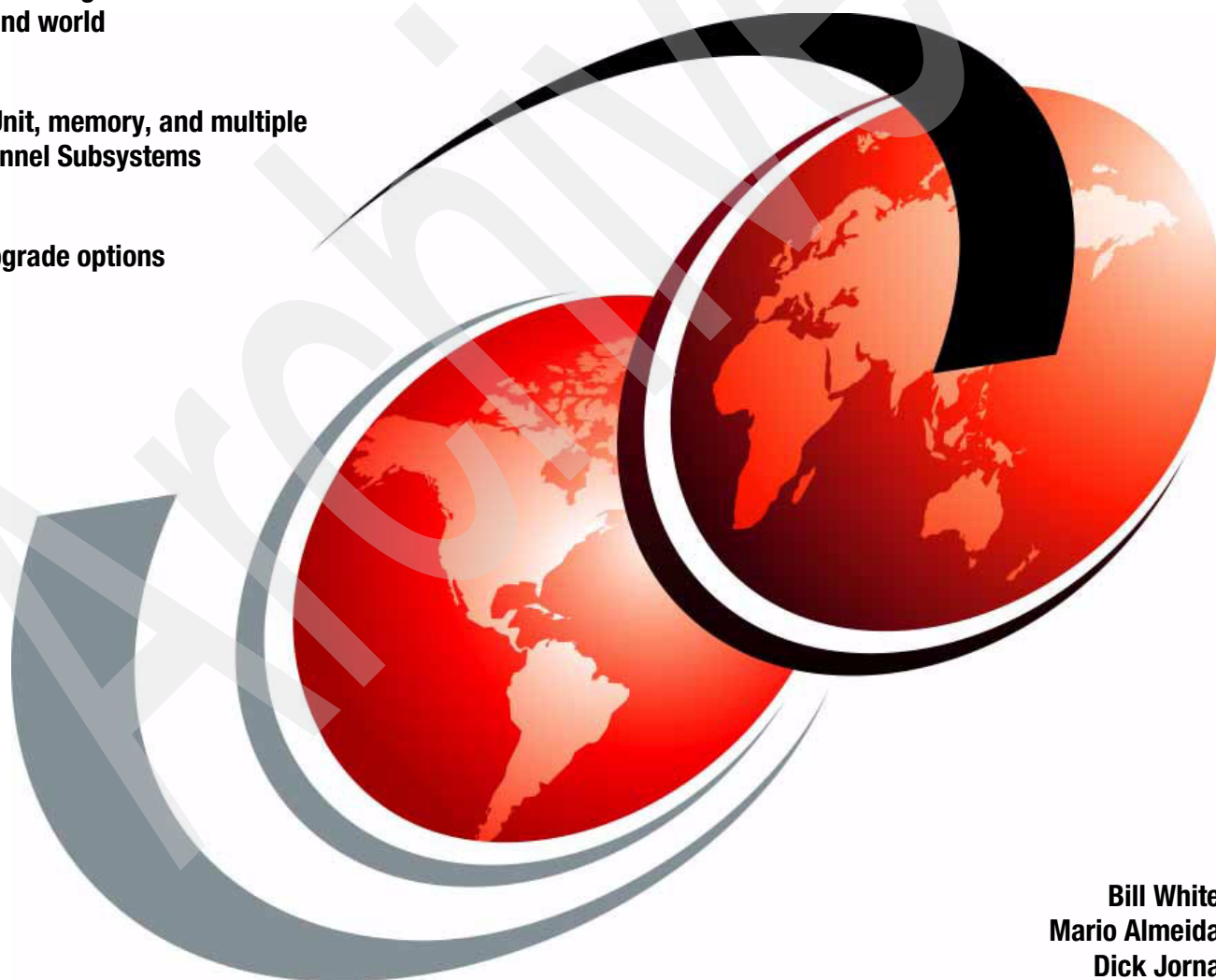


IBM **@**server zSeries 990 Technical Guide

Structure and design - A scalable server for
an on demand world

Processor Unit, memory, and multiple
Logical Channel Subsystems

Capacity upgrade options



Bill White
Mario Almeida
Dick Jorna

Redbooks



International Technical Support Organization

IBM @server zSeries 990 Technical Guide

May 2004

Archived

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

Archived

Second Edition (May 2004)

This edition applies to the IBM @serverzSeries 990 server at hardware Driver Level 55.

© Copyright International Business Machines Corporation 2003, 2004. 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	vii
Trademarks	viii
 Preface	ix
The team that wrote this redbook.	ix
Become a published author	x
Comments welcome.	x
 Chapter 1. zSeries 990 overview	1
1.1 Introduction	3
1.2 z990 models	4
1.3 System functions and features	5
1.3.1 Processor	5
1.3.2 Memory	6
1.3.3 Self-Timed Interconnect (STI)	6
1.3.4 Channel Subsystem (CSS)	6
1.3.5 Physical Channel IDs (PCHIDs) and CHPID Mapping Tool	7
1.3.6 Spanned channels	7
1.3.7 I/O connectivity	8
1.3.8 Cryptographic	12
1.3.9 Parallel Sysplex support	13
1.3.10 Intelligent Resource Director (IRD)	15
1.3.11 Hardware consoles	15
1.3.12 Concurrent upgrades	15
1.3.13 Performance	17
1.3.14 Reliability, Availability, and Serviceability (RAS)	17
1.3.15 Software	18
1.3.16 Software support	19
1.3.17 Summary	21
 Chapter 2. System structure and design	23
2.1 System structure	24
2.1.1 Book concept	24
2.1.2 Models	26
2.1.3 Memory	27
2.1.4 Ring topology	29
2.1.5 Connectivity	31
2.1.6 Frames and cages	33
2.1.7 The MCM	35
2.1.8 The PU, SC, and SD chips	36
2.1.9 Summary	37
2.2 System design	38
2.2.1 Design highlights	38
2.2.2 Book design	39
2.2.3 Processor Unit design	41
2.2.4 Processor Unit functions	46
2.2.5 Memory design	53
2.2.6 Modes of operation	56
2.2.7 Model configurations	62

2.2.8 Storage operations	67
2.2.9 Reserved storage	70
2.2.10 LPAR storage granularity	70
2.2.11 LPAR Dynamic Storage Reconfiguration (DSR)	71
2.2.12 I/O subsystem	71
2.2.13 Channel Subsystem	72
Chapter 3. I/O system structure	73
3.1 Overview	74
3.2 I/O cages	75
3.2.1 Self-Timed Interconnect (STI)	77
3.2.2 STIs and I/O cage connections	77
3.2.3 Balancing I/O connections	79
3.3 I/O and cryptographic feature cards	84
3.3.1 I/O feature cards	84
3.3.2 Cryptographic feature cards	85
3.3.3 Physical Channel IDs (PCHIDs)	86
3.4 Connectivity	89
3.4.1 I/O and cryptographic features support and configuration rules	89
3.4.2 ESCON channel	93
3.4.3 FICON channel	97
3.4.4 OSA-Express adapter	99
3.4.5 Coupling Facility links	104
3.4.6 External Time Reference (ETR) feature	107
3.4.7 Cryptographic features	108
Chapter 4. Channel Subsystem	109
4.1 Multiple Logical Channel Subsystem (LCSS)	110
4.1.1 Logical Channel Subsystem structure	110
4.1.2 Physical Channel ID (PCHID)	113
4.1.3 Channel spanning	114
4.2 LCSS configuration management	115
4.2.1 z990 configuration management	116
4.3 LCSS-related numbers	117
Chapter 5. Cryptography	119
5.1 Cryptographic function support	120
5.1.1 Cryptographic Synchronous functions	120
5.1.2 Cryptographic Asynchronous functions	120
5.2 z990 Cryptographic processors	122
5.2.1 CP Assist for Cryptographic Function (CPACF)	122
5.2.2 PCIX Cryptographic Coprocessor (PCIXCC)	123
5.2.3 PCI Cryptographic Accelerator (PCICA) feature	124
5.3 Cryptographic hardware features	125
5.3.1 PCIX Cryptographic Coprocessor feature	125
5.3.2 The PCICA feature	125
5.3.3 Configuration rules	126
5.3.4 z990 cryptographic feature codes	127
5.3.5 TKE workstation feature	128
5.4 Cryptographic features comparison	128
5.5 Software requirements	129
Chapter 6. Software support	133
6.1 Operating system support	134

6.2 z/OS software support	134
6.2.1 Compatibility Support for z/OS	134
6.2.2 Exploitation Support for z/OS	137
6.2.3 HCD support	140
6.2.4 Automation changes	140
6.2.5 SMF support	140
6.2.6 RMF support	141
6.2.7 ICKDSF requirements	141
6.2.8 ICSF support	141
6.2.9 Additional Exploitation Support considerations	142
6.3 z/VM software support	145
6.4 z/VSE and VSE/ESA software support	146
6.5 TPF software support	146
6.6 Linux software support	147
6.7 Summary of software requirements	147
6.7.1 Summary of z/OS and OS/390 software requirements	147
6.7.2 Summary of z/VM, z/VSE, VSE/ESA, TPF, and Linux software requirements	148
6.8 Workload License Charges	150
6.9 Concurrent upgrades considerations	151
Chapter 7. Sysplex functions	155
7.1 Parallel Sysplex	156
7.1.1 Parallel Sysplex described	156
7.1.2 Parallel Sysplex summary	159
7.2 Sysplex and Coupling Facility considerations	159
7.2.1 Sysplex configurations and Sysplex Timer considerations	159
7.2.2 Coupling Facility and CFCC considerations	162
7.2.3 CFCC enhanced patch apply	162
7.2.4 Coupling Facility link connectivity	164
7.2.5 Coupling Facility Resource Manager (CFRM) policy considerations	166
7.2.6 ICF processor assignments	166
7.2.7 Dynamic CF dispatching and dynamic ICF expansion	168
7.3 System-managed CF structure duplexing	169
7.3.1 Benefits	169
7.3.2 CF structure duplexing	170
7.3.3 Configuration planning	170
7.4 Geographically Dispersed Parallel Sysplex	172
7.4.1 GDPS/PPRC	172
7.4.2 GDPS/XRC	176
7.4.3 GDPS and Capacity Backup (CBU)	177
7.5 Intelligent Resource Director	178
7.5.1 LPAR CPU management	179
7.5.2 Dynamic Channel Path Management	180
7.5.3 Channel Subsystem Priority Queueing	182
7.5.4 WLM and Channel Subsystem priority	183
7.5.5 Special considerations and restrictions	184
7.5.6 References	185
Chapter 8. Capacity upgrades	187
8.1 Concurrent upgrades	188
8.2 Capacity Upgrade on Demand (CUoD)	190
8.3 Customer Initiated Upgrade (CIU)	196
8.4 On/Off Capacity on Demand (On/Off CoD)	202

8.5 Capacity BackUp (CBU)	206
8.6 Nondisruptive upgrades	210
8.6.1 Upgrade scenarios	211
8.6.2 Planning for nondisruptive upgrades	217
8.7 Capacity planning considerations	219
8.7.1 Balanced system design	219
8.7.2 Superscalar processors	222
8.7.3 Integrated hardware and system assists	222
8.8 Capacity measurements	223
8.8.1 Large Systems Performance Reference (LSPR)	224
Chapter 9. Environmental requirements	231
9.1 Introduction	232
9.1.1 Power and cooling requirements	232
9.1.2 Power consumption	232
9.1.3 Internal Battery Feature	232
9.1.4 Emergency power-off	233
9.1.5 Cooling requirements	233
9.2 Weights	233
9.3 Dimensions	234
Appendix A. Hardware Management Console (HMC)	235
z990 Hardware Management Console	238
Token ring only wiring scenario	238
Ethernet only - one-path wiring scenario	240
Ethernet only - two-path wiring scenario	242
Token ring and Ethernet wiring scenario	243
Remote operations	244
Support Element	245
z990 HMC enhancements	246
Appendix B. Fiber optic cabling services	249
Fiber optic cabling services from IBM	250
Summary	251
Glossary	253
Related publications	261
IBM Redbooks	261
Other publications	261
Online resources	262
How to get IBM Redbooks	263
Index	265

Notices

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

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

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not 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 illustrates 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. 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 IBM's application programming interfaces.

Trademarks

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

CICS®	HyperSwap™	S/360™
DB2®	IBM®	S/370™
developerWorks®	IMS™	S/390®
DRDA®	Multiprise®	Sysplex Timer®
e-business on demand™	MQSeries®	System/360™
Enterprise Storage Server®	MVS™	System/370™
Enterprise Systems	NetView®	ThinkPad®
Architecture/390®	OS/2®	Tivoli®
ECKD™	OS/390®	TotalStorage®
ES/9000®	Parallel Sysplex®	VisualAge®
ESCON®	Processor Resource/Systems	VM/ESA®
@server®	Manager™	VSE/ESA™
eServer™	PR/SM™	VTAM®
FlashCopy®	pSeries®	Wave®
FICON®	Redbooks™	WebSphere®
Geographically Dispersed Parallel	Redbooks (logo)  ™	z/Architecture™
Sysplex™	Resource Link™	z/OS®
GDDM®	RACF®	z/VM®
GDPS®	RETAIN®	zSeries®
HiperSockets™	RMF™	

The following terms are trademarks of other companies:

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

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

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

Preface

The IBM @server® zSeries® 990 scalable server provides major extensions to the existing zSeries architecture and capabilities. The concept of Logical Channel Subsystems is added, and the maximum number of Processor Units and logical partitions is increased. These extensions provide the base for much larger zSeries servers.

This IBM® Redbook is intended for IBM systems engineers, consultants, and customers who need to understand the zSeries 990 features, functions, availability, and services.

This publication is part of a series. For a complete understanding of the z990 scalable server capabilities, also refer to our companion Redbooks™:

- ▶ *IBM @server zSeries 990 Technical Introduction*, SG24-6863
- ▶ *IBM @server zSeries Connectivity Handbook*, SG24-5444

Note that the information in this book includes features and functions announced on April 7, 2004, and that certain functionality is not available until hardware Driver Level 55 is installed on the z990 server.

The team that wrote this redbook

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

Bill White is a Project Leader and Senior Networking Specialist at the International Technical Support Organization, Poughkeepsie Center.

Mario Almeida is a Certified Consulting IT Specialist in Brazil. He has 29 years of experience in IBM Large Systems. His areas of expertise include zSeries and S/390® servers technical support, large systems design, data center and backup site design and configuration, and FICON® channels.

Dick Jorna is a Certified Senior Consulting IT Specialist in the Netherlands. He has 35 years of experience in IBM Large Systems. During this time, he has worked in various roles within IBM, and currently provides pre-sales technical support for the IBM @server zSeries product portfolio. In addition, he is a zSeries product manager, and is responsible for all zSeries activities in his country.

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

Franck Injey
International Technical Support Organization, Poughkeepsie Center

Mike Scoba
zSeries Hardware Product Planning, IBM Poughkeepsie

First Edition authors

Franck Injey, Mario Almeida, Parwez Hamid, Brian Hatfield, Dick Jorna

Become a published author

Join us for a two- to six-week residency program! Help write an IBM Redbook dealing with specific products or solutions, while getting hands-on experience with leading-edge technologies. You'll team with IBM technical professionals, Business Partners and/or customers.

Your efforts will help increase product acceptance and customer satisfaction. As a bonus, you'll 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 Redbooks to be as helpful as possible. Send us your comments about this or other Redbooks in one of the following ways:

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

ibm.com/redbooks

- Send your comments in an Internet note to:

redbook@us.ibm.com

- Mail your comments to:

IBM Corporation, International Technical Support Organization
Dept. HYJ Mail Station P099
2455 South Road
Poughkeepsie, NY 12601-5400

zSeries 990 overview

This chapter gives a high-level view of the IBM *@server* zSeries 990. All the topics mentioned in this chapter are discussed in greater detail later in this book.

The legacy of zSeries goes back more than 40 years. Actually, on April 7th, 2004, it was 40 years ago that IBM introduced its S/360™. Since then, mainframes have followed a path of innovation with a focus on evolution to help protect investments made through the years.

The proliferation of servers in the last decade or so has increased complexity in IT management and operations and decreased the overall efficiency of resource use. On top of this came the need for business solutions to support business pressures on demand, which requires an on demand operating environment capable of being supportive, adaptive, and responsive to on demand business objectives and offering infrastructure simplification with the values of the mainframe technology as set forward with the zSeries 990.

The zSeries 990 is designed for any enterprise that needs the qualities of service required to sustain and expand their on demand computing environment. Customers requiring the ability to meet mission-critical requirements that include unexpected demands, high numbers of transactions, a heterogeneous application environment, and the ability to consolidate a number of servers will find the z990 an attractive solution since it leverages the current application portfolio with Linux and z/OS®, and simplifies the operation and management of business applications by consolidating both Linux and mainframe applications onto the same platform.

Customers with 9672s and z900s should consider using this server to consolidate servers and workloads, add capacity, or expand their Linux workloads in a more cost-effective manner. The increased capacity, bandwidth, number of channels, and logical partitions provide customers with the ability to reduce costs, while positioning them for future expansion.

The z990 is based on the proven IBM z/Architecture™, which was first introduced with the z900 family of servers. It is the continuation of the zSeries z/Architecture evolution and extends key platform characteristics with enhanced dynamic and flexible resource management, scalability, and partitioning of predictable and unpredictable workload environments. Additionally, the z990 availability, clustering, and Qualities of Service are built on the superior foundation of the current zSeries technologies.

The z990 servers can be configured in numerous ways to offer outstanding flexibility in the deployment of e-business on demand™ solutions. Each z990 server can operate independently, or as part of a Parallel Sysplex® cluster of servers. In addition to z/OS, the z990 can host tens to hundreds of Linux images running identical or different applications in parallel, based on z/VM® virtualization technology.

The z990 supports a high scalable standard of performance and integration by expanding on the balanced system approach of the IBM z/Architecture. It is designed to eliminate bottlenecks through its virtually unlimited 64-bit addressing capability, providing plenty of “headroom” for unpredictable growth in enterprise applications.

The z990 provides a significant increase in system scalability and opportunity for server consolidation by providing a “multi-book” system structure that supports configurations of one to four books. Each book consists of 12 Processor Units (PUs) and associated memory, for a maximum of 48 processors in a four-book system. All books are interconnected with a very high-speed internal communications links via the L2 cache, which allows the system to be operated and controlled by the PR/SM™ facility as a symmetrical, memory-coherent multiprocessor. The logical partitioning facility provides the ability to configure and operate as many as 30 logical partitions, which have processors, memory, and I/O resources assigned from any of the installed books.

The chart in Figure 1-1 shows growth improvements along all axes. While some of the previous generation of servers have grown more along one axis for a given family, later families focus on the other axes. Now, with the z990, the balanced design achieves improvement equally along all four axes.

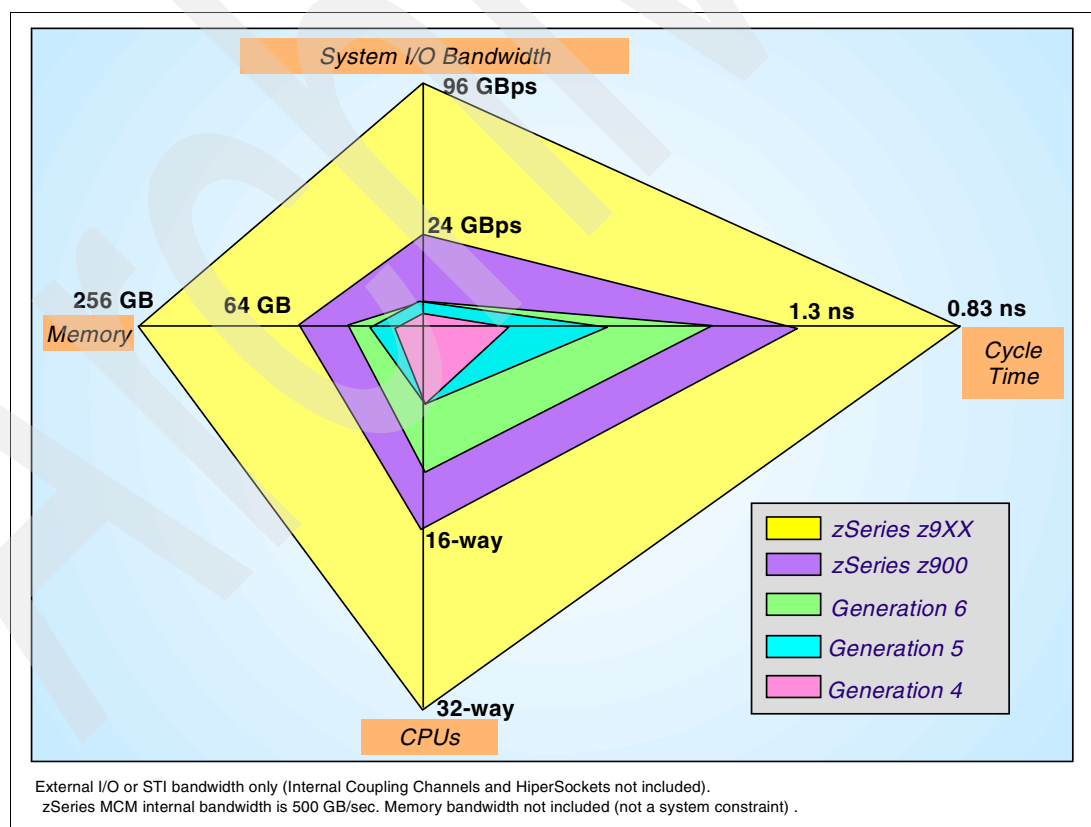


Figure 1-1 Balanced system design

1.1 Introduction

The z990 further extends and integrates key platform characteristics: dynamic and flexible partitioning, resource management in mixed and unpredictable workload environments, availability, scalability, clustering, and systems management with emerging e-business on demand application technologies (for example, WebSphere®, Java™, and Linux).

The zSeries 990 family provides a significant increase in performance over the previous zSeries servers. The z990 introduces a different design from its predecessor, the zSeries 900. One noteworthy change is to the CEC cage, which is capable of housing up to four books. This multi-book design provides enough Processor Units to improve total system capacity by nearly three times over that provided by z900.

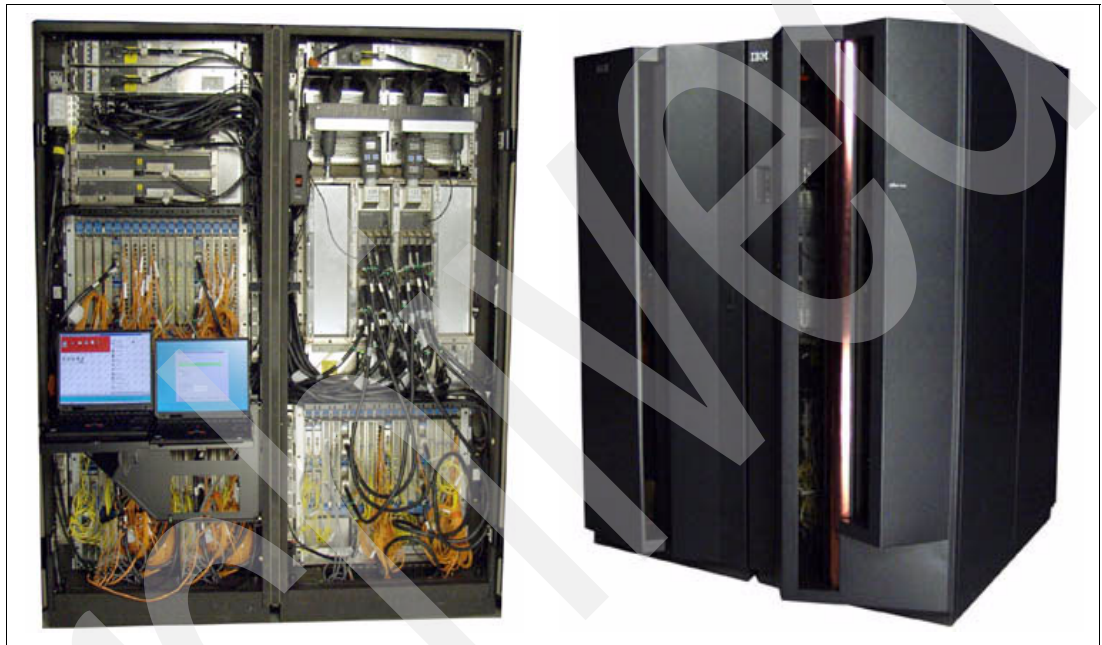


Figure 1-2 Introducing the z990 - internal and external view

The z990 introduced the superscalar microprocessor architecture. This design, and the exploitation of the CMOS 9SG-SOI technology, improves the uniprocessor performance by 54% to 61%, compared to z900 Model 2C1. However, the true capacity increase of the system is driven by the increased number of Processor Units per system: from 20 in the z900 to 48 Processor Units in the z990. The 48 Processor Units are packaged in four MCMs with 12 Processor Units each, plus up to 64 GB of memory and 12 STI links per book. All books are connected via a super-fast redundant ring structure and can be individually upgraded.

The I/O infrastructure has been redesigned to handle the large increase in system performance. The multiple Logical Channel Subsystems (LCSS) architecture on the z990 allows up to four LCSS, each with 256 channels. Channel types supported on the z990 are:

- ▶ FICON Express
- ▶ Coupling Links
- ▶ OSA-Express
- ▶ ESCON®

The following channel types, or channel cards, are *not* supported on the z990:

- ▶ Parallel channels
- ▶ 4-port ESCON cards
- ▶ OSA-2 cards
- ▶ OSA-Express ATM cards
- ▶ Pre-FICON Express cards
- ▶ PCICC cards

The logical partitioning facility, PR/SM, provides the ability to configure and operate as many as 30 logical partitions. PR/SM manages all the installed and enabled resources (processors and memory) of the installed books as a single large SMP. Each logical partition has access to physical resources (processors, memory, and I/O) in the whole system across multiple books.

1.2 z990 models

The z990 has a machine type of 2084 and has four models: A08, B16, C24, and D32. The model naming is representative of the book design of the z990, as it indicates the number of books and the number of Processor Units available in the configuration. PUs are delivered in single increments, orderable by feature code. A Processor Unit (PU) can be characterized as a Central Processor (CP), Integrated Facility for Linux (IFL), Internal Coupling Facility (ICF), zSeries Application Assist Processor (zAAP), or System Assist Processor (SAP).

The development of a multi-book system provides an opportunity for customers to increase the capacity and/or requirements of the system in three areas:

- ▶ You can add capacity by activating more CPs, IFLs, ICFs, or zAAPs on an existing book concurrently.
- ▶ You can add a new book concurrently and activate more CPs, IFLs, ICFs, or zAAPs.
- ▶ You can add a new book to provide additional memory and/or STIs to support increasing storage and/or I/O requirements. The ability to LICCC-enable more memory concurrently to existing books is dependent on enough physical memory being present. Upgrades requiring more memory than physically available are disruptive, requiring a planned outage.

General rules

All models utilize a 12 PU MCM, of which eight are available for PU characterization. The remaining four are reserved as two standard SAPs and two standard spares.

Model upgrades, from A08 to B16, from B16 to C24, or from C24 to D32, are achieved by single book adds.

The model number designates the *maximum* number of PUs available for an installation to use. Using feature codes, customers can order CPs, IFLs, ICFs, zAAPs, and optional SAPs, unassigned CPs, and/or unassigned IFLs up to the maximum number of PUs for that model. Therefore, an installation may order a model B16 with 13 CP features and three IFL features, or a model B16 with only one CP feature.

Unlike prior processor model names, which indicate the number of purchased CPs, z990 model names indicate the maximum number of Processor Units *potentially* orderable, and not the actual number that have been ordered as CPs, IFLs, ICFs, zAAPs, or additional SAPs. A software model notation is also used to indicate how many CPs are purchased and software should be charged for. See “Software models” on page 63 for more information.

Model upgrade paths

With the exception of the z900 Model 100, any z900 model may be upgraded to a z990 model. With the advancement of Linux for S/390 and Linux on zSeries, customers may choose to change the PU characterization of the server they are upgrading. In addition, customers who are consolidating may not be increasing total capacity, and/or they may wish to take advantage of the multiple Logical Channel Subsystems offered. z990-to-z990 model upgrades and feature adds may be completed concurrently.

Model downgrades

There are no model downgrades offered. Customers may purchase unassigned CPs or IFLs for future use. This avoids the placement of RPQ orders and subsequent sequential MES activity, and paying software charges for capacity that is not in use.

Concurrent Processor Unit (PU) conversions

z990 servers support concurrent conversion between different PU types, providing flexibility to meet changing business environments. Assigned CPs, unassigned CPs, assigned IFLs, unassigned IFLs, and ICFs may be converted to assigned CPs, assigned IFLs or ICFs, or to unassigned CPs or unassigned IFLs.

1.3 System functions and features

The 990 system offers the following functions and features, as shown in Figure 1-3.

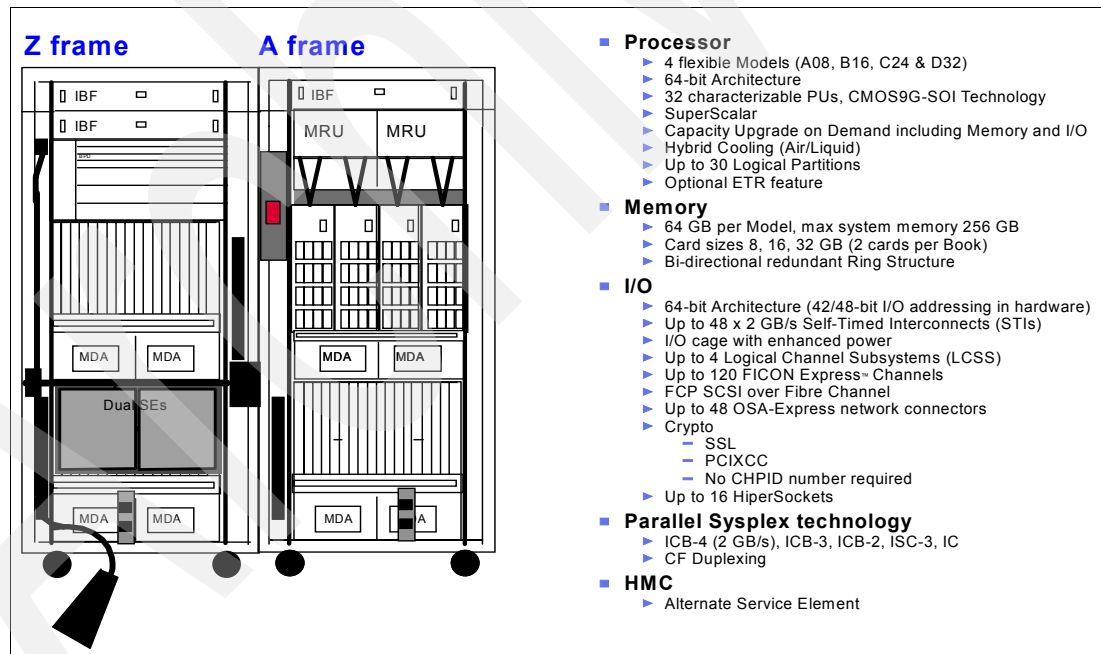


Figure 1-3 System overview

1.3.1 Processor

IBM introduced the Processor Resource/Systems Manager™ (PR/SM) feature in February 1988, supporting a maximum of four logical partitions. In June 1992, IBM introduced support for a maximum of 10 logical partitions and announced the Multiple Image Facility (MIF, also known as EMIF), which allowed sharing of ESCON channels across logical partitions, and since that time, has allowed sharing of more channels across logical partitions (such as

Coupling Links, FICON, and OSA). In June 1997, IBM announced increased support - up to 15 logical partitions on Generation 3 and Generation 4 servers.

The evolution continues and IBM is announcing support for 30 logical partitions. This support is exclusive to z990 and z890 models.

MCM technology

The z990 12-PU MCM is smaller and more capable than the z900's 20-PU MCM. It has 16 chips, compared to 35 for the z900. The total number of transistors is over 3 billion, compared with approximately 2.5 billion for the z900. With this amazing technology integration comes improvements in chip-to-substrate and substrate-to-board connections.

The z990 module uses a connection technology, Land Grid Arrays (LGA), pioneered by the pSeries® in the p690 and the i890. LGA technology enables the z990 substrate, with only 53% of the surface area of the z900 20 PU MCM substrate, to have 23% more I/Os from the logic package.

Both the z900 and z990 have 101 layers in the glass ceramic substrate. The z990's substrate is thinner, shortening the paths that signals must travel to reach their destination (another chip or exiting the MCM). Inside the low dielectric glass ceramic substrate is 0.4 km of internal wiring that interconnects the 16 chips that are mounted on the top layer of the MCM. The internal wiring provides power and signal paths into and out of the MCM.

The MCM on the z990 offers flexibility in enabling spare PUs via the Licensed Internal Code Configuration Control (LIC-CC) to be used for a number of different functions. These are:

- ▶ A Central Processor (CP)
- ▶ A System Assist Processor (SAP)
- ▶ An Internal Coupling Facility (ICF)
- ▶ An Integrated Facility for Linux (IFL)
- ▶ A zSeries Application Assist Processor (zAAP)

The number of CPs and SAPs assigned for particular general purpose models depends on the configuration. The number of spare PUs is dependent on how many CPs, SAPs, ICFs, zAAPs, and IFLs are present in a configuration.

1.3.2 Memory

The minimum system memory on any model is 16 GB. Memory size can be increased in 8 GB increments to a maximum of 64 GB per book or 256 GB for the entire CPC. Each book has two memory cards, which come in three physical size cards: 8 GB, 16 GB, and 32 GB.

The z990 continues to employ storage size selection by Licensed Internal Code introduced on the G5 processors. Memory cards installed may have more usable memory than required to fulfill the machine order. LICCC will determine how much memory is used from each card.

1.3.3 Self-Timed Interconnect (STI)

An STI is an interface to the Memory Bus Adaptor (MBA), used to gather and send data. 12 STIs per z990 physical book is supported. Each of these STIs has a bidirectional bandwidth of 2 GBps. The maximum instantaneous bandwidth per book is 24 GBps.

1.3.4 Channel Subsystem (CSS)

A new Channel Subsystem (CSS) structure was introduced with z990 to "break the barrier" of 256 channels. With the introduction of the new system structure and all of its scalability

benefits, it was essential that the Channel Subsystem also be scalable and allow “horizontal” growth. This is facilitated by multiple Logical Channel Subsystems (LCSSs) on a single zSeries server. The CSS has increased connectivity and is structured to provide the following:

- ▶ Four Logical Channel Subsystems (LCSS).
- ▶ Each LCSS may have from one to 256 channels.
- ▶ Each LCSS can be configured with 1 to 15 logical partitions.
- ▶ Each LCSS supports 63K I/O devices.

Note: There is no change to the operating system maximums. One operating system image continues to support a maximum of 256 channels, and has a maximum of 63K subchannels available to it.

The I/O subsystem continues to be viewed as a single Input/Output Configuration Data Set (IOCDs) across the entire system with multiple LCSS. Only one Hardware System Area (HSA) is used for the multiple LCSSs.

A three-digit Physical Channel Identifier (PCHID) is being introduced to accommodate the mapping of 1024 channels to four LCSSs, with 256 CHPIDs each. CHPIDs continue to exist and will be associated with PCHIDs. An updated CHPID Mapping Tool (CMT) is being introduced and the CHPID report from e-config is replaced by a PCHID report. The CHPID Mapping Tool is available from Resource Link™ as a stand-alone PC-based program.

1.3.5 Physical Channel IDs (PCHIDs) and CHPID Mapping Tool

A z990 can have up to 1024 physical channels, or PCHIDs. In order for an operating system to make use of that PCHID, it must be mapped to a CHPID within the IOCDs. Each CHPID is uniquely defined with an LCSS and mapped to an installed PCHID. A PCHID is eligible for mapping to any CHPID in any LCSS.

The z990 CHPID Mapping Tool (CMT) provides a method of customizing the CHPID assignments for a z990 system to avoid attaching critical channel paths to single points of failure. It should be used after the machine order is placed and before the system is delivered for installation. The tool can also be used to remap CHPIDs after hardware upgrades that increase the number of channels.

The tool maps the CHPIDs from an IOCP file to Physical Channel Identifiers (PCHIDs) that are assigned to the I/O ports. The PCHID assignments are fixed and cannot be changed.

A list of PCHID assignments for each hardware configuration is provided in the PCHID Report available when the z990 hardware is ordered. Unlike previous zSeries systems, there are no default CHPID assignments. CHPIDs are assigned when the IOCP file is built. When upgrading an existing zSeries configuration to z990, CHPIDs can be mapped by importing the IOCP file into the z990 CHPID Mapping Tool.

1.3.6 Spanned channels

As part of the z990 LCSS, the Channel Subsystem is extended to provide the high-speed, transparent sharing of some channel types in a manner that extends the MIF shared channel function. Internal Channel types such as HiperSocket (IQD) and Internal Coupling Channels (ICP) can be configured as “spanned” channels. External channels such as FICON channels, OSA features, and External Coupling Links can be defined as spanned channels. Spanned channels will allow the channel to be configured to multiple LCSSs, thus enabling them to be shared by any/all of the configured logical partitions, regardless of the LCSS in which the partition is configured.

Note: Spanned channels are not supported for ESCON channels, FICON conversion channels (FCV), and Coupling Receiver links (CBR, CFR).

1.3.7 I/O connectivity

Here we discuss I/O connectivity.

I/O cage

Each book provides 12 STI links (48 STI maximum with four books) for I/O and coupling connectivity, and for cryptographic feature cards. Each of these links can either be configured for ICBs, or be connected to STI distribution cards in the I/O cage(s). The data rate for the STI is 2 GBps.

Note: The z900 compatibility I/O cage is not supported on the z990.

The z990 I/O cage contains seven STI domains. Each domain has the capability of four I/O slots. A subset of previous zSeries 900 I/O and cryptographic cards is supported by the I/O cages in the z990.

Note: Parallel channels, OSA-2, OSA-Express ATM, pre-FICON Express channels, and PCICC feature cards are not supported in the z990.

The installation of an I/O cage remains a disruptive MES, so the Plan Ahead feature remains an important consideration when ordering a z990 system.

The z990 is a two-frame server. The z990 has a minimum of one CEC cage and one I/O cage in the A frame. The Z frame can accommodate an additional two I/O cages, making a total of three for the whole system. Figure 1-4 shows the layout of the frames and I/O cages.

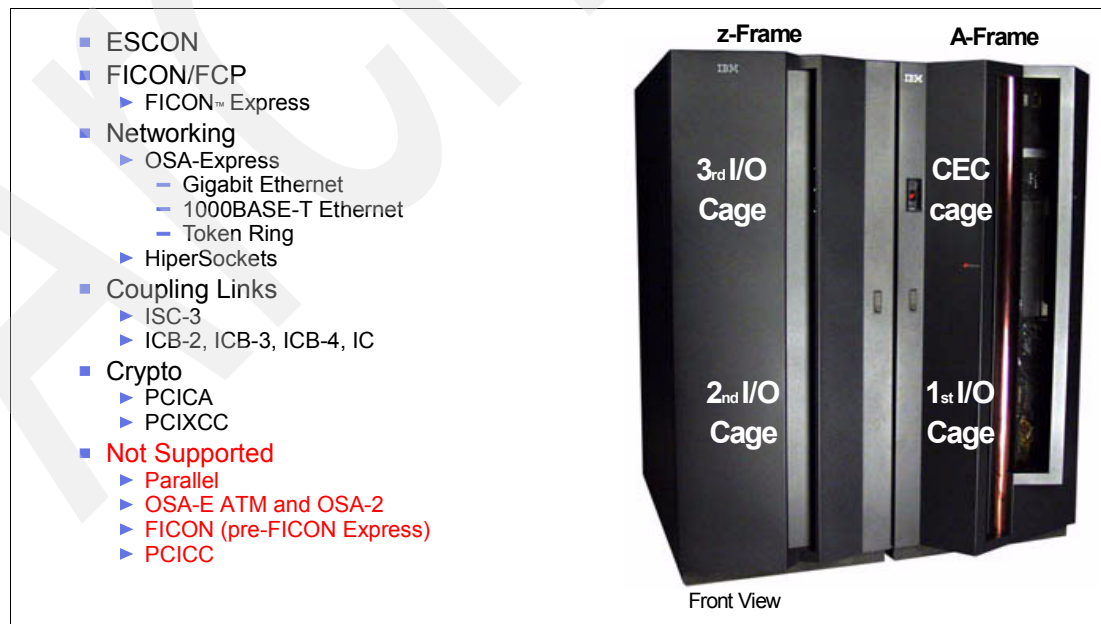


Figure 1-4 I/O cage layout and supported cards and coupling links

Up to 1024 ESCON channels

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

This is not an orderable feature. The configuration tool will select the quantity of features based upon the order quantity of ESCON FC2324 ports, distributing the ports across features for high availability. After the first pair, ESCON FC2323 are installed in increments of one.

ESCON channels are available on a port basis in increments of four. The port quantity is selected and LIC CC is shipped to activate the desired quantity of ports on the 16-port ESCON FC2323. Each port utilizes a light emitting diode (LED) as the optical transceiver, and supports use of a 62.5/125-micrometer multimode fiber optic cable terminated with a small form factor, industry standard MT-RJ connector.

Up to 120 FICON Express channels

An increased number of FICON Express features per z990 leads the way in distinguishing this server family, further setting it apart as enterprise class in terms of the number of simultaneous I/O connections available for FICON Express features. z990 supports 60 FICON Express features to be plugged, providing a total of 120 available channels. This is a 25% growth over what was available on z900. These channels are available in long wave (LX) and short wave (SX).

The FICON Express LX and SX channel cards have two ports. LX and SX ports are ordered in increments of two. The maximum number of FICON Express cards is 60, installed in the three I/O cages.

The same FICON Express channel card used for FICON channels is also used for FCP channels. FCP channels are enabled on these cards as a microcode load with an FCP mode of operation and CHPID type definition. As with FICON, FCP is available in long wavelength (LX) and short wavelength (SX) operation, though the LX and SX cannot be intermixed on a single card.

zSeries supports FCP channels, switches and FCP/SCSI devices with full fabric connectivity under Linux on zSeries. Support for FCP devices means that z990 servers will be capable of attaching to select FCP/SCSI devices, and may access these devices from Linux on zSeries. This expanded attach ability means that customers have more choices for storage solutions, or may have the ability to use existing storage devices, thus leveraging existing investments and lowering total cost of ownership for their Linux implementation.

The 2 Gb capability on the FICON Express channel cards means that 2 Gb link speeds are available for FCP channels as well.

The Fibre Channel Protocol (FCP) capability, supporting attachment to SCSI devices in Linux environments, was made generally available in conjunction with IBM TotalStorage® Enterprise Tape System 3590, IBM TotalStorage Enterprise Tape Drive 3592, and IBM TotalStorage Enterprise Tape Library 3494. For VM guest mode, z/VM Version 4 Release 3 is required to support Linux/FCP. When configured as a CHPID type FCP, FICON allows concurrent patching of Licensed Internal Code without have to configure the channel off and on.

The required Linux level for this function is SLES 8 from SUSE. This support allows a z990 system to access industry standard devices for Linux, using SCSI control block-based Input/Output (I/O) devices. These industry standard devices utilize Fixed Block rather than Extended Count Key Data (ECKD™) format. For more information, consult the IBM I/O Connectivity Web page:

<http://www.ibm.com/servers/eserver/zseries/connectivity/#fcp>

FICON CTC function

Native FICON channels support CTC on the z990, z890, z900, and z800. G5 and G6 servers can connect to a zSeries FICON CTC, as well. This FICON CTC connectivity will increase bandwidth between G5, G6, z990, z890, z900, and z800 systems.

Because the FICON CTC function is included as part of the native FICON (FC) mode of operation on zSeries, a FICON channel used for FICON CTC is not limited to intersystem connectivity but will also support multiple device definitions. For example, ESCON channels that are dedicated as CTC cannot communicate with any other device, whereas native FICON (FC) channels are not dedicated to CTC only. Native mode can support both device and CTC mode definition concurrently, allowing for greater connectivity flexibility.

FICON Cascaded Directors

Some time ago, IBM made the FICON Cascaded Director function generally available. This means that a native FICON (FC) channel or a FICON CTC can connect a server to a device or other server via two (same vendor) FICON Directors in between.

This type of cascaded support is important for disaster recovery and business continuity solutions because it can provide high availability and extended distance connectivity, and (particularly with the implementation of 2 Gb Inter Switch Links) has the potential for fiber infrastructure cost savings by reducing the number of channels for interconnecting the two sites.

The following directors and switches are supported:

- ▶ CNT (INRANGE) FC/9000 64-port and 128-port models (IBM 2042)
- ▶ McDATA Intrepid 6064 (IBM 2032)
- ▶ McDATA Intrepid 6140 (IBM 2032)
- ▶ McDATA Sphereon 4500 Fabric Switch (IBM 2031-224)
- ▶ IBM TotalStorage SAN Switches 2109-F16, S16, and S08
- ▶ IBM TotalStorage Director 2109-M12

FICON Cascaded Directors have the added value of ensuring high integrity connectivity. Transmission data checking, link incidence reporting, and error checking are integral to the FICON architecture, thus providing a true enterprise fabric.

For more information on Cascaded Directors, consult the I/O Connectivity Web page:

http://www.ibm.com/servers/eserver/zseries/connectivity/ficon_cascaded.html

OSA-Express

With the introduction of z990 and its increased processing capacity, and the availability of multiple LCSSs, the Open Systems Adapter family of local area network (LAN) adapters is also expanding by offering a maximum of 24 features per system (versus the maximum of 12 features per system on prior generations). The z990 can have 48 ports of LAN connectivity.

You can choose any combination of OSA features: the OSA-Express Gigabit Ethernet LX (FC1364), the OSA-Express Gigabit Ethernet SX (FC1365), the OSA-Express 1000BASE-T Ethernet (FC1366), or OSA-Express Token Ring (FC2367). You can also carry forward your current z900 OSA-Express features to z990, OSA-Express Gigabit Ethernet LX (FC 2364), OSA-Express Gigabit Ethernet SX (FC 2365), OSA-Express Fast Ethernet (FC 2366), and OSA-Express Token Ring (FC 2367).

Gigabit Ethernet

The OSA-Express GbE features (FC1364 and FC1365) have an LC Duplex connector type, replacing the current SC Duplex connector. This conforms to the fiber optic connectors currently in use for ISC-3 and the FICON Express features shipped after October 30, 2001.

1000BASE-T Ethernet

The z990 supports a copper Ethernet feature: 1000BASE-T Ethernet. This feature is offered on new builds and replaces the current OSA-Express Fast Ethernet (FC 2366), which can be brought forward to z990 on an upgrade from z900.

1000BASE-T Ethernet is capable of operating at 10, 100, or 1000 Mbps (1 Gbps) using the same Category-5 copper cabling infrastructure that is utilized for Fast Ethernet. The Gigabit over copper adapter allows a migration to gigabit speeds wherever there is a copper cabling infrastructure instead of a fiber optic cabling infrastructure.

OSA-Express Integrated Console Controller (OSA-ICC)

An additional function of the OSA-Express 1000BASE-T Ethernet feature is its support as an OSA-Express 100BASE-T Ethernet Integrated Console Controller. This function supports TN3270E and non-SNA DFT 3270 emulation and means that 3270 emulation for console session connections are integrated in the z990 via a port of the 1000BASE-T Ethernet feature.

Checksum Offload for Linux and z/OS when in QDIO mode

A function introduced for the Linux on zSeries and z/OS environments, called checksum offload, provides the capability of calculating the Transmission Control Protocol/User Datagram Protocol (TCP/UDP) and Internet Protocol (IP) header checksums.

Checksum verifies the correctness of files. By moving the checksum calculations to a Gigabit or 1000BASE-T Ethernet feature, host CPU cycles are reduced.

Improved performance can be realized by taking advantage of the checksum offload function of the OSA-Express Gigabit Ethernet, and OSA-Express GbE or the 1000BASE-T Ethernet (when operating at 1000 Mbps (1 Gbps)) features by offloading checksum processing to OSA-Express (in QDIO mode, CHPID type OSD)) for most IPv4 packets. This support is available with z/OS V1R5 and later as well as Linux on zSeries.

Token Ring

The OSA-Express Token Ring feature has two independent ports, each supporting attachment to either a 4 Mbps, 16 Mbps, or 100 Mbps Token Ring Local Area Network (LAN).

The OSA-Express Token Ring feature supports autosensing as well as any of the following settings: 4 Mbps half- or full-duplex, 16 Mbps half- or full-duplex, or 100 Mbps full-duplex.

Note: The demand for Token Ring on mainframe continues to decline. Migration from Token Ring to an Ethernet infrastructure is recommended as part of long term planning for Local Area Network support.

OSA-Express ATM

The OSA-Express Asynchronous Transfer Mode (ATM) features are not supported on z990. They are not offered as a new build option and are not offered on an upgrade from z900. This satisfies the Statement of General Direction in the hardware announcement dated April 30, 2002.

If ATM connectivity is still desired, a multiprotocol switch or router with the appropriate network interface (for example, 1000BASE-T Ethernet, Gigabit Ethernet) can be used to provide connectivity between the z990 and an ATM network.

OSA-2 FDDI

The OSA-2 Fiber Distributed Data Interface (FDDI) feature is not supported on z990. It is not offered as a new build option and is not offered on an upgrade from z900. This satisfies the Statement of General Direction in the hardware announcement dated October 4, 2001.

If FDDI connectivity is still desired, a multiprotocol switch or router with the appropriate network interface (for example, 1000BASE-T Ethernet, Gigabit Ethernet) can be used to provide connectivity between the z990 and a FDDI LAN.

Parallel channels and converters

Parallel channels are not supported on z990. Customers who wish to use parallel-attached devices with z990 must obtain a parallel channel converter box such as the IBM 9034, which may be available through IBM Global Services (IGS), or obtain a third-party parallel channel converter box such as the Optica 34600 FXBT. In both cases, these are connected to an ESCON channel.

For more information about Optica offerings, contact Optica directly:

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

1.3.8 Cryptographic

Here we discuss cryptographic functions and features.

CP Assist for cryptographic function

The zSeries cryptography is further advanced with the introduction of the Cryptographic Assist Architecture implemented on every z990 PU. The z990 processor provides a set of symmetric cryptographic functions, synchronously executed, which enormously enhance the performance of the encrypt/decrypt function of SSL, Virtual Private Network (VPN), and data storing applications that do not require FIPS 140-2 level 4 security. The on-processor crypto functions run at z990 processor speed.

These cryptographic functions are implemented in every PU; the affinity problem of pre-z990 systems is eliminated. The Crypto Assist Architecture includes DES and T-DES data en/decryption, MAC message authentication, and SHA-1 secure hashing. These functions are directly available to application programs (zSeries Architecture instructions). SHA-1 is always enabled, but other cryptographic functions are available only when the Crypto enablement feature (FC 3863) is installed.

PCI Cryptographic Accelerator feature (PCICA)

The Peripheral Component Interconnect Cryptographic Accelerator (PCICA) feature has two accelerator cards per feature and is an optional addition, along with the Peripheral Component Interconnect X Cryptographic Coprocessor (PCIXCC) FC0868. The PCICA is a very fast cryptographic processor designed to provide leading-edge performance of the

complex RSA cryptographic operations used with the Secure Sockets layer (SSL) protocol supporting e-business. The PCICA feature is designed specifically for maximum speed SSL acceleration.

Each zSeries PCI Cryptographic Accelerator feature (PCICA) contains two accelerator cards and can support up to 2100 SSL handshakes per second.

Note: To enable the function of the PCICA feature, the CP Assist feature (feature code 3863) must be installed.

PCI X-Cryptographic Coprocessor (PCIXCC) feature

The Peripheral Component Interconnect X Cryptographic Coprocessor (PCIXCC) feature has one coprocessor and is an optional addition, containing support to satisfy high-end server security requirements by providing full checking and fully programmable functions and User Defined Extension (UDX) support.

The PCIXCC adapter is intended for applications demanding high security. The PCIXCC feature is designed for the FIPS 140-2 Level 4 compliance rating for secure cryptographic hardware.

Note: To enable the function of the PCIXCC feature, the CP Assist feature (feature code 3863) must be installed.

1.3.9 Parallel Sysplex support

Here we discuss Parallel Sysplex support.

ISC-3

A 4-port ISC-3 card structure is provided on the z900 family of processors. It consists of a Mother Card with two Daughter Cards that have two ports each. Each Daughter Card is capable of operating at 1 Gbps in compatibility mode (HiPerLink) or 2 gigabits/sec in peer mode and up to 10 km. The mode is selected for each port via the CHPID type in the IOCDs.

InterSystem Coupling Facility-3 (ISC-3) channels provide the connectivity required for data sharing between the Coupling Facility and the CPCs directly attached to it. ISC-3 channels are point-to-point connections that require a unique channel definition at each end of the channel. ISC-3 channels operating in peer mode provide connections between z990, z890, and z900 general purpose models and z900-based Coupling Facility images. ISC-3 channels operating in compatibility mode provide connections between z990 models and ISC HiperLink channels on 9672 G5/G6 models.

ICB-2 (Integrated Cluster Bus 2)

The Integrated Cluster Bus-2 (ICB-2) link is a member of the family of Coupling Link options available on z990. Like the ISC-3 link, it is used by coupled systems to pass information back and forth over high speed links in a Parallel Sysplex environment. ICB-2 or ISC-3 links are used to connect 9672 G5/G6 to z990 servers.

An STI-2 resides in the I/O cage and provides two output ports to support the ICB-2 connections. The STI-2 card converts the 2 GBps input into two 333 MBps ICBs. The ICB-2 is defined in compatibility mode and the link speed is 333 MBps.

One feature is required for each end of the link. Ports are ordered in increments of one.

ICB-3 (Integrated Cluster Bus 3)

The Integrated Cluster Bus-3 (ICB-3) link is a member of the family of Coupling Link options available on z990. Like the ISC-3 link, it is used by coupled systems to pass information back and forth over high speed links in a Parallel Sysplex environment. ICB-3 or ISC-3 links are used to connect z900, z800, or z890 servers (2064, 2066, or 2086) to z990 servers.

An STI-3 card resides in the I/O cage and provides two output ports to support the ICB-3 connections. The STI-3 card converts the 2 GBps input into two 1 GBps ICBs. The ICB-3 is defined in peer mode and the link speed is 1 GBps.

One feature is required for each end of the link. Ports are ordered in increments of one.

ICB-4 (Integrated Cluster Bus 4)

The Integrated Cluster Bus-4 (ICB-4) link is a member of the family of Coupling Link options available on z990. ICB-4 is a “native” connection used between z990 and or z890 processors. An ICB-4 connection consists of one link that attaches directly to an STI port in the system, does not require connectivity to a card in the I/O cage, and operates at 2 GBps. The ICB-4 works in peer mode and the link speed is 2 GBps.

One feature is required for each end of the link. Ports are ordered in increments of one.

Internal Coupling (IC)

The Internal Coupling-3 (IC) channel emulates the Coupling Facility functions in LIC between images within a single system. No hardware is required; however, a minimum of two CHPID numbers must be defined in the IOCDs for each connection.

System-Managed CF Structure Duplexing

System-Managed Coupling Facility (CF) Structure Duplexing provides a general purpose, hardware-assisted, easy-to-exploit mechanism for duplexing CF structure data. This provides a robust recovery mechanism for failures (such as loss of a single structure or CF or loss of connectivity to a single CF) through rapid failover to the other structure instance of the duplex pair.

The following three structure types can be duplexed using this architecture:

- ▶ Cache structures
- ▶ List structures
- ▶ Locking structures

Support for these extensions is included in Coupling Facility Control Code (CFCC) Levels 11 12, and 13 and in z/OS V1.2, V1.3, V1.4, and V1.5 and later.

For those CF structures that support the use of System-Managed CF Structure Duplexing, customers have the ability to dynamically enable or disable, selectively by structure, the use of System-Managed CF Structure Duplexing.

Customers interested in deploying System-Managed CF Structure Duplexing in their test, development, or production Parallel Sysplex will need to read the technical paper *System-Managed CF Structure Duplexing*, GM13-0100 and analyze their Parallel Sysplex environment to understand the performance and other considerations of using this function.

System-Managed CF Structure Duplexing, GM13-0100 is available at these Web sites:

<http://www.ibm.com/server/eserver/zSeries/ps0>
<http://www.ibm.com/servers/eserver/zSeries/library/techpapers/gm130103.html>

1.3.10 Intelligent Resource Director (IRD)

Exclusive to the IBM z/Architecture is Intelligent Resource Director (IRD), a function that optimizes processor and channel resource utilization across logical partitions based on workload priorities. IRD combines the strengths of the PR/SM, Parallel Sysplex clustering, and z/OS Workload Manager.

Intelligent Resource Director uses the concept of an “LPAR cluster”, the subset of z/OS systems in a Parallel Sysplex cluster that are running as logical partitions on the same z900 server. In a Parallel Sysplex environment, Workload Manager directs work to the appropriate resources, based on business policy. With IRD, resources are directed to the priority work. Together, Parallel Sysplex technology and IRD provide flexibility and responsiveness to e-business workloads that are unrivaled in the industry.

IRD has three major functions: LPAR CPU Management, Dynamic Channel Path Management, and Channel Subsystem Priority Queuing, which are explained in the following sections.

Channel Subsystem Priority Queuing

Channel Subsystem Priority Queuing on the z900 allows priority queueing of I/O requests within the Channel Subsystem, and the specification of relative priority among logical partitions. WLM in goal mode sets priorities for a logical partition, and coordinates this activity among clustered logical partitions.

Dynamic Channel Path Management

This feature enables customers to have channel paths that dynamically and automatically move to those ESCON I/O devices that have a need for additional bandwidth due to high I/O activity. The benefits are enhanced by the use of goal mode and clustered logical partitions.

LPAR CPU Management

Workload Manager (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.

1.3.11 Hardware consoles

Here we discuss the Hardware Management Console and Support Element interface.

Hardware Management Console and Support Element interface

On z990 servers, the Hardware Management Console (HMC) provides the platform and user interface that can control and monitor the status of the system via the two redundant Support Elements installed in each z990.

The z990 server implements two fully redundant interfaces, known as the Power Service Control Network (PSCN), between the two Support Elements and the CPC. Error detection and automatic switchover between the two redundant Support Elements provides enhanced reliability and availability.

1.3.12 Concurrent upgrades

The z990 servers have concurrent upgrade capability via the Capacity Upgrade on Demand (CUoD) function. This function is also used by Customer Initiated Upgrades (CIUs) and by the Capacity BackUp (CBU) feature implementation; more details follow.

Capacity Upgrade on Demand (CUoD)

Capacity Upgrade on Demand offers server upgrades via Licensed Internal Code (LIC) enabling. CUoD can concurrently add processors (CPs, IFLs, ICFs, or zAAPs), and memory to an existing configuration when no hardware changes are required, resulting in an upgraded server. Also, I/O features can be added concurrently.

However, adequate planning is required. Proper models and memory card sizes must be used, and the Plan Ahead feature with concurrent conditioning enablement is recommended in order to ensure that all required infrastructure components are available.

Customer Initiated Upgrade (CIU)

Customer Initiated Upgrades are Web-based solutions for customers ordering and installing upgrades via IBM Resource Link and the z990 Remote Support Facility (RSF). A CIU requires a special contract and registration with IBM. The CIU uses the CUoD function to allow concurrent upgrades for processors (CPs, IFLs, ICFs, and zAAPs), and memory, resulting in an upgraded server.

As a CUoD, it also requires proper planning with respect to z990 model and memory card sizes. CIU is *not* available for I/O upgrades.

On/Off Capacity Upgrade on Demand (On/Off CoD)

On/Off Capacity on Demand (On/Off CoD) for z990 gives the customer the ability to temporarily turn on unowned PUs available within the current model. This capability allows customers to add capacity (CPs, IFLs, ICFs, and zAAPs) temporarily to meet peak workload demands.

Note: The On/Off CoD capability can coexist with Capacity BackUp (CBU) enablement. Both On/Off CoD and CBU LIC-CC can be installed on a z990 server, but the On/Off CoD activation and CBU activation are mutually exclusive.

The customer has to accept contractual terms for On/Off CoD to use this capability; activation of the additional capacity uses the CIU process. The usage is monitored and customer incurs additional charges for both the hardware and software until the added capacity is deactivated.

Capacity BackUp (CBU)

Capacity BackUp (CBU) is a *temporary* upgrade for customers who have a requirement for a robust disaster/recovery solution. It requires a special contract with IBM. CBU can concurrently add CPs to an existing configuration when another customer's servers are experiencing unplanned outages.

Note: The CBU capability can coexist with On/Off CoD enablement. Both On/Off CoD and CBU LIC-CC can be installed on a z990 server, but the On/Off CoD activation and CBU activation are mutually exclusive.

The proper number of CBU features, one for each "backup" CP, must be ordered and installed to restore the required capacity under disaster situations. The CBU activation can be tested for disaster/recovery procedures validation and testing.

Since this is a temporary upgrade, the original configuration must be restored after a test or disaster recovery situation via a concurrent CBU deactivation.

1.3.13 Performance

The IBM Large Systems Performance Reference method provides comprehensive z/Architecture processor capacity data for different configurations of central processing units across a wide variety of system control program and workload environments. For zSeries z990, z/Architecture processor capacity is defined with a 3xx notation, where xx is the number of installed Central Processor (CP) units.

The actual throughput that any user will experience will vary, depending upon 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 will achieve throughput improvements equivalent to the performance ratios shown in Figure 1-5.

For more detailed performance information, consult the Large Systems Performance Reference (LSPR), found at:

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

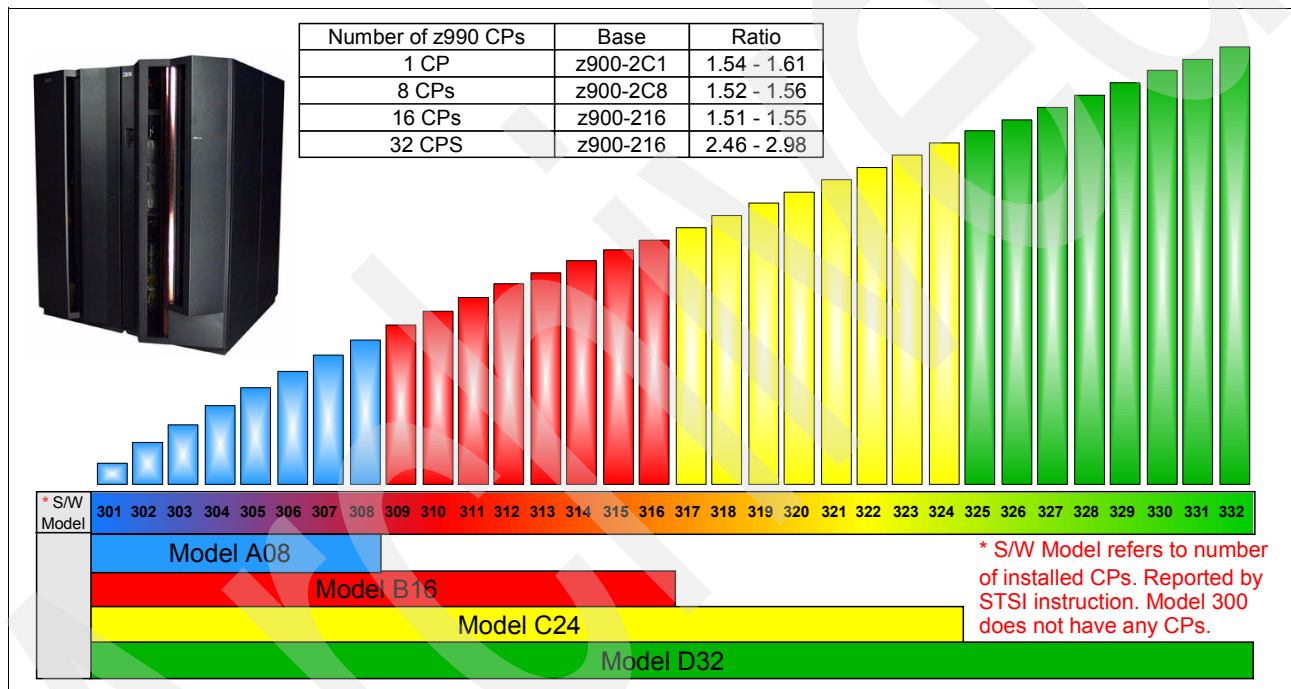


Figure 1-5 Performance comparison, SW models, and HW models

MSU values for all z990 software models are found at:

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

1.3.14 Reliability, Availability, and Serviceability (RAS)

The z990 RAS strategy is a building-block approach developed to meet the customer's stringent requirements of achieving Continuous Reliable Operation (CRO). 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 usually accomplished by using "Hi-Rel" (highest reliability) components from our technology suppliers, using screening, sorting, burn-in, run-in, and by taking advantage of technology integration. For Licensed Internal Code (LIC) and hardware design, failures are eliminated

through rigorous design rules, design walk-throughs, peer reviews, element/subsystem/system simulation, and extensive engineering and manufacturing testing.

The z990 RAS strategy is focused on a recovery design that is necessary to mask errors and make them “transparent” to customer operations. There is an extensive hardware recovery design implemented to be able to detect and correct array faults. In cases where total transparency cannot be achieved, the capability exists for the customer to restart the server with the maximum possible capacity.

1.3.15 Software

By supporting the Application Framework for e-business and Linux on zSeries, IBM provides organizations with the choices and flexibility needed to develop a robust infrastructure that provides the end-to-end qualities of service, speed of innovation, and affordability required for successful e-business.

It also enables a higher degree of integration among the three classes of workload—data transaction applications, Web applications and special function applications—that are the basis of providing a seamless business transaction over the Web (see Figure 1-6).

The result is an infrastructure that supports a more rapid move into advanced e-business, and a better chance of recognizing a lasting competitive advantage.

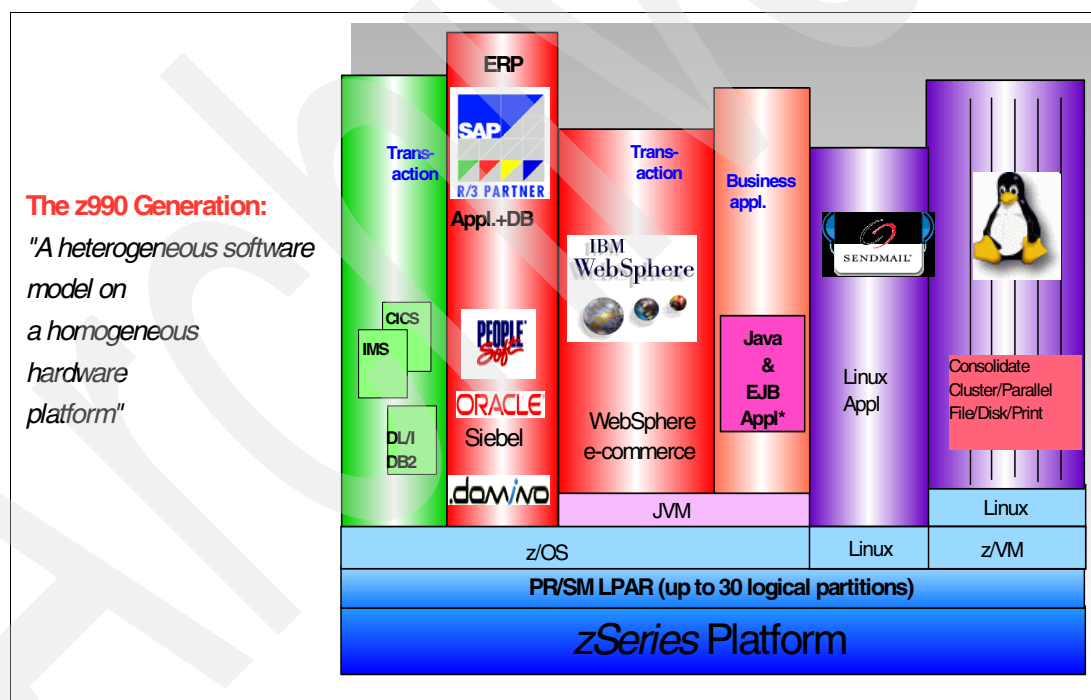


Figure 1-6 z990, the versatile server

Traditional database/transaction workloads

z/OS is well positioned as the deployment platform for e-business data transaction workloads. The traditional S/390 strengths—scalability, high availability, low total cost of ownership, and robust security—are all necessary elements for a company seeking to create the kind of flexible computing infrastructure required for enterprise-wide e-business solutions.

In addition, batch workloads, which never go away, even in an e-business environment, are a strength for z/OS, particularly with its ability to run concurrent batch jobs with online and Web

workloads; and batch jobs get resources as they are available. The strength of zSeries I/O subsystem and I/O balancing capabilities are also key reasons for this platform's excellent support of batch workloads.

UNIX® System Services

UNIX System Services include shell functions, numerous utilities, and UNIX file systems. Perhaps most significantly, what distinguished z/OS UNIX from other variants is that the UNIX services were integrated into the z/OS base, not added as middleware or shipped as a separate S/390 operating system. As a result, not only can UNIX applications be written for zSeries, but they can also take advantage of z/OS facilities and zSeries hardware to obtain true enterprise server qualities of service.

The importance of UNIX System Services is growing because most new applications and middleware build upon them.

zSeries File System (zFS)

To meet the changing needs of new workloads, a different file system is available with OS/390® R10 and beyond. This file system is complementary to the existing Hierarchical File System (HFS), but provides enhanced performance and easier management for many different file usage patterns often encountered with new workloads.

Linux on zSeries (and Linux for S/390)

Linux on zSeries offers a number of advantages compared to other platforms. First, it puts the Linux applications close to the enterprise data and applications, thus reducing the chance for bottlenecks. And since a Linux application runs in its own partition, with its own dedicated resources, it does not impact the availability or security of the rest of the system.

A second advantage is the extremely reliable zSeries hardware, which can support up to 15 logical partitions on z900/z800 and up to 30 Linux logical partitions on z990—and thousands of Linux images if Linux is run as a guest operating system under z/VM. Consolidation of multiple Linux images on a single server can greatly simplify systems management.

With the large number of applications available in the open source community, many customers will find Linux gives them a relatively low-cost way to deliver and integrate new applications quickly. For others, Internet enablement may be quicker and easier when they extend existing applications. Either way, zSeries support for the Application Framework for e-business and Linux gives that choice and flexibility.

The choice of application source depends on a number of factors, including the source, the required qualities of service, and the time allowed for development.

1.3.16 Software support

Here we discuss software support.

Compatibility and exploitation

Generally speaking, software support for the z990 comes in two steps: Compatibility support and Exploitation support. However, there are variations specific to each operating system.

Compatibility support provides no additional functionality over and above a z900 or z800. Compatibility only provides PTFs that allow the operating system to run on z990 or coexist in a sysplex with z990.

Exploitation support provides the operating systems with the ability to take advantage of greater than 15 logical partitions and multiple Logical Channel Subsystems (LCSS).

OS/390 and z/OS

OS/390 R10 and z/OS 1.2 to z/OS 1.5 and later all provide Compatibility support. In addition, z/OS 1.4, and z/OS 1.5 or later will have Exploitation support for up to 30 logical partitions and four LCSSs. OS/390 supports both 31-bit and 64-bit modes, while z/OS 1.2 to z/OS 1.4 have the bimodal migration accommodation available for fallback to 31-bit mode for a period of up to 6 months.

Exploitation support can be installed on any z/OS 1.4 system, regardless of the hardware it is running on, but obviously uses functionality as provided by that hardware. The Exploitation support is included on z/OS 1.5 and later releases.

When z/OS or OS/390 has only Compatibility support applied, it is limited to running in LCSS0 and may not have an LPAR ID greater than 'F'. In a z/OS only environment, if z/OS does not have exploitation support, then it is limited to only 15 active logical partitions and up to 256 CHPIDs. More partitions may be defined and may be physically installed, but they cannot be used.

With Exploitation support on z/OS 1.4 or later, it is possible to fully exploit the z990 capabilities. It is possible to define up to 4 LCSS with up to 256 CHPIDs in each. Up to 30 logical partitions may be defined across the 4 LCSS. z/OS 1.5 has Exploitation support delivered as part of the base function, but does not support 31-bit mode on z990.

Note: z/OS 1.1 is *not* supported on z990 or on any zSeries server participating in a sysplex that includes a z990 server.

Linux

Linux for S/390 is available in 31-bit mode and will support Exploitation mode. Linux on zSeries is available in 64-bit mode and will support Exploitation mode.

z/VM

All versions of z/VM support both 31-bit and 64-bit mode.

z/VM 4.4 and z/VM 5.1 and later are capable of exploiting up to 30 logical partitions and up to four LCSSs. VM/ESA® is not supported on z990.

z/VSE

z/VSE Version 3.1 supports 31-bit mode only and Exploitation mode.

VSE/ESA™

VSE/ESA Versions 2.6 and 2.7 will support z990 with the appropriate maintenance.

TPF

TPF V4R1 is supported in 31-bit mode with Compatibility. TPF does not provide Exploitation support.

Software pricing

The z990 product line qualifies for the same software pricing structure and software terms and conditions that are currently available for the z900 and zSeries servers. Workload License Charge (WLC) pricing is available when z/OS is running on the z990 server and all other qualifying terms and conditions for WLC are met. WLC pricing is enhanced on the variable charge products to provide greater granularity on z900 and z990 through lowering the base charge from 45 MSUs to 3 MSUs.

Parallel Sysplex License Charges (PSLC) apply for OS/390 software products and may apply with the PSLC price option if the customer elects PSLC pricing. When servers currently priced under the PSLC structure are upgraded to a z990 server, the customer may elect to continue using PSLC pricing. However, if the z990 server is an upgrade from a z900 server that has already converted to WLC pricing, the products running on the upgraded server must be charged under the WLC structure.

VM and VSE products running on z990 servers qualify for the same pricing structures currently available for other z900 servers. The Extended License Charge (ELC) applies for servers over 80 MSUs. The Graduated Monthly License Charge (GMLC) applies for servers under 80 MSUs. Customers who select WLC pricing for the z/OS environment must license VM and VSE products with the Flat Workload License Charge (FWLC). z/VM Version 4 License, and Subscription and Support (S&S) charges, are priced per processor based on terms and conditions.

Software charging for z990 will be based on the number of active CPs, except when customers qualify, and elect WLC sub-capacity pricing. The z990 software MSU values are determined when the order is placed; software charging is based on the full capacity of the selected model. Full capacity is determined by the number of active CPs within the model.

The MSU performance ratings for the z990 server are available on the Web:

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

1.3.17 Summary

On a physical resource level, z990 is a S/390 architecture server with a maximum of 48 PUs and 256 GB of main memory structured in a four-book configuration. The books are interconnected by a high speed memory coherence ring, thus building a large and very efficient SMP. The I/O adapters are housed in three I/O cages, and provide a maximum of 1024 (maximum 256 per LCSS) channel ports.

The z900 to z990 is a "Frame Roll" MES. The z990 A and Z frames are shorter and deeper than the z900's frames, actually taking up less space than the z900. The z900 I/O cards, 16-port ESCON, FICON Express, PCICA, and OSA Express will move to different I/O slots in the new z990 system I/O cages.

PR/SM handles these physical resources as one contiguous space and provides, on a logical level, up to 30 logical partitions and up to four Logical Channel Subsystems with 256 channel paths each. Thus, a large and scalable physical resource pool is generated, managed by PR/SM in a highly efficient way, taking advantage of LPAR clustering.

Dynamic CHPID management and I/O priority queuing were introduced with z900. These capabilities are extended to support Linux (for S/390 and on zSeries) partitions with the same efficiency. Since the maximum number of supported logical PUs within a partition is 24 (z/OS V1.6, and z/VM V5.1 plans to have support for up to 24 processors in a single LPAR), multiple copies of operating systems (z/OS and Linux) run "side-by-side" on the same hardware platform. z990 also continues to support the latest Ethernet technology, to provide the highest bandwidth connections to external servers.

Archived

System structure and design

This chapter introduces the IBM @server zSeries 990 system structure. Significant functions and features are described, along with their characteristics and options.

The goal is to explain how the z990 is structured, what its main components are, and how these components interconnect from a physical and logical point of view. This information is useful for planning purposes and will help you define the configuration that best fits your requirements.

The following topics are included:

- ▶ 2.1, “System structure” on page 24
- ▶ 2.2, “System design” on page 38

2.1 System structure

The z990 structure and design are the result of the continuous evolution of S/390 and zSeries since CMOS servers were introduced in 1994. The structure and design have been continuously improved, adding more capacity, performance, functionality, and connectivity.

The objective of z990 system structure and design is to offer a flexible infrastructure to accommodate a wide range of operating systems and applications, whether they be traditional or emerging e-business applications based on WebSphere, Java, and Linux, for integration and deployment in heterogeneous business solutions.

For that purpose, the z990 introduces a superscalar microprocessor architecture, improving uniprocessor performance, and providing an increase in the number of usable processors per system. In order to keep a balanced system, the I/O bandwidth and available memory sizes have been increased accordingly.

2.1.1 Book concept

The z990 Central Processor Complex (CPC) introduces a packaging concept based on *books*. A book contains processors (PUs), memory, and connectors to I/O cages and ICB-4 links. Books are located in the CEC cage in Frame A. A z990 server (CPC) has at least one book, but may have up to four books installed.

A book and its components is shown in Figure 2-1. Each book contains:

- ▶ 12 Processor Units (PUs). The PUs reside on microprocessor chips located on a Multi-Chip Module (MCM).
- ▶ 16 GB to 64 GB physical memory. There are always two memory cards, each containing 8, 16, or 32 GB.
- ▶ Three Memory Bus Adapters (MBAs), supporting up to 12 Self-Timed Interconnects (STIs) to the I/O cages and/or ICB-4 channels.

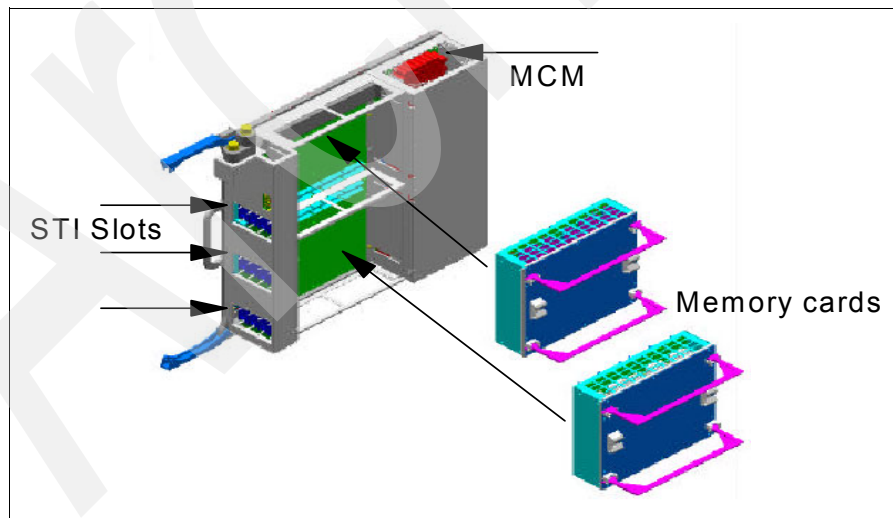


Figure 2-1 Book structure and components

Up to four books can reside in the CEC cage. Books plug into cards, which plug into slots of the CEC cage board.

Power

Each book gets its power from two Distributed Converter Assemblies (DCA) that reside on the opposite side of the CEC board. The DCAs provide the required power for the book. Each book is supported by two DCAs. The N+1 power supply design means that there is more DCA capacity than is required for the book. If one DCA fails, the power requirement for a book can still be satisfied from the remaining DCA. The DCAs can be concurrently maintained, which means that replacement of one DCA can be done without taking the book down.

Between two sets of DCAs you find the location of the oscillator cards (OSC) and the optional external time reference cards (ETR). If installed, there are two ETR ports to which an optional Sysplex Timer® can be connected. Seen from the top, the packaging of a four-book system appears as shown (schematically) in Figure 2-2.

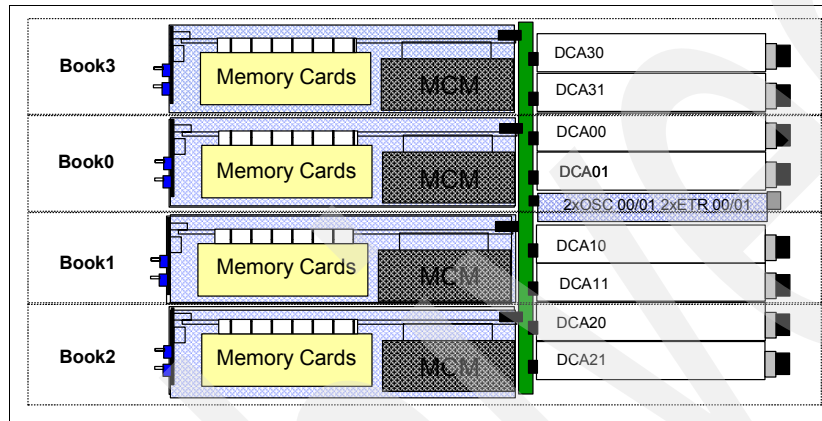


Figure 2-2 Book and power packaging (top view)

Located within each book is a card on which the Memory Bus Adapters (MBAs) are located. The card has three MBAs, each driving four STIs (see Figure 2-8 on page 32).

Figure 2-2 also illustrates the order of book installation:

- ▶ In a one-book model, only book 0 is present.
- ▶ A two-book model has books 0 and 1.
- ▶ A three-book model has books 0, 1, and 2.
- ▶ A four-book model has books 0, 1, 2, and 3.

Book installation to up to four books can be concurrent.

Cooling

The z990 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. Its components and functions are found throughout the cages, and are made up of three subsystems:

1. The Modular Refrigeration Units (MRU)
 - One or two MRUs (MRU0 and MRU1), located in the front of the A-cage above the books, provide refrigeration to the content of the books together with Motor Drive Assemblies in (MDAs) in the rear.
 - A one-book system has MRU0 installed. Upgrading to a two-book system causes MRU1 to be installed, providing all refrigeration needs 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.

2. The Motor Scroll Assembly (MSA)
3. The Motor Drive Assembly (MDA)
 - MDAs are found throughout the frames to 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, one in combination with the MSAs.

Hybrid cooling system

The z990 has a hybrid cooling system that is a breakthrough in lowering power consumption. Normal cooling is provided by one or two MRUs connected to the heat sinks of all MCMs in all books.

If one of the MRUs fails, backup MSAs are switched in to compensate for the lost refrigeration capability with additional air cooling. At the same time, the oscillator card will be set to a slower cycle time, slowing the system down by up to 8% of its maximum capacity, to allow the degraded cooling capacity to maintain the proper temperature range. Running at a slower cycle time, the MCMs will produce less heat. The slowdown process is done in steps, based on the temperature in the books.

Figure 2-3 shows the refrigeration scope of MRU0 and MRU1.

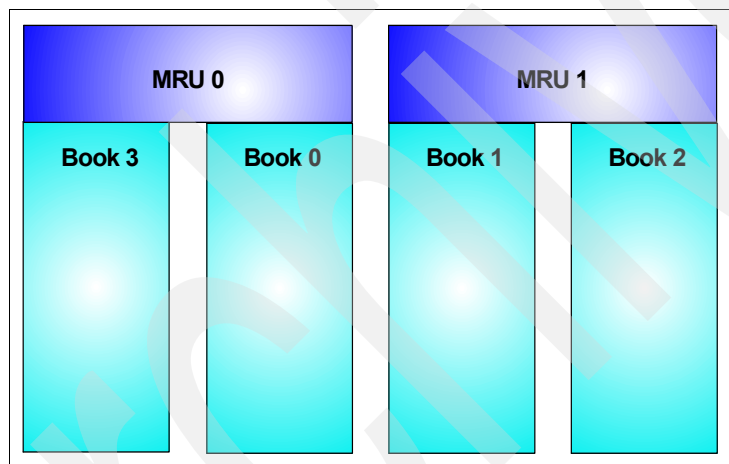


Figure 2-3 MRU scope

2.1.2 Models

The z990 has four orderable models. The model numbers are directly related to the number of books in the system and the maximum number of PUs that can be characterized by the installation. For customer use, PUs can be characterized as CPs, IFLs, ICFs, zAAPs, or if need be, additional SAPs.

