through the server call-home facility and an invoice is generated if the capacity has been enabled during the calendar month. The customer will continue to be billed for use of temporary capacity until the server is returned to the original configuration. If the On/Off CoD support is no longer needed, the enablement code must be removed.

On/Off CoD orders can be pre-staged in Resource Link to allow multiple optional configurations. The pricing of the orders is done at the time of the order, and the pricing can vary from quarter to quarter. Staged orders can have different pricing. When the order is downloaded and activated, the daily costs are based on the pricing at the time of the order. The staged orders do not have to be installed in order sequence. If a staged order is installed out of sequence, and later an order that was staged that had a higher price is downloaded, the daily cost will be based on the lower price.

Another possibility is to store unlimited On/Off CoD LICCC records on the Support Element with the same or different capacities at any given time, giving greater flexibility to quickly enable needed temporary capacity. Each record is easily identified with descriptive names, and you may select from a list of records that can be activated.

Resource Link provides the interface that allows you to order a dynamic upgrade for a specific server. You are able to create, cancel, and view the order. Configuration rules are enforced, and only valid configurations are generated based on the configuration of the individual server. After completing the prerequisites, orders for the On/Off CoD can be placed. The order process is to use the CIU facility on Resource Link.

You may order temporary capacity for CPs, ICFs, zAAPs, zIIPs, IFLs, or SAPs. Memory and channels are not supported on On/Off CoD. The amount of capacity is based on the amount of owned capacity for the different types of resources. An LICCC record is established and staged to Resource Link for this order. After the record is activated, it has no expiration date. However, an individual record can only be activated once. Subsequent sessions require a new order to be generated, producing a new LICCC record for that specific order.

8.5.3 On/Off CoD testing

Each On/Off CoD-enabled server is entitled to one no-charge 24-hour test. No IBM charges are assessed for the test, including no IBM charges associated with temporary hardware capacity, IBM software, or IBM maintenance. The test can be used to validate the processes to download, stage, install, activate, and deactivate On/Off CoD capacity.

This test can last up to a maximum duration of 24 hours, commencing upon the activation of any capacity resource contained in the On/Off CoD record. Activation levels of capacity can change during the 24-hour test period. The On/Off CoD test automatically terminates at the end of the 24-hour period.

Figure 8-10 is an example of an On/Off CoD order on the Resource Link Web page.

The screenshot shows the IBM Resource Link web interface for configuring an On/Off CoD order. The page title is "Order On/Off CoD record" and it is "Step 1 of 2: Configure the record". The breadcrumb trail is "Machine profiles > Machine 2097 - RED01 >".

Machine summary:

Type:	2097 E26
Model:	703
Serial number:	RED01

Supported upgrades:

- Show upgrades
- Show upgrade prices

Replenishment due date* 07/27/2008 (mm/dd/yyyy)

Enable upgrades for up to:

Model capacity	93%	more model capacity
ICF	4	more ICF engines
zAAP	2	more zAAP engines
zIIP	2	more zIIP engines
IFL	2	more IFL engines
SAP	6	more SAP engines

[Continue](#)

Figure 8-10 On/Off CoD order example

The example order in Figure 8-10 is a On/Off CoD order for 93% more CP capacity and for four ICFs, two zAAPs, two zIIPs, two IFLs, and six SAPs. The maximum number of CPs, ICFs, zAAPs, zIIPs, and IFLs is limited by the current number of available unused PUs of the installed books. The maximum number of SAPs is determined by the model number and the number of available PUs on the already installed books.

8.5.4 Activation and deactivation

When a previously ordered On/Off CoD is retrieved from Resource Link, it is downloaded and stored on the SE hard disk. You may activate the order when the capacity is needed, either manually or through automation.

If the On/Off CoD offering record does not contain resource tokens, you must take action to deactivate the temporary capacity. Deactivation is accomplished from the Support Element and is nondisruptive. Depending on how the additional capacity was added to the logical partitions, you might be required to perform tasks at the logical partition level in order to remove the temporary capacity. For example, you might have to configure offline CPs that had been added to the partition, or deactivate additional logical partitions created to use the temporary capacity, or both.

On/Off CoD orders can be staged in Resource Link so that multiple orders are available. An order can only be downloaded and activated one time. If a different On/Off CoD order is required or a permanent upgrade is needed, it can be downloaded and activated without having to restore the system to its original purchased capacity.

In support of automation, an API is provided that allows the activation of the On/Off CoD records. The activation is performed from the HMC and requires specifying the order number. With this API, automation code can be used to send an activation command along with the order number to the HMC to enable the order.

8.5.5 Termination

A customer is contractually obligated to terminate the On/Off CoD right-to-use feature when a transfer in asset ownership occurs. A customer may also choose to terminate the On/Off CoD right-to-use feature without transferring ownership. Application of FC 9898 terminates the right to use the On/Off CoD. This feature cannot be ordered if a temporary session is already active. Similarly, the CIU enablement feature cannot be removed if a temporary session is active. Any time the CIU enablement feature is removed, the On/Off CoD right-to-use is simultaneously removed. Reactivating the right-to-use feature subjects the customer to the terms and fees that apply at that time.

Upgrade capability during On/Off CoD

Upgrades involving physical hardware are supported while an On/Off CoD upgrade is active on a particular z10 EC. LICCC-only upgrades can be ordered and retrieved from Resource Link and applied while an On/Off CoD upgrade is active. LICCC-only memory upgrades can be retrieved and applied while a On/Off CoD upgrade is active.

Repair capability during On/Off CoD

If the z10 EC requires service while an On/Off CoD upgrade is active, the repair can take place without affecting the temporary capacity.

Monitoring

When you activate an On/Off CoD upgrade, an indicator is set in vital product data. This indicator is part of the call-home data transmission, which is sent on a scheduled basis. A time stamp is placed into call-home data when the facility is deactivated. At the end of each calendar month, the data is used to generate an invoice for the On/Off CoD that has been used during that month.

Maintenance

The maintenance price is adjusted as a result of an On/Off CoD activation.

Software

Software Parallel Sysplex License Charge (PSLC) customers are billed at the MSU level represented by the combined permanent and temporary capacity. All PSLC products are billed at the peak MSUs enabled during the month, regardless of usage. Customers with WLC licenses are billed by product at the highest four-hour rolling average for the month. In this instance, temporary capacity does not necessarily increase the software bill until that capacity is allocated to logical partitions and actually consumed.

Results from the STSI instruction reflect the current permanent and temporary CPs. See “Store system information (STSI) instruction” on page 275 for more details.

8.5.6 z/OS capacity provisioning

The z10 EC provisioning capability combined with CPM functions in z/OS provides a flexible, automated process to control the activation of On/Off Capacity on Demand. The z/OS provisioning environment is shown in Figure 8-11.

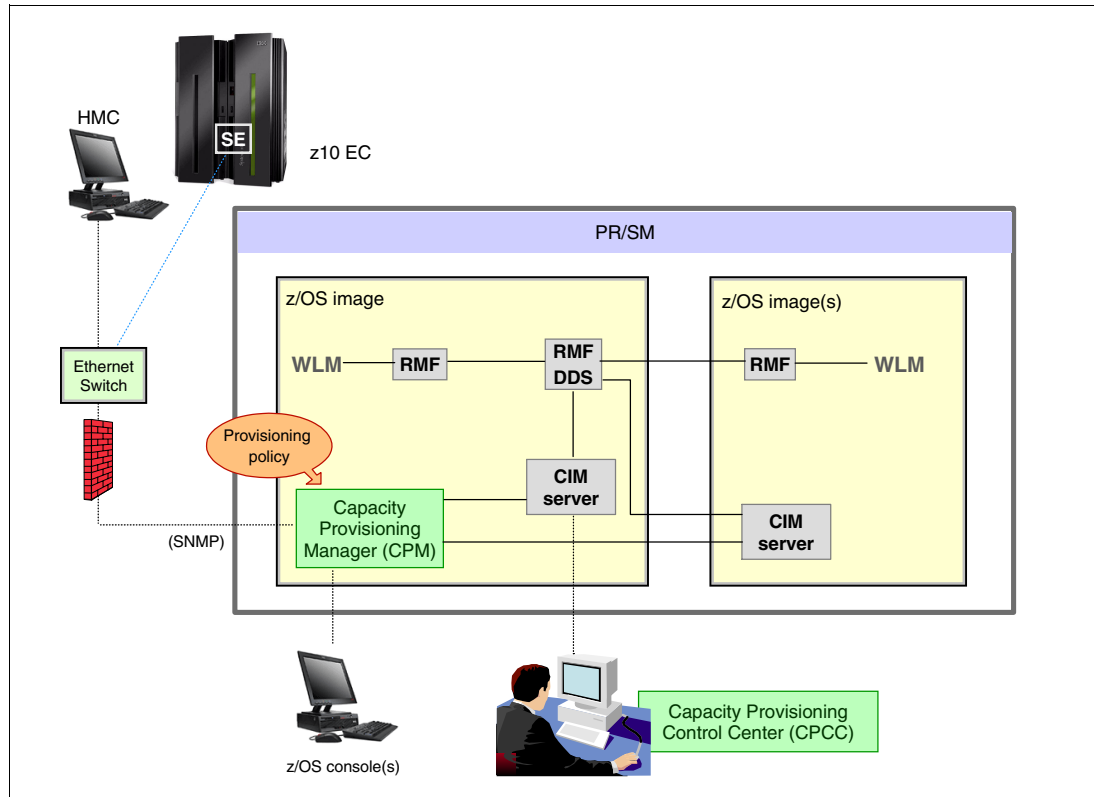


Figure 8-11 The capacity provisioning infrastructure

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 Resource Measurement Facility (RMF) Monitor III, with one RMF gatherer for each z/OS system.

Sysplex-wide data aggregation and propagation occur in the RMF distributed data server (DDS). The RMF Common Information Model (CIM) providers and associated CIM models publish the RMF Monitor III data.

The Capacity Provisioning Manager (CPM), a function inside z/OS, retrieves critical metrics from one or more z/OS systems through the Common Information Model (CIM) structures and protocol. CPM communicates to (local or remote) Support Elements and HMCs through the SNMP protocol.

CPM has visibility of the resources in the individual offering records, and the capacity tokens. When CPM decides to activate resources, a check is performed to determine whether enough capacity tokens remain for the specified resource to be activated for at least 24 hours. If not enough tokens remain, no resource from the On/Off CoD record is activated.

If a capacity token is completely consumed during an activation driven by the CPM, the corresponding On/Off CoD record is deactivated prematurely by the system, even if the CPM has activated this record, or parts of it. You do, however, receive warning messages if

capacity tokens are getting close to being fully consumed. You receive the messages five days before a capacity token is fully consumed. The five days are based on the assumption that the consumption will be constant for the 5 days. The recommendation is to put operational procedures in place to handle these situations. You may either deactivate the record manually, let it happen automatically, or replenish the specified capacity token by using the Resource Link application.

The Capacity Provisioning Control Center (CPCC), which resides on a workstation, provides an interface to administer capacity provisioning policies. The CPCC is not required for regular CPM operation.

The control over the provisioning infrastructure is executed by the CPM through the Capacity Provisioning Domain (CPD) controlled by the Capacity Provisioning Policy (CPP).

The Capacity Provisioning Domain is shown in Figure 8-12.

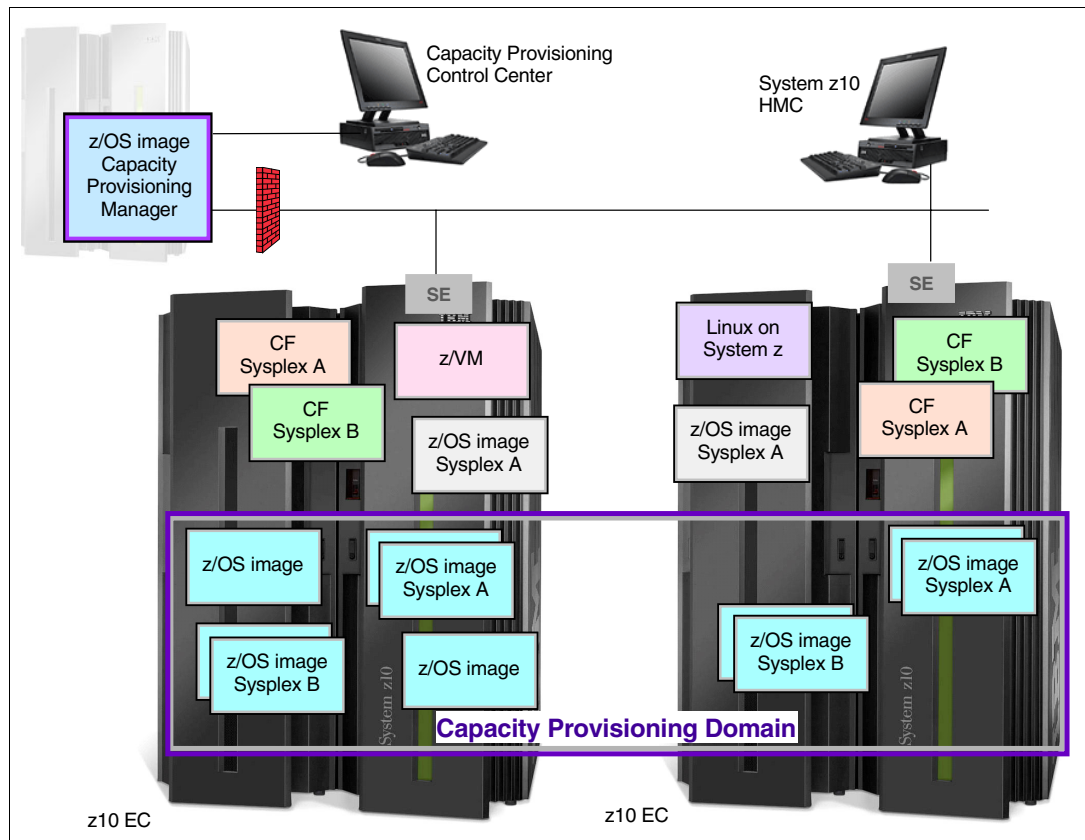


Figure 8-12 The Capacity Provisioning Domain

The Capacity Provisioning Domain represents the central processor complexes (CPCs) that are controlled by the Capacity Provisioning Manager. The HMCs of the CPCs within a CPD must be connected to the same processor LAN. Parallel Sysplex members can be part of a CPD. There is no requirement that all members of a Parallel Sysplex must be part of the CPD, but participating members must all be part of the same CPD.

The Capacity Provisioning Control Center (CPCC) is the user interface component. Administrators work through this interface to define domain configurations and provisioning policies, but it is not needed during production. The CPCC is installed on a Microsoft Windows® workstation.

CPM operates in four modes, allowing for 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.
- 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 proposed by the CPM needs to be confirmed by the operator.

▶ Autonomic mode

This mode is similar to the confirmation mode, but no operator confirmation is required.

A number of reports are available in all modes, containing information about workload and provisioning status and the rationale for provisioning recommendations. User interfaces are through the z/OS console and the CPCC application.

The provisioning policy defines the circumstances under which additional capacity may be provisioned (when, which, and how). The three elements in the criteria are:

▶ A time condition is *when* provisioning is allowed, as follows:

- Start time indicates when provisioning can begin.
- Deadline indicates that provisioning of additional capacity no longer allowed
- End time indicates that deactivation of additional capacity should begin.

▶ A workload condition is *which* work qualifies for provisioning. Parameters include:

- The z/OS systems that may execute eligible work
- Importance filter indicates eligible service class periods, identified by WLM importance
- Performance indicator (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 should not be considered.

Note: If no workload condition is specified, the full capacity described in the policy will be activated and deactivated at the start and end times specified in the policy.

▶ Provisioning scope is *how* much additional capacity may be activated, expressed in MSUs.

Specified in MSUs, number of zAAPs, and number of zIIPs must be one specification per CPC that is part of the Capacity Provisioning Domain.

The maximum provisioning scope is the maximum additional capacity that may be activated for all the rules in the Capacity Provisioning Domain.

The provisioning rule is:

In the specified time interval, *if* the specified workload is behind its objective, *then* up to the defined additional capacity may 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 server. Changing from one to another requires that the active one be stopped before the inactive one can be activated. This operation decreases the current capacity during the change.

The provisioning management routines can interrogate the installed offerings, their content, and the status of the content of the offering. To avoid the decrease in capacity, we recommend that only one On/Off CoD offering be created on the server by specifying the maximum allowable capacity. The Capacity Provisioning Manager can then, at the time when an activation is needed, activate a subset of the contents of the offering sufficient to satisfy the demand. If, at a later time, more capacity is needed, 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 content of the offerings if necessary is also possible.

Attention: As previously mentioned, the CPM has control over capacity tokens for the On/Off CoD records. In a situation where a capacity token is completely consumed, the server deactivates the corresponding offering record. Therefore, a strong recommendation is that you prepare routines for catching the warning messages about capacity tokens being consumed, and have administrative routines in place for such a situation. The messages from the system begin five days before a capacity token is fully consumed. To avoid capacity records from being deactivated in this situation, replenish the necessary capacity tokens before they are completely consumed.

In a situation where a CBU offering is active on a server and that CBU offering is 100% or more of the base capacity, activating any On/Off CoD is not possible because the On/Off CoD offering is limited to the 100% of the base configuration.

The Capacity Provisioning Manager operates based on Workload Manager (WLM) indications, and the construct used is the performance index (PI) of a service class period. It is extremely 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 executing through several service class periods, where the first period might be the important one. The application might be defined as importance level 2 or 3, but might depend on other work executing with importance level 1. Therefore, considering which workloads to control, and which service class periods to specify is very important.

8.6 Capacity for Planned Event

Capacity for Planned Event (CPE) is offered with the z10 EC to provide replacement backup capacity for planned down-time events. For example, if a server room requires an extension

or repair work, replacement capacity can be installed temporarily on another z10 EC in the customer's environment.

Note: CPE is for planned replacement capacity only and *cannot* be used for peak workload management.

The feature codes are:

- ▶ FC 6833 Capacity for Planned Event enablement
- ▶ FC 0116 - 1 CPE Capacity Unit
- ▶ FC 0117 - 100 CPE Capacity Unit
- ▶ FC 0118 - 10000 CPE Capacity Unit
- ▶ FC 0119 - 1 CPE Capacity Unit-IFL
- ▶ FC 0120 - 100 CPE Capacity Unit-IFL
- ▶ FC 0121 - 1 CPE Capacity Unit-ICF
- ▶ FC 0122 - 100 CPE Capacity Unit-ICF
- ▶ FC 0123 - 1 CPE Capacity Unit-zAAP
- ▶ FC 0124 - 100 CPE Capacity Unit-zAAP
- ▶ FC 0125 - 1 CPE Capacity Unit-zIIP
- ▶ FC 0126 - 100 CPE Capacity Unit-zIIP
- ▶ FC 0127 - 1 CPE Capacity Unit-SAP
- ▶ FC 0128 - 100 CPE Capacity Unit-SAP

The feature codes are calculated automatically when the CPE offering is configured. Whether using the eConfig tool or the Resource Link, a target configuration must be ordered consisting of a model identifier or a number of speciality engines, or both. Based on the target configuration, a number of feature codes from the list above is calculated automatically and a CPE offering record is constructed. See Figure 8-13 for an example.

