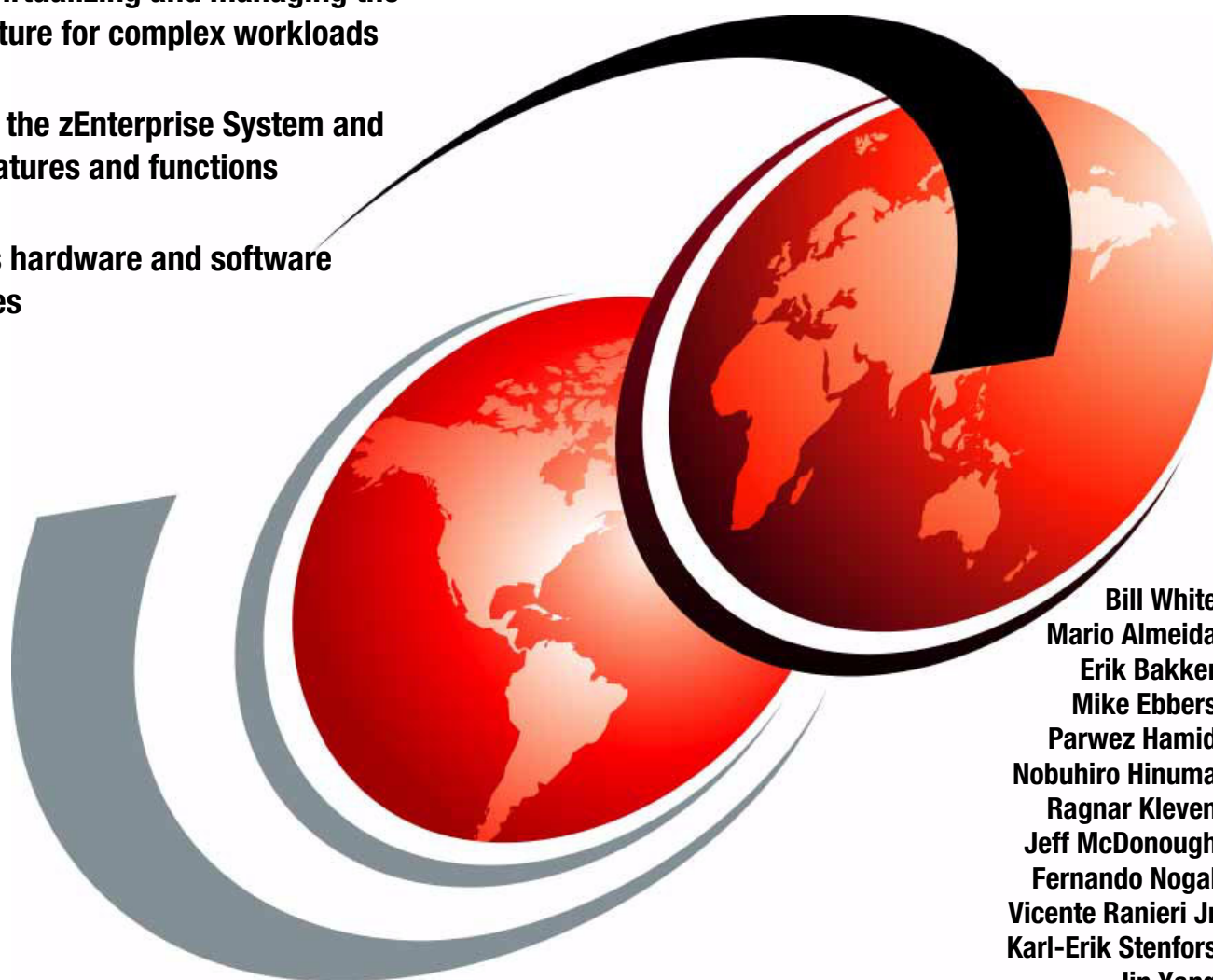**IBM**

IBM zEnterprise System Technical Guide

Explains virtualizing and managing the infrastructure for complex workloads

Describes the zEnterprise System and related features and functions

Discusses hardware and software capabilities

Bill White
Mario Almeida
Erik Bakker
Mike Ebbers
Parwez Hamid
Nobuhiro Hinuma
Ragnar Kleven
Jeff McDonough
Fernando Nogal
Vicente Ranieri Jr
Karl-Erik Stenfors
Jin Yang

Redbooks

ibm.com/redbooks

International Technical Support Organization

IBM zEnterprise System Technical Guide

August 2010

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

First Edition (August 2010)

This edition applies to the IBM zEnterprise System.

This document created or updated on September 21, 2010.

Contents

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

Trademarks

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

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

1-2-3®	IBM Systems Director Active Energy	Resource Link™
AIX®	Manager™	Resource Measurement Facility™
BladeCenter®	IBM®	RETAIN®
CICS®	IMS™	RMF™
Cool Blue™	Language Environment®	Solid®
DataPower®	Lotus®	Sysplex Timer®
DB2 Connect™	MQSeries®	System p®
DB2®	MVS™	System Storage®
Distributed Relational Database	Parallel Sysplex®	System x®
Architecture™	Power Systems™	System z10®
Domino®	POWER7™	System z9®
DRDA®	PowerPC®	System z®
DS8000®	PowerVM™	VM/ESA®
ECKD™	POWER®	WebSphere®
ESCON®	PR/SM™	z/Architecture®
FICON®	Processor Resource/Systems	z/OS®
FlashCopy®	Manager™	z/VM®
GDPS®	RACF®	z/VSE™
Geographically Dispersed Parallel	Redbooks®	z10™
Sysplex™	Redpaper™	z9®
HiperSockets™	Redbooks (logo) ®	zSeries®

The following terms are trademarks of other companies:

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

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

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

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

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

Preface

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

This IBM® Redbooks® publication discusses the IBM zEnterprise System, an IBM scalable mainframe server. IBM is taking a revolutionary approach by integrating different platforms under the well-proven System z hardware management capabilities, while extending System z qualities of service to those platforms.

The zEnterprise System consists of the IBM zEnterprise 196 central processor complex, the IBM zEnterprise Unified Resource Manager, and the IBM zEnterprise BladeCenter® Extension. The z196 is designed with improved scalability, performance, security, resiliency, availability, and virtualization. The z196 Model M80 provides up to 1.6 times the total system capacity of the z10™ EC Model E64, and all z196 models provide up to twice the available memory of the z10 EC. The zBX infrastructure works with the z196 to enhance System z virtualization and management through an integrated hardware platform that spans mainframe and POWER7™ technologies. Through the Unified Resource Manager, the zEnterprise System is managed as a single pool of resources, integrating system and workload management across the environment.

This book provides an overview of the zEnterprise System and its functions, features, and associated software support. Greater detail is offered in areas relevant to technical planning. This book is intended for systems engineers, consultants, planners, and anyone wanting to understand the zEnterprise System functions and plan for their usage. It is not intended as an introduction to mainframes. Readers are expected to be generally familiar with existing IBM System z technology and terminology.

The team who wrote this book

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

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

Mario Almeida is an IBM Certified Consulting IT Specialist working as an STG technical consultant in Brazil. He has more than 30 years of experience working with IBM large systems. Mario has co-authored several IBM Redbooks publications. His areas of expertise include System z hardware, Parallel Sysplex®, GDPS® and capacity planning.

Erik Bakker is a Senior IT Specialist working for IBM Server and Technology Group in the Netherlands. During the past 24 years he has worked in various roles within IBM and with a large number of mainframe customers. For many years he worked for Global Technology Services as a systems programmer providing implementation and consultancy services at many customer sites. He currently provides pre-sales System z technical consultancy in

support of large and small System z customers. His areas of expertise include Parallel Sysplex, z/OS® and System z.

Mike Ebbers is a Consulting IT Specialist and Project Leader at the International Technical Support Organization, Poughkeepsie Center. He has worked with IBM mainframe hardware and software products since 1974 in the field, in education, and in the ITSO.

Parwez Hamid is an Executive IT Consultant working for IBM's Server and Technology Group. During the past 37 years he has worked in various IT roles within IBM. Since 1988 he has worked with a large number of IBM's mainframe customers and spent much of his time introducing new technology. Currently, he provides pre-sales technical support for IBM's System z product portfolio and is the lead System z technical specialist for UK and Ireland. Parwez co-authors a number of ITSO Redbooks and prepares technical material for the world-wide announcement of System z Servers. Parwez works closely with System z product development in Poughkeepsie and provides input and feedback for 'future' product plans. Additionally, Parwez is a member of IBM's IT Specialist profession certification board in the UK and is also a Technical Staff member of the IBM's UK Technical Council which is made of senior technical specialist representing all of IBM's Client, Consulting, Services and Product groups. Parwez teaches and presents at numerous IBM user group and IBM internal conferences.

Nobuhiro Hinuma is an IT Specialist at STG Systems Technical Sales in Japan. He has 15 years of experience in the System z technical sales support field including Parallel Sysplex customer support in Japan and AP countries and z/OS Early Support Program. He is a member of zChamipons team since 2006. His areas of expertise include Parallel Sysplex, z/OS and System z.

Ragnar Kleven is a Client IT Architect in Norway, supporting Finance Services Sector and Public Sector customers. He has 35 years of experience in IT which of 8 are with IBM, and has been working with mainframes in many different roles. He holds a Bachelor degree in Engineering. The areas of expertise include System z, z/OS and z/OS middleware software stack and transactional banking solutions in general.

Jeff McDonough is an IT Architect in the USA. He holds a degree in Computing Analysis from Missouri Southern State College, and has 33 years experience in IT. He has experience in the manufacturing, transportation, and retail industries. He has specialized in z/OS and parallel sysplex environments for 17 years and has previously written about Parallel Sysplex using InfiniBand, Server Time Protocol, and other z/OS topics

Fernando Nogal is an IBM Certified Consulting IT Specialist working as an STG Technical Consultant for the Spain, Portugal, Greece, and Israel IMT. He specializes in on-demand infrastructures and architectures. In his 28 years with IBM, he has held a variety of technical positions, mainly providing support for mainframe customers. Previously, he was on assignment to the Europe Middle East and Africa (EMEA) zSeries® Technical Support group, working full time on complex solutions for e-business on zSeries. His job included, and still does, presenting and consulting in architectures and infrastructures, and providing strategic guidance to System z customers regarding the establishment and enablement of e-business technologies on System z, including the z/OS, z/VM®, and Linux® environments. He is a zChampion and a core member of the System z Business Leaders Council. An accomplished writer, he has authored and co-authored over 20 Redbooks and several technical papers. Other activities include chairing a Virtual Team from IBM interested in e-business on System z, and serving as a University Ambassador. He travels extensively on direct customer engagements and as a speaker at IBM and customer events, and trade shows.

Vicente Ranieri is an Executive IT Specialist at STG Advanced Technical Support (ATS) team supporting System z in Latin America. He has more than 30 years of experience

working for IBM. Vicente is a member of zChampions team, a worldwide IBM team to participate in the creation of System z technical roadmap and value proposition materials. Besides co-authoring several redbooks, he has been an ITSO guest speaker since 2001, teaching the System z security update workshops worldwide. Vicente also presents in several IBM internal and external conferences. His areas of expertise include System z security, Parallel Sysplex, System z hardware and z/OS. Vicente is a member of Technology Leadership Council – Brazil, an IBM Academy of Technology Affiliate.

Karl-Erik Stenfors is a Senior IT Specialist in the PSSC Customer Center in Montpellier, France. He has more than 40 years of working experience in the Mainframe environment, as a systems programmer, as a consultant with IBM's customers, and, since 1986 with IBM. His areas of expertise include IBM System z hardware and operating systems. He teaches at numerous IBM user group and IBM internal conferences, and he is a member of the zChampions work group. His current responsibility is to execute System z Early Support Programs in Europe and Asia.

Jin Yang is a Senior System Service Representative at the IBM Global Technical Services in Beijing, China. He joined IBM in 1999 to support and maintain System z products for clients throughout China. Jin has been working in the Technical Support Group (TSG) providing second level support to System z clients since 2009. His areas of expertise include System z hardware, Parallel Sysplex, and FICON® connectivity.

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

Ivan Bailey, Connie Beuselinck, Patty Driever, Jeff Frey, Steve Fellenz, Michael Jordan, Gary King, Bill Kostenko, Jeff Kubala, Kelly Ryan, Lisa Schloemer, Jaya Srikrishnan, Peter Yocom, Martin Ziskind
IBM Poughkeepsie

Gwendolyn Dente, Harv Emery, Gregory Hutchison
IBM Advanced Technical Skills (ATS), North America

Friedemann Baitinger, Klaus Werner
IBM Germany

Brian Tolan, Brian Valentine, Eric Weinmann
IBM Endicott

Garry Sullivan
IBM Rochester

Jerry Stevens
IBM Raleigh

John P. Troy
IBM Hartford

International Technical Support Organization:

Robert Haimowitz
IBM Raleigh

Ella Buslovich
IBM Poughkeepsie

Now you can become a published author, too!

Here's an opportunity to spotlight your skills, grow your career, and become a published author - all at the same time! Join an ITSO residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

Find out more about the residency program, browse the residency index, and apply online at:

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

We want our books to be as helpful as possible. Send us your comments about this book or other IBM Redbooks publications in one of the following ways:

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

ibm.com/redbooks

► Send your comments in an e-mail to:

redbooks@us.ibm.com

► Mail your comments to:

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

1

Introducing the IBM zEnterprise System

The zEnterprise System is the first system of its kind. It was designed to help overcome problems in today's IT infrastructures and provide a foundation for the future. The zEnterprise System represents both a revolution and an evolution of mainframe technology. IBM is taking a bold step by integrating heterogeneous platforms under the well-proven System z hardware management capabilities, while extending System z qualities of service to those platforms.

The zEnterprise 196 central processor complex (CPC) has a newly designed quad-core chip, the fastest in the industry at 5.2 GHz. The z196 can be configured with up to 80 processors running concurrent production tasks with up to 3 TB of memory. It offers hot pluggable I/O drawers that complement the I/O cages, and continues the utilization of advanced technologies such as InfiniBand.

The z196 goes beyond previous designs while continuing to enhance the traditional mainframe qualities, delivering unprecedented performance and capacity growth. The z196 is a well-balanced general-purpose server that is equally at ease with compute-intensive and I/O-intensive workloads.

The integration of heterogeneous platforms is based on IBM's BladeCenter technology. The IBM zEnterprise BladeCenter Extension (zBX) Model 002 houses general purpose blades as well as specialized solutions, such as the IBM Smart Analytics Optimizer.

Another key element is the zEnterprise Unified Resource Manager firmware. Together, z196, zBX, and Unified Resource Manager constitute a *node* in a zEnterprise *ensemble*. An zEnterprise ensemble is a collection of highly virtualized heterogeneous systems, managed as a single logical entity, where diverse workloads can be deployed.

A zEnterprise ensemble can have a maximum of eight nodes, comprised of up to eight z196 servers and 896 blades (housed in eight zBXs). The ensemble has dedicated integrated networks for management and data and the Unified Resource Manager functions.

1

Figure 1-1 shows the elements of the zEnterprise System.
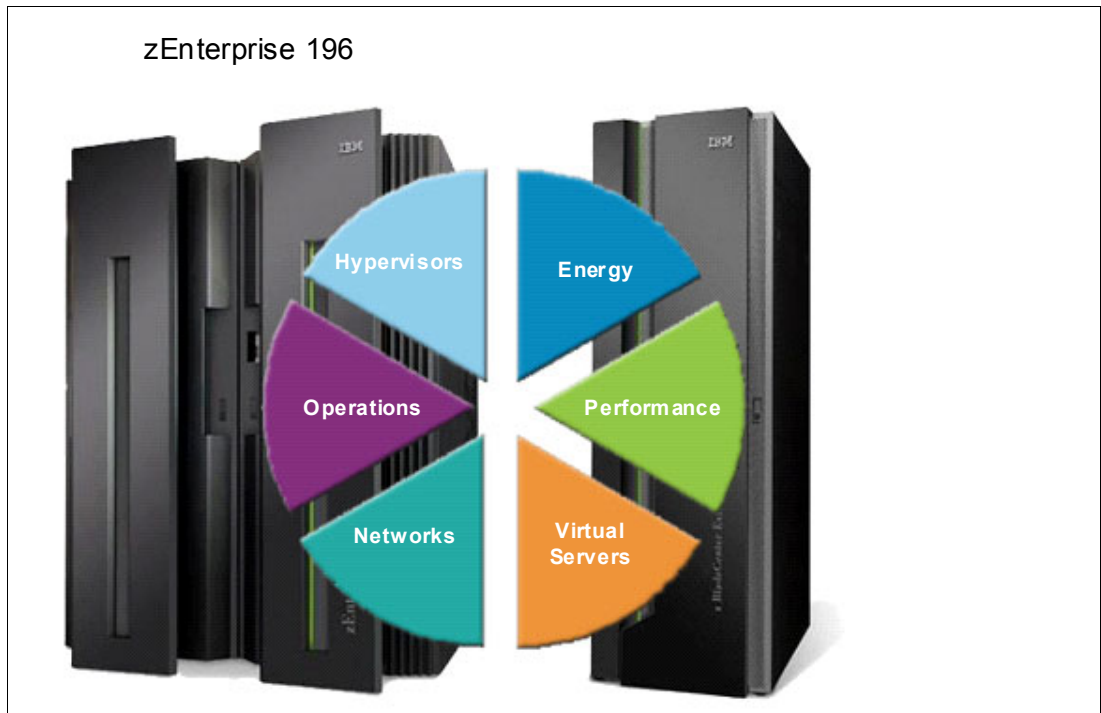
Figure 1-1   Elements of the zEnterprise System

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

Workloads have changed. Now many applications, including mission-critical ones, are deployed in heterogeneous infrastructures and the System z design has adapted to this change. The z196 design can simultaneously support a large number of diverse workloads while providing the highest qualities of service.

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

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

Careful engineering of the application's several tiers is required to provide the robustness, scaling, consistent response time, and other characteristics demanded by the users and lines of business. These infrastructures do not scale well. What is a feasible setup with a few servers becomes difficult to handle with dozens and a nightmare with hundreds. When it is doable it is expensive. Often, by the end of the distributed equipment's life cycle, its residual value is nil, requiring new acquisitions, software licenses, and re-certification. It is like going back to square one. In today's resource-constrained environments there is a better way.

To complete this picture on the technology side, performance gains from increasing the frequency of chips are becoming smaller. Thus, special-purpose compute acceleration will be required for greater levels of workload performance and scalability.

The zEnterprise is following an evolutionary path that directly addresses those infrastructure problems. Over time, it will provide increasingly complete answers to the smart infrastructure requirements. zEnterprise, with its heterogeneous platform management capabilities, already provides many of these answers, offering great value in a scalable solution that integrates and simplifies hardware and firmware management and support, as well as the definition and management of a network of virtualized servers, across multiple heterogeneous platforms.

The z196 expands the subcapacity settings offer with three different subcapacity levels for the first 15 processors, giving a total of 125 distinct capacity settings in the system, and providing for a range of over 1:200 in processing power. The z196 delivers scalability and granularity to meet the needs of medium-sized enterprises, while also satisfying the requirements of large enterprises having large-scale, mission-critical transaction and data processing requirements. The zEnterprise 196 continues to offer all the specialty engines available with System z10®.

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

1.1 zEnterprise 196 highlights

The z196 CPC provides a record level of capacity over the previous System z servers. This is achieved both by increasing the performance of the individual processor units and by increasing the number of processor units (PUs) per server. The increased performance and the total system capacity available, along with possible energy savings, offer the opportunity to consolidate diverse applications on a single platform, with real financial savings. New features help to ensure that the zEnterprise 196 is an innovative, security-rich platform that can help maximize resource exploitation and utilization, and can help provide the ability to integrate applications and data across the enterprise IT infrastructure.

IBM continues its technology leadership with the z196. The server is built using IBM modular multibook design that supports one to four books per server. The book contains a Multi-Chip Module (MCM), which hosts the newly designed CMOS 12s processor units, storage control chips, and connectors for I/O. The superscalar processor has out-of-order instruction execution for better performance.

This approach provides many high-availability and nondisruptive operations capabilities that differentiate it from other servers. In addition, the system I/O bus takes advantage of the InfiniBand technology, which is also exploited in coupling links. The Parallel Sysplex cluster takes the commercial strengths of the z/OS platform to improved levels of system management, competitive price/performance, scalable growth, and continuous availability.

The z196 has five model offerings ranging from one to 80 configurable processor units (PUs). The first four models (M15, M32, M49, and M66) have 20 PUs per book, and the high capacity model (the M80) has four 24 PU books. Model M80 is estimated to provide up to 60% more total system capacity than the z10 Model E64, with up to two times the available memory. This comparison is based on the Large Systems Performance Reference (LSPR) mixed workload analysis.

Flexibility in customizing traditional capacity to meet individual needs led to the introduction of subcapacity processors. The z196 has increased the number of subcapacity CPs available in a server to 15. When the capacity backup (CBU) function is invoked, the number of total subcapacity processors cannot exceed 15.

Depending on the model, the z196 can support from a minimum of 32 GB to a maximum of 3056 GB of useable memory, with up to 768 GB per book. In addition, a fixed amount of 16 GB is reserved for Hardware System Area (HSA) and is not part of customer-purchased memory. Memory is implemented as a Redundant Array of Independent Memory (RAIM). Up to 960 GB are installed per book, for a system total of 3840 GB.

The traditional I/O cages are complemented by I/O drawers, which were introduced with the IBM z10 BC. There are up to 48 high-performance fanouts for data communications between the server and the peripheral environment. The multiple channel subsystems (CSS) architecture allows up to four CSSs, each with 256 channels. I/O constraint relief, using three subchannel sets, allows access to a greater number of logical volumes.

Processor Resource/Systems Manager™ (PR/SM™) manages all the installed and enabled resources (processors and memory) as a single large SMP system. It enables the configuration and operation of up to 60 logical partitions, which have processors, memory, and I/O resources assigned from the installed books. PR/SM dispatching has been redesigned to work together with the z/OS dispatcher in a function called HiperDispatch. HiperDispatch provides work alignment to logical processors, and alignment of logical processors to physical processors. This alignment optimizes cache utilization, minimizes inter-book communication, and optimizes z/OS work dispatching, with the end result of increasing throughput. HiperSockets™ has been enhanced (z196 supports 32 HiperSockets).

The z196 continues the mainframe reliability, availability, and serviceability (RAS) tradition of reducing all sources of outages with continuous focus by IBM on keeping the system running. It is a design objective to provide higher availability with a focus on reducing planned and unplanned outages. With a properly configured z196, further reduction of outages can be attained through improved nondisruptive replace, repair, and upgrade functions for memory, books, and I/O adapters, as well as extending nondisruptive capability to download Licensed Internal Code (LIC) updates.

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

▶ Maximum configuration of 60 LPARs, four LCSSs and three MSSs
▶ Dynamic add/remove of a new logical partition (LPAR) to new or existing logical channel subsystem (LCSS)
▶ Dynamic addition and removal of Crypto Express3 features
▶ Dynamic I/O enabled as a default
▶ Add/change number of logical CPs, IFLs, ICFs, zAAPs, and zIIPs processors per partition
▶ Dynamic LPAR PU assignment optimization CPs, ICFs, IFLs, zAAPs, and zIIPs

Capacity on Demand

On demand enhancements enable customers to have more flexibility in managing and administering their temporary capacity requirements. The z196 supports the architectural approach for temporary offerings introduced with z10, that has the potential to change the thinking about on demand capacity. Within the z196, one or more flexible configuration definitions can be available to solve multiple temporary situations and multiple capacity configurations can be active simultaneously.

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

1.2 zEnterprise 196 models

The z196 is a two-frame server: the A frame and the Z frame. The frames contain the components, including:

- ► The CPC cage with up to four books
- ► Combinations of up to four I/O drawers and up to two I/O cages (a third I/O cage requires an RPQ)
- ► Power supplies
- ► An optional internal battery feature (IBF)
- ► Modular cooling units for either air or water cooling
- ► Support Elements

The zEnterprise 196 has a machine type of 2817. Five models are offered: M15, M32, M49, M66, and M80. The last two digits of each model indicate the maximum number of PUs available for purchase. A PU is the generic term for the z/Architecture® processor on the Multi-Chip Module (MCM) that can be characterized as any of the following items:

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

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

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

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

The I/O features or channel types supported are:

- ► ESCON® (Enterprise Systems Connection) up to 240 channels
- ► FICON Express8 (Fibre Channel connection)
- ► FICON Express4 (only when carried forward from a previous System z server)
- ► OSA-Express3
- ► OSA-Express2 (only when carried forward from a previous System z server; the OSA-Express2 10 GbE LR feature is not supported)
- ► Crypto Express3

► Coupling Links - peer mode only (ISC-3)
► The Parallel Sysplex InfiniBand coupling link (PSIFB)

1.2.1 Model upgrade paths

Any z196 can be upgraded to another z196 hardware model. Upgrade of models M15, M32, M49, and M66 to M80 is disruptive (that is, the machine is unavailable during this upgrade). Any z9® EC or z10 EC model may be upgraded to any z196 model. Figure 1-2 on page 6 presents a diagram of the upgrade path.
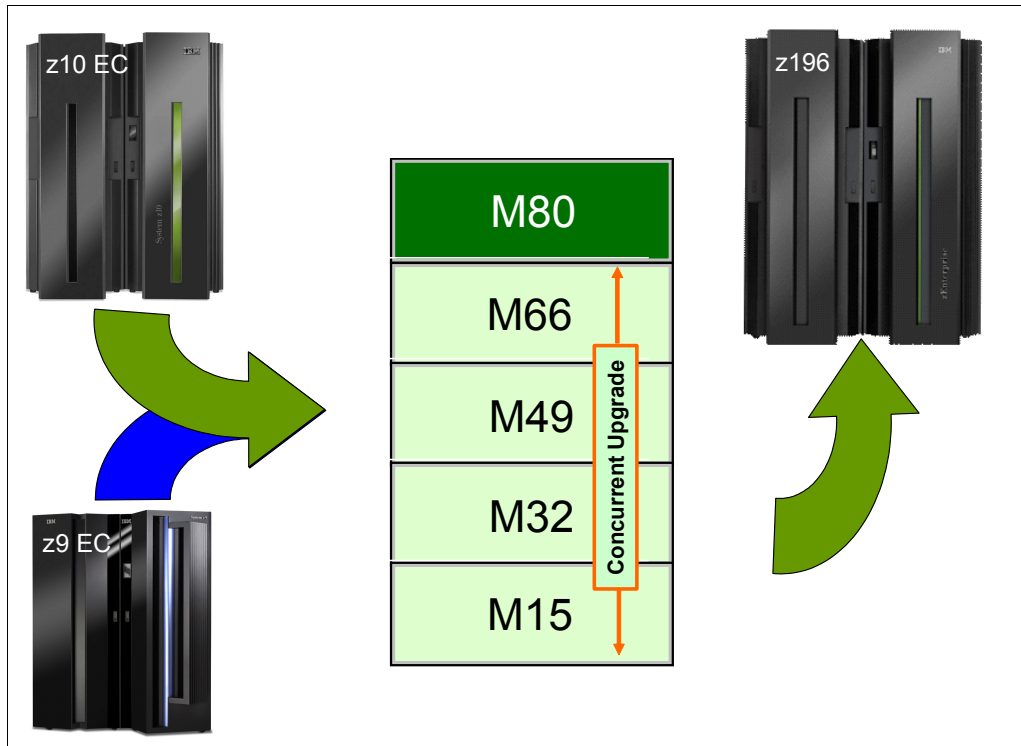
Figure 1-2   System z upgrades

1.2.2 Concurrent processor unit conversions

The z196 supports concurrent conversion between different PU types, providing flexibility to meet changing business environments. CPs, IFLs, zAAPs, zIIPs, ICFs, or optional SAPs may be converted to CPs, IFLs, zAAPs, zIIPs, ICFs, or optional SAPs.

1.3 System functions and features

The z196 is a two-frame server. The frames contain the key components.

Functions and features

These include many features described in this chapter, plus the following:

► Single processor core sparing
► Large page (1 MB)
► Redundant 100 Mb Ethernet Service Network with VLAN

- ► 12x InfiniBand coupling links for local connections and 1x InfiniBand coupling links for extended distance connections
- ► Increased flexibility for Capacity on Demand just-in-time offerings with ability for more temporary offerings installed on the central processor complex (CPC) and ways to acquire capacity backup

Design highlights

The z196 provides:

- ► Increased bandwidth between memory and I/O
- ► Reduction in the impact of planned and unplanned server outages:
    - – Enhanced book availability
    - – Hot pluggable I/O drawers
    - – Redundant I/O interconnect
    - – Enhanced driver maintenance
    - – Concurrent Host Channel Adapter (HCA-O and HCA-C) fanout card hot-plug
- ► Up to three subchannel sets that are designed to allow improved device connectivity for Parallel Access Volumes (PAVs), PPRC secondaries, and FlashCopy® devices. This third subchannel set allows the user to extend the amount of addressable external storage
- ► More capacity over native FICON channels for programs that process data sets, which exploit striping and compression (such as DB2®, VSAM, PDSE, HFS, and zFS) by reducing channel, director, and control unit overhead when using the Modified Indirect Data Address Word (MIDAW) facility
- ► Improved access to data for OLTP applications with High Performance FICON for System z (zHPF) on FICON Express8 and FICON Express4 channels
- ► Enhanced problem determination, analysis, and manageability of the storage area network (SAN) by providing registration information to the fabric name server for both FICON and FCP

1.3.1 Processor

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

Processor features

The z196 book has a Multi-Chip Module with six chips. The processor chip has a quad-core design, with either three or four active cores, and operates at 5.2 GHz. Depending on the MCM version (20 PU or 24 PU), from 20 to 96 PUs are available, on one to four books.

The MCM provides a significant increase in system scalability and an additional opportunity for server consolidation. All books are interconnected with very high-speed internal communications links, in a fully connected star topology through the L4 cache, which allows the system to be operated and controlled by the PR/SM facility as a memory- and cache-coherent symmetric multiprocessor (SMP).

The PU configuration is made up of two spare PUs per server and a variable number of system assist processors (SAPs), which scale with the number of books installed in the server, such as three SAPs with one book installed and up to 14 when four books are installed. The remaining PUs can be characterized as central processors (CPs), Integrated Facility for Linux (IFL) processors, z196 Application Assist Processors (zAAPs), z196 Integrated Information Processors (zIIPs), Internal Coupling Facility (ICF) processors, or additional SAPs.

The PU chip includes data compression and cryptographic functions, such as the CP Assist for Cryptographic Function (CPACF). Hardware data compression can play a significant role

in improving performance and saving costs over doing compression in software. Standard clear key cryptographic processors right on the processor translate to high-speed cryptography for protecting data in storage, integrated as part of the PU.

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

Increased flexibility with z/VM-mode partitions

z196 provides for the definition of a z/VM-mode logical partition (LPAR) containing a mix of processor types including CPs and specialty processors such as IFLs, zIIPs, zAAPs, and ICFs.

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

1.3.2 Memory subsystem

A buffered DIMM has been developed for the z196. For this purpose IBM has developed a chip that controls communication with the PU and drives address and control from DIMM to DIMM. The DIMM capacities are 4, 16, and 32 GB.

Memory topology

Memory topology provides:

► Redundant array of independent memory (RAIM) for protection at the DRAM, DIMM, and memory channel level
► Maximum of 3.0 TB of user configurable memory with a maximum of 3840 GB of physical memory (with a maximum of 1 TB configurable to a single logical partition)
► One memory port for each CP chip; up to three independent memory ports per book
► Asymmetrical memory size and DRAM technology across books
► Key storage
► Storage protection key array kept in physical memory
► Storage protection (memory) key is also kept in every L2 and L3 cache directory entry
► Large (16 GB) fixed-size HSA eliminates having to plan for HSA

1.3.3 Central processor complex cage

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

MCM technology

The z196 is built on a proven superscalar microprocessor architecture. In each book, there is one MCM. The MCM has six PU chips and two SC chips. The PU chip has four cores, with either three or four cores active, which can be characterized as CPs, IFLs, ICFs, zIIPs, zAAPs, or SAPs. Two MCM sizes are offered, which are 20 or 24 cores.

The z196 offers a water cooling option for increased system and data center energy efficiency. For a water cooled system the MCM is cooled by a cold plate connected to the internal water cooling loop instead of the evaporator/heat sink for an air cooled system with the modular refrigeration unit (MRU) and air backup.

Out-of-order execution

The z196 has a superscalar microprocessor with out-of-order (OOO) execution to achieve faster throughput. With OOO, instructions may not execute in the original program order, although results are presented in the original order. OOO allows, for instance, several instructions to complete while another is waiting. Up to three instructions can be decoded per cycle and up to five instructions can be executed per cycle.

Host channel adapter fanout hot-plug

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

1.3.4 I/O connectivity

The z196 offers several improved features and exploits technologies such as InfiniBand and Ethernet. In this section we briefly review the most relevant I/O capabilities.

InfiniBand

The z196 takes advantage of InfiniBand to implement:

► An I/O bus, which includes the InfiniBand infrastructure. This replaces the self-timed interconnect bus found in System z servers prior to z9.
► Parallel Sysplex coupling using InfiniBand (PSIFB). This link has a bandwidth of 6 GBps between two z196 or z10 servers and 3 GBps between System z196 or z10 and System z9 servers.

1.3.5 I/O subsystems

The I/O subsystem draws on developments from z10. The I/O subsystem is supported by a new I/O bus similar to z10's, and includes the InfiniBand Double Data Rate (IB-DDR) infrastructure (replacing self-timed interconnect found in the prior System z servers). This infrastructure is designed to reduce overhead and latency, and provide increased throughput. The I/O expansion network uses the InfiniBand Link Layer (IB-2, Double Data Rate).

z196 also offers an I/O infrastructure element called an I/O drawer which is a companion to the I/O cage.

I/O drawer

I/O drawers provide increased I/O granularity and capacity flexibility and can be concurrently added and removed in the field, an advantage over I/O cages. This also eases planning. The z196 server can have up to four I/O drawers, two in the A frame and two on the Z frame. I/O drawers were first offered with the z10 BC and can accommodate up to eight I/O features in any combination.

I/O cage

The z196 has a CPC cage and, optionally, one I/O cage in the A frame. The Z frame can accommodate two additional I/O cages, bringing the total for the system to three. Adding the third I/O cage requires an RPQ. Each I/O cage can accommodate up to 28 I/O features in any combination.

I/O features

The z196 supports the following I/O features, which can be installed in both the I/O drawers and I/O cages:

► ESCON
► FICON Express8
► FICON Express4 (when carried forward on a migration)
► OSA-Express3
► OSA-Express2 (when carried forward on a migration, except OSA-Express2 10 GbE LR)
► Crypto Express3
► ISC-3 coupling links

ESCON channels

The high-density ESCON feature (FC 2323) has 16 ports, of which 15 can be activated. One port is always reserved as a spare in the event of a failure of one of the other ports. Up to 16 features are supported (24 with an RPQ). With the z196, 240 channels are supported.

FICON channels

Up to 72 features with up to 288 FICON Express8 or FICON Express4 channels are supported:

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

The z196 supports FICON, High Performance FICON for System z (zHPF), channel-to-channel (CTC), and Fibre Channel Protocol (FCP).

With an RPQ for the third I/O cage, up to 84 features with up to 336 channels are supported.

Open Systems Adapter

The z196 can have up to 24 features of the Open Systems Adapter (OSA) family of Ethernet features, for a maximum of 96 ports of LAN connectivity. With the exception of OSA-Express2 10 GbE LR, which is not supported, any combination of the supported OSA-Express2 or OSA-Express3 features can be selected.

OSM and OSX CHPID types

The z196 introduces two OSA-Express3 CHPID types:

► OSA-Express for Unified Resource Manager (OSM): Connectivity to the intranode management network (INMN). Connections from z196 to the Bulk Power Hubs (BPHs) for use of the Unified Resource Manager functions in the HMC. Uses OSA-Express3 1000BASE-T Ethernet exclusively.
► OSA-Express for zBX (OSX): Connectivity to the intraensemble data network (IEDN). Connections from z196 to zBX. Uses OSA-Express3 10 GbE exclusively.

OSA-Express3 feature highlights

The z196 has five OSA-Express3 features. When compared to similar OSA-Express2 features, which they replace, OSA-Express3 features provide the following important benefits:

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

Performance enhancements are the result of the data router function present in all OSA-Express3 features. What previously was performed in firmware, the OSA-Express3 now performs in hardware. Additional logic in the IBM ASIC handles packet construction,

inspection, and routing, thereby allowing packets to flow between host memory and the LAN at line speed without firmware intervention.

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

For more information about the OSA-Express3 features refer to 4.7.4, "OSA-Express3" on page 134.

HiperSockets

The HiperSockets function, also known as internal queued direct input/output (internal QDIO or iQDIO), is an integrated function of the z196 that provides users with attachments to up to 32 high-speed virtual LANs with minimal system and network overhead.

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

HiperSockets eliminates having to use I/O subsystem operations and to traverse an external network connection to communicate between logical partitions in the same z196 server. HiperSockets offers significant value in server consolidation by connecting many virtual servers, and can be used instead of certain coupling link configurations in a Parallel Sysplex.

1.3.6 Cryptography

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

CP Assist for Cryptographic Function

The z196 uses the Common Cryptographic Architecture (CCA). The CP Assist for Cryptographic Function (CPACF) offers the full complement of the Advanced Encryption Standard (AES) algorithm and Secure Hash Algorithm (SHA). Support for CPACF is also available by using the Integrated Cryptographic Service Facility (ICSF). ICSF is a component of z/OS, and can transparently use the available cryptographic functions, CPACF, or Crypto Express3, to balance the workload and help address the bandwidth requirements of your applications.

The enhancements to CPACF are exclusive to the z196 and supported by z/OS, z/VM, z/VSE, z/TPF, and Linux on System z.

Configurable Crypto Express3 feature

The Crypto Express3 feature has two PCIe adapters, which can each be configured as a coprocessor or an accelerator:

► Crypto Express3 Coprocessor is for secure key encrypted transactions (default).
► Crypto Express3 Accelerator is for Secure Sockets Layer (SSL) acceleration.

A recently added functions include:

► ANSI X9.8 PIN security

- ► Enhanced Common Cryptographic Architecture (CCA) - a key wrapping to comply with ANSI X9.24-1 key bundling requirements
- ► Secure Keyed-Hash Message Authentication Code (HMAC)
- ► Elliptical Curve Cryptography Digital Signature Algorithm
- ► Modulus Exponent (ME) and Chinese Remainder Theorem (CRT)

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

TKE workstation, migration wizard and support for Smart Card Reader

The Trusted Key Entry (TKE) workstation (FC 0841) and the TKE 7.0 LIC (FC 0860) are optional features on the z196. The TKE workstation offers security-rich local and remote key management, providing authorized personnel a method of operational and master key entry, identification, exchange, separation, and update. Recent enhancements include support for the AES encryption algorithm, audit logging, and an infrastructure for payment card industry data security standard (PCIDSS), as well as:

- ► EEC Master Key Support
- ► CBC Default Settings Support
- ► TKE Audit Record Upload Configuration Utility Support
- ► USB Flash Memory Drive Support
- ► Stronger PIN Strength Support
- ► Stronger Password Requirements for TKE Passphrase user Profile Support

TKE has a wizard to allow users to collect data, including key material, from a Crypto Express coprocessor and migrate the material to a different Crypto Express coprocessor. The target coprocessor must have the same or greater capabilities. This wizard is intended to help migrate from Crypto Express 2 to Crypto Express3. Crypto Express2 is *not* supported on z196.

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

1.3.7 Parallel Sysplex support

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

Coupling links support

Coupling connectivity in support of Parallel Sysplex environments is provided as stated in the following list. The z196 does not support ICB4 connectivity. Parallel Sysplex connectivity now supports:

- ► Internal Coupling Channels (ICs) operating at memory speed
- ► InterSystem Channel-3 (ISC-3) operating at 2 Gbps and supporting an unrepeated link data rate of 2 Gbps over 9 μm single mode fiber optic cabling with an LC Duplex connector.
- ► 12x InfiniBand coupling links offer up to 6 GBps of bandwidth between z10 and z196 servers and up to 3 GBps of bandwidth between z10 or z196 severs and z9 servers for a distance up to 150 m (492 feet).
- ► 1x InfiniBand up to 5 Gbps connection bandwidth between z10 and z196 servers for a distance up to 10 km (6.2 miles).

All coupling link types can be used to carry Server Time Protocol (STP) messages.

Coupling Facility Control Code Level 17

Coupling Facility Control Code (CFCC) Level 17 is available for the IBM System z196. Enhancements include:

▶ Greater than 1024 CF Structures

   The limit has been increased to 2047 structures. Greater than 1024 CF Structures requires a new version of the CFRM CDS.

   – All systems in the sysplex need to be at z/OS V1.12 or have the coexistence/preconditioning PTF installed.
   – Falling back to a previous level (without coexistence PTF installed) is NOT supported without sysplex IPL.

▶ Greater than 32 Connectors

   Limits have been increased, depending on structure type. The new limits are: 255 for cache, 247 for lock, or 127 for serialized list structures. Greater than 32 Connectors is only usable when all CFs are at or above CF Level 17.

▶ Improved CFCC diagnostics & Link Diagnostics
▶ The number of available CHPID has been increased from 64 to 128 CHPIDs.

Server Time Protocol facility

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

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

Network Time Protocol (NTP) client support is available to the STP code on the z196, z10 and z9. With this functionality, the z196, z10 and z9 can be configured to use an NTP server as an external time source (ETS).

This implementation answers the need for a single time source across the heterogeneous platforms in the enterprise, allowing an NTP server to become the single time source for the z196, z10 and z9, as well as other servers that have NTP clients (UNIX®, NT, and so on). NTP can only be used for an STP-only CTN where no server can have an active connection to a Sysplex Timer®.

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

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

A System z196 can not be connected to a Sysplex Timer. We recommend migration to a STP-only Coordinated Time Network (CTN) for existing environments. It is possible to have a System z196 as a Stratum 2 or Stratum 3 server in a Mixed CTN, as long as there are at least two System z10s or System z9s attached to the Sysplex Timer operating as Stratum 1 servers.

1.4 IBM zEnterprise BladeCenter Extension

The IBM zEnterprise BladeCenter Extension (zBX) is available as an option with the z196 servers and consists of the following:

► Up to four IBM Enterprise racks.
► Up to eight BladeCenter chassis[1] with up to 14 blades[2] each
► Blades, up to 112[3]
► Management TOR switches for the intranode management network (INMN). The INMN provides connectivity between the z196 Support Elements and the zBX for management purposes.
► Intraensemble data network (IEDN) TOR switch. The IEDN is used for data paths between the z196 and the zBX, and the other ensemble members.
► 8 Gbps Fibre Channel switch modules for connectivity to an SAN.
► Power Distribution Units (PDUs) and cooling fans.

The zBX is configured with redundant components to provide qualities of service similar to those of System z, such as the capability for concurrent upgrades and repairs.

The zBX provides a foundation for the future. Based on IBM's judgement of the market's needs, additional specialized or general purpose blades might be introduced.

1.4.1 IBM blades

IBM offers a selected subset of IBM POWER7 blades that can be installed and operated on the zBX. These blades have been thoroughly tested to ensure compatibility and manageability in the z196 environment.

These blades are virtualized using PowerVM™ and the virtual servers run the AIX® operating system.

1.4.2 IBM Smart Analytics Optimizer solution

The IBM Smart Analytics Optimizer solution is a defined set of software and hardware that provides a cost optimized solution for running Data Warehouse and Business Intelligence queries against DB2 for z/OS, with fast and predictable response times, while retaining the data integrity, data management, security, availability and other qualities of service of the z/OS environment. It exploits special purpose blades, hosted in a zBX.

The offering is comprised of hardware and software. The software consists of the IBM Smart Analytics Optimizer for DB2 for z/OS, Version 1.1 (Program Product 5697-AQT). The hardware is offered in five different sizes based on the amount of DB2 data (DB2 tables, number of indexes, number of AQTs[4]) to be queried.

---

[1] A maximum of four BladeCenter chassis for the IBM Smart Analytics Optimizer solution.
[2] Depending on the IBM Smart Analytics Optimizer configuration this can be either 7 or 14 blades in the first BladeCenter chassis.
[3] A maximum of 56 blades for the IBM Smart Analytics Optimizer solution.
[4] Eligible queries for the IBM Smart Analytics Optimizer solutions will be executed on data marts specified as Accelerator Query Table (AQT) in DB2 for z/OS. An AQT is based on the same principles as a Materialized Query Table (MQT). MQTs are tables whose definitions are based on query results. The data in those tables is derived from the table or tables on which the MQT definition is based. See the article at:
http://www.ibm.com/developerworks/data/library/techarticle/dm-0509melnyk

The offering includes from 7 to 56 blades that are housed in one to four dedicated BladeCenter chassis. One or two standard 19 inch 42U IBM Enterprise racks and Top-of-Rack switches might be required.

In addition, a customer supplied external disk (IBM DS5020) is required for storing the compressed data segments. The data segments are read into blade memory for DB2 queries.

1.5 Unified Resource Manager

Unified Resource Manager is a part of the IBM System Director family. It is an integrated System z management facility for zEnterprise platform management. The functions are grouped as follows:

► Defining and managing virtual environments. This includes the automatic discovery as well as the definition of I/O and other hardware components across z196 and zBX, and the definition and management of LPARs, virtual machines, and virtualized LANs.
► Defining and managing workloads and workload policies.
► Receiving and applying corrections and upgrades to the Licensed Internal Code.
► Performing temporary and definitive z196 capacity upgrades.

These functions in support of an ensemble are provided by the Hardware Management Console (HMC) and Support Elements (SE) and exploit the intraensemble data network.

1.5.1 Hardware Management Console and Support Element

The Hardware Management Consoles (HMC) and Support Elements (SE) are appliances which together provide hardware platform management for System z. The HMC is used to manage, monitor, and operate one or more IBM System z servers and their associated logical partitions. The HMC[5] has a global (ensemble) management function, whereas the SE has local node management responsibility. When tasks are performed on the HMC, the commands are sent to one or more SEs, which then issue commands to their CPCs and zBXs. In order to promote high availability, an ensemble configuration requires a pair of HMCs in primary and alternate roles.

1.6 Reliability, availability, and serviceability

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

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

The RAS strategy is focused on a recovery design that is necessary to mask errors and make them transparent to customer operations. An extensive hardware recovery design has been

---

[5] From Version 2.11 on with feature codes 0090, 0025, 0019 and optionally 0020.

implemented to detect and correct array faults. In cases where total transparency cannot be achieved, you may restart the server with the maximum possible capacity.

1.7 Performance

The z196 Model M80 is designed to offer approximately 1.6 times more capacity than the z10 EC Model E64 system. Uniprocessor performance has also increased significantly. A z196 Model 701 offers, on average, performance improvements of about 1.35 to 1.5 times the z10 EC Model 701. Figure 1-3, "z196 to z10 EC performance comparison" on page 16 shows the estimated capacity ratios for z196 and z10 EC.
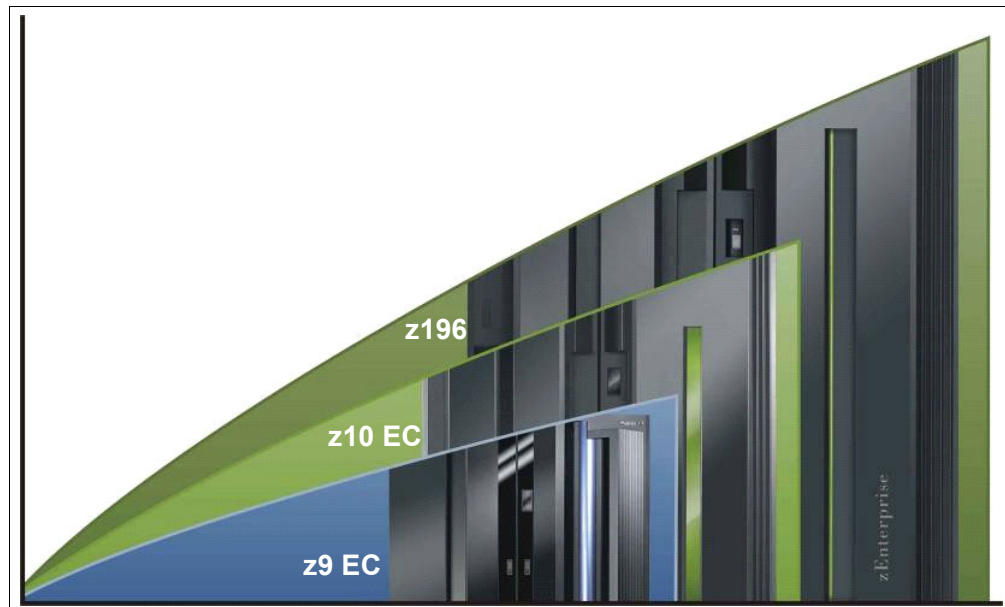
Figure 1-3   z196 to z10 EC performance comparison

On average, the z196 can deliver up to 60% more performance in an n-way configuration than an IBM System z10 EC n-way. However, variations on the observed performance increase are dependent upon the workload type.

The LSPR contains the internal throughput rate ratios (ITRRs) for the z196 and the previous generation processor families, based upon measurements and projections that use standard IBM benchmarks in a controlled environment. The actual throughput that any user experiences can vary depending on considerations, such as the amount of multiprogramming in the user's job stream, the I/O configuration, and the workload processed. Therefore, no assurance can be given that an individual user can achieve throughput improvements equivalent to the performance ratios stated.

Consult the Large System Performance Reference (LSPR) when you consider performance on the z196. The range of performance ratings across the individual LSPR workloads is likely to have a large spread. More performance variation of individual logical partitions exists because the impact of fluctuating resource requirements of other partitions can be more pronounced with the increased numbers of partitions and additional PUs available.

For detailed performance information, see the LSPR Web site:

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

The MSU ratings are available from:

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

LSPR workload suite

Historically, LSPR workload capacity curves (primitives and mixes) have had application names or been identified by a *software* characteristic. For example, past workload names have included CICS®, IMS™, OLTP-T, CB-L, LoIO-mix[6] and TI-mix[7]. However, capacity performance has always been more closely associated with how a workload uses and interacts with a particular processor *hardware* design. With the availability of CPU MF (SMF 113) data on z10, the ability to gain insight into the interaction of workload and hardware design in production workloads has arrived. The knowledge gained is still evolving, but the first step in the process is to produce LSPR workload capacity curves based on the underlying hardware sensitivities. Thus the LSPR introduces three new workload capacity categories which replace all prior primitives and mixes.

Fundamental components of workload capacity performance

Workload capacity performance is sensitive to three major factors: instruction path length, instruction complexity, and memory hierarchy. Let us examine each of these three.

Instruction path length

A transaction or job will need to execute a set of instructions to complete its task. These instructions are composed of various paths through the operating system, subsystems and application. The total count of instructions executed across these software components is referred to as the transaction or job path length. Clearly, the path length will be different for each transaction or job depending on the complexity of the task(s) that must be performed. For a particular transaction or job, the application path length tends to stay the same presuming the transaction or job is asked to perform the same task each time. However, the path length associated with the operating system or subsystem may vary based on a number of factors including:

► Competition with other tasks in the system for shared resources – as the total number of tasks grows, more instructions are needed to manage the resources
► The Nway (number of logical processors) of the image or LPAR – as the number of logical processors grows, more instructions are needed to manage resources serialized by latches and locks.

Instruction complexity

The type of instructions and the sequence in which they are executed will interact with the design of a micro-processor to affect a performance component we can define as "instruction complexity." There are many design alternatives that affect this component some of which are:

► Cycle time (GHz)
► Instruction architecture
► Pipeline
► Superscalar
► Out-of-order execution
► Branch prediction

As workloads are moved between micro-processors with different designs, performance will likely vary. However, once on a processor this component tends to be quite similar across all models of that processor.

---

[6] Low I/O Content Mix Workload
[7] Transaction Intensive Mix Workload

Memory hierarchy and "nest"

The memory hierarchy of a processor generally refers to the caches, data buses, and memory arrays that stage the instructions and data needed to be executed on the micro-processor to complete a transaction or job. There are many design alternatives that affect this component such as:

► Cache size
► Latencies (sensitive to distance from the micro-processor)
► Number of levels, MESI (management) protocol, controllers, switches, number and Bandwidth of data buses and others

Some of the cache(s) are "private" to the micro-processor which means only that micro-processor may access them. Other cache(s) are shared by multiple micro-processors. We will define the term memory "nest" for a System z processor to refer to the shared caches and memory along with the data buses that interconnect them.

Workload capacity performance will be quite sensitive to how deep into the memory hierarchy the processor must go to retrieve the workload's instructions and data for execution. Best performance occurs when the instructions and data are found in the cache(s) nearest the processor so that little time is spent waiting prior to execution; as instructions and data must be retrieved from farther out in the hierarchy, the processor spends more time waiting for their arrival.

As workloads are moved between processors with different memory hierarchy designs, performance will vary as the average time to retrieve instructions and data from within the memory hierarchy will vary. Additionally, once on a processor this component will continue to vary significantly as the location of a workload's instructions and data within the memory hierarchy is affected by many factors including, but not limited to:

► Locality of reference
► IO rate
► Competition from other applications and/or LPARs

Relative nest intensity

The most performance sensitive area of the memory hierarchy is the activity to the memory nest, namely, the distribution of activity to the shared caches and memory. We introduce a new term, "Relative Nest Intensity (RNI)" to indicate the level of activity to this part of the memory hierarchy. Using data from CPU MF, the RNI of the workload running in an LPAR may be calculated. The higher the RNI, the deeper into the memory hierarchy the processor must go to retrieve the instructions and data for that workload.

Many factors influence the performance of a workload. However, for the most part what these factors are influencing is the RNI of the workload. It is the interaction of all these factors that result in a net RNI for the workload which in turn directly relates to the performance of the workload.

It should be emphasized that these are simply tendencies and not absolutes. For example, a workload may have a low IO rate, intensive CPU use, and a high locality of reference – all factors that suggest a low RNI. But, what if it is competing with many other applications within the same LPAR and many other LPARs on the processor which tend to push it toward a higher RNI? It is the net effect of the interaction of all these factors that determines the RNI of the workload which in turn greatly influences its performance.

Note that there is little one can do to affect most of these factors. An application type is whatever is necessary to do the job. Data reference pattern and CPU usage tend to be inherent in the nature of the application. LPAR configuration and application mix are mostly a

function of what needs to be supported on a system. IO rate can be influenced somewhat through buffer pool tuning.

However, one factor that can be affected, s**oftware configuration tuning,** is often overlooked but can have a direct impact on RNI. Here we refer to the number of address spaces (such as CICS AORs or batch initiators) that are needed to support a workload. This factor has always existed but its sensitivity is higher with today's high frequency microprocessors. Spreading the same workload over a larger number of address spaces than necessary can raise a workload's RNI as the working set of instructions and data from each address space increases the competition for the processor caches. Tuning to reduce the number of simultaneously active address spaces to the proper number needed to support a workload can reduce RNI and improve performance. In the LSPR, we tune the number of address spaces for each processor type and Nway configuration to be consistent with what is needed to support the workload. Thus, the LSPR workload capacity ratios reflect a presumed level of software configuration tuning. This suggests that re-tuning the software configuration of a production workload as it moves to a bigger or faster processor may be needed to achieve the published LSPR ratios.

LSPR Workload Categories Based on Relative Nest Intensity

As discussed above, a workload's relative nest intensity is the most influential factor that determines workload performance. Other more traditional factors such as application type or IO rate have RNI tendencies, but it is the net RNI of the workload that is the underlying factor in determining the workload's capacity performance. With this in mind, the LSPR now runs various combinations of former workload primitives such as CICS, DB2, IMS, OSAM, VSAM, WebSphere®, COBOL and utilities to produce capacity curves that span the typical range of RNI. The three new workload categories represented in the LSPR tables are described below:

- ► **LOW** (relative nest intensity):
  - A workload category representing light use of the memory hierarchy. This would be similar to past high scaling primitives.
- ► **AVERAGE** (relative nest intensity):
  - A workload category representing average use of the memory hierarchy.   This would be similar to the past LoIO-mix workload and is expected to represent the majority of production workloads.
- ► **HIGH** (relative nest intensity):
  - A workload category representing heavy use of the memory hierarchy.   This would be similar to the past TI-mix workload.

Relating Production Workloads to LSPR Workloads

Historically, there have been a number of techniques used to match production workloads to LSPR workloads such as:

- ► Application name (a customer running CICS would use the CICS LSPR workload)
- ► Application type (create a mix of the LSPR online and batch workloads)
- ► IO rate (low IO rates used a mix of the low IO rate LSPR workloads)

However, as discussed in the "LSPR Workload Categories" section, the underlying performance sensitive factor is how a workload interacts with the processor hardware. These past techniques were simply trying to approximate the hardware characteristics that were not available through software performance reporting tools. Beginning with the z10 processor, the hardware characteristics can now be measured using CPU MF (SMF 113) COUNTERS data. Thus, the opportunity exists to be able to match a production workload to an LSPR workload category via these hardware characteristics (see the "LSPR Workload Categories" section for a discussion about RNI – Relative Nest Intensity).

The AVERAGE RNI LSPR workload is intended to match the majority of customer workloads. When no other data is available, it should be used for a capacity analysis.

DASD IO rate has been used for many years to separate workloads into two categories: those whose DASD IO per MSU (adjusted) is <30 (or DASD IO per PCI <5) and those higher than these values. The majority of production workloads fell into the "low IO" category and a LoIO-mix workload was used to represent them. Using the same IO test, these workloads would now use the AVERAGE RNI LSPR workload. Workloads with higher IO rates may use the HIGH RNI workload or the AVG-HIGH RNI workload that is included with zPCR.

For z10 and newer processors, the CPU MF data may be used to provide an additional "hint" as to workload selection. When available, this data allows the RNI for a production workload to be calculated. Using the RNI and another factor from CPU MF, the L1MP (percentage of data and instruction references that miss the L1 cache), a workload may be classified as LOW, AVERAGE or HIGH RNI. This classification and resulting "hint" is automated in the zPCR tool. It is highly recommended to use zPCR for capacity sizing.

The LSPR workloads, updated for z196 are considered to reasonably reflect current and growth workloads of the customer. The set contains three generic workload categories based on z/OS R1V11 supporting up to 80 processors in a single image.

Workload performance variation

Because of the nature of the z196 multi-book system and resource management across those books, performance variability from application to application, similar to that seen on the z9 EC and z10 EC, is expected. This variability can be observed in several ways. The range of performance ratings across the individual workloads is likely to have some spread, but not as large as with the z10 EC.

The new memory and cache designs affect different workloads in a number of ways. All workloads are improved, with cache-intensive loads benefiting the most. When comparing moving from z9 EC to z10 EC with moving from z10 EC to z196, it is likely that the relative benefits per workload are different. Those workloads which benefited more than the average when moving from z9 EC to z10 EC will benefit less than the average when moving from z10 EC to z196, and vice-versa.

The customer impact of this variability is seen as increased deviations of workloads from single-number metric-based factors such as MIPS, MSUs, and CPU time charge back algorithms.

1.8 Operating systems and software

The z196 is supported by a large set of software, including ISV applications. This section lists only the supported operating systems. Exploitation of some features might require the latest releases. Further information is contained in Chapter 8, "Software support" on page 209.

The z196 supports any of the following operating systems:

- ► z/OS Version 1 Release 10 and later releases
- ► z/OS Version 1 Release 7 with IBM Lifecycle Extension
- ► z/OS Version 1 Release 8 with IBM Lifecycle Extension
- ► z/OS Version 1 Release 9 with IBM Lifecycle Extension
- ► z/VM Version 5 Release 4 and later
- ► z/VSE Version 4 Release 1 and later
- ► z/TPF Version 1 Release 1
- ► Linux on System z distributions:

- Novell SUSE: SLES 10, and SLES 11[8]
- Red Hat: RHEL 5[9]

Operating system support for the zBX blades includes:

► AIX Version 5 Release 3 or later
► Linux on System x® (Statement of Direction)

Finally, a large software portfolio is available to the zEnterprise 196, including an extensive collection of middleware and ISV products that implement the most recent proven technologies.

With support for IBM WebSphere software, full support for SOA, Web services, J2EE, Linux, and Open Standards, the zEnterprise 196 is intended to be a platform of choice for integration of a new generation of applications with existing applications and data.

---

[8] SLES is the abbreviation for Novell SUSE Linux Enterprise Server.
[9] RHEL is the abbreviation for Red Hat Enterprise Linux.

2

CPC hardware components

This chapter introduces zEnterprise 196 (z196) hardware components along with significant features and functions with their characteristics and options. Our objective is to explain the z196 hardware building blocks and how these components interconnect from a physical point of view. This information can be useful for planning purposes and can help to define configurations that fit your requirements.

This chapter discusses the following topics:

2.1 Frames and cage

System z frames are enclosures built to Electronic Industry Association (EIA) standards. The z196 CPC server has two 42U EIA frames, shown in Figure 2-1. The two frames, A and Z, are bolted together and have positions for one CPC cage and a combination of I/O cages and I/O drawers:

► Frame A has the CPC cage at the top together with one of the following configurations at the bottom:

– One I/O cage
– One or two I/O drawers

► Frame Z can have one of the following configurations:

– No I/O cages and no I/O drawers
– Combinations of up to two I/O cages and up to two I/O drawers. An RPQ is available for installing a second I/O cage in the Z frame.

All books, including the distributed converter assemblies (DCAs) on the books and the cooling components, are located in the CPC cage in the top half of the A frame. Figures 2-1 and 2-2 show the front view of frame A (with four books installed) and frame Z.
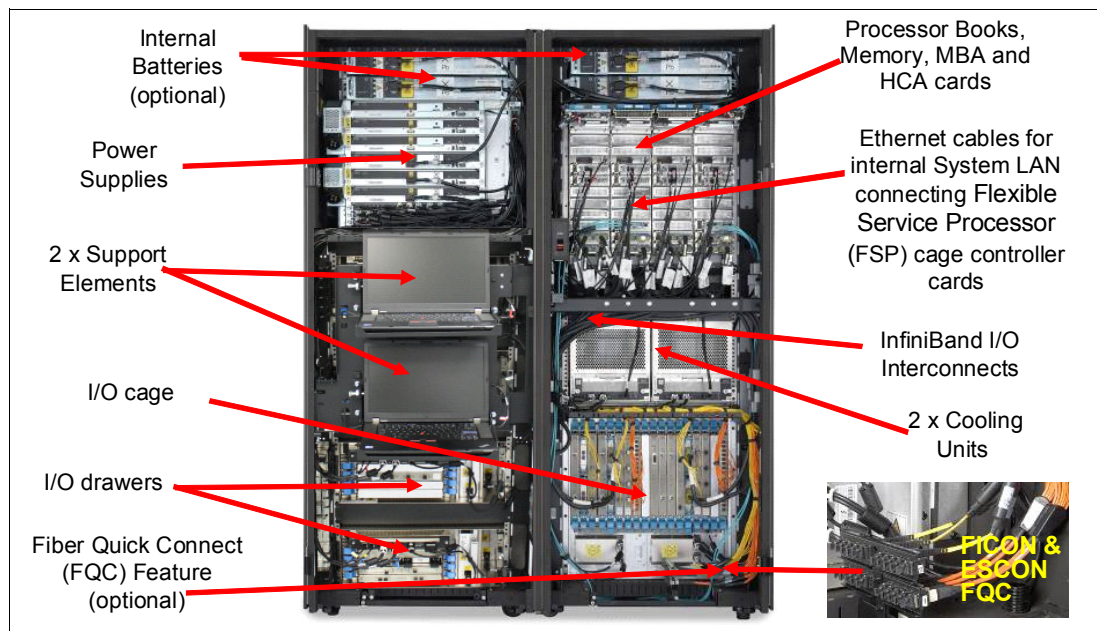
Figure 2-1   CPC cage, I/O drawers, and I/O cage locations - air-cooled system
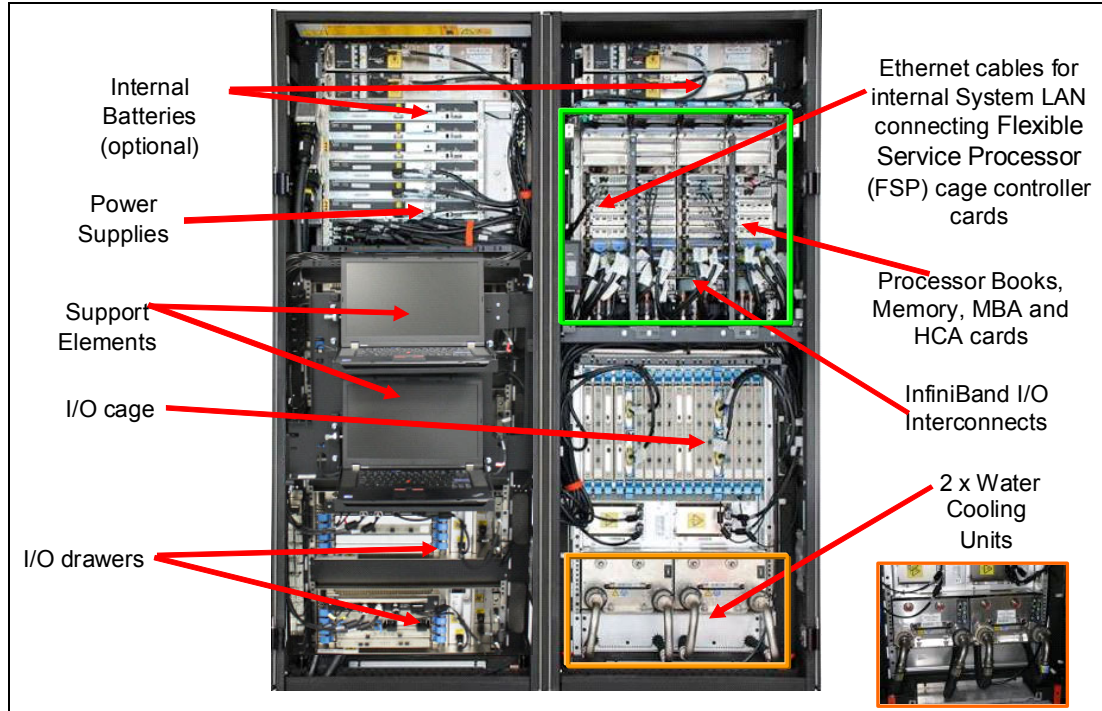
Figure 2-2   CPC cage, I/O drawers, and I/O cage locations - water cooled system

2.1.1 Frame A

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

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

The IBF further enhances the robustness of the power design, increasing power line disturbance immunity. It provides battery power to preserve processor data in case of a loss of power on all four AC feeds from the utility company. The IBF provides battery power to preserve full system function despite the loss of power at all system power cords. It allows continuous operation through intermittent losses, brownouts and power source switching or can provide time for an orderly shutdown in case of a longer outage.

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

Table 2-1   IBF estimated power time

Model	I/O configuration		
	One I/O cage	Two I/O cages	Three I/O cages
M15	9 minutes	10 minutes	10 minutes
M32	9 minutes	6 minutes	6 minutes
M49	6 minutes	4.5 minutes	4.5 minutes
M66	4.5 minutes	3.5 minutes	3.5 minutes
M80	4.5 minutes	3.5 minutes	3.5 minutes

The batteries are installed in pairs. Two to six battery units can be installed. The number is based on the z196 model and power requirements.

► Two modular refrigeration units (MRUs), which are air-cooled by their own internal cooling fans. For an air-cooled system

► In place of the MRUS, the customer can specify two Water Conditioning Unit (WCUs) connected to customer supplied water.

► CPC cage, which contains up to four books, connected to the appropriate cooling system.

► One or two I/O drawers each containing up to eight I/O cards of any type

► I/O cage, which can house all supported types of channel cards

  An I/O cage has 28 I/O card slots for installation of ESCON channels, FICON Express8 channels, OSA-Express2, OSA- Express3, and Crypto Express3 features. Up to two I/O cages are supported. If a third I/O cage is required order RPQ 8P2506.

► Air-moving devices (AMD), which provide N+1 redundant cooling for the fanouts, memory, and DCAs.

2.1.2 Frame Z

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

► Two optional Internal Battery Features (IBFs)

► Bulk Power Assemblies (BPAs)

► I/O cage 2 (bottom). The I/O cages can house all supported types of I/O features.

► Instead of, or in addition to, I/O cages, the customer can order one or two I/O drawers. Each contains up to eight I/O features of any type.

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

2.1.3 I/O cages and drawers

Each book has up to eight dual port fanouts for data transfer, each port with bidirectional bandwidth of 6 GBps. The HCA2 drives two ports. Up to 16 InfiniBand fanout connections provide an aggregated bandwidth of up to 96 GBps per book.

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

► ESCON channels (16 port cards, 15 usable ports, and one spare)

► FICON channels (FICON or FCP modes)

  – FICON Express4 channels (four port cards); carried forward during an upgrade only

  – FICON Express8 channels (four port cards)

► ISC-3 links (up to four coupling links, two links per daughter card). Two daughter cards (ISC-D) plug into one mother card (ISC-M).

► OSA-Express channels:

  – OSA-Express3 10 Gb Ethernet Long Reach and Short Reach (two ports per feature, LR and SR)

  – OSA-Express3 Gb Ethernet (four port cards, LX and SX)

  – OSA-Express3 1000BASE-T Ethernet (four port cards)

  – OSA-Express2 Gb Ethernet (two port cards, SX, LX, when carried forward on an upgrade)

– OSA-Express2 1000BASE-T Ethernet (two port card, when carried forward on an upgrade)

► Crypto Express3, with two PCI Express adapters per feature. A PCI Express adapter can be configured as a cryptographic coprocessor for secure key operations or as an accelerator for clear key operations.

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

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

2.1.4 Top exit I/O cabling

On z196 you now have the option of ordering the infrastructure to support top exit of your fiber optic cables (ESCON, FICON, OSA, 12x InfiniBand, 1x InfiniBand, and ISC-3) as well as your copper cables for the 1000BASE-T Ethernet features.

Top exit I/O cabling is designed to provide you with an additional option. Instead of all of your cables exiting under the server and/or under the raised floor, you now have the flexibility to choose the option that best meets the requirements of your data center.

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

2.2 Book concept

The central processor complex (CPC) uses a packaging concept for its processors based on books. A book contains a multi-chip module (MCM), memory, and connectors to I/O cages and other servers. Books are located in the CPC cage in frame A. The z196 has from one book to four books installed. A book and its components are shown in Figure 2-3 on page 28.
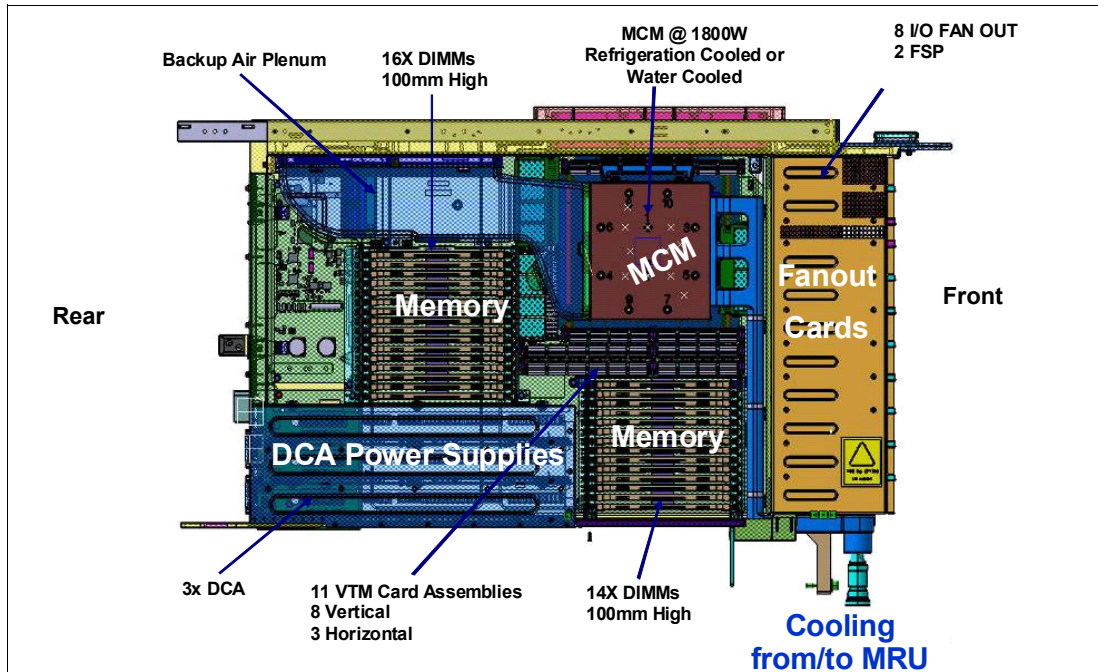
Figure 2-3   Book structure and components

Each book contains:

► One multi-chip module (MCM) with six quad-core microprocessor chips, having either 20 or 24 processor units (PUs), depending on the model, and two storage control chips with 192 MB of Level 4 cache.

► Memory DIMMs plugged into 30 available slots, providing from 60 GB to 960 GB of physical memory installed in a book.

► A combination of up to eight InfiniBand Host Channel Adapter (HCA2-Optical or HCA2-Copper) fanout cards. Each of the cards has two ports, thereby supporting up to 16 connections. HCA2-Copper connections are for links to the I/O cages in the server, and the HCA2-Optical connections are to external servers (coupling links).

► Three distributed converter assemblies (DCAs) that provide power to the book. Loss of a DCA leaves enough book power to satisfy the book's power requirements (2+1 redundancy). The DCAs can be concurrently maintained.

► Two flexible service processor (FSP) cards for system control.

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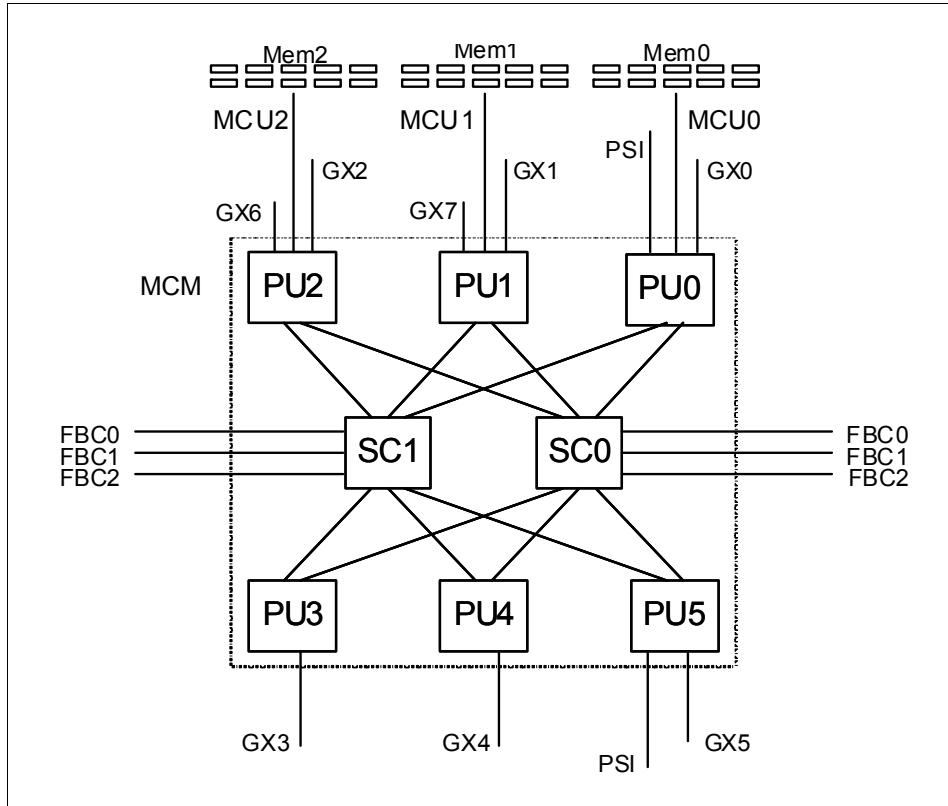
Figure 2-4   Book logical structure

Memory is connected to MCM through three memory control units (MCUs). GX0 to GX7 are the I/O buses interfaces to HCAs, with full store buffering, maximum of 10 GB/s per bus direction, and support added for PCIe.

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

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

2.2.1 Book interconnect topology

Figure 2-5 is showing the point-to-point topology for book communication. Each book communicates directly to all other books in the CPC.
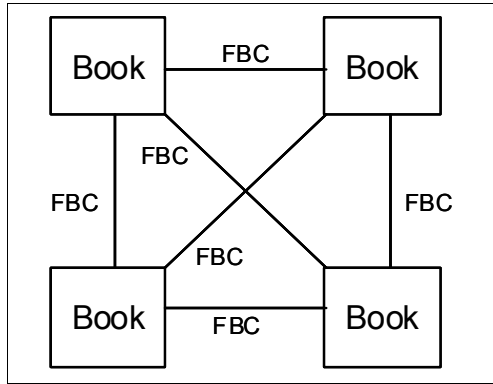
Figure 2-5   Book-to-book communication

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

The location of books is as follows:

► In a one-book model, the first book slides in the second slot from the left (CPC cage slot location LG06).

► In a two-book model, the second book slides in the right-most slot (CPC cage slot location LG15).

► In a three-book model, the third book slides in the third slot from the left (CPC cage slot location LG10).

► In a four-book model, the fourth book slides into the left-most slot (CPC cage slot location LG01).

Table 2-2 indicates the order of book installation and position in cage:

Table 2-2   Book installation order and position in cage

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

Book installation is concurrent, and concurrent book replacement requires a minimum of two books.

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

2.2.2 Dual external clock facility

Two external clock facility (ECF) cards are already installed and shipped with the server and provide a dual-path interface for pulse per second (PPS). This redundancy allows continued operation even if a single ECF card fails. This redundant design also allows concurrent maintenance. The two connectors to PPS output of an NTP server are located above the books and are on the mid-plane to which the books are connected.

Support exists for a simple network time protocol (SNTP) client on the support element. When server time protocol (STP) is used, the time of an STP-only coordinated timing network (CTN) can be synchronized with the time provided by a network time protocol (NTP) server, allowing a heterogeneous platform environment to have the same time source.

The time accuracy of an STP-only CTN is improved by adding an NTP server with the pulse per second output signal (PPS) as the external time source (ETS) device. ETS is available from several vendors that offer network timing solutions. A cable connection from the PPS port on the ECF card to the PPS output of the NTP server is required when the z196 is using STP and configured in an STP-only CTN using NTP with pulse per second as the external time source.

STP tracks the highly stable accurate PPS signal from the NTP server and maintains and accuracy of 10 μs as measured at the PPS input of the System z server.

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

ECF	OSC	OSC	ECF
Book 0	Book 1	Book 2	Book 3
LG01	LG06	LG10	LG15
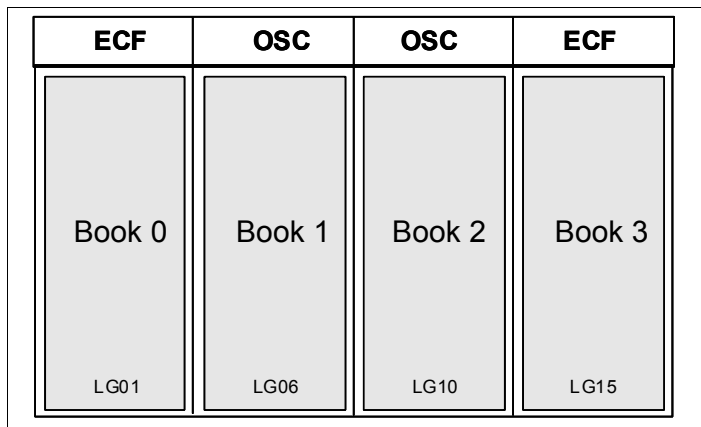
Figure 2-6   ECF and OSC cards

Figure 2-6 shows the location of the two ECF cards on the CPC, above book 0 and book 3 locations.

> **Note:** Server time protocol (STP) is available as FC 1021. STP is implemented in the licensed internal code (LIC) and is designed for multiple servers to maintain time synchronization with each other. See the following publications for more information:
>
> ► *Server Time Protocol Planning Guide*, SG24-7280
> ► *Server Time Protocol Implementation Guide*, SG24-7281

2.2.3 Oscillator

The z196 has two oscillator cards (OSC), a primary and a backup. Although not part of the book design, they are found above the books, connected to the same mid-plane to which the books are connected. If the primary fails, the secondary detects the failure, takes over transparently, and continues to provide the clock signal to the server.

Figure 2-6 shows the location of the two OSC cards on the CPC, above book 1 and book 2 locations.

2.2.4 System control

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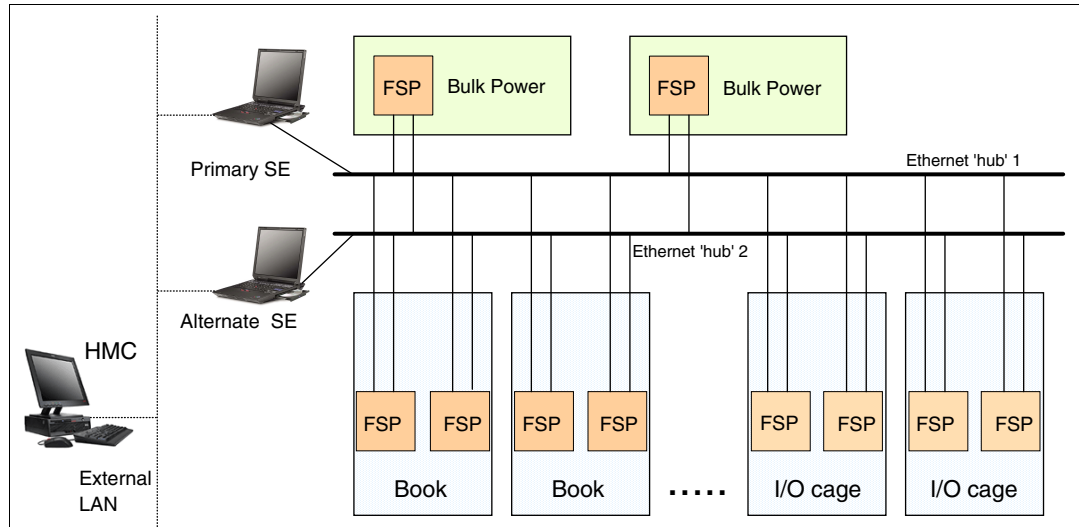
Figure 2-7   Conceptual overview of system control elements

A typical FSP operation is to control a power supply. An SE sends a command to the FSP to bring up the power supply. The FSP (using SSI connections) cycles the various components of the power supply, monitors the success of each step and the resulting voltages, and reports this status to the SE.

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

The SEs, in turn, are connected to one or two (external) LANs (Ethernet only), and the hardware management consoles (HMCs) are connected to these external LANs. One or more HMCs can be used, but two (a primary and an alternate[1]) are mandatory with an ensemble. Additional HMCs can operate a zEnterprise 196 server when it is not a member of an ensemble.

Note: The primary HMC and its alternate must be connected to the same subnetwork to allow the alternate HMC to take over the IP address of the primary HMC during failover processing.

If the zEnterprise 196 server is not a member of an ensemble, the controlling HMCs are stateless (they do not keep any system status) and therefore would not affect system operations if they are disconnected. At that time, the system can be managed from either SE.

However, if the zEnterprise 196 server is defined as a node of an ensemble, its HMC will be the authoritative owning (stateful) component for platform management, configuration, and policies that have a scope that spans all of the managed nodes (CPCs and zBXs) in the collection (ensemble). It will no longer simply be a console/access point for configuration and

---

[1] These HMCs must be running with Version 2.11 or above with feature codes 0090, 0025, 0019, and optionally 0020

policies that is owned by each of the managed CPCs. Related to this, it will also have an active role in ongoing system monitoring and adjustment. This requires the HMC to be paired with an active backup (alternate) HMC[2].

2.2.5 Book power

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

2.3 Multi-chip module

The multi-chip module (MCM) is a 103-layer glass ceramic substrate (size is 96 x 96 mm) containing eight chip sites and 7356 land grid array (LGA) connections. There are six processor unit (PU) chips and two storage control (SC) chips. Figure 2-8 illustrates the chip locations. The total number of transistors on all chips on the MCM is more than 11 billion.
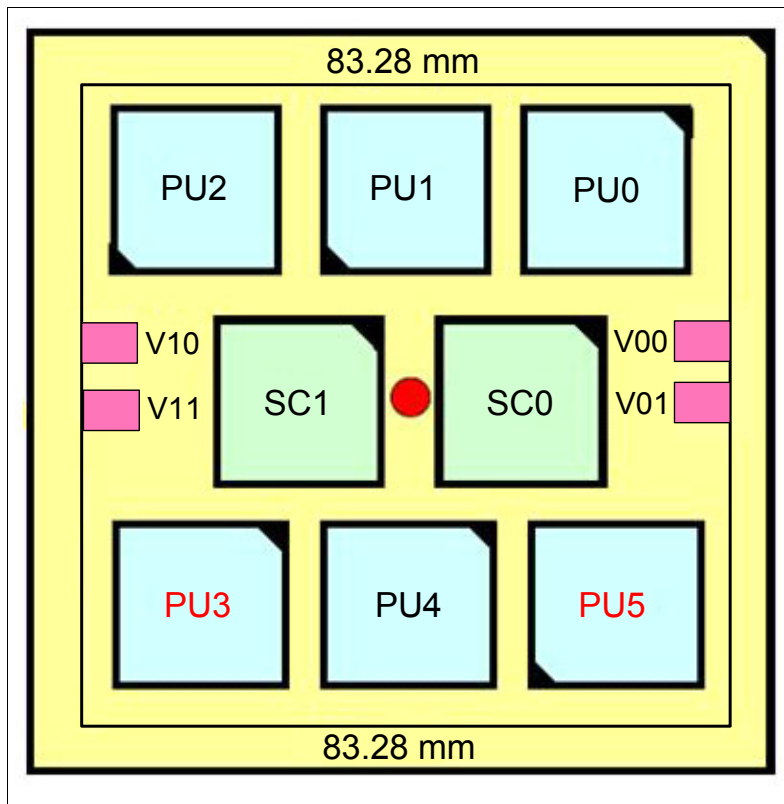
Figure 2-8   z196 Multi-Chip Module

---

[2] These HMCs must be running with Version 2.11 or above with feature codes 0090, 0025, 0019 and optionally 0020.