- ▶ The IBM 2084 model A08 has one book (A) with 12 PUs, of which eight can be characterized by the customer. The four remaining PUs are two system assist processors (SAPs) and two spares.
- ▶ The IBM 2084 model B16 has two books (B) with 12 PUs in each book for a total of 24 PUs, of which 16 can be characterized by the customer. The four remaining PUs are four system assist processors (SAPs) and four spares, two of each in each book.
- ▶ The IBM 2084 model C24 has three books (C) with 12 PUs in each book for a total of 36 PUs, of which 24 can be characterized by the customer. The remaining PUs are six system assist processors (SAPs) and six spares, two of each in each book.

- ▶ The IBM 2084 model D32 has four books (D) with 12 PUs in each book for a total of 48 PUs, of which 32 can be characterized by the customer. The remaining PUs are eight system assist processors (SAPs) and eight spares, two of each in each book.

The last two digits of the model number reflect the maximum number of PUs that can be characterized for installation use. The PUs can be characterized as CPs, IFLs, ICFs, zAAPs or additional SAPs. The characters A, B, C, and D in the model number reflect the number of books installed.

Whether one, two, three, or four books are present, to the user, all books together appear as one Symmetric Multi Processor (SMP) with a certain number CPs, IFL, ICFs, and zAAPs a certain amount of memory, and bandwidth to drive the I/O channels and devices. The packaging is designed to scale to a 32-PU Symmetric Multi-Processor (SMP) server in four books.

2.1.3 Memory

Maximum physical memory sizes are directly related to the number of books in the system. Each book may contain a maximum of 64 GB of physical memory. The amount of memory on each of the two memory cards in a book must be the same. The memory sizes in each book do not have to be similar; different books may contain different amounts of memory. The minimum orderable amount of memory is 16 GB, system-wide.

- ▶ A one-book system (IBM 2084-A08) may contain 16 GB, 32 GB, or 64 GB of physical memory. Memory is orderable in 8 GB increments for customer use.
- ▶ A two-book system (IBM 2084-B16) may contain up to a maximum of 128 GB of physical memory. For all memory card distribution variations in newly built two-book systems, refer to Table 2-1 on page 28. Memory is orderable in 8 GB increments for customer use.
- ▶ A three-book system (IBM 2084-C24) may contain up to a maximum of 192 GB of physical memory. For some memory card distribution variation in a newly built three-book system, refer to Table 2-1 on page 28. Memory is orderable in 8 GB increments for customer use.
- ▶ A four-book system (IBM 2084-D32) may contain up to a maximum of 256 GB of physical memory. For some memory card distribution variation in a newly built four-book system, refer to Table 2-1 on page 28. Memory is orderable in 8 GB increments for customer use.

The system physical memory is the sum of all book memories. Not all books need to contain the same amount of memory, and not all installed memory is necessarily configured for use.

Memory sizes

The minimum orderable amount of usable memory for all models is 16 GB. Memory upgrades are available in 8 GB increments:

- ▶ IBM 2084 Model A08, from 16 to 64 GB
- ▶ IBM 2084 Model B16, from 16 to 128 GB
- ▶ IBM 2084 Model C24, from 16 to 192 GB
- ▶ IBM 2084 Model D32, from 16 to 256 GB

Physically, the memory cards are organized as follows:

- ▶ Each book always contains two memory cards. A memory card can come in three sizes:
 - 8 GB
 - 16 GB
 - 32 GB
- ▶ Within a given book, the card sizes must be equal, but all books do not necessarily need to have the same amount of physical memory installed.

- ▶ A book may have more memory installed than enabled. The excess amount of memory can be installed by a Licensed Internal Code code load (sometimes called “dial-a-Gig”), when required by the installation.
- ▶ On initial installation, the amount of physical memory in a given model is nearest to the smallest possible size.

Memory upgrades are satisfied from already installed unused memory capacity until exhausted. When no more unused memory is available from the installed memory cards, cards have to be upgraded to a higher capacity, or the addition of a book with additional memory is necessary.

Table 2-1 shows examples of memory configurations (not all possible combinations are shown). It shows that an IBM 2084 model A08 may have 16 GB of usable memory out of a minimum of 16 GB physically installed, and that an IBM 2084 model D32, though unlikely, may have 16 GB of usable memory out of a minimum of 64 GB physical memory.

Table 2-1 New build 2084 physical memory card distribution

Available Capacity	IBM 2084-A08 Physical Cards	IBM 2084-B16 Physical Cards	IBM 2084-C24 Physical Cards	IBM 2084-D32 Physical Cards
16 GB	2 x 8 GB	Book1: 2 x 8 GB Book2: 2 x 8 GB	Book1: 2 x 8 GB Book2: 2 x 8 GB Book3: 2 x 8 GB	Book1: 2 x 8 GB Book2: 2 x 8 GB Book3: 2 x 8 GB Book4: 2 x 8 GB
24 GB	2 x 16 GB	Book1: 2 x 8 GB Book2: 2 x 8 GB	Book1: 2 x 8 GB Book2: 2 x 8 GB Book3: 2 x 8 GB	Book1: 2 x 8 GB Book2: 2 x 8 GB Book3: 2 x 8 GB Book4: 2 x 8 GB
48 GB	2 x 32 GB	Book1: 2 x 16 GB Book2: 2 x 8 GB	Book1: 2 x 8 GB Book2: 2 x 8 GB Book3: 2 x 8 GB	Book1: 2 x 8 GB Book2: 2 x 8 GB Book3: 2 x 8 GB Book4: 2 x 8 GB
64 GB	2 x 32 GB	Book1: 2 x 16 GB Book2: 2 x 16 GB	Book1: 2 x 16 GB Book2: 2 x 8 GB Book3: 2 x 8 GB	Book1: 2 x 8 GB Book2: 2 x 8 GB Book3: 2 x 8 GB Book4: 2 x 8 GB
80 GB	n/a	Book1: 2 x 32 GB Book2: 2 x 8 GB	Book1: 2 x 16 GB Book2: 2 x 16 GB Book3: 2 x 8 GB	Book1: 2 x 16 GB Book2: 2 x 16 GB Book3: 2 x 8 GB Book4: 2 x 8 GB
96 GB	n/a	Book1: 2 x 32 GB Book2: 2 x 16 GB	Book1: 2 x 16 GB Book2: 2 x 16 GB Book3: 2 x 16 GB	Book1: 2 x 16 GB Book2: 2 x 16 GB Book3: 2 x 16 GB Book4: 2 x 8 GB
128 GB	n/a	Book1: 2 x 64 GB Book2: 2 x 64 GB	Book1: 2 x 64 GB Book2: 2 x 16 GB Book3: 2 x 16 GB	Book1: 2 x 16 GB Book2: 2 x 16 GB Book3: 2 x 16 GB Book4: 2 x 16 GB

Note: The amount of memory available for use in the server is the sum of all enabled memory on all memory cards 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 if so allocated. Despite the book structure, the z990 is a Symmetric Multi-Processor (SMP).

Each memory card has two memory ports, and each port can access 128 bits (16 bytes). Actually, the data access path is 140 bits wide, allowing for sophisticated sparing and error-checking functions. Each port is capable of four fetch and four store operations concurrently.

Memory upgrade is concurrent when it requires no change of the physical memory card. A memory card change is disruptive.

Memory sparing

The z990 does not contain spare memory DIMMs. Instead, it has redundant memory distributed throughout its operational memory and these are used to bypass failing memory. Replacing memory cards requires the removal of a book and this is disruptive. The extensive use of redundant elements in the operational memory greatly minimizes the possibility of a failure that requires memory card replacement.

Memory upgrades

For a model upgrade that results in the addition of a book, a minimum of additional memory is added to the system. Remember, the minimum physical memory size in a book is 16 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, it becomes available to an active logical partition if this partition has reserved storage defined (see 2.2.9, “Reserved storage” on page 70 for more detailed information). Or, it may be used by an already defined logical partition that is activated after the memory addition.

Book replacement and memory

When a book must be replaced, for example, due to an unlikely MCM failure, the memory in the failing book is removed as well. Until the failing book is replaced, Power-on Reset of the CPC with the remaining books is not supported.

2.1.4 Ring topology

Concentric loops or rings are constructed such that in a four-book system, each book only is connected to two others, which means that only data transfers or data transactions to the third book require passing through one of the other books.

Book-to-book communications are organized as shown in Figure 2-4 on page 30. Book 0 communicates with book 2 and book 3; communication to book 1 must go through another book (either book 2 or book 3).

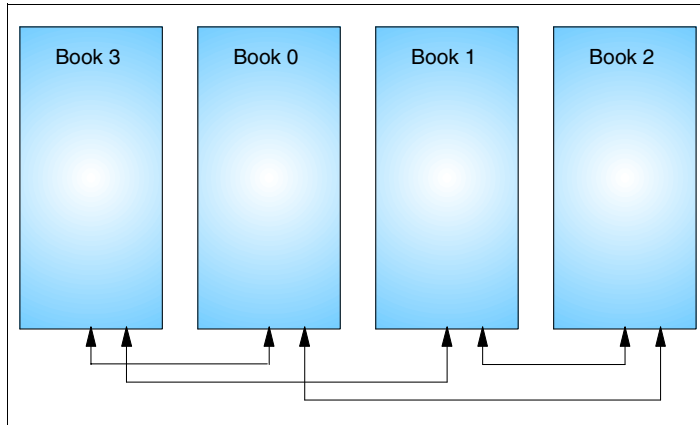


Figure 2-4 Concentric ring structure

A memory-coherent director optimizes ring traffic and filters out cache traffic by not looking on the ring for cache hits in other books if it is certain that the resources for a given logical partition exists in the same book.

The Level 2 (L2) cache is implemented on four cache (SD) chips. Each SD chip holds 8 MB, resulting in a 32 MB L2 cache per book. The L2 cache is shared by all PUs in the book and has a store-in buffer design. The connection to processor memory is done through four high-speed memory buses.

There is a ring structure within which the books maintain interbook communication at the L2 cache level. Additional books extend the function of the ring structure for interbook communication. The ring topology is shown in Figure 2-5 and Figure 2-6 on page 31, and in Figure 2-7 on page 31.

A book jumper completes the ring in order to be able to insert additional books into the ring non-disruptively.

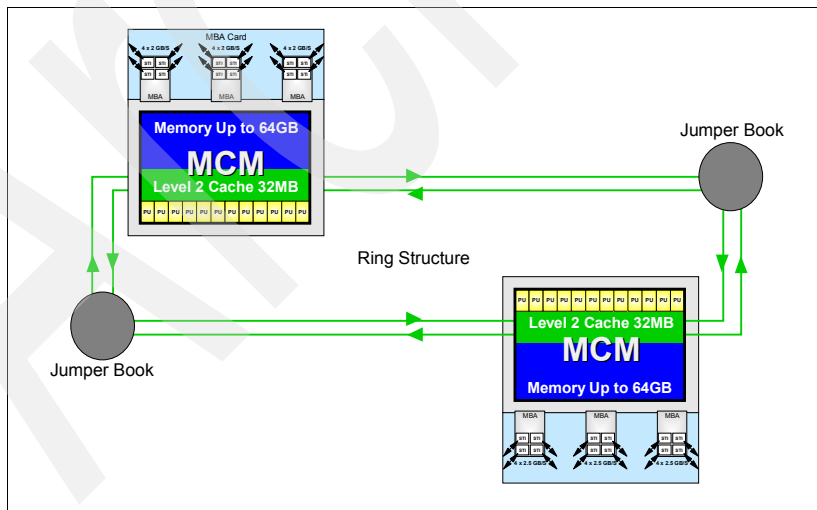


Figure 2-5 Two-book system ring structure

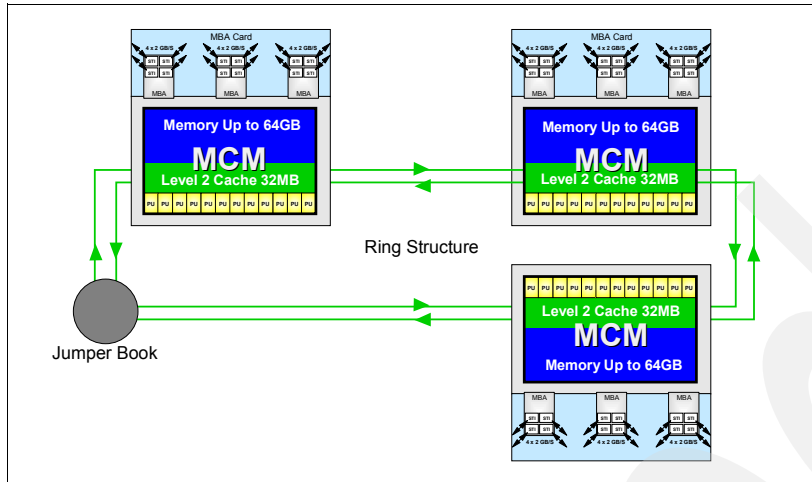


Figure 2-6 Three-book system ring structure

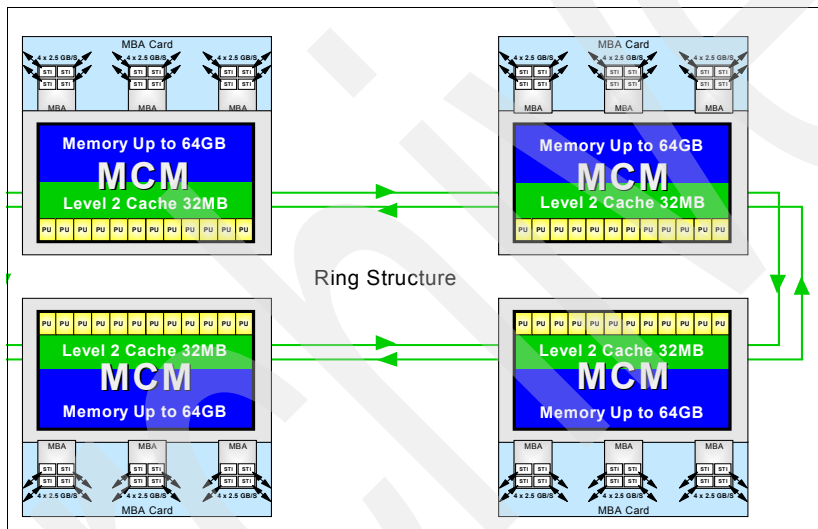


Figure 2-7 Four-book system ring structure

2.1.5 Connectivity

STI connections to I/O cages and ICB-4 links are driven from the Memory Bus Adapters (MBAs) that are located on a separate card in the book. Figure 2-8 on page 32 shows the location of the STI connectors and the MBA card.

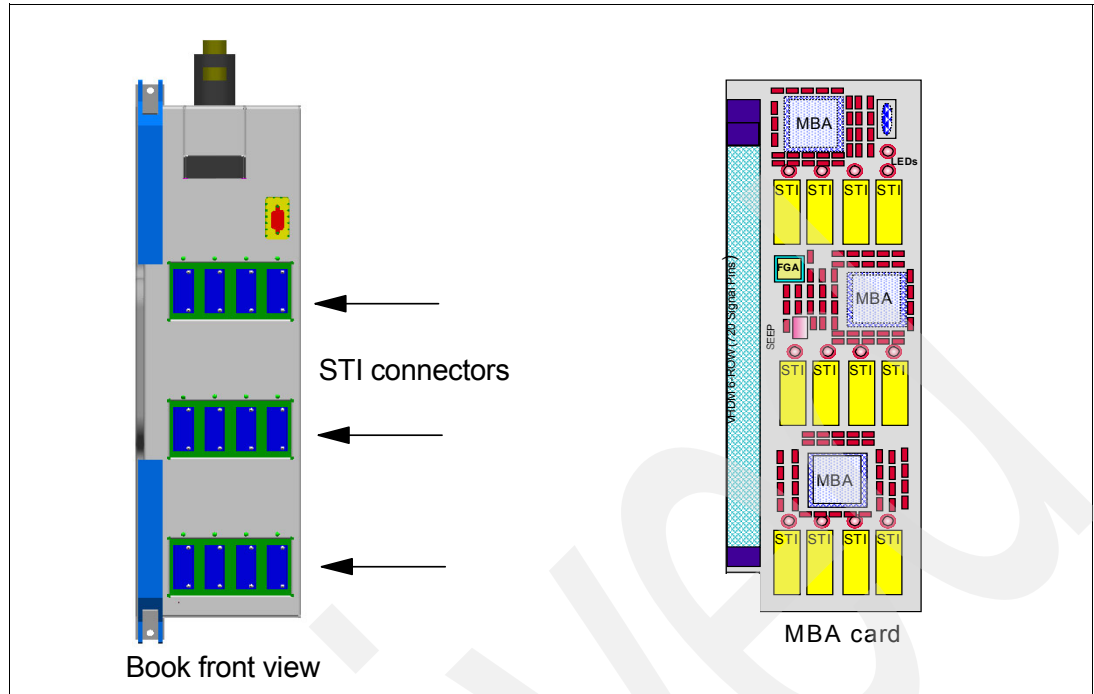


Figure 2-8 STI connectors and MBA card

Each book has three MBAs, each driving four STIs, resulting in 12 STIs per book.

All 12 STIs per book have a data rate of 2.0 GBps, resulting in a sustained bandwidth of 24 GBps per book. Consequently, the total instantaneous internal bandwidth of a four-book system is 4 x 24 GBps or 96 GBps. Depending on the channel types installed, a maximum of 512 channels per CPC is currently supported.

Four STIs are related to one MBA. When configuring for availability, you should balance channels, links, and OSAs across books, MBAs, and STIs. For z990, enhancements have been made such that, in the unlikely event of a catastrophic failure of an MBA chip, the failure is contained to that chip, while the other two MBAs on that book continue to operate. In a system configured for maximum availability, alternate paths will maintain access to critical I/O devices.

In the configuration reports, books are numbered 0, 1, 2, and 3, MBAs are numbered from 0 to 2, and the STIs are identified as jacks numbered from J.00 to J.11.

Book upgrade

As a result of a concurrent book upgrade, additional MBA and STIs connectors become available. Since now more external connections to the I/O are potentially available, there may be circumstances in which it might be beneficial to rebalance the total I/O configuration across all available MBA/STIs.

Not all book upgrades will necessitate a rebalance of the I/O configuration, since the number of STIs of the original configuration may well be able to service all existing I/O in an efficient and balanced way.

However, if the result of the upgrade is an unbalanced I/O configuration, you should consider rebalancing the configuration by using the additional MBA/STIs. An I/O distribution over books, MBAs, STIs, I/O cages, and I/O cards is often desirable for both performance and

availability purposes. Reports from the CHPID Mapping tool can help you validate your I/O configuration.

Book upgrades with substantial additions of I/O cards may require the additional STIs to be used. In that case, it is a good practice to consider rebalancing the STI configuration (FC 2400). For more information about I/O balancing, see 3.2.3, “Balancing I/O connections” on page 79. Be aware that rebalancing of the STI configuration as a result of the addition of one or more books is disruptive.

Book replacement and connectivity

When a book must be replaced, for example, due to an unlikely MCM failure, the MBA/STIs connectors from the failing book are unavailable. Until the failing book is replaced, Power-On Reset of the CPC with the remaining books is not supported.

2.1.6 Frames and cages

The z990 frames are enclosures built to Electronic Industry Association (EIA) standards. The z990 server always has two frames that are composed of two 40 EIA frames. The A and Z frames 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 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, the DCAs for the books, and the cooling components are located in the CEC cage in the top half of the A-frame of the z990. In Figure 2-9, the arrows point to the front view of the CEC cage in which four books are shown as being installed.

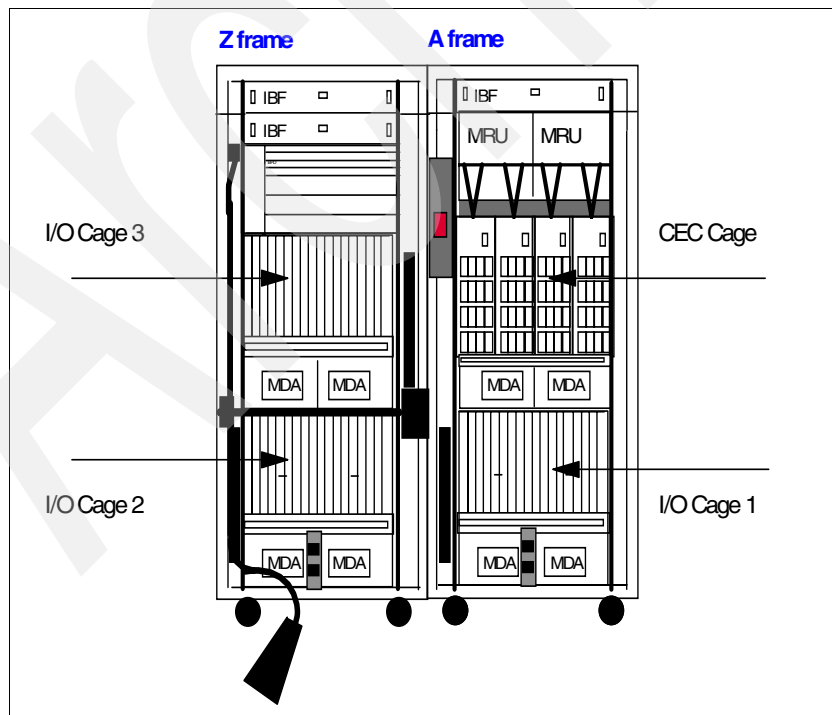


Figure 2-9 CEC cage and I/O cage locations

A frame

As shown in Figure 2-9 on page 33, the main components in the A-frame are:

1. Two Internal Battery Features (IBFs)

The optional Internal Battery Feature provides 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 both of the AC feeders 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 20 minutes of full power, depending on I/O configuration.

2. One or two Modular Refrigeration Units (MRUs) that are air-cooled by their own internal cooling fans.
3. The CEC cage, containing up to four books, each with two insulated refrigeration lines to an MRU.
4. I/O Cage 1, which can house all supported types of channel cards. An I/O cage accommodates up to 420 ESCON channels or up to 40 channels in the absence of any other card.
5. Air-moving devices (AMD) providing N+1 cooling for the MBAs, memory, and DCAs.

Z frame

As shown in Figure 2-9 on page 33, the main components in the Z-frame are:

1. Two Internal Battery Features (IBFs) (see IBF in A frame for more information).
2. The Bulk Power Assemblies (BPAs).
3. I/O cage 2 (bottom) and I/O cage 3 (top). Note that both I/O cages are the same as the one in the A frame, and can house all supported types of channel card.

The Z frame holds no cages, 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).
4. The Service Element (SE) tray, which is located in front of I/O cage 2, contains the two SEs.

I/O cages

There are 12 STI buses per book to transfer data, with a bi-directional bandwidth of 2.0 GBps each. An STI is driven off an MBA. There are three MBAs per book, each driving four STIs, providing an aggregated bandwidth of 24 GBps per book.

The STIs connect to I/O cages that may contain a variety of channel, coupling link, OSA-Express, and Cryptographic feature cards:

- ▶ ESCON channels (16 port cards).
- ▶ FICON channels (FICON or FCP modes, two port cards).
- ▶ ISC-3 links (up to four coupling links, two links per Daughter Card (ISC-D). Two Daughter Cards plug into one Mother Card (ISC-M).
- ▶ Integrated Cluster Bus (ICB) channels, both ICB-2 (333 MBps) and ICB-3 (1 GBps). Both ICB-2 and ICB-3 (compatibility mode ICBs) require an STI extender card in the I/O cage.
- ▶ OSA-Express channels:
 - OSA-E Gb Ethernet
 - Fast Ethernet
 - 1000BASE-T Ethernet

- High Speed Token Ring
- ▶ PCI Cryptographic Accelerator (PCICA, two processors per feature).
- ▶ PCIX Cryptographic Coprocessor (PCIXCC, one processor per feature).
- ▶ The STI-2 card provides two output ports to support the ICB-2 links. The STI-3 card converts the output into two 333 MBps links
- ▶ The STI-3 card provides two output ports to support the ICB-3 links. The STI-3 card converts the output into two 1GBps links.

The ICB-4 channels are unique to the z990. They do not require a slot in the I/O cage and attach directly to the STI of the communicating CPC with a bandwidth of 2.0 GBps.

2.1.7 The MCM

The z990 MultiChip Module (MCM) contains 16 chips: eight are processor chips (12 PUs), four are System Data cache (SD) chips, one is the Storage Control (SC) chip, two chips carry the Memory Subsystem Control function (MSC), and there is one chip for the clock (CLK-ETR) function.

The 93 x 93 mm glass ceramic substrate on which these 16 chips are mounted has 101 layers of glass ceramic with 400 meters of internal wiring. The total number of transistors on all chips amounts to more than 3.2 billion.

The MCM plugs into a card that is part of the book packaging, as shown in Figure 2-10. The book itself is plugged into the CEC board to provide interconnectivity between the books, so that a multibook system appears as a Symmetric Multi Processor (SMP). The MCM is connected to its environment by 5184 Land Grid Arrays (LGA) connectors. Figure 2-11 on page 36 shows the chip locations.

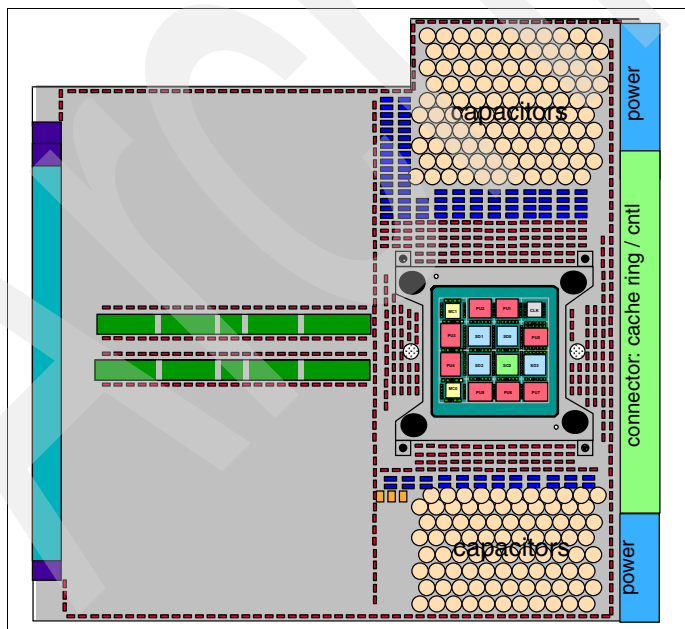


Figure 2-10 MCM card

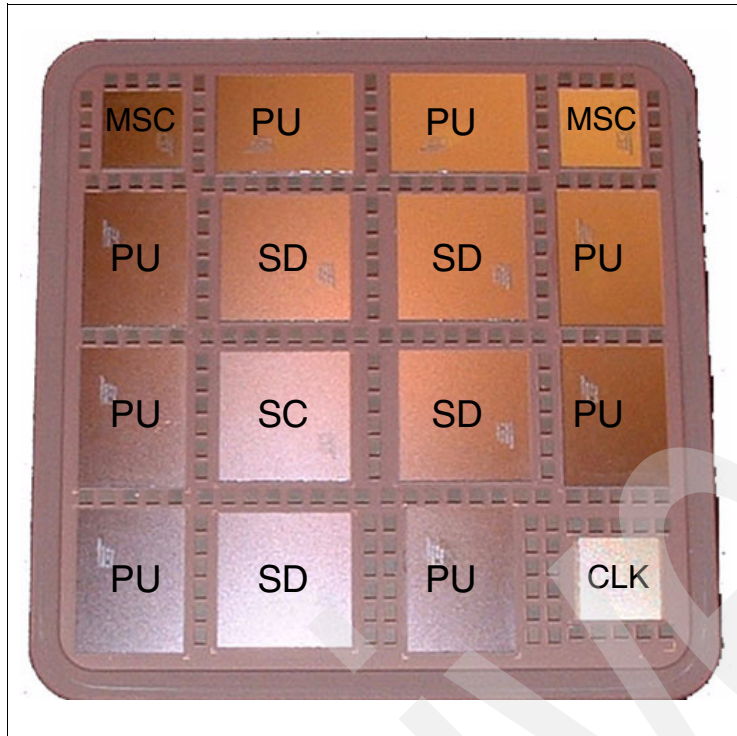


Figure 2-11 MCM chip layout

2.1.8 The PU, SC, and SD chips

All chips use CMOS 9SG technology, except for the clock chip (CMOS 8S). CMOS 9SG is state-of-the-art microprocessor technology based on eight-layer Copper Interconnections and Silicon-On Insulator technologies. The chip's lithography line width is 0.125 micron.

The eight PU chips come in two versions. The Processor Units (PUs) on the MCM in each book are implemented with a mix of single-core and dual-core PU chips. Four single-core and four dual-core chips are used, resulting in 12 PUs per MCM.

Eight PUs may be characterized for customer use, one per PU chip. The two standard SAPs and two standard spares are initially allocated to the dual-core processor chips. Each core on the chip runs at a cycle time of 0.83 nanoseconds. Each dual-core PU chip measures 14.1 x 18.9 mm and has 122 million transistors.

Each PU has a 512 KB on-chip Level 1 cache (L1) that is split into a 256 KB L1 cache for instructions and a 256 KB L1 cache for data, providing large bandwidth.

SC chip

The L1 caches communicate with the L2 caches (SD chips) by two bi-directional 16-byte data buses. There is a 2:1 bus/clock ratio between the L2 cache and the PU, controlled by the Storage Controller (SC chip), that also acts as an L2 cache cross-point switch for L2-to-L2 ring traffic, L2-to-MS traffic, and L2-to-MBA traffic. The L1-to-L2 interface is shared by two P PU cores on a dual core PU chip.

SD chip

The level 2 cache (L2) is implemented on the four System Data (SD) cache chips, each with a capacity of 8 MB, providing a cache size of 32 MB. These chips measure 17.5 x 17.5 mm and carry 521 million transistors, making them the world's densest chips.

The dual-core PU chips share the path to the SC chip (L2 control) and the clock chip (CLK).

2.1.9 Summary

Table 2-2 summarizes all aspects of the z990 system structure.

Table 2-2 System structure summary

	IBM 2084-A08	IBM 2084-B16	IBM 2084-C24	IBM 2084-D32
Number of MCMs	1	2	3	4
Total number of PUs	12	24	36	48
Maximum number of characterized PUs	8	16	24	32
Number of CPs	0 - 8	0 - 16	0 - 24	0 - 32
Number of IFLs	0 - 8	0 - 16	0 - 24	0 - 32
Number of ICFs	0 - 8	0 - 16	0 - 16	0 - 16
Number of zAAPs	0 - 4	0 - 8	0 - 12	0 - 16
Standard SAPs	2	4	6	8
Standard spare PUs	2	4	6	8
Number of memory cards	2	4	6	8
Enabled Memory Sizes (multiples of 8 GB)	16 - 64 GB	16 - 128 GB	16 - 192 GB	16 - 256 GB
L1 Cache per PU	256/256 KB	256/256 KB	256/256 KB	256/256 KB
L2 Cache	32 MB	64 MB	96 MB	128 MB
Cycle time (ns)	0.83	0.83	0.83	0.83
Maximum number of STIs	12	24	36	48
STI bandwidth/STI	2.0 GBps	2.0 GBps	2.0 GBps	2.0 GBps
Max STI bandwidth	24 GBps	48 GBps	72 GBps	96 GBps
Maximum number of I/O cages	3	3	3	3
Number of Support Elements	2	2	2	2
External power	3 phase	3 phase	3 phase	3 phase
Internal Battery Feature	optional	optional	optional	optional

2.2 System design

The IBM z990 Symmetrical Multi Processor (SMP) design is the next step in an evolutionary trajectory stemming from the introduction of CMOS technology back in 1994. Over time, 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 z990, with its superscalar processor and flexible configuration options, is the next implementation to address this ever-changing environment.

2.2.1 Design highlights

The physical packaging is a deviation from previous packaging methods in that 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 may contain even larger capacities in the future.

The main objectives of the z990 system design, which are covered in this chapter and in the following ones, are:

- ▶ To offer a flexible infrastructure to concurrently accommodate a wide range of operating systems and applications, from the traditional S/390 and zSeries systems to the new world of Linux and e-business.
- ▶ To have state-of-the-art *integration* capability for server consolidation, offering virtualization techniques, such as:
 - Logical partitioning, which allows up to 30 logical servers
 - z/VM, which can virtualize hundreds of servers as Virtual Machines
 - HiperSockets™, which implements virtual LANs between logical and/or virtual servers within a z990 server

This allows logical and virtual server coexistence and maximizes system utilization by sharing hardware resources.

- ▶ To have *high performance* to achieve the outstanding response times required by e-business applications, based on z990 superscalar processor technology, architecture, and high bandwidth channels, which offer high data rate connectivity.
- ▶ To offer the *high capacity* and *scalability* required by the most demanding applications, both from single system and clustered systems points of view.
- ▶ To have the capability of *concurrent upgrades* for processors, memory, and I/O connectivity, avoiding server outages even in such planned situations.
- ▶ To 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.
- ▶ To have a broad connectivity offering, supporting open standards such as Gigabit Ethernet (GbE) and Fibre Channel Protocol (FCP) for Small Computer System Interface (SCSI).
- ▶ To provide the highest level of *security*, each CP has a CP Assist for Cryptographic Function (CPACF). Optional PCIX Cryptographic Coprocessors and PCI Cryptographic Accelerators for Secure Sockets Layer (SSL) transactions of e-business applications can be added.
- ▶ To be *self-managing*, adjusting itself on workload changes to achieve the best system throughput, through the Intelligent Resource Director and the Workload Manager functions.

- To 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 z990 system structure, showing a logical representation of the data flow from PUs, L2 cache, memory cards, and MBAs, which connect I/O through Self-Timed Interconnect (STI).

2.2.2 Book design

A book has 12 PUs, two memory cards and three MBAs connected by the System Controller (SC). Each memory card has a capacity of 8 GB, 16 GB, or 32 GB, resulting in up to 64 GB of memory Level 3 (L3) per book. A four-book z990 can have up to 256 GB memory. The Storage Controller, shown as SCC CNTLR in Figure 2-12 on page 40, acts as a cross-point switch between Processor Units (PUs), Memory Controllers (MSCs), and Memory Bus Adapters (MBAs).

The SD chips, shown as SCD in Figure 2-12 on page 40, also incorporate a Memory Coherent Controller (MCC) function.

Each PU chip has its own 512 KB Cache Level 1 (L1), split into 256 KB for data and 256 KB for instructions. The L1 cache is designed as a store-through cache, meaning that altered data is also stored to the next level of memory (L2 cache). The z990 models A08, B16, C24, and D32 use the CMOS 9SG PU chips running at 0.83 ns.

The MCC controls a large 32 MB L2 cache, and is responsible for the interbook communication in a ring topology connecting up to four books through two concentric loops, called the ring structure. The MCC optimizes cache traffic and will 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.

The L2 cache is the aggregate of all cache space on the SD chips, resulting in a 32 MB L2 cache per book. The SC chip (SCC) controls the access and storing of data in the four SD chips. The L2 cache is shared by all PUs within a book and shared across books through the ring topology, providing the communication between L2 caches across books in systems with more than one book installed; the L2 has a store-in buffer design.

The interface between the L2 cache and processor memory (L3) is accomplished by four high-speed memory buses and controlled by the memory controllers (MSC). Storage access is interleaved between the storage cards, which tends to equalize storage activity across the cards. Each memory card has two ports that each have a maximum bandwidth of 8 GBps. Each port contains a control and a data bus, in order to further reduce any contention by separating the address and command from the data bus.

The memory cards support store protect key caches to match the key access bandwidth with that of the memory bandwidth.

The logical book structure is shown in Figure 2-12 on page 40.

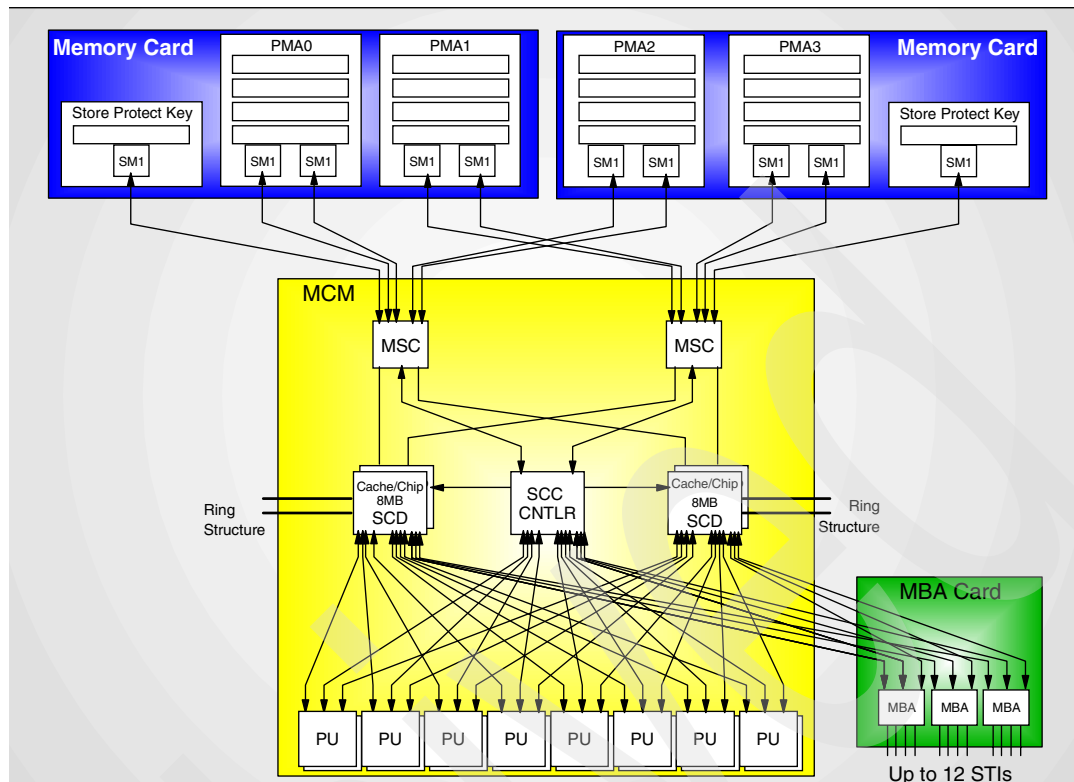


Figure 2-12 Logical book structure

There are up to 12 STI buses per book to transfer data and each STI has a bidirectional bandwidth of 2.0 GBps. A four-book z990 server may have up to 48 STIs.

An STI is an interface from the Memory Bus Adapter (MBA) to:

- ▶ An eSTI-M card in an I/O cage, to connect to:
 - ESCON channels (16 port cards)
 - FICON-Express channels (FICON or FCP modes, two port cards)
 - OSA-Express channels (all on two port cards)
 - OSA-Express Gb Ethernet
 - OSA-Express Fast Ethernet
 - OSA-Express 1000BASE-T Ethernet
 - OSA-Express High Speed Token Ring
 - ISC-3 links (up to four coupling links, two links per Daughter Card (ISC-D). Two Daughter Cards plug into one Mother Card (ISC-M).
 - PCIX Cryptographic Coprocessors (PCIXCC) in an I/O cage. Each PCIX Cryptographic Coprocessor feature contains one cryptographic coprocessor.
 - PCI Cryptographic Accelerator (PCICA) in an I/O cage. Each PCI Cryptographic Accelerator feature contains two cryptographic accelerator cards.
- ▶ An STI-2 card in an I/O cage, connecting to ICB-2 channels in 9672 G5/G6 servers.
- ▶ An STI-3 card in an I/O cage, connecting to ICB-3 channels in z800 or z900 servers.
- ▶ ICB-4, directly attached to the 2.0 GBps STI interface between z990 or z890 servers.

Data transfer between the CEC memory and attached I/O devices or CPCs is done through the Memory Bus Adapter. The physical path includes the Channel card (except for STI connected CPCs), the Self-Timed Interconnect bus, and possibly a STI extender card, the Storage Control, and the Storage Data chips.

More detailed information about I/O connectivity and channel types can be found in Chapter 2.2.12, “I/O subsystem” on page 71.

Dual External Time Reference

The optional ETR connections, although not part of the book design, are found adjacent to the books on the opposite side of the CEC board. The z990 servers implement an Enhanced ETR Attachment Facility (EEAF) designed to provide a dual External Time Reference (ETR) attachment facility. Two ETR cards are automatically shipped when Coupling Links are ordered and provide a dual path interface to the IBM Sysplex Timers, which are used for timing synchronization between systems in a Sysplex environment. This allows continued operation even if a single ETR card fails. This redundant design also allows concurrent maintenance.

2.2.3 Processor Unit design

Each PU is optimized to meet the demands of new e-business workloads, without compromising the performance characteristics of traditional workloads. The PUs in the z990 have a superscalar design.

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 their 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 may depend on data in another pipeline. A superscalar design therefore demands careful consideration of which instruction sequences can successfully operate in a multi-pipeline environment.

As an example, consider the following: If the branch prediction logic of the microprocessor makes the wrong prediction, it might be necessary to remove all instructions in the parallel pipelines also (refer to “Processor Branch History Table (BHT)” on page 44 for more details).

There are challenges in creating an efficient superscalar processor. The superscalar design of the z990 PU has made big strides in avoiding address generation interlock situations. Instructions requiring information from memory locations may suffer multi cycle delays to get the memory content. The superscalar design of the z990 PU tries to overcome these delays by continuing to execute (single cycle) instructions that do not cause delays. The technique used is called “out-of-order operand fetching”. This means that some instructions in the instruction stream are already underway, while earlier instructions in the instruction stream that cause delays due to storage references take longer. Eventually, the delayed instructions catch up with the already fetched instructions and all are executed in the designated order. The z990 PU gets much of its superscalar performance benefits from avoiding address generation interlocks.

It is not only the processor that contributes to the capability of the successful execution of instructions in parallel. Given a superscalar design, compilers and interpreters must create code that benefit optimally from the particular superscalar processor implementation. Work is

under way to update the C++ compiler and Java Virtual Machine for z/OS to better exploit the z990 microprocessor superscalar implementation. The intent is improve the performance advantage for e-business workloads such as WebSphere and Java applications.

By the time the Java Virtual Machine (JVM) and compilers are available, more improvement in the throughput of the superscalar processor is expected. In order to create instruction sequences that are least affected by interlock situations, instruction grouping rules are enforced to create instruction streams that benefit most from the superscalar processor. It is expected that e-business workloads will primarily benefit from this design since they tend to use more computational instructions.

A WebSphere Application Server workload environment that runs a mix of Java and DB2 code will greatly benefit from the superscalar processor design of the z990. Measurements already show a larger than 20% performance improvement for these types of workloads, on top of the improvements attributed to the cycle time decrease from 1.09 ns on a z900 Turbo model to 0.83 ns on a z990.

The superscalar design of the z990 microprocessor means that some instructions are processed immediately and that processing steps of other instructions may occur out of the normal sequential order, called “pipelining”. The superscalar design of the z990 offers:

- ▶ Decoding of two instructions per cycle
- ▶ Execution of three instructions per cycle (given that the oldest instruction is a branch)
- ▶ In-order execution
- ▶ Out-of-order operand fetching

Other features of the microprocessor, aimed at improving the performance of the emerging e-business application environment, are:

- ▶ Floating point performance for IEEE Binary Floating Point arithmetic is improved to assist further exploitation of Java application environments.
- ▶ A secondary cache for Dynamic Address Translation, called the Secondary level Translation Lookaside Buffer (TLB), is provided for both L2 instruction and data caches, increasing the number of buffer entries by a factor of eight.
- ▶ The CP Assist for Cryptographic Function (CPACF) accelerates the encryption and decryption of SSL transactions and VPN encrypted data transfers. The assist function uses five new instructions for symmetrical clear key cryptographic encryption and encryption operations.

Asymmetric mirroring for error detection

Each PU in the z990 servers uses mirrored instruction execution as a simple error detection mechanism. The mirroring is dependent on a dual instruction processor design with dual I-units, and E-units and floating point function. It is asymmetric because the mirrored execution is delayed from the actual operation. The benefit of the asymmetric design is that the mirrored units do not have to be closely located to the units where the actual operation takes place, thus allowing for optimization for performance (see Figure 2-13 on page 43).

Processor Branch History Table (BHT)

The Branch History Table (BHT) implementation on processors has a key performance improvement effect. The BHT was originally introduced on the IBM ES/9000® 9021 in 1990 and has been improved ever since.

The z990 server BHT offers significant branch performance benefits. The BHT allows each CP to take instruction branches based on a stored BHT, which improves processing times for calculation routines. Using a 100-iteration calculation routine as an example (Figure 2-14), the hardware preprocesses the branch incorrectly 99 times without a BHT. With a BHT, it preprocesses branch correctly 98 times.

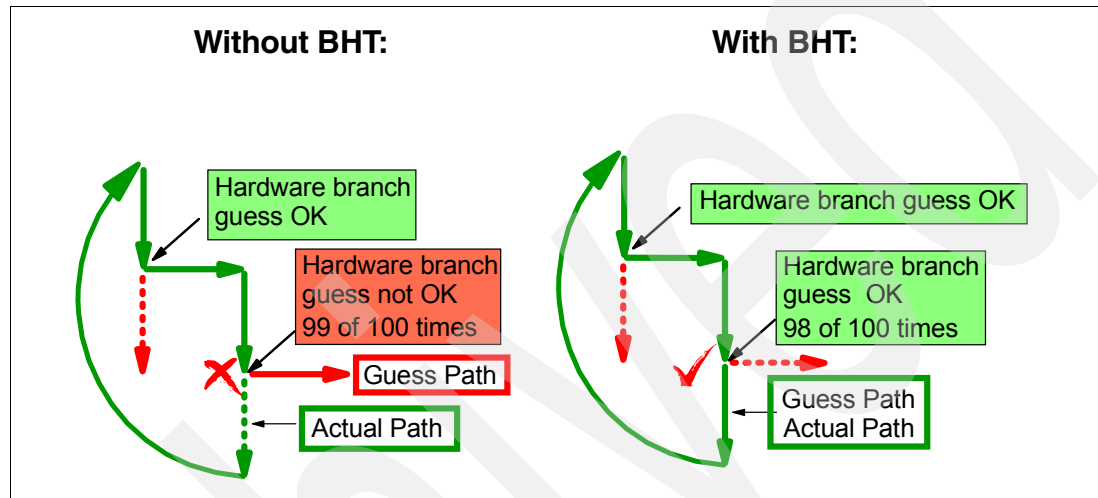


Figure 2-14 Branch History Table (BHT)

- ▶ Without BHT, the processor:
 - Makes an incorrect branch guess the first time through the loop (at the second branch point in Figure 2-14).
 - Preprocesses instructions for the guessed branch path.
 - Starts preprocessing a new path if the branch not equal to the guess.
 - Repeats this 98 more times until the last time, when the guess matches the actual branch taken
- ▶ With BHT, the processor:
 - Makes an incorrect branch guess the first time through the loop (at the second branch point in Figure 2-14).
 - Preprocesses instructions for the guessed branch path.
 - Starts preprocessing a new path if the branch is not equal to the guess.
 - Updates the BHT to indicate the last branch action taken at this address.
 - The next 98 times, the branch path comes from the BHT.
 - The last time, the guess is wrong.

The key point is that, with the BHT, the table is updated to indicate the last branch action taken at branch addresses. Using the BHT, if a hardware branch at an address matches a BHT entry, the branch direction is taken from the BHT. Therefore, in the diagram, the branches are correct for the remainder of the loop through the program routine, except for the last one.

The success rate that the BHT design offers contributes a great deal to the superscalar aspects of the z990, given the fact that the architecture rules prescribe that for successful parallel execution of an instruction stream, the correctly predicted result of the branch is essential.

IEEE Floating Point

The inclusion of the IEEE Standard for Binary Floating Point Arithmetic (IEEE 754-1985) in S/390 was made to further enhance the value of this platform for this type of calculation. The initial implementation had 121 floating-point instructions over prior S/390 CMOS models (Hexadecimal Floating Point had 54 instructions). Later, with the introduction of the 64-bit architecture, 12 additional instructions were added for IEEE Binary Floating Point Arithmetic 64-bit integer conversion.

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

On earlier systems, the emphasis has been on the traditional hexadecimal floating point arithmetic. The z990 has a Binary Floating Point unit that matches the performance of the traditional hexadecimal floating point unit by halving the number of cycles required earlier.

Translation Lookaside Buffer

The Translation Lookaside Buffer (TLB) in the Instruction and Data L1 caches now have a secondary TLB to enhance performance. In addition, a translator unit is added to translate misses in the secondary TLB.

Instruction fetching and instruction decode

The superscalar design of the z990 microprocessor allows for the decoding of up to two instructions per cycle and the execution of three instructions per cycle. Execution takes place in order, but storage accesses for instruction and operand fetching may occur out of sequence.

Instruction fetching

Instruction fetch in non-z990 models tries to get as far ahead of instruction decode and execution as possible because of the relatively large instruction buffers available. In the z990 microprocessor, smaller instruction buffers are used. The operation code is fetched from the I-cache and put in instruction buffers that hold pre-fetched data awaiting decode.

Instruction decoding

The processor can decode one or two instruction 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 redbook.

It is the compiler's responsibility to select instructions that best fit with the z990 superscalar microprocessor and abide by the grouping rules to create code that best exploits the superscalar implementation.

Extended Translation Facility

The Extended Translation Facility adds 10 instructions to the zSeries instruction set. They enhance the performance for data conversion operations for data encoded in Unicode, making applications enabled for Unicode and/or Globalization 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 will be the first to support the Extended Translation Facility instructions.

2.2.4 Processor Unit functions

One of the key components of the z990 server is the Processor Unit (PU). This is the microprocessor chip 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 512 KB Level 1 cache, split into 256 KB for data and 256 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 L2 cache. Each PU has multiple processors inside and instructions are executed twice, asynchronously, on both processors.

This asymmetric mirroring of instruction execution runs one cycle behind the actual operation. This allows the circuitry on the chip to be optimized for performance and does not compromise the simplified error detection process that is inherent to a mirrored execution unit design.

One or two Processor Units are contained on one processor chip. All PUs of a z990 server reside in a MultiChip Module. An MCM holds 12 PUs, of which eight are available for customer use, two are SAPs, and two are spares. Up to four MCMs, each contained in a book, may be available in one z990 server.

This approach allows a z990 server to have more PUs than required for a given initial configuration. This is a key point of the z990 design and is the foundation for the configuration flexibility and scalability of a single server.

All PUs in a z990 server are physically identical, but at initialization time PUs can be characterized to specific functions: CP, IFL, ICF, zAAP, or SAP. The function assigned to a PU is set by the Licensed Internal Code loaded when the system is initialized (Power-on Reset) and the PU is "characterized". Only characterized PUs have a designated function; non-characterized PUs are considered spares.

This design brings an outstanding flexibility to the z990 server, as any PU can assume any available characterization. This also plays an essential role in z990 system availability, as these PU assignments can be done dynamically, with no server outage, allowing:

- 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 via Capacity Upgrade on Demand (CUoD), Customer Initiated Upgrade (CIU), On/Off Capacity on Demand (On/Off CoD), or Capacity BackUp (CBU). More information about capacity upgrades is provided in 8.1, "Concurrent upgrades" on page 188.

- ▶ **PU sparing**

In the rare case of a PU failure, the failed PU's characterization is dynamically and transparently reassigned to a spare PU. More information on PU sparing is provided in “Sparing rules” on page 52.

A minimum of one PU per z990 server must be ordered as one of the following:

- ▶ A Central Processor (CP)
- ▶ An Integrated Facility for Linux (IFL)
- ▶ An Internal Coupling Facility (ICF)

The number of CPs, IFLs, ICFs, zAAPs, or SAPs assigned to particular models depends on the configuration. The z990 12-PU MCMs have two SAPs as standard. The standard number of SAPs in a model A08 is two; there are four in a B16, six in a C24, and eight in a D32. Optional additional SAPs may be purchased, up to two per book.

The z990 12-PU MCMs have two spares PUs as standard. The standard number of spares in a model A08 is two; there are four in a B16, six in a C24, and eight in a D32. The number of additional spare PUs is dependent on the number of books in the configuration and how many PUs are non-characterized.

Central Processors

A Central Processor is a PU that has the z/Architecture and ESA/390 instruction sets. It can run z/Architecture, ESA/390, Linux, and TPF operating systems, and the Coupling Facility Control Code (CFCC).

The z990 can only be used in LPAR mode. In LPAR mode, CPs can be defined as dedicated or shared to a logical partition. Reserved CPs can be defined to a logical partition, to allow for non-disruptive *image* upgrades. The z990 can have up to 32 CPs.

All CPs within a z990 configuration are grouped into a CP pool. Any z/Architecture, ESA/390, Linux, and TPF operating systems, and CFCC 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 via Capacity Upgrade on Demand (CUoD), Customer Initiated Upgrade (CIU), On/Off Capacity on Demand (On/Off CoD), or Capacity BackUp (CBU). More information about all forms of concurrent CP adds are found in Chapter 8, “Capacity upgrades” on page 187.

If the MCMs in the installed books have no available PUs left, the assignment of the next CP may result in the need for a model upgrade and the installation of an additional book. Book installation is a non-disruptive action, but will take more time than a simple Licensed Internal Code upgrade. Only if reserved processors have been defined to a logical partition—and when the operating system supports the function—can additional CP capacity be allocated to the logical partition dynamically.

Integrated Facilities for Linux

An Integrated Facility for Linux (IFL) is a PU that can be used to run Linux on zSeries, Linux for S/390, or Linux guests on z/VM operating systems. Up to 32 PUs may be characterized as IFLs, depending on the z990 model. IFL processors can be dedicated to a Linux or a z/VM logical partition, or be shared by multiple Linux guests and/or z/VM logical partitions running on the same z990 server. Only z/VM, Linux on zSeries, and Linux for S/390 operating systems can run on IFL processors.

All PUs characterized as IFL processors within a configuration are grouped into the ICF/IFL/zAAP processor pool. The ICF/IFL/zAAP processor pool appears on the hardware console as ICF processors. The number of ICFs shown is the sum of IFL, ICF, and zAAP processors on the server.

IFLs do not change the software model number of the z990 server. Software product license charges based on the software model number are not affected by the addition of IFLs.

Within the limit of all non-characterized PUs available in the installed configuration, IFLs can be concurrently added to an existing configuration via Capacity Upgrade on Demand (CUoD), Customer Initiated Upgrade (CIU), On/Off Capacity on Demand (On/Off CoD), but IFLs *cannot* be assigned via CBU. For more information about CUoD, CIU or On/Off CoD see Chapter 8, “Capacity upgrades” on page 187. If the installed books have no unassigned PUs left, the assignment of the next IFL may require the installation of an additional book.

Internal Coupling Facilities

An Internal Coupling Facility (ICF) is a PU used to run the IBM Coupling Facility Control Code (CFCC) 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 z990 model. You need at least an IBM 2084 model B16 to assign 16 ICFs. ICFs can be concurrently assigned to an existing configuration via Capacity Upgrade on Demand (CUoD), On/Off Capacity on Demand (On/Off CoD), or Customer Initiated Upgrade (CIU), but ICFs *cannot* be assigned via CBU.

For more information about CUoD, CIU, or On/Off CoD, see Chapter 8, “Capacity upgrades” on page 187. If the installed books have no non-characterized PUs left, the assignment of the next ICF may require the installation of an additional book.

The ICF processors can only be used by Coupling Facility logical partitions. ICF processors can be dedicated to a CF logical partition, or shared by multiple CF logical partitions running in the same z990 server.

All ICF processors within a configuration are grouped into the ICF/IFL/zAAP processor pool. The ICF/IFL/zAAP processor pool appears on the hardware console as ICF processors. The number of ICFs shown is the sum of IFL, ICF, and zAAP processors on the system.

Only Coupling Facility Control Code (CFCC) can run on ICF processors; ICFs do not change the model type of the z990 server. This is important because software product license charges based on the software model number are not affected by the addition of ICFs.

Dynamic ICF Expansion

Dynamic ICF Expansion is a function that allows a CF logical partition running on dedicated ICFs to acquire additional capacity from the LPAR pool of shared CPs or shared ICFs. The trade-off between using ICF features or CPs in the LPAR shared pool is the exemption from software license fees for ICFs. Dynamic ICF Expansion is available on any z990 model that has at least one ICF.

Dynamic ICF Expansion requires that the Dynamic CF Dispatching be turned on (DYNDISP ON). For more information, see 7.2.7, “Dynamic CF dispatching and dynamic ICF expansion” on page 168.

Dynamic Coupling Facility Dispatching

The Dynamic Coupling Facility Dispatching function has an enhanced dispatching algorithm that lets you define a backup Coupling Facility in a logical partition on your system. While this logical partition is in backup mode, it uses very little processor resource. When the backup CF becomes active, only the resource necessary to provide coupling is allocated.

The CFCC command DYNDISP controls the Dynamic CF Dispatching (use DYNDISP ON to enable the function). For more information, see 7.2.7, “Dynamic CF dispatching and dynamic ICF expansion” on page 168.

zSeries Application Assist Processors

The zSeries Application Assist Processor (zAAP) is a PU that is used exclusively for running Java application workloads under z/OS. One CP must be installed with or prior to any zAAP being installed. The number of zAAPs in a machine cannot exceed the number of CPs plus unassigned CPs in that machine. Within the capacity of the sum of all unassigned PUs in up to four books, up to 16 zAAPs can be characterized, depending on the z990 model. Up to four zAAPs can be characterized per book. You need an IBM 2084 model D32 with a total of 16 assigned and unassigned CPs to assign 16 zAAPs.

Within the limit of all non-characterized PUs available in the installed configuration, zAAPs can be concurrently added to an existing configuration via Capacity Upgrade on Demand (CUoD), Customer Initiated Upgrade (CIU), On/Off Capacity on Demand (On/Off CoD), but zAAPs *cannot* be assigned via CBU.

With On/Off CoD, you may concurrently install temporary zAAP capacity by ordering On/Off CoD Active zAAP features up to the number of current zAAPs that are permanently purchased. Also, the total number of On/Off CoD Active zAAPs plus zAAPs cannot exceed the number of On/Off Active CPs plus the number of CPs plus the number unassigned CPs on a z990 server.

For more information about CUoD, CIU, or On/Off CoD, see Chapter 8, “Capacity upgrades” on page 187. If the installed books have no unassigned PUs left, the assignment of the next zAAP may require the installation of an additional book.

PUs characterized as zAAPs within a configuration are grouped into the ICF/IFL/zAAP processor pool. The ICF/IFL/zAAP processor pool appears on the hardware console as ICF processors. The number of ICFs shown is the sum of IFL, ICF, and zAAP processors on the server.

zAAPs are orderable by feature code (FC 0520). Up to one zAAP can be ordered for each CP or unassigned CP configured in the machine.

Important: The zAAP is a specific example of an assist processor that is known generically as an Integrated Facility for Applications (IFA). The generic term IFA often appears in panels, messages, and other online information relating to the zAAP.

zAAPs and LPAR definitions

zAAP processors can be defined as dedicated or shared processors in a logical partition and are always related to CPs of the same partition. For a logical partition image, both CPs and zAAPs logical processors are either dedicated or shared.

Purpose of a zAAP

zAAPs are designed for z/OS Java code execution. When Java code must be executed (that is, 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, the z/OS dispatcher redispaches the JVM task on an available CP, after which normal processing is resumed. This reduces the CP time needed to run WebSphere applications, freeing capacity for other workloads.

A zAAP only executes Java Virtual Machine (JVM) code and is the only authorized user of a zAAP in association with some z/OS infrastructural code 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 like IPL.

Java application code can either run on a CP or an zAAP. The user can manage the use of CPs such that Java application code runs only on a CP, only on an zAAP, or on both when zAAPs are busy.

For the logical flow of a Java code execution on a zAAP, see Figure 6-3 on page 139.

Software support

zAAPs do not change the software model number of the z990 server. IBM software product license charges based on the software model number are not affected by the addition of zAAPs.

z/OS Version 1.6 is a prerequisite for supporting zAAPs, together with IBM SDK for z/OS Java 2 Technology Edition V1.4.1.

Exploiters of zAAPs include:

- ▶ WebSphere Application Server 5.1
- ▶ CICS/TS 2.3
- ▶ DB2® Version 8
- ▶ IMS™ Version 8
- ▶ WebSphere WBI for z/OS

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 z990 models have standard SAPs configured. The IBM 2084 model A08 has two SAPs, the model B16 has four SAPs, the model C24 has six SAPs, and the model D32 has eight SAPs as the standard configuration.

Channel cards are assigned across SAPs to balance SAP utilization and improve I/O subsystem performance.

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 z990 servers, the number of SAPs can be greater than the number of CPs and the number of used STIs.

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 TPF environments. The maximum number of optional additional orderable SAPs depends on the model and the number of available uncharacterized PUs in the configuration:

- ▶ IBM 2084-A08: Maximum additional orderable SAPs is two.
- ▶ IBM 2084-B16: Maximum additional orderable SAPs is four.

- ▶ IBM 2084-C24: Maximum additional orderable SAPs is six.
- ▶ IBM 2084-D32: Maximum additional orderable SAPs is eight.

Optionally assignable SAPs

Assigned CPs may be optionally reassigned as SAPs instead of CPs, 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 will reduce the number of logical processors you can activate. Activation of a logical partition will fail if the number of logical processors you attempt to activate exceeds the number of CPs available. To avoid a logical partition activation failure, you should verify that the number of logical processors assigned to a logical partition does not exceed the number of CPs available.

Note: Concurrent upgrades are not supported with CPs defined as additional SAPs.

Reserved processors

Reserved processors can be defined to a logical partition. Reserved processors are defined by the Processor Resource/System Manager (PR/SM) to allow non-disruptive *capacity* upgrade. Reserved processors are like “spare *logical* processors.” They can be defined as Shared or Dedicated.

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

Reserved processors also provide the capability of defining to a logical partition more logical processors than the number of available CPs, IFLs, ICFs, and zAAPs in the configuration. This makes it possible to configure online, non-disruptively, more logical processors after additional CPs, IFLs, ICFs, and zAAPs have been made available concurrently, via CUoD, CIU, and On/Off CoD for CPs, IFLs, ICFs, and zAAPs, or CBU for CPs. See 8.1, “Concurrent upgrades” on page 188 for more details.

When no reserved processors are defined to a logical partition, a processor upgrade in 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 upon the number of logical processors that are defined. For an ESA/390 mode logical partition the sum of defined and reserved logical processors is limited to 32, including CPs and zAAPs. However up to 24 processors, including CPs and zAAPs, are planned to be supported by z/OS 1.6. The z/VM 5.1 is planned to support up to 24 processors, either all CPs or all IFLs.

For more information about logical processors and reserved processors definition, see “Logical Partitioning overview” on page 57.

Processor Unit characterization

Processor Unit (PU) characterization is done at Power-on Reset time when the server is initialized. The z990 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, and zAAPs. 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 Processor Unit allocation rules take into consideration the newly installed book.

Note: Even in a multi-book system, a book failure is a CPC system failure. Until the failing book is repaired or replaced, Power-On Reset of the z990 with the remaining books is not supported.

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

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

The z990 server comes with two, four, six, or eight standard spare PUs, depending on the model. CP, IFL, ICF, zAAP, and SAP sparing is completely transparent and requires no operating system or operator intervention.

With transparent sparing, the application that was running on the failed processor is preserved and will continue processing on a newly assigned CP, IFL, ICF, zAAP, or SAP (allocated to one of the spare PUs) without customer intervention. If no spare PU is available, Application preservation is invoked.

Application preservation

Application preservation is used in the case where a processor fails and there are no spare PU available. 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.

Dynamic SAP sparing and reassignment

Dynamic recovery is provided in case of failure of the System Assist Processor (SAP). In the event of a SAP failure, if a spare PU is available, the spare PU will be dynamically assigned as a new SAP. If there is no spare PU available, and more than one CP is characterized, a characterized CP is reassigned as a SAP. In either case, there is no customer intervention required. This capability eliminates an unplanned outage and permits a service action to be deferred to a more convenient time.

Sparing rules

The sparing rules for the allocation of spare CPs, IFLs, ICFs, zAAPs, and SAPs depend on the type of processor chip on which the failure occurs. On each MCM, two standard spare PUs are available. The two standard SAPs and two standard spares are initially allocated to dual core processor chips. Table 2-3 on page 53 illustrates the default PU-to-chip mapping.

Table 2-3 PU chip allocation

	Core	Core	X = CP, IFL, ICF, or zAAP	
Single core	0	-	X	-
	2	-	X	-
	4	-	X	-
	6	-	X	-
Dual core	8	9	X	Spare
	A	B	X	Spare
	C	D	X	SAP
	E	F	X	SAP

- ▶ On a single-book configuration, model A08:
 - When a PU failure occurs on a dual-core chip, the two standard spares PUs are used to recover the failing chip, even though only one of the PUs has failed.
 - When a failure occurs on a PU on a single-core chip, one standard spare PU is used.

The system does not issue an RSF call in either of the above circumstances.

When a non-characterized PU is used as a spare, in case the system has run out of the standard spares, or when all PUs have been assigned and no non-characterized PU remains available, an RSF call occurs to request a book repair.
- ▶ On a multi-book configuration, models B16, C24, or D32:
 - In a first step, the standard spare PUs of the MCM where the failing PU resides is assigned as spare, in the same manner as for a one-book system.
 - In a second step, when there are not enough spares in the book with the failing PU, non-characterized PUs in other books are used for sparing. When “cross-book” sparing occurs, the book closest to the one with the failing PU will be used.

For example, if a PU failure in Book-1 cannot be solved within locally, spares in Book-2 or Book-0 are then selected. When no spares are available in any adjacent book, Book-3 is approached for a spare PU.

2.2.5 Memory design

As for PUs and the I/O subsystem designs, the z990 memory design equally provides great flexibility and high availability, allowing:

- ▶ Concurrent Memory upgrades (except when the physically installed capacity is reached.)

The z990 servers 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 via Capacity Upgrade on Demand or Customer Initiated Upgrade. Note that memory upgrades *cannot* be done via Capacity BackUp (CBU); see Table 8-1 on page 190 for more information.

► **Dynamic Memory sparing**

The z990 does not contain spare memory DIMMs. Instead, it has redundant memory distributed throughout its operational memory and this is used to bypass failing memory. Replacing memory cards requires the removal of a book and this is disruptive. The extensive use of redundant elements in the operational memory greatly minimizes the possibility of a failure that requires memory card replacement.

► **Partial Memory Restart**

In the rare event of a memory card failure, Partial Memory Restart enables the system to be restarted with only part of the original memory. In a one-book system, the failing card will be deactivated, after which the system can be restarted with the memory on the remaining memory card.

In a system with more than one book, all physical memory in the book containing the failing memory card is taken offline, allowing 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.

Memory error-checking and correction code detects and corrects single-bit errors, or 2-bit errors from a chipkill failure, using the Error Correction Code (ECC). Also, because of the memory structure design, errors due to a single memory chip failure are corrected.

Memory background scrubbing provides continuous monitoring of storage for the correction of detected faults before the storage is used.

The memory cards use the latest fast 256 Mb, and 512 Mb, synchronous DRAMs. Memory access is interleaved between the memory cards to equalize memory activity across the cards.

Memory cards have 8 GB, 16 GB, or 32 GB of capacity. All memory cards installed in one book must have the same capacity. Books may have different memory sizes, but the card size of the two cards per book must always be the same.

The total capacity installed may have more usable memory than required for a configuration, and Licensed Internal Code Configuration Control (LIC-CC) will determine how much memory is used from each card. The sum of the LIC-CC 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. Actually, PR/SM is responsible for the memory assignments; it is PR/SM that controls the resource allocation of the CPC. Table 2-1 on page 28 shows the distribution of physical memory across books when a system initially is installed with the amounts of memory shown in the first column. However, the table gives no indication of *where* the initial memory is allocated. Memory allocation is done as evenly as possible across all installed books.

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 non-disruptively 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.

Due to the memory allocation algorithm, systems that undergo a number of MES upgrades for memory can have a variety of memory card mixes in all books of the system. If, however

unlikely, memory should fail, it is technically feasible to Power-on Reset the system with the remaining memory resources (see “Partial Memory Restart” on page 54). After Power-on Reset, the memory distribution across the books is now different, as is the amount of memory.

Capacity Upgrade on Demand (CUoD) for memory can be used to order more memory than needed on the initial model, but that is required on the target model; see “Memory upgrades” on page 29. For more information about CUoD for memory, refer to “CUoD for memory” on page 193.

Processor memory, even though physically the same, can be configured as both Central storage and Expanded storage.

Central storage (CS)

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 the 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.

A logical partition can have more than 2 GB defined as central storage, but 31-bit operating systems cannot use central storage above 2 GB; refer to 2.2.8, “Storage operations” on page 67 for more detail.

Expanded storage (ES)

Expanded storage (ES) can optionally be defined on z990 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. As they operate in 64-bit addressing mode, they can have all the required storage capacity allocated as central storage. z/VM is an exception since, even when operating in 64-bit mode, it can have guest virtual machines running in 31-bit addressing mode, which can use expanded storage.

It is *not* possible to define expanded storage to a Coupling Facility image. 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 z990 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.

The Storage Assignment function of a Reset Profile on the Hardware Management Console just shows the total “Installed Storage” and the “Customer Storage”, which is the total installed storage minus the Hardware System Area (HSA). Logical partitions are still defined to have Central Storage and optional Expanded Storage. Activation of logical partitions, as well as dynamic storage reconfiguration, will cause the storage to be converted to the type needed.

Activation of logical partitions as well as dynamic storage reconfiguration will cause the storage to be assigned to the type needed (CS or ES). This does not require a Power-on Reset. No new software support is required to take advantage of this function.

Hardware System Area (HSA)

The Hardware System Area (HSA) is a non-addressable storage area that contains the CPC Licensed Internal Code and configuration-dependent control blocks. The HSA size varies according to:

- ▶ The number of defined logical partitions.
- ▶ If dynamic I/O is not enabled, the size and complexity of the system I/O configuration. The HSA may hold the configuration information for up to 63 K devices per LCSS.
- ▶ If dynamic I/O is enabled, the MAXDEV value specified in HCD or IOCP, in support of dynamic I/O configuration.

Note: The HSA is always allocated in the physical memory of Book 0.

2.2.6 Modes of operation

Figure 2-15 on page 57 shows the z990 modes of operation diagram, summarizing all available mode combinations that are discussed in this section: z990 mode, image modes and their processor types, operating system versions and releases, and architecture modes.

Note: The z990 models only operate in Logically Partitioned Mode. ESA/390 TPF mode is now only available as an image mode in a logical partition.

There is no special operating mode for the 64-bit z/Architecture mode, as the architecture mode is not an attribute of the definable images operating mode.

The 64-bit operating systems are IPLed into 31-bit mode and, optionally, can change to 64-bit mode during their initialization. It is up to the operating system to take advantage of the addressing capabilities provided by the architectural mode.

The operating systems supported on z990 servers are shown in Chapter 6, “Software support” on page 133.

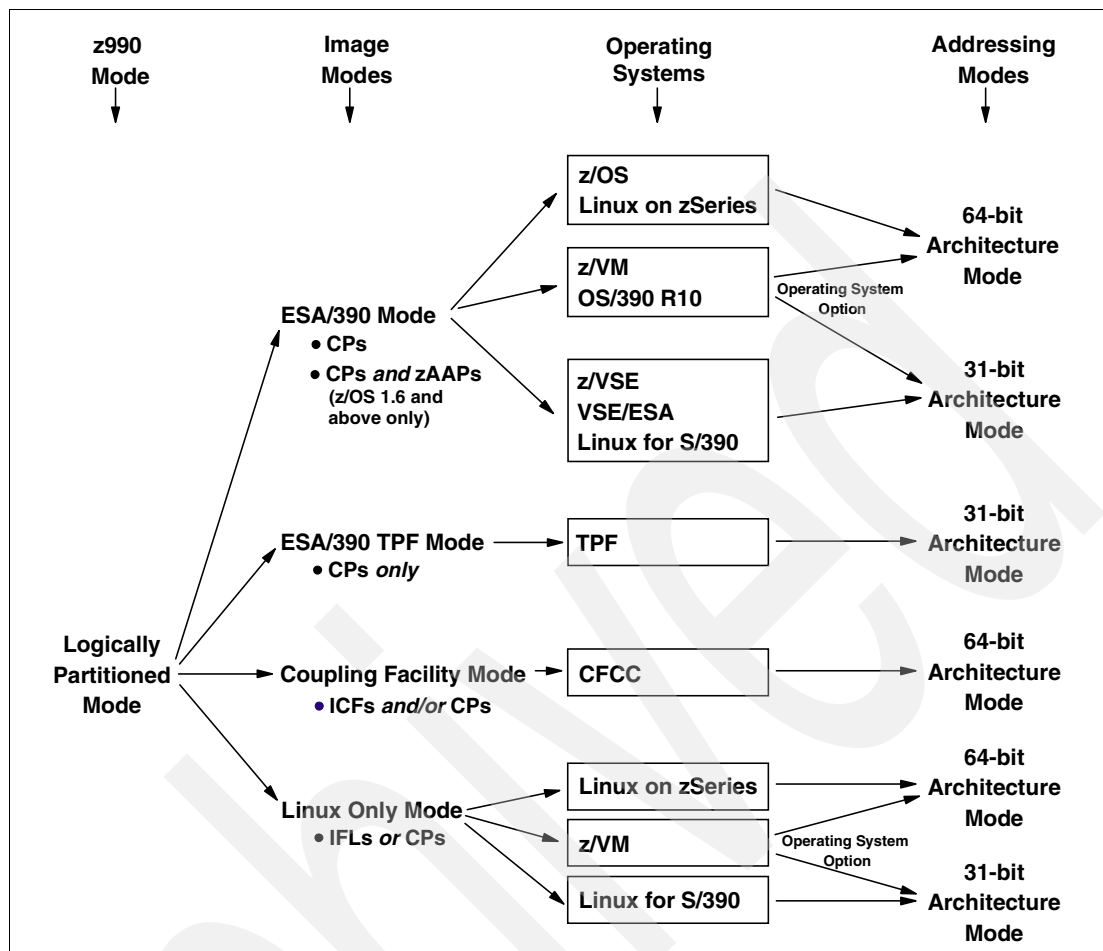


Figure 2-15 z990 Modes of operation diagram

Logical Partitioning overview

Logical Partitioning is a function implemented by the Processor Resource/Systems Manager (PR/SM), available on all z990 servers.

The z990 only runs in LPAR mode. This means that virtually all system aspects are now controlled by PR/SM functions.

PR/SM is very much aware of the book structure introduced on the z990. However, logical partitions do not have this awareness. Logical partitions have resources allocated to them coming from a variety of physical resources, and have no control over these physical resources from a systems standpoint—but the PR/SM functions do have.

PR/SM manages and optimizes allocation and dispatching work on the physical topology. Most physical topology knowledge that was previously handled to the operating systems is now 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 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 re-dispatch a logical PU on the same physical PU to assure that as much as possible of the L1 cache content can be reused.

PR/SM enables z990 servers to be initialized for logically partitioned operation, supporting up to 30 logical partitions. Each logical partition can run its own operating system image in any image mode, independently from the other logical partitions.

A logical partition can be activated or deactivated at any time, but changing the number of defined logical partitions is disruptive, as it requires a Power-on Reset (POR). Some facilities may not be available to all operating systems, as they may have software corequisites.

Each logical partition has the same resources as a “real” CPC, which are:

► Processor(s)

Called *Logical Processor(s)*, they can be defined as CPs, IFLs, ICFs, or zAAPs. They can be *dedicated* to a partition or *shared* between 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.

For z/OS Workload License Charge (WLC), a logical partition “Defined Capacity” can be set, enabling the *soft* capping function.

ESA/390 mode logical partitions can have CPs and zAAPs logical processors. Both logical processor types can be defined as either all dedicated or all shared. The zAAP support is planned to be introduced by z/OS 1.6.

Only Coupling Facility (CF) partitions can have both dedicated *and* shared logical processors defined.

Figure 2-16 on page 59 shows the logical processor assignment screen of the Customize Image Profile on the Hardware Management Console (HCM), for an ESA/390 mode image. This panel allows the definition of:

- Dedicated or shared logical processors, including CPs and zAAPs (remember that zAAPs initially appear as “Integrated Facility for Applications” on HCM panels)
- The initial weight, optional capping, and Workload Manager options for shared CPs (a shared zAAP’s weight equals a CP’s weight, but share calculation is based on the sum of ICFs’, IFLs’ and zAAPs’ weights)
- The number of initial and optional reserved processors (CPs)
- The optional number of initial and reserved integrated facility for application (zAAPs)

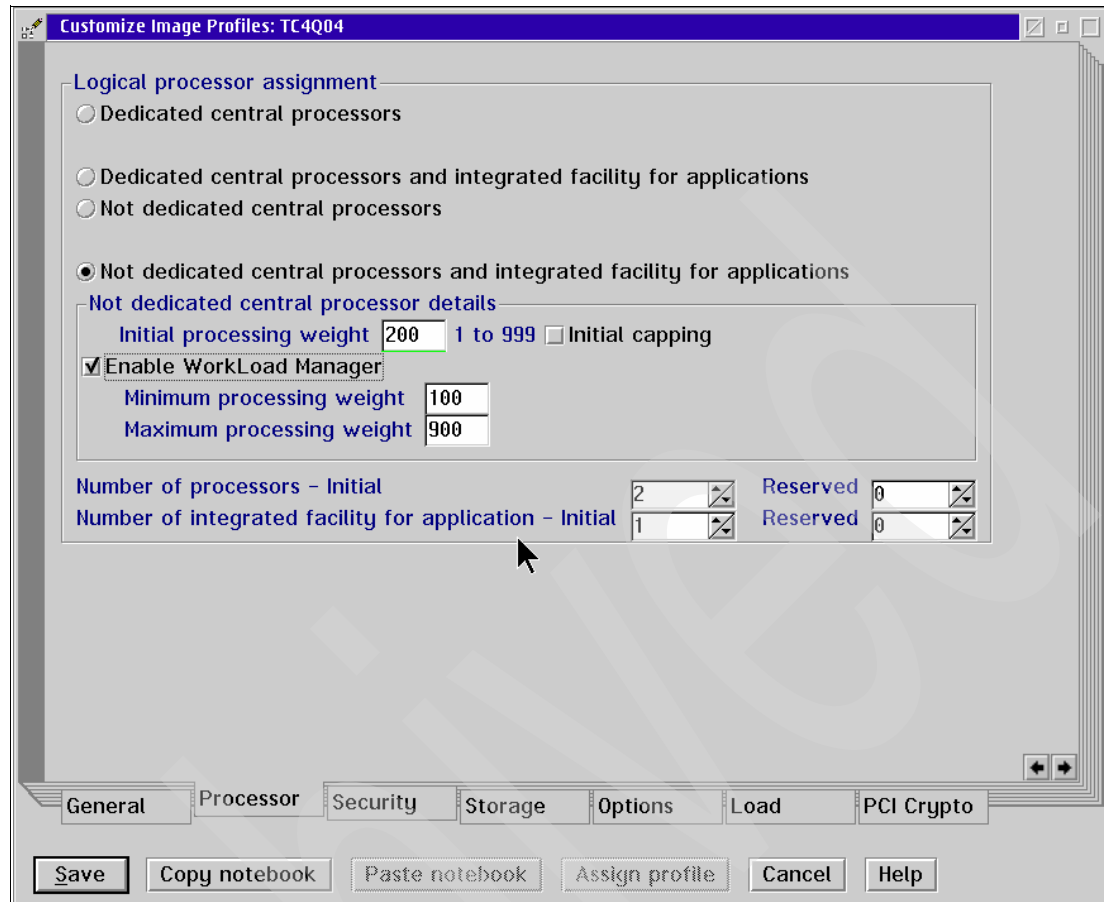


Figure 2-16 Logical processor assignment (HMC- Image Profile)

On the z990, the sum of defined and reserved logical processors for an ESA/390 mode logical partition is limited to 32. However, z/OS 1.6 and z/VM 5.1 operating systems are planned to support up to 24 processors. For z/OS, the 24 processors limit applies to the sum of CPs and zAAPs logical processors. 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.

► Memory

Memory, either Central Storage or Expanded Storage, must be dedicated to a logical partition. The defined storage(s) must be available during the logical partition activation; otherwise, the activation fails.

Reserved storage can be defined to a logical partition, enabling non-disruptive memory add to and removal from a logical partition, using the LPAR Dynamic Storage Reconfiguration. Refer to 2.2.11, “LPAR Dynamic Storage Reconfiguration (DSR)” on page 71 for more information.

► 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 I/O 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, OS/390, z/VM, VM/ESA, z/VSE, and VSE/ESA operating systems, and as a stand-alone program on the z990 hardware console. HCD is available on z/OS, z/VM, and OS/390 operating systems.

ESCON channels (CHPID type CNC or FCV) 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.

Logically partitioned mode

The z990 server can only run in LPAR Mode; up to 30 logical partitions can be defined on any z990 server. A logical partition can be defined to operate in one of the following image modes:

- ▶ ESA/390 mode, to run:
 - A z/Architecture operating system image, on dedicated *or* shared CPs
 - An ESA/390 operating system image, on dedicated *or* shared CPs
 - A Linux operating system image, on dedicated *or* shared CPs
 - A z/OS 1.6 or later operating system image, on any of the following:
 - Dedicated *or* shared CPs
 - Dedicated CPs *and* dedicated zAAPs
 - Shared CPs *and* shared zAAPs

Note: zAAPs can be defined to any ESA/390 mode image (see Figure 2-16 on page 59). However, zAAPs are supported only by z/OS 1.6 and later operating systems. Any other operating system cannot use zAAPs, even if they are defined to the logical partition. zAAPs are not supported for a z/OS guest under z/VM.

- ▶ ESA/390 TPF mode, to run:
 - A TPF operating system image, only on dedicated *or* shared CPs
- ▶ Coupling Facility mode, to run a CF image, by loading the CFCC into this logical partition. The CF image can run any of the following definitions:
 - Dedicated *or* shared CPs
 - Dedicated *or* shared ICFs
 - Dedicated *and* shared ICFs
 - ICFs dedicated *and* CPs shared
- ▶ Linux-only mode, to run:
 - A Linux operating system image, on either:
 - Dedicated *or* shared IFLs
 - Dedicated *or* shared CPs
 - A z/VM operating system image, on either:
 - Dedicated *or* shared IFLs
 - Dedicated *or* shared CPs

Table 2-4 on page 61 shows all LPAR modes, required characterized PUs, and operating systems, and which PU characterizations can be configured to a logical partition image. The available combinations of dedicated (DED) and shared (SHR) processors are also shown. For

all combinations, an image can also have Reserved Processors defined, allowing non-disruptive image upgrades.

Table 2-4 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	CPs DED <i>or</i> CPs SHR
	CPs <i>and</i> zAAPs	z/OS (1.6 and later)	CPs DED <i>and</i> zAAPs DED, <i>or</i> CPs SHR <i>and</i> zAAPs SHR
ESA/390 TPF	CPs	TPF	CPs DED <i>or</i> CPs SHR
Coupling Facility	ICFs <i>and/or</i> CPs	CFCC	ICFs DED <i>or</i> ICFs SHR, <i>or</i> CPs DED <i>or</i> CPs SHR, <i>or</i> ICFs DED <i>and</i> ICFs SHR, <i>or</i> ICFs DED <i>and</i> CPs SHR
Linux Only	IFLs <i>or</i> CPs	Linux z/VM	IFLs DED <i>or</i> IFLs SHR, <i>or</i> CPs DED <i>or</i> CPs SHR

Dynamic Add/Delete of a logical partition name

The ability to add meaningful logical partition names to the configuration without a Power-On Reset is being introduced. Prior to this support, extra logical partitions were defined by adding reserved names in the Input/Output Configuration Data Set (IOCDS), but one may not have been able to predict what might be meaningful names in advance.