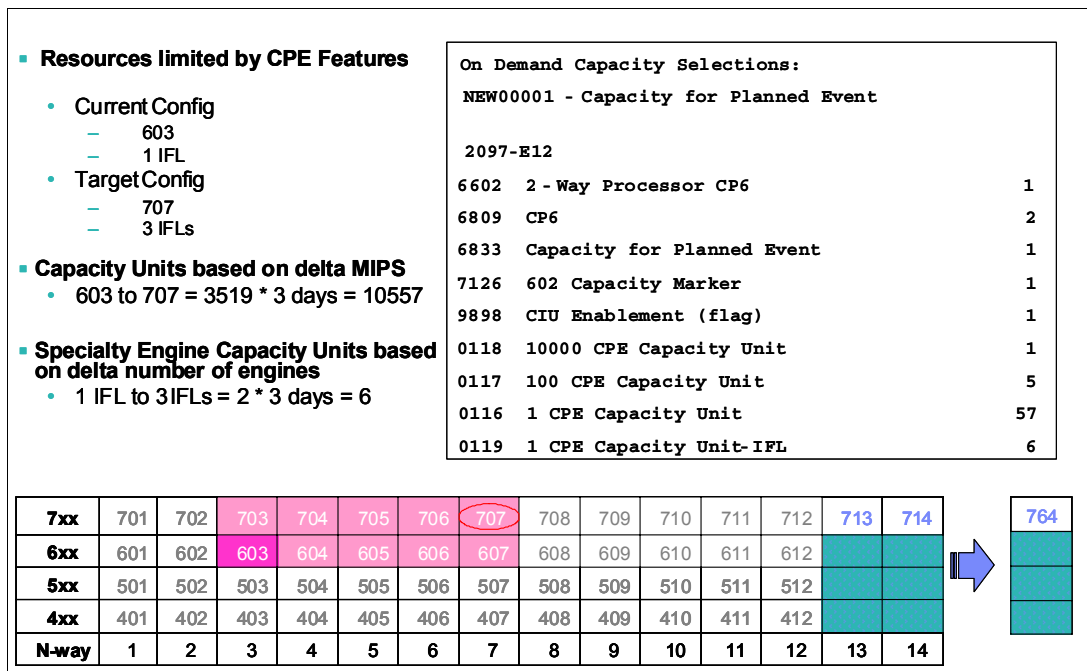


Figure 8-13 CPE feature code calculation

CPE is intended to replace capacity lost within the enterprise because of a planned event such as a facility upgrade or system relocation. CPE is intended for short duration events lasting up to a maximum of three days. Each CPE record, after it is activated, gives the you

access to dormant PUs on the server that you have a contract for as described above by the feature codes. Processing units can be configured in any combination of CP or specialty engine types (zIIP, zAAP, SAP, IFL, and ICF). At the time of CPE activation the contracted configuration will be activated. The general rule of one zIIP and one zAAP for each configured CP will be controlled for the contracted configuration.

The processors that can be activated by CPE come from the available unassigned PUs on any installed book. CPE features can be added to an existing z10 EC non-disruptively. A one-time fee is applied for each individual CPE event depending on the contracted configuration and its resulting feature codes. Only one CPE contract can be ordered at a time.

The base server configuration must have sufficient memory and channels to accommodate the potential requirements of the large CPE-configured server. It is important to ensure that all required functions and resources are available on the server where CPE is activated, including CF LEVELs for coupling facility partitions, memory, cryptographic functions, and including connectivity capabilities.

The CPE configuration is activated *temporarily* and provides additional PUs in addition to the server's original, *permanent* configuration. The number of additional PUs is predetermined by the number and type of feature codes configured as described above by the feature codes. The number PUs that can be activated is limited by the unused capacity available on the server. For example:

- ▶ A model E26 with 16 CPs, no IFLs, ICFs, or zAAPs, has 10 unassigned PUs available.
- ▶ A model E40 with 28 CPs, 1 IFL, and 1 ICF has 7 unassigned PUs available.

When the planned event is over, the server must be returned to its original configuration. You may 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 server. CPE features can be added to an existing z10 EC nondisruptively.

8.7 Capacity Backup

Capacity Backup (CBU) provides reserved emergency backup processor capacity for unplanned situations in which capacity is lost in another part of your enterprise and you want to recover by adding the reserved capacity on a designated z10 EC.

CBU is the quick, *temporary* activation of PUs and is available as follows:

- ▶ For up to 90 contiguous days, in case of a loss of processing capacity as a result of an emergency or disaster recovery situation
- ▶ For 10 days for testing your disaster recovery procedures

Note: 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, zAAPs, zIIPs, IFLs, and SAPs. To be able to use the CBU process a CBU enablement feature (FC 9910) must be ordered and installed. You must order the quantity and type of PU that you require. The feature codes are:

- ▶ 6805: Additional test activations
- ▶ 6817: Total CBU years ordered
- ▶ 6818: CBU records ordered
- ▶ 6820: Single CBU CP-year
- ▶ 6822: Single CBU IFL-year
- ▶ 6826: Single CBU zAAP-year
- ▶ 6828: Single CBU zIIP-year
- ▶ 6830: Single CBU SAP-year

The CBU entitlement record (6818) contains an expiration date that is established at the time of order and is dependent upon the quantity of CBU years (6817). You have the capability to extend your CBU entitlements through the purchase of additional CBU years. The number of 6817 per instance of 6818 remains limited to five and fractional years are rounded up to the near whole integer when calculating this limit. For instance, if there are two years and eight months to the expiration date at the time of order, the expiration date can be extended by no more than two additional years. One test activation is provided for each additional CBU year added to the CBU entitlement record.

Feature code 6805 allows for ordering additional tests in increments of one. The total number of tests allowed is 15 for each feature code 6818.

The processors that can be activated by CBU come from the available unassigned PUs on any installed book. The maximum number of CBU features that can be *ordered* is 64. The number of features that can be *activated* is limited by the number of unused PUs on the server. For example:

- ▶ A model E12 with Capacity Model Identifier 410 can activate up to 12 CBU features: ten to change the capacity setting of the existing CPs and two to activate unused PUs.
- ▶ A model E26 with 15 CPs, four IFLs, and one ICF has six unused PUs available. It can *activate* up to six CBU features.

However, the ordering system allows for over-configuration in the order itself. You may *order* up to 64 CBU features regardless of the current configuration, however at *activation*, only the capacity already installed can be *activated*. Note that at activation, you can decide to activate only a sub-set of the CBU features that are ordered for the system.

Subcapacity makes a difference in the way the CBU features are done. On the full-capacity models, the CBU features indicate the amount of additional capacity needed. If the amount of necessary CBU capacity is equal to four CPs, then the CBU configuration would be four CBU CPs.

The subcapacity models have multiple capacity settings of 4xx, 5xx, or 6xx; the standard models have capacity setting 7xx. The number of CBU CPs must be equal to or greater than the number of CPs in the base configuration, and all the CPs in the CBU configuration must have the same capacity setting. For example, if the base configuration is a 2-way 402, then 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, then going from model capacity identifier 402 to a CBU configuration of a 4-way 504 would require four CBU feature codes with a capacity setting of 5xx.

If the capacity setting of the CPs is changed, then more CBU features are required, not more physical PUs. This means that your CBU contract requires more CBU features if the capacity setting of the CPs is changed.

Note that CBU can add CPs through LICCC-only, and the z10 EC must have the proper number of books installed to allow the required upgrade. CBU can change the model capacity identifier to a *higher* value than the base setting, 4xx, 5xx, or 6xx, but does not change the *server* model 2097-Evv. 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 server. CBU features can be added to an existing z10 EC nondisruptively. For each machine enabled for CBU, the authorization to use CBU is available for a definite number of years of 1-5 years.

The installation of the CBU code provides an alternate configuration that can be activated in case of an actual emergency. Five CBU tests, lasting up to 10 days each, and one CBU activation, lasting up to 90 days for a real disaster and recovery, are typically allowed in a CBU contract.

The alternate configuration is activated *temporarily* and provides additional capacity greater than the server's original, *permanent* configuration. At activation time, you determine the capacity required for a given situation, and you can decide to activate only a sub-set of the capacity specified in the CBU contract.

Note: Do not run *production* on a server that has an active test CBU. Instead, run only a *copy* of production.

The base server configuration must have sufficient memory and channels to accommodate the potential requirements of the large CBU target server. Ensure that all required functions and resources are available on the backup servers, including CF LEVELs for coupling facility partitions, memory, and cryptographic functions, as well as connectivity capabilities.

When the emergency is over (or the CBU test is complete), the server must be taken back to its original configuration. The CBU features can be deactivated by the customer at any time before the expiration date. Failure to deactivate the CBU feature before the expiration date can cause the system to degrade gracefully back to its original configuration. The system does *not* deactivate dedicated engines, or the last of in-use shared engines.

Note: CBU for processors provides a concurrent upgrade, resulting in more enabled processors or changed capacity settings available to a server configuration, or both. You decide, at activation time, to activate a sub-set of the CBU features ordered for the system. Thus, additional planning and tasks are required for *nondisruptive* logical upgrades. See “Recommendations to avoid disruptive upgrades” on page 277.

For detailed instructions, see the *System 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 a customer responsibility and does not require on-site presence of IBM service personnel. The CBU function is activated/deactivated concurrently from the HMC using the API. On the SE, CBU is be activated either using the Perform Model Conversion task or through the API (API enables task automation).

CBU activation

CBU is activated from the SE by using the Perform Model Conversion task or through automation by using API on the SE or the HMC. In case of real disaster, use the Activate CBU option to activate the 90-day period.

Image upgrades

After the CBU activation, the z10 EC can have more capacity, more active PUs, or both. The additional resources go into the resource pools and are available to the logical partitions. If the logical partitions have to increase their share of the resources, the logical partition weight can be changed or the number of logical processors can be concurrently increased by configuring reserved processors online. The operating system must have the capability to concurrently configure more processors online. If necessary, additional logical partitions can be created to use the newly added capacity.

CBU deactivation

To deactivate the CBU, the additional resources have to be released from the logical partitions by the operating systems. In some cases, this is a matter of varying the resources offline. In other cases, it can mean shutting down operating systems or deactivating logical partitions. After the resources have been released, the same facility on the SE is used to turn off CBU. To deactivate CBU, click the **Undo temporary upgrade** option from the Perform Model Conversion task on the SE.

CBU testing

Test CBUs are provided as part of the CBU contract. CBU is activated from the SE by using the Perform Model Conversion task. Select the test option to initiate a 10-day test period. A standard contract allows five tests of this type. However, you may order additional tests in increments of one up to a maximum of 15 for each CBU order. The test CBU has a 10-day limit and must be deactivated in the same way as the real CBU, using the same facility through the SE. Failure to deactivate the CBU feature before the expiration date can cause the system to degrade gracefully back to its original configuration. The system does *not* deactivate dedicated engines, or the last of in-use shared engine. Testing can be accomplished by ordering a diskette, calling the support center, or using the facilities on the SE. The customer has the possibility of purchasing additional tests.

CBU example

Figure 8-14 shows an example of a capacity Backup operation.

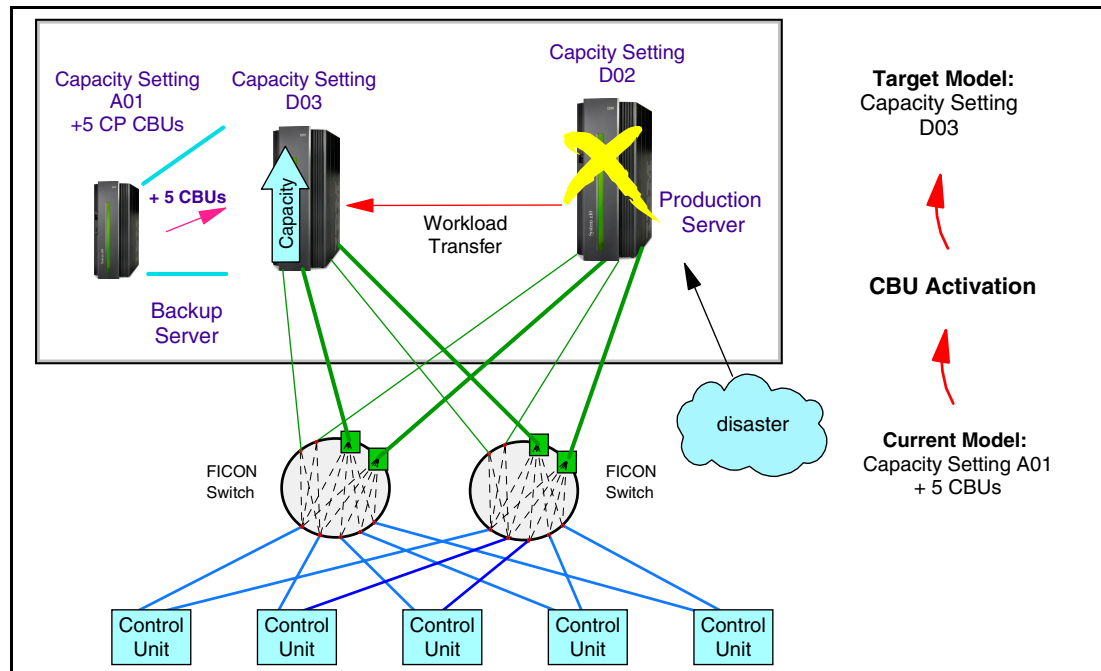


Figure 8-14 Capacity Backup operation example

In this example, 12 CBU features are installed on the backup model E26 with model capacity identifier 708. When the production model E12 with model capacity identifier 708 has an unplanned outage, the backup server can be temporarily upgraded from model capacity identifier 708 to 720 so that the capacity can take over the workload from the failed production site.

Furthermore, you may configure systems to back up each other. For example, if you use two models of E12 model capacity identifier 705 for the production environment, each can have five or more features installed. If one server suffers a disaster, the other one uses a temporary upgrade to recover approximately the total original capacity.

8.7.3 Automatic CBU enablement for GDPS

The intent of the Geographically Dispersed Parallel Sysplex™ (GDPS®) CBU is to enable automatic management of the PUs provided by the CBU feature in the event of a server or site failure. Upon detection of a site failure or planned disaster test, GDPS will concurrently add CPs to the servers in the take-over site to restore processing power for mission-critical production workloads. GDPS automation does the following tasks:

- ▶ Performs the analysis required to determine the scope of the failure. This minimizes operator intervention and the potential for errors.
- ▶ Automates authentication and activation of the reserved CPs
- ▶ Automatically restarts the critical applications after reserved CP activation.
- ▶ Reduces the outage time to restart critical workloads from several hours to minutes.

The GDPS service is for z/OS only, or for z/OS in combination with Linux on System z.

8.8 Nondisruptive upgrades

Continuous availability is an increasingly important requirement for most customers, and even planned outages are no longer acceptable. Although Parallel Sysplex clustering technology is the best continuous availability solution for z/OS environments, nondisruptive upgrades within a single server can avoid system outages and are suitable to additional operating system environments.

The z10 EC allows *concurrent* upgrades, meaning that dynamically adding more capacity to the server is possible. If operating system images running on the upgraded server do not require disruptive tasks in order to use the new capacity, the upgrade is also *nondisruptive*. This means that power-on reset (POR), logical partition deactivation, and IPL do not have to take place.

If the concurrent upgrade is intended to satisfy an *image upgrade* to a logical partition, the operating system running in this partition must also have the capability 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 for z/VM V5R4 and later releases, memory can be dynamically added to z/VM partitions.

If the concurrent upgrade is intended to satisfy the need for more operating system images, additional logical partitions can be created *concurrently* on the z10 EC server, including all resources needed by such logical partitions. These additional logical partitions can be activated *concurrently*.

These enhanced configuration options are made available through the separate HSA, which is introduced on the System z10 EC.

Linux operating systems in general do *not* have the capability of adding more resources concurrently. However, Linux, and other types of virtual machines running under z/VM, can benefit from the z/VM capability to nondisruptively configure more resources online (processors and I/O).

Starting with z/VM V5R4, Linux guests can manipulate their logical processors through the use of the Linux CPU hotplug daemon. The daemon can start and stop logical processors based on the Linux average load value. The daemon is available in Linux SLES 10 SP2. IBM is working with our Linux distribution partners to have the daemon available in other distributions for the System z servers.

Important: If the STI rebalance feature (FC 2400) is selected at server upgrade configuration time, and effectively results in HCA rebalancing for ICB-4s, it will also change the PCHID number of ICB-4 links, requiring a corresponding update on the server's I/O definition through HCD/HCM. The STI rebalance feature is disruptive. This feature does not apply to model E64.

Processors

CPs, ICFs, zAAPs, zIIPs, IFLs, and SAPs can be concurrently added to a z10 EC if unassigned PUs are available on any installed book. The number of zAAPs cannot exceed the number of CPs plus unassigned CPs. The same holds true for the zIIPs.

Additional books can also be installed concurrently, allowing further processor upgrades.

Concurrent upgrades are not supported with PUs defined as additional SAPs.

If necessary, additional logical partitions 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 logical partitions by using the CFCC image operations window.

Memory

Memory can be concurrently added up to the physical installed memory limit. Additional books can also be installed concurrently, allowing further memory upgrades by LICCC, enabling memory capacity on the new books.

Using the previously defined reserved memory, z/OS operating system images, and z/VM V5R4 partitions, can dynamically configure more memory online, allowing nondisruptive memory upgrades. Linux on System z supports Dynamic Storage Reconfiguration.

I/O

I/O cards can be added concurrently if all the required infrastructure (I/O slots and HCAs) is present on the configuration. The plan-ahead process can assure that an initial configuration will have all the infrastructure required for the target configuration.

I/O ports can be concurrently added by LICCC, enabling available ports on ESCON and ISC-3 daughter cards.

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 server is not possible because there is no operating system with this capability running on this server.

Cryptographic adapters

Crypto Express2 and Crypto Express3 features can be added concurrently if all the required infrastructure, I/O slots, and STIs are present on the configuration. The plan-ahead process can assure that an initial configuration will have all the infrastructure required for the target configuration.

Concurrent upgrade considerations

By using MES upgrade, On/Off CoD, CBU, or CPE, a z10 EC can be concurrently upgraded from one model to another, either temporarily or permanently.

Enabling and using the additional processor capacity is transparent to most applications. However, certain programs depend on processor model-related information, for example, Independent Software Vendor (ISV) products. You should consider the effect on the software running on a z10 EC when you perform any of these configuration upgrades.