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

2.4 Processor units and storage control chips

Both processor unit (PU) and storage control (SC) chips on the MCM use CMOS 12s chip technology. CMOS 12s is state-of-the-art microprocessor technology based on 13-layer copper interconnections and silicon-on insulator (SOI) technologies. The chip lithography line width is 0.045 μm (45 nm). On the MCM, four serial electrically erasable programmable ROM (SEEPROM) chips are rewritable memory chips that hold data without power, use the same technology, and are used for retaining product data for the MCM and relevant engineering information.

Figure 2-9 is the MCM structure diagram, showing the PUs and SCs and their connections, which will be detailed on following sections.
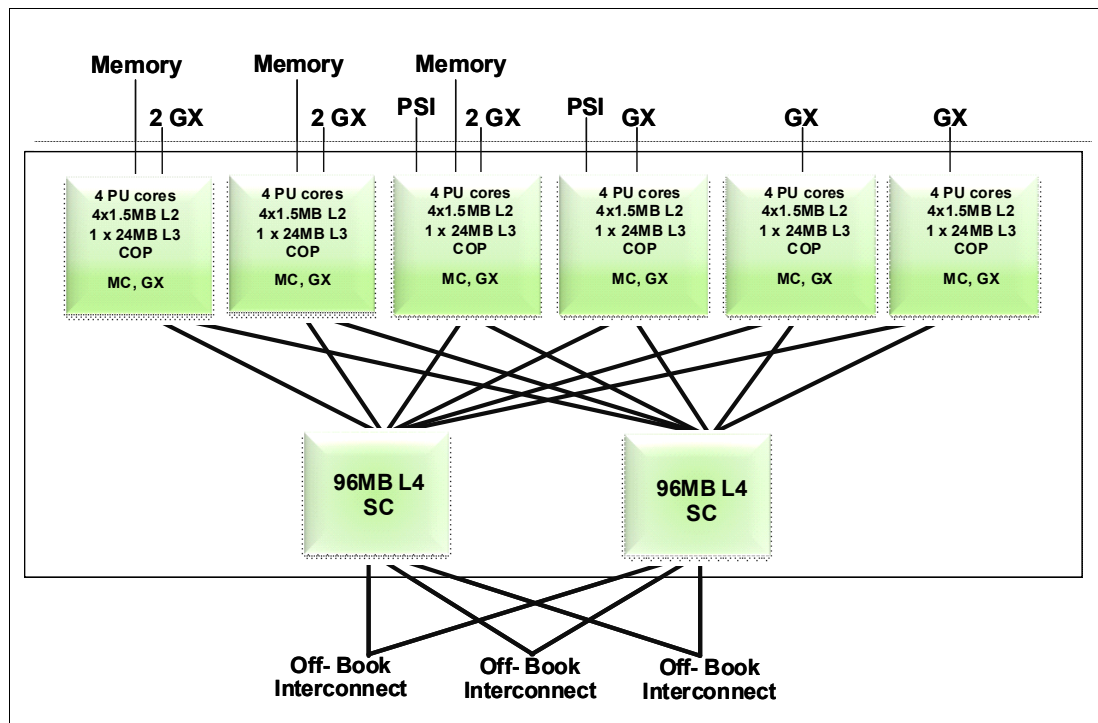
Figure 2-9   PU MCM structure

2.4.1 PU chip

The z196 PU chip is an aggressive derivative of the z10 core design, using 12s technology, out-of-order instruction processing, higher clock frequency, and larger caches. Compute intensive workloads can achieve additional performance improvements through compiler enhancements, and larger caches can improve system performance on many production workloads.

Each PU chip has up to four cores running at 5.2 GHz, which means a 0.19 ns cycle time. There are six PU chips on each MCM. The PU chips come in two versions, having three active cores or all four active cores. For models M15, M32, M49, and M66, the processor units

in the MCM in each book are implemented with a mix of four PU chips with three active cores (PU0, PU1, PU2, and PU4) and two PU chips with four active cores (PU3 and PU5), resulting in 20 active cores per MCM. This means that model M15 has 20, model M32 has 40, model M49 has 60, and model E66 has 80 active cores.

For the model M80, each processor chip has four active cores, resulting in 24 active cores per MCM. This means that there are 96 active cores on model M80.

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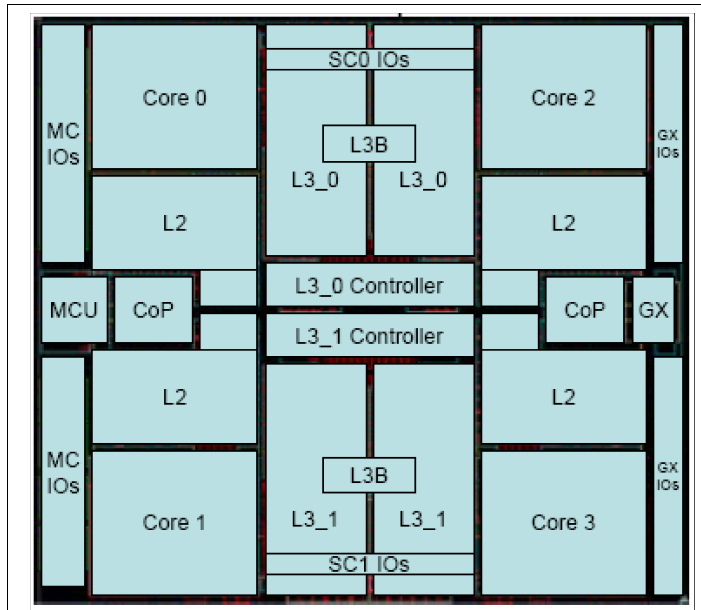
Figure 2-10   PU chip diagram

Each PU chip has 1.4 billion transistors. Each one of the four cores has its own L1 with 64 KB for instructions and 128 KB for data. Next to each core resides its private L2 cache, with 1.5 MB.

There is one L3 cache, with 24 MB. This 24 MB L3 cache is a store-in shared cache across all four cores in the PU chip. It has 192 512Kb eDRAM macros, dual address-sliced and dual store pipe support, an integrated on-chip coherency manager, cache and cross-bar switch. The L3 directory filters queries from local L4. Both L3 slices can deliver up to 160 GB/s bandwidth to each core simultaneously. The L3 cache interconnects the four cores, GX I/O buses, and memory controllers (MCs) with storage control (SC) chips.

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

There are also two co-processors (CoP) for data compression and encryption functions, each one shared by two cores. Figure 2-11 on page 36 shows the logical diagram of the compression and cryptographic coprocessor.
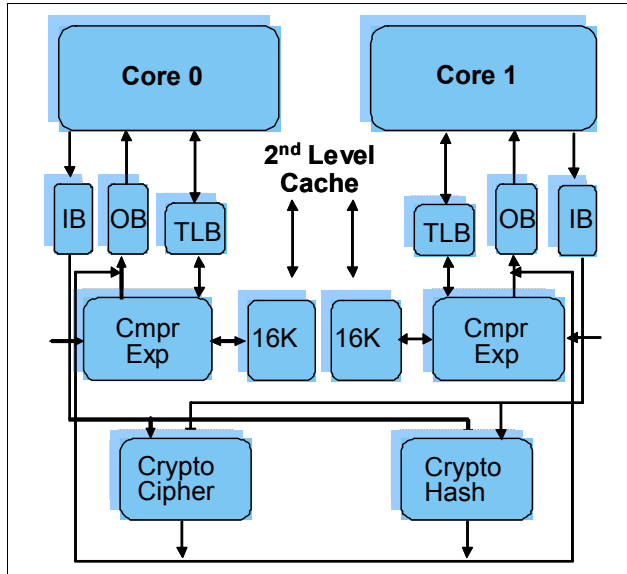
Figure 2-11   Compression and cryptographic coprocessor

The compression unit is integrated with the CP assist for cryptographic function (CPACF), benefiting from combining (or sharing) the use of buffers and interfaces. The assist provides high-performance hardware encrypting and decrypting support for clear key operations.

2.4.2 Processor unit (core)

Each processor unit, or core, is a superscalar, out of order processor, having six RISC-like execution units as follows:

► two fixed point (integer)

► two load/store

► one binary floating point

► one decimal floating point

Up to three instructions can be decoded per cycle and up to five instructions/operations can be executed per cycle. The instructions execution can occur out of program order, as well as memory address generation and memory accesses can also occur out of program order. Each core has special circuitry to make execution and memory accesses appear in order to software. There are 246 complex instructions executed by millicode and another 211 complex instructions cracked into multiple RISC like operations.

The following functional areas are on each core, as shown in Figure 2-12:

► Instruction sequence unit (ISU)

  This new unit (ISU) enables the out-of-order (OOO) pipeline. It keeps track of register names, OOO instruction dependency, and handling of instruction resource dispatch.

  This unit is also central to performance measurement through a function called instrumentation.

► Instruction fetch and branch (IFB) (prediction) and Instruction cache & merge (ICM)

  These two sub units (IFB and ICM) contain the instruction cache, branch prediction logic, instruction fetching controls, and buffers. Its relative size is the result of the elaborate

branch prediction design, which is further described in 3.3.2, "Superscalar processor" on page 73.

► Instruction decode unit (IDU)

The IDU is fed from the IFU buffers and is responsible for parsing and decoding of all z/Architecture operation codes.

► Load-store unit (LSU)

The LSU contains the data cache and is responsible for handling all types of operand accesses of all lengths, modes and formats as defined in the z/Architecture.

► Translation unit (XU)

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

► Fixed-point unit (FXU)

The FXU handles fixed point arithmetic.

► Binary floating-point unit (BFU)

The BFU handles all binary and hexadecimal floating-point, and fixed-point multiplication and division operations.

► Decimal unit (DU)

The DU executes both floating- point and fixed-point decimal operations.

► Recovery unit (RU)

The RU keeps a copy of the complete state of the system, including all registers, collects hardware fault signals, and manages the hardware recovery actions.
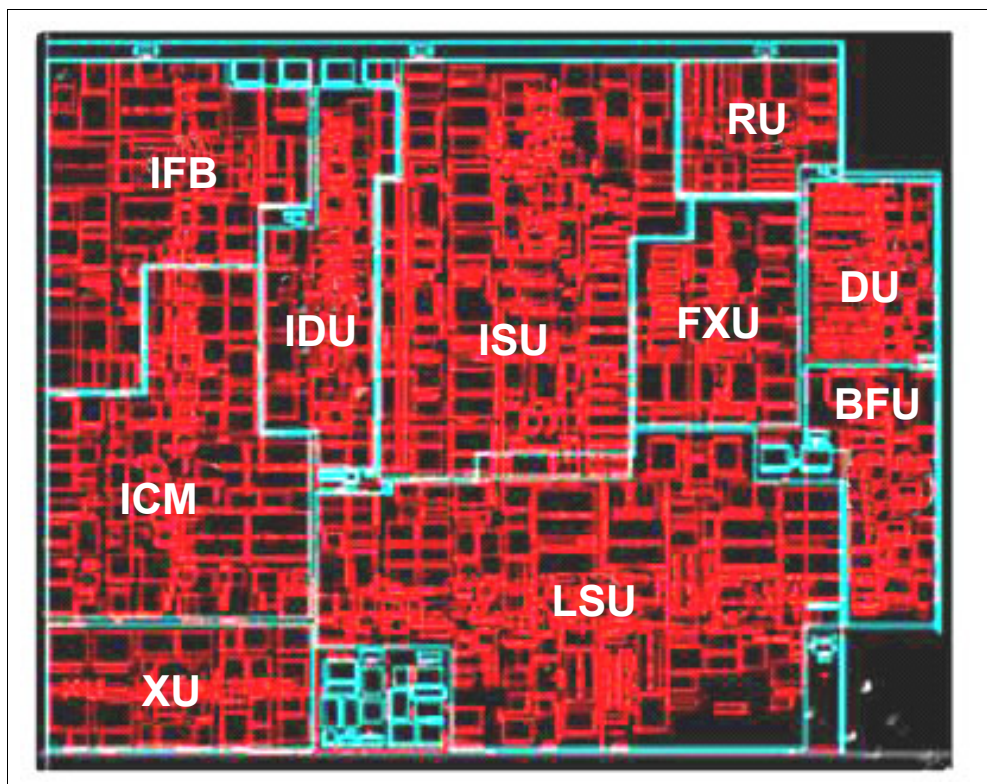
Figure 2-12   Core layout

2.4.3 PU characterization

In each MCM, some PUs may be characterized for customer use. The characterized PUs may be used for general purpose to run supported operating systems (as z/OS, z/VM, Linux on System z), or specialized to run specific workloads (as Java, XML services, IPSec, some DB2 workloads) or functions (as Coupling Facility Control Code). For more information about PU characterization, see Chapter 3.4, "Processor unit functions" on page 77.

The maximum number of characterized PUs depends on the z196 model. Some PUs are characterized by the system as standard system assist processors (SAPs), to run the I/O processing. Also as standard, there are at least two spare PUs per system, which are used to assume the function of a failed PU. The remaining installed PUs can be characterized for customer use. A z196 model nomenclature includes a number which represents this maximum number of PUs that can be characterized for customer use, as shown on Table 2-3.

Table 2-3   Number of PUs per z196 model

Model	Books	Installed PUs	Standard SAPs	Min Spare PUs	Max characterized PUs
M15	1	20 (1 x 20)	3	2	15
M32	2	40 (2 x 20)	6	2	32
M49	3	60 (3 x 20)	9	2	49
M66	4	80 (4 x 20)	12	2	66
M80	4	96 (4 x 24)	14	2	80

2.4.4 Storage control (SC) chip

The storage control (SC) chip uses the CMOS 12s 45nm SOI technology, with 13 layers of metal. It measures 24.4 x 19.6 mm, has 1.5 billion transistors and 1 billion cells for eDRAM. Each MCM has two SC chips. The L4 cache on each SC chip has 96 MB, resulting on 192 MB of L4 cache shared per book.

Figure 2-13 shows a schematic representation of the SC chip with its elements.
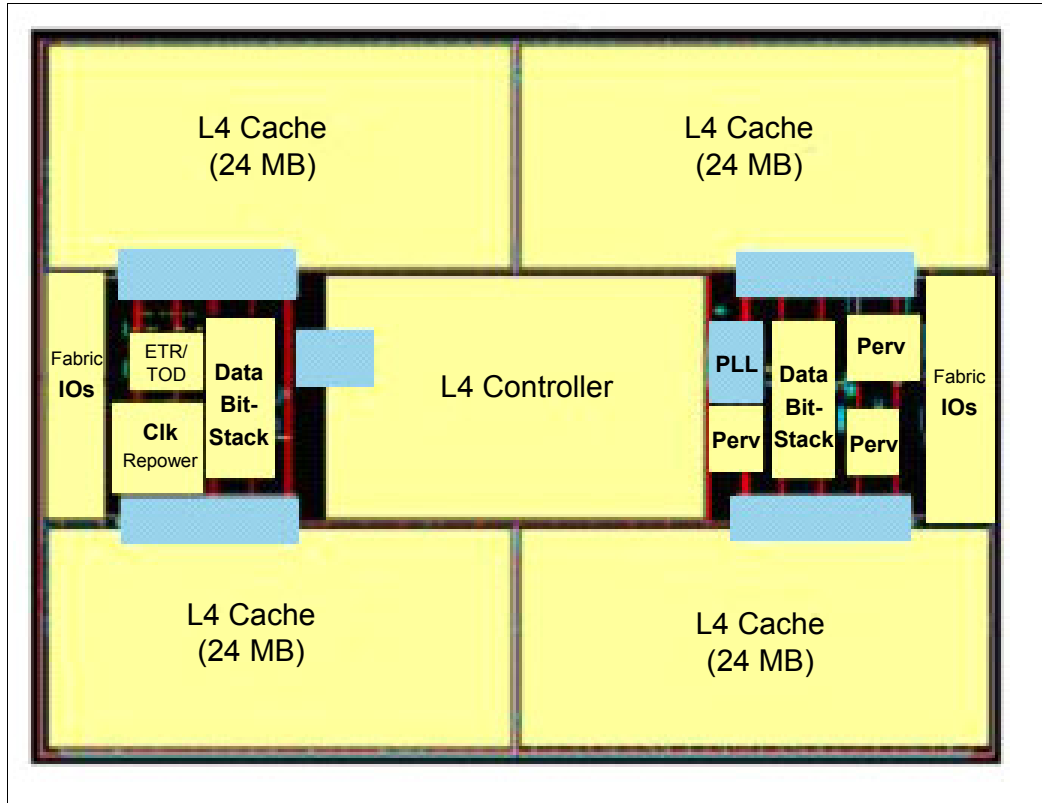
Figure 2-13   SC chip diagram

Most of the space is taken by the L4 controller and the L4 cache, which consists of four 24 MB eDRAM, a 16-way cache banking, 24-way set associative and a single pipeline design with split address-sliced directories. There are 768 1 MB eDRAM macros and eight 256 B cache banks per logical directory.

The L3 caches on PU chips communicate with the L4 caches on SC chips by six bidirectional data buses. The bus/clock ratio between the L4 cache and the PU is controlled by the storage controller on the SC chip.

The SC chip also acts as an L4 cache cross-point switch for L4-to-L4 traffic to up to three remote books by three bidirectional data buses. The integrated SMP fabric transport and system coherency manager use the L4 directory to filter snoop traffic from remote books, with a enhanced synchronous fabric protocol for improved latency and cache management. There are two clock domains and the clock function is distributed among both SC chips.

2.4.5 Cache level structure

The z196 server implements a four level cache structure, as shown on Figure 2-14.
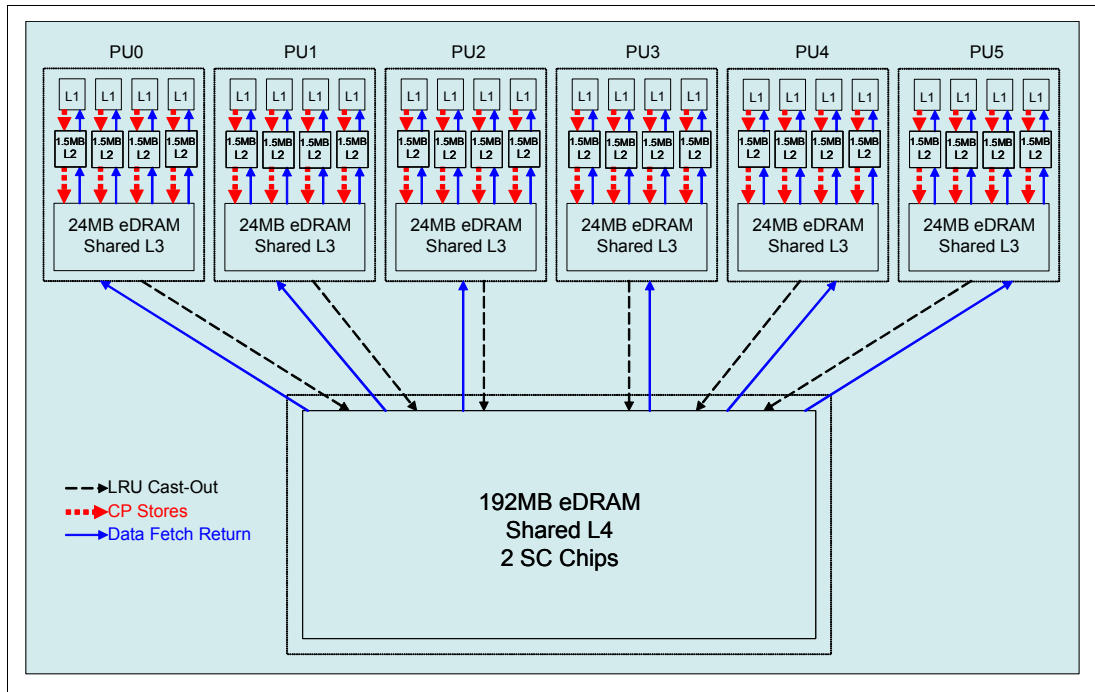
Figure 2-14   Cache levels structure

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

The next level is the private cache Level 2 (L2) located on each core, having 1.5 MB and also designed as a store-through cache.

The cache Level 3 (L3) is also located on the PUs chip and shared by the four cores, having 24 MB and designed as a store-in cache.

Cache levels L2 and L3 are implemented on the PU chip to reduce the latency between the processor and the large shared cache L4, which is located on the two SC chips. Each SC chip has 96 MB, resulting in 192 MB of L4 cache, which is shared by all PUs on the MCM. The L4 cache uses a store-in design.

2.5 Memory

Maximum physical memory size is directly related to the number of books in the system. Each book may contain up to 960 GB of physical memory, for a total of 3840 GB (3.75 TB) of installed memory per system.

A z196 server has more memory installed than ordered. Part of the physical installed memory is used to implement the redundant array of independent memory (RAIM) design, resulting on up to 768 GB of available memory per book and up to 3072 GB (3 TB) per system.

Table 2-4 shows the maximum and minimum memory sizes customer can order for each z196 model.

Table 2-4   z196 servers memory sizes

Model	Number of books	Customer memory (GB)
M15	1	32 - 704
M32	2	32 - 1520
M49	3	32 - 2288
M66	4	32 - 3056
M80	4	32 - 3056

The minimum physical installed memory is 40 GB per book except for the model M15, and the minimum initial amount of memory that can be ordered is 32 GB for all z196 models. The maximum customer memory size is based on the physical installed memory minus RAIM and minus HSA memory.

Table 2-5 shows the memory granularity based on the installed customer memory.

Table 2-5   Memory granularity

Granularity (GB)	Customer memory (GB)
32	32 - 256
64	320 - 512
96	608 - 896
112	1008
128	1136 - 1520
256	1776 - 3056

On z196 servers the memory granularity varies from 32 GB (for customer memory sizes from 32 to 256 GB) up to 256 GB (for servers having from 1776 GB to 3056 GB of customer memory).

2.5.1 Memory subsystem topology

The z196 memory subsystem uses high speed, differential ended communications memory channels to link a host memory to the main memory storage devices.

Figure 2-15 shows an overview of the book memory topology of a z196 server.
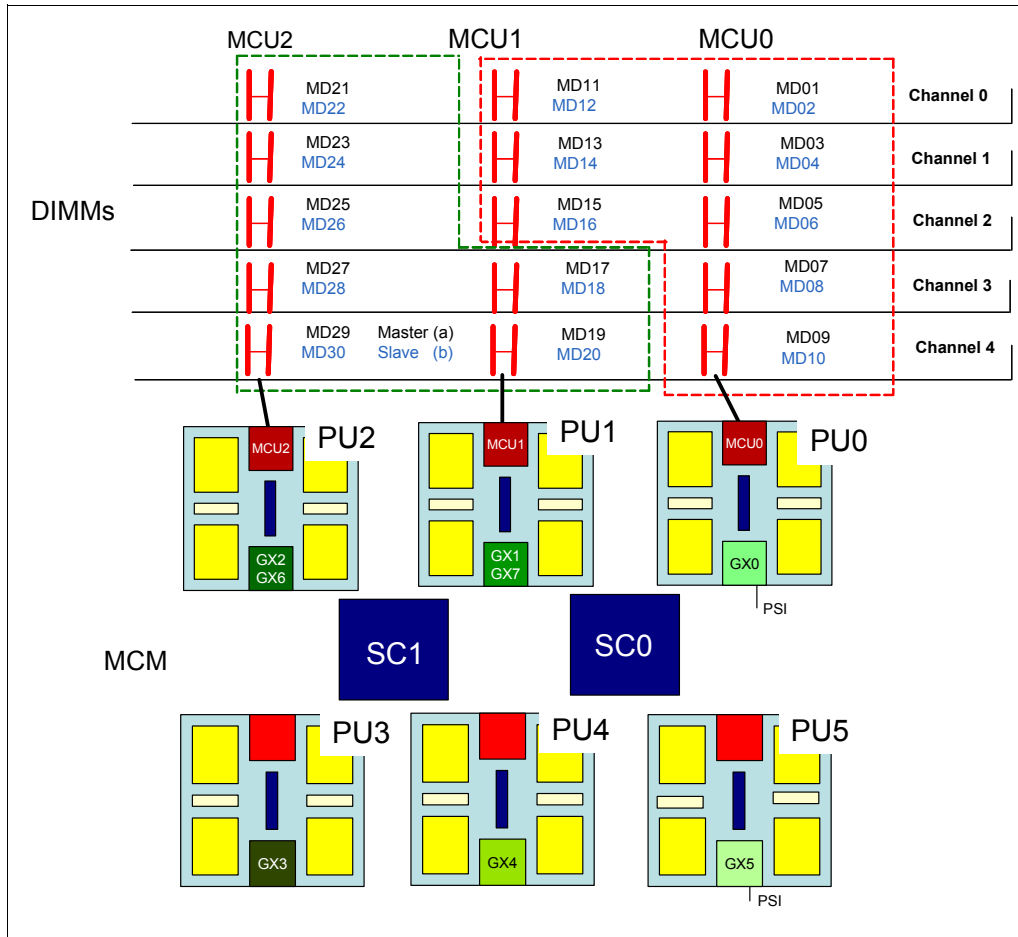
Figure 2-15   Book memory topology

Each book has from 10 to 30 dual in-line memory modules (DIMMs). DIMMs are connected to the MCM through three memory control units (MCUs) located on PU0, PU1 and PU2. Each MCU uses five channels, one of them for RAIM implementation, on a 4 +1 (parity) design. Each channel has one or two chained DIMMs, so a single MCU can have five or ten DIMMs. Each DIMM has 4, 16 or 32 GB, and there is no mixing of DIMM sizes on a book.

2.5.2 Redundant array of independent memory (RAIM)

z196 introduces the redundant array of independent memory (RAIM). The RAIM design detects and recovers from DRAM, socket, memory channel or DIMM failures.

The RAIM design requires the addition of one memory channel that is dedicated for RAS, as shown on Figure 2-16.
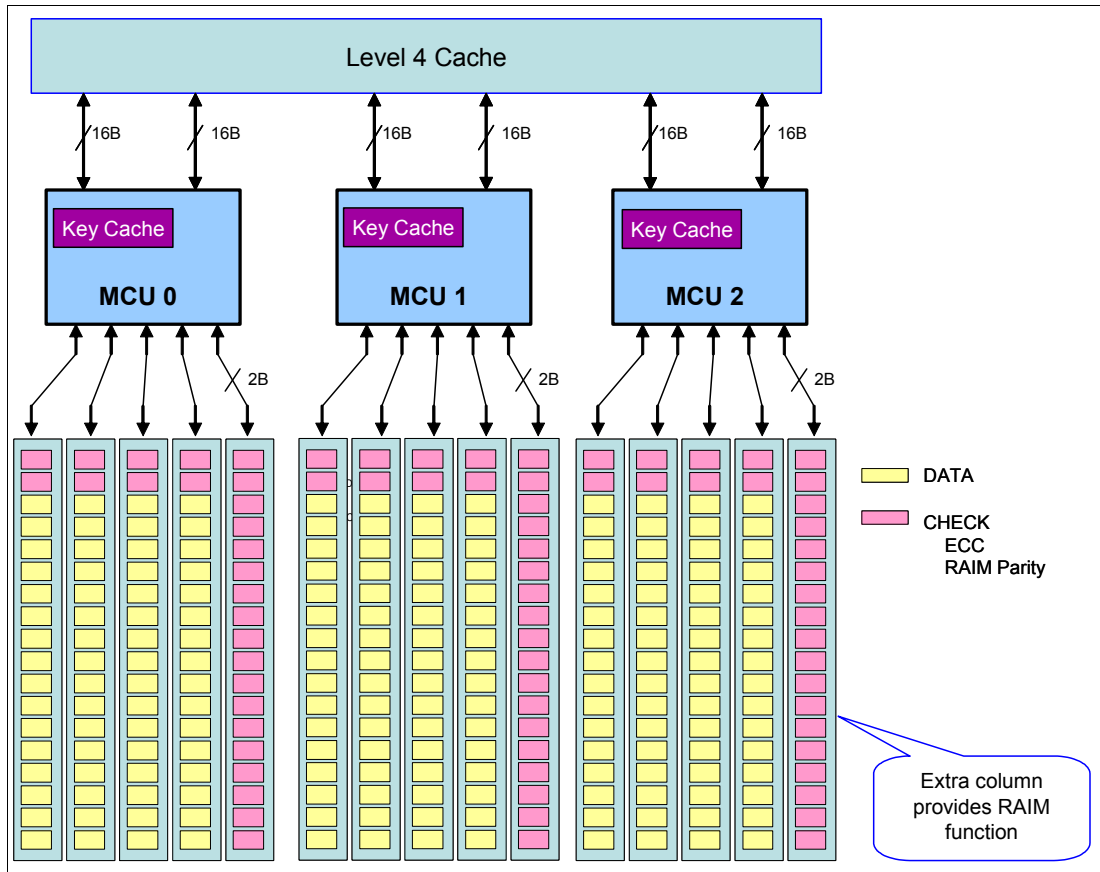
Figure 2-16   RAIM DIMMs

The parity of the four "data" DIMMs are stored in the DIMMs attached to the fifth memory channel. Any failure in a memory component can be detect and corrected dynamically. This design takes the RAS of the memory subsystem to another level, making it essentially a fully fault tolerant "N+1" design.

2.5.3 Memory configurations

Memory sizes in each book do not have to be similar. Different books can contain different amounts of memory. Table 2-6 shows the physically installed memory on each book for all z196 server models.

Table 2-6   Physically installed memory

Memory	Model M15	Model M32		Model M49			Model M66 Model M80			
(GB)	Book 1	Book 1	Book 3	Book 1	Book 2	Book 3	Book 0	Book 1	Book 2	Book 3
32	60[a]	40	40	40	40	40	40	40	40	40
64	100	60	40	40	40	40	40	40	40	40
96	160	80	60	60	40	40	40	40	40	40
128	240	100	80	60	60	60	60	40	40	40
160	240	120	100	80	80	60	60	60	60	40

Memory	Model M15	Model M32		Model M49			Model M66 Model M80			
(GB)	Book 1	Book 1	Book 3	Book 1	Book 2	Book 3	Book 0	Book 1	Book 2	Book 3
192	320	160	120	100	80	80	80	60	60	60
224	320	160	160	100	100	100	80	80	80	60
256	400	240	160	120	120	100	100	80	80	80
320	480	240	240	160	160	100	120	100	100	100
384	640	320	240	240	160	160	160	120	120	100
448	640	320	320	240	240	160	160	160	160	100
512	800	400	320	240	240	240	240	160	160	160
608	800	400	400	320	240	240	240	240	160	160
704	960	480	480	320	320	320	240	240	240	240
800	N/A	640	480	400	320	320	320	240	240	240
896	N/A	640	640	400	400	400	320	320	320	240
1008	N/A	640	640	480	400	400	320	320	320	320
1136	N/A	800	640	480	480	480	400	400	320	320
1264	N/A	800	800	640	480	480	400	400	400	400
1392	N/A	960	800	640	640	480	480	480	400	400
1520	N/A	960	960	640	640	640	480	480	480	480
1776	N/A	N/A	N/A	800	800	640	640	640	480	480
2032	N/A	N/A	N/A	960	800	800	640	640	640	640
2288	N/A	N/A	N/A	960	960	960	800	800	640	640
2544	N/A	N/A	N/A	N/A	N/A	N/A	800	800	800	800
2800	N/A	N/A	N/A	N/A	N/A	N/A	960	960	800	800
3056	N/A	N/A	N/A	N/A	N/A	N/A	960	960	960	960

a. 60 GB for a one book system. However if System is ordered with >1 book, 40 GB installed

Physically, memory is organized as follows:

► A book always contains a minimum of 10 DIMMs of 4 GB each (40 GB).
► A book has more memory installed than enabled. The amount of memory that can be enabled by the customer is the total physically installed memory minus the RAIM amount and minus the 16 GB HSA memory.
► A book may have available unused memory, which can be ordered on a memory upgrade.

Figure 2-17 illustrates how the physical installed memory is allocated on a z196 server, showing HSA memory, RAIM, customer memory, and the remaining available unused memory that can be enabled by a licensed internal code (LIC) code load when required.
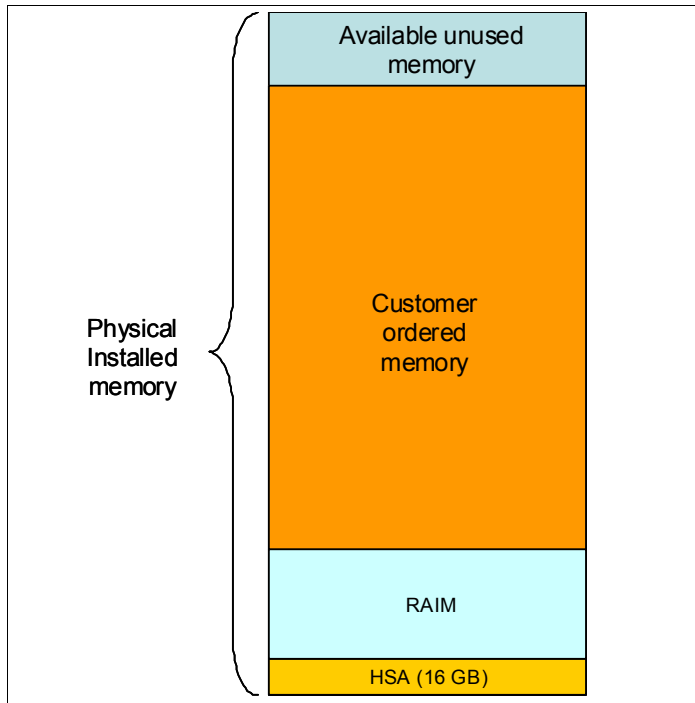
Figure 2-17   Memory allocation diagram

As an example, a z196 server model M32 (two books) ordered with 192 GB of memory would have memory sizes as follows. Refer to Figure 2-17.

► Physical installed memory is 280 GB: 160 GB on book 1 and 120 GB on book 3.

► Book 1 has the 16 GB HSA memory and up to 112 GB for customer memory, while book 2 has up to 96 GB for customer memory, resulting in 208 GB of available memory for the customer.

► As customer ordered 192 GB, provided the granularity rules are met, there is still 16 GB (208 - 192 GB) available to be used in conjunction with additional memory for future upgrades by LIC.

Memory upgrades are satisfied from already installed unused memory capacity until it is exhausted. When no more unused memory is available from the installed memory cards (DIMMs), one of the following additions must occur:

– Memory cards have to be upgraded to a higher capacity.
– An additional book with additional memory is necessary.
– Memory cards (DIMMs) must be added.

A memory upgrade is concurrent when it requires no change of the physical memory cards. A memory card change is disruptive when no use is made of enhanced book availability. See 2.7.2, "Enhanced book availability" on page 51.

When activated, a logical partition can use memory resources located in any book. No matter in which book the memory resides, a logical partition has access to that memory for up to a maximum of 1 TB. Despite the book structure, the z196 is still a symmetric multiprocessor (SMP). For more information see Chapter 3.6, "Logical partitioning" on page 93.

2.5.4 Memory upgrades

For a model upgrade that results in the addition of a book, the minimum memory increment is added to the system. As previously mentioned, the minimum physical memory size in a book is 40 GB. During a model upgrade, the addition of a book is a concurrent operation. The addition of the physical memory that is in the added book is also concurrent. If all or part of the additional memory is enabled for installation use (if it has been purchased), it becomes available to an active logical partition if this partition has reserved storage defined. For more information, see 3.6.2, "Reserved storage" on page 101. Alternately, additional memory may be used by an already-defined logical partition that is activated after the memory addition.

2.5.5 Book replacement and memory

With enhanced book availability as supported for z196 (see 2.7.2, "Enhanced book availability" on page 51), sufficient resources must be available to accommodate resources that are lost when a book is removed for upgrade or repair. Most of the time, removal of a book results in removal of active memory. With the flexible memory option (see 2.5.6, "Flexible memory option" on page 46), evacuating the affected memory and reallocating its use elsewhere in the system is possible. This requires additional available memory to compensate for the memory lost with the removal of the book.

2.5.6 Flexible memory option

With the flexible memory option, additional physical memory is supplied to support activation of the actual purchased memory entitlement in the event of a single book failure, or to be available during an enhanced book availability action.

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

Flexible memory is available only on the M32, M49, M66, and M80 models. Table 2-7 shows the flexible memory sizes available for z196 servers.

Table 2-7　 z196 servers memory sizes

Model	Standard memory (GB)	Flexible memory (GB)
M15	32 - 704	N/A
M32	32 - 1520	32 - 704
M49	32 - 2288	32 - 1520
M66	32 - 3056	32 - 2288
M80	32 - 3056	32 - 2288

Table 2-8 is showing the memory granularity for the flexible memory option.

Table 2-8　 Flexible memory granularity

Granularity (GB)	Flexible memory (GB)
32	32 - 256
64	320 - 512

Granularity (GB)	Flexible memory (GB)
96	608 - 896[a]
112	1008
128	1136 - 1520[b]
256	1776 - 2288

a. Model M32 limit is 704 GB
b. Model M49 limit is 1520 GB

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

2.5.7 Plan-ahead memory

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

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

► The IBM Resource Link™ (login is required):

http://www.ibm.com/servers/resourcelink/

► An IBM representative

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

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

Table 2-9   Feature codes for plan-ahead memory

Memory	z196 feature code
Pre-planned memory Charged when physical memory is installed. Used for tracking the quantity of physical increments of plan-ahead memory capacity.	FC1996
Pre-planned memory activation Charged when plan-ahead memory is enabled. Used for tracking the quantity of increments of plan-ahead memory that is being activated.	FC1997

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

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

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

2.6 Reliability, availability, serviceability (RAS)

IBM System z continues to deliver enterprise RAS with the IBM zEnterprise 196. Patented error correction technology in the memory subsystem provides IBM's most robust error correction to date. Two full DRAM failures per rank can be spared and a third full DRAM failure corrected. DIMM level failures, including components such as the controller ASIC, the power regulators, the clocks and the board, can be corrected. Channel failures such as signal lines, control lines and drivers/receivers on the MCM can be corrected. Up stream and down stream data signals can be spared using two spare wires on both the upstream and downstream paths. One of these signals can be used to spare a clock signal line (one up stream and one down stream). Taken together this provides System z's strongest memory subsystem.

The IBM zEnterprise 196 family of servers has improved chip packaging (encapsulated chip connectors) and is uses soft error rate (SER) hardened latches throughout the design.

z196 introduces fully fault protected N+2 voltage transformation module (VTM) power conversion in the processor book. This redundancy protects processor workloads from loss of voltage due to VTM failures. System z uses triple redundancy on the environmental sensors (humidity and altitude) for reliability.

System z delivers robust server designs through exciting new technologies, hardening and classic redundancy.

2.7 Connectivity

Connections to I/O cages, I/O drawers and Parallel Sysplex InfiniBand coupling (PSIFB) are driven from the host channel adapter fanouts that are located on the front of the book. Figure 2-18 shows the location of the fanouts and connectors for a two-book system. In the figure, ECF is the External Clock Facility card for the Pulse Per Second (PPS) connectivity, OSC is the oscillator card; FSP is flexible service processor; and LG is location code for logic card.

ECF	OSC	OSC	ECF
filler	D1 I/O D2 I/O D3 FSP D4 FSP D5 I/O D6 I/O D7 I/O D8 I/O D9 I/O DA I/O	filler	D1 I/O D2 I/O D3 FSP D4 FSP D5 I/O D6 I/O D7 I/O D8 I/O D9 I/O DA I/O
LG01	LG06	LG10	LG15
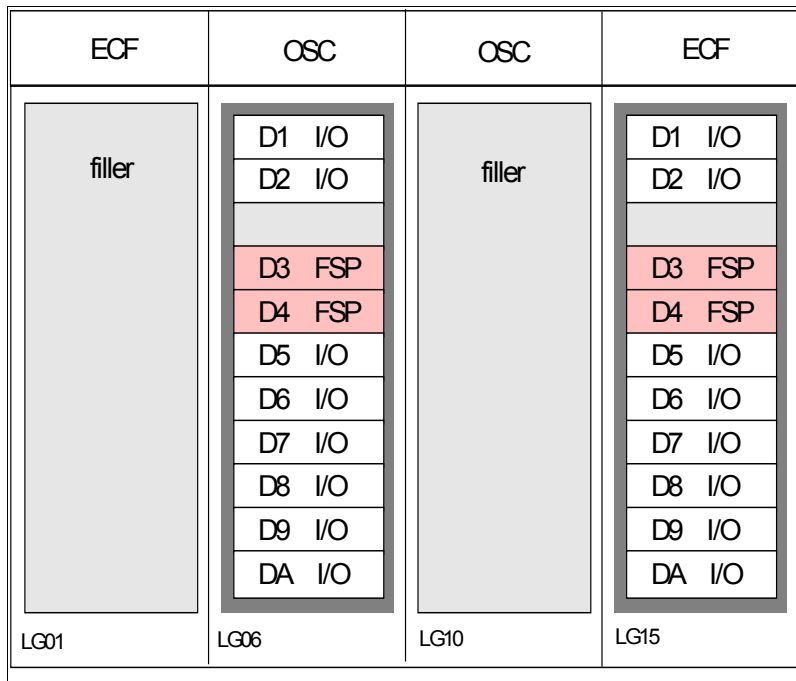
Figure 2-18   Location of the host channel adapter fanouts

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

A fanout can be repaired concurrently with the use of redundant I/O interconnect. See 2.7.1, "Redundant I/O interconnect" on page 51.

The three types of two-port fanouts are:

► Host Channel Adapter2-C (HCA2-C) provides copper connections for InfiniBand I/O interconnect to all I/O, ISC-3, and Crypto Express cards in I/O cages and I/O drawers.

► Host Channel Adapter2-O (HCA2-O) provides optical connections for 12x InfiniBand for coupling links (PSIFB). The HCA2-O provides a point-to-point connection over a distance of up to 150 m (492 feet), using four 12x MPO fiber connectors and OM3 fiber optic cables (50/125 µm).

z196 to z196 or System z10 connections use 12-lane InfiniBand Double Data Rate (12 x IB-DDR) link at 6 GBps. If the connection is from z196 to a System z9®, 12-lane InfiniBand Single Data Rate (12 x IB-SDR) at 3 GBps is used.

► The HCA2-O LR fanout provides optical connections for 1x InfiniBand and supports PSIFB Long Reach (PSIFB LR) coupling links for distances of up to 10 km and up to 100 km when repeated through a System z qualified DWDM. This fanout is supported on z196 and System z10 only.

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

> **Note:** The InfiniBand link data rates (6 GBps, 3 GBps, 5 Gbps, or 2.5 Gbps) do not represent the actual performance of the link. The actual performance depends on several factors, such as latency, cable lengths, and the type of workload.

Note the following information about models and fanout positions:

► On a model M15, all fanout positions can be populated, for up to 16 I/O connections of any type. On a model M32, all fanout positions can be populated, for up to 32 I/O connections.

► On a model M49, all fanout positions can be populated only on the first book. Positions D1 and D2 must remain free of fanouts on both the second and third books, for up to 40 I/O connections of any type.

► On models M66 and M80, all D1 and D2 positions must remain free of any fanout. This results in up to 48 I/O connections of any type.

Figure 2-19 shows the InfiniBand connectors used for each of the two optical HCA types.
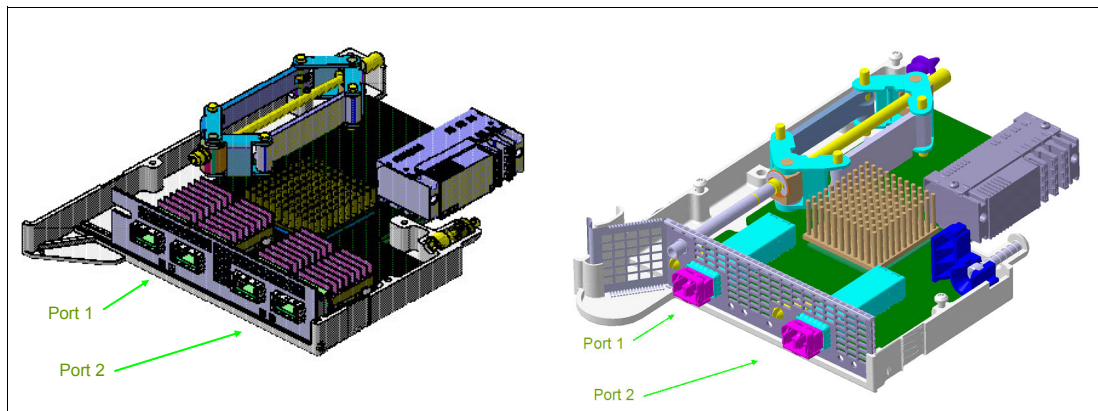
Figure 2-19   Infiniband HCA2 - 12x optical interface and 1x optical interface

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

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

In the configuration report, fanouts are identified by their locations in the CPC cage. Fanout locations are numbered from D3 through D8. The jacks are numbered J01 and J02 for each HCA2-C, or HCA2-O LR fanout port. Jack numbering for HCA2-O fanout ports for transmit and receive jacks is JT1 and JR1, and JT2 and JR2.

2.7.1 Redundant I/O interconnect

Redundant I/O interconnect is accomplished by the facilities of the InfiniBand I/O connections to the InfiniBand Multiplexer (IFB-MP) card. Each IFB-MP card is connected to a jack located in the InfiniBand fanout of a book. IFB-MP cards are half-high cards and are interconnected with cards called STI-A8 and STI-A4, allowing redundant I/O interconnect in case the connection coming from a book ceases to function, as happens when, for example, a book is removed. A conceptual view of how redundant I/O interconnect is accomplished is shown in Figure 2-20.
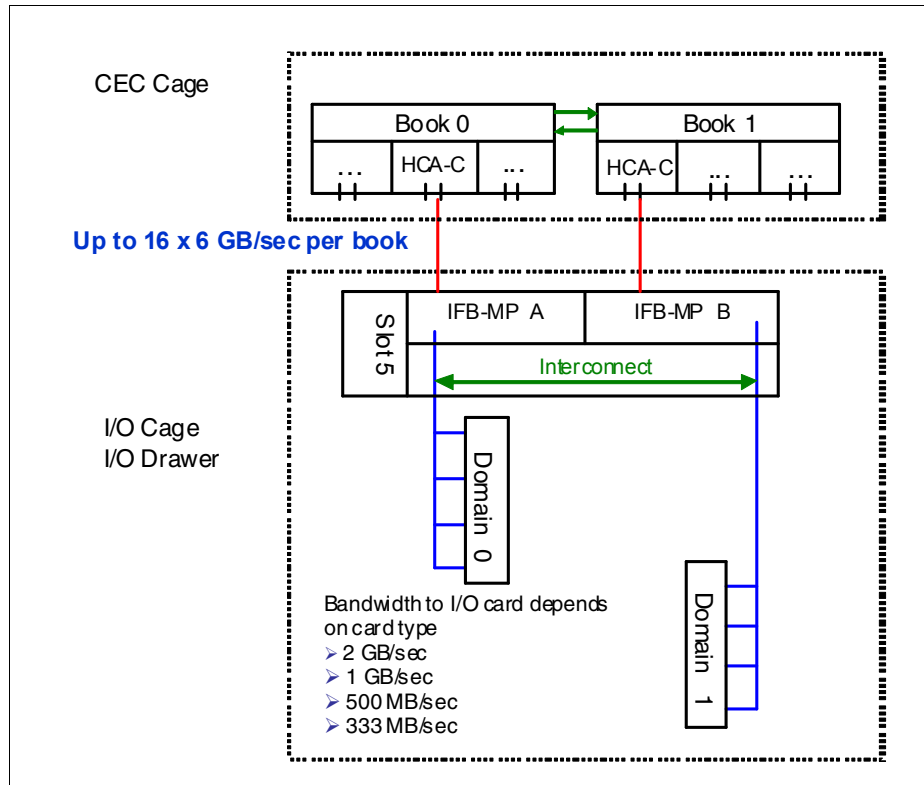
Figure 2-20   Redundant I/O interconnect

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

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

2.7.2 Enhanced book availability

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

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

Removal of a book also removes the book connectivity to the I/O cages. The impact of the removal of the book on the system is limited by the use of redundant I/O interconnect, which is described in 2.7.1, "Redundant I/O interconnect" on page 51. However, all PSIFBs on the removed book must be configured offline.

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

2.7.3 Book upgrade

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

2.8 Model configurations

When a z196 order is configured, PUs are characterized according to their intended usage. They can be ordered as any of the following items:

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

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

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

Unassigned IFL   A processor purchased for future use as an IFL. It is offline and cannot be used until an upgrade for the IFL is installed. It does not affect software licences or maintenance charges.

ICF   An internal coupling facility (ICF) processor purchased and activated for use by the Coupling Facility Control Code.

zAAP   A z196 Application Assist Processor (zAAP) purchased and activated to run eligible workloads such as Java code under control of z/OS JVM or z/OS XML System Services.

zIIP   A z196 Integrated Information Processor (zIIP) purchased and activated to run eligible workloads such as DB2 DRDA or z/OS[3] Communication Server IPSec.

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

---
[3] z/VM V5R4 and above support zAAP and zIIP processors for guest configurations.

A minimum of one PU characterized as a CP, IFL, or ICF is required per system. The maximum number of CPs is 80, the maximum number of IFLs is 80, and the maximum number of ICFs is 16. The maximum number of zAAPs is 40, but requires an equal or greater number of characterized CPs. The maximum number of zIIPs is also 40 and requires an equal or greater number of characterized CPs. The sum of all zAAPs and zIIPs cannot be larger than two times the number of characterized CPs.

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

The z196 model nomenclature is based on the number of PUs available for customer use in each configuration. The models are summarized in Table 2-10.

Table 2-10   z196 configurations

Model	Books	PUs per MCM	Active PUs			zAAPs	zIIPs	Add. SAPs	Std. SAPs	Spares
			CPs	IFLs/ uIFL	ICFs					
M15	1	20	0–15	0–15	0–12	0–7	0–7	0–3	3	2
M32	2	20	0–32	0–32	0–16	0–16	0–16	0–7	6	2
M49	3	20	0–49	0–49	0–16	0–24	0–24	0–11	9	2
M66	4	20	0–66	0–66	0–16	0–33	0–33	0–18	12	2
M80	4	24	0–80	0–80	0–16	0–40	0–40	0–18	14	2

A capacity marker identifies that a certain number of CPs have been purchased. This number of purchased CPs is higher than or equal to the number of CPs actively used. The capacity marker marks the availability of purchased but unused capacity intended to be used as CPs in the future. They usually have this status for software-charging reasons. Unused CPs are not a factor when establishing the MSU value that is used for charging MLC software, or when charged on a per-processor basis.

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

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

2.8.1 Upgrades

Concurrent CP, IFL, ICF, zAAP, zIIP, or SAP upgrades are done within a z196. Concurrent upgrades require available PUs. Concurrent processor upgrades require that additional PUs are installed (at a prior time) but not activated.

Spare PUs are used to replace defective PUs. There are always two spare PUs on a z196.

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

Although upgrades from one z196 model to another z196 model are concurrent, meaning that one or more books can be added, there is one exception. Upgrades from any z196 (model M15, M32, M49, M66) to a model M80 is disruptive because this upgrade requires the

replacement of four books. Table 2-11 shows the possible upgrades within the z196 configuration range.

Table 2-11   z196 to z196 upgrade paths

To 2817 From 2817	Model M15	Model M32	Model M49	Model M66	Model M80[a]
Model M15	-	Yes	Yes	Yes	Yes
Model M32	-	-	Yes	Yes	Yes
Model M49	-	-	-	Yes	Yes
Model M66	-	-	-	-	Yes

a. Disruptive upgrade

You may also upgrade a System z10 EC or a System z9 EC to a z196, preserving the server serial number (S/N). The I/O cards are also moved up (with certain restrictions).

> **Note:** Upgrades from System z10 and System z9 are disruptive.

Upgrade paths from any z9 EC to any z196 are supported as listed in Table 2-12.

Table 2-12   z9 EC to z196 upgrade paths

To 2817 From 2094	Model M15	Model M32	Model M49	Model M66	Model M80
Model S08	Yes	Yes	Yes	Yes	Yes
Model S18	Yes	Yes	Yes	Yes	Yes
Model S28	Yes	Yes	Yes	Yes	Yes
Model S38	Yes	Yes	Yes	Yes	Yes
Model S54	Yes	Yes	Yes	Yes	Yes

Upgrades from any z10 EC to any z196 are supported as listed in Table 2-13 on page 54.

Table 2-13   z10 EC to z196 upgrade paths

To 2817 From 2097	Model M15	Model M32	Model M49	Model M66	Model M80
Model E12	Yes	Yes	Yes	Yes	Yes
Model E26	Yes	Yes	Yes	Yes	Yes
Model E40	Yes	Yes	Yes	Yes	Yes
Model E56	Yes	Yes	Yes	Yes	Yes
Model E64	Yes	Yes	Yes	Yes	Yes

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

2.8.2 Concurrent PU conversions

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

Most conversions are not disruptive. In exceptional cases, the conversion can be disruptive, for example, when a model M15 with 15 CPs is converted to an all IFL system. In addition, a logical partition might be disrupted when PUs must be freed before they can be converted. Conversion information is summarized in Table 2-14.

Table 2-14   Concurrent PU conversions

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

2.8.3 Model capacity identifier

To recognize how many PUs are characterized as CPs, the store system information (STSI) instruction returns a value that can be seen as a model capacity identifier (MCI), which determines the number and speed of characterized CPs. Characterization of a PU as an IFL, an ICF, a zAAP, or a zIIP is not reflected in the output of the STSI instruction, because these have no effect on software charging. More information about the STSI output is in "Processor identification" on page 305.

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

► For full-capacity engines, model capacity identifiers 701 to 780 are used. They express the 80 possible capacity settings from one to 80 characterized CPs.

► Three model capacity identifier ranges offer a unique level of granular capacity at the low end. They are available when no more than fifteen CPs are characterized. These three subcapacity settings applied to up to fifteen CPs offer 45 additional capacity settings. See "Granular capacity" on page 55.

Granular capacity

The z196 offers 45 capacity settings at the low end of the processor. Only 15 CPs can have granular capacity. When subcapacity settings are used, other PUs, beyond 15, can only be characterized as specialty engines.

The three defined ranges of subcapacity settings have model capacity identifiers numbered from 401 to 415, 501 to 515, and 601 to 615.

Note: Within a z196, all CPs have the same capacity identifier. Specialty engines (IFLs, zAAPs, zIIPs, ICFs) operate at full speed.

List of model capacity identifiers

Table 2-15 shows that regardless of the number of books, a configuration with one characterized CP is possible. For example, model M80 may have only one PU characterized as a CP.

Table 2-15   Model capacity identifiers

z196	Model capacity identifier
Model M15	701–715, 601–615, 501–515, 401–415
Model M32	701–732, 601–615, 501–515, 401–415
Model M49	701–749, 601–615, 501–515, 401–415
Model M66	701–766, 601–615, 501–515, 401–415
Model M80	701–780, 601–615, 501–515, 401–415

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

2.8.4 Model capacity identifier and MSU values

All model capacity identifiers have a related MSU value (millions of service units) that is used to determine the software license charge for MLC software. Table 2-16 and Table 2-17 on page 57 show MSU values for each model capacity identifier.

Table 2-16   Model capacity identifier and MSU values

Model capacity identifier	MSU	Model capacity identifier	MSU	Model capacity identifier	MSU
701	150	728	2704	755	4656
702	281	729	2780	756	4726
703	408	730	2855	757	4795
704	531	731	2927	758	4864
705	650	732	2998	759	4933
706	766	733	3068	760	5001
707	879	734	3134	761	5069
708	988	735	3207	762	5136
709	1091	736	3279	763	5203
710	1191	737	3351	764	5270
711	1286	738	3418	765	5336
712	1381	739	3489	766	5402
713	1473	740	3564	767	5466
714	1562	741	3639	768	5528
715	1648	742	3713	769	5588
716	1731	743	3788	770	5646

Model capacity identifier	MSU	Model capacity identifier	MSU	Model capacity identifier	MSU
717	1816	744	3861	771	5702
718	1899	745	3935	772	5757
719	1983	746	4009	773	5810
720	2064	747	4082	774	5862
721	2144	748	4155	775	5912
722	2224	749	4228	776	5960
723	2306	750	4300	777	6007
724	2388	751	4372	778	6053
725	2469	752	4444	779	6097
726	2550	753	4515	780	6140
727	2627	754	4586	-	-

Table 2-17   Model capacity identifier and MSU values for subcapacity models

Model capacity identifier	MSU	Model capacity identifier	MSU	Model capacity identifier	MSU
401	30	501	74	601	97
402	58	502	140	602	182
403	85	503	204	603	263
404	110	504	265	604	344
405	135	505	326	605	421
406	160	506	385	606	497
407	183	507	442	607	569
408	207	508	498	608	640
409	229	509	552	609	709
410	252	510	604	610	777
411	274	511	655	611	843
412	296	512	705	612	907
413	318	513	754	613	969
414	339	514	806	614	1028
415	359	515	849	615	1084

2.8.5 Capacity Backup

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

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

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

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

CBU and granular capacity

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

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

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

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

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

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

Example 2-1   Simple CBU record and related configuration features

```
On Demand Capacity Selecions:
NEW00001 - CBU - CP(1) - Years(5) - Tests(0)
          Expiration(09/10/2012)


Resulting feature numbers in configuration:

6817  Total CBU Years Ordered                5
```

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

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

Example 2-2   Second CBU record and resulting configuration features

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

Resulting cumulative feature numbers in configuration:

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

CBU for CP rules

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

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

(503 + 2 = 505)

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

(3 + 3 = 6)

► The CBU cannot decrease the number of CPs.

► The CBU cannot lower the capacity setting.

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

CBU for specialty engines

Specialty engines (ICFs, IFLs, zAAPs, and zIIPs) run at full capacity for all capacity settings. This also applies to CBU for specialty engines. Table 2-18 shows the minimum and maximum (min-max) numbers of all types of CBUs that might be activated on each of the models. Note that the CBU record can contain larger numbers of CBUs than can fit in the current model.

Table 2-18   Capacity BackUp matrix

Model	Total PUs available	CBU CPs min-max	CBU IFLs min-max	CBU ICFs min-max	CBU zAAPs min-max	CBU zIIPs min-max	CBU SAPs min-max
Model M15	15	0–15	0–15	0–12	0–7	0–7	0-3
Model M32	32	0–32	0–32	0–16	0–16	0–16	0-7
Model M49	49	0–49	0–49	0–16	0–24	0–24	0-11
Model M66	66	0–66	0–66	0 –16	0–33	0–33	0-20
Model M80	80	0–80	0–80	0–16	0–40	0–40	0-18

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

2.8.6 On/Off Capacity on Demand and CPs

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

On/Off CoD and granular capacity

When temporary capacity requested by On/Off CoD for CPs matches the model capacity identifier range of the permanent CP feature, the total number of active CP equals the sum of the number of permanent CPs plus the number of temporary CPs ordered. For example, when a model capacity identifier 504 has two CP5s added temporarily, it becomes a model capacity identifier 506.

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

On/Off CoD guidelines

When you request temporary capacity, consider the following guidelines

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

Table 9-3 on page 289 shows possible On/Off CoD CP upgrades for granular capacity models. For more information about temporary capacity increases, see Chapter 9, "System upgrades" on page 263.

2.9 Cooling

zEnterprise 196 is an air-cooled system assisted by refrigeration. A chilled water cooling option is also available. Selection of the cooling method is done at ordering and the appropriate equipment is factory installed.

2.9.1 Air cooled models

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

Subsystems

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

► The modular refrigeration units (MRU)

   – One (or two) MRUs (MRU0 and MRU1), located in the front of the frame A below the books, provide refrigeration to the content of the books together with two large blower assemblies are at the rear of the CPC cage, one for each MRU. The assemblies, which are the motor scroll assembly (MSA) and the motor drive assembly (MDA), are connected to the bulk power regulator (BPR) that regulates cooling by increasing the blower speed in combination with an air-moving assembly located in the top part of the CPC cage.

   – A one-book system has MRU0 installed. MRU1 is installed when you upgrade to a two-book system, providing all refrigeration requirements for a four-book system. Concurrent repair of an MRU is possible by taking advantage of the hybrid cooling implementation described in the next section.

► The motor drive assembly (MDA)

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

Hybrid cooling system

The z196 air cooled models have a hybrid cooling system that is designed to lower power consumption. Normal cooling is provided by one or two MRUs connected to the evaporator heat sinks mounted on all MCMs in all books.

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

Figure 2-21 on page 62 shows the refrigeration scope of MRU0 and MRU1.

Fourth Book	**First Book**	**Third Book**	**Second Book**
MRU 0		**MRU 1**	
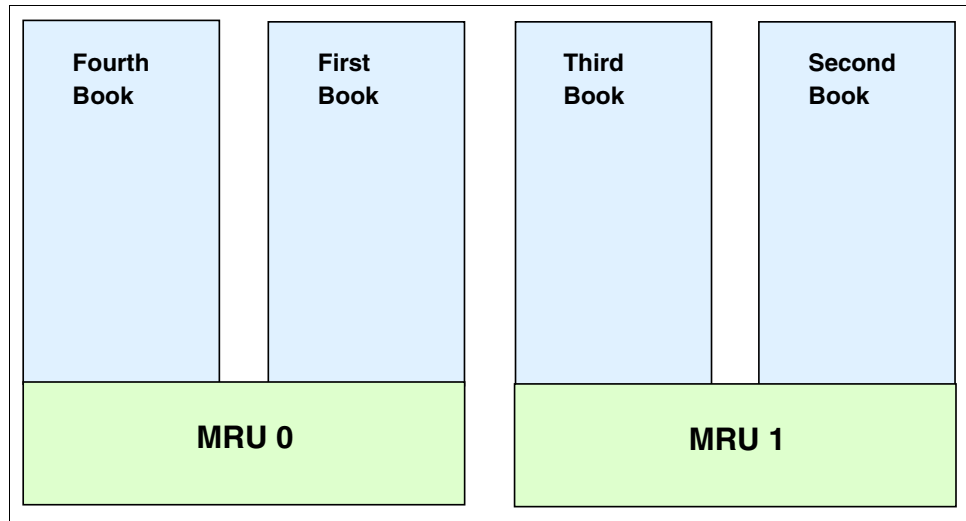
Figure 2-21   MRU scope

2.9.2 Water cooled models

The z196 introduces the ability to cool systems with customer chilled water by introducing the water cooling unit (WCU) technology.

Water supply

Following are some general conditions that your facility must meet prior to installation of the water-cooled models of the server:

► Allowable system inlet water temperature range is 6-16 $_0$C (43-61 $_0$F), using standard building chilled water (BCW). A special water system for the server is typically not required.
► Required flow rate to the frame is 3.7 – 79.4 lpm (1 -21 gpm), depending on inlet water temperature and the number of nodes populated in the server. Colder inlet water temperatures require less flow then warmer water temperatures. Fewer nodes require less flow then maximum populated processors.
► Minimum water pressure required across the IBM hose ends is 0.34 – 2.32BAR (5 – 33.7 psi), depending on the minimum flow required.

For a more detailed description of the water-cooling option see *Installation Manual for Physical Planning, 2817 All Models* GC28-6897

If a water-cooled model of the server is ordered, the server will come with two rear-door heat exchangers - one for each of the two frames, as shown in Figure 2-22 on page 63. For availability reasons each water cooled server will have two WCUs installed, although the system will operate with a single WCU.

Figure 2-22   Water-cooled model - rear view

The water cooling option is recommended as it can substantially lower the total power consumption of the z196 and thus positively influence the total cost of ownership for the server. This is particularly true for the bigger models of the server (see Table 2-19). The water cooling option cannot be installed in the field, so careful consideration of present and future computer room and server configuration options should be exercised before deciding what cooling option to order.

Table 2-19   Power consumption based on temperature

Temperature	3 book typical configuration	4 book typical configuration	4 book maximum power configuration
Water cooled system power in normal room/hot room -est.			
	12.9 kW / 14.1 kW	17.4 kW / 19.0 kW	24.7 kW / 26.3 kW
Inlet air temperature Heat to water and as % of total system heat load			
18 C	7.3 kW (57%)	9.8 kW (56%)	12.6 kW (51%)
23 C	9.5 kW (74%)	12.6 kW (72%)	15.6 kW (63%)
27 C	11.5 kW (89%)	14.8 kW (85%)	18.0 kW (73%)
32 C (hot room)	14.8 kW (105%)	18.2 kW (96%)	21.6 kW (82%)

Chapter 2. CPC hardware components   **63**

2.10 Summary of z196 structure

Table 2-20 summarizes all aspects of the z196 structure.

Table 2-20   System structure summary

Description	Model M15	Model M32	Model M49	Model M66	Model M80
Number of MCMs	1	2	3	4	4
Total number of PUs	20	40	60	80	96
Maximum number of characterized PUs	15	32	49	66	80
Number of CPs	0–15	0–32	0–49	0–66	0–80
Number of IFLs	0–15	0–32	0–49	0–66	0–80
Number of ICFs	0–15	0–16	0–16	0–16	0–16
Number of zAAPs	0–7	0–16	0–24	0–33	0–40
Number of zIIPs	0–7	0–16	0–24	0–33	0–40
Standard SAPs	3	6	9	12	14
Additional SAPs	0–3	0–7	0–11	0–18	0–18
Standard spare PUs	2	2	2	2	2
Enabled memory sizes	32–752 GB	32–1520 GB	32–2280 GB	32–3056 GB	32–3056 GB
L1 cache per PU	64-I/128-D KB	64-I/128-D KB	64-I/128-D KB	64-I/128-D KB	64-I/128-D KB
L2 cache per PU	1.5 MB	1.5 MB	1.5 MB	1.5 MB	1.5 MB
L3 shared cache per PU chip	24 MB	24 MB	24 MB	24 MB	24 MB
L4 shared cache	192 MB	384 MB	576 MB	768 MB	768 MB
Cycle time (ns)	0.19	0.19	0.19	0.19	0.19
Clock frequency	5.2 GHz	5.2 GHz	5.2 GHz	5.2 GHz	5.2 GHz
Maximum number of fanout ports	16	32	40	48	48
I/O interface per IFB cable	6 GBps	6 GBps	6 GBps	6 GBps	6 GBps
Maximum I/O cages	2	3[a]	3[a]	3[a]	3[a]
Maximum I/O drawers	4	4	4	4	4
Number of support elements	2	2	2	2	2
External AC power	3 phase	3 phase	3 phase	3 phase	3 phase
Optional external DC	570V/380V	570V/380V	570V/380V	570V/380V	570V/380V
Internal Battery Feature	Optional	Optional	Optional	Optional	Optional

a. Requires RPQ 8P2506

3

CPC system design

The objective of this chapter is to explain how the z196 is designed. This information can be used to understand the functions that make the z196 a server that suits a broad mix of workloads for large enterprises.

This chapter discusses the following topics:

The design of the z196 symmetric multiprocessor (SMP) 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 new types of applications that customers are becoming more and more dependent on.

The z196 offers very high levels of serviceability, availability, reliability, resilience, and security, and fits in the IBM strategy in which mainframes play a central role in creating an intelligent, energy efficient, integrated infrastructure. The z196 is designed in such a way that not only is the server considered important for the infrastructure, but also everything around it (operating systems, middleware, storage, security, and network technologies supporting open standards) helps customers achieve their business goals.

The modular book design aims to reduce planned and unplanned outages by offering concurrent repair, replace, and upgrade functions for processors, memory, and I/O. The z196 with its ultra-high frequency, very large high speed buffers (caches) and memory, superscalar processor design, out-of-order core execution, and flexible configuration options is the next implementation to address the ever-changing IT environment.

3.1 Design highlights

The physical packaging of the z196 is comparable to the packaging used for z10 EC and z9 EC systems. Its modular book design creates the opportunity to address the ever-increasing costs related to building systems with ever-increasing capacities. The modular book design is flexible and expandable, offering unprecedented capacity to meet consolidation needs, and might contain even larger capacities in the future.

z196 continues the line of upward-compatible mainframe processors, having 246 complex instructions executed by millicode and another 211 complex instructions cracked into multiple RISC like operations. It uses 24, 31, and 64-bit addressing modes, multiple arithmetic formats, and multiple address spaces robust inter-process security.

The main objectives of the z196 system design, which are discussed in this and subsequent chapters, are as follows:

► Offer a *flexible infrastructure* to concurrently accommodate a wide range of operating systems and applications, from the traditional systems (for example z/OS and z/VM) to the world of Linux and e-business.

► Offer state-of-the-art *integration* capability for server consolidation, offering virtualization techniques, such as:

  – Logical partitioning, which allows 60 independent logical servers

  – z/VM, which can virtualize hundreds to thousands of servers as independently running virtual machines

  – HiperSockets, which implement virtual LANs between logical partitions within a server

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

► Offer *high performance* to achieve the outstanding response times required by new workload-type applications, based on high frequency, superscalar processor technology, out-of-order core execution, large high speed buffers (cache) and memory, architecture, and high bandwidth channels, which offer second-to-none data rate connectivity.

► Offer the *high capacity* and *scalability* required by the most demanding applications, both from single-system and clustered-systems points of view.

► Offer the capability of *concurrent upgrades* for processors, memory, and I/O connectivity, avoiding server outages in planned situations.

► Implement a system with *high availability* and *reliability,* from the redundancy of critical elements and sparing components of a single system, to the clustering technology of the Parallel Sysplex environment.

► Have broad internal and external *connectivity* offerings, supporting open standards such as Gigabit Ethernet (GbE), and Fibre Channel Protocol (FCP) for Small Computer System Interface (SCSI).

► Provide the highest level of *security* in which every two PUs share a CP Assist for Cryptographic Function (CPACF). Optional Crypto Express features with Cryptographic Coprocessors and Cryptographic Accelerators for Secure Sockets Layer (SSL) transactions of applications can be added.

► Be *self-managing* and *self-optimizing*, adjusting itself on workload changes to achieve the best system throughput, through the Intelligent Resource Director or the Workload Manager functions, assisted by HiperDispatch.

► Have a *balanced system* design, providing large data rate bandwidths for high performance connectivity along with processor and system capacity.

The following sections describe the z196 system structure, showing a logical representation of the data flow from PUs, caches, memory cards, and a variety of interconnect capabilities.

3.2 Book design

A z196 system has up to four books on a fully connected topology, up to 80 processor units can be characterized, and up to 3 TB of memory capacity. Memory has up to 12 memory controllers, using 5-channel redundant array of independent memory (RAIM) protection, with DIMM bus cyclic redundancy check (CRC) error retry. The 4-level cache hierarchy is implemented with eDRAM (embedded) caches. Up until recently eDRAM was considered to be too slow for this kind of use, however, a break-through in IBM technology has negated that. In addition eDRAM offers higher density, less power utilization, fewer soft-errors, and better performance. Concurrent maintenance allows dynamic book add and repair.

The z196 server uses 45nm chip technology, with advancing low latency pipeline design, leveraging high speed yet power efficient circuit designs. The multichip module (MCM) has a dense packaging, allowing modular refrigeration units (MRUs) cooling, or as an option, water cooling. The water cooling option is recommended as it can lower the total power consumption of the server. This is particularly true for the larger configurations as discussed in "Cooling" on page 61.

3.2.1 Cache levels and memory structure

The z196 memory subsystem focuses on keeping data "closer" to the processor unit, implementing new chip-level shared cache (L3) and much larger book-level shared cache (L4).

Figure 3-1 shows the z196 cache levels and memory hierarchy.
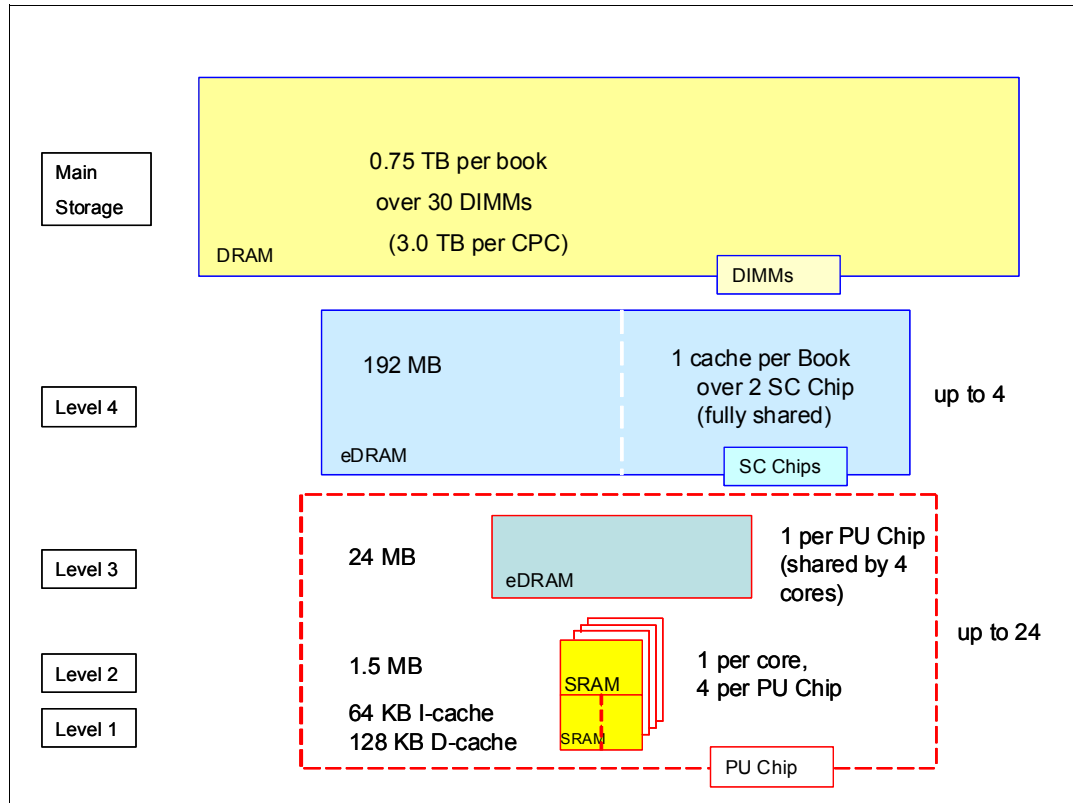
Figure 3-1   z196 cache levels and memory hierarchy

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

L1 and L2 caches use static random access memory (SRAM) and are private for each core. L3 cache uses embedded dynamic static random access memory (eDRAM) and is shared by all four cores within the PU chip. Each book has six L3 caches and a four-book system has 24 of them, resulting on 576 MB (24 x 24 MB) of this shared PU chip level cache. L4 cache also uses eDRAM and is shared by all PU chips on the MCM. A four-book server has 768 MB (4 x 192 MB) of shared L4 cache. Main storage has up to 0.75 TB per book, using up to 30 DIMMs. A four-book server can have up to 3 TB of main storage.

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

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

In order to overcome this and avoid potential latency, z196 introduces two additional cache levels (L2 and L3) within the PU chip, with denser packaging. This design reduces traffic to and from the shared L4 cache, which is located on another chip (SC chip). Only when there is a cache miss in L1, L2, and L3, the request is sent to L4. L4 is the coherence manager, meaning that all memory fetches must be in the L4 cache before that data can be used by the processor.

Another approach is available for avoiding L4 cache access delays (latency) as much as possible. The L4 cache straddles up to four books. This means relatively large distances exist between the higher-level caches in the processors and the L4 cache content. To overcome the delays that are inherent to the book design and to save cycles to access the *remote* L4 content, it is beneficial to keep instructions and data as close to the processors as possible by directing as much work of a given logical partition workload on the processors located in the same book as the L4 cache. This is achieved by having the PR/SM scheduler and the z/OS dispatcher work together to keep as much work as possible within the boundaries of as few processors and L4 cache space (which is best within a book boundary) as can be achieved without affecting throughput and response times. Preventing PR/SM and the dispatcher from scheduling and dispatching a workload on any processor available, and keeping the workload in as small a portion of the server as possible, contributes to overcoming latency in a high-frequency processor design such as the z196. The cooperation between z/OS and PR/SM has been bundled in a function called HiperDispatch. HiperDispatch exploits the new z196 cache topology, with reduced cross-book "help", and better locality for multi-task address spaces. More information about HiperDispatch is in 3.6, "Logical partitioning" on page 93.

Figure 3-2 compares the cache structures of the z196 with the System z previous generation server, the z10 EC.
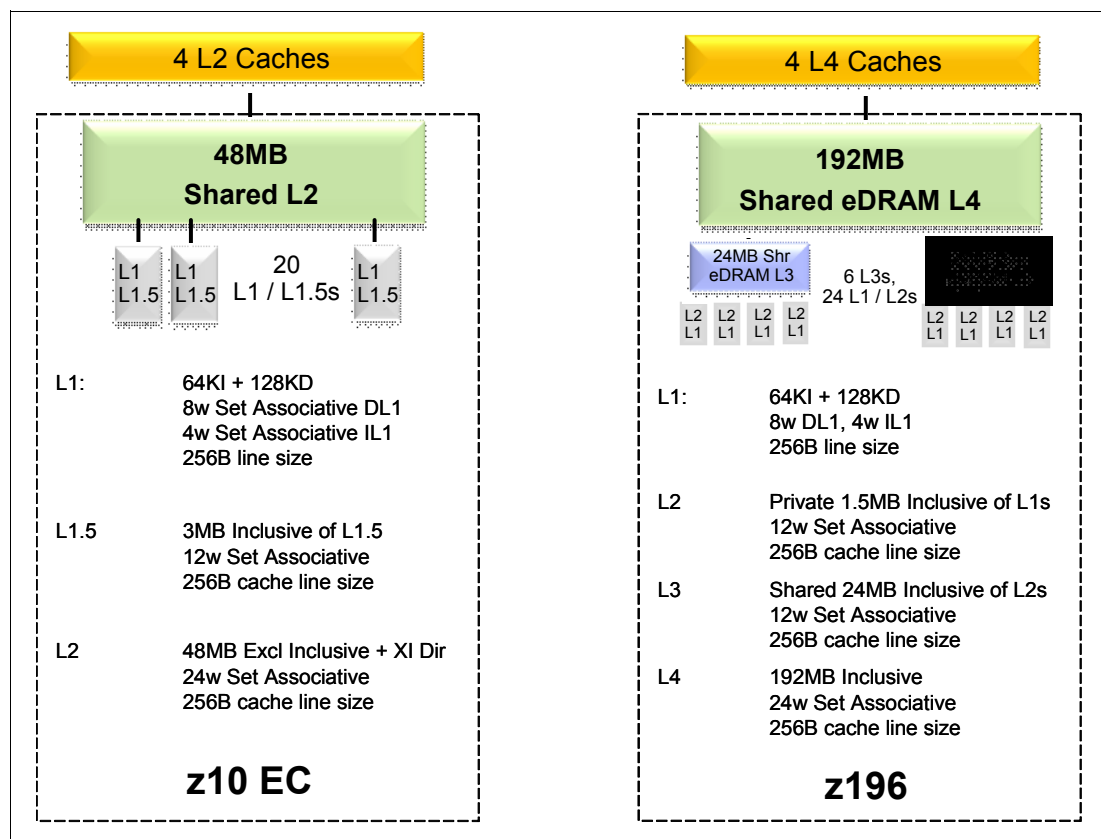
Figure 3-2   z196 and z10 EC cache levels comparison

Compared to z10 EC, the z196 cache design has one more shared level and larger cache level sizes, except for the L1 private cache on each core (as its access time should occur in one cycle).

The z196 cache level structure is focused on keeping more data closer to the processor unit. This design can improve system performance on many production workloads.

3.2.2 Book interconnect topology

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

Figure 3-3 shows a simplified topology for a four-book system.
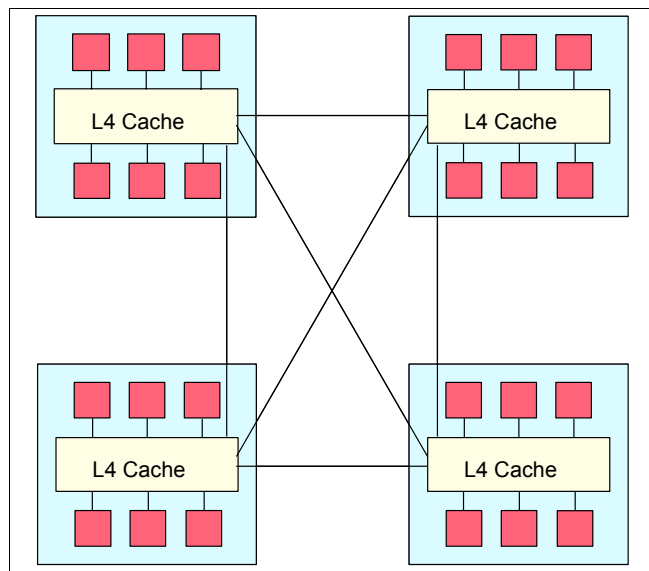
Figure 3-3　Point-to-point topology for book-to-book communication

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

3.3 Processor unit design

Today's systems design is driven by processor cycle time, though this does not automatically mean that the performance characteristics of the system improve. Processor cycle time is especially important for CPU-intensive applications. The System z previous generation server (z10 EC) introduced a dramatic PU cycle time improvement, and z196 is going even further, reaching 200 picoseconds (5.2 GHz).

Besides the cycle time, other processor design aspects, like pipeline, execution order, branch prediction, and high speed buffers (caches) are also very important for the performance of the system. Each z196 processor unit core is a superscalar, out of order processor, having six RISC-like execution units. There are 246 complex instructions executed by millicode and another 211 complex instructions cracked into multiple RISC like operations.

z196 introduces architectural extensions, with new instructions to allow reduced processor quiesce effects, reduced cache misses, and reduced pipeline disruption. The z196 new PU architecture includes the following:

► PER4 Zero Address Detect with improved debug capability to detect uninitialized pointers
► New trunc and OR inexactness Binary Floating Point rounding mode
► New Decimal Floating Point quantum exception, which eliminates need for test data group for every operation

- ► Virtual Architecture Level, that allows the z/VM Live Guest Relocation Facility to make a z196 behave architecturally like a z10 system, and facilitates moving work transparently between z196 and z10 systems for backup and capacity reasons.
- ► On Non-quiescing set storage key extended (SSKE) instruction, with significant performance improvement for systems with large number of PUs. It also improves multi-processing (MP) ratio for larger systems, and performance increases when exploited by the operating system (exploited by z/OS 1.10 with PTFs and above, and planned to be exploited by Linux and z/VM in the future).
- ► Other minor architecture: RRBM, Fast-BCR-Serialization Facility, Fetch-Store-Access Exception Indicator, CMPSC Enhancement Facility.

The z196 new instruction set architecture (ISA) includes 110 new instructions added to improve compiled code efficiency:

- ► High-Word Facility (30 new instructions), with independent addressing to high word of 64 bit GPRs, and effectively provides compiler/ software with 16 additional 32 bit registers.
- ► Interlocked-Access Facility (12 new instructions), including interlocked (atomic) load, value update and store operation in a single instruction, with immediate exploitation by Java.
- ► Load/Store-on-Condition Facility (6 new instructions), with load or store conditionally executed based on condition code, achieving dramatic improvement in certain codes with highly unpredictable branches.
- ► Distinct-Operands Facility (22 new instructions), with independent specification of result register (different than either source register), reducing register value copying.
- ► Population-Count Facility (1 new instruction), which is a hardware implementation of bit counting, achieving up to five times faster than prior software implementations.
- ► Integer to/from Floating point converts (39 new instructions).

This results on optimized processor units to meet the demands of a wide variety of business workload types without compromising the performance characteristics of traditional workloads.

3.3.1 Out-of-order execution

The z196 is the first System z CMOS to implement an out-of-order (OOO) core. OOO yields significant performance benefit for compute intensive applications through re-ordering instruction execution, allowing later (younger) instructions to be executed ahead of a stalled instruction, and re-ordering storage accesses and parallel storage accesses. OOO maintains good performance growth for traditional applications. Out-of-order (OOO) execution can improve performance by:

- ► Re-ordering instruction execution

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

- ► Re-ordering storage accesses

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

- ► Hiding storage access latency

  Many instructions access data from storage. Storage accesses can miss the L1 and require 10 to 500 additional cycles to retrieve the storage data. In an in-order core, later

instructions in the code stream are stalled. In an out-of-order core, later instructions that are not dependent on this storage data are allowed to execute.