Dynamic add/delete of a logical partition name allows reserved logical partition 'slots' to be created in an IOCDS in the form of extra Logical Channel Subsystem (CSS), Multiple Image Facility (MIF) image ID pairs. A reserved partition is defined with the partition name placeholder '*', and cannot be assigned to an access or candidate list of channel paths or devices. These extra Logical Channel Subsystem MIF image ID pairs (CSSID/MIFID) can be later assigned an logical partition name for use (or later removed) via dynamic I/O commands using the Hardware Configuration Definition (HCD). The IOCDS still must have the extra I/O slots defined in advance since many structures are built based upon these major I/O control blocks in the Hardware System Area (HSA). This support is exclusive to the z990 and z890 and is applicable to z/OS V1.6, which is planned to be available in September 2004.

When a logical partition is renamed, its name can be changed from 'NAME1' to '*' and then changed again from '*' to 'NAME2'; the logical partition number and MIFID are retained across the logical partition name change. However, the master keys in PCIXCC that were associated with the old logical partition 'NAME1' are retained. There is no explicit action taken against a cryptographic component for this.

Attention: Cryptographic cards are not tied to partition numbers or MIFIDs. 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 now and continues to have responsibility to clear them out if he changes who is using them.

2.2.7 Model configurations

The z990 server model nomenclature is based on the number of PUs available for customer use in each configuration. Four models of the z990 server are available:

- ▶ IBM 2084 model A08: Eight PUs are available for characterization as CPs, IFLs, ICFs, zAAPs (up to four), or additional SAPs.
- ▶ IBM 2084 model B16: 16 PUs are available for characterization as CPs, IFLs, ICFs, zAAPs (up to eight), or additional SAPs.
- ▶ IBM 2084 model C24: 24 PUs are available for characterization as CPs, IFLs, ICFs, zAAPs (up to 12), or additional SAPs.
- ▶ IBM 2084 model D32: 32 PUs are available for characterization as CPs, IFLs, ICFs, zAAPs (up to 16), or additional SAPs.

When a z990 order is configured, PUs are selected according to their intended usage. They can be ordered as:

CP	The Processor Unit purchased and activated supporting the z/OS, OS/390, z/VSE, VSE/ESA, z/VM, and Linux operating systems, which can also run the Coupling Facility Control Code (CFCC).
Unassigned CP	A Central Processor purchased for future use as a CP. It is offline and unavailable for use.
IFL	The Integrated Facility for Linux (IFL) is a Processor Unit that is purchased and activated for exclusive use by the z/VM and Linux operating systems.
Unassigned IFL	A Processor Unit purchased for future use as an IFL. It is offline and unavailable for use.
ICF	A Processor Unit purchased and activated for exclusive use by the Coupling Facility Control Code (CFCC).
zAAP	A Processor Unit purchased and activated for exclusive use to run Java code under control of z/OS JVM.
Additional SAP	The Optional System Assist Processor (SAP) is a Processor Unit that is purchased and activated for use as a SAP.

Unassigned CPs are purchased PUs with the intention to be used as CPs, and usually have this status for software charging reasons. Unassigned CPs do not count in establishing the MSU value to be used for MLC software charging, or when charged on a per Processor Unit basis.

Unassigned IFLs are purchased IFLs with the intention to be used as IFLs, and usually have this status for software charging reasons. Unassigned IFLs do not count in establishing the charge for either z/VM or Linux.

This method prevents RPQ handling in case a temporary downgrade is required. When the capacity need arises, the unassigned CPs and IFLs can be assigned nondisruptively.

Upgrades

Concurrent CP, IFL, ICF, or zAAP upgrades are done within a z990 model. Concurrent upgrades require PU spares. PU spares are PUs that are *not* the two standard spares on each MCM and are not characterized as a CP, IFL, ICF, zAAP, or SAP.

If the upgrade request cannot be accomplished within the given model, a model upgrade is required. A model upgrade will cause the addition of one or more books to accommodate the desired capacity. Additional books can be installed concurrently.

Upgrades from one IBM 2084 model to another are concurrent and mean that one or more books are added. Table 2-5 shows the possible model upgrades within the IBM 2084 model range.

Table 2-5 z990 upgrade paths

z990 Models	2084-A08	2084-B16	2084-C24	2084-D32
2084-A08	-	X	X	X
2084-B16	-	-	X	X
2084-C24	-	-	-	X
2084-D32	-	-	-	-

Upgrade paths from IBM 2064 models (z900) offer a virtually unrestricted upgrade capability. Upgrades from any z900 to any z990 server are supported (with the exception of the IBM 2064 model 100, which can only be upgraded to another z900 model). There are also no upgrade paths from IBM 9672 G5 and G6 models, nor is there an upgrade path from the IBM 2066 models (z800).

There are limited upgrade paths from the 2086-A04 model (z890) to the 2084-A08 model (z990). For details, contact your IBM representative or IBM Business Partner.

Figure 2-17 shows all upgrade paths from z900 to z990 models, and all upgrade paths within the z990 model range.

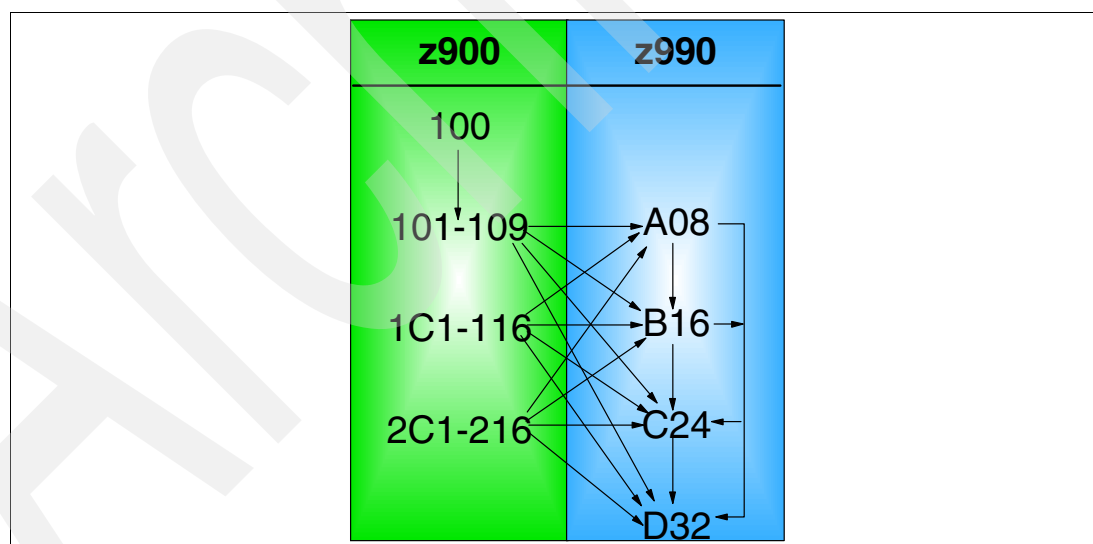


Figure 2-17 Supported z900-to-z990 upgrade paths

Software models

In order to recognize how many PUs are characterized as a CP, the STSI instruction returns a value that can be seen as a software model number to determine the number of assigned CPs. Characterization of a PU as an IFL, an ICF, or an zAAP is not reflected in the output of the STSI instruction, since they have no effect on MLC software charging.

Table 2-6 shows that regardless of the number of books, a configuration with one characterized CP is possible (for example, an IBM 2084 model D32 may have only one PU characterized as a CP for customer use).

Table 2-6 z990 software models

z990 Models	Software Models
IBM 2084-A08	301 - 308
IBM 2084-B16	301 - 316
IBM 2084-C24	301 - 324
IBM 2084-D32	301 - 332

Note: Software model number 300 is used for IFL or ICF only models.

This structure enables a different approach to downgrading the system in cases where a larger system is installed on which, for software charging reasons, temporarily less CP capacity must be assigned. It is now possible (with the use of the IBM internal e-Config tool) to order a simultaneous downgrade.

Consider, for example, an IBM 2084-A08 ordered with six PUs for customer use. Feature code (F/C 0716) specifies the number of PUs characterized as CPs (assume four), and a different feature code (F/C 1716) specifies the number of unassigned CPs (assume two). The unassigned CPs are part of the order, but cannot be used and will not be charged for MLC software charges. Later, when the capacity need requires it, the unassigned CPs can be assigned and will become active assigned CPs.

An unassigned CP is a PU that is purchased as a CP, but is not active in the model configuration. An unassigned IFL is a PU that is purchased as an IFL, but is not active in the current model configuration.

A minimum of one PU characterized as a CP, IFL, or ICF is required per system. PUs can be characterized as CPs, IFLs, ICFs, or zAAPs. The maximum number of CPs is 32, the maximum number of IFLs is 32, the maximum number of ICFs is 16, and the maximum number of zAAPs amounts to 16 (up to four zAAPs per book). Not all PUs available on a model are required to be characterized as a CP, IFL, ICF, or zAAP. Only purchased PUs are identified by a feature code.

Feature codes related to CPs and unassigned CPs, IFLs and unassigned IFLs, ICFs, SAPs, and zAAPs are:

- ▶ Feature code 0716 for a CP
- ▶ Feature code 1716 for an unassigned CP
- ▶ Feature code 0516 for an IFL
- ▶ Feature code 0517 for an unassigned IFL
- ▶ Feature code 0518 for an ICF
- ▶ Feature code 0519 for an optional SAP
- ▶ Feature code 0520 for a zAAP

PU conversions

Assigned CPs, unassigned CPs, assigned IFLs, unassigned IFLs, and ICFs may be converted to other assigned, or unassigned feature codes. Valid conversion paths are:

- ▶ Conversion of feature code 0716 to 1716, 0516 or 0518 for conversion of a CP to an unassigned CP, an IFL, or an ICF

- Conversion of feature code 1716 to 0716, for conversion of an unassigned CP to a CP
- Conversion of feature code 0516 to 0517, 0518 or 0716, for conversion of an IFL to an unassigned IFL, an ICF, or a CP
- Conversion of feature code 0517 to 0516, for conversion of an unassigned IFL to an IFL
- Conversion of feature code 0518 to 0516 or 0716, for conversion of an ICF to an IFL or a CP

All listed conversions are usually non-disruptive. In exceptional cases the conversion may be disruptive, for example, when an IBM 2084 model A08 with eight CPs is converted to an all IFL system. In addition, LPAR disruption may occur when PUs must be freed before they can be converted.

This information is also summarized in Table 2-7.

Table 2-7 *PU conversions*

From	To	CP (0716)	Unassigned CP (1716)	IFL (0516)	Unassigned IFL (0517)	ICF (0518)
CP (0716)	-	Yes	Yes	Yes	No	Yes
Unassigned - CP (1716)	Yes	-	No	No	No	No
IFL (0516)	Yes	No	-	Yes	Yes	Yes
Unassigned IFL (0517)	No	No	Yes	-	No	No
ICF (0518)	Yes	No	Yes	No	-	-

Capacity Backup (CBU)

CBUs deliver temporary capacity (feature code 7800) on top of what a customer might have installed in numbers of assigned CPs, IFLs, ICFs, zAAPs and additional SAPs. The total number of active PUs (the sum of all assigned CPs, IFLs, ICFs, zAAPs, and additional SAPs) plus the number of CBUs cannot exceed the total number of PUs available on the MCMs in all books.

To determine the number of CBUs that can be added to a given configuration, some rules must be considered:

- The number of assigned CPs + IFLs + ICFs + zAAPs + additional SAPs + CBUs equals less than 8 on an IBM 2084-A08.
- The number of assigned CPs + IFLs + ICFs + zAAPs + additional SAPs + CBUs equals less than 16 on an IBM 2084-B16.
- The number of assigned CPs + IFLs + ICFs + zAAPs + additional SAPs + CBUs equals less than 24 on an IBM 3084-C24.
- The number of assigned CPs + IFLs + ICFs + zAAPs + additional SAPs + CBUs equals less than 32 on an IBM 3084-D32.

Unassigned CPs and IFLs are ignored. In fact, they are considered spares and are available for use as a CBU. When an unassigned CP or IFL is converted into an assigned CP or IFL, or when additional PUs are characterized as CPs or IFLs, then the number of CBUs that can be activated is decreased.

Software model MSU values

All software models have an MSU value that is used to determine the software license charge for all MLC software. Table 2-8 on page 66 shows all MSU values for all software models.

The Mainframe Charter, announced in August 2003, offers lower software charges for selected IBM software on z990. This price/performance improvement is achieved by lowering the original established MSU value for each software model by approximately 10%, resulting in a list of software MSU values (Pricing MSUs) to be used for software charging.

Table 2-8 Software model MSU values

Software model	MSU/ Pricing MSU	Software model	MSU/ Pricing MSU
301	77/70	317	886/799
302	147/132	318	927/837
303	213/191	319	973/878
304	277/248	320	1018/919
305	337/302	321	1062/959
306	395/352	322	1106/999
307	451/402	323	1149/1037
308	503/448	324	1192/1076
309	551/492	325	1234/1114
310	601/538	326	1276/1151
311	647/580	327	1317/1188
312	691/620	328	1358/1225
313	733/661	329	1398/1261
314	772/696	330	1436/1296
315	810/730	331	1474/1332
316	844/761	332	1512/1365

Hardware Management Console and Support Elements

All z990 models include a Hardware Management Console (HMC) and two internal Support Elements (SEs) that are located in the Z frame.

On z990 servers, the Hardware Management Console provides the platform and user interface that can control and monitor the status of the system. The SEs are basically used by IBM service representatives.

Up to four Hardware Management Consoles can be ordered per z990 server, providing more flexibility and additional points of control. The Hardware Management Console can also provide a single point of control and single system image for a number of CPCs configured to it.

The internal SEs for each CPC are attached by local area network (LAN) to the Hardware Management Console, and allow the Hardware Management Console to monitor the CPC by providing status information. Each internal SE provides the Hardware Management Console

with operator controls for its associated CPC, so you can target operations in parallel to multiple or all CPCs, or to a single CPC.

The second SE, called Alternate SE, is standard on all z990 models and serves as a backup to the primary SE. Error detection and automatic switch-over between the two redundant SEs provides enhanced reliability and availability. There are also two fully redundant interfaces, known as the Power Service Control Network (PSCN), between the two SEs and the CPC.

Note: The z990 (and the z890) are the last zSeries servers offering Token Ring adapters on the Hardware Management Consoles and Support Elements. Timely planning is advised in preparation of migration to the Ethernet environment.

Note: Hardware Management Consoles are to become closed platforms with the next zSeries server and will only support the HMC application. Other applications, such as for the IBM ESCON Director and the IBM Sysplex Timer, will no longer be supported from the HMC. Timely planning for needed console equipment for Directors and Timers is recommended. When available, these HMCs can only communicate with Generation 5 servers and later (Multiprise® 3000, G5, G6, z800, z900, z890, z990) and TCP/IP will be the only communication protocol supported.

For further information on the Hardware Management Console and SEs, refer to Appendix A, “Hardware Management Console (HMC)” on page 235.

Dual External Time Reference

The z990 implements a dual External Time Reference (ETR). The optional ETR cards provide the interface to the IBM Sysplex Timers, which are used for timing synchronization between systems in a Sysplex environment.

If z990 models have coupling links, then two ETR cards with dual paths to each book are installed, allowing continued operation even if a single ETR card fails. This redundant design also allows concurrent maintenance.

2.2.8 Storage operations

In z990 servers, memory can be assigned as a combination of central storage and expanded storage, supporting up to 30 logical partitions.

Before you activate a logical partition, central storage (and optional 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 128 GB 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) (refer to “LPAR single storage pool” on page 55 for further details).

Memory *cannot* be shared between system images. You can dynamically reallocate storage resources for z/Architecture and ESA/390 Architecture mode logical partitions running operating systems that support Dynamic Storage Reconfiguration (DSR) (refer to 2.2.11, “LPAR Dynamic Storage Reconfiguration (DSR)” on page 71 for further details).

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’ memories.

Figure 2-18 shows the z990 modes and memory diagram, summarizing all image modes, with their processor types and the Central Storage (CS) and Expanded Storage (ES) definitions allowed for each mode.

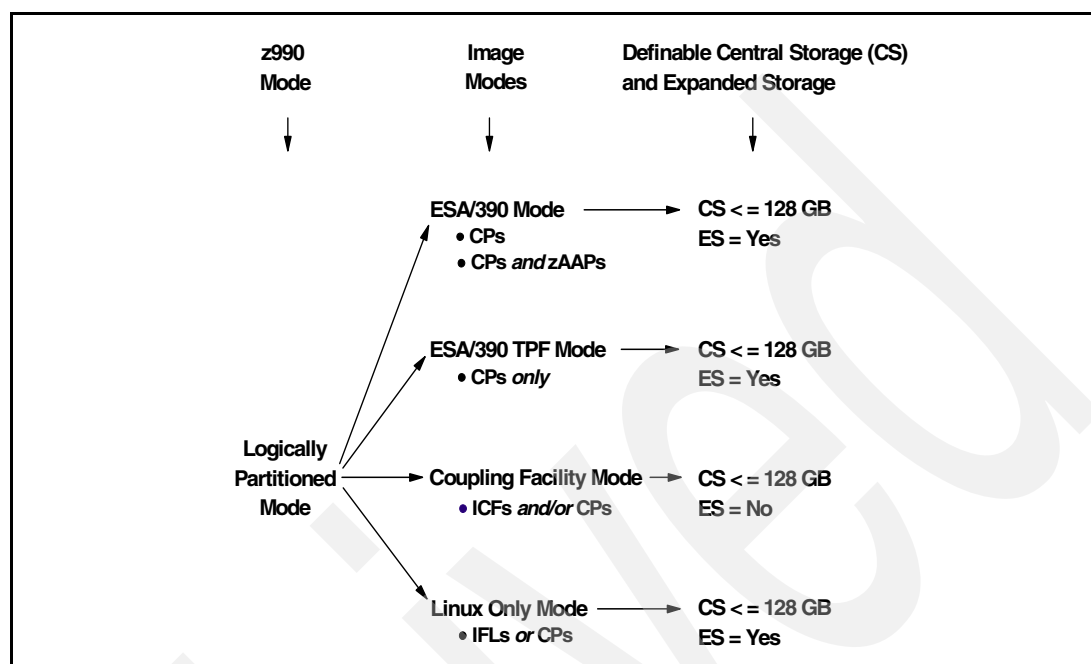


Figure 2-18 Modes and memory diagram

Table 2-9 shows the z990 storage *allocation* and *usage* possibilities, which depend upon the image and architecture modes.

Table 2-9 Storage definition and usage possibilities

Image mode	Architecture mode (addressability)	Maximum central storage		Expanded storage	
		Architecture	z990 definition	z990 definable	Operating system usage
ESA/390	z/Architecture (64-bit)	16 EB	128 GB	yes	<i>only</i> by z/VM
	ESA/390 (31-bit)	2 GB	128 GB	yes	yes
ESA/390 TPF	ESA/390 (31-bit)	2 GB	128 GB	yes	yes
Coupling Facility	CFCC (64-bit)	16 EB	128 GB	no	no
Linux Only	z/Architecture (64-bit)	16 EB	128 GB	yes	<i>only</i> by z/VM
	ESA/390 (31-bit)	2 GB	128 GB	yes	yes

Remember that either a z/Architecture mode or an ESA/390 architecture mode operating system can run in an ESA/390 image mode on a z990. 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.

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, like z/OS, will *always* change this addressing mode and operate in 64-bit mode. The z/OS Bimodal Migration Accommodation Offering allows for a limited amount of time to run z/OS in 31-bit mode. This offering provides fallback support to 31-bit mode in the event it is required during migration to z/OS in 64-bit mode. Beginning with z/OS V1.5, the z/OS Bimodal Migration Accommodation Offering is no longer available.

Other z/Architecture operating systems, like the z/VM and the OS/390 Version 2 Release 10, can be configured to change to 64-bit mode or to stay in 31-bit mode and operate in the ESA/390 architecture mode.

z/Architecture mode

In z/Architecture mode, storage addressing is 64-bit, allowing for an addressing range of up to 16 exabytes (16 EB). The 64-bit architecture allows a maximum of 16 EB to be used as central storage. However, the current z990 definition limit for logical partitions is 128 GB of central storage.

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 (like a z/OS or a Linux for zSeries image) *will 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 an addressing range of up to 2 GB. A maximum of 2 GB can be used for central storage. Since 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, and permits existing application programs to run with existing control programs.

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

ESA/390 TPF mode

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

Coupling Facility mode

In Coupling Facility mode, storage addressing is 64-bit for a Coupling Facility image running CFLEVEL 12 or above, allowing for an addressing range up to 16 EB. However, the current z990 definition limit for logical partitions is 128 GB of storage.

Expanded storage cannot be defined for a Coupling Facility image.

Only IBM Coupling Facility Control Code can run in Coupling Facility mode.

Linux Only mode

In Linux Only mode, storage addressing can be 31- 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:

- ▶ Linux for zSeries uses 64-bit addressing and operates in the z/Architecture mode.
- ▶ Linux for S/390 uses 31-bit addressing and operates in the ESA/390 Architecture mode.
- ▶ z/VM can be configured to use 64-bit addressing and operate in the z/Architecture mode, or to use 31-bit addressing and operate in the ESA/390 architecture mode.

2.2.9 Reserved storage

Reserved storage can optionally be defined to a logical partition allowing a non-disruptive 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: an initial value and a reserved value:

- ▶ 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 just predefined storage sizes to allow a storage increase from the logical partition point of view.

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

- ▶ Partition deactivation
- ▶ An initial storage size definition change
- ▶ 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 CPC memory upgrade

A concurrent logical partition storage upgrade uses Dynamic Storage Reconfiguration (DSR) and the operating system must use the Reconfigurable Storage Units (RSUs) definition to be able to add or remove storage units. Currently, only z/OS and OS/390 operating systems have this support.

2.2.10 LPAR storage granularity

Storage granularity for Central Storage and Expanded Storage in LPAR mode varies as a function of the total installed storage, as shown in Table 2-10 on page 71.

This information is required for Logical Partition Image setup and for z/OS and OS/390 Reconfigurable Storage Units definition.

Table 2-10 LPAR storage granularity

Total installed memory	Partition storage granularity (CS and ES)
Installed memory <= 32 GB	64 MB
32 GB < Installed memory <= 64 GB	128 MB
64 GB < Installed memory <= 128 GB	256 MB
128 GB < Installed memory <= 256 GB	512 MB

Remember that logical partitions are currently limited to a maximum size of 128 GB of storage.

2.2.11 LPAR Dynamic Storage Reconfiguration (DSR)

Dynamic Storage Reconfiguration (DSR) on z990 servers allows an operating system running in a logical partition to add (non-disruptively) 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 enhanced DSR, 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, PR/SM will check if there are any free storage increments and will dynamically bring the storage online.

PR/SM will dynamically take offline a storage increment and will make it available to other partitions when an operating system running in a logical partition releases a storage increment.

2.2.12 I/O subsystem

All models have one I/O subsystem. The I/O subsystem should be considered as the physical entity that encompasses all control functions and all connections to all devices.

The z990 I/O subsystem provides great flexibility and high availability and performance, allowing:

- ▶ High bandwidth

The z990 I/O subsystem can handle up to 96 GBps; this is four times the z900 server's bandwidth. Individual channels can have up to 2 GBps data rates.

- ▶ Wide connectivity

A z990 server can be connected to an extensive range of interfaces, using protocols such as Fibre Channel Protocol (FCP) for Small Computer System Interface (SCSI), Gigabit Ethernet (GbE), Fast Ethernet (FENET), 1000Base-T Ethernet, and High Speed Token Ring, along with FICON, ESCON, and coupling link channels.

- ▶ Concurrent channel upgrades

It is possible to concurrently add channels to a z990 server provided there are unused channel positions in an I/O cage. Additional I/O cages can be previously installed on an initial configuration via Plan Ahead, to provide greater capacity for concurrent upgrades. This capability may help eliminate an outage to upgrade the channel configuration. For more information about concurrent channel upgrades, see "CUoD for I/O" on page 195.

► Dynamic I/O reconfiguration

Dynamic I/O reconfiguration enhances system availability by supporting the dynamic addition, removal, or modification of channel paths, control units, I/O devices, and I/O configuration definitions to both hardware and software (if it has this support), without requiring a planned outage.

► ESCON port sparing and upgrading

The ESCON 16-port I/O card includes one unused port 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, enabling concurrent upgrades via Licensed Internal Code.

For detailed information about the I/O system structure, see Chapter 3, “I/O system structure” on page 73.

2.2.13 Channel Subsystem

The representation of all connections and devices is called the Channel Subsystem. The z990 introduces the concept of Multiple Logical Channel Subsystems. Up to two Logical Channel Subsystems (LCSSs) can be defined in the IOCDS, and this allows for the definition of up to 512 channels. One IOCDS describes the complete I/O configuration and one HSA after Power-On Reset (POR). The HSA is always located in the physical memory of Book 0.

Logical Channel Subsystem (LCSS)

A Logical Channel Subsystem (LCSS) is a logical collection of up to 256 CHPIDs that are mapped to physical channels with the assistance of HCD, the Channel Mapping Tool (CMT), and IOCP. Physical channels are represented in the system by Physical Channel IDs (PCHIDs). The z990 supports up to two LCSSs (512 CHPIDs), but the Multiple CSS architecture allows for more LCSSs.

Physical Channel ID (PCHID)

PCHIDs identify the physical ports on cards located in I/O cages and follow the numbering scheme listed in Table 2-11.

Table 2-11 PCHID locations

Cage	Front PCHID ##	Rear PCHID ##
I/O Cage 1	100 - 1FF	200 - 2FF
I/O Cage 2	300 - 3FF	400 - 4FF
I/O Cage 3	500 - 5FF	600 - 6FF
CEC Cage	000 - 0FF reserved for ICB-4	

Introduction of PCHIDs means that CHPIDs are no longer pre-assigned. It is the responsibility of the user to assign the CHPID numbers through the use of HCD/IOCP, and the CHPID mapping tool. Assigning a CHPID means that the CHPID number is associated with a physical channel port location (PCHID) and an LCSS (LCSS0 or LCSS1). CHPID numbers still range from 00 to FF and must be unique within an LCSS.

For more detailed information about the Logical Channel Subsystem structure, see 4.1.1, “Logical Channel Subsystem structure” on page 110.

I/O system structure

This chapter describes the I/O system structure, the connectivity and the cryptographic options available on the zSeries 990 server.

The z990 server I/O and cryptographic features are also discussed, including configuration options for each feature.

The following topics are included:

- ▶ 3.1, “Overview” on page 74
- ▶ 3.2, “I/O cages” on page 75
- ▶ 3.3, “I/O and cryptographic feature cards” on page 84
- ▶ 3.4, “Connectivity” on page 89

3.1 Overview

The z990 I/O system design provides great flexibility, high availability and performance, allowing:

High bandwidth

The z990 I/O system can handle up to 96 GBps, which is four times the z900 server's bandwidth. Individual channels can have up to 2 GBps and individual Coupling Facility links up to 2 GBps data rates.

Wide connectivity

A z990 server can be connected to an extensive range of interfaces, using protocols such as Fibre Channel Protocol (FCP) for Small Computer System Interface (SCSI), Gigabit Ethernet (GbE), 1000BASE-T Ethernet, 100BASE-T Ethernet, 10BASE-T Ethernet, Token Ring along with FICON Express, ESCON, and coupling links channels.

Cryptographic functions

The z990 I/O system also supports optional cryptographic cards to complement the standard CP Assist for Cryptographic Function (CPACF) that is implemented in every PU, enhancing the performance of cryptographic processing.

Concurrent I/O upgrades

It is possible to concurrently add I/O cards to a z990 server provided there are unused slot positions in an I/O cage. Additional I/O cages can be previously installed on an initial configuration, via CUoD, to provide greater capacity for concurrent upgrades. This capability may help eliminate an outage to upgrade the I/O configuration. See more information about concurrent upgrades on Chapter 8, "Capacity upgrades" on page 187.

Dynamic I/O configuration

Dynamic I/O configuration enhances system availability by supporting the dynamic addition, removal, or modification of channel paths, control units, I/O devices, and I/O configuration definitions to both hardware and software (if it has this support) without requiring a planned outage.

ESCON port sparing and upgrading

The ESCON 16-port I/O card includes one unused port 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, enabling concurrent upgrades via Licensed Internal Code (LIC).

The following I/O feature *ports* are supported in the zSeries 990 server:

- ▶ Up to 1024 ESCON (up to 720 ESCON on model A08)
- ▶ Up to 120 Fibre Connection (FICON) Express (up to 96 FICON on model A08)
- ▶ Up to 48 Open Systems Adapter (OSA) Express
- ▶ Up to 16 Integrated Cluster Bus-4 (ICB-4) (up to 12 ICB-4 on model A08)
- ▶ Up to 16 Integrated Cluster Bus-3 (ICB-3)
- ▶ Up to eight Integrated Cluster Bus-2 (ICB-2)
- ▶ Up to 48 Inter-System Channel-3 (ISC-3) in peer mode (up to 32 ISC-3 in compatibility mode)
- ▶ Up to two External Time Reference (ETR)

Note: The maximum number of Coupling Links combined (IC, ICB-2, ICB-3, ICB-4, and active ISC-3 links) cannot exceed 64 per z990 server.

The following cryptographic feature cards are supported in the zSeries 990 server:

- ▶ Up to four Peripheral Component Interconnect X Cryptographic Coprocessor (PCIXCC)
- ▶ Up to 12 Peripheral Component Interconnect Cryptographic Accelerator (PCICA)

All z990 servers have two frames. The A frame holds the CEC cage on top and one I/O cage on the bottom. The Z frame holds up to two optional I/O cages, which may be needed to accommodate the I/O configuration requirements.

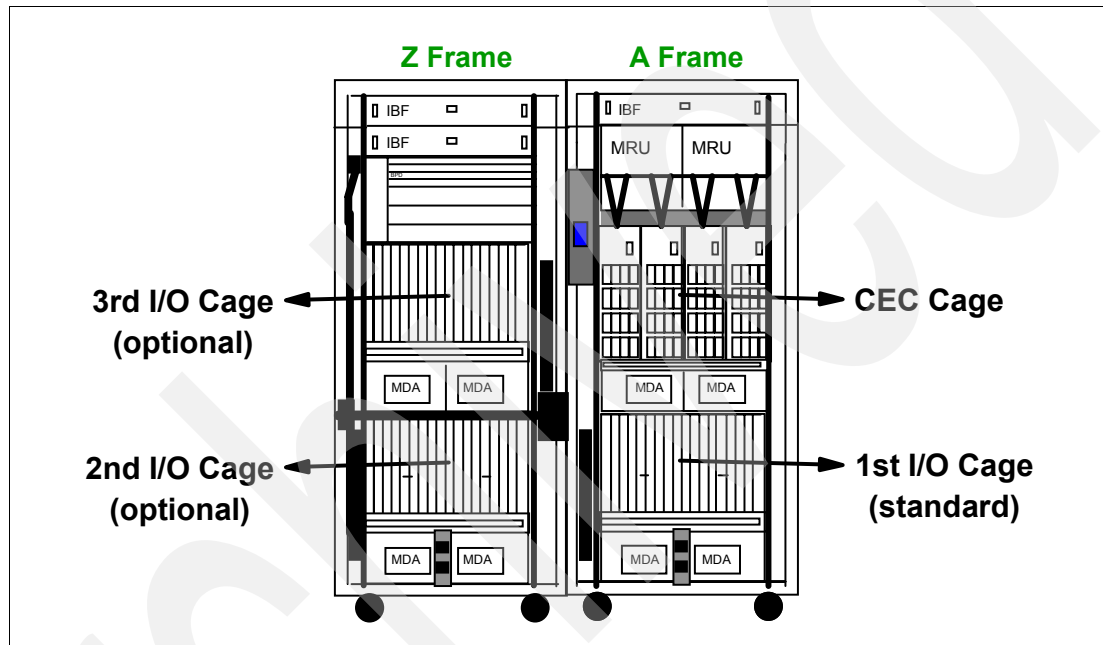


Figure 3-1 z990 frames and cages

Additional optional I/O cages may be required to install additional I/O and cryptographic cards during an upgrade. The first optional I/O cage is placed at the bottom of the Z frame, and the second optional I/O cage is at the top, as shown in Figure 3-1. Although I/O or cryptographic card installation is concurrent, an I/O cage installation requires an outage.

3.2 I/O cages

As mentioned, the z990 server can have up to three I/O cages to house the I/O cards and cryptographic cards required by a configuration.

Each I/O cage has 28 I/O slots available for I/O cards and cryptographic cards installation and up to seven I/O domains. Each I/O domain is made up of up to four I/O slots, as shown in Figure 3-2 on page 76.

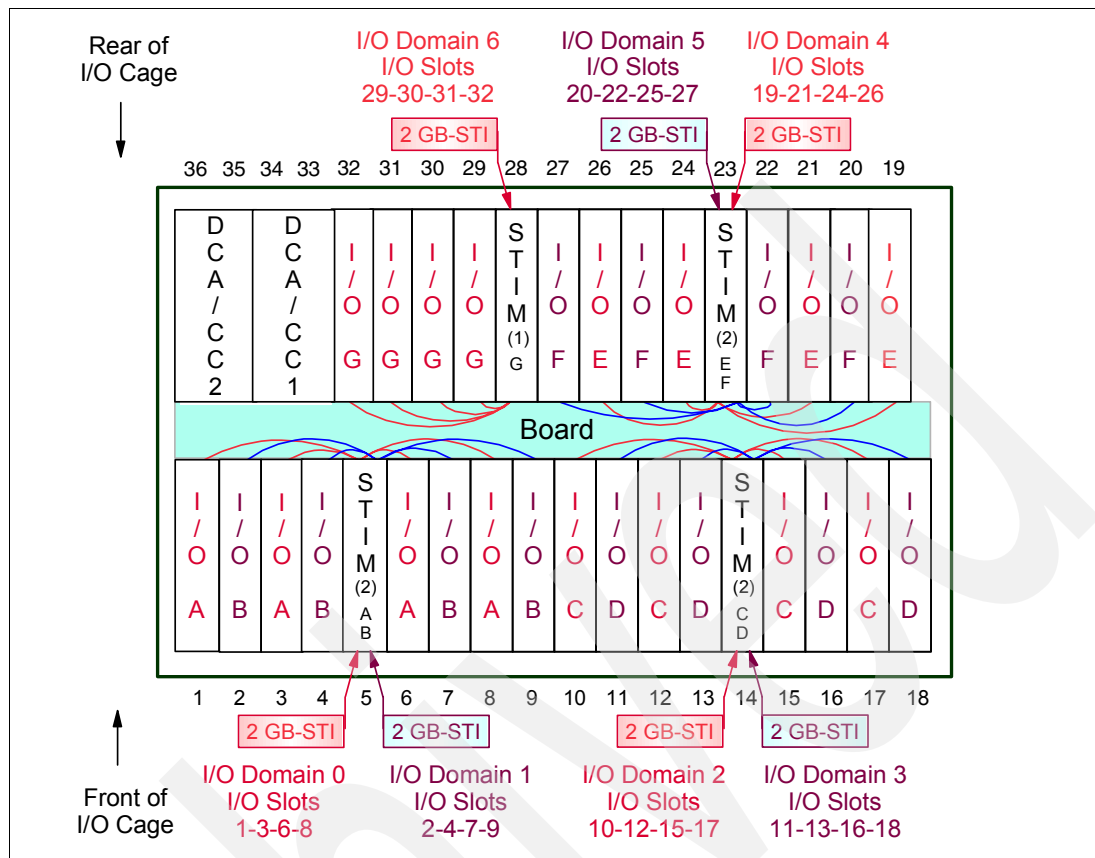


Figure 3-2 z990 I/O cage

Each I/O domain requires one Self-Timed Interconnect Multiplexer (eSTI-M) card. All I/O cards within an I/O domain are connected to its eSTI-M card via the back plane board. A full I/O cage requires seven eSTI-M cards, which are half-high cards, using three and a half slots.

In addition, two Distributed Converter Assembly-Cage Controller (DCA-CC) cards plug into the I/O cage and the “other half” of slot 28 may be used for a Power Sequence Control (PSC) card.

If one I/O domain is fully populated with ESCON cards (15 available ports and one spare per card), up to 60 (four cards x 15 ports) ESCON channels can be installed and used. A fully populated I/O cage with ESCON cards can have up to 420 (60 x 7 domains) ESCON channels.

Table 3-1 lists the I/O domain-to-I/O slot relationships within an I/O cage.

Table 3-1 I/O domain-to-I/O slot relationships

Domain	I/O slots in domain
0	01, 03, 06, 08
1	02, 04, 07, 09
2	10, 12, 15, 17
3	11, 13, 16, 18
4	19, 21, 24, 26

Domain	I/O slots in domain
5	20, 22, 25, 27
6	29, 30, 31, 32

Each eSTI-M card is connected to an STI jack located in a book's Memory Bus Adapter (MBA) via an STI cable. As each eSTI-M card requires one STI, up to seven STIs are required to support one I/O cage. A fully populated three-I/O cage system requires 21 STIs.

IBM selects which slots are used for I/O cards and supplies the appropriate number of I/O cages and STI cables, either for a new build server or for an existing server upgrade.

Important: Installing an additional I/O cage to an existing z990 server configuration is disruptive. The Plan Ahead process allows you to avoid this outage by including, in the initial z990 server order, the number of optional I/O cages required by a future I/O configuration.

3.2.1 Self-Timed Interconnect (STI)

There are three Memory Bus Adapters (MBAs) on each z990 book. Each MBA has four Self-Timed Interconnects (STIs), resulting in a total of 12 STIs on each z990 book. Each STI has a bandwidth of 2 GBps full-duplex, resulting in a maximum bandwidth of 24 GBps per z990 book.

Depending on the number of books in the configuration, there will be 12, 24, 36, or 48 STIs in a z990 server, as shown in Table 3-2.

Table 3-2 Number of MBAs and STIs

z990 Model	Number of books	Number of MBAs	Number of STIs
2084-A08	1	3	12
2084-B16	2	6	24
2084-C24	3	9	36
2084-D32	4	12	48

The z990 model D32 has a maximum bandwidth of 96 GBps.

3.2.2 STIs and I/O cage connections

Figure 3-3 on page 78 shows the STI connections from the server's CEC cage to an I/O cage, and to an Integrated Cluster Bus-4 (ICB-4) link.

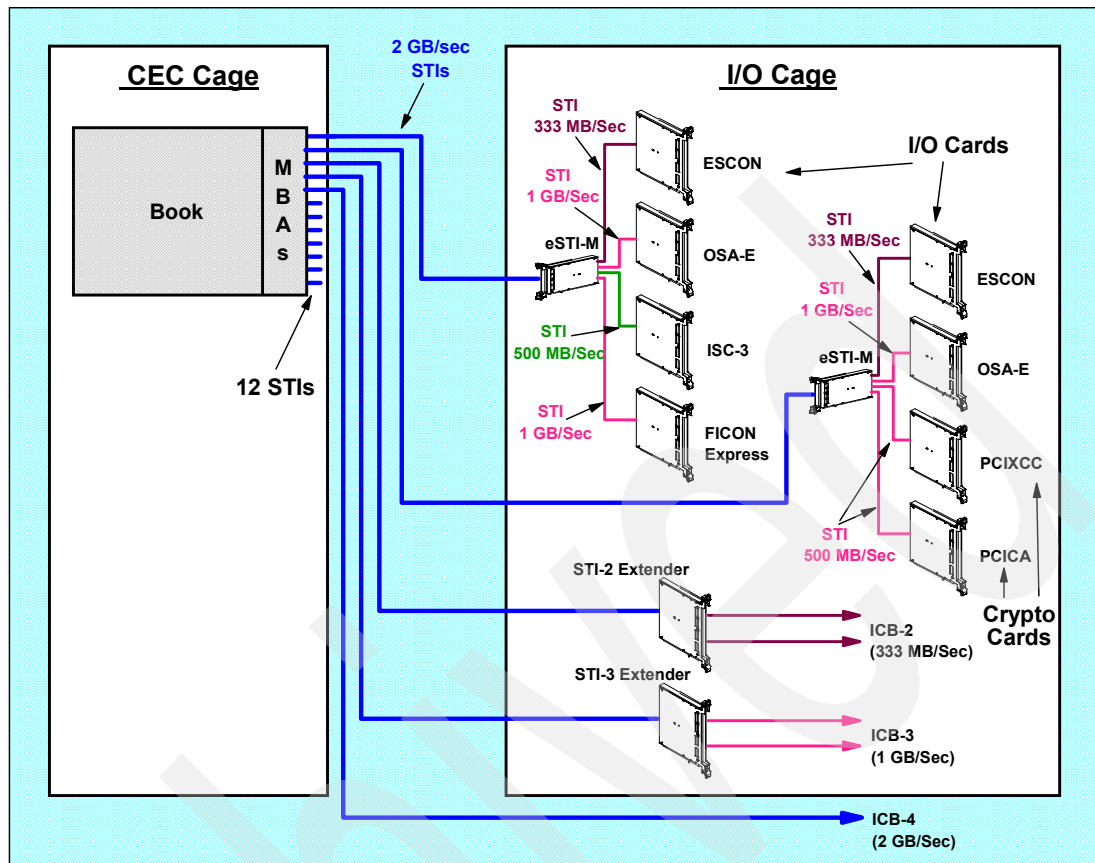


Figure 3-3 STIs and I/O cage connections

A Memory Bus Adapter (MBA) STI connector, located in a book, can be connected to one of the following:

- ▶ An eSTI-M card, which creates up to four secondary STI links to connect I/O cards
- ▶ An STI-2 Extender card, which has up to two ICB-2 links
- ▶ An STI-3 Extender card, which has up to two ICB-3 links
- ▶ An ICB-4 link, which attaches directly to an STI port

eSTI-M card

For each z990 I/O cage domain, the MBA-to-I/O card connectivity is achieved using an eSTI-M card and a 2 GBps STI cable. These half-high eSTI-M cards plug into specific slots (5, 14, 23, and 28) in the z990 I/O cage. Physical slot locations 5, 14, and 23 house two half-high eSTI-M cards, while slot 28 has only one half-high card plugged in the top.

The eSTI-M card (Feature Code 0322) takes the 2 GBps link from an MBA's STI and creates four secondary STI links, which are connected to the I/O and cryptographic cards through the I/O cage board. The bandwidth of the secondary link is determined by the feature card it is attached to:

- ▶ 333 MBps for ESCON
- ▶ 500 MBps for ISC-3
- ▶ 1 Gbps for FICON, OSA-E, PCIXCC and PCICA

Depending on the number of I/O slots plugged into the cage, there may be from one to seven eSTI-M cards plugged into a z990 I/O cage. The eSTI-M card can be installed or replaced concurrently.

STI-2 Extender card

The STI-2 Extender card (Feature Code 3992) takes the 2 GBps link from an MBA's STI and creates two secondary 333 MBps STI links, which are used to connect ICB-2 links. ICB-2 is supported only for connection to G5/G6 servers.

The number of STI-2 Extender cards depends on the number of ICB-2 links in a configuration. Usually, the number of STI-2 Extender cards is half the number of ICB-2 links, but for availability reasons, two ICB-2 links are connected to two STI-2 Extender cards, each one having one active ICB-2 link port.

The maximum number of STI-2 Extender cards in a z990 server is four cards, resulting in up to eight ICB-2 ports. All of them can be installed in a single I/O cage. The STI-2 Extender card can be installed or replaced concurrently.

STI-3 Extender card

The STI-3 Extender card (Feature Code 3993) takes the 2 GBps link from an MBA's STI and creates two secondary 1 GBps STI links, which are used to connect ICB-3 links.

The number of STI-3 Extender cards depends on the number of ICB-3 links in a configuration. Usually, the number of STI-3 Extender cards is half the number of ICB-3 links, but for availability reasons, two ICB-3 links are connected to two STI-3 Extender cards, each one having one active ICB-3 link port.

The maximum number of STI-3 Extender cards in a z990 server is eight cards, resulting in up to 16 ICB-3 ports. All of them can be installed in a single I/O cage. The STI-3 Extender card can be installed or replaced concurrently.

3.2.3 Balancing I/O connections

The z990 server's multi-book structure results in multiple MBAs; therefore, there are multiple STI sets. This means that an I/O distribution over books, MBAs, STIs, I/O cages, and I/O cards is desirable for both performance and availability purposes.

The STI links balancing across a book's MBAs, I/O cages, and I/O cards is done by IBM at the server's initial configuration time. Follow-on upgrades of the initial server configuration, including additional book(s) and/or I/O cage(s), may undo the balance of the original STI links distribution.

The optional upgrade feature STI Rebalance (Feature Code 2400) can be requested at upgrade configuration time to rebalance STI links across the new total number of books and I/O cages. However, STI rebalancing is disruptive, requiring a server outage.

The processor I/O ports balancing across I/O cards, I/O cages, STI links, and a book's MBAs is done by the customer at I/O definition time. This is done by either the use of the CHPID Mapping Tool to assign CHPIDs to PCHIDs, or manually by assigning installed PCHIDs to CHPIDs. The use of the CHPID Mapping tool is strongly recommended.

The balancing may also be affected by the STI Rebalance feature (FC 2400) after a server upgrade.

STI links balancing across books and MBAs

Figure 3-4 shows a 2084-B16 server's initial configuration example with two fully populated I/O cages (seven I/O domains on each one).

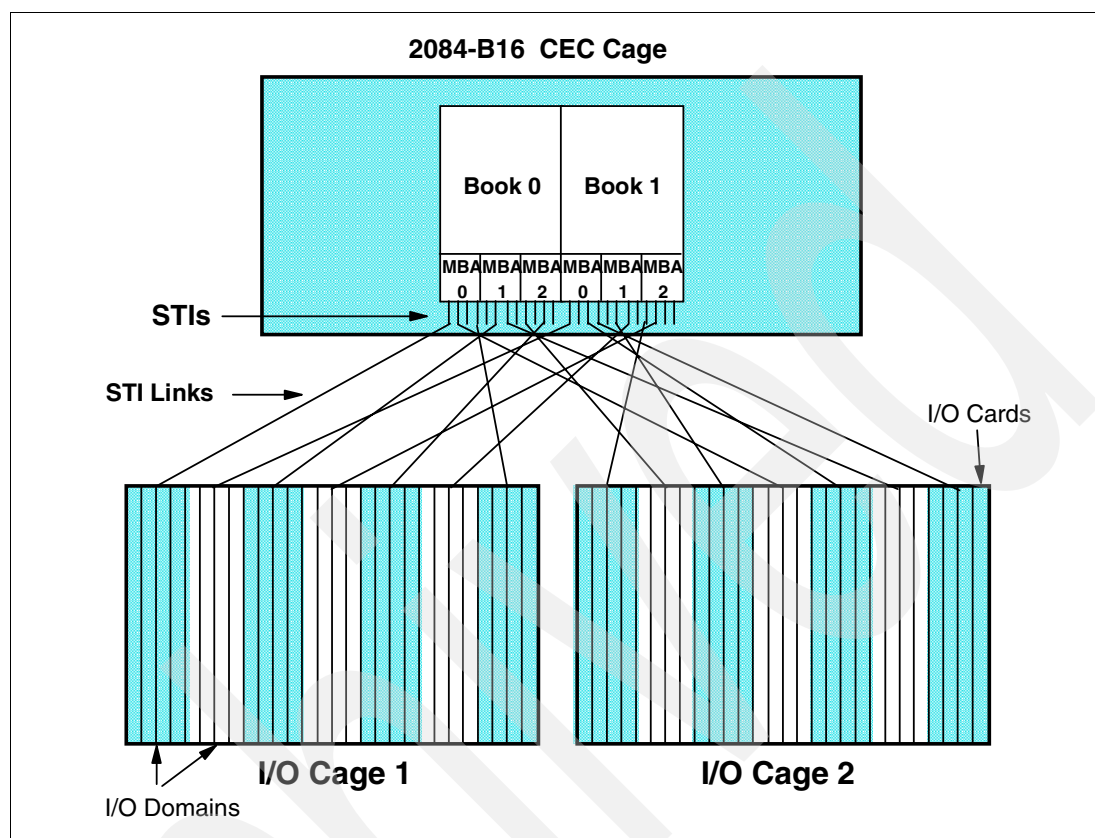


Figure 3-4 2084-B16 initial configuration example

The 2084-B16 server has two books in the CEC cage. The STI links are distributed across books, MBAs, and I/O cages, as a result of the initial server configuration balancing.

Nearly the same number of STIs of each book's MBA are used and spread across the two I/O cages, resulting in the best STI link distribution for both performance and availability.

Figure 3-5 on page 81 shows an upgrade from this 2084-B16 server to a model D32, maintaining the same I/O cages and I/O cards.

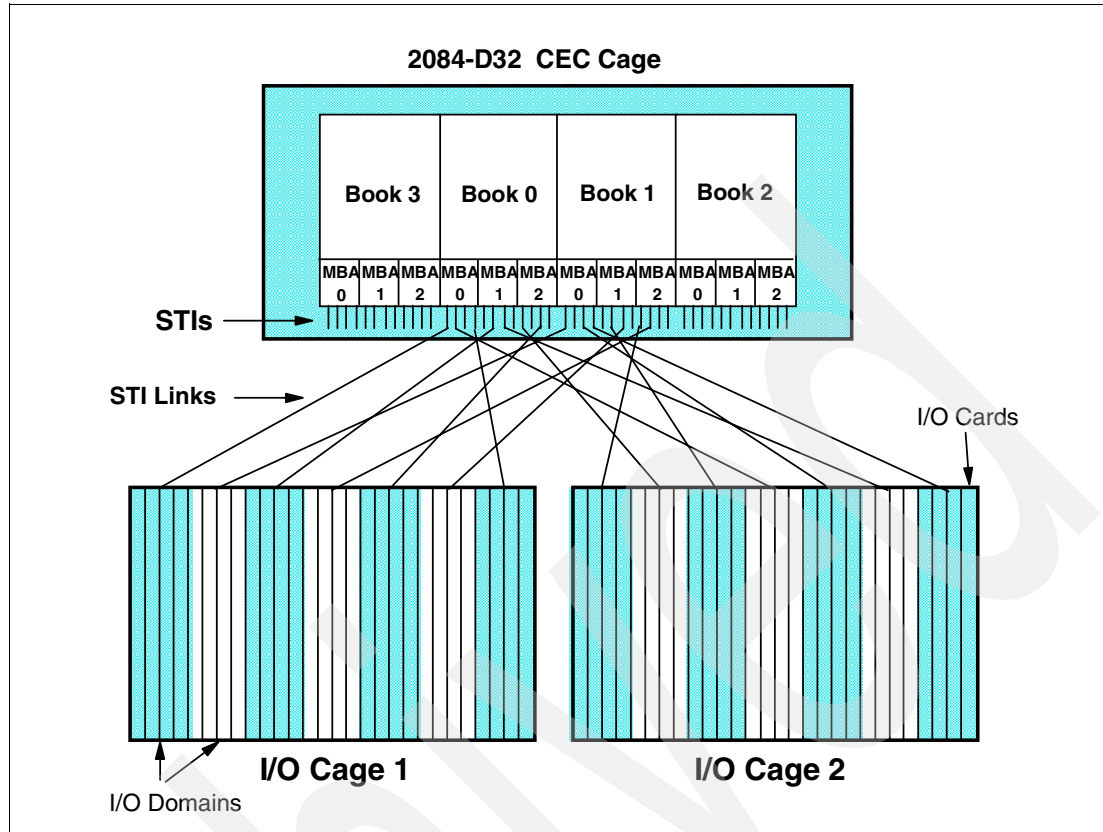


Figure 3-5 2084-B16-to-D32 upgrade example

This upgrade adds, concurrently, two more books in the CEC cage, and the standard upgrade configuration will *not* change the STI links' original distribution and connections. The new books 2 and 3 do not have any STI connection and all STI links remain in the original books 0 and 1, resulting in unbalanced STI connections across books.

To optimize Reliability, Availability, and Serviceability (RAS) characteristics of the server, the STI Rebalance feature (Feature Code 2400) can be ordered on server upgrades, including additional books.

STI Rebalance feature (Feature Code 2400)

Figure 3-6 on page 82 shows the previous server upgrade example, from a 2084-B16 to a model D32, with the STI Rebalance feature (FC 2400) selected for the upgrade of the configuration.

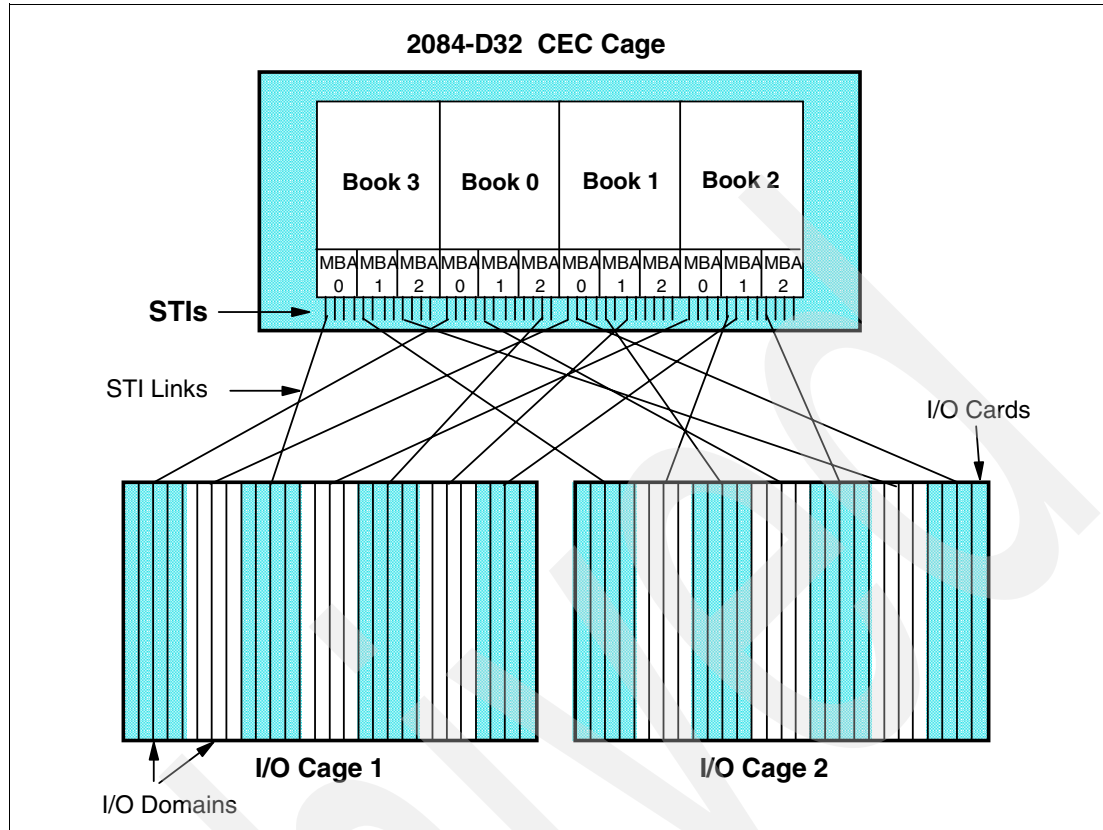


Figure 3-6 Upgrade example with the STI Rebalance feature (FC 2400)

Now you can see that the required number of STI links is spread across all books' MBAs, including the two newly installed books 2 and 3, and redistributed across the existing I/O cages. The result is a balanced STI system, as if a new build 2084-D32 server was initially configured.

On the other hand, an upgrade *including* the STI Rebalance feature (FC 2400) is disruptive, as STI cables must be reconnected to other STI locations, affecting the corresponding I/O domains. After the Power-on Reset, SAPs are assigned to channel cards using the new STI links configuration.

Important: If the z990 STI Rebalance feature (FC 2400) is selected at server upgrade configuration time, and effectively results in STI rebalancing, the server upgrade will be disruptive.

The z990 STI Rebalance feature may also change the Physical Channel ID (PCHID) number of ICB-4 links (see 3.3.3, "Physical Channel IDs (PCHIDs)" on page 86), requiring a corresponding update on the server's I/O definition via HCD/HCM.

Adding a book via MES will result in STI plugging that is different from new build STI plugging with the same number of books. FC 2400 can be ordered to replug the STIs as new builds. The concurrent addition of a single book is supported, but be aware that regardless of how the customer planned the previous configuration, the CHPID Mapping Tool (CMT) can be used to evaluate the effects of FC2400 on the current configuration.

- If you take the current IOCP statements and the current CFReport (provided by the IBM Account Representative) and input these via the availability option in the CMT, it will be

possible to see any places where a control unit, or group of control units, have single points of failure (SPOF); in this case, books and MBAs are of interest.

- For the next step, use the CFReport for FC2400 along with the same IOCP statements and repeat the availability option in the CMT. This will potentially show a different set of SPOFs.

By comparing the two reports, you can determine if FC2400 is the right choice and what, if any, other configuration changes will need to be made in conjunction with the install of FC2400.

I/O port balancing across MBAs and books

At I/O definition time, the customer is able to select I/O ports for different paths of a multi-path control unit that come from different I/O cards, different I/O domains (including different eSTI-M cards and different STI links), different I/O cages, and different MBAs from different books. This improves I/O throughput and system availability by avoiding single point of failure paths.

Figure 3-7 shows a simplified example of multi-path device connectivity.

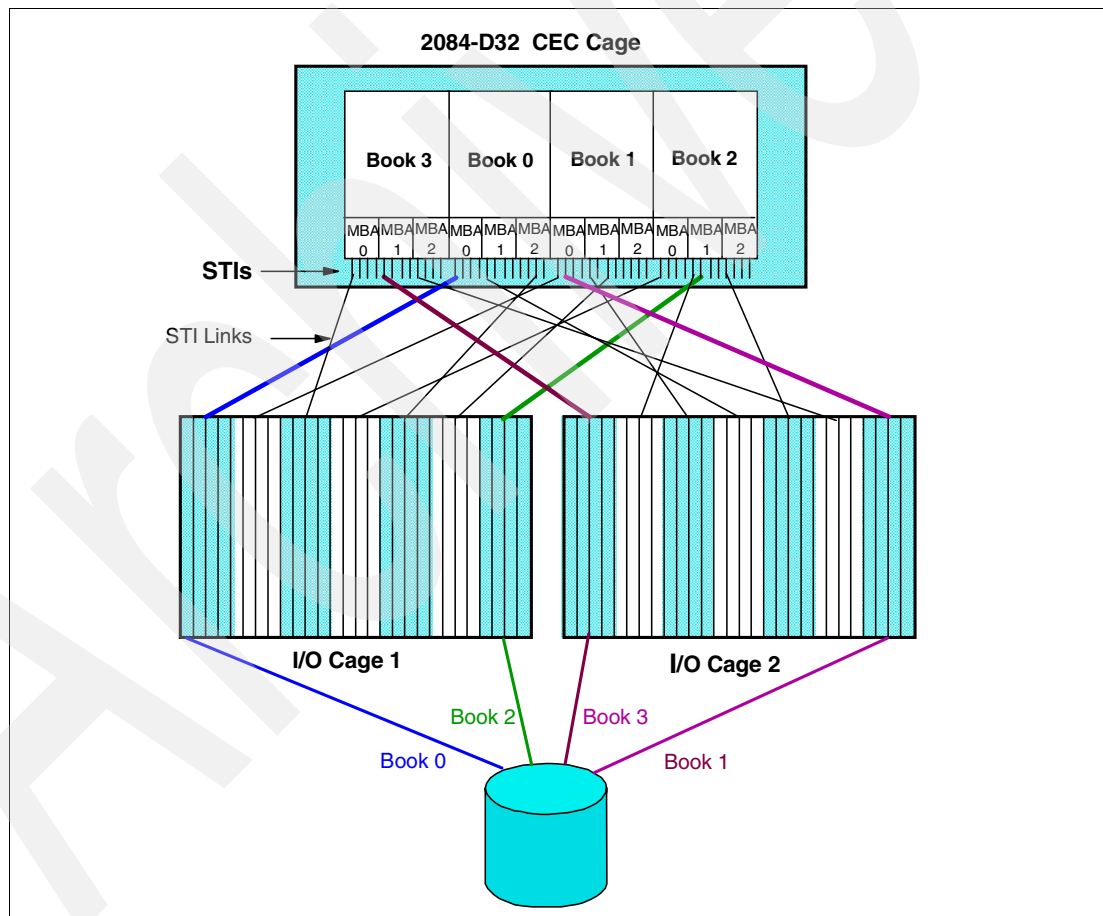


Figure 3-7 Balancing multi-path device connectivity example

Of course, this example assumes that there are enough I/O cards available for such connectivity distribution, and this may not be true for all channel types on a given real configuration. However, the overall goal is to avoid, as much as possible, connectivity single points of failures.

The z990 CHPID Mapping Tool (CMT) can help you plan for the best I/O port selection for high availability purposes. For more information about the z990 CMT, see “IBM z990 CHPID Mapping Tool (CMT)” on page 116.

3.3 I/O and cryptographic feature cards

I/O cards have the I/O port(s) to connect the z990 server to external devices, networks, or to other servers. I/O cards are plugged into I/O slots in an I/O cage, and their specific locations are based on z990 configuration rules. There are different types of I/O cards, one for each channel or link type. I/O cards can be installed or replaced concurrently.

Optional cryptographic cards are also plugged into an I/O slot in an I/O cage, and have coprocessors and accelerator cards for cryptographic functions. There are two different types of cryptographic cards, and they can be installed or replaced concurrently.

3.3.1 I/O feature cards

Table 3-3 gives a summary of all I/O feature cards that are supported on z990 servers.

Table 3-3 I/O feature cards

I/O card types	Feature Codes (FC)
ESCON	2323
FICON Express LX	2319
FICON Express SX	2320
OSA-E GbE LX	1364 2364 (*)
OSA-E GbE SX	1365 2365 (*)
OSA-E 1000BASE-T Ethernet	1366
OSA-E Fast Ethernet	2366 (*)
OSA-E Token Ring	2367
ISC-3	0218 (ISC-D), 0217 (ISC-M)
ISC-3 up to 20 Km	RPQ 8P2197 (ISC-D)
ETR	6154

(*) OSA-E Feature Codes 2364, 2365, and 2366 are brought forward on an upgrade only.

I/O feature cards no longer supported

The following I/O feature cards are no longer supported on z990 servers:

- Parallel channel cards (z900's FC 2304)

Parallel channel cards are not offered as a new build option and are not offered on an upgrade from z900. Parallel control units can be connected to ESCON channels of the z990 server through the following ESCON Converters:

- IBM 9034 (which has been withdrawn from marketing)

- Optica Technologies 34600 FXBT ESCON Converter. For more information, check the Optica Technologies Web site:

<http://www.opticatech.com/34600.asp>

- ▶ ESCON 4-port channel cards (z900 FC 2313)

ESCON 4-port channel cards are not offered as a new build option and are replaced with new 16-Port ESCON cards (FC 2323) during an upgrade from z900.

The 16-Port ESCON card has MT-RJ connectors.

- ▶ FICON channel cards (pre-FICON Express) (z900 FC 2315 and FC 2318)

FICON channel cards (FC 2315 and FC 2318), the original pre-FICON Express cards, are not offered as a new build option and are replaced with new FICON Express cards (FC 2319 or FC 2320) during an upgrade from z900.

The FICON Express cards have LC Duplex connectors.

- ▶ OSA-2 adapter cards (z900 FC 5201 and FC 5202)

The OSA-2 Token Ring (z900's FC 5201) and OSA-2 Fiber Distributed Data Interface (FDDI) (z900's FC 5202) features are not offered as a new build option and are not offered on an upgrade from z900.

For Token Ring connectivity, use the equivalent OSA-Express adapter.

If FDDI connectivity is still desired, a multiprotocol switch or router with the appropriate network interface (for example, 1000BASE-T Ethernet, Gigabit Ethernet) can be used to provide connectivity between the z990 server and a FDDI LAN, via an OSA-Express adapter.

- ▶ OSA-Express ATM adapters (z900's FC 2362 and FC 2363)

The OSA-Express Asynchronous Transfer Mode (ATM) features are not offered as a new build option and are not offered on an upgrade from z900.

If ATM connectivity is still desired, a multiprotocol switch or router with the appropriate network interface (for example, 1000BASE-T Ethernet, Gigabit Ethernet) can be used to provide connectivity between the z990 server and an ATM LAN, via an OSA-Express adapter.

3.3.2 Cryptographic feature cards

Table 3-4 gives a summary of all cryptographic feature cards that are supported on z990 servers.

Table 3-4 Cryptographic feature cards

Cryptographic card types	Feature Codes (FC)
PCIXCC	0868
PCICA	0862

Cryptographic feature card no longer supported

The following cryptographic feature card is no longer supported on z990 servers:

- ▶ PCI Cryptographic Coprocessor (PCICC) (z900's FC 0861)

The PCI Cryptographic Coprocessor (PCICC) (FC 0861) is replaced with the PCIX Cryptographic Coprocessor (PCIXCC) (FC 0868) and the CMOS Cryptographic Coprocessor Facility that were offered on z900. In addition, functions from the

Cryptographic Coprocessor Facility used by known applications have also been implemented in the PCIXCC feature.

3.3.3 Physical Channel IDs (PCHIDs)

A Physical Channel ID (PCHID) is the number assigned to a port of an I/O or cryptographic card. Each enabled port has its own PCHID number, which is based on its I/O slot location in the I/O cage (except for ESCON sparing).

In the case of an ICB-4 link, its PCHID number is based on its CEC cage location.

Figure 3-8 shows the rear view of the first I/O cage (bottom of the A frame), including some I/O cards in slots 01 to 05, and the PCHID numbers of each port.

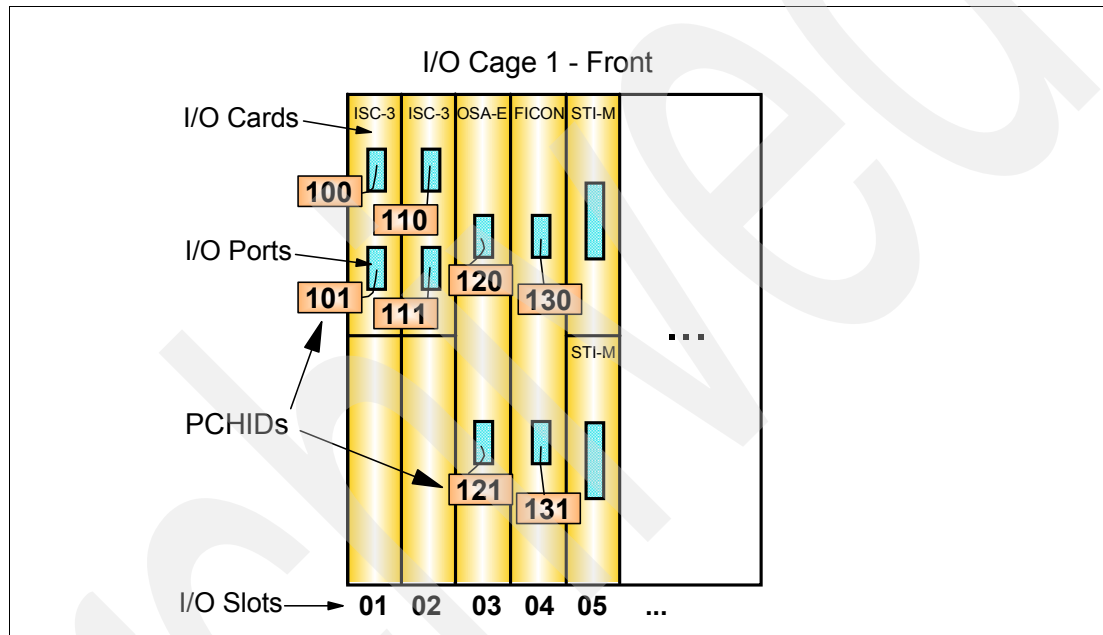


Figure 3-8 Physical Channel IDs (PCHIDs)

Figure 3-9 on page 87 contains the corresponding PCHID Report of the configuration example shown in Figure 3-8.

Machine: 2084-A08 NEW1						
Book/Jack/MBA	Cage	Slot	F/C	PCHID/Ports		Comment
0/J.00/0	A01B	D101	0218	100/J00	101/J01	
0/J.08/2	A01B	D102	0218	110/J00	111/J01	
0/J.00/0	A01B	03	1364	120/J00	121/J01	
0/J.08/2	A01B	04	2319	130/J00	131/J01	
Legend:						
A19B	Top of A frame					
A01B	Bottom of A frame					
D1xx	Half high card in top of slot xx					
0218	ISC D <10KM					
1364	OSA Express GbE LX					
2319	FICON Express LX					

Figure 3-9 PCHID Report example

I/O slot 01 has an ISC-3 Daughter (ISC-D) half-high card (FC 0218) in the top, connected to STI 0 (Jack J.00) from MBA 0 of book 0. Its two enabled ports have PCHID numbers 100 and 101.

I/O slot 04 has a FICON Express LX card (FC 2319), connected to STI 8 (Jack J.08) from MBA 2 of book 0, and its two ports have PCHID numbers 130 and 131.

The pre-assigned PCHID number of each I/O port relates directly to its physical location (jack location in a specific slot), except for ESCON sparing; refer to Figure 3-11 on page 96 for an ESCON sparing example.

Table 3-5 shows the PCHID numbers range for each I/O slot of each I/O cage.

Table 3-5 PCHID numbers and locations

I/O cage slot	PCHID numbers		
	1st I/O cage	2nd I/O cage	3rd I/O cage
01 (front)	100 - 10F	300 - 30F	500 - 50F
02 (front)	110 - 11F	310 - 31F	510 - 51F
03 (front)	120 - 12F	320 - 32F	520 - 52F
04 (front)	130 - 13F	330 - 33F	530 - 53F
06 (front)	140 - 14F	340 - 34F	540 - 54F
07 (front)	150 - 15F	350 - 35F	550 - 55F
08 (front)	160 - 16F	360 - 36F	560 - 56F
09 (front)	170 - 17F	370 - 37F	570 - 57F
10 (front)	180 - 18F	380 - 38F	580 - 58F
11 (front)	190 - 19F	390 - 39F	590 - 59F

I/O cage slot	PCHID numbers		
	1st I/O cage	2nd I/O cage	3rd I/O cage
12 (front)	1A0 - 1AF	3A0 - 3AF	5A0 - 5AF
13 (front)	1B0 - 1BF	3B0 - 3BF	5B0 - 5BF
15 (front)	1C0 - 1CF	3C0 - 3CF	5C0 - 5CF
16 (front)	1D0 - 1DF	3D0 - 3DF	5D0 - 5DF
17 (front)	1E0 - 1EF	3E0 - 3EF	5E0 - 5EF
18 (front)	1F0 - 1FF	3F0 - 3FF	5F0 - 5FF
19 (rear)	200 - 20F	400 - 40F	600 - 60F
20 (rear)	210 - 21F	410 - 41F	610 - 61F
21 (rear)	220 - 22F	420 - 42F	620 - 62F
22 (rear)	230 - 23F	430 - 43F	630 - 63F
24 (rear)	240 - 24F	440 - 44F	640 - 64F
25 (rear)	250 - 25F	450 - 45F	650 - 65F
26 (rear)	260 - 26F	460 - 46F	660 - 66F
27 (rear)	270 - 27F	470 - 47F	670 - 67F
29 (rear)	280 - 28F	480 - 48F	680 - 68F
30 (rear)	290 - 29F	490 - 49F	690 - 69F
31 (rear)	2A0 - 2AF	4A0 - 4AF	6A0 - 6AF
32 (rear)	2B0 - 2BF	4B0 - 4BF	6B0 - 6BF

Note that I/O cage slot numbers 05, 14, 23, and 28 are reserved for eSTI-M cards.

The PCHID number range from 000 to 0FF is reserved for ICB-4 links. As ICB-4 links are directly connected to a book, Table 3-6 shows the ICB-4 PCHID numbers range for each book.

Table 3-6 PCHID numbers for ICB-4 links

CEC cage book	PCHID numbers
0	010 - 01B
1	020 - 02B
2	030 - 03B
3	000 - 00B

Important: If the STI Rebalance feature (Feature Code 2400) is selected on a z990 server upgrade, the current ICB-4 PCHID numbers may change. This requires the corresponding update of the ICB-4 link definition in the z990 server I/O configuration.

The server's PCHID Report has all the installed PCHID numbers. At definition time, PCHIDs are assigned to Channel Path IDs (CHPIDs) using the CHPID Mapping Tool, or HCD/HCM, or

IOCP. The CHPID assignment associates the CHPID number to a physical channel port location (PCHID).

HiperSockets (IQD) and IC links (ICP) do not have PCHIDs, as they are virtual and not physical links, but they do require CHPID numbers.

The PCIX Cryptographic Coprocessor (PCIXCC) and PCI Cryptographic Accelerator (PCICA) features do not require CHPID numbers but are assigned PCHIDs.

z990 CHPID Mapping Tool (CMT)

The z990 CHPID Mapping Tool (CMT) is available, and it is highly recommended for PCHID-to-CHPID assignments. For complex configurations and configurations using multiple Logical Channel Subsystems, using the CMT is strongly recommended.

CMT has a manual mapping and availability mapping function, which does the PCHID-to-CHPID assignments for the best availability. For more information about the z990 CMT, see “IBM z990 CHPID Mapping Tool (CMT)” on page 116.

3.4 Connectivity

Input/output (I/O) channels are components of the z990 server Channel Subsystem (CSS). They provide a pipeline through which data is exchanged between processors, or between a processor and external devices or networks. The most common type of device attached to a channel is a control unit (CU). The CU controls I/O devices such as disk and tape drives.

Server-to-server communications are most commonly implemented using Inter-System Channels (ISC-3), Integrated Cluster Bus (ICB4, ICB-3 or ICB-2) links, or channel-to-channel (CTC) connections.

Local Area Network (LAN) connectivity can be done by the OSA Express cards. There are specific OSA Express cards to support Gigabit Ethernet (GbE), 1000BASE-T Ethernet, 100BASE-T Ethernet, 10BASE-T Ethernet, and Token Ring networks. For additional, detailed information about z990 connectivity, see the IBM Redbook *IBM @server zSeries Connectivity Handbook*, SG24-5444.

3.4.1 I/O and cryptographic features support and configuration rules

Table 3-7 on page 90 summarizes all available I/O and cryptographic features on z990 servers, the maximum number of ports for each type, the number of I/O slots required to achieve this number, and the port increments.

The I/O configuration rules are also provided, including the configuration requirements and limitations.

Table 3-7 I/O and cryptographic features support

I/O feature	Feature codes	Number of		Max. number of		PCHID	CHPID definition	Config. rules notes
		Ports per card	Ports increments	Ports	I/O slots			
ESCON	2323 2324 (ports)	16 (one spare)	4 (LIC-CC)	1024	69	yes	CNC, CVC, CTC, CBY	1, 2
FICON Express LX/SX	2319/2320	2	2	120	60	yes	FC, FCV, FCP	3, 4
OSA-E Gbit Ethernet LX/SX	1364/1365 (2364/2365)	2	2	48 (24)	24 (12)	yes	OSD	4, 5, 6
OSA-E 1000BASE-T Ethernet	1366	2	2	48	24	yes	OSE, OSD, OSC	4, 5
OSA-E Fast Ethernet	2366	2	2	24	12	yes	OSE, OSD	4, 5, 6
OSA-E Token Ring	2367	2	2	48	24	yes	OSE, OSD	4, 5
ICB-2 (333 MBps)	0992	2	1	8	4	yes	CBS, CBR	7
ICB-3 (1 GBps)	0993	2	1	16	8	yes	CBP	7
ICB-4 (2.0 GBps)	3393	-	1	16	0	yes	CBP	7, 8
ISC-3 at 10km (1 or 2 Gbps)	0217 (ISC-M) 0218 (ISC-D) 0219 (ports)	4/ISC-M 2/ISC-D	1 (LIC-CC)	48	12	yes	CFP CFS, CFR	7, 9, 10
ISC-3 20km support (1 Gbps)	RPQ 8P2197 (ISC-D)	4/ISC-M 2/ISC-D	2	48	12	yes	CFP CFS, CFR	7, 9, 10
HiperSockets	-	-	1	16	0	no	IQD	11
IC	-	-	2	32	0	no	ICP	7, 11
ETR	6154	1	-	2	-	no	-	12
PCIXCC	0868	1	1	4	4	yes	-	4, 13, 14
PCICA	0862	2	2	12	6	yes	-	4, 13, 14, 15

Configuration rules notes

1. The ESCON 16-port card feature code is 2323, while individual ESCON ports are ordered in increments of four using feature code 2324. The ESCON card has one spare port and up to 15 usable ports.
2. The maximum number of ESCON ports on a 2084-A08 is 720.
3. The maximum number of FICON Express ports on a 2084-A08 model is 96.
4. The total number of FICON Express, OSA-Express, PCIXCC, and PCICA cards cannot exceed 20 per I/O cage.
5. The sum of OSA-Express GbE, 1000BASE-T Ethernet, Fast Ethernet, and Token Ring cards cannot exceed 24.
6. OSA-Express GbE LX/SX (FC 2364 and 2365) and OSA-Express Fast Ethernet (FC 2366) can only be brought forward on an upgrade; new adapters cannot be ordered.

7. The sum of IC, ICB-2, ICB-3, ICB-4, active ISC-3, and RPQ 8P2197 links supported on a 2084 server is limited to 64.
8. The maximum number of ICB-4 links on a 2084-A08 model is 12.
9. There are two feature codes for the ISC-3 card:
 - Feature code 0217 is for the ISC Mother Card (ISC-M).
 - Feature code 0218 is for the ISC Daughter Card (ISC-D).

One ISC Mother Card supports up to two ISC Daughter Cards, and each ISC Daughter Card contains two ports. Port activation must be ordered using feature code 0219. RPQ 8P2197 is available to extend the distance of ISC-3 links to 20 km at 1Gbps. When RPQ 8P2197 is ordered, both ports (links) in the card are activated.
10. The maximum number of ISC-3 ports in peer mode is 48, and the maximum number in compatibility mode is 32.
11. There are two types of “virtual” links that can be defined and that require CHPID numbers, but do not have PCHID numbers:
 - Internal Coupling (IC) links; each IC link pair requires two CHPID numbers.
 - HiperSockets, also called Internal Queued Direct I/O (iQDIO). Up to 16 virtual LANs can be defined, each one requiring a CHPID number.
12. Two ETR cards are automatically included in a server configuration if any coupling link I/O feature (ISC-3, ICB-2, ICB-3 or ICB-4) is selected.
13. The sum of PCIXCC and PCICA cards on a z990 server is limited to eight.
14. The PCIXCC and PCICA features do not require CHPIDs but have PCHIDs.
15. The maximum number of PCICA cards per I/O cage is two.

At least one channel I/O feature (FICON Express or ESCON) or one Coupling Facility link I/O feature (ISC-3, ICB-2, ICB-3, or ICB-4) must be present in a configuration.

The maximum number of configurable CHPIDs is 256 per Logical Channel Subsystem (LCSS) and per operating system image.

Spanned and shared channels

The Multiple Image Facility (MIF) allows channels to be shared among multiple logical partitions in a server:

- ▶ MIF shared channels can be shared by logical partitions within a Logical Channel Subsystem (LCSS).
- ▶ MIF spanned channels can be shared by logical partitions within and across LCSSs.

Table 3-8 on page 92 shows which channel types can be defined as MIF shared and/or MIF spanned channels on a z990 server.

Table 3-8 Spanned and shared channels

Channel type		CHPID definition	MIF shared channel	MIF spanned channel
ESCON	external	CNC, CTC	yes	no
		CVC, CBY	no	no
FICON Express	external	FC, FCP	yes	yes
		FCV	yes	no
OSA-Express	external	OSC, OSD, OSE	yes	yes
ICB-4	external	CBP	yes	yes
ICB-3	external	CBP	yes	yes
ICB-2	external	CBS	yes	yes
		CBR	no	no
ISC-3	external	CFP	yes	yes
		CFS	yes	yes
		CFR	no	no
IC	internal	ICP	yes	yes
HiperSockets	internal	IQD	yes	yes

Note that while the PCICA and PCIXCC cryptographic features do not have CHPID types and are not identified as spanned channels, all logical partitions in all LCSSs have access to the PCICA feature, up to 30 logical partitions per feature, and all logical partitions in all LCSSs have access to the PCIXCC feature, up to 16 logical partitions per feature.

I/O features cables and connectors

Attention: All fiber optic cables, cable planning, labeling, and installation are customer responsibilities for new z990 installations and upgrades. Fiber optic conversion kits and Mode Conditioning Patch (MCP) cables are not orderable as features on z990 servers; only ICB (copper) cables are orderable. All other cables have to be sourced separately.

IBM Fiber Cabling Services offer a total cable solution service to help with your cable ordering needs, and is highly recommended. These services take into consideration the requirements for all of the protocols/media types supported on zSeries (for example, ESCON, FICON, Coupling Links, 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 employ the use of a proven modular cabling system, the Fiber Transport System (FTS), which includes trunk cables, zone cabinets, and panels for your servers, directors, and storage devices.

FTS supports Fiber Quick Connect (FQC), a fiber harness integrated in the zSeries frame for “quick” connect, which is offered as a feature on zSeries for connection to ESCON channels.

Whether you choose a packaged service or a custom service, high quality components are used to facilitate moves, adds, and changes in the enterprise to prevent extending your maintenance “window”.

Table 3-9 lists the required connectors and cable types for each I/O feature on z990 servers.

Table 3-9 I/O features connectors and cables types

Feature code	Feature name	Connector type	Cable type
0219	ISC-3 link	LC Duplex	9 micron SM ¹
6154	ETR	MT-RJ	62.5 micron MM ²
2324	ESCON channel	MT-RJ	62.5 micron MM
2319	FICON Express LX	LC Duplex	9 micron SM
2320	FICON Express SX	LC Duplex	50, 62.5 micron MM
1364	OSA-E GbE LX ³	LC Duplex	9 micron SM
2364 ⁴	OSA-E GbE LX	SC Duplex ⁵	9 micron SM
1365	OSA-E GbE SX	LC Duplex	50, 62.5 micron MM
2365 ⁴	OSA-E GbE SX	SC Duplex ⁵	50, 62.5 micron MM
1366	OSA-E 1000BASE-T ⁶	RJ-45	Category 5 UTP ⁷
2366 ⁴	OSA-E Fast Ethernet ⁶	RJ-45	Category 5 UTP
2367	OSA-E Token Ring	RJ-45	UTP or STP ⁸

1. SM is single mode fiber.
2. MM is multimode fiber.
3. OSA-E refers to OSA-Express.
4. Brought forward to z990 on an upgrade only.
5. The OSA-Express GbE features brought forward from an upgrade have a different connector (SC Duplex) than the new OSA-E GbE features.
6. 1000BASE-T is the new Ethernet feature.
7. UTP is Unshielded Twisted Pair.
8. STP is Shielded Twisted Pair.

3.4.2 ESCON channel

What follows are z990 connectivity options in the ESCON I/O interface environment.

z990 16-port ESCON feature

The z990 16-port ESCON feature (feature code 2323) occupies one I/O slot in the z990 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 micron 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. There is a minimum of one spare port per feature, 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 new fiber optic connector called MT-RJ. The MT-RJ is an industry standard connector which has a much smaller profile compared with 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.

Note: The z990 16-port ESCON feature does *not* support a multimode fiber optic cable terminated with an ESCON Duplex connector. However, 62.5 micron 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 plant.

Fiber optic conversion kits and Mode Conditioning Patch (MCP) cables are not orderable as features on z990. Fiber optic cables, cable planning, labeling, and installation are all customer responsibilities for new z990 installations and upgrades.

As mentioned, IBM Fiber Cabling Services offer a total cable solution service to help with your cable ordering needs, and is highly recommended.

ESCON channel port enablement feature

The 15 active ports on each 16-port ESCON feature are activated in groups of four ports via Licensed Internal Code - Control Code (LIC-CC), by using the ESCON channel port feature (feature code 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 ports 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 LIC-CC enabled. This is not only because the last ESCON port (J15) of every 16-port ESCON channel card is a spare, but also because several physically installed channels are typically inactive (LIC-CC protected). These inactive channel ports are available to satisfy future channel adds.

If there is a requirement to increase the number of ESCON channel ports (minimum increment is four), and there are sufficient unused ports already available to fulfill this requirement, then IBM manufacturing provides an LIC-CC diskette to concurrently enable the number of additional ESCON ports ordered. This is illustrated in Figure 3-10 on page 95. In this case, no additional hardware is installed.