Processor identification

Two instructions are used to obtain processor information:

- ▶ Store System Information instruction (STSI)

STSI reports the processor model and model capacity identifier for the base configuration and for any additional configuration changes through temporary upgrade actions. It fully supports the concurrent upgrade functions and is the preferred way to request processor information.

- ▶ Store CPU ID instruction (STIDP)

STIDP is provided for purposes of backward compatibility.

Store system information (STSI) 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 is key to the functioning of Capacity on Demand offerings.

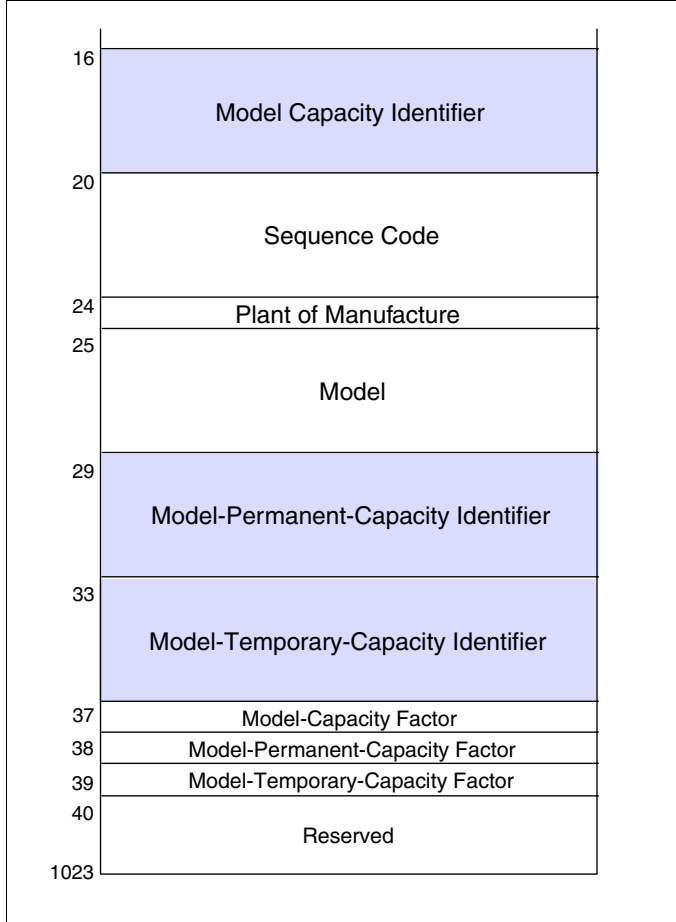


Figure 8-15 STSI output on z10 EC

The model capacity identifier contains the base capacity, the On/Off CoD, and the CBU. The model permanent capacity identifier and the Model Permanent Capacity Rating contain the base capacity of the system, and the model temporary capacity identifier and model temporary capacity rating contain the base capacity and the On/Off CoD.

Store CPU ID instruction

The STIDP instruction provides information about the processor type, serial number, and logical partition identifier. See Table 8-4. The logical partition identifier field is a full byte to support greater than 15 logical partitions.

Table 8-4 STIDP output for z10 EC

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	Logical partition ID ^b	6-digit number derived from the CPC serial number	x'2097'	x'8000' ^c

a. The version code for System z10 is x00.

b. The logical partition identifier is a two-digit number in the range of 00 - 3F. It is assigned by the user on the image profile through the Support Element or HMC.

c. High order bit on indicates that the logical partition ID value returned in bits 8 - 15 is a two-digit value.

When issued from an operating system running as a guest under z/VM, the result depends on whether the SET CPUID command has been used, as follows:

- ▶ Without the use of the SET CPUID command, bits 0 - 7 are set to FF by z/VM, but the remaining bits are unchanged, which means that they are exactly as they would have been without running as a z/VM guest.
- ▶ If the SET CPUID command has been issued, bits 0 - 7 are set to FF by z/VM and bits 8 - 31 are set to the value entered in the SET CPUID command. Bits 32 - 63 are the same as they would have been without running as a z/VM guest.

Table 8-5 lists the possible output returned to the issuing program for an operating system running as a guest under z/VM.

Table 8-5 VM guest STIDP output for z10 EC EC

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'	Logical partition ID	4-digit number derived from the CPC serial number	x'2097'	x'8000'
With SET CPUID command	x'FF'	6-digit number as entered by the command SET CPUID = nnnnnn		x'2097'	x'8000'

Planning for nondisruptive upgrades

Online permanent upgrades, On/Off CoD, CBU, and CPE can be used to concurrently upgrade a z10 EC. However, certain situations require a disruptive task in order to enable the new capacity that was recently added to the server. Some of these situation can be avoided if planning is done in advance. Planning ahead is a key factor for nondisruptive upgrades.

The following list describes main reasons for disruptive upgrades. However, by carefully planning and by reviewing “Recommendations to avoid disruptive upgrades” on page 277, you can minimize the need for these outages:

- ▶ z/OS logical partition processor upgrades when reserved processors were not previously defined are disruptive to image upgrades.
- ▶ Logical partition memory upgrades when reserved storage was not previously defined are disruptive to image upgrades. z/OS and z/VM V5R4 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. Linux, z/VSE, TPF, z/TPF, and CFCC do not support dynamic I/O configuration.

Recommendations to avoid disruptive upgrades

Based on the previous list of reasons for disruptive upgrades (“Planning for nondisruptive upgrades” on page 276), here are several recommendations for avoiding or at least minimizing these situations, increasing the possibilities for nondisruptive upgrades:

- ▶ For z/OS logical partitions configure as many reserved processors (CPs, ICFs, zAAPs, and zIIPs) as possible.

Configuring reserved processors for all logical z/OS partitions *before* their activation enables them to be nondisruptively upgraded. The operating system running in the logical partition must have the ability to configure processors online. The total number of defined and reserved CPs cannot exceed the number of CPs supported by the operating system. z/OS V1R7 can support up to 32 processors. z/OS V1R8 can support up to 54 processors including CPs, zAAPs, and zIIPs. z/OS V1R9 can support up to 64 processors including CPs, zAAPs, and zIIPs. z/VM V5R3 and later releases support up to 32 processors.

- ▶ Configure reserved storage to logical partitions.

Configuring reserved storage for all logical partitions *before* their activation enables them to be nondisruptively upgraded. The operating system running in the logical partition must have the ability to configure memory online. The amount of reserved storage can be above the book threshold limit (384 GB), even if no other book is already installed. The current partition storage limit is 1TB. z/OS and z/VM V5R4 support this function.

- ▶ Consider the flexible and plan-ahead memory options.

Use a convenient entry point for memory capacity and consider the memory options to allow future upgrades within the memory cards already installed on the books. For details about the offerings, see

- 2.5.4, “Flexible memory option” on page 39
- 2.5.5, “Plan-ahead memory” on page 39

- ▶ Use the plan-ahead concurrent condition for I/O.

Use the plan-ahead concurrent condition process to include in the initial configuration all the I/O cages required by future I/O upgrades, allowing the planned concurrent I/O upgrades.

Considerations when installing additional books

During an upgrade, additional books can be installed concurrently. Depending on the number of additional books in the upgrade and your I/O configuration, an STI and HCA2 rebalancing for ICB-4s might be recommended for availability reasons. It will change ICB-4 PCHID numbers, requiring an I/O definition update.

8.9 Summary of Capacity on Demand offerings

The capacity on demand infrastructure and its offerings are major features introduced with the System z10 EC server. The reasons for the introduction of these features are based on numerous customer requirements for more flexibility, granularity, and better business control over the System z infrastructure, operationally and financially.

One major customer requirement is to dismiss the necessity for a customer authorization connection to IBM Resource Link system at the time of activation of any offering. This requirement is being met by the z10 EC. After the offerings have been installed on the z10 EC, they can be activated at any time, completely at the customer's discretion. No intervention through IBM or IBM personnel is necessary. In addition, the activation of the Capacity Backup does not require a password.

The z10 can have up to eight offerings installed at the same time, with the limitation that only one of them can be an On/Off Capacity on Demand offering; the others can be any combination. The installed offerings can be activated fully or partially, and in any sequence and any combination. The offerings can be controlled manually through command interfaces on the HMC, or programmatically through a number of APIs, so that IBM applications, ISV programs, or customer-written applications, can control the usage of the offerings.

Resource consumption (and thus financial exposure) can be controlled by using capacity tokens in On/Off CoD offering records.

The Capacity Provisioning Manager (CPM) is an example of an application that uses the Capacity on Demand APIs to provision On/Off CoD capacity based on the requirements of the executing workload. The CPM cannot control other offerings.



RAS

This chapter describes several of the reliability, availability, and serviceability (RAS) features of the z10 EC server.

The z10 EC design is focused on providing higher availability by reducing planned and unplanned outages. RAS can be accomplished with improved concurrent replace, repair, and upgrade functions for processors, memory, books, and I/O. RAS also extends to the nondisruptive capability for downloading Licensed Internal Code (LIC) updates. In most cases a capacity upgrade can be concurrent without a system outage.

To make the delivery and transmission of microcode (LIC), fixes and restoration/backup files are digitally signed. Any data transmitted to IBM Support is encrypted.

The design goal for the z10 EC has been to remove all sources of planned outages.

This chapter discusses the following topics:

- ▶ 9.1, “z10 Availability characteristics” on page 280
- ▶ 9.2, “z10 RAS functions” on page 281
- ▶ 9.3, “z10 Enhanced book availability” on page 283
- ▶ 9.4, “z10 Enhanced driver maintenance” on page 292

9.1 z10 Availability characteristics

The following functions include availability characteristics on the z10 EC:

- ▶ Enhanced book availability (EBA)

The z10 EC allows a single book in a multibook server to be concurrently removed from the server and reinstalled during an upgrade or repair action.

- ▶ Concurrent memory upgrade or replacement

Memory can be upgraded concurrently using LICCC if physical memory is available on the books. If the physical memory cards have to be changed in a multibook configuration, which would require the book to be removed, the enhanced book availability function can be useful. It requires the availability of additional resources on other books or reducing the need for resources during this action. To help ensure that the appropriate level of memory is available in a multiple-book configuration, consider the selection of the flexible memory option for providing additional resources to exploit EBA when repairing a book or memory on a book, or when upgrading memory where larger memory cards might be required.

Memory can be upgraded concurrently by using LICCC if physical memory is available as previously explained. The plan-ahead memory function available with the System z10 server provides the ability to plan for nondisruptive memory upgrades by having the system pre-plugged based on a target configuration. Pre-plugged memory is enabled when you place an order through LICCC.

- ▶ Enhanced driver maintenance (EDM)

One of the greatest contributors to downtime during planned outages is Licensed Internal Code driver updates performed in support of new features and functions. The z10 EC is designed to support activating a selected new driver level concurrently.

- ▶ Concurrent HCA/MBA fanout addition or replacement

A Host Channel Adapter (HCA)/Memory Bus Adapter (MBA) fanout card provides the path for data between memory and I/O using InfiniBand (IFB) cables. With the z10 EC, a hot-pluggable and concurrently upgradeable HCA/MBA fanout card is available. Up to eight HCA/MBA fanout cards are available per book for a total of up to 32 HCA/MBA fanout cards when four books are installed. In the event of an outage, an HCA/MBA fanout card, used for I/O, may be concurrently repaired while redundant I/O interconnect ensures that no I/O connectivity is lost.

- ▶ Redundant I/O interconnect

Redundant I/O interconnect helps maintain critical connections to devices. The z10 EC allows a single book, in a multibook server, to be concurrently removed and reinstalled during an upgrade or repair, continuing to provide connectivity to the server I/O resources using a second path from a different book.

- ▶ Dynamic oscillator switch-over

The z10 EC has two oscillator cards, a primary and a backup. In the event of a primary card failure, the backup card is designed to transparently detect the failure, switch-over, and provide the clock signal to the server.

9.2 z10 RAS functions

Hardware RAS function improvements focus on addressing all sources of outages. Sources of outages have three classifications:

- Unscheduled** This outage occurs because of an unrecoverable malfunction in a hardware component of the server.
- Scheduled** This outage is caused by changes or updates that have to be done to the server in a timely fashion. A scheduled outage can be caused by a disruptive patch that has to be installed, or other changes that have to be done to the system. A scheduled outage is usually requested by the vendor (IBM).
- Planned** This outage is also caused by changes or updates that have to be done to the server. A planned outage can be caused by a capacity upgrade or a driver upgrade. A planned outage is usually requested by the customer and often requires pre-planning. The System z10 design phase focused on this pre-planning effort and was able to simplify or eliminate it.

Unscheduled, scheduled, and planned outages have been addressed for the mainframe family of servers for many years. Figure 9-1 shows a summary. Planned outages have been specifically addressed with the z9 EC and pre-planning requirements are specifically addressed with the z10 EC server.

	Prior Servers	z9 EC	z10
Unscheduled Outages	✓	✓	✓
Scheduled Outages	✓	✓	✓
Planned Outages		✓	✓
Pre planning requirements			✓

Figure 9-1 RAS focus

Pre-planning requirements have been reduced for the z10 EC server. A fixed size HSA of 16 GB has been introduced to help eliminate pre-planning requirements for HSA and to provide flexibility to dynamically update the configuration.

Performing the following tasks dynamically is possible:

- ▶ Add a logical partition.
- ▶ Add a logical channel subsystem (LCSS).
- ▶ Add a subchannel set.
- ▶ Add a logical CP to a logical partition.
- ▶ Add a cryptographic coprocessor.

- ▶ Remove a cryptographic coprocessor.
- ▶ Enable I/O connections.
- ▶ Swap processor types.

In addition, by addressing the elimination of planned outages, the following tasks are also possible:

- ▶ Concurrent driver upgrades
- ▶ CBU activation without previous unnecessary passwords
- ▶ Concurrent and flexible customer-initiated upgrades

For a description of the flexible customer-initiated upgrades see Chapter 8, “System upgrades” on page 233.

As previously described, scheduled outages are most often requested by the vendor. Concurrent hardware upgrades, concurrent parts replacement, concurrent driver upgrade, and concurrent firmware fixes available with the z10 EC, all address elimination of scheduled outages. Furthermore, the following indicators and functions that address scheduled outages are included:

- ▶ Dual in-line memory module (DIMM) field replaceable unit (FRU) indicators

These indicators imply that a memory module is not error free and could fail sometime in the future. This gives IBM a warning, and the possibility and time to concurrently repair the storage module. To do this, first fence-off the book, then remove the book, replace the failing storage module, and then add the book. The flexible memory option might be necessary to maintain sufficient capacity while repairing the storage module.
- ▶ Single processor core checkstop and sparing

This indicator implies that a processor core has malfunctioned and has been *spared*. IBM has to consider what to do and also take into account the history of the server by asking the question: Has this type of incident happened previously on this server?
- ▶ Point-to-point fabric for the SMP (no longer a ring)

Having fewer components that can fail is an advantage. In a two-book or three-book machine the need to complete a ring has been removed. In addition, a book can always be added concurrently.
- ▶ Hot swap ICB-4 and InfiniBand (IFB) hub cards

When properly configured for redundancy, hot swapping (replacing) the ICB-4 (MBA) and the IFB (HCA2-O) hub cards is possible, thereby avoiding any kind of interruption when the need for replacing these types of cards occurs.
- ▶ Redundant 100 Mbps Ethernet service network with VLAN

The service network in the machine gives the machine code the capability to monitor each single internal function in the machine. This helps to identify problems, maintain the redundancy, and provides assistance in concurrently replacing a part. Through the implementation of the VLAN to the redundant internal Ethernet service network, these advantages are improving even more, as it makes the service network itself easier to handle and more flexible.

An unscheduled outage occurs because of an unrecoverable malfunction in a hardware component of the server.

The following improvements can minimize unscheduled outages:

- ▶ **Reduced chip count on the Multi-Chip Module (MCM)**

The number of chips on the MCM is reduced compared to the z9 EC. Statistics collected over several years indicate that fewer parts in the hardware lead to fewer malfunctions. This is also valid for the MCM. Fewer chips imply fewer unrecoverable malfunctions.
- ▶ **Continued focus on firmware quality**

For Licensed Internal Code 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**

As for previous servers, error detection and recovery in the memory subsystem hardware are implemented by error correction code (ECC). The memory subsystem has been enhanced with additional robust data ECC. The connection between the memory DIMMs and the memory controller is now also ECC protected. This ECC provides failure protection for virtually every type of packet transfer failure that can be corrected spontaneously. The data portion of the packet-transfers especially benefit because they are now protected.
- ▶ **Soft-switch firmware**

The capabilities of soft-switching firmware have been enhanced. Enhanced logic in this function ensures that every affected circuit is powered off during soft-switching of firmware components. For example, if you must upgrade the microcode of a FICON feature, enhancements have been implemented to avoid any unwanted side-effects detected on previous servers.

9.3 z10 Enhanced book availability

With the z10 EC, a single book in a multibook server, can be concurrently removed from the server and reinstalled during an upgrade or repair action, while continuing to provide connectivity to the server I/O resources by using a second path from a different book.

With enhanced book availability (EBA), and with proper planning to ensure that all the resources are still available to run critical applications in a (n-1) book configuration, you might be able to avoid planned outages. Consider, also, the selection of the flexible memory option to provide additional memory resources when replacing a book.

To minimize affecting current workloads, ensure that there are sufficient inactive physical resources on the remaining books to complete a book removal. Also consider non-critical system images, such as test or development logical partitions. After these non-critical logical partitions have been stopped and their resources freed, you might find sufficient inactive resources to contain critical workloads while completing a book replacement.

Before you configure the z10 EC read 9.3.1, “Planning considerations” on page 284.

9.3.1 Planning considerations

To take advantage of the enhanced book availability function, configure enough physical memory and engines so that the loss of a single book does not result in any degradation to critical workloads during the following occurrences:

- ▶ A degraded restart in the rare event of a book failure
- ▶ A book replacement for repair or physical memory upgrade

We recommend the following configurations that enable exploitation of the enhanced book availability function. The PU and models suggested have enough unused capacity so that 100% of the customer-owned PUs can be activated even when one book within a model is fenced.

- ▶ A maximum of 12 customer PUs are configured on the E26.
- ▶ A maximum of 26 customer PUs are configured on the E40.
- ▶ A maximum of 40 customer PUs are configured on the E56.
- ▶ A maximum of 46 customer PUs are configured on the E64.
- ▶ No special feature codes are required for PU and model configuration.
- ▶ For the four book models, the number of SAPs in a book is not the same for all books. This is a planning consideration. For the exact distribution of SAPs and spares over the four books, see Table 2-3 on page 34.
- ▶ Flexible memory option, which delivers physical memory so that 100% of the purchased memory increment can be activated even when one book is fenced.