OOO execution does not change any program results. Execution can occur out of (program) order, but all program dependencies are honored, ending up with same results of the in order (program) execution. This implementation requires special circuitry to make execution and memory accesses appear in order to software. The logical diagram of a z196 PU core is shown on Figure 3-4.
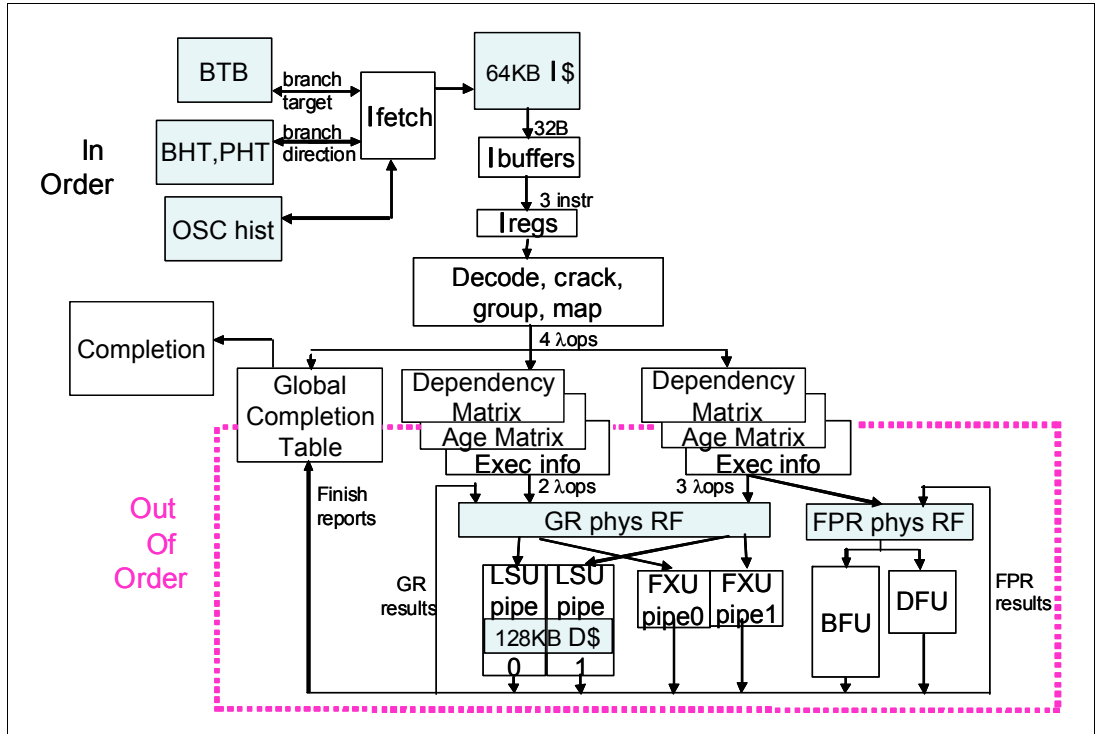

Figure 3-4　z196 PU core logical diagram

Memory address generation and memory accesses can occur out of (program) order. This capability can provide a greater exploitation of the z196 superscalar core, and can improve system performance. Figure 3-5 shows how OOO core execution can reduce the execution time of a program.
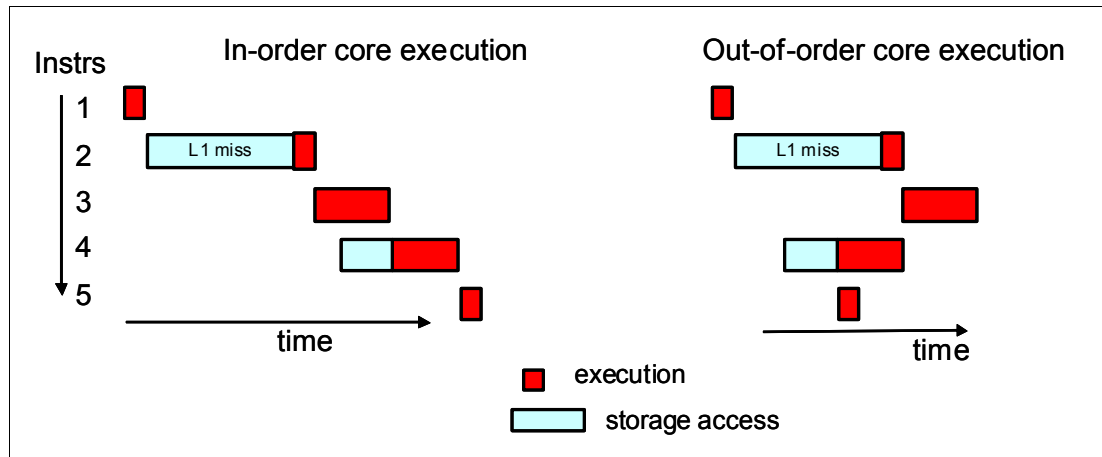

Figure 3-5　In-order and out-of-order core execution

On the example, the left side is showing an in-order core execution. Instruction 2 has a big delay due to an L1 miss, and next instructions wait until instruction 2 finishes. On usual in-order execution, next instruction waits until the previous one finishes. Using OOO core execution, shown on the right side of the example, instruction 4 could start its storage access and execution while the instruction 2 is waiting for data, if no dependencies exist between both instructions. So, when L1 miss is solved, instruction 2 could also start its execution while instruction 4 is executing, Instruction 5 could need the same storage data required by instruction 4, and as soon as this data is on L1, instruction 5 starts execution at the same time. The z196 superscalar PU core can execute up to five instructions/operations per cycle.

3.3.2 Superscalar processor

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

A superscalar processor is based on a multi-issue architecture. In such a processor, where multiple instructions can be executed at each cycle, a higher level of complexity is reached, because an operation in one pipeline stage might depend on data in another pipeline stage. Therefore, a superscalar design demands careful consideration of which instruction sequences can successfully operate in a long pipeline environment.

On the z196, up to three instructions can be decoded per cycle and up to five instructions/operations can be executed per cycle. Execution can occur out of (program) order.

Example of branch prediction

If the branch prediction logic of the microprocessor makes the wrong prediction, removing all instructions in the parallel pipelines also might be necessary. Obviously, the cost of the wrong branch prediction is more costly in a high-frequency processor design, as we discussed previously. Therefore, the branch prediction techniques used are very important to prevent as many wrong branches as possible. For this reason, a variety of history-based branch prediction mechanisms are used, as shown on the in-order part of the z196 PU core logical diagram on Figure 3-4. The branch target buffer (BTB) runs ahead of instruction cache pre-fetches to prevent branch misses in an early stage. Furthermore, a branch history table (BHT) in combination with a pattern history table (PHT) and the use of tagged multi-target prediction technology branch prediction offer an extremely high branch prediction success rate.

Challenges of creating a superscalar processor

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

3.3.3 Compression and cryptography accelerators on a chip

There are two coprocessor units for compression and cryptography on each PU chip. Each coprocessor accelerator is shared by two cores of the PU chip. The compression engines are independent and the cryptography engines are shared. The compression engine uses static dictionary compression and expansion. The dictionary size is up to 64KB, with 8K entries,

and has a local 16 KB cache per core for dictionary data. The cryptography engine is used for CP assist for cryptographic function (CPACF), which implements enhancements for the new NIST standard.

CP assist for cryptographic function

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

http://www.ibm.com/servers/resourcelink/

For more information about cryptographic on z196, see Chapter 6., "Cryptography" on page 163.

3.3.4 Decimal floating point accelerator

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

Base 10 arithmetic is used for most business and financial computation. Floating point computation that is used for work typically done in decimal arithmetic has involved frequent necessary data conversions and approximation to represent decimal numbers. This has made floating point arithmetic complex and error-prone for programmers using it for applications in which the data is typically decimal data.

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

Benefits of DFP accelerator

The DFP accelerator offers the following benefits:

► Avoids rounding issues such as those happening with binary-to-decimal conversions.

► Has better functionality over existing binary coded decimal (BCD) operations.

► Follows the standardization of the dominant decimal data and decimal operations in commercial computing supporting industry standardization (IEEE 745R) of decimal floating point operations. Instructions are added in support of the Draft Standard for Floating-Point Arithmetic, which is intended to supersede the ANSI/IEEE Std 754-1985.

Software support

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

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

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

3.3.5 Processor error detection and recovery

The PU uses something called transient recovery as an error recovery mechanism. When an error is detected, the instruction unit retries the instruction and attempts to recover the error. If the retry is not successful (that is, a permanent fault exists), a relocation process is started that restores the full capacity by moving work to another PU. Relocation under hardware control is possible because the R-unit has the full architected state in its buffer. The principle is shown in Figure 3-6.
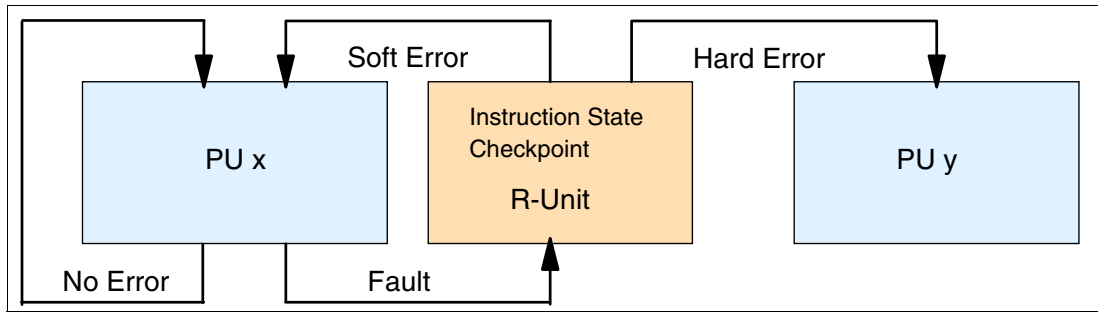
Figure 3-6   PU error detection and recovery

3.3.6 Branch prediction

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

The branch history table (BHT) implementation on processors has a large performance improvement effect. Originally introduced on the IBM ES/9000 9021 in 1990, the BHT has been continuously improved.

The BHT offers significant branch performance benefits. The BHT allows each PU to take instruction branches based on a stored BHT, which improves processing times for calculation routines. Besides the BHT, the z196 uses a variety of techniques to improve the prediction of the correct branch to be executed. The techniques include:

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

The success rate of branch prediction contributes significantly to the superscalar aspects of the z196. This is because the architecture rules prescribe that, for successful parallel execution of an instruction stream, the correctly predicted result of the branch is essential.

3.3.7 Wild branch

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

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

collect data from that storage location. This approach decreases the many debugging steps necessary when looking for where the branch came from.

3.3.8 IEEE floating point

Over 130 binary and hexadecimal floating-point instructions are present in z196. They incorporate IEEE Standards into the platform.

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

3.3.9 Translation look-aside buffer

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

The size of the TLB is kept as small as possible because of its low access time requirements and hardware space limitations. Because memory sizes have recently increased significantly, as a result of the introduction of 64-bit addressing, a smaller working set is represented by the TLB. To increase the working set representation in the TLB without enlarging the TLB, large page support is introduced and can be used when appropriate. See "Large page support" on page 91.

3.3.10 Instruction fetching, decode, and grouping

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

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

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

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

The compilers and JVMs are responsible for selecting instructions that best fit with the superscalar microprocessor and abide by the rules to create code that best exploits the superscalar implementation. All the System z compilers and the JVMs are under constant change to benefit from new instructions as well as advances in microprocessor designs.

3.3.11 Extended translation facility

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

3.3.12 Instruction set extensions

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

► Hexadecimal floating point instructions for various unnormalized multiply and multiply-add instructions.

► Immediate instructions, including various add, compare, OR, exclusive OR, subtract, load, and insert formats. Use of these instructions improves performance.

► Load instructions for handling unsigned half words (such as those used for Unicode).

► Cryptographic instructions, extended with AES, SHA-256, and functions for random number generation

► Extended Translate Facility-3 instructions, enhanced to conform with the current Unicode 4.0 standard

► Assist instructions, help eliminate hypervisor overhead

3.4 Processor unit functions

All PUs on a z196 server are physically identical. When the system is initialized, PUs can be characterized to specific functions: CP, IFL, ICF, zAAP, zIIP, or SAP. The function assigned to a PU is set by the licensed internal code (LIC), which is loaded when the system is initialized (at power-on reset) and the PU is *characterized*. Only characterized PUs have a designated function. Non-characterized PUs are considered spares. At least one CP, IFL, or ICF must be ordered on a z196.

This design brings outstanding flexibility to the z196 server, because any PU can assume any available characterization. This also plays an essential role in system availability, because PU characterization can be done dynamically, with no server outage, allowing the actions discussed in the following sections.

Refer also to Chapter 8., "Software support" on page 209 for information about software level support on functions and features.

Concurrent upgrades

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

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

If the MCMs in the installed books have no available remaining PUs, an upgrade results in a model upgrade and the installation of an additional book (up to the limit of 4 books). Book installation is nondisruptive, but can take more time than a simple LIC upgrade.

For more information about Capacity on Demand, see Chapter 9, "System upgrades" on page 263.

PU sparing

In the rare event of a PU failure, the failed PU's characterization is dynamically and transparently reassigned to a spare PU. There are at least two spare PUs on a z196 server. PUs not characterized on a server configuration are also used as additional spare PUs. More information about PU sparing is provided in "Sparing rules" on page 89.

PU pools

PUs defined as CPs, IFLs, ICFs, zIIPs, and zAAPs are grouped together in their own pools, from where they can be managed separately. This significantly simplifies capacity planning and management for logical partitions. The separation also has an effect on weight management because CP, zAAP, and zIIP weights can be managed separately. For more information, see "PU weighting" on page 79.

All assigned PUs are grouped together in the PU pool. These PUs are dispatched to online logical PUs. As an example, consider a z196 with ten CPs, three zAAPs, two IFLs, two zIIPs, and one ICF. This system has a PU pool of 18 PUs, called the *pool width*. Subdivision of the PU pool defines:

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

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

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

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

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

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

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

PU weighting

Because zAAPs, zIIPs, IFLs, and ICFs have their own pools from where they are dispatched, they can be given their own weights. For more information about PU pools and processing weights, see *zEnterprise 196 Processor Resource/Systems Manager Planning Guide*, SB10-7155.

3.4.1 Central processors

A central processor (CP) is a PU that uses the full z/Architecture instruction set. It can run z/Architecture-based operating systems (z/OS, z/VM, TPF, z/TPF, z/VSE, Linux) and the Coupling Facility Control Code (CFCC). Up to 80 PUs can be characterized as CPs, depending on the configuration.

The z196 can only be initialized in LPAR mode. CPs are defined as either dedicated or shared. Reserved CPs can be defined to a logical partition to allow for nondisruptive *image* upgrades. If the operating system in the logical partition supports the *logical processor add* function, reserved processors are no longer needed. Regardless of the installed model, a logical partition can have up to 80 logical CPs defined (the sum of active and reserved logical CPs). We recommend defining no more CPs than the operating system supports.

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

Granular capacity

The z196 recognizes four distinct capacity settings for CPs. Full-capacity CPs are identified as CP7. In addition to full-capacity CPs, three subcapacity settings (CP6, CP5, and CP4), each for up to 15 CPs, are offered. The four capacity settings appear in hardware descriptions, as follows:

- ▶ CP7 feature code 1880
- ▶ CP6 feature code 1879
- ▶ CP5 feature code 1878
- ▶ CP4 feature code 1877

Granular capacity adds 45 subcapacity settings to the 80 capacity settings that are available with full capacity CPs (CP7). Each of the 45 subcapacity settings applies only to up to 15 CPs, independently of the model installed.

Information about CPs in the remainder of this chapter applies to all CP capacity settings, CP7, CP6, CP5, and CP4, unless indicated otherwise. See 2.8, "Model configurations" on page 52, for more details about granular capacity.

Capacity marker

A capacity marker indicates the presence of purchased capacity. For example, a model M15 can be configured with five full capacity CP7s, of which one has been purchased but is not used. The capacity level is identified by FC 1880, and model capacity marker 705 (FC 7251) indicates the purchased capacity. As one of the purchased CPs is not activated, FC 9004 (Downgraded PUs Per Request) is also included.

The same applies to the subcapacity models. For example, when a model M15 is configured with four subcapacity CP5s, of which one CP has been purchased but is not used. The capacity level is identified with FC 1878, and capacity marker 504 (FC 7220) indicates the purchased capacity. As on previous example, FC 9004 is also included.

> **Note:** Capacity settings smaller than the full-capacity setting (CP6, CP5, and CP4) only apply to up to 15 PUs characterized as CPs. Specialty engines such as IFLs, ICFs, zAAPs, and zIIPs always run at full capacity.

3.4.2 Integrated facility for Linux

An integrated facility for Linux (IFL) is a PU that can be used to run Linux or Linux guests on z/VM operating systems. Up to 80 PUs may be characterized as IFLs, depending on the configuration. IFLs can be dedicated to a Linux or a z/VM logical partition, or can be shared by multiple Linux guests or z/VM logical partitions running on the same z196 server. Only z/VM and Linux on System z operating systems and designated software products can run on IFLs. IFLs are orderable by feature code (FC 1881).

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

IFLs do not change the model capacity identifier of the z196. Software product license charges based on the model capacity identifier are not affected by the addition of IFLs.

Unassigned IFLs

An IFL that is purchased but not activated is registered as an unassigned IFL (FC 1886). When the system is subsequently upgraded with an additional IFL, the system recognizes that an IFL was already purchased and is present.

3.4.3 Internal coupling facilities

An internal coupling facility (ICF) is a PU used to run the coupling facility control code (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 model. At least a model M32 is necessary to characterize 16 ICFs (model M15 supports only up to 15 ICFs). ICFs are orderable by feature code (FC 1882).

Only CFCC can run on ICFs. ICFs do not change the model capacity identifier of the z196. Software product license charges based on the model capacity identifier are not affected by the addition of ICFs.

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

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

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

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

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

In Figure 3-7, 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.
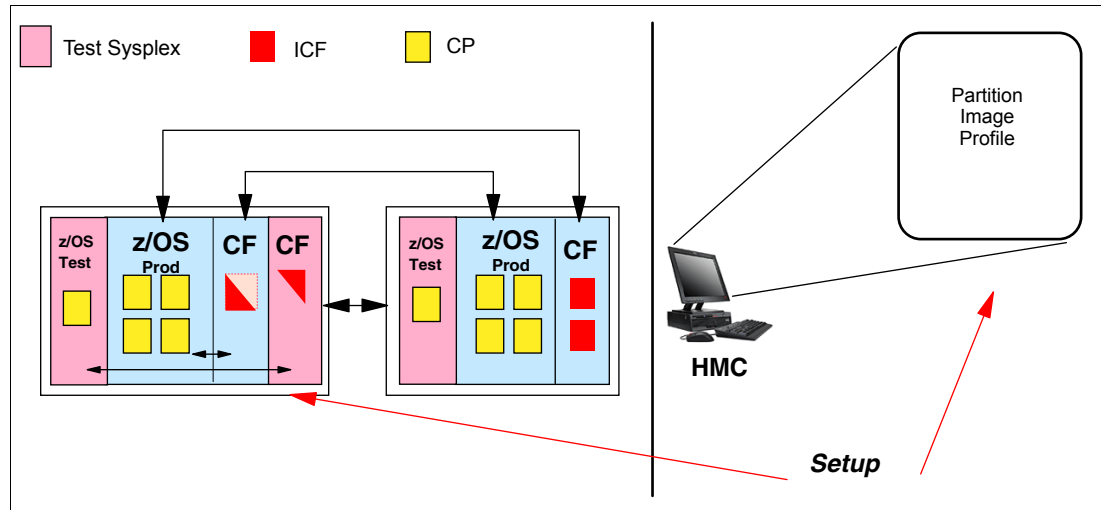
Figure 3-7   ICF options; shared ICFs

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

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

Dynamic coupling facility dispatching

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

3.4.4 System z application assist processors

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

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

zAAP benefits include:

▶   Potential cost savings

► Simplification of infrastructure as a result of the integration of new applications with their associated database systems and transaction middleware (such as DB2, IMS, or CICS). Simplification can happen, for example, by introducing a uniform security environment, reducing the number of TCP/IP programming stacks and server interconnect links

► Prevention of processing latencies that would occur if Java application servers and their database servers were deployed on separate server platforms

One CP must be installed with or prior to installing a zAAP. The number of zAAPs in a server cannot exceed the number of purchased CPs. Within the capacity of the sum of all unassigned PUs in up to four books, up to 40 zAAPs on a model M80 can be characterized. Table 3-1 shows the allowed number of zAAPs for each model.

Table 3-1   Number of zAAPs per model

Model	M15	M32	M49	M66	M80
zAAPs	0 – 7	0 – 16	0 – 24	0 – 33	0 – 40

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

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

zAAPs are orderable by feature code (FC 1884). Up to one zAAP can be ordered for each CP or marked CP configured in the server.

zAAPs and logical partition definitions

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

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

How zAAPs work

zAAPs are designed for z/OS Java code execution. When Java code must be executed (for example, under control of WebSphere), the z/OS Java Virtual Machine (JVM) calls the function of the zAAP. The z/OS dispatcher then suspends the JVM task on the CP it is running on and dispatches it on an available zAAP. After the Java application code execution is finished, z/OS redispatches the JVM task on an available CP, after which normal processing is resumed. This process reduces the CP time needed to run Java WebSphere applications, freeing capacity for other workloads.

Figure 3-8 shows the logical flow of Java code running on a z196 that has a zAAP available. When JVM starts execution of a Java program, it passes control to the z/OS dispatcher that will verify the availability of a zAAP:

► If a zAAP is available (not busy), the dispatcher suspends the JVM task on the CP, and assign the Java task to the zAAP. When the task returns control to the JVM, it passes control back to the dispatcher that reassigns the JVM code execution to a CP.

▶ If no zAAP is available (all busy) at that time, the z/OS dispatcher may allow a Java task to run on a standard CP, depending on the option used in the OPT statement in the IEAOPTxx member of SYS1.PARMLIB.
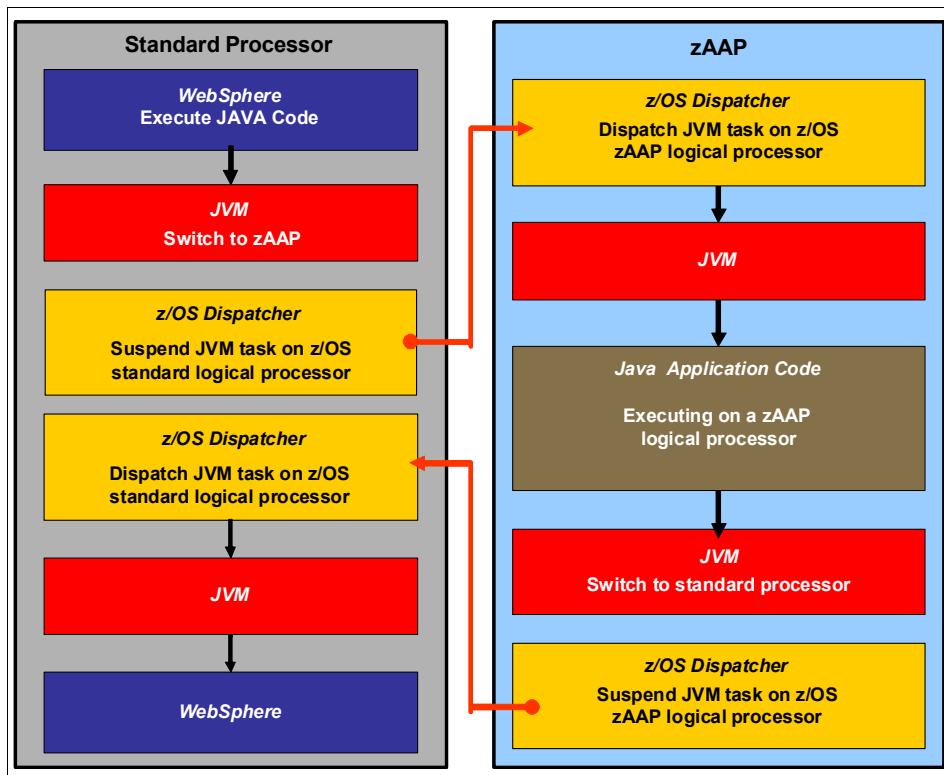
Figure 3-8   Logical flow of Java code execution on a zAAP

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

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

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

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

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

This option has the effect of running as much zAAP-eligible work on zAAPs as possible and only allowing it to spill over onto the CPs when the zAAPs are overcommitted.

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

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

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

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

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

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

3.4.5 System z integrated information processor

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

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

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

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

DRDA enables relational data to be distributed among multiple platforms. It is native to DB2 for z/OS, thus reducing the need for additional gateway products that can affect performance and availability. The application uses the DRDA requestor or server to access a remote database. (DB2 Connect™ is an example of a DRDA application requester.)

► Star schema query processing, mostly used in Business Intelligence (BI) work

A star schema is a relational database schema for representing multidimensional data. It stores data in a central fact table and is surrounded by additional dimension tables holding information about each perspective of the data. A star schema query, for example, joins several dimensions of a star schema data set.

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

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

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

z/OS communications server exploits the zIIP for eligible internet protocol security (IPSec) network encryption workloads. This requires z/OS V1R8 with PTFs or later releases. Portions of IPSec processing take advantage of the zIIPs, specifically end-to-end encryption with IPSec. The IPSec function moves a portion of the processing from the general-purpose

processors to the zIIPs. In addition to performing the encryption processing, the zIIP also handles cryptographic validation of message integrity and IPSec header processing.

z/OS Global Mirror, formerly known as Extended Remote Copy (XRC), exploits the zIIP too. Most z/OS DFSMS system data mover (SDM) processing associated with zGM is eligible to run on the zIIP. This requires z/OS V1R8 with PTFs or later releases.

The first IBM exploiter of z/OS XML system services is DB2 V9. With regard to DB2 V9 prior to the z/OS XML system services enhancement, z/OS XML system services non-validating parsing was partially directed to zIIPs when used as part of a distributed DB2 request through DRDA. This enhancement benefits DB2 V9 by making all z/OS XML system services non-validating parsing eligible to zIIPs when processing is used as part of any workload running in enclave SRB mode.

z/OS communications server also allows the HiperSockets Multiple Write operation for outbound large messages (originating from z/OS) to be performed by a zIIP. Application workloads based on XML, HTTP, SOAP, Java, etc as well as traditional file transfer, can benefit.

For business intelligence, IBM Scalable Architecture for Financial Reporting provides a high-volume, high performance reporting solution by running many diverse queries in z/OS batch and can also be eligible for zIIP.

For more information, see the IBM zIIP Web site:

http://www-03.ibm.com/systems/z/advantages/ziip/about.html

zIIP installation information

One CP must be installed with or prior to any zIIP being installed. The number of zIIPs in a server cannot exceed the number of CPs and unassigned CPs in that server. Within the capacity of the sum of all unassigned PUs in up to four books, up to 40 zIIPs on a model M80 can be characterized. Table 3-2 shows the allowed number of zIIPs for each model.

Table 3-2   Number of zIIPs per model

Model	M15	M32	M49	M66	M80
Maximum zIIPs	0 – 7	0 – 16	0 – 24	0 – 33	0 – 40

zIIPs are orderable by feature code (FC 1885). Up to one zIIP can be ordered for each CP or marked CP configured in the server. If the installed books have no remaining unassigned PUs, the assignment of the next zIIP may require the installation of an additional book.

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

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

zIIPs and logical partition definitions

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

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

3.4.6 zAAP on zIIP capability

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

► The combined eligible workloads may make the zIIP acquisition more cost effective.

► When zIIPs are already present, investment is maximized by running the Java and z/OS XML System Services-based workloads on existing zIIPs.

This capability does not eliminate the need to have one or more CPs for every zIIP processor in the server. Support is provided by z/OS. See 8.3.2, "zAAP support" on page 222.

When zAAPs are present[1] this capability is not available, as it is neither intended as a replacement for zAAPs, which continue to be available, nor as an overflow possibility for zAAPs. IBM does not recommend converting zAAPs to zIIPs in order to take advantage of the zAAP to zIIP capability:

► Having both zAAPs and zIIPs maximizes the system potential for new workloads.

► zAAPs have been available for over five years and there may exist applications or middleware with zAAP-specific code dependencies. For example, the code may use the number of installed zAAP engines to optimize multithreading performance.

It is a good idea to plan and test before eliminating all zAAPs, as there may be application code dependencies that may affect performance.

3.4.7 System assist processors

A system assist processor (SAP) is a PU that runs the channel subsystem licensed internal code (LIC) to control I/O operations.

All SAPs perform I/O operations for all logical partitions. All models have standard SAPs configured. The number of standard SAPs depends on the z196 server model, as shown in Table 3-3.

Table 3-3   SAPs per model

Model	M15	M32	M49	M66	M80
Standard SAPs	3	6	9	12	14

SAP configuration

A standard SAP configuration provides a very well-balanced system for most environments. However, there are application environments with very high I/O rates (typically some TPF environments). In this case, optional additional SAPs can be ordered. Assignment of

---

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

additional SAPs can increase the capability of the channel subsystem to perform I/O operations. In z196 servers, the number of SAPs can be greater than the number of CPs.

Optional additional orderable SAPs

An option available on all models is additional orderable SAPs (FC 1883). These additional SAPs increase the capacity of the channel subsystem to perform I/O operations, usually suggested for Transaction Processing Facility (TPF) environments. The maximum number of optional additional orderable SAPs depends on the configuration and the number of available uncharacterized PUs. The number of SAPs are listed in Table 3-4.

Table 3-4   Optional SAPs per model

Model	M15	M32	M49	M66	M80
Optional SAPs	0 – 3	0 – 7	0 – 11	0 – 18	0 – 18

Optionally assignable SAPs

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

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

3.4.8 Reserved processors

Reserved processors are defined by the Processor Resource/Systems Manager (PR/SM) to allow for a nondisruptive *capacity* upgrade. Reserved processors are like spare *logical* processors, and can be shared or dedicated. Reserved CPs should be defined to a logical partition to allow for nondisruptive *image* upgrades.

> **Note:** If the operating system in the logical partition supports the logical processor add function, reserved processors are no longer needed.

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

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

When no reserved processors are defined to a logical partition running an operating system that does not support the logical processor add function, an addition of a processor to that logical partition is disruptive, requiring the following tasks:

1. Partition deactivation
2. A logical processor definition change

3. Partition activation

The maximum number of reserved processors that can be defined to a logical partition depends on the number of logical processors that are already defined. The maximum number of logical processors plus reserved processors is 80.

Do not define more active and reserved processors than the operating system for the logical partition can support. For more information about logical processors and reserved processors definition see 3.6, "Logical partitioning" on page 93.

3.4.9 Processor unit assignment

Processor unit assignment of characterized PUs is done at power-on reset (POR) time, when the server is initialized. The intention of this initial assignment rules is to keep PUs of the same characterization type grouped together as much as possible regarding PU chips and books boundaries, to optimize shared cache usage.

The PU assignment is based on book plug ordering. This defines the low order and the high order books, as follows:

▶ Book 0: plug order 4 (when plugged this is the low order book).

▶ Book 1: plug order 1 (when Book 0 is not plugged this is the low order book).

▶ Book 2: plug order 3.

▶ Book 3: plug order 2.

The assignment rules follow this order:

▶ Spares: book 1 and 3 get assigned one spare each on the high PU chip, core 1, then core 0. If not available, then look at the next highest PU chip, core 1, then core 0.

▶ SAPs: spread across books and high PU chips. Start with high PU chip high core, then next highest PU chip high core. This prevents all the SAPs from being assigned on one PU chip.

▶ CPs: fill PU chip and spill into next chip on low order book first before spilling over into next book.

▶ ICFs: fill high PU chip on high book.

▶ IFLs: fill high PU chip on high book.

▶ zAAPs: attempts are made to align these close the CPs.

▶ zIIPs: attempts are made to align these close the CPs.

This implementation is to isolate as much as possible on different books (and even on different PU chips) processors that are used by different operating systems, so they do not use the same shared caches. CPs, zAAPs, and zIIPs are all used by z/OS, and can benefit by using the same shared caches. IFLs are used by z/VM and Linux, and ICFs are used by CFCC, so for performance reasons the assignment rules prevent them to share L3 and L4 caches with z/OS processors.

This initial PU assignment done at POR can be dynamically rearranged by LPAR, to improve system performance (see , "LPAR dynamic PU reassignment" on page 98).

When an additional book is added concurrently after POR 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. Only after the next POR that the processor unit assignment rules take into consideration the newly installed book.

3.4.10 Sparing rules

On a z196 system, two PUs are reserved as spares. The reserved spares are available to replace two characterized PUs, whether it is a CP, IFL, ICF, zAAP, zIIP, or SAP.

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

Sparing rules are as follows:

► When a PU failure occurs on a chip that has four active cores, the two standard spare PUs are used to recover the failing PU and the parent PU that shares function (for example, the compression unit and CPACF) with the failing PU, even though only one of the PUs has failed.

► When a PU failure occurs on a chip that has three active cores, one standard spare PU is used to replace the PU that does not share any function with another PU.

► When no spares are left, non-characterized PUs are used for sparing, following the previous two rules.

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

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

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

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

Application preservation

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

Dynamic SAP sparing and reassignment

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

3.4.11 Increased flexibility with z/VM-mode partitions

z196 provides a capability for the definition of a z/VM-mode logical partition that contains a mix of processor types including CPs and specialty processors, such as IFLs, zIIPs, zAAPs, and ICFs.

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

► Manage guests that exploit Linux on System z on IFLs, z/VSE and z/OS on CPs.

► Execute designated z/OS workloads, such as parts of DB2 DRDA processing and XML, on zIIPs.

► Provide an economical Java execution environment under z/OS on zAAPs.

3.5 Memory design

The z196 memory design also provides flexibility and high availability, allowing:

► Concurrent memory upgrades (if the physically installed capacity is not yet reached)

The z196 may have more physically installed memory than the initial available capacity. Memory upgrades within the physically installed capacity can be done concurrently by LIC, and no hardware changes are required. Concurrent memory upgrades can be done through Capacity on Demand. Note that memory upgrades *cannot* be done through Capacity BackUp (CBU) or On/Off CoD.

► Concurrent memory upgrades (if the physically installed capacity is reached)

Physical memory upgrades require a book to be removed and re-installed after having replaced the memory cards in the book. Except for a model M15, the combination of enhanced book availability and the flexible memory option allow you to concurrently add memory to the system. For more information see 2.5.5, "Book replacement and memory" on page 46, and 2.5.6, "Flexible memory option" on page 46.

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

Memory allocation

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

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

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

Large page support

By default, page frames are allocated with a 4 KB size. The z196 supports a large page size of 1 MB. The first z/OS release that supports large pages is z/OS V1R9. Linux on System z support for large pages is available in Novell SUSE SLES 10 SP2 and Red Hat RHEL 5.2.

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

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

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

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

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

Very short-running work does not overcome the overhead; short processes with small working sets are expected to provide little or no improvement. Long-running work with high memory-access frequency is the best candidate to benefit from large pages.

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

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

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

3.5.1 Central storage

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

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

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

3.5.2 Expanded storage

Expanded storage can optionally be defined on z196. Expanded storage is physically a section of processor storage. It is controlled by the operating system and transfers 4 KB pages to and from central storage.

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

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

The z196 only runs in LPAR mode. Storage is placed into a single storage pool called LPAR Single storage pool, which can be dynamically converted to expanded storage and back to central storage as needed when partitions are activated or de-activated.

LPAR single storage pool

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

On the hardware management console (HMC), the storage assignment tab of a reset profile shows the *customer storage,* which is the total installed storage minus the 16 GB hardware system area. Logical partitions are still defined to have central storage and, optionally, expanded storage.

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

3.5.3 Hardware system area

The hardware system area (HSA) is a non-addressable storage area that contains server Licensed Internal Code and configuration-dependent control blocks. The HSA has a fixed size of 16 GB and is not part of the purchased memory that you order and install.

The fixed size of the HSA eliminates planning for future expansion of the HSA because HCD/IOCP always reserves:

► Four channel subsystems (CSSs)
► Fifteen logical partitions in each CSS for a total of 60 logical partitions
► Subchannel set 0 with 63.75 K devices in each CSS
► Subchannel set 1 with 64 K devices in each CSS
► Subchannel set 2 with 64 K devices in each CSS

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

3.6 Logical partitioning

Logical partitioning (LPAR) is a function implemented by the Processor Resource/Systems Manager (PR/SM) on all z196 servers. The z196 runs only in LPAR mode. This means that all system aspects are controlled by PR/SM functions.

PR/SM is aware of the book structure on the z196. Logical partitions, however, do not have this awareness. Logical partitions have resources allocated to them from a variety of physical resources. From a systems standpoint, logical partitions have no control over these physical resources, but the PR/SM functions do.

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

As seen at "Processor unit assignment" on page 88, the initial PU assignment is done during power-on-reset (POR), using rules to optimize cache usage. This is the "physical" step, where CPs, zIIPs, zAAPs, IFLs, ICFs, and SAPs are allocated on books.

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

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

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

Book level assignments are more important, as this optimizes L4 cache usage. So logical processors from a given logical partition are packed into a book (or books) as much as possible.

Then PR/SM optimizes chip assignments within the assigned book (or books), to maximize L3 cache efficiency. So logical processors from a logical partition are dispatched on physical processors on the same PU chip as much as possible. Note that the number of processors per chip (four) matches the number of z/OS processor affinity queues (also four), used by HiperDispatch, achieving optimal cache usage within an affinity node.

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

HiperDispatch

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

Performance can be optimized by redispatching units of work to same processor group, keeping processes running near their cached instructions and data, and minimizing transfers of data ownership among processors/books.

The nested topology is returned to z/OS by the Store System Information (STSI) 15.1.3 instruction, and HiperDispatch utilizes the information to concentrate logical processors around shared caches (L3 at PU chip level, and L4 at book level), and dynamically optimizes assignment of logical processors and units of work.

z/OS dispatcher manages multiple queues, called affinity queues, with a target number of four processors per queue, which fits nicely into a single PU chip. These queues are used to assign work to as few logical processors as are needed for a given logical partition workload. So, even if the logical partition is defined with a large number of logical processors, HiperDispatch optimizes this number of processors nearest to the required capacity. The optimal number of processors to be used are kept within a book boundary where possible.

Logical partitions

PR/SM enables z196 servers to be initialized for a logically partitioned operation, supporting up to 60 logical partitions. Each logical partition can run its own operating system image in any image mode, independent from the other logical partitions.

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

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

▶ Processors

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

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

The weight and the number of online logical processors of a logical partition can be dynamically managed by the LPAR CPU Management function of the Intelligent Resource Director to achieve the defined goals of this specific partition and of the overall system. The provisioning architecture of the z196, described in Chapter 9, "System upgrades" on page 263, adds another dimension to dynamic management of logical partitions.

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

- In support of WLC, the user can specify a defined capacity in millions of service units (MSUs) per hour. The defined capacity sets the capacity of an individual logical partition when soft capping is selected.

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

- WLM keeps a 4-hour rolling average of the CPU usage of the logical partition, and when the 4-hour average CPU consumption exceeds the defined capacity limit, WLM dynamically activates LPAR capping (soft capping). When the rolling 4-hour average returns below the defined capacity, the soft cap is removed.

For more information regarding WLM, see *System Programmer's Guide to: Workload Manager*, SG24-6472.

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

► Memory

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

Reserved storage can be defined to a logical partition, enabling nondisruptive memory addition to and removal from a logical partition, using the LPAR dynamic storage reconfiguration (z/OS and z/VM). For more information see 3.6.4, "LPAR dynamic storage reconfiguration" on page 102.

► Channels

Channels can be shared between logical partitions by including the partition name in the partition list of a channel path identifier (CHPID). I/O configurations are defined by the input/output configuration program (IOCP) or the hardware configuration 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 identifiers (PCHIDs) that represent the physical location of a port on a card in an I/O cage or I/O drawer.

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

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

Modes of operation

Table 3-5 shows the modes of operation, summarizing all available mode combinations: operating modes and their processor types, operating systems, and addressing modes.

Table 3-5   z196 modes of operation

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

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

For information about operating system support see Chapter 8, "Software support" on page 209.

Logically partitioned mode

The z196 only runs in LPAR mode. Each of the 60 logical partitions can be defined to operate in one of the following image modes:

► ESA/390 mode, to run:

 – A z/Architecture operating system, on dedicated *or* shared CPs

 – An ESA/390 operating system, on dedicated *or* shared CPs

 – A Linux on System z operating system, on dedicated *or* shared CPs

 – z/OS, on any of the following processor units:

 • Dedicated *or* shared CPs
 • Dedicated CPs *and* dedicated zAAPs *or* zIIPs
 • Shared CPs *and* shared zAAPs *or* zIIPs

> **Note:** zAAPs and zIIPs can be defined to an ESA/390 mode or z/VM-mode image (Table 3-5 on page 96). However, zAAPs and zIIPs are supported only by z/OS. Other operating systems cannot use zAAPs or zIIPs, even if they are defined to the logical partition. z/VM V5R3 and later can provide zAAPs or zIIPs to a guest z/OS.

► ESA/390 TPF mode, to run TPF or z/TPF operating system, on dedicated *or* shared CPs

► Coupling facility mode, by loading the CFCC code into the logical partition defined as:

 – Dedicated *or* shared CPs

- – Dedicated *or* shared ICFs

► Linux-only mode, to run:

- – A Linux on System z operating system, on either:
  - • Dedicated *or* shared IFLs
  - • Dedicated *or* shared CPs

- – A z/VM operating system, on either:
  - • Dedicated *or* shared IFLs
  - • Dedicated *or* shared CPs

► z/VM-mode to run z/VM on dedicated *or* shared CPs or IFLs, plus zAAPs, zIIPs, and ICFs.

Table 3-6 shows all LPAR modes, required characterized PUs, operating systems, and the PU characterizations that can be configured to a logical partition image. The available combinations of dedicated (DED) and shared (SHR) processors are also shown. For all combinations, a logical partition can also have reserved processors defined, allowing nondisruptive logical partition upgrades.

Table 3-6   LPAR mode and PU usage

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

Dynamic add or delete of a logical partition name

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

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

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

Add crypto feature to a logical partition

You can preplan the addition of Crypto Express3 features to a logical partition on the crypto page in the image profile by defining the Cryptographic Candidate List, Cryptographic Online List and usage, and Control Domain Indices in advance of installation. By using the Change LPAR Cryptographic Controls task, adding crypto dynamically to a logical partition without an outage of the logical partition is possible. Also, dynamic deletion or moving of these features no longer requires pre-planning. Support is provided in z/OS, z/VM, z/VSE, and Linux on System z.

LPAR group capacity limit

The group capacity limit feature allows the definition of a capacity limit for a group of logical partitions on z196 servers. This feature allows a capacity limit to be defined for each logical partition running z/OS, and to define a group of logical partitions on a server. This allows the system to manage the group in such a way that the sum of the LPAR group capacity limits in MSUs per hour will not be exceeded. To take advantage of this, you must be running z/OS V1.8 or later and all logical partitions in the group have to be z/OS V1.8 and later.

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

LPAR dynamic PU reassignment

System configuration has been enhanced to optimize the PU-to-book assignment of physical processors dynamically. The initial assignment of customer usable physical processors to physical books, as described on 3.4.9, "Processor unit assignment", can change dynamically to better suit the actual logical partition configurations that are in use. Swapping of specialty engines and general processors with each other, with spare PUs, or with both, can occur as the system attempts to compact logical partition configurations into physical configurations that span the least number of books.

LPAR dynamic PU reassignment can swap customer processors of different types between books. For example, reassignment can swap an IFL on book 1 with a CP on book 2. Swaps can also occur between PU chips within a book, and includes spare PUs in what can be swapped within books. The goals are to further pack the logical partition on fewer books and also on fewer PU chips, based on the z196 book topology. The effect of this is evident in dedicated and shared logical partitions that use HiperDispatch.

LPAR dynamic PU reassignment is transparent to operating systems.

3.6.1 Storage operations

In z196 servers, memory can be assigned as a combination of central storage and expanded storage, supporting up to 60 logical partitions. Expanded storage is only used by the z/VM operating system.

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

Central storage can be dynamically assigned to expanded storage and back to central storage as needed without a power-on reset (POR). For details see "LPAR single storage pool" on page 92.

Memory *cannot* be shared between system images. It is possible to dynamically reallocate storage resources for z/Architecture logical partitions running operating systems that support dynamic storage reconfiguration (DSR). This is supported by z/OS, and z/VM V5R4 and later releases. z/VM in turn virtualizes this support to its guests. For details see 3.6.4, "LPAR dynamic storage reconfiguration" on page 102.

Operating systems running under z/VM can exploit the z/VM capability of implementing virtual memory to guest virtual machines. The z/VM dedicated *real* storage can be *shared* between guest operating systems.

Table 3-7 shows the z196 storage *allocation* and *usage* possibilities, depending on the image mode.

Table 3-7   Storage definition and usage possibilities

Image mode	Architecture mode (addressability)	Maximum central storage		Expanded storage	
		Architecture	z196 definition	z196 definable	Operating system usage[1]
ESA/390	z/Architecture (64-bit)	16 EB	1 TB	Yes	Yes
	ESA/390 (31-bit)	2 GB	128 GB	Yes	Yes
z/VM[2]	z/Architecture (64-bit)	16 EB	256 GB	Yes	Yes
ESA/390 TPF	ESA/390 (31-bit)	2 GB	2 GB	Yes	No
Coupling facility	CFCC (64-bit)	1.5 TB	1 TB	No	No
Linux only	z/Architecture (64-bit)	16 EB	256 GB	Yes	*Only by z/VM*
	ESA/390 (31-bit)	2 GB	2 GB	Yes	*Only by z/VM*

1. z/VM supports the use of expanded storage.
2. z/VM-mode is supported by z/VM V5R4 and later.

ESA/390 mode

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

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

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

The modes are:

▶ z/Architecture mode

In z/Architecture mode, storage addressing is 64-bit, allowing for virtual addresses up to 16 exabytes (16 EB). The 64-bit architecture theoretically allows a maximum of 16 EB to be used as central storage. However, the current central storage limit for logical partitions is 1 TB of central storage. The operating system that runs in z/Architecture mode has to be

able to support the real storage. Currently, z/OS for example, supports up to 4 TB of real storage (z/OS V1.8 and higher releases).

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

► ESA/390 architecture mode

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

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

> **Note:** Either a z/Architecture mode or an ESA/390 architecture mode operating system can run in an ESA/390 image on a z196. Any ESA/390 image can be defined with more than 2 GB of central storage *and* can have expanded storage. These options allow you to configure more storage resources than the operating system is capable of addressing.

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

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

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

Coupling facility mode
In coupling facility mode, storage addressing is 64-bit for a coupling facility image running CFCC Level 12 or later, allowing for an addressing range up to 16 EB. However, the current z196 definition limit for logical partitions is 1 TB of storage.

CFCC Level 17 is available for the z196. CFCC Level 17 allows:

► Greater than 1024 CF Structures
  – New limit 2047
► Greater than 32 Connectors
  – New limits: 255 cache, 247 lock, or 127 serialized list
► Improved CFCC diagnostics & Link Diagnostics
► An increase from 64 to 128 CHPIDs

For details see .

Expanded storage cannot be defined for a coupling facility image. Only IBM CFCC can run in coupling facility mode.

Linux-only mode

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

Only Linux and z/VM operating systems can run in Linux-only mode. Linux on System z 64-bit distributions (Novell SUSE SLES 10 and later, Red Hat RHEL 5 and later) use 64-bit addressing and operate in the z/Architecture mode. z/VM also uses 64-bit addressing and operates in the z/Architecture mode.

3.6.2 Reserved storage

Reserved storage can optionally be defined to a logical partition, allowing a nondisruptive image memory upgrade for this partition. Reserved storage can be defined to both central and expanded storage, and to any image mode, except the coupling facility mode.

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

► The initial value is the storage size allocated to the partition when it is activated.

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

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

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

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

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

A concurrent logical partition storage upgrade uses dynamic storage reconfiguration (DSR). z/OS uses the reconfigurable storage unit (RSU) definition to add or remove storage units in a nondisruptive way.

z/VM V5R4 and later releases support the dynamic addition of memory to a running logical partition by using reserved storage, and also virtualizes this support to its guests. Removal of storage from the guests or z/VM is disruptive.

SUSE Linux Enterprise Server (SLES) 11 supports both concurrent add and remove.

3.6.3 Logical partition storage granularity

Granularity of central storage for a logical partition depends on the largest central storage amount defined for either initial or reserved central storage, as shown in Table 3-8.

Table 3-8 Logical partition main storage granularity

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

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

Expanded storage granularity is fixed at 256 MB.

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

3.6.4 LPAR dynamic storage reconfiguration

Dynamic storage reconfiguration on z196 servers allows an operating system running in a logical partition to add (nondisruptively) its reserved storage amount to its configuration, if any unused storage exists. This unused storage can be obtained when another logical partition releases some storage or when a concurrent memory upgrade takes place.

With dynamic storage reconfiguration, the unused storage does not have to be continuous.

When an operating system running in a logical partition assigns a storage increment to its configuration, Processor Resource/Systems Manager (PR/SM) determines whether any free storage increments are available and dynamically brings the storage online.

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

3.7 Intelligent resource director

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

IRD is a feature that extends the concept of goal-oriented resource management by allowing grouping system images that are resident on the same System z running in LPAR mode, and in the same Parallel Sysplex, into an *LPAR cluster*. This gives Workload Manager the ability to manage resources, both processor and I/O, not just in one single image, but across the entire cluster of system images.

Figure 3-9 shows an LPAR cluster. It contains three z/OS images, and one Linux image managed by the cluster. Note that included as part of the entire Parallel Sysplex is another z/OS image, and a coupling facility image. In this example, the scope that IRD has control over is the defined LPAR cluster.
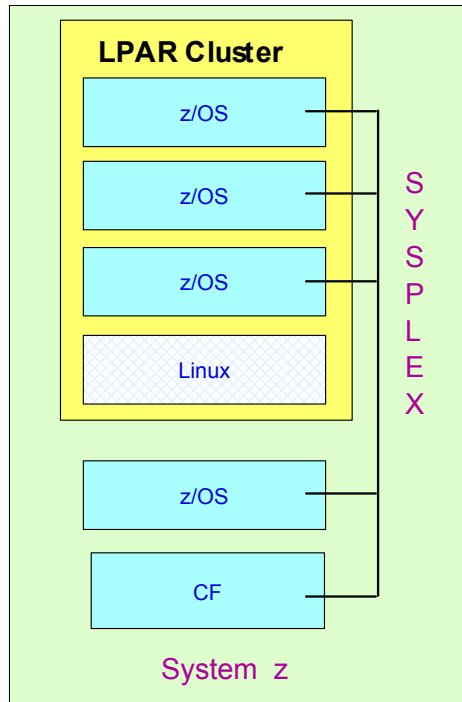

Figure 3-9   IRD LPAR cluster example

IRD addresses three separate but mutually supportive functions:

► LPAR CPU management

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

HiperDispatch was introduced in 3.6, "Logical partitioning" on page 93.

HiperDispatch manages the number of logical CPs in use. It adjusts the number of logical processors within a logical partition in order to achieve the optimal balance between CP resources and the requirements of the workload in the logical partition. When HiperDispatch is active the LPAR CPU management part of IRD is automatically deactivated.

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

► Dynamic channel path management (DCM)

DCM moves ESCON and FICON channel bandwidth between disk control units to address current processing needs. The z196 supports DCM within a channel subsystem.

► Channel subsystem priority queuing

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

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

3.8 Clustering technology

Parallel Sysplex continues to be the clustering technology used with z196 servers. Figure 3-10 illustrates the components of a Parallel Sysplex as implemented within the System z architecture. The figure is intended only as an example. It shows one of many possible Parallel Sysplex configurations. Many other possibilities exist.
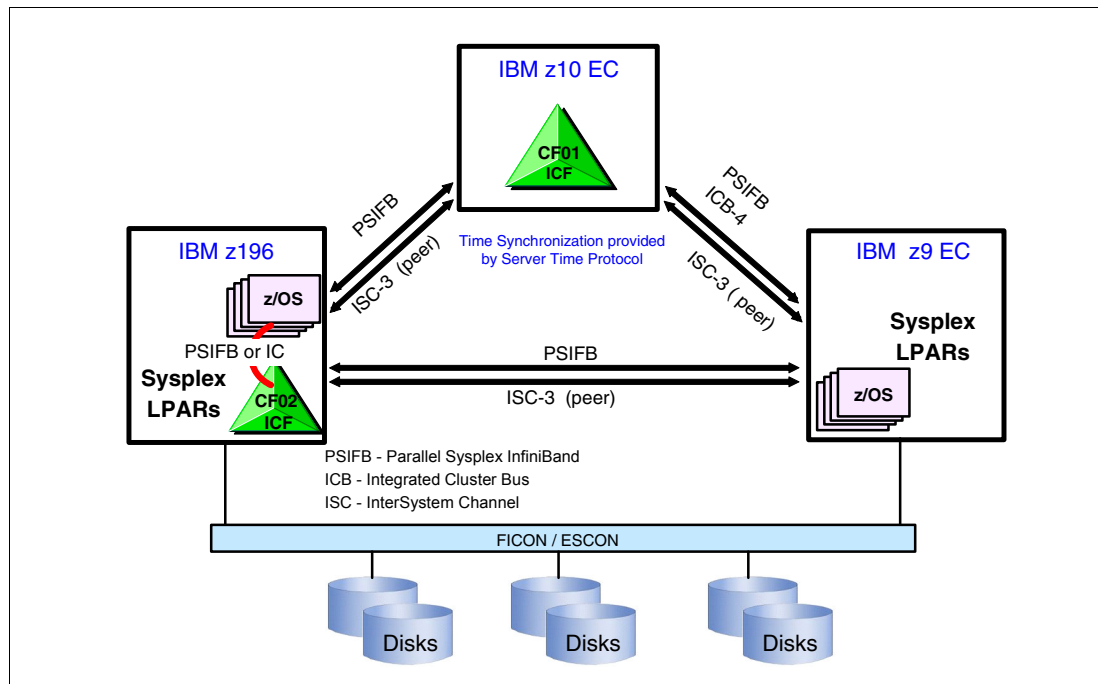
Figure 3-10   Sysplex hardware overview

Figure 3-10 shows a z196 containing multiple z/OS sysplex partitions and an internal coupling facility (CF02), a z10 EC containing a stand-alone ICF (CF01), and a z9 EC containing multiple z/OS sysplex partitions. STP over coupling links provides time synchronization to all servers. CF link technology (PSIFB, ICB-4, ISC-3) selection depends on sever configuration. Link technologies are described in , "Coupling link features" on page 143.

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

► Continuous (application) availability: Changes can be introduced, such as software upgrades, one image at a time, while the remaining images continue to process work. For details, see *Parallel Sysplex Application Considerations*, SG24-6523.

► High capacity: Scales can be from 2 to 32 images.

► Dynamic workload balancing: Viewed as a single logical resource, work can be directed to any similar operating system image in a Parallel Sysplex cluster having available capacity.

► Systems management: Architecture provides the infrastructure to satisfy customer requirements for continuous availability, and provides techniques for achieving simplified systems management consistent with this requirement.

► Resource sharing: A number of base (z/OS) components exploit coupling facility shared storage. This exploitation enables sharing of physical resources with significant improvements in cost, performance, and simplified systems management.

► Single system image: The collection of system images in the Parallel Sysplex appears as a single entity to the operator, the user, the database administrator, and so on. A single system image ensures reduced complexity from both operational and definition perspectives.

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

Coupling facility control code

Coupling facility control code (CFCC) Level 17 is made available on the z196.

CFCC Level 17 allows an increase in the number of CHPID from 64 to 128. (This applies to IC, 12x IFB, 1x IFB, and active ISC-3 links.) This constraint relief can help supporting better CF link throughput, since each coupling CHPID carries with it only 7 primary command link buffers, each capable of performing a CF operation. z/OS maps these buffers to subchannels. By allowing more subchannels, more parallel CF operations can be serviced and therefore CF link throughput can increase.

CFCC level 17 now supports up to 2048 structures. When sysplex was first implemented, only 64 structures were supported in the sysplex. Before very long, sysplex exploitation took off, and customers levied requirements up to 1024 structures with CFCC level 16. New exploiters demand for more structures, some examples are:

► Logical groupings such as DB2, IMS, and MQ datasharing groups, for which multiple group instances may exist in the same sysplex (each potentially with many structures)

► "Service provider" customers who provide IT services for many customers, and define large numbers of individual small datasharing groups, one per customer

► Customer mergers, acquisitions, sysplex consolidations often grow the requirements in quantum leaps rather than slow/steady "compound" growth

CFCC level 17 now supports more than 32 connectors. A connector to a structure is a specific instance of the exploiting product or subsystem, who is running on a particular system in the sysplex. A sysplex can contain at most 32 z/OS system images. In situations where subsystem-specific constraints on the amount of capacity or throughput that can be achieved within a single exploiter instance (e.g. threading constraints, virtual storage constraints, common storage constraints) can be relieved by defining two or more instances of the exploiter, the demand for structure connectors can increase above 32. CFCC level 17 now supports 255 connectors for cache structures, 247 for lock structures, or 127 for serialized list structures.

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

Dynamic CF dispatching

Dynamic CF dispatching provides the following function on a coupling facility:

1. If there is no work to do, CF enters a wait state (by time).

2. After an elapsed time, CF wakes up to see whether there is any new work to do (requests in the CF Receiver buffer).

3. If there is no work, CF sleeps again for a longer period of time.

4. If there is new work, CF enters into the normal active wait until there is no more work, starting the process all over again.

This function saves processor cycles and is an excellent option to be used by a production backup CF or a testing environment CF. This function is activated by the CFCC command DYNDISP ON.

The CPs can run z/OS operating system images and CF images. For software charging reasons, using only ICF processors to run coupling facility images is better.
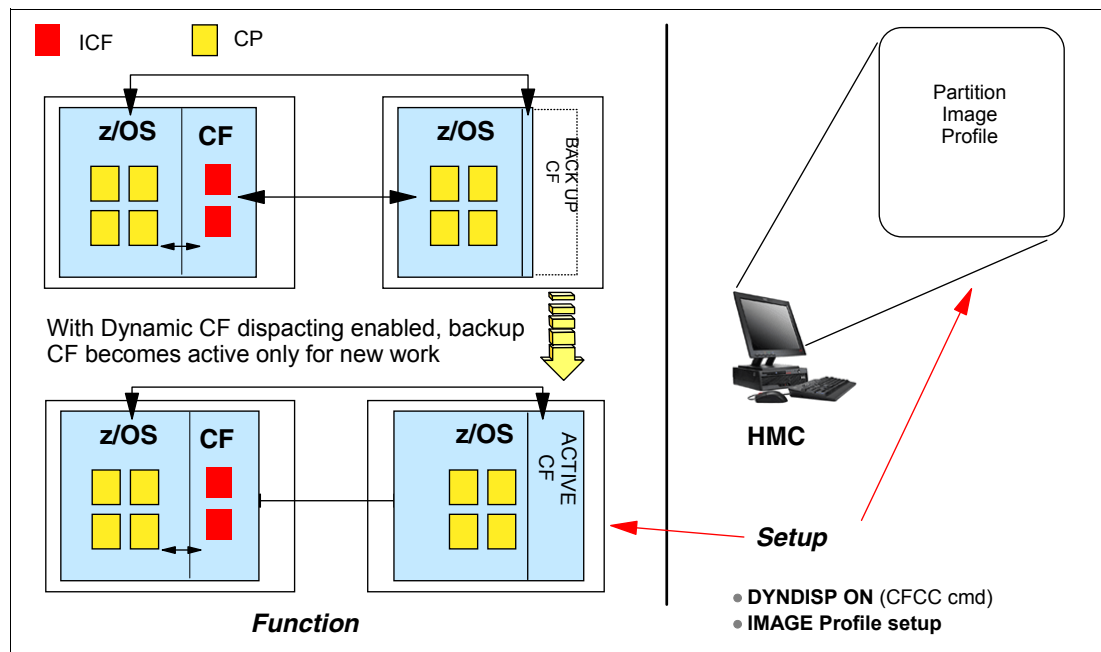
Figure 3-11 shows the dynamic CF dispatching.

Figure 3-11   Dynamic CF dispatching (shared CPs or shared ICF PUs)

For additional details regarding CF configurations, see *Coupling Facility Configuration Options,* GF22-5042, also available from the Parallel Sysplex Web site:

http://www.ibm.com/systems/z/advantages/pso/index.html

4

CPC I/O system structure

This chapter describes the I/O system structure and the connectivity options available on the zEnterprise 196 (z196).

This chapter discusses the following topics:

- ▶ 4.1, "Introduction" on page 108
- ▶ 4.2, "I/O system overview" on page 109
- ▶ 4.3, "I/O cages" on page 110
- ▶ 4.4, "I/O drawers" on page 113
- ▶ 4.5, "Fanouts" on page 115
- ▶ 4.6, "I/O feature cards" on page 123
- ▶ 4.7, "Connectivity" on page 127
- ▶ 4.8, "Parallel Sysplex connectivity" on page 141

4.1 Introduction

The z196 I/O system structure uses InfiniBand as the interconnect protocol for various connectivity types, to support both I/O cages and I/O drawers, and to satisfy different requirements of I/O connectivity.

Before describing the InfiniBand implementation on the z196, we provide a short general introduction to InfiniBand.

> **Note:** Not all properties and functions offered by InfiniBand are implemented on the z196. Only a subset is used to fulfill the interconnect requirements that have been defined for z196.

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

Similar to Fibre Channel, PCI Express, Serial ATA, and many other contemporary interconnects, InfiniBand is a point-to-point, bidirectional serial link intended for the connection of processors with high-speed peripherals, such as disks. InfiniBand supports several signalling rates and, as with PCI Express, links can be bonded together for additional bandwidth.

The serial connection's signalling rate is 2.5 Gbps on one lane in each direction (SDR)[1], per physical connection. InfiniBand also supports double (DDR) and quad speeds (QDR), for 5 Gbps or 10 Gbps, respectively.

4.1.1 Data, signalling, and link rates

Links use 8b/10b encoding (every ten bits sent carries eight bits of data), so that the useful data transmission rate is four-fifths of the signalling rate (signalling rate equals raw bit rate). Thus, single, double, and quad rates carry 2, 4, or 8 Gbps of useful data, respectively.

Links can be aggregated in units of 4 or 12, indicated as 4x[2] or 12x. A quad-rate 12x (12x QDR) link therefore carries 120 Gbps raw or 96 Gbps of payload (useful) data. Larger systems with 12x links are typically used for cluster and supercomputer interconnects, as implemented on the z196, and for inter-switch connections.

Table 4-1 lists the effective theoretical InfiniBand data throughput in different configurations.

Table 4-1   Effective data rates of aggregated links

Number of links	Single (SDR)	Double (DDR)	Quad (QDR)
1X	2 Gbps	4 Gbps	8 Gbps
4X	8 Gbps	16 Gbps	32 Gbps
12X	24 Gbps	48 Gbps	96 Gbps

Throughout this chapter the following terminology is used:

Data rate   The data transfer rate is expressed in bytes; one byte equals eight bits.

---

[1] SDR is Single Data Rate, DDR is Dual Data Rate, QDR is Quad Data Rate
[2] z196 does not support this data rate.

Signalling rate   The raw bit rate is expressed in bits.

Link rate   The rate is equal to the signalling rate expressed in bits.

For details and the standard for InfiniBand, see the InfiniBand Web site:

http://www.infinibandta.org

4.2 I/O system overview

This section lists characteristics and a summary of features that are supported.

4.2.1 Characteristics

The z196 I/O subsystem design provides great flexibility, high availability, and excellent performance characteristics, as follows:

► High bandwidth

  The z196 uses InfiniBand as the internal interconnect protocol to drive ESCON and FICON channels, OSA ports, and ISC-3 coupling links. As a connection protocol, InfiniBand supports InfiniBand coupling (PSIFB[3]) with a link rate of up to 6 GBps.

► Connectivity options

  The z196 can be connected to an extensive range of interfaces such as ESCON, FICON/Fibre Channel Protocol for storage area network connectivity, 10 Gigabit Ethernet, Gigabit Ethernet, and 1000BASE-T Ethernet for local area network connectivity, and ISC-3 coupling links.

► Concurrent I/O upgrade

  You may concurrently add I/O cards to the server if an unused I/O slot position is available. Additional I/O cages can be installed in advance to provide greater capacity for concurrent upgrades.

► Concurrent I/O drawer upgrade

  Additional I/O drawers can be installed concurrently without preplanning.

► Dynamic I/O configuration

  Dynamic I/O configuration supports the dynamic addition, removal, or modification of channel path, control units, and I/O devices without a planned outage.

► Pluggable optics

  The FICON Express8 and FICON Express4 features have Small Form Factor Pluggable (SFP) optics to permit each channel to be individually serviced in the event of a fiber optic module failure. The traffic on the other channels on the same feature can continue to flow if a channel requires servicing.

► Concurrent I/O card maintenance

  Each I/O card plugged in an I/O cage supports concurrent card replacement in case of a repair action.

4.2.2 Summary of supported I/O features

The following I/O features are supported:

---

[3] Parallel Sysplex InfiniBand

- ► Up to 240 ESCON channels (up to 360 with an RPQ 8P2507)
- ► Up to 288 FICON Express4 channels (when carried forward on upgrade only)
- ► Up to 288 FICON Express8 channels (up to 336 with an RPQ 8P2506)
- ► Up to 48 OSA-Express2 ports (when carried forward on upgrade only)
- ► Up to 96 OSA-Express3 ports
- ► Up to 48 ISC-3 coupling links
- ► Up to 32 InfiniBand coupling links (12x InfiniBand, 1x InfiniBand)

> **Note:** The maximum number of external coupling links combined (ISC-3, and PSIFB coupling links) cannot exceed 80 for each z196 server.

4.3 I/O cages

The z196 supports up to four I/O drawers, two I/O cages, or combinations of each. Installation of a third I/O cage requires RPQ 8P2506 support.

Each cage supports up to seven I/O domains (named A to G) for a total of 28 I/O card slots. Each I/O domain supports four I/O card slots, as shown in Figure 4-1 on page 111.
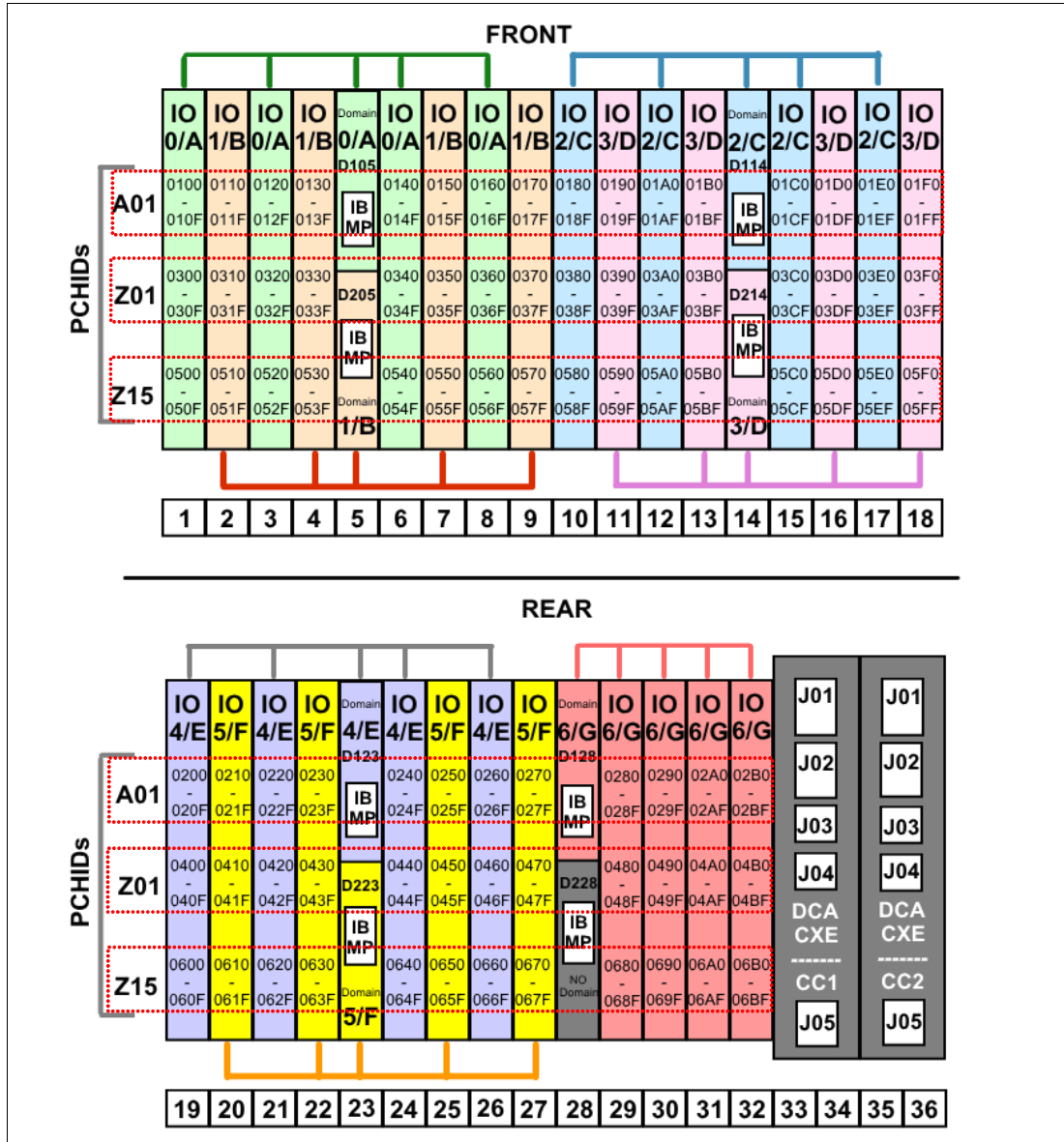
Figure 4-1   I/O cage

Each I/O domain uses an IFB-MP card (IB MP in Figure 4-1 on page 111) in the I/O cage and a copper cable connected to an Host Channel Adapter (HCA) fanout in the CPC cage. A maximum of seven I/O domains are available in each cage. An eighth IFB-MP card is installed to provide an alternate path to I/O cards in slots 29, 30, 31, and 32 in case of a repair action.

Domain number 6 (G) is not used until all other domains are full in all installed I/O cages and I/O drawers. If more than 32 or 44 I/O cards are required, a new I/O drawer or I/O cage must be installed. Only when more than 72 I/O cards are required will the third I/O cage be installed, which requires an RPQ. Table 4-3 on page 113 list the combinations of I/O cages and I/O drawers.

Figure 4-2 on page 112 illustrates the I/O structure of a z196. An InfiniBand (IFB) cable connects the HCA2-C fanout to an IFB-MP card in the I/O cage. The passive connection between two IFB-MP cards allows for redundant I/O interconnection. The IFB cable between

an HCA2-C fanout in a book, and each IFB-MP card in the I/O cage supports a 6 GBps bandwidth.
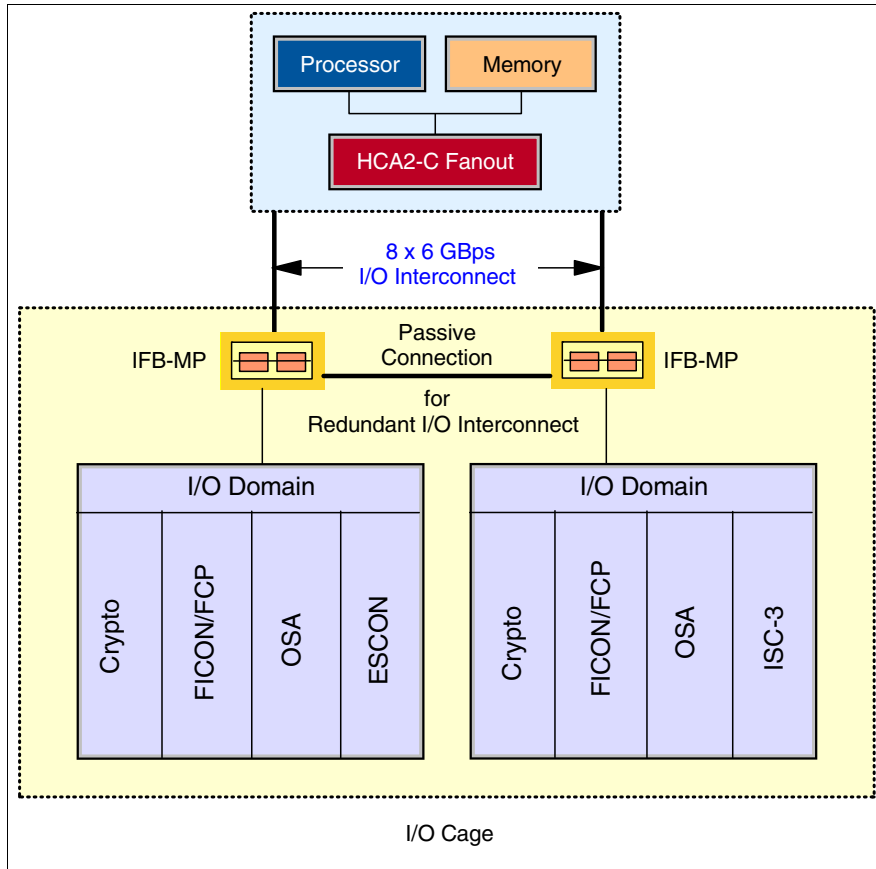
Figure 4-2   z196 I/O structure when using I/O cages

> **Note:** Installing an additional I/O cage is disruptive.

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

A fully populated system with two I/O cages and two I/O drawers installed has a total of 72 available I/O card slots. With an RPQ, the third I/O cage can be installed to provide a total of 84 available I/O card slots.

Table 4-2 lists the I/O domains and their related I/O slots (see also Figure 4-1 on page 111).

Table 4-2   I/O domains

Domain number (name)	I/O slot in domain
0 (A)	01, 03, 06, 08
1 (B)	02, 04, 07, 09
2 (C)	10, 12, 15, 17
3 (D)	11, 13, 16, 18
4 (E)	19, 21, 24, 26

Domain number (name)	I/O slot in domain
5 (F)	20, 22, 25, 27
6 (G)	29, 30, 31, 32

The configuration process selects which I/O slots are used for I/O cards and provides the required number of I/O cages, I/O drawers, HCA2-C fanout cards, IFB-MP cards, and IFB cables, either for a new build server or a server upgrade.

If you order the Power Sequence Controller (PSC) feature, the PSC24V card is always plugged into slot 29 of domain G. Installing a PSC24V card is always disruptive.

Note: It is intended that the z196 is the last high end server to support Power Sequence Controller feature.

4.4 I/O drawers

The z196 supports up to four I/O drawers, up to two I/O cages, or combinations of each. The I/O drawer is five EIA units high and supports up to eight I/O feature cards. Combinations of I/O drawers and I/O cages can be installed to support up to 72 I/O feature cards (with two I/O drawers and two I/O cages), or 84 I/O feature cards with RPQ 8P2506, supporting a third I/O cage, as shown in Table 4-3. Installing an I/O drawer is nondisruptive and can be done concurrently in a running system.

Table 4-3   Combinations of I/O drawers and cages

Number of I/O feature cards	Number of I/O drawers	Number of I/O cages
1-8	1	0
9-16	2	0
17-24	3	0
25-32	4	0
33-36	1	1
37-44	2	1
45-52	3	1
53-60	4	1
45-56	0	2
57-64	1	2
65-72	2	2
73-84 (RPQ required)	0	3

Each drawer supports two I/O domains (A and B) for a total of eight I/O card slots. Each I/O domain uses an IFB-MP card in the I/O drawer and a copper cable to connect to a Host Channel Adapter (HCA) fanout in the CPC cage. The link between the HCA in the CPC and the IFB-MP in the I/O drawer supports a link rate of up to 6 GBps. All cards in the I/O drawer

are installed horizontally. The two distributed converter assemblies (DCAs) distribute power to the I/O drawer. The locations of the DCAs, I/O feature cards, and IFB-MP card in the I/O drawer are shown in Figure 4-3.
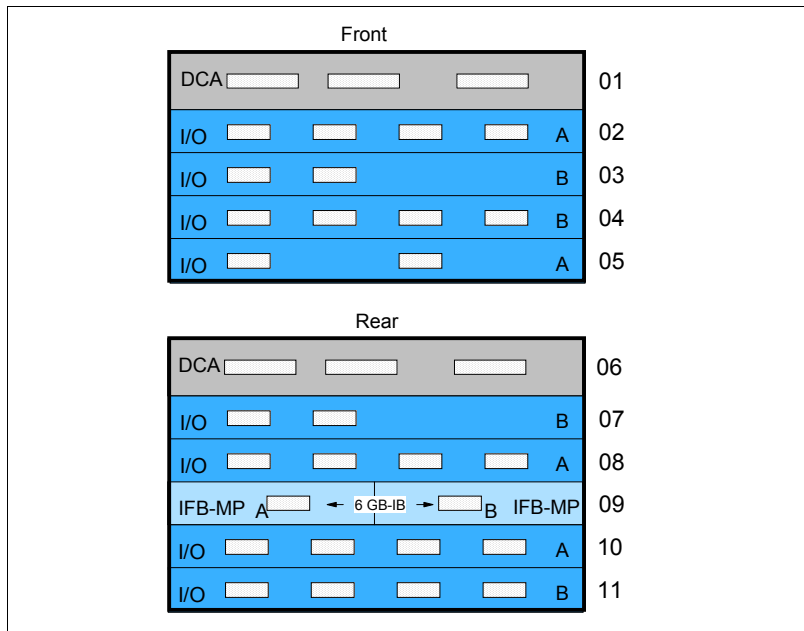

Figure 4-3   I/O feature cards plugging locations in an I/O drawer

The IFB-MP cards are installed at location 09 at the rear side of the I/O drawer. The I/O cards are installed from the front and rear side of the I/O drawer. Two I/O domains (A and B) are supported. Each I/O domain has up to four I/O feature cards of any type (ESCON, FICON, ISC or OSA). The I/O cards are connected to the IFB-MP card through the backplane board.

The I/O structure in a z196 server is illustrated in Figure 4-4. An IFB cable connects the HCA fanout card to an IFB-MP card in the I/O drawer. The passive connection between two IFB-MP cards allows redundant I/O interconnection. This provides connectivity between an HCA fanout card, and I/O cards in case of concurrent fanout card or IFB cable replacement. The IFB cable between an HCA fanout card and each IFB-MP card supports a 6 GBps link rate.
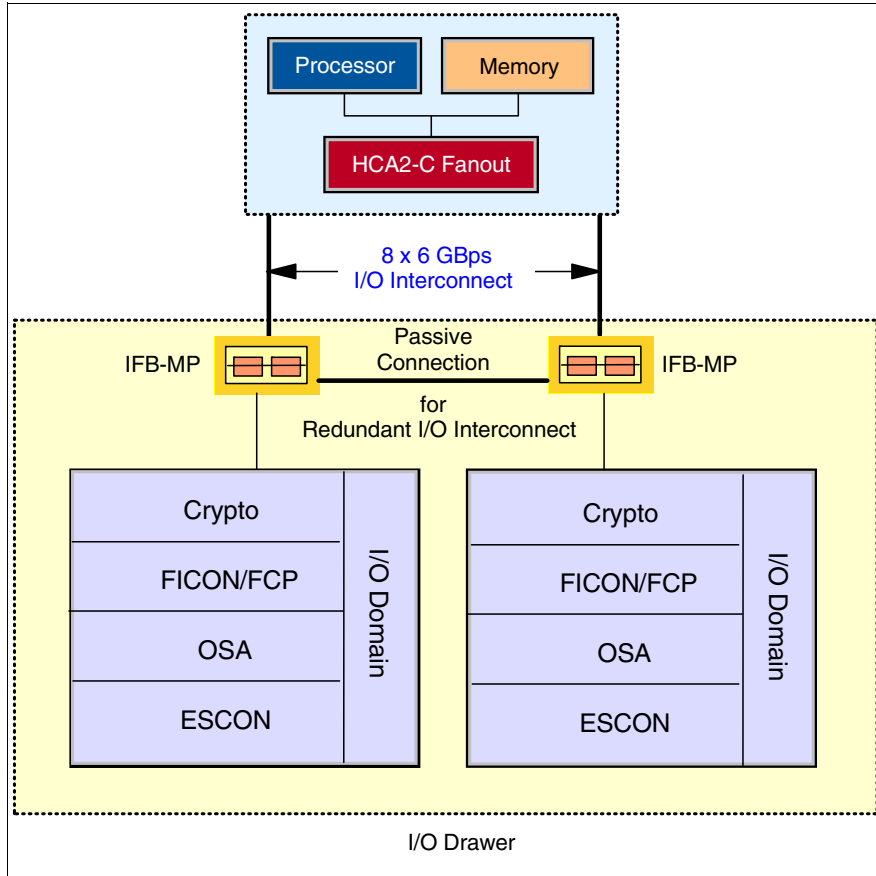
Figure 4-4   z196 I/O structure when using I/O drawers

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

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

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

If the Power Sequence Controller (PSC) feature is ordered, the PSC24V card is always plugged into slot 11 of the first I/O drawer. Installing the PSC24V card is always disruptive.

> **Note:** It is intended that the z196 is the last high end server to support Power Sequence Controller feature.

4.5 Fanouts

InfiniBand offers a point-to-point bidirectional serial, high-bandwidth, low-latency link that is used for the connection of processors. Its use is introduced for the connection to other systems in a Parallel Sysplex, and for the internal connection to the I/O cages and

I/O drawers in which the cards for the connection to peripheral devices and networks reside. The InfiniBand fanouts are located in the front of each book.

Each book has eight fanout slots. They are named D1 to DA, top to bottom; slots D3 and D4 are not used for fanouts. Each fanout has two ports to connect an InfiniBand (IFB) cable, depending on the type of fanout. There are three types of Host Channel Adapters (HCAs). One uses a copper cable (HCA2-C) to connect to an I/O cage or I/O drawer, the other two use optical connections (HCA2-O, HCA2-O LR). The optical cabling exits the front of the book. Each slot holds one of the following three fanouts:

► Host Channel Adapter (HCA2-C): This copper fanout provides connectivity to the IFB-MP card in the I/O cage and I/O drawer.

► Host Channel Adapter (HCA2-O): This optical fanout provides 12x InfiniBand coupling link connectivity up to 150 meters (492 feet) distance to a z196 and other z10 and z9 servers.

► Host Channel Adapter (HCA2-O LR): This optical long range fanout provides 1x InfiniBand coupling link connectivity up to 10 km (6.2 miles) unrepeated distance to a z196 and other z10 servers.

Figure 4-5 illustrates the IFB connection from the CPC cage to an I/O cage, an I/O drawer, and coupling using InfiniBand (PSIFB).
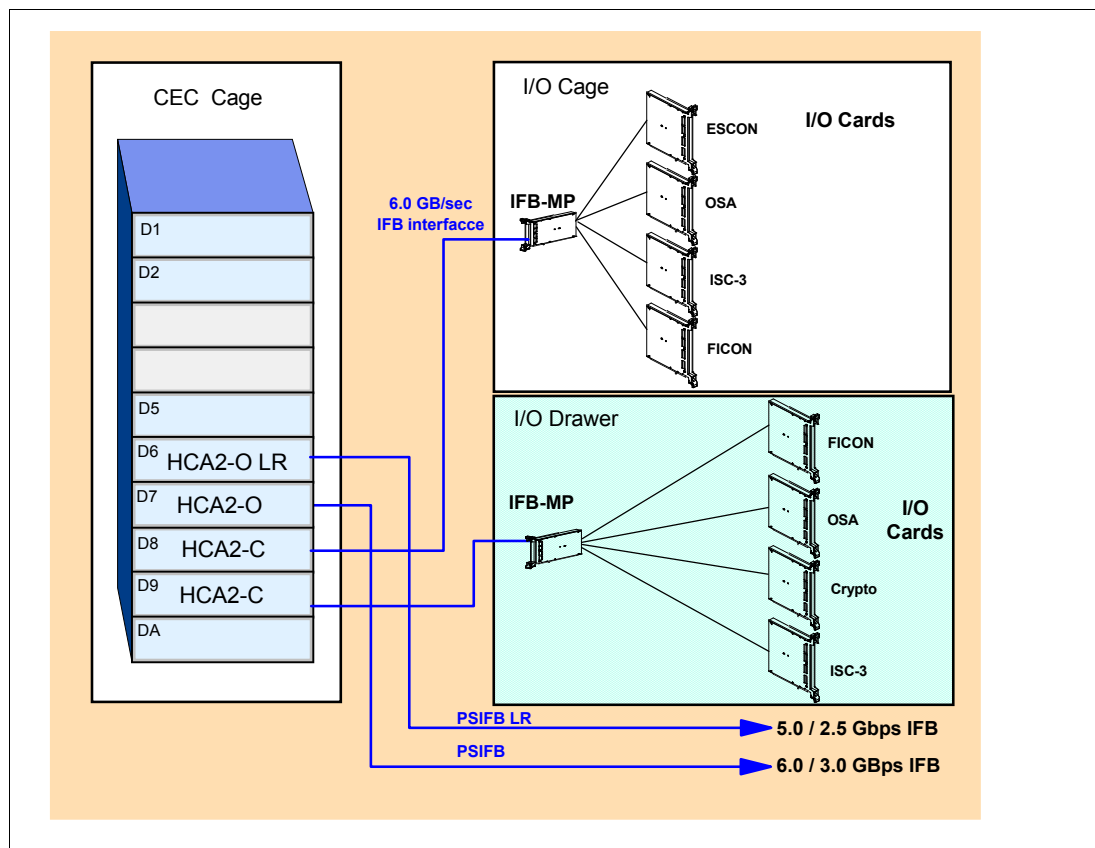
Figure 4-5   PSIFB, IFB I/O drawer and IFB I/O cage interface connections

4.5.1 HCA2-C fanout

The HCA2-C fanout is used to connect to an I/O cage or an I/O drawer using a copper cable. The two ports on the fanout are dedicated to I/O. The bandwidth of each port on the HCA2-C fanout supports a link rate of up to 6 GBps.

A 12x InfiniBand DDR copper cable of 1.5 to 3.5 meters long is used for connection to the IFB-MP card in the I/O cage or the I/O drawer. For a z196 having two I/O drawers and two I/O cages fully populated, 10 HCA2-C fanouts (20 ports) are required.

> **Note:** The HCA2-C fanout is used exclusively for I/O and cannot be shared for any other purpose.

4.5.2 HCA2-O fanout

The HCA2-O fanout for 12x InfiniBand DDR provides an optical interface used for coupling links. The two ports on the fanout are dedicated to coupling links to connect to z196, z10, and z9 servers, or to connect to a coupling port in the same server by using a fiber cable. Each fanout has an optical transmitter and receiver module and allows dual simplex operation. Up to 16 HCA2-O fanouts are supported and provide up to 32 ports for coupling links. The maximum of all PSIFB links (12x InfiniBand, 1x InfiniBand) is 32.

The HCA2-O fanout supports InfiniBand double data rate (12x IB-DDR) and InfiniBand single data rate (12x IB-SDR) optical links that offer longer distance, configuration flexibility, and high bandwidth for enhanced performance of coupling links. There are 12 lanes (two fibers per lane) in the cable, which means 24 fibers are used in parallel for data transfer.

The fiber optic cables are industry standard OM3 (2000 MHz-km) 50 μm multimode optical cables with Multi-Fiber Push-On (MPO) connectors. The maximum cable length is 150 meters (492 feet). There are 12 pairs of fibers, 12 fibers for transmitting, and 12 fibers for receiving.

Each fiber supports a link rate of 6 GBps (12x IB-DDR) if connected to a z196 server or z10 server, and 3 GBps (12x IB-SDR) when connected to a System z9 server. The link rate is auto-negotiated to the highest common rate.

> **Note:** Ports on the HCA2-O fanout are exclusively used for coupling links and cannot be used or shared for any other purpose.

A fanout has two ports for optical link connections and supports up to 16 CHPIDs across both ports. These CHPIDs are defined in IOCDS as coupling links.

> **Note:** The recommendation is to define only four CHPIDs for each port.

Each HCA2-O fanout used for coupling links has an assigned adapter ID (AID) number that must be used for definitions in IOCDS to create a relationship between the physical fanout location and the CHPID number. For details about AID numbering, see "Adapter ID number assignment" on page 121.