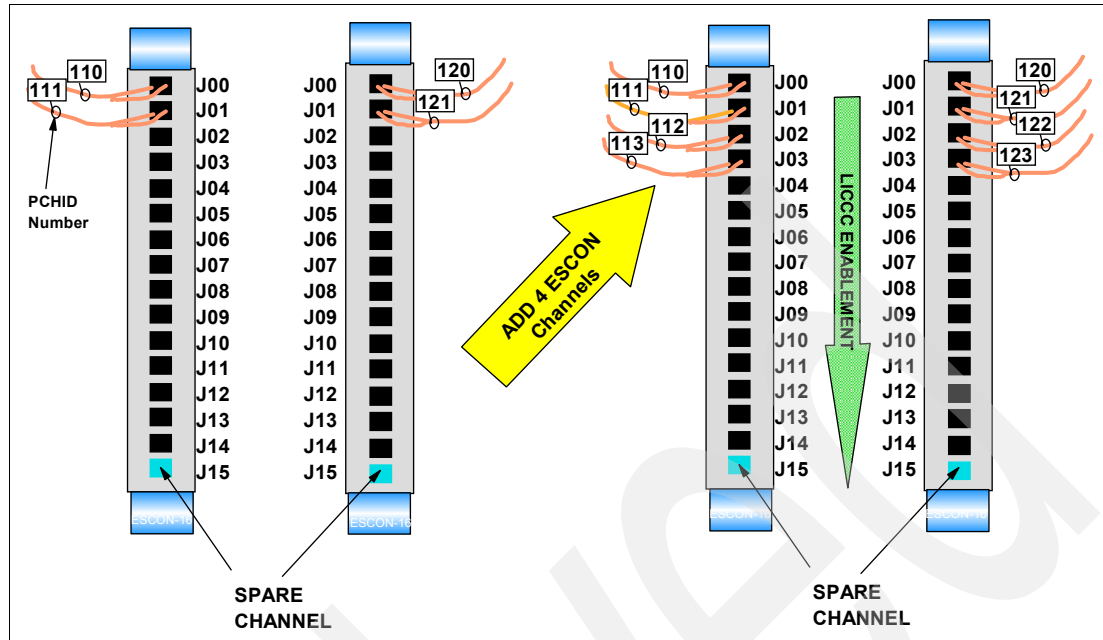


Figure 3-10 16-port ESCON - LIC-CC

An ESCON channel add will never activate the spare channel port. However, if the spare port on a card was previously used, then the add may activate all the remaining ports on that card.

If there are not enough inactive ports on existing 16-port ESCON cards installed to satisfy the additional channel order, then IBM manufacturing ships additional 16-port ESCON channel card(s) and an LIC-CC diskette.

If there has been multiple sparing on a 16-port ESCON card and, by replacing that card the additional channel add can be satisfied, the card will be replaced.

A maximum of 1024 ESCON ports can be activated on a z990 server. This maximum requires 69 16-port ESCON channel cards to be installed. The z990 model A08 can have up to 720 ESCON ports, on 48 channel cards, which is limited by the number of STIs available on the A08 model.

16-port ESCON channel sparing

The last ESCON port on a 16-port ESCON channel card (normally J15) is assigned as a spare port. Should an LIC-CC-enabled ESCON port on the card fail, the spare port is used to replace it, as shown in Figure 3-11 on page 96.

If the initial first spare port (J15) is already in use and a second LIC-CC-enabled port fails, then the highest *LIC-CC-protected* port (for example, J14) is used to spare the failing port.

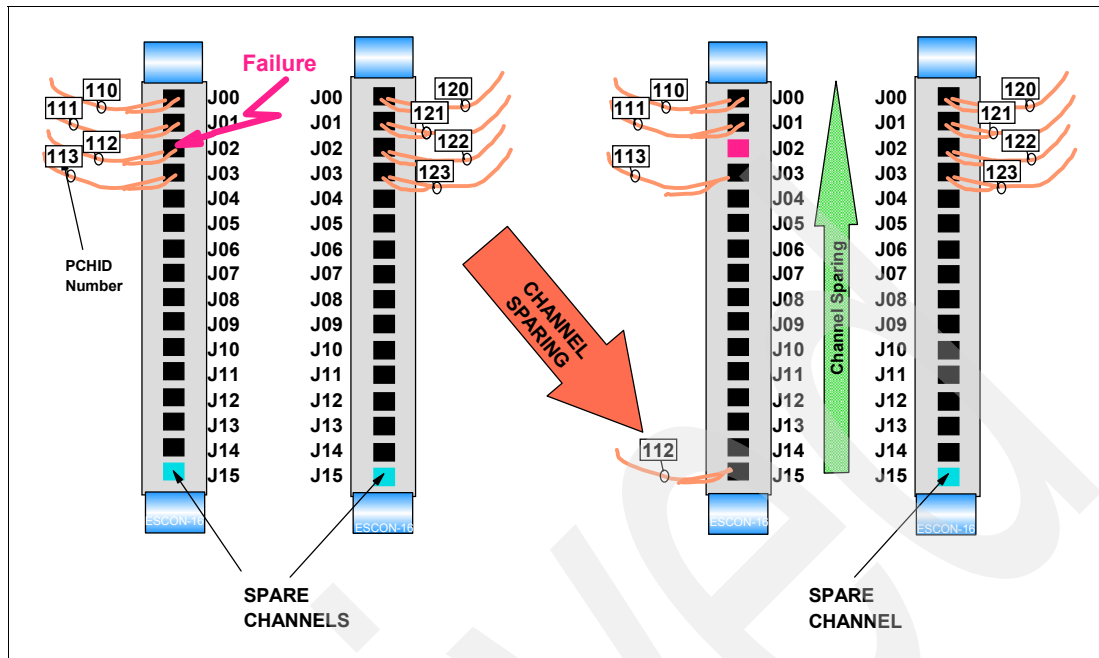


Figure 3-11 16-port ESCON channel sparing

Channel port sparing can only occur between ports on the same 16-port ESCON card, that is, a failing ESCON channel port cannot be spared with another port on a different 16-port ESCON card.

Channel sparing is a service repair action performed by an IBM service representative (SR) using the z990 server maintenance package Repair and Verify procedure. If sparing can take place, the IBM SR moves the external fiber optic cable from the failing port to the spare port. When sparing occurs, the PCHID moves to the spare port (PCHID 112 in Figure 3-11). If sparing cannot be performed, the 16-port ESCON card is replaced.

Fiber Quick Connect (FQC) for ESCON “Quick Connect”

The Fiber Quick Connect (FQC) features are optional features for factory installation of the IBM Fiber Transport System (FTS) fiber harnesses for connection to ESCON channels in the I/O cage. Each direct-attach fiber harness connects to six ESCON channels at one end and one coupler in a Multi-Terminated Push-On Connector (MTP) coupler bracket at the opposite end. When ordered, the features support all of the installed ESCON features in all of the installed I/O cages. FQC cannot be ordered on a partial or one cage basis.

FQC supports all of the ESCON channels in the I/O cage. FQC cannot be ordered for selected channels.

3.4.3 FICON channel

What follows are the connectivity options in the FICON I/O interface environment.

FICON Express features

There are two types of FICON channel transceivers supported on the z990, a long wavelength (LX) laser version, and a short wavelength (SX) laser version:

- ▶ z990 FICON Express LX feature (feature code 2319) with two ports per feature, supporting LC Duplex connectors
- ▶ z990 FICON Express SX feature (feature code 2320) with two ports per feature, supporting LC Duplex connectors

Note: Only FICON Express feature cards are available for FICON connectivity on z990. FICON channel cards (z900's feature codes 2315 and 2318), the original pre-FICON Express cards, are not offered as a new build option and are replaced with new FICON Express feature cards (feature cards 2319 or 2320) during an upgrade from z900.

FICON channel features can be installed in the z990 server. The features can be connected to a FICON-capable control unit, either point-to-point or switched point-to-point, through a Fibre Channel switch. FICON Express LX (at 1Gbps) can also be connected to the FICON LX Bridge port feature of an IBM 9032 ESCON Director.

Up to 120 FICON channels (60 features) can be installed in the z990. The model A08 can have up to 96 FICON channels (48 features), which is limited by the number of STIs available on model A08 servers.

FICON Express LX feature

The z990 FICON Express LX feature (feature code 2319) occupies one I/O slot in the z990 I/O cage. The feature has two Peripheral Component Interconnect (PCI) cards. Each PCI card has a single port supporting an LC duplex connector, with one PCHID associated to each port, and supports link speeds of 1 Gbps or 2 Gbps.

Each port supports attachment to the following:

- ▶ FICON LX Bridge one port feature of an IBM 9032 ESCON Director at 1Gbps *only*
- ▶ Fibre Channel Switch that supports 1 Gbps/2 Gbps Fibre Channel/FICON LX
- ▶ Control unit that supports 1 Gbps/2 Gbps Fibre Channel/FICON LX
- ▶ FICON channel in Fibre Channel Protocol (FCP) mode

Each port of the z990 FICON Express LX feature uses a 1300 nanometer (nm) fiber bandwidth transceiver. The port supports connection to a 9 micron single-mode fiber optic cable terminated with an LC Duplex connector.

Note: Mode Conditioning Patch (MCP) cables are for use with 1 Gbps (100 MBps) links *only*.

Multimode (62.5 or 50 micron) fiber optic cable may be used with the z990 FICON Express LX feature for 1 Gbps *only*. The use of this multimode cable type requires a Mode Conditioning Patch (MCP) 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 optical distance of the 1 Gbps link to an end-to-end maximum of 550 meters.

Fiber optic conversion kits and Mode Conditioning Patch (MCP) cables are not orderable as features on z990. Fiber optic cables, cable planning, labeling, and installation are all customer responsibilities for new z990 installations and upgrades.

IBM Fiber Cabling Services offer total a cable solution service to help with your cable ordering needs, and is highly recommended.

FICON Express SX feature

The z990 FICON Express SX feature (feature code 2320) occupies one I/O slot in the z990 I/O cage. The feature has two Peripheral Component Interconnect cards. Each PCI card has a single port supporting an LC Duplex connector, with one PCHID associated with each port, and supports link speeds of 1 Gbps and 2 Gbps.

Each port supports attachment to the following:

- ▶ Fibre Channel Switch that supports 1 Gbps/2 Gbps Fibre Channel/FICON SX
- ▶ Control unit that supports 1 Gbps/2 Gbps Fibre Channel/FICON SX

Each port of the z990 FICON Express SX feature uses an 850 nanometer (nm) fiber bandwidth transceiver. The port supports connection to a 62.5 micron or 50 micron multimode fiber optic cable terminated with an LC Duplex connector.

FICON channel in Fibre Channel Protocol (FCP) mode

When configured for FCP mode, the FICON Express features can access FCP devices either:

- ▶ Via a FICON channel in FCP mode through a single Fibre Channel switch or multiple switches to an FCP device
- ▶ Via a FICON channel in FCP mode through a single Fibre Channel switch or multiple switches to a Fibre Channel-to-SCSI bridge

Note: z990 FCP channel direct attachment in point-to-point and arbitrated loop topologies is not supported as part of the zSeries FCP enablement.

z990 adapter interruptions enhancement for FCP

The z990 servers, Linux for zSeries, and z/VM work together to provide performance improvements by exploiting extensions to the Queued Direct Input/Output (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 - FICON Express, when using the FCP CHPID type.

In extending the use of adapter interruptions to FCP channels, the programming overhead to process a traditional I/O interruption is reduced. This benefits FCP support in Linux for zSeries.

Adapter interruptions apply to a z990 FICON Express channel when in FCP mode (FCP CHPID type), which supports attachment of SCSI devices in a Linux for zSeries environment.

z990 FCP SCSI IPL feature enabler (FC 9904)

This optional z990 feature (FC 9904) allows Linux on zSeries operating system IPL from a SCSI or FCP disk. Both IPL of logical partition images and z/VM guests are supported.

Using this feature, Linux logical partitions can be started and run completely from SCSI or FCP disks. Further, a stand-alone dump program can be loaded from such a SCSI or FCP disk in order to dump the contents of a logical partition, and the dump data can be written to this same disk.

z990 FCP concurrent patch

FICON channels, when configured as CHPID type FCP, support concurrent patches allowing the application of a new Licensed Internal Code (LIC) without requiring a configuration of off/on. This is a zSeries exclusive FCP availability feature, available with FICON Express feature codes 2319 and 2320.

3.4.4 OSA-Express adapter

What follows is a discussion of the connectivity options in the OSA-Express environment. The z990 supports the following OSA-Express features:

- ▶ OSA-Express Gigabit Ethernet (GbE) Long Wavelength (LX), feature code 1364
- ▶ OSA-Express Gigabit Ethernet (GbE) Short Wavelength (SX), feature code 1365
- ▶ OSA-Express Gigabit Ethernet (GbE) Long Wavelength (LX), feature code 2364
 - Brought forward on an upgrade only
- ▶ OSA-Express Gigabit Ethernet (GbE) Short Wavelength (SX), feature code 2365
 - Brought forward on an upgrade only
- ▶ OSA-Express 1000BASE-T Ethernet, feature code 1366
- ▶ OSA-Express Fast Ethernet, feature code 2366
 - Brought forward on an upgrade only
- ▶ OSA-Express Token Ring, feature code 2367

Note: If FDDI or ATM connectivity is desired, a multiprotocol switch or router with the appropriate network interface (for example, 1000BASE-T Ethernet, Gigabit Ethernet) can be used to provide connectivity between the z990 server and an FDDI or ATM network.

A z990 server can support a maximum of 24 OSA-Express features (48 ports).

Table 3-10 OSA-Express features support

I/O feature	Feature codes	Number of		Maximum number		PCHID	CHPID definition	Config. rules notes
		Ports per card	Ports increments	Ports	I/O slots			
OSA-E Gbit Ethernet LX/SX	1364/1365 (2364/2365)	2	2	48 (24)	24 (12)	yes	OSD	1, 2, 3
OSA-E 1000BASE-T Ethernet	1366	2	2	48	24	yes	OSE, OSD, OSC	1, 2

I/O feature	Feature codes	Number of		Maximum number		PCHID	CHPID definition	Config. rules notes
		Ports per card	Ports increments	Ports	I/O slots			
OSA-E Fast Ethernet	2366	2	2	24	12	yes	OSE, OSD	1, 2, 3
OSA-E Token Ring	2367	2	2	48	24	yes	OSE, OSD	1, 2

Notes

1. The total number of FICON Express, OSA-Express, PCIXCC, and PCICA cards cannot exceed 20 per I/O cage.
2. The sum of OSA-Express GbE, 1000BASE-T Ethernet, Fast Ethernet and Token Ring cards cannot exceed 24.
3. OSA-Express GbE LX/SX (FC 2364 and 2365) and OSA-Express Fast Ethernet (FC 2366) can only be brought forward on an upgrade; new adapters cannot be ordered.

OSA-Express GbE LX (feature code 1364)

The z990 OSA-Express Gigabit (GbE) Long Wavelength (LX) feature (feature code 1364) occupies one slot in the z990 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 via a 9 micron single-mode fiber optic cable terminated with an LC Duplex connector. This feature utilizes a long wavelength (LX) laser as the optical transceiver.

Multimode (62.5 or 50 micron) fiber cable may be used with the z990 OSA-Express 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 z990 OSA-Express GbE LX feature (feature code 1364) supports Queued Direct Input/Output (QDIO) mode only, full-duplex operation, jumbo frames, and checksum offload. It is defined with CHPID type OSD.

OSA-Express GbE SX (feature code 1365)

The z990 OSA-Express Gigabit (GbE) Short Wavelength (SX) feature (feature code 1365) occupies one slot in the z990 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 via a 62.5 micron or 50 micron multimode fiber optic cable terminated with an LC Duplex connector. The feature utilizes a short wavelength (SX) laser as the optical transceiver.

The z990 OSA-Express GbE SX feature (feature code 1365) supports Queued Direct Input/Output (QDIO) mode only, full-duplex operation, jumbo frames, and checksum offload. It is defined with CHPID type OSD.

OSA-Express GbE LX (feature code 2364, upgrade only)

The z990 OSA-Express GbE LX feature (feature code 2364) *can only be brought forward on an upgrade*. Its replacement adapter for new orders is the z990 OSA-Express GbE LX feature (feature code 1364).

The z990 OSA-Express GbE LX feature occupies one slot in the z990 I/O cage and has two independent ports with one PCHID associated with each port. This feature supports the 1000BASE-SX standard transmission scheme.

Each port supports connection to a 1 Gbps Ethernet LAN via a 9 micron single-mode fiber optic cable terminated with an SC Duplex connector.

Multimode (62.5 or 50 micron) fiber cable may be used with the z990 OSA-Express 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 z990 OSA-Express GbE LX feature (feature code 2364) only supports QDIO mode and TCP/IP. It is defined with CHPID type OSD. The Enterprise Extender (EE) function of Communications Server for z/OS and OS/390 allows you to run SNA applications and data on IP networks and IP-attached clients.

OSA-Express GbE SX (feature code 2365, upgrade only)

The z990 OSA-Express GbE SX feature (feature code 2365) *can only be brought forward on an upgrade*. Its replacement adapter for new orders is the z990 OSA-Express GbE SX feature (feature code 1365).

The z990 OSA-Express GbE SX feature occupies one slot in the z990 I/O cage and has two independent ports with one PCHID associated with each port. This feature supports the 1000BASE-SX standard transmission scheme.

Each port supports connection to a 1 Gbps Ethernet LAN via a 62.5 micron or 50 micron multimode fiber optic cable terminated with an SC Duplex connector.

The z990 OSA-Express GbE SX feature (feature code 2365) only supports QDIO mode and TCP/IP. It is defined with CHPID type OSD. The Enterprise Extender (EE) function of Communications Server for z/OS and OS/390 allows you to run SNA applications and data on IP networks and IP-attached clients.

OSA-Express 1000BASE-T Ethernet (feature code 1366)

The z990 OSA-Express 1000BASE-T Ethernet, feature code 1366, occupies one I/O slot in the z990 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 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-Express 1000BASE-T Ethernet feature supports auto-negotiation and automatically adjusts to 10 Mbps, 100 Mbps, or 1000 Mbps, depending upon the LAN.

LAN speed and/or the duplex mode can be set explicitly, using OSA/SF or the OSA Advanced Facilities function of the z990 server hardware management console (HMC). The explicit settings will override the OSA-Express feature port ability to auto-negotiate with its attached Ethernet switch.

You can choose any one of the following settings for the OSA-Express 1000BASE-T Ethernet feature:

- ▶ Auto-negotiate
- ▶ 10 Mbps half-duplex or full-duplex
- ▶ 100 Mbps half-duplex or full-duplex
- ▶ 1000 Mbps/1Gbps full-duplex

LAN speed and duplexing mode default to auto negotiation. The OSA-Express 1000BASE-T 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, or OSE.

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 (OSD CHPID type), carries TCP/IP packets only, operates in full-duplex mode only, supports jumbo frames, and supports checksum offload.

OSA-Express Integrated Console Controller (OSA-ICC)

The 1000BASE-T Ethernet feature also provides the OSA-Express Integrated Console Controller (OSA-ICC) function, which supports TN3270E (RFC 2355) and non-SNA DFT 3270 emulation. The OSA-ICC function uses a new definition as OSC CHPID 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 z990 via a port on the OSA-Express 1000BASE-T Ethernet feature. This can help eliminate the requirement for external console controllers, like 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 port-by-port basis, and is supported at any of the feature settings (10, 100, 1000 Mbps, half or full-duplex).

OSA-Express Fast Ethernet (feature code 2366, upgrade only)

The z990 OSA-Express Fast Ethernet (FENET), feature code 2366, *can only be brought forward on an upgrade*. The replacement adapter for new orders is the z990 OSA-Express 1000BASE-T Ethernet feature (feature code 1366).

The z990 OSA-Express FENET feature occupies one I/O slot in the z990 I/O cage and has two independent ports, with one PCHID associated with each port.

Each port supports connection to either a 100 Mbps or 10 Mbps Ethernet LAN. The LAN must conform either to the IEEE 802.3 (ISO/IEC 8802.3) standard or the Ethernet V2.0 specifications, and the 10BASE-T or 100BASE-TX standard transmission schemes.

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).

You can choose any one of the following settings for the OSA-Express FENET feature:

- ▶ Auto negotiate
- ▶ 10 Mbps half-duplex or full-duplex
- ▶ 100 Mbps half-duplex or full-duplex

LAN speed and/or the duplex mode can be set explicitly, using OSA/SF or the OSA Advanced Facilities function of the z990 server hardware management console (HMC). The explicit

settings will override the OSA-Express feature port ability to auto-negotiate with its attached Ethernet switch.

The OSA-Express FENET feature supports auto-negotiation with its attached Ethernet hub, router, or switch. If you allow the LAN speed to default to auto-negotiation, the FENET OSA-Express and the attached hub, router, or switch auto-negotiates the LAN speed setting between them. If the attached Ethernet hub, router, or switch does not support auto-negotiation, the OSA enters the LAN at the default speed of 100 Mbps in half-duplex mode.

If you are not using auto-negotiate, the OSA will attempt to join the LAN at the specified speed/mode; however, the speed/mode settings are only used when the OSA is first in the LAN. If this fails, the OSA will attempt to join the LAN as if auto-negotiate were specified.

The OSA-Express FENET feature can be defined with CHPID type OSD or OSE. The HPDT MPC mode is no longer available on the FENET for z990.

OSA-Express Token Ring (feature code 2367)

The z990 OSA-Express Token Ring feature (feature code 2367) occupies one I/O slot in the z990 I/O cage. The feature has two independent ports, with one PCHID associated with each port. The OSA-Express Token Ring feature can be defined with CHPID type OSD or OSE.

The OSA-Express Token Ring feature supports auto-sensing as well as any of the following settings: 4 Mbps half- or full-duplex, 16 Mbps half- or full-duplex, and 100 Mbps full-duplex. Regardless of the choice made, the network switch settings must agree with those of the OSA-Express Token Ring feature. If the LAN speed defaults to auto-sense, the OSA-Express Token Ring feature will sense the speed of the attached switch and insert into the LAN at the appropriate speed. If the Token Ring feature is the first station on the LAN and the user specifies auto-sense, it will default to a speed of 16 Mbps and will attempt to open in full-duplex mode. If unsuccessful, it will default to half-duplex mode. The OSA-Express Token Ring feature conforms to the IEEE 802.5 (ISO/IEC 8802.5) standard.

Each port has an RJ-45 receptacle and a DB-9 D shell receptacle for cabling to a Token Ring MAU or Token Ring switch that is appropriate for the LAN speed. Only one of the port's two receptacles can be used at any time.

The RJ-45 receptacle is required to be attached using an EIA/TIA category 5 unshielded twisted pair (UTP) cable that does not exceed 100 m (328 ft), or a shielded twisted pair (STP) cable with a DB-9 D Shell connector.

Checksum offload for IPv4 packets when in QDIO mode

A function, called Checksum Offload, offered for the OSA-Express GbE and 1000BASE-T Ethernet features, was introduced for the Linux for zSeries and z/OS environments. Checksum Offload provides the capability of calculating the Transmission Control Protocol (TCP), User Datagram Protocol (UDP), and Internet Protocol (IP) header checksums. Checksum verifies the correctness 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. This function applies to packets that actually go onto the Local Area Network (LAN) or come in 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 sharing the OSA-Express, OSA-Express 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.

This function does not apply to IPv6 packets. TCP/IP will continue to perform all checksum processing for IPv6 packets. This function also does not apply to ICMP checksum processing. TCP/IP will continue to perform processing for ICMP checksum.

Checksum offload is supported by the OSA-Express GbE features (FC 1364, FC 1365) and the 1000BASE-T Ethernet feature (FC 1366) when operating at 1000 Mbps (1 Gbps). This is applicable to the QDIO mode only (channel type OSD).

z/OS support for Checksum offload is planned to be available in z/OS V1.5.

For Linux for zSeries support, refer to the following Web site for further information:

<http://www.ibm.com/developerworks>

z990 adapter interruptions enhancement for QDIO

The z990 servers, Linux for zSeries, and z/VM work together to provide performance improvements by exploiting extensions to the Queued Direct Input/Output (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-Express when using the OSD CHPID type.

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 both Linux for zSeries and z/VM.

Adapter interruptions apply to all of the OSA-Express features available on z990, whether offered as a new build or on an upgrade from z900 when in QDIO mode (CHPID type OSD).

HiperSockets function

The HiperSockets function, also known as internal Queued Direct Input/Output (iQDIO) or internal QDIO, is an integrated function of the z990 server that provides users with attachment to up to sixteen high-speed “virtual” Local Area Networks (LANs) with minimal system and network overhead.

HiperSockets eliminates the need to utilize I/O subsystem operations and the need to traverse an external network connection to communicate between logical partitions in the same z990 server. HiperSockets offers significant value in server consolidation connecting many virtual servers, and can be used instead of certain XCF link configurations in a Parallel Sysplex.

HiperSockets are customizable to accommodate varying traffic sizes. Since HiperSockets does not use an external network, it can free up system and network resources, eliminating attachment costs while improving availability and performance.

3.4.5 Coupling Facility links

What follows are Coupling Facility links connectivity options in the Parallel Sysplex environment. For more information about Parallel Sysplex connectivity, see 7.2.4, “Coupling Facility link connectivity” on page 164.

z990 coupling link features

z990 supports the following coupling link features:

- ▶ Inter-System Channel-3, ISC-3 (Peer and Compatibility modes), feature codes 0217, 0218, and 0219

- ▶ Integrated Cluster Bus-4, ICB-4 (Peer mode), feature code 3393
- ▶ Integrated Cluster Bus-3, ICB-3 (Peer mode) feature code 0993
- ▶ Integrated Cluster Bus-2, ICB-2 (Compatibility mode), feature code 0992
- ▶ Internal Channel, IC (Peer mode): No feature code; Licensed Internal Code (LIC) function defined via HCD/IOCP

ISC-3 link

The z990 ISC-3 feature is made up of the following feature codes:

- ▶ ISC-3 Mother Card (feature code 0217)
- ▶ ISC-3 Daughter Card (feature code 0218)
- ▶ ISC-3 Port (feature code 0219)

The z990 ISC-3 Mother Card occupies one slot in the I/O cage. The ISC-3 Mother Card supports up to two ISC-3 Daughter Cards. Each ISC-3 Daughter Card has two independent ports with one PCHID associated with each active port. The ISC-3 ports are activated via Licensed Internal Code Configuration Control (LIC-CC).

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 supports connection to a 2 Gbps (ISC-3 Peer mode) or 1 Gbps (ISC-3 Compatibility mode) Coupling link via 9 micron single mode fiber optic cable terminated with an LC-Duplex connector.

ISC features on G5/G6 and earlier servers have Fiber Optic Sub Assemblies (FOSA) that support SC-Duplex cable connectors. These existing single mode HiPerLink cables can be reused by attaching a single mode fiber LC-Duplex to SC-Duplex conversion cable. This is a 2 m cable that is connected between the z990 server ISC-3 port and the existing HiPerlink cable from the G5/G6 server.

Note: Existing SC-Duplex 50 micron multimode fiber cable infrastructure may be reused with the z990 ISC-3 port features in Compatibility mode (1 Gbps) only. The use of these multimode cable types requires a Mode Conditioner 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 550 meters.

Fiber optic conversion kits and Mode Conditioning Patch (MCP) cables are not orderable as features on z990. Fiber optic cables, cable planning, labeling, and installation are all customer responsibilities for new z990 installations and upgrades.

IBM Fiber Cabling Services offer a total cable solution service to help with your cable ordering needs, and is highly recommended.

RPQ 8P2197: Extended distance option

The RPQ 8P2197 ISC-3 Daughter Card has two links per card. Both links are active when installed and do not need to be activated via LIC.

This RPQ card supports Peer mode and Compatibility mode at 1 Gbps *only*. It extends the maximum distance of the ISC-3 link to 20 km. For Peer mode, one RPQ Daughter Card is required at each end of the link between the z990, z900, z890, or z800 servers. For Compatibility mode, the equivalent G5/G6 server extended distance RPQ Daughter Card is required on the G5/G6 server end of the link.

Table 3-11 lists the various ISC-3 link characteristics.

Table 3-11 ISC-3 link characteristics

Mode of operation	IOCP definition	Bandwidth	Open Fiber Control (OFC)	Intended attachment	Maximum distance
Peer	CFP	2 Gbps	No	z990, z900, z890, or z800	10 km
Compatibility	CFS/CFR	1 Gbps	Emulation	9672 G5/G6	10 km
Peer with RPQ 8P2197	CFP	1 Gbps	No	z990, z900, z890, or z800	20 km
Compatibility with RPQ 8P2197	CFS/CFR	1 Gbps	Emulation	9672 G5/G6	20 km

ICB-4 link

The Integrated Cluster Bus-4 (ICB-4) link (feature code 3393) is a member of the family of Coupling link options available on z990 servers. ICB-4 is a “native” connection used to connect z990/z890 servers (2084/2086) to z990 servers. An ICB-4 connection consists of one link that attaches directly to an STI port in the CEC cage, does not require connectivity to a card in the I/O cage, and operates at 2 GBps. One ICB-4 feature is required for each end of the link. Each end of the ICB-4 link has a PCHID number.

The ICB-4 cable, feature code 0228, is a unique 10 meter (33 feet) 2.0 GB copper cable to be used with ICB-4 links only.

ICB-4 cables (FC 0228) are automatically ordered to match the quantity of ICB-4 feature code 3393 on order. Order one cable per connection, not per feature. The quantity of ICB cables can be reduced, but cannot exceed the quantity of ICB features on order.

ICB-3 link

The Integrated Cluster Bus-3 (ICB-3) link (feature code 0993) is a member of the family of Coupling Link options available on z990 servers. ICB-3 links are used to connect z900/z800 servers (2064/2066) to z990 servers. There is an STI-3 card that resides in the I/O cage and provides two ports to support the ICB-3 connections. The STI-3 card converts the 2 GBps input into two 1 GBps ICB-3s. One ICB-3 feature is required for each end of the link. Each ICB-3 link at the z990 end has a PCHID number.

The ICB-3 cable (feature code 0227) is a unique 10 meter (33 feet) 1.0 GB copper cable to be used with ICB-3 links. Existing 10 meter 1.0 GB ICB-3 cables can be reused.

ICB-3 cables (feature code 0227) will be automatically ordered to match the quantity of ICB-3s (feature code 0993) on order. Order one cable per connection, not per feature. The quantity of ICB cables can be reduced, but cannot exceed the quantity of ICB features on order.

ICB-2 link

The Integrated Cluster Bus-2 (ICB-2) link (feature code 0992) is a member of the family of Coupling Link options available on z990 servers. ICB-2 links are used to connect 9672 G5/G6 to a z990 server. There is an STI-2 card that resides in the I/O cage and provides two output

ports to support the ICB-2 connections. The STI-2 card converts the 2 GBps input into two 333 MBps ICB-2s. One ICB-2 feature is required for each end of the link. Each ICB-2 link at the z990 end has a PCHID number. ICB-2 only support connection to 9672 G5/G6 servers.

The ICB-2 cable (feature code 0226) is a unique 10 meter (33 feet) 333 MB copper cable to be used with ICB-2 links. Existing 10 meter 333 MB ICB-2 cables can be reused.

ICB-2 cables (feature code 0226) will be automatically ordered to match the quantity of ICB-2s (feature code 0992) on order. Order one cable per connection, not per feature. The quantity of ICB cables can be reduced, but cannot exceed the quantity of ICB features on order.

z990 ICB links summary

Table 3-12 shows a summary of all z990 Integrated Cluster Bus (ICB) link types, including some of their characteristics.

Table 3-12 z990 ICB links summary

ICB link type	Feature code	IOCP definition	Bandwidth	Intended attachment	Maximum cable length
ICB-4	3393	CBP	2 GBps	z990, z890	10 meters
ICB-3	0993	CBP	1 GBps	z900, z800	10 meters
ICB-2	0992	CBS, CBR	333 MBps	9672 G5/G6 only	10 meters

IC links

IC links are used when an ICF logical partition is on the same CPC as other system images participating in the sysplex. An IC link is the fastest Coupling link, using just memory-to-memory data transfers. IC links do not have PCHID number, but do require CHPIDs.

IC links require ICP channel path definition at the OS/390 or z/OS and the CF end of a channel connection to operate in peer mode. They are always defined and connected in pairs.

3.4.6 External Time Reference (ETR) feature

The External Time Reference (ETR), feature code 6154, is an optional z990 feature.

Each ETR feature consists of one ETR card and each card has one port. When a quantity of one is ordered, two features are shipped. The two ETR features are automatically ordered if any coupling link feature (ISC-3, ICB-2, ICB-3 or ICB-4) is ordered.

These cards provide attachment to the Sysplex Timer in the CPC cage. Each ETR card should connect to a different 9037 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 one CEC cage card slot in the rear and provide attachment to a 9037 Sysplex Timer; it can be either a 9037 Model 1¹ or 9037 Model 2. The 9037 Sysplex Timer provides the synchronization for the Time-of-Day (TOD) clocks of multiple CPCs, and thereby allows events started by different CPCs to be properly sequenced in time. When multiple CPCs update the same database and database reconstruction is necessary, all updates are required to be time-stamped in proper sequence.

¹ Note that 9037-1 goes End of Service on Dec. 31, 2003.

Important: The z990 timer support has been enhanced and requires a customer to assign a timer netID that must match the Sysplex Timer (9037) netID before the CPC clock will step to the 9037 time signals; refer to 7.2.1, “Sysplex configurations and Sysplex Timer considerations” on page 159 for more information.

The port cards support concurrent maintenance. The ETR card port has a small form factor optical transceiver that supports an MT-RJ connector only.

Note: The ETR card does not support a multimode fiber optic cable terminated with an ESCON Duplex connector.

However, 62.5 micron multimode ESCON Duplex jumper cables 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 on z990. Fiber optic cables, cable planning, labeling, and installation are all customer responsibilities for new z990 installations and upgrades.

IBM Fiber Cabling Services offer a total cable solution service to help with your cable ordering needs, and is highly recommended.

3.4.7 Cryptographic features

What follows are the cryptographic features available on z990 servers. To enable any cryptographic function on the z990, the CP Assist for Cryptographic Function (feature code 3863) must be installed and enabled. For more information about cryptographic functions, see Chapter 5, “Cryptography” on page 119.

PCIX Cryptographic Coprocessor (PCIXCC) feature

The Peripheral Component Interconnect X Cryptographic Coprocessor (PCIXCC) feature (feature code 0868) occupies one I/O slot in the z990 I/O cage.

Each PCIXCC feature has one coprocessor. It does not require a CHPID number, but is assigned to a PCHID.

The CP Assist for Cryptographic Function (feature code 3863) must be installed and enabled to use PCIXCC.

PCI Cryptographic Accelerator (PCICA) feature

The Peripheral Component Interconnect Cryptographic Accelerator (PCICA) feature (feature code 0862) occupies one I/O slot in the z990 I/O cage.

Each PCICA feature has two accelerator cards. The PCICA feature does not require CHPID numbers, but is assigned to PCHIDs.

The CP Assist for Cryptographic Function (feature code 3863) must be enabled to use PCICA.

Channel Subsystem

This chapter describes how the Channel Subsystem (CSS) is implemented on the z990. Each server has a Channel Subsystem. Its role 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 Input/Output (I/O) operations. The CSS provides the server communications to external devices via 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.

The architecture and functionality of the CSS is significantly enhanced on the z990. The implementation of the Channel Subsystem is significantly different from those on previous zSeries servers.

This chapter introduces the concept of multiple Logical Channel Subsystems (LCSSs) implemented on the z990. The goal of this chapter is to make you familiar with the technology, terminology, and implementation aspects of the z990 Channel Subsystem design. The multiple Logical Channel Subsystem (LCSS) and related components are described in:

- ▶ 4.1, “Multiple Logical Channel Subsystem (LCSS)” on page 110
- ▶ 4.1.1, “Logical Channel Subsystem structure” on page 110
- ▶ 4.2, “LCSS configuration management” on page 115
- ▶ 4.3, “LCSS-related numbers” on page 117

4.1 Multiple Logical Channel Subsystem (LCSS)

The concept of Logical Channel Subsystem (LCSS) is new to the z990. The z990 supports up to four Logical Channel Subsystems, hence the term multiple Logical Channel Subsystem. The design of the z990 offers a considerable increase in processing power, memory sizes, and I/O connectivity. In support of the larger I/O capability, the Channel Subsystem has been scaled up correspondingly and the LCSS concept is designed to do just that. This concept is introduced to facilitate the architectural change that provides more logical partitions and more channels than before.

The structure provides up to four Logical Channel Subsystems. Each LCSS may have from one to 256 CHPIDs, and may in turn be configured with up to 15 logical partitions that relate to that particular Logical Channel Subsystem. LCSSs are numbered from 0 to 3, and are sometimes referred to as the CSS Image ID (CSSID 0, 1, 2, and 3).

Note: The z990 provides for four Logical Channel Subsystems, 1024 CHPIDs, and up to 30 logical partitions for the total system.

Table 4-1 shows the number of logical partitions and CHPIDs supported.

Table 4-1 Logical partitions and CHPID numbers support

LCSS supported	Number of active LPARs	Number of defined LPARs	Number of server CHPIDs supported
LCSS 0, 1, 2, 3	30	30	1024

4.1.1 Logical Channel Subsystem structure

The provision for multiple LCSSs is an extension to that provided on previous z/Architecture servers. It provides channel connectivity to the defined logical partitions in a manner that is transparent to subsystems and application programs.

The LCSS introduces new components and terminology that differs from previous server generations. These components are explained in the following sections.

Logical Channel Subsystem (LCSS)

The z990 provides the ability to define more than 256 CHPIDs because of the introduction of Logical Channel Subsystem concept. An LCSS is a logical replication of CSS facilities (subchannels, CHPIDs, controls, and so on). This enables the definition of a balanced configuration for the processor, and I/O capabilities. The LCSSs of the z990 introduce significant changes to the I/O configuration.

For ease of management, it is strongly recommended that the Hardware Configuration Dialog (HCD) be used to build and control your z990 Input/Output Definition file (IODF). HCD support for z990 multiple Logical Channel Subsystems is available beginning with z/VM 4.4 and is available on all current OS/390 and z/OS levels. HCD provides the capability to make both dynamic hardware and software I/O configuration changes.

A z990 must have at least one LCSS defined. No logical partitions can exist without at least one defined LCSS. Logical partitions are defined to an LCSS, not to a processor. Up to four LCSSs are supported on the z990, and a logical partition is associated with one and only one LCSS. CHPIDs are unique within an LCSS and range from 00 to FF. The same CHPID number range is used again for the other LCSSs.

It is important to note that an IBM 2084 is one processor with logical extensions. All Channel Subsystem Images (CSS Image or LCSS) are defined within a single IOCDS. The IOCDS is loaded and initialized into the Hardware System Area at Power-on Reset.

There is no HSA expansion support for dynamic I/O on the z990 Support Element. The HSA allocation is controlled by the “maximum number of devices” field on the HCD Channel Subsystem List panel. This value can only be changed by a Power-on Reset. Figure 4-1 shows a logical view of the relationships. It must be noted that each LCSS supports up to 15 logical partitions, but system wide, a total of up to 30 logical partitions are supported.

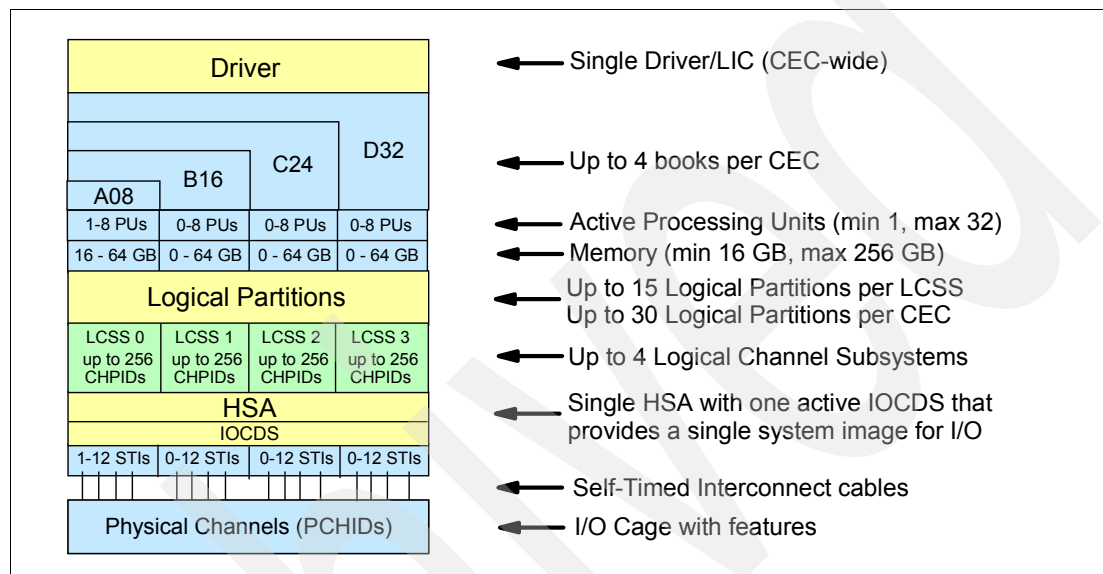


Figure 4-1 Logical view of IBM 2084 models, LCSSs, IOCDS and HSA¹

Note: HSA is always allocated in the physical memory of Book 0.

The channel definitions of an LCSS are not bound to a single book. An LCSS may define resources that are physically connected to all STIs of all books in any multi-book IBM 2084 model.

Multiple Image Facility (MIF)

Multiple Image Facility (MIF) enables resource sharing across logical partitions within a single LCSS or across the LCSSs. When a channel resource is shared across logical partitions in multiple LCSS, this is known as “spanning”; refer to 4.1.3, “Channel spanning” on page 114 for more information about spanning.

With the introduction of multiple LCSSs, the IOCDS logical partition MIF Image ID is no longer unique within the z990 server. Therefore, the logical partition identifier value has been changed to provide a unique value for each logical partition within the same z990 server. The following terminology applies:

Logical partition number

The logical partition number cannot be specified by the user; actually, it is not even visible to the user. On the z990, it is assigned at Power-on Reset by PR/SM and is based on the total

¹ Although each LCSS supports up to 15 logical partitions, it should be noted that, system wide, up to 30 logical partitions are supported.

number of partitions defined in the RESOURCE statements in the IOCDs. It is unique for each logical partition.

Logical partition identifier

The logical partition identifier is a number in the range from '00' to '3F'. It is assigned by the user on the image profile through the Support Element (SE) or the Hardware Management Console. It is unique across the z990 and may also be referred to as the User Logical Partition ID (UPID).

MIF Image ID (MIFID)

The Multiple Image facility enables channel sharing among logical partitions pertaining to the same Logical Channel Subsystem.

The MIF Image ID is a number that is defined through Hardware Configuration Dialog (HCD) or directly via the IOCP. It is a number that is specified in the RESOURCE statement in the configuration definitions. It is in the range '1' to 'F' and is unique within an LCSS, but it is not unique within the z990. Multiple LCSSs may specify the same MIF Image ID (the MIFID was called the logical partition number in previous zSeries).

Logical partition name

This name is user defined through HCD or the IOCP and is the partition name in the RESOURCE statement in the configuration definitions. The names must be unique across all LCSSs defined for the z990.

Figure 4-2 summarizes all the identifiers and how they are defined, using a three LCSS configuration example.

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	
Logical Partition Number			Logical Partition Number			Log Part Number	Logical Partition Number		Assigned by PR/SM at POR ←
MIF ID			MIF ID			MIF ID	MIF ID		Specified in HCD / IOCP ←
2	4	A	4	6	D	2	5	A	
LCSS0			LCSS1			LCSS2	LCSS3		Specified in HCD / IOCP ←

Figure 4-2 LCSS, logical partition, and identifiers

We suggest you establish a naming convention for the logical partition identifiers. As shown in Figure 4-2 on page 112, you could use the LCSS number concatenated to the MIF Image ID, which means logical partition ID 3A is in LCSS 3 with MIF ID A. 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

In order to have a partition defined for future use, such a dynamic partition must be reserved beforehand in the IOCDS that is used for Power-On Reset. A reserved partition is defined with a partition name placeholder, a MIF ID, a usage type, and, optionally, may contain a description. The reserved partition can be assigned a logical partition name to be later used in I/O commands of HCD.

Important: Some HCD and HCM panels may still refer the user to the definition of a “logical partition number”. For a z990 configuration, this is incorrect, and the user should understand that the panel refers to the definition of a “MIF ID”

As previously mentioned, on a z990 the “logical partition number” is assigned by PR/SM during Power-on Reset and cannot be modified, nor visualized, by the user.

4.1.2 Physical Channel ID (PCHID)

A Physical Channel ID, or 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 on a z990, and may be arbitrarily assigned. A hardware channel is now identified by a PCHID, or Physical Channel Identifier.

You can address 256 CHPIDs within a single Logical Channel Subsystem. That gives a maximum of 1024 CHPIDs when four LCSSs are defined. Each CHPID 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 follows the numbering scheme shown in Table 4-2.

Table 4-2 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 - 0FF reserved for ICB-4s	

CHPIDs are not pre-assigned. It is the responsibility of the user to assign the CHPID numbers through the use of the CHPID Mapping Tool (CMT) or HCD/IOCP. Assigning CHPIDs means that the CHPID number is associated with a physical channel port location (PCHID), and a LCSS. The CHPID number range is still from ‘00’ to ‘FF’ and must be unique within an LCSS. Any CHPID not connected to a PCHIDs will fail validation when an attempt is made to build a production IODF or an IOCDS.

A pictorial view of a z990 with multiple LCSS is shown in Figure 4-3 on page 114. Two Logical Channel Subsystems are defined (LCSS0 & LCSS1). Each LCSS has three logical partitions with their associated MIF Image Identifiers.

In each LCSS, the CHPIDs are shared across all logical partitions. The CHPIDs in each LCSS 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. See 4.2.1, “z990 configuration management” on page 116 for further details on the CMT.

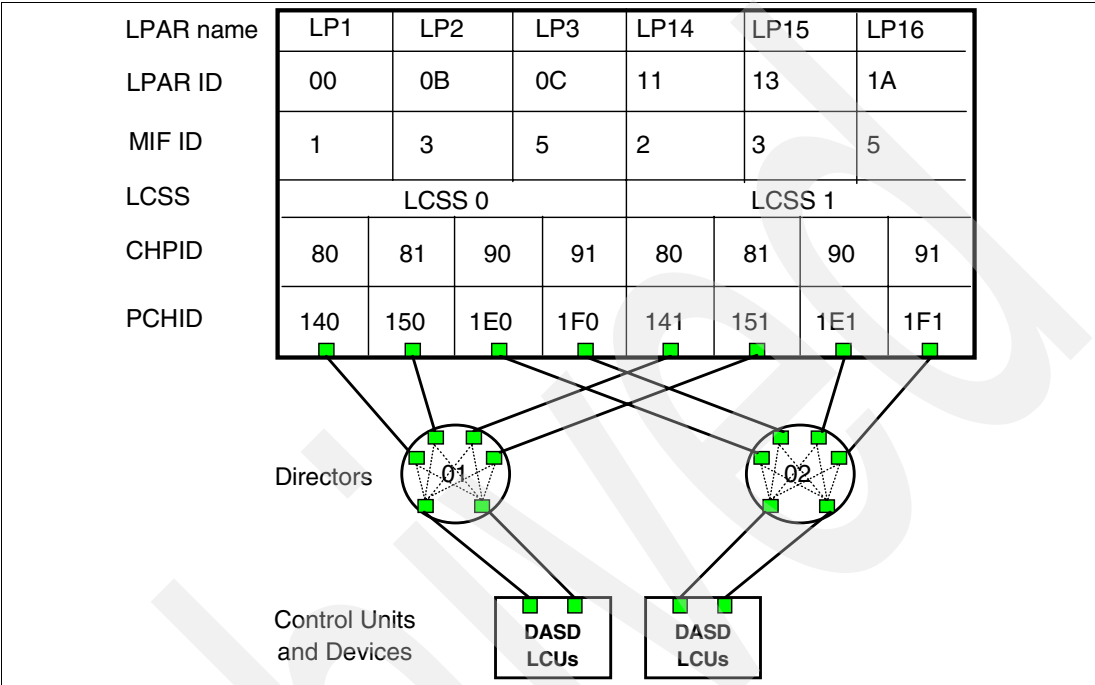


Figure 4-3 z990 LCSS connectivity

4.1.3 Channel spanning

Channel spanning extends the MIF concept of sharing channels across logical partitions to sharing channels across logical partitions *and* Logical Channel Subsystems.

Spanning is the ability for the channel to be configured to multiple Logical Channel Subsystems. When defined that way, the channels can be transparently shared by any or all of the configured logical partitions, regardless of the Logical Channel Subsystem to which the logical partition is configured.

A channel is considered a spanned channel if the same CHPID number in different LCSSs 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 there is no PCHID association. They are defined with the same CHPID number in multiple LCSSs.

CHPIDs that span LCSSs reduce the total number of channels available on the z990. The total is reduced, since no LCSS can have more than 256 CHPIDs. For a z990 with two LCSSs, a total of 512 CHPIDs are supported. If all CHPIDs are spanned across the two LCSSs, then only 256 channels can be supported. For a z990 with four LCSSs, a total of 1024 CHPIDs are supported. If all CHPIDs are spanned across the four LCSSs, then only 256 channels can be supported.

Channel spanning is supported for internal links (HiperSockets, and Internal Coupling (IC) links), and for some external links (FICON Express channels, OSA-Express, and coupling links). For a complete list of supported spanned channels, see Table 3-8 on page 92.

Note: Spanning of ESCON channels, FICON converter (FCV) channels, and receiver coupling links are not supported.

In Figure 4-4, CHPID 04 is spanned to LCSS0 and LCSS1. Since it is not an external channel link, there is no PCHID assigned. CHPID 06 is an external spanned channel and has a PCHID assigned.

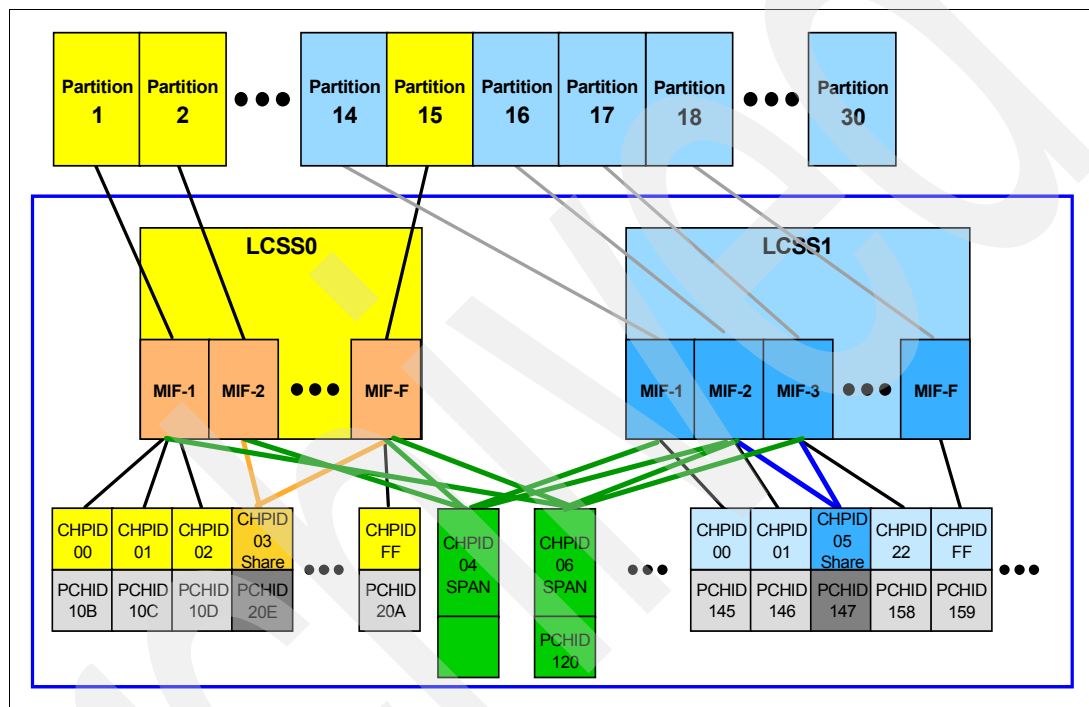


Figure 4-4 z990 CSS - two Logical Channel Subsystems with channel spanning

4.2 LCSS configuration management

Tools are provided to maintain and optimize the I/O configuration of a z990.

IBM Configurator for e-business (e-Config)

The e-Config tool is available to your IBM representative. It is used to configure new configurations or upgrades of existing configuration, and maintains installed features of those configurations.

Hardware Configuration Dialog (HCD)

HCD supplies an interactive dialog to generate your I/O definition file (IODF) and subsequently your Input/Output Configuration Data Set (IOCDS). It is strongly recommended that HCD or HCM be used to generate your I/O configuration, as opposed to writing your own IOCP. The validation checking that HCD performs as you enter data helps eliminate errors before you implement your I/O configuration.

IBM z990 CHPID Mapping Tool (CMT)

The z990 CHPID Mapping Tool provides a mechanism to map CHPIDs onto PCHIDs as required on a z990. Additional enhancements have been built into the CMT to cater for the new requirements of the z; it provides the best availability recommendations for the installed z990 features and defined configuration.

For further details on the CMT, refer to 4.2.1, “z990 configuration management” on page 116 for a brief introduction, and to *IBM @server zSeries Connectivity Handbook*, SG24-5444, for a comprehensive explanation.

4.2.1 z990 configuration management

The architectural enhancements of the z990 enforce a new approach to configuration management. Every CHPID is mapped to a PCHID; it is mandatory that every CHPID has a PCHID associated with it. For internal channels, such as IC links, and HiperSockets, CHPIDs are not assigned a PCHID.

The z990 does *not* have default CHPIDs assigned to channel ports as part of the initial configuration process. CHPIDs are assigned to physical channel path identifiers (PCHIDs) in the IOCP input file. It is the customer's responsibility to perform these assignments by using the HCD/IOCP definitions or by importing the output of the CHPID Mapping Tool.

It is recommended that the CMT be used for all new build z990 configurations, or when upgrading from a z900. It can also be used as part of standard hardware changes to your installed z990.

The Channel Mapping Tool takes input from two sources:

1. The Configuration Report file (CFreport) produced by the IBM order tool (e-Config) can be obtained from your IBM representative, or the Hardware Configuration File produced by IBM manufacturing can be obtained from IBM Resource Link.
2. An IOCP statement file.

The following output is produced by the CMT:

- ▶ Tailored reports. All reports should be saved for reference. The Port Report sorted by CHPID number and location should be supplied to your IBM hardware service representative for the z990 installation.
- ▶ An IOCP input file with PCHIDs mapped to CHPIDs. This IOCP input file can then be migrated back into HCD from which a production IODF can be built.

Important: When an IOCP statement file is exported from a Validated Work IODF using HCD, it must be imported back to HCD for the process to be valid. The IOCP file cannot be used directly by the IOCP program.

The configuration management process is reflected in Figure 4-5 on page 117.

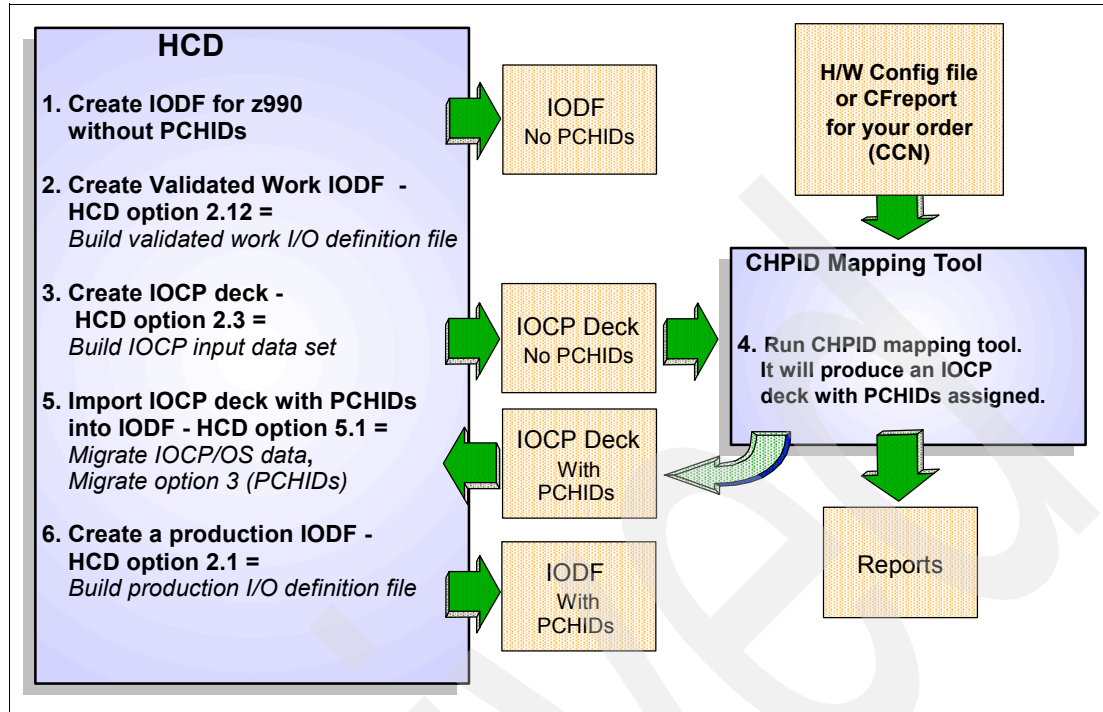


Figure 4-5 z990 I/O configuration definition flow

4.3 LCSS-related numbers

Table 4-3 lists LCSS-related information in terms of maximum values for devices, subchannels, logical partitions, and CHPIDs.

Table 4-3 z990 LCSS at a glance

	z990
Maximum number of LCSSs	4
Maximum number of CHPIDs	1024
Max. number of LPs supported per LCSS	15
Max. number of LPs supported per system	30
Maximum number of HSA subchannels	1890K (63K per partition * 30 partitions)
Maximum number of devices	252K (4 LCSSs * 63K devices)
Maximum number of CHPIDs per LCSS	256
Maximum number of CHPIDs per logical partition	256
Maximum number of devices/subchannels per logical partition	63K

Archived

Cryptography

This chapter describes the Cryptography functions of the z990. On the z990, the Cryptographic Assist Architecture (CAA), along with the CP Assist for Cryptographic Function, offers a balanced use of resources and unmatched scalability.

Included in this chapter are:

- ▶ 5.1, “Cryptographic function support” on page 120”
- ▶ 5.2.1, “CP Assist for Cryptographic Function (CPACF)” on page 122
- ▶ 5.2.2, “PCI Cryptographic Coprocessor (PCIXCC)” on page 123
- ▶ 5.2.3, “PCI Cryptographic Accelerator (PCICA) feature” on page 124
- ▶ 5.3, “Cryptographic hardware features” on page 125
- ▶ 5.4, “Cryptographic features comparison” on page 128
- ▶ 5.5, “Software requirements” on page 129

5.1 Cryptographic function support

The z990 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 delivering the only integrated cryptographic hardware in a server to achieve the US Government's highest FIPS 140-2 Level 4 rating for secure cryptographic hardware.

The z990 cryptographic functions include the full range of cryptographic operations needed for e-business, e-commerce, and financial institution applications. In addition, custom cryptographic functions can be added to the set of functions that the z990 offers.

Today, e-business applications are increasingly relying on cryptographic techniques to provide the confidentiality and authentication required in this environment. Secure Sockets Layer (SSL) 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 and performance.

5.1.1 Cryptographic Synchronous functions

For clear key functions only, the hardware includes implementation of the following:

- ▶ Data encryption/decryption algorithms
 - Data Encryption Standard (DES)
 - Double length-key DES
 - Triple length- key DES (TDES)
- ▶ Hashing algorithms SHA-1
- ▶ Message authentication code (MAC):
 - Single-key MAC
 - Double-key MAC

5.1.2 Cryptographic Asynchronous functions

For secured key functions, Cryptographic Asynchronous functions process messages that are passed to it.

- ▶ Data encryption/decryption algorithms
 - Data Encryption Standard (DES)
 - Double length-key DES
 - Triple length- key DES
- ▶ DES key generation and distribution
- ▶ PIN generation, verification, and translation functions
- ▶ Pseudo Random Number (PRN) Generator
- ▶ Public Key Algorithm (PKA) Facility

These commands are intended for application programs using public key algorithms, including:

- Importing RSA public-private key pairs in clear and encrypted forms
- Rivest-Shamir-Adelman (RSA)

- Key generation, up to 2048-bit
- Signature Verification, up to 2048-bit
- Import and export of DES keys under an RSA key, up to 2048-bit
- Public Key Encrypt (PKE)

Public Key Encrypt service is provided for the Mod_Raised_to Power (MRP) function. MRP is used to offload compute intensive portions of the Diffie-Hellman protocol onto the PCICA or PCIXCC features of the z990.
- Public Key Decrypt (PKD)

Public Key Decrypt supports a zero-pad option for clear RSA private keys. PKD is used as an accelerator for raw RSA private operations, including the use of CRT format keys. The function may be exploited on Linux to allow use of the PCICC and PCIXCC features of the z990 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 PCIXCC feature.
- Europay Mastercard VISA (EMV) 2000 standard

Applications may be written to comply with the EMV 2000 standard for financial transactions between heterogeneous hard- and software. Support for EMV 2000 applies only to the PCIXCC feature of the z990.

Other key functionalities of the PCIXCC serve to enhance the security of public/private key encryption processing:

- ▶ Retained key support (RSA private keys generated and kept stored within the secure hardware boundary)
- ▶ Support for 4753 Network Security Processor migration
- ▶ User Defined Extensions (UDX) support enhancements, including:
 - For 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 program that executes within the PCIX Cryptographic Coprocessor will be supported via an IBM Service Offering.

For unique customer applications, the PCIX Cryptographic Coprocessor will support the loading of customized cryptographic functions on z990. Support is available via ICSF and the z990 Cryptographic Support.

More information can be found in the publication *IBM @server zSeries CCA User Defined Extensions Reference and Guide*, available on the cryptocards Web site:

<http://www.ibm.com/security/cryptocards>

The Web site will direct the customer's request to an IBM Global Services (IGS) location appropriate for the customer's geographic location. A special contract will be negotiated between IGS and the customer, covering development of the UDX by IGS per the customer's specifications, as well as an agreed-upon level of the UDX.

Under a special contract with IBM, PCIX Cryptographic Coprocessor customers will gain the flexibility to define and load custom cryptographic functions themselves. This service offering can be requested via the IBM Cryptocards Web site by selecting the **Custom Programming** option.

5.2 z990 Cryptographic processors

Three types of cryptographic hardware features are available on z990. The cryptographic features are usable only when explicitly enabled through IBM.

- ▶ CP Assist for Cryptographic Function (CPACF)

The CP Assist for Cryptographic Function feature provides hardware acceleration for DES, TDES, MAC, and SHA-1 cryptographic services. Cryptographic keys must be protected by the application system.

- ▶ PCIX Cryptographic Coprocessor (PCIXCC)

The PCIX Cryptographic Coprocessor provides a replacement for both the PCICC and the CMOS Cryptographic Coprocessor Facility (CCF). The PCIXCC on z990 provides equivalent PCICC functions at higher performance. It also includes functions that were implemented in the CCF. The PCIXCC supports highly secure cryptographic functions, use of secure encrypted key values, and user-defined extensions.

- ▶ PCI Cryptographic Accelerator (PCICA)

Secure Web transactions frequently employ the secure Socket Layer (SSL) protocol. The IBM e-business PCI Cryptographic Accelerator offloads your server from compute-intensive public-key cryptographic operations employed in the protocol. This cost-effective solution often enables significantly greater server throughput

5.2.1 CP Assist for Cryptographic Function (CPACF)

Each CP has an assist processor on the chip in support of cryptography. The CP Assist for Cryptographic Function (CPACF) provides high performance hardware encryption and decryption support. To that end, the following five new instructions are introduced with 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

The CP Assist for Cryptographic Function provides high performance hardware encryption and decryption support.

The CP Assist for Cryptographic Function 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 security. The cryptographic architecture includes DES, T-DES data encryption and decryption, MAC message authorization, and SHA-1 hashing. These functions are directly available to application programs, diminishing programming overhead.

The CP Assist for Cryptographic Function complements but does not execute public key (PKA) functions and is a prerequisite for the secure cryptographic operations provided by the PCIX Cryptographic Coprocessor (PCIXCC) feature, and the PCI Cryptographic Accelerator (PCICA) feature. The CP Assist for Cryptographic Function runs at z990 processor speed, and since the facility is available on every CP in the system, there are no affinity issues as in earlier CMOS processors.

The functions of the CP Assist for Cryptographic Function must be enabled or disabled by the manufacturing process to conform to United States export requirements.

5.2.2 PCIX Cryptographic Coprocessor (PCIXCC)

The optional Peripheral Component Interconnect Extended Cryptographic Coprocessor (PCIXCC) provides a high performance cryptographic environment with added function. In fact, the PCIX Cryptographic Coprocessor consolidates the functions previously offered on the z900 by the Cryptographic Coprocessor feature (CCF) and the PCI Cryptographic Coprocessor (PCICC) feature. CCF and PCICC features are not available on the z990. The PCIXCC feature provides asynchronous functions only.

The PCIXCC feature is designed for FIPS 140-2 Level 4 compliance rating for secure cryptographic hardware. Unauthorized removal of the card or feature “zeroizes” its content.

The PCIX Cryptographic Coprocessor features on the z990 enable the user to do the following:

- ▶ Encrypt and decrypt data utilizing secret-key algorithms. Triple-length key DES and double-length key DES algorithms are supported.
- ▶ Generate, install, and distribute cryptographic keys securely using both public and secret key cryptographic methods.
- ▶ Generate, verify, and translate personal identification numbers (PINs).
- ▶ Ensure the integrity of data by using message authentication codes (MACs), hashing algorithms, and Rivest-Shamir-Adelman (RSA) public key algorithm (PKA) digital signatures.

Three methods of master key entry are provided by ICSF for the PCIX Cryptographic Coprocessor features:

1. A pass phrase initialization method that generates and enters all master keys that are necessary to fully enable the cryptographic system in a minimal number of steps.
2. A simplified master key entry procedure provided through a series of Clear Master Key Entry panels from a TSO terminal.
3. In enterprises that require enhanced key-entry security, a Trusted Key Entry (TKE) workstation is available as an optional feature.

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.

The PCIXCC features operate with the Integrated Cryptographic Service Facility (ICSF) and IBM Resource Access Control Facility (RACF®), or equivalent software products, in a z/OS or

OS/390 operating environment to provide data privacy, data integrity, cryptographic key installation and generation, electronic cryptographic key distribution, and personal identification number (PIN) processing.

IBM Processor Resource/System Manager (PR/SM) fully supports the PCIX Cryptographic Coprocessor features to establish a logically partitioned environment in 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 keys, are provided for each of 16 cryptographic domains.

Via the dynamic add/delete of a logical partition name, a logical partition can be renamed: its name can be changed from 'NAME1' to '*' and then changed again from '*' to 'NAME2'. In this case, the logical partition number and MIF ID are retained across the logical partition name change. However, the master keys in PCIXCC that were associated with the old logical partition 'NAME1' are retained. There is no explicit action taken against a cryptographic component for this dynamic change.

Note: Cryptographic cards 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 can assign them to the partitions and clear them if needed.

5.2.3 PCI Cryptographic Accelerator (PCICA) feature

The Peripheral Component Interconnect Cryptographic Accelerator (PCICA) is an orderable feature on z990. This optional feature is a reduced-function, performance-enhanced addition to the CPACF and the PCIX Cryptographic Coprocessor with reduced functional characteristics. It does not have FIPS 140-2 level 4 certification and is non-programmable.

The z990 also supports the optional PCICA. The PCICA feature is used for the acceleration of modular arithmetic operations, in particular the complex RSA cryptographic operations used with the SSL protocol.

This is a unique cryptographic feature for SSL encryption. It has a very fast cryptographic processor designed to provide leading-edge performance of the complex Rivest-Shamir-Adelman (RSA) cryptographic operations used in the SSL protocol. In essence, it is for SSL acceleration rather than for specialized financial applications for secure, long-term storage of keys or secrets. SSL is an essential and widely used protocol in secure e-business applications.

Since the PCI Cryptographic Accelerator is only involved in clear key operations, it does not need the tamper-proof design of the PCIXCC feature.

The PCICA feature provides functions designed for maximum acceleration of the complex RSA cryptographic operations used with the SSL protocol, including:

- ▶ High-speed RSA cryptographic accelerator
- ▶ 1024- and 2048-bit RSA operations for the Modulus Exponent (ME) and Chinese Remainder Theorem (CRT) formats.

The maximum number of SSL transactions per second that can be supported on a z990 by any combination of CPACF, and PCICA features is limited by the amount of cycles available to perform the software portion of the SSL transactions. An IBM 2084 model B16 with 16 CPs active and six PCICA features is designed to provide increased secure Web transaction performance by supporting greater than 11,000 SSL handshakes per second.

In the z990, there can be a maximum of six PCI Cryptographic Accelerator (PCICA) features (two per I/O cage), along with a maximum of four PCIX Cryptographic Coprocessor features. The combined number of PCIXCC and PCICA features on a z990 cannot exceed eight. Within these parameters, the PCIXCC and PCICA features can coexist in any combination. This scalability provides increasing cryptographic processing capacity as customers expand their use of e-business applications requiring cryptographic processing.

5.3 Cryptographic hardware features

This section describes the three cryptographic hardware features and the feature codes associated with the cryptographic functions of the z990.

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 whenever moving, selling, or transferring these products.

5.3.1 PCIX Cryptographic Coprocessor feature

Each PCIX Cryptographic Coprocessor feature contains one cryptographic coprocessor card. The z990 allows for up to four PCIXCC features (or cards) to be installed.

The card is attached to an STI and has no other external interfaces. Removal of the card or feature “zeroizes” the content. A physical layout of the PCIXCC card is shown in Figure 5-1.

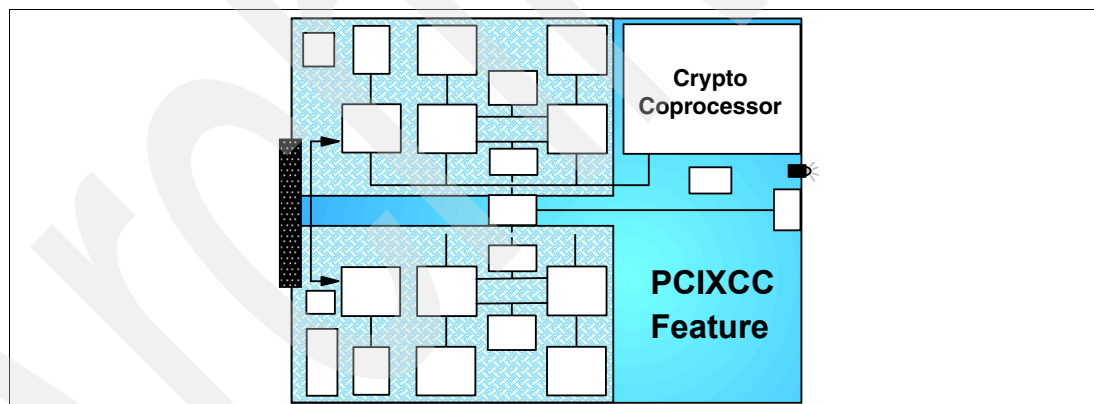


Figure 5-1 PCIX Cryptographic Coprocessor feature

Each PCIXCC feature has one cryptographic coprocessor card embedded in an adapter package for installing in I/O slots of the z990 cage. The PCIXCC feature does not have ports and does not use fiber optic or other cables. The PCIXCC cryptographic coprocessor card can be shared by any logical partition defined in the system up to 15 logical partitions per card.

The PCIX Cryptographic Coprocessor feature does not use CHPID, but is assigned a PCHID.

5.3.2 The PCICA feature

Each PCI Cryptographic Accelerator feature contains up to two cryptographic accelerator cards. The physical layout of the PCICA card is illustrated in Figure 5-2 on page 126.

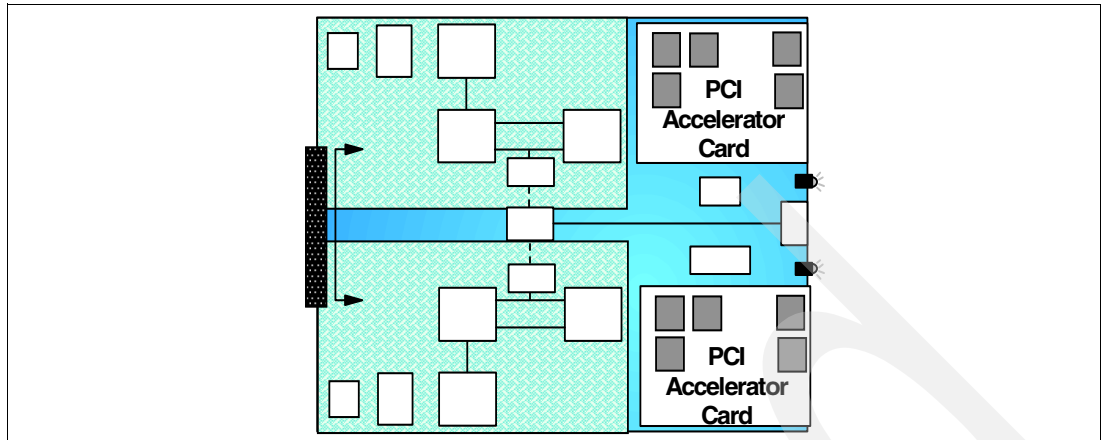


Figure 5-2 PCI Cryptographic Accelerator feature

Each PCICA feature has up to two cryptographic accelerator cards embedded in an adapter package for installing in I/O slots of the z990 cage. The PCICA feature does not have ports and does not use fiber optic or other cables. The PCICA cards can be shared by any logical partition defined in the system up to 15 logical partitions per card, or 30 logical partitions per feature.

The PCICA feature does not use CHPIDs, but is assigned two PCHIDs (one per accelerator card).

Note: While PCICA and PCIXCC features have no CHPID type and are not identified as external spanned channels, all logical partitions in all LCSSs have access to the PCICA feature (up to 15 logical partitions per cryptographic accelerator card). Even so, all logical partitions in all LCSSs have access to the PCIXCC feature (up to 15 logical partitions per feature).

5.3.3 Configuration rules

The following rules apply:

- ▶ The z990 allows for up to two PCICA features per I/O cage. This allows for a maximum of six PCICA features or twelve PCICA coprocessors per z990 server.
- ▶ The maximum number of PCIXCC features (or cryptographic coprocessors) per I/O cage is four; the maximum number of PCIXCC features per system is also four.
- ▶ The total number of cryptographic features may not exceed eight per z990 for any combination of PCIXCC and PCICA features.
- ▶ In addition, any combination of PCIXCC, PCICA, OSA-Express, and FICON-Express features may not exceed 16 features per I/O cage, or 48 features per z990 server.

Table 5-1 on page 127 summarizes support of partitions on z990 for PCICA and PCIXCC features.

Table 5-1 PCI Cryptography features

	Maximum number of features per z990 server	Number of cryptographic coprocessors per feature	Maximum number of cryptographic coprocessors per z990 server	Number of cryptographic domains per coprocessor ^a	Number of logical partitions per z990 server (Defined/Active)
PCICA	6	2	12	16	30/30
PCIXCC	4	1	4	16	30/30

a. Two partitions, defined to the same LCSS or to different LCSSs, can use the same domain number when assigned to different cryptographic coprocessors.

5.3.4 z990 cryptographic feature codes

What follows is a list of the cryptographic features available with the z990.

Feature code	Description
3863	Crypto enablement Crypto enablement feature. Prerequisite to use the CPACF, PCIXCC, and PCICA hardware features.
0868	PCIX Cryptographic Coprocessor (PCIXCC) feature
0862	PCI Cryptographic Accelerator (PCICA) feature
0886	TKE hardware for Token Ring TKE workstation hardware with Token Ring connection, DVD drive, and 17-inch monitor.
0851	TKE 4.0 code
0852	TKE 4.1 code (available May 2004)
0889	TKE hardware for Ethernet TKE workstation hardware with Token Ring connection, DVD drive, and 17-inch monitor.
0851	TKE 4.0 code
0852	TKE 4.1 code (available May 2004)

The z990 requires TKE Version 4.0 or 4.1 code. In case a z900 with a TKE 3.x workstation is upgraded to a z990, the TKE Version 3.x workstations can be retained to control other servers, but cannot be used on the z990. To control the z990, TKE Version 4.0 or 4.1 code is required. TKE 3.x workstations that are carried forward are automatically updated to the 4.1 TKE code when a PCIXCC feature is ordered or present on the sever.

The TKE Version 4.0 or 4.1 code supports all models back to CCF support on a G3 CMOS processor.

TKE 4.1 code provides local and a remote key entry with a master key.

Note: The z990 (and the z890) are the last zSeries servers offering Token Ring adapters on the Trusted Key Entry (TKE) workstations. Timely planning is advised to prepare for migration to the Ethernet environment.

5.3.5 TKE workstation feature

A TKE workstation is part of a customized solution for using the Integrated Cryptographic Service Facility for z/OS program product to manage cryptographic keys of a z990 that has PCIX Cryptographic Coprocessor features installed and configured for using Data Encryption Standard (DES) and Public Key Algorithm (PKA) cryptographic keys.

The TKE workstation provides secure control of the PCIX Cryptographic Coprocessor features, including loading of master keys.

If one or more logical partitions are customized for using PCIX Cryptographic Coprocessors, the TKE workstation can be used to manage DES master keys and PKA master keys for all cryptographic domains of each PCIX Cryptographic Coprocessor feature assigned to logical partitions defined to the TKE workstation.

Each logical partition using a domain managed through a TKE workstation connection is either a TKE host or a TKE target. A logical partition with TCP/IP connection to the TKE is referred to as TKE host; all other partitions are TKE targets.

The cryptographic controls set for a logical partition, through the z990 Support Element, determine whether it can be a TKE host or TKE target.

5.4 Cryptographic features comparison

Table 5-2 summarizes the functions and attributes of the cryptographic hardware features.

Table 5-2 Cryptographic features comparison

Functions or attributes	CPACF	PCIXCC	PCICA
Supports z/OS applications using ICSF	X	X	X
Supports OS/390 applications using ICSF	X	X	X
Encryption and decryption using secret-key algorithm		X	
Provides highest SSL handshake performance			X ⁽¹⁾
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		(2)	(2)
Requires IOCDS definition			
Uses CHPID numbers			
Is assigned PCHIDs		X ⁽⁴⁾	X ⁽⁴⁾
Physically embedded on each Central Processor (CP)	X		
Requires CP Assist for Cryptographic Function enablement	X	X	X ⁽³⁾
Requires ICSF to be active		X	X
Offers user programming function support (UDX)		X	
Usable for data privacy - encryption and decryption processing	X	X	

Functions or attributes	CPACF	PCIXCC	PCICA
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	
Designed for FIPS 140-2 Level 4 certification		X	
Supports SSL functions	X	X	X
Supports Linux applications doing SSL handshakes			X
RSA functions		X	X
High performance SHA-1, Hash function	X		
Clear key DES/T-DES	X		
Clear key RSA		X	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

Notes

1. Requires CPACF enablement.
2. In order to make addition of PCIXCC and PCICA features nondisruptive, the logical partition must be predefined with the appropriate PCI cryptographic processor number selected in its candidate list, on the partition image profile.
3. Not required for Linux
4. PCIXCC is assigned one PCHID per feature. PCICA is assigned two PCHIDs per feature (one per accelerator processor).

5.5 Software requirements

Cryptographic support for z990 for OS/390 V2.10, z/OS V1.2 and later is made available as a Web deliverable.

z990 Cryptographic Support is planned to be available through May 28, 2004. It is a Web deliverable to provide exploitation support for the CPACF, PCICA, and PCIXCC features on behalf of OS/390 V2.10 and z/OS V1.2 and later releases. On May 28, 2004, this support is replaced by *z990 and z890 Enhancements to Cryptographic Support*.

z990 and z890 Enhancements to Cryptographic Support is planned to be available on May 28, 2004. It is a Web deliverable to provide exploitation support for CPACF, PCICA, and PCIXCC

features on behalf of OS/390 V2.10, z/OS V1.3, z/OS V1.4, and z/OS V1.5, and in addition supports the following functions:

- ▶ Double length Derived Unique Key Per Translation (DUKPT) on PCIXCC
- ▶ EMV 2000 Standard on PCIXCC
- ▶ Public Key Decrypt (PKD) enhancements on PCICA and PCIXCC
- ▶ Public Key Encrypt (PKE) enhancements on PCICA and PCIXCC

The Web deliverable is found at:

<http://www.ibm.com/eserver/zseries/zos/downloads>

The CP Assist for Cryptographic Function (CPACF), PCIX Cryptographic Coprocessor (PCIXCC), and PCI Cryptographic Accelerator (PCICA) features have specific software requirements.

The Integrated Cryptographic Service Facility (ICSF) is the support program for the cryptographic features CPCPACF, PCIXCC, and PCICA. ICSF is integrated into z/OS.

The minimum cryptographic software requirements are:

- ▶ CP Assist for Cryptographic Function (CPACF):
 - OS/390 V2.10 or z/OS V1.2 and later with z990 Cryptographic Support, or z990 and z890 Enhancements to Cryptographic Support.
 - z/VM 3.1 and V4.3 and later.
 - Linux distributions with the most recent cryptographic libraries, found at:
<http://www-124.ibm.com/developerworks/projects/libica>
- ▶ PCI Cryptographic Accelerator (PCICA)
 - OS/390 V2.10 or z/OS V1.2 and later with z990 Cryptographic Support, or z990 and z890 Enhancements to Cryptographic Support.
 - z/VM V5.1 for z/OS and Linux guests.
 - z/VMV4.3 and later for Linux guests.
 - Linux for zSeries: Red Hat RHEL 3.0 with Cryptographic modules, SUSE SLES 7, Turbolinux TLES8, and Conectiva CLEE.
- ▶ PCIX Cryptographic Coprocessor
 - OS/390 V2.10 or z/OS V1.2 and later with z990 Cryptographic Support, or z990 and z890 Enhancements to Cryptographic Support.
 - z/VM V5.1 for z/OS and Linux guests.
 - Including dedicated queue support for secure key and clear key cryptographic functions for z/OS guests
 - Including shared queue and dedicated queue support for clear key cryptographic functions for Linux guests
 - Linux on zSeries support is delivered as an Open Source contribution. See:
<http://www10.software.ibm.com/developerworks/opensource/linux390/index.shtml>
- ▶ PCIX Cryptographic Coprocessor User-Defined Extensions (UDX)
 - OS/390 V2.10 or z/OS V1.2 and later with z990 Cryptographic Support, or z990 and z890 Enhancements to Cryptographic Support. Table 5-2 on page 126 summarizes the software support requirements by operating system.

Table 5-3 Software requirements to support cryptographic features

Operating system	CPACF	PCIXCC	PCICA
OS/390 V2.10 and z/OS V1.2 and later with z990 Cryptographic Support or z990 and z890 Enhancements to Cryptographic Support	Y	Y	Y
z/VM V3.1 and V4.3 and later	Y		
z/VM V4.3 and later for Linux guests	Y		Y
z/VM V5.1 for z/OS and Linux guests	Y	Y	Y
Linux on zSeries	Y	Y	Y
z/VSE V3.1			Y
VSE/ESA V2.7 and later			Y

Archived

Software support

This chapter describes the software support available on the z990, including z/OS, OS/390, z/VM, z/VSE, VSE/ESA, TPF, and Linux operating systems. Software migration considerations and workload license charges are also discussed. The following topics are included:

- ▶ 6.1, “Operating system support” on page 134
- ▶ 6.2, “z/OS software support” on page 134
- ▶ 6.3, “z/VM software support” on page 145
- ▶ 6.4, “z/VSE and VSE/ESA software support” on page 146
- ▶ 6.5, “TPF software support” on page 146
- ▶ 6.6, “Linux software support” on page 147
- ▶ 6.7, “Summary of software requirements” on page 147
- ▶ 6.8, “Workload License Charges” on page 150
- ▶ 6.9, “Concurrent upgrades considerations” on page 151

6.1 Operating system support

There are many significant changes in the z990 architecture and hardware features when compared to the z900 processor. Extensive software support has been made available to existing OS levels via compatibility and Exploitation Support to accommodate these changes in the OS/390, z/OS, z/VSE, VSE/ESA, TPF, z/VM and Linux on zSeries operating systems. Table 6-1 summarizes supported software on the z990.