The main point here is that the server configuration should have sufficient *dormant* resources on the remaining books in the system for the *evacuation* of the book to be replaced or upgraded. Dormant resources can be:

- ▶ Unused PUs or memory are not enabled by LICCC
- ▶ Inactive resources that are enabled by LICCC (memory that is not being used by any activated logical partitions)
- ▶ Memory purchased with the flexible memory option
- ▶ Additional books, as discussed previously

The I/O connectivity must also support book removal. The majority of the path to the I/O has redundant I/O interconnect support in the I/O cages and that enable connection through multiple IFBs. You must ensure that all the ICBs have redundant paths from different books.

If sufficient resources are not present on the remaining books, certain non-critical logical partitions might have to be deactivated, and one or more CPs, specialty engines, or storage might have to be configured offline to reach the required level of available resources. Planning that addresses these possibilities can help to reduce operational errors.

Note: Single-book systems cannot make use of the EBA function.

The planning should be included as part of the initial installation and any follow-on upgrade that modifies the operating environment. The eConfig report can be used to determine the number of books, active PUs, memory configuration, and the channel layout.

If the z10 EC is installed, you may click the **Prepare for Enhanced Book Availability** option in the Perform Model Conversion panel of the EBA process. This task helps you determine

the resources required to support the removal of a book with acceptable degradation to the operating system images.

The EBA process determines which resources, including memory, PUs, and I/O paths will have to be freed to allow for the removal of a given book. You may run this preparation on each book to determine which resource changes are necessary; use the results as input to the planning stage to help identify critical resources.

With this planning information, you can examine the logical partition configuration and workload priorities to determine how resources might be reduced and allow for the book to be removed.

Planning should include the following tasks:

- ▶ Review of the z10 EC configuration to determine:
 - Number of books installed and the number of PUs enabled. Note the following information:
 - Use the eConfig output or the HMC to determine the model, number and types of PUs (CPs, IFL, ICF, zAAP, and zIIP).
 - Determine the amount of memory, both physically installed and LICCC-enabled.
 - Work with your IBM service personnel to determine the memory card size in each book. The memory card sizes and the number of cards installed for each installed book can be viewed from the SE under the CPC configuration task list, using the view hardware configuration option.
 - Channel layouts, IFB to channel connections.
Use eConfig output to review channel configuration including the IFB path. This is a normal part of the I/O connectivity planning. The alternate paths should be separated as far into the system as possible.
- ▶ Review the system image configurations to determine the resources for each.
- ▶ Determine the importance and relative priority of each logical partition.
- ▶ Identify the logical partition or workloads and the actions to be taken:
 - Deactivate the entire logical partition.
 - Configure PUs.
 - Reconfigure memory, which might require the use of Reconfigurable Storage Units (RSU) value.
 - Vary off of the channels.
- ▶ Review the channel layout and determine whether any changes are necessary to address single paths.
- ▶ Develop the plan to address the requirements.

When performing the review, document the resources that could be made available if the EBA were to be used. The resources on the books are allocated during a power-on reset (POR) of the system and can change during a POR. Perform a review when changes are made to the z10 EC, such as adding books, CPs, memory, or channels, or when workloads are added or removed. The Prepare for Enhanced Book Availability function can be used ahead of any EBA action to determine whether the system can be conditioned to allow for book removal or what actions are required to support the removal.

9.3.2 Enhanced book availability processing

To use the EBA, certain conditions must be satisfied:

- ▶ Free the used processors (PUs) on the book that will be removed.
- ▶ Free the used memory on the book.
- ▶ For all I/O domains connected to this book, ensure that alternate paths exist, otherwise place the I/O paths offline.

For the EBA process, this is the preparation phase, and is started from the SE, either directly or on the HMC by using the single object operation option in the Perform Model Conversion panel from the CPC configuration task list. See Figure 9-2 on page 287.

Processor availability

Processor resource availability for reallocation or deactivation is affected by the type and quantity of resources in use, as follows:

- ▶ Total number of PUs that are enabled through LICCC
- ▶ PU definitions in the profiles and that can be
 - Dedicated and dedicated reserved
 - Shared
- ▶ Active logical partitions with dedicated resources at the time of book repair or replacement

To maximize the PU availability option, ensure that there are sufficient inactive physical resources on the remaining books to complete a book removal.

Memory availability

Memory resource availability for reallocation or deactivation depends on:

- ▶ Physically installed storage
- ▶ Image profile storage allocations
- ▶ Amount of memory enabled through LICCC
- ▶ Flexible memory option

See 2.6.2, “Enhanced book availability” on page 44.

HCA-2 to IFB-MP connectivity requirements

The optimum approach is to maintain maximum I/O connectivity during book removal. The redundant I/O interconnect (RII) function provides for redundant IFB connectivity to all installed I/O domains in all I/O cages, as follows:

- ▶ ICB-4s connect directly from the IFB port on the book to an IFB on another server.

Note: The ICB-4s do *not* take advantage of the RII function. In this case, configure the associated CHPIDS offline.

- ▶ Always ensure that there are redundant ICBs to the same CF or z/OS image, from different books.

Preparing for enhanced book availability

The Prepare Concurrent Book replacement option validates that enough dormant resources exist for this operation. If enough resources are not available on the remaining books to complete the EBA process, the process identifies the resources and guides you through a series of steps to select and free up resources. The preparation process does not complete until all memory and I/O conditions have been successfully resolved.

Note: The preparation step does not reallocate any resources. It is only used to record customer choices and to produce a configuration file on the SE that will be used by the *perform concurrent book* replacement operation.

The preparation step can be done in advance. However, if any changes to the configuration occur between the time the preparation is done and when the book is physically removed, you must rerun the preparation phase.

The process can be run multiple times, because it does not move any resources. To view results of the last preparation operation, select **Display Previous Prepare Enhanced Book Availability Results** from the Perform Model Conversion panel (in SE).

The preparation step can be run several times without actually performing a book replacement. It enables you to dynamically adjust the operational configuration for book repair or replacement prior to Customer Engineer (CE) activity. Figure 9-2 shows the Perform Model Conversion panel where you select the **Prepare for Enhanced Book Availability** option.

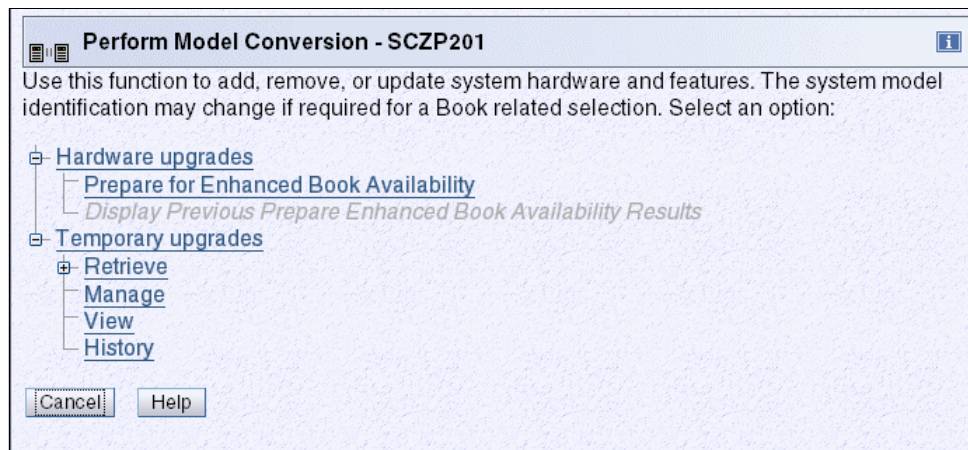


Figure 9-2 Perform Model Conversion; select Prepare for Enhanced Book Availability

After you select **Prepare for Enhanced Book Availability**, the Enhanced Book Availability panel opens. Select the book that is to be repaired or upgraded and click **OK**. See Figure 9-3. Only one target book can be selected each time.

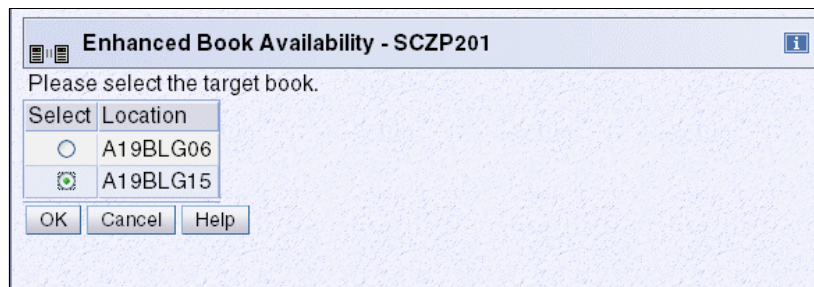


Figure 9-3 Enhanced Book Availability, selecting the target book

The system verifies the resources required for the removal, determines the required actions, and presents the results for review. Depending on the configuration, the task can take approximately 30 minutes.

The preparation step determines the readiness of the system for the removal of the targeted book. The configured processors and the memory that is in are evaluated against unused resources available across the remaining books.

If not enough resources are available, the system identifies the conflicts so you can take action to free other resources. The system analyzes I/O connections associated with the removal of the targeted book for any single path I/O connectivity.

Three states can result from the preparation step:

- ▶ The system is ready to perform the enhanced book availability for the targeted book with the original configuration.
- ▶ The system is not ready to perform the enhanced book availability because of conditions indicated from the preparation step.
- ▶ The system is ready to perform the enhanced book availability for the targeted book. However, to continue with the process, processors are reassigned from the original configuration. Review the results of this reassignment relative to your operation and business requirements. The reassignments can be changed on the final window that is presented. However, before making changes or approving reassignments, ensure that the changes have been reviewed and approved by the correct level of support, based on your organization's business requirements.

Preparation tabs

The results of the preparation are presented for review in a tabbed format. Each tab indicates conditions that prevent the EBA option from being performed. Tabs are for processors, memory, and various single path I/O conditions. See Figure 9-4 on page 289. Possible tab selections are:

- ▶ Processors
- ▶ Memory
- ▶ Single I/O
- ▶ Single Domain I/O
- ▶ Single Alternate Path I/O

Only the tabs that have conditions that prevent the book from being removed are displayed. Each tab indicates what the specific conditions are and possible options to correct the conditions.

Example panels from the preparation phase

The figures in this section are examples of panels that are displayed, during the preparation phase, when a condition requires further actions to prepare the system for the book removal.

Figure 9-4 shows the results of the preparation phase for removing book 0. The tabs labeled Memory and Single I/O indicate the conditions that were found in preparing the book to be removed. In the figure, the Memory tab indicates the amount of memory in use, the amount of memory available, and the amount of memory that must be made available. The amount of in-use memory is indicated in megabytes for each partition name. After the required amount of memory has been made available, rerun the preparation to verify the changes.

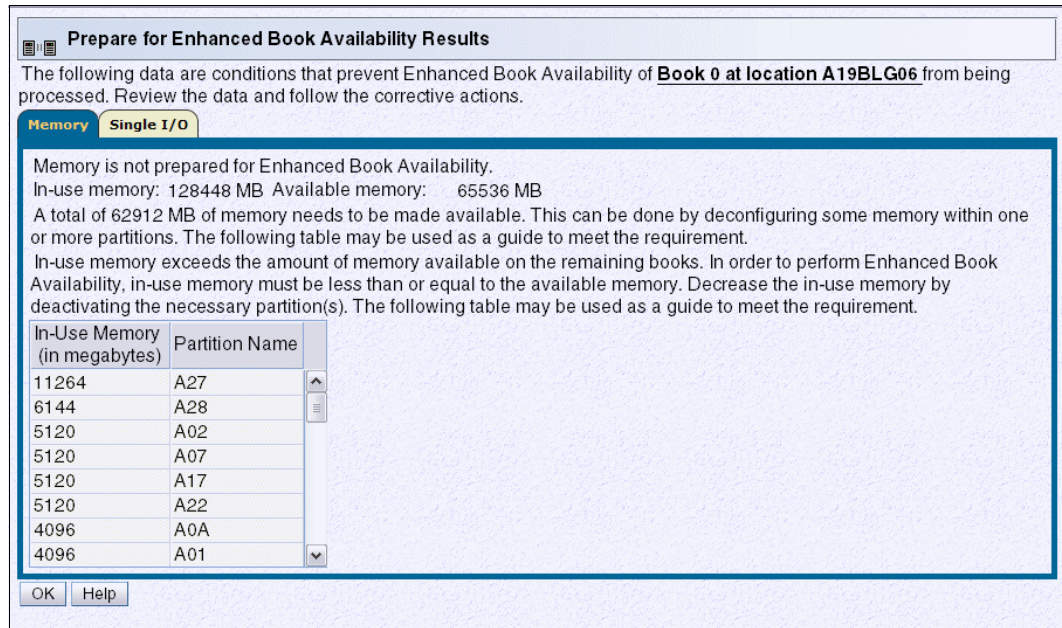


Figure 9-4 Prepare for EBA: Memory conditions

Figure 9-5 shows the Single I/O tab. The preparation has identified single I/O paths associated with the removal of book 0. The paths have to be placed offline to perform the book removal. After addressing the condition, rerun the preparation step to ensure that all the required conditions have been met.

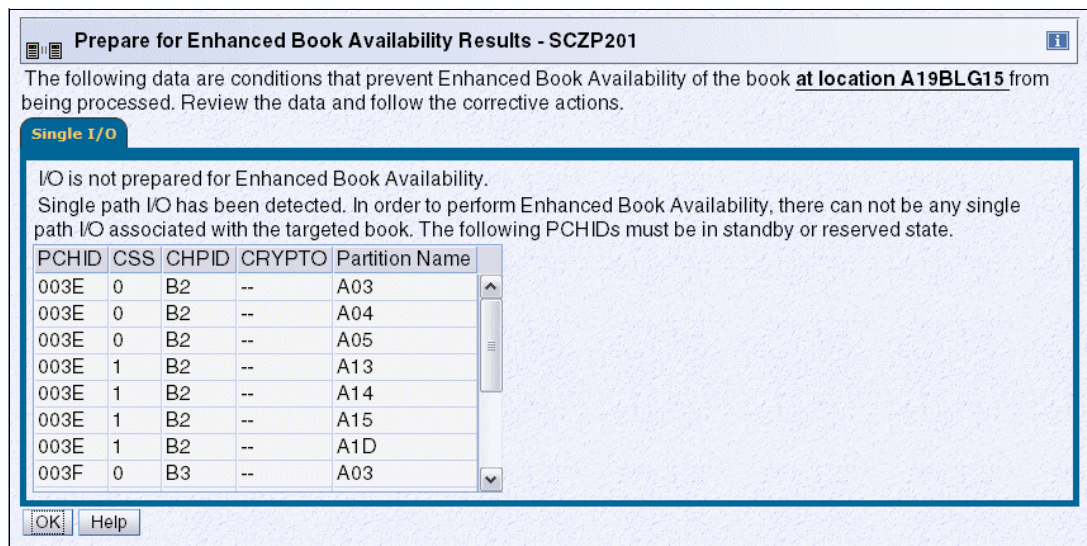


Figure 9-5 Prepare for EBA: Single I/O conditions

Preparing the server to perform enhanced book availability

During the preparation, the system determines the CP configuration that is required to remove the book. Figure 9-6 shows the results and provides the option to change the assignment on non-dedicated processors.

Processor Type	Dedicated Count	Non-Dedicated Count	Processor Totals	LICCC Count
CPU	0	7	7	12
ICF	1	0	1	4
IFL	0	0	0	0
IFA	0	1	1	2
SAP	3	0	3	4
Available to use		0	0	
Remaining Book Totals	4	8	12	

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 some cases, the solution might be to defer the reassignments to another time that would have less impact on the production system images.

After reviewing the reassignment results, and making adjustments if necessary, click **OK**.

The final results of the reassignment, which include changes made as a result of the review, are displayed. See Figure 9-7. These results will be the assignments when the book removal phase of the EBA is completed.

Reassign Non-Dedicated Processors

The following processor allocation will be made if OK is selected.
Select CANCEL if you wish to not make changes or abort the allocation.

Number of CPUs = 7
 Number of ICFs = 0
 Number of IFLs = 0
 Number of IFAs = 1
 Number of zIIPs = 0
 PUs not assigned = {5}

ACT37294

Figure 9-7 Reassign Non-Dedicated Processors, message ACT37294

Summary of the book removal process steps

This section steps through the process of a concurrent book replacement.

To remove a book, the following resources must be moved to the remaining active books:

- ▶ PUs: Enough PUs must be available on the remaining active books, including all types of characterizable PUs (CPs, IFLs, ICFs, zAAPs, zIIPs, and SAPs).
- ▶ Memory: Enough installed memory must be available on the remaining active books.
- ▶ I/O connectivity: Alternate paths to other books must be available on the remaining active books or the I/O path must be taken offline.

By understanding both the server configuration and the LPAR allocation for memory, PUs, and I/O, you should be able to make the best decision about how to free necessary resources and allow for book removal.

To concurrently replace a book:

1. Run the preparation task to determine the necessary resources.
2. Review the results.
3. Determine the actions to perform in order to meet the required conditions for EBA.
4. When you are ready for the book removal, free the resources that are indicated in the preparation steps.
5. Rerun the step in Figure 9-2 on page 287 (Prepare for Enhanced Book Availability task) to ensure that the required conditions are all satisfied.
6. Upon successful completion (Figure 9-8), the system is ready for the removal of the book.

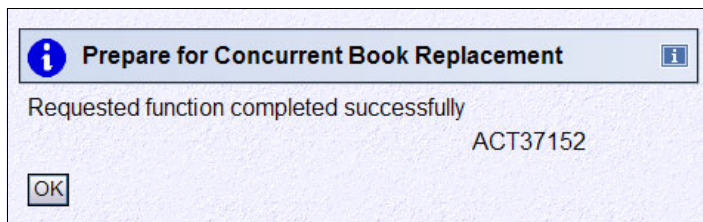


Figure 9-8 Preparation completed successfully, message ACT37152

The preparation process can be run multiple times to ensure that all conditions have been met. It does not reallocate any resources. All it does is to produce a report. The resources are not reallocated until the Perform Book Removal process is invoked.

Rules during EBA

Processor, memory, and single I/O rules during EBA are as follows:

▶ Processor rules

All processors in any remaining books are available to be used during EBA. This includes the two spare PUs or any available PU that is non-LICCC.

The EBA process also allows conversion of one PU type to another PU type. One example is converting a zAAP to a CP for the duration of the EBA function. The preparation for concurrent book replacement task indicates whether any SAPs have to be moved to the remaining books.

► Memory rules

All physical memory installed in the system, including flexible memory, is available for the duration of the EBA function. Any physical installed memory, whether purchased or not, is available to be used by the EBA function.

► Single I/O rules

Alternate paths to other books must be available or the I/O path must be taken offline.

Note: ICB-4s are connected directly to the physical book. Removal of the book requires that all cables be disconnected. Connectivity for ICB-4 is lost during this time. Make sure that redundant paths are available.