For detailed information about how the AID is used and referenced in HCD, see *Getting Started with InfiniBand on System z10 and System z9,* SG24-7539.

4.5.3 HCA2-O LR fanout

The HCA2-O LR fanout for 1x InfiniBand provides an optical interface used for coupling links. The two ports on the fanout are dedicated to coupling links to connect to z196 and z10 servers. Up to 16 HCA2-O LR fanouts are supported and provide 32 ports for coupling link. The maximum of all PSIFB links is 32.

The HCA-O LR fanout supports InfiniBand double data rate (1x IB-DDR) and InfiniBand single data rate (1x IB-SDR) optical links that offer longer distance of coupling links. The cable has one lane containing two fibers; one fiber is used for transmitting and one fiber used for receiving data.

Each fiber supports a link rate of 5 Gbps (1x IB-DDR) if connected to a z196 server, a z10 server or to a repeater (System z qualified DWDM[4]) supporting IB-DDR, and a data link rate of 2.5 Gbps (1x IB-SDR) when connected to a repeater (System z qualified DWDM) that supports IB-SDR. The link rate is auto- negotiated to the highest common rate.

> **Note:** Ports on the HCA2-O LR fanout are used exclusively for coupling links and cannot be used or shared for any other purpose.

The fiber optic cables are 9 μm single mode (SM) optical cables terminated with an LC Duplex connector. The maximum unrepeated distance is 10 km (6.2 miles) and up to 100 km (62 miles) with repeaters (System z qualified DWDM).

A fanout has two ports for optical link connections and supports up to 16 CHPIDs across both ports. These CHPIDs are defined in IOCDS as coupling links and require a fiber cable to connect to other z196 or z10 servers.

> **Note:** It is recommended that you define up to four CHPIDs per port.

Each HCA2-O LR fanout used for coupling links has an assigned adapter ID (AID) number that must be used for definitions in IOCDS to create a relationship between the physical fanout location and the CHPID number. See "Adapter ID number assignment" on page 121 for details about AID numbering.

4.5.4 Fanout considerations

Because fanout slots in each book can be used to plug different fanouts, where each fanout is designed for a special purpose, some restrictions might apply to the number of available channels located in the I/O cage and I/O drawer.

A fully populated server has two I/O drawers and two I/O cages. Each drawer requires two connections to support all eight slots, and each cage requires eight connections to support all 28 slots for I/O cards in an I/O cage. This is a total of 20 connections required, which is equivalent to 10 HCA2-C fanouts (20 ports) dedicated to I/O links.

If fewer than 10 HCA-C fanouts are available, the number of supported I/O cards and the number of CHPIDs available can decrease. The number of HCA2-C fanouts for cage and drawer connections depends on the number of HCA2-O LR and HCA2-O fanouts used for coupling links, and vice versa. Also, the fanouts for I/O are always plugged in pairs.

Depending on the model, the number of fanouts varies. The following sections show the relationship between number of fanouts used for coupling links and the remaining available I/O domains and CHPIDs for each model. The plugging rules for fanouts for each model are illustrated in Figure 4-6.

---

[4] dense wavelength division multiplexing

Figure 4-6   Fanout plugging rules

An addition or removal of an I/O cage is disruptive; however, an addition or removal of an I/O drawer can be performed concurrently. So one option is to populate a z196 with I/O drawers and no more than two I/O cages. A configuration with 64 or less I/O slots is ideal to leave an empty location for future expansion.

Fanouts in a model M15 (one book)

Model M15 has one book installed, supporting eight fanouts. The maximum number of (FICON) channels supported is 256 (two I/O drawers and one I/O cage in Z frame, one I/O cage in A frame). This number is decreased by each fanout used for a coupling link. For example, if four fanouts of any type designated for coupling links are installed, the maximum number of I/O domains is eight, supporting up to 128 (FICON) CHPIDs in four I/O drawers, as shown in Table 4-5 on page 120.

A maximum of eight HCA2-O LR and HCA2-O fanouts used for coupling links is supported.

Table 4-5   Available CHPIDs in I/O cage and drawer (one book)

Maximum of eight fanouts									
Number of HCA2-O LR and HCA2-O fanouts	0	1	2	3	4	5	6	7	8
Available I/O domains	16	12	12	8	8	4	4	0	0
Available I/O slots	64	48	48	32	32	16	16	0	0
Number of I/O cages (C) and drawers (D)	2 C 2 D	2 C	2 C	4 D	4 D	2 D	2 D	0	0
Maximum number of CHPIDS (FICON)	256	192	192	128	128	64	64	0	0

Fanouts in a model M32 (two books)

Model M32 has two books installed, supporting 16 fanouts. The maximum number of (FICON) channels supported is 288 using 72 features (across 18 domains). This number can decrease if fewer than 10 fanouts for I/O connectivity remain. See Table 4-6 for details.

A maximum of 16 HCA2-O LR and HCA2-O fanouts, used for coupling links, is supported.

Table 4-6   Available CHPIDs in I/O cage and drawer (two books)

Maximum of 16 fanouts										
Number of HCA2-O LR and HCA2-O fanouts	0 - 6	7	8	9	10	11	12	13	14	15	16
Available I/O domains	18	16	16	12	12	8	8	4	4	0	0
Available I/O slots	72	64	64	48	48	32	32	16	16	0	0
Number of I/O cages (C) and drawers (D)	2 C 2 D	2 C 2 D	2 C 2 D	2 C	2 C	4 D	4 D	2 D	2 D	0	0
Maximum number of CHPIDS (FICON)	288	256	256	192	192	128	128	64	64	0	0

Fanouts in model M49 (three books)

Model M49 has three books, supporting 20 fanouts. The maximum number of (FICON) channels supported is 288 using 72 features (across 18 domains). This number can decrease if fewer than 10 fanouts for I/O connectivity remain. See Table 4-7 for details.

A maximum of 16 HCA2-O LR and HCA2-O fanouts, used for coupling links, is supported.

Table 4-7   Available CHPIDs in I/O cage and drawer (three books)

Maximum of 20 fanouts							
Number of HCA2-O LR and HCA2-O fanouts	0 - 10	11	12	13	14	15	16
Available I/O domains		18	16	16	12	12	8	8
Available I/O slots		72	64	64	48	48	32	32
Number of I/O cages (C) and drawers (D)	2 C 2 D	2 C 2 D	2 C 2 D	2 C	2 C	4 D	4 D

Maximum of 20 fanouts							
Maximum number of CHPIDS (FICON)	288	256	256	192	192	128	128

Fanouts in models M66 and M80 (four books)

Models M66 and M80 have four books, supporting 24 fanouts. The maximum number of (FICON) channels supported is 288 using 72 features (across 18 domains). This number can decrease if fewer than 10 fanouts used for I/O connectivity remain. See Table 4-8 for details.

A maximum of 16 HCA2-O LR and HCA2-O fanouts, used for coupling links, is supported.

Table 4-8   Available CHPIDs in I/O cage and drawer (four book)

Maximum of 24 fanouts			
Number of HCA2-O LR and HCA2-O fanouts	0 - 14	15	16
Available I/O domains	18	16	16
Available I/O slots	72	64	64
Number of I/O cages (C) and drawers (D)	2 C 2 D	2 C 2 D	2 C 2 D
Maximum number of CHPIDS (FICON)	288	256	256

Fanouts with the 3rd I/O cage (RPQ 8P2506)

With RPQ 8P2506, three I/O cages can be installed in a z196, supporting up to 336 (FICON) channels using 84 features (256 on M15). See Table 4-9 for details.

Table 4-9   Available CHPIDs with three I/O cages

	M15	M32	M49	M66 M80
Number of books	1	2	3	4
Maximum Number of Fanouts	8	16	20	24
Available I/O domains	16	21	21	21
Available I/O slots	64	84	84	84
Maximum number of CHPIDS (FICON)	256	336	336	336

Adapter ID number assignment

Unlike channels installed in an I/O cage, which are identified by a PCHID number related to their physical location, PSIFB fanouts and ports are identified by an adapter ID (AID), initially dependent on their physical locations. This AID must be used to assign a CHPID to the fanout in the IOCDS definition. The CHPID assignment is done by associating the CHPID to an AID port.

Table 4-10 illustrates the AID assignment for each fanout slot relative to the book location on a new build system.

Table 4-10   AID number assignment

Book	Slot	Fanout slot	AIDs
First	6	D1, D2, D5–DA	08, 09, 0A –0F
Second	15	D1, D2, D5–DA	18, 19, 1A–1F
Third	10	D1, D2, D5–DA	10, 11, 12–17
Fourth	1	D1, D2, D5–DA	00, 01, 02–07

The fanout slots are numbered D1 to DA top to bottom, as shown in Table 4-11. All fanout locations and their AIDs for all four books are shown in the table for reference only. Fanouts in locations D1 and D2 are not available on all models. Slots D3 and D4 will never have a fanout installed (dedicated for FSPs).

> **Note:** Slots D1 and D2 are not used in a 4-book server, and only partially in a 3-book server.

Table 4-11   Fanout AID numbers

Fanout location	Fourth book	First book	Third book	Second book
D1	00	08	10	18
D2	01	09	11	19
D3	-	-	-	-
D4	-	-	-	-
D5	02	0A	12	1A
D6	03	0B	13	1B
D7	04	0C	14	1C
D8	05	0D	15	1D
D9	06	0E	16	1E
DA	07	0F	17	1F

> **Important Note:** The AID numbers in Table 4-11 are valid only for a new build server or for new books added. If a fanout is moved, the AID follows the fanout to its new physical location.

The AID assigned to a fanout is found in the PCHID REPORT provided for each new server or for MES upgrade on existing servers.

Example 4-1 shows part of a report, named PCHID REPORT, for a model M32. In this example, one fanout is installed in the first book (location 06) and one fanout is installed in the second book (location 15), both in location D5. The assigned AID for the fanout in the first book is 0A; the AID assigned to the fanout in the second book is 1A.

Example 4-1   AID assignment in PCHID report

```
CHPIDSTART
  08348295                    PCHID REPORT                    Mar 24,2010
```

```
Machine: 2817-M32  SNXXXXXXX
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
Source          Cage  Slot  F/C   PCHID/Ports or AID          Comment

06/D5           A25B  D506  0163   AID=0A

15/D5           A25B  D515  0163   AID=1A
```

4.5.5 Fanout summary

Fanout features supported by the z196 server are shown in Table 4-12. The table provides the feature type and code, total maximum (Max.) number of features and ports, and information about the link supported by the fanout feature.

Table 4-12   Fanout summary

Fanout feature	Feature code	Max. features	Max. ports	Use	Cable type	Connector type	Max. distance	Link data rate
HCA2-C	0162	12	24	Connect to I/O cage or drawer	Copper	n/a	3.5 m	6 GBps
HCA2-O	0163	16[a]	32	Coupling link	50 μm MM OM3 (2000 MHz-km)	MPO	150 m	6 GBps[b]
HCA2-O LR	0168	16[a]	32	Coupling link	9 μm SM	LC Duplex	10 km[c]	5.0 Gbps 2.5 Gbps[d]

a. A maximum of 16 combined of PSIFB features (FC 0163, 0168) in any combination is supported
b. 3 GBps link data rate if connected to a System z9 server
c. Up to 100 km with repeaters (System z qualified DWDM)
d. Auto-negotiated, depending on DWDM equipment

4.6 I/O feature cards

I/O cards have ports to connect the z196 to external devices, networks, or other servers. I/O cards are plugged into the I/O cage and I/O drawer based on the configuration rules for the server. Different types of I/O cards are available, one for each channel or link type. I/O cards can be installed or replaced concurrently.

4.6.1 I/O feature card types

The I/O features listed in Table 4-13 on page 123 can be ordered for newly built servers.

Table 4-13   I/O feature codes

Card type	Feature code
ESCON (16-port)	2323
FICON Express8 LX (10 km)	3325
FICON Express8 SX	3326
OSA-Express3 10 GbE LR	3370

Card type	Feature code
OSA-Express3 10 GbE SR	3371
OSA-Express3 GbE LX	3362
OSA-Express3 GbE SX	3363
OSA-Express3 1000BASE-T	3367
ISC-3	0217 (ISC-M) 0218 (ISC-D)
ISC-3 up to 20 km	RPQ 8P2197 (ISC-D)
Crypto Express3	0864

Table 4-14 lists I/O features that are available only if carried over during an upgrade.

Table 4-14　I/O feature codes

Card type	Feature code
FICON Express4 LX (4 km)	3324
FICON Express4 LX (10 km)	3321
FICON Express4 SX	3322
OSA-Express2 1000BASE-T	3366
OSA-Express2 GbE LX	3364
OSA-Express2 GbE SX	3365

4.6.2 PCHID report

A physical channel ID (PCHID) number is assigned to each I/O card port and the Crypto Express3 card plugged in the I/O cage and I/O drawer. Each enabled port has a PCHID number assigned, depending on the physical I/O slot location of where the card is plugged in, and on the physical port on the card.

Table 4-15 lists the PCHID assignments for slots in the I/O cages. Table 4-16 on page 125 lists the PCHID assignments for slots in the I/O drawers. Only the active ports on an installed card are actually assigned a PCHID. The remainder are unused.

Table 4-15　PCHID assignments for I/O cages

I/O cage slot[a]	PCHID numbers[b]		
	First I/O cage frame A bottom	Second I/O cage frame Z bottom	Third I/O cage frame Z top
1	100-10F	300-30F	500-50F
2	110-11F	310-31F	510-51F
3	120-12F	320-32F	520-52F
4	130-13F	330-33F	530-53F
6	140-14F	340-34F	540-54F
7	150-15F	350-35F	550-55F

I/O cage slot[a]	PCHID numbers[b]		
	First I/O cage frame A bottom	**Second I/O cage frame Z bottom**	**Third I/O cage frame Z top**
8	160-16F	360-36F	560-56F
9	170-17F	370-37F	570-57F
10	180-18F	380-38F	580-58F
11	190-19F	390-39F	590-59F
12	1A0-1AF	3A0-3AF	5A0-5AF
13	1B0-1BF	3B0-3BF	5B0-5BF
15	1C0-1CF	3C0-3CF	5C0-5CF
16	1D0-1DF	3D0-3DF	5D0-5DF
17	1E0-1EF	3E0-3EF	5E0-5EF
18	1F0-1FF	3F0-3FF	5F0-5FF
19	200-20F	400-40F	600-60F
20	210-21F	410-41F	610-61F
21	220-22F	420-42F	620-62F
22	230-23F	430-43F	630-63F
24	240-24F	440-44F	640-64F
25	250-25F	450-45F	650-65F
26	260-26F	460-46F	660-66F
27	270-27F	470-47F	670-67F
29	280-28F	480-48F	680-68F
30	290-29F	490-49F	690-69F
31	2A0-2AF	4A0-4AF	6A0-6AF
32	2B0-2BF	4B0-4BF	6B0-6BF

a. Slots 5, 14, 23, and 28 are reserved for IFB-MP cards.
b. The PCHID number range from 000 to 03F is reserved

Table 4-16　PCHID assignments for I/O drawers

Slot	PCHID Range					
	Drawer 1 Z22B	**Drawer 2 Z15B**	**Drawer 3 Z08B**	**Drawer 4 Z01B**	**Drawer 5 A08B(MRU) A16B(MWU)**	**Drawer 6 A01B(MRU) A09B(MWU)**
2	580-58F	500-50F	380-38F	300-30F	180-18F	100-10F
3	590-59F	510-51F	390-39F	310-31F	190-19F	110-11F
4	5A0-5AF	520-52F	3A0-3AF	320-32F	1A0-1AF	120-12F
5	5B0-5BF	530-53F	3B0-3BF	330-33F	1B0-1BF	130-13F

Slot	PCHID Range					
	Drawer 1 Z22B	Drawer 2 Z15B	Drawer 3 Z08B	Drawer 4 Z01B	Drawer 5 A08B(MRU) A16B(MWU)	Drawer 6 A01B(MRU) A09B(MWU)
7	5C0-5CF	540-54F	3C0-3CF	340-34F	1C0-1CF	140-14F
8	5D0-5DF	550-55F	3D0-3DF	350-35F	1D0-1DF	150-15F
10	5E0-5EF	560-56F	3E0-3EF	360-36F	1E0-1EF	160-16F
11	5F0-5FF	570-57F	3F0-3FF	370-37F	1F0-1FF	170-17F

A PCHID report is created for each new build server and for upgrades on existing servers. The report lists all I/O features installed, the physical slot location, and the assigned PCHID. Example 4-2 shows a portion of a sample PCHID report.

The AID numbering rules for InfiniBand coupling links are described in "Adapter ID number assignment" on page 121.

Example 4-2   PCHID report

```
CHPIDSTART
  08348295                    PCHID REPORT                    Mar 24,2010
  Machine: 2817-M32  SN2
  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
  Source          Cage  Slot  F/C    PCHID/Ports or AID          Comment

  06/D7           A25B  D706  0168   AID=0C

  15/D5           A25B  D515  0163   AID=1A

  06/DA/J01       A01B  04    3363   130/J00J01 131/J02J03

  15/DA/J01       A01B  08    2323   160/J00 161/J01 162/J02 163/J03
                                     164/J04 165/J05 166/J06 167/J07
                                     168/J08 169/J09 16A/J10 16B/J11
                                     16C/J12 16D/J13

  06/D8/J01       A01B  D120  0218   210/J00 211/J01

  06/D8/J01       A01B  D220  0218   218/J00 219/J01

  06/D9/J02       Z15B  04    3325   520/D1 521/D2 522/D3 523/D4

  06/DA/J02       Z22B  02    0864   580/P00 581/P01

  15/DA/J02       Z22B  03    3371   590/J00 591/J01

  15/DA/J02       Z22B  04    3367   5A0/J00J01 5A1/J02J03

  06/DA/J02       Z22B  05    3367   5B0/J00J01 5B1/J02J03
```

The following list explains the content of the sample PCHID REPORT:

► Feature code 0168 (HCA2-O LR) is installed in the first book (cage A25B, slot 06) location D7 and has AID 0C assigned.

► Feature code 0163 (HCA2-O) is installed in the second book (cage A25B, slot 15) location D5 and has AID 1A assigned.

► Feature code 3363 (OSA-Express3 GbE SX) is installed in cage A01B slot 4 and has PCHIDs 130 and 131 assigned. PCHID 130 is shared by port 00 and 01, PCHID 131 is shared by port 02 and 03.

► Feature code 2323 (ESCON 16-port) is installed in cage A01B slot 8 and has PHCHIDs 160 to 16D for the 14 ports enabled on that adapter card.

► Feature code 0218 (ISC-3) is installed in cage A01B slot 20 and has PCHID 210 and 211 assigned to the two ports on the upper daughter card, and PCHID 218 and 219 to the two ports on the lower daughter card.

► Feature code 3325 (FICON Express8 LX 10 km) is installed in drawer Z15B slot 4 and has PCHIDs 520, 521, 522, and 523 assigned.

► Feature code 0864 (Crypto Express3) is installed in drawer Z22B slot 2 and has PCHIDs 580 and 581 assigned.

► Feature code 3371 (OSA-Express3 10 GbE SR) is installed in drawer Z22B slot 3 and has PCHIDs 590 and 591 assigned.

► Feature code 3367 (OSA-Express3 1000BASE-T) is installed in drawer Z22B slot 4 and has PCHIDs 5A0 and 5A1 assigned. PCHID 5A0 is shared by port 00 and 01, PCHID 5A1 is shared by port 02 and 03.

The pre-assigned PCHID number of each I/O port relates directly to its physical location (jack location in a specific slot). For PCHID numbers and their locations, see Table 4-15 and Table 4-16 on page 125.

4.7 Connectivity

I/O channels are part of the channel subsystem (CSS). They provide connectivity for data exchange between servers, or between servers and external control units (CU) and devices, or networks.

Communication between servers is implemented by using InterSystem Channel-3 (ISC-3), coupling using InfiniBand (IFB), or channel-to-channel connections (CTC).

Communication to local area networks (LANs) is provided by the OSA-Express2 and OSA-Express3 features.

Connectivity to I/O subsystems to exchange data is provided by ESCON and FICON channels.

4.7.1 I/O feature support and configuration rules

Table 4-17 lists the I/O features supported. The table shows the feature code numbers, number of ports per card, port increments, and the maximum number of feature cards and the maximum of channels for each feature type. Also, the CHPID definitions used in the IOCDS are listed.

Table 4-17   Supported I/O features

I/O feature	Feature codes	Number of		Max. number of		PCHID	CHPID definition
		Ports per card	Port increments	Ports	I/O slots		
ESCON	2323[a] 2324[a]	16 (1 spare)	4 (LICCC)	240 360[b]	16 24	Yes	CNC, CVC, CTC, CBY
FICON Express4 LX/SX	3324/3321/3322	4	4	288 336[c]	72 84[c]	Yes	FC, FCP
FICON Express8 LX/SX	3325/3326	4	4	288 336[c]	72 84[c]	Yes	FC, FCP
OSA- Express2 GbE LX/SX	3364/3365	2	2	48	24[d]	Yes	OSD, OSN
OSA- Express2 1000BASE-T	3366	2	2	48	24[d]	Yes	OSE, OSD, OSC, OSN
OSA- Express3 10 GbE LR/SR	3370/3371	2	2	48	24[d]	Yes	OSD, OSX
OSA-Express3 GbE LX/SX	3362/3363	4	4	96	24[d]	Yes	OSD, OSN
OSA-Express3 1000BASE-T	3367	4	4	96	24[d]	Yes	OSE, OSD, OSC, OSN, OSM
ISC-3 2 Gbps (10 km)[e]	0217,0218,0219	2 / ISC-D	1	48	12	Yes	CFP
ISC-3 1 Gbps (20 km)[e]	RPQ 8P2197	2 / ISC-D	2	48	12	Yes	CFP
InfiniBand coupling (IFB)[e]	0163	2	2	32[f]	-	No	CIB
InfiniBand coupling (IFB LR)[e]	0168	2	2	32[f]	-	No	CIB

a. Feature code 2323 is the ESCON 16-port card; feature code 2324 is for the amount of ESCON ports ordered in increments of four. Each ESCON card has 15 usable ports and one spare port.

b. With RPQ 8P2507

c. With three I/O cages installed (RPQ 8P2506)

d. The maximum number of combined OSA features is 24.

e. For a z196, the maximum number of combined external coupling links (ISC, and IFB) is 80, and the maximum number of coupling CHPIDs is 128 (ICs, IFBs, and ISC-3: ICP, CIB, CFP).

f. A maximum of 16 combined of FC 0163 and FC 0168 is supported

At least one I/O feature (FICON or ESCON) or one coupling link feature (IFB or ISC-3) must be present in the minimum configuration. A maximum of 256 channels is configurable per channel subsystem and per operating system image.

Spanned and shared channels

The multiple image facility (MIF) allows sharing channels within a channel subsystem, as follows:

► Shared channels are shared by logical partitions within a channel subsystem (CSS).
► Spanned channels are shared by logical partitions within and across CSSs.

The following channel *cannot* be shared or spanned: ESCON-to-parallel channel conversion (defined as CVC and CBY).

The following channels can be shared but *cannot* be spanned:

► ESCON channels defined as CNC or CTC

The following channels can be shared and spanned:

► FICON channels defined as FC or FCP
► OSA-Express2 defined as OSC, OSD, OSE, or OSN
► OSA-Express3 defined as OSC, OSD, OSE, OSM, OSN or OSX
► Coupling links defined as CFP, ICP, or CIB
► HiperSockets defined as IQD

The Crypto Express3 features do not have a CHPID type, but logical partitions in all CSSs have access to the features. Each adapter on a Crypto Express3 feature can be defined to up to 32 logical partitions.

I/O feature cables and connectors

> **Note:** All fiber optic cables, cable planning, labeling, and installation are customer responsibilities for new z196 installations and upgrades. Fiber optic conversion kits and mode conditioning patch (MCP) cables are not orderable as features on z196 servers. All other cables have to be sourced separately.

IBM Facilities Cabling Services - fiber transport system offers a total cable solution service to help with cable ordering requirements, and is highly recommended. These services consider the requirements for all of the protocols and media types supported (for example, ESCON, FICON, Coupling Links, and OSA), whether the focus is the data center, the storage area network (SAN), local area network (LAN), or the end-to-end enterprise.

The Enterprise Fiber Cabling Services make use of a proven modular cabling system, the Fiber Transport System (FTS), which includes trunk cables, zone cabinets, and panels for servers, directors, and storage devices. FTS supports Fiber Quick Connect (FQC), a fiber harness integrated in the frame of a z196 for *quick* connection, which is offered as a feature on z196 servers for connection to FICON LX and ESCON channels.

Whether you choose a packaged service or a custom service, high quality components are used to facilitate moves, additions, and changes in the enterprise to prevent having to extend the maintenance window.

Table 4-18 lists the required connector and cable type for each I/O feature and the ETR feature on the z196.

Table 4-18   I/O features connector and cable types

Feature code	Feature name	Connector type	Cable type
0163	InfiniBand coupling (PSIFB)	MPO	50 µm MM[a] OM3 (2000 MHz-km)
0168	InfiniBand coupling (PSIFB LR)	LC Duplex	9 µm SM[b]
0219	ISC-3	LC Duplex	9 µm SM
2324	ESCON	MT-RJ	62.5 µm MM
3321	FICON Express4 LX 10 km	LC Duplex	9 µm SM

Feature code	Feature name	Connector type	Cable type
3322	FICON Express4 SX	LC Duplex	50, 62.5 µm MM
3324	FICON Express4 LX 4 km	LC Duplex	9 µm SM
3325	FICON Express8 LX 10 km	LC Duplex	9 µm SM
3326	FICON Express8 SX	LC Duplex	50, 62.5 µm MM
3364	OSA-Express2 GbE LX	LC Duplex	9 µm SM
3365	OSA-Express2 GbE SX	LC Duplex	50, 62.5 µm MM
3366	OSA-Express2 1000BASE-T	RJ-45	Category 5 UTP[c]
3370	OSA-Express3 10 GbE LR	LC Duplex	9 µm SM
3371	OSA-Express3 10 GbE SR	LC Duplex	50, 62.5 µm MM
3362	OSA-Express3 GbE LX	LC Duplex	9 µm SM
3363	OSA_Express3 GbE SX	LC Duplex	50, 62.5 µm MM
3367	OSA-Express3 1000BASE-T	RJ-45	Category 5 UTP[c]

a. MM is multimode fiber.
b. SM is single mode fiber.
c. UTP is unshielded twisted pair.

4.7.2 ESCON channels

ESCON channels support the ESCON architecture and directly attach to ESCON-supported I/O devices.

Sixteen-port ESCON feature

The 16-port ESCON feature (FC 2323) occupies one I/O slot in an I/O drawer or I/O cage. Each port on the feature uses a 1300 nanometer (nm) light-emitting diode (LED) transceiver, designed to be connected to 62.5 µm multimode fiber optic cables only.

The feature has 16 ports with one PCHID and one CHPID associated with each port, up to a maximum of 15 active ESCON channels per feature. Each feature has a minimum of one spare port to allow for channel-sparing in the event of a failure of one of the other ports.

The 16-port ESCON feature port utilizes a small form factor optical transceiver that supports a fiber optic connector called MT-RJ. The MT-RJ is an industry standard connector that has a much smaller profile compared to the original ESCON Duplex connector. The MT-RJ connector, combined with technology consolidation, allows for the much higher density packaging implemented with the 16-port ESCON feature.

Notes:

► The 16-port ESCON feature does *not* support a multimode fiber optic cable terminated with an ESCON Duplex connector. However, 62.5 µm multimode ESCON Duplex jumper cables *can* be reused to connect to the 16-port ESCON feature. This is done by installing an MT-RJ/ESCON conversion kit between the 16-port ESCON feature MT-RJ port and the ESCON Duplex jumper cable. This protects the investment in the existing ESCON Duplex cabling infrastructure.

► Fiber optic conversion kits and mode conditioning patch (MCP) cables are not orderable as features. Fiber optic cables, cable planning, labeling, and installation are all customer responsibilities for new installations and upgrades.

► IBM Facilities Cabling Services - fiber transport system offers a total cable solution service to help with cable ordering needs, and are highly recommended.

ESCON channel port enablement feature

The 15 active ports on each 16-port ESCON feature are activated in groups of four ports through Licensed Internal Code Control Code (LICCC) by using the ESCON channel port feature (FC 2324).

The first group of four ESCON ports requires two 16-port ESCON features. After the first pair of ESCON cards is fully allocated (by seven ESCON port groups, using 28 ports), single cards are used for additional ESCON ports groups.

Ports are activated equally across all installed 16-port ESCON features for high availability. In most cases, the number of physically installed channels is greater than the number of active channels that are LICCC-enabled. The reason is because the last ESCON port (J15) of every 16-port ESCON channel card is a spare, and because several physically installed channels are typically inactive (LICCC-protected). These inactive channel ports are available to satisfy future channel adds.

Note: The zEnterprise 196 is planned to be the last high end server to offer ordering of ESCON channels on new builds, migration offerings, upgrades, and System z exchange programs. Enterprises should begin migrating from ESCON to FICON. Alternate solutions are available for connectivity to ESCON devices.

IBM Global Technology Services (through IBM Facilities Cabling Services), offers ESCON to FICON migration services. For more information see:

http://www-935.ibm.com/services/us/index.wss/offering/its/c337386u66547p02

The PRIZM Protocol Converter Appliance from Optica Technologies Incorporated provides a FICON-to-ESCON conversion function that has been System z qualified. For more information see:

http://www.opticatech.com

Note: IBM cannot confirm the accuracy of compatibility, performance, or any other claims by vendors for products that have not been System z qualified. Questions regarding these capabilities and device support should be addressed to the suppliers of those products.

4.7.3 FICON channels

The FICON Express8 and FICON Express4 features conform to the Fibre Connection (FICON) architecture, the High Performance FICON on System z (zHPF) architecture, and the Fibre Channel Protocol (FCP) architecture, providing connectivity between any combination of servers, directors, switches, and devices (control units, disks, tapes, printers) in a Storage Area Network (SAN). FICON Express8 provides increased performance over FICON Express4.

> **Note:** FICON Express and FICON Express2 features installed in previous servers are *not* supported on a z196 and cannot be carried forward on an upgrade.

Each FICON Express8 or FICON Express4 feature occupies one I/O slot in the I/O cage or I/O drawer. Each feature has four ports, each supporting an LC Duplex connector, with one PCHID and one CHPID associated with each port.

All FICON Express8 and FICON Express4 features use small form-factor pluggable (SFP) optics that allow for concurrent repair or replacement for each SFP. The data flow on the unaffected channels on the same feature can continue. A problem with one FICON port no longer requires replacement of a complete feature.

FICON channels (CHPID type FC or FCP) can be shared among logical partitions and can be defined as spanned. All ports on a FICON feature must be of the same type, either LX or SX. The features are connected to a FICON-capable control unit, either point-to-point or switched point-to-point, through a Fibre Channel switch.

Up to 288 FICON Express8 or FICON Express4 channels (up to 72 features) can be installed in the z196 server. The Model M15 can have up to 256 FICON channels (64 features). The number is limited by the number of available IFB connections (based on HCA2-C fanouts) on the M15 model.

> **Note:** If required, using RPQ 8P2506, up to 84 features and 336 FICON channels can be installed.

FICON Express8

The FICON Express8 features are designed to support a link data rate of 8 Gbps with auto-negotiation to 2 or 4 Gbps to support existing devices, delivering increased performance compared with the FICON Express4 features. For more information about FICON channel performance see the technical papers on the System z I/O connectivity Web site at:

http://www-03.ibm.com/systems/z/hardware/connectivity/ficon_performance.html

The two types of FICON Express8 transceivers supported are the long wavelength (LX) laser version and the short wavelength (SX) LED version:

► FICON Express8 10km LX feature FC 3325, with four ports per feature, supporting LC Duplex connectors

► FICON Express8 SX feature FC 3326, with four ports per feature, supporting LC Duplex connectors

Each port of FICON Express8 10 km LX feature uses a 1300 nanometer (nm) fiber bandwidth transceiver, supports an unrepeated distance of 10 km (6.2 miles) using 9 µm single-mode fiber, Use of MCP cables limits the link speed to 1 Gbps and the unrepeated distance to 550 meters (1804 feet).

Each port of FICON Express8 SX feature uses an 850 nanometer (nm) fiber bandwidth SX transceiver. supports varying distances depending on the fiber used (50 or 62.5 μm multimode fiber) and the link speed (2 Gbps, 4 Gbps, or 8 Gbps).

> **Note:** FICON Express8 features do not support auto-negotiation to a data link rate of 1 Gbps.

FICON Express4

The three types of FICON Express4 transceivers supported (only if carried over during an upgrade) are the two long wavelength (LX) laser versions and one short wavelength (SX) LED version:

► FICON Express4 10km LX feature FC 3321, with four ports per feature, supporting LC Duplex connectors

► FICON Express4 4km LX feature FC 3324, with four ports per feature, supporting LC Duplex connectors

► FICON Express4 SX feature FC 3322, with four ports per feature, supporting LC Duplex connectors

> **Note:** It is intended that the z196 is the last server to support FICON Express4 features, We recommend that customers review the usage of their installed FICON Express4 channels and where possible migrate to FICON Express8 channels.

Both FICON Express4 LX features use 1300 nanometer (nm) fiber bandwidth transceivers. One supports an unrepeated distance of 10 km, and the other an unrepeated distance of 4 km, using 9 μm single-mode fiber. Use of MCP cables limits the link speed to 1 Gbps and the unrepeated distance to 550 meters (1804 feet).

The FICON Express4 SX feature use 850 nanometer (nm) fiber bandwidth SX transceivers. supports varying distances depending on the fiber used (50 or 62.5 μm multimode fiber) and the link speed (1 Gbps, 2 Gbps, or 4 Gbps).

> **Note:** FICON Express4 is the last FICON family able to negotiate link speed down to 1 Gbps.

FICON feature summary

Table 4-19 shows the FICON card feature codes on a z196 and their respective specifications, such as connector and cable type, maximum unrepeated distance, and the link data rate.

Table 4-19   FICON Channel specifications

Feature code	Feature name	Connector type	Cable type[a]	Unrepeated max. distance	Link data rate
3321	FICON Express4 10KM LX	LC Duplex	SM 9 μm	10 km	1, 2, or 4 Gbps
			SM 9 μm with MCP	550 m (1,804 feet)[b]	

Feature code	Feature name	Connector type	Cable type[a]	Unrepeated max. distance	Link data rate
3322	FICON Express4 SX	LC Duplex	MM 62.5 µm	55 m (180 feet)[c] at 160 MHz-km 70 m (230 feet)[c] at 200 MHz-km	1, 2, or 4 Gbps
			MM 50 µm	150 m (492 feet)[c] at 500 MHz-km 270 m (886 feet)[c] at 2,000 MHz-km	
3324	FICON Express4 4KM LX	LC Duplex	SM 9 µm	4 km	1, 2, or 4 Gbps
			SM 9 µm with MCP	550 m (1804 feet)[b]	
3325	FICON Express8 10KM LX	LC Duplex	SM 9 µm	10 km	2, 4, or 8 Gbps
3326	FICON Express8 SX	LC Duplex	MM 62.5 µm	21 m (69 feet)[d] at 200 MHz-km	2, 4, or 8 Gbps
			MM 50 µm	50 m (164 feet)[d] at 500 MHz-km 150 m (492 feet)[d] at 1500 MHz-km	

a. MM is multimode; SM is single mode
b. Maximum unrepeated distance with a mode conditioning patch (MCP) cable
c. Maximum unrepeated distance at 4 Gbps
d. Maximum unrepeated distance at 8 Gbps

4.7.4 OSA-Express3

This section discusses the connectivity options offered by the OSA-Express3 features.

The OSA-Express3 features provide improved performance by reducing latency at the TCP/IP application. Direct access to the memory allows packets to flow directly from the memory to the LAN without firmware intervention in the adapter.

The following OSA-Express3 features can be installed on z196 servers:

- ► OSA-Express3 10 Gigabit Ethernet (GbE) Long Range (LR), feature code 3370
- ► OSA-Express3 10 Gigabit Ethernet (GbE) Short Reach (SR), feature code 3371
- ► OSA-Express3 Gigabit Ethernet (GbE) Long wavelength (LX), feature code 3362
- ► OSA-Express3 Gigabit Ethernet (GbE) Short wavelength (SX), feature code 3363
- ► OSA-Express3 1000BASE-T Ethernet, feature code 3367

All OSA-Express3 Ethernet features are available on newly built servers. Up to 24 OSA-Express3 features are supported on the z196, which is a total of 48 ports when on 2-port OSA-Express3 features and up to 96 ports on 4-port OSA-Express3 features.

Note that the maximum number of OSA-Express2 and OSA-Express3 features, in combination, is 24 system-wide.

Table 4-20 lists the OSA-Express3 features.

Table 4-20   OSA-Express3 features

I/O feature	Feature code	Number of ports per feature	Port increment	Maximum number of ports (CHPIDs)	Maximum number of features	PCHID	CHPID type
OSA-Express3 10 GbE LR	3370	2	2	48	24	Yes	OSD, OSX
OSA-Express3 10 GbE SR	3371	2	2	48	24	Yes	OSD, OSX
OSA-Express3 GbE LX	3362	4	4	96 (48)	24	Yes	OSD, OSN
OSA-Express3 GbE SX	3363	4	4	96 (48)	24	Yes	OSD, OSN
OSA-Express3 1000BASE-T	3367	4	4	96 (48)	24	Yes 2 ports	OSC, OSD, OSE, OSN, OSM

OSA-Express3 data router

OSA-Express3 features help reduce latency and improve throughput by providing a data router. What was previously done in firmware (packet construction, inspection, and routing) is now performed in hardware. With the data router, there is now direct memory access. Packets flow directly from host memory to the LAN without firmware intervention. OSA-Express3 is also designed to help reduce the round-trip networking time between systems. Up to a 45% reduction in latency at the TCP/IP application later has been measured.

The OSA-Express3 features are also designed to improve throughput for standard frames (1492 byte) and jumbo frames (8992 byte) to help satisfy bandwidth requirements for applications. Up to a 4x improvement has been measured (compared to OSA-Express2).

These statements are based on OSA-Express3 performance measurements performed in a laboratory environment and do not represent actual field measurements. Results can vary.

OSA-Express3 10 GbE LR (FC 3370)

The OSA-Express3 10 GbE LR feature occupies one slot in the I/O cage or I/O drawer and has two ports that connect to a 10 Gbps Ethernet LAN through a 9 μm single mode fiber optic cable terminated with an LC Duplex connector. Each port on the card has a PCHID assigned. The feature supports an unrepeated maximum distance of 10 km.

Compared to the OSA-Express2 10 GbE LR feature, the OSA-Express3 10 GbE LR feature has double port density (two ports for each feature) and improved performance for standard and jumbo frames.

The OSA-Express3 10 GbE LR feature does not support auto-negotiation to any other speed and runs in full-duplex mode only. It supports 64B/66B encoding, whereas GbE supports 8B/10B encoding. Therefore, auto-negotiation to any other speed is not possible.

The OSA-Express3 10 GbE LR feature has two CHPIDs, with each CHPID having one port, and supports CHPID types OSD (QDIO mode) and OSX.

CHPID type OSD is supported by z/OS, z/VM, z/VSE, TPF, and Linux on System z to provide customer managed external network connections.

CHPID type OSX is dedicated for connecting the z196 to an intraensemble data network (IEDN), providing a private data exchange path across ensemble nodes.

OSA-Express3 10 GbE SR (FC 3371)

The OSA-Express3 10 GbE SR feature (FC 3371) occupies one slot in the I/O cage or I/O drawer and has two CHPIDs, with each CHPID having one port.

External connection to a 10 Gbps Ethernet LAN is done through a 62.5 μm or 50 μm multimode fiber optic cable terminated with an LC Duplex connector. The maximum supported unrepeated distance is 33 meters (108 feet) on a 62.5 μm multimode (200 MHz) fiber optic cable, 82 meters (269 feet) on a 50 μm multi mode (500 MHz) fiber optic cable, and 300 meters (984 feet) on a 50 μm multimode (2000 MHz) fiber optic cable.

The OSA-Express3 10 GbE SR feature does not support auto-negotiation to any other speed and runs in full-duplex mode only. OSA-Express3 10 GbE SR supports 64B/66B encoding, whereas GbE supports 8B/10 encoding, making auto-negotiation to any other speed impossible.

The OSA-Express3 10 GbE SR feature supports CHPID types OSD (QDIO mode) and OSX.

CHPID type OSD is supported by z/OS, z/VM, z/VSE, TPF, and Linux on System z to provide customer managed external network connections.

CHPID type OSX is dedicated for connecting the z196 to an intraensemble data network (IEDN), providing a private data exchange path across ensemble nodes.

OSA-Express3 GbE LX (FC 3362)

Feature code 3362 occupies one slot in the I/O cage or I/O drawer. It has four ports that connect to a 1 Gbps Ethernet LAN through a 9 μm single mode fiber optic cable terminated with an LC Duplex connector, supporting an unrepeated maximum distance of 5 km (3.1 miles). Multimode (62.5 or 50 μm) fiber optic cable can be used with this features.

> **Note:** The use of these multimode cable types requires a mode conditioning patch (MCP) cable at each end of the fiber optic link. Use of the single mode to multimode MCP cables reduces the supported distance of the link to a maximum of 550 meters (1084 feet).

The OSA-Express3 GbE LX feature does not support auto-negotiation to any other speed and runs in full-duplex mode only.

The OSA-Express3 GbE LX feature has two CHPIDs, with each CHPID (OSD or OSN) having two ports for a total of four ports per feature. Exploitation of all four ports requires operating system support. See 8.2, "Support by operating system" on page 210.

OSA-Express3 GbE SX (FC 3363)

Feature code 3363 occupies one slot in the I/O cage or I/O drawer. It has four ports that connect to a 1 Gbps Ethernet LAN through a 50 μm or 62.5 μm multimode fiber optic cable terminated with an LC Duplex connector over an unrepeated distance of 550 meters (for 50 μm fiber) or 220 meters (for 62.5 μm fiber).

The OSA-Express3 GbE SX feature does not support auto-negotiation to any other speed and runs in full-duplex mode only.

The OSA-Express3 GbE SX feature has two CHPIDs (OSD or OSN) with each CHPID having two ports for a total of four ports per feature. Exploitation of all four ports requires operating system support. See section 8.2, "Support by operating system" on page 210.

OSA-Express3 1000BASE-T Ethernet feature (FC 3367)

Feature code 3367 occupies one slot in the I/O cage or I/O drawer. It has four ports that connect to a 1000 Mbps (1 Gbps), 100 Mbps, or 10 Mbps Ethernet LAN. Each port has an RJ-45 receptacle for cabling to an Ethernet switch. The RJ-45 receptacle is required to be attached using EIA/TIA category 5 unshielded twisted pair (UTP) cable with a maximum length of 100 meters (328 feet).

The OSA-Express3 1000BASE-T Ethernet feature supports auto-negotiation when attached to an Ethernet router or switch. If you allow the LAN speed and duplex mode to default to auto-negotiation, the OSA-Express port and the attached router or switch auto-negotiate the LAN speed and duplex mode settings between them and connect at the highest common performance speed and duplex mode of interoperation. If the attached Ethernet router or switch does not support auto-negotiation, the OSA-Express port examines the signal it is receiving and connects at the speed and duplex mode of the device at the other end of the cable.

The OSA-Express3 1000BASE-T Ethernet feature can be configured as CHPID type OSC, OSD, OSE, OSN or OSM. Non-QDIO operation mode requires CHPID type OSE.

The following settings are supported on the OSA-Express3 1000BASE-T Ethernet feature port:

► Auto-negotiate
► 10 Mbps half-duplex or full-duplex
► 100 Mbps half-duplex or full-duplex
► 1000 Mbps full-duplex

If you are not using auto-negotiate, the OSA-Express port will attempt to join the LAN at the specified speed and duplex mode. If this does not match the speed and duplex mode of the signal on the cable, the OSA-Express port will not connect.

4.7.5 OSA-Express2

This section discusses the connectivity options offered by the OSA-Express2 features.

The following three types of OSA-Express2 features are supported only if carried over during an upgrade:

► OSA-Express2 Gigabit Ethernet (GbE) Long Wavelength (LX), feature code 3364
► OSA-Express2 Gigabit Ethernet (GbE) Short Wavelength (SX), feature code 3365
► OSA-Express2 1000BASE-T Ethernet, feature code 3366

OSA-Express and OSA-Express2 Gigabit Ethernet 10 GbE LR (FC 3368) features installed in previous servers are *not* supported on a z196 and cannot be carried forward on an upgrade.

A z196 supports up to 24 OSA-Express2 features (48 ports). The maximum number of combined OSA-Express2 and OSA-Express3 features is 24.

Table 4-21 on page 138 lists the OSA-Express2 features.

Table 4-21   OSA-Express2 features

I/O feature	Feature code	Number of		Max. number of		PCHID	CHPID type
		Ports per feature	**Port increments**	**Ports**	**I/O slots**		
OSA-Express2 GbE LX/SX	3364 3365	2	2	48	24	Yes	OSD, OSN
OSA-Express2 1000BASE-T	3366	2	2	48	24	Yes	OSE, OSD, OSC, OSN

Note: It is intended that the z196 is the last server to support OSA-Express2 features, We recommend to review the usage of installed OSA-Express2 features and where possible migrate to OSA-Express3 features.

OSA-Express2 GbE LX (FC 3364)

The OSA-Express2 Gigabit (GbE) Long Wavelength (LX) feature occupies one slot in an I/O cage or I/O drawer and has two independent ports, with one CHPID associated with each port.

Each port supports a connection to a 1 Gbps Ethernet LAN through a 9 µm single-mode fiber optic cable terminated with an LC Duplex connector. This feature uses a long wavelength laser as the optical transceiver.

A multimode (62.5 or 50 µm) fiber cable may be used with the OSA-Express2 GbE LX feature. The use of these multimode cable types requires a mode conditioning patch (MCP) cable to be used at each end of the fiber link. Use of the single-mode to multimode MCP cables reduces the supported optical distance of the link to a maximum end-to-end distance of 550 meters.

The OSA-Express2 GbE LX feature supports Queued Direct Input/Output (QDIO) and OSN modes only, full-duplex operation, jumbo frames, and checksum offload. It is defined with CHPID types OSD or OSN.

OSA-Express2 GbE SX (FC 3365)

The OSA-Express2 Gigabit (GbE) Short Wavelength (SX) feature occupies one slot in an I/O cage or I/O drawer and has two independent ports, with one CHPID associated with each port.

Each port supports a connection to a 1 Gbps Ethernet LAN through a 62.5 µm or 50 µm multimode fiber optic cable terminated with an LC Duplex connector. The feature uses a short wavelength laser as the optical transceiver.

The OSA-Express2 GbE SX feature supports Queued Direct Input/Output (QDIO) and OSN mode only, full-duplex operation, jumbo frames, and checksum offload. It is defined with CHPID types OSD or OSN.

OSA-Express2 1000BASE-T Ethernet (FC 3366)

The OSA-Express2 1000BASE-T Ethernet occupies one slot in the I/O cage or I/O drawer and has two independent ports, with one CHPID associated with each port.

Each port supports connection to either a 1000BASE-T (1000 Mbps), 100BASE-TX (100 Mbps), or 10BASE-T (10 Mbps) Ethernet LAN. The LAN must conform either to the IEEE 802.3 (ISO/IEC 8802.3) standard or to the DIX V2 specifications.

Each port has an RJ-45 receptacle for cabling to an Ethernet switch that is appropriate for the LAN speed. The RJ-45 receptacle is required to be attached using EIA/TIA category 5 unshielded twisted pair (UTP) cable with a maximum length of 100 m (328 ft).

The OSA-Express2 1000BASE-T Ethernet feature supports auto-negotiation and automatically adjusts to 10 Mbps, 100 Mbps, or 1000 Mbps, depending upon the LAN.

The OSA-Express2 1000BASE-T Ethernet feature supports CHPID types OSC, OSD, OSE, and OSN.

You may choose any of the following settings for the OSA-Express2 1000BASE-T Ethernet and OSA-Express2 1000BASE-T Ethernet features:

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

LAN speed and duplexing mode default to auto-negotiation. The feature port and the attached switch automatically negotiate these settings. If the attached switch does not support auto-negotiation, the port enters the LAN at the default speed of 1000 Mbps and full-duplex mode.

4.7.6 OSA-Express3 for ensemble connectivity

The following three types of OSA-Express3 features are used to connect the z196 central processor complex (CPC) to its attached IBM zEnterprise BladeCenter Extension (zBX), and other ensemble nodes:

- ► OSA-Express3 10 Gigabit Ethernet (GbE) Long Range (LR), feature code 3370
- ► OSA-Express3 10 Gigabit Ethernet (GbE) Short Reach (SR), feature code 3371
- ► OSA-Express3 1000BASE-T Ethernet, feature code 3367

Intraensemble data network (IEDN)
The IEDN is a private and secure 10 Gbps Ethernet network that connects all elements of an ensemble and is *access-controlled* using integrated virtual LAN (VLAN) provisioning. No customer managed switches or routers are required. The IEDN is managed by a primary HMC[5]

IEDN requires two OSA-Express3 10 GbE ports (one port from two OSA-Express3 10 GbE features), configured as CHPID type OSX. The connection is from the z196 to the IEDN top of rack (TOR) switches on zBX. Or with a stand-alone z196 node the OSA Express-3 10 GbE ports are connected to one another in the z196 with client provided 10 GbE loop back cables (either SR or LR, depending on the OSA feature).

Intranode management network (INMN)
The INMN is a private and physically isolated 1000BASE-T Ethernet internal management network, operating at 1 Gbps. It connect all resources (z196 and zBX components) of an ensemble node for management purposes. It is prewired, internally switched, configured, and managed with full redundancy for high availability.

---

[5] This HMC must be running with Version 2.11 or above with feature codes 0090, 0025, 0019 and optionally 0020.

The INMN requires two ports (one port from two OSA-Express3 1000BASE-T features), configured as CHPID type OSM. The connection is via port J07 of the bulk power hubs (BPHs) in the z196. The INMN top of rack (TOR) switches on zBX also connect to the BPHs.

For detailed information about OSA-Express3 in an ensemble network, see "zBX connectivity" on page 190.

4.7.7 HiperSockets

The HiperSockets function of z196 is improved to provide up to 32 high-speed virtual LAN attachments. Previous servers provide 16 attachments.

HiperSockets can be customized to accommodate varying traffic sizes. Because HiperSockets does not use an external network, it can free up system and network resources, which can help eliminate attachment costs, and improve availability and performance.

HiperSockets eliminates having to use I/O subsystem operations and having to traverse an external network connection to communicate between logical partitions in the same z196 server. HiperSockets offers significant value in server consolidation connecting many virtual servers, and can be used instead of certain coupling link configurations in a Parallel Sysplex.

HiperSockets internal networks on z196 servers support two transport modes:

- ▶ Layer 2 (link layer)
- ▶ Layer 3 (network or IP layer)

Traffic can be IPv4 or IPv6, or non-IP such as AppleTalk, DECnet, IPX, NetBIOS, or SNA.

HiperSockets devices are protocol and Layer 3-independent. Each HiperSockets device (Layer 2 and Layer 3 mode) has its own MAC address designed to allow the use of applications that depend on the existence of Layer 2 addresses, such as DHCP servers and firewalls. Layer 2 support helps facilitate server consolidation, can reduce complexity, can simplify network configuration, and allows LAN administrators to maintain the mainframe network environment similarly as for non-mainframe environments.

Packet forwarding decisions are based on Layer 2 information instead of Layer 3. The HiperSockets device can perform automatic MAC address generation to create uniqueness within and across logical partitions and servers. The use of Group MAC addresses for multicast is supported as well as broadcasts to all other Layer 2 devices on the same HiperSockets networks.

Datagrams are delivered only between HiperSockets devices that use the same transport mode. A Layer 2 device cannot communicate directly to a Layer 3 device in another logical partition network. A HiperSockets device can filter inbound datagrams by VLAN identification, the destination MAC address, or both.

Analogous to the Layer 3 functions, HiperSockets Layer 2 devices can be configured as primary or secondary connectors or multicast routers. This enables the creation of high-performance and high-availability link layer switches between the internal HiperSockets network and an external Ethernet or to connect to the HiperSockets Layer 2 networks of different servers.

HiperSockets Layer 2 on z196 is supported by Linux on System z, and by z/VM for Linux guest exploitation.

4.8 Parallel Sysplex connectivity

Coupling links are required in a Parallel Sysplex configuration to provide connectivity from the z/OS images to the coupling facility. A properly configured Parallel Sysplex provides a highly reliable, redundant, and robust System z technology solution to achieve near-continuous availability. A Parallel Sysplex comprises one or more z/OS operating system images coupled through one or more coupling facilities.

The type of coupling link that is used to connect a coupling facility (CF) to an operating system logical partition is important because of the effect of the link performance on response times and coupling overheads. For configurations covering large distances, the time spent on the link can be the largest part of the response time.

The types of links that are available to connect an operating system logical partition to a coupling facility are:

► ISC-3

The InterSystem Channel-3 (ISC-3) type is available in peer mode only. ISC-3 links can be used to connect to z196, z10 or z9 servers. They are optic fiber links that support a maximum distance of 10 km, 20 km with RPQ 8P2197, and 100 km with a System z qualified dense wave division multiplexer (DWDM). ISC-3s support 9 um single mode fiber optic cabling. The link data rate is 2 Gbps at distances up to 10 km, and 1 Gbps when RPQ 8P2197 is installed. Each port operates at 2 Gbps. Ports are ordered in increments of one. The maximum number of ISC-3 links per z196 is 48. ISC-3 supports transmission of Server Time Protocol (STP) messages.

> **Note:** It is intended that the z196 is the last server for ordering of ISC-3 coupling links. Customers should review the usage of their installed ISC-3 coupling links and where possible migrate to PSIFB (FC 0163) or PSIFB LR (FC 0168) coupling links.

► PSIFB

Parallel Sysplex using Infinband (PSIFB) connects a z196 to a z196, z10, z9 EC or z9 BC. 12x InfiniBand coupling links are fiber optic connections that support a maximum distance of up to 150 meters. PSIFB coupling links are defined as CHPID type CIB. The maximum number of PSIFB links is 32 for each z196 (12x and 1x InfiniBand combined). PSIFB supports transmission of STP messages.

► PSIFB LR

PSIFB LR (Long Reach) connects a z196 to another z196 or z10 server. 1x InfiniBand coupling links are fiber optic connections that support a maximum unrepeated distance of up to 10 km and up to 100 km with a System z qualified dense wave division multiplexor (DWDM). PSIFB LR coupling links are defined as CHPID type CIB. The maximum number of PSIFB LR links is 32 for each z196 server (12x and 1x InfiniBand combined). PSIFB LR supports transmission of STP messages.

► IC

CHPIDs (type ICP) defined for internal coupling can connect a CF to a z/OS logical partition in the same z196. IC connections require two CHPIDs to be defined, which can only be defined in peer mode. The bandwidth is greater than 2 GBps. A maximum of 32 IC CHPIDs (16 connections) can be defined.

Table 4-22 shows the coupling link options.

Table 4-22   Coupling link options

Type	Description	Use	Link rate	Distance	z196 maximum
ISC-3	InterSystem Channel-3	z196 to z196, z10, z9	2 Gbps	10 km unrepeated (6.2 miles) 100 km repeated	48
PSIFB	12x IB-DDR InfiniBand	z196 to z196, z10	6 GBps	150 meters (492 feet)	32
	12x IB-SDR InfiniBand	z196 to z9	3 GBps[a]		
PSIFB LR	1x IB-SDR[b] InfiniBand	z196 to z196, z10	2.5 Gbps 5.0 Gbps	10 km unrepeated (6.2.miles) 100 km repeated	32
IC	Internal coupling channel	Internal communication	Internal speeds	N/A	32

a. When connected to a System z9 EC or System z9 BC.
b. Double data rate (1x IB-DDR) is supported if connected to a System z qualified DWDM supporting DDR.

The maximum PSIFB links is 32. The maximum number of external coupling links combined cannot exceed 80 per server (active ISC-3 links, PSIFB, PSIFB LR). There is a maximum of 128 coupling CHPIDs limitation, including ICs - ICP, PSIFB/PSIFB LR - CIB, ISC-3 - CFP per server.

The z196 supports several connectivity options depending on the connected z10 or z9 server. Figure 4-7 shows z196 coupling link support for z10 and z9 servers.
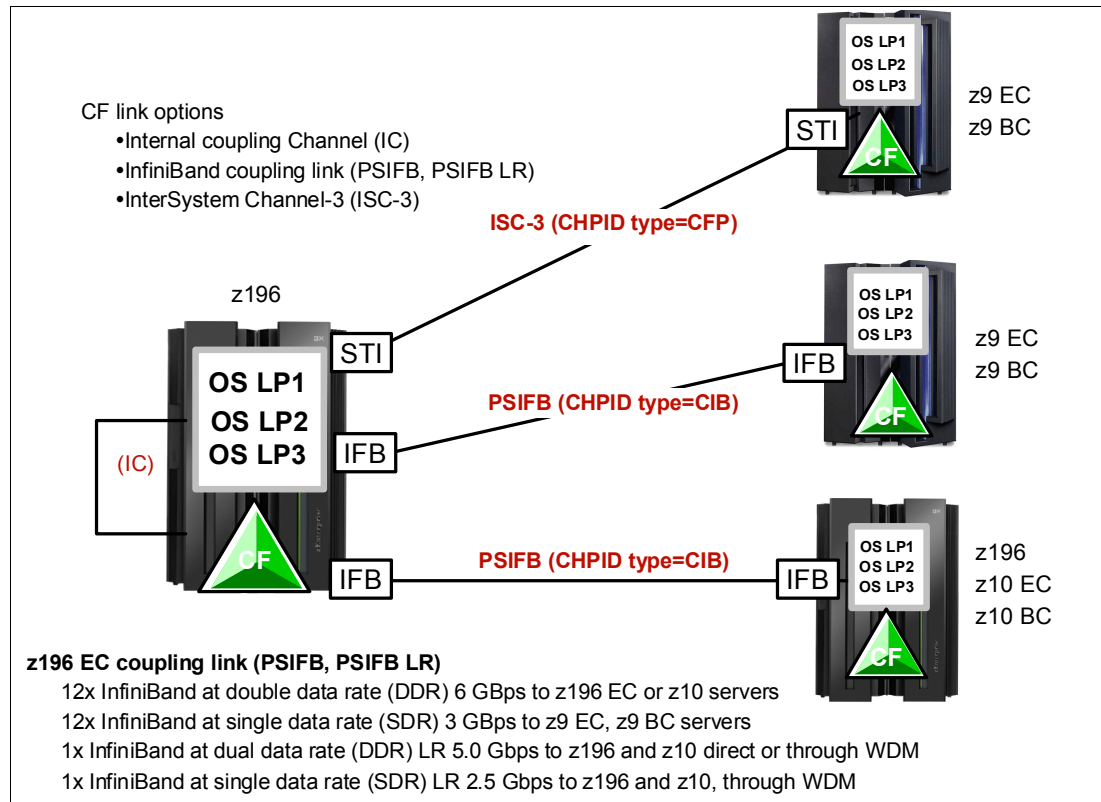
Figure 4-7   z196 to System z CF connectivity options

z/OS and coupling facility images may be running on the same or on separate servers. There must be at least one CF connected to all z/OS images, although there can be other CFs that are connected only to selected z/OS images. Two coupling facility images are required for system-managed CF structure duplexing and, in this case, each z/OS image must be connected to both duplexed CFs.

To eliminate any single-points of failure in a Parallel Sysplex configuration, there should be at least:

▶ Two coupling links between the z/OS and coupling facility images

▶ Two coupling facility images not running on the same server

▶ One stand-alone coupling facility. If you are using system-managed CF structure duplexing or running with *resource sharing* only, then a stand-alone coupling facility is not mandatory.

Coupling link features

The z196 supports three types of coupling link options:

▶ InterSystem Channel-3 (ISC-3) FC 0217, FC 0218, and FC 0219
▶ Parallel Sysplex using InfiniBand (PSIFB) coupling link, FC 0163
▶ Parallel Sysplex using InfiniBand Long Reach (PSIFB LR) coupling link, FC 0168

The coupling link features available on the z196 connect z196 servers to the identified System z servers by various link options:

- ► ISC-3 at 2 Gbps to z196, z10 and z9
- ► PSIFB at 6 GBps to z196 and z10, or 3 GBps to z9 EC and z9 BC
- ► PSIFB LR at 5.0 or 2.5 Gbps to z196 and z10 servers

ISC-3 coupling links

Three feature codes are available to implement ISC-3 coupling links:

- ► FC 0217, ISC-3 mother card
- ► FC 0218, ISC-3 daughter card
- ► FC 0219, ISC-3 port

The ISC mother card (FC 0217) occupies one slot in the I/O cage or I/O drawer and supports up to two daughter cards. The ISC daughter card (FC 0218) provides two independent ports with one CHPID associated with each enabled port. The ISC-3 ports are enabled and activated individually (one port at a time) by Licensed Internal Code.

When the quantity of ISC links (FC 0219) is selected, the quantity of ISC-3 port features selected determines the appropriate number of ISC-3 mother and daughter cards to be included in the configuration, up to a maximum of 12 ISC-M cards. Additional ISC-M cards can be ordered, up to the number of ISC-D features or twelve, whichever is smaller.

Each active ISC-3 port in peer mode supports a 2 Gbps (200 MBps) connection through 9 μm single mode fiber optic cables terminated with an LC Duplex connector. The maximum unrepeated distance for an ISC-3 link is 10 km. With repeaters the maximum distance extends to 100 km. ISC-3 links can be defined as *timing-only links* when STP is enabled. Timing-only links are coupling links that allow two servers to be synchronized using STP messages when a CF does not exist at either end of the link.

RPQ 8P2197 extended distance option

The RPQ 8P2197 daughter card provides two ports that are active and enabled when installed and do not require activation by LIC.

This RPQ allows the ISC-3 link to operate at 1 Gbps (100 MBps) instead of 2 Gbps (200 MBps). This lower speed allows an extended unrepeated distance of 20 km. One RPQ daughter is required on both ends of the link to establish connectivity to other servers. This RPQ supports STP if defined as either a coupling link or timing-only.

InfiniBand coupling links (FC 0163)

The Parallel Sysplex using InfiniBand (PSIFB) coupling option uses InfiniBand over an optical interface provided by the HCA2-O fanout (FC 0163). Each fanout has two ports. Both ports on the HCA2-O fanout are exclusively used for coupling links and *cannot* be shared for other functions. Up to 16 CHPIDs are supported for each HCA2-O fanout. However, we recommend no more than 4 CHPIDs per port or no more than 8 CHPIDs across the two ports.

The maximum distance for PSIFB coupling links is 150 meters. The maximum number of features is 16, supporting 32 optical links. The maximum number of links to a System z9 is 16.

The InfiniBand coupling link is defined as channel type CIB in the IOCDS. The coupling links can be defined as shared between images within a channel subsystem and they can be also be spanned across multiple CSSs in a server.

Each HCA2-O fanout for optical links can be used for link definitions to another server or a link from one port to a port in another fanout on the same server.

When connected to an external server the source and the target operating system or CF image must be defined in the IOCDS.

Figure 4-8 shows an optical link connection between two servers.
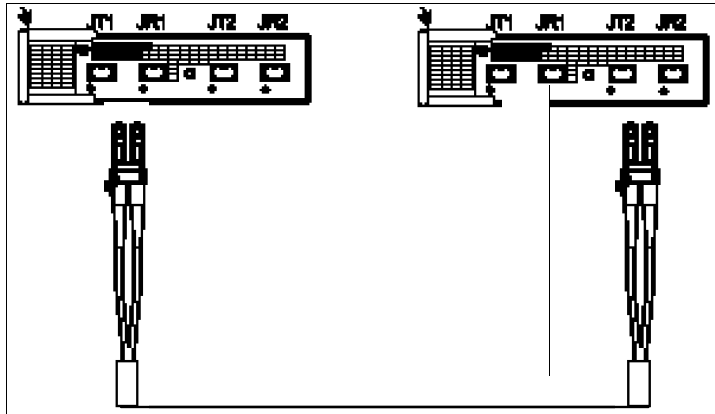
Figure 4-8   Infiniband link - need some fixing

The link rate for coupling links to System z9 servers is auto-negotiated to the highest maximum rate, which is 3 GBps.

When STP is enabled, PSIFB coupling links can be defined as timing-only links to other z196, z10 and z9 servers.

InfiniBand coupling links LR (FC 0168)

The Parallel Sysplex using InfiniBand Long Reach (PSIFB LR) coupling option uses InfiniBand over an optical interface provided by the HCA2-O LR fanout (FC 0168). Each fanout provides two ports. Both ports on the HCA2-O LR fanout are exclusively used for coupling links and *cannot* be shared for other functions. Up to 16 CHPIDs are supported per HCA2-O LR fanout. However, we recommend no more than 4 CHPIDs per port or no more than 8 CHPIDs across the two ports.

The maximum unrepeated distance for PSIFB LR coupling links is 10 km, and up to 100 km if using System z qualified DWDMs.

The maximum number of FC 0168s on a z196 is 16, supporting 32 optical links. The PSIFB coupling links are intended to replace the existing ISC-3 coupling links. It uses the same fiber optic cabling as is used for ISC-3 coupling links, which is a 9 µm single mode (SM) cable terminated with an LC Duplex connector.

The InfiniBand coupling link is defined as channel type CIB in the IOCDS. The coupling links can be defined as shared between images within a channel subsystem and they can be also be spanned across multiple CSSs in a server.

Each HCA2-O LR fanout can be used for link definitions to another server or a link from one port to a port in another fanout on the same server.

Definitions of the source and target operating system image, CF image, and the CHPIDs used on both ports in both servers, are defined in IOCDS.

When STP is enabled, PSIFB LR coupling links can be defined as timing-only links to other z196 and z10 servers.

The PSIFB LR feature is exclusive to z196 and z10 servers. PSIFB LR coupling link connectivity to other servers is not supported.

Internal coupling links

IC links are Licensed Internal Code-defined links to connect a CF to a z/OS logical partition in the same server. These links are available on all System z servers. The IC link is a System z server coupling connectivity option that enables high-speed, efficient communication between a CF partition and one or more z/OS logical partitions running on the same server. The IC is a linkless connection (implemented in Licensed Internal Code) and so does not require any hardware or cabling.

An IC link is a fast coupling link, using memory-to-memory data transfers. IC links do not have PCHID numbers, but do require CHPIDs.

IC links require an ICP channel path definition at the z/OS and the CF end of a channel connection to operate in peer mode. They are always defined and connected in pairs. The IC link operates in peer mode and its existence is defined in HCD/IOCP.

IC links have the following attributes:

► On System z servers, IC links operate in peer mode (channel type ICP).

► Provide the fastest connectivity, significantly faster than any external link alternatives.

► Result in better coupling efficiency than with external links, effectively reducing the server cost associated with Parallel Sysplex technology.

► Can be used in test or production configurations, and reduce the cost of moving into Parallel Sysplex technology while enhancing performance and reliability.

► Can be defined as spanned channels across multiple CSSs.

► Are free of charge (no feature code). Employing ICFs with IC channels will result in considerable cost savings when configuring a cluster.

IC links are enabled by defining channel type ICP. A maximum of 32 IC channels can be defined on a System z server.

Coupling link migration considerations

The following restrictions apply to customers that are running a Parallel Sysplex including z990 and z890 servers, and are installing a z196 server:

► The z196 cannot be added to the Parallel Sysplex.

► Rolling IPLs cannot be performed to introduce the z196.

► If the sysplex also includes any z9 EC, z9 BC or z10 that is being upgraded, then the z990 and z890 in the sysplex must either be upgraded or removed from the sysplex.

► When the z990 or z890 is being used as a coupling facility, the coupling facility *must* be moved to a z9 EC or z9 BC, or later, *before* introducing a z196 for a z/OS image or ICF.

Note: The InfiniBand link data rates of 6 GBps, 3 GBps, 2.5 Gbps, or 5 Gbps do not represent the performance of the link. The actual performance depends on many factors including latency through the adapters, cable lengths, and the type of workload.

For a more specific explanation of when to continue using the current ISC-3 technology versus migrating to InfiniBand coupling links, see the *Coupling Facility Configuration Options* white paper, available at:

  http://www.ibm.com/systems/z/advantages/pso/whitepaper.html

Coupling links and Server Time Protocol

All external coupling links can be used to pass time synchronization signals by using Server Time Protocol (STP). Server Time Protocol is a message-based protocol in which STP messages are passed over data links between servers. The same coupling links can be used to exchange time and coupling facility messages in a Parallel Sysplex.

Using the coupling links to exchange STP messages has the following advantages:

► By using the same links to exchange STP messages and coupling facility messages in a Parallel Sysplex, STP can scale with distance. Servers exchanging messages over short distances, such as PSIFB links, can meet more stringent synchronization requirements than servers exchanging messages over long ISC-3 links (distances up to 100 km). This advantage is an enhancement over the IBM Sysplex Timer implementation, which does not scale with distance.

► Coupling links also provide the connectivity necessary in a Parallel Sysplex. Therefore, there is a potential benefit of minimizing the number of cross-site links required in a multi-site Parallel Sysplex.

Between any two servers that are intended to exchange STP messages, we recommend that each server be configured so that at least two coupling links exist for communication between the servers. This configuration prevents the loss of one link, causing the loss of STP communication between the servers. If a server does not have a CF logical partition, timing-only links can be used to provide STP connectivity.

The z196 no longer supports attachment to the IBM Sysplex Timer. A z196 can be added into a Mixed CTN only when there is a z10 or z9 attached to the Sysplex Timer operating as Stratum 1 server. Connection to two Stratum 1 servers are recommended to provide redundancy and avoid a single point of failure.

> **Note:** A Parallel Sysplex in a ETR network *must* migrate to Mixed CTN or STP-only CTN *before* introducing a z196.

For Sysplex Timer connectivity and configuration information see *IBM System z Connectivity Handbook*, SG24-5444.

For STP configuration information, see *Server Time Protocol Planning Guide*, SG24-7280, and *Server Time Protocol Implementation Guide*, SG24-7281.

4.8.1 External clock facility

The external clock facility (ECF) card located in CPC cage provides a Pulse Per Second (PPS) connection to external time sources (ETS). Two ECF cards are installed in the card slots above the books, to provide redundancy for continued operation and concurrent maintenance when a single ECF card fails. Each ECF card has a BNC connector for PPS connection support, attaching to two different ETSs.

The time accuracy of an STP-only CTN is improved by adding an NTP server with the pulse per second output signal (PPS) as the ETS device. STP tracks the highly stable accurate PPS signal from the NTP server and maintains and accuracy of 10 µs as measured at the PPS input of the System z server. If STP uses a dial-out time service or an NTP server without PPS a time accuracy of 100 ms to the ETS is maintained. NTP servers with PPS output are available from several vendors that offer network timing solutions.

PPS connection from a NTP server with PPS output to the ECF card to is required when the z196 is configured in an STP-only CTN using NTP with pulse per second as the external time

source. Two PPS connections from two different NTP servers are recommended for redundancy.

4.8.2 Cryptographic feature

Cryptographic functions are provided by CP Assist for Cryptographic Function (CPACF) and the Crypto Express3 feature. Feature code (FC) 3863 is required to enable CPACF functions.

Crypto Express3 feature (FC 0864)

Crypto Express3 is an optional feature. On the initial order, the minimum of two features are installed. After the initial configuration, the number of features increase one at a time up to a maximum of eight.

Each Crypto Express3 feature holds two PCI Express cryptographic adapters that can be configured as coprocessors or accelerators. Either of the adapters can be configured by the installation as a coprocessor or accelerator.

Each Crypto Express3 feature occupies one I/O slot in an I/O cage and has no CHPIDs assigned, but uses two PCHIDS.

Cryptographic functions are described in Chapter 6, "Cryptography" on page 163.

5

CPC channel subsystem

This chapter describes the concepts of the System z channel subsystem, including multiple channel subsystems. It also discusses the technology, terminology, and implementation aspects of the channel subsystem.

This chapter discusses the following topics:

5.1 Channel subsystem

The role of the channel subsystem (CSS) is to control communication of internal and external channels to control units and devices. The CSS configuration defines the operating environment for the correct execution of all system I/O operations.

The CSS provides the server communications to external devices through channel connections. The channels execute transfer of data between main storage and I/O devices or other servers under the control of a channel program. The CSS allows channel I/O operations to continue independently of other operations within the central processors (CPs) and IFLs.

The building blocks that make up a channel subsystem are shown in Figure 5-1.
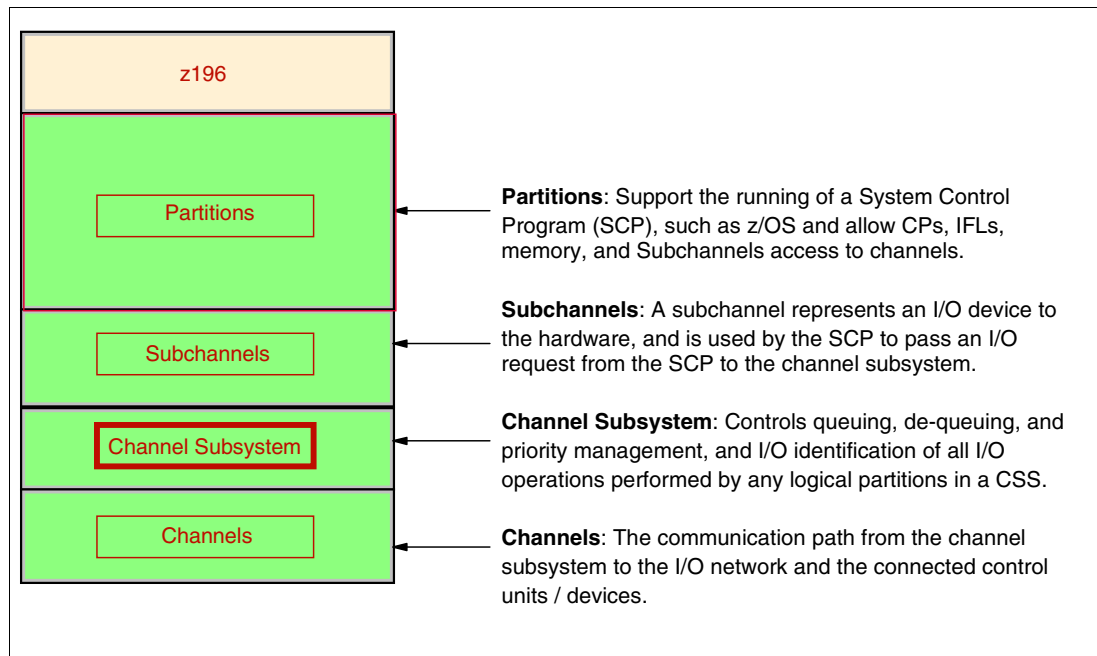
Figure 5-1   Channel subsystem overview

5.1.1 Multiple CSSs concept

The design of System z servers offers considerable processing power, memory sizes, and I/O connectivity. In support of the larger I/O capability, the CSS concept has been scaled up correspondingly to provide relief for the number of supported logical partitions, channels, and devices available to the server.

A single channel subsystem allows the definition of up to 256 channel paths. To overcome this limit the multiple channel subsystems concept was introduced. The architecture provides up to four channel subsystems. The structure of the multiple CSSs provides channel connectivity to the defined logical partitions in a manner that is transparent to subsystems and application programs, enabling the definition of a balanced configuration for the processor and I/O capabilities.

Each CSS may have from 1 to 256 channels and be configured with 1 to 15 logical partitions. Therefore, four CSSs support a maximum of 60 logical partitions. CSSs are numbered from 0 to 3 and are sometimes referred to as the CSS image ID (CSSID 0, 1, 2 or 3).

5.1.2 CSS elements

The elements that encompass the CSS are described in this section.

Subchannels

A subchannel provides the logical representation of a device to a program and contains the information required for sustaining a single I/O operation. A subchannel is assigned for each device defined to the logical partition.

Multiple subchannel sets, described in 5.1.3, "Multiple subchannel sets" on page 151 are available to increase addressability. Three subchannel sets per CSS are supported on z196. Subchannel set 0 can have up to 63.75 K subchannels, and subchannel sets 1 and 2 can have up to 64 K subchannels each.

Channel paths

Each CSS can have up to 256 channel paths. A channel path is a single interface between a server and one or more control units. Commands and data are sent across a channel path to perform I/O requests.

Channel path identifier

Each channel path in the system is assigned a unique identifier value known as a channel path identifier (CHPID). A total of 256 CHPIDs are supported by the CSS, and a maximum of 1024 are supported per system (CPC).

The channel subsystem communicates with I/O devices by means of channel paths between the channel subsystem and control units. On System z, a CHPID number is assigned to a physical location (slot/port) by the customer, through the hardware configuration definition (HCD) tool or IOCP.

Control units

A control unit provides the logical capabilities necessary to operate and control an I/O device and adapts the characteristics of each device so that it can respond to the standard form of control provided by the CSS. A control unit may be housed separately, or it may be physically and logically integrated with the I/O device, the channel subsystem, or within the server itself.

I/O devices

An I/O device provides external storage, a means of communication between data-processing systems, or a means of communication between a system and its environment. In the simplest case, an I/O device is attached to one control unit and is accessible through one channel path.

5.1.3 Multiple subchannel sets

Do not confuse the multiple subchannel set (MSS) functionality with multiple channel subsystems.

In most cases, a subchannel represents an addressable device. For example, a disk control unit with 30 drives uses 30 subchannels (for base addresses), and so forth. An addressable device is associated with a device number and the device number is commonly (but incorrectly) known as the device address.

Subchannel numbers (including their implied path information to a device) are limited to four hexadecimal digits by the architecture (0x0000 to 0xFFFF). Four hexadecimal digits provide

64 K addresses, known as a *set*. IBM has reserved 256 subchannels, leaving over 63 K subchannels for general use[1].

Again, addresses, device numbers, and subchannels are often used as synonyms, although this is not technically correct. We may hear that there is *a maximum of 63.75 K addresses* or *a maximum of 63.75 K device numbers*.

The processor architecture allows for *sets* of subchannels (addresses), with a current implementation of three sets. Each set provides 64 K addresses. Subchannel set 0, the first set, still reserves 256 subchannels for IBM use. Each of subchannel sets 1 and 2 provides the full range of 64 K subchannels. In principle, subchannels in either set could be used for any device-addressing purpose. Base addresses must be in subchannel set 0. Parallel Access Volumes alias devices, PPRC secondaries, and FlashCopy devices may be in subchannel sets 0, 1 or 2. These are referred to as special devices below.

Figure 5-2 summarizes the multiple channel subsystems and multiple subchannel sets.
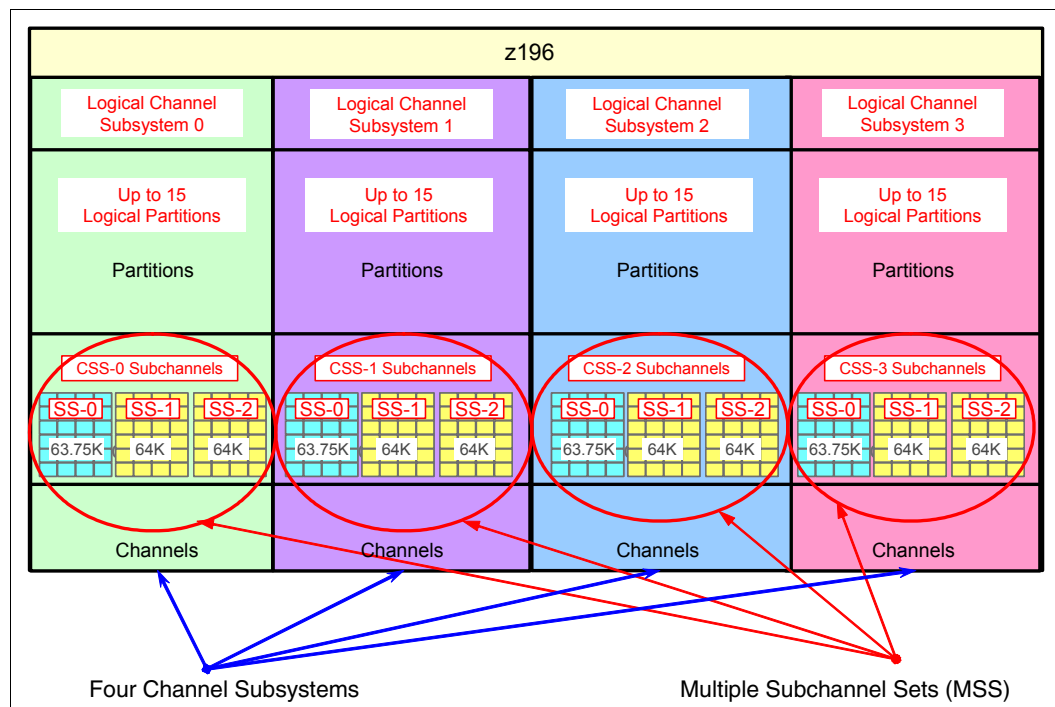
Figure 5-2   Multiple channel subsystems and multiple subchannel sets

The additional subchannel sets, in effect, add an extra high-order digit (either 0, 1 or 2) to existing device numbers. For example, we might think of an address as 08000 (subchannel set 0), or 18000 (subchannel set 1) or 28000 (subchannel set 2). Adding a digit is not done in system code or in messages because of the architectural requirement for four-digit addresses (device numbers or subchannels). However, some messages do contain the subchannel set number, and you can mentally use that as a high-order digit for device numbers. There should be few requirements to refer to the subchannel sets 1 and 2, since they are only used for these special devices. JCL, messages, and programs rarely refer directly these special devices.

---

[1] The number of reserved subchannels is 256. We abbreviate this to 63.75 K in this discussion to easily differentiate it from the 64 K subchannels available in subchannel sets 1 and 2. The informal name, 63.75 K subchannel, represents the following equation: (63 x 1024) + (0.75 x 1024) = 65280

Moving these special devices into an alternate subchannel set creates additional space for device number growth. The appropriate subchannel set number must be included in IOCP definitions or in the HCD definitions that produce the IOCDS. The subchannel set number defaults to zero.

The display ios,config command

The **display ios,config(all)** command, shown in Figure 5-3, includes information about the MSSs.

```
D IOS,CONFIG(ALL)
IOS506I 18.21.37 I/O CONFIG DATA 610
ACTIVE IODF DATA SET = SYS6.IODF45
CONFIGURATION ID = TEST2097 EDT ID = 01
TOKEN:  PROCESSOR DATE     TIME     DESCRIPTION
 SOURCE: SCZP201  10-03-04 09:20:58 SYS6     IODF45
ACTIVE CSS:  0    SUBCHANNEL SETS CONFIGURED: 0, 1, 2
CHANNEL MEASUREMENT BLOCK FACILITY IS ACTIVE
HARDWARE SYSTEM AREA AVAILABLE FOR CONFIGURATION CHANGES
PHYSICAL CONTROL UNITS              8131
CSS  0 - LOGICAL CONTROL UNITS     4037
 SS  0   SUBCHANNELS              62790
 SS  1   SUBCHANNELS              61117
 SS  2   SUBCHANNELS              60244
CSS  1 - LOGICAL CONTROL UNITS     4033
 SS  0   SUBCHANNELS              62774
 SS  1   SUBCHANNELS              61117
 SS  2   SUBCHANNELS              60244
CSS  2 - LOGICAL CONTROL UNITS     4088
 SS  0   SUBCHANNELS              65280
 SS  1   SUBCHANNELS              65535
 SS  2   SUBCHANNELS              62422
CSS  3 - LOGICAL CONTROL UNITS     4088
 SS  0   SUBCHANNELS              65280
 SS  1   SUBCHANNELS              65535
 SS  2   SUBCHANNELS              62422
ELIGIBLE DEVICE TABLE LATCH COUNTS
        0 OUTSTANDING BINDS ON PRIMARY EDT
```

Figure 5-3   Display ios,config(all) with MSS

5.1.4 Parallel access volumes and extended address volumes

Parallel access volume (PAV) support enables a single System z server to simultaneously process multiple I/O operations to the same logical volume, which can help to significantly reduce device queue delays. Dynamic PAV allows the dynamic assignment of aliases to volumes to be under WLM control.

With the availability of HyperPAV, the requirement for PAV devices is greatly reduced. HyperPAV allows an alias address to be used to access any base on the same control unit image per I/O base. It also allows different HyperPAV hosts to use one alias to access different bases, which reduces the number of alias addresses required. HyperPAV is designed to enable applications to achieve equal or better performance than possible with the original PAV feature alone, while also using the same or fewer z/OS resources. HyperPAV is an optional feature on the IBM DS8000® series.

To further reduce the complexity of managing large I/O configurations System z introduces Extended Address Volumes (EAV). EAV is designed to build very large disk volumes using

virtualization technology. By being able to extend the disk volume size a customer may potentially need fewer volumes to hold the data, therefore making systems management and data management less complex.

5.1.5 Logical partition name and identification

No logical partitions can exist without at least one defined CSS. Logical partitions are defined to a CSS, not to a server. A logical partition is associated with one CSS only.

A logical partition is identified through its name, its identifier, and its multiple image facility (MIF) image ID (MIF ID). The logical partition name is user defined through HCD or the IOCP and is the partition name in the RESOURCE statement in the configuration definitions. Each name must be unique across the CPC.

The logical partition identifier is a number in the range of 00 - 3F assigned by the user on the image profile through the support element (SE) or the hardware management console (HMC). It is unique across the CPC and may also be referred to as the user logical partition ID (UPID).

The MIF ID is a number that is defined through the HCD tool or directly through the IOCP. It is specified in the RESOURCE statement in the configuration definitions. It is in the range of 1 - F and is unique within a CSS. However, because of the multiple CSSs, the MIF ID is not unique within the CPC.

The multiple image facility enables resource sharing across logical partitions within a single CSS or across the multiple CSSs. When a channel resource is shared across logical partitions in multiple CSSs, this is known as *spanning*. Multiple CSSs may specify the same MIF image ID. However, the combination CSSID.MIFID is unique across the CPC.

Dynamic addition or deletion of a logical partition name

All undefined logical partitions are reserved partitions. They are automatically predefined in the HSA with a name placeholder and a MIF ID.