Table 6-1 z990 software support summary

Operating system	ESA/390 (31-bit)	z/Arch. (64-bit)	Compatibility	Exploitation
OS/390® Version 2 Release 10	Yes	Yes	Yes	No
z/OS Version 1 Release 2	No ^a	Yes	Yes	No
z/OS Version 1 Release 3	No ^a	Yes	Yes	No
z/OS Version 1 Release 4	No ^a	Yes	Yes	Yes
z/OS Version 1 Release 5 and 6 ^b	No	Yes	Included	Included
Linux for S/390	Yes	No	Yes	Yes
Linux® on zSeries	No	Yes	Yes	Yes
z/VM Version 3 Release 1	Yes	Yes	Yes	No
z/VM™ Version 4 Release 3	Yes	Yes	Yes	No
z/VM Version 4 Release 4	Yes	Yes	Included	Included
z/VM Version 5 Release 1	No	Yes	Included	Included
VSE/ESA™ Version 2 Release 6 and 7	Yes	No	Yes	Yes
z/VSE Version 3 Release 1 ^c	Yes	No	Included	Included
TPF Version 4 Release 1 (ESA mode only)	Yes	No	Yes	No

a. 31-bit mode is only available as part of the z/OS Bimodal Migration Accommodation software program. The program is intended to provide fallback support to 31-bit mode in the event that it is required during migration to z/OS in z/Architecture mode (64-bit).

b. z/OS 1.6 is planned to be available in September 2004.

c. The z/VSE operating system can execute in 31-bit mode only. It does not implement z/Architecture, and specifically does not implement 64-bit mode capabilities. The z/VSE operating system is designed to exploit select features of IBM @server zSeries hardware.

6.2 z/OS software support

z/OS software support has been designed at two levels: *Compatibility Support* and *Exploitation Support*.

6.2.1 Compatibility Support for z/OS

Compatibility Support for z/OS software is delivered in several ways, depending on the version of release of z/OS.

z990 compatibility for selected OS/390 and z/OS releases

OS/390 V2.10, z/OS V1.2, and z/OS V1.3 require the Web delivered Compatibility Support to run on a z990.

Attention: Compatibility and Exploitation Support is *not* available for z/OS 1.1

Compatibility Support allows these releases to:

- ▶ Define a z990 environment with HCD
- ▶ Run on a z990 processor in a logical partition in LCSS-0, using an LPAR ID equal to or less than x'F'
- ▶ Coexist in a sysplex that contains a z990 processor
- ▶ Coexist with z990 processors sharing disk devices outside of a sysplex

Compatibility Support requirements

Compatibility Support is required under the following circumstances:

- ▶ Compatibility Support is required for *all* images running on a z990.
- ▶ Compatibility Support is required on *any* image that is used for defining the I/O configuration for the z990.
- ▶ Compatibility Support is required on *all* images in a sysplex, whether running on a z990 or not, if a Coupling Facility logical partition for that sysplex is running on a z990 and has an LPAR identifier greater than 15 (x'F').

Compatibility Support allows the supported operating systems to run in LCSS 0 on the z990 processor. It is not possible to run z/OS or OS/390 in LCSSs 1, 2, and 3, even with the compatibility maintenance. This is supported on z/OS 1.4 with the Exploitation Support and subsequent releases.

Note: A Coupling Facility logical partition can reside in any Logical Channel Subsystem.

Compatibility Support can run on any processor that is already supported by one of the listed operating systems. For example, you can install Compatibility Support on a z/OS V1.2 system that is running on an IBM 9672 G5 processor.

Figure 6-1 on page 136 shows a situation where the Compatibility Support is *not* required for systems not on the z990.

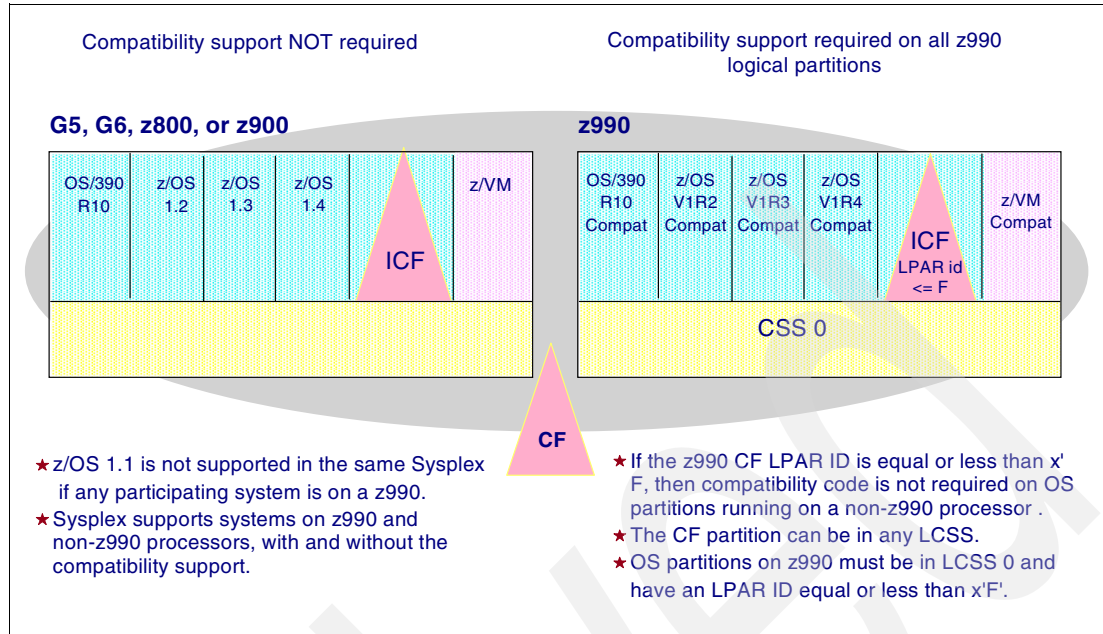


Figure 6-1 Conditions where compatibility maintenance is not required

Figure 6-2 shows the situation where the Compatibility Support *is* required.

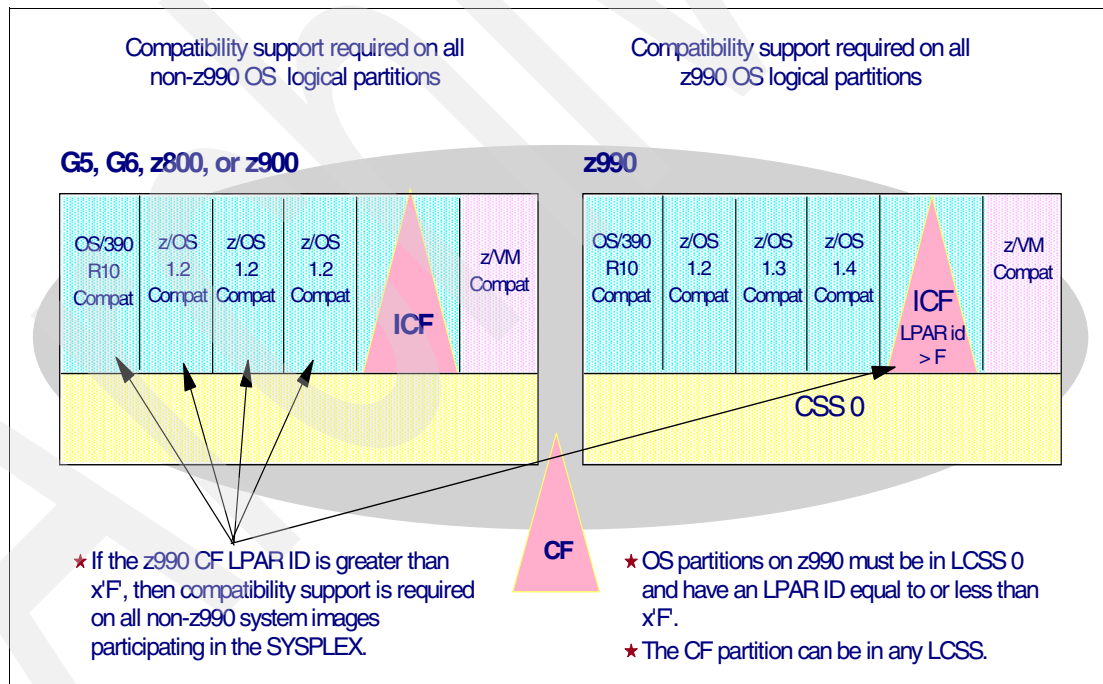


Figure 6-2 Conditions where compatibility maintenance is required

Compatibility Support restrictions

Compatibility Support does not allow you to make full use of all the capabilities of the z990. Some restrictions that apply when running in this mode are:

- z/OS with Compatibility Support must be IPLed in a partition defined to LCSS 0. If it is IPLed in a partition in another LCSS, it will terminate with a 07C-01 wait state.

- ▶ z/OS with Compatibility Support must be IPLd in a partition that has an LPAR identifier in the range 0-F. If the LPAR identifier is outside of this range, then z/OS will terminate with a 07C-02 wait state.
- ▶ Dynamic activates for hardware changes can only be done for LCSS 0. A Power-on Reset is required for changes to other Logical Channel Subsystems. Dynamic activate for hardware changes within LCSS 0 cannot be done if the resource is also defined in any other LCSS. For example, if a DASD control unit has connections to LCSS 1, 2, or 3, then additional connections cannot be added to LCSS 0 dynamically while in compatibility mode. Software activates can be done regardless of the number of LCSSs defined.

z/OS V1.4 z990 Compatibility Support feature

The Compatibility Support feature is an optional, unpriced orderable feature that is required to allow z/OS Version 1 Release 4 to run on a z990. It offers the capability to execute on the hardware in a way that is compatible with earlier processor models, but without exploitation of new functions, although some new features (such as the increased number of HiperSockets) can be used with this z/OS Version. This feature is no longer orderable and has been replaced by the z/OS V1.4 z990 Exploitation Support feature since February 24, 2004.

6.2.2 Exploitation Support for z/OS

Exploitation Support for the z990 is delivered for z/OS V1.4. Follow-on z/OS releases will have Exploitation Support included. For z/OS V1.4, it is shipped as a separately orderable feature.

z/OS Exploitation Support allows:

- ▶ z/OS to run in a partition defined to any Logical Channel Subsystem
- ▶ z/OS to run in a partition with an LPAR identifier greater than x'F'
- ▶ z/OS dynamic activation for hardware changes to any Logical Channel Subsystem

z/OS V1.4 z990 Exploitation Support feature

The Exploitation Support feature is an optional unpriced orderable feature that provides Exploitation Support for two Logical Channel Subsystems and 30 logical partitions. This is a mandatory feature when ordering z/OS V1.4 as of February 24, 2004.

Other exploitation items are:

- ▶ Dynamic I/O support to dynamically add, change, and delete channel paths, control units, and devices in multiple Logical Channel Subsystems
- ▶ Up to four Logical Channel Subsystems, with HCD PTFs
- ▶ External spanned channels

z/OS V1.5 support

Exploitation Support (and Compatibility Support) is integrated into the base of z/OS V1.5 and later releases. Exploitation Support in z/OS V1.5 offers:

- ▶ Up to 30 Logical Partitions
- ▶ Up to four Logical Channel Subsystems, with HCD PTFs

Planned z/OS V1.6 support

Exploitation Support (and Compatibility Support) is integrated into the base of z/OS V1.6 (planned to be available in September 2004). Some unique support offered by this release is:

- ▶ Dynamic addition and deletion of a logical partition name
- ▶ 24 processors (sum of CPs and zAAPs) within a single logical partition
- ▶ zSeries Application Assist Processors (zAAPs)

Dynamic addition and deletion of a logical partition name

z/OS V1.6 supports dynamic naming of a reserved logical partition. Reserved logical partitions are defined with a name placeholder ‘ * ’ and can be dynamically named or removed from the list of named logical partitions.

A dynamic partition must be reserved in the IOCDS and will be established when a Power-On Reset with this IOCDS is executed. A reserved partition has a MIF ID and usage type assigned.

24 processors within a single logical partition

z/OS V1.6 supports up to 24 processors (the sum of CPs and zAAPs). Note that the sum of initial and reserved processors, including CPs and zAAPs, for an ESA/390 mode logical partition can go up to 32 processors.

zSeries Application Assist Processor (zAAP)

Support for the zSeries Application Assist Processor is introduced in z/OS V1.6. This z/OS release is planned to be available in September 2004.

Note: zAAPs are not supported for a z/OS guest under z/VM.

A zAAP reduces the standard processor (CP) capacity requirements for Java 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 fee.

zAAPs only run Java code. The IBM SDK for z/OS Java 2 Technology Edition (the Java Virtual Machine), in cooperation with z/OS and PR/SM, directs JVM processing from CPs to zAAPs. Apart from the cost savings this may realize, the integration of Java based applications with their associated data base systems such as DB2, IMS, or CICS®, may simplify the infrastructure, for example, reducing the number of TCP/IP programming stacks and server interconnect links. Furthermore, processing latencies that would occur if Java application servers and their data base servers were deployed on separate server platforms are prevented.

Figure 6-3 on page 139 shows the logical flow of Java code running on a z990 server that has a zAAPs available. The Java Virtual Machine (JVM), when it starts execution of a Java program, passes control to the z/OS dispatcher that will verify the availability of a zAAP:

- ▶ If a zAAP is available (not busy), the dispatcher will suspend 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 will reassign the JVM code execution to a CP.
- ▶ If there is no zAAP available 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).

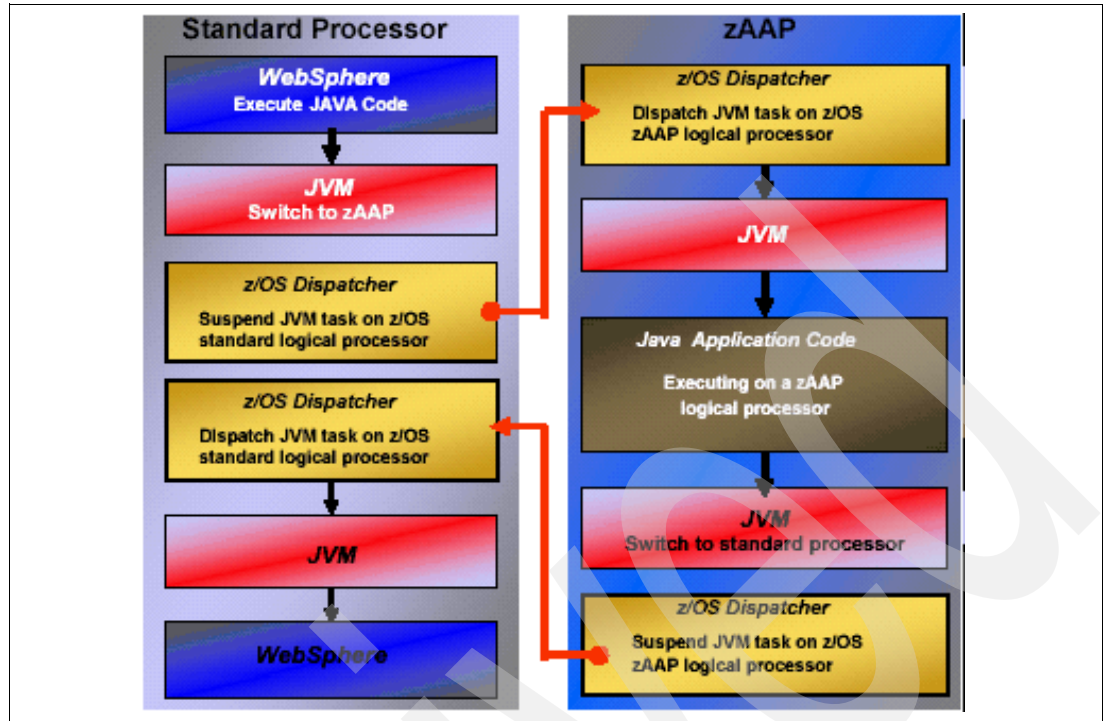


Figure 6-3 Logical flow of Java code execution on a zAAP

zAAPs do not affect the overall MSU or capacity rating of a system or logical partition, that is, adding zAAPs to a system or defining to a logical partition does not affect the software license charges. There is no additional z/OS charge for zAAPs.

Subsystems that exploit zAAPs are:

- ▶ WebSphere Application Server V5.1
- ▶ CICS/TS V2.3
- ▶ DB2 V8
- ▶ IMS V8
- ▶ WebSphere WBI for z/OS

The functioning of a zAAP is transparent to all IBM and ISV Java programming on JVM 1.4.1 and above.

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.

Option 1: Java dispatching by priority (honor_priority=yes)

Option 1 is the default option and specifies that standard CPs execute both Java and non-Java work in priority order when zAAPs are not configured. When zAAPs are configured, they only execute Java work in priority order while the CPs execute normal tasks and JVM tasks in priority order. This option is oriented towards servicing work with the highest priority first, regardless of the type of work.

Option 2: Java discretionary crossover (crossover=yes)

Standard CPs execute Java work in priority order only if no non-Java (standard) work is available to be dispatched. This way, Java work may execute on a CP as if it has a lower priority than non-Java work. This option is oriented towards environments where not enough

zAAP capacity is available and the Java work has no need for priority over non-Java work. When executed on a zAAP, Java work is handled in priority order.

Option 3: No Java crossover (crossover=no)

This option is set to prevent Java work being executed on a CP. If, for example, Sub Capacity Workload License Charging is applicable, Java work that is executed on a CP will increase CP utilization and consequently may increase the software charges for non-Java work. Also, if ample zAAP capacity is available, this option assures that all Java work is done on a zAAP. Only if the last available zAAP would fail is crossover to a standard CP enforced.

6.2.3 HCD support

HCD support for both Compatibility Support and Exploitation Support allows the definition of the z990 processor and I/O configuration from an existing system. If the z990 processor is part of a S/390 microprocessor cluster, the HCD support also allows the IOCDS to be written directly from an existing system. The new HCD elements apply to z/VM and z/OS.

HCD uses a concept: the validated work IODF. This is a new status for an IODF data set. It contains a complete set of validated processor, LCSS, partition, channel, control unit, and I/O device definitions. A validated work IODF would not normally contain the physical channel identifiers (PCHIDs) for channels.

Closely allied with this function, support is added to allow HCD to work with the z990 CHPID Mapping Tool (CMT) to assign PCHIDs. This support allows an IOCP source statement data set to be created from a validated work IODF; it also allows the data from the CHPID Mapping Tool to be merged with the validated work IODF to complete the PCHID assignments.

Compatibility Support for HCD allows an installation to do the following:

- ▶ Define a z990 environment with multiple Logical Channel Subsystems.
- ▶ Make dynamic hardware changes to LCSS 0 only. Devices cannot be added, modified, or deleted if they are also defined to another LCSS. This requires a Power-on Reset.

Exploitation Support for HCD has full dynamic change support for all LCSSs.

6.2.4 Automation changes

Several commands and messages have been adapted to accommodate the two-digit LPAR identifier.

6.2.5 SMF support

CPU and PR/SM activity data is recorded by RMF in the SMF type 70 subtype 1 records. Prior to the z990, these records were always shorter than 32 KB. With the increased number of logical partitions and logical CPs, these records could potentially increase in size beyond the 32 KB limit.

To accommodate this, each record is now broken into pieces where each piece is shorter than 32 KB. Each piece is self-containing, that is, the record can be processed without re-assembling the broken pieces. If you have any site-specific processing of this data outside of RMF, you may need to review that application to ensure that it is no longer dependent upon all this data being contained within a single record.

RMF Monitor 1 Device Activity reporting is recorded in the SMF 74, subtype 1 records. These have been updated to support an extended device data section. This section now includes the

Initial Command Response time for the device. A similar change has also been made to the RMF Monitor II Device Activity recording in the SMF 79, subtype 9 records.

The SMF records for the SRM Decisions data stored in type 99 (subtypes 8 and 9) have also been extended. These records now contain the LCSS ID for the WLM LPAR management and I/O subsystem information.

6.2.6 RMF support

There are several changes to RMF reports to accommodate the enhanced I/O subsystem and improved collection of channel measurement data.

The I/O Activity report no longer shows the Control Unit Busy (CUB) and Director Port Busy (DPB) times. (The corresponding percentage and pending reason fields for CUB and DPB have also been removed from the Monitor III reports.) This information was already available at an LCU level anyway and is much more useful than the figures being broken out for individual devices.

Furthermore, the Director Port Busy field would only show a non-zero value for the events when *all* director ports were busy. If an individual director port was found to be busy, but a connection was established through an alternate path, then this figure was not updated. With FICON connections, a Director Port was never reported busy, since that type of channel allows multiple data transfers to occur simultaneously.

A better measure of fabric contention has now been provided with the Initial Command Response Time. This is a measure of the time taken from sending a command to a device, to it responding that it has accepted the command. This new metric (AVG CMR DLY) has now replaced the older AVG CUB and AVG DPB columns on the Monitor I Device Activity Report. Corresponding DELAY CMR% and PENDING REASON CMR displays have been introduced into the Monitor III reports. Monitor III exception reporting has also been updated to replace the previous CUBDL and DPBDL conditions with the new condition CMRDL.

6.2.7 ICKDSF requirements

ICKDSF Release 17 is required on all systems that share DASD with a z990 processor.

ICKDSF 17 supports the new format of the CPU information field, which now contains a two-digit LP 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. In order to prevent any possible data corruption, ICKDSF must be able to determine all sharing systems that may potentially run ICKDSF; therefore, this support is required for z990.

Important: The need for ICKDSF Release 17 applies even to systems that are not part of the same sysplex, or that are running a non-MVS based operating system, such as z/VM.

6.2.8 ICSF support

If you use z990 cryptographic hardware functions with ICSF, then you must install Compatibility Support for this feature.

Important: The ICSF Compatibility Support is available for OS/390 V2.10, z/OS V1.3, and V1.4. Later releases have this support incorporated.

If you are running on an earlier level of OS/390 or z/OS, then you *must* upgrade the logical partition calling ICSF to either OS/390 V2.10, z/OS V1.3, or V1.4 before you can move that environment onto a z990 processor.

The PCIXCC features are the cryptographic hardware for the z990 that support Secure Keys. Clear Key support is provided by the PCICA cards and the CP Assist for Cryptographic Function.

6.2.9 Additional Exploitation Support considerations

The following areas need to be considered when running z/OS in exploitation mode.

SMF

The SMF type 89 record is used for recording Product Usage data and is extended for Exploitation mode support. A new field, SMF89LP3, allows an 8-bit LPAR ID to be stored. This field is marked valid by the new flag bit SMF89LPM. When an LPAR ID is less than or equal to x'F', the LPAR ID is stored in both the new field and the old 4-bit SMF89LP2 field to maintain compatibility.

Stand-alone dump

z/OS systems that have Exploitation Support installed must generate a new version of the stand-alone dump program. This stand-alone dump program cannot be used for dumping systems at earlier releases of z/OS or z/OS V1.4 systems that have only the Compatibility Support installed.

Automation

The output from the D M=CPU command shows the two-digit LPAR ID (set in the image profile), the LCSS ID associated with the logical CPUs associated with the logical partition in that LCSS, and the MIF ID (see Figure 6-4 on page 143)s. The logical CPU address no longer appears in the first digit of the serial number, as a result of the change to the STIDP instruction.

```

D M=CPU
IEE174I 14.45.55 DISPLAY M 159
PROCESSOR STATUS
ID  CPU              SERIAL
0   +               1293052084
1   +               1293052084
2   +               1293052084

CPC ND = 002084.R01.IBM.02.000000049305
CPC SI = 2084.R01.IBM.02.000000000049305
CPC ID = 00
CPC NAME = XXXXXXXX
LP NAME = SC66, LP ID = 12
CSS ID   = 1
MIF ID   = D

+ ONLINE  - OFFLINE  . DOES NOT EXIST
CPC ND    CENTRAL PROCESSING COMPLEX NODE DESCRIPTOR
CPC SI    SYSTEM INFORMATION FROM STSI INSTRUCTION
CPC ID    CENTRAL PROCESSING COMPLEX IDENTIFIER
CPC NAME  CENTRAL PROCESSING COMPLEX NAME...

```

Figure 6-4 D M=CPU command output

The output from the D IOS,CONFIG(HSA) and D IOS,CONFIG(ALL) commands no longer has references to SHARED and UNSHARED control units. Previously, this command would have been used to determine the HSA space available for dynamically adding control units and I/O devices. On the z990 processor, the number of additional devices that can be added dynamically is determined by the MAXDEV value associated with each LCSS. This parameter is specified via HCD and is set in the IOCDS. The new output from the D IOS command is shown in Figure 6-5.

```

D IOS,CONFIG(HSA)
IOS506I hh.mm.ss I/O CONFIG DATA
HARDWARE SYSTEM AREA AVAILABLE FOR CONFIGURATION CHANGES
PHYSICAL CONTROL UNITS                50
CSS 0 - LOGICAL CONTROL UNITS          100
SUBCHANNELS                          2000
CSS 1 - LOGICAL CONTROL UNITS          120
SUBCHANNELS                          3240

```

Figure 6-5 Output from the Display IOS command on z990

The IEE174I, IOS050I, and IOS051I messages display not only the CHPID number, but also the associated PCHID. This addition assists IBM service representatives and operations staff with the diagnosis of hardware problems. The new output is shown in Figure 6-6 on page 144.

```

D M=CHP(58)
IEE174I 17.00.35 DISPLAY M 263
CHPID 58: TYPE=1B, DESC=FICON SWITCHED, ONLINE
DEVICE STATUS FOR CHANNEL PATH 58
    0 1 2 3 4 5 6 7 8 9 A B C D E F
022 + . . . . . . . . . . . . . .
023 + . . . . . . . . . . . . . .
098 . + + . . . . . . . . . . . .
SWITCH DEVICE NUMBER = B000
DEFINED ENTRY SWITCH - LOGICAL SWITCH ID = 03
ATTACHED ND = 006064.001.MCD.01.000000040012
PHYSICAL CHANNEL ID = 034D
***** SYMBOL EXPLANATIONS *****
+ ONLINE    @ PATH NOT VALIDATED  - OFFLINE    . DOES NOT EXIST
* PHYSICALLY ONLINE  $ PATH NOT OPERATIONAL

IOS050I CHANNEL DETECTED ERROR ON dddd,yy,op,stat,PCHID=pppp
IOS051I INTERFACE TIMEOUT DETECTED ON dddd,yy,op,stat,PCHID=pppp

```

Figure 6-6 PCHID support for channel messages

EREP

The PCHID value associated with a particular CHPID is shown in the EREP Subchannel Logout record, as can be seen in Figure 6-7.

DEVICE NUMBER:	0A02	REPORT:	SLH EDIT	DAY YEAR	JOB			
		SCP:	VS 2 REL. 3	DATE:	258 02			
DEVICE TYPE:	CACA							
		CPU MODEL:	2084	HH MM SS.TH				
CHANNEL PATH ID:	3E	LOGICAL CPU ID:	232920	TIME:	12 50 29.84			
PHYSICAL CHAN ID:	XXXX	PHYSICAL CPU ID:	612920	PHYSICAL CPU ADDRESS:	00			
	CC	CA	FL	CT				
FAILING CCW	02	01DA2500	24	5000	VOLUME SERIAL			
					SUBCHANNEL ID NUMBER			
	K	FLAGS	CA	US	SS	CT	ERROR TYPE	
SCSW	64	C24017	01DA2318	00	02	5000		OTHER
---UNIT STATUS--- SUB-CHANNEL STATUS -----SCSW FLAGS-----								
							FLAG 0	FLAG 1
								FLAG 2

Figure 6-7 EREP support of PCHIDs

Extended Channel Measurement Block

The z990 processor supports the new Extended Channel Measurement Blocks. It also supports the original XA I/O architecture CMBs for compatibility reasons.

In z/OS with Exploitation Support, the ECMBs are placed in a system common area dataspace. You should, therefore, review the setting of the MAXCAD value in IEASYSxx. The CMB parameter in IEASYSxx is now redundant and will be ignored.

Dynamic activation for hardware changes

If a z990 processor running z/OSs with exploitation and Compatibility Support and a new hardware configuration is to be activated, the H/W activation must be done on one of the z/OS systems that run with Exploitation Support if both of the following conditions exist:

- ▶ There is more than one Logical Channel Subsystem defined.
- ▶ A non-zero LCSS is being changed, or resources affected by the change are defined to non-zero CSSs.

Dynamic CHPID management

On a z990, systems that are part of the same LPAR cluster may be in different LCSSs.

Dynamic CHPID management is supported even if the LPAR cluster spans multiple Logical Channel Subsystems. Movement of CHPIDs within the LPAR cluster is confined to movement within that LCSS, since ESCON channels cannot span multiple Logical Channel Subsystems.

If an LPAR cluster spans multiple LCSSs, then the DCM command VARY SWITCH must be issued from one of the systems with Exploitation Support. The DCM command SETIOS DCM=ON/OFF can be issued from any system, whether in compatibility mode or exploitation mode.

Greater than 15 logical partitions

You must define more than one LCSS if you plan to use more than 15 logical partitions. An individual Logical Channel Subsystem can only support up to 15 logical partitions.

In z/OS V1.2 and above and OS/390 V2.10, logical partitions running in compatibility mode can only reside in LCSS 0.

Coupling Facilities, z/VM V4R4 and follow-on releases, and Linux on zSeries can reside in any Logical Channel Subsystem.

6.3 z/VM software support

Compatibility Support for z/VM on the z990 is provided for z/VM V3R1 and V4R3 (it is available through the normal service stream for z/VM). See the appropriate PSP bucket.

z/VM Compatibility Support is almost identical to that for z/OS in the function that it provides. The z990 Compatibility Support allows:

- ▶ Support for up to four LCSSs
- ▶ Dynamic I/O support for LCSS 0 only
- ▶ Up to 15 Logical Partitions (LPAR identifiers < 15 (x'F'))
- ▶ Internal and External spanned channels

For z/VM V4R4 and z/VM V5R1, Compatibility and Exploitation Support is included. Some items of interest are:

- ▶ Up to 30 Logical Partitions
- ▶ Up to four LCSSs
- ▶ Dynamic I/O support for LCSS 0, 1, 2, and 3
- ▶ 24 processors (all CPs or all IFLs) within a single logical partition (in z/VM 5.1 only)

- Note that the sum of initial and reserved processors for an ESA/390 mode logical partition can go up to 32 processors.
- ▶ Adapter Interruption applying to performance assist for FICON Express features (type FCP) and OSA-Express features (type OSD)
- ▶ V=V guest support applying to performance assist for FICON Express features (type FCP), OSA-Express features (type OSD), and HiperSockets (type IQD)

See also 6.2.3, “HCD support” on page 140 for I/O configuration support under z/VM.

ICKDSF Release 17 is required on all systems that share DASD with a z990 processor (see 6.2.7, “ICKDSF requirements” on page 141).

6.4 z/VSE and VSE/ESA software support

z/VSE and VSE/ESA support for z990 is provided on VSE/ESA V2.6, VSE/ESA V2.7, z/VSE V3.1, and later. For the latest information on support requirements, refer to the VSE subset of the 2084DEVICE Preventative Service Planning (PSP) bucket prior to installing the z990 server.

Some z/VSE and VSE/ESA supported functions are:

- ▶ Support for up to 30 logical partitions
- ▶ Up to four Logical Channel Subsystems
- ▶ Spanned channel support (in VSE/ESA V2.7 and z/VSE V3.1 for HiperSockets CHPID type IQD only)
- ▶ Adapter Interruption applying to performance assist (in VSE/ESA V2.7 and z/VSE V3.1 for OSA-Express CHPID type OSD)
- ▶ HiperSockets support for CHPID type IQD (in VSE/ESA V2.7 and z/VSE V3.1 only)

z/VSE V3R1 is planned to have support for the Fiber Channel Protocol (FCP attached to a SCSI disk).

Important: The z/VSE operating system can execute in 31-bit mode only. It does not implement z/Architecture, and specifically does not implement 64-bit mode capabilities. The z/VSE operating system is designed to exploit select features of IBM @server zSeries hardware.

ICKDSF Release 17 is required on all systems that share DASD with a z990 processor (see 6.2.7, “ICKDSF requirements” on page 141).

6.5 TPF software support

TPF support for z990 is provided by TPF Version 4.1, which also includes the support for up to 30 logical partitions.

ICKDSF Release 17 is required on all systems that share DASD with a z990 processor (see 6.2.7, “ICKDSF requirements” on page 141).

6.6 Linux software support

The currently available distributions SUSE SLES 7, SUSE SLES 8, Red Hat 7.1 and Red Hat RHEL 3.0, Turbolinux TLES 8, and Conectiva CLEE support compatibility and exploitation of 30 logical partitions and two LCSSs. Support to further exploit z990 functions will be delivered as an Open Source Contribution via:

<http://www.software.ibm.com/developerworks/opensource/linux390>

Commercial distributions of Linux on zSeries are available from Linux distributors, such as Red Hat, SUSE, and Turbolinux. To learn more about distributor offerings, contact these distributors through their representatives or through the following Web sites:

- ▶ Red Hat
<http://www.redhat.com>
- ▶ SUSE
<http://www.suse.com>
- ▶ Turbolinux
<http://www.turbolinux.com>

z990 Exploitation Support is also delivered by developerWorks®. For details regarding functions that are not yet available via distributor offerings, information on Linux support for FCP, VLAN, IPv6, SNMP, TCP/IP Broadcast, Query ARP, and Purge ARP, refer to:

<http://www.ibm.com/developerworks>

For Linux support, visit the Web site:

<http://www.software.ibm.com/developerworks/opensource/linux390>

6.7 Summary of software requirements

The following tables provide summaries of software requirements for z990 functions and features. Software requirements for the Cryptographic functions of the z990 are found in 5.5, “Software requirements” on page 129.

6.7.1 Summary of z/OS and OS/390 software requirements

Table 6-2 Minimum z/OS and OS/390 software requirements

Functions	Software Requirements	z/OS V1.6 ^a	z/OS V1.5	z/OS V1.4 Exploit	z/OS V1.3	z/OS V1.2	OS/390 V2.10
16 to 30 logical partitions		X	X	X			
Two Logical Channel Subsystems (LCSSs)		X	X	X			
Four Logical Channel Subsystems (LCSSs)		X	X ^b	X ^b			
Dynamic I/O support for multiple LCSSs		X	X	X			
Dynamic Add/Delete Logical Partition Name		X					
Extended Translation Facility		X					
zSeries Application Assist Processor (zAAP)		X ^c					
24 processors within a single logical partition		X					

Software Requirements	z/OS V1.6 ^a	z/OS V1.5	z/OS V1.4 Exploit	z/OS V1.3	z/OS V1.2	OS/390 V2.10
Functions						
Internal spanned channels	X	X	X	X	X	
External spanned channels	X	X	X			
HiperSockets	X	X	X	X	X	
Broadcast for IPv4 packets	X	X				
16-port ESCON feature	X	X	X	X	X	X
FICON Express (type FCV)	X	X	X	X	X	X
FICON Express (type FC)	X	X	X	X	X	X
Cascaded FICON Directors (CHPID types FC and FCP) including CTC	X	X	X	X		X
OSA-Express GbE (CHPID type OSD) ^d	X	X	X	X	X	X
OSA-Express 1000BASE-T Ethernet	X	X	X	X	X	X
OSA-Express Integrated Console Controller (OSA-ICC)	X	X	X	X		
OSA-Express Token Ring	X	X	X	X	X	X
Checksum offload for IPv4 packets ^e	X	X				
z/OS Full VLAN (IEEE 802.1q) support	X	X				
Intrusion Detection Services (CHPID type OSD)	X	X				
OSA/SF Java GUI	X	X	X	X	X	X
OSA-Express Direct SNMP subagent support ^f	X	X	X			

a. z/OS 1.6 is planned to be available in September 2004.

b. With HCD PTFs.

c. z/OS V1.6 plus IBM SDK for z/OS Java 2 Technology Edition V1.4.

d. CHPID type OSE (non-QDIO) supports TCP/IP and SNA.

e. For OSA-E GbE and 1000BASE-T EN with CHPID type OSD.

f. z/OS V1.5 for 'Traps and Set', and z/OS V1.6 for 'Direct SNMP for LCS' support all other SNMP subagent support z/OS V1.4 and later.

6.7.2 Summary of z/VM, z/VSE, VSE/ESA, TPF, and Linux software requirements

Table 6-3 Minimum z/VM, z/VSE, VSE/ESA, TPF, and Linux on zSeries requirements

Software Requirements	z/VM V5.1	z/VM V4.4	z/VM V3.1 and V4.3	VSE/ ESA V2.7 and z/VSE V3.1	VSE/ ESA V2.6	TPF V4.1	Linux on zSeries ^a
Functions							
16 to 30 logical partitions	X	X		X	X	X	X
Two Logical Channel Subsystems (LCSSs)	X	X	X ^b	X	X		X
Four Logical Channel Subsystems (LCSSs)	X	X	X ^b	X	X		X

Software Requirements	z/VM V5.1	z/VM V4.4	z/VM V3.1 and V4.3	VSE/ ESA V2.7 and z/VSE V3.1	VSE/ ESA V2.6	TPF V4.1	Linux on zSeries ^a
Functions							
Dynamic I/O support for multiple LCSSs	X	X	X ^b				X
Dynamic Add/Delete Logical Partition Name							X
24 processors within a single logical partition	X						X
Internal spanned channels	X	X	X	X			X
External spanned channels	X	X	X				X
Adapter Interruption (CHPID types FCP and OSD)	X	X		X ^c			X
V=V support for CHPID types FCP, OSD, and IQD	X	X					
HiperSockets	X	X	X ^d				X
VLAN (IEEE 802.1q)							X
Broadcast for IPv4 packets	X	X					X
HiperSockets Network Concentrator	X	X	X ^e				X
16-port ESCON feature	X	X	X	X	X	X	X
FICON Express (CHPID type FCV)	X	X	X	X	X	X ^f	
FICON Express (CHPID type FC)	X	X	X	X	X	X ^f	X
FICON Express (CHPID type FCP)	X ^g	X	X ^h	X ⁱ			X
FCP SAN Management							X ^j
SCSI IPL for FCP	X ^k	X ^l					X
Cascaded FICON Directors (CHPID types FC and FCP) including CTC	X	X		X	X	X ^f	X
OSA-Express GbE (CHPID type OSD)	X	X	X	X	X	X ^m	X
OSA-Express 1000BASE-T Ethernet	X	X	X	X	X		X
OSA-Express Integrated Console Controller (OSA-ICC)	X	X ⁿ		X	X	X	
OSA-Express Token Ring	X	X	X	X	X		X
Checksum offload for IPv4 packets							X
z/VM VLAN (IEEE 802.1q) support	X ^o	X ^p					
Linux on zSeries VLAN (IEEE 802.1q) support							X ^q
Intrusion Detection Services							X ^r
OSA/SF Java GUI	X	X	X	X	X		
OSA port name relief	X	X	X ^s				
OSA-Express Direct SNMP subagent support							X ^t

- a. Current distributions are SUSE SLES7 and SLES8, Red Hat RHEL 3.0, Turbolinux TLES 8, and Conectiva CLEE.
- b. Dynamic I/O configuration for LCSS 0 only.
- c. CHPID type OSD only.
- d. z/VM V4.3 only.
- e. z/VM 4.3 only with PTF for APAR VM63397.
- f. PUT 16.
- g. For z/VM install, IPL, and operation from SCSI disks.
- h. z/VM V4.3 only. For Linux as a guest.
- i. Planned for z/VSE V3.1 only (for FCP attached SCSI disks on the IBM ESS).
- j. See <http://www10.software.ibm.com/devel/operworks/opensource/linux390>.
- k. z/VM IPL from SCSI disks.
- l. For Linux as a guest.
- m. PUT 13 with PTF for APAR PJ2733.
- n. With PTF for APAR VM63405.
- o. For one global VLAN ID for IPv6 (applies to OSA-Express 1000BASE-T Ethernet, Fast Ethernet, and GbE).
- p. For one global VLAN ID for IPv4 (applies to OSA-Express 1000BASE-T Ethernet, Fast Ethernet, and GbE).
- q. Applies to OSA-Express 1000BASE-T Ethernet, Fast Ethernet, and GbE with CHPID type OSD.
- r. Applies to all OSA-Express features in QDIO mode (CHPID type OSD).
- s. z/VM V4.3 only, applies to all OSA-Express features in QDIO mode (CHPID type OSD).
- t. Applies to performance data, Get and GetNext, and Ethernet data for dot3StatsTable in QDIO mode, CHPID Type OSD.

6.8 Workload License Charges

Workload License Charges (WLC) is a software license charge method introduced with the z/Architecture.

WLC requires zSeries server(s) running z/OS operating system(s) in 64-bit mode. All MVS™-type operating system images running in the zSeries server *must* be z/OS. Any mix of z/OS, z/VM, Linux, VM/ESA, VSE/ESA, and TPF images is allowed, but no OS/390 image can exist.

There are two WLC license types:

- ▶ Flat WLC (FWLC): Software products licensed under FWLC are charged per copy basis, one copy for each zSeries server, independently of the server's capacity (MSUs).
- ▶ Variable WLC (VWLC): VWLC software products can be charged in two different ways:
 - Full-capacity. The server's total number of MSUs is used for charging. Full-capacity is applicable when the server is not eligible for Sub-capacity.
 - Sub-capacity Software charges are based on the logical partition's utilization where the product is running.

WLC Sub-capacity allows software charges based on logical partition utilizations instead of the server's total number of MSUs. Sub-capacity removes the dependency between software charges and server (hardware) installed capacity.

Sub-capacity is based on the logical partition's rolling 4-hour average utilization. It is *not* based on the utilization of each product, but on the utilization of the logical partition or partitions where it runs. The VWLC licensed products running on a logical partition will be charged by the maximum value of this partition's rolling 4-hour average utilization within a month.

The logical partitions' rolling 4-hour average utilization can be limited by a "Defined Capacity" definition on the partitions' image profiles. This 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.

Even using the soft capping option, the partition’s utilization can reach up to its maximum, based on the number of logical processors and weights, as usual. Only the rolling 4-hour average utilization is tracked, allowing utilization peaks above the defined capacity value.

As in the Parallel Sysplex License Charges (PSLC) software license charge type, the aggregation of servers’ capacities within a same Parallel Sysplex is also possible in WLC, following the same prerequisites.

For further information about WLC and details how to combine logical partitions utilization, see *z/OS Planning for Workload License Charges*, SA22-7506.

6.9 Concurrent upgrades considerations

Using Capacity Upgrade on Demand (CUoD), On/Off Capacity on Demand (On/Off CoD), Customer Initiated Upgrade (CIU) or Capacity Backup (CBU), you can concurrently upgrade the z990 from one model to another, either temporarily or permanently. You need to consider the effect on the software running on a z990 when performing these upgrades on a z990 processor.

Enabling and using the additional processor capacity should be transparent to all applications. There may be, however, a small class of applications that obtains the processor model-related information, for example, software monitors or applications that use the processor model information as a means of validation.

Processor identification

There are two instructions used to obtain the processor model information:

► STIDP Store CPU ID instruction

STIDP instruction provides a 1-byte hexadecimal version code, which is x'00' for zSeries servers. The STIDP instruction also provides information on the processor type (2084), serial number and LPAR identifier, as shown on Table 6-4.

On the z990, the LPAR identifier field has been expanded to a full byte to support greater than 15 logical partitions.

Table 6-4 STIDP output for z990

	Version code	CPU identification number		Machine type number	LPAR 2-digit indicator
Bit position	0-7	8-15	16-31	32-48	48-63
Value	x'00' ^a	LPAR ID ^b	6-digit number derived from the CPC serial number	x'2084'	x'8000' ^c

a. Version code is zero for zSeries processors.

b. The logical partition identifier is a two-digit number in the range from '00' to '3F'. It is assigned by the user on the image profile through the Support Element (SE) or Hardware Management Console (HMC).

c. High order bit on indicates that the LPAR ID value returned in bits 8-15 is now a two-digit value. zSeries processors prior to z990 return x'0000'.

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 or not.

- Without the use of the **set CPUID** command, bits 0-7 are set to 'FF' by z/VM but remaining bits are unchanged, meaning 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 6-5 shows the possible output returned to the issuing program for an operating system running as a guest under z/VM.

Table 6-5 STIDP output for z990, VM guest

	Version code	CPU identification number		Machine type number	LPAR 2-digit indicator
Bit position	0-7	8-15	16-31	32-48	48-63
Without set CPUID command	x'FF'	LPAR ID	4-digit number derived from the CPC serial number	x'2084'	x'8000'
With set CPUID command	x'FF'	6-digit number as entered by the command SET CPUID = nnnnnn		x'2084'	x'8000'

► STSI Store System Information instruction

The STSI instruction returns the processor software model as a 16-byte character field. It also returns the same processor type that is returned by the STIDP instruction and the full serial number information.

The STSI instruction always returns the latest processor software model information, including information about the new processor model after a concurrent model upgrade has occurred. This is key to the functioning of CUoD, On/Off CoD, CIU, and CBU.

Channel to channel links

After a concurrent upgrade, the channel CPC Node-Descriptor (NED) information is not updated until after a processor POR.

Additional planning may be required in a multisystem environment with CTCs linking different processors. NED information, which includes serial number, machine type, and model, is exchanged between systems on the CTC link. As a way to prevent cabling errors, CTCs will go into a "boxed" state if the NED information changes without having taken the proper actions. Boxed CTCs may impact XCF, VTAM®, IMS, and other software products.

Dealing with boxed CTCs in a multisystem environment is not new; it occurs during the POR after traditional disruptive upgrades. However, in the case of a concurrent upgrade, the node-descriptor information (model number) will not change until the next POR, which may be months after the actual upgrade. At that time, the customer needs to be prepared to deal with the boxed CTCs. It is important to consider and prepare for the case where, during an unplanned POR of the upgraded process, the CTCs become boxed.

The boxing of CTCs can be avoided if, during the concurrent upgrade, the CTC links between systems are deallocated and then varied offline. However, when alternate communication links are not available, this may be disruptive to applications.

The alternative is to be prepared for the boxed CTCs to occur during the next POR of the upgraded system. In most cases, using the UNCOND option of the VARY ONLINE command will un-box the CTCs in a nondisruptive manner.

The implications of boxed CTCs, particularly on ISV products, should be investigated during the planning process prior to a concurrent upgrade.

Archived

Archived

Sysplex functions

This chapter describes the capabilities of the z990 to support coupling functions, including Parallel Sysplex, Geographically Dispersed Parallel Sysplex™ (GDPS®), and Intelligent Resource Director.

The following topics are included:

- ▶ 7.1, “Parallel Sysplex” on page 156
- ▶ 7.2, “Sysplex and Coupling Facility considerations” on page 159
- ▶ 7.3, “System-managed CF structure duplexing” on page 169
- ▶ 7.4, “Geographically Dispersed Parallel Sysplex” on page 172
- ▶ 7.5, “Intelligent Resource Director” on page 178

7.1 Parallel Sysplex

Figure 7-1 illustrates the components of a Parallel Sysplex as implemented within the zSeries architecture. Shown is a z900 model 2xx ICF (CF01) connected to two z990 servers running in Sysplex. There is a second Integrated Coupling Facility (CF02) defined within one of the z990s, containing Sysplex logical partitions running z/OS.

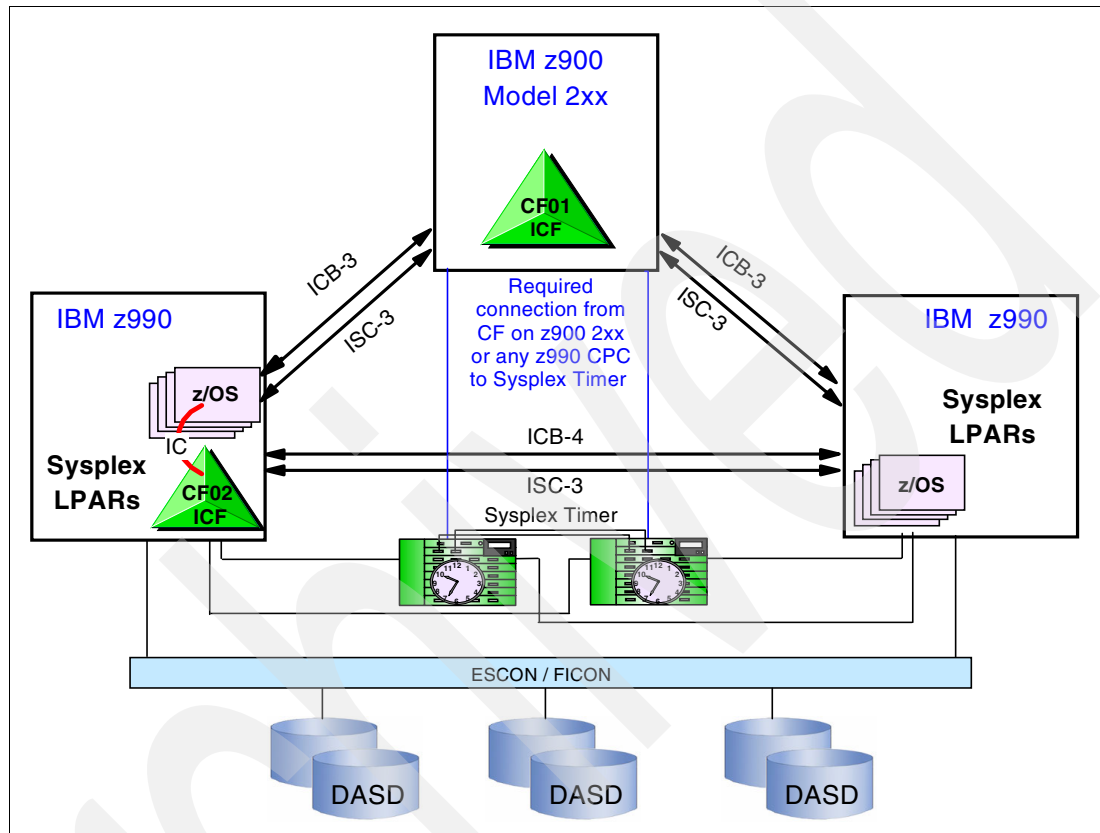


Figure 7-1 Sysplex hardware overview

Also shown is the required connection between the Coupling Facility (CF01) defined on a z900 turbo model (2064-2xx) or any z990 Server, and the Sysplex Timer, to support Message Time Ordering.

7.1.1 Parallel Sysplex described

Parallel Sysplex technology is an enabling technology, allowing highly reliable, redundant, and robust zSeries technologies to achieve near-continuous availability. A Parallel Sysplex is comprised of one or more z/OS and/or OS/390 operating system images coupled via one or more Coupling Facilities. The images can be combined together to form clusters. A properly configured Parallel Sysplex cluster is designed to maximize availability. For example:

- ▶ Hardware and software components provide for concurrent planned maintenance, like adding additional capacity to a cluster via additional images, without disruption to customer workloads.
- ▶ Networking technologies that deliver functions like VTAM Generic Resources, Multi-Node Persistent Sessions, Virtual IP Addressing, and Sysplex Distributor to provide fault-tolerant network connections.

- ▶ z/OS and OS/390 software components allow new software releases to coexist with lower levels of that software component to facilitate rolling maintenance.
- ▶ Business applications are “data sharing enabled” and cloned across images to allow workload balancing and to prevent loss of application availability in the event of an outage.
- ▶ Many operational and recovery processes can be automated, reducing the need for human intervention.

Parallel Sysplex is a way of managing a multi-system environment, providing the benefits of:

- ▶ Continuous (application) availability
- ▶ High capacity
- ▶ Dynamic workload balancing
- ▶ Simplified systems management
- ▶ Resource sharing
- ▶ Single system image

Continuous (application) availability

Within a Parallel Sysplex cluster, it is possible to construct a parallel processing environment with high availability. This environment is composed of multiple images that provide concurrent access to all critical applications and data.

You can introduce changes (such as software upgrades) one image at a time, while remaining images continue to process work. This allows you to roll changes through your images at a pace that makes sense for your business.

High capacity

The Parallel Sysplex environment can scale, in a nearly linear fashion, from two to 32 images. This can be a mix of any server or operating system that supports the Parallel Sysplex environment. The aggregated capacity of this configuration meets every processing requirement known today.

Dynamic workload balancing

The entire Parallel Sysplex cluster can be viewed as a single logical resource to end users and business applications. Work can be directed to any like operating system image in a Parallel Sysplex cluster having available capacity. This avoids the need to partition data or applications among individual images in the cluster or to replicate databases across multiple servers.

Workload management permits you to run diverse applications across a Parallel Sysplex cluster while maintaining the response levels critical to your business. You select the service level agreements required for each workload, and the z/OS or OS/390 Workload Manager (WLM), along with the subsystems such as CP/SM or WebSphere, automatically balances tasks across all the resources of the Parallel Sysplex cluster to meet your business goals. Whether the work is coming from batch, SNA, TCP/IP, DRDA®, or MQSeries® (non-persistent) messages, dynamic session balancing gets the business requests into the system best able to process the transaction. This provides the performance and flexibility you need to achieve the responsiveness your customers demand, and it is invisible to users.

Systems management

The Parallel Sysplex architecture provides the infrastructure to satisfy a customer requirement for continuous availability, while providing techniques for achieving simplified systems management consistent with this requirement. Some of the features of the Parallel Sysplex solution that contribute to increased availability also help to eliminate some systems management tasks. Examples include:

- ▶ **z/OS or OS/390 Workload Manager**

The Workload Manager (WLM) component of z/OS or OS/390 provides sysplex-wide workload management capabilities based on installation-specified performance goals and the business importance of the workloads. The Workload Manager tries to attain the performance goals through dynamic resource distribution. WLM provides the Parallel Sysplex cluster with the intelligence to determine where work needs to be processed and in what priority. The priority is based on the customer's business goals and is managed by sysplex technology.

- ▶ **Sysplex Failure Manager**

The Sysplex Failure Management component of z/OS or OS/390 allows the installation to specify failure detection intervals and recovery actions to be initiated in the event of the failure of an image in the sysplex.

- ▶ **Automatic Restart Manager**

The Automatic Restart Manager (ARM), a component of z/OS or OS/390, enables fast recovery of the subsystems that might hold critical resources at the time of failure. If other instances of the subsystem in the Parallel Sysplex need any of these critical resources, fast recovery will make these resources available more quickly. Even though automation packages are used today to restart the subsystem to resolve such deadlocks, ARM can be activated closer to the time of failure.

- ▶ **Cloning/symbolics**

Cloning refers to replicating the hardware and software configurations across the different physical servers in the Parallel Sysplex, that is, an application that is going to take advantage of parallel processing might have identical instances running on all images in the Parallel Sysplex. The hardware and software supporting these applications could also be configured identically on all images in the Parallel Sysplex to reduce the amount of work required to define and support the environment.

Resource sharing

A number of base z/OS or OS/390 components exploit Coupling Facility shared storage, providing an excellent medium for sharing component information for the purpose of multi-image resource management. This exploitation, called IBM @server zSeries Resource Sharing, enables sharing of physical resources such as files, tape drives, consoles, catalogs, and so forth, with significant improvements in cost, performance, and simplified systems management. The zSeries Resource Sharing delivers immediate value, even for customers who are not leveraging data sharing, through exploitation delivered with the base z/OS or OS/390 software stack.

Single system image

Even though there could be multiple servers and z/OS or OS/390 images in the Parallel Sysplex cluster, it is essential that the collection of images in the Parallel Sysplex appear as a single entity to the operator, the end user, the database administrator, and so on. A single system image ensures reduced complexity from both operational and definition perspectives.

Regardless of the number of images and the underlying hardware, the Parallel Sysplex cluster appears as a single system image from several perspectives:

- ▶ Data access, allowing dynamic workload balancing and improved availability
- ▶ Dynamic transaction routing, providing dynamic workload balancing and improved availability
- ▶ End-user interface, allowing access to an application as opposed to a specific image
- ▶ Operational interfaces that allow Systems Management across the sysplex from a single point

7.1.2 Parallel Sysplex summary

Through this state-of-the-art cluster technology, the power of multiple z/OS and/or OS/390 images can be harnessed to work in concert on common workloads. The zSeries Parallel Sysplex cluster takes the commercial strengths of the z/OS or OS/390 platform to improved levels of system management, competitive price/performance, scalable growth, and continuous availability.

7.2 Sysplex and Coupling Facility considerations

Described here are the supported Parallel Sysplex configurations, required set-up information when connected to an Sysplex Timer, different forms of Coupling Facilities (CFs) supported on z990 servers, CFRM policy considerations, and ICF processor assignments. The z990 models support both Central Processors (CPs) and Internal Coupling Facility (ICF) processors.

The z990 family of servers does not provide a special model for a CF-only processor. You can, however, have a z990 server with up to 16 PUs defined as ICFs.

7.2.1 Sysplex configurations and Sysplex Timer considerations

Parallel Sysplex configurations today can have system images and Coupling Facilities located across multiple servers. This can be anything from S/390 to zSeries servers. However, z990 brings some additional considerations to these types of configurations.

Message Time Ordering

As server and Coupling Facility link technologies have improved over the years, the synchronization tolerance between operating systems in a Parallel Sysplex has become more rigorous. In order to ensure that any exchange of timestamped information between operating systems in a sysplex involving the Coupling Facility observe the correct time ordering, timestamps are now included in the message-transfer protocol between the server operating systems and the Coupling Facility. This is known as Message Time Ordering.

Message Time Ordering requires a connection between the z990 CPC and the Sysplex Timer whenever a Coupling Facility is located in a logical partition on a z990.

Therefore, when a Coupling Facility is configured as an ICF on any z990 model, the Coupling Facility will require connectivity to the same 9037 Sysplex Timer that the systems in its Parallel Sysplex are using for the time synchronization. If the ICF is on the same server as one or more a member of its Parallel Sysplex, no additional Sysplex Timer connectivity is required, since the server already has connectivity to the Sysplex Timer. However, when an ICF is configured on any z990 model that does not host any systems in the same Parallel Sysplex, it is necessary to attach the z990 server to the 9037 Sysplex Timer.

Even though multiple servers can connect to only one Sysplex Timer unit, the typical configuration is usually connected to two different Sysplex Timer units called an Expanded Availability configuration. Refer to *IBM @server zSeries Connectivity Handbook*, SG24-5444, for IBM 9037 Sysplex Timer connectivity information.

External Reference ID

A Sysplex Timer unit is assigned a unique two-digit ID at installation time, called a Unit ID. This Unit ID is referenced as an External Time Reference ID (ETR ID) in the output of the z/OS command D ETR and Support Element panels.

A function is introduced with the z990 server, implemented in the server's Support Element code, which now requires the ETR Network ID of the attached Sysplex Timer Network to be manually set in the Support Element at installation time. This function checks that the ETR Network ID being received in the timing signals via each of the server's two ETR ports matches the ETR Network ID manually set in the server's Support Element (SE).

Up to two Sysplex Timer units can be configured in an Expanded Availability configuration, each one with a unique ETR ID. When in Expanded Availability configuration, a network ID (Net ID) is also assigned at installation time to identify that these two Sysplex Timer units belong to the same Sysplex Timer configuration.

The z990 requires that the ETR Network ID of the attached Sysplex Timers be entered in a panel on the Support Element (SE). As part of the installation of a Sysplex Timer in either Basic or Expanded Availability configuration, each IBM 9037 Sysplex Timer Unit is assigned an ETR Network ID (0 to 31 decimal) and an ETR Unit ID (0 to 31 decimal). Within the valid range, the ETR Network ID and ETR Unit ID values are arbitrary and can be chosen by the customer to uniquely identify an ETR network and a unique ETR unit (Sysplex Timer) within the ETR network.

In addition, on the same panel, the ETR ports have to be enabled for stepping the TOD. This function checks that the ETR Network ID being received in the timing signals via each of the two ETR ports matches the ETR Network ID manually set in the z990 Support Element. This provides greater checking, helping eliminate cabling errors where a z990 ETR port may be incorrectly connected to a Sysplex Timer Unit in an incorrect Sysplex Timer ETR network, and allows verification of cabling connectivity from the Sysplex Timer to the z990 server prior to IPLing z/OS or OS/390.

To get the appropriate SE panel, log on to the SE directly or via the HMC single object operations task. Navigate to the CPC Configuration task list and invoke the System Complex (Sysplex) Timer task. Once this task is invoked, a notebook is displayed with two panels. The first panel contains configuration information, as shown in Figure 7-2 on page 161; the ETR Network ID (0-31) is entered on this configuration panel.

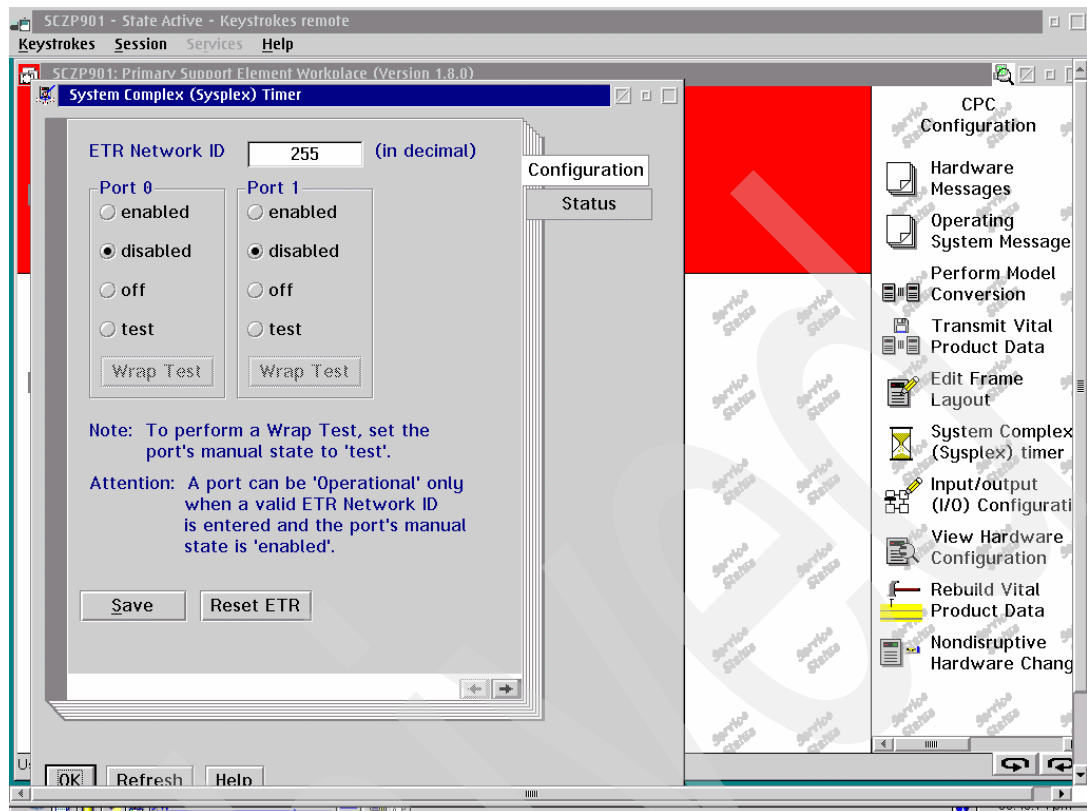


Figure 7-2 z990 SE workpace: External Timer Reference Configuration panel

The network ID configured on the z990 must match the actual Sysplex Timer network ID to which the server is connected. If the network ID entered on the SE panel does not match the network ID that was assigned to the Sysplex Timer, the timer port enters a semi-operation state. In this state, the port is disabled from stepping, but still receives configuration data from the attached Sysplex Timer.

The ETR status information panel, shown in Example 7-1, displays the configuration status.

After IPL, the configuration can be identified to any attached z/OS image by issuing the Display ETR command. The command output identifies the ETR configuration, including ETR ID and ETR Network ID (see the following example).

Example 7-1 Display ETR

```
D ETR
IEA282I 16.30.19 ETR STATUS 686
SYNCHRONIZATION MODE = ETR      CPC SIDE = 0
  CPC PORT 0  <== ACTIVE      CPC PORT 1
  OPERATIONAL                                OPERATIONAL
  ENABLED                                    ENABLED
  ETR NET ID=00                        ETR NET ID=00
  ETR PORT=11                          ETR PORT=11
  ETR ID=04                            ETR ID=03
```

7.2.2 Coupling Facility and CFCC considerations

The z990 can participate in a Parallel Sysplex when the Coupling Facility resides in a G5 or G6 or any zSeries Server. When system images in the sysplex reside across z990 and non-z990 Servers, consideration must be given for compatibility support; see 6.2.1, “Compatibility Support for z/OS” on page 134. When any system image in a Parallel Sysplex resides on a z990 server, all other system images and Coupling Facilities must be on 9672 G5 or G6 or zSeries hardware.

The location of the Coupling Facility and the level of the Coupling Facility Control Code (CFCC) must also be considered. See Table 7-1 for Coupling Facility Control Code requirements when the Coupling Facility resides on a non-z990 server and is connected to a z/OS image on a z990, or when CF duplexing is used and one Coupling Facility resides on the z990 server.

Table 7-1 z990 CF code level considerations

CF with connections to z990 z/OS image or z990 (CF duplexing)	Connected to a z990 with 15 or less LPARs defined	Connected to a z990 with more than 15 LPARs defined
Pre-G5 CPC	Not supported	Not supported
G5/G6 CPC	CFCC level 11	CFCC level 11
z800 or z900 CPC	CFCC level 12 (not recommended ^{a b})	CFCC level 12 ^b with Compatibility patch
z990 or z890 CPC	CFCC level 13	CFCC level 13

a. It is recommended that the Compatibility patch be installed prior to the z990 installation. Not installing the patch opens up the possibility that a z990 will have more than 15 partitions defined later without the Compatibility patch being installed.

b. Installation of the Compatibility patch is disruptive.

The initial support of the CFCC on the z990 Server is level 12. Typically each new level of CFCC introduces additional functionality. Coupling Facilities within a Parallel Sysplex should be at equivalent levels, and the precise CFCC level supported can depend on server type. CFCC level 13 is current as of May 28, 2004. CFCC level 13 introduces some availability and performance enhancements.

To support migration from one Coupling Facility level to the next, you can run different levels of the Coupling Facility concurrently as long as the Coupling Facility logical partitions are running on different servers.

CF logical partitions running on the same server share the same Coupling Facility Control Code EC level. A single server cannot support multiple Coupling Facility levels. Table 7-2 on page 163 summarizes the CFCC CFLEVELs supported on the z990 servers.

7.2.3 CFCC enhanced patch apply

Patch application to CFCC level 13 code is designed to eliminate the need for a Power-on Reset of the z990 when a ‘disruptive’ patch must be applied.

This method of patch application enables you to:

- Apply the patch on one of the available Coupling Facilities. A good place would be the test Coupling Facility of a test sysplex if one is available. If the test of the CFCC code is successful, it can be applied to production Coupling Facility on the same z990. To use the updated CFCC code to a CF logical partition, simply deactivate and reactivate the

partition. When the CF comes up, it displays its version on the OPRMSG panel for that partition.

- ▶ Continue to run other LPARs on the z990 where a 'disruptive' CFCC patch is applied without being impacted by the application of the patch.

Table 7-2 CFCC levels supported on a z990

CFLEVEL	Minimum software levels ^a
CFLEVEL 12 <ul style="list-style-type: none"> ▶ 64-bit support, removal of the 2 GB line ▶ System Managed CF Structure Duplexing ▶ Support for Message Time Ordering 	<ul style="list-style-type: none"> ▶ z/OS 1.4 and above is required to fully exploit the functions. ▶ z/OS 1.2 and above with APAR OW41617 is the minimum for System Managed CF Structure Duplexing. ▶ All supported levels of z/OS^b and OS/390 2.10 with PTFs can be used with CFLEVEL 12, but may not take advantage of the enhancements.
CFLEVEL 13 <ul style="list-style-type: none"> ▶ CFCC enhanced patch apply ▶ CFCC performance improvements benefitting castout processing against large DB2 group buffer pool structures 	<ul style="list-style-type: none"> ▶ All supported levels of z/OS^b and OS/390 2.10 with PTFs can be used with CFLEVEL 13, but may not take advantage of the enhancements.

a. Always consult the latest PSP bucket for 2084DEVICE and the appropriate subset for the latest maintenance information.

b. z/OS1.1 is not supported on the z990.

Note: When migrating to a new CFCC level, lock, list, and cache structure sizes will typically increase to support new functions.

This adjustment can have an impact when the system allocates structures or copies structures from one Coupling Facility to another at different CFCC levels.

The Coupling Facility structure sizer tool can size structures for you and takes into account the amount of space needed for the current CFCC levels.

The CFSIZER tool can be found at:

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

CFLEVEL 13 has improvements that benefit some software environments in a Parallel Sysplex. DB2 data sharing may especially expect a performance improvement for castout processing against large DB2 group buffer pool structures.

For additional details on CF code levels, see the following link:

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

For additional details regarding CF configurations, see the paper *Coupling Facility Configuration Options*, GF22-5042, available from the Parallel Sysplex Web site:

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

7.2.4 Coupling Facility link connectivity

The type of CF links you can use to connect a CF to an operating system logical partition is important because of the impact 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 (mainly if the CF is defined on a z990 server).

The types of links that are available to connect an operating system logical partition to a Coupling Facility are:

- ▶ **IC:** Licensed Internal Code defined links to connect a CF to a z/OS logical partition in the same z990 processor. IC links require two CHPIDs to be defined and can only be defined in Peer mode. The link bandwidth is greater than 2 GBps. A maximum of 32 IC links can be defined per z990.
- ▶ **ICB-4:** Copper links are available to connect z990 to z990, or z890 processors; the maximum distance between the two processors is 7 meters (maximum cable length is 10 meters). The link bandwidth is 2 GBps. ICB-4 links can only be defined in Peer mode. Maximum number of ICB-4 links is 16 per z990.
- ▶ **ICB-3:** Copper links are available to connect z990 to z990, z890, z900, or z800 processors; the maximum distance between the two processors is 7 meters (maximum cable length is 10 meters). The link bandwidth is 1 GBps. ICB3 links can only be defined in Peer mode. Maximum number of ICB-3 links is 16 per z990.
- ▶ **ICB-2:** Copper links are available to connect a z990 to 9672 G5/G6 processors (you cannot use ICB-2 to connect a z990 to a z900 or to another z990); the maximum distance between the two processors is 7 meters (maximum cable length is 10 meters). The link bandwidth is 333 MBps. Maximum number of ICB-2 links is eight per z990.
- ▶ **ISC-3:** The z990 ISC-3 feature is compatible with Coupling Links/ISCs (referred to as a Hiperlink on G5/G6 Servers) on S/390 generation 5 and generation 6 servers as well as zSeries 890, zSeries 900, and zSeries 800. Each port is capable of 1 Gbps or 2 Gbps, depending upon the mode of operation selected in the Hardware Configuration Definition (HCD) tool or IOCP. Ports are ordered in increments of one. The maximum number of ISC-3 links per z990 is 48 in peer mode, and 32 ISC-3 links in compatibility mode.
 - There are fiber links available to connect z990 to z990, z890, z900, or z800 processors; the maximum distance is 10 km, 20 km with RPQ 8P2197, and 40 km with Dense Wave® Division Multiplexing (DWDM). ISC-3s operate in single mode only and link bandwidth is 200 MBps for distances up to 10 km, and 100MBps when RPQ 8P2197 is installed. ISC-3 links should be defined in peer mode. The peer mode is used between zSeries servers only.
 - There are fiber links available to connect z990 to G5 and G6 servers. The z990 ISC-3 feature is compatible with Coupling Links/ISCs (referred to as a Hiperlink) on G5/G6 servers. Compatibility mode is used between the z990 and G5/G6 servers. The port is defined as a sender/receiver (CFS/CFR) channel and the link is capable of 1 Gbps.

Table 7-3 z990 Coupling Link maximums

Link Type	z990 Max
IC	32
ISC-3	48 ^a
ICB-2	8
ICB-3	16
ICB-4	16

Link Type	z990 Max
Maximum number of links per z990	64 ^b

a. A maximum of 32 ISC-3s can be defined in compatibility mode, which operates up to 1 Gb/s.

b. The maximum number of external and internal Coupling Links combined (ICB-2, ICB-3, ICB-4, ISC-3, and active IC links) cannot exceed 64 per system.

Refer to Table 7-4 for an overview of the CF link connectivity options for the various supported servers.

Table 7-4 z990 CF link connectivity

Connectivity options	z990 ISC-3	z990 ICB-2	z990 ICB-3	z990 ICB-4
G5/G6 ISC	1 Gbps Compat Mode	N/A	N/A	N/A
z800/z900 ISC-3	2 Gbps Peer Mode ^a	N/A	N/A	N/A
z890/z990 ISC-3	2 Gbps Peer Mode ^a	N/A	N/A	N/A
G5/G6 ICB	N/A	333 MBps Compat Mode	N/A	N/A
z900 ICB-2	N/A	Not supported	N/A	N/A
z990 ICB-2	N/A	Not supported	N/A	N/A
z900 ICB-3	N/A	N/A	1 GByte/sec Peer Mode	N/A
z990/z890 ICB-3	N/A	N/A	1 GByte/sec, Peer mode, Recommendation use ICB-4	N/A
z990/z890 ICB-4	N/A	N/A	N/A	2 GBps Peer Mode

a. 1 Gbps when 20km RPQ 8P2197 is installed.

Peer mode links

There are several advantages in using peer mode links. First of all, peer mode links operate on a higher bandwidth than the equivalent compatibility mode link. A single CHPID (one side of the link) can be both sender and receiver; this means that a single CHPID can be shared between multiple OS logical partitions and one Coupling Link logical partition. The number of link buffers provided when peer mode links are used is seven per link compared to two per link in compatibility mode; this is particularly important with System-Managed CF Structure Duplexing. Peer links have 224 KB data buffer space compared to 8 KB on compatibility mode links; this is especially important for long distances, as it reduces the handshaking for large data transfers.

z/OS and/or OS/390 images and Coupling Facility images may be running on the same or on separate servers. Every z/OS or OS/390 image in a Parallel Sysplex must have at least one coupling link to each CF image.

For availability reasons, there should be:

- ▶ At least two coupling links between z/OS and/or OS/390 and Coupling Facility images
- ▶ At least two Coupling Facility images (not running on the same server)
- ▶ At least one stand-alone Coupling Facility (if using system-managed CF structure duplexing or running with “Resource Sharing” only, then a stand-alone Coupling Facility is not mandatory)
- ▶ At least two Coupling Facility images are required for system-managed CF structure duplexing.

7.2.5 Coupling Facility Resource Manager (CFRM) policy considerations

Because the z990 is capable of having greater than 15 logical partitions, there is a change in the usage of the PARTITION keyword when defining the CFRM policy via the administration utility. Support has been added to allow for two digits in the keyword PARTITION(nn). The meaning of keyword also depends upon if the CF location resides on a z990 or non-z990 (see Figure 7-3 for more details).

If the CF LPAR ID on the z990 is equal or less than x' F', then z/OS compatibility support is not required on z/OS partitions running on non-z990 processors (see 6.2.1, “Compatibility Support for z/OS” on page 134).

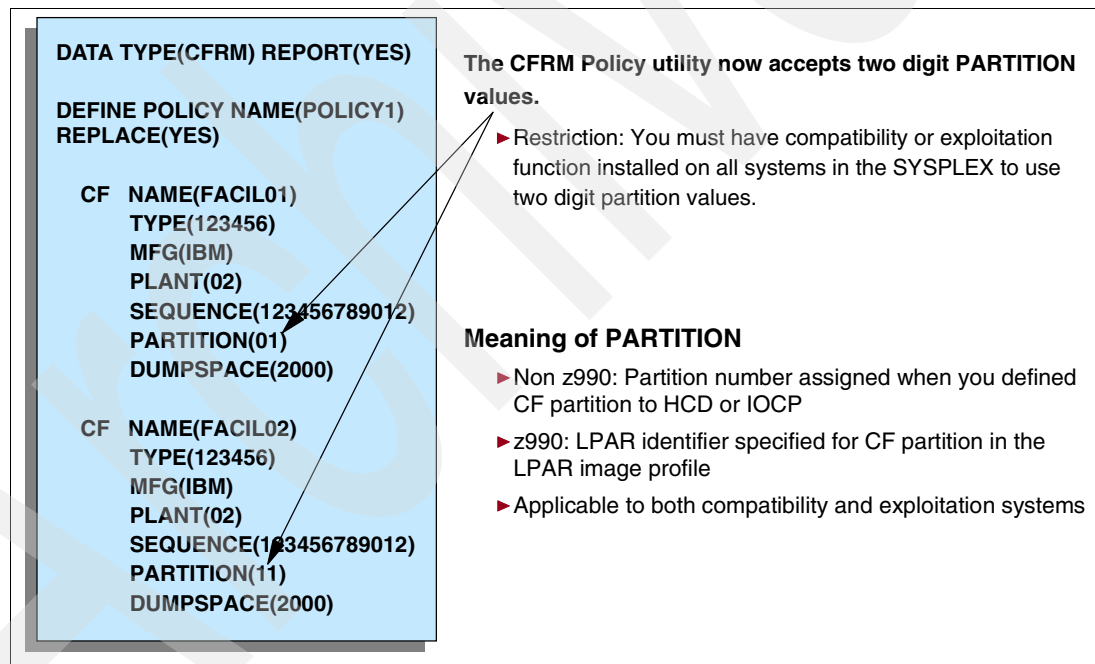


Figure 7-3 CFRM Policy changes

7.2.6 ICF processor assignments

The advantage of using ICF Processors (PUs characterized as ICFs) instead of CPs for Coupling Facility images is that because an ICF cannot run any z/OS or OS/390 operating systems, software licenses are not charged for those processors.

Note: There must be spare PUs to order dedicated ICFs. Refer to 2.2, “System design” on page 38 for details.

CPs are Processor Units used to process z/OS, OS/390, CFCC, z/VM, Linux, TPF, VSE/ESA, or z/VSE instructions. The logical partition can use dedicated *or* shared CPs. However, it is not possible to have a logical partition with dedicated *and* shared CPs at the same time.

ICFs are PUs dedicated to process the CF Control Code (CFCC) on a Coupling Facility image, which is always running on a logical partition. A CF image can use dedicated *and/or* shared ICFs. It can also use dedicated or shared CPs. With Dynamic ICF Expansion, a Coupling Facility image can also use dedicated ICFs and shared CPs.

The z990 can have ICF processors defined to CF Images.

A z990 server Coupling Facility image can have one of the following defined in the image profile:

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

Shared ICFs add flexibility. However, running with shared Coupling Facility Processor Units (ICFs or CPs) adds overhead to the coupling environment and is not a recommended production configuration.

In Figure 7-4, 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. 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.

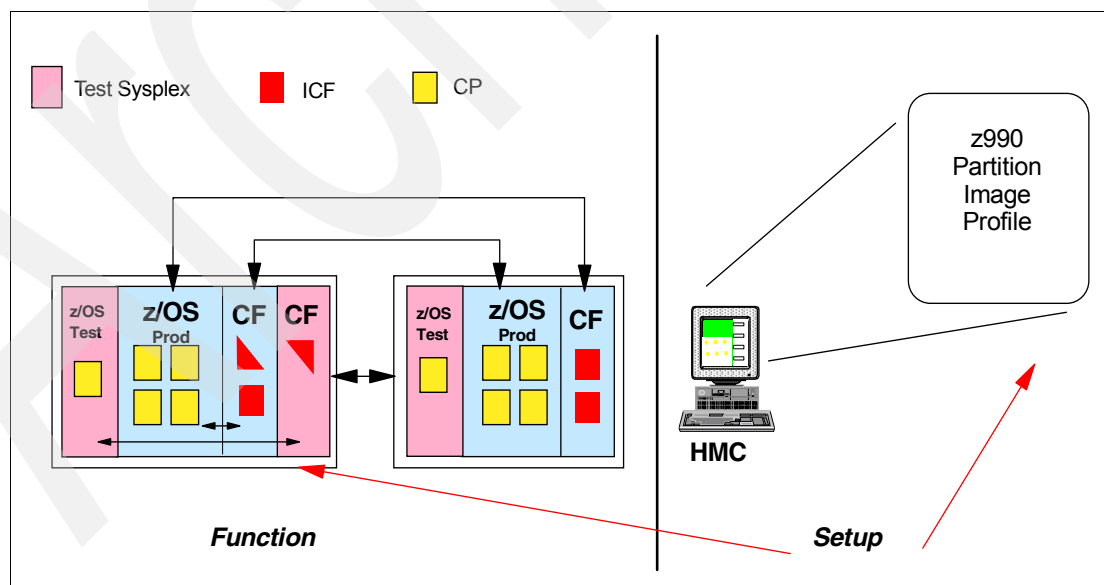


Figure 7-4 z990 ICF options - Shared ICFs

7.2.7 Dynamic CF dispatching and dynamic ICF expansion

The CF Control Code (CFCC), the “CF Operating System,” is implemented using the *Active Wait* technique. This means it is always running (processing or searching for service) and never enters into a wait state. This also means that it gets all the processor capacity (cycles) available for the Coupling Facility logical partition. If this logical partition uses only dedicated processors (CPs or ICFs), this is not a problem. But this may not be desirable when it uses shared processors (CPs or ICFs).

Dynamic CF dispatching provides the following function on a Coupling Facility: If there is no work to do, it enters into a wait state (by time). After an elapsed time, it wakes up to see if there is any new work to do (requests in the CF Receiver buffer). If there is no work, it will sleep again for a longer period of time. If there is new work, it enters into the normal Active Wait until there is no more work, starting the process all over again. This 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 z990 CPs can run z/OS and/or OS/390 operating system images and CF Images. For software charge reasons, it is better to use ICF processors to run Coupling Facility images.

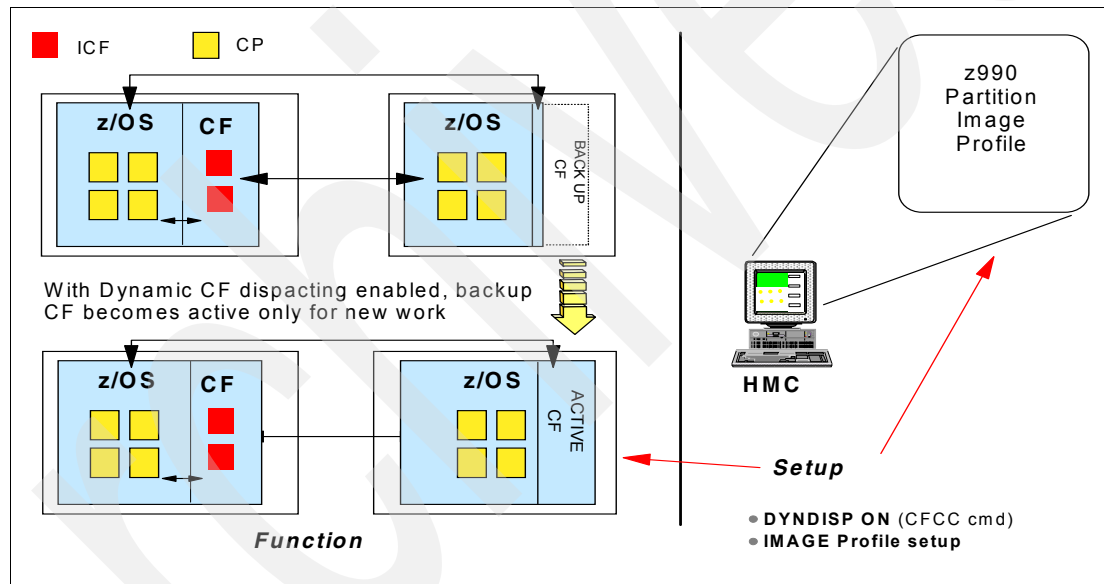


Figure 7-5 z990 Dynamic CF Dispatching (shared CPs or shared ICF PUs)

With Dynamic ICF Expansion, a Coupling Facility image using one or more *dedicated* ICFs can also use one or more *shared* CPs of this same server. The Coupling Facility image uses the shared CPs only when needed, that is, when its workload requires more capacity than its dedicated ICFs have. This may be necessary during peak periods or during recovery processes.