Review the results. The result of the preparation task is a list of resources that need to be made available before the book replacement can take place.

Free any resources

At this stage, create a plan to free up these resources. The resources and actions to free them are in the following list:

► To free any PUs:

- Vary the CPs offline, reducing the number of CP in the shared CP pool.
- Use any spare CP.
- Deactivate logical partitions.
- Convert the PU. For example, convert a zAAP to a CP.

► To free memory:

- Deactivate a logical partition.
- Vary offline a portion of the reserved (online) memory. For example, in z/OS issue the command:

```
CONFIG_STOR(E=1),<OFFLINE/ONLINE>
```

This command enables a storage element to be taken offline. Note that the size of the storage element depends on the RSU value. In z/OS, the following command enables you to configure offline smaller amounts of storage than what has been set for the storage element:

```
CONFIG_STOR(nnM),<OFFLINE/ONLINE>
```

- A combination of both logical partition deactivation and varying memory offline.

Note: If you plan to use the EBA function with z/OS logical partitions, we recommend setting up reserved storage and setting 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, allowing for storage elements to be varied offline.

9.4 z10 Enhanced driver maintenance

Enhanced driver maintenance (EDM) is another step in reducing both the necessity and the eventual duration of a planned outage. One of the contributors to planned outages is Licensed Internal Code (LIC) updates that are performed in support of new features and functions.

When properly configured, the z10 EC supports concurrently activating a selected new LIC level. Concurrent activation of the selected new LIC level was previously supported only at specific sync points. Concurrently activating a selected new LIC level anywhere in the maintenance stream is possible. Certain LIC updates are still not supported this way.

The key points of EDM are:

- ▶ HMC is capable of querying whether a system is ready for a concurrent driver upgrade.
- ▶ Firmware upgrades, which require an initial machine load (IML) of System z10 be activated, might not block concurrent driver upgrade.
- ▶ An icon on the Support Element (SE) allows you or IBM support personnel to define the concurrent driver upgrade sync point that is planned for use.
- ▶ Concurrent driver upgrade is supported.
- ▶ The ability to concurrently install and activate a new driver can eliminate a planned outage.
- ▶ Concurrent crossover from driver level N to driver level N+1, to driver level N+2 must be done serially; no composite moves are allowed.
- ▶ Disruptive driver upgrades are permitted at any time.
- ▶ No concurrent back-off is possible. The driver level must move forward to driver level N+1 after enhanced driver maintenance is initiated. Catastrophic errors during an update can lead to an outage.

The EDM function does not completely eliminate the need for planned outages for driver-level upgrades. Although very infrequent, certain circumstances require that the system must be scheduled for an outage. The following circumstances require a planned outage:

- ▶ Specific complex code changes might dictate a disruptive driver upgrade. You are alerted in advance so you can plan for the following changes:
 - Design data fixes
 - CFCC level change
- ▶ Non-QDIO OSA CHPID types, CHPID type OSC, and CHPID type OSE require CHPID Vary OFF/ON in order to activate new code.
- ▶ Cryptographic code loading on a Crypto Express2 feature requires a Config OFF/ON in order to activate new code. For a Crypto Express3 feature the code load is concurrent.
- ▶ FICON and FCP code changes involving code loads require a CHPID *reset* to activate.



Environmental requirements

This chapter briefly describes the environmental requirements for the z10 EC. We list the dimensions, weights, power, and cooling requirements as an overview of what is needed to plan for the installation of a z10 EC server.

For comprehensive physical planning information see *System z10 Enterprise Class Installation Manual for Physical Planning*, GC28-6865.

This chapter discusses the following topics:

- ▶ 10.1, “z10 Power and cooling” on page 296
 - 10.2.1, “Weights” on page 298
 - 10.2.2, “Dimensions” on page 299
- ▶ 10.3, “Power estimation tool” on page 300

10.1 z10 Power and cooling

The z10 EC is always a two-frame system. The frames are shipped separately and are fastened together when installed. Installation is always on a raised floor; the number of cables to be expected for most configurations might be so large that installation is only possible with space underneath.

10.1.1 Power consumption

The power system for z10 EC features front-end bulk power supplies and DCA technology, each of which provides the increased power needed to meet the packaging and power requirements.

The z10 EC requires at least two power feeds and uses up to two three-phase line-cord pairs for the larger models, allowing the system to survive the loss of power to either one of one pair or either two of two pairs. In case of a line-cord power failure, the remaining line-cord is able to take over the entire load to keep the system operating without interruption. For a system with two line-cord pairs, loss of one pair leaves sufficient power to keep the system running. The z10 EC is installed with three-phase wiring and operates with 50/60 Hz ac power, and voltages with a range of 200 - 480 V. For ancillary equipment such as the Hardware Management Console, its display, and its modem, additional single-phase outlets are required.

Table 10-1 shows the line-cord pair requirements for books and cages.

Table 10-1 Line-cord pair requirements

Number of books	One I/O cage	Two I/O cages	Three I/O cages
One book	One pair	One pair	One pair
Two books	One pair	Two pairs	Two pairs
Three books	Two pairs	Two pairs	Two pairs
Four books	Two pairs	Two pairs	Two pairs

Actual power consumption is dependent on the server configuration in terms of the number of books and the number of I/O cages installed. Table 10-2 assumes a maximum configuration.

Table 10-2 Power consumption and heat output

Model	One I/O cage	Two I/O cages	Three I/O cages
E12	9.52 kW 32.5 kBTU/hr	13.26 kW 45.24 kBTU/hr	13.56 kW 46.27 kBTU/hr
E26	13.77 kW 46.98 kBTU/hr	17.51 kW 59.75 kBTU/hr	21.17 kW 72.23 kBTU/hr
E40	16.92 kW 57.73 kBTU/hr	20.66 kW 70.49 kBTU/hr	24.40 kW 83.26 kBTU/hr
E56	19.55 kW 66.71 kBTU/hr	23.29 kW 79.47 kBTU/hr	27.00 kW 92.13 kBTU/hr
E64	19.55 kW 66.71 kBTU/hr	23.29 kW 79.47 kBTU/hr	27.00 kW 92.13 kBTU/hr

Input power in kVA is equal to the output power in kW. Heat output expressed in kBTU per hour is derived from multiplying the table entries by a factor of approximately 3.4.

Larger configurations require up to four line cords, of 60 A, to be used for all power feeds where 208 - 240 V is applicable. We recommend separate circuit breakers for all power feeds, as follows:

- ▶ 32 A for 380 - 415 V
- ▶ 30 A for 480 V

Systems that specify two line cords can be brought up with one line cord and will continue to run without power redundancy. The larger machines that specify four line cords can be brought up with two line cords and will continue to run without power redundancy. The four line cords offer power redundancy, so that when a line cord fails, the remaining cords deliver sufficient power to keep the system up and running.

The same power plug as on z990 and z9 EC is used on z10 EC. Depending on the size of the configuration, four or two power cables are required, as shown in Table 10-3.

Table 10-3 Power cables

Number of books	One cage	Two cages	Three cages
One book	2 x 60 A	2 x 60 A	2 x 60 A
Two books	2 x 60 A	4 x 60 A	4 x 60 A
Three books	4 x 60 A	4 x 60 A	4 x 60 A
Four books	4 x 60 A	4 x 60 A	4 x 60 A

10.1.2 Internal Battery Feature

The optional Internal Battery Feature (IBF) provides sustained system operations for a relatively short period of time, allowing for orderly shutdown. In addition, an external uninterruptible power supply system can be connected, allowing for longer periods of sustained operation.

If the batteries are not older than three years and have been discharged regularly, the IBF is capable of providing emergency power for the periods of time listed in Table 10-4.

Table 10-4 Internal Battery Feature emergency power times

Model	One I/O cage	Two I/O cages	Three I/O cages
E12	10.8 minutes	12 minutes	12 minutes
E26	10.8 minutes	7.2 minutes	5.4 minutes
E40	7.2 minutes	5.4 minutes	4.2 minutes
E56	5.4 minutes	4.2 minutes	2.4 minutes
E64	5.4 minutes	4.2 minutes	2.4 minutes

10.1.3 Emergency power-off

On the front of frame A (shown in Figure 2-1 on page 24) is an emergency power-off switch that, when activated, immediately disconnects utility and battery power from the server. This causes all volatile data in the server to be lost.

If the server is connected to a machine-room's emergency power-off switch, and the Internal Battery Feature is installed, the batteries take over if the switch is engaged.

To avoid take-over, connect the machine-room emergency power-off switch to the server power-off switch. Then, when the machine-room emergency power-off switch is engaged, all power will be disconnected from the line cords and the Internal Battery Features. However, all volatile data in the server will be lost.

10.1.4 Cooling requirements

The z10 EC is an air cooled system. It requires chilled air to fulfill the air-cooling requirements, ideally coming from under the raised floor. The chilled air is usually provided through perforated floor tiles. The amount of chilled air required for a variety of underfloor temperatures in the computer room is indicated in *System z10 Enterprise Class Installation Manual for Physical Planning*, GC28-6865, also known as the IMPP.

At an underfloor temperature of 20°C (68°F), the cooling airflow requirements in cubic meters per minute (CM/M) and cubic feet per minute (CF/M) with maximum populated I/O cages are listed in Table 10-5.

Table 10-5 Underfloor cooling airflow requirements showing CM/M (CF/M)

Model	One I/O cage	Two I/O cages	Three I/O cages
E12	30.3 (1080)	50.97 (1800)	62.30 (2200)
E26	39.62 (1400)	62.30 (2200)	72.16 (2550)
E40	50.97 (1800)	62.30 (2200)	83.43 (2950)
E56	62.30 (2200)	72.16 (2550)	83.43 (2950)
E64	62.30 (2200)	72.16 (2550)	83.43 (2950)

10.2 z10 Physical specifications

This section describes weights and dimensions.

10.2.1 Weights

Because a large number of cables can be connected to a z10 EC installation, a raised floor is mandatory. In the *System z10 Enterprise Class Installation Manual for Physical Planning*, GC28-6864, weight distribution and floor loading tables are published, to be used together with the maximum frame weight, frame width, and frame depth to calculate the floor loading.

Minimum and maximum system weights

Table 10-6 indicates the minimum and maximum system weights for all models. The weight ranges are based on configuration models with one I/O frame and three I/O cages, and with and without IBF.

Table 10-6 System weights

Configuration	Weight in kg (lb) without IBF	Weight in kg (lb) with IBF
E12	1273 - 1674 (2807 - 3692)	1478 - 1983 (3258 - 4372)
E26	1439 - 1856 (3173 - 4092)	1748 - 2165 (3853 - 4772)
E40	1534 - 1933 (3382 - 4261)	1842 - 2241 (4062 - 4941)
E56	1585 - 2010 (3495 - 4430)	1894 - 2318 (4175 - 5110)
E64	1585 - 2010 (3495 - 4430)	1894 - 2318 (4175 - 5110)

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 Table 10-6 shows, the weight of a frame can be substantial, causing a concentrated load on a caster or leveling foot to 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.

Always consult the floor tile manufacturer to determine the load rating of the tile and pedestal structure. Additional panel support might be required to improve the structural integrity, because cable cutouts significantly reduce the floor tile rating.

10.2.2 Dimensions

The z10 EC always has two frames, which are frame A and frame Z. The external dimensions of both frames of a z10 EC, with and without covers, are listed in Table 10-7.

Table 10-7 Frame dimensions

Frames	Width mm (in)	Depth mm (in)	Height mm (in)
Frame A without covers	750 (29.5)	1270 (50.0)	1994 (78.5)
Frame A with covers	770 (30.3)	1803 (71.0)	2015 (79.3)
Frame Z without covers	750 (29.5)	1270 (50.0)	1994 (78.5)
Frame Z with covers	770 (30.3)	1803 (71.0)	2015 (79.3)

Note: The total machine room area required is 2.82 square meters (29.88 square feet). With service clearance, 7.28 square meters (78.26 square feet) are required.

Product comparison dimensions

Compared to the z9 EC, the dimensions of IBM System z10 Enterprise Class differ slightly. For example, you might have to order the height reduction feature (FC #9975) to clear elevator or other entrance doors. When you plan the floor location, consider the differences, which are listed in Table 10-8.

Table 10-8 Product dimensions

Description	z10 EC	Difference compared to z9 EC
Height with covers	201.5 cm (79.3 inches)	+ 7.4 cm (+ 2.9 inches)
Width (two frames with side covers)	154.0 cm (60.6 inches)	+ 0.0 cm (+ 0 inches)
Depth with covers	180.3 cm (71.0 inches)	+ 22.6 cm (+ 8.9 inches)

The front and the rear of the two-frame z10 EC dissipate different amounts of heat. Most of the heat comes from the rear of the server. In the planning phase, consider the proper placement regarding the cooling capabilities of the data center if the physical location requires it. Ensure that the top of the server is kept clear to prevent blocking normal air-cooling output and to prevent the catapulting of items lying on the top when the emergency blowers engage (when the MRUs are having difficulty cooling the server).

Frame tie-down for raised floor and non-raised floor

A bolt-down kit for raised floor environments can be ordered only for the z10 EC frames. The kit provides hardware to enhance the ruggedness of the frames and to tie down the frames to a concrete floor beneath a raised floor of 228–330 mm or 304–558 mm (9–13 inches or 12–22 inches). The kit is offered in the following configurations:

- ▶ The Bolt-Down Kit for Low-Raised Floor (FC 7993) provides frame stabilization and bolt-down hardware for securing a frame to a concrete floor beneath a 235–298 mm (9.25–11.75 in) raised floor.
- ▶ The Bolt-Down Kit for High-Raised Floor (FC 7994) provides frame stabilization and bolt-down hardware for securing a frame to a concrete floor beneath a 298–405 mm (11.75–16.0 in) raised floor.

These kits help to 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 weighing less than 1632 kg (3600 lbs) per frame. Two bolt-down kits are required.

10.3 Power estimation tool

Several aids are available to monitor the power consumption and heat dissipation of the z10 EC. This section summarizes the tools that are available to estimate the energy consumption of the z10 EC. The following tools are available:

- ▶ Power estimation tool
- ▶ System activity display
- ▶ IBM Systems Director Active Energy Manager™

Power estimation tool

The power estimation tool for z10 EC is available through the IBM Resource Link Web site: <http://www.ibm.com/servers/resourceLink>

The tool provides an estimate of the anticipated power consumption of a particular machine model given its configuration. For the z10 EC, you input the machine model, memory size, number of I/O cages, and quantity of each type of I/O feature card. The tool outputs an estimate of the power requirements for your configuration.

The tool helps with power and cooling planning for installed/planned z10s servers.

System activity display with power monitoring

Actual power consumption of the system can be seen on a system activity display (SAD) panel of HMC, shown in Figure 10-1.

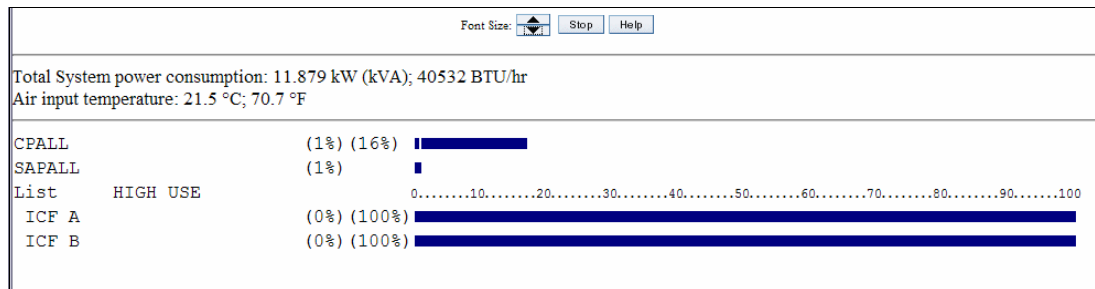


Figure 10-1 Power consumption on SAD

IBM Systems Director Active Energy Manager

IBM Systems Director Active Energy Manager is an energy management solution building-block that returns true control of energy costs to the customer. Active Energy Manager is an industry-leading cornerstone of the IBM energy management framework and is part of the IBM Cool Blue™ portfolio.

Active Energy Manager Version 4.1.1 is a plug-in to IBM Systems Director Version 6.1 and is available for installation on Linux on System z. It can also run on Windows, Linux on IBM System x, Linux, and IBM System p. For more specific information see *Implementing IBM Systems Director Active Energy Manager 4.1.1*, SG24-7780.

Active Energy Manager is an energy management software tool that can provide a single view of the actual power usage across multiple platforms as opposed to the benchmarked or rated power consumption. It can effectively monitor and control power in the data center at the system, chassis, or rack level. By enabling these power management technologies, data center managers can more effectively power manage their systems while lowering the cost of computing.

The following power management functions are available with Active Energy Manager:

- ▶ Power trending

Power trending allows you to monitor, in real time, the consumption of power by a supported power managed object. You use this data to track the actual power consumption of monitored devices and to determine the maximum value over time. The data can be presented either graphically or in tabular form.

- ▶ Thermal trending

Thermal trending allows you to monitor, in real-time, the heat output and ambient temperature of a supported power managed object. Use this data to help avoid situations

where overheating might cause damage to computing assets, and to study how the thermal signature of various monitored devices varies with power consumption. The data can be presented either graphically or in tabular form.

The following data is available from System z HMC:

- ▶ System name, machine type, model, serial number, firmware level
- ▶ Ambient temperature
- ▶ Exhaust temperature
- ▶ Average power usage over a one minute period
- ▶ Peak power usage over a one minute period
- ▶ Limited status and configuration information. This information helps explain changes to the power consumption, called Events, which can be:
 - Changes in fan speed
 - MRU failures
 - Changes between power-off, power-on, and IML-complete states
 - Number of books and I/O cages
 - CBU records expiration(s)

Figure 10-2 shows a sample chart of the data that is available from Active Energy Manager and System z10.