Summary of identifiers

We recommend establishing a naming convention for the logical partition identifiers. As shown in Figure 5-4, which summarizes the identifiers and how they are defined, you could use the CSS number concatenated to the MIF ID, which means that logical partition ID 3A is in CSS 3 with MIF ID A. This fits within the allowed range of logical partition IDs and conveys useful information to the user.

CSS0	CSS1	CSS2	CSS3	Specified in HCD / IOCP

Logical Partition Name			Logical Partition Name			Log Part Name	Logical Partition Name		Specified in HCD / IOCP
TST1	PROD1	PROD2	TST2	PROD3	PROD4	TST3	TST4	PROD5	

Logical Partition ID			Logical Partition ID			Log Part ID	Logical Partition ID		Specified in HMC Image Profile
02	04	0A	14	16	1D	22	35	3A	

MIF ID 2	MIF ID 4	MIF ID A	MIF ID 4	MIF ID 6	MIF ID D	MIF ID 2	MIF ID 5	MIF ID A	Specified in HCD / IOCP

Figure 5-4   CSS, logical partition, and identifiers example

5.1.6 Physical channel ID

A physical channel ID (PCHID) reflects the physical identifier of a channel-type interface. A PCHID number is based on the I/O drawer or I/O cage location, the channel feature slot number, and the port number of the channel feature. A hardware channel is identified by a PCHID. In other words, the physical channel, which uniquely identifies a connector jack on a channel feature, is known by its PCHID number.

PCHIDs identify the physical ports of the features located in the I/O drawers and I/O cages. The PCHID numbering scheme for the I/O drawers is shown in Table 5-1.

Table 5-1   PCHID numbering scheme for I/O drawers

Slot	PCHID Range					
	Drawer 1 Z22B	Drawer 2 Z15B	Drawer 3 Z08B	Drawer 4 Z01B	Drawer 5 A08B(MRU) A16B(MWU)	Drawer 6 A01B(MRU) A09B(MWU)
2	580-58F	500-50F	380-38F	300-30F	180-18F	100-10F
3	590-59F	510-51F	390-39F	310-31F	190-19F	110-11F
4	5A0-5AF	520-52F	3A0-3AF	320-32F	1A0-1AF	120-12F
5	5B0-5BF	530-53F	3B0-3BF	330-33F	1B0-1BF	130-13F
7	5C0-5CF	540-54F	3C0-3CF	340-34F	1C0-1CF	140-14F
8	5D0-5DF	550-55F	3D0-3DF	350-35F	1D0-1DF	150-15F
10	5E0-5EF	560-56F	3E0-3EF	360-36F	1E0-1EF	160-16F
11	5F0-5FF	570-57F	3F0-3FF	370-37F	1F0-1FF	170-17F

The PCHID numbering scheme for the I/O cages is shown in Table 5-2 on page 156.

Table 5-2   PCHID numbering scheme for I/O cages

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

Do not confuse PCHIDs with CHPIDs. A CHPID does not directly correspond to a hardware channel port, and may be arbitrarily assigned. Within a single channel subsystem, 256 CHPIDs can be addressed. That gives a maximum of 1,024 CHPIDs when four CSSs are defined. Each CHPID number is associated with a single channel.

CHPIDs are not pre-assigned. The installation is responsible to assign the CHPID numbers through the use of the CHPID mapping tool (CMT) or HCD/IOCP. Assigning CHPIDs means that a CHPID number is associated with a physical channel/port location and a CSS. The CHPID number range is still from 00 - FF and must be unique within a CSS. Any non-internal CHPID that is not defined with a PCHID can fail validation when an attempt is made to build a production IODF or an IOCDS.

5.1.7 Channel spanning

Channel spanning extends the MIF concept of sharing channels across logical partitions to sharing channels across logical partitions *and* channel subsystems.

Spanning is the ability for a physical channel (PCHID) to be mapped to CHPIDs defined in multiple channel subsystems. When defined that way, the channels can be transparently shared by any or all of the configured logical partitions, regardless of the channel subsystem to which the logical partition is configured.

A channel is considered a spanned channel if the same CHPID number in different CSSs is assigned to the same PCHID in IOCP, or is defined as *spanned* in HCD.

In the case of internal channels (for example, IC links and HiperSockets), the same applies, but with no PCHID association. They are defined with the same CHPID number in multiple CSSs.

In Figure 5-5, CHPID 04 is spanned to CSS0 and CSS1. Because it is not an external channel link, no PCHID is assigned. CHPID 06 is an external spanned channel and has a PCHID assigned.
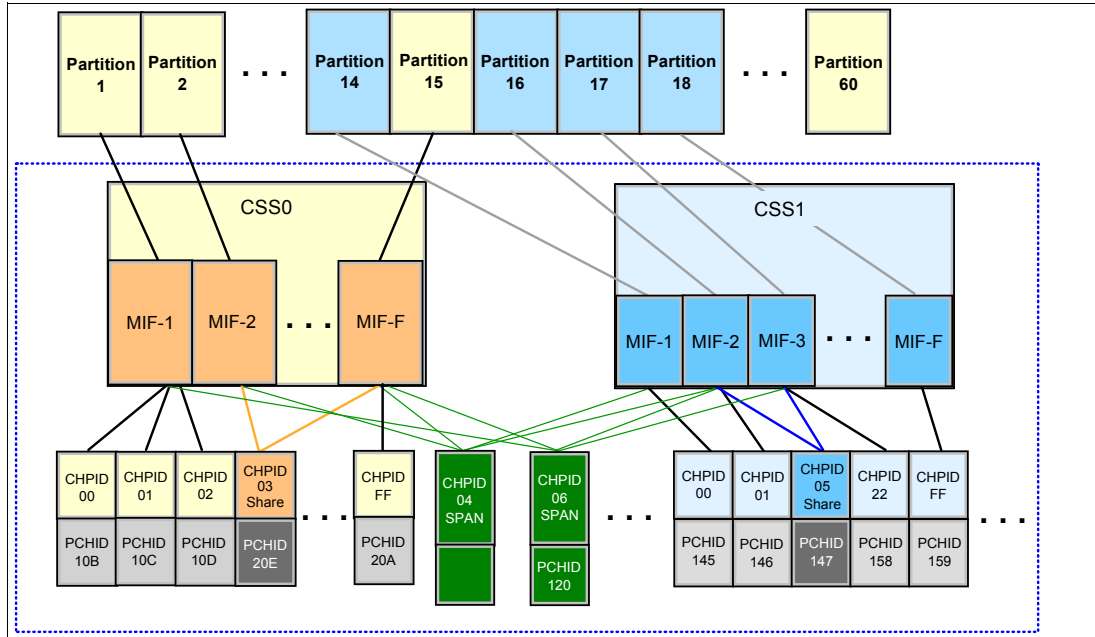
Figure 5-5   z196 CSS: two channel subsystems with channel spanning

CHPIDs that span CSSs reduce the total number of channels available. The total is reduced, because no CSS can have more than 256 CHPIDs. For a z196 with two CSSs defined, a total of 512 CHPIDs is supported. If all CHPIDs are spanned across the two CSSs, then only 256 channels are supported. For a z196 with four CSSs defined, a total of 1024 CHPIDs is supported. If all CHPIDs are spanned across the four CSSs, then only 256 channels can be supported.

Channel spanning is supported for internal links (HiperSockets and Internal Coupling (IC) links) and for certain external links (FICON Express8, and FICON Express4 channels, OSA-Express2, OSA-Express3, and Coupling Links).

> **Note:** Spanning of ESCON channels is not supported.

5.1.8 Multiple CSS construct

A pictorial view of a z196 with multiple CSSs defined is shown in Figure 5-6. In this example, two channel subsystems are defined (CSS0 and CSS1). Each CSS has three logical partitions with their associated MIF image identifiers.
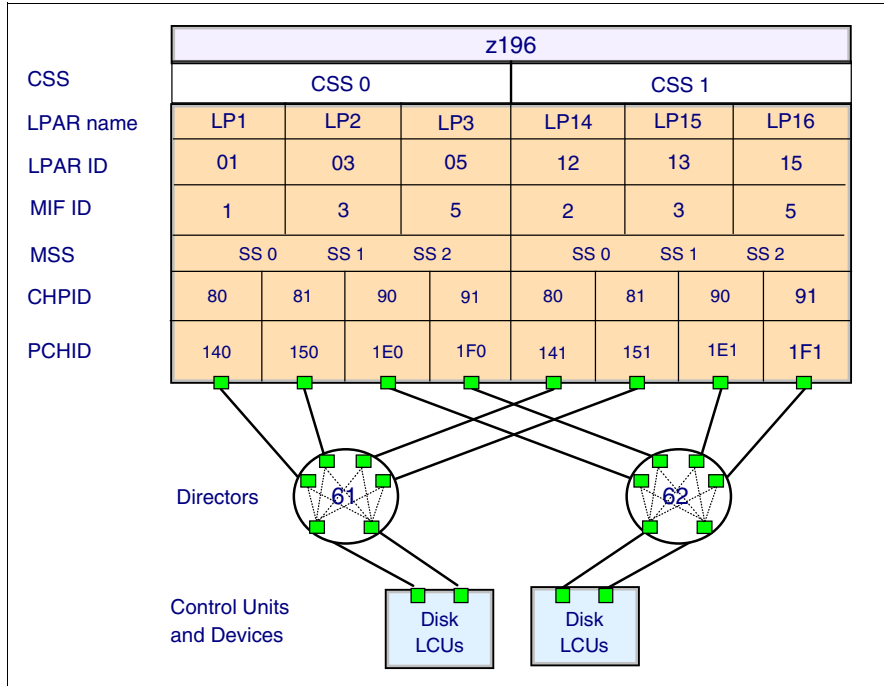
Figure 5-6   z196 CSS connectivity

In each CSS, the CHPIDs are shared across all logical partitions. The CHPIDs in each CSS can be mapped to their designated PCHIDs using the CHPID Mapping Tool (CMT) or manually using HCD or IOCP. The output of the CMT is used as input to HCD or the IOCP to establish the CHPID to PCHID assignments.

5.1.9 Adapter ID

When using HCD or IOCP to assign a CHPID to a Parallel Sysplex over InfiniBand (PSIFB) coupling link port an adapter ID (AID) number is required.

The AID is a number from 00 through 1F. If the fanout is moved to another slot, the AID changes for that specific fanout and it might be necessary to readjust the IOCDS.

The AID is bound to the serial number of the fanout. If the fanout is moved, the AID moves with it. No IOCDS update is required if adapters are moved to a new physical location.

Table 5-32 shows the assigned AID numbers for a newly built z196.

Table 5-3   Fanout AID numbers

Fanout location	Fourth book	First book	Third book	Second book
D1	00	08	10	18
D2	01	09	11	19
D3	N/A	N/A	N/A	N/A
D4	N/A	N/A	N/A	N/A
D5	02	0A	12	1A
D6	03	0B	13	1B

Fanout location	Fourth book	First book	Third book	Second book
D7	04	0C	14	1C
D8	05	0D	15	1D
D9	06	0E	16	1E
DA	07	0F	17	1F

The AIDs are shown in the PCHID report provided by an IBM representative for newly built z196 servers or for upgrades. Part of a PCHID report is shown in Example 5-1.

Example 5-1   AID assignment in a PCHID report

```
CHPIDSTART
  19756694                      PCHID REPORT
  Machine: 2817-M32 SNxxxxxxx
  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
  Source          Cage  Slot  F/C    PCHID/Ports or AID          Comment
  06/D6           A25B  D606  0163   AID=0B

  15/D6           A25B  D615  0163   AID=1B
```

For more information regarding PSIFB coupling link features, see Chapter 4, "CPC I/O system structure" on page 107.

5.2 I/O configuration management

For ease of management, we strongly recommend that HCD be used to build and control the I/O configuration definitions. HCD support for multiple channel subsystems is available with z/VM and z/OS. HCD provides the capability to make both dynamic hardware and software I/O configuration changes.

Tools are provided to help maintain and optimize the I/O configuration:

► IBM Configurator for e-business (eConfig)

   The eConfig tool is available to your IBM representative. It is used to create new configurations or upgrades of an existing configuration, and maintains installed features of those configurations. Reports produced by eConfig are helpful in understanding the changes being made for a system upgrade and what the final configuration will look like.

► Hardware configuration definition (HCD)

   HCD supplies an interactive dialog to generate the I/O definition file (IODF) and subsequently the input/output configuration data set (IOCDS). We strongly recommend that HCD or HCM be used to generate the I/O configuration, as opposed to writing IOCP statements. The validation checking that HCD performs as data is entered helps minimize the risk of errors before the I/O configuration is implemented.

► Hardware configuration management (HCM)

   HCM is a priced optional feature that supplies a graphical interface to HCD. It is installed on a PC and allows managing both the physical and the licitly aspects of a mainframe server's hardware configuration.

► CHPID mapping tool (CMT)

The CHPID Mapping Tool provides a mechanism to map CHPIDs onto PCHIDs as required. Additional enhancements have been built into the CMT to cater to the requirements of the z196. It provides the best availability recommendations for the installed features and defined configuration. CMT is a workstation-based tool available for download from the IBM Resource Link site:

http://www.ibm.com/servers/resourcelink

The health checker function in z/OS V1.10 introduces a health check in the I/O Supervisor that can help system administrators identify single points of failure in the I/O configuration.

5.3 Channel subsystem summary

The z196 provides support for the full architecture. Table 5-4 shows CSS-related information in terms of maximum values for devices, subchannels, logical partitions, and CHPIDs.

Table 5-4   z196 CSS overview

Setting	z196
Maximum number of CSSs	4
Maximum number of CHPIDs	1024
Maximum number of LPARs supported per CSS	15
Maximum number of LPARs supported per system	60
Maximum number of HSA subchannels	11505 K (191.75 K per partition x 60 partitions)
Maximum number of devices	255 K (4 CSSs x 63.75 K devices)
Maximum number of CHPIDs per CSS	256
Maximum number of CHPIDs per logical partition	256
Maximum number of subchannels per logical partition	191.75 K (63.75 K + 2 x 64 K)

All channel subsystem images (CSS images) are defined within a single I/O configuration data set (IOCDS). The IOCDS is loaded and initialized into the hardware system area (HSA) during system power-on reset.The HSA is pre-allocated in memory with a fixed size of 16 GB. This eliminates planning for HSA and pre-planning for HSA expansion, because HCD/IOCP always reserves the following items by the IOCDS process:

► Four CSSs
► Fifteen LPARs in each CSS
► Subchannel set 0 with 63.75 K devices in each CSS
► Subchannel set 1 with 64 K devices in each CSS
► Subchannel set 2 with 64 K devices in each CSS

All these are designed to be activated and used with dynamic I/O changes.

Figure 5-7 shows a logical view of the relationships. Note that each CSS supports up to 15 logical partitions. System-wide, a total of up to 60 logical partitions are supported.
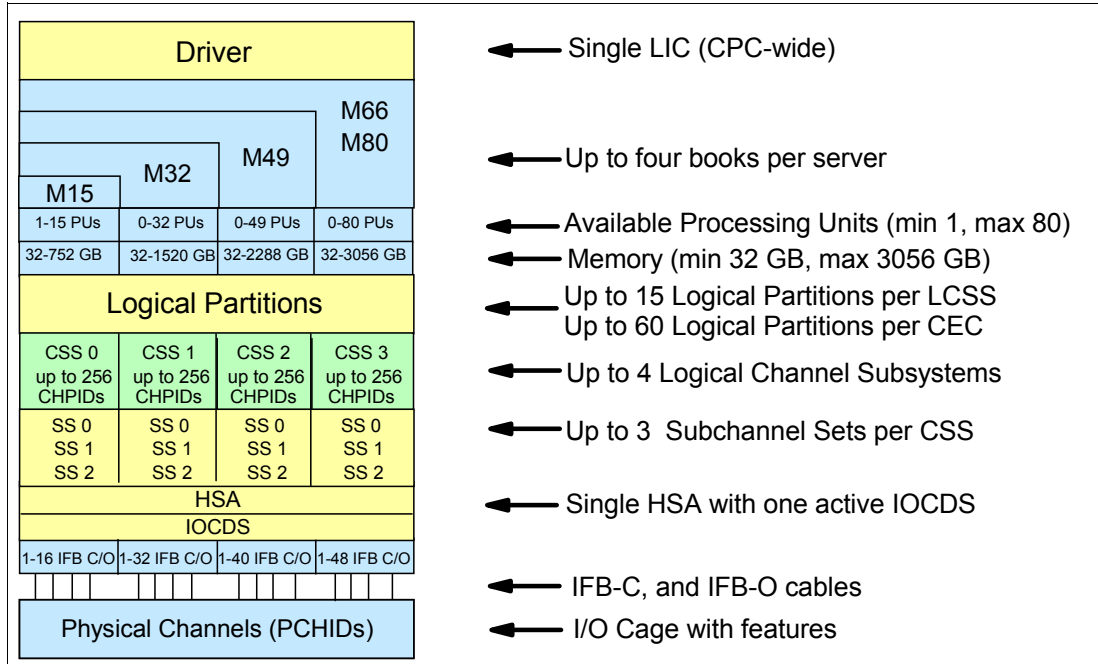
Figure 5-7   Logical view of z196 models, CSSs, IOCDS, and HSA

> **Note:** The HSA can be moved from one book to a different book in an enhanced availability configuration as part of a concurrent book repair action.

The channel definitions of a CSS are not bound to a single book. A CSS can define resources that are physically connected to any InfiniBand cable of any book in a multibook CPC.

5.4 System-initiated CHPID reconfiguration

The system-initiated CHPID reconfiguration function is designed to reduce the duration of a repair action and minimize operator interaction when an ESCON or FICON channel, an OSA port, or an ISC-3 link is shared across logical partitions on an z196 server. When an I/O card is to be replaced for a repair, it usually has some failed channels and some that are still functioning.

To remove the card, all channels must be configured offline from all logical partitions sharing those channels. Without system-initiated CHPID reconfiguration, this means that the CE must contact the operators of each affected logical partition and have them set the channels offline, and then after the repair, contact them again to configure the channels back online.

With system-initiated CHPID reconfiguration support, the support element sends a signal to the channel subsystem that a channel needs to be configured offline. The channel subsystem determines all the logical partitions sharing that channel and sends an alert to the operating systems in those logical partitions. The operating system then configures the channel offline without any operator intervention. This cycle is repeated for each channel on the card. When the card is replaced, the Support Element sends another signal to the channel subsystem for each channel. This time, the channel subsystem alerts the operating system that the channel should be configured back online. This process minimizes operator interaction to configure channels offline and online.

System-initiated CHPID reconfiguration is supported by z/OS.

5.5 Multipath initial program load

Multipath initial program load (IPL) helps increase availability and helps eliminate manual problem determination during IPL execution. This happens by allowing IPL to complete, if possible, using alternate paths when executing an IPL from a device connected through ESCON and FICON channels. If an error occurs, an alternate path is selected.

Multipath IPL is applicable to ESCON channels (CHPID type CNC) and to FICON channels (CHPID type FC). z/OS supports multipath IPL.

6

Cryptography

This chapter describes the hardware cryptographic functions available on the z196. The CP Assist for Cryptographic Function (CPACF) along with the PCIe Cryptographic Coprocessors offer a balanced use of resources and unmatched scalability.

The z196 includes both standard cryptographic hardware and optional cryptographic features for flexibility and growth capability. IBM has a long history of providing hardware cryptographic solutions, from the development of Data Encryption Standard (DES) in the 1970s to have the Crypto Express tamper-resistant features designed to meet the U.S. Government's highest security rating FIPS 140-2 Level 4[1].

The cryptographic functions include the full range of cryptographic operations necessary for e-business, e-commerce, and financial institution applications. Custom cryptographic functions can also be added to the set of functions that the z196 offers.

Today, e-business applications increasingly rely on cryptographic techniques to provide the confidentiality and authentication required in this environment. Secure Sockets Layer/Transport Layer Security (SSL/TLS) is a key technology for conducting secure e-commerce using Web servers, and it has being adopted by a rapidly increasing number of applications, demanding new levels of security, performance, and scalability.

This chapter discusses the following topics:

---

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

6.1 Cryptographic synchronous functions

Cryptographic synchronous functions are provided by the CP Assist for Cryptographic Function (CPACF). For IBM and customer written programs, CPACF functions can be invoked by instructions described in the z/Architecture Principles of Operation. As a group, these instructions are known as the Message-Security Assist (MSA). z/OS Integrated Cryptographic Service Facility (ICSF) callable services and z90crypt device driver running at Linux on System z also invoke CPACF synchronous functions.

The z196 hardware includes the implementation of algorithms as hardware synchronous operations, which means holding the PU processing of the instruction flow until the operation has completed. The synchronous functions are:

► Data encryption and decryption algorithms for data privacy and confidentially

   Data Encryption Standard (DES), which includes:

   – Single-length key DES

   – Double-length key DES

   – Triple-length key DES (also known as Triple-DES)

   Advanced Encryption Standard (AES) for 128-bit, 192-bit, and 256-bit keys.

► Hashing algorithms for data integrity, such as SHA-1, and SHA-2 support for SHA-224, SHA-256, SHA-384, and SHA-512.

► Message authentication code (MAC)

   – Single-length key MAC
   – Double-length key MAC

► Pseudo Random Number Generation (PRNG) for cryptographic key generation.

> **Note:** Keys must be provided in clear form only.

SHA-1, and SHA-2 support for SHA-224, SHA-256, SHA-384, and SHA-512 are shipped enabled on all servers and do not require the CPACF enablement feature. The CPACF functions are supported by z/OS, z/VM, z/VSE, and Linux on System z.

6.2 Cryptographic asynchronous functions

Cryptographic asynchronous functions are provided by the PCI Express (PCIe) cryptographic adapters.

6.2.1 Secure key functions

The following secure key functions are provided as cryptographic asynchronous functions. System internal messages are passed to the cryptographic coprocessors to initiate the operation, then messages are passed back from the coprocessors to signal completion of the operation.

► Data encryption and decryption algorithms

   Data Encryption Standard (DES), which includes:

   – Single-length key DES
   – Double-length key DES

- Triple-length key DES (Triple-DES)

► DES key generation and distribution

► PIN generation, verification, and translation functions

► Random number generator

► Public key algorithm (PKA) facility

Supported callable services intended for application programs that use PKA include:

- Importing RSA public-private key pairs in clear and encrypted forms

- Rivest-Shamir-Adelman (RSA), which can provide:
  - Key generation, up to 4,096-bit
  - Signature verification, up to 4,096-bit
  - Import and export of DES keys under an RSA key, up to 4,096-bit

- Public key encryption (PKE)

  The PKE service is provided for assisting the SSL/TLS handshake. PKE is used to offload compute-intensive portions of the protocol onto the cryptographic adapters.

- Public key decryption (PKD)

  PKD supports a zero-pad option for clear RSA private keys. PKD is used as an accelerator for raw RSA private operations, such as those required by the SSL/TLS handshake and digital signature generation. The Zero-Pad option is exploited by Linux on System z to allow the use of cryptographic adapters for improved performance of digital signature generation.

- Europay Mastercard VISA (EMV) 2000 standard

  Applications may be written to comply with the EMV 2000 standard for financial transactions between heterogeneous hardware and software. Support for EMV 2000 requires PCIe feature at z196.

The Crypto Express3 card, a PCI Express cryptographic adapter, offers SHA-2 functions similar to those functions offered in the CPACF. This is in addition to the functions mentioned above.

6.2.2 Protected key

The z196 supports the protected key implementation. Since PCIXCC deployment, secure keys are processed on the PCI-X and PCIe cards, requiring an asynchronous operation to move the data and keys from the general purpose CP to the crypto cards. Clear keys process faster than secure keys because the process is done synchronously on the CPACF. Protected keys blend the security of Crypto Express3 (CEX3) and the performance characteristics of the CPACF, running closer to the speed of clear keys.

An enhancement to CPACF facilitate the continued privacy of cryptographic key material when used for data encryption. In Crypto Express3, a secure key is encrypted under a master key, while a protected key is encrypted under a wrapping key that is unique to each LPAR. Once the wrapping key is unique to each LPAR, a protected key cannot be shared with another LPAR. CPACF, using key wrapping, ensures that key material is not visible to applications or operating systems during encryption operations.

CPACF code generates the wrapping key and stores it in the protected area of hardware system area (HSA).Wrapping key is accessible only by firmware. It cannot be accessed by operating systems or applications. DES/T-DES and AES algorithms were implemented in CPACF code with support of hardware assist functions. Two variations of wrapping key are

generated, one for DES/T-DES keys and another for AES keys. Wrapping keys are generated during the clear reset each time an LPAR is activated or reset. There is no customizable option available at SE or HMC that permits or avoids the wrapping key generation. Figure 6-1 shows this function.
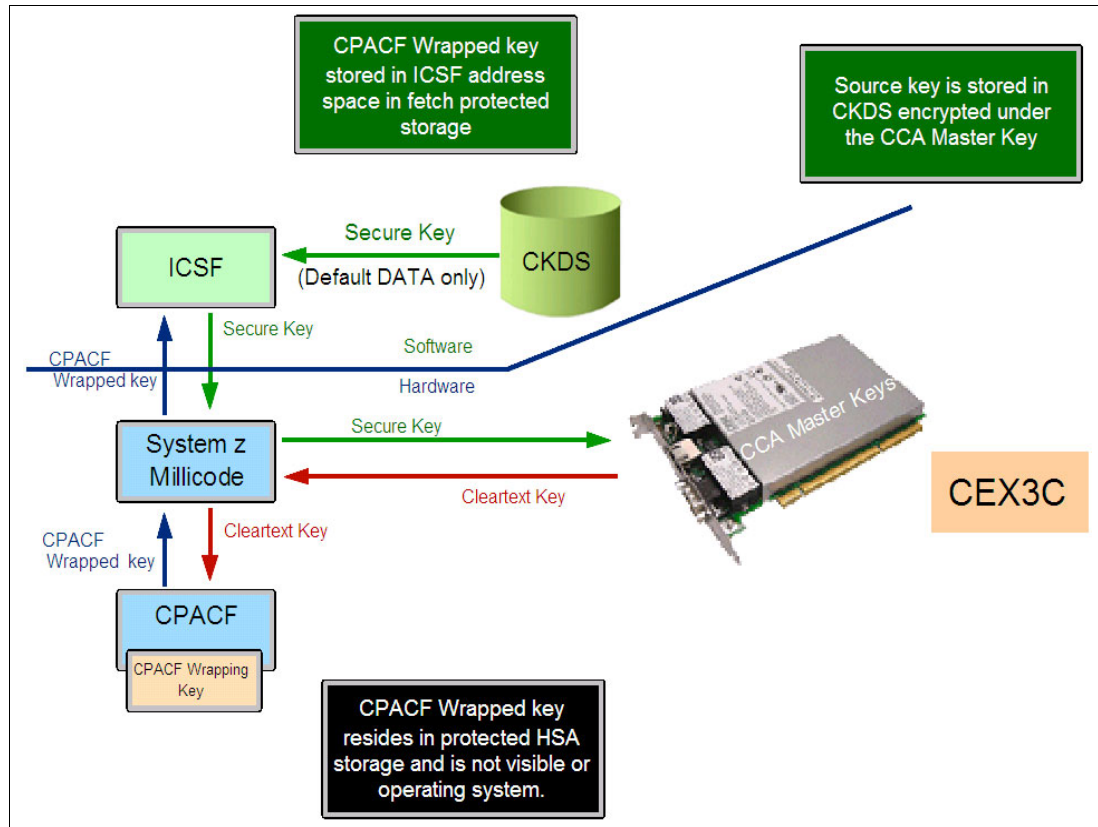
Figure 6-1   CPACF key wrapping

If a CEX3 coprocessor is available, a protected key may begin its life as a secure key. Otherwise, an application is responsible for creating or loading a clear key value and then use the new PCKMO instruction to wrap the key. ICSF is not called by application if CEX3C is not available.

A new segment in profiles at CSFKEYS class in RACF® restrict which secure keys can be used as protected keys. By default, all secure keys are considered not eligible to be used as protected keys. The process described in Figure 6-1 considers a secure key being the source of a protected key.

At Figure 6-1, the source key is already stored in CKDS as a secure key (encrypted under the master key). This secure key is sent to CEX3C to be deciphered and sent to CPACF in clear text. At CPACF, the key is wrapped under the LPAR wrapping key and them it is returned to ICSF. Once the key is wrapped, ICSF can keep the protected value in memory, passing it to the CPACF when the key will be unwrapped for each encryption/decryption operation.

Protected key is designed to provide substantial throughput improvements for large volume of data encryption as well as low latency for encryption of small blocks of data. High performance secure key solution, also know as protected key solution, requires ICSF HCR7770 and it is highly recommended Crypto Express3 card.

6.2.3 Other key functions

Other key functions of the Crypto Express features serve to enhance the security of public and private key encryption processing:

► Remote loading of initial ATM keys

This function provides the ability to remotely load the initial keys for capable Automated Teller Machines (ATM) and Point of Sale (POS) systems. Remote key loading refers to the process of loading DES keys to ATM from a central administrative site without requiring someone to manually load the DES keys on each machine. The process uses ICSF callable services along with the Crypto Express features to perform the remote load.

ICSF has added two callable services, Trusted Block Create (CSNDTBC) and Remote Key Export (CSNDRKX). CSNDTBC is a callable service that is used to create a trusted block containing a public key and certain processing rules. The rules define the ways and formats in which keys are generated and exported. CSNDRKX is a callable service that uses the trusted block to generate or export DES keys for local use and for distribution to an ATM or other remote device. The PKA Key Import (CSNDPKI), PKA Key Token Change (CSNDKTC), and Digital Signature Verify (CSFNDFV) callable services support remote key loading.

► Key exchange with non-CCA cryptographic systems

This function allows the exchange of operational keys between the Crypto Express3 and non-CCA systems, such as the Automated Teller Machines (ATM). IBM Common Cryptographic Architecture (CCA) employs control vectors to control usage of cryptographic keys. Non-CCA systems use other mechanisms, or can use keys that have no associated control information. Enhancements to key exchange functions added to CCA the ability to exchange keys between CCA systems and systems that do not use control vectors.It allows the CCA system owner to define permitted types of key import and export while preventing uncontrolled key exchange that can open the system to an increased threat of attack.

► Retained key support

Retained keys are RSA keys generated within the secure boundary of the card and never leave the secure boundary. Only the domain that created the retained key can access it.

► User-Defined Extensions (UDX) support.

UDX allows the user to add customized operations to a cryptographic processor. User-Defined Extensions to the Common Cryptographic Architecture (CCA) support customized operations that execute within the Crypto Express features when defined as coprocessor.

UDX is supported under a special contract through an IBM or approved third-party service offering. The CryptoCards Web site directs your request to an IBM Global Services location appropriate for your geographic location. A special contract is negotiated between you and IBM Global Services. The contract is for development of the UDX by IBM Global Services according to your specifications and an agreed-upon level of the UDX.

It is not possible to mix and match UDX definitions across Crypto Express2 and Crypto Express3 features. Panels on the HMC and SE ensure that UDX files are applied to the appropriate crypto card type.

An UDX toolkit for System z is available for the Crypto Express3 feature. In addition, there is a migration path for customers with UDX on a previous feature to migrate their code to the Crypto Express3 feature. An UDX migration is no more disruptive than a normal MCL or ICSF release migration.

More information can be found on the IBM CryptoCards Web site:

http://www.ibm.com/security/cryptocards

6.2.4 Cryptographic feature codes

Table 6-1 lists the cryptographic features available.

Table 6-1   Cryptographic features for System z CPC

Feature code	Description
3863	CP Assist for Cryptographic Function (CPACF) enablement This feature is a prerequisite to use CPACF (except for SHA-1, SHA-224, SHA-256, SHA-384, and SHA-512) and Crypto Express features.
0864	Crypto Express3 feature A maximum of eight features may be ordered. Each feature contains two PCI Express cryptographic adapters (adjunt processors).
0841	Trusted Key Entry (TKE) workstation This feature is optional. TKE provides basic key management (key identification, exchange, separation, update, backup), as well as security administration. The TKE workstation has one Ethernet port and supports connectivity to an Ethernet Local Area Network (LAN) operating at 10, 100, or 1000 Mbps. Up to ten (10) features per z196 may be installed
0860	TKE 7.0 Licensed Internal Code (TKE 7.0 LIC) The 7.0 LIC requires Trusted Key Entry workstation feature code 0841. It is required to support z196. The 7.0 LIC can also be used to control z10 EC, z10 BC, z9 EC, z9 BC, z990, and z890 servers.
0885	TKE Smart Card Reader Access to information about the smart card is protected by a personal identification number (PIN). One (1) feature code includes two Smart Card Readers, two cables to connect to the TKE 7.0 workstation, and 20 smart cards.
0884	TKE additional smart cards When one feature code is ordered a quantity of 10 smart cards are shipped. Order increment is one up to 99 (990 blank Smart Cards).

TKE includes support for the AES encryption algorithm with 256-bit master keys and key management functions to load or generate master keys to the cryptographic coprocessor.

If the TKE workstation is chosen to operate the Crypto Express features, a TKE workstation with the TKE 7.0 LIC or later is required. See 6.5, "TKE workstation feature" on page 175 for a more detailed description.

> **Important:** Products that include any of the cryptographic feature codes contain cryptographic functions that are subject to special export licensing requirements by the United States Department of Commerce. It is the customer's responsibility to understand and adhere to these regulations when moving, selling, or transferring these products.

6.3 CP Assist for Cryptographic Function

The CP Assist for Cryptographic Function (CPACF) offers a set of symmetric cryptographic functions that enhance the encryption and decryption performance of clear key operations for SSL, VPN, and data-storing applications that do not require FIPS 140-2 level 4 security[2].

CPACF is designed to facilitate the privacy of cryptographic key material when used for data encryption through key wrapping implementation. It ensures that key material is not visible to applications or operating systems during encryption operations

The CPACF feature provides hardware acceleration for DES, Triple-DES, MAC, AES-128, AES-192, AES-256, SHA-1, SHA-224, SHA-256, SHA-384, and SHA-512 cryptographic services. It provides high-performance hardware encryption, decryption, and hashing support.

The following instructions support the cryptographic assist function:

KMAC	Compute Message Authentic Code.
KM	Cipher Message.
KMC	Cipher Message with Chaining.
KMF	Cipher Message with CFB.
KMCTR	Cipher Message with Counter.
KMO	Cipher Message with OFB.
KIMD	Compute Intermediate Message Digest.
KLMD	Compute Last Message Digest.
PCKMO	Provide Cryptographic Key Management Operation.

New function codes for existing instructions were introduced at z196:

► Compute intermediate Message Digest (KIMD) adds KIMD- GHASH

These functions are provided as problem-state z/Architecture instructions, directly available to application programs. They are known as Message-Security Assist (MSA). When enabled, the CPACF runs at processor speed for every CP, IFL, zIIP, and zAAP.

The cryptographic architecture includes DES, Triple-DES, MAC message authentication, AES data encryption and decryption, SHA-1, and SHA-2 support for SHA-224, SHA-256, SHA-384, and SHA-512 hashing.

The functions of the CPACF must be explicitly enabled using FC 3863 by the manufacturing process or at the customer site as a MES installation, except for SHA-1, and SHA-2 support for SHA-224, SHA-256, SHA-384, and SHA-512, which are always enabled.

6.4 Crypto Express3

The Crypto Express3 feature (FC 0864) has two Peripheral Component Interconnect Express (PCIe) cryptographic adapters. Each of the PCI Express cryptographic adapters can be configured as a cryptographic coprocessor or a cryptographic accelerator.

The Crypto Express3 feature is the newest state-of-the-art generation cryptographic feature. Like its predecessors it is designed to complement the functions of CPACF. This feature is tamper-sensing and tamper-responding. It provides dual processors operating in parallel supporting cryptographic operations with high reliability.

The CEX3 uses the 4765 PCIe Coprocessor. It holds a secured subsystem module, batteries for backup power and a full-speed USB 2.0 host port available through a mini-A connector. On System z these USB ports are not used. The securely encapsulated subsystem contains two 32-bit PowerPC® 405D5 RISC processors running in lock-step with cross-checking to detect malfunctions as well as a separate service processor used to manage self-test and firmware updates, RAM, flash memory, and battery-powered memory, cryptographic-quality

---

[2] Federal Information Processing Standard

random number generator, AES, DES, TDES, SHA-1, SHA-224, SHA-256, SHA-384, SHA-512 and modular-exponentiation (for example, RSA, DSA) hardware, and full-duplex DMA communications. Figure 6-2 shows the physical layout of the Crypto Express3 feature.
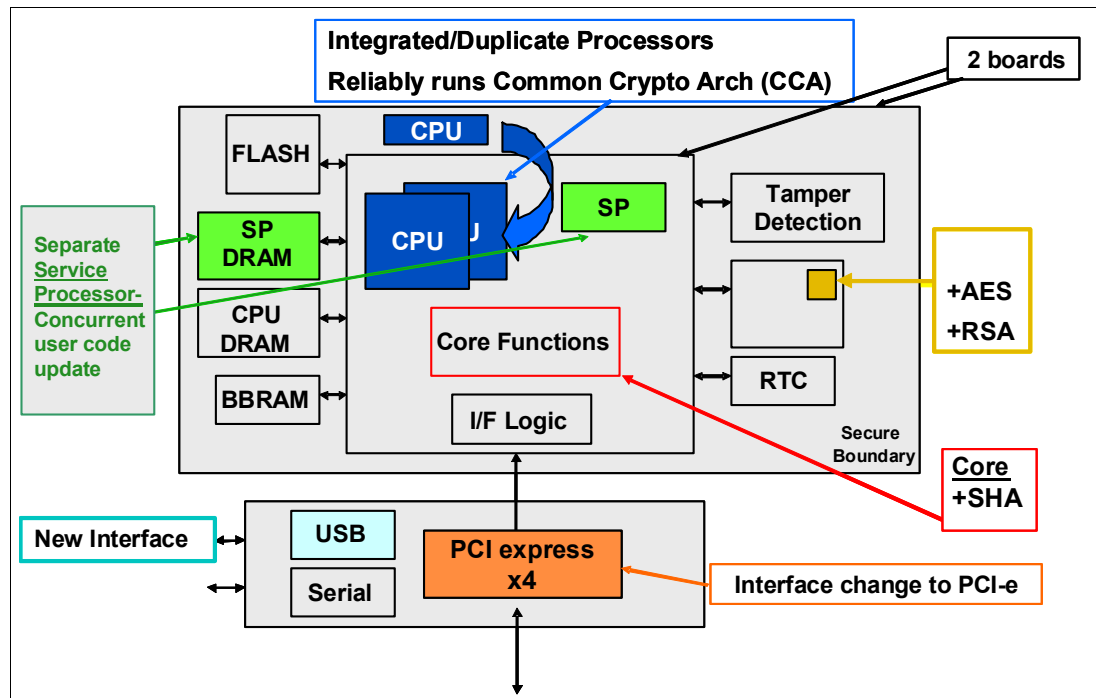
Figure 6-2   Crypto Express3 feature layout

The Crypto Express3 feature does not have external ports and does not use fiber optic or other cables. It does not use CHPIDs, but requires one slot in the I/O cage and one PCHID for each PCI-e cryptographic adapter. Removal of the feature or card *zeroizes* the content.

The z196 supports a maximum of eight Crypto Express3 features, offering a combination of up to 16 coprocessor and accelerators. Access to the PCI-e cryptographic adapter is controlled through the setup in the image profiles on the SE.

> **Note:** Although PCI-e cryptographic adapters have no CHPID type and are not identified as external channels, all logical partitions in all channel subsystems have access to the adapter (up to 16 logical partitions per adapter). Having access to the adapter requires setup in the image profile for each partition. The adapter must be in the candidate list.

The Crypto Express3 feature, residing in the I/O cage of the z196, continues to support all of the cryptographic functions available on Crypto Express3 on System z10. When one or both of the two PCIe adapters are configured as a coprocessor, the following cryptographic enhancements introduced at z196 are supported:

► ANSI X9.8 PIN security.

– It facilitates compliance with the processing requirements defined in the new version of the ANSI X9.8 and ISO 9564 PIN Security Standards and provides added security for transactions that require Personal Identification Numbers (PIN).

► Enhance CCA key wrapping to comply with ANSI X9.24-1 key bundling requirements.

– A new Common Cryptographic Architecture (CCA) key token wrapping method uses Cipher Block Chaining (CBC) mode in combination with other techniques to satisfy the key bundle compliance requirements in standards including ANSI X9.24-1 and the

recently published Payment Card Industry Hardware Security Module (PCI HSM) standard.

► Secure key HMAC (Keyed-Hash Message Authentication Code)

  – HMAC is a method for computing a message authentication code using a secret key and a secure hash function. It is defined in the standard FIPS (Federal Information Processing Standard) 198, "The Keyed-Hash Message Authentication Code (HMAC)". The new CCA functions support HMAC using SHA-1, SHA-224, SHA-256, SHA-384, and SHA-512 hash algorithms. The HMAC keys are variable-length and are securely encrypted so that their values are protected. This Crypto function is supported by z/OS, z/VM and Linux on System z.

► Elliptic Curve Cryptography (ECC) Digital Signature Algorithm

  – Elliptic Curve Cryptography is an emerging public-key algorithm to eventually replace RSA cryptography in many applications. ECC is capable of providing digital signature functions and key agreement functions. The new CCA functions provide ECC key generation and key management and provide digital signature generation and verification functions compliant with the ECDSA method described in ANSI X9.62 "Public Key Cryptography for the Financial Services Industry: The Elliptic Curve Digital Signature Algorithm (ECDSA)". ECC uses keys that are shorter than RSA keys for equivalent strength-per-key-bit; RSA is impractical at key lengths with strength-per-key-bit equivalent to AES-192 and AES-256. So the strength-per-key-bit is substantially greater in an algorithm that uses elliptic curves. This Crypto function is supported by z/OS, z/VM and Linux on System z.

> **Note:** Elliptical Curve Cryptography technology (ECC) is delivered through the machine's Machine Code (also called Licensed Internal Code, or LIC), and requires license terms in addition to the standard IBM License Agreement for Machine Code (LMC). These additional terms are delivered through the LMC's Addendum for Elliptical Curve Cryptography. This ECC Addendum will be delivered with the machine along with the LMC when a cryptography feature is included in the z196 order, or when a cryptography feature is carried forward as part of an MES order into z196.

► Concurrent Driver Upgrade (CDU) and Concurrent Path Apply (CPA)

  – It is a process to eliminate or reduce cryptographic coprocessor card outages for new Cryptographic function releases. With concurrent driver upgrade and concurrent patch apply, new cryptographic functions can be applied without configuring the Cryptographic coprocessor card off / on. It is now possible to upgrade Common Cryptographic Architecture (CCA), segment 3, licensed internal code without any performance impact during the upgrade. However, some levels of Common Cryptographic Architecture (CCA) or hardware changes will still require cryptographic coprocessor card vary off / on. This Crypto function is exclusive to z196.

Additional key features of Crypto Express3 include:

► Dynamic power management to maximize RSA performance while keeping the CEX3 within temperature limits of the tamper-responding package.

► All logical partitions (LPARs) in all Logical Channel Subsystems (LCSSs) have access the Crypto Express3 feature, up to 32 LPARs per feature.

► Secure code loading that enables the updating of functionality while installed in application systems.

- ► Lock-step checking of dual CPUs for enhanced error detection and fault isolation of cryptographic operations performed by a coprocessor when a PCI-E adapter is defined as a coprocessor.

- ► Improved RAS over previous crypto features due to dual processors and the service processor.

- ► Dynamic addition and configuration of the Crypto Express3 features to LPARs without an outage.

The Crypto Express3 feature is designed to deliver throughput improvements for both symmetric and asymmetric operations.

A Crypto Express3 migration wizard is available to make the migration easier. The wizard allows the user to collect configuration data from a Crypto Express2 or Crypto Express3 feature configured as a coprocessor and migrate that data to a different Crypto Express coprocessor. The target for this migration must be a coprocessor with equivalent or greater capabilities.

6.4.1 Crypto Express3 coprocessor

The Crypto Express3 coprocessor is a PCI-e cryptographic adapter configured as a coprocessor and provides a high-performance cryptographic environment with added functions.

The Crypto Express3 coprocessor provides asynchronous functions only.

The Crypto Express3 feature contains two PCI-e cryptographic adapters. The two adapters provide the equivalent (plus additional) functions as the PCIXCC and Crypto Express2 features with improved throughput.

PCI-e cryptographic adapters, when configured as coprocessors, are designed for FIPS 140-2 Level 4 compliance rating for secure cryptographic hardware modules. Unauthorized removal of the adapter or feature *zeroizes* its content.

The Crypto Express3 coprocessor enables the user to:

- ► Encrypt and decrypt data by using secret-key algorithms. Triple-length key DES and double-length key DES as well as AES algorithms are supported.

- ► Generate, install, and distribute cryptographic keys securely by using both public and secret-key cryptographic methods.

- ► Generate, verify, and translate personal identification numbers (PINs).

- ► CEX3C supports 13 through 19-digit personal account numbers (PANs).

- ► Ensure the integrity of data by using message authentication codes (MACs), hashing algorithms, and Rivest-Shamir-Adelman (RSA) public key algorithm (PKA) digital signatures as well as Elliptic Curve Cryptography (ECC) digital signatures.

The Crypto Express3 coprocessor also provides the functions listed for the Crypto Express3 accelerator, however, with a lower performance than the Crypto Express3 accelerator can provide.

Three methods of master key entry are provided by Integrated Cryptographic Service Facility (ICSF) for the Crypto Express3 feature coprocessor:

- ► A pass-phrase initialization method, which generates and enters all master keys that are necessary to fully enable the cryptographic system in a minimal number of steps.

- ► A simplified master key entry procedure provided through a series of Clear Master Key Entry panels from a TSO terminal.

- ► A Trusted Key Entry (TKE) workstation, which is available as an optional feature in enterprises that require enhanced key-entry security.

Linux on System z also permits the master key entry through panels or through TKE workstation.

The security-relevant portion of the cryptographic functions is performed inside the secure physical boundary of a tamper-resistant card. Master keys and other security-relevant information are also maintained inside this secure boundary.

A Crypto Express3 coprocessor operates with the Integrated Cryptographic Service Facility (ICSF) and IBM Resource Access Control Facility (RACF), or equivalent software products, in a z/OS operating environment to provide data privacy, data integrity, cryptographic key installation and generation, electronic cryptographic key distribution, and personal identification number (PIN) processing. These functions are also available at CEX3 coprocessor running in a Linux for System z environment.

The Processor Resource/Systems Manager (PR/SM) fully supports the Crypto Express3 coprocessor feature to establish a logically partitioned environment on which multiple logical partitions can use the cryptographic functions. A 128-bit data-protection symmetric master key, a 256-bit AES master key, a 256-bit ECC master key and one 192-bit public key algorithm (PKA) master key are provided for each of 16 cryptographic domains that a coprocessor can serve.

Use the dynamic addition or deletion of a logical partition name to rename a logical partition. Its name can be changed from NAME1 to * (single asterisk) and then changed again from * to NAME2. The logical partition number and MIF ID are retained across the logical partition name change. The master keys in the Crypto Express3 feature coprocessor that were associated with the old logical partition NAME1 are retained. No explicit action is taken against a cryptographic component for this dynamic change.

Note: Cryptographic coprocessors are not tied to logical partition numbers or MIF IDs. They are set up with PCI-e adapter numbers and domain indices that are defined in the partition image profile. The customer can dynamically configure them to a partition and change or clear them when needed.

6.4.2 Crypto Express3 accelerator

The Crypto Express3 accelerator is a coprocessor that is reconfigured by the installation process so that it uses only a subset of the coprocessor functions at a higher speed. Note the following information about the reconfiguration:

- ► It is done through the Support Element.

- ► It is done at the PCI-e cryptographic adapter level. A Crypto Express3 feature can host a coprocessor and an accelerator, two coprocessors, or two accelerators.

- ► It works both ways, from coprocessor to accelerator and from accelerator to coprocessor. Master keys in the coprocessor domain can be optionally preserved when it is reconfigured to be an accelerator.

- ► Reconfiguration is disruptive to coprocessor and accelerator operations. The coprocessor or accelerator must be deactivated before engaging the reconfiguration.

► FIPS 140-2 certification is not relevant to the accelerator because it operates with clear keys only.

► The function extension capability through UDX is not available to the accelerator.

The functions that remain available when CEX3 is configured as an accelerator are used for the acceleration of modular arithmetic operations (that is, the RSA cryptographic operations used with the SSL/TLS protocol), as follows:

► PKA Decrypt (CSNDPKD), with PKCS-1.2 formatting
► PKA Encrypt (CSNDPKE), with zero-pad formatting
► Digital Signature Verify

The RSA encryption and decryption functions support key lengths of 512 bit to 4,096 bit, in the Modulus Exponent (ME) and Chinese Remainder Theorem (CRT) formats.

6.4.3 Configuration rules

Each z196 supports up to eight Crypto Express3 features, which equals up to a maximum of 16 PCI-e cryptographic adapters. In a one-book system up to eight features may be installed and configured. Table 6-2 summarizes configuration information for Crypto Express3.

Table 6-2   Crypto Express3 feature

Minimum number of orderable features for each server[a]	2
Order increment above two features	1
Maximum number of features for each server	8
Number of PCI-e cryptographic adapters for each feature (coprocessor or accelerator)	2
Maximum number of PCI-e adapters for each server	16
Number of cryptographic domains for each PCI-e adapter[b]	16

a. The minimum initial order of Crypto Express3 features is two. After the initial order, additional Crypto Express3 can be ordered one feature at a time up to a maximum of eight.
b. More than one partition, defined to the same CSS or to different CSSs, can use the same domain number when assigned to different PCI-e cryptographic adapters.

The concept of *dedicated processor* does not apply to the PCI-e cryptographic adapter. Whether configured as coprocessor or accelerator, the PCI-e cryptographic adapter is made available to a logical partition as directed by the domain assignment and the candidate list in the logical partition image profile, regardless of the shared or dedicated status given to the CPs in the partition.

When installed non-concurrently, Crypto Express3 features are assigned PCI-e cryptographic adapter numbers sequentially during the power-on reset following the installation. When a Crypto Express3 feature is installed concurrently, the installation can select an out-of-sequence number from the unused range. When a Crypto Express3 feature is removed concurrently, the PCI-e adapter numbers are automatically freed.

The definition of domain indexes and PCI-e cryptographic adapter numbers in the candidate list for each logical partition should be planned ahead to allow for nondisruptive changes, as follows.

► Operational changes can be made by using the Change LPAR Cryptographic Controls task from the Support Element, which reflects the cryptographic definitions in the image

profile for the partition. With this function, adding and removing the cryptographic feature without stopping a running operating system can be done dynamically.

► The same usage domain index may be defined more than once across multiple logical partitions. However, the PCI-e cryptographic adapter number coupled with the usage domain index specified must be unique across all active logical partitions.

The same PCI-e cryptographic adapter number and usage domain index combination may be defined for more than one logical partition, for example to define a configuration for backup situations. Note that only one of the logical partitions can be active at any one time.

The z196 allows for up to 60 logical partitions to be active concurrently. Each PCIe adapter supports 16 domains, whether it is configured as a Crypto Express3 accelerator or a Crypto Express3 coprocessor. The server configuration must include at least two Crypto Express3 (four PCI-e adapters and 16 domains per PCI-e adapter) when all 60 logical partitions require concurrent access to cryptographic functions. More Crypto Express3 features may be needed to satisfy application performance and availability requirements.

6.5 TKE workstation feature

The TKE, Trusted Key Entry, workstation is an optional feature that offers key management functions. The TKE workstation, feature code 0841, contains a combination of hardware and software. Included with the system unit are a mouse, keyboard, flat panel display, PCIe adapter and a writeable USB media to install TKE Licensed Internal Code (LIC). The TKE workstation feature code 0841 will be the first to have Crypto Express3 installed. TKE LIC V7.0 requires CEX3 and it will not be supported on TKE workstation feature code 0840.

> **Note:** The TKE workstation supports Ethernet adapters only to connect to a LAN.

A TKE workstation is part of a customized solution for using the Integrated Cryptographic Service Facility for z/OS program product (ICSF for z/OS) or the Linux for System z to manage cryptographic keys of a z196 that has Crypto Express features installed and that is configured for using DES, AES, ECC and PKA cryptographic keys.

The TKE provides a secure, remote and flexible method of providing Master Key Part Entry, and to remotely manage PCIe Cryptographic Coprocessor(s). The cryptographic functions on the TKE are performed by one PCIe Cryptographic Coprocessor. The TKE workstation communicates with the System z server using a TCP/IP connection. The TKE workstation is available with Ethernet LAN connectivity only. Up to ten TKE workstations can be ordered. TKE feature number 0841 can be used to control the z196 and it can also be used to control z10 EC, z10 BC, z9 EC, z9 BC, z990, and z890 servers.

The TKE workstation feature code 0841 along with LIC 7.0 offers a significant number of enhancements:

► ECC Master Key Support

ECC keys will be protected using a new ECC master key (256-bit AES key). From the TKE, administrators can generate key material, load or clear the new ECC master key register, or clear the old ECC master key register. The ECC key material can be stored on the TKE or on a smart card.

► CBC Default Settings Support

The TKE provides function that allows the TKE user to set the default key wrapping method used by the host crypto module.

► TKE Audit Record Upload Configuration Utility Support

The TKE Audit Record Upload Configuration Utility allows Trusted Key Entry (TKE) workstation audit records to be sent to a System z host and saved on the host as z/OS System Management Facilities (SMF) records. The SMF records have a record type of 82 (ICSF) and a subtype of 29. TKE workstation audit records are sent to the same TKE Host Transaction Program that is used for Trusted Key Entry operations.

► USB Flash Memory Drive Support

The TKE workstation now supports a USB flash memory drive as a removable media device. When a TKE application displays media choices, the application allows you to choose a USB flash memory drive if the IBM supported drive is plugged into a USB port on the TKE and it has been formatted for the specified operation.

► Stronger Pin Strength Support

TKE smart cards created on TKE 7.0 require a 6-digit pin rather than a 4-digit pin. TKE smart cards that were created prior to TKE 7.0 will continue to use 4-digit pins and will work on TKE 7.0 without changes. You can take advantage of the stronger pin strength by initializing new TKE smart cards and copying the data from the old TKE smart cards to the new TKE smart cards.

► Stronger Password Requirements for TKE Passphrase User Profile Support

New rules are required for the passphrase used for passphrase logon to the TKE workstation crypto adapter. The passphrase must:

– be 8 to 64 characters long

– contain at least 2 numeric and 2 non-numeric characters

– not contain the user ID

These rules are enforced when you define a new user profile for passphrase logon, or when you change the passphrase for an existing profile. Your current passphrases will continue to work.

► Simplified TKE usability with Crypto Express3 migration wizard

A wizard is now available to allow users to collect data, including key material, from a Crypto Express coprocessor and migrate the data to a different Crypto Express coprocessor. The target Crypto Express coprocessor must have the same or greater capabilities. This wizard is an aid to help facilitate migration from Crypto Express2 to Crypto Express3. Crypto Express2 is not supported on z196. Benefits of using this wizard include:

– Reduces migration steps, thereby minimizing user errors

– Minimizes the number of user clicks

– Significantly reduces migration task duration

Logical partition, TKE host, and TKE target

If one or more logical partitions are customized for using Crypto Express coprocessors, the TKE workstation can be used to manage DES, AES, ECC and PKA master keys for all cryptographic domains of each Crypto Express coprocessor feature assigned to the logical partitions defined to the TKE workstation.

Each logical partition in the same system using a domain managed through a TKE workstation connection is either a TKE host or a TKE target. A logical partition with a TCP/IP connection to the TKE is referred to as TKE host. All other partitions are TKE targets.

The cryptographic controls as set for a logical partition through the Support Element determine whether the workstation is a TKE host or TKE target.

Optional smart card reader

Adding an optional smart card reader (FC 0885) to the TKE workstation is possible. One (1) feature code 0885 includes two Smart Card Readers, two cables to connect to the TKE 7.0 workstation, and 20 smart cards. The reader supports the use of smart cards that contain an embedded microprocessor and associated memory for data storage that can contain the keys to be loaded into the Crypto Express features. Access to and use of confidential data on the smart card is protected by a user-defined personal identification number (PIN). Up to 990 additional smart cards can be ordered for backup. The additional smart card feature code is FC 0884 and one feature code is ordered a quantity of ten smart cards are shipped. Order increment is one up to 99 (990 blank Smart Cards).

6.6 Cryptographic functions comparison

Table 6-3 lists functions or attributes on z196 of the three cryptographic hardware features. In the table, X indicates the function or attribute is supported.

Table 6-3   Cryptographic functions on z196

Functions or attributes	CPACF	Crypto Express3 Coprocessor	Crypto Express3 Accelerator
Supports z/OS applications using ICSF	X	X	X
Supports Linux on System z CCA applications	X	X	X
Encryption and decryption using secret-key algorithm	-	X	-
Provides highest SSL/TLS handshake performance	-	-	X[a]
Provides highest symmetric (clear key) encryption performance	X	-	-
Provides highest asymmetric (clear key) encryption performance	-	-	X
Provides highest asymmetric (encrypted key) encryption performance	-	X	-
Disruptive process to enable	-	Note[b]	Note[b]
Requires IOCDS definition	-	-	-
Uses CHPID numbers	-	-	-
Uses PCHIDs		X [c]	X[c]
Requires CPACF enablement (FC 3863)	X[d]	X[d]	X[d]
Requires ICSF to be active	-	X	X
Offers user programming function (UDX)	-	X	-
Usable for data privacy: encryption and decryption processing	X	X	-

Functions or attributes	CPACF	Crypto Express3 Coprocessor	Crypto Express3 Accelerator
Usable for data integrity: hashing and message authentication	X	X	-
Usable for financial processes and key management operations	-	X	-
Crypto performance RMF™ monitoring	-	X	X
Requires system master keys to be loaded	-	X	-
System (master) key storage	-	X	-
Retained key storage	-	X	-
Tamper-resistant hardware packaging	-	X	X[e]
Designed for FIPS 140-2 Level 4 certification	-	X	-
Supports SSL functions	X	X	X
Supports Linux applications doing SSL handshakes	-	-	X
RSA functions	-	X	X
High performance SHA-1 and SHA2	X	X	-
Clear key DES or triple DES	X	-	-
Advanced Encryption Standard (AES) for 128-bit, 192-bit, and 256-bit keys	X	X	-
Pseudorandom number generator (PRNG)	X	-	-
Clear key RSA	-	-	X
Europay Mastercard VISA (EMV) support	-	X	-
Public Key Decrypt (PKD) support for Zero-Pad option for clear RSA private keys	-	X	X
Public Key Encrypt (PKE) support for MRP function	-	X	X
Remote loading of initial keys in ATM	-	X	-
Improved key exchange with non CCA system	-	X	-
ISO 16609 CBC mode triple DES MAC support	-	X	-

a. Requires CPACF enablement feature code 3863.
b. To make the addition of the Crypto Express features nondisruptive, the logical partition must be predefined with the appropriate PCI Express cryptographic adapter number selected in its candidate list in the partition image profile.
c. One PCHID is required for each PCI-e cryptographic adapter.
d. This is not required for Linux if only RSA clear key operations are used. DES or triple DES encryption requires CPACF to be enabled.
e. This is physically present but is not used when configured as an accelerator (clear key only).

6.7 Software support

The software support levels are listed in 8.4, "Cryptographic support" on page 245.

7

zEnterprise BladeCenter Extension Model 002

IBM has extended the role of the mainframe by adding new infrastructure based on the IBM BladeCenter. It is called the zEnterprise BladeCenter Extension (zBX) Model 002.

The zBX brings computing capacity of systems in blade form-factor to the zEnterprise System. It is designed to provide a redundant hardware infrastructure that supports the multi-platform environment of the zEnterprise System in a seamless integrated way.

Key to the zEnterprise System is also the Unified Resource Manager, which helps deliver end-to-end virtualization and management, as well as the ability to optimize multi-platform technology deployment according to individual workload requirements. For more information on Unified Resource Manager refer to Chapter 13, "Unified Resource Manager" on page 363.

In this chapter we introduce the zBX Model 002 and describe its hardware components. We also explain the basic concepts and building blocks for zBX connectivity.

The information in this chapter can be used for planning purposes and to help define the configurations that best fit your requirements.

This chapter discusses the following topics:

7.1 zBX concepts

IBM zEnterprise System represents a new height for mainframe functionality and qualities of service. It has been rightly portrayed as a cornerstone for the IT infrastructure, especially when flexibility for rapidly changing environments is called for.

IBM zEnterprise System characteristics make it especially valuable for mission critical workloads. Today, most of these applications have multi-tiered architectures that span several hardware and software platforms. However, there are differences in the qualities of service offered by the platforms. There are also various configuration procedures for their hardware and software, operational management, software servicing, failure detection and correction, and so on. These in turn require personnel with several distinct skill sets, several sets of operational procedures, and an integration effort that is not trivial and, therefore, not often achieved. Failure in achieving integration translates to lack of flexibility and agility, which can impact the bottom line.

IBM mainframe systems have been providing specialized hardware and dedicated computing capabilities for a long time. In addition to the machine instruction assists, another example is the vector facility of the IBM 3090. Other such specialty hardware includes the System Assist Processor for I/O handling (that implemented the 370-XA architecture), the Coupling Facility, and the Cryptographic processors. Furthermore, all the I/O cards are specialized dedicated hardware components, with sophisticated software, that offload processing from the System z processor units (PUs).

The common theme with all of these specialized hardware components is their seamless integration within the mainframe. The zBX components are also configured, managed, and serviced the same way as the other components of the System z server. Despite the fact that the zBX processors may not be System z PUs, the zBX is in fact, handled by System z management firmware called the IBM zEnterprise Unified Resource Manager. The zBX hardware features are part of the mainframe, not add-ons.

System z has long been an integrated heterogeneous platform. With zBX, that integration reaches a new level. zBX provides within the System zEnterprise infrastructure a solution for running AIX workloads with IBM POWER®-based blades. Also, zBX provides an optimized solution for running Data Warehouse and Business Intelligence queries against DB2 for z/OS, with fast and predictable response times, while retaining the data integrity, data management, security, availability, and other qualities of service of System z. This solution is known as the IBM Smart Analytics Optimizer.

> **Statement of Direction:** In the first half of 2011, IBM intends to offer both an x86 blade running Linux and a WebSphere DataPower® Appliance for the zBX Model 002.

7.2 zBX hardware description

The zBX has a machine type of 2458-002 and is exclusive to the z196. It is capable of hosting integrated multi-platform systems and heterogeneous workloads, with integrated advanced virtualization management. The zBX Model 002 is configured with the following key components:

► One to four standard 19 inch 42U IBM zEnterprise racks with required network and power infrastructure
► One to eight BladeCenter chassis[1] with a combination of up to 112 different blades
► Redundant infrastructure for fault tolerance and higher availability
► Management support through the z196 Hardware Management Console (HMC) and Support Element (SE)

The zBX can be ordered with a new z196 or as an MES to an existing z196. Either way, the zBX is treated as an extension to a z196 and cannot be ordered as a standalone feature.

Figure 7-1 shows a z196 with a maximum zBX configuration. The first rack (Rack B) in the zBX is the primary rack where one or two BladeCenter chassis and four top of rack (TOR) switches reside. The other three racks (C, D, and E) are expansion racks with one or two BladeCenter chassis each.
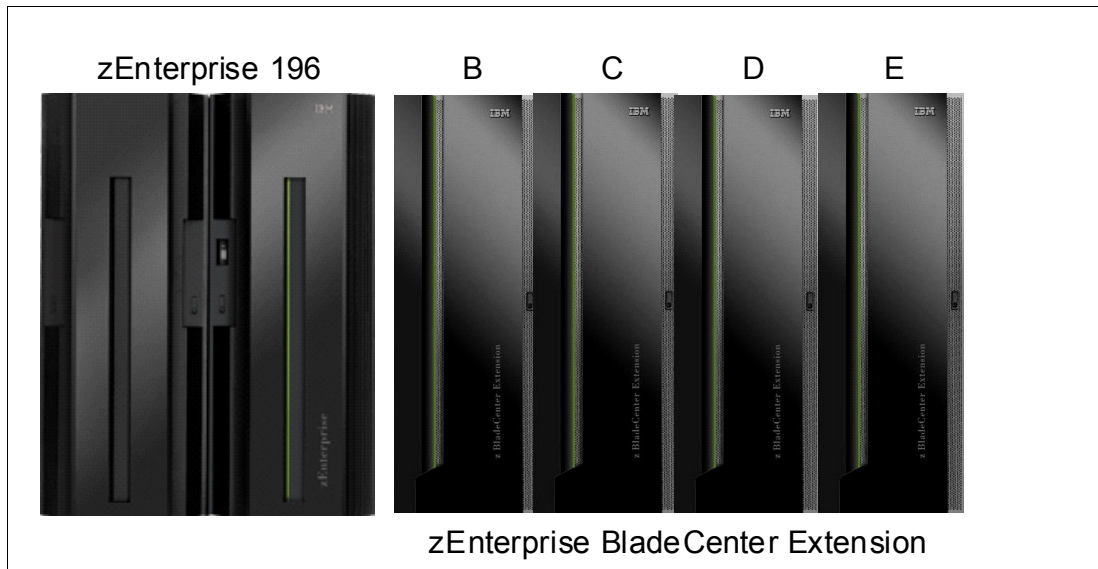
Figure 7-1   z196 with a maximum zBX configuration

7.2.1 zBX racks

The zBX Model 002 (2458-002) hardware is housed in up to four IBM zEnterprise racks. Each rack is 42U high (industry-standard 19") and has four sidewall compartments to support installation of power distribution units (PDUs) and switches, with additional space for cable management.

Figure 7-2 on page 184 shows the rear view of a two rack zBX configuration. The racks include:

---

[1] The IBM Smart Analytics Optimizer solution has a maximum of two zBX racks (B and C) and up to four BladeCenter chassis

- ► Two TOR 1000BASE-T switches (Rack B only) for the intranode management network (INMN)
- ► Two TOR 10 GbE switches (Rack B only) for the intraensemble data network (IEDN)
- ► Up to two BladeCenter Chassis in each rack with:
  - – Up to 14 blades[1] (POWER7 or x86[2])
  - – Advance management modules (AMM)
  - – Ethernet switch modules (ESM)
  - – High speed switch (HSS) modules
  - – 8 Gbps Fiber Channel switches for connectivity of client supplied disks
  - – Blower modules
- ► Power Distribution Units (PDUs)
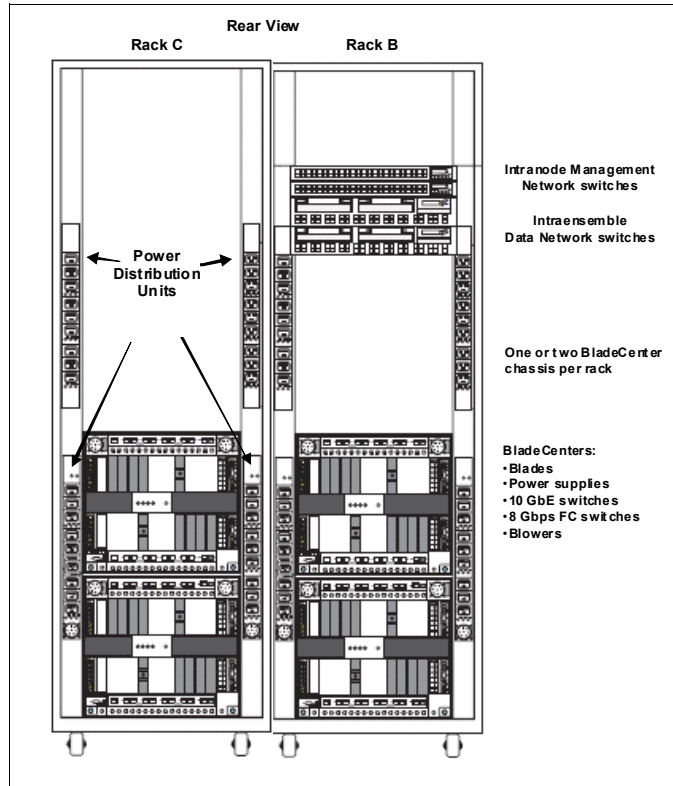
Figure 7-2   zBX racks - rear view with BladeCenter chassis

A zBX rack can support a maximum of two BladeCenter chassis. Each rack is designed for enhanced air flow and is shipped loaded with the initial configuration. It can be upgraded on-site to a larger configuration.

The zBX racks are shipped with lockable standard non-acoustic doors and side panels. The following optional features are also available:

- ► IBM rear door heat eXchanger (FC 0540) reduces the heat load of the zBX emitted into ambient air. The rear door heat eXchanger is an air-to-water heat exchanger that diverts the waste heat of the zBX to chilled water. The rear door heat eXchanger requires external conditioning units for its use. For more information, refer to: http://www.rdhxinfo.com

- ► IBM acoustic door (FC 0543) can be used to reduce the acoustical noise from the zBX.

---

[1] Depending on the IBM Smart Analytics Optimizer configuration, this can be either 7 or 14 blades in the first BladeCenter chassis.

[2] In the first half of 2011, IBM intends to offer an x86 blade running Linux in the IBM zEnterprise System on zBX Model 002

► Height reduction (FC 0570) reduces the rack height to 36U high and accommodates doorway openings as low as 1832 mm (72.1 inches). It should be ordered if you have doorways with openings less than 1941 mm (76.4 inches) high.

7.2.2 Top of rack (TOR) switches

The four top-of-rack (TOR) switches are installed in the first rack (Rack B). Adding expansion racks (Rack C, D, and E) do not require additional TOR switches.

The TOR switches are located near the top of the rack and are mounted from the rear of the rack. From top to bottom are two 1000BASE-T switches for the intranode management network (INMN) and two 10 GbE switches for the intraensemble data network (IEDN).

A zBX can only be managed by one z196 via the INMN connections. Each VLAN-capable 1000BASE-T switch has 48 ports. The switch ports are reserved as follows:

► One port for each of the two bulk power hubs (BPH) on the owning z196
► One port for each of the advance management modules (AMM) and Ethernet switch modules (ESM), in each zBX BladeCenters chassis
► Two management ports, one port for each of the two IEDN 10 GbE TOR switches.
► One port each for interconnecting the two switches

Both switches have the same connections to the corresponding redundant components (the BPH, AMM, ESM and IEDN TOR switch) to avoid any single point of failure.

> **Important:** Although IBM provides a 26m cable for the INMN connection, it is recommended that the zBX is installed next to or near the *owning* z196 server, for easy access to the zBX for service related activities or tasks.

Each VLAN-capable 10 GbE TOR switch has 40 ports dedicated to the IEDN. The switch ports have the following connections:

► Up to eight ports for connections to a HSS module of each BladeCenter chassis in the same zBX (as part of IEDN), to provide data paths to blades
► Up to eight ports for OSA-Express3 10 GbE (LR or SR) connections to the ensemble CPCs (as part of IEDN), to provide data paths between the ensemble CPCs and the blades in a zBX
► Up to seven ports for zBX to zBX connections within a same ensemble (as part of the IEDN)
► Up to eight ports for the customer managed data network. These connections are not part of IEDN, and cannot be managed or provisioned by the Unified Resource Manager. The Unified Resource Manager will recognize them as migration connections and provide access control for their connection to the 10 GbE TOR switches.
► Two ports are used for interconnections between two switches as a failover path
► One port is the management port connected to INMN 1000BASE-T TOR switch
► Two Direct Attached Cables (DAC) interconnect both switches.

For more information about the connectivity options for the INMN and the IEDN, as well as the connectivity rules, can be found in 7.4, "zBX connectivity" on page 190.

7.2.3 zBX BladeCenter chassis

In keeping with the System z QoS, each zBX BladeCenter chassis is designed with additional components installed for high levels of resiliency.

The front of a zBX BladeCenter chassis has the following components:

► **Blade server slot**

There are 14 blade server slots (BS01 to BS14) available in a zBX BladeCenter chassis. Each slot is capable of housing any zBX supported blades. In the future, some supported blade types will be double-wide. Slot 14 cannot hold a double-wide blade.

Blades should be sequentially plugged into each BladeCenter (first slot and go linearly out, slots should not be skipped).

► **Power module**

The power module includes a power supply and a three pack of fans, two of three fans are needed for a power module operation. Power module 1 and 2 (PM01 and PM02) are installed as a pair to provide power supply for the seven blade server slots from BS01 to BS07, and power module 3 and 4 (PM03 and PM04) support the BS08 to BS14.

In Figure 7-3 on page 186, the two colors indicate different power sources for the power modules (PM) and blade server slots. PM01 and PM04 are connected to power source 1, while PM02 and PM03 are connected to power source 2. Thus each blade server slot can have a fully redundant power supply from a different power module connected to a different power source.

Figure 7-3 on page 186 shows the rear view of a zBX BladeCenter chassis.
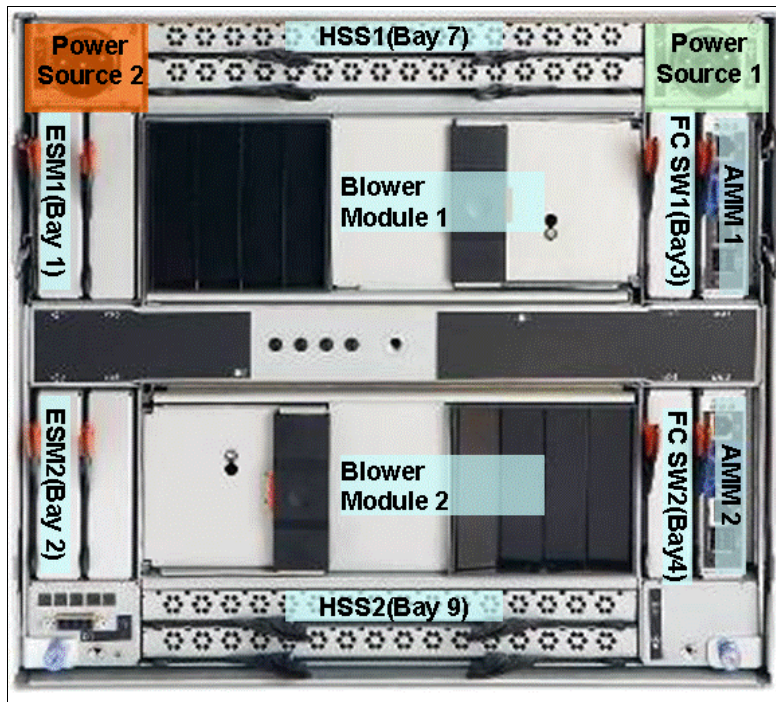
Figure 7-3   zBX BladeCenter chassis rear view

The rear of a zBX BladeCenter chassis has following components:

► **Advance management module**

Advance management module (AMM) provides systems-management functions and keyboard/video/mouse (KVM) multiplexing for all of the blade servers in the BladeCenter unit that support KVM. It controls the external keyboard, mouse, and video connections, for use by a local console, and a 10/100 Mbps Ethernet remote management connection.

The management module communicates with all components in the BladeCenter unit, detecting their presence or absence, reporting their status, and sending alerts for error conditions when required.

The service processor in the management module communicates with the service processor (iMM) in each blade server to support features such as blade server power-on requests, error and event reporting, KVM requests, and requests to use the BladeCenter shared media tray.

The AMMs are connected to the INMN through the 1000BASE-T TOR switches. Thus, firmware and configuration for the AMM is controlled by the SE on owning z196, with all the service management and reporting function of AMMs.

Two AMMs (AMM1 and 2) are installed in the zBX BladeCenter chassis. Only one AMM has primary control of the chassis, the second module is in passive or standby mode. If the active or primary module fails, the second module is automatically enabled with all of the configuration settings of the primary module.

► **Ethernet switch module**

Two 1000BASE-T (1 Gbps) Ethernet switch modules (ESM1 and ESM2) are installed in switch bay 1 and 2 in the chassis. Each ESM has 14 internal full-duplex Gigabit ports, one connected to each of the blade servers in the BladeCenter chassis, two internal full-duplex 10/100 Mbps ports connected to the AMM modules, and six 1000BASE-T copper RJ-45 connections for INMN connections to the TOR 1000BASE-T switches.

The ESM port 00 is connected to one of the 1000BASE-T TOR switches. As part of the INMN, configuration and firmware of ESM is controlled by the owning z196 Support Element (SE).

► **High speed switch module**

Two high-speed switch modules (HSS1 and 2) are installed to the switch bay 7 and 9. The HSS provide 10 GbE uplinks the 10 GbE TOR switches and 10 GbE downlinks to the blades in the chassis.

Port 01 is connected to one of the 10 GbE TOR switch. Port 10 is used to interconnect HSS in bay 7 and 9 as a failover path.

► **8 Gbps Fiber Channel switch module**

Two 8 Gbps Fibre Channel (FC) switches (FC SW1 and 2) are installed in switch bays 3 and 4. Each switch has 14 internal ports reserved for the blade servers in the chassis, and six external fiber channel ports to provide connectivity to client supplied external disk for Smart Analytics Optimizer or blade solutions.

For the Smart Analytics Optimizer solutions, ports 00 and 15 of FC SW1 and FC SW2 in the first two Smart Analytics Optimizer chassis are used to connect the client provided FC disk. Ports 16 to 19 are reserved for cascade connections from other Smart Analytics Optimizer chassis to access the Smart Analytics Optimizer FC disk.

For blade solutions, ports 00 and 15 of FC SW1 and FC SW2 can provide connections to client provided FC storage disks for the blade servers in the same chassis.

► **Blower module**

There are two hot swap blower modules installed. The blower speeds vary depending on the ambient air temperature at the front of the BladeCenter unit and the temperature of internal BladeCenter components. And if a blower fails, the remaining blowers will run full speed.

► **BladeCenter mid-plane fabric connections**

The BladeCenter mid-plane provide redundant power, control and data connections to a blade server by internally routed chassis components (power modules, AMMs, switch

modules, media tray) to connectors in a blade server slot. There are six connectors in a blade server slot on the mid-plane, from top to bottom:

– Top 1X fabric connects blade to AMM1, ESM1 and FC SW1
– Power connector from power module 1 (slot 1 to 7) or power module 3 (slot 8 to 14)
– Top 4X fabric connects blade to HSS1
– Bottom 4X fabric connects blade to HSS2
– Bottom 1X fabric connects blade to AMM2, ESM2 and FC SW2
– Power connector from power module 2 (slot 1 to 7) or power module 4 (slot 8 to 14)

Thus, each blade server has redundant power, data and control links from separate components.

7.2.4 zBX blades

The zBX Model 002 currently supports two types of blades:

► Up to 56 IBM Smart Analytics Optimizer blades

The number of Smart Analytics Optimizer blades depends on the Smart Analytics Optimizer solution size. Five Smart Analytics Optimizer solution sizes are available, corresponding to quantities of 7, 14, 28, 42, and 56 Smart Analytics Optimizer blades.

► Up to 112 POWER7 blades

Three configurations of POWER blades are supported, depending on their memory size (see Table 7-1 on page 189). The number of blades can be from 1 to 112. There is a Statement of Direction (SoD)[1] that support for IBM x86 blades and IBM WebSphere DataPower integration blades will be added at a later time.

For a zBX configuration, Smart Analytics Optimizer blades can be separated in non-adjacent chassis, but cannot share a chassis with other types of blades. However, mixed installation of POWER7, x86, and DataPower blades in a chassis will be supported.

All zBX blades are connected to AMMs and ESMs through the chassis mid-plane. The AMMs are connected to the INMN.

zBX blade expansion cards

Each zBX blade has two PCI Express connectors, combination input output vertical (CIOv) and combination form factor horizontal (CFFh). I/O expansion cards are attached to these connectors and connected to the mid-plane fabric connectors. Thus a zBX blade can expand its I/O connectivity via the mid-plane to the high speed switches and switch modules in the chassis.

Depending on the blade type, 10 GbE CFFh expansion cards, 1GbE CFFh expansion cards and 8 Gbps Fibre Channel CIOv expansion cards provide I/O connectivity to the IEDN, the INMN or client supply FC storage disks.

IBM Smart Analytics Optimizer blade

IBM Smart Analytics Optimizer blade uses the x86 technology and has the following specifications:

► Two 2.93 GHz quad core processors
► Twelve 4 GB dual in-line memory modules (DIMMs) giving a total of 48 GB (24 GB for each quad core processor)
► A 32 GB Solid® State Drive (SDD)

---

[1] All statements regarding IBM future direction and intent are subject to change or withdrawal without notice, and represents goals and objectives only.

▶ 2-port 10 GbE CFFh expansion card for IEDN connection
▶ 8 Gbps Fiber Channel CIOv expansion card for client provide DS5020 connection

POWER7 blade

The POWER7 blade is a single width blade, which includes a POWER7 processor, up to 16 DIMMs, and a HDD. The POWER7 blade supports 10 GbE connection to IEDN, and 8 Gbps FC connections to client provide Fibre Channel storage via the FC SW1 and FC SW2 in the chassis.

The POWER7 blade is loosely integrated to a zBX, so that you can acquire supported blades through existing channels or IBM. The primary HMC and SE of the owning z196 perform entitlement management for installed POWER7 blades on a one-blade basis.

Table 7-1   Supported configuration of POWER7 blades

	Config 1	Config 2	Config 3
Processors	3.0GHz@150W	3.0GHz@150W	3.0GHz@150W
DIMMs	32 GB (8 x 4 GB)	64 GB (16 x 4 GB)	128 GB (16 x 8 GB)
Internal Disks	1 x 300 GB HDD	1 x 300 GB HDD	1 x 300 GB HDD
CFFh I/O expansion	10 GbE	10 GbE	10 GbE
CIOv I/O expansion	8 Gbps FC	8 Gbps FC	8 Gbps FC

7.2.5 Power distribution unit (PDU)

The power distribution units (PDUs) provide the connection to the main power source, the power connection to the intranode management network and intraensemble data network top of rack switches, and the power connection to the BladeCenter. The number of power connections needed is based on the zBX configuration. A rack contains two PDUs if one BladeCenter is installed or four PDUs if two BladeCenters are installed.

7.3 zBX entitlements and firmware

When ordering a zBX, the owning z196 server will have the entitlements feature for the IBM Smart Analytics Optimizer blade (FC 0610) and/or the POWER7 Blade (FC0612). The entitlements are similar to a high water mark or max purchased flag and only a blade quantity equal to or less than installed in the zBX can communicate with the CPC.

Also, Unified Resource Manager has two management suites, Manage suite (FC 0019) and Automate suite (FC 0020).

If owning z196 server has Manage suite (FC 0019), then the same quantity entered for FC 0610 or FC 0612 will be used for Manage Firmware IBM Smart Analytics Optimizer (FC 0039) and/or Manage Firmware POWER7 blade (FC 0041).

If owning z196 server has Automate Firmware suite (FC 0020), then the same quantity entered for FC 0610 will be used for Manage Firmware IBM Smart Analytics Optimizer (FC 0039) and Automate Firmware IBM Smart Analytics Optimizer (FC 0043). The same quantity entered for FC 0612 will be used for Manage Firmware POWER blade (FC 0041) and Automate Firmware POWER blade (FC 0045).

The minimum quantity for FC 0039 and FC0043 is 7 and the maximum is 56. The minimum quantity for FC 0041 and FC0045 is one and the maximum is 112. FC 0039, FC0041, FC0043, and FC0045 are priced feature.

> **Note:** If any attempt is made to install additional blades that exceed the FC 0610 or FC 0612 count, those blades will be not be powered on by the system. The blades will also be checked for minimum hardware requirements.

7.3.1 zBX management

One key feature of the zBX is its integration under the System z management umbrella. Thus, initial firmware installation as well as updates and patches follow the already familiar pattern of System z. The same reasoning applies to the configuration and definitions.

Similar to channels and processors, the SE has a view for the zBX blades. This view shows icons for each of the zBX component's objects including an overall status (power, operational, and so on).

The following functions and actions are managed and controlled from the z196 HMC/SE:

- ▶ View firmware information for the BladeCenter and blades
- ▶ Retrieve firmware changes
- ▶ Change firmware level
- ▶ Backup/restore critical data
  - – zBX configuration data is backed up as part of System z196 SE backup and restored on replacement of a blade.

For more details refer to Chapter 13, "Unified Resource Manager" on page 363.

zBX firmware

The firmware for the zBX is managed, controlled, and delivered in the same way as for the z196 server. It is packaged and tested with System z microcode and changes are supplied and applied with MCL bundle releases.

Benefits of the zBX firmware packaged with System z microcode:

- ▶ Tested together with System z driver code and MCL bundle releases
- ▶ Retrieve code as same integrated process of System z (IBM RETAIN® or media)
- ▶ No need to use separate tools and connect to Web sites to obtain code
- ▶ Utilize new upcoming System z firmware features, such as Digitally Signed Firmware
- ▶ Infrastructure incorporates System z concurrency controls where possible
- ▶ zBX firmware update fully concurrent, blades similar to Config Off/On controls
- ▶ Audit trail of all code changes in security log
- ▶ Automatic back out of changes to previous working level on code apply failures
- ▶ Optimizer firmware

The IBM Smart Analytics Optimizer application, supporting operating system and management agent code, released as firmware with the rest of the code, is automatically downloaded (if necessary) from DB2 on the first connection.

7.4 zBX connectivity

There are three types of LANs (each with redundant connections) that attach to the zBX: the INMN, the IEDN, and the customer managed data network. The INMN is fully isolated and

only established between the owning z196 server and the zBX. The IEDN connects the zBX to a maximum of eight z196 servers. Each z196 server must have a minimum of two connections to the zBX. The IEDN is used to connect a zBX to a maximum of seven other zBXs. The IEDN is a VLAN-capable network that allows enhanced security by isolating data traffic between virtual servers.

Figure 7-4 shows a high-level summary of the connectivity required for the zBX environment. It shows the z196 connections through two OSA-Express3 1000BASE-T features (CHPID type OSM) to the INMN TOR switches. The OSA-Express3 10 GbE features (CHPID type OSX) connect to the two IEDN TOR switches. Depending on workload requirements, any OSA-Express2 or OSA-Express features (CHPID type OSD) can connect to the customer managed data network.

The Fibre Channel (FC) connections are only required between the zBX and the attached Fibre Channel disk or storage area network (SAN).