Figure 7-6 on page 169 shows an example where the server on the left has a production and a test Coupling Facility that has dedicated and shared ICF PUs. This configuration enables the Coupling Facilities to utilize the shared ICF PUs when workload becomes excessive. Additionally, if the alternate production Coupling Facility goes down (for maintenance, for example) and the allocated ICFs' capacity on the left server is not big enough to maintain its own workload plus that of the other Coupling Facility, then with Dynamic ICF Expansion, the remaining Coupling Facility image can be expanded over shared ICF PUs.

Dynamic ICF expansion can also be configured using dedicated ICF PUs and shared CPs from the z/OS image. The z/OS image *must* have all CPs defined as *shared* and the Dynamic CF Dispatch function must be activated. Dynamic ICF Expansion is available on z990 models that have at least one ICF.

Dynamic ICF Expansion requires that Dynamic CF Dispatching be activated (DYNDISP ON).

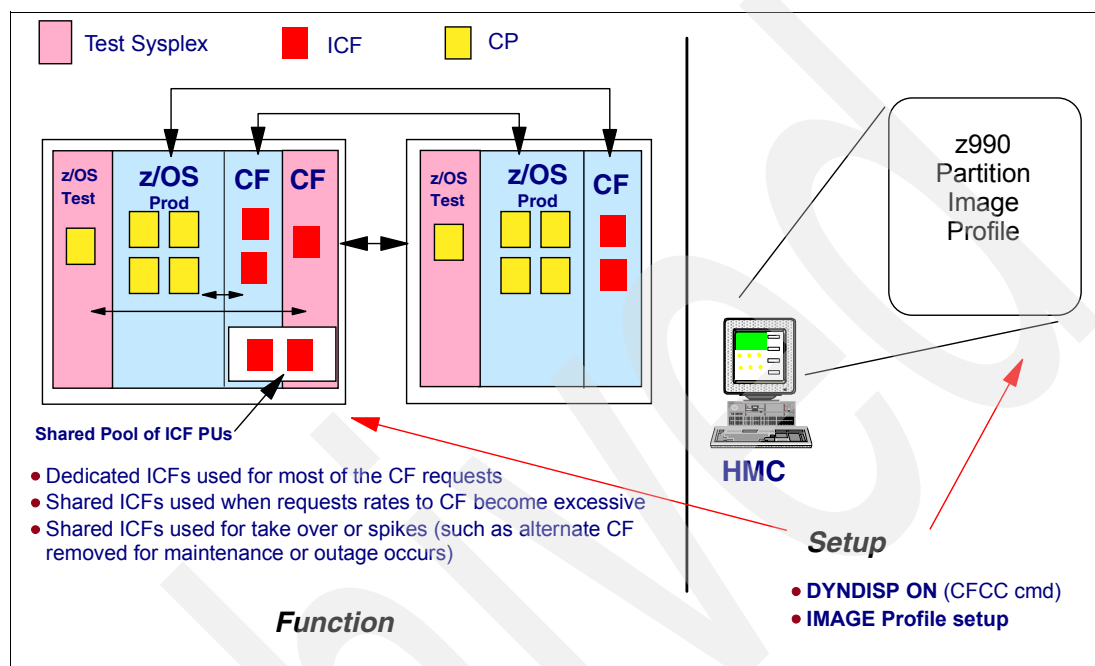


Figure 7-6 z990 CF options - Dynamic ICF Expansion

7.3 System-managed CF structure duplexing

System-managed Coupling Facility structure duplexing provides a general purpose, hardware-assisted, easy-to-exploit mechanism for duplexing CF structure data. This provides a robust recovery mechanism for failures, such as loss of a single structure or Coupling Facility, or loss of connectivity to a single Coupling Facility, through rapid failover to the other structure instance of the duplex pair.

7.3.1 Benefits

Benefits of system-managed CF structure duplexing include:

► Availability

Faster recovery of structures is provided by having the data already in the second Coupling Facility when a failure occurs. Furthermore, if a potential IBM, vendor, or customer Coupling Facility exploitation were being prevented, due to the effort required, from providing alternative recovery mechanisms such as structure rebuild, log recovery, and so forth, system-managed duplexing could provide the necessary recovery solution.

► Manageability and usability

These benefits are achieved by a consistent procedure to set up and manage structure recovery across multiple exploiters.

- **Cost benefits**

Cost benefits are realized by enabling the use of non-standalone Coupling Facilities (for example, ICFs) for all resource sharing and data sharing environments.

7.3.2 CF structure duplexing

System-managed Coupling Facility structure duplexing creates a duplexed copy of the structure in advance of any failure, providing a robust failure recovery capability through failover to the unaffected structure instance. This results in:

- An easily-exploited common framework for duplexing the structure data contained in any type of CF structure, with installation control over which structures are duplexed
- Minimized overhead of duplexing during mainline operation via hardware-assisted serialization and synchronization between the primary and secondary structure updates
- Maximized availability in failure scenarios by providing a rapid failover to the unaffected structure instance of the duplexed pair, with very little disruption to the ongoing execution of work by the exploiter and applications

System-managed duplexing rebuild provides robust failure recovery capability via the redundancy of duplexing, and low exploitation cost via system-managed, internalized processing. Structure failures, CF failures, or losses of CF connectivity can be handled by:

1. Hiding the observed failure condition from the active connectors to the structure, so that they do not perform unnecessary recovery actions
2. Switching over to the structure instance that did not experience the failure
3. Re-establishing a new duplex copy of the structure if appropriate as the Coupling Facility becomes available again, or on a third CF in the Parallel Sysplex

System messages are generated as the structure falls back to simplex mode for monitoring and automation purposes. The structure operates in simplex mode until a new duplexed structure can be established, and can be recovered using whatever existing recovery techniques are supported by the exploiter.

System-managed duplexing's main focus is providing this robust recovery capability for structures whose users do not support user-managed duplexing rebuild processes, or do not even support user-managed rebuild at all.

7.3.3 Configuration planning

A new connectivity requirement for system-managed CF structure duplexing is that there must be bi-directional CF-to-CF connectivity between each pair of CFs in which duplexed structure instances reside. With peer links, this connectivity can be provided by a single bi-directional link (two with redundancy).

CF-to-CF links can either be dedicated or shared via MIF. They can be shared with z/OS-to-CF links between z/OS and Coupling Facility images in the pair of servers they connect. When planning sharing links, remember that receiver links cannot be shared, and peer links can only be shared by one Coupling Facility partition.

Figure 7-7 on page 171 gives an overview of system-managed CF structure duplexing.

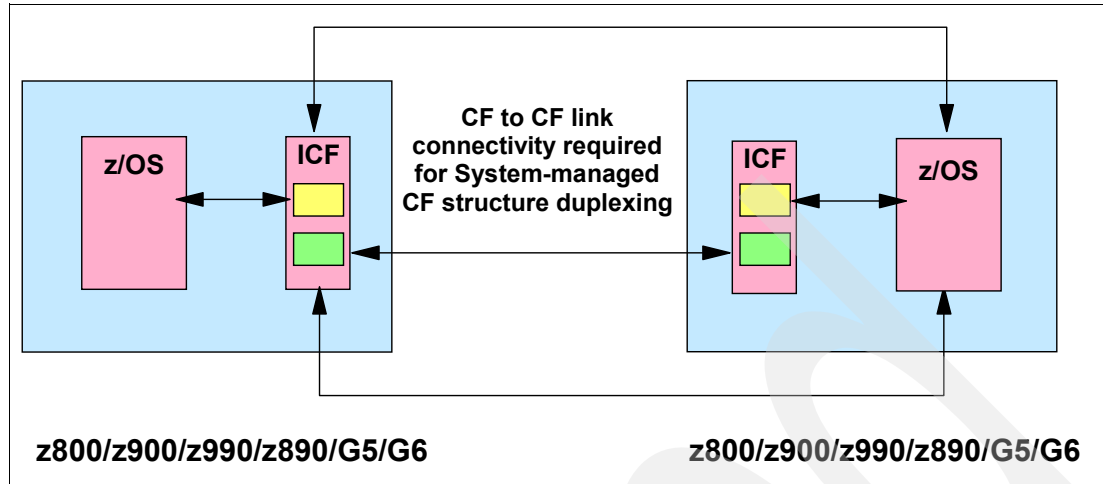


Figure 7-7 System-managed CF structure duplexing

Whenever possible, try to comply with the following recommendations:

- ▶ Provide two or more physical CF-to-CF links (peer mode), or two or more physical CF-to-CF links in each direction (sender/receiver mode), between each pair of CFs participating in duplexing. The physical CF-to-CF links may be shared by a combination of z/OS-to-CF links and CF-to-CF links.
- ▶ For redundancy, provide two or more z/OS-to-CF links from each system to each CF. Provide dedicated z/OS-to-CF links if possible. If z/OS-to-CF links are shared between z/OS partitions, the occurrence of path busy conditions should be limited to at most 10 to 20 percent of total requests. If path busy exceeds this guideline, either provide dedicated links, or provide additional shared links, to eliminate or reduce the contention for these link resources. Use peer links whenever possible.
- ▶ You can provide either dedicated or shared z/OS CPs when using system-managed CF structure duplexing. Dedicated Coupling Facility CPs are highly recommended for system-managed CF structure duplexing.
- ▶ Be prepared to provide additional z/OS CPU capacity when the workload's CF operations become duplexed. Provide sufficient Coupling Facility CP resources so that Coupling Facility CP utilization remains below 50% in all CF images.
- ▶ Provide "balanced" Coupling Facility CP capacity between duplexed pairs of CFs. Avoid significant imbalances such as one CF with shared CPs and the other CF with dedicated CPs, CFs with wildly disparate numbers of CPs, CFs of different machine types with very different raw processor speed, and so forth.
- ▶ As z/OS-to-CF and CF-to-CF distances increase, monitor the Coupling Facility link subchannel and path busy status. If more than 10% of all messages are being delayed on the CF link due to subchannel or path busy conditions, either migrate to peer mode links to increase the number of subchannels for each link, or configure an additional link.
- ▶ In a GDPS/PPRC multi-site configuration, do not duplex CF structure data between coupling facilities located in different sites; rather, if desired, duplex the structures between two coupling facilities located at the same site. CF structure data is not preserved in GDPS site fail-over situations, regardless of duplexing.

A technical paper on system-managed CF structure duplexing is available at:

<http://www-1.ibm.com/servers/eserver/zseries/library/techpapers/gm130103.html>

It includes a sample migration plan, describes how to monitor this new Parallel Sysplex technology and how to determine its cost/benefit in your environment, and gives setup recommendations.

7.4 Geographically Dispersed Parallel Sysplex

IBM Installation Services for GDPS is a total end-to-end solution that manages availability within a site and across multiple sites. It provides the automation to manage not only unplanned exception conditions, but also the many planned exception conditions that are faced as a part of normal everyday processing in any I/T environment. The GDPS solution can be tailored to specific Business Continuance requirements, and is based on either the synchronous Peer to Peer Remote Copy (PPRC) or the asynchronous Extended Remote Copy (XRC).

GDPS also supports the Peer-to-Peer Virtual Tape Server (PtP VTS) form of remote copying tape data. By extending GDPS support to data resident on tape, the GDPS solution is designed to provide continuous availability and near transparent business continuity benefits for both disk- and tape-resident data. Enterprises should no longer be forced to develop and utilize processes that create duplex tapes and maintain the tape copies in alternate sites.

GDPS is application independent and is enabled by means of key IBM technologies and architectures:

- ▶ Parallel Sysplex
- ▶ Tivoli® Netview for z/OS or OS/390
- ▶ System Automation for z/OS or OS/390
- ▶ Enterprise Storage Server™ (ESS)
- ▶ Peer-to-Peer Virtual Tape Server (PtP VTS)
- ▶ Optical Dense or Coarse Wavelength Division Multiplexer
- ▶ PPRC (Peer-to-Peer Remote Copy) architecture
- ▶ XRC (Extended Remote Copy) architecture
- ▶ Virtual Tape Server Remote Copy architecture

All GDPS images are running GDPS automation based upon Tivoli Netview for z/OS or OS/390 and System Automation for z/OS or OS/390. Each image will monitor the base or Parallel Sysplex cluster, Coupling Facilities, and storage subsystems; and maintain GDPS status. GDPS automation can coexist with an enterprise existing automation product.

For more detailed information on GDPS, see the white paper *GDPS: The e-business Availability Solution*, GF22-5114 at:

<http://www.ibm.com/servers/eserver/zseries/library/whitepapers/gf225114.html>

7.4.1 GDPS/PPRC

PPRC is a hardware solution that synchronously mirrors data residing on a set of disk volumes, called primary volumes in Site 1, to secondary disk volumes on a second system at Site 2. Only when the application site storage subsystem receives write complete from the recovery site storage subsystem is the I/O signaled as completed.

The physical topology of a GDPS/PPRC consists of a base or Parallel Sysplex cluster spread across two sites (site 1 and site 2), with one or more z/OS and/or OS/390 systems at each sites. The maximum distance between sites is 100 km (62 miles), using the GDPS/PPRC Cross-site Extended Distance for Parallel Sysplex RPQ (see “GDPS/PPRC Cross-site extended distance for Parallel Sysplex” on page 175). Without the extended distance RPQ, the distance between sites is up to 40 km, as shown in Figure 7-8 on page 173. The multisite

Parallel Sysplex cluster must be configured with redundant hardware (for example, a Coupling Facility and a Sysplex Timer in each site), and the cross-site connections must be redundant.

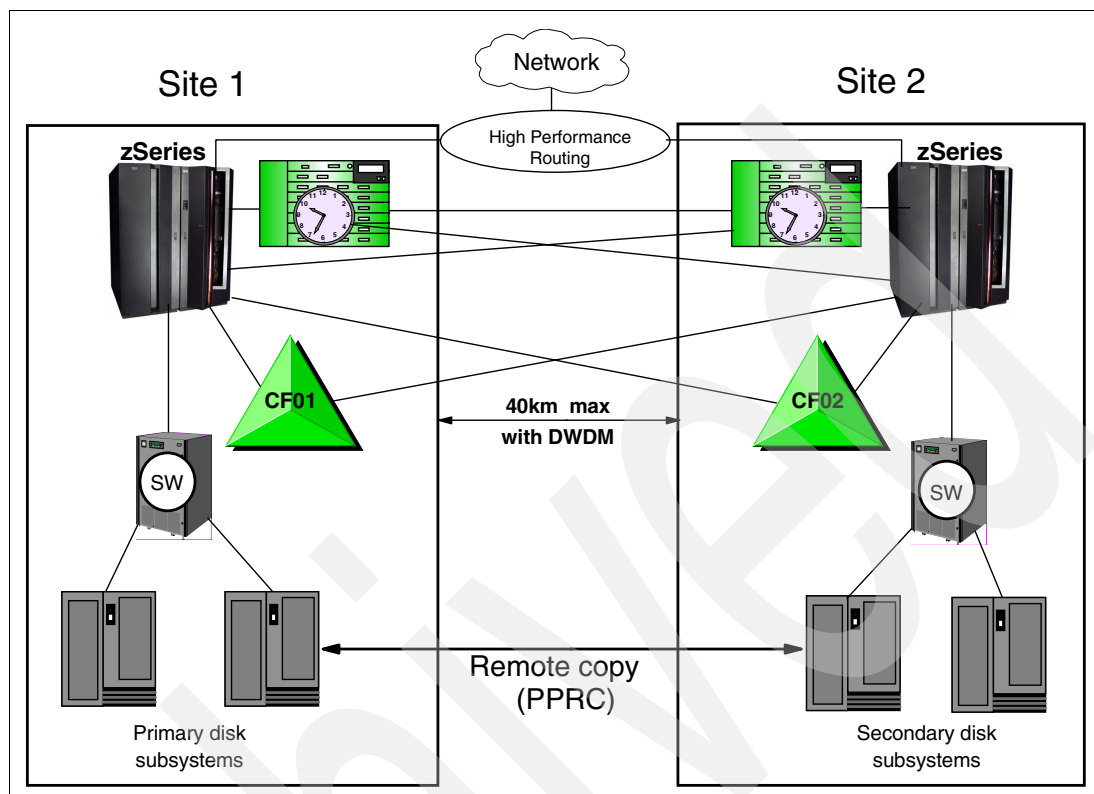


Figure 7-8 GDPS/PPRC (no extended distance RPQ)

All critical data resides on storage subsystems in site 1 (the primary copy of data) and is mirrored to site 2 (the secondary copy of data) via PPRC synchronous remote copy.

GDPS/PPRC is capable of the following attributes:

- ▶ Continuous availability
- ▶ Near-transparent disaster recovery
- ▶ Recovery Time Objective (RTO) less than an hour
- ▶ Recovery Point Objective (RPO) of zero (optional)
- ▶ Protects against localized area disasters

GDPS/PPRC HyperSwap™

The GDPS/PPRC HyperSwap function is designed to broaden the continuous availability attributes of GDPS/PPRC by extending the Parallel Sysplex redundancy to disk subsystems. The HyperSwap function can help significantly reduce the time needed to switch disks between sites and the time to switch sites. GDPS/PPRC HyperSwap provides the ability to transparently switch all primary PPRC disk subsystems with the secondary PPRC disk subsystems for a *planned* or *unplanned* reconfiguration, as shown on Figure 7-9 on page 174.

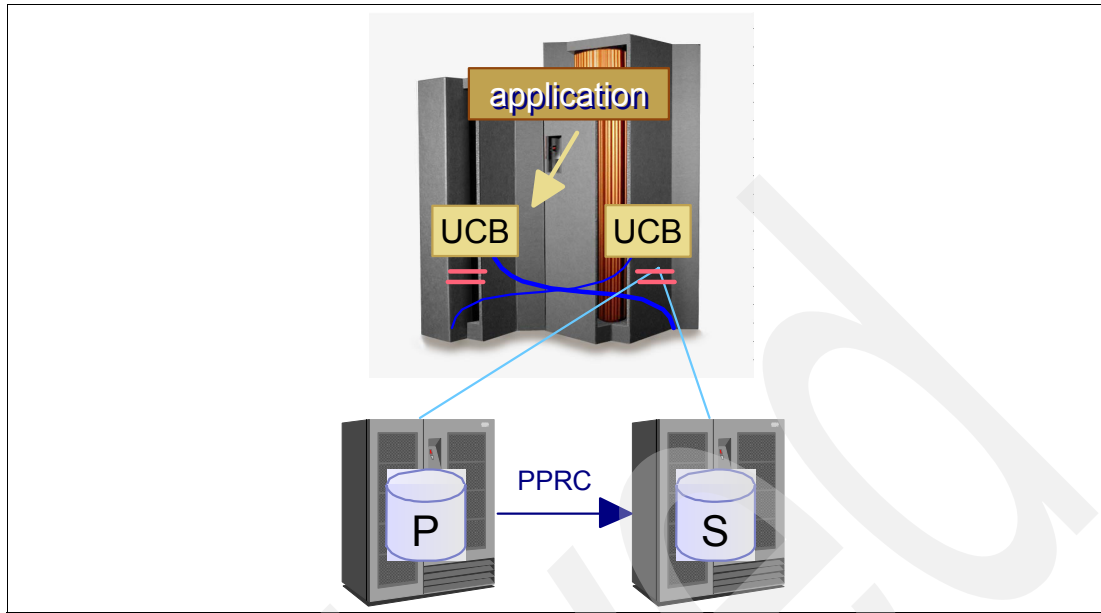


Figure 7-9 HyperSwap

Planned HyperSwap

GDPS/PPRC planned HyperSwap provides:

- ▶ The ability to switch all primary PPRC disk subsystems with the secondary PPRC disk subsystems for a planned reconfiguration and enables disk configuration maintenance and planned site maintenance without requiring any applications to be quiesced.
- ▶ Production systems and workload to remain active during maintenance of the site containing the primary PPRC disk subsystems, if applications are cloned and exploiting data sharing across the two sites.

See *GDPS: The e-business Availability Solution*, GF22-5114 for benchmark measurements of switch times, found at:

<http://www.ibm.com/servers/eserver/zseries/library/whitepapers/gf225114.html>

Unplanned HyperSwap

GDPS/PPRC unplanned HyperSwap allows:

- ▶ Production systems and workload to remain active during a disk subsystem failure. Disk subsystem failures will no longer constitute a single point of failure for an entire sysplex.
- ▶ Production systems to remain active during a failure of the site containing the primary PPRC disk subsystems, if applications are cloned and exploiting data sharing across the two sites. Even though the workload in the second site will need to be restarted, an improvement in the Recovery Time Objective (RTO) will be accomplished.

See *GDPS: The e-business Availability Solution*, GF22-5114 for benchmark measurements of switch times, found at

<http://www.ibm.com/servers/eserver/zseries/library/whitepapers/gf225114.html>

The GDPS/PPRC HyperSwap function is an integration of IBM @server and TotalStorage technologies. It integrates enhancements to GDPS code, z/OS, and ESS Licensed Internal Code.

The HyperSwap function is designed to be controlled by complete automation, allowing all aspects of the site switch to be controlled via GDPS.

GDPS/PPRC Management for Open Systems LUNs

GDPS/PPRC technology has been extended to manage a heterogeneous environment of z/OS and Open Systems data. If installations share their disk subsystems between the z/OS and Open Systems platforms, GDPS/PPRC, running in a z/OS system, can manage the PPRC status of devices that belong to the other platforms and are not even defined to the z/OS platform. GDPS/PPRC will also provide data consistency across both z/OS and Open Systems data.

GDPS/PPRC over Fiber Channel links

The IBM TotalStorage Enterprise Storage Server (ESS) supports PPRC over Fiber Channel for ESS Model 800. It is designed to improve throughput as compared to PPRC over ESCON links and reduces cross-site connectivity (two PPRC Fiber Channel links are considered sufficient for most workloads). The benefit of this support is the opportunity to increase the distance between sites without losing performance.

GDPS FlashCopy® V2 support

Previously, source and target volumes needed to reside on the same Logical Subsystem (LSS) within the disk subsystem. With FlashCopy V2, a flash copy can be created from a source in one LSS to a target in a different LSS in the same disk subsystem.

Business Continuity for Linux guests

GDPS plans to exploit the HyperSwap function in z/VM V5.1 to provide business continuity for z/OS and Linux guests. z/VM HyperSwap swaps the virtual device associated with one real disk to another and can be used to switch to a secondary disk storage subsystem mirrored by PPRC. This is a useful function for those users that share data and storage subsystems between z/OS and Linux. A SAP application server running on Linux and a SAP data base server running on a z/OS is an example of an environment that will benefit from the z/VM V5.1 HyperSwap functionality.

Much of the functionality is similar to that for z/OS systems and data. The following recovery actions are designed to support planned and unplanned outages:

- ▶ In place re-IPL of failing operating system images
- ▶ Site takeover of a production site
- ▶ Transparent planned and unplanned HyperSwap of disk subsystem

Near continuous availability and disaster recovery solutions require IBM Tivoli System Automation for Linux, and z/VM V5.1, in addition to other GDPS/PPRC prerequisites.

GDPS/PPRC Cross-site extended distance for Parallel Sysplex

Via a RPQ, the capability to configure GDPS/PPRC or a multi-site Parallel Sysplex up to a distance of 100 kilometers (62 miles) is made possible. Support for the following has been extended for up to 100 kilometers from the previous limitation of 50 kilometers (31 miles), through use of Dense Wavelength Division Multiplexer (DWDM) equipment for:

- ▶ External Timer Reference (ETR) links to a Sysplex Timer
- ▶ ISC-3 links in peer mode

This support is consistent with other technologies that support the same distance, such as FICON, Peer-to-Peer Remote Copy (PPRC), and Peer-to-Peer Virtual Tape Server (PtP VTS).

Restriction: The maximum distance between a pair of Sysplex Timers in an Expanded Availability configuration remains at 40 kilometers (25 miles). To achieve 100 kilometers distance between sites, one option is to consider an intermediate site at less than 40 kilometers from one or the other site or to place two Sysplex Timers in one site.

Figure 7-10 shows how a GDPS/PPRC Cross-site extended distance Parallel Sysplex can be established with RPQ 8P2263.

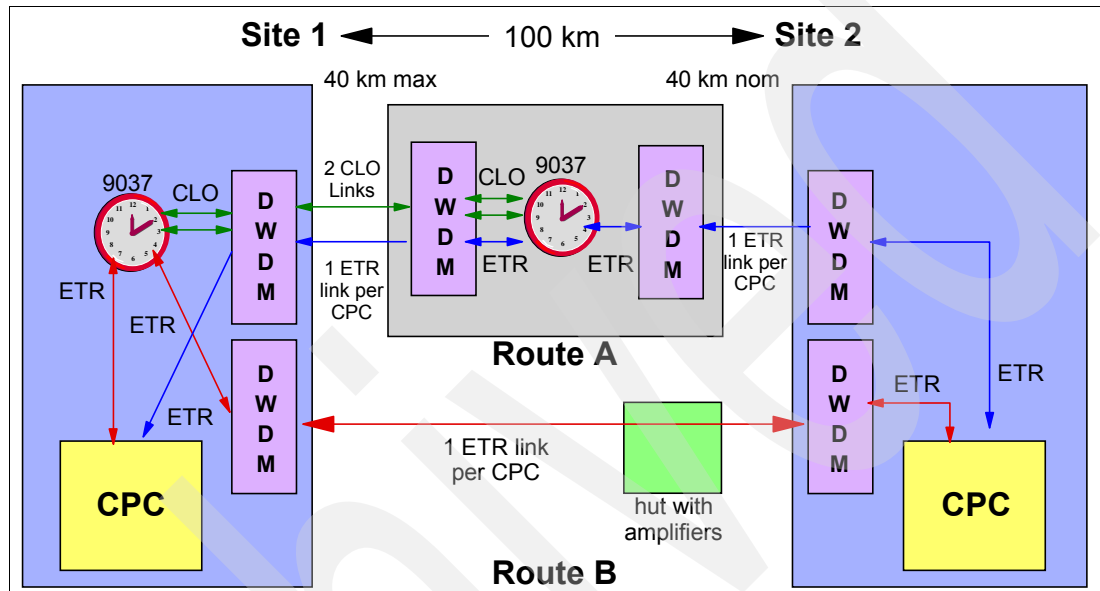


Figure 7-10 Cross-site extended distance for Parallel Sysplex and GDPS/PPRC

Note: In Figure 7-10, the midspan 9037 can also be located within 40 km of site 2 or on the South path; All ETR and CLO links are provisioned as one channel per wavelength.

7.4.2 GDPS/XRC

XRC is a combined hardware and software asynchronous remote copy solution. The application I/O is signalled completed when the data update to the primary storage is completed. A DFSMS component, called System Data Mover (SDM), asynchronously offloads data from the primary storage subsystem's cache and updates the secondary disk volumes in the recovery site.

The GDPS solution based on XRC, referred to as GDPS/XRC, has the attributes of a Disaster Recovery solution.

In GDPS/XRC, the production system(s) can be a single system, multiple systems sharing a disk, or a base or Parallel Sysplex cluster¹. GDPS/XRC provides a single, automated solution to dynamically manage storage subsystem mirroring (disk and tape) to allow a business to attain "near-transparent" disaster recovery with minimal data loss. GDPS/XRC is designed to provide the ability to perform a controlled site switch for an unplanned site outage, maintaining full data integrity across multiple volumes and storage subsystems, and the ability to perform a normal Data Base Management System (DBMS) restart - not DBMS recovery - at the opposite site. GDPS/XRC is application-independent and therefore covers the customer's complete application environment.

The physical topology of a GDPS/XRC, shown on Figure 7-11, consists of a production site (Site 1) and a recovery site (Site 2) located virtually at any distance from Site 1.

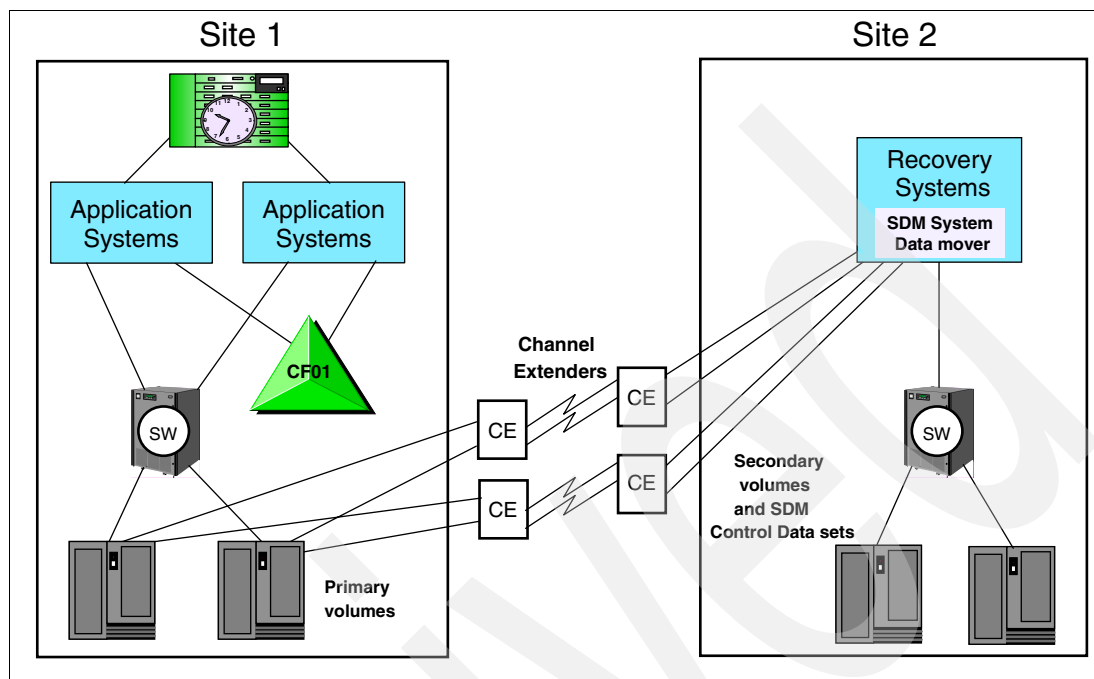


Figure 7-11 GDPS/XRC

GDPS/XRC is capable of the following attributes:

- ▶ Disaster recovery
- ▶ RTO between one and two hours
- ▶ RPO less than two minutes
- ▶ Protects against metropolitan as well as regional disasters (distance between sites is unlimited)
- ▶ Minimal remote copy performance impact

In a GDPS/XRC configuration, it is often necessary to have multiple System Data Movers (SDMs). The number of SDMs is based on many factors, such as the number of volumes being copied, the I/O rate, and so on. Functions are now capable of being executed in parallel across multiple SDMs, thus providing improved scalability for a coupled SDM configuration.

7.4.3 GDPS and Capacity Backup (CBU)

GDPS consists of production images and controlling images. The production images execute the mission-critical workload. There must be sufficient processing resource capacity, such as processor capacity, main storage, and channel paths available, that can quickly be brought online to restart an image's or site's critical workload. Typically, this is accomplished by terminating one or more systems executing expendable (non-critical) work and acquiring their processing resource.

The Capacity Backup (CBU) feature, available on the zSeries, provides a significant cost savings. The CBU feature has the ability to increment capacity temporarily, when capacity is lost elsewhere in the enterprise. CBU adds Central Processors (CPs) to the available pool of processors and is activated only in an emergency.

GDPS-CBU management automates the process of dynamically adding reserved Central Processors, thereby minimizing manual customer intervention and the potential for errors. The outage time for critical workloads can be reduced from hours to minutes.

Concurrent activation of Capacity Backup (CBU) can be performed in parallel across multiple servers, which results in an improved RTO. This applies to both the GDPS/PPRC and GDPS/XRC configurations.

Similarly, GDPS-CBU management can also automate the process of dynamically returning the reserved CPs when the temporary period has expired.

7.5 Intelligent Resource Director

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

IRD overview

The Intelligent Resource Director (IRD) is a new feature introduced in z/OS, extending the concept of goal-oriented resource management by allowing you to group system images that are resident on the same zSeries server running in LPAR mode, and in the same Parallel Sysplex, into an “LPAR cluster.” This gives Workload Management 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 7-12 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 an OS/390 image, as well as a Coupling Facility image. In this example, the scope that IRD has control over is the defined LPAR cluster.

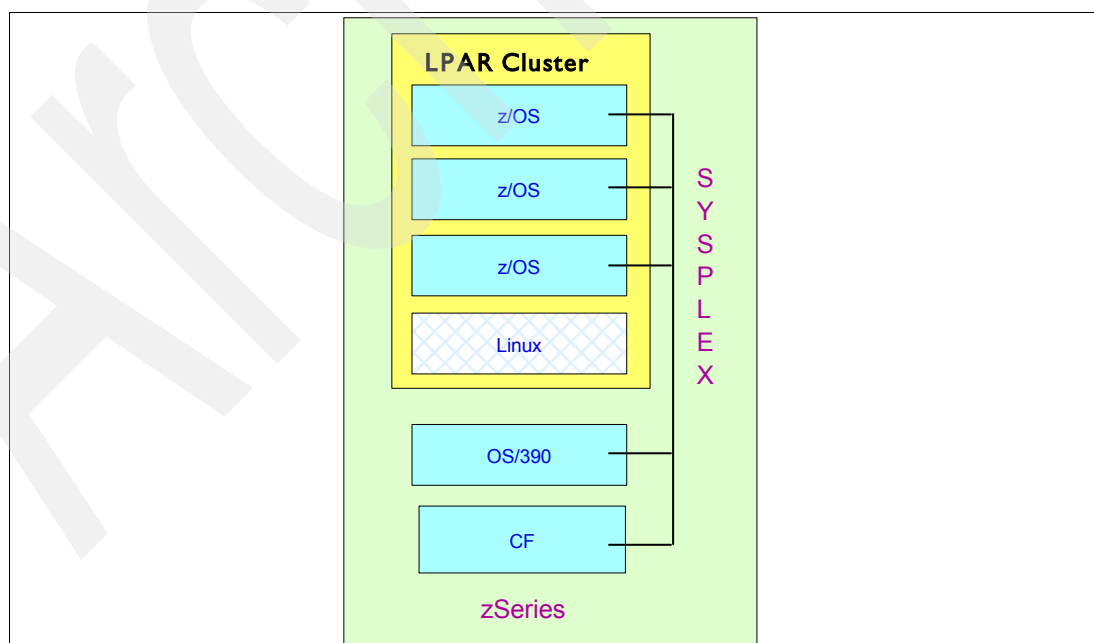


Figure 7-12 IRD LPAR cluster example

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.

- ▶ Dynamic channel path management (DCM)

DCM moves channel bandwidth between disk control units to address current processing needs. The z990 supports DCM within a Logical Channel Subsystem.

- ▶ Channel Subsystem Priority Queuing

This feature on the zSeries allows the priority queueing 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.

7.5.1 LPAR CPU management

LPAR CPU management allows WLM working in goal mode to manage the processor weighting and logical processors across an LPAR cluster.

LPAR CPU management was enhanced in z/OS 1.2 to dynamically manage non-z/OS operating systems, such as Linux and z/VM. This function allows z/OS WLM to manage the CPU resources given to these partitions based on their relative importance compared to the other workloads running in the same LPAR cluster.

Note: In order to manage non-z/OS images, such as Linux, z/VM, VM/ESA, TPF, z/VSE, or VSE/ESA, at least one image in the LPAR Cluster must be running z/OS 1.2 or higher.

Workload Manager distributes processor resources across an LPAR cluster by dynamically adjusting the LPAR weights in response to changes in the workload requirements. When important work is not meeting its goals, WLM will raise the weight of the partition where that work is running, thereby giving it more processing power. As the LPAR weights change, the number of online logical CPUs may also be changed to maintain the closest match between logical CPU speed and physical CPU speed.

LPAR CPU management runs on a zSeries server in z/Architecture mode, and in LPAR mode only. The participating z/OS system images must be running in goal mode. It also requires a CF level 9 or above Coupling Facility structure.

Enabling LPAR CPU management involves defining the Coupling Facility structure and then performing several operations on the hardware management console: defining logical CPs, and setting initial, minimum, and maximum processing weights for each logical partition.

CPU resources are automatically moved toward logical partitions with the most need by adjusting the partition's weight. The sum of the weights for the participants in an LPAR cluster is viewed as a pooled resource that can be apportioned among the participants to meet the goal mode policies. The installation can place limits on the processor weight value.

WLM will also manage the available processors by varying off unneeded CPs (more logical CPs implies more parallelism, and less weight per CP).

Value of CPU management

The benefits of CPU management include the following:

- ▶ Logical CPs perform at the fastest uniprocessor speed available.

This results in the number of logical CPs tuned to the number of physical CPs of service being delivered by the logical partition current weight. If the logical partition is getting four equivalent physical CPs of service and has eight logical CPs online to z/OS, then each logical CP only gets half of an equivalent physical CP. For example, if a CP delivers 200 MIPS, half of it will deliver 100 MIPS. This occurs because each logical CP gets fewer time slices.

- ▶ Reduced PR/SM overhead.

There is a PR/SM overhead for managing a logical CP. The higher the number of logical CPs in relation to the number of equivalent physical CPs, the higher the PR/SM overhead. This is because PR/SM has to do more processing to manage the number of logical CPs that exceeds the number of equivalent physical CPs.

- ▶ z/OS gets more control over how CP resources are distributed.

Using CPU management, z/OS is able to manage CP resources in relation to WLM goals for work. This was not possible in the past when a logical partition had CP resources assigned and used these as best it could in one logical partition. Now, z/OS is able to change the assigned CP resources (LPAR weights) and place them where they are required for the work. CPU management does the following:

- Identifies what changes are needed and when.
- Projects the likely results on both the work it is trying to help and the work that it will be taking the resources from.
- Performs the changes.
- Analyzes the results to ensure the changes have been effective.

There is also the question of the speed at which an operator can perform these actions. WLM can perform these actions every Policy Adjustment interval, which is normally ten seconds, as determined by WLM. It is not possible for an operator to perform all the tasks in this time.

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

7.5.2 Dynamic Channel Path Management

There is no such thing as a “typical” workload. The requirements for processor capacity, I/O capacity, and other resources vary throughout the day, week, month, and year.

Dynamic Channel Path Management (DCM) provides the ability to have the system automatically manage the number of paths available to disk subsystems. By making additional paths available where they are needed, the effectiveness of your installed channels is increased, and the number of channels required to deliver a given level of service is potentially reduced.

DCM also provides availability benefits by attempting to ensure that the paths it adds to a control unit have as few points of failure in common with existing paths as possible, and configuration management benefits by allowing the installation to define a less specific configuration. On a z990 where paths can be shared by Multiple Image Facility (MIF), DCM will coordinate its activities across logical partitions within a single Logical Channel Subsystem on a server within a single sysplex.

Where several channels are attached from a z990 LCSS to a switch, they can be considered a resource pool for accessing any of the control units attached to the same switch. To achieve this without DCM would require deactivating paths, performing a dynamic I/O reconfiguration, and activating new paths. DCM achieves the equivalent process automatically, using those same mechanisms.

Channels managed by DCM are referred to here as “managed” channels. Channels not managed by DCM are referred to as “static” channels.

Workload Manager dynamically moves channel paths through the ESCON Director from one I/O control unit to another in response to changes in the workload requirements. By defining a number of channel paths as managed, they become eligible for this dynamic assignment.

By moving more bandwidth to the important work that needs it, your disk I/O resources are used much more efficiently. This may decrease the number of channel paths you need in the first place, and could improve availability because, in the event of a hardware failure, another channel could be dynamically moved over to handle the work requests.

Dynamic Channel Path Management runs on a zSeries server in z/Architecture mode, in both basic and LPAR mode. The participating z/OS system images can be defined as XCFLOCAL, MONOPLEX, or MULTISYSTEM.

If a system image running Dynamic Channel Path Management in LPAR mode is defined as being part of a multisystem sysplex, it also requires a CF level 9 Coupling Facility structure, even if it is the only image currently running on the system.

Dynamic Channel Path Management operates in two modes:

- ▶ **Balance mode**

In balance mode, DCM will attempt to equalize performance across all of the managed control units.

- ▶ **Goal mode**

In goal mode, which is available only when WLM is operating in goal mode on systems in an LPAR cluster, DCM will still attempt to equalize performance, as in balance mode. In addition, when work is failing to meet its performance goals due to I/O delays, DCM will take additional steps to manage the channel bandwidth accordingly, so that important work meets its goals.

Enabling Dynamic Channel Path Management involves defining managed channels and control units via HCD. On the Hardware Management Console, you then need to ensure that all of the appropriate logical partitions are authorized to control the I/O configuration.

For additional information on implementing Dynamic Channel Path Management under IRD, see *z/OS Intelligent Resource Director*, SG24-5952.

Value of Dynamic Channel Path Management

Dynamic Channel Path Management provides the following benefits:

- ▶ **Improved overall image performance**

Improved image performance is achieved by automatic path balancing (WLM compatibility and goal mode) and Service Policy (WLM goal mode).

- ▶ **Maximum utilization of installed hardware**

Channels will be automatically balanced, providing opportunities to use fewer I/O paths to service the same workload.

- ▶ Simplified I/O definition

The connection between managed channels and managed control units does not have to be explicitly defined.

- ▶ Reduced skills required to manage z/OS

Managed channels and control units are automatically monitored, balanced, tuned, and reconfigured.

- ▶ Enhanced availability

A failing or hung channel path will result in reduced throughput on the affected control unit. DCM will rapidly detect the symptom and augment the paths, automatically bypassing the problem. The problem will still have to be analyzed and corrected by site personnel.

DCM will automatically analyze and minimize single points of failure on an I/O path by selecting appropriate paths. DCM is sensitive to single points of failure, such as:

- ESCON or FICON channel cards
- I/O CHA cards
- Processor Self-Timed Interconnect
- Director port cards
- Control Unit I/O bay
- Control Unit Interface card
- ESCON Director

7.5.3 Channel Subsystem Priority Queueing

Channel Subsystem (CSS) Priority Queueing is a new function available on zSeries processors in either¹ basic or LPAR mode. It allows the z/OS operating system to specify a priority value when starting an I/O request. When there is contention causing queueing in the Channel Subsystem, the request is prioritized by this value.

If important work is missing its goals due to I/O contention on channels shared with other work, it will be given a higher Channel Subsystem I/O priority than the less important work. This function goes hand in hand with the Dynamic Channel Path Management described previously: as additional channel paths are moved to control units to help an important workload meet goals, Channel Subsystem Priority Queueing ensures that the important work receives greater access to additional bandwidth than less important work that happen to be using the same channel.

Channel Subsystem Priority Queueing runs on a zSeries server in z/Architecture mode, in both basic and LPAR mode. The participating z/OS system images can be defined as XCFLOCAL, MONOPLEX, or MULTISYSTEM. It is optimized when WLM is running in goal mode. It does not require a Coupling Facility structure.

Enabling Channel Subsystem Priority Queueing involves defining a range of I/O priorities for each logical partition on the hardware management console, and then turning on the “Global input/output (I/O) priority queueing” switch. (You also need to specify “YES” for WLM’s I/O priority management setting.)

z/OS will set the priority based on a goal mode WLM policy. This complements the goal mode priority management that sets I/O priority for IOS UCB queues, and for queueing in the 2105 ESS disk subsystem.

CSS Priority Queueing uses different priorities calculated in a different way from the I/O priorities used for UCB and control unit queueing.

¹ The z990 operates in LPAR mode only

Value of Channel Subsystem Priority Queueing

The benefits proved by Channel Subsystem Priority Queueing include the following:

- Improved performance

I/O from work that is not meeting its goals may be given priority over I/O from work that is meeting its goals, providing Workload Manager with an additional method for adjusting I/O performance. Channel Subsystem Priority Queueing is complementary to UCB priority queueing and control unit priority queueing, each addressing a different queueing mechanism that may affect I/O performance.

- Reduced skills required to manage z/OS

Monitoring and tuning requirements are reduced because of the self-tuning abilities of the Channel Subsystem.

7.5.4 WLM and Channel Subsystem priority

WLM assigns the highest to lowest CSS priority, as given in Table 7-5. It assigns eight priority levels.

Table 7-5 WLM-assigned CSS I/O priorities

Workload type	Priority
System work.	FF
Importance of 1 and 2 missing goals.	FE
Importance of 3 and 4 missing goals.	FD
Meeting goals. Adjust by ratio of connect time to elapsed time.	F9-FC
Discretionary.	F8

Work that is meeting its WLM target is assigned CSS priorities between F9 and FC, depending on its execution profile. Work that has a light I/O usage has its CSS priority moved upwards.

When an I/O operation is started by a CP on the Server, it can be queued by the Channel Subsystem for several reasons, including Switch port busy, Control unit busy, Device busy, and All channel paths busy. Queued I/O requests are started or restarted when an I/O completes or the Control unit indicates the condition has cleared. Where two or more I/O requests are queued in the Channel Subsystem, the CSS LIC on the zSeries selects the requests in priority order. The LIC also ages requests to ensure that low priority requests are not queued for excessive periods.

In the LPAR image profile for the z/OS image, there are two specifications that relate to the Channel Subsystem I/O Priority Queueing. They are:

- The range of priorities that will be used by this image
- The default Channel Subsystem I/O priority

For images running operating systems that do not support Channel Subsystem priority, the customer can prioritize all the Channel Subsystem requests coming from that image against the other images by specifying a value for the default priority.

Within an LPAR cluster, the prioritization is managed by WLM goal mode and coordinated across the cluster. Hence, the range should be set identically for all logical partitions in the same LPAR cluster.

WLM sets priorities within a range of eight values that will be mapped to the specified range. If a larger range is specified, WLM uses the top eight values. If a smaller range is specified, WLM maps its values into the smaller range, retaining as much function as possible within the allowed range. Note that the WLM calculated priority is still a range of 8. The mapped priority is shown in Table 7-6.

A range of eight values is recommended for CSS I/O priority-capable logical partitions. If the logical partition is run in compatibility mode or with I/O priority management disabled, the I/O priority is set to the middle of the specified range.

Table 7-6 WLM CSS priority range mapping with specified range less than 8

WLM CSS priorities (range width)	Calculated range (8)	Specified range (7)	(6)	(5)	(4)	(3)	(2)
System work.	FF	FF	FF	FF	FF	FF	FF
Importance of 1 and 2 missing goals.	FE	FE	FE	FE	FE	FE	FF
Importance of 3 and 4 missing goals.	FD	FE	FE	FE	FE	FE	FF
Meeting goals. Adjust by ratio of connect time to elapsed time.	FC-F9	FD-FA	FD-FB	FD-FC	FD	FE	FF
Discretionary.	F8	F9	FA	FB	FC	FD	FE

7.5.5 Special considerations and restrictions

Here we discuss some special considerations and restrictions of Sysplex functions.

Unique LPAR cluster names

LPAR clusters, running on a 2064, 2066, 2084, or 2086 server, must be uniquely named. This is the sysplex name that is associated with the LPAR cluster. Managed channels have an affinity (are owned by) a specific LPAR cluster. Non-unique naming creates problems in terms of scope of control.

Disabling Dynamic Channel Path Management

To disable Dynamic Channel Path Management within an LPAR cluster running z/OS, turning off the function by using the SETIOS DCM=OFF command is not sufficient. Although a necessary step, this does not ensure that the existing configuration is adequate to handle your workload needs, since it leaves the configuration in the state it was at the time the function was disabled. During your migration to DCM, we would recommend that you continue to maintain your old IODF until you are comfortable with DCM. This will allow you to back out of DCM by activating a known configuration.

Automatic I/O interface reset

When going through all of the steps to enable Dynamic Channel Path Management, also ensure that the “Automatic input/output (I/O) interface reset” option is enabled on the Hardware Management Console. This will allow Dynamic Channel Path Management to continue functioning in the event that one participating system image fails.

This is done by enabling the option in the reset profile used to activate the server. Using the “Customize/Delete Activation Profiles task” available from the “Operational Customization tasks list,” open the appropriate reset profile and then open the Options page to enable the option.

System automation - I/O operations

When using system automation, take care when using PROHIBIT or BLOCK on a port that is participating in Dynamic Channel Path Management.

When blocking a managed channel port, configuring the CHPID OFFLINE to all members of the LPAR Cluster is all that is required. Dynamic Channel Path Management will ensure that if the CHPID is configured to managed subsystems, then the CHPID will be deconfigured from all subsystems to which it is currently configured.

When blocking a port connected to a managed subsystem, the port must first be disabled for Dynamic Channel Path Management usage. This is done using the VARY SWITCH command to take the port OFFLINE to Dynamic Channel Path Management. This command should be issued on all partitions that are running DCM. Disabling the port for Dynamic Channel Path Management usage will deconfigure all managed channels which are connected to the subsystem through that port. Once the port is disabled to Dynamic Channel Path Management, it can then be blocked.

When prohibiting a set of ports, if any of the ports are connected to managed subsystems, then the PROHIBIT operation must be preceded by the VARY SWITCH command(s) to disable the managed subsystem ports to Dynamic Channel Path Management. As in the blocking case, this will cause any managed channels currently connected to the subsystem port(s) to be deconfigured. Once the subsystem ports are disabled to Dynamic Channel Path Management, the PROHIBIT function can be invoked. This must then be followed by the VARY SWITCH command(s) to re-enable the prohibited subsystem ports to Dynamic Channel Path Management.

When ports are unprohibited or unblocked, these operations need to be followed, as necessary, by VARY SWITCH commands to bring ports ONLINE to Dynamic Channel Path Management.

7.5.6 References

For more detailed information on Intelligent Resource Director, see *z/OS MVS Planning Workload Management*, SA22-7602, and the IBM Redbook *z/OS Intelligent Resource Director*, SG24-5952.

Archived

Capacity upgrades

This chapter describes the zSeries 990 server's capacity upgrade functions and features. It also includes capacity measurements and performance considerations.

The z990 servers have the capability of *concurrent* upgrades, without a server outage, in both planned and unplanned situations.

In most cases, a z990 capacity upgrade can also be *nondisruptive*, without a system outage.

The following sections are included:

- ▶ 8.1, "Concurrent upgrades" on page 188
- ▶ 8.2, "Capacity Upgrade on Demand (CUoD)" on page 190
- ▶ 8.3, "Customer Initiated Upgrade (CIU)" on page 196
- ▶ 8.4, "On/Off Capacity on Demand (On/Off CoD)" on page 202
- ▶ 8.5, "Capacity BackUp (CBU)" on page 206
- ▶ 8.6, "Nondisruptive upgrades" on page 210
- ▶ 8.7, "Capacity planning considerations" on page 219
- ▶ 8.8, "Capacity measurements" on page 223

8.1 Concurrent upgrades

The z990 servers have the capability of concurrent upgrades, providing additional capacity with no *server* outage. In most cases, with prior planning and operating system support, a concurrent upgrade can also be nondisruptive, that is, without *system* outage (Power-on Resets (PORs), logical partition deactivations, and IPLs do not have to take place).

Given today's business environment, the benefits of the concurrent capacity growth capabilities provided by z990 servers are plentiful, and include:

- ▶ 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

This capability is based on the flexibility of the z990 system design and structure, which allows configuration control by the Licensed Internal Code (LIC) and concurrent hardware installation.

Licensed Internal Code (LIC)-based upgrades

The LIC - Configuration Control (LIC-CC) provides for server upgrade with no hardware changes by enabling the activation of additional, previously installed capacity. Concurrent upgrades via LIC-CC can be done for:

- ▶ Processors (CPs, IFLs, ICFs, and zAAPs)
Requires available spare PUs on installed book(s).
- ▶ Memory
Requires available capacity on installed memory cards.
- ▶ I/O cards ports (ESCON channels and ISC-3 links)
Requires available ports on installed I/O cards.

Concurrent hardware installation upgrades

Configuration upgrades can also be concurrent by installing additional:

- ▶ Books (which contain processors, memory, and STIs)
Requires available book slot(s) in the installed CEC cage.
- ▶ I/O cards
Requires available slots on installed I/O cage(s). I/O cages *cannot* be installed concurrently.

The concurrent upgrade capability can be better exploited when a future target configuration is considered in the initial configuration. Using this Plan Ahead concept, the required number of I/O cages for concurrent upgrades, up to the target configuration, can be included in the z990 server's initial configuration.

Concurrent PU conversions

z990 servers support concurrent conversion between different PU types, providing flexibility to meet changing business environments.

These LIC-CC based PU conversions, as listed in Table 2-7 on page 65, require that at least one PU (CP, ICF or IFL) remain untouched; otherwise, the conversion is disruptive. The PU conversion generates a new LIC-CC that can be installed concurrently in two steps. First, the assigned PU is removed from the z990 configuration. Second, the newly available PU is activated as the new PU type.

Logical partitions may also need to “free” PUs to be converted, and the operating systems must have the “configure offline/online” capability to do the PU conversion nondisruptively.

Model upgrades

The z990 servers have both a server model (2084-xxx) and a software model (3xx).

The server model indicates how many books are present on the configuration, while the software model indicates how many CPs are present for software billing-related purposes.

z990 model upgrades always require physical hardware (books) addition. z990 servers upgrades can change either, or both, the server and the software models:

- ▶ On LIC-only upgrades:
 - May change the server’s software model (3xx) if additional CPs are included.
 - Cannot change the server’s model (2084-xxx), as no additional books can be included.
- ▶ On hardware installation upgrades:
 - May change the server’s model (2084-xxx) if additional books are included.
 - May change the server’s software model (3xx) if additional CPs are included.

Both the server and the software models can be concurrently upgraded.

Concurrent upgrades can be accomplished in both *planned* and *unplanned* upgrade situations.

Important: If the z990 STI Rebalance feature (FC 2400) is selected and effectively results in STI rebalancing, the server upgrade will be disruptive and this outage must be planned. The z990 STI Rebalance feature may also change the Physical Channel ID (PCHID) of ICB-4 links, requiring a corresponding update on the server I/O definition via HCD or HCM.

Planned upgrades

Planned upgrades can be done by the Capacity Upgrade on Demand (CUoD), the Customer Initiated Upgrade (CIU), or the On/Off Capacity on Demand (On/Off CoD) functions.

CUoD and CIU are functions available on z990 servers that enable *concurrent* and *permanent* capacity growth of a z990 server.

CUoD can concurrently add processors (CPs, IFLs, ICFs, and zAAPs), memory, and I/O ports to an existing server. CUoD requires IBM service personnel for the upgrade.

CIU can concurrently add processors (CPs, IFLs, ICFs, and zAAPs) and memory up to the limit of the installed book(s) of an existing server. CIU is initiated by the customer via the Web using IBM Resource Link, and makes use of CUoD techniques. CIU requires a special contract.

On/Off CoD is a function available on z990 servers that enables *concurrent* and *temporary* capacity growth of a z990 server. On/Off CoD *can* be used for customer peak workload requirements, for any length of time.

On/Off CoD can concurrently add processors (CPs, IFLs, ICFs, and zAAPs) up to the limit of the installed book(s) of an existing server, and is restricted to double the current installed capacity. On/Off CoD uses the CIU ordering process, initiated by the customer via the Web using IBM Resource Link, and makes use of CUoD techniques. On/Off CoD requires a special contract.

Unplanned upgrades

Unplanned upgrades can be done by the Capacity BackUp (CBU) for emergency or disaster/recovery situations.

CBU is a *concurrent* and *temporary* activation of Central Processors (CPs) in the face of a loss of customer processing capacity due to an emergency in any customer's zSeries or S/390 server or servers at any sites or locations. CBU *cannot* be used for peak load management of customer workload. A CBU activation can last up to 90 days when a disaster/recovery situation occurs.

CBU features, one for each "standby" CP, are optional on z990 servers and require spare PUs on installed book(s) of the existing server. A CBU contract must be in place before the special code that enables this capability can be loaded on the customer server.

Capacity upgrade functions

Table 8-1 summarizes the capacity upgrade functions available for z990 servers.

Table 8-1 Capacity upgrade functions summary

Function	Upgrades	Via	Type	Process
CUoD	CPs, IFLs, ICFs, and zAAPs Memory I/O	LIC or hardware installation	Concurrent and permanent	Ordered as a normal upgrade and activated by IBM
CIU	CPs, IFLs, ICFs, and zAAPs Memory	LIC-only cannot add book	Concurrent and permanent	Initiated via Web and activated by customer
On/Off CoD	CPs, IFLs, ICFs, and zAAPs	LIC-only cannot add book	Concurrent and temporary (no time limit)	Initiated via Web and activated by customer
CBU	CPs	LIC-only cannot add book	Concurrent and temporary (up to 90 days)	Ordered for backup/recovery only and activated by customer

8.2 Capacity Upgrade on Demand (CUoD)

Capacity Upgrade on Demand (CUoD) is a function available on z990 servers that enables *concurrent* and *permanent* capacity growth.

The CUoD function is based on the Configuration Reporting Architecture, which provides detailed information on system-wide changes, such as the number of configured Processor Units, system serial numbers, and other information.

CUoD provides the ability to concurrently add processors (CPs, IFLs, ICFs, and zAAPs), memory capacity, and I/O ports. The concurrent upgrade can be done by Licensed Internal

Code Configuration Control (LIC-CC) only or also by installing additional book(s) and/or I/O card(s):

- ▶ CUoD upgrades for processors are done by either:
 - LIC-CC assigning and activating spare PUs up to the limit of the current installed book(s)
 - Installing additional book(s) and LIC-CC assigning and activating spare PUs on installed book(s)
- ▶ CUoD upgrades for memory are done by either:
 - LIC-CC activating additional memory capacity up to the limit of the memory cards on the current installed book(s)
 - Installing additional book(s) and LIC-CC activating additional memory capacity on installed book(s)
- ▶ CUoD upgrades for I/O are done by either:
 - LIC-CC activating additional ports on already installed ESCON and ISC-3 cards
 - Installing additional I/O card(s) and supporting infrastructure if required on already installed I/O cage(s)

Important: If the z990 STI Rebalance feature (FC 2400) is selected at server upgrade configuration time, and effectively results in STI rebalancing, the server upgrade will be disruptive and this outage must be planned. The STI rebalancing operation may also be done independently of a model upgrade.

The z990 STI Rebalance feature may also change the Physical Channel ID (PCHID) number of ICB-4 links, requiring a corresponding update on the server's I/O definition via HCD or HCM.

CUoD is ordered as a “normal” upgrade, also known as Miscellaneous Equipment Specification (MES).

CUoD does not require any special contract, but requires IBM service personnel for the upgrade. In most cases, a very short period of time is required for the IBM personnel to install the LIC-CC and complete the upgrade.

To better exploit the CUoD function, an initial configuration should be carefully planned to allow a concurrent upgrade up to a target configuration.

You need to consider planning, positioning, and other issues to allow a CUoD *nondisruptive* upgrade. By planning ahead, it is possible to enable nondisruptive capacity and I/O growth for the z990 with no system power down and no associated POR or IPLs.

The Plan Ahead feature involves pre-installation of additional I/O cage(s), as it is not possible to install an I/O cage concurrently.

Note: CUoD basically provides a “physical” concurrent upgrade, resulting in more enabled processors, memory, and/or I/O ports available to a server configuration. Thus, additional planning and tasks are required for *nondisruptive* “logical” upgrades (see “Recommendations to avoid disruptive upgrades” on page 217).

CUoD for processors

CUoD for processors can add, *concurrently*, more CPs, IFLs, ICFs, and zAAPs to a z990 server by assigning available spare PUs via LIC-CC. Depending on the quantity of the

additional CPs, IFLs, ICFs, and zAAPs in the upgrade, additional book(s) may be required and can be concurrently installed before the LIC-CC enablement.

Note: The sum of CPs, unassigned CPs, IFLs, unassigned IFLs, ICFs, and zAAPs cannot exceed eight PUs per book. The number of zAAPs cannot exceed four zAAPs per book. The total number of zAAPs cannot exceed the number of CPs plus unassigned CPs on a z990 server.

Important: CUoD for processors is not supported when CBU or On/Off CoD is *activated* on a z990 server. CUoD for processors can be applied after the temporary capacity upgrade via CBU or On/Off CoD is deactivated.

Figure 8-1 is an example of CUoD for processors, showing the eight PUs per book that can be assigned as CPs, IFLs, ICFs, or zAAPs.

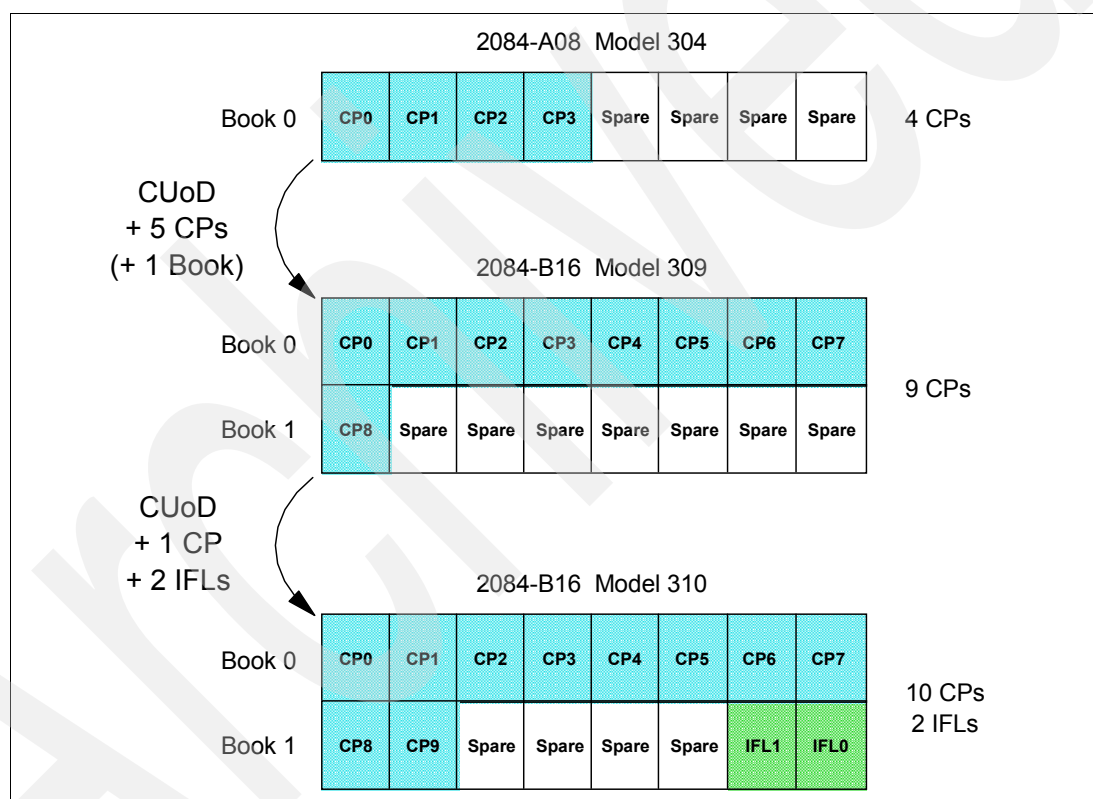


Figure 8-1 CUoD for processor example

An initial z990 server 2084-A08 (one book), software model 304 (four CPs) is concurrently upgraded to a 2084-B16 (two books), software model 309 (nine CPs). The model upgrade requires adding a book and assigning and activating five PUs as CPs.

Then the 2084-B16, software model 309, is concurrently upgraded to a software model 310 (10 CPs) with two IFLs by assigning and activating three more spare PUs (one as CP and two as IFLs).

Additional logical processors can be concurrently configured online to logical partitions by the operating system when reserved processors are previously defined, resulting in image

upgrades. The operating system must have the capability to concurrently configure more processors online.

Attention: Up to 32 logical processors, including reserved processors, can be defined to a logical partition. z/OS 1.6 is planned to support up to 24 processors, as a combination of CPs and zAAPs. z/VM 5.1 is planned to support up to 24 processors, which can be either all CPs or all IFLs.

Software charges based on the total capacity of the server on which the software is installed would be adjusted to the maximum capacity after the CUoD upgrade. Refer to Table 6-3 on page 148 to check software implications for CUoD.

Software products using Workload License Charge (WLC) may not be affected by the server upgrade, as their charges are based on partition utilization and not based on the server total capacity. Refer to 6.8, “Workload License Charges” on page 150 for more information about WLC.

CUoD for memory

CUoD for memory can add, *concurrently*, more memory to a z990 server by enabling, via LIC-CC, additional capacity up to the limit of the current installed memory cards, and/or by installing concurrently additional book(s) and LIC-CC enabling memory capacity on the new book(s).

The memory card sizes on the z990 are 8, 16, or 32 GB, and each book has two memory cards with the same physical storage capacity.

Note: Upgrades requiring memory card changes on any installed book are disruptive.

Table 2-1 on page 28 lists the range of system memory associated with a given memory card size and the number of memory cards for each server model. Figure 8-2 on page 194 shows an example of CUoD for memory of a 2084-A08 server with 24 GB of available memory.

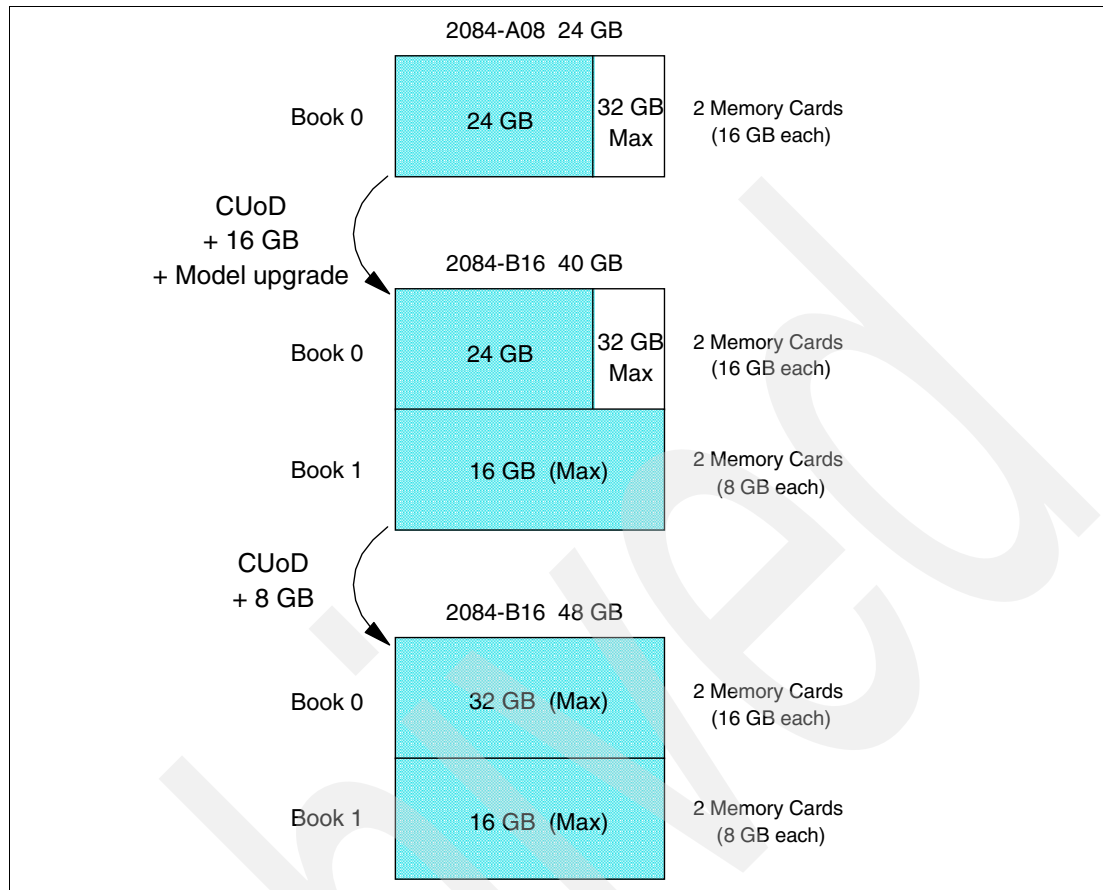


Figure 8-2 CUoD for memory example

This one-book z990 model has two 16 GB memory cards, resulting in 32 GB of installed memory in total. Therefore, a concurrent memory upgrade within this model A08 can be done up to the 32 GB limit, via LIC-CC, but a memory upgrade to 40 GB would require the book's memory cards replacement by two 32 GB memory cards and is *disruptive*.

However, as shown in the example, the upgrade of this model A08 with 24 GB of memory to a model B16 with 40 GB is *concurrent*, as the additional book comes with two memory cards (in this case, two 8 GB memory cards). The additional 16 GB memory capacity is enabled by LIC-CC on the second book (Book 1).

In the last part of this example, this model B16 server is concurrently upgraded to 48 GB, by LIC-CC enabling all the installed memory.

For a logical partition's memory upgrade, reserved storage must have been previously defined to that logical partition. It makes use of the LPAR Dynamic Storage Reconfiguration (DSR) function. DSR allows a z/OS or OS/390 operating system running in a partition to add its reserved storage to its configuration, if any unused storage exists. When the operating system running in a partition requests an assignment of a storage increment to its configuration, PR/SM checks for any free storage and brings it online dynamically.

Concurrent memory upgrades also require that:

- Memory must not be running in degraded mode.

Upgrades are disruptive until failing memory cards have been replaced.

- The new amount of installed memory cannot cause the storage granularity or increment to change.

However, a new Reset Profile (to allow the customer to potentially select a higher storage increment to Plan Ahead for concurrent memory upgrade) will be available.

The Minimum Storage Granularity will be the required storage granularity based on what memory is currently LIC-CC installed. The Maximum Concurrent Upgrade Value will be the smaller of the amount of storage that is physically installed and the maximum storage allowed for the Minimum Storage Granularity (see 2.2.10, “LPAR storage granularity” on page 70).

CUoD for I/O

CUoD for I/O can add, *concurrently*, more I/O ports to a z990 server by either:

- Enabling additional ports on the already installed I/O cards via LIC-CC.

LIC-CC-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 cage(s) must provide the number of I/O slots required by the target configuration.

Note: I/O cages *cannot* be installed concurrently.

Figure 8-3 shows an example of CUoD for I/O via LIC-CC.

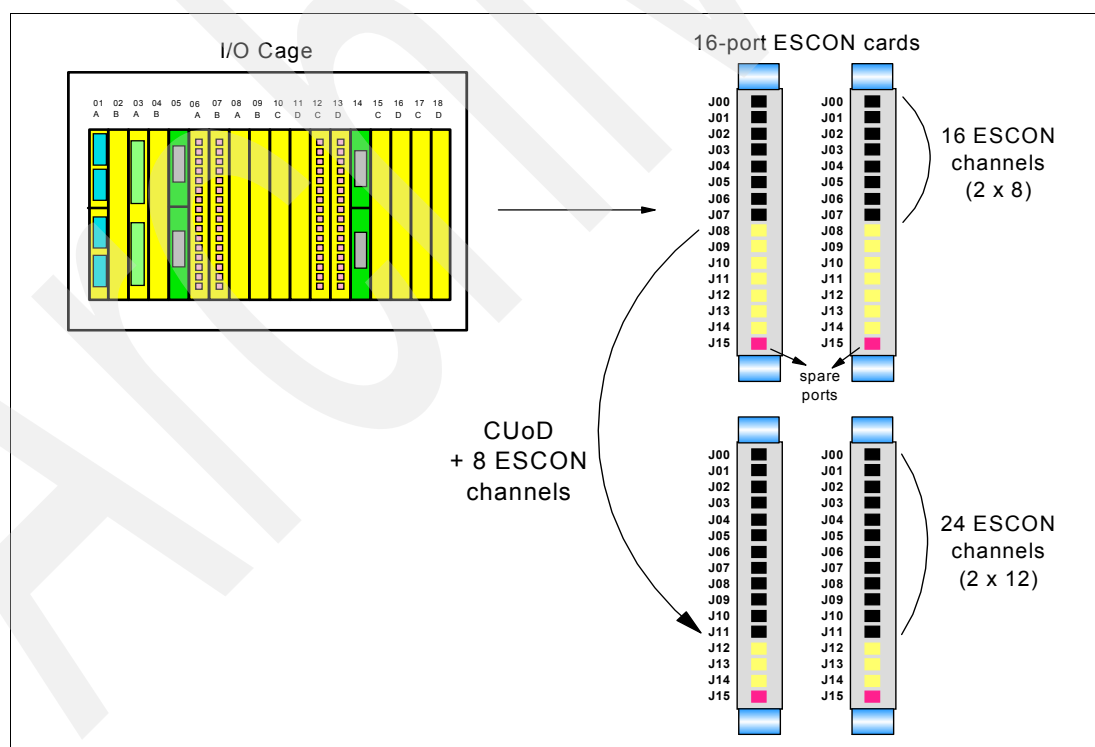


Figure 8-3 CUoD for I/O LIC-CC upgrade example

A z990 server 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, via LIC-CC, using four unused ports on each ESCON channel card.

The additional channels installed concurrently to the hardware can also be concurrently activated to an operating system using the Dynamic I/O configuration function. Dynamic I/O configuration can be used by the z/OS, OS/390, or z/VM operating systems. Linux and CFCC do *not* provide Dynamic I/O configuration support.

To better exploit the CUoD for I/O capability, an initial configuration should be carefully planned to allow concurrent upgrades up to a target configuration. Plan Ahead concurrent conditioning process can include, in the initial configuration, the shipment of additional I/O cages required for future I/O upgrades.

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 zSeries 990's order process configurator for concurrent I/O upgrades at some future time.

The Plan Ahead feature identifies the content of the target configuration, which cannot be concurrently installed, 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.

Accurate planning and definition of the target configuration is vital in maximizing the value of this feature.

8.3 Customer Initiated Upgrade (CIU)

Customer Initiated Upgrade (CIU) is the capability for the z990 *user* to initiate a *permanent* upgrade for CPs, ICFs, IFLs, zAAPs, and/or memory via the Web, using IBM Resource Link. CIU is similar to CUoD, but the capacity growth can be added by the customer. The customer also has the ability to unassign previously purchased CPs and IFLs processors via CIU.

CIU requires the CIU Enablement feature (FC 9898) installed.

The customer will then be able to download and apply the upgrade using functions on the HMC via the Remote Support Facility, without requiring the assistance of IBM service personnel. Once all the prerequisites are in place, the whole process from ordering to activation of the upgrade is performed by the customer. The actual upgrade process is fully automated and does not require any onsite presence of IBM service personnel.

CIU supports LIC-CC upgrades only and does not support I/O upgrades. All additional capacity required by a CIU upgrade must be previously installed. This means that additional books and/or I/O cards cannot be installed via CIU. The sum of CPs, unassigned CPs, IFLs, unassigned IFLs, ICFs, and zAAPs cannot exceed eight PUs per book. The number of zAAPs cannot exceed four zAAPs per book. The total number of zAAPs cannot exceed the number of CPs plus unassigned CPs on a z990 server.

Important: CIU for processors cannot be completed when CBU or On/Off CoD is *activated* on a z990 server. In this case, the CIU for processors can be ordered and retrieved by the customer, but *cannot* be applied until the temporary capacity upgrade via CBU or On/Off CoD is deactivated.

CIU may change the server's *software* model (3xx) if additional CPs are requested, but it cannot change the z990 *server* model (2084-xxx).

Additional logical processors can be concurrently configured online to logical partitions by the operating system when reserved processors are previously defined, resulting in image upgrades. The operating system must have the capability to concurrently configure more processors online.

Note: CIU for processors provides 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 217 for more information.

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 CIU upgrade. See Table 6-3 on page 148 to check software implications for CIU.

Software products using Workload License Charge (WLC) may not be affected by the server upgrade, as their charges are based on a partition’s utilization and not based on the server total capacity. See 6.8, “Workload License Charges” on page 150 for more information about WLC.

CIU registration and agreed contract for CIU

Before customers are able to use the CIU function, they have to be registered. Once they are registered, customers gain access to the CIU application by ordering the CIU Registration feature from their sales person.

This capability requires a CIU contract, which gives huge benefits to the customer because the upgrade can happen much faster than waiting for a normal MES to be processed. It allows the customer to be ready to accommodate new workload peaks in a very timely manner.

Ordering and activation of the upgrade is accomplished by the customer logging on to IBM Resource Link and executing the CIU application to upgrade a machine for CPs, ICFs, IFLs, zAAPs, and/or memory. It is possible to require a customer secondary order approval to conform to customer operation policies.

Figure 8-4 on page 198 illustrates the simplicity of the CIU ordering process on the IBM Resource Link.

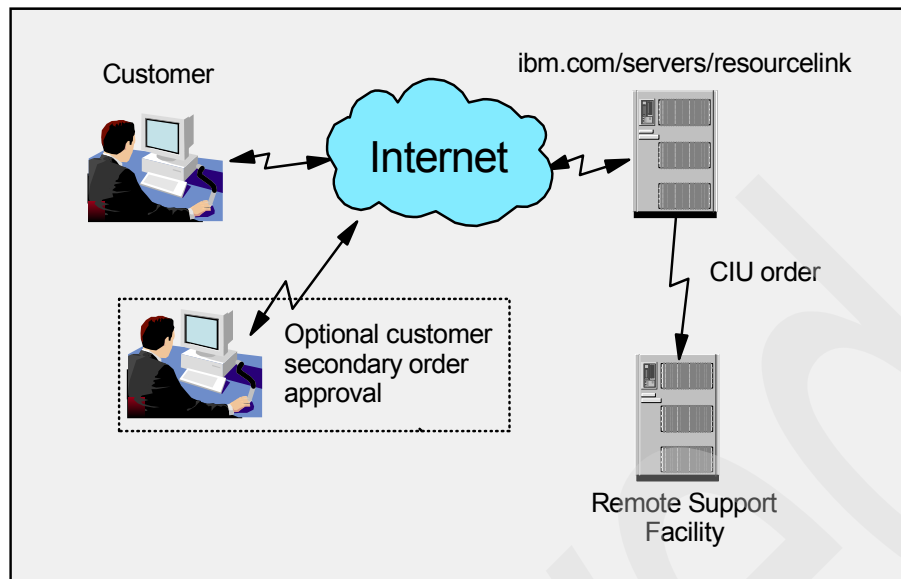


Figure 8-4 CIU ordering example

The following is a sample list of the screen sequences a customer must follow on Resource Link to initiate an order:

1. Sign on to Resource Link.
2. Select the CIU option from the main Resource Link page.
3. Customer and machine details associated with the User ID are listed.
4. The current configuration (PU allocation and memory) is shown for the selected server serial number.
5. Create a target configuration step-by-step for each upgradeable option. Resource Link limits options to those that are valid/possible for this z990 configuration.
6. The target configuration is verified.
7. The customer has the option to accept or reject.
8. An order is created and verified against the pre-established Agreement.
9. A price is quoted for the order; customer signals acceptance/rejection.
10. A customer secondary order approval is optional.
11. On confirmation of acceptance, the order is processed.
12. LIC-CC for the upgrade should be available within few hours.

Figure 8-5 on page 199 shows the CIU activation process. IBM Resource Link communicates with the Remote Support Facility to stage the CIU order and prepare it for download. The customer is automatically notified when the order is ready for download.

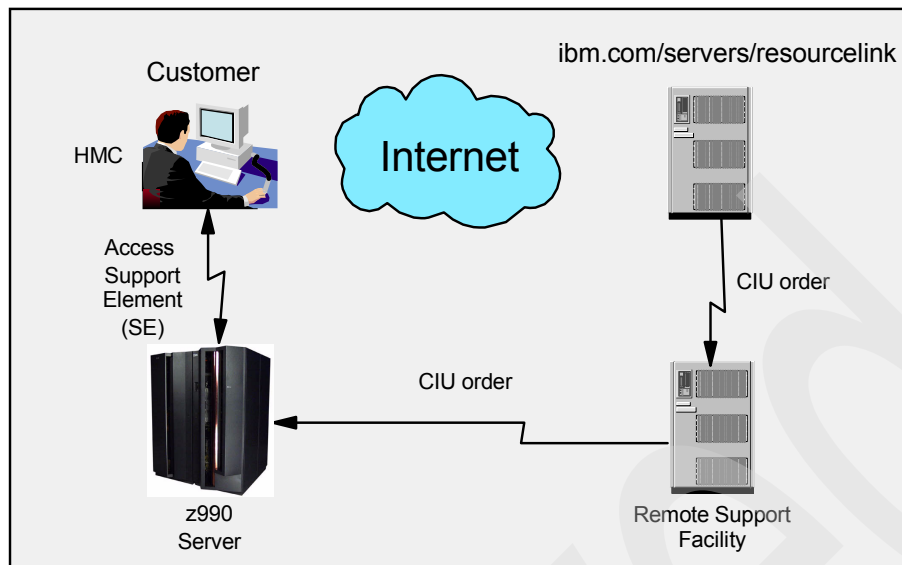


Figure 8-5 CIU activation example

Order and fulfillment process

By using the CIU process, associated systems allow the customer to order increased capacity for CPs, ICFs, IFLs, zAAPs, and/or memory. Resource Link is responsible for delivering the price or lease agreement to the customer. The interface handles the order differently based on whether the customer is leasing the server or not. The customer profile associated with the machine serial number will contain an indicator that Resource Link uses to make the determination. If the customer chooses to accept this agreement, then it will be forwarded to the correct billing system. Only Resource Link users who accept this feature will be able to access to this CIU application.

The two major components in the process are *Ordering* and *Activation*.

Ordering

Resource Link provides the interface that allows the customer to order a dynamic upgrade for a specific server. The customer is able to create, cancel, and view the order. The customer also is able to view the history of orders that were placed through this interface. Configuration rules will enforce only valid configurations being generated within the limits of the individual server. Warning messages will be issued when certain invalid upgrade options are selected.

Figure 8-6 on page 200 shows a Resource Link Web page that displays a CIU order example.

The screenshot shows the IBM Machine profile page. The top navigation bar includes the IBM logo, 'United States', a search bar, and links for Home, Products & services, Support & downloads, My account, and a phone number 1-888-SHOP-IBM. A sidebar on the left lists various resource links like Site search, Planning, Education, Library, Forums, Fixes, Problem solving, Services, Tools, Customer Initiated Upgrade, and Feedback. The main content area is titled 'Machine profile' and displays customer information (Customer user ID, name, number), machine details (Machine type, serial, model), and approval information (Approval ID, GEO, Country, CIU Express, Machine name, Sold as model, On/Off CoD). A table compares the 'Current Configuration' and 'Ordered Configuration' for various components: CPs, ICFs, zAAPs, IFLs, SAPs, Memory, and Additional CBU CPs. Below the table, there are three bullet points providing status information. At the bottom, an 'Order history information' table shows the order number, type, status, and date updated.

	Current Configuration	Ordered Configuration
CP:	3 CPs	4 CPs
ICF:	0	0
zAAP:	2	2
IFL:	0	1
SAP:	4	4
Memory:	16	16
Additional CBU CPs:	N/A	N/A
Unassigned CPs:	4	3
Unassigned IFLs:	1	0

Order number	Order type	Order status	Date status updated
LC5XSQWZ	Permanent	Needs customer approval	04/06/2004 03:33:10 PM

Figure 8-6 CIU order example

The number of CPs, ICFs, zAAPs, IFLs, SAPs, memory size, CBU features, unassigned CPs, and unassigned IFLs (Linux) on the current configuration are displayed on the left side. On the right side are the corresponding updated values of the ordered configuration. This CIU order is requesting a 2084-B16 server upgrade from three CPs (software model 303) to four CPs (software model 304) plus one IFL, by assigning previously unassigned (but owned) CPs and IFLs (note that the number of unassigned CPs and IFLs decreases).

On *processor upgrades*, Resource Link will offer the customer the ability to upgrade only to those configurations that are deemed valid by the Order Process, within the already installed number of books.

Note that the CBU features count does not change. It will only be adjusted if the customer creates an order that requires the CBU feature count be decremented because there are no other spare PUs left.

On *memory upgrades*, Resource Link will retrieve and store relevant data associated with the installed memory cards for the specific server. It will allow you to select only those upgrade options that are deemed valid by the Order Process. Resource Link will only allow the customer to upgrade memory within the given bounds of the currently installed hardware. It will not allow for the ordering of memory or books not attainable within the current configuration.