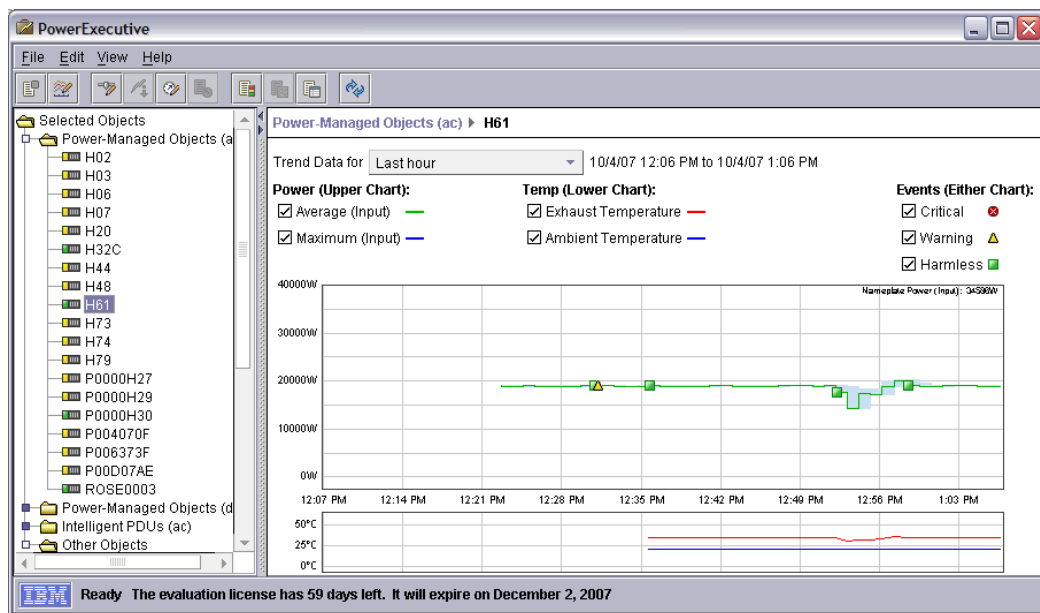


Figure 10-2 Active Energy Manager example chart for z10

IBM Systems Director Active Energy Manager is the first solution on the market that provides customers with the intelligence necessary to effectively manage power consumption in the data center. Active Energy Manager, which is an extension to IBM Director systems management software, enables you to *meter* actual power usage and trend data for any single physical system or group of systems. Developed by IBM Research, Active Energy Manager uses monitoring circuitry, developed by IBM, to help identify how much actual power is being used and the temperature of the system.



Hardware Management Console

In the past several years the Hardware Management Console (HMC) has been enhanced to support many functions and tasks to extend the management capabilities of System z. This is also true with the System z10 servers and will continue in the future. Therefore, the HMC becomes more important in the overall management of the data center infrastructure.

This chapter describes the z10 EC HMC and Support Element (SE). It is intended to give an overview of the HMC and SE functions.

This chapter discusses the following topics:

- ▶ 11.1, “HMC and SE introduction” on page 304
- ▶ 11.2, “HMC and SE connectivity” on page 304
- ▶ 11.3, “Remote Support Facility” on page 308
- ▶ 11.4, “HMC remote operations” on page 308
- ▶ 11.5, “z10 EC HMC and SE key capabilities” on page 309

11.1 HMC and SE introduction

The Hardware Management Console (HMC) is a combination of a stand alone computer and 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 IBM System z servers. It manages System z hardware, its logical partitions, and provides support applications.

An HMC is required to operate a System z10 server. A single console can manage multiple System z servers and can be located in local or remote site.

The Support Elements (SEs) are a pair of integrated ThinkPads that are supplied with the System z server. One of them is always the active SE and the other is strictly the alternate element. 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 System z server.

One HMC can control up to 100 SEs and one SE can be controlled by up to 32 HMCs.

At the time of this writing, the z10 EC is shipped with HMC version 2.10.2, which is capable of supporting different System z server types. Many functions that are available on Version 2.10.0 and later are only supported when connected to a System z10 server. HMC Version 2.10.2 supports the servers and SE versions shown in Table 11-1.

Table 11-1 System z10 HMC server support summary

Server	Machine type	Minimum firmware driver	Minimum SE version
z10 BC and z10 BC	2098	76	2.10.1
z10 EC	2097	73	2.10.0
z9 BC	2096	67	2.9.2
z9 EC	2094	67	2.9.2
z890	2086	55	1.8.2
z990	2084	55	1.8.2
z800	2066	3G	1.7.3
z900	2064	3G	1.7.3
9672 G6	9672/9674	26	1.6.2
9672 G5	9672/9674	26	1.6.2

11.2 HMC and SE connectivity

Although the HMC has two Ethernet adapters, each SE has one Ethernet adapter and both are connected to the same Ethernet switch. The Ethernet switch (FC 0089) is supplied with every system order. Additional Ethernet switches (up to a total of ten) may be added.

The switch is a standalone unit located outside the frame and it operates on building AC power. A customer-supplied switch may be used if it matches IBM specifications.

The internal LAN for the SEs (on the System z10 server) connects to the Bulk Power Hub. The HMC must be connected to the Ethernet switch through one of its Ethernet ports. Only the switch may be connected to the customer ports J01 and J02 on the Bulk Power Hub. Other server's SEs may also be connected to the switches. To provide redundancy for the HMCs, two switches are recommended, as shown in Figure 11-1.

For more information see *System z10 Enterprise Class Installation Manual for Physical Planning*, GC28-6865.

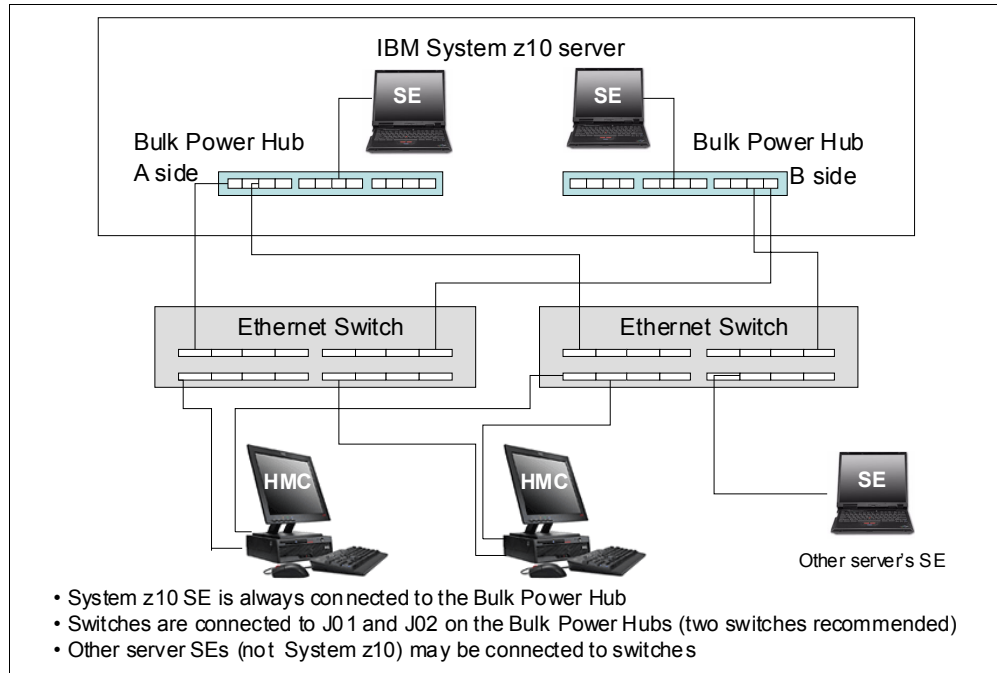


Figure 11-1 HMC to SE connectivity

The HMC and SE have several exploiters that either require or can take advantage of broadband connectivity to the Internet and your corporate intranet.

Several methods are available for setting up the network to allow access to the HMC from your corporate intranet or to allow the HMC to access the Internet. The method you select depends on your connectivity and security requirements.

One example is to connect the second Ethernet port of the HMC to a separate switch that has access to the intranet or Internet, as shown in Figure 11-2.

Also, the HMC has built-in firewall capabilities to protect the HMC and SE environment. The HMC firewall can be set up to allow certain types of TCP/IP traffic between the HMC and permitted destinations in your corporate intranet or the Internet.

Note: Configuration of network components, such as routers or firewall rules, is beyond the scope of this document. Anytime networks are interconnected, security exposure can exist. Network security is the customer's responsibility.

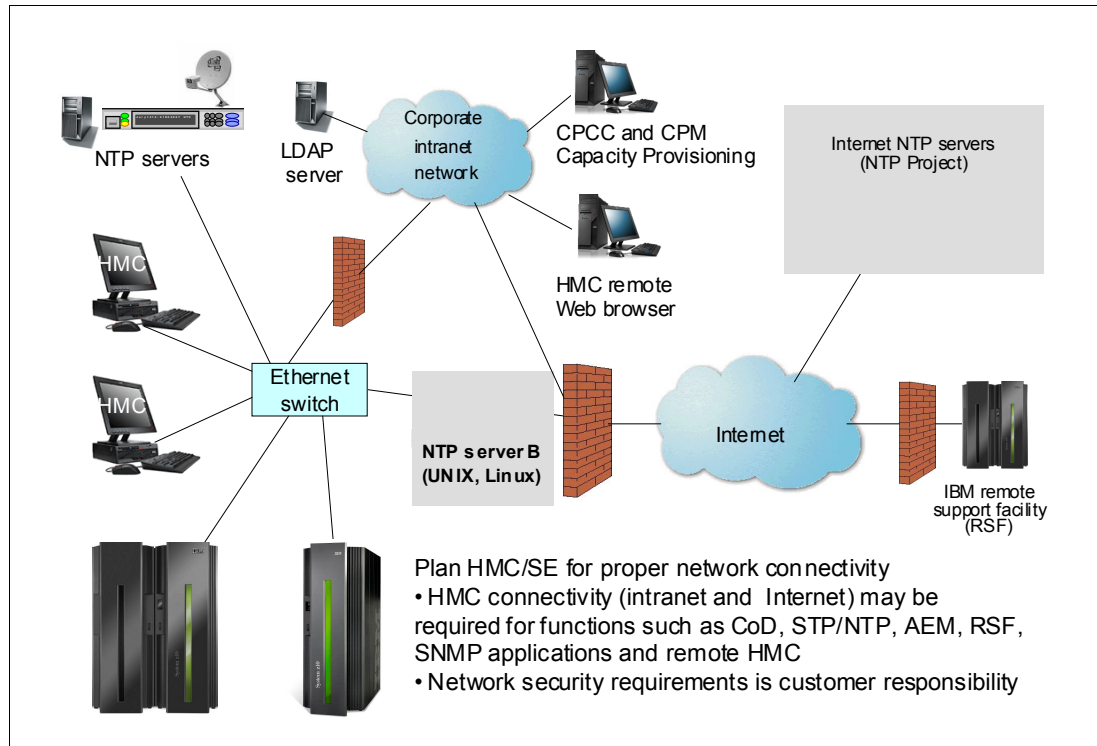


Figure 11-2 HMC connectivity

The HMC and SE network connectivity should be planned carefully to allow for current and future use. Many of the System z capabilities benefit from the various network connectivity options available. For example, several functions available to the HMC that depend on the HMC connectivity are:

- ▶ LDAP support that can be used for HMC user authentication
- ▶ STP and NTP client/server support
- ▶ RSF is available through the HMC with an Internet-based connection, providing increased performance as compared to dial-up
- ▶ Enablement of the SNMP and CIM APIs to support automation or management applications such as Capacity Provisioning Manager and Active Energy Manager (AEM)

TCP/IP Version 6 on HMC and SE

HMC and SE have been enhanced to support IPv6. The HMC and SE can communicate using IPv4, IPv6 or both IPv4 and IPv6. Assigning a static IP address to an SE is

unnecessary if the SE only has to communicate with HMCs on the same subnet. An HMC and SE can use IPv6 link-local addresses to communicate with each other.

IPv6 link-local addresses characteristics are:

- ▶ Every IPv6 network interface is assigned a link-local IP address.
- ▶ A link-local address is for use 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.

Assigning addresses to HMC and SE

An HMC can have the following IP addresses:

- ▶ Statically assigned IPv4 or statically assigned IPv6
- ▶ DHCP assigned IPv4 or DHCP assigned IPv6
- ▶ Autoconfigured IPv6:
 - Link-local is assigned to every network interface.
 - Router-advertised, which is broadcast from the router, can be combined with a MAC address to create a totally unique address.
 - Privacy extensions can be enabled for these addresses as a way to avoid using MAC address as part of address to ensure uniqueness.

An SE can have the following IP addresses:

- ▶ Statically assigned IPv4 or statically assigned IPv6
- ▶ Autoconfigured 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 to automatically discover SEs. The HMC Network Diagnostic Information task may 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, written as eight 16-bit hex blocks 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. This is where the leading zeros can be omitted and consecutive zeros can be replaced with a double colon. The address in the previous example can also be written as:

```
2001:db8::202:B3FF:fe1e:8329
```

For remote operations using a Web browser, if an IPv6 address is assigned to the HMC, navigate to it by specifying that address. The address must be surrounded with square brackets in the browser's address field:

```
https://[fdab:1b89:fc07:1:201:6cff:fe72:ba7c]
```

Using link-local addresses must be supported by browsers.

11.3 Remote Support Facility

The HMC Remote Support Facility (RSF) provides communication to a centralized IBM support network for hardware problem reporting and service. The types of communication provided include:

- ▶ Problem reporting and repair data
- ▶ Fix delivery to the service processor and HMC
- ▶ Hardware inventory data
- ▶ On-demand enablement

The HMC can be configured to send hardware service related information to IBM by using a dialup connection over a modem or using an Internet connection. The advantages of using an Internet connection include:

- ▶ Significantly faster transmission speed
- ▶ Ability to send more data on an initial problem request, potentially resulting in more rapid problem resolution
- ▶ Reduced customer expense (for example, the cost of a dedicated analog telephone line)
- ▶ Greater reliability

Unless the enterprise's security policy prohibits any connectivity from the HMC over the Internet, an Internet connection is recommended.

If both types of connections are configured, the Internet will be tried first, and if it fails, then the modem is used.

The following security characteristics are in effect regardless of the connectivity method chosen:

- ▶ Remote Support Facility requests are always initiated from the HMC to IBM. An inbound connection is never initiated from the IBM Service Support System.
- ▶ All data transferred between the HMC and the IBM Service Support System is encrypted in a high-grade Secure Sockets Layer (SSL) encryption.
- ▶ When initializing the SSL encrypted connection the HMC validates the trusted host by its digital signature issued for the IBM Service Support System.
- ▶ Data sent to the IBM Service Support System consists solely of hardware problems and configuration data. No application or customer data is transmitted to IBM.

11.4 HMC remote operations

The z10 EC HMC application simultaneously supports one local user and any number of remote users. Remote operations provide the same interface used by a local HMC operator. The two ways to perform remote manual operations are:

- ▶ Using a Remote HMC

A remote HMC is an HMC that is on a different subnet from the SE, therefore the SE cannot be automatically discovered with IP multicast.

- ▶ Using a Web browser to connect to an HMC

The choice between a remote HMC and a Web browser connected to a local HMC is determined by the scope of control needed. A remote HMC can control only a specific set of

objects, but a Web browser connected to a local HMC controls the same set of objects as the local HMC.

In addition, consider communications connectivity and speed. LAN connectivity provides acceptable communications for either a remote HMC or Web browser control of a local HMC, but dialup connectivity is only acceptable for occasional Web browser control.

Using a remote HMC

Although a remote HMC is a complete HMC, its connection configuration differs from a local HMC. As a complete HMC, it requires the same setup and maintenance as other HMCs (Figure 11-2 on page 306).

A remote HMC requires TCP/IP connectivity to each SE to be managed. Therefore, any existing customer-installed firewall between the remote HMC and its managed objects must permit communications between the HMC and SE. For service and support, the remote HMC also requires connectivity to IBM, or to another HMC with connectivity to IBM.

Using a Web browser

Each HMC contains a Web server that can be configured to allow remote access for a specified set of users. When properly configured, an HMC can provide a remote user with access to all the functions of a local HMC except those that require physical access to the diskette or DVD media. The user interface in the browser is the same as the local HMC and has the same functionality as the local HMC.

The Web browser can be connected to the local HMC by using either a LAN TCP/IP connection or a switched, dial-up, or network PPP TCP/IP connection. Both connection types use only encrypted (HTTPS) protocols, as configured in the local HMC. If a PPP connection is used, the PPP password must be configured in the local HMC and in the remote browser system. 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.

11.5 z10 EC HMC and SE key capabilities

The z10 EC comes with HMC application Version 2.10.2. For a complete list of HMC functions see *System z HMC Operations Guide Version 2.10.2, SC28-6881*.

Version 2.10.2 includes a number of enhancements:

- ▶ **Digitally signed firmware**

One critical issue with firmware upgrades is security and data integrity. Procedures are in place to use a process to digitally sign the firmware update files on the HMC, the SE, and the TKE. Using a hash-algorithm, a message digest is generated that is then encrypted with a private key to produce a digital signature. This operation ensures that any changes made to the data will be detected during the upgrade process. It helps ensure that no malware can be installed on System z products during firmware updates. It enables, with other existing security functions, System z10 CPACF functions to comply with Federal Information Processing Standard (FIPS) 140-2 Level 1 for Cryptographic Licensed Internal Code (LIC) changes. The enhancement follows the System z focus of security for the HMC and the SE.

- ▶ **Serviceability enhancements for FICON channels:**

Simplified problem determination to more quickly detect fiber optic cabling problems in a Storage Area Network.

All FICON channel error information is forwarded to the HMC, thus facilitating detection and reporting trends and thresholds for the channels with aggregate views including data from multiple systems.

- ▶ Optional user password on disruptive confirmation

The requirement to supply a user password on disruptive confirmation is optional. The general recommendation remains to require a password.

- ▶ Improved consistency of confirmation panels on the HMC and the SE

Attention indicators are on the top of panels, and there will be a list of objects affected by the action, target, and secondary objects, for example, LPARs if the target is CPC.

11.5.1 CPC management

The HMC is the primary place for central processor complex (CPC) control. For example, to define hardware to z10 EC, I/O configuration data set (IOCDS) must be defined. The IOCDS contains definitions of logical partitions, channel subsystems, control units and devices and their accessibility from logical partitions. IOCDS can be created and put into production from the HMC.

The z10 EC server is powered on and off from the HMC. HMC is used to initiate power-on reset (POR) of the server. During the POR, among other things, PUs are characterized and placed into their respective pools, memory is put into a single main storage pool and IOCDS is loaded and initialized into the hardware system area.

The Hardware messages task displays hardware-related messages on the CPC level, a logical partition level, SE level, or hardware messages related to the HMC itself.

11.5.2 LPAR management

Use HMC to define logical partition properties, such as how many processors of each type, how many are reserved, or how much memory is assigned to it. These parameters are defined in logical partition profiles and they are stored on the SE.

Because PR/SM has to manage logical partition access to processors and initial weights of each partition, weights are used to prioritize partition access to processors.

A Load task on the HMC enables you to IPL an operating system. It causes a program to be read from a designated device and initiates that program. The operating system can be IPLed from disk, HMC CD-ROM/DVD, or FTP server.

When a logical partition is active and an operating system is running in it, you may use the HMC to change certain logical partition parameters dynamically. The HMC also provides an interface to change partition weight, add logical processors to partitions, and add memory.

LPAR weights can be also changed through a scheduled operation. Use the HMC's Customize Scheduled Operations task to define the weights that will be set to logical partitions at the scheduled time.

Channel paths can be configured on and off dynamically, as needed, for each partition from an HMC.