> **Note:** It is the client's responsibility to supply the cables for the IEDN, the customer managed network, and the connection between the zBX and the FC disk.
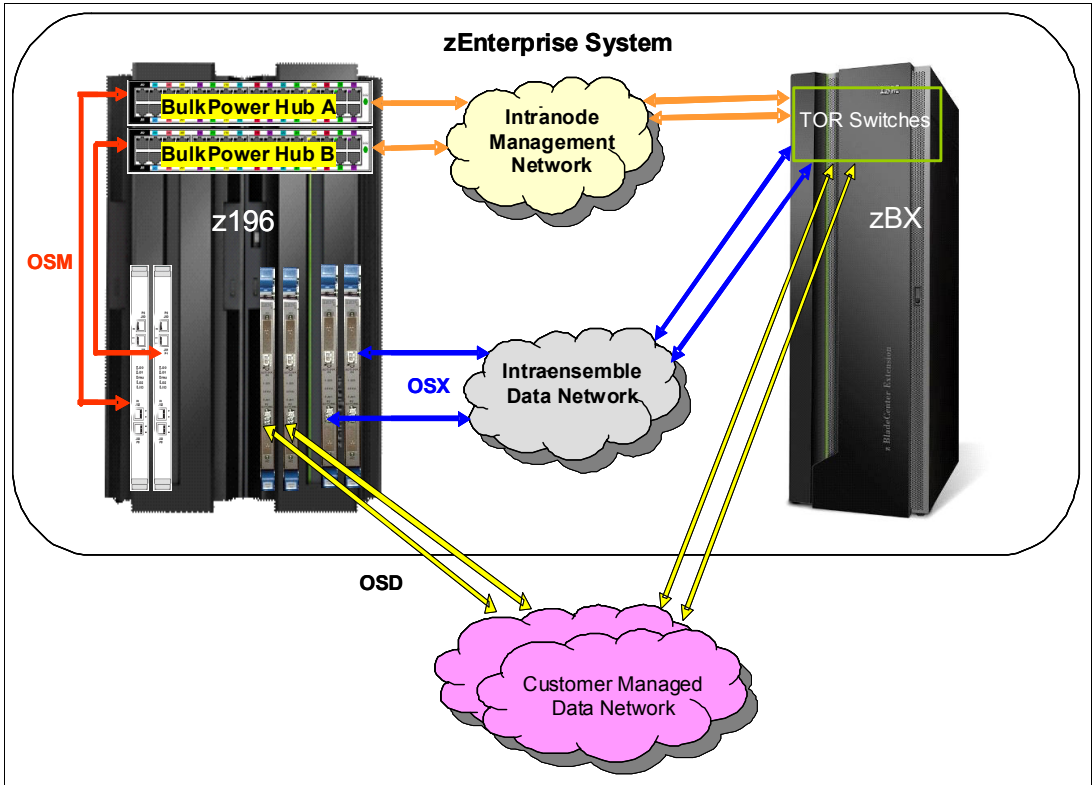
Figure 7-4   INMN, IEDN, customer managed data network, and Fibre Channel connectivity

The IEDN provides private and secure 10 GbE high speed data paths between all elements of a zEnterprise ensemble (up to eight z196s with optional zBXs).

The zBX is managed by the HMC through the physically isolated INMN, which interconnects all resources of the zEnterprise System (z196 and zBX components).

7.4.1 Intranode management network

The scope of the intranode management network (INMN) is within an ensemble *node*. A node consists of a z196 CPC and its optional zBX. INMNs in different nodes are not connected to each other. The INMN connects the Support Element (SE) of the z196 to the hypervisor, optimizer, and guest management agents within the node. Communication across the INMN is exclusively for the purpose of enabling the Unified Resource Manager of the HMC to perform its various management disciplines (for example, performance management, network virtualization management, or energy management) for the node. The z196 connection to the INMN is achieved through the definition of a CHPID type OSM, which can be defined over an OSA-Express3 1000BASE-T Ethernet feature. There is also a 1 GbE infrastructure within the zBX.

INMN configuration

The key points to consider for an INMN are:

► Each z196 server must have two OSA-Express3 1000BASE-T ports connected to the Bulk Power Hub in the same z196:

– The two ports provide a redundant configuration for failover purposes in case one link fails.

– For availability, each connection should be from two different OSA-Express3 1000BASE-T features within the same z196 server.

Figure 7-5 shows the OSA-Express3 1000BASE-T feature and required cable type.
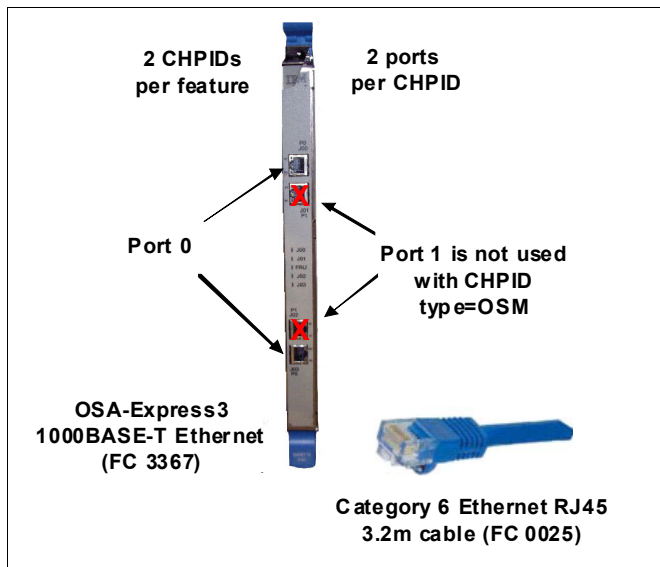
Figure 7-5   OSA-Express3 1000BASE-T feature and cable type

► OSA-Express3 1000BASE-T ports can be defined in the IOCDS as SPANNED, SHARED, or DEDICATED:

– DEDICATED, restricts the OSA-Express3 1000BASE-T port to a single LPAR

– SHARED allows the OSA-Express3 1000BASE-T port to be used by all or selected LPARS in the same z196 server

– SPANNED allows the OSA-Express3 1000BASE-T port to be used by all or selected LPARS across multiple Channel Subsystems in the same z196 server

– SPANNED and SHARED port can be restricted by the PARTITION keyword in the CHPID statement to allow only a subset of LPARs in the z196 server to use the OSA-Express3 1000BASE-T port

– OSA-Express3 1000BASE-T ports should be defined as SPANNED for all the ensemble member LPARs

– SPANNED, SHARED, and DEDICATED links pairs can be defined within the maximum of 16 links supported by the zBX

► z/OS Communication server TCPIP stack must be enabled for IPv6. The CHPID type OSM related definitions will be dynamically created.

– No IPv4 address is needed, a IPv6 link local address will be dynamically applied

► z/VM virtual switch types provide INMN access:

– Up-link can be virtual machine NIC
– Ensemble membership conveys UUID and MAC prefix

► Two 1000BASE-T top of rack switches in the zBX (Rack B) are used for the INMN, no additional 1000BASE-T Ethernet switches are required. Figure 7-6 shows the 1000BASE-T TOR switches.

Figure 7-6   Two 1000BASE-T TOR switches

The port assignments for both 1000BASE-T TOR switches are listed in Table 7-2.

Table 7-2   Port assignments for the 1000BASE-T TOR switches

Ports	Description
J01-J03	Management for BladeCenters located in zBX Rack-B
J04-J07	Management for BladeCenters located in zBX Rack-C
J08-J11	Management for BladeCenters located in zBX Rack-D
J12-J15	Management for BladeCenters located in zBX Rack-E
J16-J43	not used
J44-J45	INMN switch B36P(Top) to INMN switch B35P(Bottom)
J46	INMN-A to IEDN-A port J41 / INMN-B to IEDN-B port J41
J47	INMN-A to z196 BPH-A port J06 / INMN-B to z196 BPH-B port J06

► 1000BASE-T supported cable:

– 3.2 meter Category 6 Ethernet cables are shipped with the z196 ensemble management flag feature (FC 0025). Those cables connect the OSA-Express3 1000BASE-T ports to the Bulk Power Hubs (port 7).

– 26 meter Category 5 Ethernet cables are shipped with the zBX. Those cables are used to connect the z196 Bulk Power Hubs (port 6) and the zBX top of rack switches (port J47).

7.4.2 Primary and alternate HMCs

The zEnterprise System Hardware Management Console (HMC) that has management responsibility for a particular zEnterprise ensemble is called a primary HMC. Only one primary HMC is active for a given ensemble. This HMC has an alternate HMC to provide redundancy; it is not available for use until it becomes the primary HMC in a failover situation. To manage ensemble resources, the primary HMC for that ensemble must be used. A primary HMC can of course perform all HMC functions.

An HMC network configuration consists of the following:

► Primary and alternate HMCs must be connected to the same LAN segment

– The z196 can be on a different LAN segment

► One pair of HMCs (primary and alternate) is required per ensemble

– The primary HMC can control up to eight z196 with one zBX each
– The two HMCs must be on the same LAN segment

► The primary and alternate HMCs use Category 6 Ethernet cables

– Maximum distance is 100 meters, but can be extended with an Ethernet switch

► An additional 16-port Ethernet switch (FC 0070) can be ordered for the HMC network

► zEnterprise ensemble can be configured without zBX hardware (using only z196 nodes)

Figure 7-7 shows the primary and alternate HMC configuration connecting into the two bulk power hubs (BPHs) in the z196. The 1000BASE-T TOR switches in the zBX is also connected to the BPHs in the z196.

> **Note:** All ports on the z196 BPH are reserved for specific connections. Any deviations or mis-cabling will affect the operation of the z196 system.
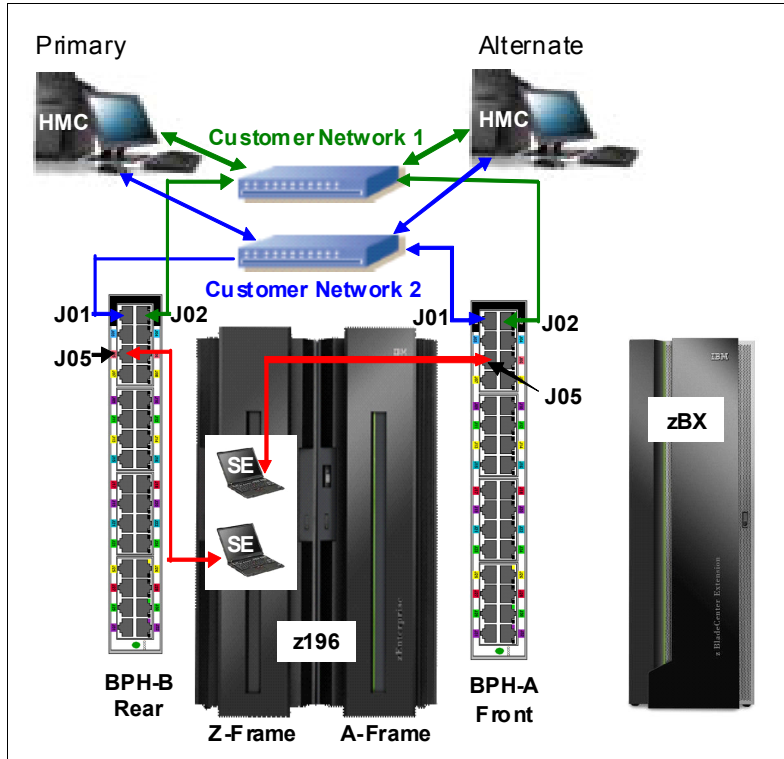
Figure 7-7   HMC configuration in an ensemble node

Table 7-3 lists the port assignments for both bulk power hubs (BPHs).

Table 7-3   Port assignments for the BPHs

Ports	Description
J01	Customer Network 2 - HMC to SE Network
J02	Customer Network 1 - HMC to SE Network
J03	BPH-A to BPH-B
J04	BPH-A to BPH-B
J05	BPH-A Top SE / BPH-B Bottom SE
J06	zBX TOR INMN-A / TOR INMA-B (port 47)
J07	OSA-Express3 1000BASE-T (CHPID type OSM)
J08	unused
J09-J32	Used for internal z196 components

Refer to Chapter 12, "Hardware Management Console" on page 343 and Chapter 13, "Unified Resource Manager" on page 363 for more information.

7.4.3 Intraensemble data network

The intraensemble data network (IEDN) is the main application data path provisioned and managed by the Unified Resource Manager of the owning z196. Data communications for ensemble-defined workloads flow over the IEDN between nodes of an ensemble. All of the

physical and logical resources of the IEDN are configured and managed by the Unified Resource Manager. The IEDN extends from the z196 through the OSA-Express3 10 GbE ports when defined as CHPID type OSX. The minimum number of OSA-Express3 10 GbE features is two per z196 server. Similarly a 10 GbE networking infrastructure within the zBX is used for IEDN access.

IEDN configuration

The IEDN connections can be configured in a number of ways. The key points to consider for an IEDN are:

▶ Each z196 server must have a minimum of two OSA-Express3 10 GbE ports connected to the zBX via the IEDN:

– The two ports provide a redundant configuration for failover purposes in case one link fails.
– For availability, each connection should be from two different OSA-Express3 10 GbE features within the same z196 server.
– The zBX can have a maximum of 16 IEDN connections (8 pairs of OSA-Express3 10 GbE ports).

Figure 7-8 shows the OSA-Express3 10 GbE feature (long reach or short reach) and the required fiber optic cable types.
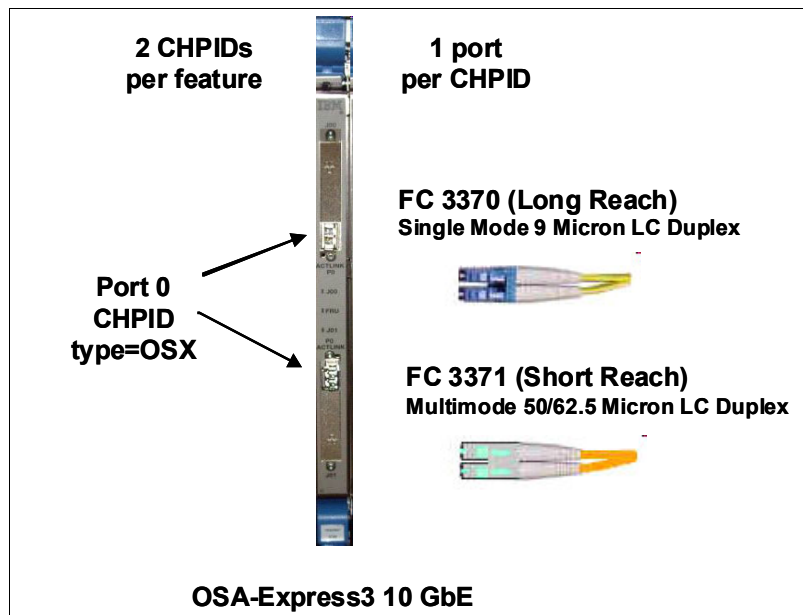
Figure 7-8   OSA-Express3 10 GbE feature and cable

▶ OSA-Express3 10 GbE ports can be defined in the IOCDS as SPANNED, SHARED, or DEDICATED:

– DEDICATED restricts the OSA-Express3 10 GbE port to a single LPAR
– SHARED allows the OSA-Express3 10 GbE port to be used by all or selected LPARS in the same z196 server
– SPANNED allows the OSA-Express3 10 GbE port to be used by all or selected LPARS across multiple Channel Subsystems (CSSs) in the same z196 server
– SHARED and SPANNED ports can be restricted by the PARTITION keyword in the CHPID statement to allow only a subset of LPARs on the z196 server to use OSA-Express3 10 GbE port

- – OSA-Express3 10 GbE ports should be defined as SPANNED for all the ensemble member LPARs
- – SPANNED, SHARED, and DEDICATED links pairs can be defined within the maximum of 16 links supported by the zBX

► z/OS Communication Server requires minimal configuration

- – IPv4 or IPv6 addresses
- – VLAN must be configured to match HMC (zManager) configuration

► z/VM virtual switch types provide IEDN access

- – Up-link can be virtual machine NIC
- – Ensemble membership conveys Ensemble UUID and MAC prefix

► IEDN network definition are completed from the primary HMC —"Manage Virtual Network" task.

► Two 10 GbE top of rack switches in the zBX (Rack B) are used for the IEDN, no additional Ethernet switches are required. Figure 7-9 on page 197 shows the 10 GbE TOR switches.
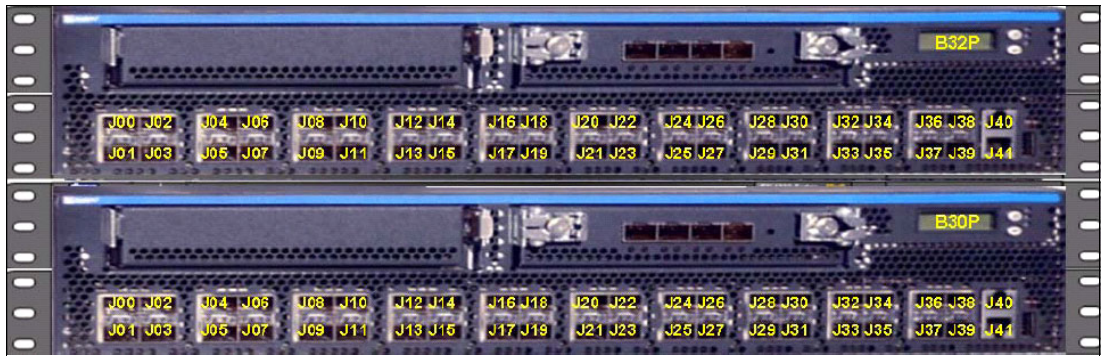
Figure 7-9   Two 10 GbE TOR switches

The port assignments for both 10 GbE TOR switches are listed in Table 7-4.

Table 7-4   Port assignments for the 10 GbE TOR switches

Ports	Description
J00 - J07	SFP and reserved for z196 (OSX) IEDN connections
J08 - J21[1]	DAC reserved for BladeCenter SM07/SM09 IEDN connections
J22 / J23	DAC for TOR switch-to-TOR switch IEDN communication
J24 - J30	SFP reserved for zBX-to-zBX IEDN connections
J31 - J39	SFP reserved for client IEDN connections
J40	RJ-45 (not used)
J41	RJ-45 IEDN Switch Management Port to INMN TOR switch port 46

1. Only eight of the 14 ports are are currently used

► All IEDN connections must be point to point to the 10 GbE switch

- – IEDN connection uses MAC address, not IP address (Layer 2 connection)
- – No additional switches or routers are needed
- – This limits the distances the CPCs can be from the 10 GbE switches in an ensemble

► The 10 GbE TOR switches utilize small form factor plugable (SFP) optics for the external connections and direct attach cables (DAC) for connections as follows:

Ports J00-J07 are reserved for the z196 OSX IEDN connections. These ports utilize SFPs (Small Form Factor Pluggable Modules) plugged according to the zBX order.

– FC 0632 LR SFP to FC3370 OSA-Express3 10 GbE LR
– FC 0633 SR SFP to FC3371 OSA-Express3 10 GbE SR

Ports J08-J23 are reserved for IEDN to BladeCenter attachment. The cables used are Direct Attached Cables (DAC) and are included with the zBX. These are hard wired 10 GbE SFP cables. The feature codes indicate the length of the cable:

– FC 0626 - 1m for Rack B BladeCenters and IEDN to IEDN
– FC 0627 - 5m for Rack C BladeCenter
– FC 0628 - 7m for Racks D and E BladeCenters

► 10 GbE fiber optic cable types and maximum distance:

– Client provides all IEDN cables (except for zBX internal connections)

– Multimode fiber:

• 50 micron fiber at 2000 MHz-km: 300 meters (984 feet)
• 50 micron fiber at 500 MHz-km: 82 meters (269 feet)
• 62.5 micron fiber at 200 MHz-km: 33 meters (108 feet)

– Single mode fiber:

• 10km (6.2 miles)

7.4.4 Network connectivity rules with zBX

The network connectivity rules for interconnecting a zBX are as follows:

► Only one zBX is allowed per owning z196 server

► The zBX can be installed next to the owning z196 server or within the limitation of the 26 meter cable

► Although z10 servers do not support CHPID type OSX, a z10 can attach to the zBX (2458-002) with OSA connections (CHPID type OSD), and can access the IBM Smart Analytics Optimizer solution

► Customer managed data networks are outside the ensemble. A customer managed data network is connected with:

– CHPID type OSD from z196
– IEDN TOR switch ports J32 to J39 from zBX

► Ensemble nodes must be located in the same site

7.4.5 Network security considerations with zBX

The private networks involved in connecting the z196 to the zBX are constructed with extreme security in mind, for example:

► The INMN is entirely private and can be accessed only through the HMC, via its connection to the SE. (standard HMC security still applies). There are also additions to Unified Resource Manager "role-based" security, so that not just any user can reach the Unified Resource Manager panels even if that user can perform other functions of the HMC. Very strict authorizations for users and programs control who is allowed to take advantage of the INMN.

► The INMN network is using "link-local" IP addresses. "Link-local" addresses are not advertised and are accessible only within a single LAN segment. There is no routing in this network, as it is a "flat network" with all Virtual Servers residing on the same IPv6 network.

The SE communicates through the Unified Resource Manager with Virtual Servers over the INMN; even Virtual Servers that reside on the INMN cannot communicate with each other over this internal network; they can only communicate with the SE.

► Only authorized programs or agents can take advantage of the INMN; currently the Performance Agent can to do so. However, there could be other platform management applications in the future, these must be authorized to access the INMN.

► The IEDN is built on a flat network design (same IPv4 or IPv6 network) and each server accessing the IEDN must be an authorized Virtual Server and must belong to an authorized Virtual LAN within the physical IEDN. VLAN enforcement sits within the hypervisor functions of the ensemble - controls reside in the OSA (CHPID type OSX), in the z/VM VSWITCH, and in the VSWITCH hypervisor function of the blades on the zBX. The VLAN IDs and the Virtual MACs that are assigned to the connections from the Virtual Servers are tightly controlled through the Unified Resource Manager and thus there is no chance of either MAC or VLAN spoofing for any of the servers on the IEDN. If a you decide to attach to the TOR switches of the zBX in order to send data to Virtual Servers on the zBX blade, the permitted VLAN IDs and VMACs must be authorized in the TOR switches. Although the TOR switches will enforce the VMACs and VLAN IDs here, you must take the usual network security measures to ensure that the attaching devices in the Customer Managed Data Network are not subject to MAC or VLAN spoofing; the Unified Resource Manager functions cannot control the assignment of VLAN IDs and VMACs in those devices. In other words, whenever you decide to interconnect the external network to the secured IEDN, the security of that external network should involve all the usual layers of the IBM Security Framework: physical security, platform security, application and process security, data and information security, and so on.

► The INMN and the IEDN are both subject to Network Access Controls as implemented in z/OS and in z/VM so that not just any Virtual Server on the z196 can utilize these networks.

► Although we deem it unnecessary to implement firewalls, IP filtering, or encryption for data flowing over the IEDN, if company policy or security mandates require such measures to be taken, then these are supported. You could implement any of the security technologies available: for example, SSL/TLS, or IP Filtering.

► The centralized and internal network design of both the INMN and the IEDN limit the scope of vulnerability to security breaches. Both networks reduce the amount of network equipment and processes and routing hops that are under the control of multiple individuals and subject to security threats. Both require the use of IBM-only equipment (switches, blades) which have been tested previously and in some cases pre-installed.

In summary, many technologies are architected in a more robust, secure fashion in the client network than they have been implemented in the past. This is because of either their implementation through the Unified Resource Manager, or because of additional SAF controls specific to zEnterprise System and the ensemble, such as:

– MAC filtering
– VLAN enforcement
– ACCESS control
– Role-based security
– The following standard security implementations are still available for use in the IEDN:
   • Authentication
   • Authorization and access control (including Multilevel Security (MLS); also firewall IP filtering. Only stateless firewalls or IP filtering implementations can be installed in a Virtual Server in the ensemble)
   • Confidentiality
   • Data integrity
   • Non-repudiation

7.4.6 zBX storage connectivity

Each BladeCenter chassis in the zBX has two 20-port 8 Gbps Fibre Channel (FC) switch modules. Each switch has 14 internal ports and six shortwave (SX) external ports. The internal ports are reserved for the blades in the chassis. Figure 7-10 on page 200 shows the external ports image.

Those external ports are used to connect the client provided external Fibre Channel disk (two ports—port 0 and port 15) and cascaded connection between BladeCenter chassis (four ports 16 to 19) to access the Smart Analytics Optimizer FC disk. Note that cascaded connections between BladeCenter chassis is only required for IBM Smart Analytics Optimizer solutions with 3 or 4 BladeCenter chassis.
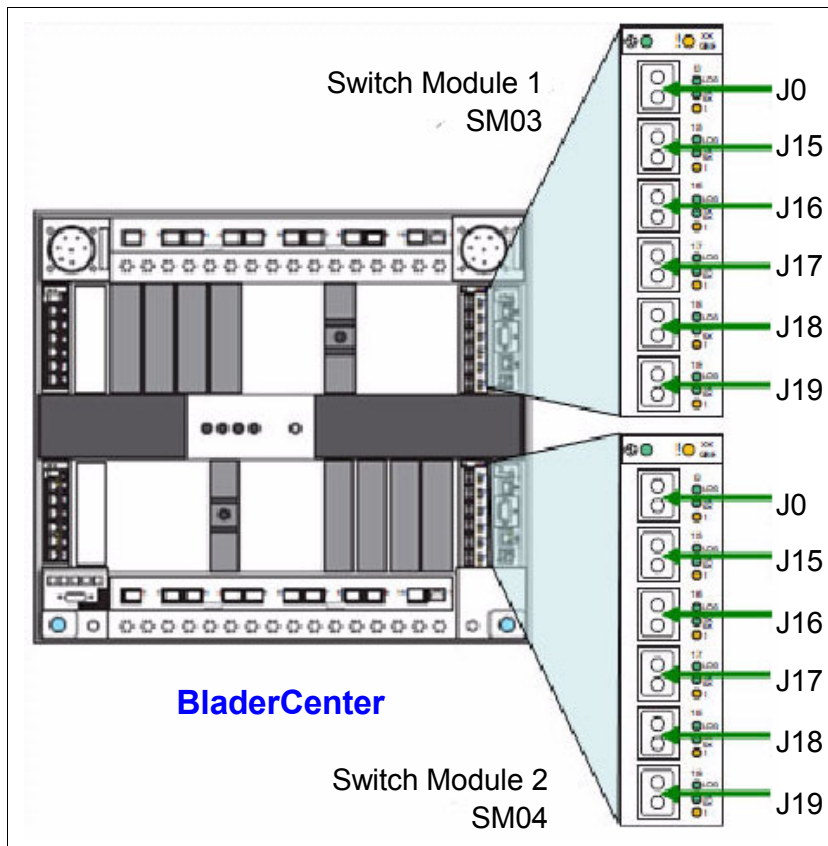
Figure 7-10   8 Gb FC switch external ports

Client provided multi-mode LC duplex cables are used for FC disk connections to support a speeds of 8 Gbps, 4 Gbps, or 2 Gbps. (A speed of 1Gbps is not supported.) Maximum distance depends on the speeds, fiber diameter, and signal frequency.

Cabling specifications are defined by the Fibre Channel - Physical Interface - 4 (FC-PI-4) standard. Table 7-5 identifies cabling types and link data rates that are supported in the zBX SAN environment, including their allowable maximum distances and link loss budget.

The link loss budget is derived from the channel insertion loss budget defined by the FC-PI-4 standard (Revision 8.00).

Table 7-5   Fiber optic cabling for zBX FC disk - maximum distances and link loss budget

FC-PI-4	2 Gbps		4 Gbps		8 Gbps	
Fiber core (light source)	Distance in meters (in feet)	Link loss budget in dB	Distance in meters (in feet)	Link loss budget in dB	Distance in meters (in feet)	Link loss budget in dB
50 µm MM[1] (SX laser)	500 (1640)	3.31	380 (1247)	2.88	150 (492)	2.04
50 µm MM[2] (SX laser)	300 (984)	2.62	150 (492)	2.06	50 (164)	1.68
62.5 µm MM[3] (SX laser)	150 (492)	2.1	70 (230)	1.78	21 (69)	1.58

1. OM3: 50/125 µm laser optimized multimode fiber with a minimum overfilled launch bandwidth of 1500 MHz-km at 850nm as well as an effective laser launch bandwidth of 2000 MHz-km at 850 nm in accordance with IEC 60793-2-10 Type A1a.2 fiber
2. OM2: 50/125 µm multimode fiber with a bandwidth of 500 MHzkm at 850 nm and 500 MHz-km at 1300 nm in accordance with IEC 60793-2-10 Type A1a.1 fiber.
3. OM1: 62.5/125 µm multimode fiber with a minimum overfilled launch bandwidth of 200 MHzkm at 850 nm and 500 MHz-km at 1300 nm in accordance with IEC 60793-2-10 Type A1b fiber.

> **Note:** IBM does not support a mix of 50 µm and 62.5 µm fiber optic cabling in the same physical link.

IBM blade storage connectivity

IBM blades use ports J0 and J15 in both FC switch modules of the BladeCenter chassis to connect either directly to FC disk storage or via a SAN switch to FC disk storage (see Figure 7-11). Cascaded connection between BladeCenter chassis in the zBX is not necessary.
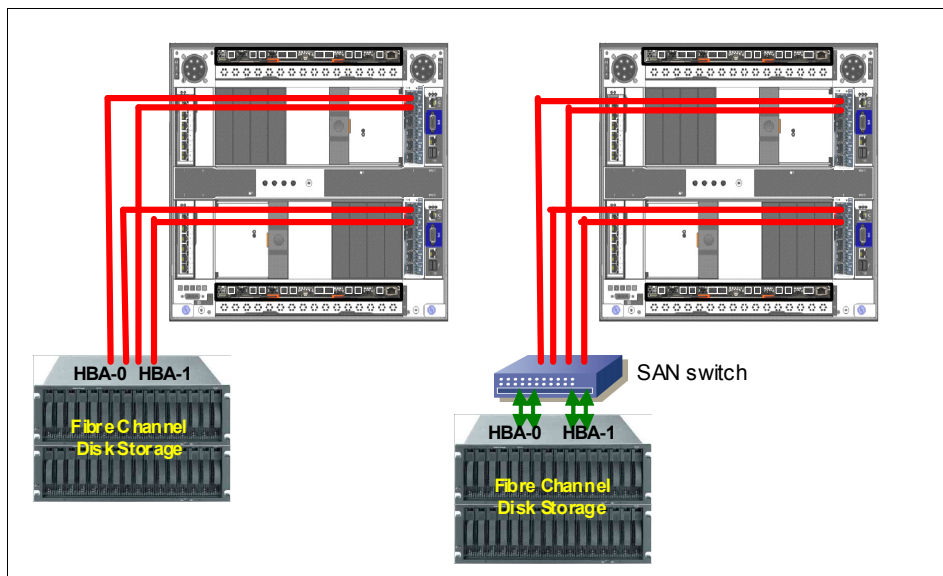
Figure 7-11   BladeCenter chassis storage connectivity options

The client provides all cables, FC disk storage, and SAN switches. It is also the client's responsibility to configure and cable the FC disk storage.

> **Note:** FC disk storage cannot be shared with the IBM Smart Analytics Optimizer solution.

IBM Smart Analytics Optimizer storage connectivity

In an IBM Smart Analytics Optimizer configuration that has one or two BladeCenters, both FC switch modules of each BladeCenter chassis connect directly to FC disk storage using ports J0 and J15. The number of connections between the DS5020 and each BladeCenter is four.

> **Note:** The IBM Smart Analytics Optimizer requires an IBM System Storage® DS5020. The DS5020 must be attached directly to the BladeCenter chassis.

In an IBM Smart Analytics Optimizer configuration that has three or four BladeCenters, the first and second BladeCenter chassis are connected to the DS5020 as stated above.

The third and fourth BladeCenter chassis must have cascaded connections between the first and second BladeCenter chassis using ports 16 and 17 of their FC switch modules. First-Second-Third-Fourth sequence is decided with top-to-bottom, left-to-right rules. (top-to-bottom in the same rack, and left-to-right from racks B through E).

Figure 7-12 shows a cascaded connection example with four BladeCenter chassis for the IBM Smart Analytics Optimizer.
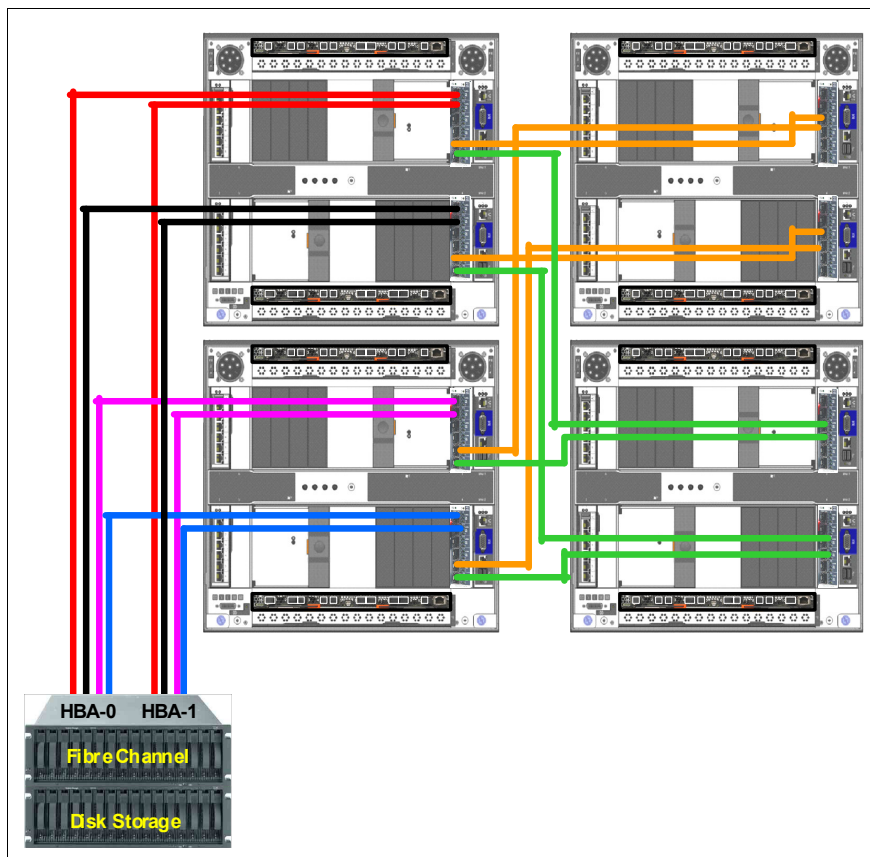
Figure 7-12   IBM Smart Analytics Optimizer and DS5020 connectivity

The above rules are also applied when a zBX has a mix of Smart Analytics Optimizer and POWER7 BladeCenter chassis, but the cascaded connections are only required between the IBM Smart Analytics Optimizer BladeCenter chassis.

Supported FC disk storage

Supported FC disk types and vendors with IBM blades is listed on the IBM System Storage Interoperation Center (SSIC) website, at:

http://www-03.ibm.com/systems/support/storage/config/ssic/displayesssearchwithoutjs.wss?start_over=yes

For an IBM Smart Analytics Optimizer solution, a client supplied IBM System Storage DS5020 with the appropriate configuration and fiber optic cables are required.

The FC 7802 DS5020 Linux/intel license feature is needed to support the Smart Analytics Optimizer.

The storage specifications for each Smart Analytics Optimizer solution size are listed in Table 7-6.

Table 7-6   Storage specifications (DS5020) based on Smart Analytics Optimizer solution size

Solution size	A1-7	A1-14	A1-28	A1-42	A1-56
Number of Fibre Channel ports	4	4	8	8	8
Number of disks[1]	16	16	16	32	32

1. Each disk is a 1000 GB/7.2K SATA II E-DDM

The DS5020 needs to be prepared and connections to the zBX must be in place before it can be used by the Smart Analytics Optimizer application.

Based on the data mart definitions provided by the DB2 administrator, the IBM Smart Analytics Optimizer application sends data from DB2 for z/OS to the zBX. The zBX firmware compresses and the data, and then stores it to the DS5020. From the DS5020, the compressed data is read into blade memory during blade initialization. Processing for DB2 queries is done entirely in blade memory.

The DS5020 also stores the zBX firmware, which is provided by z196 MCLs and loaded from the z196 HMC/SE. Once the firmware is stored, each blade boots from the DS5020.

7.5 zBX connectivity examples

This section shows various ensemble configuration examples containing a zBX and the necessary connectivity for operation. Subsequent configuration diagrams build on the previous configuration and only additional connections will be noted. For a connectivity example of a z196 node without a zBX, refer to Figure 12-7 on page 360.

7.5.1 A single node ensemble with a zBX

Figure 7-13 on page 204 shows a single node ensemble with a zBX. The necessary components include the owning z196 CPC1 and the attached zBX, switches, and fibre channel disk storage.
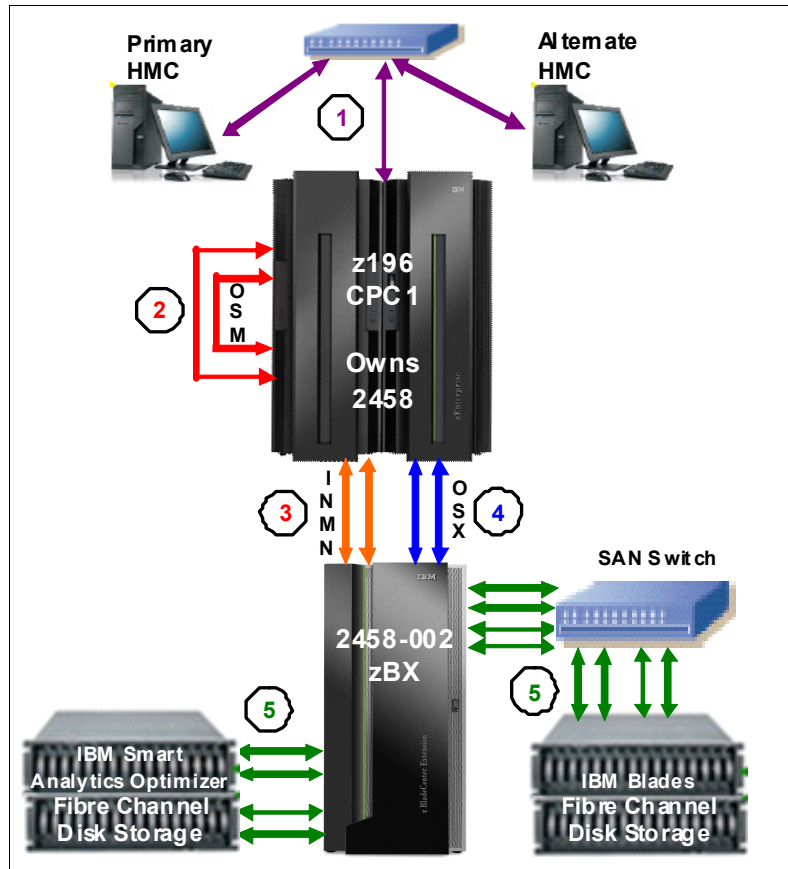
Figure 7-13   Single node ensemble with zBX

1. Client provided management network

    – IBM supplies a 50 feet Ethernet RJ-45 cable with the 1000BASE-T (1GbE) switch (FC 0070)

    – The switch connects to the reserved client network ports in the z196 Z29BPS11/PS31-J02. A second switch connects to Z29BPS11/PS31-J01

2. Intranode management network

    – Two CHPIDs from two different OSA-Express3 1000BASE-T features configured as CHPID type OSM

    – IBM supplies two 3.2 meter Ethernet Category 6 cables from the OSM CHPIDs (ports) to z196 Bulk Power Hub (BPH) Z29BPS11/PS31-J07 (this is a z196 internal connection supplied with feature code 0025)

3. Intranode management network - extension

    – IBM supplies two 26 meter Category 5 Ethernet cables (chrome gray plenum rated cables) from zBX Rack B INMN-A/B switches port J47 to z196 Bulk Power Hub (BPH) Z29BPS11/PS31-J06

4. Intraensemble data network

    – Two ports from two different OSA-Express3 10 GbE (SR Short Reach or LR Long Reach) features configured as CHPID type OSX

    – Client supplies the fiber optic cables (single mode or multimode)

5. 8 Gbps Fibre Channel attached disk storage

– Client supplies all Fibre Channel cables (multimode) from the zBX to the attached disk storage

– Client supplies switches if necessary

– Client is responsible for the configuration and management of the FC attached disk storage.

7.5.2 A dual node ensemble with a single zBX

A second z196 CPC2 (node) is introduced in Figure 7-14, showing the additional hardware. Up to eight additional nodes (z196 servers) can be added in the same fashion.
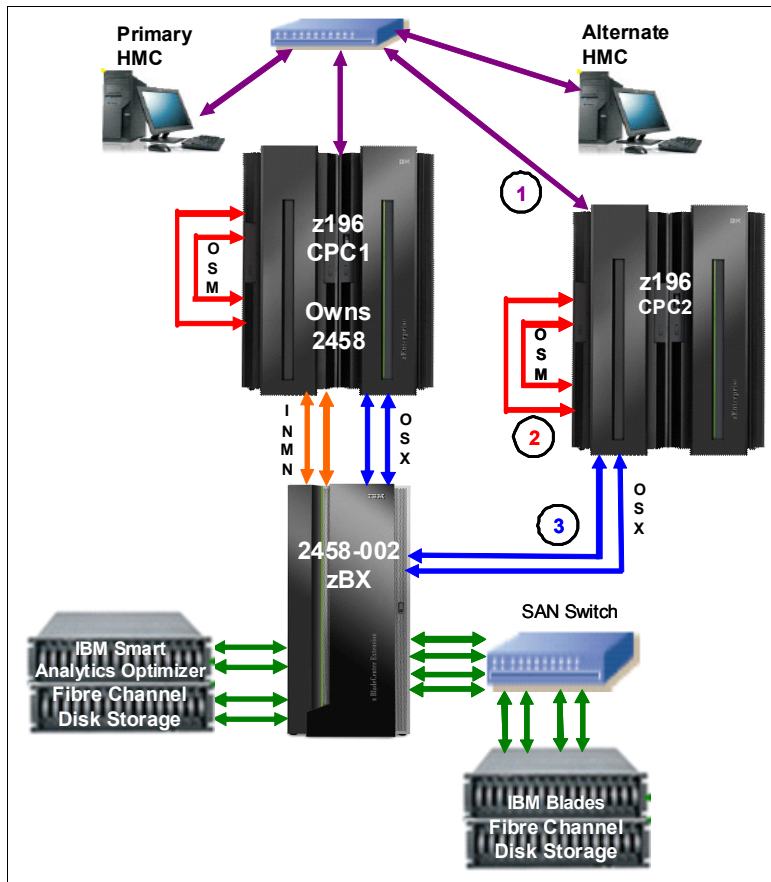
Figure 7-14   Dual node ensemble with a single zBX

1. Client provided management network

– Client supplies an Ethernet RJ-45 cable

– The 1000BASE-T switch (FC 0070) connects to the reserved client network ports in the z196 Z29BPS11/PS31-J02. A second switch connects to Z29BPS11/PS31-J01.

2. Intranode management network

– Two ports from two different OSA-Express3 1000BASE-T features configured as CHPID type OSM

– IBM supplies two 3.2 meter Ethernet Category 6 cables from the OSM CHPIDs (ports) to the z196 Bulk Power Hub (BPH) Z29BPS11/PS31-J07 (this is a z196 internal connection)

3. Intraensemble data network

– Two ports from two different OSA-Express3 10 GbE (SR Short Reach or LR Long Reach) features configured as CHPID type OSX

– Client supplies the fiber optic cables (single mode or multimode)

7.5.3 A dual node ensemble with two zBXs

Figure 7-15 introduces a second zBX added to the original configuration. The two zBXs are interconnected via fiber optic cables to SFPs in the IEDN switches for isolated communication (SR or LR) over the IEDN network.
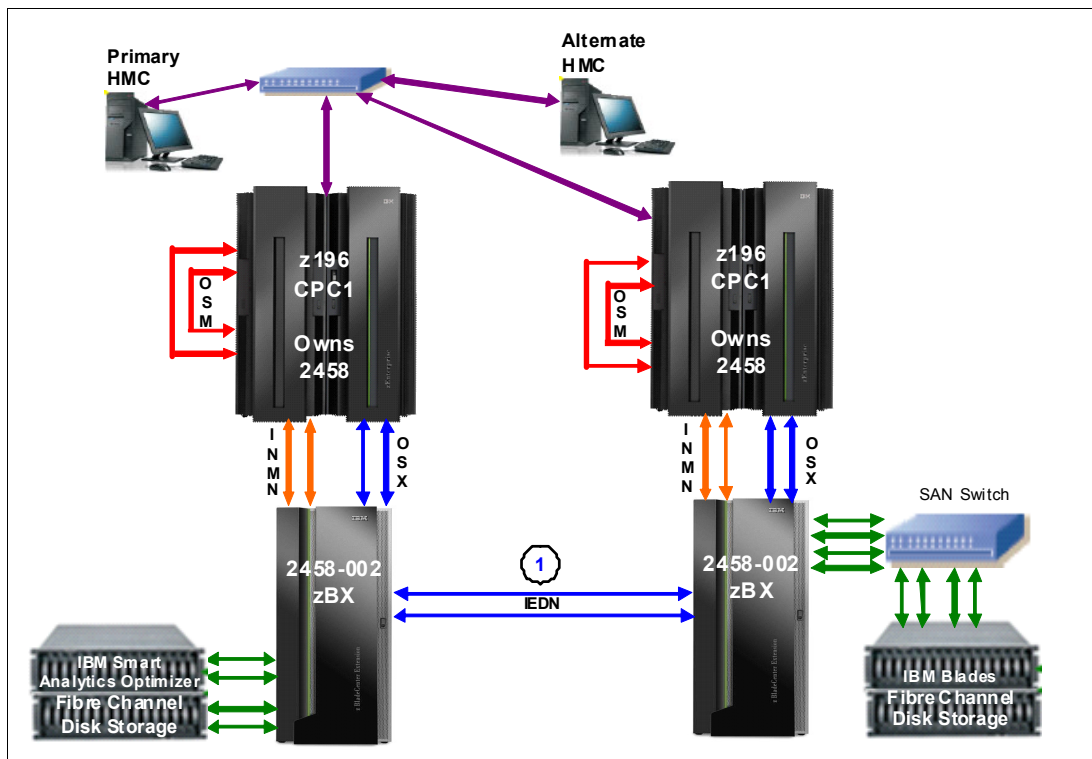
Figure 7-15   Dual node ensemble

1. Intraensemble data network

– Two 10 GbE ports in the TORs are used to connect the two zBXs (10 GbE TOR switch to 10GbE TOR switch).

The maximum of z196s that can be connected to a zBX using the IEDN is eight. Additional z196 CPCs are added and connected to the zBX via the OSA-Express3 10 GbE (SR Short Reach or LR Long Reach) features configured as CHPID type OSX.

Note: z10 servers cannot participate in an ensemble; however they can utilize the zBX environment (applications).

z196 and z10 servers that are not part of an ensemble can connect to a zBX Model 002 via OSA-Express2 or OSA-Express3 features when defined as CHPID type OSD. The OSA ports can be connected either directly to the IEDN or through client supplied Ethernet switches connected to the IEDN.

7.6 References

Installation details can be found in *IBM zEnterprise BladeCenter Extension Model 002 Installation Manual for Physical Planning*, GC27-2611 and *IBM zEnterprise BladeCenter Extension Model 002 Installation Manual*, GC27-2610.

For details about the BladeCenter components, refer to *IBM BladeCenter Products and Technology,* SG24-7523.

Additional documentation is available on IBM Resource Link at:

http://www.ibm.com/servers/resourcelink

8

Software support

This chapter lists the minimum operating system requirements and support considerations for the z196 and its features. It discusses z/OS, z/VM, z/VSE, z/TPF, and Linux on System z. Because this information is subject to change, see the Preventive Service Planning (PSP) bucket for 2817DEVICE for the most current information. Also discussed is generic software support for zEnterprise BladeCenter Extension.

Support of zEnterprise 196 functions is dependent on the operating system, version, and release.

This chapter discusses the following topics:

8.1 Operating systems summary

Table 8-1 lists the minimum operating system levels required on the z196. For zBX see 8.11, "zEnterprise BladeCenter Extension software support" on page 258

Note that operating system levels that are no longer in service are not covered in this publication. These older levels may provide support for some features.

Table 8-1   z196 minimum operating systems requirements

Operating systems	ESA/390 (31-bit mode)	z/Architecture (64-bit mode)	Notes
z/OS V1R7[a]	No	Yes	Service is required. See the following shaded Note box.
z/VM V5R4[b]	No	Yes[c]	
z/VSE V4	No	Yes	
z/TPF V1R1	Yes	Yes	
Linux on System z	See Table 8-2 on page 211.	See Table 8-2 on page 211.	Novell SUSE SLES 10 Red Hat RHEL 5

a. Regular service support for z/OS V1R7 ended in September 2008. However, by ordering the IBM Lifecycle Extension for z/OS V1.7 product, fee-based corrective service can be obtained for up to two years after withdrawal of service (September 2010). Similarly, the IBM Lifecycle Extension for z/OS V1.8 product provides corrective service up to September 2011.

b. z/VM V5R4 provides compatibility support only. z/VM V6 provides both compatibility and exploitation items.

c. z/VM supports both 31-bit and 64-bit mode guests.

Note: Exploitation of certain features depends on a particular operating system. In all cases, PTFs might be required with the operating system level indicated. Check the z/OS, z/VM, z/VSE, and z/TPF subsets of the 2817DEVICE Preventive Service Planning (PSP) buckets. The PSP buckets are continuously updated and contain the latest information about maintenance.

Hardware and software buckets contain installation information, hardware and software service levels, service recommendations, and cross-product dependencies.

8.2 Support by operating system

System z196 introduces several new functions. In this section, we discuss support of those by the current operating systems. Also included are some of the functions introduced in previous System z servers and carried forward or enhanced in the z196. Features and functions available on previous servers but no longer supported by z196 have been removed.

For a list of supported functions and the z/OS and z/VM minimum required support levels, see Table 8-3 on page 213. For z/VSE, Linux on System z, and z/TPF see Table 8-4 on page 217. The tabular format is intended to help determine, by a quick scan, which functions are supported and the minimum operating system level required.

8.2.1 z/OS

z/OS Version 1 Release 9 is the earliest in-service release supporting the z196. Although service support for z/OS Version 1 Release 8 ended in September of 2009, a fee-based extension for defect support (for up to two years) can be obtained by ordering the IBM Lifecycle Extension for z/OS V1.8. Similarly, IBM Lifecycle Extension for z/OS V1.7 provides fee-based support for z/OS Version 1 Release 7 until September 2010, respectively. Support for z/OS Version 1 Release 6 ended on September 30, 2007. Also note that z/OS.e is not supported on z196 and that z/OS.e Version 1 Release 8 was the last release of z/OS.e.

See Table 8-3 on page 213 for a list of supported functions and their minimum required support levels.

8.2.2 z/VM

At general availability:

► z/VM V5R4 provides compatibility only support and z/VM V6R1 provides both compatibility support and exploitation items.

See Table 8-3 on page 213 for a list of supported functions and their minimum required support levels.

Notes: We recommend that the capacity of any z/VM logical partitions, and any z/VM guests, in terms of the number of IFLs and CPs, real or virtual, be adjusted to accommodate the PU capacity of the z196.

8.2.3 z/VSE

Support is provided by z/VSE V4. Note that z/VSE:

► Executes in z/Architecture mode only
► Exploits 64-bit real memory addressing
► Does not support 64-bit virtual addressing

See Table 8-4 on page 217 for a list of supported functions and their minimum required support levels.

8.2.4 Linux on System z

Linux on System z distributions are built separately for the 31-bit and 64-bit addressing modes of the z/Architecture. The newer distribution versions are built for 64-bit only. You can run 31-bit applications in the 31-bit emulation layer on a 64-bit Linux on System z distribution. None of the current versions of Linux on System z distributions (Novell SUSE SLES 10, SLES 11, and Red Hat RHEL 5)[1] require z196 toleration support. Table 8-2 shows the most recent service levels of the current SUSE and Red Hat releases at the time of writing.

Table 8-2   Current Linux on System z distributions as of October 2010

Linux on System z distribution	z/Architecture (64-bit mode)
Novell SUSE SLES 10 SP3	Yes
Novell SUSE SLES 11	Yes
Red Hat RHEL 5.4	Yes

IBM is working with its Linux distribution partners to provide further exploitation of selected z196 functions in future Linux on System z distribution releases.

We recommend that:

► Novell SUSE SLES 11 or Red Hat RHEL 5 be used in any new projects for the z196.

► Any Linux distributions be updated to their latest service level before migration to z196.

► The capacity of any z/VM and Linux on System z logical partitions guests, as well as z/VM guests, in terms of the number of IFLs and CPs, real or virtual, be adjusted according to the PU capacity of the z196.

8.2.5 z/TPF

See Table 8-4 on page 217 for a list of supported functions and their minimum required support levels.

8.2.6 z196 functions support summary

In the following tables, although we attempt to note all functions requiring support, the PTF numbers are not given. Therefore, for the most current information, see the Preventive Service Planning (PSP) bucket for 2817DEVICE.

The following two tables summarize the z196 functions and their minimum required operating system support levels:

► Table 8-3 on page 213 is for z/OS and z/VM.

► Table 8-4 on page 217 is for z/VSE, Linux on System z, and z/TPF.

Information about Linux on System z refers exclusively to appropriate distributions of Novell SUSE and Red Hat.

Both tables use the following conventions:

    Y   The function is supported.
    N   The function is not supported.
    -   The function is not applicable to that specific operating system.

---

[1] SLES is Novell SUSE Linux Enterprise Server
   RHEL is Red Hat Enterprise Linux

Table 8-3   z196 functions minimum support requirements summary, part 1

Function	z/OS V1R12	z/OS V1R11	z/OS V1R10	z/OS V1R9	z/OS V1R8	z/OS V1R7	z/VM V6R1	z/VM V5R4
z196	Y	Y	Y	Y	Y	Y	Y	Y
Greater than 64 PUs single system image	Y	Y	Y^i	N	N	N	N	N
Greater than 54 PUs single system image	Y	Y	Y	Y	N	N	N^a	N^a
zIIP	Y	Y	Y	Y	Y	Y	Y^b	Y^b
zAAP	Y	Y	Y	Y	Y	Y	Y^b	Y^b
zAAP on zIIP	Y	Y	Y^i	Y^i	N	N	Y^c	Y^c
Large memory (> 128 GB)	Y	Y	Y	Y	Y	N	Y^d	Y^d
Large page support	Y	Y	Y	Y	N	N	N^e	N^e
Out-of-order execution	Y	Y	Y	Y	Y	Y	Y	Y
Guest support for execute-extensions facility	-	-	-	-	-	-	Y	Y
Hardware decimal floating point	Y^f	Y^f	Y^f	Y^f	Y^f	Y^f	Y^b	Y^b
Zero address detection	Y	N	N	N	N	N	N	N
60 logical partitions	Y	Y	Y	Y	Y	Y	Y	Y
LPAR group capacity limit	Y	Y	Y	Y	Y	N	-	-
CPU measurement facility	Y	Y	Y^i	Y^i	Y^i	N	N	N
Separate LPAR management of PUs	Y	Y	Y	Y	Y	Y	Y	Y
Dynamic add and delete logical partition name	Y	Y	Y	Y	Y	Y	Y	Y
Capacity provisioning	Y	Y	Y	Y^i	N	N	N^e	N^e
Enhanced flexibility for CoD	Y^f	Y^f	Y^f	Y^f	Y^f	Y^f	Y^f	Y^f
HiperDispatch	Y	Y	Y	Y	Y	Y^g	N^e	N^e
63.75 K subchannels	Y	Y	Y	Y	Y	Y	Y	Y
Four logical channel subsystems (LCSS)	Y	Y	Y	Y	Y	Y	Y	Y
Dynamic I/O support for multiple LCSS	Y	Y	Y	Y	Y	Y	Y	Y
Third subchannel set	Y	Y^i	Y^i	N	N	N	N^e	N^e
Multiple subchannel sets	Y	Y	Y	Y	Y	Y	N^e	N^e
MIDAW facility	Y	Y	Y	Y	Y	Y	Y^b	Y^b
Cryptography								
CPACF protected public key	Y^h	Y^h	Y^h	Y^h	N	N	N^e	N^e
CPACF enhancements	Y^h	Y^h	Y^h	Y^h	Y^h	Y^h	Y^b	Y^b
CPACF AES, PRNG, and SHA-256	Y	Y	Y	Y	Y	Y^h	Y^b	Y^b
CPACF	Y	Y	Y	Y	Y	Y^h	Y^b	Y^b

Function	z/OS V1R12	z/OS V1R11	z/OS V1R10	z/OS V1R9	z/OS V1R8	z/OS V1R7	z/VM V6R1	z/VM V5R4
Personal Account Numbers of 13 to 19 digits	Y[h]	Y[h]	Y[h]	Y[h]	Y[h]	Y[h]	Y[b]	Y[b]
Crypto Express3 enhancements	Y[h]	Y[h]	Y[h]	N	N	N	Y[b]	Y[b]
Crypto Express3	Y[h]	Y[h]	Y[h]	Y[h]	N	N	Y[b]	Y[b]
Remote key loading for ATMs, ISO 16609 CBC mode triple DES MAC	Y	Y	Y	Y	Y	Y[h]	Y[b]	Y[b]
HiperSockets								
32 Hipersockets	Y	Y[i]	Y[i]	N	N	N	Y[i]	Y[i]
HiperSockets multiple write facility	Y	Y	Y	Y[i]	N	N	N[e]	N[e]
HiperSockets support of IPV6	Y	Y	Y	Y	Y	Y	Y	Y
HiperSockets Layer 2 Support	Y	N	N	N	N	N	Y[b]	Y[b]
HiperSockets	Y	Y	Y	Y	Y	Y	Y	Y
ESCON (Enterprise Systems CONnection)								
16-port ESCON feature	Y	Y	Y	Y	Y	Y	Y	Y
FICON (FIber CONnection) and FCP (Fibre Channel Protocol)								
z/OS Discovery and auto configuration (zDAC)	Y	N	N	N	N	N	N	N
zHPF enhanced multitrack support	Y	Y[i]	Y[i]	N	N	N	N	N
High Performance FICON for System z (zHPF)	Y	Y	Y[i]	Y[i]	Y[i]	N[e]	N[e]	N[e]
FCP - increased performance for small block sizes	N	N	N	N	N	N	Y	Y
Request node identification data	Y	Y	Y	Y	Y	Y	N	N
FICON link incident reporting	Y	Y	Y	Y	Y	Y	N	N
N_Port ID Virtualization for FICON (NPIV) CHPID type FCP	N	N	N	N	N	N	Y	Y
FCP point-to-point attachments	N	N	N	N	N	N	Y	Y
FICON SAN platform & name server registration	Y	Y	Y	Y	Y	Y	Y	Y
FCP SAN management	N	N	N	N	N	N	N	N
SCSI IPL for FCP	N	N	N	N	N	N	Y	Y
Cascaded FICON Directors CHPID type FC	Y	Y	Y	Y	Y	Y	Y	Y
Cascaded FICON Directors CHPID type FCP	H	N	N	N	N	N	Y	Y
FICON Express8 and FICON Express4 support of SCSI disks CHPID type FCP	N	N	N	N	N	N	Y	Y

Function	z/OS V1R12	z/OS V1R11	z/OS V1R10	z/OS V1R9	z/OS V1R8	z/OS V1R7	z/VM V6R1	z/VM V5R4
FICON Express8	Y	Y[j]	Y[j]	Y[j]	Y[j]	Y[j]	Y[j]	Y[j]
FICON Express4[k]	Y	Y	Y	Y	Y	Y	Y	Y
OSA (Open Systems Adapter)								
VLAN management	Y	Y	Y	Y	Y	Y	Y	Y
VLAN (IEE 802.1q) support	Y	Y	Y	Y	Y	Y	Y	Y
QDIO data connection isolation for z/VM virtualized environments	-	-	-	-	-	-	Y	Y[i]
OSA Layer 3 Virtual MAC	Y	Y	Y	Y	Y	N	Y[b]	Y[b]
OSA Dynamic LAN idle	Y	Y	Y	Y	Y	N	Y[b]	Y[b]
OSA/SF enhancements for IP, MAC addressing (CHPID=OSD)	Y	Y	Y	Y	Y	Y	Y	Y
QDIO diagnostic synchronization	Y	Y	Y	Y	Y	N	Y[b]	Y[b]
OSA-Express2 Network Traffic Analyzer	Y	Y	Y	Y	Y	N	Y[b]	Y[b]
Broadcast for IPv4 packets	Y	Y	Y	Y	Y	Y	Y	Y
Checksum offload for IPv4 packets	Y	Y	Y	Y	Y	Y	Y[l]	Y[l]
OSA-Express3 10 Gigabit Ethernet LR CHPID type OSX	Y	Y[i]	Y[i]	N	N	N	Y[i]	Y[im]
OSA-Express3 10 Gigabit Ethernet SR CHPID type OSX	Y[i]	Y[i]	Y[i]	N	N	N	Y[i]	Y[im]
OSA-Express3 10 Gigabit Ethernet LR CHPID type OSD	Y[i]	Y	Y	Y	N	N	Y	Y
OSA-Express3 10 Gigabit Ethernet SR CHPID type OSD	Y	Y	Y	Y	Y	Y	Y	Y
OSA-Express3 Gigabit Ethernet LX (using four ports) CHPID types OSD, OSN	Y	Y	Y	Y[i]	Y[i]	N	Y	Y
OSA-Express3 Gigabit Ethernet LX using two ports. CHPID types OSD, OSN	Y	Y	Y	Y	Y	Y	Y	Y
OSA-Express3 Gigabit Ethernet SX (using four ports) CHPID types OSD, OSN	Y	Y	Y	Y[i]	Y[i]	N	Y	Y
OSA-Express3 Gigabit Ethernet SX (using 2 ports) CHPID types OSD, OSN	Y	Y	Y	Y	Y	Y	Y	Y
OSA-Express3 1000BASE-T (using two ports) CHPID type OSM	Y[i]	Y[i]	Y[i]	N	N	N	Y[i]	Y[im]
OSA-Express3 1000BASE-T (using 1 + 1 port) CHPID type OSC	Y	Y	Y	Y	Y	Y	Y	Y

Function	z/OS V1R12	z/OS V1R11	z/OS V1R10	z/OS V1R9	z/OS V1R8	z/OS V1R7	z/VM V6R1	z/VM V5R4
OSA-Express3 1000BASE-T (using four ports) CHPID type OSD	Y	Y	Y	Y[i]	Y[i]	N	Y	Y
OSA-Express3 1000BASE-T (using two ports) CHPID type OSD	Y	Y	Y	Y	Y	Y	Y	Y
OSA-Express3 1000BASE-T (using two or four ports) CHPID type OSE	Y	Y	Y	Y	Y	Y	Y	Y
OSA-Express3 1000BASE-T CHPID type OSN	Y	Y	Y	Y	Y	Y	Y	Y
OSA-Express2 Gigabit Ethernet LX and SX[m] CHPID type OSD	Y	Y	Y	Y	Y	Y	Y	Y
OSA-Express2 Gigabit Ethernet LX and SX[m] CHPID type OSN	Y	Y	Y	Y	Y	Y	Y	Y
OSA-Express2 1000BASE-T Ethernet CHPID type OSC	Y	Y	Y	Y	Y	Y	Y	Y
OSA-Express2 1000BASE-T Ethernet CHPID type OSD	Y	Y	Y	Y	Y	Y	Y	Y
OSA-Express2 1000BASE-T Ethernet CHPID type OSE	Y	Y	Y	Y	Y	Y	Y	Y
OSA-Express2 1000BASE-T Ethernet CHPID type OSN	Y	Y	Y	Y	Y	Y	Y	Y
Parallel Sysplex and other								
z/VM integrated systems management	-	-	-	-	-	-	Y	Y
System-initiated CHPID reconfiguration	Y	Y	Y	Y	Y	Y	-	-
Program-directed re-IPL	-	-	-	-	-	-	Y	Y
Multipath IPL	Y	Y	Y	Y	Y	Y	N	N
STP enhancements	Y	Y	Y	Y	Y	Y	-	-
Server Time Protocol	Y	Y	Y	Y	Y	Y	-	-
Coupling over InfiniBand CHPID type CIB	Y	Y	Y	Y	Y	Y	Y[m]	Y[m]
InfiniBand coupling links (1x IB-SDR or 1xIB DDR) at an unrepeated distance of 10 km	Y	Y	Y	Y[i]	Y[i]	N	Y[m]	Y[m]
Dynamic I/O support for InfiniBand CHPIDs	-	-	-	-	-	-	Y	Y
CFCC Level 17	Y	Y	Y	Y	Y	Y	Y[b]	Y[b]

a. A maximum of 32 PUs per system image is supported. Guests can be defined with up to 64 virtual PUs. z/VM V5R4 and above support up to 32 PUs.

b. Support is for guest use only.

c. Available for z/OS on virtual machines without virtual zAAPs defined when the z/VM LPAR does not have zAAPs defined.

d. 256 GB of central memory are supported by z/VM V5R4 and later. z/VM V5R4 and later are designed to support more than 1 TB of virtual memory in use for guests.

e. Not available to guests.

f. Support varies by operating system and by version and release.

g. This requires support for zIIP.

h. FMIDs are shipped in a Web Deliverable.

i. PTFs are required.

j. Support varies with operating system and level. See "FCP provides increased performance" on page 234"FCP provides increased performance" on page 234 for details.

k. FICON Express4 10KM LX, 4KM LX, and SX features are withdrawn from marketing.

l. Supported for dedicated devices only.

m. Support is for dynamic I/O configuration only.

Table 8-4   z196 functions minimum support requirements summary, part 2

Function	z/VSE V4R3	z/VSE V4R2	z/VSE V4R1	Linux on System z	z/TPF V1R1
z196	Y	Y	Y	Y	Y
Greater than 64 PUs single system image	N	N	N	N	Y
Greater than 54 PUs single system image	N	N	N	Y	Y
zIIP	-	-	-	-	-
zAAP	-	-	-	-	-
zAAP on zIIP	-	-	-	-	-
Large memory (> 128 GB)	N	N	N	Y	Y
Large page support	Y	N	N	Y	N
Out-of-order execution	Y	Y	Y	Y	Y
Guest support for Execute-extensions facility	-	-	-	-	-
Hardware decimal floating point[a]	N	N	N	Y[b]	N
Zero address detection	N	N	N	N	N
60 logical partitions	Y	Y	Y	Y	Y
CPU measurement facility	N	N	N	N	N
LPAR group capacity limit	-	-	-	-	-
Separate LPAR management of PUs	Y	Y	Y	Y	Y
Dynamic add/delete logical partition name	N	N	N	Y	N
Capacity provisioning	-	-	-	-	N
Enhanced flexibility for CoD	-	-	-	-	N
HiperDispatch	N	N	N	N	N
63.75 K subchannels	N	N	N	Y	N
Four logical channel subsystems	Y	Y	Y	Y	N
Dynamic I/O support for multiple LCSS	N	N	N	Y	N
Third subchannel set	N	N	N	N	N

Function	z/VSE V4R3	z/VSE V4R2	z/VSE V4R1	Linux on System z	z/TPF V1R1
Multiple subchannel sets	N	N	N	Y	N
MIDAW facility	N	N	N	N	N
Cryptography					
CPACF protected public key	N	N	N	Y	N
CPACF enhancements	Y	Y	Y	N	N
CPACF AES, PRNG, and SHA-256	Y	Y	Y	Y	N
CPACF	Y	Y	Y	Y	Y
Personal Account Numbers of 13 to 19 digits	N	N	N	-	
Crypto Express3 enhancements	Y	Y[c]	N	Y[d]	Y
Crypto Express3	Y	Y[c]	N	Y[d]	Y
Remote key loading for ATMs, ISO 16609 CBC mode triple DES MAC	N	N	N	-	N
HiperSockets					
32 Hipersockets	N	N	N	N	N
HiperSockets multiple write facility	N	N	N	N	N
HiperSockets support of IPV6	N	N	N	Y	N
HiperSockets Layer 2 Support	N	N	N	Y	N
HiperSockets	Y	Y	Y	Y	N
ESCON (Enterprise System CONnection)					
16-port ESCON feature	Y	Y	Y	Y	Y
FIber CONnection (FICON) and Fibre Channel Protocol (FCP)					
z/OS Discovery and auto configuration (zDAC)	N	N	N	N	N
zHPF enhanced multitrack support	N	N	N	N	N
High Performance FICON for System z (zHPF)	N	N	N	N	N
FCP - increased performance for small block sizes	Y	Y	Y	Y	N
Request node identification data	-	-	-	-	-
FICON link incident reporting	N	N	N	N	N
N_Port ID Virtualization for FICON (NPIV) CHPID type FCP	Y	Y	Y	Y	N
FCP point-to-point attachments	Y	Y	Y	Y	N
FICON SAN platform and name registration	Y	Y	Y	Y	Y
FCP SAN management	N	N	N	Y	N
SCSI IPL for FCP	Y	Y	Y	Y	N

Function	z/VSE V4R3	z/VSE V4R2	z/VSE V4R1	Linux on System z	z/TPF V1R1
Cascaded FICON Directors CHPID type FC	Y	Y	Y	Y	Y
Cascaded FICON Directors CHPID type FCP	Y	Y	Y	Y	N
FICON Express8 and FICON Express4 support of SCSI disks CHPID type FCP	Y	Y	Y	Y	N
FICON Express8	Y[e]	Y[e]	Y[e]	Y[e]	Y[e]
FICON Express4[f]	Y	Y	Y	Y	Y
Open Systems Adapter (OSA)					
VLAN management	N	N	N	N	N
VLAN (IEE 802.1q) support	N	N	N	Y	N
QDIO data connection isolation for z/VM virtualized environments	-	-	-	-	-
OSA Layer 3 Virtual MAC	N	N	N	N	N
OSA Dynamic LAN idle	N	N	N	N	N
OSA/SF enhancements for IP, MAC addressing (CHPID=OSD)	N	N	N	N	N
OSA-Express2 QDIO Diagnostic Synchronization	N	N	N	N	N
OSA-Express2 Network Traffic Analyzer	N	N	N	N	N
Broadcast for IPv4 packets	N	N	N	Y	N
Checksum offload for IPv4 packets	N	N	N	Y	N
OSA-Express3 10 Gigabit Ethernet LR CHPID type OSX	N	N	N	N	N
OSA-Express3 10 Gigabit Ethernet SR CHPID type OSX	N	N	N	N	N
OSA-Express3 10 Gigabit Ethernet LR CHPID type OSD	Y	Y	Y	Y	Y
OSA-Express3 10 Gigabit Ethernet SR CHPID type OSD	Y	Y	Y	Y	Y
OSA-Express3 Gigabit Ethernet LX 4-ports CHPID types OSD	Y	Y	Y	Y	Y
OSA-Express3 Gigabit Ethernet SX 4-ports CHPID types OSD	Y	Y	Y	Y	Y
OSA-Express3 1000BASE-T (using two or four ports) CHPID type OSM	N	N	N	N	N

Function	z/VSE V4R3	z/VSE V4R2	z/VSE V4R1	Linux on System z	z/TPF V1R1
OSA-Express3 1000BASE-T CHPID type OSC	Y	Y	Y	-	N
OSA-Express3 1000BASE-T 4-ports CHPID type OSD	Y	Y	Y	Y	Y
OSA-Express3 1000BASE-T 4-ports CHPID type OSE	Y	Y	Y	N	N
OSA-Express3 1000BASE-T Ethernet CHPID type OSN	Y	Y	Y	Y	Y
OSA-Express2 Gigabit Ethernet LX and SX[g] CHPID type OSD	Y	Y	Y	Y	Y
OSA-Express2 Gigabit Ethernet LX and SX[g] CHPID type OSN	Y	Y	Y	Y	Y
OSA-Express2 1000BASE-T Ethernet CHPID type OSC	Y	Y	Y	N	N
OSA-Express2 1000BASE-T Ethernet CHPID type OSD	Y	Y	Y	Y	Y
OSA-Express2 1000BASE-T Ethernet CHPID type OSE	Y	Y	Y	N	N
OSA-Express2 1000BASE-T Ethernet CHPID type OSN	Y	Y	Y	Y	Y
Parallel Sysplex and other					
z/VM integrated systems management	-	-	-	-	-
System-initiated CHPID reconfiguration	-	-	-	Y	-
Program-directed re-IPL	Y[g]	Y[g]	Y[g]	Y	-
Multipath IPL	-	-	-	-	-
STP enhancements	-	-	-	-	-
Server Time Protocol	-	-	-	-	-
Coupling over InfiniBand CHPID type CIB	-	-	-	-	Y
InfiniBand coupling links (1x IB-SDR or IB-DDR) at unrepeated distance of 10 km	-	-	-	-	-
Dynamic I/O support for InfiniBand CHPIDs	-	-	-	-	-
CFCC Level 17	-	-	-	-	Y

a. Support varies with operating system and level.
b. Supported by Novell SUSE SLES 11.
c. Service is required.
d. Toleration support only. Requires Novell SUSE SLES 10 SP3 or Red Hat RHEL 5.4.
e. See "FCP provides increased performance" on page 234 for details.
f. FICON Express4 10KM LX, 4KM LX, and SX features are withdrawn from marketing.
g. This is for FCP-SCSI disks.

8.3 Support by function

In this section, we discuss operating system support by function.

8.3.1 Single system image

A single system image can control several processor units such as CPs, zIIPs, zAAPs, or IFLs, as appropriate.

Maximum number of PUs

Table 8-5 shows the maximum number of PUs supported for each operating system image.

Table 8-5   Single system image software support

Operating system	Maximum number of (CPs+zIIPs+zAAPs)[a] or IFLs per system image
z/OS V1R12	80
z/OS V1R11	80
z/OS V1R10	80[b]
z/OS V1R9	64
z/OS V1R8	32
z/OS V1R7	32[c]
z/VM V6R1	32
z/VM V5R4	32
z/VSE V4	z/VSE Turbo Dispatcher can exploit up to 4 CPs and tolerates up to 10-way LPARs
Linux on System z	Novell SUSE SLES 10: 64 CPs or IFLs
	Novell SUSE SLES 11: 64 CPs or IFLs
	Red Hat RHEL 5: 64 CPs or IFLs
z/TPF V1R1	84 CPs; z196 limits this number to 80

a. The number of purchased zAAPs and the number of purchased zIIPs each cannot exceed the number of purchased CPs. A logical partition can be defined with any number of the available zAAPs and zIIPs. The total refers to the sum of these PU characterizations.
b. Service is required
c. z/OS V1R7 requires IBM zIIP support for z/OS V1R7 Web deliverable to be installed to enable HiperDispatch.

The z/VM-mode logical partition

z196 supports a logical partition (LPAR) mode, named z/VM-mode, which is exclusive for running z/VM. The z/VM-mode requires z/VM V5R4 or later and allows z/VM to utilize a wider variety of specialty processors in a single LPAR. For instance, in a z/VM-mode LPAR, z/VM can manage Linux on System z guests running on IFL processors while also managing z/VSE and z/OS on central processors (CPs), and allowing z/OS to fully exploit IBM z196 Integrated Information Processors (zIIPs) and IBM z196 Application Assist Processors (zAAPs).

8.3.2 zAAP support

zAAPs do not change the model capacity identifier of the z196. IBM software product license charges based on the model capacity identifier are not affected by the addition of zAAPs. On a z196, z/OS Version 1 Release 7 is the minimum level for supporting zAAPs, together with IBM SDK for z/OS Java 2 Technology Edition V1.4.1.

Exploiters of zAAPs include:

► Any Java application that is using the current IBM SDK.

► WebSphere Application Server V5R1and later, and products based on it, such as WebSphere Portal, WebSphere Enterprise Service Bus (WebSphere ESB), WebSphere Business Integration (WBI) for z/OS and so on.

► CICS/TS V2R3 and later.

► DB2 UDB for z/OS Version 8 and later.

► IMS Version 8 and later.

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

In order to exploit zAAPs DB2 V9 has the following prerequisites:

► DB2 V9 for z/OS in new function mode

► The C API for z/OS XML System Services, available with z/OS V1R9 with rollback APARs to z/OS V1R7, and z/OS V1R8

► One of the following items:

– z/OS V1R9 has native support.

– z/OS V1R8 requires an APAR for zAAP support.

– z/OS V1R7 requires an APAR for zAAP support and an APAR for rollback of z/OS XML System Services.

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

Use the PROJECTCPU option of the IEAOPTxx parmlib member to help determine whether zAAPs can be beneficial to the installation. Setting PROJECTCPU=YES directs z/OS to record the amount of eligible work for zAAPs and zIIPs in SMF record type 72 subtype 3. Field APPL% AAPCP of the Workload Activity Report listing by WLM service class indicates what percentage of a processor is zAAP eligible. Because of zAAPs lower prices, as compared to CPs, an utilization as low as 10% may provide benefit.

8.3.3 zIIP support

zIIPs do not change the model capacity identifier of the z196. IBM software product license charges based on the model capacity identifier are not affected by the addition of zIIPs. On a z196, z/OS Version 1 Release 7 with the zIIP web deliverable is the minimum level for supporting zIIPs, together with IBM SDK for z/OS Java 2 Technology Edition V1.4.1.

No changes to applications are required to exploit zIIPs. Exploiters of zIIPs include:

► DB2 V8 and above for z/OS Data serving, for applications using data DRDA over TCP/IP, such as data serving and data warehousing, and selected utilities

- ► z/OS XML services

- ► z/OS CIM Server

- ► z/OS Communications Server for network encryption (IPSec) and for large messages sent via HiperSockets

- ► IBM GBS Scalable Architecture for Financial Reporting

- ► z/OS Global Mirror (formerly XRC) and System Data Mover

The functioning of a zIIP is transparent to application programs.

Use the PROJECTCPU option of the IEAOPTxx parmlib member to help determine whether zIIPs can be beneficial to the installation. Setting PROJECTCPU=YES directs z/OS to record the amount of eligible work for zAAPs and zIIPs in SMF record type 72 subtype 3. Field APPL% IIPCP of the Workload Activity Report listing by WLM service class indicates what percentage of a processor is zIIP eligible. Because of zIIPs lower prices, as compared to CPs, an utilization as low as 10% may provide benefit.

8.3.4 zAAP on zIIP capability

This capability, first made available on System z9 servers under defined circumstances, enables workloads eligible to run on Application Assist Processors (zAAPs) to run on Integrated Information Processors (zIIP). It is intended as a means to optimize the investment on existing zIIPs and not as a replacement for zAAPs. The rule of at least one CP installed per zAAP and zIIP installed still applies.

Exploitation of this capability is by z/OS only, and is only available when zIIPs are installed and one of the following situations occurs:

- ► There are no zAAPs installed on the server.

- ► z/OS is running as a guest of z/VM V5R4 or later and there are no zAAPs defined to the z/VM LPAR. The server may have zAAPs installed. Because z/VM can dispatch both virtual zAAPs and virtual zIIPs on real CPs[2], the z/VM partition does not require any real zIIPs defined to it, although we recommend using real zIIPs due to software licensing reasons.

Support is available on z/OS V1R11 and this capability is enabled by default (ZAAPZIIP=YES). To disable it, specify NO for the ZAAPZIIP parameter in the IEASYSxx PARMLIB member.

On z/OS V1R10 and z/OS V1R9 support is provided by PTF for APAR OA27495 and the default setting in the IEASYSxx PARMLIB member is ZAAPZIIP=NO. Enabling or disabling this capability is disruptive. After changing the parameter, z/OS must be re-IPLed for the new setting to take effect.

---

[2] The z/VM system administrator can use the SET CPUAFFINITY command to influence the dispatching of virtual specialty engines on CPs or real specialty engines.

8.3.5 Maximum main storage size

Table 8-6 on page 224 lists the maximum amount of main storage supported by current operating systems. Expanded storage, although part of the z/Architecture, is currently exploited only by z/VM. A maximum of 1 TB of main storage can be defined for a logical partition.

Table 8-6   Maximum memory supported by operating system

Operating system	Maximum supported main storage
z/OS	z/OS V1R12 supports 4 TB and up to 3 TB per server[a] z/OS V1R11 supports 4 TB and up to 3 TB per server[a] z/OS V1R10 supports 4 TB and up to 3 TB per server[a] z/OS V1R9 supports 4 TB and up to 3 TB per server[a] z/OS V1R8 supports 4 TB and up to 3 TB per server[a] z/OS V1R7 supports 128 GB
z/VM	z/VM V6R1 supports 256 GB z/VM V5R4 supports 256 GB
Linux on System z (64-bit)	Novell SUSE SLES 11 supports 4 TB[a] Novell SUSE SLES 10 supports 4 TB[a] Red Hat RHEL 5 supports 64 GB
z/VSE	z/VSE V4R2 supports 32 GB z/VSE V4R1 supports 8 GB
z/TPF	z/TPF supports 4 TB[a]

a. System z196 restricts the maximum LPAR memory size to 1 TB.

8.3.6 Large page support

In addition to the existing 4 KB pages and page frames, z196 supports large pages and large page frames that are 1 MB in size, as described in "Large page support" on page 91. Table 8-7 lists large page support requirements.

Table 8-7   Minimum support requirements for large page

Operating system	Support requirements
z/OS	z/OS V1R9
z/VM	Not supported; not available to guests
z/VSE	z/VSE V4R3; supported for data spaces
Linux on System z	Novell SUSE SLES 10 SP2 Red Hat RHEL 5.2

8.3.7 Guest support for execute-extensions facility

The execute-extensions facility contains several new machine instructions. Support is required in z/VM so that guests can exploit this facility. Table 8-8 lists the minimum support requirements.

Table 8-8   Minimum support requirements for execute-extensions facility

Operating system	Support requirements

z/VM	z/VM V5R4: support is included in the base

8.3.8 Hardware decimal floating point

Industry support for decimal floating point is growing, with IBM leading the open standard definition. Examples of support for the draft standard IEEE 754r include Java BigDecimal, C#, XML, C/C++, GCC, COBOL, and other key software vendors such as Microsoft® and SAP.

Decimal floating point support was introduced with the z9 EC. However, the z196 has inherited the decimal floating point accelerator feature introduced with the z10 EC and described in 3.3.4, "Decimal floating point accelerator" on page 74.

Table 8-9 lists the operating system support for decimal floating point. See also 8.5.7, "Decimal floating point and z/OS XL C/C++ considerations" on page 251.

Table 8-9   Minimum support requirements for decimal floating point

Operating system	Support requirements
z/OS	z/OS V1R9: Support includes XL, C/C++, HLASM, Language Environment®, DBX, and CDA RTLE. z/OS V1R8: Support includes HL ASM, Language Environment, DBX, and CDA RTLE. z/OS V1R7: Support is for the High Level Assembler (HLASM) only.
z/VM	z/VM V5R4: Support is for guest use.
Linux on System z	Novell SUSE SLES 11.

8.3.9 Up to 60 logical partitions

This feature, first made available in the z9 EC, allows the system to be configured with up to 60 logical partitions. Because channel subsystems can be shared by up to 15 logical partitions, it is necessary to configure four channel subsystems to reach 60 logical partitions. Table 8-10 lists the minimum operating system levels for supporting 60 logical partitions.

Table 8-10   Minimum support requirements for 60 logical partitions

Operating system	Support requirements
z/OS	z/OS V1R7
z/VM	z/VM V5R4
z/VSE	z/VSE V4R1
Linux on System z	Novell SUSE SLES 10 Red Hat RHEL 5
z/TPF	z/TPF V1R1

8.3.10 Separate LPAR management of PUs

The z196 uses separate PU pools for each optional PU type. The separate management of PU types enhances and simplifies capacity planning and management of the configured logical partitions and their associated processor resources. Table 8-11 on page 226 lists the support requirements for separate LPAR management of PU pools.

Table 8-11   Minimum support requirements for separate LPAR management of PUs

Operating system	Support requirements
z/OS	z/OS V1R7
z/VM	z/VM V5R4
z/VSE	z/VSE V4R1
Linux on System z	Novell SUSE SLES 10 Red Hat RHEL 5
z/TPF	z/TPF V1R1

8.3.11 Dynamic LPAR memory upgrade

A logical partition can be defined with both an initial and a reserved amount of memory. At activation time the initial amount is made available to the partition and the reserved amount can be added later, partially or totally. Those two memory zones do not have to be contiguous in real memory but appear as *logically contiguous* to the operating system running in the LPAR.

z/OS is able to take advantage of this support and nondisruptively acquire and release memory from the reserved area. z/VM V5R4 and higher are able to acquire memory nondisruptively, and immediately make it available to guests. z/VM virtualizes this support to its guests, which now can also increase their memory nondisruptively, if supported by the guest operating system. Releasing memory from z/VM is a disruptive operation to z/VM. Releasing memory from the guest depends on the guest's operating system support.

8.3.12 Capacity Provisioning Manager

The provisioning architecture, described in 9.8, "Nondisruptive upgrades" on page 303, enables you to better control the configuration and activation of the On/Of Capacity on Demand. The new process is inherently more flexible and can be automated. This capability can result in easier, faster, and more reliable management of the processing capacity.

The Capacity Provisioning Manager, a function first available with z/OS V1R9, interfaces with z/OS Workload Manager (WLM) and implements capacity provisioning policies. Several implementation options are available: from an analysis mode, that only issues recommendations to an autonomic mode providing fully automated operations.

Replacing manual monitoring with autonomic management or supporting manual operation with recommendations can help ensure that sufficient processing power will be available with the least possible delay. Support requirements are listed on Table 8-12.

Table 8-12   Minimum support requirements for capacity provisioning

Operating system	Support requirements
z/OS	z/OS V1R9
z/VM	Not supported; not available to guests

8.3.13 Dynamic PU add

z/OS has long been able to define reserved PUs to an LPAR for the purpose of non-disruptively bringing online the additional computing resources when needed.

Starting with z/OS V1R10, z/VM V5R4, and z/VSE V4R3, an enhanced capability, the ability to dynamically define and change the number and type of reserved PUs in an LPAR profile can be used for that purpose. No pre-planning is required.

The new resources are immediately made available to the operating systems and, in the z/VM case, to its guests. However, z/VSE, when running as a z/VM guest does not support this capability.

8.3.14 HiperDispatch

HiperDispatch, which is exclusive to z196 and System z10, represents a cooperative effort between the z/OS operating system and the z196 hardware. It improves efficiencies in both the hardware and the software in the following ways:

► Work may be dispatched across fewer logical processors, therefore reducing the multiprocessor (MP) effects and lowering the interference among multiple partitions.