Activation

The customer's system stores all the LIC-CC records associated with the z990 that has the CIU option. These LIC-CC records will only be available for download when they are externally activated by Resource Link. Upon submission of the order, Resource Link will dynamically enable the appropriate LIC-CC records and make them available to the customer, via the Remote Support Facility, to be downloaded by the Hardware Management Console. When the order is available for download, the customer will be given an activation number. Once Resource Link has notified the customer that the upgrade is ready for download, the customer can go to any of the Hardware Management Consoles attached to the system and do a Single Object Operations to the Support Element where the upgrade is to be applied by selecting the Perform Model Conversion task. Using the Model Conversion screen, select the CIU Options (see Figure 8-7) to start the process.

Model Conversion

This function is used to add, remove or update system hardware and features. The system model identification may change if a Book related selection warrants it. Select the appropriate option.

Hardware or LICCC upgrades

☐ Book Diskette Upgrade

☐ Channel Diskette Upgrade

☒ Customer Initiated Upgrade

Temporary upgrades

☐ Test Capacity Backup feature using Password Panel

☐ Test Capacity Backup feature using IBM Service Support System

☐ Temporary Capacity Backup feature Upgrade using Password Panel

☐ Temporary Capacity Backup feature Upgrade using IBM Service Support System

☐ Undo Temporary Upgrade

Features

☐ Add Capacity Backup feature

Figure 8-7 z990 Model Conversion screen

A new Configuration panel, shown in Figure 8-8 on page 202, offers the option Retrieve and Apply CIU. It prompts the customer to enter the order activation number to begin the code download process. Once downloaded, the system will check if the upgrades can be retrieved and applied.

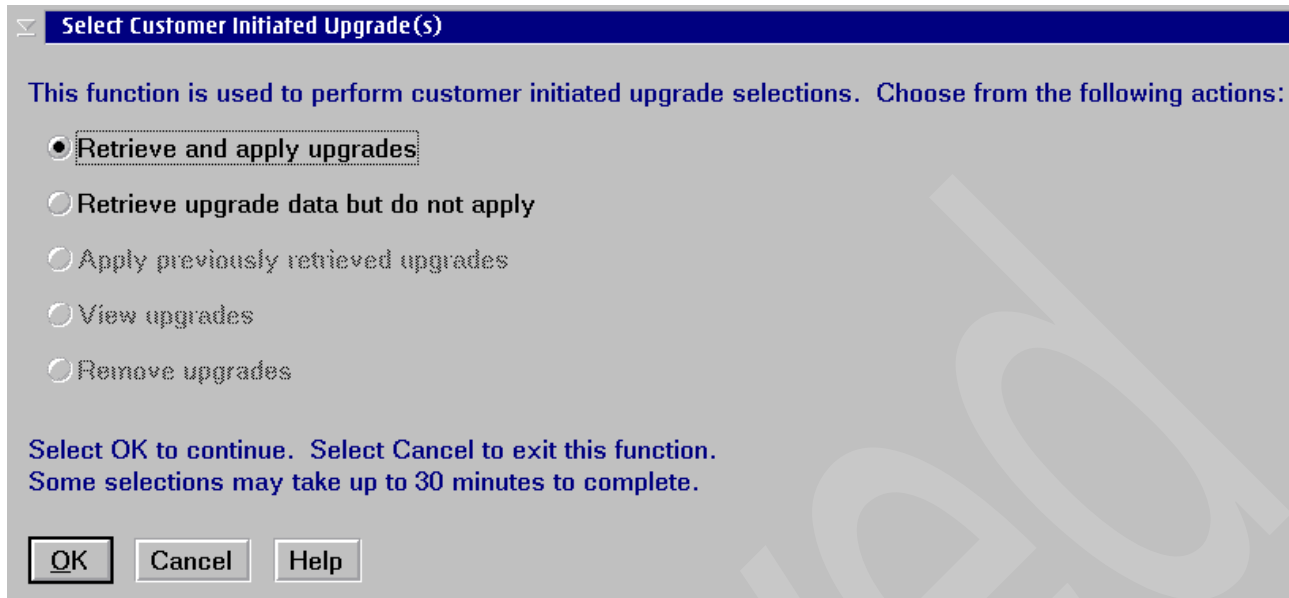


Figure 8-8 CIU upgrade selection screen

An On/Off CoD upgrade for processors cannot be applied while a previous On/Off CoD or a CBU is activated. In those cases, the requested upgrade can be retrieved, but can be applied only after the current temporary upgrade is deactivated. On/Off CoD upgrades for memory can be retrieved and applied, even while a CBU or a previous On/Off CoD activation is in place.

8.4 On/Off Capacity on Demand (On/Off CoD)

The On/Off Capacity on Demand (On/Off CoD) for z990 servers is the ability for the z990 *user* to *temporarily* turn on unowned PUs, unassigned CPs, and unassigned IFLs available within the current model, to help meet customer's peak workload requirements. On/Off CoD uses the Customer Initiated Upgrade (CIU) process to request the upgrade via the Web, using IBM Resource Link.

On/Off CoD requires the CIU Enablement feature (FC 9898) and the On/Off CoD Enablement feature (FC 9896) installed.

Important: The On/Off CoD capability can coexist with Capacity BackUp (CBU) enablement. Both On/Off CoD and CBU LIC-CC can be installed on a z990 server, but the On/Off CoD activation and CBU activation are mutually exclusive.

The resources eligible for temporary use are CPs, ICFs, IFLs, and/or zAAPs. Temporary use of memory and I/O ports is not supported. Spare PUs that are currently unassigned and unowned can be temporarily and concurrently activated as CPs, ICFs, IFLs, or zAAPs via LIC-CC, up to the double of the current installed capacity, and up to the limits of the physical server size. This means that an On/Off CoD upgrade cannot change the z990 *server* model (2084-xxx), as additional book installation is not supported. However, On/Off CoD may change the server's *software* model (3xx) if additional CPs are requested.

The On/Off CoD upgrade features are:

- On/Off CoD Active CP (FC 9897)
- On/Off CoD Active IFL (FC 9888)
- On/Off CoD Active ICF (FC 9889)
- On/Off CoD Active zAAP (FC 9893)

You may concurrently install temporary capacity by ordering On/Off CoD Active CP features up to the number of current CPs, On/Off CoD Active IFL features up to the number of current IFLs, On/Off CoD Active ICF features up to the number of current ICFs, and On/Off CoD Active zAAP features up to the number of current zAAPs that are permanently purchased. In other words, the upgrade configuration capacity is limited to the double of the current installed capacity, for each individual processor type (CPs, IFLs, ICFs, and zAAPs).

Also, the total number of On/Off CoD Active zAAPs plus zAAPs cannot exceed the number of On/Off Active CPs plus the number of CPs plus the number unassigned CPs on a z990 server. The sum of CPs, unassigned CPs, IFLs, unassigned IFLs, ICFs, and zAAPs cannot exceed eight PUs per book. The number of zAAPs cannot exceed four zAAPs per book.

The CIU process will continue to bill for the upgrade on a daily basis until it detects the On/Off CoD has been deactivated. When the temporary capacity is no longer required, its removal is nondisruptive. If On/Off CoD is activated on a z990 server, other hardware upgrades/MES are restricted. With the exception of memory and channels, LIC-CC enabled features, such as CPs, ICFs, IFLs, and zAAPs, can be ordered but not enabled until the On/Off CoD upgrade is deactivated.

The customer's user will then be able to download and apply the upgrade using functions on the HMC via the Remote Support Facility, without requiring the assistance of IBM service personnel. Once all the prerequisites are in place, the whole process from ordering to activation of the upgrade is performed by the customer. The actual upgrade process is fully automated and does not require any on-site presence of IBM service personnel.

Additional logical processors can be concurrently configured online to logical partitions by the operating system when reserved processors are previously defined, resulting in image upgrades. The operating system must have the capability to concurrently configure more processors online.

Note: On/Off CoD provides 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 217.

To participate in this offering, customers must have accepted contractual terms for On/Off CoD (in addition to Customer Initiated Upgrade (CIU)), established a CIU profile, and installed an On/Off CoD “right to use” feature on the machine. Subsequently, the customer may concurrently install temporary capacity in any amount and use it for an indeterminate time. The customer will be billed on a monthly basis. If the customer installs more than one On/Off CoD order within a billing month, they will be billed for the greater of all orders installed within that month. Monitoring will occur through the server call home facility and an invoice will be generated if the capacity has been enabled for any portion of a calendar month.

The customer will continue to be billed for use of temporary capacity until they return the server to the original state. After concurrently returning to the original state, the customer may choose to activate a new temporary session that can be different from the previous session.

When the customer disposes of the server, or decides that they want to disable future On/Off CoD, the customer is required to apply a termination feature that disables the right to use.

Initiation

Before a customer can order temporary capacity, they must have a signed agreement for the Customer Initiated Upgrade (CIU) facility. In addition to this agreement, they will need to acknowledge and agree to additional specific terms that govern the use of temporary capacity. At the completion of signing a contract, an order is placed through CIU to install an On/Off CoD right to use feature. This feature cannot be installed if CBU is already active on the z990 server. Once installed, the customer is free to order and activate temporary capacity.

Ordering

Typically, On/Off CoD can only be ordered through CIU; however, there will be an RPQ available for customers who do not have an Remote Support Facility (RSF) connection.

If an order is attempted prior to establishing the right to use, or if the configuration has no permanent CPs, that order will not be processed. Similar to a permanent upgrade, the customer will order one to n CPs, ICFs, IFLs, and/or zAAPs worth of temporary capacity. A LIC record is established and staged to RETAIN® for this order, where it remains available for 30 days. This record, once activated, has no expiration date; however, an individual record can only be activated once. Subsequent sessions will require a new order to be generated producing a new LIC record for that specific order.

Figure 8-9 on page 205 shows a Resource Link Web page that displays an On/Off CoD order example.

IBM Resource Link: Create machine upgrade order - Microsoft Internet Explorer

IBM

Create On/Off Capacity on Demand order

Customer number: 5555556 Machine type: 2084
Order number: LC5XTH44 Machine serial: WLK01

	Current configuration	Upgrade configuration	Upgrade price
CP:	4 CPs	5 CPs	\$.....
ICF:	2	2	\$0.00
zAAP:	0	2	Not Negotiated
IFL:	2	2	\$0.00
		2	
		3	
		4	
		5	
		6	
		7	
Total daily price:			Not Negotiated

Submit Cancel

Figure 8-9 On/Off CoD order example

This On/Off CoD example is ordering an upgrade from four CPs to five CPs plus two zAAPs to the current server. The maximum number of CPs, ICFs, zAAPs, and IFLs is limited by the current number of available spare PUs of the installed books on the z990 server. The upgrade configuration capacity is limited to the double of the current installed capacity, for each individual processor type (CPs, IFLs, ICFs, and zAAPs). The total number of zAAPs cannot exceed the total number of CPs plus unassigned CPs on a z990 server.

Activation/Deactivation

When a previously ordered LIC record is retrieved from RETAIN, it is downloaded and immediately activated. The customer does not have the ability to stage the record at their site. When the customer has finished using temporary capacity, they must take action to deactivate the session. This deactivation uses the same facility as CBU undo and is non-disruptive to the customer operation. Depending on the use of the extra capacity, customers may be required to perform tasks at the logical partition level in order to remove the temporary CPs. An example of this would be removal of temporary CPs allocated to a logical partition.

Termination

A customer will be contractually required to terminate the On/Off CoD right to use feature whenever there is a transfer in asset ownership. A customer may also choose to terminate the

On/Off CoD right to use feature without transferring ownership. This might be desirable if the customer wants to put CBU on the z990 server, or simply wants to put business controls in place that prevent On/Off CoD from being used in the future. Application of feature code 9898 will terminate the right to use On/Off CoD. This feature cannot be ordered if a temporary session is already active. Similarly, CIU cannot be removed if a temporary session is active. Anytime CIU is removed, the On/Off CoD right to use will be simultaneously removed. Reactivating of the right to use feature will subject the customer to whatever terms and fees apply at that time.

Upgrade Capability during On/Off CoD

No upgrades involving physical hardware will be supported while an On/Off CoD upgrade is active on a particular z990 server. However, LIC-only upgrades can be ordered and retrieved from RETAIN but not applied while an On/Off CoD upgrade is active. LIC-only memory upgrades can be retrieved and applied while an On/Off CoD upgrade is active.

Repair capability during On/Off CoD

If the z990 server requires service while an On/Off CoD upgrade is active, the repair and verify code (R&V) will automatically deactivate the On/Off CoD upgrade. At the end of the repair, R&V will retrieve a new LIC record from RETAIN to replace the record that was deactivated. The On/Off CoD upgrade will be activated to the state prior to R&V, including restoration of the original activation date.

Monitoring

When the customer activates an On/Off CoD upgrade, an indicator is set in VPD 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, CIU will send a note to generate an invoice for On/Off CoD used during that month.

Software

Software PSLC customers will be billed at the MSU level represented by the combined permanent and temporary capacity. All PSLC products will be billed at the peak MSUs enabled during the month, regardless of usage. Customers with WLC licenses will be billed by product at the highest four hour rolling average for the month. In this instance, temporary capacity will not necessarily increase the customer's software bill until that capacity is allocated to logical partitions and actually consumed.

Results from the STSI instruction will reflect the current permanent plus temporary CPs. See Table 6-3 on page 148 for more details.

8.5 Capacity BackUp (CBU)

Capacity BackUp (CBU) is offered with the z990 servers to provide reserved emergency backup processor capacity for unplanned situations where customers have lost capacity in another part of their establishment and want to recover by adding the reserved capacity on a designated z990 server.

CBU is the quick, *temporary* activation of Central Processors (CPs), up to 90 days, in the face of a loss of customer processing capacity due to an emergency or disaster/recovery situation.

Note: CBU is for disaster/recovery purposes only and *cannot* be used for peak load management of customer workload.

Important: The CBU capability can coexist with On/Off CoD enablement. Both CBU and On/Off CoD LIC-CC can be installed on a z990 server, but the CBU activation and On/Off CoD activation are mutually exclusive.

CBU can only add CPs to an existing z990 server, but note that CPs can assume any kind of workload that could be running on IFLs, zAAPs, and ICF processors at the failed system or systems. z/VM, Linux, Java code, and CFCC (for Coupling Facility partitions) can also run on CPs.

When the CBU activated capacity is no longer required, its removal is nondisruptive. If CBU is activated on a z990 server, other hardware upgrades/MES are restricted. With exception of memory and channels, LIC-CC enabled features, such as CPs, ICFs, IFLs, and zAAPs, can be ordered but not enabled, until the CBU upgrade is deactivated.

The CPs that can be activated by CBU come from the available spare PUs on any installed book of the designated z990 server. So the number of CBU features (FC 7800), one for each “stand-by” CP, that can be ordered is limited by the number of spare PUs on the server. Some examples:

- ▶ A 2084-B16 server with eight CPs (no IFLs, ICFs, or zAAPs) has eight available spare PUs available; this server can have up to eight CBU features.
- ▶ A 2084-B16 server with eight CPs and two zAAPs (no IFLs or ICFs) has six available spare PUs available; this server can have up to six CBU features.
- ▶ A 2084-C24 server with 12 CPs, two IFLs, and one ICFs has nine available spare PUs available; this server can have up to nine CBU features.

Note that CBU can add CPs via LIC-CC only and the z990 server must have the proper number of books installed to allow the required upgrade by LIC-CC. CBU changes the server's *software* model (3xx) but cannot change the *z990 server* model (2084-xxx).

A CBU contract must be in place before the special code that enables this capability can be loaded on the customer's server. CBU features can be added to an existing z990 server nondisruptively.

The installation of the CBU code provides an alternate configuration that can be activated in the face 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/recovery, are usually allowed in each CBU contract.

A CBU system normally operates with a “base” PU configuration having a pre-configured number of additional spare PUs reserved for activation as CPs in case of an emergency. One CBU feature is required for each “stand-by” CP that can be activated. A CBU activation enables the *total* number of CBU features installed.

The base CBU configuration must have sufficient memory and channels to accommodate the potential needs of the large CBU target server. When capacity is needed in an emergency, the customer can activate the emergency CBU configuration with the reserved spare PUs added into the configuration as CPs. It is very important to ensure that all required functions are available on the “backup” server(s), including CFLEVELs for Coupling Facility partitions, as well as cryptographic and connectivity capabilities.

This upgraded configuration is activated *temporarily* and provides additional CPs above and beyond the server's original, *permanent* configuration. The number of additional CPs is predetermined by the alternate configuration, which has been stated in the CBU contract.

When the emergency is over (or the CBU test is complete), the server must be taken back to its original, permanent configuration. The CBU features can be deactivated by the customer at any time before the expiration date. Otherwise, the performance of the system will be degraded after expiration, unless CBU is deactivated.

Note: CBU for processors provides 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 217.

Software charges based on the total capacity of the server on which the software is installed would be adjusted to the maximum capacity after the CBU upgrade. See Table 6-3 on page 148 to check software implications of CUoD, which is used by the CBU upgrade.

Software products using Workload License Charge (WLC) may not be affected by the server upgrade, as their charges are based on partition’s utilization and not based on the server total capacity. See 6.8, “Workload License Charges” on page 150 for more information about WLC.

For detailed instructions, refer to the *IBM @server zSeries Capacity Backup User’s Guide*, SC28-6810, available on the IBM Resource Link.

Activation/Deactivation of CBU

The activation and deactivation of the CBU function can be initiated by the customer without the need for the onsite presence of IBM service personnel. The CBU function is activated and deactivated from the HMC, and in each case it is a nondisruptive task.

CBU activation

Upon request from the customer, IBM can remotely activate the emergency configuration, eliminating the time associated with waiting for an IBM service person to arrive on site to perform the activation.

A fast electronic activation is available through the Hardware Management Console (HMC) and Remote Support Facility (RSF) and could drive activation time down to minutes. The z990 server invokes the RSF to trigger an automatic verification of CBU authentication at IBM. This will initiate an automatic sending of the authentication to the customer’s server, automatic unlocking of the reserved capacity, and activation of the target configuration.

In situations where the RSF cannot be used, CBU can be activated through a password panel. In this case, a request by telephone to the IBM support center usually enables activation within a few hours.

The CBU activation cannot be done when an On/Off CoD upgrade is already activated.

Image upgrades

After the CBU activation, the z990 server has more physical CPs available to the operating system image(s). The logical partition image(s) can concurrently increase the number of logical CPs by configuring reserved processors online. The operating system must have the capability to concurrently configure more processors online. If a nondisruptive CBU upgrade is needed, the same principles of nondisruptive CUoD should be applied.

CBU deactivation

The process of deactivating CBU is simple and straightforward. The process starts by quiescing the added CPs (normally the highest numbered) from all the logical partitions, and varying them offline from the operating systems. Then from the HMC CBU activation panel, perform a concurrent CBU undo.

CBU testing

Testing of disaster/recovery plans is easy with CBU. Testing can be accomplished by ordering a diskette, calling the support center, or using the fast activation icon on the HMC.

Capacity BackUp operation example

Figure 8-10 shows an example of a 2084-B16 software model 304 to a 2084-B16 software model 312 Capacity BackUp operation.

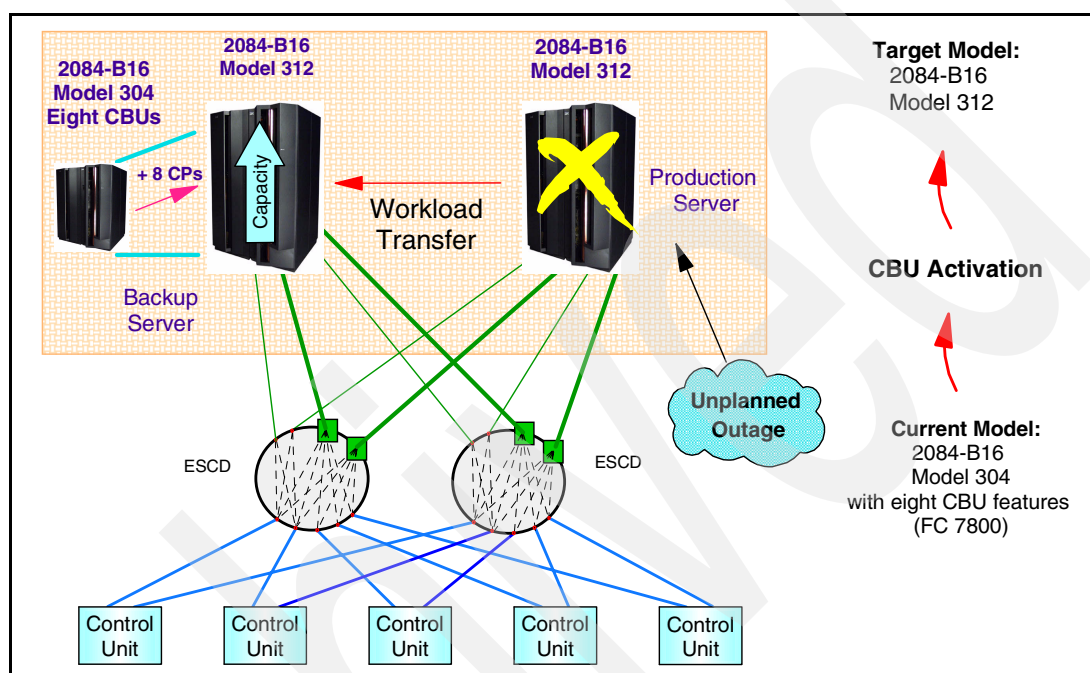


Figure 8-10 Capacity BackUp operation example

The PUs associated with Capacity BackUp are reserved for future use with CBU features (FC 7800) installed on the backup server. In this example, there should be eight CBU features installed on the backup server 2084-B16 software model 304. When the production server 2084-B16 software model 312 has an unplanned outage, the backup server can be temporarily upgraded to the target model planned, 2084-B16 software model 312, to get the capacity to take over the workload on the failed production server.

Furthermore, customers can configure systems to back each other up. For example, if a customer uses two 2084-A08 software model 303s for the production environment, both can have three CBU features installed (or even more). If one server has a disaster, the other one can be upgraded up to the approximately total original CP capacity.

Automatic CBU enablement for GDPS

The intent of the GDPS CBU is to enable automatic management of the reserved PUs provided by the CBU feature in the event of a server failure and/or a 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 will:

- ▶ Perform the analysis required to determine the scope of the failure; this minimizes operator intervention and the potential for errors.
- ▶ Automate authentication and activation of the reserved CPs.
- ▶ Automatically restart the critical applications after reserved CP activation.

- Reduce the outage time to restart critical workloads from several hours to minutes.

8.6 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 and OS/390 environments, nondisruptive upgrades within a single server can avoid system outages and are suitable to further operating system environments.

The z990 servers allow *concurrent* upgrades, meaning they can dynamically add more capacity to the server. If operating system images running on the upgraded server need no disruptive tasks to use the new capacity, the upgrade is also *nondisruptive*. This means that Power-on Resets (PORs), logical partition deactivations, and IPLs 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 and OS/390 operating systems have this capability. z/VM can concurrently configure new processors and I/O devices online, but it does not support dynamic storage reconfiguration.

Linux operating systems do *not* have the capability of adding more resources concurrently. However, Linux virtual machines running under z/VM can take advantage of the z/VM capability to nondisruptively configure more resources online (processors and I/O).

Important: Dynamic add/delete of a logical partition name allows reserved partition ‘slots’ to be created in an IOCDS in the form of extra Logical Channel Subsystem, Multiple Image Facility (MIF) image pairs, which can be later assigned a logical partition name for use (or later removed) via HCD concurrently.

Important: If the z990 STI Rebalance feature (FC 2400) is selected at server upgrade configuration time, and effectively results in STI rebalancing, the server upgrade will be disruptive and this outage must be planned. The z990 STI Rebalance feature may also change the Physical Channel ID (PCHID) number of ICB-4 links, requiring a corresponding update on the server’s I/O definition via HCD/HCM.

Processors

CPs, IFLs, ICFs, and/or zAAPs processors can be concurrently added to a z990 server if there are spare PUs available on any installed book. The number of zAAPs cannot exceed the number of CPs plus unassigned CPs on a z990 server.

Additional book(s) can also be installed concurrently, allowing further processor upgrades.

A processor upgrade cannot be performed when CBU or On/Off CoD is activated.

Concurrent upgrades are not supported with CPs defined as additional SAPs.

If reserved processors are defined to a logical partition, then z/OS, OS/390, and z/VM operating system images can dynamically configure more processors online, allowing nondisruptive processor upgrades. The Coupling Facility Control Code (CFCC) can also configure more processors online to Coupling Facility logical partitions using the CFCC image operations panel.

Memory

Memory can be concurrently added to a z990 server up to the physical installed memory limit. Additional book(s) can also be installed concurrently, allowing further memory upgrades by LIC-CC enabling memory capacity on the new book(s).

Using the previously defined reserved memory, z/OS and OS/390 operating system images can dynamically configure more memory online, allowing nondisruptive memory upgrades.

I/O

I/O cards can be added concurrently to a z990 server if all the required infrastructure (I/O slots and STIs) 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.

Also I/O ports can be concurrently added by LIC-CC, enabling available ports on ESCON and ISC-3 Daughter Cards.

Dynamic I/O configurations are supported by some operating systems (z/OS, OS/390, and z/VM), allowing nondisruptive I/O upgrades. However, it is not possible to have dynamic I/O reconfigurations on a stand-alone Coupling Facility server, because there is no operating system with this capability running on this server. Dynamic I/O configurations require additional space in the HSA for expansion.

PCI Cryptographic coprocessors

PCI cryptographic (PCIXCC and PCICA) cards can be added concurrently to a z990 server if all the required infrastructure, I/O slots and STIs, is present on the configuration. The Plan Ahead process can assure that an initial configuration will have all the infrastructure required by the target configuration.

In order to make the addition of PCIXCC and/or PCICA cards nondisruptive, logical partitions must be predefined with the appropriate PCI cryptographic processor number selected in its candidate list on the partition image profile. To maximize concurrent upgrade possibilities in this area, it is recommended that all logical partitions define all possible PCI cryptographic coprocessors as candidates for the logical partition. This is possible even if there are no PCI cryptographic coprocessors currently installed on the machine.

8.6.1 Upgrade scenarios

The following scenarios are examples of nondisruptive upgrades, showing the hardware (z990 server) upgrades and the image (logical partitions) upgrades. Only the images previously configured with Reserved Processors and/or Reserved Memory can be nondisruptively upgraded. Spare PUs are used for hardware upgrades and “spare logical processors” (Reserved Processors) are used for image upgrades.

Tip: Configure as many as possible reserved CPs, IFLs, ICFs, zAAPs, and memory to a logical partition to allow concurrent image upgrades.

All scenarios show the hardware (physical) and the logical partition configurations *before* and *after* the upgrade. Only the eight PUs available for CPs, IFLs, ICFs, and zAAPs are shown on each server's book.

Shared logical partitions upgrade

Figure 8-11 shows a 2084-A08 software model 307 server. This one-book server configuration has seven CPs and one spare PU.

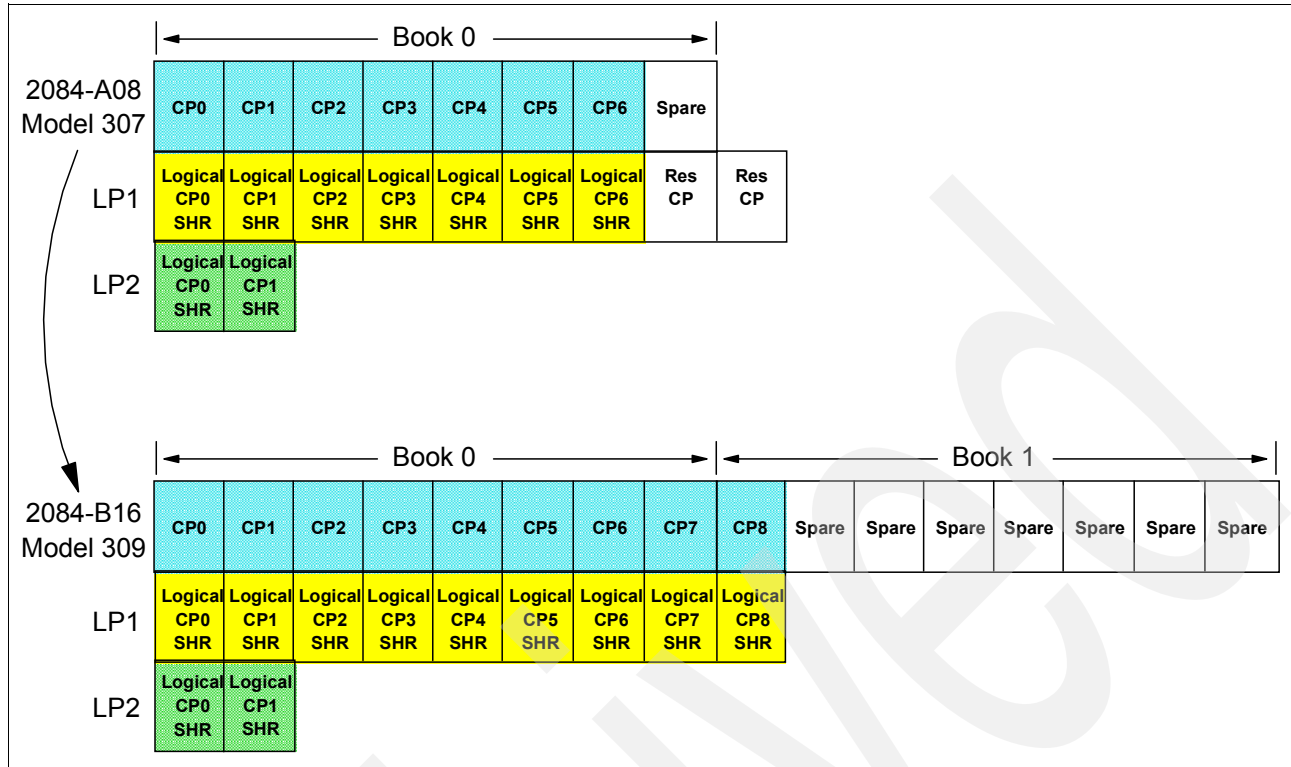


Figure 8-11 Shared logical partitions upgrade example

There are two activated logical partitions: LP1, having seven shared (SHR) logical CPs and two reserved CPs defined, and LP2, having only two shared (SHR) logical CPs defined. Note that the number of reserved CPs for a logical partition can be higher than the number of installed PUs. This allows nondisruptive image upgrades even when new books are installed.

This example shows an upgrade to the 2084-B16 software model 309, achieved by adding concurrently, one more book and LIC-CC, enabling two more CPs to the physical configuration. Seven spare PUs are available for future processor upgrades. The physical upgrade ends here.

At this point, before any other configuration changes are made, the images with shared logical CPs running on this server can experience performance improvements. As there is now more available capacity (physical processors) to be used by all logical shared CPs, the “logical-to-physical processors ratio” is reduced. In this example, before the upgrade, there were nine shared logical CPs to be dispatched into seven physical CPs. If all nine logical CPs have tasks to run, two of them have to wait. After the physical upgrade, up to nine logical CPs can run at the same time.

Now let us see the logical upgrades. Since there is no activated partition with dedicated CP, any partition can have up to nine activated logical CPs.

Partition LP1 has two reserved CPs defined and, if the operating system running on it has the capability of configuring processors online, this partition can be nondisruptively upgraded to nine CPs, as shown in this example.

Partition LP2 has no reserved CPs, so it cannot be nondisruptively upgraded. If any upgrade for this partition is required, it will require deactivation to configure more CPs.

Dedicated and shared logical partitions upgrade

Figure 8-12 shows a 2084-B16 software model 309 server. This two-book server configuration has nine CPs and seven spare PUs.

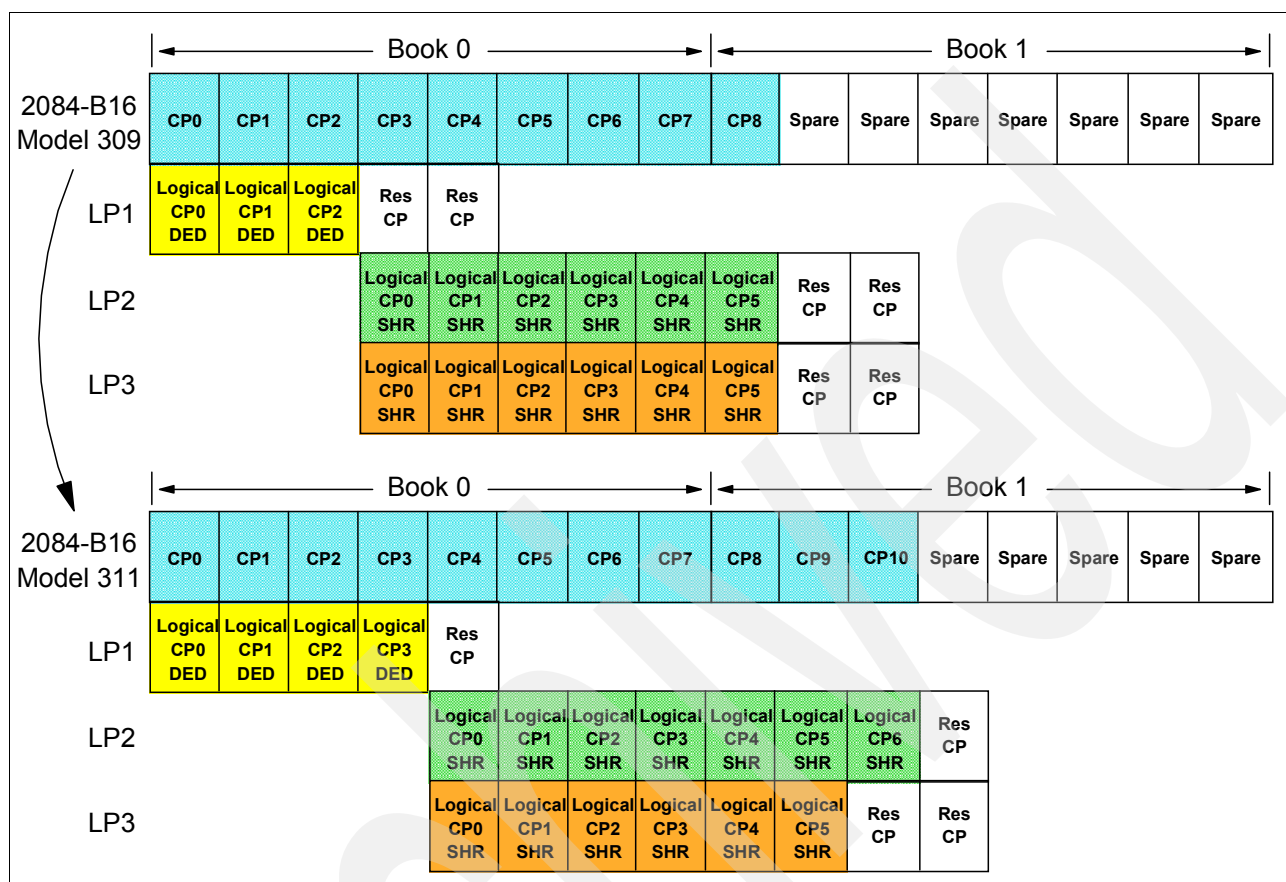


Figure 8-12 Dedicated and shared logical partitions upgrade example

This 2084-B16 software model 309 (with no ICFs or IFLs) can be concurrently upgraded to the software model 316 via LIC-CC, with no model upgrade.

There are three activated logical partitions: LP1 has three dedicated (DED) logical CPs and two reserved CPs defined, LP2 has six shared (SHR) logical CPs and two reserved CPs defined, and LP3 also has six shared (SHR) logical CPs and two reserved CPS defined.

This example shows an upgrade to the 2084-B16 software model 311, by adding, concurrently, two more CPs to the physical configuration. Two available spare PUs are used by LIC-CC for this upgrade. The physical upgrade ends here.

At this point, even with no partition configuration changes, LP2 and LP3 (shared) partitions may experience performance improvements. As there is now more available capacity (physical processors) to be used by all logical shared CPs, the "logical-to-physical processors ratio" is reduced. In this example, before the upgrade, there were twelve shared logical CPs (six for LP2 and six for LP3) to be dispatched into six physical CPs. If all twelve logical CPs have tasks to run, six of them have to wait. After the physical upgrade, up to seven shared logical CPs can run at the same time.

Now let us consider the logical upgrades, assuming that all operating systems running in these partitions have the capability of configuring processors online. Partition LP1, which has two reserved CPs defined, can configure up to two more CPs online. In this example, LP1 is

configuring one more CP online, with one reserved CP for a future image upgrade. LP1 now has four dedicated CPs, enabled by doing a nondisruptive upgrade.

From the eleven physical CPs available, seven CPs are left to be shared by LP2 and LP3 shared logical CPs.

- ▶ LP2 has six logical CPs and two reserved CPs defined, but only one more can be configured online, as the current configuration has only seven CPs to be shared. After the LP2 logical upgrade, one reserved CP remains offline.
- ▶ LP3 has the same configuration as LP2, and in this example, it is not being changed. However, it could have one more reserved CP configured online.
- ▶ LP1 and LP2 remain with one reserved CP each, but they cannot be configured online in the current configuration. However, if LP1 configures one dedicated CP *offline*, then LP2 can activate its last reserved CP.

Shared partitions and zAAP upgrade

Figure 8-13 is an example of a nondisruptive upgrade, adding three zSeries Application Assist Processors (zAAPs) to a 2084-B16 software model 309 server.

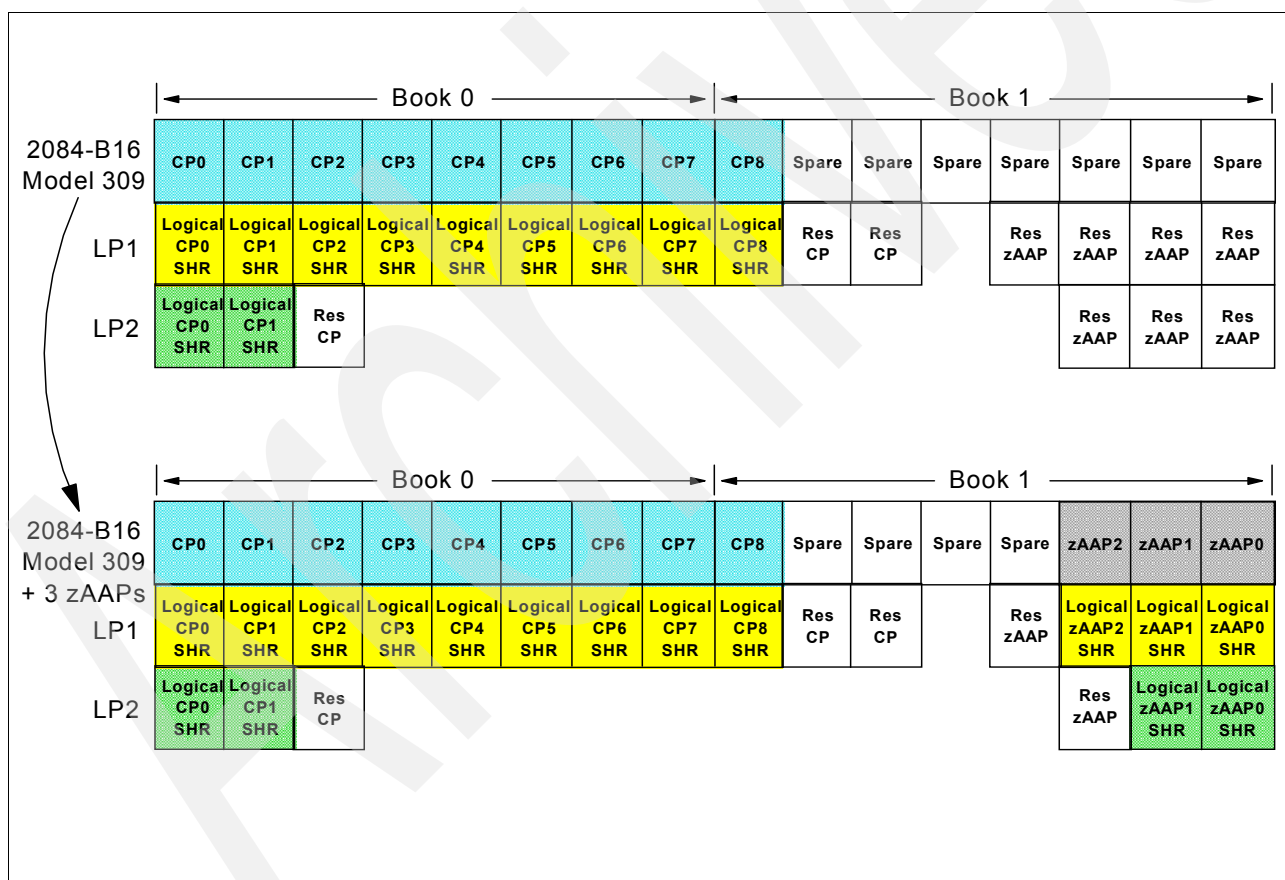


Figure 8-13 Shared logical partitions and zAAP upgrade example

This two-book server configuration has nine CPs and seven spare PUs. The concurrent hardware upgrade adds three zAAP processors, using three available spare PUs on the installed books. There are two activated logical partitions, both running z/OS V1.6 or higher:

- ▶ LP1 has eight shared (SHR) logical CPs, two reserved CPs, and four reserved zAAPs defined.
- ▶ LP2 has two shared (SHR) logical CPs, one reserved CP, and three reserved zAAPs defined.

The z/OS logical partition LP1, which has four reserved zAAPs defined, can configure online all of the three zAAP processors added, with one reserved zAAPs and two reserved CPs remaining for future image upgrades.

The z/OS logical partition LP2, which has three reserved zAAPs defined, is configuring two zAAPs online. After this logical partition upgrade, LP2 has one reserved CP and one reserved zAAP remaining for future upgrades.

Dedicated, shared partitions, and IFL upgrade

Figure 8-14 is an example of a nondisruptive upgrade, adding two Integrated Facility for Linux (IFL) processors to a 2084-B16 software model 309 server.

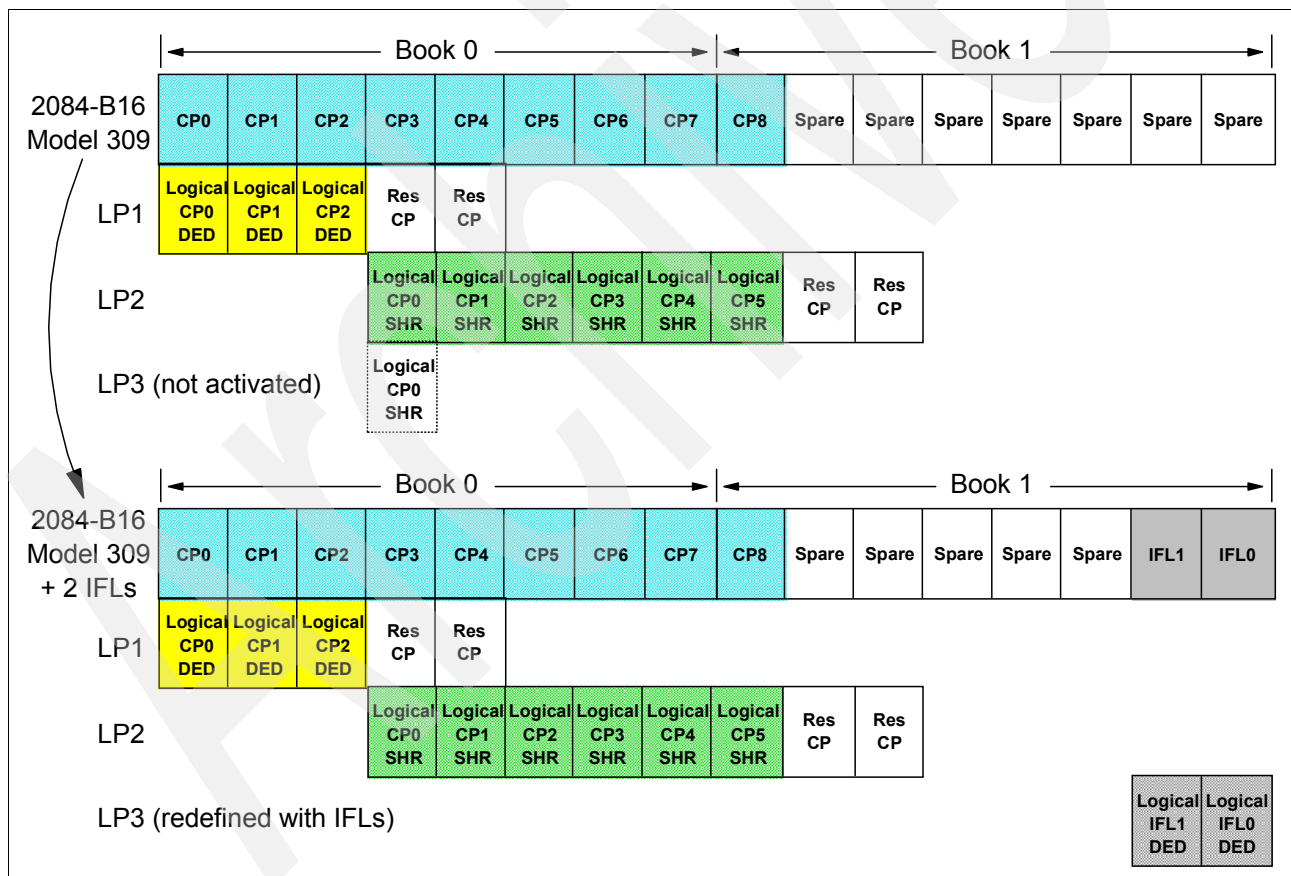


Figure 8-14 Dedicated, shared logical partitions, and IFL upgrade example

This two-book server configuration has nine CPs and seven spare PUs. The concurrent hardware upgrade adds two IFL processors, using two available spare PUs on the installed books. There are two activated logical partitions, and one non-activated:

- ▶ LP1 has three dedicated (DED) logical CPs and two reserved CPs defined.

- After the upgrade, including two IFL processors, the partition LP3 is redefined to have two dedicated logical IFLs. The I/O definitions for this logical partition can be dynamically done by LP1 or LP2, if any of them has the Dynamic I/O Configuration OS support. Then partition LP3 can be activated and use the IFL processors. By using a previously defined logical partition, this upgrade is non-disruptive.

Figure 8-15 shows a 2084-B16 software model 309 with one Internal Coupling Facility (ICF). This example is similar to the previous one, but now includes a Coupling Facility (CF) partition.



There are three activated logical partitions: LP1 has three dedicated (DED) logical CPs and two reserved CPs defined, LP2 has six shared (SHR) logical CPs and two reserved CPs defined, and LP3 is a CF partition with one dedicated (DED) ICF and one reserved ICF defined.

Now let us see the logical upgrade. LP3 has one reserved ICF defined, and since the server now has one more physical ICF, the reserved ICF can be configured online by the CF image operator function. This CF image is nondisruptively upgraded to two ICFs.

8.6.2 Planning for nondisruptive upgrades

CUoD, CIU, On/Off CoD, and CBU can be used to concurrently upgrade a z990 server. But there are some situations that require a disruptive task to use the new capacity just added to the server. Some of these can be avoided if planning is done in advance. Planning ahead is a key factor for nondisruptive upgrades. Refer to Table 6-3 on page 148 for more discussion about nondisruptive planning.

Reasons for disruptive upgrades

These are the current main reasons for disruptive upgrades:

- ▶ Changing the number of logical partitions defined to a z990 server.
The only way to add or delete a logical partition is by a POR using a new IOCDs including or excluding the new partition.
- ▶ Changing the number of LCSS on a server.
- ▶ Changing the number of subchannels supported on a LCSS.
- ▶ Logical partition processor upgrades when reserved processors were not previously defined are disruptive to image upgrades.
- ▶ Memory capacity upgrades are disruptive when memory cards replacement is required.
- ▶ Logical partition memory upgrades when reserved storage was not previously defined are disruptive to image upgrades.
- ▶ Installation of I/O cages is disruptive.
- ▶ An I/O upgrade when the operating system cannot use the Dynamic I/O configuration function.
 - Linux and CFCC do not support Dynamic I/O configuration.
 - If there is no space available in the reserved HSA for the required I/O expansion.
- ▶ An STI rebalancing, when the STI Rebalance feature (Feature Code 2400) is ordered at the server's upgrade configuration time.
- ▶ Adding a PCIXCC or a PCICA coprocessor to a logical partition, if not predefined with the appropriate PCI cryptographic processor number selected in the PCI Cryptographic Candidate List of the logical partition's image profile.

Recommendations to avoid disruptive upgrades

Based on the previous list of reasons for disruptive upgrades, here are some recommendations to avoid or at least minimize these situations, increasing the possibilities for nondisruptive upgrades:

- ▶ Define spare or reserved logical partitions.
A z990 server can have up to 30 logical partitions defined. It is possible to define more partitions than you need in the initial configuration, just by:
 - Including more partition names in the IOCP statement RESOURCE. The *spare partitions* do not need to be activated, so any valid partition configuration can be used during their definitions. The initial definitions (LPAR mode, processors, and so on) can be changed later to match the image type requirements.

The only resource that spare partitions will use is subchannels, so careful planning must be done here, keeping in mind that z990s can have up to 1890 K subchannels (63 K per logical partition * 30 partitions) total in HSA.

- Defining *reserved* logical partitions, if you are running z/OS V1.6 or higher. The dynamic logical partition name definition allows reserved partition ‘slots’ to be created in an IOCDS in the form of extra Logical Channel Subsystem, Multiple Image Facility (MIF) image pairs. These extra Logical Channel Subsystem MIF image ID pairs (CSSID/MIFID) can be later assigned a logical partition name for use (or later removed) via dynamic I/O commands using the Hardware Configuration Definition (HCD), concurrently. A reserved partition is defined with the partition name placeholder ‘*’, and cannot be assigned to access our candidate list of channel paths or devices. The IOCDS still must have the extra I/O slots defined in advance, since structures are built in the Hardware System Area (HSA).
- ▶ Define an appropriate number of Logical Channel Subsystems (LCSSs).

Define the appropriate number of LCSSs (maximum is four), based on the required number of logical partitions (maximum is 30) and the number of CHPIDs (maximum is 256 per image and per LCSS) that a future configuration may have. Spare and reserved logical partitions, which can have partition name ‘*’ for future renaming, as described in the previous item, can help to define additional LCSSs for future use.

Spanned channels can help spread logical partitions across LCSSs while maintaining physical channels sharing for some channel types (see Table 3-8 on page 92 for a list of supported spanned channels).
- ▶ Define the maximum supported number of subchannels on each LCSS.

The z990s can have up to 1890 K subchannels (63 K per logical partition * 30 partitions) total in HSA (the current maximum number of subchannels on a z900 is 63 K).
- ▶ Configure as many Reserved Processors (CPs, IFLs, ICFs, and zAAPs) as possible.

Configuring Reserved Processors 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 processors online.
- ▶ 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 (64 GB), even if no other book is already installed. The current partition storage limit is 128 GB.
- ▶ Start with a convenient memory size.

Use a convenient entry point memory capacity to allow future concurrent memory upgrades within the same memory cards already installed on the books.
- ▶ 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 cage(s) required by future I/O upgrades, allowing the planned concurrent I/O upgrades.
- ▶ Define all possible PCI cryptographic coprocessors as candidates for all logical partitions.

You can select PCI cryptographic processor numbers in the PCI Cryptographic Candidate List of an image profile even if there are no PCIXCC or PCICA coprocessors currently installed on the server.

Considerations when installing additional books

During a z990 server upgrade, additional books can be concurrently installed. Depending on the number of additional books in the upgrade and the customer's I/O configuration, an STI Rebalancing may be recommended for availability reasons. However the z990 STI Rebalancing, via the STI Rebalance feature, requires STI recabling and results in a server outage. It may also change ICB-4 PCHID numbers, requiring an I/O definition update.

8.7 Capacity planning considerations

The z990 servers represent a major evolution for the Series platform, extending the key platform characteristics and embracing e-business on demand requirements.

The z990 has significant performance improvements over the previous z900 servers. The z990 server includes enhanced processor and system designs, and also introduces new building blocks, such as the multi-book structure.

z990 introduces a new microprocessor architecture exploiting the CMOS9S-SOI technology while improving uniprocessor performance. A significant capacity and throughput increase has been achieved with the introduction of:

- ▶ Up to 32 Processor Units as CPs, IFLs, ICFs, or zAAPs, for operating systems
- ▶ Up to eight Processor Units, as standard System Assist Processors (SAPs), for I/O processing
- ▶ Up to 30 logical partitions
- ▶ Up to 256 GB of memory
- ▶ Up to 96 GBps of bandwidth for data communication via up to 48 Self-Timed Interconnect (STI) host buses
- ▶ A new Channel Subsystem (CSS): four Logical Channel Subsystems (LCSSs) can exist for horizontal growth, supporting up to 256 CHPIDs per CSS for a total of 1024 CHPIDs per system
- ▶ Increased channel maximums for ESCON, FICON Express, and OSA-Express
- ▶ Three cryptographic features:
 - New CP Assist for Cryptographic Function (CPACF)
 - New PCIX Cryptographic Coprocessor (PCIXCC)
 - PCI Cryptographic Accelerator (PCICA)
- ▶ Integrated Cluster Bus-4 (ICB-4), capable of up to 2 GBps

Some of the most important performance related topics are the balanced system design, the superscalar processors, and the integrated hardware and system assists.

8.7.1 Balanced system design

One of the most important design objectives of a zSeries server is to build a balanced system. This means a system with no specific constraints, where individual parts, such as processor speed, memory bandwidth, and I/O bandwidth, are designed for the server's best performance and throughput.

The balanced system design is also based on the fact that no single component just by itself, like processor frequency (expressed in GHz), can improve the system's overall performance for a wide range of workloads and applications.

The z990 servers have performance improvements on all workload environments, from traditional to e-business on demand. Comparing to the z900 turbo servers, the z990 has improved all major components:

- ▶ Maximum number of assigned processors, from 16 to 32
- ▶ Processor cycle time, from 1.09 ns to 0.83 ns
- ▶ L2 caches, from 32 MB per 20 PUs to 32 MB per 12 PUs
- ▶ Maximum memory size, from 64 GB to 256 GB
- ▶ STI bandwidth, from 1 GBps to 2 GBps per STI
- ▶ Maximum number of STIs, from 24 to 48 STIs
- ▶ Maximum I/O bandwidth, from 24 GBps to 96 GBps
- ▶ Maximum number of channels, from 256 to 1024

Additional performance improvements for e-business

z990 has also further performance improvements for e-business application environments:

- ▶ zSeries Application Assist Processors (zAAPs), which are designed to operate asynchronously with the CPs to execute Java programming under control of IBM Java Virtual Machine (JVM) for logical partitions running z/OS. The IBM JVM processing cycles can be executed on the configured zAAPs with no anticipated modifications to the Java applications.
- ▶ IEEE Floating Point: Used by Java and C/C++ applications, the new Binary Floating Point unit halves the number of cycles required on previous servers.
- ▶ Secondary level Translation Lookaside Buffer (TLB): A secondary cache for Dynamic Address Translation, for both the instruction and data caches, increases the number of buffer entries by a factor of eight.
- ▶ CP Assist for Cryptographic Function (CPACF): Implemented on each PU, the assist function uses five new instructions for symmetrical clear key cryptographic encryption and decryption operations, to accelerate the encryption and decryption of SSL transactions, and VPN encrypted data transfers.

In addition, the following improvements for specific areas are also implemented on z990:

- ▶ Compression Unit:
The Compression Unit is integrated with the CP Assist for Cryptographic Function, benefiting from combining the use of buffers and interfaces. It is implemented on each PU and provides excellent hardware compression performance.
- ▶ Checksum offload for IPV4 packets when in QDIO mode for Linux and z/OS:
Checksum Offload provides the capability of calculating the Transmission Control Protocol (TCP), User Datagram Protocol (UDP), and Internet Protocol (IP) header checksums. Checksum verifies the correctness of files. By moving the checksum calculations to a Gigabit or 1000BASE-T Ethernet feature, host CPU cycles are reduced and performance is improved. It is supported by the OSA-Express GbE and 1000BASE-T Ethernet features when operating at 1 Gbps.

Multi-book structure

The multiple book structure introduced with the z990 servers offers more flexibility, capacity, and scalability to the system.

Each book has its own MCM (which contains PUs and an L2 cache), memory cards, and MBAs with their STIs. Up to four books are connected through L2 caches by concentric rings, resulting in a single integrated system.

Previous zSeries servers have PU clusters, or PU sets, which are also connected to each other through L2 caches. But in those cases, all PUs and L2 caches reside in a single MCM.

The z990 multi-MCM design introduces two types of PU to L2 cache access: a “local” access, when the PU and L2 cache are located in the same MCM (or book), and a “remote” access, when PU and L2 cache are located in different books.

Figure 8-16 shows a two-book z990 server logical view.

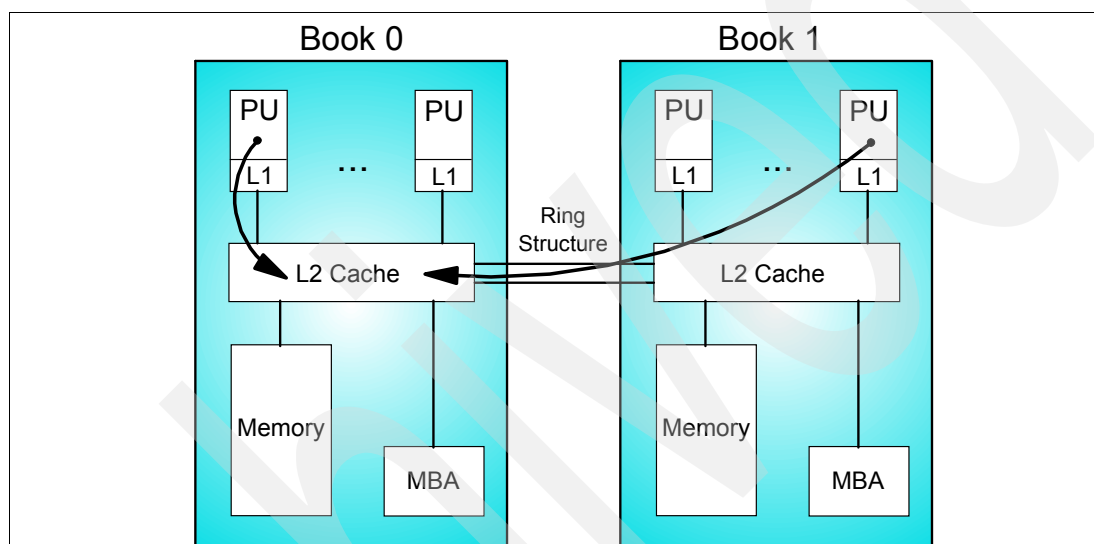


Figure 8-16 Two-book system logical view

In this example, a local access is done by the first PU on Book 0 to its local L2 cache, and a remote access is done by the last PU on Book 1 to the L2 cache on Book 0. As the access time of the remote connection is higher than a local one, the performance of such a system would not be as consistent as a single MCM system, being dependent on the remote access rates. To avoid this effect, z990 has implemented some optimizations.

The L2 cache is implemented as a processor cache, not as a memory cache. This means that data (and instructions) are normally residents in the L2 cache on the book where it is being used by a PU, and not in the book where the associated memory address resides. So in the previous example, the L2 cache in Book 1 will have the data/instructions used by its local PU after the remote L2 cache access.

Along with the PU allocation and assignment algorithm during IML, described in “Processor Unit characterization” on page 52, the PR/SM has a major role in the z990 system optimization.

The z990 PR/SM has changed to support the multi-book structure and to provide optimal system performance. PR/SM is aware of the physical book structure, while the logical partitions do not require awareness about this design. The PR/SM hypervisor manages and optimizes allocation and dispatching for the underlying physical topology, providing a transparent multi-book implementation to operating systems. The PR/SM main objective is to allocate all processors and storage for a logical partition to the same book, and to redispach a logical processor back to the same physical processor.

This implementation provides optimal performance and a more linear scalability to the z990 server. The results can be observed in the LSPR's ITR values from the uniprocessor to the 32-way server.

8.7.2 Superscalar processors

The z990 server is the first generation of zSeries servers that uses superscalar processors. A superscalar processor can execute multiple instructions per cycle, potentially providing better performance than a sequential processor running at the same cycle time (or processor frequency).

To exploit this parallel execution capability, the z990 server has implemented an improved instruction scheduling. The instruction execution order is optimized to provide instruction sequences that can operate in a multi-pipeline environment. The major enhancements are in the e-business applications, such as WebSphere and other Java and C/C++ code.

Further exploitation can also be done on the software side. Compilers can be changed to provide an optimized instruction scheduling to better exploit the superscalar design.

8.7.3 Integrated hardware and system assists

To achieve the required throughput and implement new functions while maintaining balanced usage of system resources, integrated hardware and system assists are key.

zSeries Application Assist Processors (zAAPs)

The zSeries Application Assist Processors (zAAPs) are designed to operate asynchronously with the CPs to execute Java programming under control of IBM Java Virtual Machine (JVM) for logical partitions running z/OS. This can help reduce the demands and capacity requirements on CPs that may be available for relocation to other zSeries workloads.

The IBM JVM processing cycles can be executed on the configured zAAPs with no anticipated modifications to the Java applications. Execution of the JVM processing cycles on a zAAP is a function introduced by the Software Developer's kit (SDK) 1.4.1 for zSeries, z/OS V1.6, and the Processor Resource/Systems Manager (PR/SM).

Cryptographic function on every Processor Unit (PU)

The z990 introduces the Cryptographic Assist Architecture (CAA) along with the CP Assist for Cryptographic Function (CPACF), delivering cryptographic support on every Processor Unit (PU) with DES and TDES data encryption/decryption and SHA-1 hashing.

This offers balanced use of system resources and provides unprecedented scalability (a z990 can have from one to 32 PUs, depending upon model) and data rates at 2X or more faster than the CMOS Cryptographic Coprocessor Facility (CCF).

Since these cryptographic functions are implemented in each and every PU, the association of cryptographic functions to specific PUs, as was done with previous generations of zSeries, is eliminated.

Secure encrypted transactions with higher performance

PCIX Cryptographic Coprocessor (PCIXCC) FC 0868 is a replacement for the PCI Cryptographic Coprocessor (PCICC) and the CMOS Cryptographic Coprocessor Facility that were offered on z900. All of the equivalent PCICC functions that are implemented offer higher performance. In addition, the functions on the CMOS Cryptographic Coprocessor Facility used by known applications have also been implemented in the PCIXCC feature.

The PCIxCC feature supports secure cryptographic functions, use of secure encrypted key values, and User-Defined Extensions.

Continued cryptographic support for the e-business environment

PCI Cryptographic Accelerator (PCICA) FC 0862 is also available on the z990. This hardware-based cryptographic solution continues to address the high Secure Sockets Layer (SSL) performance needs of the on demand environment.

The SSL and Transport Layer Security (TLS) protocols are essential and widely used protocols to support secure e-business applications. Compute-intensive public key cryptographic processes that use SSL/TLS can be offloaded from the host to the PCICA feature to reduce CP and IFL usage, thus increasing system throughput.

Performance assists for Linux and z/VM

- ▶ z990 adapter interruptions for Linux and z/VM: The z990, Linux on zSeries, and z/VM work together to provide performance improvements by exploiting extensions to the Queued Direct Input/Output (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: FICON Express when using the FCP CHPID type, and OSA-Express when using the OSD CHPID type.

In extending the use of adapter interruptions to FCP and OSD (QDIO) channels, the programming overhead to process a traditional I/O interruption is reduced. This benefits OSA-Express TCP/IP support in both Linux on zSeries and z/VM, and FCP support in Linux on zSeries.

Adapter interruptions apply to a z990 FICON Express channel when in FCP mode (FCP CHPID type), which supports attachment of SCSI devices in a Linux on zSeries environment. This support is exclusive to z990 and applies to all of the OSA-Express features available on z990, when in QDIO mode (OSD CHPID type).

- ▶ Performance assist for V=V guests in the z/VM environment: z990's support of virtual machine technology has been enhanced to include a new performance assist for virtualization of Adapter Interruptions. This new z990 performance assist is available to V=V guests (only pageable guests) that support QDIO (Queued Direct Input/Output) on z/VM V4.4 and later. The deployment of adapter interruptions improves efficiency and performance by reducing overhead.

The z990 performance assist for V=V guests is a passthrough architecture that reduces host programming overhead by avoiding the need to stop guest processing when adapter interruptions are presented. Without the assist, the z/VM control program must intercede to process and route the adapter interruptions.

The z990 performance assist improves Linux on zSeries performance under z/VM by allowing guest I/O (FICON (FCP CHPID type), HiperSockets (IQD CHPID type), and OSA-Express (OSD CHPID type)) to be handled with minimal z/VM overhead through direct presentation of adapter interruptions by the server to a pageable guest, boosting I/O performance.

8.8 Capacity measurements

The result of all performance related characteristics of a server should be evaluated by their combined effect when estimating the system performance. Some design and/or implementation aspects can improve the performance of specific workloads and this should also be evaluated.

Measurements are the most accurate source for processor capacity data. Modeling techniques may produce reasonably accurate processor capacity data, assuming all pertinent workload and hardware design and implementation factors are considered.

There is no reasonable way to construct a benchmark that simulates instruction paths and storage reference patterns typical of a production workload without using actual production software and activities.

IBM utilizes the Large Systems Performance Reference (LSPR) method to provide relative capacity information, which takes into account processor design sensitivities to workload type.

LSPR benchmarks are laboratory controlled tests of representative workload environments, objectively measured and analyzed.

8.8.1 Large Systems Performance Reference (LSPR)

The IBM Large Systems Performance Reference (LSPR) method is intended to provide comprehensive S/370™, S/390, and zSeries architecture processor capacity data across a wide variety of operating systems, or System Control Programs (SCPs), and workload environments.

To assure that the processor is the LSPR's primary focus, the processor capacity data reported assumes sufficient external resources, such as storage size, number of channels, control units, and I/O devices, so as to prevent any significant external resource constraints.

LSPR data is based on a set of measured benchmarks and analysis, and is intended to be used to estimate the capacity expectation for a given production workload when considering a move to a new processor.

The average rate that processors execute instructions is quoted as Millions of Instructions Per Second (MIPS). With today's high-performance processors, the actual MIPS rate achieved is extremely sensitive to the workload type being run, and its relationship to underlying processor design. Therefore, the relative capacity of one processor to other will be very dependent on the type of the work being run.

For this reason, IBM has chosen to provide capacity data in terms of work accomplished, or throughput, in various operating systems and workloads environments, rather than in MIPS or instruction execution rates.

Internal Throughput Rate (ITR) and ITR Ratio (ITRR)

LSPR uses the Internal Throughput Rate (ITR) metric to measure work done on different workloads environments. ITR is computed as:

$$\text{ITR} = \text{Units of Work} / \text{Processor Busy Time}$$

The "Processor Busy Time" is normalized to 100% utilization. "Units of Work" are normally expressed as jobs for batch workloads, and as transactions or commands for online workloads.

ITR characterizes processor capacity, since it is a CPU busy time measurement. ITRs are useful for determining the relative capacity between two processors running the exact same workload environment. However, the absolute ITR values from one workload *cannot* be compared to those of a different workload environment.

The relative capacity between processors for a given workload is done by dividing the ITR of one processor by the ITR of another to produce an ITR ratio, called ITRR. For example, to

determine the capacity of processor B relative to that of processor A, the ITR Ratio (ITRR) is calculated as follows:

$$\text{ITRR} = \text{ITR for Processor B} / \text{ITR for Processor A}$$

ITR values used in this calculation *must* be for identical workload environments.

Each individual LSPR workload is designed to focus on a single major type of activity, such as interactive, online database, or batch. The LSPR does not focus on individual pieces of work, such as a specific job or application. Instead, each LSPR workload includes a broad mix of activity related to that workload type.

The ITR value for each workload environment is measured for IBM (and some non-IBM) processors and the results are published via LSPR tables. An LSPR table shows the ITR Ratios (ITRRs) of each processor within a group related to a base processor. The base processor is set as ITR = 1 for all workloads, so all other processors in the table are compared to the base one, for each workload environment.

To obtain a single number that could estimate the *average* capacity of a given processor, a mixed workload is also calculated. A mixed workload consists of a mix of selected LSPR workloads. Remember that single-number capacity tables may be useful for rough processor positioning, but cannot provide a precise view of relative processor capacity and should not be used for capacity planning purposes.

The LSPR is now using new predefined mixed workloads, offering better average capacity estimation for the most usual production environments, and providing more representative average numbers for relative processor capacity evaluation.

LSPR workloads prior to z990

The LSPR workloads prior to z990 are listed Table 8-2. The measured ITRs represent Basic Mode.

Table 8-2 LSPR workloads prior to z990 (Basic Mode)

Operating system	Workload type	Workload description
OS/390	FCP1	Engineering and Scientific batch (Floating Point)
	CBW2	CPU-intensive commercial batch
	CB84	I/O-intensive commercial batch
	TSO	Interactive TSO user population
	CICS/DB2	Traditional OLTP using CICS and DB2
	IMS	Traditional OLTP using IMS
	R3-DB	EAS DB server (SAP SD benchmark)
VM/ESA	CMS1	Interactive CMS user population
VSE/ESA	CICS	Traditional OLTP using CICS
	CICS VM/V=R	Traditional OLTP using CICS as a VM V=R guest

The default mixed workload consists of an equal mix (25%) of CB84, TSO, CICS/DB2, and IMS workloads, running under OS/390 V2 R10 in Basic mode, in a mixture of 31-bit and 64-bit mode addressing.

For a complete description of these LSPR workloads, refer to *Large Systems Performance Reference*, SC28-1187.

LSPR workloads for z990

The z990 servers introduce many different innovations, such as multi-book design, LPAR Mode only, up to 32 CPs/IFLs/ICFs/zAAPs, and up to 30 logical partitions. Also, some new workload types are now required to better evaluate processor capacity, as the production environments are changing, including new e-business applications.

Those factors contributed to the following LSPR changes for the z990 processors:

- ▶ Newer version of operating systems (z/OS V1 R4 and z/VM V4 R3)
- ▶ Some workloads dropped; others changed or added
- ▶ New VM workloads run as Linux virtual machines
- ▶ New Linux native workloads
- ▶ All ITRs now assume LPAR mode
- ▶ Most workload ITRs now assume 64-bit addressing
- ▶ VSE/ESA workloads have been dropped

The LSPR workloads for z990 are listed in Table 8-3. The measured ITRs represent LPAR Mode.

Table 8-3 LSPR workloads for z990 (LPAR Mode)

Operating system	Workload type	Workload description
z/OS	CB-L	CPU-intensive commercial batch (formerly CBW2).
	CB-S	I/O-intensive commercial batch (formerly CB84).
	WASDB	WebSphere Application Server+DB2 (Trade2-EJB).
	OLTP-W	Web-enabled OLTP. CICS/DB2 with WebSphere Application Server front end.
	OLTP-T	Traditional OLTP (IMS).
z/VM	CMS1	Interactive CMS users.
	WASDB/LV/m	Linux guests with WebSphere app+DB (Trade2-EJB).
Linux	WASDB/L	WebSphere app+DB (Trade2-EJB).
	EAS-AS/L	Enterprise Application Solution Application Serving under Linux.

The following are descriptions of the new workloads:

- ▶ CB-L (Commercial Batch Long job steps - CPU-Intensive; formerly CBW2)

The CB-L workload is a commercial batch jobstream reflective of fairly heavy CPU processing. The jobs are more resource-intensive than jobs in the CB-S workload, use more current software, and exploit ESA features. The work done by these jobs includes various combinations of C, COBOL, FORTRAN, and PL/I compile, link-edit, and execute steps. Sorting, DFSMS, VSAM and DB2 utilities, SQL processing, SLR processing, GDDM® graphics, and FORTRAN engineering/scientific subroutine library processing are also included. This workload is heavily DB2-oriented, with about half of the processing time performing DB2-related functions.

- ▶ **CB-S (Commercial Batch Short job steps - I/O-Intensive; formerly CB84)**

The CB-S workload is a moderate commercial batch jobstream reflective of fairly I/O processing. The work done by these jobs includes various combinations of compile, link-edit, and execute steps. Utility jobs, primarily for data manipulation, are also included.

- ▶ **WASDB (WebSphere Application Serving and Data Base)**

The WASDB workload reflects a new-e-business production environment that uses WebSphere applications and a DB2 data base all running in z/OS.

WASDB is a collection of Java classes, Java Servlets, Java Server Pages, and Enterprise Java Beans integrated into a single application. It is designed to emulate an online brokerage firm. WASDB was developed using the VisualAge® for Java and WebSphere Studio tools. Each of the components is written to open Web and Java Enterprise APIs, making the WASDB application portable across J2EE-compliant application servers.

- ▶ **OLTP-W (Web-enabled On-line workload)**

The OLTP-W workload reflects a production environment that has Web-enabled access to a traditional data base. For the LSPR, this has been accomplished by placing a WebSphere front end to connect to the LSPR CICS/DB2 workload.

The J2EE application for legacy CICS transactions was created using the CICS Transaction Gateway (CTG) external call interface (ECI) connector enabled in a J2EE server in WebSphere for z/OS V4.0.1. The application uses the J2EE architected Common Client Interface (CCI). Clients access WebSphere services using HTTP Transport Handler. Then the appropriate servlet is run through the Web container, which calls EJBs in the EJB Container. Using the CTG external call interface (ECI,) CICS is called to invoke DB2 to access the database and obtain the information for the client.

- ▶ **OLTP-T (Traditional On-line workload - formerly IMS)**

The OLTP-T workload consists of light-to-moderate IMS transactions from DLI applications covering diverse business functions. These applications all make use of IMS functions, such as logging and recovery. Conversational and wait-for-input transactions are included in the workload.

- ▶ **CMS1 (CMS workload used for z/VM)**

The CMS workload is designed to represent a VM/CMS end-user community. Processor time per command, I/Os per command, T/V ratio, and think time distribution are similar to those observed for actual VM production systems running large numbers of CMS users.

- ▶ **WASDB/LVm (Linux guests under z/VM running WebSphere Application Serving and Data Base)**

The WASDB/LVm workload reflects a server consolidation environment where the servers being consolidated were running a full function application. For LSPR, this was accomplished by taking the WASDB workload, splitting it across a pair of Linux guests (one guest for application and one guest for database), and then replicating the Linux pair many times to reflect the consolidation of many independent servers. The software levels used were Linux SLES 7, WebSphere Application Server 4.0.4, and UDB 7.0.

- ▶ **HSRV/LV (Linux guests under z/VM performing HTTP Serving)**

The HSRV/LV workload reflects a server consolidation environment where the servers being consolidated were performing HTTP serving.

The workload simulates browsers accessing Web pages of mainly HTML files and their graphics. The bulk of the files range in size from 1 KB to 100 KB, with a small number of files greater than 1 MB being accessed. This last set simulates a large file being downloaded. A driving system is used to send requests to the system under test. Client subprocesses, or threads, generate an independent stream of HTTP requests, pausing in

between requests so that on average it generates the specified number of requests per second.

► **WASDB/L (WebSphere Application Serving and Data Base under Linux)**

The WASDB/L workload reflects an e-business environment where a full function application is being run under Linux in logical partition. For LSPR, this was accomplished by taking the WASDB workload, and converting it to run both application and data base servers in a single Linux image.

The WASDB/L workload is basically the same as the WASDB workload for z/OS, with the exception of being enabled for Linux. UDB 7.0 is used instead of DB2 v7.0, and WebSphere AE 4.0.4 is used instead of WebSphere 4.0.1.400.

► **EAS-AS/L (Enterprise Application Solution Application Serving under Linux)**

The EAS-AS/L workload reflects the Application Server (AS) portion of the Enterprise Application Solution running in a Linux environment. The AS resides in a Linux on zSeries image while the database server and presentation server reside outboard. The EAS application used for this workload is the SAP R/3 product using DB2 for z/OS. The workload is derived from the SAP AG defined Sales and Distribution (SD) environment, but is not based on SAP AG certified benchmarks results. Specifically, the results show an SAP R/3 AS in a 3-tier client/server configuration. The other two tiers, database server and presentation server, are not represented in this data.

This workload is similar to the previous EAS-DB (OS/390 R3-DB) workload description. The software levels used are SUSE Linux Enterprise Server 8 for zSeries and SAP R/3 Release 4.6D.

New predefined z/OS workload mixes

For better accuracy when projecting capacity with LSPR workload data and improved consistency when working across multiple LSPR releases, new predefined z/OS workload mixes are also introduced.

There are six predefined z/OS workload mixes:

- **LSPR-Mix:** LSPR generic mix (60% online, 40% other), the default
- **TI-Mix:** Transaction intensive mix (60% online, 40% other)
- **TD-Mix:** Transaction dominant mix (40% online, 60% other)
- **TM-Mix:** Transaction moderate mix (30% online, 70% other)
- **CB-Mix:** Commercial batch mix (100% other)
- **LoIO-Mix:** Low I/O content (special; for use when less than 30 DASD I/Os/sec per MSU)

Table 8-4 lists the LSPR workload types and percentages used for each predefined z/OS mixed workload.

Table 8-4 New z/OS predefined workload mixes

Workload type	z/OS Workload Mixes					
	LSPR-Mix (Default)	TI-Mix	TD-Mix	TM-Mix	CB-Mix	LoIO-Mix
CB-L	20%	30%	45%	52.5%	75%	60%
CB-S	20%	10%	15%	17.5%	25%	-
WASDB	20%	-	-	-	-	20%
OLTP-W	20%	30%	20%	15%	-	20%

Workload type	z/OS Workload Mixes					
	LSPR-Mix (Default)	TI-Mix	TD-Mix	TM-Mix	CB-Mix	LoIO-Mix
OLTP-T	20%	30%	20%	15%	-	-

The new default mixed workload is the LSPR-Mix, which is calculated with equal mix (20%) of CB-L, CB-S, WASDB, OLTP-W and OLTP-T workloads, running under z/OS V1 R4, in LPAR mode.

z990 LSPR tables

The current LSPR tables, including all the z990 ITR Ratios, based on a z990 uniprocessor for all workload environments, can be found at:

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

Archived

Environmental requirements

This chapter introduces the IBM eServer™ zSeries 990 environmental requirements. We list its dimensions, weights, power, and cooling requirements as an overview of what is needed to plan for the installation of a z990 server.

For more comprehensive physical planning information, refer to *IBM @server zSeries 990 Installation Manual for Physical Planning*, GC28-6824.

We cover the following topics:

- ▶ 9.1.1, “Power and cooling requirements” on page 232
- ▶ 9.2, “Weights” on page 233
- ▶ 9.3, “Dimensions” on page 234

9.1 Introduction

The z990 is always a two-frame system. The frames are shipped separately and are fastened together when installed.

Installation of a z990 is always on a raised floor. The number of cables to be expected for most configurations may be so large that installation is only possible with space underneath. The dimensions of a z990 are slightly smaller than that of a two-frame z900 and its maximum weight is slightly higher.

9.1.1 Power and cooling requirements

The z990 requires at least two power feeds and uses two redundant three-phase line cords, allowing the system to survive the loss of power to either one. In case of a power failure of one of the line cords, the other one is able to take over the entire load to keep the system operating without interruption. The z990 is installed with three-phase wiring and operates with 50/60Hz AC power, and voltages ranging from 200V to 480V. For ancillary equipment (like the Hardware Management Console, its display, and the modem), additional single-phase outlets are required.

9.1.2 Power consumption

Actual power consumption is dependent on the server configuration in terms of the number of books and the number of I/O cages installed. The figures listed in Table 9-1 assume the maximum configuration.

Table 9-1 Power consumption and heat load

Model	One I/O cage	Two I/O cages	Three I/O cages
IBM 2084 model A08	6.74 kW	10.64 kW	13.81 kW
IBM 2084 model B16	9.57 kW	13.27 kW	16.98 kW
IBM 2084 model C24	11.82 kW	15.53 kW	19.23 kW
IBM 2084 model D32	13.98 kW	17.68 kW	21.39 kW

Input power in kVA is equal to the output power in kW. Heat output expressed in kBTU per hour is derived by multiplying the table entries by a factor of 3.4.

The maximum allowed circuit breaker rating is 60 Amps, which is to be used for both power feeds where 200-240V is applicable. Where 380-480 Volts is applicable, 30 Amps are recommended for both power feeds.

9.1.3 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 UPS system can be connected to the z990, allowing for longer periods of sustained operation.

The Internal Battery Feature, given that the batteries are not older than three years and have been discharged regularly, are capable of providing emergency power for the periods of time shown in Table 9-2 on page 233.

Table 9-2 Internal Battery Feature emergency power times

Model	One I/O cage	Two I/O cages	Three I/O cages
IBM 2084 model A08	8 minutes	13 minutes	12 minutes
IBM 2084 model B16	13 minutes	8.5 minutes	10.5 minutes
IBM 2084 model C24	8.5 minutes	11 minutes	9 minutes
IBM 2084 model D32	13 minutes	8.5 minutes	7.5 minutes

9.1.4 Emergency power-off

On the front of frame A is an emergency power-off switch that will immediately disconnect utility and battery power from the server when activated. This causes all volatile data in the server to be lost.

In case a server is connected to a machine room emergency power-off switch, and the Internal Battery Feature is installed, the batteries will take over if the switch is engaged.

It is possible to connect the machine room emergency power-off switch to the server power-off switch. In that case, when the machine room emergency power-off switch is engaged, all power will be disconnected from the line cords and the Internal Battery Features. This causes all volatile data in the server to be lost.

9.1.5 Cooling requirements

The z990 requires chilled air from under the raised floor to fulfill the air-cooling requirements. The chilled air usually is provided through perforated floor tiles. The amount of chilled air needed in the computer room is indicated in the IMPP for a variety of underfloor temperatures.

At an underfloor temperature of 20° Celsius (68° Fahrenheit), the cooling airflow requirements with maximum populated I/O cages are listed in Table 9-3.

Table 9-3 Underfloor cooling airflow requirements (C/FM)

Model	One I/O cage	Two I/O cages	Three I/O cages
IBM 2084-A08	1050	1350	1750
IBM 2084-B16	1350	1750	2200
IBM 2084-C24	1750	2200	2600
IBM 2084-D32	1750	2200	2600

9.2 Weights

Since there may be a large number of cables connected to a z990 installation, a raised floor is mandatory. In the IMPP, 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 for the z990 system.

Table 9-4 on page 234 indicates the minimum and maximum system weights for all models. The weight ranges are base on configuration models with one and three I/O cages.

Table 9-4 System weights

Configuration	Weight in kg (lb) without IBF	Weight in kg (lb) with IBF
IBM 2084-A08	1174 (2582) to 1534 (3376)	1263 (2779) to 1714 (3770)
IBM 2084-B16	1281 (2818) to 1642 (3612)	1460 (3212) to 1910 (4203)
IBM 2084-C24	1329 (2924) to 1690 (3718)	1508 (3318) to 1959 (4309)
IBM 2084-D32	1401 (3082) to 1738 (3824)	1669 (3673) to 2007 (4415)

9.3 Dimensions

The z990 always has two frames: frame A and frame Z. The external dimensions of both frames of a z990, with and without covers, are listed in Table 9-5.

Table 9-5 Frame dimensions

Frames	Width mm (in)	Depth mm (in)	Height mm (in)
Frame A without covers	750 (29.5)	1172 (46.1)	1921 (75.6)
Frame A with covers	767 (30.2)	1577 (62.1)	1941 (76.4)
Frame Z without covers	750 (29.5)	1172 (46.1)	1921 (75.6)
Frame Z with covers	767 (30.2)	1519 (58.1)	1941 (76.4)
Frame A or Z with height reduction	n/a	n/a	1785 (70.3)

Note: The total machine room area required is 2.49 square meters (26.78 square feet). With service clearance, 5.45 square meters (58.69 square feet) are needed.

Hardware Management Console (HMC)

In this appendix, we discuss the z990 Hardware Management Console, and provide you with some configuration guidelines.

The HMC is a PC/ISA bus PC running OS/2®, Communications Server for OS/2, a remote control systems management product, and the Hardware Management Console Application (HMCA). It also contains configuration information about its own configuration and about the Support Element (SE) defined to it.