11.5.3 Operating system communication

The Operating system messages task displays messages from a logical partition. You may also enter operating system commands and interact with the system.

HMC also provides integrated 3270 and ASCII consoles so you can access an operating system without requiring other network or network devices (such as TCP/IP or control units).

11.5.4 SE access

To use an SE, being physically close to it is not necessary. Use the HMC to remotely access the SE; the same interface is provided on the SE.

The HMC enables you to:

- ▶ Synchronize content of the primary SE to the alternate SE
- ▶ Determine if a switch from primary to the alternate can be performed
- ▶ Switch between primary and alternate SEs

11.5.5 Monitoring

Use the System Activity Display (SAD) task on the HMC to monitor the activity of one or more CPCs. The task monitors processor and channel usage. You may define multiple activity profiles. The task also includes power monitoring information, the power being consumed, and the air input temperature for the server.

For HMC users with Service authority, SAD shows information about each power cord. Power cord information should only be used by those with extensive knowledge about System z10 internals and three-phase electrical circuits. Weekly call-home data includes power information for each power cord.

IBM Systems Director Active Energy Manager

As discussed in “IBM Systems Director Active Energy Manager” on page 301, the Active Energy Manager is an energy management solution building-block that returns true control of energy costs to the customer. It is a software tool that provides a single view of the actual power usage across multiple platforms and helps to increase energy efficiency by controlling power use across the data center.

Active Energy Manager runs on Windows, Linux on IBM System x®, Linux on IBM System p, and Linux on IBM System z.

How Active Energy Manager works

Active Energy Manager interacts with systems as follows:

- ▶ Hardware, firmware, and systems management software in servers and blades provide information to Active Energy Manager.
- ▶ Active Energy Manager calculates the power consumption for each component and tracks power usage over time.
- ▶ When power is constrained, Active Energy Manager allows power to be allocated on a server-by-server basis.

- ▶ Active Energy Manager ensures that limiting the power consumption does not affect performance
- ▶ Sensors and alerts warn the user if limiting power to a particular server can affect performance

Data available from z10 EC HMC

The following data is available from the z10 EC HMC:

- ▶ System name, machine type, model, serial number, firmware level
- ▶ Ambient temperature
- ▶ Exhaust temperature
- ▶ Average power (over a one-minute period)
- ▶ Peak power (over a one-minute period)
- ▶ Limited status and configuration information. This information helps explain changes to the power consumption, called Events, which can be:
 - Changes in fan speed
 - Changes between power-off, power-on, and IML-complete states
 - Number of I/O drawers
 - CBU records expiration(s)

11.5.6 HMC Console Messenger

The Console Messenger task provides basic messaging capabilities between users of the HMC and the SE.

Console Messenger provides:

- ▶ Basic messaging capabilities that allow system operators or administrators to coordinate their activities
- ▶ Messaging capability to HMC and SE users
- ▶ Messaging between local and remote users by using existing HMC and SE interconnection protocols
- ▶ Interactive chats between two partners with send and receive messages, and chat history displayed in the task panel
- ▶ Broadcast message to all sessions on a selected console, with ability to send one-shot message to all sessions on a selected console
- ▶ Plain text messages in chats and broadcast messages

Console Messenger also allows messaging between sessions on remote consoles that can be reached by using existing communication facilities (Figure 11-3 on page 313).

From an HMC, the reachable console set consists of:

- ▶ Other HMCs that are in the same security domain and that are automatically discovered through console framework communication discovery
- ▶ Any additional HMCs manually configured as data replication partners
- ▶ Any SEs that are being managed by this HMC
- ▶ Any HMCs that are also managing those SEs, even if not discovered or reachable by using normal HMC framework communication (indirect path)

and enters CoD requests. For this reason, SNMP must be configured and enabled on the HMC.

For additional information about using and setting up CPM, see the publications:

- ▶ *z/OS MVS Capacity Provisioning User's Guide, SA33-8299*
- ▶ *IBM System z10 Enterprise Class Capacity On Demand, SG24-7504*

11.5.8 Server Time Protocol support

Server Time Protocol (STP) is supported on System z servers. With the STP functions, the role of the HMC has been extended to provide the user interface for managing the Coordinated Timing Network (CTN).

In a mixed CTN (one containing both STP and Sysplex Timer), the HMC can be used to:

- ▶ Initialize or modify the CTN ID and ETR port states.
- ▶ Monitor the status of the CTN.
- ▶ Monitor the status of the coupling links initialized for STP message exchanges.

In an STP-only CTN, the HMC can be used to:

- ▶ Initialize or modify the CTN ID.
- ▶ Initialize the time, manually or by dialing out to a time service, so that the Coordinated Server Time (CST) can be set to within 100 ms of an international time standard, such as UTC.
- ▶ Initialize the time zone offset, daylight saving time offset, and leap second offset.
- ▶ Schedule periodic dial-outs to a time service so that CST can be steered to the international time standard.
- ▶ Assign the roles of preferred, backup, and current time servers, as well as arbiter.
- ▶ Adjust time by up to plus or minus 60 seconds.
- ▶ Schedule changes to the offsets listed. STP can automatically schedule daylight saving time, based on the selected time zone.
- ▶ Monitor the status of the CTN.
- ▶ Monitor the status of the coupling links initialized for STP message exchanges.

For additional planning and setup information, see the following publications:

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

11.5.9 NTP client/server support on HMC

The Network Time Protocol (NTP) client support allows an STP-only Coordinated Timing Network (CTN) to use an NTP server as an External Time Source (ETS) that addresses the requirement for:

- ▶ Customers who want time accuracy for the STP-only CTN
- ▶ Using a common time reference across heterogeneous platforms

NTP client allows the same accurate time across an enterprise comprised of heterogeneous platforms.

NTP server becomes the single time source, ETS for STP, as well as other servers that are not System z (such as UNIX, Windows NT®, and others) that have NTP clients.

When the HMC is configured to have an NTP client running, the HMC time will be continuously synchronized to an NTP server instead of synchronizing to the SE.

HMC can also act as an NTP server. With this support, z10 EC can get time from HMC without accessing 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 separate LAN is recommended from the HMC/SE LAN.

The NTP client support can be used to connect to other NTP servers that can potentially receive NTP through the Internet. When using another NTP server, then the NTP server becomes the single time source, ETS for STP, and other servers that are not System z servers (such as UNIX, Windows NT, and others) that have NTP clients.

When the HMC is configured to have an NTP client running, the HMC time will be continuously synchronized to an NTP server instead of synchronizing to a support element.

For additional planning and setup information for STP and NTP check the following manuals:

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

11.5.10 System Input/Output Configuration Analyzer on the SE/HMC

A System Input/Output Configuration Analyzer task is provided that supports the system I/O configuration function.

The information necessary to manage a system's I/O configuration has to be obtained from many separate applications. A System Input/Output Configuration Analyzer task enables the system hardware administrator to access, from one location, the information from these many sources. Managing I/O configurations then becomes easier, particularly across multiple servers.

The System Input/Output Configuration Analyzer task performs 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 FICON channels support remote node ID information, which is also collected.

The System Input/Output Configuration Analyzer is a view-only tool. It does not offer any options other than viewing options. With the tool, data is formatted and displayed in five different views, various sort options, are available, and data can be exported to a USB flash drive for a later viewing.

The five views are:

- ▶ PCHID Control Unit View, which shows PCHIDs, CSS, CHPIDs and their control units
- ▶ PCHID Partition View, which shows PCHIDS, CSS, CHPIDs and the partitions they are in
- ▶ Control Unit View, which shows the control units, their PCHIDs, and their link addresses in each CSS

- ▶ Link Load View, which shows the Link address and the PCHIDs that use it
- ▶ Node ID View, which shows the Node ID data under the PCHIDs

11.5.11 Network Analysis Tool for SE Communication

The Network Analysis Tool tests that communication between the HMC and SE is available.

The tool performs five tests:

- ▶ HMC pings SE.
- ▶ HMC connects to SE and also verifies the SE is at the correct level.
- ▶ HMC sends a message to SE and receives a response.
- ▶ SE connects back to HMC.
- ▶ SE sends a message to HMC and receives a response.

11.5.12 Automated operations

As an alternative to manual operations, a computer can interact with the consoles through an application programming interface (API). The interface allows a program to monitor and control the hardware components of the system in the same way a human can monitor and control the system. The HMC APIs provide monitoring and control functions through TCP/IP, SNMP, and CIM to an HMC. These APIs provide the ability to get and set a managed object's attributes, issue commands, receive asynchronous notifications, and generate SNMP traps.

The HMC supports Common Information Model (CIM) as an additional systems management API. The focus is on attribute query and operational management functions for System z, such as CPCs, images, activation profiles. The System z10 contains a number of enhancements to the CIM systems management API. The function is similar to that provided by the SNMP API.

For additional information about APIs, see the *System z Application Programming Interfaces*, SB10-7030.

11.5.13 Cryptographic support

The z10 EC includes both standard cryptographic hardware and optional cryptographic features for flexibility and growth capability.

The HMC/SE interface provides the capability to:

- ▶ Define the cryptographic controls
- ▶ Dynamically add a Crypto to a partition for the first time
- ▶ Dynamically add a Crypto to a partition already using Crypto
- ▶ Dynamically remove Crypto from a partition

A Usage Domain Zeroize task is provided to clear the appropriate partition crypto keys for a given usage domain when removing a crypto card from a partition. For detailed set-up information, see *IBM System z10 Enterprise Class Configuration Setup*, SG24-7571.

11.5.14 z/VM virtual machine management

HMC can be used for basic management of z/VM and its virtual machines. HMC exploits the z/VM Systems Management Application Programming Interface (SMAPI) and provides a graphical user interface (GUI)-based alternative to the 3270 interface.

Monitoring the status information and changing the settings of z/VM and its virtual machines are possible. From the HMC interface, virtual machines can be activated, monitored, and deactivated.

Authorized HMC users can obtain various status information, such as:

- ▶ Configuration of the particular z/VM virtual machine
- ▶ z/VM image-wide information about virtual switches and guest LANs
- ▶ Virtual Machine Resource Manager (VMRM) configuration and measurement data

The activation and deactivation of z/VM virtual machines is integrated into the HMC interface. You can select the Activate and Deactivate tasks on CPC and CPC image objects, and for virtual machines management.

An event monitor is a trigger that is listening for events from objects managed by HMC. When z/VM virtual machines change their status, they generate such a events. You can create event monitors to handle the events coming from z/VM virtual machines. For example, selected users can be notified by an e-mail message if the virtual machine changes status from Operating to Exceptions, or any other state.

In addition, in z/VM V5.4, the APIs can perform the following functions:

- ▶ Create, delete, replace, query, lock, and unlock directory profiles
- ▶ Manage and query LAN access lists (granting and revoking access to specific user IDs)
- ▶ Define, delete, and query virtual CPUs, within an active virtual image and in a virtual image's directory entry
- ▶ Set a maximum number of virtual processors that can be defined in a virtual image's directory entry

11.5.15 Installation support for z/VM using the HMC

The traditional way of installing Linux on System z in the z/VM virtual machine requires a network connection to a file server that is hosting the installation files of the Linux distribution.

Starting with z/VM 5.4 and System z10, Linux on System z can be installed in a z/VM virtual machine from the HMC workstation DVD drive. This Linux on System z installation can exploit the existing communication path between the HMC and the SE, where *no external network and no additional network setup is necessary* for the installation. This simplification can eliminate potential customer concerns and additional configuration efforts.

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 publications

For information about ordering these publications, see “How to get Redbooks publications” on page 321. Note that several documents referenced here might be available in softcopy only.

- ▶ *IBM System z10 Enterprise Class Technical Introduction*, SG24-7515
- ▶ *IBM System z10 Enterprise Class Technical Guide*, SG24-7516
- ▶ *Getting Started with InfiniBand on System z10 and System z9*, SG24-7539
- ▶ *IBM System z Connectivity Handbook*, SG24-5444
- ▶ *Server Time Protocol Planning Guide*, SG24-7280
- ▶ *Server Time Protocol Implementation Guide*, SG24-7281
- ▶ *IBM System z10 Enterprise Class Configuration Setup*, SG24-7571
- ▶ *IBM System z10 Enterprise Class Capacity On Demand*, SG24-7504

Other publications

These publications are also relevant as further information sources:

- ▶ *Hardware Management Console Operations Guide Version 2.10.0*, SC28-6867
- ▶ *Support Element Operations Guide V2.10.0*, SC28-6868
- ▶ *IOCP User's Guide*, SB10-7037
- ▶ *Stand-Alone Input/Output Configuration Program User's Guide*, SB10-7152
- ▶ *Planning for Fiber Optic Links*, GA23-0367
- ▶ *System z10 Enterprise Class Capacity on Demand User's Guide*, SC28-6871
- ▶ *CHPID Mapping Tool User's Guide*, GC28-6825
- ▶ *Common Information Model (CIM) Management Interfaces*, SB10-7154
- ▶ *System z10 Enterprise Class Installation Manual*, GC28-6865
- ▶ *System z10 Enterprise Class Installation Manual for Physical Planning*, GC28-6865
- ▶ *System z10 Enterprise Class Processor Resource/Systems Manager Planning Guide*, SB10-7153
- ▶ *System z10 Enterprise Class System Overview*, SA22-1084
- ▶ *System z10 Enterprise Class Service Guide*, GC28-6866
- ▶ *System z Functional Matrix*, ZSW0-1335
- ▶ *z/Architecture Principles of Operation*, SA22-7832

- ▶ *z/OS Cryptographic Services Integrated Cryptographic Service Facility Administrator's Guide, SA22-7521*
- ▶ *z/OS Cryptographic Services Integrated Cryptographic Service Facility Application Programmer's Guide, SA22-7522*
- ▶ *z/OS Cryptographic Services Integrated Cryptographic Service Facility Messages, SA22-7523*
- ▶ *z/OS Cryptographic Services Integrated Cryptographic Service Facility Overview, SA22-7519*
- ▶ *z/OS Cryptographic Services Integrated Cryptographic Service Facility System Programmer's Guide, SA22-7520*

Online resources

These Web sites are also relevant as further information sources:

- ▶ Resource Link
<http://www.ibm.com/servers/resourceLink>
- ▶ IBM Crypto Cards Web site
<http://www.ibm.com/security/cryptocards>
- ▶ Large Systems Performance Reference (LSPR) for IBM System z
<http://www.ibm.com/servers/eserver/zseries/lSpr/>
- ▶ System z Application Assist Processor (zAAP)
<http://www.ibm.com/systems/z/advantages/zaap/index.html>
- ▶ System z Integrated Information Processor (zIIP)
<http://www.ibm.com/systems/z/advantages/ziip/about.html>
- ▶ Parallel Sysplex
<http://www.ibm.com/systems/z/advantages/ps0/index.html>
- ▶ CFSizer
<http://www.ibm.com/systems/support/z/cfsizer/>
- ▶ I/O Connectivity
<http://www.ibm.com/systems/z/hardware/connectivity/index.html>
- ▶ System z New Application License Charges
<http://www.ibm.com/servers/eserver/zseries/swprice/zna1c/>
- ▶ IBM System z Software Pricing: Other Monthly License Charge Metrics
<http://www.ibm.com/servers/eserver/zseries/swprice/other/>
- ▶ Green Data Center
<http://www.ibm.com/systems/greendc/>
- ▶ InfiniBand Trade Association
<http://www.infinibandta.org/home>

- ▶ For the most current planning information about each operating system:
 - z/OS
<http://www.ibm.com/systems/support/z/zos/>
 - z/VM
<http://www.ibm.com/systems/support/z/zvm/>
 - z/TPF
<http://www.ibm.com/software/htp/tpf/pages/maint.htm>
 - z/VSE
<http://www.ibm.com/servers/eserver/zseries/zvse/support/preventive.html>
 - Linux on System z
<http://www.ibm.com/systems/z/os/linux/>

How to get Redbooks publications

You can search for, view, or download Redbooks publications, Redpapers publications, Technotes, draft publications and Additional materials, as well as order hardcopy Redbooks publications, at this Web site:

ibm.com/redbooks

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

Index

Numerics

- 50.0 μm 138
- 60 logical partitions support 205
- 62.5 μm 138
- 63.75K subchannels 165, 206

A

- A frame 24
- activated capacity 235
- Active Energy Manager 300–301, 311
- Advanced Encryption Standard (AES) 68, 172, 187
- application preservation 86

B

- billable capacity 235
- book 235
 - channel definition 161
 - logical structure 62
 - ring topology 64
 - upgrade 45
- branch history table (BHT) 70
- Bulk Power Assembly 26

C

- cage
 - CEC cage 14, 24
 - I/O cage 29, 179, 241, 250–251, 296
- capacity 235
- Capacity Backup
 - See CBU
- Capacity for Planned Events
 - See CPE
- Capacity marked CP 46
- capacity marker 46
- Capacity on Demand (CoD) 73–75, 78, 236, 246, 248, 250
- Capacity Provisioning Control Center 264
- Capacity Provisioning Domain 264–265
- Capacity Provisioning Manager 206, 236, 263, 266
- Capacity Provisioning Policy 266
- capacity setting 75, 235–236
- CBU 73–74, 87, 235, 238, 244–245, 254, 266, 268–269
 - activation 270
 - contract 245
 - conversions 53
 - deactivation 271
 - example 272
 - testing 271
- CBU for CP 51
- CBU for IFL 51
- central processor
 - See CP

- central processor complex
 - See CPC
- central storage (CS) 89, 96
- CFCC 9, 74, 94
- CFLEVEL 98
- Channel Data Link Control (CDLC) 214
- channel path identifier
 - See CHPID
- channel spanning 167
- channel subsystem
 - See CSS
- Chinese Remainder Theorem (CRT) 181
- chip lithography 31
- CHPID 93, 159–160, 167, 169
 - mapping tool 93, 166, 169
- CIU facility 236, 251
- Common Cryptographic Architecture 175, 187
- compression unit 67–68
- concurrent book add (CBA) 235
- concurrent book replacement 286–287, 291
- concurrent hardware upgrade 241
- concurrent memory upgrade 87, 280
- configuration report 44
- Configurator for e-business 169
- configuring for availability 43
- Console Messenger 312
- control unit 160
- cooling requirements 298
- coupling facility (CF) 9, 73–74, 76, 89, 98, 268, 270
 - mode 94
- Coupling Facility Control Code
 - See CFCC
- coupling link 9, 26, 63, 149
 - peer mode 9
- CP 54, 58, 67–68, 70, 72, 74, 158, 171–172, 177, 240
 - assigned 48
 - conversion 10
 - CP4 feature 75
 - CP5 feature 75
 - CP6 feature 75
 - CP7 feature 75
 - enhanced book availability 45
 - logical processors 84
 - pool 73–74
 - sparing 85
- CP Cryptographic Assist Facility (CPACF) 68
- CPACF 177
 - cryptographic capabilities 16
 - definition of 68
 - design highlights 58
 - feature code 177
 - instructions 72
 - PU design 68
- CPC
 - logical partition resources 91