► Specific z/OS tasks may be dispatched to a small subset of logical processors that Processor Resource/Systems Manager (PR/SM) will tie to the same physical processors, thus improving the hardware cache reuse and locality of reference characteristics such as reducing the rate of cross-book communication.

For more information, see 3.6, "Logical partitioning" on page 93. Table 8-13 lists HiperDispatch support requirements.

Table 8-13   Minimum support requirements for HiperDispatch

Operating system	Support requirements
z/OS	z/OS V1R7 and later with PTFs (z/OS V1R7 requires IBM zIIP support for z/OS V1R7 Web deliverable)
z/VM	Not supported; not available to guests

8.3.15 The 63.75 K subchannels

Servers prior to the z9 EC reserved 1024 subchannels for internal system use out of the maximum of 64 K subchannels. Starting with the z9 EC, the number of reserved subchannels has been reduced to 256, thus increasing the number of subchannels available. Reserved subchannels exist only in subchannel set 0. No subchannels are reserved in subchannel set 1 or set 2. The informal name, *63.75 K subchannels*, represents 65280 subchannels, as shown in the following equation:

`63` x 1024 + 0`.75` x 1024 = 65280

Table 8-14 lists the minimum operating system level required on z196.

Table 8-14   Minimum support requirements for 63.75 K subchannels

Operating system	Support requirements
z/OS	z/OS V1R7
z/VM	z/VM V5R4

Operating system	Support requirements
Linux on System z	Novell SUSE SLES 10 Red Hat RHEL 5

8.3.16 Multiple subchannel sets

Multiple subchannel sets, first introduced in z9 EC, provide a mechanism for addressing more than 63.75 K I/O devices and aliases for ESCON (CHPID type CNC) and FICON (CHPID types FC) on the z9 EC and z196. z196 introduces the third subchannel set (SS2).

Multiple subchannel sets are not supported for z/OS running as a guest of z/VM.

Table 8-15 lists the minimum operating systems level required on the z196.

Table 8-15   Minimum software requirement for MSS

Operating system	Support requirements
z/OS	z/OS V1R7
Linux on System z	Novell SUSE SLES 10 Red Hat RHEL 5

8.3.17 Third subchannel set

With z196 a third subchannel set (SS2) is introduced. It applies to ESCON (CHPID type CNC) and FICON (CHPID type FC for both FICON and zHPF paths) channels.

Together with the second set (SS1) it can be used for disk alias devices of both primary and secondary devices, and as Metro Mirror secondary devices. This should help facilitate storage growth and complements other functions such as EAV and HyperPAV.

Table 8-16 lists the minimum operating systems level required on the z196.

Table 8-16   Minimum software requirement for SS2

Operating system	Support requirements
z/OS	z/OS V1R10 with PTFs

8.3.18 MIDAW facility

The modified indirect data address word (MIDAW) facility improves FICON performance. The MIDAW facility provides a more efficient CCW/IDAW structure for certain categories of data-chaining I/O operations.

Support for the MIDAW facility when running z/OS as a guest of z/VM requires z/VM V5R4 or higher. See 8.7, "MIDAW facility" on page 253.

Table 8-17 lists the minimum support requirements for MIDAW.

Table 8-17   Minimum support requirements for MIDAW

Operating system	Support requirements
z/OS	z/OS V1R7

Operating system	Support requirements
z/VM	z/VM V5R4 for guest exploitation

8.3.19 Enhanced CPACF

Cryptographic functions are described in 8.4, "Cryptographic support" on page 245.

8.3.20 HiperSockets multiple write facility

This capability allows the streaming of bulk data over a HiperSockets link between two logical partitions. Multiple output buffers are supported on a single SIGA write instruction.The key advantage of this enhancement is that it allows the receiving logical partition to process a much larger amount of data per I/O interrupt. This is transparent to the operating system in the receiving partition. HiperSockets Multiple Write Facility with fewer I/O interrupts is designed to reduce CPU utilization of the sending and receiving partitions.

Support for this function is required by the sending operating system. See 4.7.7, "HiperSockets" on page 140.

Table 8-18   Minimum support requirements for HiperSockets multiple write

Operating system	Support requirements
z/OS	z/OS V1R9 with PTFs

8.3.21 HiperSockets IPv6

IPv6 is expected to be a key element in future networking. The IPv6 support for HiperSockets permits compatible implementations between external networks and internal HiperSockets networks.

Table 8-19 lists the minimum support requirements for HiperSockets IPv6 (CHPID type IQD).

Table 8-19   Minimum support requirements for HiperSockets IPv6 (CHPID type IQD)

Operating system	Support requirements
z/OS	z/OS V1R7
z/VM	z/VM V5R4
Linux on System z	Novell SUSE SLES 10 SP2 Red Hat RHEL 5.2

8.3.22 HiperSockets Layer 2 support

For flexible and efficient data transfer for IP and non-IP workloads, the HiperSockets internal networks on z196 can support two transport modes, which are Layer 2 (Link Layer) and the current Layer 3 (Network or IP Layer). Traffic can be Internet Protocol (IP) Version 4 or Version 6 (IPv4, IPv6) or non-IP (AppleTalk, DECnet, IPX, NetBIOS, or SNA).

HiperSockets devices are protocol-independent and Layer 3 independent. Each HiperSockets device has its own Layer 2 Media Access Control (MAC) address, which allows the use of

applications that depend on the existence of Layer 2 addresses such as Dynamic Host Configuration Protocol (DHCP) servers and firewalls.

Layer 2 support can help facilitate server consolidation. Complexity can be reduced, network configuration is simplified and intuitive, and LAN administrators can configure and maintain the mainframe environment the same as they do a non-mainframe environment.

Table 8-20 show the requirements for HiperSockets Layer 2 support.

Table 8-20   Minimum support requirements for HiperSockets Layer 2

Operating system	Support requirements
z/VM	z/VM V5R4 for guest exploitation
Linux on System z	Novell SUSE SLES 10 SP2 Red Hat RHEL 5.2

8.3.23 HiperSockets network traffic analyzer for Linux on System z

HiperSockets network traffic analyzer (HS NTA) is an enhancement to HiperSockets architecture on z196, with support to trace Layer2 and Layer3 HiperSockets network traffic in Linux on System z. This allows Linux on System z to control the trace for the internal virtual LAN, to capture the records into host memory and storage (file systems).

Linux on System z tools can be used to format, edit, and process the trace records for analysis by system programmers and network administrators.

8.3.24 FICON Express8

FICON Express8 is the newest generation of FICON features. They provide a link rate of 8 Gbps, with autonegotiation to 4 or 2 Gbps, for compatibility with previous devices and investment protection. Both 10KM LX and SX connections are offered (in a given feature all connections must have the same type).

With FICON Express 8 customers may be able to consolidate existing FICON, FICON Express2 and FICON Express4 channels, while maintaining and enhancing performance.

Table 8-21 lists the minimum support requirements for FICON Express8.

Table 8-21   Minimum support requirements for FICON Express8

Operating system	z/OS	z/VM	z/VSE	Linux on System z	z/TPF
Native FICON and Channel-to-Channel (CTC) CHPID type FC	V1R7	V5R4	V4R1	Novell SUSE SLES 10 Red Hat RHEL 5	V1R1
zHPF single track operations CHPID type FC	V1R7[a]	NA	NA	NA	NA
zHPF multitrack operations CHPID type FC	V1R9[a]	NA	NA	NA	NA

| Support of SCSI devices CHPID type FCP | NA | V5R4 | V4R1 | Novell SUSE SLES 10 Red Hat RHEL 5 | NA |

a. PTFs required

8.3.25 z/OS discovery and autoconfiguration (zDAC)

z/OS discovery and autoconfiguration for FICON channels (zDAC) is designed to automatically perform a number of I/O configuration definition tasks for new and changed disk and tape controllers connected to a switch or director, when attached to a FICON channel.

The zDAC function is integrated into the existing Hardware Configuration Definition (HCD). Customers can define a policy which can include preferences for availability and bandwidth including parallel access volume (PAV) definitions, control unit numbers, and device number ranges. Then, when new controllers are added to an I/O configuration or changes are made to existing controllers, the system is designed to discover them and propose configuration changes based on that policy.

zDAC provides real-time discovery for the FICON fabric, subsystem and I/O device resource changes from z/OS. By exploring the discovered control units for defined logical control units (LCU) and devices, zDAC compares the discovered controller information with the current system configuration to determine delta changes to the configuration for a proposed configuration.

All new added or changed logical control units and devices will be added into the proposed configuration, with proposed control unit and device numbers, and channel paths base on the defined policy. zDAC use channel path chosen algorithm to minimize single point of failure. The zDAC proposed configurations are created as work I/O definition files (IODF), that can be converted to production IODF and activated.

zDAC is designed to perform discovery for all systems in a sysplex that support the function. Thus, zDAC helps simplifying I/O configuration on z196 servers running z/OS and reduces complexity and setup time.

zDAC applies to all FICON features supported on z196 when configured as CHPID type FC. Table 8-22 lists the minimum support requirements for zDAC.

Table 8-22   Minimum support requirements for zDAC

Operating system	Support requirements
z/OS	z/OS V1R12

8.3.26 High performance FICON (zHPF)

High performance FICON (zHPF), first provided on System z10, is a FICON architecture for protocol simplification and efficiency, reducing the number of information units (IUs) processed. Enhancements have been made to the z/Architecture and the FICON interface architecture to provide optimizations for on line transaction processing (OLTP) workloads.

When exploited by the FICON channel, the z/OS operating system, and the DS8000 control unit or other subsystems (new levels of Licensed Internal Code are required) the FICON channel overhead can be reduced and performance can be improved. Additionally, the

changes to the architectures provide end-to-end system enhancements to improve reliability, availability, and serviceability (RAS).

zHPF is compatible with:

► Fibre Channel Physical and Signaling standard (FC-FS)
► Fibre Channel Switch Fabric and Switch Control Requirements (FC-SW)
► Fibre Channel Single-Byte-4 (FC-SB-4) standards

The zHPF channel programs can be exploited, for instance, by z/OS OLTP I/O workloads; DB2, VSAM, PDSE and zFS.

At announcement zHPF supported the transfer of small blocks of fixed size data (4 K). This has been extended on z10 EC to multitrack operations (limited to 64 k bytes) and z196 removes the 64k byte data transfer limit on multitrack operations. This improvement allows the channel to fully exploit the bandwidth of FICON channels, results in higher throughputs and lower response times.

The multi-track operations extension applies exclusively to the FICON Express8s and FICON Express4s on z196, when configured as CHPID type FC, and connecting to z/OS. zHPF requires matching support by the DS8000 series, otherwise the extended multitrack support is transparent to the control unit.

From the z/OS point of view, the existing FICON architecture is called *command mode* and zHPF architecture is called *transport mode*. During link initialization, the channel node and the control unit node indicate whether they support zHPF.

> **Note:** All FICON channel paths (CHPIDs) defined to the same Logical Control Unit (LCU) must support zHPF. The inclusion of any non-compliant zHPF features in the path group will cause the entire path group to support command mode only.

The mode used for an I/O operation depends on the control unit supporting zHPF and settings in the z/OS operating system. For z/OS exploitation there is a parameter in the IECIOSxx member of SYS1.PARMLIB (ZHPF=YES or NO) and in the SETIOS system command to control whether zHPF is enabled or disabled. The default is ZHPF=NO.

Support is also added for the D IOS,ZHPF system command to indicate whether zHPF is enabled, disabled, or not supported on the server.

Similar to the existing FICON channel architecture, the application or access method provides the channel program (channel command words, CCWs). The way that zHPF (transport mode) manages channel program operations is significantly different from the CCW operation for the existing FICON architecture (command mode). While in command mode, each single CCW is sent to the control unit for execution. In transport mode, multiple channel commands are packaged together and sent over the link to the control unit in a single control block. Less overhead is generated compared to the existing FICON architecture. Certain complex CCW chains are not supported by zHPF.

The zHPF is exclusive to z196 and System z10. The FICON Express8 and FICON Express4[3] (CHPID type FC) concurrently support both the existing FICON protocol and the zHPF protocol in the server Licensed Internal Code.

Table 8-23 lists the minimum support requirements for zHPF.

---

[3] FICON Express4 10KM LX, 4KM LX and SX features are withdrawn from marketing. All FICON Express2 and FICON features are withdrawn from marketing.

Table 8-23   Minimum support requirements for zHPF

Operating system	Support requirements
z/OS	Single track operations: z/OS V1R7 with PTFs Multitrack operations: z/OS V1R10 with PTFs 64K enhancement: z/OS V1R10 with PTFs
z/VM	Not supported; not available to guests
Linux	SLES 11 SP1 supports zHPF. IBM continues to work with its Linux distribution partners on exploitation of appropriate z196 functions be provided in future Linux on System z distribution releases.

For more information about FICON channel performance, see the performance technical papers on the System z I/O connectivity Web site at:

http://www-03.ibm.com/systems/z/hardware/connectivity/ficon_performance.html

8.3.27 Request node identification data

First offered on z9 EC, the request node identification data (RNID) function for native FICON CHPID type FC allows isolation of cabling-detected errors.

Table 8-24 lists the minimum support requirements for RNID.

Table 8-24   Minimum support requirements for RNID

Operating system	Support requirements
z/OS	z/OS V1R7

8.3.28 Extended distance FICON

An enhancement to the industry standard FICON architecture (FC-SB-3) helps avoid degradation of performance at extended distances by implementing a new protocol for *persistent* information unit (IU) pacing. Extended distance FICON is transparent to operating systems and applies to all the FICON Express8 and FICON Express4 features carrying native FICON traffic (CHPID type FC).

For exploitation, the control unit must support the new IU pacing protocol. IBM System Storage DS8000 series supports extended distance FICON for IBM System z environments. The channel defaults to current pacing values when it operates with control units that cannot exploit extended distance FICON.

8.3.29 Platform and name server registration in FICON channel

The FICON Express8, FICON Express4 features on the z196 servers support platform and name server registration to the fabric for both CHPID type FC and FCP.

Information about the channels connected to a fabric, if registered, allows other nodes or storage area network (SAN) managers to query the name server to determine what is connected to the fabric.

The following attributes are registered for the z196 servers:

► Platform information
► Channel information

- ► World Wide Port Name (WWPN)
- ► Port type (N_Port_ID)
- ► FC-4 types supported
- ► Classes of service supported by the channel

The platform and name server registration service are defined in the Fibre Channel - Generic Services 4 (FC-GS-4) standard.

8.3.30 FICON link incident reporting

FICON link incident reporting allows an operating system image (without operator intervention) to register for link incident reports. Table 8-25 lists the minimum support requirements for this function.

Table 8-25   Minimum support requirements for link incident rreporting

Operating system	Support requirements
z/OS	z/OS V1R7

8.3.31 FCP provides increased performance

The Fibre Channel Protocol (FCP) Licensed Internal Code has been modified to help provide increased I/O operations per second for both small and large block sizes and to support 8 Gbps link speeds.

For more information about FCP channel performance, see the performance technical papers on the System z I/O connectivity Web site at:

http://www-03.ibm.com/systems/z/hardware/connectivity/fcp_performance.html

8.3.32 N_Port ID virtualization

N_Port ID virtualization (NPIV) provides a way to allow multiple system images (in logical partitions or z/VM guests) to use a single FCP channel as though each were the sole user of the channel. This feature, first introduced with z9 EC, can be used with earlier FICON features that have been carried forward from earlier servers.

Table 8-26 lists the minimum support requirements for NPIV.

Table 8-26   Minimum support requirements for NPIV

Operating system	Support requirements
z/VM	z/VM V5R4 provides support for guest operating systems and VM users to obtain virtual port numbers. Installation from DVD to SCSI disks is supported when NPIV is enabled.
z/VSE	z/VSE V4R1
Linux on System z	Novell SUSE SLES 10 SP3 Red Hat RHEL 5.4

8.3.33 OSA-Express3 10 Gigabit Ethernet LR and SR

The OSA-Express3 10 Gigabit Ethernet features offer two ports, defined as CHPID type OSD or OSX. CHPID type OSD supports the queued direct input/output (QDIO) architecture for high-speed TCP/IP communication. The z196 introduces the CHPID type OSX, see 8.3.39, "Intraensemble data network (IEDN)".

Table 8-27 lists the minimum support requirements for OSA-Express3 10 Gigabit Ethernet LR and SR features.

Table 8-27   Minimum support requirements for OSA-Express3 10 Gigabit Ethernet LR and SR

Operating system	Support requirements
z/OS	OSD: z/OS V1R7 OSX: z/OS V1R12; z/OS V1R10 and z/OS V1R11, with service
z/VM	OSD: z/VM V5R4 OSX: z/VM V5R4 for dynamic I/O only; z/VM V6R1 with service
z/VSE	OSD: z/VSE V4R1; service required
z/TPF	OSD: z/TPF V1R1
Linux on System z	OSD: Novell SUSE SLES 10 OSD: Red Hat RHEL 5

8.3.34 OSA-Express3 Gigabit Ethernet LX and SX

The OSA-Express3 Gigabit Ethernet features offer two cards with two PCI Express adapters each. Each PCI Express adapter controls two ports, giving a total of four ports per feature. Each adapter has its own CHPID, defined as either OSD or OSN, supporting the queued direct input/output (QDIO) architecture for high-speed TCP/IP communication. Thus, a single feature can support both CHPID types, with two ports for each type.

Operating system support is required in order to recognize and use the second port on each PCI Express adapter. Minimum support requirements for OSA-Express3 Gigabit Ethernet LX and SX features are listed in Table 8-28 (four ports) and Table 8-29 on page 236 (two ports).

Table 8-28   Minimum support requirements for OSA-Express3 Gigabit Ethernet LX and SX, four ports

Operating system	Support requirements when using four ports
z/OS	z/OS V1R8; service required
z/VM	z/VM V5R4; service required
z/VSE	z/VSE V4R1; service required
z/TPF	z/TPF V1R1; service required
Linux on System z	Novell SUSE SLES 10 SP2 Red Hat RHEL 5.2

Table 8-29   Minimum support requirements for OSA-Express3 Gigabit Ethernet LX and SX, two ports

Operating system	Support requirements when using two ports
z/OS	z/OS V1R7
z/VM	z/VM V5R4
z/VSE	z/VSE V4R1; service required
z/TPF	z/TPF V1R1
Linux on System z	Novell SUSE SLES 10 Red Hat RHEL 5

8.3.35 OSA-Express3 1000BASE-T Ethernet

The OSA-Express3 1000BASE-T Ethernet features offer two cards with two PCI Express adapters each. Each PCI Express adapter controls two ports, giving a total of four ports for each feature. Each adapter has its own CHPID, defined as one of OSC, OSD, OSE, OSM or OSN. A single feature can support two CHPID types, with two ports for each type. The OSM CHPID type is new with the z196, see 8.3.38, "Intranode management network (INMN)" on page 238.

Each adapter can be configured in the following modes:

► QDIO mode, with CHPID types OSD and OSN
► Non-QDIO mode, with CHPID type OSE
► Local 3270 emulation mode, including OSA-ICC, with CHPID type OSC
► Ensemble management, with CHPID type OSM.

Operating system support is required in order to recognize and use the second port on each PCI Express adapter. Minimum support requirements for OSA-Express3 1000BASE-T Ethernet feature are listed in Table 8-30 (four ports) and Table 8-32 on page 237.

Table 8-30   Minimum support requirements for OSA-Express3 1000BASE-T Ethernet, four ports

Operating system	Support requirements when using four ports[a,b]
z/OS	OSD: z/OS V1R8; service required OSE: z/OS V1R7 OSM: z/OS V1R12; z/OS V1R10 and z/OS V1R11, with service OSN[b]: z/OS V1R7
z/VM	OSD: z/VM V5R4; service required OSE: z/VM V5R4 OSM: z/VM V5R4 for dynamic I/O only; z/VM V6R1 with service OSN[b]: z/VM V5R4
z/VSE	OSD: z/VSE V4R1; service required OSE: z/VSE V4R1 OSN[b]: z/VSE V4R1; service required
z/TPF	OSD and OSN[b]: z/TPF V1R1; service required
Linux on System z	OSD: ► Novell SUSE SLES 10 SP2 ► Red Hat RHEL 5.2 OSN: ► Novell SUSE SLES 10 SP2 ► Red Hat RHEL 5.2

a. Applies to CHPID types OSC, OSD, OSE, OSM and OSN. For support, see Table 8-32 on page 237.

b. Although CHPID type OSN does not use any ports (because all communication is LPAR to LPAR), it is listed here for completeness.

8.3.36 OSA-Express2 1000BASE-T Ethernet

The OSA-Express2 1000BASE-T Ethernet adapter can be configured in:

▶ QDIO mode, with CHPID type OSD or OSN
▶ Non-QDIO mode, with CHPID type OSE
▶ Local 3270 emulation mode with CHPID type OSC

Table 8-31 lists the support for OSA-Express2 1000BASE-T.

Table 8-31   Minimum support requirements for OSA-Express2 1000BASE-T

Operating system	CHPID type OSC	CHPID type OSD	CHPID type OSE
z/OS V1R7	Supported	Supported	Supported
z/VM V5R4	Supported	Supported	Supported
z/VSE V4R1	Supported	Supported	Supported
z/TPF V1R1	Supported	Supported	Not supported
Linux on System z	Not supported	Supported	Not supported

Table 8-32 on page 237 lists the minimum support requirements for OSA-Express3 1000BASE-T Ethernet.

Table 8-32   Minimum support requirements for OSA-Express3 1000BASE-T Ethernet, two ports

Operating system	Support requirements when using two ports
z/OS	OSD, OSE, OSM and OSN; V1R7
z/VM	OSD, OSE, OSM and OSN: V5R4
z/VSE	V4R1
z/TPF	OSD, OSN, and OSC: V1R1
Linux on System z	OSD: ▶ Novell SUSE SLES 10 ▶ Red Hat RHEL 5 OSN: ▶ Novell SUSE SLES 10 SP3 ▶ Red Hat RHEL 5.4

8.3.37 Open System Adapter for Ensemble

The following three types of OSA-Express3 features are used to connect the z196 central processor complex (CPC) to its attached IBM System z BladeCenter Extension (zBX), and other ensemble nodes:

▶ OSA-Express3 10 Gigabit Ethernet (GbE) Long Range (LR), feature code 3370
▶ OSA-Express3 10 Gigabit Ethernet (GbE) Short Reach (SR), feature code 3371
▶ OSA Express3 1000BASE-T Gigabit Ethernet (GbE), feature code 3367

These connections are part of the ensemble's two private and secure internal networks.

For detailed information about OSA-Express3 in an ensemble network, see "zBX connectivity" on page 190.

8.3.38 Intranode management network (INMN)

The intranode management network (INMN) is one of the ensemble's two private and secure internal networks. INMN is used by the Unified Resource Manager functions.

The INMN is a private and physically isolated 1000Base-T ethernet internal platform management network, operating at 1 Gbps, that connect all resources (CPC and zBX components) of a zEnterprise ensemble node, for management purposes. It is prewired, internally switched, configured, and managed with fully redundancy for high availability.

The z196 introduces the OSA-Express3 OSA Direct-Express Management (OSM) CHPID type. INMN requires two OSA Express3 1000BASE-T ports, from two different OSA-Express3 1000Base-T features, configured as CHPID type OSM.

The OSA connection is via the Bulk Power Hub (BPH) port J07 on the z196 to the Top of the Rack (TORs) switches on zBX.

8.3.39 Intraensemble data network (IEDN)

The intraensemble Data Network (IEDN) is one of the ensemble's two private and secure internal networks. IEDN provides an application data exchanging path between ensemble nodes. More specifically it is used for communications across the virtualized images (LPARs, z/VM's virtual machines, and blades' LPARs).

The IEDN is a private and secure 10 Gbps ethernet network that connects all elements of a z196 ensemble and is access-controlled using integrated virtual LAN (VLAN) provisioning. No customer managed switches or routers are required. IEDN is managed by the primary HMC that controls the ensemble, helping to reduce the need of firewall and encryption, and simplifying network configuration and management, with full redundancy for high availability.

The z196 introduces the OSA-Express3 OSA Direct-Express zBX (OSX) CHPID type. The OSA connection is from the z196 to the Top of the Rack (TORs) switches on zBX.

IEDN requires two OSA Express3 10 GbE ports configured as CHPID type OSX.

8.3.40 OSA-Express3 and OSA-Express2 NCP support (OSN)

OSA-Express3 GbE, OSA-Express3 1000BASE-T Ethernet, OSA-Express2 GbE, and OSAExpress2 1000BASE-T Ethernet features can provide channel connectivity from an operating system in a z196 to IBM Communication Controller for Linux on System z (CCL) with the Open Systems Adapter for NCP (OSN), in support of the Channel Data Link Control (CDLC) protocol. OSN eliminates the requirement for an external communication medium for communications between the operating system and the CCL image.

With OSN, using an external ESCON channel is unnecessary. Data flow of the logical-partition to the logical-partition is accomplished by the OSA-Express3 or OSA-Express2 feature without ever exiting the card. OSN support allows multiple connections between the same CCL image and the same operating system (such z/OS or z/TPF). The operating system must reside in the same physical server as the CCL image.

For CCL planning information see *IBM Communication Controller for Linux on System z V1.2.1 Implementation Guide*, SG24-7223. For the most recent CCL information, see:

http://www-01.ibm.com/software/network/ccl/

Channel Data Link Control (CDLC), when used with the Communication Controller for Linux, emulates selected functions of IBM 3745/NCP operations. The port used with the OSN support appears as an ESCON channel to the operating system. This support can be used with OSA-Express3 GbE and 1000BASE-T, and OSA-Express2 GbE[4] and 1000BASE-T features.

Table 8-33 lists the minimum support requirements for OSN.

Table 8-33   Minimum support requirements for OSA-Express3 and OSA-Express2 OSN

Operating system	OSA-Express3 and OSA-Express2 OSN
z/OS	z/OS V1R7
z/VM	z/VM V5R4
z/VSE	z/VSE V4R1
Linux on System z	Novell SUSE SLES 10 SP3 Red Hat RHEL 5.4
z/TPF	z/TPF V1R1

8.3.41 Integrated Console Controller

The 1000BASE-T Ethernet features provide the Integrated Console Controller (OSA-ICC) function, which supports TN3270E (RFC 2355) and non-SNA DFT 3270 emulation. The OSA-ICC function uses a definition as CHIPD type OSC and console controller, and has multiple logical partitions support, both as shared or spanned channels.

With the OSA-ICC function, 3270 emulation for console session connections is integrated in the z196 through a port on the OSA-Express3 or OSA-Express2 1000BASE-T features. This function eliminates the requirement for external console controllers, such as 2074 or 3174, helping to reduce cost and complexity. Each port can support up to 120 console session connections.

OSA-ICC can be configured on a PCHID-by-PCHID basis and is supported at any of the feature settings (10, 100, or 1000 Mbps, half-duplex or full-duplex).

8.3.42 VLAN management enhancements

Table 8-34 lists minimum support requirements for VLAN management enhancements for the OSA-Express2 and OSA-Express features (CHPID type OSD).

Table 8-34   Minimum support requirements for VLAN management enhancements

Operating system	Support requirements
z/OS	z/OS V1R7
z/VM	z/VM V5R4. Support of guests is transparent to z/VM if the device is directly connected to the guest (pass through).

---

[4] OSA Express2 GbE is withdrawn from marketing.

8.3.43 GARP VLAN Registration Protocol

All OSA-Express3 and OSA-Express2 features support VLAN prioritization, a component of the IEEE 802.1 standard. GARP[5] VLAN Registration Protocol (GVRP) support allows an OSA-Express3 or OSA-Express2 port to register or unregister its VLAN IDs with a GVRP-capable switch and dynamically update its table as the VLANs change. This simplifies the network administration and management of VLANs as manually entering VLAN IDs at the switch is no longer necessary. Minimum support requirements are listed in Table 8-35.

Table 8-35   Minimum support requirements for GVRP

Operating system	Support requirements
z/OS	z/OS V1R7
z/VM	z/VM V5R4

8.3.44 Inbound workload queueing (IWQ) for OSA-Express3

OSA-Express-3 introduces inbound workload queueing (IWQ), which creates multiple input queues and allows OSA to differentiate workloads "off the wire" and then assign work to a specific input queue (per device) to z/OS.

With each input queue representing a unique type of workload, each having unique service and processing requirements, the IWQ function allows z/OS to preassign the appropriate processing resources for each input queue. This approach allows multiple concurrent z/OS processing threads to process each unique input queue (workload) avoiding traditional resource contention. In a heavily mixed workload environment, this "off the wire" network traffic separation provided by OSA-Express3 IWQ reduces the conventional z/OS processing required to identify and separate unique workloads, which results in improved overall system performance and scalability.

A primary objective of IWQ is to provide improved performance for business critical interactive workloads by reducing contention created by other types of workloads. The types of z/OS workloads that are identified and assigned to unique input queues are:

► z/OS Sysplex Distributor traffic—network traffic which is associated with a distributed virtual internet protocol address (VIPA) is assigned to a unique input queue allowing the Sysplex Distributor traffic to be immediately distributed to the target host

► z/OS bulk data traffic—network traffic which is dynamically associated with a streaming (bulk data) TCP connection is assigned to a unique input queue allowing the bulk data processing to be assigned the appropriate resources and isolated from critical interactive workloads.

IWQ is exclusive to OSA-Express3 CHPID type OSD and the z/OS operating system. This applies to z196 and System z10. Minimum support requirements are listed in Table 8-36.

Table 8-36   Minimum support requirements for IWQ

Operating system	Support requirements
z/OS	z/OS V1R12
z/VM	z/VM V5R4 for guest exploitation only; service required

---

[5] Generic Attribute Registration Protocol

8.3.45 Query and display OSA configuration

OSA-Express3 introduces the capability for the operating system to directly query and display the current OSA configuration information (similar to OSA/SF). z/OS exploits this new OSA capability by introducing a new TCP/IP operator command called `display OSAINFO`.

Display OSAINFO allows the operator to monitor and verify the current OSA configuration which will help to improve the overall management, serviceability, and usability of OSA-Express3.

Display OSAINFO is exclusive to z/OS and applies to OSA-Express3 CHPID types OSD, OSM, and OSX.

8.3.46 Link aggregation support for z/VM

Link aggregation (IEEE 802.3ad) controlled by the z/VM Virtual Switch (VSWITCH) allows the dedication of an OSA-Express3 or OSA-Express2 port to the z/VM operating system, when the port is participating in an aggregated group configured in Layer 2 mode. Link aggregation (trunking) is designed to allow combining multiple physical OSA-Express3 or OSA-Express2 ports into a single logical link for increased throughput and for nondisruptive failover in the event that a port becomes unavailable. The target links for aggregation must be of the same type.

Link aggregation is applicable to the OSA-Express3, and OSA-Express2 features when configured as CHPID type OSD (QDIO). Link aggregation is supported by z/VM V5R4.

8.3.47 QDIO data connection isolation for z/VM

The Queued Direct I/O (QDIO) data connection isolation function provides a higher level of security when sharing the same OSA connection in z/VM environments that use the Virtual Switch (VSWITCH). The VSWITCH is a virtual network device that provides switching between OSA connections and the connected guest systems.

QDIO data connection isolation allows disabling internal routing for each QDIO connected, and provides a means for creating security zones and preventing network traffic between the zones.

VSWITCH isolation support is provided by APAR VM64281. z/VM 5R4 and later support is provided by CP APAR VM64463 and TCP/IP APAR PK67610.

QDIO data connection isolation is supported by all OSA-Express3 and OSA-Express2 features on z196.

8.3.48 QDIO interface isolation for z/OS

Some environments require strict controls for routing data traffic between severs or nodes. In certain cases, the LPAR-to-LPAR capability of a shared OSA connection can prevent such controls from being enforced. With interface isolation, internal routing can be controlled on an LPAR basis. When interface isolation is enabled, the OSA will discard any packets destined for a z/OS LPAR that is registered in the OAT as isolated.

QDIO interface isolation is supported by Communications Server for z/OS V1R11 and all OSA-Express3 and OSA-Express2 features on z196.

8.3.49 QDIO optimized latency mode

QDIO optimized latency mode (OLM) can help improve performance for applications that have a critical requirement to minimize response times for inbound and outbound data.

OLM optimizes the interrupt processing as follows:

► For inbound processing, the TCP/IP stack looks more frequently for available data to process, ensuring that any new data is read from the OSA-Express3 without requiring additional program controlled interrupts (PCIs).

► For outbound processing, the OSA-Express3 also looks more frequently for available data to process from the TCP/IP stack, thus not requiring a Signal Adapter (SIGA) instruction to determine whether more data is available.

8.3.50 Checksum offload for IPv4 packets when in QDIO mode

A function referred to as *checksum offload*, supports z/OS and Linux on System z environments. It is offered on the OSA-Express3 GbE, OSA-Express3 100BASE-T Ethernet, OSA-Express2 GbE, and OSA-Express2 1000BASE-T Ethernet features. Checksum offload provides the capability of calculating the Transmission Control Protocol (TCP), User Datagram Protocol (UDP), and Internet Protocol (IP) header checksum. Checksum verifies the accuracy of files. By moving the checksum calculations to a Gigabit or 1000BASE-T Ethernet feature, host CPU cycles are reduced and performance is improved.

When checksum is offloaded, the OSA-Express feature performs the checksum calculations for Internet Protocol Version 4 (IPv4) packets. The checksum offload function applies to packets that go to or come from the LAN. When multiple IP stacks share an OSA-Express, and an IP stack sends a packet to a next hop address owned by another IP stack that is sharing the OSA-Express, the OSA-Express then sends the IP packet directly to the other IP stack without placing it out on the LAN. Checksum offload does not apply to such IP packets.

Checksum offload is supported by the GbE features (FC 3362, FC 3363, FC 3364, and FC 3365) and the 1000BASE-T Ethernet features (FC 3366 and FC 3367) when operating at 1000 Mbps (1 Gbps). Checksum offload is applicable to the QDIO mode only (channel type OSD).

z/OS support for checksum offload is available in all in-service z/OS releases, and in all supported Linux on System z distributions.

8.3.51 Adapter interruptions for QDIO

Linux on System z and z/VM work together to provide performance improvements by exploiting extensions to the Queued Direct I/O (QDIO) architecture. Adapter interruptions, first added to z/Architecture with HiperSockets, provide an efficient, high-performance technique for I/O interruptions to reduce path lengths and overhead in both the host operating system and the adapter (OSA-Express3 and OSA-Express2 when using type OSD CHPID).

In extending the use of adapter interruptions to OSD (QDIO) channels, the programming overhead to process a traditional I/O interruption is reduced. This benefits OSA-Express TCP/IP support in Linux on System z, z/VM and z/VSE.

Adapter interruptions apply to all of the OSA-Express3 and OSA-Express2 features on z196 when in QDIO mode (CHPID type OSD).

8.3.52 OSA Dynamic LAN idle

OSA Dynamic LAN idle parameter change helps reduce latency and improve performance by dynamically adjusting the inbound blocking algorithm. System administrators can authorize the TCP/IP stack to enable a dynamic setting, which was previously a static setting.

For latency-sensitive applications, the blocking algorithm is modified to be *latency sensitive*. For streaming (throughput-sensitive) applications, the blocking algorithm is adjusted to maximize throughput. In all cases, the TCP/IP stack determines the best setting based on the current system and environmental conditions (inbound workload volume, processor utilization, traffic patterns, and so on) and can dynamically update the settings. OSA-Express3 and OSA-Express2 features adapt to the changes, avoiding thrashing and frequent updates to the OSA address table (OAT). Based on the TCP/IP settings, OSA holds the packets before presenting them to the host. A dynamic setting is designed to avoid or minimize host interrupts.

OSA Dynamic LAN idle is supported by the OSA-Express2 and OSA-Express3 features on z196 when in QDIO mode (CHPID type OSD), and is exploited by z/OS V1.8 (or higher) with program temporary fixes (PTFs).

8.3.53 OSA Layer 3 Virtual MAC for z/OS environments

To help simplify the infrastructure and to facilitate load balancing when a logical partition is sharing the same OSA Media Access Control (MAC) address with another logical partition, each operating system instance can have its own unique *logical* or *virtual* MAC (VMAC) address. All IP addresses associated with a TCP/IP stack are accessible by using their own VMAC address, instead of sharing the MAC address of an OSA port, which also applies to Layer 3 mode and to an OSA port spanned among channel subsystems.

OSA Layer 3 VMAC is supported by the OSA-Express2 and OSA-Express3 features on z196 when in QDIO mode (CHPID type OSD), and is exploited by z/OS V1R8 and later.

8.3.54 QDIO Diagnostic Synchronization

QDIO Diagnostic Synchronization enables system programmers and network administrators to coordinate and simultaneously capture both software and hardware traces. It allows z/OS to signal an OSA-Express3 or OSA-Express2 feature (by using a diagnostic assist function) to stop traces and capture the current trace records.

QDIO Diagnostic Synchronization is supported by the OSA-Express2 and OSA-Express3 features on z196 when in QDIO mode (CHPID type OSD), and is exploited by z/OS V1R8 and later.

8.3.55 Network Traffic Analyzer

With the large volume and complexity of today's network traffic, the z196 offers systems programmers and network administrators the ability to more easily solve network problems. With the availability of the OSA-Express Network Traffic Analyzer and QDIO Diagnostic Synchronization on the server, you can capture trace and trap data, and forward it to z/OS tools for easier problem determination and resolution.

The Network Traffic Analyzer is supported by the OSA-Express2 and OSA-Express3 features on z196 when in QDIO mode (CHPID type OSD), and is exploited by z/OS V1R8 and later.

8.3.56 Program directed re-IPL

First available on System z9, program directed re-IPL allows an operating system on a z196 to re-IPL without operator intervention. This function is supported for both SCSI and ECKD™ devices. Table 8-37 lists the minimum support requirements for program directed re-IPL.

Table 8-37   Minimum support requirements for program directed re-IPL

Operating system	Support requirements
z/VM	z/VM V5R4
Linux on System z	Novell SUSE SLES 10 SP3 Red Hat RHEL 5.4
z/VSE	V4R1 on SCSI disks

8.3.57 Coupling over InfiniBand

InfiniBand technology can potentially provide high-speed interconnection at short distances, longer distance fiber optic interconnection, and interconnection between partitions on the same system without external cabling. Several areas of this book discuss InfiniBand characteristics and support. For example, see 4.8, "Parallel Sysplex connectivity" on page 141.

InfiniBand coupling links

Table 8-38 lists the minimum support requirements for coupling links over InfiniBand.

Table 8-38   Minimum support requirements for coupling links over InfiniBand

Operating system	Support requirements
z/OS	z/OS V1R7
z/VM	z/VM V5R4 (dynamic I/O support for InfiniBand CHPIDs only; coupling over InfiniBand is not supported for guest use)
z/TPF	z/TPF V1R1

InfiniBand coupling links at an unrepeated distance of 10 km

Support for HCA2-O LR fanout supporting InfiniBand coupling links (1x IB-SDR or 1x IB-DDR) at an unrepeated distance of 10 KM (6.2 miles) is listed on Table 8-39.

Table 8-39   Minimum support requirements for coupling links over InfiniBand at 10 km

Operating system	Support requirements
z/OS	z/OS V1R8; service required
z/VM	z/VM V5R4 (dynamic I/O support for InfiniBand CHPIDs only; coupling over InfiniBand is not supported for guest use)

8.3.58 Dynamic I/O support for InfiniBand CHPIDs

This function refers exclusively to the z/VM dynamic I/O support of InfiniBand coupling links. Support is available for the CIB CHPID type in the z/VM dynamic commands, including the `change channel path` dynamic I/O command. Specifying and changing the system name

when entering and leaving configuration mode is also supported. z/VM does not use InfiniBand and does not support the use of InfiniBand coupling links by guests.

Table 8-40 lists the minimum support requirements of dynamic I/O support for InfiniBand CHPIDs.

Table 8-40   Minimum support requirements for dynamic I/O support for InfiniBand CHPIDs

Operating system	Support requirements
z/VM	z/VM V5R4

8.4 Cryptographic support

z196 provides two major groups of cryptographic functions:

► Synchronous cryptographic functions, provided by the CP Assist for Cryptographic Function (CPACF)
► Asynchronous cryptographic functions, provided by the Crypto Express3 feature

The minimum software support levels are listed in the following sections. Obtain and review the most recent Preventive Service Planning (PSP) buckets to ensure that the latest support levels are known and included as part of the implementation plan.

8.4.1 CP Assist for Cryptographic Function

In z196, the CP Assist for Cryptographic Function (CPACF) supports the full standard for Advanced Encryption Standard (AES, symmetric encryption) and secure hash algorithm (SHA, hashing). For a detailed description, see 6.3, "CP Assist for Cryptographic Function" on page 168. Support for this function is provided through a Web deliverable. Table 8-41 lists the support requirements for enhanced CPACF.

Table 8-41   Support requirements for enhanced CPACF

Operating system	Support requirements
z/OS[a]	z/OS V1R7 and later: The function varies by release. Protected key requires z/OS V1R9 and higher plus PTFs.
z/VM	z/VM V5R4 and higher: Supported for guest use. Protected key not supported.
z/VSE	z/VSE V4R1 and later, and IBM TCP/IP for VSE/ESA V1R5 with PTFs
Linux on System z	Novell SUSE SLES 10 and SLES 11 Red Hat RHEL 5 The z10 EC CPACF enhancements can be used with: ► Novell SUSE SLES 10 SP2 and SLES 11 ► Red Hat RHEL 5.2
z/TPF	z/TPF V1R1

a. CPACF is also exploited by several IBM Software product offerings for z/OS, such as IBM WebSphere Application Server for z/OS.

8.4.2 Crypto Express3

Support of Crypto Express3 functions varies by operating system and release. Table 8-42 lists the minimum software requirements for the Crypto Express3 features when configured

as a coprocessor or an accelerator. For a full description, see 6.4, "Crypto Express3" on page 169.

Table 8-42   Crypto Express2 and Crypto Express3 support on z196

Operating system	Crypto Express3
z/OS	V1R12: Web deliverable V1R11: Web deliverable V1R10: Web deliverable V1R9: Web deliverable V1R8: Not supported V1R7: Not supported
z/VM	V5R1: Service required; supported for guest use only V5R4: Service required; supported for guest use only
z/VSE	V4R2 with IBM TCP/IP for VSE/ESA V1R5. Service required
Linux on System z	Note[a] Novell SUSE SLES 11 Novell SUSE SLES 10 SP3 Red Hat RHEL 5.4
z/TPF V1R1	Service required (accelerator mode only)

a. Support for Crypto Express3 is provided at the same functional level as for Crypto Express2

8.4.3 Web deliverables

For Web-delivered code on z/OS, see the z/OS downloads :

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

For Linux on System z, support is delivered through IBM and distribution partners. For more information see Linux on System z on the developerWorks Web site:

http://www.ibm.com/developerworks/linux/linux390/

8.4.4 z/OS ICSF FMIDs

Integrated Cryptographic Service Facility (ICSF) is a component of z/OS, and is designed to transparently use the available cryptographic functions, whether CPACF or Crypto Express3 to balance the workload and help address the bandwidth requirements of the applications.

For a list of ICSF versions and FMID cross-references, see the Technical Documents page:

http://www.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/TD103782

Table 8-43 on page 247 lists the ICSF FMIDs and Web-delivered code for z/OS V1R7 through V1R12.

Table 8-43   z/OS ICSF FMIDs

z/OS	ICSF FMID[a]	Web deliverable name	Supported function
V1R7	HCR7731	Enhancements for cryptographic support for z/OS V1R6 and V1R7 (Web deliverable)	▸ PCI-X Adapter Coprocessor and Accelerator ▸ CPACF enhancements ▸ Remote Key Loading ▸ ISO 16609 CBC Mode TDES MAC
	HCR7750	Enhancements for Cryptographic support for z/OS V1R7 through z/OS V1R9 (Web deliverable)	▸ Cryptographic exploitation ▸ 4096-bit RSA keys ▸ CPACF support for SHA-384 and 512 ▸ Reduced support for retained private key in ICSF
V1R7 and V1R8	HCR7731	Enhancements for Cryptographic support for z/OS V1R6 and V1R7 (included in base)	▸ PCI-X Adapter Coprocessor and Accelerator ▸ CPACF enhancements ▸ Remote Key Loading and ISO 16609 CBC Mode triple DES MAC
	HCR7750	Enhancements for Cryptographic Support for z/OS V1R7 through z/OS V1R9 (Web deliverable)	▸ Cryptographic exploitation z10 BC ▸ 4096-bit RSA keys ▸ CPACF support for SHA-384 and 512 ▸ Reduced support for retained private key in ICSF
	HCR7751	Cryptographic Support for z/OS V1R8-V1R10 and z/OS.e V1R8 (Web deliverable)	▸ Secure key AES ▸ 13 through 19-digit personal account number data ▸ Crypto Query service ▸ Enhanced SAF checking
V1R9	HCR7740	Cryptographic support for z/OS V1R7 through z/OS V1R9 (included in base)	▸ Cryptographic toleration z10 BC
	HCR7750	Enhancements for Cryptographic support for z/OS V1R7 through z/OS V1R9 (Web deliverable)	▸ Cryptographic exploitation z10 BC ▸ 4096-bit RSA keys ▸ CPACF support for SHA-384 and 512 ▸ Reduced support for retained private key in ICSF
	HCR7751	Cryptographic Support for z/OS V1R8 through V1R10 and z/OS.e V1R8 (Web deliverable)	▸ Secure key AES ▸ 13 through 19-digit personal account number data ▸ Crypto Query service ▸ Enhanced SAF checking
	HCR7770	Cryptographic support for z/OS V1R9 through V1R11 (Web deliverable)	▸ Protected key for CPACF ▸ Crypto Express3 and Crypto Express3-1P

z/OS	ICSF FMID[a]	Web deliverable name	Supported function
V1R10	HCR7750	Enhancements for Cryptographic support for z/OS V1R7 through z/OS V1R9 (included in base)	▶ Cryptographic exploitation z10 BC ▶ 4096-bit RSA keys ▶ CPACF support for SHA-384 and 512 ▶ Reduced support for retained private key in ICSF
	HCR7751	Cryptographic Support for z/OS V1R8 through V1R11 and z/OS.e V1R8 (Web deliverable)	▶ Secure key AES ▶ 13 through 19-digit personal account number data ▶ New Crypto Query service ▶ Enhanced SAF checking
	HCR7770	Cryptographic support for z/OS V1R9 through V1R11 (Web deliverable)	▶ Protected key for CPACF ▶ Crypto Express3 and Crypto Express3-1P
	HCR7780	Cryptographic support for z/OS V1R10 through V1R12 (Web deliverable)	▶ X9.8 Pin, 64 Bit, z196 CPACF, HMAC, CKDS Constraint Relief, AP Interrupt, PCI Audit, ECC HW Support, CBC Key Wrap
V1R11	HCR7751	Cryptographic Support for z/OS V1R8 through V1R11 and z/OS.e V1R8 (included in base)	▶ Secure key AES ▶ 13 through 19-digit personal account number data ▶ New Crypto Query service ▶ Enhanced SAF checking
	HCR7770	Cryptographic support for z/OS V1R9 through V1R11 (Web deliverable)	▶ Protected key for CPACF ▶ Crypto Express3 and Crypto Express3-1P
	HCR7780	Cryptographic support for z/OS V1R10 through V1R12 (Web deliverable)	▶ X9.8 Pin, 64 Bit, z196 CPACF, HMAC, CKDS Constraint Relief, AP Interrupt, PCI Audit, ECC HW Support, CBC Key Wrap
V1R12	HCR7770	Cryptographic support for z/OS V1R9 through V1R11 (included in base)	▶ Protected key for CPACF ▶ Crypto Express3 and Crypto Express3-1P
	HCR7780	Cryptographic support for z/OS V1R10 through V1R12 (Web deliverable)	▶ X9.8 Pin, 64 Bit, z196 CPACF, HMAC, CKDS Constraint Relief, AP Interrupt, PCI Audit, ECC HW Support, CBC Key Wrap

a. PTF information is located in z10 EC PSP bucket: upgrade 2097DEVICE, subset 2097/ZOS.

Note the following FMID information:

▶ FMID HCR7730 is available as a Web download for z/OS V1R7

▶ FMID HCR7731 is available as a Web download for z/OS V1R8 in support of the PCI-X cryptographic coprocessor and accelerator functions, and the CPACF AES, PRNG, and SHA support.

▶ FMID HCR7740 is integrated in the base of z/OS V1R9, so no download is necessary.

▶ FMID HCR7750 must be downloaded and installed for support of the SHA-384 and SHA-512 function on z/OS V1R7, V1R8, and V1R9.

- ► FMID HCR7751, which is available for z/OS V1R8 and later, supports functions such as Secure Key AES, Crypto Query Service, enhanced IPv6 support, and enhanced SAF Checking and Personal Account Numbers with 13-19 digits.
- ► FMID HCR7770, which is available for z/OS V1R9 and later, supports Crypto Express3, Crypto Express3-1P and CPACF protected key.
- ► FMID HCR7780, which has a planned availability date of September 2010, provides support for z/OS V1R10 and later, including CPACF new message cipher instructions and new codes for message digest, and numerous Crypto Express3 functions.

8.4.5 ICSF migration considerations

Consider the following points about the Web-delivered ICSF code:

- ► Increased size of the PKDS file is required in order to allow 4096-bit RSA keys to be stored.

   If you use the PKDS for asymmetric keys, copy your PKDS to a larger VSAM data set before using the new version of ICSF. The ICSF options file must be updated with the name of the new data set. ICSF can then be started.

   A toleration PTF must be installed on any system that is sharing the PKDS with a system running HCR7750 ICSF. The PTF allows the PKDS to be larger and prevents any service from accessing 4096-bit keys stored in a HCR7750 PKDS.

- ► Support is reduced for retained private keys.

   Applications that make use of the retained private key capability for key management are no longer able to store the private key in the cryptographic coprocessor card. The applications will continue to be able to list the retained keys and to delete them from the cryptographic coprocessor cards.

8.5 z/OS migration considerations

With the exception of base processor support, z/OS software changes do not require the new z196 functions. Equally, the new functions do not require functional software. The approach has been, where applicable, to let z/OS automatically decide to enable a function based on the presence or absence of the required hardware and software.

8.5.1 General recommendations

The zEnterprise 196 introduces the latest System z technology. Although support is provided by z/OS starting with z/OS V1R7, exploitation of z196 is dependent on the z/OS release. The z/OS.e is *not* supported on z196.

In general, we have the following recommendations:

- ► Do not migrate software releases and hardware at the same time.
- ► Keep members of sysplex at same software level, except during brief migration periods.
- ► Review z196 restrictions and migration considerations prior to creating an upgrade plan.

8.5.2 HCD

When using the hardware configuration definition (HCD) on z/OS V1R6 to create a definition for z196, *all* subchannel sets must be defined or the VALIDATE task can fail. On z/OS V1R7, HCD or the Hardware Configuration Manager (HCM) assist in the definitions.

8.5.3 InfiniBand coupling links

Each system can use, or not use, InfiniBand coupling links independently of what other systems are doing, and do so in conjunction with other link types.

InfiniBand coupling connectivity can only be obtained with other systems that also support InfiniBand coupling.

8.5.4 Large page support

The large page support function must not be enabled without the software support. If large page is not specified, page frames are allocated at the current size of 4 K.

In z/OS V1R9 and later, the amount of memory to be reserved for large page support is defined by using parameter LFAREA in the IEASYSxx member of SYS1.PARMLIB, as follows:

`LFAREA=xx%│xxxxxxM│xxxxxxG`

The parameter indicates the amount of storage, in percentage, megabytes, or gigabytes. The value cannot be changed dynamically.

8.5.5 HiperDispatch

The HIPERDISPATCH=YES/NO parameter in the IEAOPTxx member of SYS1.PARMLIB and on the SET OPT=xx command can control whether HiperDispatch is enabled or disabled for a z/OS image. It can be changed dynamically, without an IPL or any outage.

The default is that HiperDispatch is disabled on all releases, from z/OS V1R7 (requires PTFs for zIIP support) through z/OS V1R12.

To effectively exploit HiperDispatch, the Workload Manager (WLM) goal adjustment might be required. We recommend that you review WLM policies and goals, and update them as necessary. You may want to run with the new policies and HiperDispatch on for a period, turn it off and use the older WLM policies while analyzing the results of using HiperDispatch, re-adjust the new policies and repeat the cycle, as needed. In order to change WLM policies, turning HiperDispatch off then on is not necessary.

A health check is provided to verify whether HiperDispatch is enabled on a system image that is running on z196.

8.5.6 Capacity Provisioning Manager

Installation of the capacity provision function on z/OS requires:

► Setting up and customizing z/OS RMF, including the Distributed Data Server (DDS)
► Setting up the z/OS CIM Server (included in z/OS base because V1R7)
► Performing capacity provisioning customization as described in the publication *z/OS MVS Capacity Provisioning User's Guide*, SA33-8299

Exploitation of the capacity provisioning function requires:

► TCP/IP connectivity to observed systems.
► RMF Distributed Data Server must be active.
► CIM server must be active.
► Security and CIM customization.
► Capacity Provisioning Manager customization.

In addition, the Capacity Provisioning Control Center has to be downloaded from the host and installed on a PC server. This application is only used to define policies. It is not required for regular operation.

Customization of the capacity provisioning function is required on the following systems:

► Observed z/OS systems. These are the systems in one or multiple sysplexes that are to be monitored. For a description of the capacity provisioning domain, see 9.8, "Nondisruptive upgrades" on page 303.

► Runtime systems. These are the systems where the Capacity Provisioning Manager is running, or to which the server can fail over after server or system failures.

8.5.7 Decimal floating point and z/OS XL C/C++ considerations

The following two C/C++ compiler options require z/OS V1R9:

► The ARCHITECTURE option, which selects the minimum level of machine architecture on which the program will run. Note that certain features provided by the compiler require a minimum architecture level. ARCH(8) and ARCH(9) exploit instructions available respectively on the z10 EC and z196.

► The TUNE option, which allows optimization of the application for a specific machine architecture, within the constraints imposed by the ARCHITECTURE option. The TUNE level must not be lower than the setting in the ARCHITECTURE option.

For more information about the ARCHITECTURE and TUNE compiler options, see the *z/OS V1R9.0 XL C/C++ User's Guide*, SC09-4767.

> **Note:** The ARCHITECTURE or TUNE options for C++ programs should be used if the same applications should run on both the z196 as well as on previous System z servers. However, if C++ applications will only run on z196 servers the latest ARCHITECTURE and TUNE options should be used assuring that the best performance possible is delivered through the latest instruction set additions.

8.6 Coupling facility and CFCC considerations

Coupling facility connectivity to a z196 is supported on the z10EC, z10 BC, z9 EC, z9 BC, or another z196. The logical partition running the Coupling Facility Control Code (CFCC) can reside on any of the supported servers previously listed. See Table 8-44 on page 253 for Coupling Facility Control Code requirements for supported servers.

> **Note:** Because coupling link connectivity to z890, z990, and previous servers is *not* supported, this could affect the introduction of z196 into existing installations, and require additional planning. Also consider the level of CFCC. For more information, see "Coupling link migration considerations" on page 146.

The initial support of the CFCC on the z196 is level 17. CFCC level 17 is available and is exclusive to z196. CFCC level 17 offers the following enhancements:

▶ Availability improvements with non-disruptive CFCC dumps

The Coupling Facility is now designed to collect a serialized, time-consistent dump without disrupting CFCC operations. This improves serviceability, availability and system management for the CF images participating in a Parallel Sysplex.

▶ Scalability improvements with up to 2047 structures

CFCC Level 17 increases the number of structures that can be allocated in a CFCC image from 1023 to 2047. Allowing more CF structures to be define and used in a sysplex permits more discrete data sharing groups to operate concurrently, and can help environments requiring many structure to be defined, such as to support SAP or service providers.

> **Note:** Having more than 1024 structures requires a new version of the CFRM CDS. In addition, all systems in the sysplex need to be at z/OS V1R12 or have the coexistence/preconditioning PTFs installed. Falling back to a previous level, without the coexistence PTF installed, is *not supported without at sysplex IPL.*

▶ Increased number of lock and list structure connectors

z196 supports 247 connectors to a lock structure and 127 connectors to list structure, up from 32. This can specifically help many IMS and DB2 environments where the subsystems can now be split to provide virtual storage and thread constraint reduction. IBM has supported 255 connector to a cache structure for several years.

Enhancements available with previous level (CFCC level 16):

▶ CF Duplexing enhancements

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

▶ List notification improvements

Prior to CFCC level 16, when a list changed state from empty to non-empty, it notified its connectors. The first one to respond would read the new message, but when the others read, they would find nothing, paying the cost for the *false scheduling*.

CFCC level 16 can help improve CPU utilization for IMS Shared Queue and WebSphere MQ Shared Queue environments. The coupling facility only notifies one connector in a round-robin fashion. If the shared queue is read within a fixed period of time, the other connectors do not have to be notified, saving the cost of the false scheduling. If a list is not read within the time limit, then the other connectors are notified as they are prior to CFCC level 16.

As storage requirements may increase when moving to CFCC level 17, we strongly recommend using the CFSizer Tool, located on the Web at:

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

z196 servers with CFCC level 17 require z/OS V1R7 or later, and z/VM V5R4 or later for guest virtual coupling.

The current CFCC level for z196 servers is CFCC level 17, see Table 8-44. To support migration from one CFCC level to the next, different levels of CFCC can be run concurrently

while the coupling facility logical partitions are running on different servers (CF logical partitions running on the same server share the same CFCC level).

Table 8-44   System z CFCC code level considerations

z196	CFCC level 17 or later
z10 EC or z10 BC	CFCC level 15 or later
z9 EC or z9 BC	CFCC level 14 or later
z990 or z890	CFCC level 13 or later

Previous to migration, installation of compatibility/coexistence PTFs is highly recommended.A planned outage is required when migrating the CF or the CF LPAR to CFCC level 17.

For additional details about CFCC code levels, see the Parallel Sysplex Web site:

http://www.ibm.com/systems/z/pso/cftable.html

8.7 MIDAW facility

The modified indirect data address word (MIDAW) facility is a system architecture and software exploitation designed to improve FICON performance. This facility was first made available on System z9 servers and is exploited by the media manager in z/OS.

The MIDAW facility provides a more efficient CCW/IDAW structure for certain categories of data-chaining I/O operations:

► MIDAW can significantly improve FICON performance for extended format data sets. Non-extended data sets can also benefit from MIDAW.

► MIDAW can improve channel utilization and can significantly improve I/O response time. It reduces FICON channel connect time, director ports, and control unit overhead.

IBM laboratory tests indicate that applications using EF data sets, such as DB2, or long chains of small blocks can gain significant performance benefits by using the MIDAW facility.

MIDAW is supported on ESCON channels configured as CHPID type CNS and on FICON channels configured as CHPID types FC.

8.7.1 MIDAW technical description

An indirect address word (IDAW) is used to specify data addresses for I/O operations in a virtual environment.[6] The existing IDAW design allows the first IDAW in a list to point to any address within a page. Subsequent IDAWs in the same list must point to the first byte in a page. Also IDAWs (except the first and last IDAW) in a list must deal with complete 2 K or 4 K units of data. Figure 8-1 on page 254 shows a single channel command word (CCW) to control the transfer of data that spans non-contiguous 4 K frames in main storage. When the IDAW flag is set, the data address in the CCW points to a list of words (IDAWs), each of which contains an address designating a data area within real storage.

---

[6] There are exceptions to this statement and we skip a number of details in the following description. We assume that the reader can merge this brief description with an existing understanding of I/O operations in a virtual memory environment.
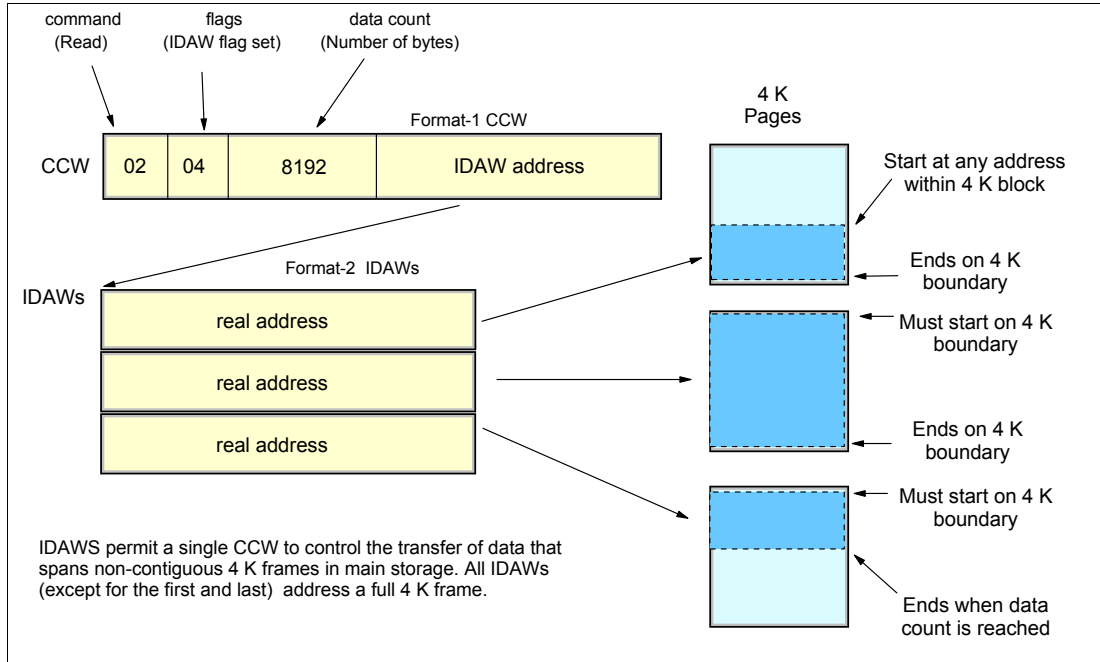
Figure 8-1   IDAW usage

The number of IDAWs required for a CCW is determined by the IDAW format as specified in the operation request block (ORB), by the count field of the CCW, and by the data address in the initial IDAW. For example, three IDAWS are required when the following three events occur:

1. The ORB specifies format-2 IDAWs with 4 KB blocks.
2. The CCW count field specifies 8 KB.
3. The first IDAW designates a location in the middle of a 4 KB block.

CCWs with *data chaining* may be used to process I/O data blocks that have a more complex internal structure, in which portions of the data block are directed into separate buffer areas (this is sometimes known as scatter-read or scatter-write). However, as technology evolves and link speed increases, data chaining techniques are becoming less efficient in modern I/O environments for reasons involving switch fabrics, control unit processing and exchanges, and others.

The MIDAW facility is a method of gathering and scattering data from and into discontinuous storage locations during an I/O operation. The modified IDAW (MIDAW) format is shown in Figure 8-2. It is 16 bytes long and is aligned on a quadword.
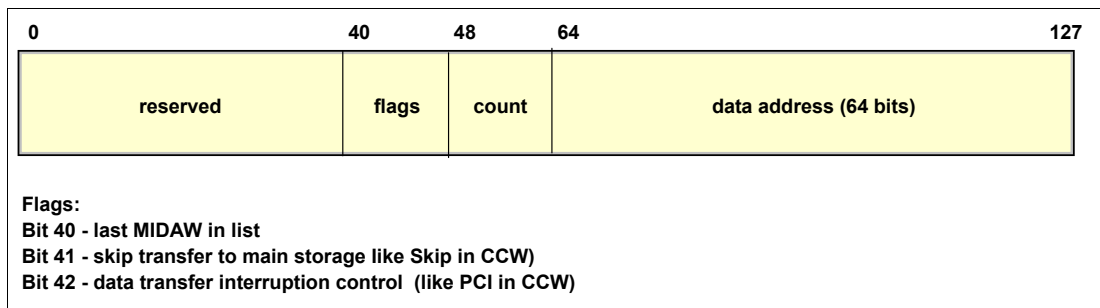
Figure 8-2   MIDAW format

An example of MIDAW usage is shown in Figure 8-3.
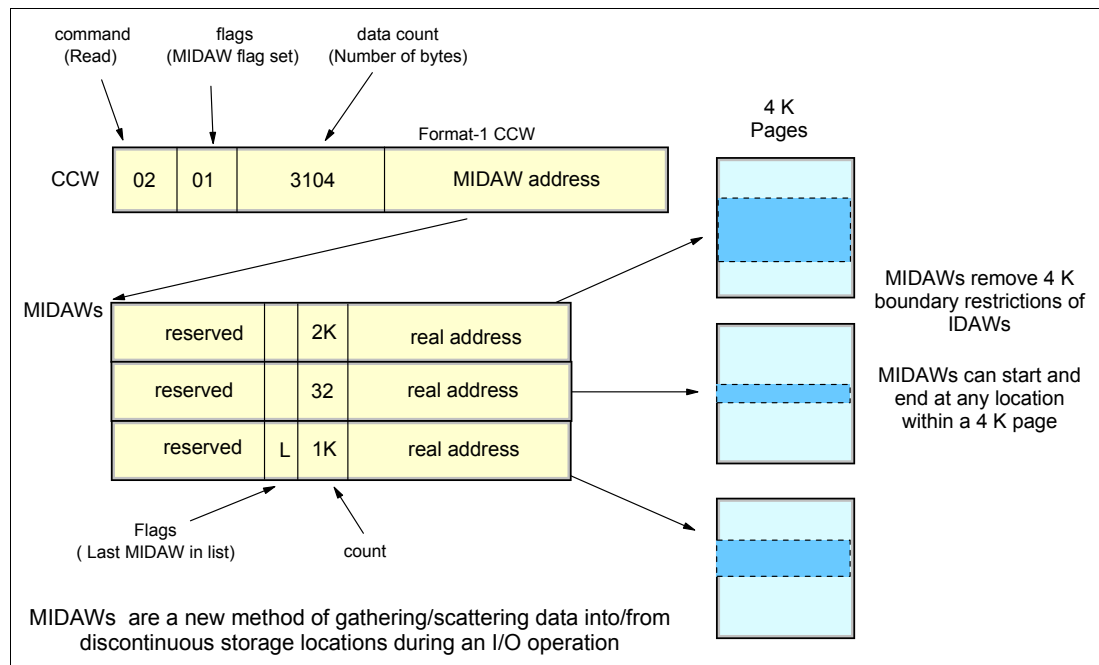
Figure 8-3   MIDAW usage

The use of MIDAWs is indicated by the MIDAW bit in the CCW. If this bit is set, then the *skip flag* cannot be set in the CCW. The skip flag in the MIDAW may be used instead. The data count in the CCW should equal the sum of the data counts in the MIDAWs. The CCW operation ends when the CCW count goes to zero or the last MIDAW (with the *last* flag) ends. The combination of the address and count in a MIDAW cannot cross a page boundary. This means that the largest possible count is 4 K. The maximum data count of all the MIDAWs in a list cannot exceed 64 K, which is the maximum count of the associated CCW.

The scatter-read or scatter-write effect of the MIDAWs makes it possible to efficiently send small control blocks embedded in a disk record to separate buffers from those used for larger data areas within the record. MIDAW operations are on a single I/O block, in the manner of data chaining. Do not confuse this operation with CCW *command* chaining.

8.7.2 Extended format data sets

z/OS extended format data sets use internal structures (usually not visible to the application program) that require scatter-read (or scatter-write) operation. This means that CCW data chaining is required and this produces less than optimal I/O performance. Because the most significant performance benefit of MIDAWs is achieved with extended format (EF) data sets, a brief review of the EF data sets is included here.

Both Virtual Storage Access Method (VSAM) and non-VSAM (DSORG=PS) can be defined as extended format data sets. In the case of non-VSAM data sets, a 32-byte suffix is appended to the end of every physical record (that is, block) on disk. VSAM appends the suffix to the end of every control interval (CI), which normally corresponds to a physical record (a 32 K CI is split into two records to be able to span tracks.) This suffix is used to improve data reliability and facilitates other functions described in the following paragraphs. Thus, for example, if the DCB BLKSIZE or VSAM CI size is equal to 8192, the actual block on DASD consists of 8224 bytes. The control unit itself does not distinguish between suffixes and user data. The suffix is transparent to the access method or database.

In addition to reliability, EF data sets enable three other functions:

► DFSMS striping
► Access method compression
► Extended addressability (EA)

EA is especially useful for creating large DB2 partitions (larger than 4 GB). Striping can be used to increase sequential throughput, or to spread random I/Os across multiple logical volumes. DFSMS striping is especially useful for utilizing multiple channels in parallel for one data set. The DB2 logs are often striped to optimize the performance of DB2 sequential inserts.

To process an I/O operation to an EF data set would normally require at least two CCWs with data chaining. One CCW would be used for the 32-byte suffix of the EF data set. With MIDAW, the additional CCW for the EF data set suffix can be eliminated.

MIDAWs benefit both EF and non-EF data sets. For example, to read twelve 4 K records from a non-EF data set on a 3390 track, Media Manager would chain 12 CCWs together using data chaining. To read twelve 4 K records from an EF data set, 24 CCWs would be chained (two CCWs per 4 K record). Using Media Manager track-level command operations and MIDAWs, an entire track can be transferred using a single CCW.

8.7.3 Performance benefits

z/OS Media Manager has the I/O channel programs support for implementing Extended Format data sets and it automatically exploits MIDAWs when appropriate. Today, most disk I/Os in the system are generated using media manager.

Users of the Executing Fixed Channel Programs in Real Storage (EXCPVR) instruction *may* construct channel programs containing MIDAWs provided that they construct an IOBE with the IOBEMIDA bit set. Users of EXCP instruction *may not* construct channel programs containing MIDAWs

The MIDAW facility removes the 4 K boundary restrictions of IDAWs and, in the case of EF data sets, reduces the number of CCWs. Decreasing the number of CCWs helps to reduce the FICON channel processor utilization. Media Manager and MIDAWs do not cause the bits to move any faster across the FICON link, but they do reduce the number of frames and sequences flowing across the link, thus using the channel resources more efficiently.

Use of the MIDAW facility with FICON Express4, operating at 4 Gbps, compared to use of IDAWs with FICON Express2, operating at 2 Gbps, showed an improvement in throughput for all reads on DB2 table scan tests with EF data sets.

The performance of a specific workload can vary according to the conditions and hardware configuration of the environment. IBM laboratory tests found that DB2 gains significant performance benefits by using the MIDAW facility in the following areas:

► Table scans
► Logging
► Utilities
► Using DFSMS striping for DB2 data sets

Media Manager with the MIDAW facility can provide significant performance benefits when used in combination applications that use EF data sets (such as DB2) or long chains of small blocks.

For additional information relating to FICON and MIDAW, consult the following resources:

► The I/O Connectivity Web site contains the material about FICON channel performance:

  http://www.ibm.com/systems/z/connectivity/

► The following publication:

  DS8000 Performance Monitoring and Tuning, SG24-7146

8.8 IOCP

The required level of I/O configuration program (IOCP) for z196 is V2R1L0 (IOCP 2.1.0) or later.

8.9 Worldwide portname (WWPN) prediction tool

A part of the installation of your z196 server is the preplanning of the Storage Area Network (SAN) environment. IBM has made available a stand alone tool to assist with this planning prior to the installation.

The capability of worldwide port name (WWPN) prediction tool has been extended to calculate and show WWPNs for both virtual and physical ports ahead of system installation.

The tool assigns WWPNs to each virtual Fibre Channel Protocol (FCP) channel/port using the same WWPN assignment algorithms a system uses when assigning WWPNs for channels utilizing N_Port Identifier Virtualization (NPIV). Thus, the SAN can be set up in advance, allowing operations to proceed much faster once the server is installed. In addition, the SAN configuration can be retained instead of altered by assigning the WWPN to physical FCP ports when a FICON feature is replaced.

The WWPN prediction tool takes a .csv file containing the FCP-specific I/O device definitions and creates the WWPN assignments which are required to set up the SAN. A binary configuration file that can be imported later by the system is also created. The .csv file can either be created manually, or exported from the Hardware Configuration Definition/Hardware Configuration Manager (HCD/HCM).

The WWPN prediction tool on z196 (CHPID type FCP) requires:

► z/OS V1R8, V1R9, V1R10, V1R11, with PTFs, or V1R12
► z/VM V5R4 or V6R1, with PTFs

The WWPN prediction tool is available for download at Resource Link and is applicable to all FICON channels defined as CHPID type FCP (for communication with SCSI devices) on z196.

http://www.ibm.com/servers/resourcelink/

8.10 ICKDSF

Device Support Facilities, ICKDSF, Release 17 is required on all systems that share disk subsystems with a z196 processor.

ICKDSF supports a modified format of the CPU information field, which contains a two-digit logical partition identifier. ICKDSF uses the CPU information field instead of CCW reserve/release for concurrent media maintenance. It prevents multiple systems from running ICKDSF on the same volume, and at the same time allows user applications to run while ICKDSF is processing. To prevent any possible data corruption, ICKDSF must be able to determine all sharing systems that can potentially run ICKDSF. Therefore, this support is required for z196.

> **Important:** The need for ICKDSF Release 17 applies even to systems that are not part of the same sysplex, or that are running an operating system other than z/OS, such as z/VM.

8.11 zEnterprise BladeCenter Extension software support

zBX house two types of blades: general purpose, such as the POWER7 blades, and solution specific, such as the IBM Smart Analytics Optimizer.

IBM Blades

IBM offers a selected subset of IBM POWER7 blades that can be installed and operated on the zBX. These blades have been thoroughly tested to ensure compatibility and manageability in the z196 environment.

The blades are virtualized and their LPARs run either AIX Version 5 Release 3 TL12 (POWER6 mode) or AIX Version 6 Release 1 TL5 (POWER7 mode). Applications supported on AIX can be deployed to blades.

> **Statement of Direction:** In the first half of 2011, IBM intends to introduce IBM x86 blades running Linux in the zBX Model 002. The blades will be virtualized and will host Linux virtual machines.

IBM Smart Analytics Optimizer solution

The IBM Smart Analytics Optimizer solution is a defined set of software and hardware that provides a cost optimized solution for running database queries such as those typically found in Data Warehouse and Business Intelligence workloads.

The queries run against DB2 for z/OS, with fast and predictable response times, while retaining the data integrity, data management, security, availability and other qualities of service of the z/OS environment. No change to the applications is required. DB2 for z/OS transparently exploits the special purpose hardware and software for query execution by sending qualified queries to the Smart Analytics Optimizer code running on zBX.

The offering is comprised of hardware and software. The software, IBM Smart Analytics Optimizer for DB2 for z/OS, Version 1.1 (Program Product 5697-AQT), exploits the zBX to provide a comprehensive Business Intelligence solution on System z.

The IBM Smart Analytics Optimizer software is implemented as a logical extension of DB2 for z/OS, and thus works deeply integrated with the DB2 for z/OS Optimizer. It requires DB2 for z/OS Version 9, plus service, as well as service to z/OS. DB2 must run in *new function* mode.

The bundled software and the DB2 Stored Procedures comprise the software product IBM Smart Analytics Optimizer for DB2 for z/OS. The IBM Smart Analytics Optimizer software is delivered via a DVD. An SMP/E installation package provides integration with the DB2 for z/OS environment.

The IBM Smart Analytics Optimizer software is installed on the zBX blades via the System z196 Service Element, using the product DVD. Once installed, updates to the software are installed as PTFs and updated on the blades by calling a DB2 Stored Procedure.

The IBM Smart Analytics Optimizer Studio software is installed from the DVD on a workstation that is attached to z196 and connected to DB2 for z/OS. The workstation must be running IBM Data Studio, which can be downloaded at no charge from the IBM developerworks Web site at:

http://www.ibm.com/developerworks/spaces/optim?pageid=649&S_TACT=105AGX01&S_CMP

A customer supplied IBM System Storage DS5020 with the appropriate configuration and fiber optic cables is required for the IBM Smart Analytics Optimizer solution.

8.12 Software licensing considerations

The z196 mainframe software portfolio includes operating system software (z/OS, z/VM, z/VSE, and z/TPF) and middleware that runs on these operating systems. It also includes middleware for Linux on System z environments. Two major metrics for software licensing are available from IBM, depending on the software product:

- ▸ Monthly License Charge (MLC)
- ▸ International Program License Agreement (IPLA)

The MLC pricing metrics have a recurring charge that applies each month. In addition to the right to use the product, the charge includes access to IBM product support during the support period. MLC metrics have several offerings that are applicable to the zEnterprise 196:

- ▸ Workload License Charge (WLC)
- ▸ System z New Application License Charge (zNALC)
- ▸ Parallel Sysplex License Charge (PSLC)
- ▸ Midrange Workload License Charge (MWLC)

IPLA metrics have an single, up-front, charge for an entitlement to use the product. Optionally, a separate annual charge called *subscription and support* entitles customers to receive future releases and versions at no additional charge, and also allows access to IBM product support during the support period.

For details, consult the *IBM System z Software Pricing Reference Guide*, G326-0594:

http://www.ibm.com/servers/eserver/zseries/library/refguides/sw_pricing.html

8.12.1 Workload License Charge

Workload License Charge (WLC) requires z/OS or z/TPF operating systems in 64-bit mode. Any mix of z/OS, z/VM, Linux on System z, VM/ESA®, z/VSE, and z/TPF images is allowed.

The two WLC license types are:

- ▸ Flat WLC (FWLC)

  Software products licensed under FWLC are charged at the same flat rate, no matter what capacity (MSUs) the server is.

- ▸ Variable WLC (VWLC)

  Products such as z/OS, DB2, IMS, CICS, MQSeries®, and Lotus® Domino® can be charged in two different ways:

- – Full-capacity is when the server's total number of MSUs is used for charging. Full-capacity is applicable when the server is not eligible for subcapacity.

- – Subcapacity is when software charges are based on the logical partition's utilization where the product is running.

WLC subcapacity allows software charges based on logical partition utilizations instead of the server's total number of MSUs. Subcapacity removes the dependency between software charges and server (hardware) installed capacity.

Subcapacity is based on the logical partition's rolling 4-hour average utilization. It is *not* based on the utilization of each product[7], but on the utilization of the logical partition or partitions where it runs. The VWLC licensed products running on a logical partition are charged by the maximum value of this partition's rolling 4-hour average utilization within a month.

The logical partition's rolling 4-hour average utilization can be limited by a *defined capacity* definition on the partition's image profiles. The defined capacity definition activates the *soft capping* function of PR/SM, avoiding 4-hour average partition utilizations above the defined capacity value. Soft capping controls the maximum rolling 4-hour average utilization (the last 4-hour average value at every five minutes interval), but does *not* control the maximum instantaneous partition utilization.

Even by using the soft-capping option, the partition's utilization can reach its maximum share based on the number of logical processors and weights in the image profile. Only the rolling 4-hour average utilization is tracked, allowing utilization peaks above the defined capacity value.

As with the Parallel Sysplex License Charge (PSLC) software license charge type, the aggregation of servers' capacities within the same Parallel Sysplex is also possible in WLC, following the same prerequisites.

Entry Workload License Charge (EWLC) is not offered for zEnterprise 196.

For further information about WLC and details about how to combine logical partitions utilization, see *z/OS Planning for Workload License Charges*, SA22-7506.

8.12.2 System z New Application License Charge

System z New Application License Charge (zNALC) offers a reduced price for the z/OS operating system on logical partitions running a qualified new workload application such as Java language business applications running under WebSphere Application Server for z/OS, Domino, SAP, PeopleSoft, and Siebel.

z/OS with zNALC provides a strategic pricing model available on the full range of System z servers for simplified application planning and deployment. zNALC allows for aggregation across a qualified Parallel Sysplex, which can provide a lower cost for incremental growth across new workloads that span a Parallel Sysplex.

For additional information see the zNALC Web site:

http://www.ibm.com/servers/eserver/zseries/swprice/znalc.html

---

[7] With the exception of products licensed using the Select Application License Charge (SALC) pricing metric.

8.12.3 Select Application License Charge

Select Application License Charge (SALC) applies only to WebSphere MQ for System z. It allows a WLC customer to license MQ under product utilization rather than the subcapacity pricing provided under WLC.

WebSphere MQ is typically a low-usage product that runs pervasively throughout the customer environment. Customers who run WebSphere MQ at a very low usage can benefit from SALC. Alternatively, one can still choose to license WebSphere MQ under WLC.

A reporting function, which IBM provides in the operating system IBM Software Usage Report Program, is used to calculate the daily MSU number. The rules to determine the billable SALC MSUs for WebSphere MQ use the following algorithm:

1. Determine the highest daily usage of a program[8] family, which is the highest of 24 hourly measurements recorded each day.

2. Determine the monthly usage of a program family, which is the fourth highest daily measurement recorded for a month.

3. Use the highest monthly usage determined for the next billing period.

For additional information about SALC, see the MWLC Web site:

http://www.ibm.com/servers/eserver/zseries/swprice/other.html

8.12.4 Midrange Workload License Charge

Midrange Workload License Charge (MWLC) applies to z/VSE V4 when it is running on z196, IBM System z10, and IBM System z9 servers. The exceptions are the z10 BC and z9 BC servers at capacity setting A01 to which zSeries Entry License Charge (zELC) applies. Similar to Workload License Charge, MWLC can be implemented in full-capacity or subcapacity mode. MWLC applies to z/VSE V4 and several IBM middleware products for z/VSE. All other z/VSE programs continue to be priced as before.

The z/VSE pricing metric is independent of the pricing metric for other systems, for instance, z/OS, that might be running on the same server. When z/VSE is running as a guest of z/VM, z/VM V5R4 or later is required.

The Subcapacity Report Tool (SCRT) is used to report utilization. One SCRT report is required for each server.

For additional information see the MWLC Web site:

http://www.ibm.com/servers/eserver/zseries/swprice/mwlc.html

8.12.5 System z International Licensing Agreement

On the mainframe, the following types of products are generally in the IPLA category:

► Data Management Tools
► CICS Tools
► Application Development Tools
► Certain WebSphere for System z products
► System z Linux middleware products
► z/VM Versions 5 and 6

---

[8] The term *program* refers to all active versions of MQ.

For additional information, see the System z IPLA Web site:

http://www.ibm.com/servers/eserver/zseries/swprice/zipla/

8.13 References

For the most current planning information, see the support Web site for each of the following operating systems:

► z/OS

  http://www.ibm.com/systems/support/z/zos/

► z/VM

  http://www.ibm.com/systems/support/z/zvm/

► z/TPF

  http://www.ibm.com/software/htp/tpf/pages/maint.htm

► z/VSE

  http://www.ibm.com/servers/eserver/zseries/zvse/support/preventive.html

► Linux on System z

  http://www.ibm.com/systems/z/os/linux/

<div align="right">

9

</div>

System upgrades

This chapter provides an overview of zEnterprise 196 upgrade capabilities and procedures, with an emphasis on Capacity on Demand offerings.

The upgrade offerings to the z196 servers have been developed from previous IBM System z servers. In response to customer demands and changes in market requirements, a number of features have been added. The changes and additions are designed to provide increased customer control over the capacity upgrade offerings with decreased administrative work and with enhanced flexibility. The provisioning environment gives the customer an unprecedented flexibility and a finer control over cost and value.

Given today's business environment, the benefits of the growth capabilities provided by the z196 are plentiful, and include, but are not limited to:

► Enabling exploitation of new business opportunities
► Supporting the growth of dynamic, smart environments
► Managing the risk of volatile, high-growth, and high-volume applications
► Supporting 24x365 application availability
► Enabling capacity growth during lock down periods
► Enabling planned-downtime changes without availability impacts

This chapter discusses the following topics:

For more information, see the following publications:

► *IBM System z10 Enterprise Class Capacity On Demand,* SG24-7504
► *IBM zEnterprise 196 Capacity on Demand User's Guide,* SC28-2605

　　　　　263

9.1 Upgrade types

Types of upgrades for a z196 are summarized in this section.

Permanent and temporary upgrades

In different situations, different types of upgrades are needed. After some time, depending on your growing workload, you might require more memory, additional I/O cards, or more processor capacity. However, in certain situations, only a short-term upgrade is necessary to handle a peak workload, or to temporarily replace a server that is down during a disaster or data center maintenance. The z196 offers the following solutions for such situations:

► Permanent

   – Miscellaneous equipment specification (MES)

     The MES upgrade order is always performed by IBM personnel. The result can be either real hardware added to the server or installation of LIC configuration control (LICCC) to the server. In both cases, installation is performed by IBM personnel.

   – Customer Initiated Upgrade (CIU)

     Using the CIU facility for a given server requires that the online CoD buying feature (FC 9900) is installed on the server. The CIU facility supports LICCC upgrades only.

► Temporary

   All temporary upgrades are LICCC-based. The one billable capacity offering is On/Off Capacity on Demand (On/Off CoD). The two replacement capacity offerings available are Capacity Backup (CBU) and Capacity for Planned Event (CPE).

For descriptions see 9.1.1, "Terminology related to CoD for System z196 servers" on page 265.

> **Note:** The MES provides system upgrade that can result in more enabled processors, different CP capacity level, but also in additional books, memory, I/O drawers, and I/O cards (physical upgrade). An MES can also upgrade the zEnterprise BladeCenter Extension. Additional planning tasks are required for nondisruptive logical upgrades. MES is ordered through your IBM representative and delivered by IBM service personnel.

Concurrent and nondisruptive upgrades

Depending on the impact on system and application availability, upgrades can be classified as:

► Concurrent

   In general, concurrency addresses the continuity of operations of the hardware part of an upgrade, for instance, whether a server (as a box) is required to be switched off during the upgrade. For details see 9.2, "Concurrent upgrades" on page 269.

► Non-concurrent

   This type of upgrade requires the stopping the system (HW). Examples of such upgrades include model upgrades from any M15, M32, M49, M66 models to the M80 model, certain physical memory capacity upgrades and adding I/O cages

Disruptive   An upgrade is disruptive when resources added to an operating system image require that the operating system be recycled to configure the newly added resources.

Nondisruptive   Nondisruptive upgrades do not require the running software or operating system to be restarted for the upgrade to take an effect.

Thus, even concurrent upgrades can be disruptive to those operating systems or programs that do not support the upgrades while at the same time being nondisruptive to others. For details see 9.8, "Nondisruptive upgrades" on page 303.

9.1.1 Terminology related to CoD for System z196 servers

Table 9-1 briefly describes the most frequently used terms related to Capacity on Demand for z196 servers.