The HMC is attached to the SE either by a local area network (LAN) through the Multistation Access Unit (MAU) or by an Ethernet LAN. The HMC resides external to the system frame. The HMC can also control and monitor status for multiple Central Processor Complexes (CPCs) configured to it, providing a single point of control and a single system image. One HMC can control 100 SEs, and one SE can be controlled by 32 HMCs.

The physical location of some HMC hardware features (standard or optional) are dictated by the specific PC. Some features may be mutually exclusive with other features, depending on the PC model.

The HMC user interface is designed to provide the functions you need to operate, monitor, and maintain your processor. Various elements of the processor hardware are represented as objects by the HMC application. Through this application, you can directly manipulate the CPC objects that are defined to the HMC and be aware of changes to hardware status as they are detected.

The Hardware Management Console and Support Elements provide the following:

- ▶ A Hardware Management Console Application object-oriented interface
- ▶ Customizable groups of hardware objects
- ▶ A customizable Hardware Management Console
- ▶ Application settings

Consolidation of:

- ▶ Operator controls
- ▶ Hardware status reporting
- ▶ Hardware message presentation
- ▶ Operating system messages
- ▶ Problem analysis and reporting
- ▶ Licensed Internal Code (LIC) control and distribution
- ▶ Remote I/O configuration and IOCDS management
- ▶ Scheduled operations

The HMC communicates with the CPC through the Support Element (SE). When tasks are performed at the HMC, the commands are sent to one or more Support Elements which then issue commands to their CPCs. CPCs can be grouped at the HMC so that a single command can be passed along to as many as all of the CPCs defined to the HMC.

One Hardware Management Console can control up to 100 Support Elements, and one Support Element can be controlled by 32 Hardware Management Consoles. Refer to the examples shown in Figure A-1 and Figure A-2 on page 237 for typical Hardware Management Console configurations.

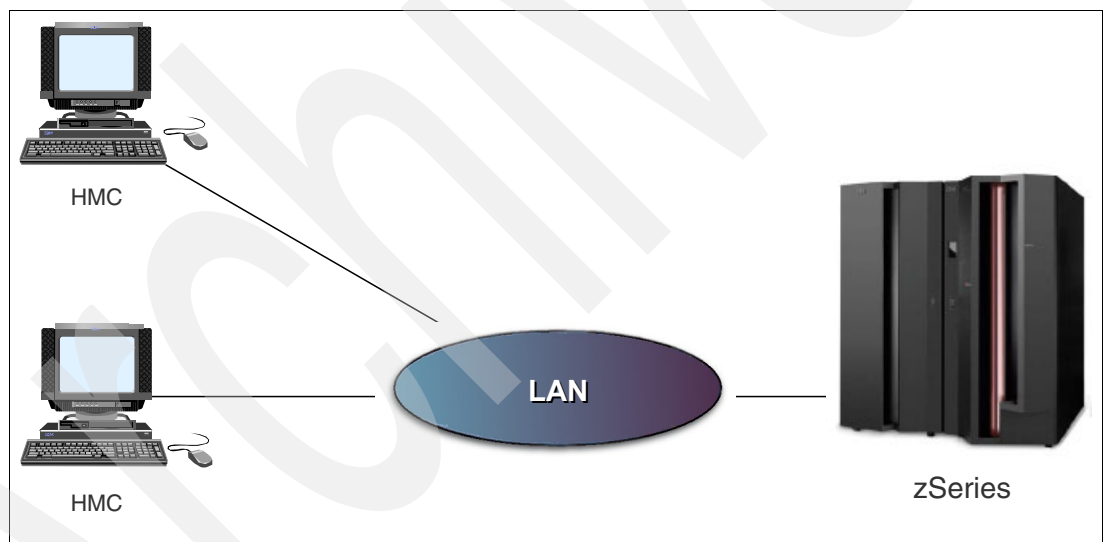


Figure A-1 Single CPC environment

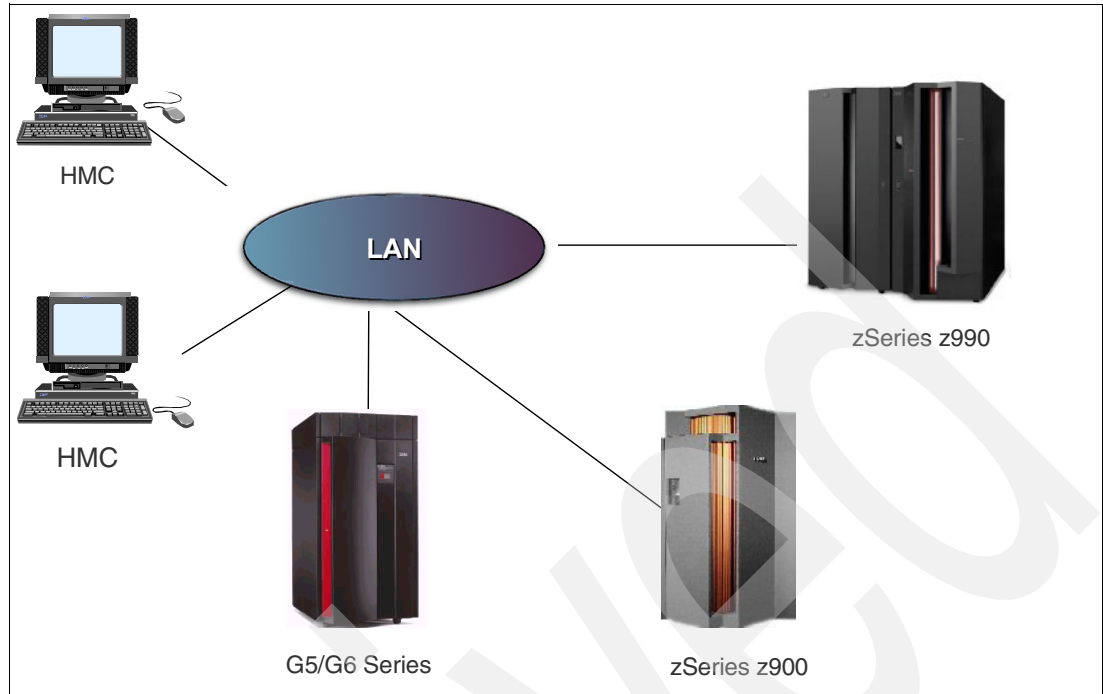


Figure A-2 Multi CPC environment

Important: Beginning with the next zSeries server, after the IBM @server zSeries 890 and 990, all new HMCs on all currently marketed zSeries servers are intended to become closed platforms. They will only support the HMC application and not the installation of other applications, such as the IBM ESCON Director and the IBM Sysplex Timer console applications.

When available, the next generation HMC is expected to communicate only with G5 Servers, and above (Multiprise 3000, G5/G6, z800, z900, z890, z990). TCP/IP is intended to be the only communications protocol supported.

z990 Hardware Management Console

A local Hardware Management Console must be connected to its Support Elements using Local Area Network (LAN) wiring. zSeries 990 provide both token ring and Ethernet options for the LAN wiring between the Hardware Management Console and the Support Elements. The necessary LAN adapters for the Support Elements and the Hardware Management Console may be specified as features on the system order.

Each Support Element has two LAN adapter options. Neither is the default; one of the following must be selected:

- ▶ Dual Ethernet SE
- ▶ Token ring/Ethernet SE

The Hardware Management Console has two LAN adapter options. Neither is the default; you must select one of the following:

- ▶ Dual Ethernet HMC
- ▶ Token ring/Ethernet HMC (traditional)

Important: The z890 and z990 will be the last zSeries servers to offer Token Ring adapter features on the Hardware Management Consoles (HMCs), Support Element (SEs), and Trusted Key Entry (TKE) workstations. The IBM 2074 Model 3 Console Support Controller will be the last controller to offer Token Ring adapter features.

IBM @server zSeries team is making these statements to allow enterprises sufficient opportunity to prepare for a migration to Ethernet environments.

Notes on wiring with multiple adapters

- ▶ It is intended that a Hardware Management Console and the SE together be connected by only *one* LAN.
- ▶ Multiple adapters in a Hardware Management Console allow that Hardware Management Console to connect with two *independent* sets of SEs: one set on a token ring and a second set on an Ethernet.

This is done to allow for migration from token ring environments to Ethernet environments.

- ▶ Multiple adapters in a ThinkPad® SE are intended to allow two *different* Hardware Management Consoles to have independent paths to the SE.

This is done so that the console can be controlled if one LAN goes down.

In the following sections, we describe the four wiring scenarios.

Token ring only wiring scenario

The token ring only wiring scenario, shown in Figure A-3 on page 239, is the standard wiring approach used in previous generations of the IBM Enterprise Server Hardware Management Console and Support Element wiring. As in previous systems, each system includes a Multistation Access Unit (MAU) that may be used to interconnect the token ring adapter wiring of the Support Elements to the Hardware Management Console. Token ring wiring may be used to interconnect the MAUs to form a larger private LAN where multiple systems are to be controlled by a single Hardware Management Console.

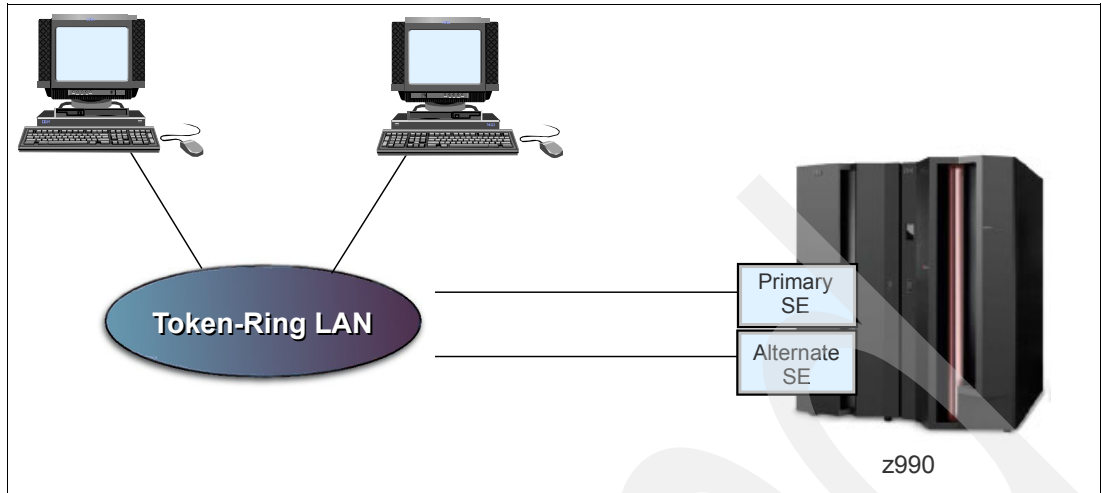


Figure A-3 Token ring only wiring scenario

Additional token ring only wiring scenario

Additional connections to the token ring LAN may be made to expand the connectivity beyond the local Hardware Management Console and Support Elements, as shown in Figure A-4 on page 240.

If connections to previous generations of Enterprise Server systems are desired, they may be connected using the MAU in the system, or they may be connected using token ring-to-token ring bridges.

If connection to the enterprise LAN is desired, it is recommended that a token ring bridge be installed to isolate the Hardware Management Console and Support Elements from other systems.

If connection to a central site focal point is desired, a local control unit can be attached to the LAN.

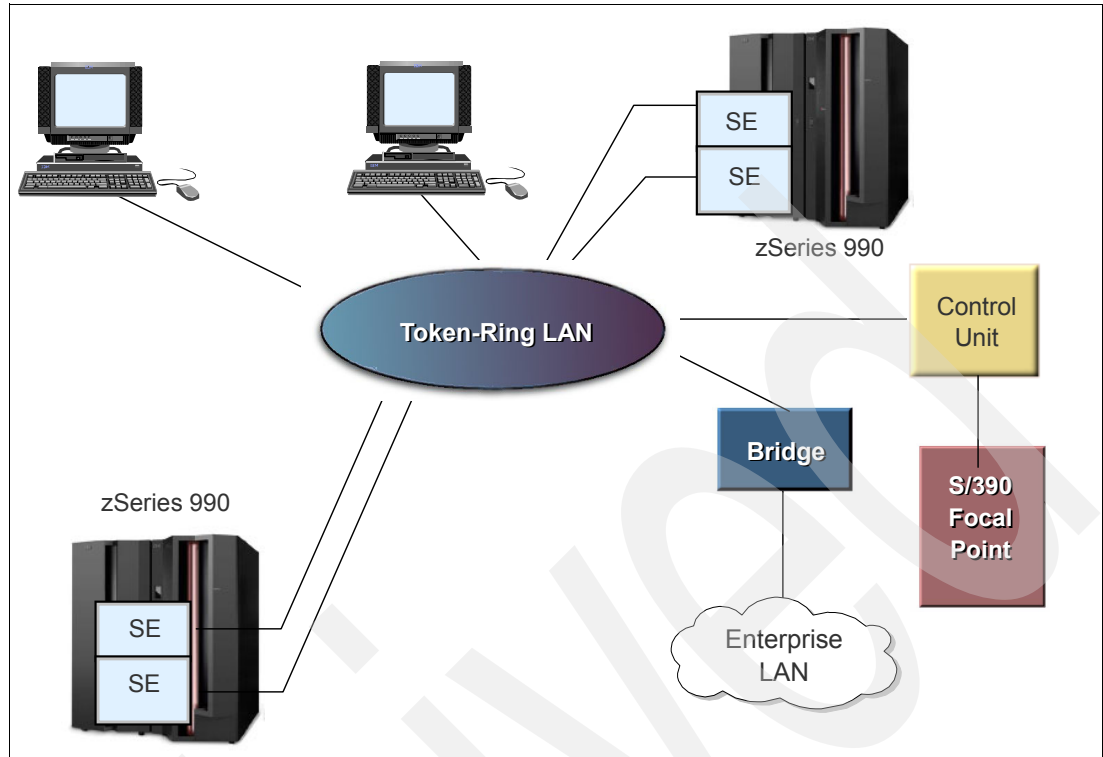


Figure A-4 Token ring only wiring with additional connections

Ethernet only - one-path wiring scenario

This Ethernet only wiring scenario is intended for enterprises that currently have Ethernet installed and do not want token ring wiring introduced into their environment. This wiring scenario requires that a second Ethernet be specified with the Support Elements, and that no token ring feature be ordered on the Hardware Management Console, such that no token ring exists in either the Support Elements or in the Hardware Management Console. (If your system is going to use FC 0075, this console is always equipped with a token ring-Ethernet communication capability. The token ring wrap plug shipped with the system must be installed in the Hardware Management Console token ring adapter to enable the console to operate properly without using token ring.)

The Ethernet features assume the use of 10/100 Mbit Ethernet facilities, requiring the use of CAT-5 Ethernet cabling.

Since the Support Element Ethernet only feature includes two Ethernet adapters, there will be two Ethernet connections available. For this scenario, only the Ethernet cable connected to the Ethernet in the first (top) PCMCIA slot of the Support Elements will be used.

The three communication protocols (SNA, TCP/IP, and NetBios) used in Support Element-to-Hardware Management Console communication are defined for both adapters in the PCMCIA slots of the Support Elements.

It will be necessary to connect the "top" Ethernet adapter cable to a customer-supplied local hub capable of 10/100 Mbit Ethernet rates. It will be necessary to connect the Ethernet from the Hardware Management Console to either the same hub as the Support Elements, or to a hub that connects to the Support Element hub.

Figure A-5 on page 241 gives an overview of the Ethernet only - one path wiring scenario.

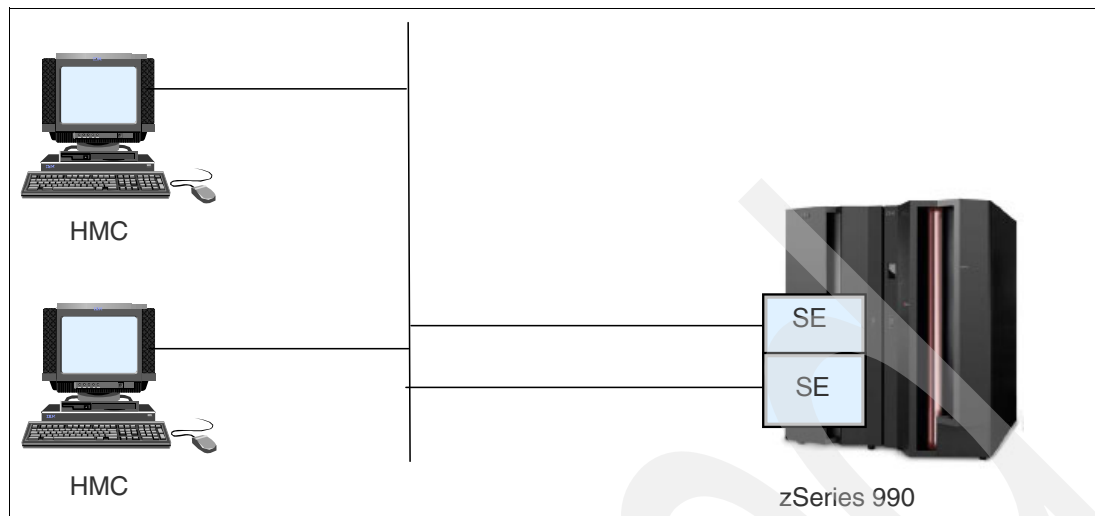


Figure A-5 Ethernet only - one path wiring scenario

Additional connections to the Ethernet LAN

Additional connections to the Ethernet LAN may be made to expand the connectivity beyond the local Hardware Management Console and Support Element (see Figure A-6).

If connections to previous generations of IBM Enterprise Server systems are desired, they may be connected using Ethernet-to-token ring bridges.

If connection to the enterprise LAN is desired, it is recommended that an Ethernet bridge or router be installed to isolate the Hardware Management Console and Support Element from other systems.

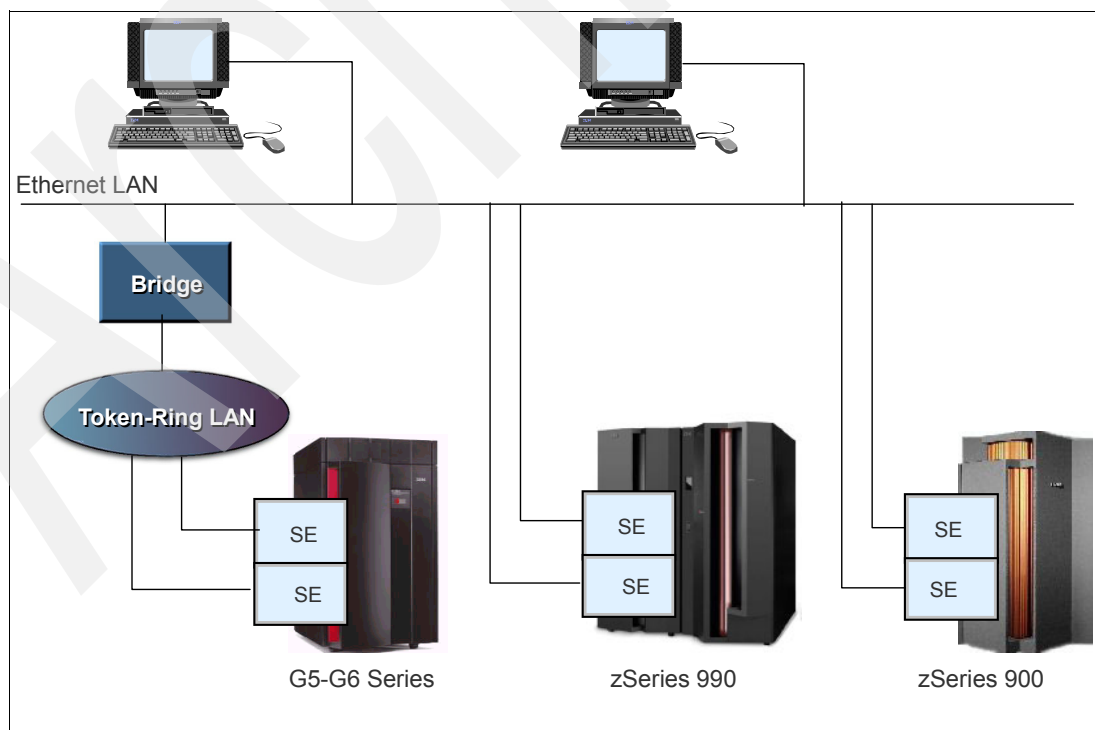


Figure A-6 Ethernet only - one-path wiring scenario with additional connections

Ethernet only - two-path wiring scenario

The Ethernet only - two-path wiring scenario is not applicable to systems with FC 0075. This scenario is included for those who may be reusing a previous Hardware Management Console.

This Ethernet only wiring scenario, shown in Figure A-7, is also intended for enterprises that currently have Ethernet wiring and do not want token ring wiring introduced into their environment. The two-path scenario is included to provide the possibility of a second, separate, and redundant path to the Support Elements.

This wiring scenario requires that FC 3063 (a second Ethernet) be specified with the Support Elements, and that the token ring feature not be ordered with the Hardware Management Console. The Ethernet features assume the use of 10/100 Mb Ethernet facilities, requiring the use of CAT-5 Ethernet cabling. Since the Support Element Ethernet only feature includes two Ethernet adapters, there are two Ethernet connections available. For this scenario, both the Ethernet cables will be used.

The three communication protocols (SNA, TCP/IP and NetBios) used in Support Element to Hardware Management Console communication are defined for both adapters in the PCMCIA slots of the Support Elements.

It is necessary to connect the “top” Ethernet adapter cable to a customer-supplied local hub capable of 10/100 Mb Ethernet rates. It is necessary to connect the Ethernet from at least one local Hardware Management Console to the same hub as the Support Elements.

It is necessary to connect the “bottom” Ethernet adapter cable to a customer-supplied local hub capable of 10/100 Mb Ethernet rates. This second adapter has to be assigned an address on a separate subnet from the first adapter. Any Hardware Management Consoles attached to either LAN is able to automatically discover the Support Elements, assuming that the LAN network allows NetBios to flow between the devices (allowing NetBios to flow is a requirement for local HMC-to-SE communication, but it is not a requirement for remote HMC-to-SE communication).

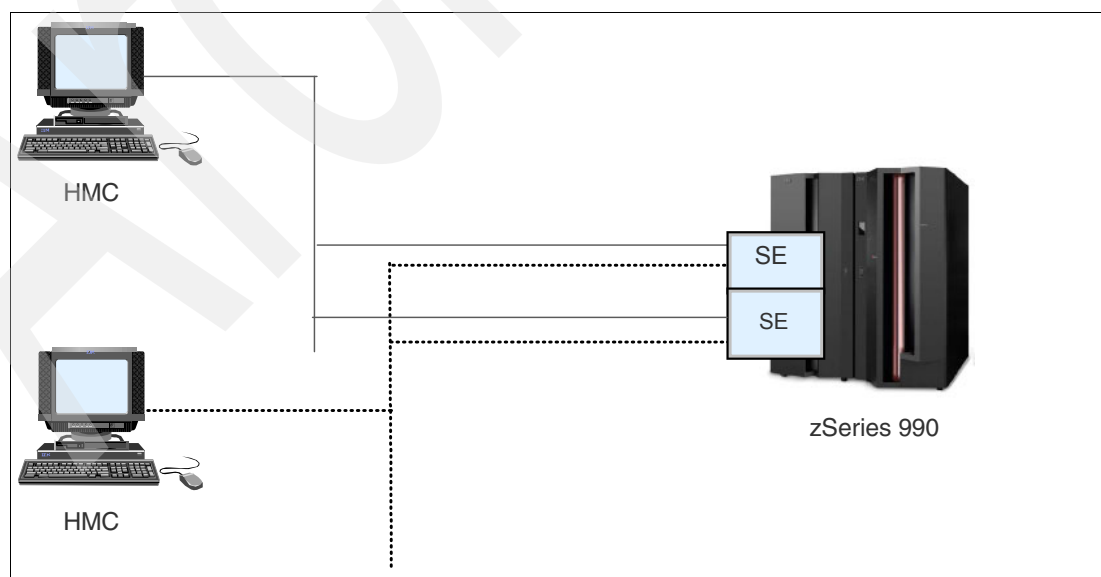


Figure A-7 Ethernet only - two-path wiring scenario

Additional connections to the Ethernet LAN

These may be made to expand the connectivity beyond the local HMC and SE (not applicable to systems with FC 0075).

This scenario is shown in Figure A-8. If connections to previous generations of Enterprise Server systems are desired, they may be connected using Ethernet-to-token ring bridges.

If connection to the enterprise LAN is desired, it is recommended that an Ethernet bridge or router be installed to isolate the HMC and SE from other systems.

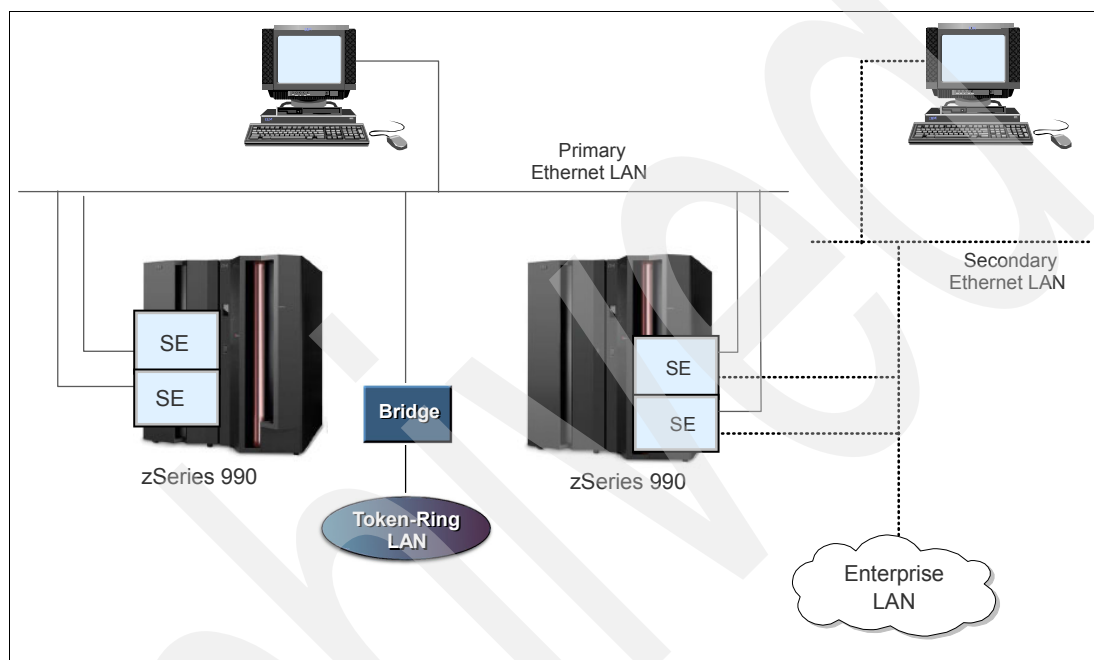


Figure A-8 Ethernet only - two-path wiring scenario with additional connections

Token ring and Ethernet wiring scenario

The token ring and Ethernet wiring scenario, shown in Figure A-9 on page 244, is intended for enterprises that have both token ring wiring and Ethernet wiring requirements. This scenario is included to provide the possibility of controlling the Support Elements from both a token ring Hardware Management Console and an Ethernet Hardware Management Console at the same time.

This wiring scenario is supported by the default set of adapters available with the Support Elements and the token ring adapter on the Hardware Management Console. The token ring adapter assumes the use of a 16 Mb token ring facility. The Ethernet features assume the use of 10/100 Mb Ethernet facilities, requiring the use of CAT-5 Ethernet cabling.

For this scenario, the three communication protocols (SNA, TCP/IP, and NetBios) used in Support Element-to-Hardware Management Console communication are defined for both the token ring and the Ethernet adapters of the Support Elements. The token ring wiring is connected using the MAU as described in "Token ring only wiring scenario" on page 238. The Ethernet wiring is connected from the Support Elements to a customer-supplied local hub capable of 10/100 Mb Ethernet rates.

The Hardware Management Consoles attached to either LAN will be able to automatically discover the Support Elements, assuming that the LAN network allows NetBios to flow

between the devices. The Ethernet adapter will have to be assigned an address on a separate subnet from the token ring adapter.

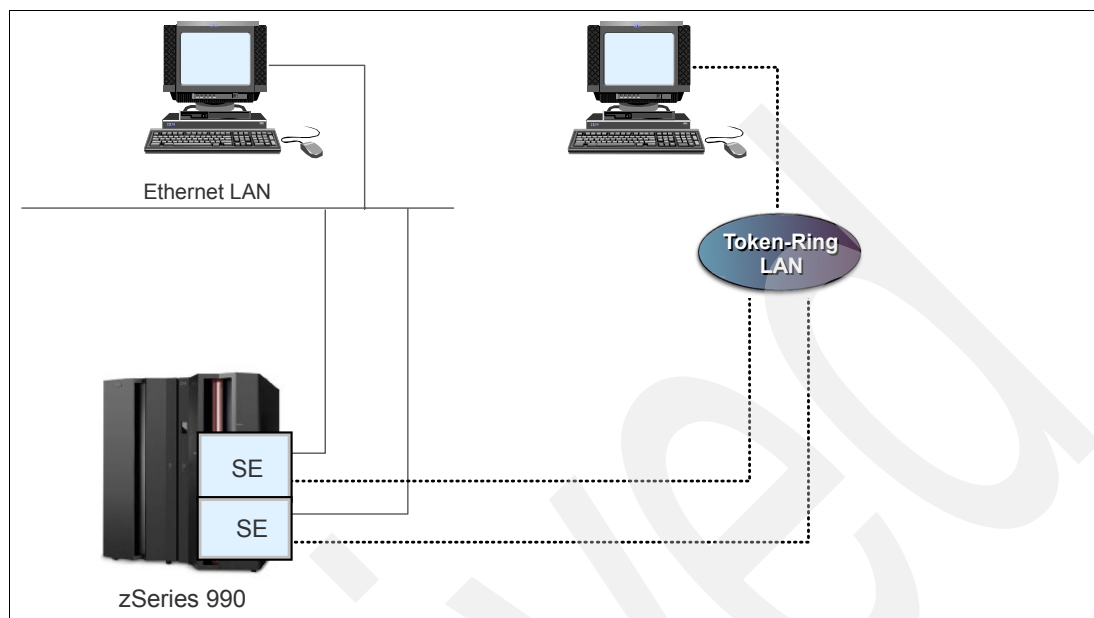


Figure A-9 Token ring and Ethernet wiring scenario

Additional connections to the Token Ring LAN

These connections, shown in Figure A-10 on page 245, may be made to expand the connectivity beyond the local Hardware Management Console and Support Elements. This would be done as in the “Token ring only wiring scenario” on page 238.

Additional connections to the Ethernet LAN may be made to expand the connectivity beyond the local Hardware Management Console and Support Elements. This would be done as in the “Ethernet only - one-path wiring scenario” on page 240.

If connections to previous generations of IBM Enterprise Server systems are desired, they may be connected using the token ring LAN.

If connection to the enterprise LAN is desired, it is recommended that an Ethernet bridge or router be installed to isolate the Hardware Management Console and Support Elements from other systems.

Remote operations

The ability to monitor or control a system from a central or remote location creates a powerful tool for problem determination and diagnosis and operations assistance. This remote capability can save time and money and increase the productivity of support staff. Technical expertise can be centralized, reducing the need for highly skilled personnel at remote locations.

There are several options available for controlling systems from a remote location:

- ▶ Hardware Management Console
- ▶ Web browser
- ▶ Remote control program management product

Choosing the best option involves understanding your remote control needs and use patterns. Figure A-10 shows an example configuration for each option.

IBM uses remote control program product facilities to assist in problem determination, and to provide operational assistance as required. IBM also uses the SDLC or TCP/IP asynchronous connection facilities to transmit service data to and from the IBM Service Support System, to gather error data, and to receive fixes. A remote operation configuration is shown in Figure A-10.

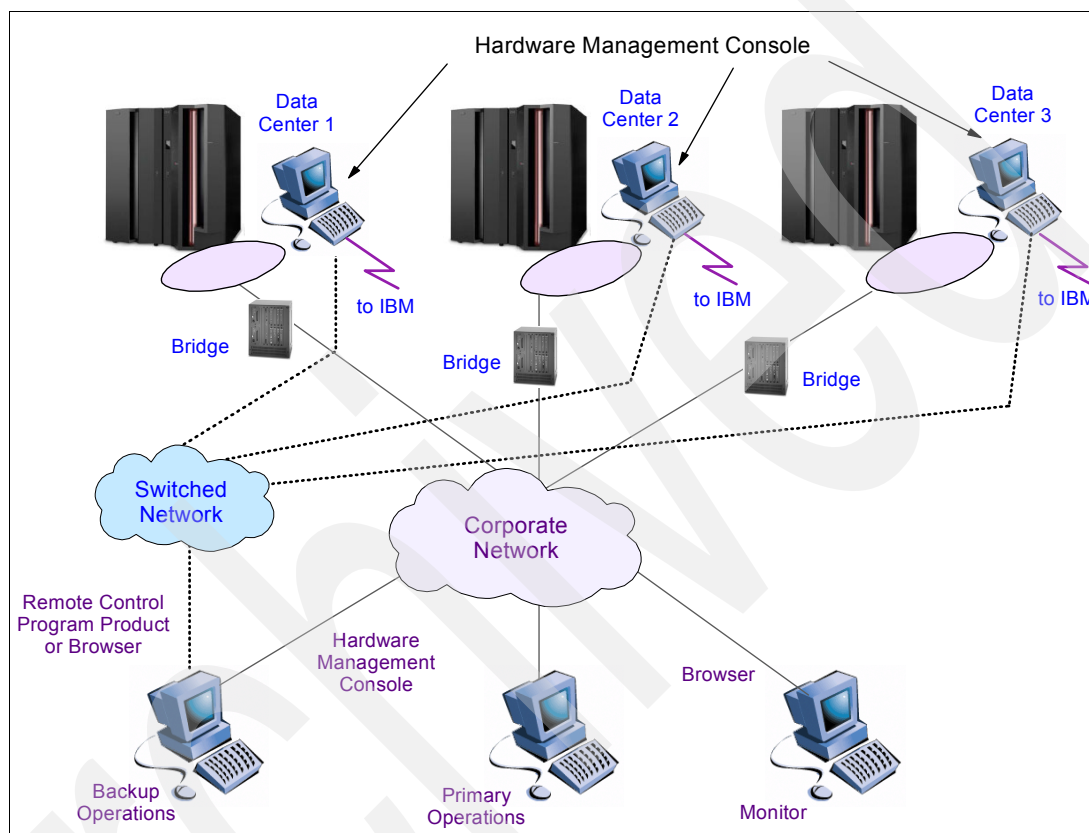


Figure A-10 Remote operation configuration

Support Element

The zSeries 990 is supplied with a pair of integrated ThinkPad Support Elements (SE). One is always active, while the other is strictly an alternate. Power for the support elements is supplied by the processor frame, and there are no additional power requirements.

Each SE comes with two communication adapters included. There are two configuration options: a Dual Ethernet SE or a Token ring/Ethernet SE. Neither is the default; you must select one or the other. Regardless of which pair of adapters you choose, both SEs will be configured the same.

If you order the Token ring/Ethernet default combination, a Multistation Access Unit (MAU) is required to operate the token ring LAN. The MAU is installed in the A frame and requires no power.

z990 HMC enhancements

Here we discuss z990 HMC enhancements.

z990 HMC Integrated 3270 Console

For many users, particularly those new to zSeries and z/VM, the requirement to invest in rather expensive console hardware such as the 2074 just have a z/VM console has been seen as overkill. Thus, a solution for those users has been to lower the cost of acquisition by offering an alternative solution. The integrated 3270 Console support meets that requirement.

On the HMC workspace, there is now an icon to open this 3270 window. It is designed for production system use. One HMC at a time can use the function; however, there is support to switch the function from one HMC to another. With the function, there is also a highly customizable keyboard mapping support, to alleviate the ASCII-to-EBCDIC mapping. This function uses the SCLP hardware interface to connect to the operating system, and support for the function is planned for z/VM 4.4.

The Integrated 3270 Console support is also available on G5/G6 systems in Driver 26, and on z800/z900 in Driver 3G.

z990 HMC Integrated ASCII Console support

As for the 3270 integrated console support, a similar requirement exists for customers wishing to run Linux in a logical partition on a zSeries machine. The Integrated ASCII Console support meets that requirement. One HMC at the time can use the function to have a Linux terminal running, and there is support for switching the function from one HMC to another.

A code drop to Open Source for inclusion in the Linux kernel is planned for the near future. The Integrated ASCII Console support is also available on G5/G6 systems in Driver 26 and on z800/z900 in Driver 3G.

Optional “strict” password rules supported

To implement a more secure interface, a user can now, optionally, switch on a stricter password rule set (the support is implemented as a check box in the user administration function). With this function, there are also messages prompting for a new password when the current password expires.

Enhanced logging facilities

With the more powerful processors, it can be expected that the amount of messages and logs generated will be larger. To meet this requirement, the log files for both the HMC and the SE has been expanded from 1.4 MB to 10 MB. There is also a usability change in handling the SE log display; the IBM CE can now utilize user-defined names for the SE log file reader.

Increased “Console tasks performed” log

The “Console tasks performed” log can now contain the last 500 actions performed.

Customizable console data mirroring

Many of our large customers have a high availability requirement for their systems, including the HMCs. To ease the “duplication” of console customization, IBM has implemented a “data mirroring” function for the HMC setup. A user can now customize an HMC, and then associate other HMCs in the configuration to the same customized data.

As a result, all associated HMCs will be started with the same data. (Earlier, that function could only be achieved by copying the customers data to a diskette, and carrying that diskette from one HMC to another.)

Minor changes to “Operating System Messages”

Based on customer requirements, minor changes have been made to the “Operating System Messages” screens:

- ▶ The command line appears on the first page.
- ▶ The Send and Receive push buttons have been removed.
- ▶ There is now a check box for indicating that the command typed is a reply to a “Priority Message”.

SNA Operations Management for Operations Automation

With the industry move to TCP/IP networks, Systems Network Architecture (SNA) Operations Management commands will no longer be supported on z990 servers. These commands were previously used by the System Automation for OS/390 product, as well as NetView®. The recommendation is to now use the Simple Network Management Protocol (SNMP) Application Programming Interfaces (APIs) for automation needs.

- ▶ If the System Automation for OS/390 product is used, it must now be at Version 2.2 or later. This will allow you to define an automation policy for SNMP APIs, rather than a policy for SNA Operations Management commands.
- ▶ If the SNA Operations Management commands on NetView are used directly, now an SNMP agent and the SNMP APIs for systems automation management should be used.

For detailed information on the SNMP APIs commands and environment requirements, see *IBM @server zSeries Application Programming Interfaces*, SB10-7030. For more information on the SNA Operations Management command support that is not offered on z990, see *Managing Your Processors*, GC38-0452. Both publications are available on IBM Resource Link.

Archived

Fiber optic cabling services

In order to address the complexities and changes over time to different type of cables/connectors and standards, IBM Networking Services provides a comprehensive set of services for your products and enterprises. This service helps you gain an Information Technology (IT) advantage by providing you with the tools you need to gain market share in this fast-paced e-business economy. It relieves you of the stress and complexity of selecting the appropriate connectors and cables to support your servers, devices, LANs, and Storage Area Networks.

When integrating a zSeries server into a data center, an IBM Installation Planning Representative (IPR) provides planning assistance to customers for equipment power, cooling, and the physical placement of the zSeries server.

However, fiber-optic cable planning and connection of zSeries server channels to I/O equipment, Coupling Facilities, networks, and other servers is the responsibility of the customer.

Customers with the resources and personnel to plan and implement their own connectivity, or those with less complex system configurations, can consult the following manuals to help them determine and order the required fiber optic cabling:

- ▶ *IBM @server zSeries 990 Installation Manual for Physical Planning*, GC28-6824
- ▶ *IBM @server zSeries 900 Installation Manual for Physical Planning*, 2064-IMPP
- ▶ *IBM @server zSeries 800 Installation Manual for Physical Planning*, 2066-IMPP
- ▶ *S/390 Installation Manual - Physical Planning Parallel Enterprise Server - Generation 5 Parallel Enterprise Server - Generation 6*, GC22-7106
- ▶ *Planning for Fiber Optic Links (ESCON, FICON, Coupling Links, and Open System Adapters)*, GA23-0367

These manuals are available on the IBM Resource Link Web site:

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

Customers, especially those with complex system integration requirements, may request connectivity assistance from IBM Global Services.

Fiber optic cabling services from IBM

As mentioned, fiber optic cables, cable planning, labeling, and installation are all customer responsibilities for new installations and upgrades. Fiber optic conversion kits and Mode Conditioning Patch (MCP) cables are not orderable as features on z990.

To better serve the cabling needs of z800, z900, and z990 customers, IBM Networking Services has enhanced their fiber optic cabling services to better match your requirements, and is introducing new services. IBM Networking Integration and Deployment Services for zSeries fiber cabling and for enterprise fiber cabling help to ensure IBM has a comprehensive set of services for all customers, from product-level to enterprise-level services geared for today and tomorrow. These services take into consideration the requirements for all of the protocols/media types supported on zSeries (for example, ESCON, FICON, Coupling Links, OSA), whether the focus is the data center, the storage area network (SAN), local area network (LAN), or the end-to-end enterprise.

There are three options to provide individual fiber optic cables (jumper cables, conversion kits, or MCP cables) for connecting to z800, z900, or z990.

- ▶ **Option 1: Fiber optic jumper cabling package (available today for z800 and z900)**
IBM does the detailed planning. This option includes planning, new cables, installation, and documentation. An analysis of the zSeries channel configuration, I/O devices, and any existing fiber optic cabling is required to determine the appropriate fiber optic cables.
- ▶ **Option 2: Fiber optic jumper cable migration and reuse for a zSeries upgrade (new option)**
This option includes planning, reuse of existing cables, and documentation. IBM organizes the existing fiber optic cables based upon the new z990 connection details. Relabeling, rerouting, and reconnection to the appropriate z990 channels is performed. New cables are not offered as a part of this option.
- ▶ **Option 3: Fiber optic jumper cables and installation (new option)**
The customer tells IBM what they need, but the customer does the detailed planning. The service includes new cables, installation, and documentation. Planning and providing the list of required cables are customer's responsibility.

Options 1 and 2 can be combined within one contract to provide complete upgrade coverage.

Under the Enterprise Fiber Cabling Services umbrella, there are two options to provide IBM Fiber Transport System (FTS) trunking commodities (fiber optic trunk cables, fiber harnesses, and panel-mount boxes) for connecting to the z800, z900, and z990:

- ▶ **Option 1: zSeries fiber optic trunk cabling package (new option)**
IBM reduces the cable clutter under the floor. An analysis of the zSeries (z800, z900, and z990) channel configuration and any existing fiber optic cabling is performed to determine the required FTS fiber optic trunking commodities (trunk cables, harnesses, and panel-mount boxes). This option includes zSeries planning, FTS fiber optic trunking commodities, installation, and documentation. This option does *not* include enterprise-level planning.
- ▶ **Option 2: Enterprise fiber cabling services**
IBM organizes the entire enterprise. This option includes enterprise planning, new cables, fiber optic trunking commodities, installation, and documentation. This is the most comprehensive set of services.

Under the zSeries Fiber Cabling Services umbrella there are two options to provide individual fiber optic cables (jumper cables, conversion kits, MCP cables) for connecting to z800, z900, and z990:

- ▶ Option 1: Fiber optic jumper cabling package (available today for z800 and z900)
IBM does the detailed planning. This option includes planning, new cables, installation, and documentation. An analysis of the zSeries channel configuration, I/O devices, and any existing fiber optic cabling is required to determine the appropriate fiber optic cables.
- ▶ Option 2: Fiber optic jumper cable migration and reuse for a zSeries upgrade (new option)
This option includes planning, reuse of existing cables, and documentation. IBM organizes the existing fiber optic cables based upon the new z990 connection details. Relabeling, rerouting, and reconnection to the appropriate z990 channels is performed. New cables are not offered as a part of this option.

Tip: Additional information is documented in the IBM Redbook *IBM @server zSeries Connectivity Handbook*, SG24-5444.

Summary

Each enterprise fiber cabling services design is unique in that it is based on physical room characteristics and equipment placement preferences. It is a long-term “connectivity solution” that provides an organized network of cabling options for future equipment reconfigurations and additions.

The zSeries fiber cabling services and enterprise fiber cabling services, offered by IBM Networking Services, are designed to help you keep pace and provide you with the optimum reliability, availability, and serviceability, as well as the scalability you need to grow your system(s).

Archived

Glossary

active configuration. In an ESCON environment, the ESCON Director configuration determined by the status of the current set of connectivity attributes. Contrast with *saved configuration*.

allowed. In an ESCON Director, the attribute that, when set, establishes dynamic connectivity capability. Contrast with *prohibited*.

American National Standards Institute (ANSI). An organization consisting of producers, consumers, and general interest groups, which establishes the procedures by which accredited organizations create and maintain voluntary industry standards in the United States.

ANSI. See *American National Standards Institute*.

APAR. See *authorized program analysis report*.

authorized program analysis report (APAR). A report of a problem caused by a suspected defect in a current, unaltered release of a program.

basic mode. A S/390 central processing mode that does not use logical partitioning. Contrast with logically partitioned (LPAR) mode.

blocked. In an ESCON Director, the attribute that, when set, removes the communication capability of a specific port. Contrast with *unblocked*.

CBY. Mnemonic for an ESCON channel attached to an IBM 9034 convertor. The 9034 converts from ESCON CBY signals to parallel channel interface (OEMI) communication operating in byte multiplex mode (Bus and Tag). Contrast with *CVC*.

chained. In an ESCON environment, pertaining to the physical attachment of two ESCON Directors (ESCDs) to each other.

channel path (CHP). A single interface between a central processor and one or more control units along which signals and data can be sent to perform I/O requests.

channel path identifier (CHPID). In a Channel Subsystem, a value assigned to each installed channel path of the system that uniquely identifies that path to the system.

Channel Subsystem (CSS). Relieves the processor of direct I/O communication tasks, and performs path management functions. Uses a collection of subchannels to direct a channel to control the flow of information between I/O devices and main storage.

channel. (1) A processor system element that controls one channel path, whose mode of operation depends on the type of hardware to which it is attached. In a Channel Subsystem, each channel controls an I/O interface between the channel control element and the logically attached control units. (2) In the ESA/390 architecture, the part of a Channel Subsystem that manages a single I/O interface between a Channel Subsystem and a set of controllers (control units).

channel-attached. (1) Pertaining to attachment of devices directly by data channels (I/O channels) to a computer. (2) Pertaining to devices attached to a controlling unit by cables rather than by telecommunication lines.

CHPID. Channel path identifier.

cladding. In an optical cable, the region of low refractive index surrounding the core. See also *core* and *optical fiber*.

CNC. Mnemonic for an ESCON channel used to communicate to an ESCON-capable device.

configuration matrix. In an ESCON environment, an array of connectivity attributes that appear as rows and columns on a display device and can be used to determine or change active and saved configurations.

connected. In an ESCON Director, the attribute that, when set, establishes a dedicated connection between two ESCON ports. Contrast with *disconnected*.

connection. In an ESCON Director, an association established between two ports that provides a physical communication path between them.

connectivity attribute. In an ESCON Director, the characteristic that determines a particular element of a port's status. See *allowed*, *blocked*, *connected*, *disconnected*, *prohibited*, and *unblocked*.

control unit. A hardware unit that controls the reading, writing, or displaying of data at one or more input/output units.

core. (1) In an optical cable, the central region of an optical fiber through which light is transmitted. (2) In an optical cable, the central region of an optical fiber that has an index of refraction greater than the surrounding cladding material. See also *cladding* and *optical fiber*.

coupler. In an ESCON environment, link hardware used to join optical fiber connectors of the same type. Contrast with *adapter*.

CPC. Central Processor Complex

CTC. (1) Channel-to-channel. (2) Mnemonic for an ESCON channel attached to another ESCON channel.

CVC. Mnemonic for an ESCON channel attached to an IBM 9034 convertor. The 9034 converts from ESCON CVC signals to parallel channel interface (OEMI) communication operating in block multiplex mode (Bus and Tag). Contrast with *CBY*.

DDM. See *disk drive module*.

dedicated connection. In an ESCON Director, a connection between two ports that is not affected by information contained in the transmission frames. This connection, which restricts those ports from communicating with any other port, can be established or removed only as a result of actions performed by a host control program or at the ESCD console. Contrast with *dynamic connection*. **Note:** The two links having a dedicated connection appear as one continuous link.

default. Pertaining to an attribute, value, or option that is assumed when none is explicitly specified.

destination. Any point or location, such as a node, station, or a particular terminal, to which information is to be sent.

device address. In the ESA/390 architecture and the z/Architecture, the field of an ESCON or FICON (FC mode) device-level frame that selects a specific device on a control-unit image.

device number. (1) In the ESA/390 architecture and the z/Architecture, a four-hexadecimal-character identifier, for example 19A0, that you associate with a device to facilitate communication between the program and the host operator. (2) The device number that you associate with a subchannel that uniquely identifies an I/O device.

device. A mechanical, electrical, or electronic contrivance with a specific purpose.

direct access storage device (DASD). A mass storage medium on which a computer stores data.

disconnected. In an ESCON Director, the attribute that, when set, removes a dedicated connection. Contrast with *connected*.

disk drive module (DDM). A disk storage medium that you use for any host data that is stored within a disk subsystem.

Disk. A physical or logical storage media on which a computer stores data (is also sometimes referred to as a magnetic disk).

distribution panel. (1) In an ESCON or FICON environment, a panel that provides a central location for the attachment of trunk and jumper cables and can be mounted in a rack, wiring closet, or on a wall.

duplex connector. In an ESCON environment, an optical fiber component that terminates both jumper cable fibers in one housing and provides physical keying for attachment to a duplex receptacle.

duplex receptacle. In an ESCON environment, a fixed or stationary optical fiber component that provides a keyed attachment method for a duplex connector.

duplex. Pertaining to communication in which data or control information can be sent and received at the same time. Contrast with *half duplex*.

dynamic connection. In an ESCON Director, a connection between two ports, established or removed by the ESCD and that, when active, appears as one continuous link. The duration of the connection depends on the protocol defined for the frames transmitted through the ports and on the state of the ports. Contrast with *dedicated connection*.

dynamic connectivity. In an ESCON Director, the capability that allows connections to be established and removed at any time.

Dynamic I/O Reconfiguration. A S/390 function that allows I/O configuration changes to be made non-disruptively to the current operating I/O configuration.

EMIF. See *ESCON Multiple Image Facility*.

Enterprise System Connection (ESCON). (1) An ESA/390 computer peripheral interface. The I/O interface uses ESA/390 logical protocols over a serial interface that configures attached units to a communication fabric. (2) A set of IBM products and services that provide a dynamically connected environment within an enterprise.

Enterprise Systems Architecture/390® (ESA/390). An IBM architecture for mainframe computers and peripherals. Processors that follow this architecture include the S/390 Server family of processors.

ESA/390. See *Enterprise Systems Architecture/390*.

ESCD console. The ESCON Director display and keyboard device used to perform operator and service tasks at the ESCD.

ESCD. Enterprise Systems Connection (ESCON) Director.

ESCON channel. A channel having an Enterprise Systems Connection channel-to-control-unit I/O interface that uses optical cables as a transmission medium. May operate in CBY, CNC, CTC, or CVC mode. Contrast with *parallel channel*.

ESCON Director. An I/O interface switch that provides the interconnection capability of multiple ESCON interfaces (or FICON FCV (9032-5) in a distributed-star topology.

ESCON Multiple Image Facility (EMIF). In the ESA/390 architecture, a function that allows LPARs to share an ESCON channel path (and other channel types) by providing each LPAR with its own channel-subsystem image.

ESCON. See *Enterprise System Connection*.

FCS. See *Fibre Channel standard*.

FCTC. FICON Channel-to-Channel

fiber optic cable. See *optical cable*.

fiber optics. The branch of optical technology concerned with the transmission of radiant power through fibers made of transparent materials, such as glass, fused silica, and plastic.

Note: Telecommunication applications of fiber optics use optical fibers. Either a single discrete fiber or a non-spatially aligned fiber bundle can be used for each information channel. Such fibers are often called optical fibers to differentiate them from fibers used in non-communication applications.

fiber. See *optical fiber*.

Fibre Channel standard. An ANSI standard for a computer peripheral interface. The I/O interface defines a protocol for communication over a serial interface that configures attached units to a communication fabric. The protocol has four layers. The lower of the four layers defines the physical media and interface, the upper of the four layers defines one or more logical protocols (for example, FCP for SCSI command protocols and FC-SB-2 for FICON for ESA/390). Refer to ANSI X3.230.1999x.

FICON channel. A channel having a Fibre Channel channel-to-control-unit I/O interface that uses optical cables as a transmission medium. The FICON channel may operate in (1) FC mode (FICON native mode - FC-SB-2/3), (2) FCV mode (FICON conversion mode to a IBM 9032-5), or (3) FCP mode (FICON channel operating in "open mode", which is FC-FCP).

FICON. (1) An ESA/390 and z/Architecture computer peripheral interface. The I/O interface uses ESA/390 and z/Architecture logical protocols over a FICON serial interface that configures attached units to a FICON communication fabric. (2) An FC4 adopted standard that defines an effective mechanism for the export of the SBCON command protocol via Fibre Channels.

field replaceable unit (FRU). An assembly that is replaced in its entirety when any one of its required components fails.

FRU. See *field replaceable unit*.

Gigabit (Gb). Usually used to refer to a data rate, the number of gigabits being transferred in one second.

Gigabyte (GB). Usually used to refer to an amount of storage space or size. One gigabyte is 10^9 , or 1,073,741,824 bytes.

half duplex. In data communication, pertaining to transmission in only one direction at a time. Contrast with *duplex*.

hard disk drive. (1) A storage media within a storage server used to maintain information that the storage server requires. (2) A mass storage medium for computers that is typically available as a fixed disk or a removable cartridge.

HDA. Head and disk assembly.

HDD. See *hard disk drive*.

head and disk assembly. The portion of an HDD associated with the medium and the read/write head.

I/O configuration. The collection of channel paths, control units, and I/O devices that attaches to the processor. This may also include channel switches (for example, an ESCON Director).

I/O. See *input/output*.

ID. See *identifier*.

Identifier. A unique name or address that identifies things such as programs, devices, or systems.

initial program load (IPL). (1) The initialization procedure that causes an operating system to commence operation. (2) The process by which a configuration image is loaded into storage at the beginning of a work day or after a system malfunction. (3) The process of loading system programs and preparing a system to run jobs.

input/output (I/O). (1) Pertaining to a device whose parts can perform an input process and an output process at the same time. (2) Pertaining to a functional unit or channel involved in an input process, output process, or both, concurrently or not, and to the data involved in such a process. (3) Pertaining to input, output, or both.

input/output configuration data set (IOCDs). The data set in the S/390 processor (in the support element) that contains an I/O configuration definition built by the input/output configuration program (IOCP).

input/output configuration program (IOCP). A S/390 program that defines the channels, I/O devices, paths to the I/O devices, and the addresses of the I/O devices to a system. The output is normally written to a S/390 IOCDs.

interface. (1) A shared boundary between two functional units, defined by functional characteristics, signal characteristics, or other characteristics as appropriate. The concept includes the specification of the connection of two devices having different functions. (2) Hardware, software, or both, that links systems, programs, or devices.

IOCDs. See *Input/Output configuration data set*.

IOCP. See *Input/Output configuration control program*.

IODF. The data set that contains the S/390 I/O configuration definition file produced during the defining of the S/390 I/O configuration by HCD. Used as a source for IPL, IOCP, and Dynamic I/O Reconfiguration.

IPL. See *initial program load*.

jumper cable. In an ESCON and FICON environment, an optical cable having two conductors that provides physical attachment between a channel and a distribution panel or an ESCON Director port or a control unit/devices, or between an ESCON Director port and a distribution panel or a control unit/device, or between a control unit/device and a distribution panel. Contrast with *trunk cable*.

LAN. See *local area network*.

laser. A device that produces optical radiation using a population inversion to provide *light amplification by stimulated emission of radiation* and (generally) an optical resonant cavity to provide positive feedback. Laser radiation can be highly coherent temporally, or spatially, or both.

LC connector. An optical fibre cable duplex connector that terminates both jumper cable fibres into one housing and provides physical keying for attachment to an LC duplex receptacle. For technical details, see the NCITS - American National Standard for Information Technology - Fibre Channel Standards document FC-P1.

LCSS. See *Logical Channel Subsystem*.

LCU. See *Logical Control Unit*.

LED. See *light emitting diode*.

licensed internal code (LIC). Microcode that IBM does not sell as part of a machine, but instead licenses it to the customer. LIC is implemented in a part of storage that is not addressable by user programs. Some IBM products use it to implement functions as an alternate to hardware circuitry.

light-emitting diode (LED). A semiconductor chip that gives off visible or infrared light when activated. Contrast *Laser*.

link address. On an ESCON or a FICON interface, the portion of a source or destination address in a frame that ESCON or FICON uses to route a frame through an ESCON or FICON director. ESCON and FICON associates the link address with a specific switch port that is on the ESCON or FICON director. **Note:** For ESCON, there is a one-byte link address. For FICON, there can be a one-byte or two-byte link address specified. One-byte link address for a FICON non-cascade topology and two-byte link address supports a FICON cascade switch topology.

See also *port address*.

link. (1) In an ESCON or FICON environment, the physical connection and transmission medium used between an optical transmitter and an optical receiver. A link consists of two conductors, one used for sending and the other for receiving, thereby providing a duplex communication path. (2) In an ESCON or FICON I/O interface, the physical connection and transmission medium used between a channel and a control unit, a channel and an ESCON or FICON Director, a control unit and an ESCON or FICON Director, or, at times, between two ESCON Directors or two FICON Directors.

local area network (LAN). A computer network located in a user's premises within a limited geographic area.

Logical Channel Subsystem (LCSS). A defined subset of the CPC hardware (subchannels, channels, and I/O interfaces) that is used to support the operation of a Logical Channel Subsystem. The LCSS relieves the processor of direct I/O communication tasks, and performs path management functions. Uses a collection of subchannels (defined to the LCSS) to direct a channel to control the flow of information between its defined I/O devices and main storage.

logical control unit (LCU). A separately addressable control unit function within a physical control unit. Usually a physical control unit that supports several LCUs. For ESCON, the maximum number of LCUs that can be in a control unit (and addressed from the same ESCON fiber link) is 16; they are addressed from x'0' to x'F'.

logical partition (LPAR). A set of functions that create a programming environment that is defined by the ESA/390 architecture. ESA/390 architecture uses this term when more than one LPAR is established on a processor. An LPAR is conceptually similar to a virtual machine environment, except that the LPAR is a function of the processor. Also, LPAR does not depend on an operating system to create the virtual machine environment.

logical switch number (LSN). A two-digit number used by the I/O Configuration Program (IOCP) to identify a specific ESCON Director.

logically partitioned (LPAR) mode. A central processor mode, available on the Configuration frame when using the PR/SM facility, that allows an operator to allocate processor hardware resources among logical partitions. Contrast with *basic mode*.

LPAR. See *logical partition*.

megabyte (MB). (1) For processor storage, real and virtual storage, and channel volume, 2^{20} or 1 048 576 bytes. (2) For disk storage capacity and communications volumes, 1 000 000 bytes.

MT-RJ. An optical fibre cable duplex connector that terminates both jumper cable fibres into one housing and provides physical keying for attachment to an MT-RJ duplex receptacle. For technical details, see the NCITS - American National Standard for Information Technology - Fibre Channel Standards document FC-PI.

multi-mode optical fiber. A graded-index or step-index optical fiber that allows more than one bound mode to propagate. Contrast with *single-mode optical fiber*.

Multiple Image Facility (EMIF). In the ESA/390 architecture and z/Architecture, a function that allows LPARs to share a channel path by providing each LPAR with its own set of subchannels for accessing a common device.

National Committee for Information Technology Standards. NCITS develops national standards and its technical experts participate on behalf of the United States in the international standards activities of ISO/IEC JTC 1, information technology.

NCITS. See *National Committee for Information Technology Standards*.

ND. See *node descriptor*.

NED. See *node-element descriptor*.

node descriptor. In an ESCON and FICON environment, a node descriptor (ND) is a 32-byte field that describes a node, channel, ESCON Director port, FICON Director port, or a control unit.

node-element descriptor. In an ESCON and FICON environment, a node-element descriptor (NED) is a 32-byte field that describes a node element, such as a DASD (Disk) device.

OEMI. See *original equipment manufacturers information*.

open system. A system whose characteristics comply with standards made available throughout the industry and that therefore can be connected to other systems complying with the same standards.

optical cable assembly. An optical cable that is connector-terminated. Generally, an optical cable that has been terminated by a manufacturer and is ready for installation. See also *jumper cable* and *optical cable*.

optical cable. A fiber, multiple fibers, or a fiber bundle in a structure built to meet optical, mechanical, and environmental specifications. See also *jumper cable*, *optical cable assembly*, and *trunk cable*.

optical fiber connector. A hardware component that transfers optical power between two optical fibers or bundles and is designed to be repeatedly connected and disconnected.

optical fiber. Any filament made of dielectric materials that guides light, regardless of its ability to send signals. See also *fiber optics* and *optical waveguide*.

optical waveguide. (1) A structure capable of guiding optical power. (2) In optical communications, generally a fiber designed to transmit optical signals. See *optical fiber*.

original equipment manufacturers information (OEMI). A reference to an IBM guideline for a computer peripheral interface. More specifically, refers to IBM S/360 and S/370 Channel to Control Unit Original Equipment Manufacture's Information. The interfaces uses ESA/390 logical protocols over an I/O interface that configures attached units in a multi-drop bus environment.

parallel channel. A channel having a System/360™ and System/370™ channel-to-control-unit I/O interface that uses bus and tag cables as a transmission medium. Contrast with *ESCON channel*.

path group. The ESA/390 and z/Architecture term for a set of channel paths that are defined to a controller as being associated with a single S/390 image. The channel paths are in a group state and are online to the host.

path. In a channel or communication network, any route between any two nodes. For ESCON or FICON, this would be the route between the channel and the control unit/device, or sometimes from the operating system control block for the device and the device itself.

path-group identifier. The ESA/390 term for the identifier that uniquely identifies a given LPAR. The path-group identifier is used in communication between the system image program and a device. The identifier associates the path-group with one or more channel paths, thereby defining these paths to the control unit as being associated with the same system image.

PCICC. (IBM's) PCI Cryptographic Coprocessor.

physical channel identifier (PCHID). A value assigned to each physically installed and enabled channel in the CPC that uniquely identifies that channel to the system (for the IBM z990, the assigned PCHID values are between 000 and 6FF).

port address. In an ESCON Director or a FICON Director, an address used to specify port connectivity parameters and to assign link addresses for attached channels and control units. See also *link address*.

port card. In an ESCON or FICON environment, a field-replaceable hardware component that provides the optomechanical attachment method for jumper cables and performs specific device-dependent logic functions.

port name. In an ESCON Director or a FICON Director, a user-defined symbolic name of 24 characters or less that identifies a particular port.

port. (1) An access point for data entry or exit. (2) A receptacle on a device to which a cable for another device is attached. See also *duplex receptacle*.

processor complex. A system configuration that consists of all the machines required for operation, for example, a Processor Unit, a processor controller, a system display, a service support display, and a power and coolant distribution unit.

program temporary fix (PTF). A temporary solution or bypass of a problem diagnosed by IBM in a current unaltered release of a program.

prohibited. In an ESCON Director or FICON Director, the attribute that, when set, removes dynamic connectivity capability. Contrast with *allowed*.

protocol. (1) A set of semantic and syntactic rules that determines the behavior of functional units in achieving communication. (2) In SNA, the meanings of and the sequencing rules for requests and responses used for managing the network, transferring data, and synchronizing the states of network components. (3) A specification for the format and relative timing of information exchanged between communicating parties.

PTF. See *program temporary fix*.

route. The path that an ESCON frame or FICON frame (Fibre Channel frame) takes from a channel through an ESCON Director or FICON Director to a control unit/device.

saved configuration. In an ESCON or FICON environment, a stored set of connectivity attributes whose values determine a configuration that can be used to replace all or part of the ESCON Director's or FICON Director's active configuration. Contrast with active configuration.

SC Connector. An optical fibre cable duplex connector that terminates both jumper cable fibres into one housing and provides physical keying for attachment to an LC duplex receptacle. For technical details, see the NCITS - American National Standard for Information Technology - Fibre Channel Standards document FC-PI.

Self-Timed Interconnect (STI). An interconnect path cable that has one or more conductors that transit information serially between two interconnected units without requiring any clock signals to recover that data. The interface performs clock recovery independently on each serial data stream and uses information in the data stream to determine character boundaries and inter-conductor synchronization.

service element (SE). A dedicated service processing unit used to service a S/390 machine (processor).

Small Computer System Interface (SCSI). (1) An ANSI standard for a logical interface to a computer peripherals and for a computer peripheral interface. The interface uses a SCSI logical protocol over an I/O interface that configures attached targets and initiators in a multi-drop bus topology. (2) A standard hardware interface that enables a variety of peripheral devices to communicate with one another.

spanning channels. MIF spanning channels have the ability to be configured to multiple Channel SubSystems, and be transparently shared by any or all of the configured LPARs without regard to the Logical Channel SubSystem to which the LPAR is configured.

subchannel. A logical function of a Channel Subsystem associated with the management of a single device.

subsystem. (1) A secondary or subordinate system, or programming support, usually capable of operating independently of or asynchronously with a controlling system.

SWCH. In ESCON Manager, the mnemonic used to represent an ESCON Director.

switch. In ESCON Manager, synonym for ESCON Director.

trunk cable. In an ESCON environment, a cable consisting of multiple fiber pairs that do not directly attach to an active device. This cable usually exists between distribution panels and can be located within, or external to, a building. Contrast with *jumper cable*.

unblocked. In an ESCON Director, the attribute that, when set, establishes communication capability for a specific port. Contrast with *blocked*.

unit address. The ESA/390 term for the address associated with a device on a given controller. On ESCON or FICON interfaces, the unit address is the same as the device address. On OEMI interfaces, the unit address specifies a controller and device pair on the interface.

z/Architecture. An IBM architecture for mainframe computers and peripherals. Processors that follow this architecture include the zSeries Server family of processors.

Archived

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this redbook.

IBM Redbooks

For information on ordering these publications, see “How to get IBM Redbooks” on page 263. Note that some of the documents referenced here may be available in softcopy only.

- ▶ *IBM @server zSeries 990 Technical Introduction*, SG24-6863
- ▶ *IBM @server zSeries Connectivity Handbook*, SG24-5444
- ▶ *z/OS Intelligent Resource Director*, SG24-5952

Other publications

These publications are also relevant as further information sources:

- ▶ *Coupling Facility Configuration Options*, GF22-5042, found at:
<http://www.ibm.com/servers/eserver/zseries/pso/>
- ▶ *Enterprise Systems Architecture/390 Principles of Operation*, SA22-7201
- ▶ *GDPS: The e-business Availability Solution*, GF22-5114, found at:
<http://www.ibm.com/servers/eserver/zseries/library/whitepapers/gf225114.html>
- ▶ *IBM @server Hardware Management Console Guide Version 1.8.0*, SC28-6819
- ▶ *IBM @server System Overview*, SA22-1032
- ▶ *IBM @server zSeries 800 Installation Manual for Physical Planning*, 2066-IMPP
- ▶ *IBM @server zSeries 900 Installation Manual for Physical Planning*, 2064-IMPP
- ▶ *IBM @server zSeries 990 Installation Manual for Physical Planning*, GC28-6824
- ▶ *IBM @server zSeries 990 Processor Resource/Systems Manager Planning Guide*, SB10-7036
- ▶ *IBM @server zSeries 990 Stand-Alone Input/Output Configuration Program User's Guide*, SB10-7040
- ▶ *IBM @server zSeries Application Programming Interfaces*, SB10-7030
- ▶ *IBM @server zSeries Capacity Backup User's Guide*, SC28-6810
- ▶ *IBM @server zSeries CCA User Defined Extensions Reference and Guide*, found at:
<http://www.ibm.com/security/cryptocards>
- ▶ *IBM @server zSeries Input/Output Configuration Program User's Guide for ICP IOCP*, SB10-7037
- ▶ *Large Systems Performance Reference*, SC28-1187
- ▶ *Managing Your Processors*, GC38-0452
- ▶ *Planning for Fiber Optic Links (ESCON, FICON, Coupling Links, and Open System Adapters)*, GA23-0367

- ▶ *S/390 Installation Manual - Physical Planning Parallel Enterprise Server - Generation 5*
Parallel Enterprise Server - Generation 6, GC22-7106
- ▶ *System-Managed CF Structure Duplexing*, GM13-0100
- ▶ *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 System Programmer's Guide*, SA22-7520
- ▶ *z/OS Cryptographic Services ICSF Trusted Key Entry Workstation User's Guide*, SA22-7524
- ▶ *z/OS Hardware Configuration Definition Planning*, GA22-7525
- ▶ *z/OS Hardware Configuration Definition: User's Guide*, SC33-7988
- ▶ *z/OS MVS Recovery and Reconfiguration Guide*, SA22-7623
- ▶ *z/OS MVS System Commands*, SA22-7627
- ▶ *z/OS MVS Planning: Workload Management*, SA22-7602
- ▶ *z/OS Planning for Workload License Charges*, SA22-7506

Online resources

These Web sites and URLs are also relevant as further information sources:

- ▶ developerWorks: IBM's resource for developers
<http://www.ibm.com/developerworks>
- ▶ IBM CFSizer coupling facility sizer web tool home page
<http://www.ibm.com/servers/eserver/zseries/cfsizer>
- ▶ IBM developerWorks: Open source
<http://www-124.ibm.com/developerworks/projects/libica>
- ▶ IBM @server zSeries server product line
<http://www.ibm.com/servers/eserver/zseries/>
- ▶ IBM @server zSeries connectivity options
<http://www.ibm.com/servers/eserver/zseries/connectivity>
- ▶ IBM @server zSeries Parallel Sysplex
<http://www.ibm.com/server/eserver/zSeries/psa>
- ▶ IBM @server zSeries Software Pricing Exhibits
<http://www.ibm.com/servers/eserver/zseries/library/swpriceinfo>
- ▶ IBM: FICON Support of Cascaded Directors
http://www.ibm.com/servers/eserver/zseries/connectivity/ficon_cascaded.html
- ▶ IBM Large Systems Performance Reference for zSeries
<http://www.ibm.com/servers/eserver/zseries/lspr>
- ▶ IBM Resource Link
<http://www.ibm.com/servers/resourceLink>

- ▶ IBM: z/OS downloads - Useful technology demos, sample code, tools, and documentation for the z/OS platform
<http://www.ibm.com/eserver/zseries/zos/downloads>
- ▶ Linux for S/390 and zSeries
<http://www10.software.ibm.com/developerworks/opensource/linux390>
<http://www10.software.ibm.com/developerworks/opensource/linux390/index.shtml>
- ▶ Optica Technologies Incorporated
<http://www.opticatech.com>
- ▶ Optica Technologies Incorporated 34600 FXBT Converter
<http://www.opticatech.com/34600.asp>
- ▶ Red Hat — Linux, Embedded Linux and Open Source Solutions
<http://www.redhat.com>
- ▶ SUSE Linux
<http://www.suse.com>
- ▶ System-Managed CF Structure Duplexing
<http://www.ibm.com/servers/eserver/zSeries/library/techpapers/gm130103.html>
<http://www-1.ibm.com/servers/eserver/zseries/library/techpapers/gm130103.html>
- ▶ Turbolinux, Inc.
<http://www.turbolinux.com>

How to get IBM Redbooks

You can search for, view, or download Redbooks, Redpapers, Hints and Tips, draft publications and Additional materials, as well as order hardcopy Redbooks or CD-ROMs, at this Web site:

ibm.com/redbooks

Archived

Index

Numerics

1000BASE-T Ethernet 11
16-port ESCON feature 93
50.0 micron 100–101
62.5 micron 98, 100–101

A

A Frame 33
architecture modes 68

B

BHT 44
book 111
 connectivity 33
 jumper 30
 replacement 33
Branch History Table (BHT) 44

C

cages 33
Capacity Backup 65
Capacity BackUp (CBU) 15–16, 65, 190, 206, 208–209
Capacity Upgrade on Demand (CUoD) 15–16, 55, 189–190
 for I/O 195
 for memory 193
 for processors 191
CBU 15–16, 65, 190, 206
 activation 208
 deactivation 208
 testing 209
CEC cage 3, 8, 24, 33, 75, 80–81
Central Processor (CP) 47
 pool 47
central storage 55
CF
 mode 69
 structure duplexing 170
CFCC 47
 enhanced patch apply 162
CFLEVEL 163
CHA 182
channel spanning 114
channel sparing 95
Channel Subsystem (CSS) 6, 61, 71, 109–110, 179, 182–184
 ID 61, 110
Checksum Offload 103
CHPID 7, 59, 79, 110
CHPID Mapping Tool (CMT) 7, 79, 84, 89, 114, 116
CIU 15–16, 189
 enablement feature 196, 202

 registration 197
CMT 7, 79, 84, 89, 114, 116
compatibility support for z/OS 134
Compression Unit 43
concurrent channel upgrades 71
concurrent conditioning 196
Concurrent Processor Unit 6, 47, 62, 64
 conversions 5, 188
concurrent upgrade 15, 46, 188, 210
Configuration Management 116
Configurator for e-business 115
connectors 92
Control for Plan-Ahead 196
cooling 25
Coupling Facility (CF)
 mode 69
CP 47
 assigned 64
 pool 47
CP Assist for Cryptographic 12
 function 38, 42, 122
CP Cryptographic Assist Facility 43
CPACF 122
CPU
 management 180
 resources 179
Cryptographic 12
 asynchronous function 120
 function support 120
 processors 122
 synchronous function 120
CSS 6, 61, 71, 109
 ID 61, 110
 Image ID 110
 priority 183–184
 priority queueing 179, 182
CUoD 15–16, 55, 189–190
 for memory 193
 for processors 191
CUoDCUoD
 for I/O 195
Customer Initiated Upgrade (CIU) 15–16, 189
 enablement feature 196, 202
 registration 197

D

Data Encryption Standard (DES) 120
DCA 25, 76
DCM 182
DES 120
director port cards 182
disruptive upgrades 217
Distributed Converter Assembly (DCA) 25, 76
dual processor design 42

- Dynamic Add/Delete
 - logical partition name 61, 113
- Dynamic CF Dispatching 168
- Dynamic Channel-path Management 179
- Dynamic CHPID Management 15
- Dynamic Coupling Facility Dispatching 48
- Dynamic I/O configuration 72
- Dynamic ICF Expansion 168
- Dynamic Storage Reconfiguration 71

E

- e-Config 115
- EE 101
- Enterprise Extender (EE) 101
- ESA/390 Architecture Mode 69
- ESA/390 mode 68
- ESA/390 TPF mode 69
- ESCON 9, 84, 182
 - port sparing 72
 - upgrading 72
- ETR 84, 91
 - port 25
- Expanded storage 55
- exploitation support for z/OS 137

F

- FCP 98
 - concurrent patch 99
 - SCSI IPL feature 99
- Feature
 - 16-port ESCON (2323) 93
- Fiber Quick Connect 92, 96
- FICON 182
 - Cascaded Directors 10
 - CTC function 10
- FICON Express 9, 84, 97
 - LX feature 97
 - SX feature 98
- Flexible Channel 9
- frames 33

G

- GDPS 172, 209
 - PPRC 172
 - XRC 176
- glass ceramic substrate 6, 35
- Goal mode 179
- Goal mode WLM policy 182

H

- hardware compression 43
- Hardware Configuration Dialog (HCD) 60–61, 112–113, 115
- Hardware Management Console (HMC) 15, 66, 112, 238
- Hardware System Area (HSA) 56, 61
- HCD 60–61, 112–113, 115
- HiperSockets
 - function 104

- HMC 15, 66, 112, 238
 - Integrated 3270 Console 246
 - Integrated ASCII Console support 246
- HSA 56, 61

I

- I/O
 - connectivity 8
 - definition file 115
 - features cables 92
 - path 182
 - performance 183
 - priority queuing 15
- I/O cage 8, 24, 75
- IBM 9034 84
- IC 164
- IC3 14
- ICB-2 13, 164
 - link 106
- ICB-3 14, 164
 - link 106
- ICB-4 14, 86, 164
 - link 106
- ICF 6, 47–48, 62, 64, 216
- IEEE Floating Point 45
- IFA 49
- IFC 48
- IFL 6, 47, 62, 64, 215
 - assigned 64
- Input/Output Configuration Dataset 115
- Input/Output Configuration Dataset (IOCDs) 61, 113, 115
- Instruction grouping 45
- Integrated Cryptographic Service Facility 123
- Integrated Facility for Applications (IFA) 49
- Integrated Facility for Linux (IFL) 47
- Intelligent Resource Director (IRD) 15
- Intergrated Facility for Applications (IFA) 49
- Intergrated Facility for Linux (IFL) 6, 47, 62, 64, 215
- Internal Battery Feature 34
- Internal Coupling Facility(ICF) 6, 47–48, 62, 64, 216
- Internal Coupling Facility (ICF) 48
- IOCDs 61, 113, 115
- IOCP 60
- IODF 115
- IRD 15
- ISC-3 13, 84, 164
 - Daughter Card 105
 - link 105
 - Mother Card 105

L

- L1 cache 39, 46
- L2 cache 30, 36, 39, 46
- Land Grid Arrays 6, 35
- Large Systems Performance Reference (LSPR) 224
- LCSS 3, 61, 72, 110
 - configuration management 115
 - structure 110

- LICCC 6
- Linux 47, 60
 - Integrated Facilities for Linux 47
 - mode 69
 - storage 69
- Linux on zSeries 19, 147
- logical book structure 39
- Logical Channel Subsystem (LCSS) 3, 61, 72, 110, 115
- Logical Processor 58
- LPAR 182–183
 - cluster 178–179
 - CPU management 15, 179
 - mode 47, 60
 - single storage pool 55
 - weights 180
- LSPR 224
- LSPR workloads for z990 226

M

- managed channels 182
- master key entry 123
- MBA 24, 31, 40, 77
- MCM 35
- MCM technology 6
- MCP 100–101
- memory 6
 - allocation 54
 - cards 27
 - upgrade 27
- Memory Bus Adapter
 - see MBA 24
- Memory Coherent Controller 39
- Message authentication code (MAC) 120
- Message Time Ordering 159
- MIF 111, 113
 - ID 61, 111, 113
- mode conditioner patch (MCP) 100–101
- model downgrades 5
- model number 4
- model upgrade paths 5
- model upgrades 189
- modes of operation 56
- Modular Refrigeration Unit 25
- Motor Drive Assembly 26
- Motor Scroll Assembly 26
- MSU 66
- MSU value 66
- Multiple Image Facility (MIF) 111
- Multiple Logical Channel Subsystems 110

N

- N+1 power supply 25
- nondisruptive upgrades 211, 217

O

- On/Off Capacity Upgrade on Demand (On/Off CoD) 16
- On/Off CoD 189, 202
 - active CP 203

- active ICF 203
- active IFL 203
- active zAAP 203
- enablement feature 202
- Optica Technologies 85
- OSA-2
 - FDDI 12
 - FENET 102
 - Token Ring 85
- OSA-Express 10
 - 1000BASE-T 84, 101
 - ATM 12, 85
 - Fast Ethernet 84, 102
 - GbE 84, 100
 - GbE LX 100
 - GbE SX 100
 - Gigabit Ethernet 11
 - Integrated Console Controller 11, 102
 - Token Ring 84, 103
- OSA-ICC 11, 102

P

- Parallel channel 12, 84
- Parallel Sysplex 13
- Partition Number 111
- PCHID 7, 59, 72, 79, 86, 113, 116
- PCI Cryptographic Accelerator 38, 108, 122
 - feature 12
- PCI Cryptographic Coprocessor 85
- PCI X-Cryptographic Coprocessor 13
- PCICA 12, 35, 85, 108, 122, 125–126, 211
- PCICC 85
- PCIX Cryptographic Coprocessor 38, 108, 122
- PCIXCC 13, 61, 85, 108, 122, 211
 - feature 125
- Performance 17
- performance 183
- Physical Channel ID 113
- physical channel path identifiers 116
- Plan-Ahead 191, 196
- Plan-ahead Concurrent Conditioning for I/O 196
- Planned upgrades 189
- Power 25
- Power and cooling requirement 232
- PR/SM 4, 54, 57
- Pricing MSUs 66
- Processor Unit (PU) 6, 24, 46, 52, 54
- processor weighting 179
- PU 6, 24, 46, 52, 54
 - characterization 60
 - chip 36
 - conversions 64
 - sparing 47
- Public Key Algorithm 120

Q

- QDIO 103

R

- Redbooks Web site 263
 - Contact us x
- refrigeration 25
- Reliability, Availability, Serviceability (RAS) 17
- Reserved Processors 51
- Reserved storage 70
- Ring Topology 29
- riple length- key DES 120
- RMF support 141
- RPQ 8P2197 105
- RSA 120

S

- SAP 6, 50
 - additional 62
 - optional 64
- SC chip 35–36
- SD chip 30, 36
- SE 34, 66, 112, 238
- Self Timed Interconnect (STI) 6, 8, 31, 34, 39, 41, 77–79, 81, 88, 182, 189, 191, 210
- Service Element (SE) 34, 66, 112, 238
- Simplified I/O definition 182
- Software model MSU values 66
- Software support 19
- SSL 43
- STI 6, 39, 77, 182
 - connector 31, 78
 - extender card 34, 41
 - link 8, 81
 - rebalance feature 79, 81, 88, 189, 191, 210
- STI-2
 - extender card 79
- STI-3
 - extender card 79
- STI-M
 - card 76
- storage
 - CF mode 69
 - ESA/390 architecture mode 69
 - ESA/390 mode 68
 - expanded 55
 - granularity 70
 - Linux only mode 69
 - operations 67
 - reserved 70
 - TPF mode 69
 - z/Architecture mode 69
- STSI instruction 63
- superscalar 24, 42, 222
 - processor 41
- System Assist Processor (SAP) 6, 50, 64
- system memory 6

T

- TDES 120
- Token Ring 11
- TPF 146

- Trusted Key Entry 123

U

- UDX 121
- unassigned
 - CP 62, 64
 - IFL 62, 64
- uniprocessor speed 180
- unplanned upgrades 190
- upgrade paths 63
- upgrades
 - disruptive 217
 - nondisruptive 217
- UPID 112
- User Logical Partition ID 112

V

- VSE/ESA 146
 - software support 146

W

- WLM 15, 183–184
 - goals 180
- Workload License Charge 20
- Workload Manager (WLM) 15, 180, 183–184

Z

- Z Frame 33–34
- z/Architecture mode 69
- z/VM 47, 145
- z/VSE 146
 - software support 146
- z990
 - configurations 62
 - frames 33
 - models 62
- zAAP 6, 49, 62, 64, 138, 214
- zSeries Application Assist Processor (zAAP) 6, 49, 62, 64, 138, 214



IBM @server zSeries 990 Technical Guide

(0.5" spine)
0.475" <-> 0.873"
250 <-> 459 pages



IBM zSeries 990 Technical Guide




Redbooks

**Structure and design -
A scalable server for
an on demand world**

**Processor Unit,
memory, multiple
Logical Channel
Subsystems**

**Capacity upgrade
options**

The IBM  zSeries® 990 scalable server provides major extensions to the existing zSeries architecture and capabilities. The concept of Logical Channel Subsystems is added, and the maximum number of Processor Units and logical partitions is increased. These extensions provide the base for much larger zSeries servers.

This IBM® Redbook is intended for IBM systems engineers, consultants, and customers who need to understand the zSeries 990 features, functions, availability, and services.

This publication is part of a series. For a complete understanding of the z990 scalable server capabilities, also refer to our companion Redbooks:

IBM  zSeries 990 Technical Introduction, SG24-6863
IBM  zSeries Connectivity Handbook, SG24-5444

Note that the information in this book includes features and functions announced on April 7, 2004, and that certain functionality is not available until hardware Driver Level 55 is installed on the z990 server.

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