- management 310
- CPE 235, 238, 266
- CPM 236
- Crypto enablement 176
- Crypto Express2 9, 14, 16, 26, 63, 154, 174–175, 178–181, 183, 274
 - accelerator 16, 179–180, 182, 186
 - coprocessor 16, 174, 176, 178–180, 182, 185–186
 - feature 176
 - support 218
- cryptographic
 - accelerator 178
 - asynchronous functions 173
 - domain 180–181
 - feature codes 176
 - features comparison 186
 - synchronous function 172
- Cryptographic Accelerator (CA) 182
- Cryptographic Coprocessor (CC) 182
- cryptology
 - Advanced Encryption Standard (AES) 16
 - Secure Hash Algorithm (SHA) 16
- CSS 83, 158, 160, 169
 - components 159
 - configuration management 169
 - definition of 8
 - ID 95
 - image ID 159
 - structure 160
- Customer Initiated Upgrade (CIU) 73
 - activation 254
 - Ordering 253
- customer profile 236

D

- data chaining 226
- Data Encryption Standard (DES) 68, 171–172, 174, 184
- DFSMS striping 227
- Digital Signature Verify (CSFNDFV) 175
- display ios.config 165
- disruptive upgrades 277
- Distributed Converter Assemblies (DCAs) 27
- double-key DES 172, 174
- double-key MAC 172
- dynamic coupling facility dispatching 77
- dynamic I/O configuration 108
- dynamic LPAR memory upgrade 205
- dynamic oscillator switchover 280
- dynamic PU exploitation 206
- dynamic SAP sparing and reassignment 86
- dynamic storage reconfiguration (DSR) 90, 100

E

- Electronic Industry Association (EIA) 24
- emergency power-off 297
- enhanced book availability (EBA) 11, 39, 43–44, 87, 236, 280, 286
 - definition of 283
 - prepare 286

- enhanced driver maintenance (EDM) 11, 280, 292
- enterprise service bus (ESB) 6
- error correction code (ECC) 38
- ESA/390 Architecture mode 97–98
- ESA/390 TPF mode 98
- ESCON 9
 - channel 25–26, 62, 93, 168, 214, 240, 250
 - port sparing 108
- ESCON feature 127
- ETR cards 28
- Europay Mastercard VISA (EMV) 2000 174
- EXCP 228
- EXCPVR 228
- expanded storage 89, 96
- Extended Address Volumes (EAV) 165
- extended addressability 227
- extended distance FICON 134
- extended format data set 227
- extended translation facility 72
- external time reference (ETR) 17, 28, 154
 - dual, two cards 63
 - receiver 30

F

- fanout rebalance 246
- feature code 251
 - CBU 267, 269
 - FC 1995 251
 - FC 28xx 292
 - flexible memory option 284
 - STI Rebalance 246
 - zAAP 78
 - zIIP 82
- Fibre Channel Physical and Signaling Standard 133
- Fibre Channel Protocol 58, 210
- Fibre Channel Switch Fabric and Control Requirements 133
- FICON channel 12, 26, 207, 210
- FICON Express 9, 14, 26, 62, 132
 - channel 26, 168
- FICON Express LX 132
- FICON Express SX 133
- FICON Express2 9, 12, 14–15, 25–26, 62, 131
- FICON Express2 LX 132
- FICON Express2 SX 132
- FICON Express4 62, 130
 - feature 128
- FICON Express4 10km LX 130
- FICON Express4 4km LX 130
- FICON Express4 SX 131
- FICON Express8 62
- FICON extended distance 134
- FIPS 140-2 Level 4 171
- five-model structure 9
- flexible memory option 39, 45, 87, 280, 283–284, 286
- flexible service processor (FSP) 65
- frames 24
- frames A and Z 24
- full capacity CP feature 236

G

GARP VLAN Registration Protocol (GVRP) 214
Geographically Dispersed Parallel Sysplex® (GDPS) 272

H

hardware configuration definition
 See HCD

Hardware Management Console
 See HMC

hardware messages 310

hardware system area
 See HSA

HCD 93, 95, 160, 164, 169

High Performance FICON for System z 133

High Performance FICON for System z10 209

high water mark 236

HiperSockets 145

 Layer 2 support in z10 145
 multiple write facility 145, 208

HMC 66, 84, 92, 254, 270–271, 304

 browser access 309

 firewall 306

 remote access 309

Host Channel Adapter 41

HSA 8, 38, 45, 89–90, 160–161

I

I/O

 cage, I/O slot 15, 250, 274

 card 240, 246, 250, 274

 connectivity 14, 58, 63

 device 43, 158, 160

 domains 44, 109

 operation 83, 158, 208, 225

 system 108

I/O Configuration Program (IOCP) 93, 164, 166–167

IBM Power PC microprocessor 65

IBM Systems Director Active Energy Manager 300

ICB-4 link 41, 63, 273

 connectivity 149

 STI rebalance 246

ICF 46, 48, 54, 73–74, 76, 168, 256

 backup capacity 77

 CBU 51

 pool 73

 sparing 85

IEEE Floating Point 71

IFC 76

IFL 9, 46, 54, 72–76, 85, 241, 254

 assigned 48

 backup capacity 76

 sparing 85

indirect address word (IDAW) 208, 225

InfiniBand

 coupling (PSIFB) 108

 coupling links LR 151

 overview of 106

 road map 106

InfiniBand coupling links 150

input/output configuration data set (IOCDS) 169

installed record 236

instruction

 decoding 71

 fetching 71

 grouping 71

 set extensions 72

Integrated Cluster Bus-4 (ICB-4) 17

Integrated Console Controller (OSA-ICC) 141

Integrated Cryptographic Service Facility (ICSF) 175, 180, 184, 218

Integrated Facility for Linux

 See IFL

Internal Battery Feature (IBF) 24, 55, 297–298

 estimated power time 25

Internal Coupling Channels 17

Internal Coupling Facility

 See ICF

InterSystem Channel-3 17

IOCDS 169

IOCP 93

IODF 169

iQDIO

 See HiperSockets

IRD 58, 92–93

 LPAR CPU Management 92

ISC-3 17

 link 26, 63, 240, 250

ISC-3 coupling links 149

ISO 16609 CBC Mode 175

ITRR 20

J

Java virtual machine (JVM) 77–78, 80

K

key exchange 175

L

L1 cache 60, 72

L2 cache 34, 61, 64

land grid array (LGA) 30

Large System Performance Reference

 See LSPR

Level 1 (L1) cache 34

Level 2 (L2) cache 64

LICCC 236, 240

 I/O 240

 memory 240

 processors 240

Licensed Internal Code (LIC) 8, 38, 72–73, 75, 239–240, 246, 280

 See also LICCC

link aggregation 142

Linux 9, 74–75, 95, 174, 187, 192

 mode 98

 storage 98

- Linux on System z 75, 189–191, 207
- Linux-only mode 95
- loading of initial ATM keys 175
- local area network (LAN) 184
 - Open Systems Adapter family 15
- logical partition 90, 180, 193, 197, 248–249
 - central storage 90
 - CFCC 94
 - dynamic add and delete 95
 - I/O operations 83
 - identifier 162
 - logical processors 92
 - mode 94
 - processor upgrade 85
 - real storage 90
 - reserved processors 277
 - reserved storage 277
- logical processor 84, 91
 - add 74, 84
- LPAR
 - management 310
 - mode 74, 85, 89–90, 94–95
 - single storage pool 89
- LSPR 8
 - default mixed workload 19
 - Web site 20

M

- machine type 9
- master key entry 180
- MBA 41, 280
 - fanout card 13, 41
- MCI 236, 247, 274
 - 701 to 754 49
 - Capacity on Demand 236
 - ICF 76
 - IFL 75
 - list of identifiers 49
 - model upgrade 240
 - sub-capacity settings 49, 51
 - updated 247
 - ZAAP 79
- MCM 9, 13, 30–31, 33, 236
- Media Manager 228
- memory
 - allocation 87
 - card 35, 38, 59, 240, 246, 249
 - physical 35, 38, 87, 284, 292
 - size 35, 54, 87
 - upgrades 87
- Memory Bus Adapter
 - See MBA
- message authentication code (MAC) 172, 179
- MIDAW facility 12, 193, 197, 208, 224, 226, 228
- MIF image ID (MIF ID) 95, 161
- miscellaneous equipment specification (MES) 88, 237, 246
- Mod_Raised_to Power (MRP) 174
- mode conditioner patch (MCP) 138
- model capacity identifier

- See MCI
- Model Permanent Capacity Identifier (MPCI) 236
- model S08 34, 54, 248
- model S54 31, 34
- Model Temporary Capacity Identifier (MTCI) 236
- model upgrade 10, 240
- modes of operation 93
- modular refrigeration unit 29
- Modulus Exponent (ME) 181
- motor drive assembly (MDA) 29
- motor scroll assembly 29
- MPCI 236
- MSS 163–164, 193, 197, 207
 - definition of 11
- MSU
 - value 20, 46, 50, 77, 81
- MTCI 236
- multiple CSS 166, 168
- multiple image facility (MIF) 163, 169
- multiple subchannel sets
 - See MSS

N

- N_Port ID virtualization (NPIV) 211
- native FICON 210
- Network Analysis Tool 316
- Network Traffic Analyzer 144
- nondisruptive upgrades 273, 276
- NPIV 211

O

- On/Off Capacity on Demand
 - See On/Off CoD
- On/Off CoD 53, 73–74, 76, 236, 238, 242, 245, 255, 274
 - contractual terms 260
 - granular capacity 53
 - Repair capability 262
 - rules 54
 - Upgrade Capability 262
- Open Systems Adapter (OSA) 9, 141
- operating system 9, 84, 189–190, 239, 241
 - messages 311
 - requirements 189
 - support 190
 - support Web page 232
- optionally assignable SAPs 84
- OSA 9, 141
- OSA Layer 3 Virtual MAC 144
- OSA-Express 9, 15
- OSA-Express2 62, 139
 - 10 Gb Ethernet LR 26, 63
 - 10 GbE LR 139
 - 1000BASE-T 138
 - 1000BASE-T Ethernet 26, 63, 140
 - GbE LX 140
 - GbE SX 140
 - OSN 214
- OSA-Express3 136, 216
 - 10 Gb Ethernet LR 137

- 10 Gb Ethernet SR 137
- Ethernet Data Router 137
- Gb Ethernet LX 138
- Gb Ethernet SX 138
- oscillator 64, 280

P

- parallel access volume (PAV) 11, 164
 - HyperPAV 164
- Parallel Sysplex 146
 - cluster 273
 - configuration 104
 - environment 58
 - license charge 231
 - Web site 224
- partial memory restart 87
- PCHID 93, 166–168, 179, 187, 246, 273
 - assignment 166
 - ICB-4 277
- PCI Cryptographic Accelerator (PCICA) 179
- PCICC 179
- PCI-X
 - cryptographic adapter 26, 63, 173, 176, 178–179, 181–182, 274
 - cryptographic coprocessor 58, 178, 182
- performance improvement, CP 11
- performance indicator (PI) 265
- permanent capacity 236–237
- permanent entitlement record (PER) 237
- permanent upgrade 237
 - retrieve and apply data 255
- personal identification number (PIN) 179–180
- physical channel ID
 - See also* PCHID 162
- physical memory 35, 38, 87, 284, 292
- PKA Encrypt 181
- PKA Key Import (CSNDPKI) 175
- PKA Key Token Change (CSNDKTC) 175
- Plan 251
- plan-ahead
 - concurrent conditioning 251, 277
 - control for plan-ahead 251
 - memory 280
- planned event 237
- pool
 - ICF 76
 - IFL 75
 - width 73
- power consumption 296
- power estimation tool 300
- power-on reset (POR) 246, 273
 - expanded storage 89
 - hardware system area (HSA) 160
- PR/SM 87, 90
- Preventive Service Planning (PSP) 189, 192
- processing unit (PU) 9–11, 31, 34, 46, 55, 64, 72–74, 85, 87, 241, 252, 269
 - characterizable PU 292
 - characterization 85, 95
 - concurrent conversion 10, 241

- conversion 48, 241
- cycle time 33
- dual-core 31
- feature code 9
- Maximum number 9
- pool 73, 205
- single-core 31
- spare 73, 86
- sparing 73
- type 93, 95, 241, 291

- program directed re-IPL 216
- pseudorandom number generator (PRNG) 68, 172, 174, 187, 220
- PSIFB 108
- public key
 - algorithm 174, 179, 184
 - decrypt 174, 187
 - encrypt 174, 187
- pulse per second (PPS) 18, 28
- purchased capacity 237

Q

- QDIO Diagnostic Synchronization 144
- QDIO interface isolation 142
- QDIO mode 143
- QDIO optimized latency mode 142
- Queued Direct Input/Output (QDIO) 213, 215

R

- reconfigurable storage unit (RSU) 99
- Red Hat RHEL 190, 205, 207
- Redbooks Web site 321
 - Contact us xvii
- redundant I/O interconnect (RII) 11, 13, 280, 286
- refrigeration 28
- reliability, availability, serviceability (RAS) 18, 20
- Remote Direct Memory Access (RDMA) 106
- Remote HMC 308
- Remote Support Facility (RSF) 253–254, 308
- replacement capacity 235–237
- request node identification data (RNID) 194, 198, 210
- reserved
 - processor 277
 - PU 268, 272
 - storage 99
- Resource Access Control Facility (RACF) 180
- Resource Link 237, 252
 - machine profile 254
- Rivest-Shamir-Adelman (RSA) 174, 179, 181, 187
- RMF distributed data server 263

S

- SAP 9, 34, 83
 - additional 46, 273
 - concurrent book replacement 291
 - definition 83
 - number of 46, 54, 240, 254, 256, 260, 269, 273
- SC chip 30, 34, 61

- SCSI disk 211
- SD chip 34, 64
- secondary approval 237
- Secure Sockets Layer (SSL) 16, 58, 68, 171, 174, 177, 181, 186
- Select Application License Charges (SALC) 231
- self-timed interconnect (STI) 246, 273, 286
- Server Time Protocol (STP) 11, 18, 153, 314
- SET CPUID command 276
- SHA-1 172
- SHA-1 and SHA-256 172
- SHA-256 172
- single storage pool 89
- single system image 201
- single-key MAC 172
- Small Computer System Interface (SCSI) 58
- soft capping 230
- software licensing 228
- software support 20, 79, 201
- sparing of CP, ICF, IFL 85
- SSL/TLS 171
- staged CoD records 9
- staged record 237
- Standard SAP 54
- STI 11, 161
 - MP card 44, 62
 - rebalance feature 273
- storage
 - CF mode 98
 - ESA/390 mode 97–98
 - expanded 89
 - Linux-only mode 98
 - operations 96
 - reserved 99
 - TPF mode 98
 - z/Architecture mode 97
- storage control (SC) 30
- store system information (STSI) instruction 49, 247, 262, 274–275
- subcapacity 237
- subcapacity models 49, 51, 256
- subchannel 159, 163, 207
- Superscalar 67
- superscalar processor 67
- Support Element (SE) 10, 26, 66, 237, 254, 276, 286, 304
- SUSE SLES 190, 205, 207, 215–216
- symmetric multiprocessor (SMP) 31
- system activity display (SAD) 300
- system assist processor
 - See also* SAP
- system image 89, 91, 96, 201, 211, 273
- System Input/Output Configuration Analyzer 315

T

- temporary capacity 236–237
- temporary entitlement record (TER) 237
- TKE 180
 - 5.3 Licensed Internal Code 176
 - additional smart cards 177

- Smart Card Reader 177
 - workstation 16, 19, 176–177, 184
 - workstation feature 184
- TPF mode 94
- translation look-aside buffer (TLB) 71
- Transport Layer Security (TLS) 171
- triple-key DES 68, 172, 174
- Trusted Key Entry
 - See also* TKE

U

- unassigned
 - CP 46, 48
 - IFL 46, 48
- unplanned upgrades 243
- upgrade 47
 - disruptive 277
 - for I/O 250
 - for memory 249
 - for processors 247
 - nondisruptive 276
 - permanent upgrade 251
- user ID 252
- user logical partition ID (UPID) 162
- User-Defined Extension (UDX) 175, 181, 186

V

- version code 276
- VLAN ID 214
- VPD 237

W

- Web 2.0 5
- WebSphere 6
- WebSphere MQ 231
- wild branch 70
- Workload License Charge (WLC) 92, 230–231, 251
 - CIU 249
 - Flat WLC (FWLC) 230
 - sub-capacity 230
 - Variable WLC (VWLC) 230
- Workload Manager (WLM) 266

Z

- Z frame 24, 26
- z/Architecture 6, 9, 72, 94–95, 97–99, 177, 190–191
- z/OS 91–92, 169, 191, 219–220
 - Capacity Provisioning Manager 9
- z/TPF 20
- z/VM 75, 95, 250
 - virtual machine management 316
- z/VSE 215
- z9 BC 184
- z900 memory design 87
- z990 63
- zAAP 46, 54, 72–74, 77
 - and LPAR definitions 78
- CBU 51

pool 73, 78
zIIP 46, 54, 72–74
pool 73–74, 82



IBM System z10 Enterprise Class Technical Guide



IBM System z10 Enterprise Class Technical Guide



Redbooks®

Describes the Enterprise Class server and related features

Addresses increasing complexity, rising costs, and energy constraints

Discusses infrastructure for the data center of the future

This IBM Redbooks publication discusses the IBM System z10 Enterprise Class, which offers a continuation of IBM scalable mainframe servers. Based on z/Architecture, the IBM System z10 Enterprise Class (z10 EC) server provides major extensions by:

- ▶ Increasing the maximum number of processor units
- ▶ Providing fixed HSA where all devices, channel subsystems, and multiple subchannel sets are defined, thus better supporting dynamic changes
- ▶ Providing a base for major server consolidation by further removing memory, processor, and channel constraints
- ▶ Increasing the flexibility of capacity upgrades

This book provides an overview of the z10 EC and its functions, features, and associated software support. Greater detail is offered in areas relevant to technical planning.

This book is intended for systems engineers, consultants, planners, and anyone wanting to understand the IBM System z10 Enterprise Class functions and plan for their usage. It is not intended as an introduction to mainframes. Readers are expected to be generally familiar with existing IBM System z technology and terminology.

INTERNATIONAL TECHNICAL SUPPORT ORGANIZATION

BUILDING TECHNICAL INFORMATION BASED ON PRACTICAL EXPERIENCE

IBM Redbooks are developed by the IBM International Technical Support Organization. Experts from IBM, Customers and Partners from around the world create timely technical information based on realistic scenarios. Specific recommendations are provided to help you implement IT solutions more effectively in your environment.

For more information:
ibm.com/redbooks

SG24-7516-02

ISBN 0738433772