Table 9-1   CoD terminology

Term	Description
Activated capacity	Capacity that is purchased and activated. Purchased capacity can be greater than activated capacity.
Billable capacity	Capacity that helps handle workload peaks, either expected or unexpected. The one billable offering available is On/Off Capacity on Demand.
Book	A physical package that contains memory, a Multi-Chip Module (MCM), and host channel adapters (HCA2s). A book plugs into one of four slots in the central processor complex (CPC) cage of the z196.
Capacity	Hardware resources (processor and memory) able to process workload can be added to the system through various capacity offerings.
Capacity Backup (CBU)	A function that allows the use of spare capacity in a CPC to replace capacity from another CPC within an enterprise, for a limited time. Typically, CBU is used when another CPC of the enterprise has failed or is unavailable because of a disaster event. The CPC using CBU replaces the missing CPC's capacity.
Capacity for planned event (CPE)	Used when temporary replacement capacity is needed for a short term event. CPE activate processor capacity temporarily to facilitate moving machines between data centers, upgrades, and other routine management tasks. CPE is an offering of Capacity on Demand.
Capacity levels	Can be full capacity or subcapacity. For the z196 server, capacity levels for the CP engine are 7, 6, 5, and 4: ▶ Full capacity CP engine is indicated by 7. ▶ Subcapacity CP engines are indicated by 6, 5, and 4.
Capacity setting	Derived from the capacity level and the number of processors. For the z196 server, the capacity levels are 7nn, 6xx, 5xx, 4xx, where xx or nn indicates the number of active CPs. The number of processors can have a range of: ▶ 0–80 for capacity level 7nn ▶ 1–15 for capacity levels 6xx, 5xx, 4xx
Concurrent book add (CBA)	Concurrently adds book hardware, including processors, physical memory, and I/O connectivity
Capacity Backup (CBU)	Provides reserved emergency backup processor capacity for unplanned situations when a loss of capacity occurs in another part of the enterprise
Central processor complex (CPC)	A physical collection of hardware that consists of main storage, one or more central processors, timers, and channels
Customer Initiated Upgrade (CIU)	A Web-based facility where you may request processor and memory upgrades by using the IBM Resource Link and the system's remote support facility (RSF) connection

Term	Description
Capacity on Demand (CoD)	The ability of a computing system to increase or decrease its performance capacity as needed to meet fluctuations in demand
Capacity Provisioning Manager (CPM)	As a component of z/OS Capacity Provisioning, CPM monitors business-critical workloads that are running on z/OS systems on z196 servers.
Customer profile	This information resides on Resource Link and contains customer and machine information. A customer profile may contain information about more than one machine.
Enhanced book availability	In a multibook configuration, the ability to have a book concurrently removed from the server and reinstalled during an upgrade or repair action
Full capacity CP feature	For z196 feature (CP7), provides full capacity. Capacity settings 7xx are full capacity settings.
High water mark	Capacity purchased and owned by the customer
Installed record	The LICCC record has been downloaded, staged to the SE, and is now installed on the CPC. A maximum of eight different records can be concurrently installed and active.
Licensed Internal Code (LIC)	LIC is microcode, basic I/O system code, utility programs, device drivers, diagnostics, and any other code delivered with an IBM machine for the purpose of enabling the machine's specified functions.
LIC Configuration Control (LICCC)	Configuration control by the LIC to provides for server upgrade without hardware changes by enabling the activation of additional previously installed capacity
Multi-Chip Module (MCM)	An electronic package where multiple integrated circuits (semiconductor dies) and other modules are packaged on a common substrate to be mounted on a PCB (printed circuit board) as a single unit.
Model capacity identifier (MCI)	Shows the current active capacity on the server, including all replacement and billable capacity. For the z196, the model capacity identifier is in the form of 7nn, 6xx, 5xx, or 4xx, where xx or nn indicates the number of active CPs. ► nn can have a range of 00 - 80. ► xx can have a range of 01-15.
Model Permanent Capacity Identifier (MPCI)	Keeps information about capacity settings active before any temporary capacity was activated
Model *Temporary Capacity* Identifier (MTCI)	Reflects the permanent capacity with billable capacity only, without replacement capacity. If no billable temporary capacity is active, Model Temporary Capacity Identifier equals Model Permanent Capacity Identifier.
On/Off Capacity on Demand (CoD)	Represents a function that allows a spare capacity in a CPC to be made available to increase the total capacity of a CPC. For example, On/Off CoD may be used to acquire additional capacity for the purpose of handling a workload peak.
Permanent capacity	The capacity that a customer purchases and activates. This amount might be less capacity than the total capacity purchased.
Permanent upgrade	LIC licensed by IBM to enable the activation of applicable computing resources, such as processors or memory, for a specific CIU-eligible machine on a permanent basis
Purchased capacity	Capacity delivered to and owned by the customer. It can be higher than permanent capacity.

Term	Description
Permanent/Temporary entitlement record	The internal representation of a temporary (TER) or permanent (PER) capacity upgrade processed by the CIU facility. An entitlement record contains the encrypted representation of the upgrade configuration with the associated time limit conditions.
Replacement capacity	A temporary capacity used for situations in which processing capacity in other parts of the enterprise is lost during either a planned event or an unexpected disaster. The two replacement offerings available are, Capacity for Planned Events and Capacity Backup.
Resource Link	IBM Resource Link is a technical support Web site included in the comprehensive set of tools and resources available from the IBM Systems technical support site: http://www.ibm.com/servers/resourcelink/
Secondary approval	An option, selected by the customer, that a second approver control each Capacity on Demand order. When a secondary approval is required, the request is sent for approval or cancellation to the Resource Link secondary user ID.
Staged record	The point when a record representing a capacity upgrade, either temporary or permanent, has been retrieved and loaded on the Support Element (SE) disk.
Subcapacity	For the z196, CP features (CP4, CP5, and CP6) provide reduced capacity relative to the full capacity CP feature (CP7).
Temporary capacity	An optional capacity that is added to the current server capacity for a limited amount of time. It can be capacity that is owned or not owned by the customer.
Vital product data (VPD)	Information that uniquely defines system, hardware, software, and microcode elements of a processing system
Miscellaneous equipment specification (MES)	An upgrade process initiated through IBM representative and installed by IBM personnel

9.1.2 Permanent upgrades

Permanent upgrades can be:

► Ordered through an IBM sales representative
► Initiated by the customer with the Customer Initiated Upgrade (CIU) on IBM Resource Link

Note: The use of the CIU facility for a given server requires that the online CoD buying feature (FC 9900) is installed on the server. The CIU facility itself is enabled through the permanent upgrade authorization Feature Code (FC 9898).

Permanent upgrades ordered through an IBM representative

Through a permanent upgrade you can:

► Add processor books.
► Add I/O cages and features.
► Add model capacity.
► Add specialty engines.
► Add memory.
► Activate unassigned model capacity or IFLs.
► Deactivate activated model capacity or IFLs.
► Activate channels.
► Activate cryptographic engines.
► Change specialty engine (re-characterization).
► Add zBX and zBX features:

- Chassis
- Racks
- Blades
- Entitlements

> **Attention:** Most of the MES can be concurrently applied, without disrupting the existing workload (see 9.2, "Concurrent upgrades" on page 269 for details). However, certain MES changes are disruptive (for example, upgrade of models M15, M32, M49, and M66 to M80, or adding I/O cages).
>
> Memory upgrades that require DIMM changes can be made nondisruptive if the flexible memory option is ordered.

Permanent upgrades initiated through CIU on IBM Resource Link

Ordering a permanent upgrade by using the CIU application through Resource Link allows you to add capacity to fit within your existing hardware, as follows:

- ▶ Add model capacity
- ▶ Add specialty engines
- ▶ Add memory
- ▶ Activate unassigned model capacity or IFLs
- ▶ Deactivate activated model capacity or IFLs

Flexible Upgrade Option

This option allows you to test and modify capacity additions to your permanent configuration using On/Off CoD until you determine exactly how much your business requires.

9.1.3 Temporary upgrades

System z196 offers three types of temporary upgrades:

- ▶ On/Off Capacity on Demand (On/Off CoD)

   This offering allows you to temporarily add additional capacity or specialty engines due to seasonal activities, period-end requirements, peaks in workload, or application testing. This temporary upgrade can only be ordered using the CIU application through Resource Link.

- ▶ Capacity Backup (CBU)

   This offering allows you to replace model capacity or specialty engines to a backup server in the event of an unforeseen loss of server capacity because of an emergency.

- ▶ Capacity for Planned Event (CPE)

   This offering allows you to replace model capacity or specialty engines due to a relocation of workload during system migrations or a data center move.

CBU or CPE temporary upgrades can be ordered by using the CIU application through Resource Link or by calling your IBM sales representative.

Temporary upgrades capacity changes can be billable or replacement.

Billable capacity

To handle a peak workload, processors can be rented temporarily on a daily basis. You may activate up to double the purchased capacity of any PU type.

The one billable capacity offering is On/Off Capacity on Demand (On/Off CoD).

Replacement capacity

When a processing capacity is lost in another part of an enterprise, replacement capacity can be activated. It allows you to activate any PU type up to authorized limit.

The two replacement capacity offerings are:

► Capacity Backup
► Capacity for Planned Event

9.2 Concurrent upgrades

Concurrent upgrades on the z196 can provide additional capacity with no server outage. In most cases, with prior planning and operating system support, a concurrent upgrade can also be nondisruptive to the operating system.

Given today's business environment, the benefits of the concurrent capacity growth capabilities provided by the z196 are plentiful, and include, but are not limited to:

► Enabling exploitation of new business opportunities
► Supporting the growth of smart environments
► Managing the risk of volatile, high-growth, and high-volume applications
► Supporting 24x365 application availability
► Enabling capacity growth during *lock down* periods
► Enabling planned-downtime changes without affecting availability

This capability is based on the flexibility of the design and structure, which allows concurrent hardware installation and Licensed Internal Code (LIC) control over the configuration.

The subcapacity models allow additional configuration granularity within the family. The added granularity is available for models configured with up to 15 CPs and provides 45 additional capacity settings. Subcapacity models provide for CP capacity increase in two dimensions that can be used together to deliver configuration granularity. The first dimension is by adding CPs to the configuration, the second is by changing the capacity setting of the CPs currently installed to a higher model capacity identifier.

The z196 allows the concurrent addition of processors to a running logical partition. As a result, you can have a flexible infrastructure, in which you may add capacity without pre-planning. This function is supported by z/VM. Planning ahead is required for z/OS logical partitions. To be able to add processors to a running z/OS, reserved processors must be specified in the logical partition's profile.

Another function concerns the system assist processor (SAP). When additional SAPs are concurrently added to the configuration, the SAP-to-channel affinity is dynamically re-mapped on all SAPs on the server to rebalance the I/O configuration.

All of the zBX and its features can be installed concurrently. For the IBM Smart Analytics Optimizer solution, the applications using the solution will continue to execute during the upgrade. However they will use the z196 resources to satisfy the application execution instead of using the zBX infrastructure.

9.2.1 Model upgrades

The z196 has a machine type and model, and model capacity identifiers:

► Machine type and model is 2817-Mvv.

The *vv* can be 15, 32, 49, 66, or 80. The model number indicates how many PUs (vv) are available for customer characterization. Model M15 has one book installed, model M32 contains two books, model M49 contains three books, and models M66 and M80 contain four books.

► Model capacity identifiers are 4xx, 5xx, 6xx, or 7yy.

The xx is a range of 01 - 15 and *yy* is a range of 00 - 80. The model capacity identifier describes how many CPs are characterized (xx or yy) and the capacity setting (4, 5, 6, or 7) of the CPs.

A hardware configuration upgrade always requires additional physical hardware (books, cages, drawers or all of them). A server upgrade can change either, or both, the server model and the model capacity identifier (MCI).

Note the following model upgrade information:

► LICCC upgrade

– Does not change the server model 2817-Mvv, because additional books are not added
– Can change the model capacity identifier, the capacity setting, or both

► Hardware installation upgrade

– Can change the server model 2817-Mvv, if additional books are included
– Can change the model capacity identifier, the capacity setting, or both

The server model and the model capacity identifier can be concurrently changed. Concurrent upgrades can be accomplished for both *permanent* and *temporary* upgrades.

> **Note:** A model upgrade can be concurrent by using concurrent book add (CBA), except for upgrades to Model M80.

Licensed Internal Code upgrades (MES ordered)

The LIC Configuration Control (LICCC) provides for server upgrade without hardware changes by activation of additional (previously installed) unused capacity. Concurrent upgrades through LICCC can be done for:

► Processors (CPs, ICFs, zAAPs, **zIIPs**, IFLs, and SAPs) if unused PUs are available on the installed books or if the model capacity identifier for the CPs can be increased.

► Memory, when unused capacity is available on the installed memory cards. Plan-ahead memory and the flexible memory option are available for customers to gain better control over future memory upgrades. See 2.5.6, "Flexible memory option" on page 46, and 2.5.7, "Plan-ahead memory" on page 47 for more details.

► I/O card ports (ESCON channels and ISC-3 links), when there are available ports on the installed I/O cards.

Concurrent hardware installation upgrades (MES ordered)

Configuration upgrades can be concurrent when installing additional:

► Books (which contain processors, memory, and HCA2s), when book slots are available in the CPC cage

► HCA2 fanouts

► InfiniBand-Multiplexer (IFB-MP) cards

► I/O cards, when slots are still available on the installed I/O cages or I/O drawers. I/O cages *cannot* be installed concurrently.

- ▶ I/O drawers and I/O cards to go into those drawers.
- ▶ All of zBX and zBX features

The concurrent I/O upgrade capability can be better exploited if a future target configuration is considered during the initial configuration.

It is highly recommended to use I/O drawers to satisfy the required I/O cards as these drawers can be installed concurrently. Should the demand for I/O cards be bigger than what can be installed in a drawer-only configuration, I/O cages (up to two) can be used. If the customers need a very high number of I/O cards, an RPQ is available to install three I/O cages.

Using the plan-ahead concept, the required number of I/O cages for concurrent upgrades, up to the target configuration, can be included in the initial configuration.

Concurrent PU conversions (MES-ordered)

The z196 supports concurrent conversion between all PU types, any-to-any PUs including SAPs, to provide flexibility to meet changing business requirements.

> **Note:** The LICCC-based PU conversions require that at least one PU, either CP, ICF, or IFL, remains unchanged. Otherwise, the conversion is disruptive. The PU conversion generates a new LICCC that can be installed concurrently in two steps:
>
> 1. The assigned PU is removed from the configuration.
> 2. The newly available PU is activated as the new PU type.

Logical partitions might also have to free the PUs to be converted, and the operating systems must have support to configure processors offline or online so that the PU conversion can be done non-disruptively.

> **Note:** Customer planning and operator action are required to exploit concurrent PU conversion. Consider the following information about PU conversion:
>
> - ▶ It is disruptive if *all* current PUs are converted to different types.
> - ▶ It might require individual logical partition outage if dedicated PUs are converted.
>
> Unassigned CP capacity is recorded by a model capacity identifier. CP feature conversions change (increase or decrease) the model capacity identifier.

9.2.2 Customer Initiated Upgrade facility

The Customer Initiated Upgrade (CIU) facility is an IBM online system through which a customer may order, download, and install permanent and temporary upgrades for System z servers. Access to and use of the CIU facility requires a contract between the customer and IBM, through which the terms and conditions for use of the CIU facility are accepted. The use of the CIU facility for a given server requires that the online CoD buying feature code (FC 9900) is installed on the server. The CIU facility itself is controlled through the permanent upgrade authorisation feature code, FC 9898.

After a customer has placed an order through the CIU facility, the customer will receive a notice that the order is ready for download. The customer may then download and apply the upgrade by using functions available through the HMC, along with the remote support facility. After all the prerequisites are met, the entire process, from ordering to activation of the upgrade, is performed by the customer.

After the download, the actual upgrade process is fully automated and does not require any on-site presence of IBM service personnel.

CIU prerequisites

The CIU facility supports LICCC upgrades only. It does not support I/O upgrades. All additional capacity required for an upgrade must be previously installed. Additional books or I/O cards cannot be installed as part of an order placed through the CIU facility. The sum of CPs, unassigned CPs, ICFs, zAAPs, zIIPs, IFLs, and unassigned IFLs cannot exceed the PU count of the installed books. The total number of zAAPs or zIIPs cannot each exceed the number of purchased CPs.

CIU registration and agreed contract for CIU

To use the CIU facility, a customer must be registered and the system must be set up. After completing the CIU registration, access the CIU application through the IBM Resource Link Web site:

> http://www.ibm.com/servers/resourcelink/

As part of the setup, the customer provides one resource link ID for configuring and placing CIU orders and, if required, a second ID as an approver. The IDs are then set up for access to the CIU support. The CIU facility is beneficial by allowing upgrades to be ordered and delivered much faster than through the regular MES process.

To order and activate the upgrade, log on to the IBM Resource Link Web site and invoke the CIU application to upgrade a server for processors, or memory. Requesting a customer order approval to conform to customer operation policies is possible. As previously mentioned, customers may allow the definition of additional IDs to be authorized to access the CIU. Additional IDs can be authorized to enter or approve CIU orders, or only view existing orders.

Permanent upgrades

Permanent upgrades can be ordered by using the CIU facility.

Through the CIU facility, you may generate online permanent upgrade orders to concurrently add processors (CPs, ICFs, zAAPs, zIIPs, IFLs, and SAPs) and memory, or change the model capacity identifier, up to the limits of the installed books on an existing server.

Temporary upgrades

The base model z196 describes permanent and dormant capacity (Figure 9-1) using the capacity marker and the number of PU features installed on the server. Up to eight temporary offerings can be present. Each offering has its own policies and controls and each can be activated or deactivated independently in any sequence and combination. Although multiple offerings can be active at any time, if enough resources are available to fulfill the offering specifications, only one On/Off CoD offering can be active at any time.
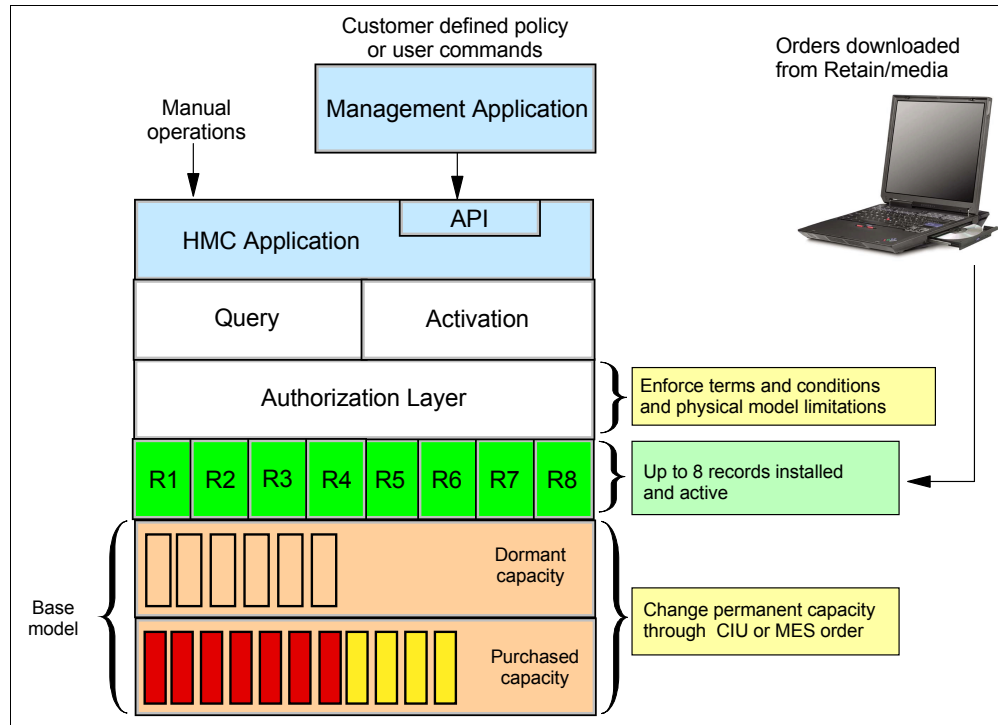
Figure 9-1   The provisioning architecture

Temporary upgrades are represented in the server by a *record*. All temporary upgrade records, downloaded from the remote support facility (RSF) or installed from portable media, are resident on the Service Element (SE) hard drive. At the time of activation, the customer can control everything locally. Figure 9-1 shows a representation of the provisioning architecture.

The authorization layer enables administrative control over the temporary offerings.

The activation and deactivation can be driven either manually or under control of an application through a documented application program interface (API).

By using the API approach, you may customize, at activation time, the resources necessary to respond to the current situation, up to the maximum specified in the order record. If the situation changes, you can add more or remove resources without having to go back to the base configuration. This eliminates the need for temporary upgrade specification for all possible scenarios. However, for CPE the ordered configuration is the only possible activation.

In addition, this approach enables you to update and replenish temporary upgrades, even in situations where the upgrades are already active. Likewise, depending on the configuration, permanent upgrades can be performed while temporary upgrades are active. Figure 9-2 shows examples of activation sequences of multiple temporary upgrades.
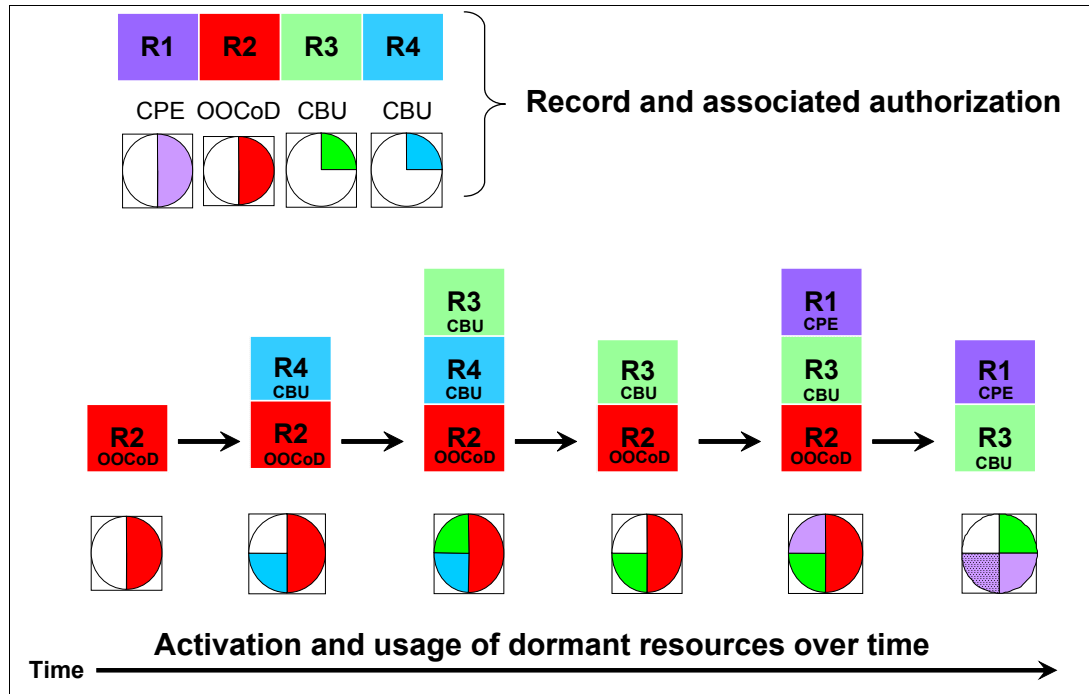
Figure 9-2   Example of temporary upgrade activation sequence

In the case of the R2, R3, and R1 being active at the same time, only parts of R1 can be activated, because not enough resources are available to fulfill all of R1. When R2 is then deactivated, the remaining parts of R1 may be activated as shown.

Temporary capacity can be billable as On/Off Capacity on Demand (On/Off CoD), or replacement as Capacity Backup (CBU) or CPE:

► On/Off CoD is a function that enables *concurrent* and *temporary* capacity growth of the server.

  On/Off CoD *can* be used for customer peak workload requirements, for any length of time, and has a daily hardware and maintenance charge. The software charges can vary according to the license agreement for the individual products. See your IBM Software Group representative for exact details.

  On/Off CoD can concurrently add processors (CPs, ICFs, zAAPs. zIIPs, IFLs, and SAPs), increase the model capacity identifier, or both, up to the limit of the installed books of an existing server, and is restricted to twice the currently installed capacity. On/Off CoD requires a contract agreement between the customer and IBM.

  You decide whether to pre-pay or post-pay On/Off CoD. Capacity tokens inside the records are used to control activation time and resources.

► CBU is a concurrent and temporary activation of additional CPs, ICFs, zAAPs, zIIPs, IFLs, and SAPs, an increase of the model capacity identifier, or both.

  CBU cannot be used for peak load management of customer workload or for CPE. A CBU activation can last up to 90 days when a disaster or recovery situation occurs.

CBU features are optional and require unused capacity to be available on installed books of the backup server, either as unused PUs or as a possibility to increase the model capacity identifier, or both. A CBU contract must be in place before the special code that enables this capability can be loaded on the server. The standard CBU contract provides for five 10-day tests and one 90-day disaster activation over a five-year period. Contact your IBM Representative for details.

► CPE is a concurrent and temporary activation of additional CPs, ICFs, zAAPs, zIIPs, IFLs, and SAPs or an increase of the model capacity identifier, or both.

The CPE offering is used to replace temporary lost capacity within a customer's enterprise for planned downtime events, for example, with data center changes. CPE cannot be used for peak load management of customer workload or for a disaster situation.

The CPE feature requires unused capacity to be available on installed books of the backup server, either as unused PUs or as a possibility to increase the model capacity identifier on a subcapacity server, or both. A CPE contract must be in place before the special code that enables this capability can be loaded on the server. The standard CPE contract provides for one three-day planned activation at a specific date. Contact your IBM representative for details.

9.2.3 Summary of concurrent upgrade functions

Table 9-2 summarizes the possible concurrent upgrades combinations.

Table 9-2   Concurrent upgrade summary

Type	Name	Upgrade	Process
Permanent	MES	CPs, ICFs, zAAPs, zIIPs, IFLs, SAPs, book, memory, and I/Os	Installed by IBM service personnel
	Online permanent upgrade	CPs, ICFs, zAAPs, zIIPs, IFLs, SAPs, and memory	Performed through the CIU facility
Temporary	On/Off CoD	CPs, ICFs, zAAPs, zIIPs, IFLs, and SAPs	Performed through the OOCoD facility
	CBU	CPs, ICFs, zAAPs, zIIPs, IFLs, and SAPs	Performed through the CBU facility
	CPE	CPs, ICFs, zAAPs, zIIPs, IFLs, and SAPs	Performed through the CPE facility

9.3 MES upgrades

Miscellaneous equipment specification (MES) upgrades enable concurrent and permanent capacity growth. MES upgrades allow the concurrent adding of processors (CPs, ICFs, zAAPs, zIIPs, IFLs, and SAPs), memory capacity, and I/O ports as well as hardware and entitlements to the zEnterprise BladeCenter Extension. Regarding subcapacity models, MES upgrades allow the concurrent adjustment of both the number of processors and the capacity level. The MES upgrade can be done using Licensed Internal Code Configuration Control (LICCC) only, by installing additional books, adding I/O cards, or a combination:

► MES upgrades for processors are done by any of the following methods:

– LICCC assigning and activating unassigned PUs up to the limit of the installed books

– LICCC to adjust the number and types of PUs or to change the capacity setting, or both

– Installing additional books and LICCC assigning and activating unassigned PUs on installed books

► MES upgrades for memory are done by either of the following methods:

– Using LICCC to activate additional memory capacity up to the limit of the memory cards on the currently installed books. Plan-ahead and flexible memory features enable you to have better control over future memory upgrades. For details about the memory features, see:

  • 2.5.7, "Plan-ahead memory" on page 47
  • 2.5.6, "Flexible memory option" on page 46

– Installing additional books and using LICCC to activate additional memory capacity on installed books

– Using the enhanced book availability (EBA), where possible, on multibook systems to add or change the memory cards

► MES upgrades for I/O are done by either of the following methods:

– Using LICCC to activate additional ports on already installed ESCON and ISC-3 cards

– Installing additional I/O cards and supporting infrastructure if required on I/O cages or I/O drawers that are already installed, or installing additional I/O drawers to hold the new cards.

► MES upgrades for the zEnterprise BladeCenter Extension can only be performed through your IBM customer representative.

An MES upgrade requires IBM service personnel for the installation. In most cases, the time required for installing the LICCC and completing the upgrade is short.

To better exploit the MES upgrade function, it is strongly recommended to carefully plan the initial configuration to allow a concurrent upgrade to a target configuration.

By planning ahead, it is possible to enable nondisruptive capacity and I/O growth with no system power down and no associated PORs or IPLs. The availability of I/O drawers has improved the flexibility to do un-planned I/O configuration changes concurrently.

The store system information (STSI) instruction gives more useful and detailed information about the base configuration and about temporary upgrades. This enables you to more easily resolve billing situations where Independent Software Vendor (ISV) products are in use.

The model and model capacity identifier returned by the STSI instruction are updated to coincide with the upgrade. See "Store system information (STSI) instruction" on page 305 for more details.

> **Note:** The MES provides the physical upgrade, resulting in more enabled processors, different capacity settings for the CPs, additional memory, and I/O ports. Additional planning tasks are required for non-disruptive logical upgrades (see "Recommendations to avoid disruptive upgrades" on page 308).

9.3.1 MES upgrade for processors

An MES upgrade for processors can concurrently add CPs, ICFs, zAAPs, zIIPs, IFLs, and SAPs to a z196 by assigning available PUs that reside on the books, through LICCC. Depending on the quantity of the additional processors in the upgrade, additional books might be required and can be concurrently installed before the LICCC is enabled. With the subcapacity models, additional capacity can be provided by adding CPs, by changing the capacity identifier on the current CPs, or by doing both.

> **Note:** The sum of CPs, inactive CPs, ICFs, zAAPs, zIIPs, IFLs, unassigned IFLs, and SAPs cannot exceed the maximum limit of PUs available for customer use. The number of zAAPs or zIIPs cannot exceed the number of purchased CPs.

Example of MES upgrade

Figure 9-3 is an example of an MES upgrade for processors, showing two upgrade steps.
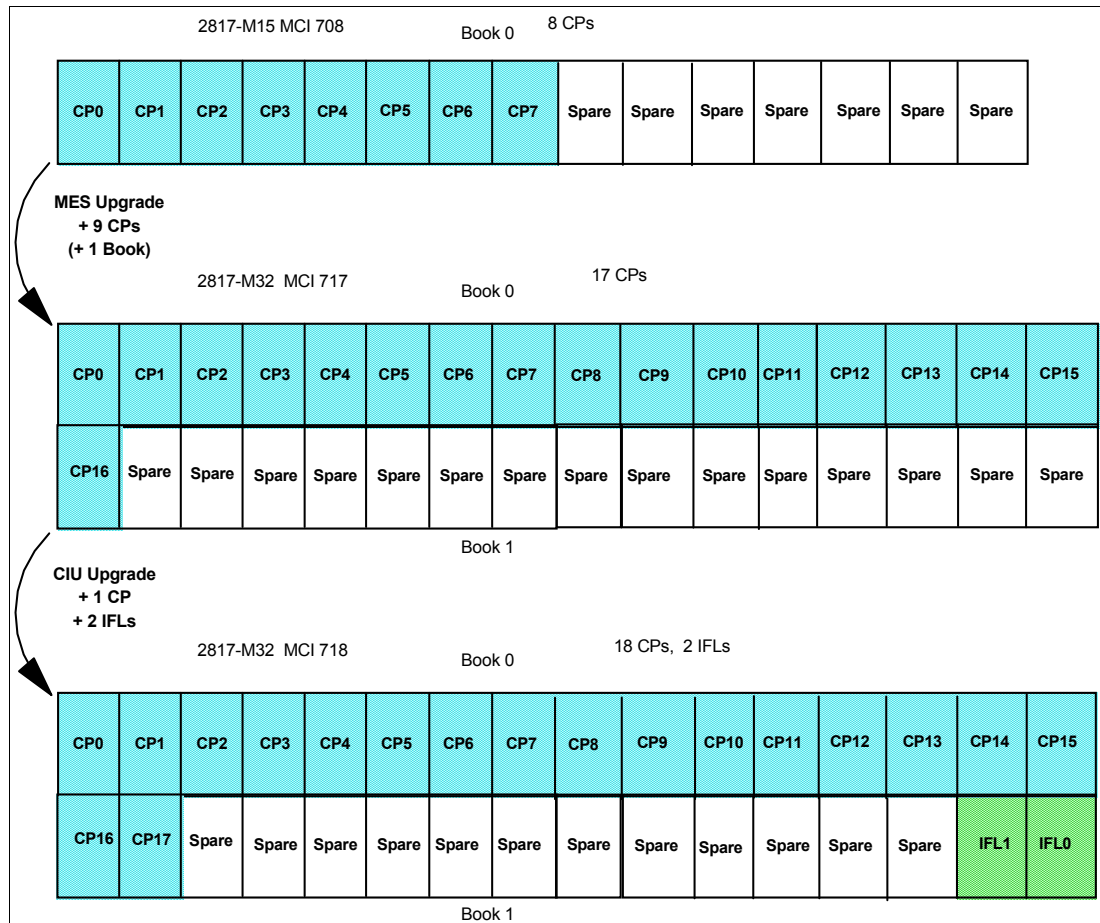
Figure 9-3   MES for processor example

A model M15 (one book), model capacity identifier 708 (eight CPs), is concurrently upgraded to a model M32 (two books), with model capacity identifier (MCI) 717 (which is 17 CPs). The model upgrade requires adding a book and assigning and activating nine PUs as CPs. Then, model M32, capacity identifier 717, is concurrently upgraded to a capacity identifier 718 (which is 18 CPs) with two IFLs by assigning and activating three more unassigned PUs (one as CP and two as IFLs). If needed, additional logical partitions can be created concurrently to use the newly added processors.

> **Note:** Up to 80 logical processors, including reserved processors, can be defined to a logical partition. You should not define more processors to a logical partition than the target operating system supports:
>
> ► z/OS V1R12, V1R11, V1R10 with PTF's support up to 80 as a combination of CPs, zAAPs, and zIIPs.
> ► z/VM supports up to 32 processors of any type.

Software charges, based on the total capacity of the server on which the software is installed, are adjusted to the new capacity after the MES upgrade.

Software products that use Workload License Charge (WLC) might not be affected by the server upgrade, because their charges are based on partition utilization and not based on the server total capacity. For more information about WLC, see 8.12.1, "Workload License Charge" on page 259.

9.3.2 MES upgrade for memory

MES upgrade for memory can concurrently add more memory by:

► Enabling, through LICCC, additional capacity up to the limit of the current installed memory cards

► Concurrently installing additional books and LICCC-enabling memory capacity on the new books.

Plan-ahead memory features are available to allow better control over future memory upgrades. See 2.5.6, "Flexible memory option" on page 46, and 2.5.7, "Plan-ahead memory" on page 47, for details about plan-ahead memory features.

If the z196 is a multiple-book configuration, using the enhanced book availability (EBA) feature to remove a book and add memory cards or to upgrade the already-installed memory cards to a larger size and then using LICCC to enable the additional memory is possible. With proper planning, additional memory can be added non-disruptively to z/OS partitions and z/VM partitions. If necessary, new logical partitions can be created non-disruptively to use the newly added memory.

> **Note:** Upgrades requiring DIMM changes can be concurrent by using the enhanced book availability feature. Planning is required to see whether this is a viable option in your configuration. The use of the flexible memory option (FC 1996) and the plan-ahead memory features (FC1991 and FC1992) is the safest way to ensure that EBA can work with the least disruption.

The one-book model has, as a minimum, fifteen 4 GB DIMMs, resulting in 60 GB of installed memory in total. The minimum customer addressable storage is 32 GB. If you require more than that, a *non-concurrent* upgrade can install up to 752 GB of memory for customer use, by changing the existing DIMM sizes and adding additional DIMMs in all available slots in the

book. Another possibility is to add memory by *concurrently* adding a second book with sufficient memory into the configuration and then using LICCC to enable that memory.

A logical partition can dynamically take advantage of a memory upgrade if reserved storage has been defined to that logical partition. The reserved storage is defined to the logical partition as part of the image profile. Reserved memory can be configured online to the logical partition by using the LPAR dynamic storage reconfiguration (DSR) function. DSR allows a z/OS operating system image, and z/VM partitions, to add reserved storage to their configuration if any unused storage exists. The nondisruptive addition of storage to a z/OS and z/VM partition necessitates that pertinent operating system parameters have been prepared. If reserved storage has not been defined to the logical partition, the logical partition must be deactivated, the image profile changed, and the logical partition reactivated to allow the additional storage resources to be available to the operating system image.

9.3.3 MES upgrades for I/O

MES upgrades for I/O can concurrently add more I/O ports by one of the following methods:

► Enabling additional ports on the already installed I/O cards through LICCC

  LICCC-only upgrades can be done for ESCON channels and ISC-3 links, activating ports on the existing 16-port ESCON or ISC-3 daughter (ISC-D) cards.

► Installing additional I/O cards on an already installed I/O cage's slots

  The installed I/O cages must provide the number of I/O slots required by the target configuration.

► Installing additional I/O cards in an existing I/O drawer, or adding a new I/O drawer to hold the new I/O cards.

> **Note:** I/O cages *cannot* be installed concurrently.

Figure 9-4 shows a z196 that has 16 ESCON channels available on two 16-port ESCON channel cards installed in an I/O cage. Each channel card has eight ports enabled. In this example, eight additional ESCON channels are concurrently added to the configuration by enabling, through LICCC, four unused ports on each ESCON channel card.
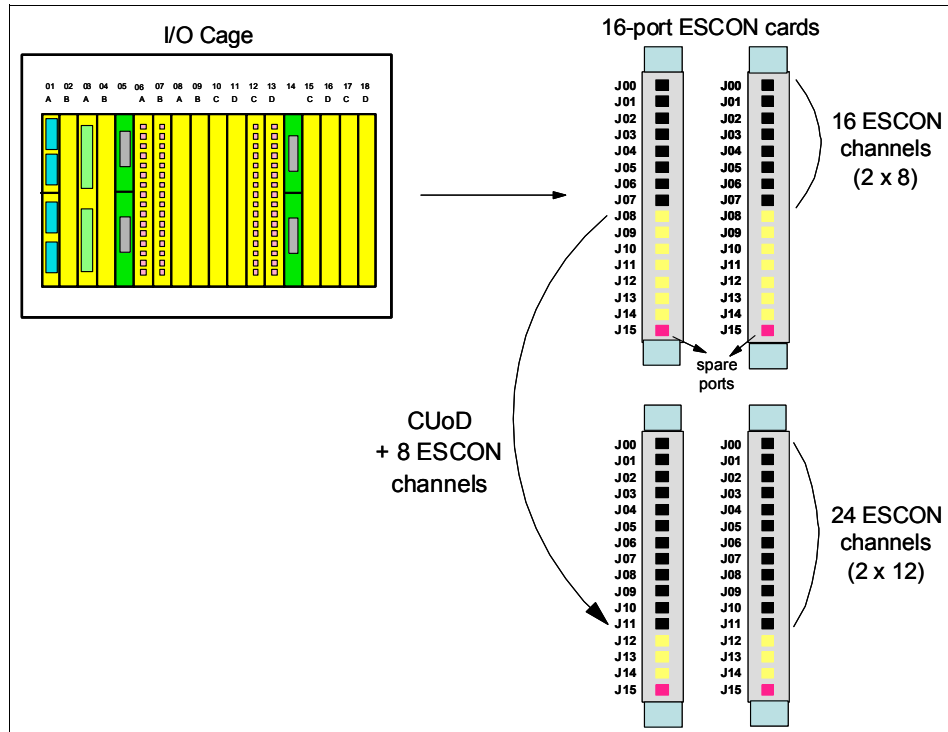
Figure 9-4   MES for I/O LICCC upgrade example

The additional channels installed concurrently to the hardware can also be concurrently defined in HSA and to an operating system by using the dynamic I/O configuration function. Dynamic I/O configuration can be used by z/OS or z/VM operating systems.

z/VSE, TPF, z/TPF, Linux on System z, and CFCC do *not* provide dynamic I/O configuration support. The installation of the new hardware is performed concurrently, but defining the new hardware to these operating systems requires an IPL.

To better exploit the MES for I/O capability, an initial configuration should be carefully planned to allow concurrent upgrades up to the target configuration. The plan-ahead concurrent conditioning process can include, in the initial configuration, the shipment of additional I/O cages required for future I/O upgrades. Another option is to configure all the I/O cards in I/O drawers which can be installed concurrently. It may not always be possible to do so, in which case a minimum number of I/O cages should be used to satisfy the customers I/O needs.

9.3.4 MES upgrades for the zBX

The MES upgrades for zBX can concurrently add blades if there are any slots available in existing blade chassis, chassis if there are any free spaces in existing racks, racks up to a maximum of 4, and entitlements for connections to the z196. For the IBM Smart Analytics Optimizer, the solution will continue to support applications using the z196 resources until the zBX has been upgraded and brought back into production status.

9.3.5 Plan-ahead concurrent conditioning

Concurrent Conditioning (FC 1999) and Control for Plan-Ahead (FC 1995) features, together with the input of a future target configuration, allow upgrades to exploit the order process configurator for concurrent I/O upgrades at a future time. If the initial configuration of a z196

can be installed with two power line-cords, order a plan-ahead feature for additional line-cords (FC 2000) if the future configuration will require additional power cords.

The plan-ahead feature identifies the content of the target configuration, thereby avoiding any down time associated with feature installation. As a result, Concurrent Conditioning may include, in the initial order, additional I/O cages to support the future I/O requirements.

Plan-ahead memory features enable you to install memory for future use:

► FC 1991 specifies memory to be installed but not used.

► FC 1992 is used to activate previously installed plan-ahead memory and can activate all the pre-installed memory or subsets of it.

Accurate planning and definition of the target configuration is vital to maximize the value of these features.

9.4 Permanent upgrade through the CIU facility

By using the CIU facility (through the IBM Resource Link on the Web), you may initiate a permanent upgrade for CPs, ICFs, zAAPs, zIIPs, IFLs, SAPs, or memory. When performed through the CIU facility, you add the resources; IBM personnel do not have to be present at the customer location. You may also unassign previously purchased CPs and IFLs processors through the CIU facility.

The capability to add permanent upgrades to a given server through the CIU facility requires that the permanent upgrade enablement feature (FC 9898) be installed on the server. A permanent upgrade might change the server model capacity identifier 4xx, 5xx, 6xx, or 7xx if additional CPs are requested or the capacity identifier is changed as part of the permanent upgrade, but it cannot change the server model, for example, 2817-Mvv. If necessary, additional logical partitions can be created concurrently to use the newly added processors.

> **Note:** A permanent upgrade of processors can provide a physical concurrent upgrade, resulting in more enabled processors available to a server configuration. Thus, additional planning and tasks are required for *nondisruptive* logical upgrades. See "Recommendations to avoid disruptive upgrades" on page 308 for more information.

Maintenance charges are automatically adjusted as a result of a permanent upgrade.

Software charges based on the total capacity of the server on which the software is installed are adjusted to the new capacity in place after the permanent upgrade is installed. Software products that use Workload License Charge (WLC) might not be affected by the server upgrade, because their charges are based on a logical partition utilization and not based on the server total capacity. See 8.12.1, "Workload License Charge" on page 259, for more information about WLC.

Figure 9-5 illustrates the CIU facility process on IBM Resource Link.
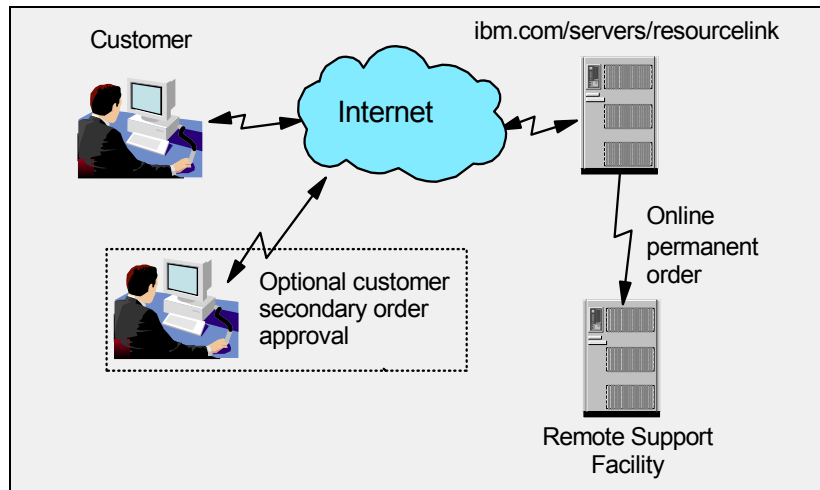
Figure 9-5   Permanent upgrade order example

The following sample sequence on IBM Resource Link initiates an order:

1. Sign on to Resource Link.

2. Select the **Customer Initiated Upgrade** option from the main Resource Link page. Customer and server details associated with the user ID are listed.

3. Select the server that will receive the upgrade. The current configuration (PU allocation and memory) is shown for the selected server.

4. Select **Order Permanent Upgrade** function. Resource Link limits options to those that are valid or possible for this configuration.

5. After the target configuration is verified by the system, accept or cancel the order.

   An order is created and verified against the pre-established agreement.

6. Accept or reject the price that is quoted. A secondary order approval is optional.

   Upon confirmation, the order is processed. The LICCC for the upgrade should be available within hours.

Figure 9-6 illustrates the process for a permanent upgrade. When the LICCC is passed to the remote support facility, you are notified through an e-mail that the upgrade is ready to be downloaded.
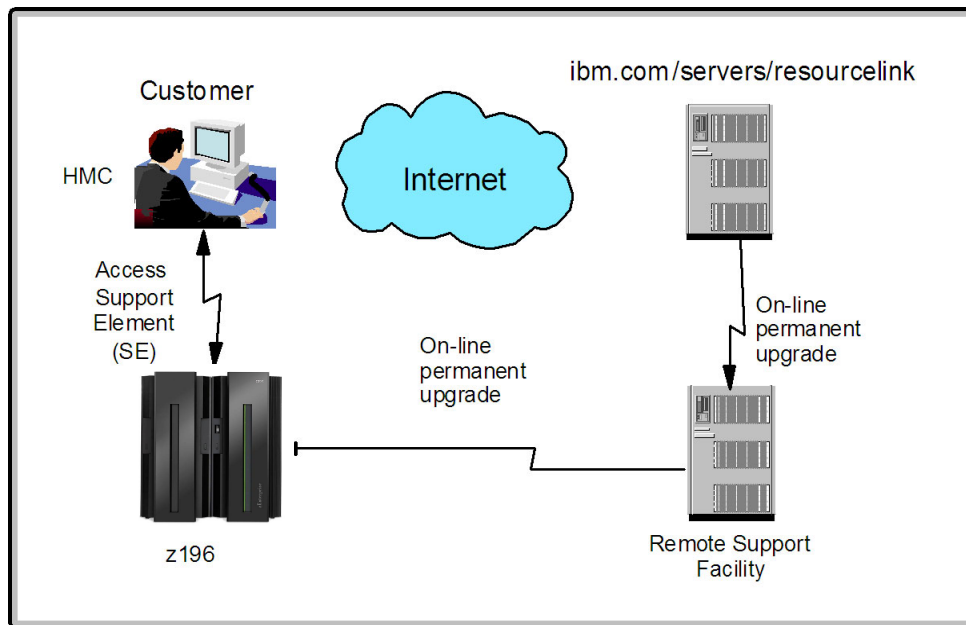
Figure 9-6   CIU-eligible order activation example

The two major components in the process are *ordering* and *retrieval* (along with activation).

9.4.1 Ordering

Resource Link provides the interface that enables you to order a concurrent upgrade for a server. You may create, cancel, view the order, and view the history of orders that were placed through this interface. Configuration rules enforce only valid configurations being generated within the limits of the individual server. Warning messages are issued if you select invalid upgrade options. The process allows only one permanent CIU-eligible order for each server to be placed at a time.

Figure 9-7 shows the initial view of the machine profile on Resource Link.

Figure 9-7   Machine profile

The number of CPs, ICFs, zAAPs, zIIPs, IFLs, SAPs, memory size, CBU features, unassigned CPs, and unassigned IFLs on the current configuration are displayed on the left side of the Web page.

Resource Link retrieves and stores relevant data associated with the processor configuration, such as the number of CPs and installed memory cards. It allows you to select only those upgrade options that are deemed valid by the order process. It allows upgrades only within the bounds of the currently installed hardware.

9.4.2 Retrieval and activation

After an order is placed and processed, the appropriate upgrade record is passed to the IBM support system for download.

When the order is available for download, you receive an e-mail that contains an activation number. You may then retrieve the order by using the Perform Model Conversion task from the Support Element (SE), or through Single Object Operation to the SE from an HMC.

In the Perform Model Conversion panel, select the **Permanent upgrades** option to start the process. See Figure 9-8.
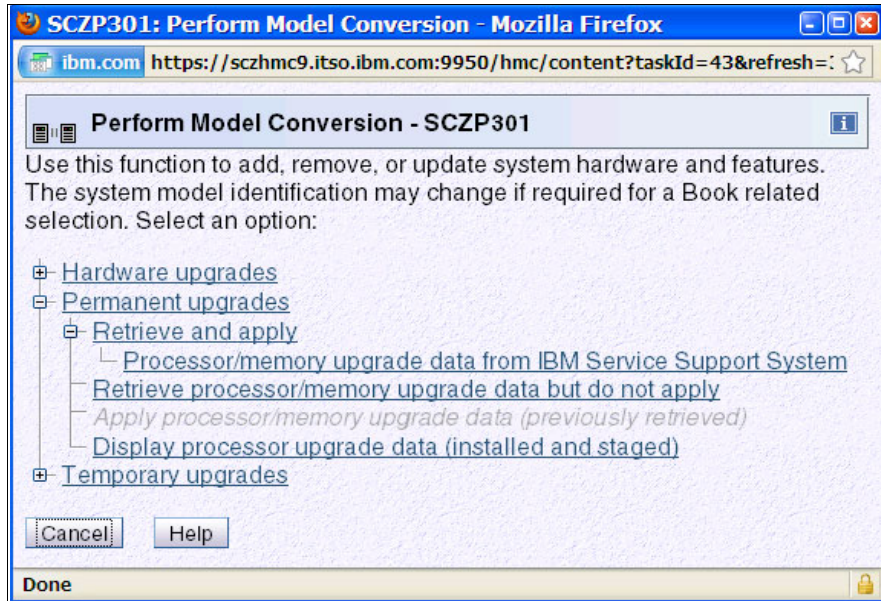
Figure 9-8   z196 Perform Model Conversion panel

The panel provides several possible options. If you select the **Retrieve and apply** data option, you are prompted to enter the order activation number to initiate the permanent upgrade. See Figure 9-9.
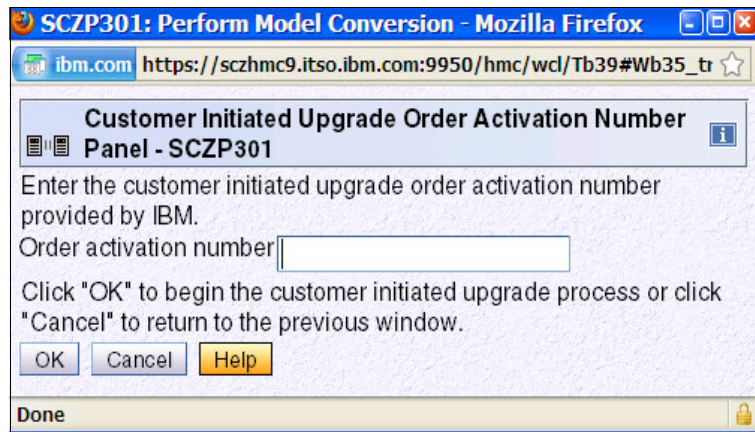
Figure 9-9   Customer Initiated Upgrade Order Activation Number Panel

9.5 On/Off Capacity on Demand

On/Off Capacity on Demand (On/Off CoD) allows you to temporarily enable PUs and unassigned IFLs available within the current model, or to change capacity settings for CPs to help meet your peak workload requirements.

9.5.1 Overview

The capacity for CPs is expressed in MSUs. Capacity for speciality engines is expressed in number of speciality engines. Capacity tokens are used to limit the resource consumption for all types of processor capacity.

Capacity tokens are introduced to provide better control over resource consumption when On/Off CoD offerings are activated. Token are represented as follows:

► For CP capacity, each token represents the amount of CP capacity that will result in one MSU of software cost for one day (an *MSU-day token*).

► For speciality engines, each token is equivalent to one speciality engine capacity for one day (an *engine-day token*).

Tokens are by capacity type, MSUs for CP capacity, and number of engines for speciality engines. Each speciality engine type has its own tokens, and each On/Off CoD record has separate token pools for each capacity type. During the ordering sessions on Resource Link, you decide how many tokens of each type should be created in an offering record. Each engine type must have tokens for that engine type to be activated. Capacity that has no tokens cannot be activated.

When resources from an On/Off CoD offering record containing capacity tokens are activated, a *billing window* is started. A billing window is always 24 hours in length. Billing takes place at the end of each billing window. The resources billed are the highest resource usage inside each billing window for each capacity type. An activation period is one or more complete billing windows, and represents the time from the first activation of resources in a record until the end of the billing window in which the last resource in a record is deactivated. At the end of each billing window, the tokens are decremented by the highest usage of each resource during the billing window. If any resource in a record does not have enough tokens to cover usage for the next billing window, the entire record will be deactivated.

On/Off CoD requires that the Online CoD Buying feature (FC 9900) be installed on the server that is to be upgraded.

On/Off CoD to Permanent Upgrade Option is a new offering, which is an offshoot of On/Off CoD and takes advantage of the aspects of the architecture. The customer is given a window of opportunity to assess capacity additions to their permanent configurations using On/Off CoD. If a purchase is made, the hardware On/Off CoD charges during this window, 3 days or less, are waived. If no purchase is made, then the customer is charged for the temporary use.

The resources eligible for temporary use are CPs, ICFs, zAAPs, zIIPs, IFLs, and SAPs. Temporary addition of memory and I/O ports is not supported. Unassigned PUs that are on the installed books can be temporarily and concurrently activated as CPs, ICFs, zAAPs, zIIPs, IFLs, SAPs through LICCC, up to twice the currently installed CP capacity and up to twice the number of ICFs, zAAPs, zIIPs, or IFLs. This means that an On/Off CoD upgrade cannot change the server Model 2817-Mvv. The addition of new books is not supported. However, activation of an On/Off CoD upgrade can increase the model capacity identifier 4xx, 5xx, 6xx, or 7xx.

9.5.2 Ordering

Concurrently installing temporary capacity by ordering On/Off CoD is possible, as follows:

► CP features equal to the MSU capacity of installed CPs

► IFL features up to the number of installed IFLs

► ICF features up to the number of installed ICFs

► zAAP features up to the number of installed zAAPs

► zIIP features up to the number of installed zIIPs

► SAPs up to three for model M15, seven for an M32, eleven for an M49, eighteen for an M66, and twenty-two for an M80.

On/Off CoD can provide CP temporary capacity in two ways:

► By increasing the number of CPs.

► For subcapacity models, capacity can be added by increasing the number of CPs or by changing the capacity setting of the CPs, or both. The capacity setting for all CPs must be the same. If the On/Off CoD is adding CP resources that have a capacity setting different from the installed CPs, then the base capacity settings are changed to match.

On/Off CoD has the following limits associated with its use:

– The number of CPs cannot be reduced.
– The target configuration capacity is limited to:

• Twice the currently installed capacity, expressed in MSUs for CPs

• Twice the number of installed IFLs, ICFs, zAAPs, and zIIPs. The number of SAPs that can be activated depends on the model described in 9.2.1, "Model upgrades" on page 269.

Table 9-3 on page 289 shows the valid On/Off CoD configurations for CPs on the subcapacity models.

On/Off CoD can be ordered as prepaid or postpaid:

► A prepaid On/Off CoD offering record contains resource descriptions, MSUs, number of speciality engines, and tokens that describe the total capacity that can be used. For CP capacity, the token contains MSU-days; for speciality engines, the token contains speciality engine-days.

► When resources on a prepaid offering are activated, they must have enough capacity tokens to allow the activation for an entire billing window, which is 24 hours. The resources remain active until you deactivate them or until one resource has consumed all of its capacity tokens. When that happens, all activated resources from the record are deactivated.

► A postpaid On/Off CoD offering record contains resource descriptions, MSUs, speciality engines, and may contain capacity tokens describing MSU-days and speciality engine-days.

► When resources in a postpaid offering record without capacity tokens are activated, those resources remain active until they are deactivated, or until the offering record expires, which is usually 180 days after its installation.

► When resources in a postpaid offering record with capacity tokens are activated, those resources must have enough capacity tokens to allow the activation for an entire billing window (24 hours). The resources remain active until they are deactivated or until one of the resource tokens are consumed, or until the record expires, usually 180 days after its installation. If one capacity token type is consumed, resources from the entire record are deactivated.

As an example, for a z196 with capacity identifier 502, two ways to deliver a capacity upgrade through On/Off CoD exist:

► The first option is to add CPs of the same capacity setting. With this option, the model capacity identifier could be changed to a 503, which would add one additional CP (making a 3-way) or to a 504, which would add two additional CPs (making a 4-way).

► The second option is to change to a different capacity level of the current CPs and change the model capacity identifier to a 602 or to a 702. The capacity level of the CPs is increased but no additional CPs are added. The 502 could also be temporarily upgraded to a 603 as indicated in the table, thus increasing the capacity level and adding another processor. The 415 does not have an upgrade path through On/Off CoD.

We recommend that you use the Large Systems Performance Reference (LSPR) information to evaluate the capacity requirements according to your workload type. LSPR data for current IBM processors is available at:

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

Table 9-3   Valid On/Off CoD upgrade examples

Capacity identifier	On/Off CoD CP4	On/Off CoD CP5	On/Off CoD CP6	On/Off CoD CP7
401	402	-	-	-
402	403, 404	-	-	-
403	404, 405, 406	-	-	-
404	405 - 408	-	-	-
405	406 - 410	-	-	-
406	407 - 413	-	-	-
407	408 - 412	-	-	-
408	409 - 415	-	-	-
409	410 - 415	-	-	-
410	411, 415	-	-	-
411	412 - 415	-	-	-
412	413 - 415	-	-	-
413	414, 415	-	-	-
414	415	-	-	-
415	-	-	-	-
501	-	502	601	-
502	-	503, 504	602, 603	-
503	-	504, 505, 506	603, 604	703
504	-	505 - 508	604 - 606	-
505	-	506 - 510	605 - 608	705
506	-	507 - 513	606 - 609	706
507	-	508 - 515	607 - 611	707
508	-	509 - 515	608 - 613	708
509	-	510 - 515	609 - 615	709
510	-	511 - 515	610 - 615	710
511	-	512 - 515	611 - 615	711
512	-	513 - 515	612 - 615	712
513	-	514, 515	613 - 615	713
514	-	515	614, 615	714
515	-	-	615	715
601	-	-	602	701
602	-	-	603, 604	702

Capacity identifier	On/Off CoD CP4	On/Off CoD CP5	On/Off CoD CP6	On/Off CoD CP7
603	-	-	604 - 606	703
604	-	-	605 - 608	704, 705
605	-	-	606 - 610	705, 706
606	-	-	607 - 613	706 - 708
607	-	-	608 - 615	707 - 709
608	-	-	609 - 615	708 - 710
609	-	-	610 - 615	709 - 712
610	-	-	611 - 615	710 - 713
611	-	-	612 - 615	711 - 715
612	-	-	613 - 615	712 - 716
613	-	-	614, 615	713 - 718
614	-	-	615	714 - 719
615	-	-	-	715 - 721
701	-	-	-	702
702	-	-	-	703, 704
703	-	-	-	704 - 706
704	-	-	-	705 - 708
705	-	-	-	706 - 711
706	-	-	-	707 - 713
707	-	-	-	708 - 716
708	-	-	-	709 - 718
709	-	-	-	710 - 721
710	-	-	-	711 - 723
711	-	-	-	712 - 726
712	-	-	-	713 - 728
713	-	-	-	714 - 731
714	-	-	-	715 - 733
715	-	-	-	716 - 736

The On/Off CoD hardware capacity is charged on a 24-hour basis. There is a grace period at the end of the On/Off CoD day. This allows up to an hour after the 24-hour billing period to either change the On/Off CoD configuration for the next 24-hour billing period or deactivate the current On/Off CoD configuration. The times when the capacity is activated and deactivated are maintained in the z196 and sent back to the support systems.

If On/Off capacity is already active, additional On/Off capacity can be added without having to return the server to its original capacity. If the capacity is increased multiple times within a 24-hour period, the charges apply to the highest amount of capacity active in the period. If additional capacity is added from an already active record containing capacity tokens, a check is made to control that the resource in question has enough capacity to be active for an entire billing window (24 hours). If that criteria is not met, no additional resources will be activated from the record.

If necessary, additional logical partitions can be activated concurrently to use the newly added processor resources.

> **Note:** On/Off CoD provides a concurrent *hardware* upgrade, resulting in more enabled processors available to a server configuration. Additional planning tasks are required for *nondisruptive* upgrades. See "Recommendations to avoid disruptive upgrades" on page 308.

To participate in this offering, you must have accepted contractual terms for purchasing capacity through the Resource Link, established a profile, and installed an On/Off CoD enablement feature on the server. Subsequently, you may concurrently install temporary capacity up to the limits in On/Off CoD and use it for up to 180 days. Monitoring occurs through the server call-home facility and an invoice is generated if the capacity has been enabled during the calendar month. The customer will continue to be billed for use of temporary capacity until the server is returned to the original configuration. If the On/Off CoD support is no longer needed, the enablement code must be removed.

On/Off CoD orders can be pre-staged in Resource Link to allow multiple optional configurations. The pricing of the orders is done at the time of the order, and the pricing can vary from quarter to quarter. Staged orders can have different pricing. When the order is downloaded and activated, the daily costs are based on the pricing at the time of the order. The staged orders do not have to be installed in order sequence. If a staged order is installed out of sequence, and later an order that was staged that had a higher price is downloaded, the daily cost will be based on the lower price.

Another possibility is to store unlimited On/Off CoD LICCC records on the Support Element with the same or different capacities at any given time, giving greater flexibility to quickly enable needed temporary capacity. Each record is easily identified with descriptive names, and you may select from a list of records that can be activated.

Resource Link provides the interface that allows you to order a dynamic upgrade for a specific server. You are able to create, cancel, and view the order. Configuration rules are enforced, and only valid configurations are generated based on the configuration of the individual server. After completing the prerequisites, orders for the On/Off CoD can be placed. The order process is to use the CIU facility on Resource Link.

You may order temporary capacity for CPs, ICFs, zAAPs, zIIPs, IFLs, or SAPs. Memory and channels are not supported on On/Off CoD. The amount of capacity is based on the amount of owned capacity for the different types of resources. An LICCC record is established and staged to Resource Link for this order. After the record is activated, it has no expiration date. However, an individual record can only be activated once. Subsequent sessions require a new order to be generated, producing a new LICCC record for that specific order. Alternatively the customer can use an auto renewal feature to eliminate the need for a manual replenishment of the On/Off CoD order. The is feature is implemented in Resource Link and the customer will have to check this feature in the machine profile. See Figure 9-10 for more details.

Figure 9-10   Order On/Off CoD record panel

9.5.3 On/Off CoD testing

Each On/Off CoD-enabled server is entitled to one no-charge 24-hour test. No IBM charges are assessed for the test, including no IBM charges associated with temporary hardware capacity, IBM software, or IBM maintenance. The test can be used to validate the processes to download, stage, install, activate, and deactivate On/Off CoD capacity.

This test can last up to a maximum duration of 24 hours, commencing upon the activation of any capacity resource contained in the On/Off CoD record. Activation levels of capacity can change during the 24-hour test period. The On/Off CoD test automatically terminates at the end of the 24-hour period.

In addition there is a possibility to perform administrative testing, through which no additional capacity is added to the server, but the customer can test all the procedures and automation for the management of the On/Off CoD facility.

Figure 9-11 is an example of an On/Off CoD order on the Resource Link Web page.

Figure 9-11   On/Off CoD order example

The example order in Figure 9-11 is a On/Off CoD order for 100% more CP capacity and for six ICFs, four zAAPs, four zIIPs, and six SAPs. The maximum number of CPs, ICFs, zAAPs, zIIPs, and IFLs is limited by the current number of available unused PUs of the installed books. The maximum number of SAPs is determined by the model number and the number of available PUs on the already installed books.

9.5.4 Activation and deactivation

When a previously ordered On/Off CoD is retrieved from Resource Link, it is downloaded and stored on the SE hard disk. You may activate the order when the capacity is needed, either manually or through automation.

If the On/Off CoD offering record does not contain resource tokens, you must take action to deactivate the temporary capacity. Deactivation is accomplished from the Support Element and is nondisruptive. Depending on how the additional capacity was added to the logical partitions, you might be required to perform tasks at the logical partition level in order to remove the temporary capacity. For example, you might have to configure offline CPs that had been added to the partition, or deactivate additional logical partitions created to use the temporary capacity, or both.

On/Off CoD orders can be staged in Resource Link so that multiple orders are available. An order can only be downloaded and activated one time. If a different On/Off CoD order is required or a permanent upgrade is needed, it can be downloaded and activated without having to restore the system to its original purchased capacity.

In support of automation, an API is provided that allows the activation of the On/Off CoD records. The activation is performed from the HMC and requires specifying the order number. With this API, automation code can be used to send an activation command along with the order number to the HMC to enable the order.

9.5.5 Termination

A customer is contractually obligated to terminate the On/Off CoD right-to-use feature when a transfer in asset ownership occurs. A customer may also choose to terminate the On/Off CoD right-to-use feature without transferring ownership. Application of FC 9898 terminates the right to use the On/Off CoD. This feature cannot be ordered if a temporary session is already active. Similarly, the CIU enablement feature cannot be removed if a temporary session is active. Any time the CIU enablement feature is removed, the On/Off CoD right-to-use is simultaneously removed. Reactivating the right-to-use feature subjects the customer to the terms and fees that apply at that time.

Upgrade capability during On/Off CoD

Upgrades involving physical hardware are supported while an On/Off CoD upgrade is active on a particular z196. LICCC-only upgrades can be ordered and retrieved from Resource Link and applied while an On/Off CoD upgrade is active. LICCC-only memory upgrades can be retrieved and applied while a On/Off CoD upgrade is active.

Repair capability during On/Off CoD

If the z196 requires service while an On/Off CoD upgrade is active, the repair can take place without affecting the temporary capacity.

Monitoring

When you activate an On/Off CoD upgrade, an indicator is set in vital product data. This indicator is part of the call-home data transmission, which is sent on a scheduled basis. A time stamp is placed into call-home data when the facility is deactivated. At the end of each calendar month, the data is used to generate an invoice for the On/Off CoD that has been used during that month.

Maintenance

The maintenance price is adjusted as a result of an On/Off CoD activation.

Software

Software Parallel Sysplex License Charge (PSLC) customers are billed at the MSU level represented by the combined permanent and temporary capacity. All PSLC products are billed at the peak MSUs enabled during the month, regardless of usage. Customers with WLC licenses are billed by product at the highest four-hour rolling average for the month. In this instance, temporary capacity does not necessarily increase the software bill until that capacity is allocated to logical partitions and actually consumed.

Results from the STSI instruction reflect the current permanent and temporary CPs. See "Store system information (STSI) instruction" on page 305 for more details.

9.5.6 z/OS capacity provisioning

The z196 provisioning capability combined with CPM functions in z/OS provides a flexible, automated process to control the activation of On/Off Capacity on Demand. The z/OS provisioning environment is shown in Figure 9-12.
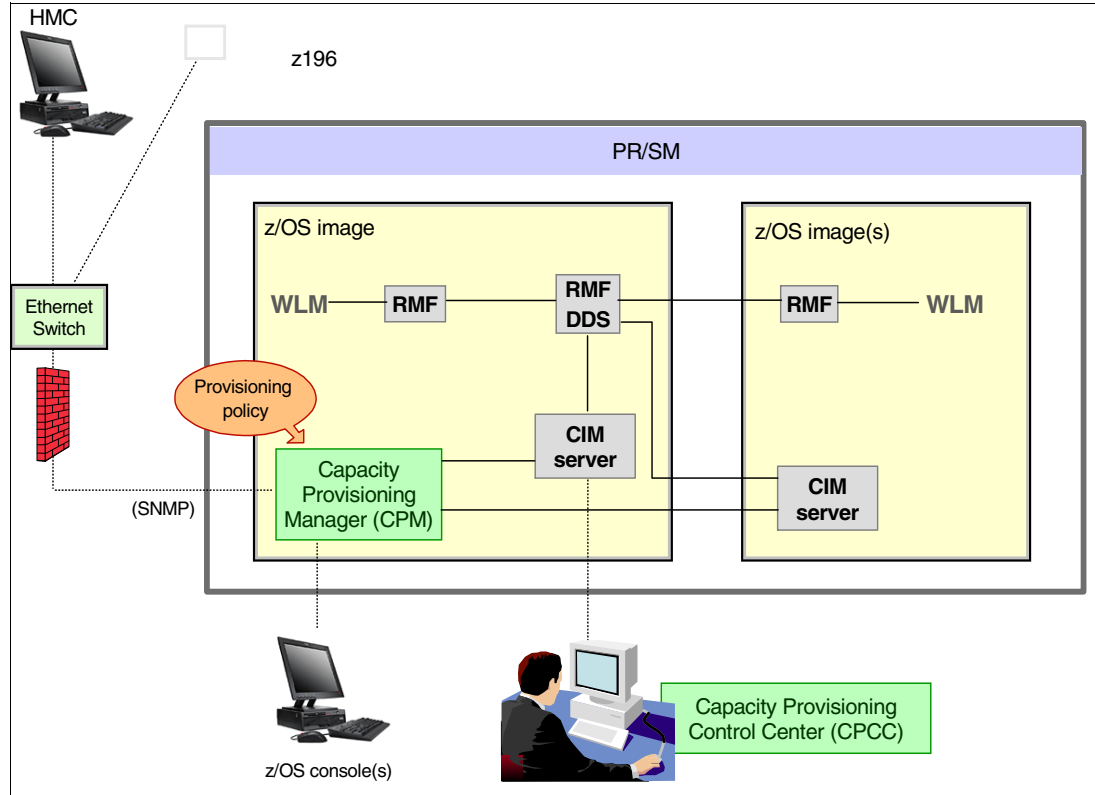
Figure 9-12   The capacity provisioning infrastructure

The z/OS WLM manages the workload by goals and business importance on each z/OS system. WLM metrics are available through existing interfaces and are reported through Resource Measurement Facility™ (RMF) Monitor III, with one RMF gatherer for each z/OS system.

Sysplex-wide data aggregation and propagation occur in the RMF distributed data server (DDS). The RMF Common Information Model (CIM) providers and associated CIM models publish the RMF Monitor III data.

The Capacity Provisioning Manager (CPM), a function inside z/OS, retrieves critical metrics from one or more z/OS systems through the Common Information Model (CIM) structures and protocol. CPM communicates to (local or remote) Support Elements and HMCs through the SNMP protocol.

CPM has visibility of the resources in the individual offering records, and the capacity tokens. When CPM decides to activate resources, a check is performed to determine whether enough capacity tokens remain for the specified resource to be activated for at least 24 hours. If not enough tokens remain, no resource from the On/Off CoD record is activated.

If a capacity token is completely consumed during an activation driven by the CPM, the corresponding On/Off CoD record is deactivated prematurely by the system, even if the CPM has activated this record, or parts of it. You do, however, receive warning messages if

capacity tokens are getting close to being fully consumed. You receive the messages five days before a capacity token is fully consumed. The five days are based on the assumption that the consumption will be constant for the 5 days. The recommendation is to put operational procedures in place to handle these situations. You may either deactivate the record manually, let it happen automatically, or replenish the specified capacity token by using the Resource Link application.

The Capacity Provisioning Control Center (CPCC), which resides on a workstation, provides an interface to administer capacity provisioning policies. The CPCC is not required for regular CPM operation.

The control over the provisioning infrastructure is executed by the CPM through the Capacity Provisioning Domain (CPD) controlled by the Capacity Provisioning Policy (CPP).

The Capacity Provisioning Domain is shown in Figure 9-13.
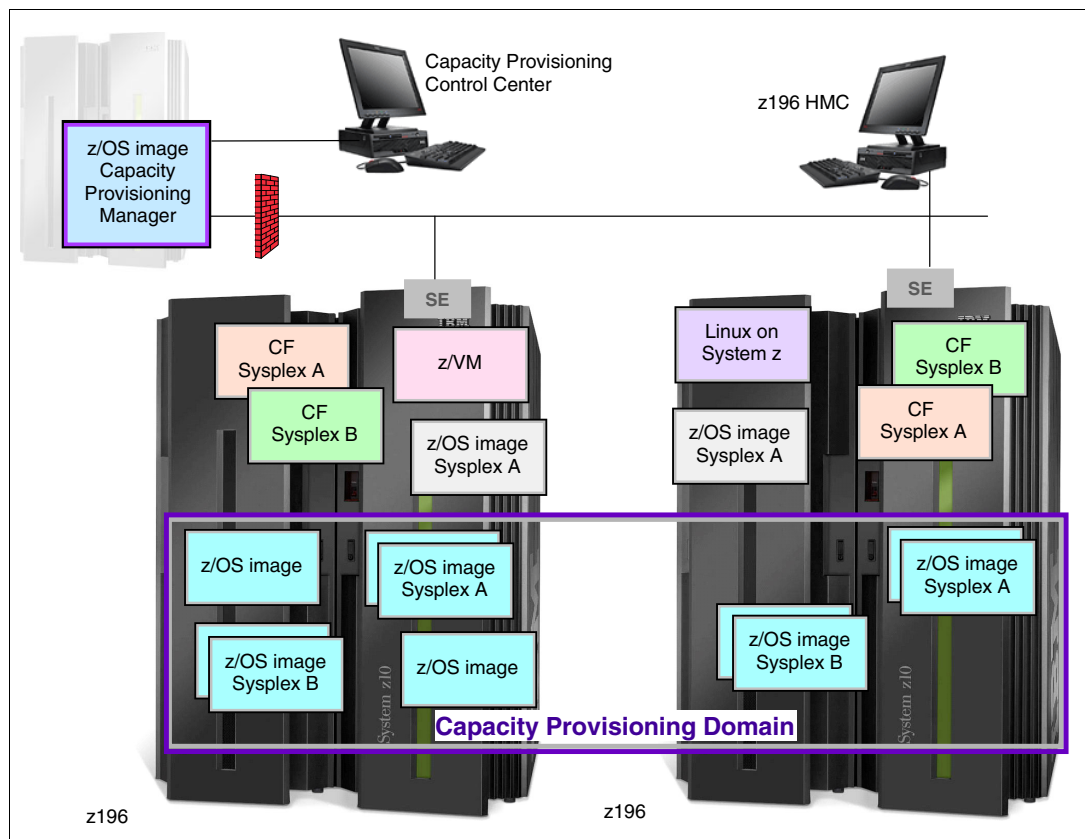
Figure 9-13   The Capacity Provisioning Domain

The Capacity Provisioning Domain represents the central processor complexes (CPCs) that are controlled by the Capacity Provisioning Manager. The HMCs of the CPCs within a CPD must be connected to the same processor LAN. Parallel Sysplex members can be part of a CPD. There is no requirement that all members of a Parallel Sysplex must be part of the CPD, but participating members must all be part of the same CPD.

The Capacity Provisioning Control Center (CPCC) is the user interface component. Administrators work through this interface to define domain configurations and provisioning policies, but it is not needed during production. The CPCC is installed on a Microsoft Windows® workstation.

CPM operates in four modes, allowing for different levels of automation:

► Manual mode

   Use this command-driven mode when no CPM policy is active.

► Analysis mode

   In analysis mode:

   – CPM processes capacity-provisioning policies and informs the operator when a provisioning or deprovisioning action is required according to policy criteria.

   – The operator determines whether to ignore the information or to manually upgrade or downgrade the system by using the HMC, the SE, or available CPM commands.

► Confirmation mode

   In this mode, CPM processes capacity provisioning policies and interrogates the installed temporary offering records. Every action proposed by the CPM needs to be confirmed by the operator.

► Autonomic mode

   This mode is similar to the confirmation mode, but no operator confirmation is required.

A number of reports are available in all modes, containing information about workload and provisioning status and the rationale for provisioning recommendations. User interfaces are through the z/OS console and the CPCC application.

The provisioning policy defines the circumstances under which additional capacity may be provisioned (when, which, and how). The three elements in the criteria are:

► A time condition is when provisioning is allowed, as follows:

   – Start time indicates when provisioning can begin.
   – Deadline indicates that provisioning of additional capacity no longer allowed
   – End time indicates that deactivation of additional capacity should begin.

► A workload condition is which work qualifies for provisioning. Parameters include:

   – The z/OS systems that may execute eligible work

   – Importance filter indicates eligible service class periods, identified by WLM importance

   – Performance indicator (PI) criteria:

      • Activation threshold: PI of service class periods must exceed the activation threshold for a specified duration before the work is considered to be suffering.

      • Deactivation threshold: PI of service class periods must fall below the deactivation threshold for a specified duration before the work is considered to no longer be suffering.

   – Included service classes are eligible service class periods.

   – Excluded service classes are service class periods that should not be considered.

   Note: If no workload condition is specified, the full capacity described in the policy will be activated and deactivated at the start and end times specified in the policy.

► Provisioning scope is how much additional capacity may be activated, expressed in MSUs.

   Specified in MSUs, number of zAAPs, and number of zIIPs must be one specification per CPC that is part of the Capacity Provisioning Domain.

   The maximum provisioning scope is the maximum additional capacity that may be activated for all the rules in the Capacity Provisioning Domain.

The provisioning rule is:

In the specified time interval, if the specified workload is behind its objective, then up to the defined additional capacity may be activated.

The rules and conditions are named and stored in the Capacity Provisioning Policy.

For more information about z/OS Capacity Provisioning functions, see *z/OS MVS Capacity Provisioning User's Guide,* SA33-8299 .

Planning considerations for using automatic provisioning

Although only one On/Off CoD offering can be active at any one time, several On/Off CoD offerings can be present on the server. Changing from one to another requires that the active one be stopped before the inactive one can be activated. This operation decreases the current capacity during the change.

The provisioning management routines can interrogate the installed offerings, their content, and the status of the content of the offering. To avoid the decrease in capacity, we recommend that only one On/Off CoD offering be created on the server by specifying the maximum allowable capacity. The Capacity Provisioning Manager can then, at the time when an activation is needed, activate a subset of the contents of the offering sufficient to satisfy the demand. If, at a later time, more capacity is needed, the Provisioning Manager can activate more capacity up to the maximum allowed increase.

Having an unlimited number of offering records pre-staged on the SE hard disk is possible; changing content of the offerings if necessary is also possible.

> **Attention:** As previously mentioned, the CPM has control over capacity tokens for the On/Off CoD records. In a situation where a capacity token is completely consumed, the server deactivates the corresponding offering record. Therefore, a strong recommendation is that you prepare routines for catching the warning messages about capacity tokens being consumed, and have administrative routines in place for such a situation. The messages from the system begin five days before a capacity token is fully consumed. To avoid capacity records from being deactivated in this situation, replenish the necessary capacity tokens before they are completely consumed.

In a situation where a CBU offering is active on a server and that CBU offering is 100% or more of the base capacity, activating any On/Off CoD is not possible because the On/Off CoD offering is limited to the 100% of the base configuration.

The Capacity Provisioning Manager operates based on Workload Manager (WLM) indications, and the construct used is the performance index (PI) of a service class period. It is extremely important to select service class periods that are appropriate for the business application that needs more capacity. For example, the application in question might be executing through several service class periods, where the first period might be the important one. The application might be defined as importance level 2 or 3, but might depend on other work executing with importance level 1. Therefore, considering which workloads to control, and which service class periods to specify is very important.

9.6 Capacity for Planned Event

Capacity for Planned Event (CPE) is offered with the z196 to provide replacement backup capacity for planned down-time events. For example, if a server room requires an extension

or repair work, replacement capacity can be installed temporarily on another z196 in the customer's environment.

> **Note:** CPE is for planned replacement capacity only and *cannot* be used for peak workload management.

The feature codes are:

► FC 6833 Capacity for Planned Event enablement
► FC 0116 - 1 CPE Capacity Unit
► FC 0117 - 100 CPE Capacity Unit
► FC 0118 - 10000 CPE Capacity Unit
► FC 0119 - 1 CPE Capacity Unit-IFL
► FC 0120 - 100 CPE Capacity Unit-IFL
► FC 0121 - 1 CPE Capacity Unit-ICF
► FC 0122 - 100 CPE Capacity Unit-ICF
► FC 0123 - 1 CPE Capacity Unit-zAAP
► FC 0124 - 100 CPE Capacity Unit-zAAP
► FC 0125 - 1 CPE Capacity Unit-zIIP
► FC 0126 - 100 CPE Capacity Unit-zIIP
► FC 0127 - 1 CPE Capacity Unit-SAP
► FC 0128 - 100 CPE Capacity Unit-SAP

The feature codes are calculated automatically when the CPE offering is configured. Whether using the eConfig tool or the Resource Link, a target configuration must be ordered consisting of a model identifier or a number of speciality engines, or both. Based on the target configuration, a number of feature codes from the list above is calculated automatically and a CPE offering record is constructed.

CPE is intended to replace capacity lost within the enterprise because of a planned event such as a facility upgrade or system relocation. CPE is intended for short duration events lasting up to a maximum of three days. Each CPE record, after it is activated, gives the you access to dormant PUs on the server that you have a contract for as described above by the feature codes. Processor units can be configured in any combination of CP or specialty engine types (zIIP, zAAP, SAP, IFL, and ICF). At the time of CPE activation the contracted configuration will be activated. The general rule of one zIIP and one zAAP for each configured CP will be controlled for the contracted configuration.

The processors that can be activated by CPE come from the available unassigned PUs on any installed book. CPE features can be added to an existing z196 non-disruptively. A one-time fee is applied for each individual CPE event depending on the contracted configuration and its resulting feature codes. Only one CPE contract can be ordered at a time.

The base server configuration must have sufficient memory and channels to accommodate the potential requirements of the large CPE-configured server. It is important to ensure that all required functions and resources are available on the server where CPE is activated, including CF LEVELs for coupling facility partitions, memory, cryptographic functions, and including connectivity capabilities.

The CPE configuration is activated temporarily and provides additional PUs in addition to the server's original, permanent configuration. The number of additional PUs is predetermined by the number and type of feature codes configured as described above by the feature codes. The number PUs that can be activated is limited by the unused capacity available on the server. For example:

► A model M32 with 16 CPs, no IFLs, ICFs, or zAAPs, has 16 unassigned PUs available.

► A model M49 with 28 CPs, 1 IFL, and 1 ICF has 19 unassigned PUs available.

When the planned event is over, the server must be returned to its original configuration. You may deactivate the CPE features at any time before the expiration date.

A CPE contract must be in place before the special code that enables this capability can be installed on the server. CPE features can be added to an existing z196 non-disruptively.

9.7 Capacity Backup

Capacity Backup (CBU) provides reserved emergency backup processor capacity for unplanned situations in which capacity is lost in another part of your enterprise and you want to recover by adding the reserved capacity on a designated z196.

CBU is the quick, temporary activation of PUs and is available as follows:

► For up to 90 contiguous days, in case of a loss of processing capacity as a result of an emergency or disaster recovery situation

► For 10 days for testing your disaster recovery procedures

> **Note:** CBU is for disaster and recovery purposes only and *cannot* be used for peak workload management or for a planned event.

9.7.1 Ordering

The CBU process allows for CBU to activate CPs, ICFs, zAAPs, zIIPs, IFLs, and SAPs. To be able to use the CBU process a CBU enablement feature (FC 9910) must be ordered and installed. You must order the quantity and type of PU that you require. The feature codes are:

► 6805: Additional test activations
► 6817: Total CBU years ordered
► 6818: CBU records ordered
► 6820: Single CBU CP-year
► 6821: 25 CBU CP-year
► 6822: Single CBU IFL-year
► 6823: 25 CBU IFL-year
► 6824: Single CBU ICF-year
► 6825: 25 CBU ICF-year
► 6826: Single CBU zAAP-year
► 6827: 25 CBU zAAP-year
► 6828: Single CBU zIIP-year
► 6829: 25 CBU zIIP-year
► 6830: Single CBU SAP-year
► 6831: 25 CBU SAP-year
► 6832: CBU replenishment

The CBU entitlement record (6818) contains an expiration date that is established at the time of order and is dependent upon the quantity of CBU years (6817). You have the capability to extend your CBU entitlements through the purchase of additional CBU years. The number of 6817 per instance of 6818 remains limited to five and fractional years are rounded up to the near whole integer when calculating this limit. For instance, if there are two years and eight months to the expiration date at the time of order, the expiration date can be extended by no

more than two additional years. One test activation is provided for each additional CBU year added to the CBU entitlement record.

Feature code 6805 allows for ordering additional tests in increments of one. The total number of tests allowed is 15 for each feature code 6818.

The processors that can be activated by CBU come from the available unassigned PUs on any installed book. The maximum number of CBU features that can be *ordered* is 80. The number of features that can be *activated* is limited by the number of unused PUs on the server. For example:

► A model M15 with Capacity Model Identifier 410 can activate up to 15 CBU features: ten to change the capacity setting of the existing CPs and five to activate unused PUs.

► A model M32 with 15 CPs, four IFLs, and one ICF has twelve unused PUs available. It can *activate* up to twelve CBU features.

However, the ordering system allows for over-configuration in the order itself. You may *order* up to 80 CBU features regardless of the current configuration, however at *activation*, only the capacity already installed can be *activated*. Note that at activation, you can decide to activate only a sub-set of the CBU features that are ordered for the system.

Subcapacity makes a difference in the way the CBU features are done. On the full-capacity models, the CBU features indicate the amount of additional capacity needed. If the amount of necessary CBU capacity is equal to four CPs, then the CBU configuration would be four CBU CPs.

The subcapacity models have multiple capacity settings of 4xx, 5xx, or 6xx; the standard models have capacity setting 7xx. The number of CBU CPs must be equal to or greater than the number of CPs in the base configuration, and all the CPs in the CBU configuration must have the same capacity setting. For example, if the base configuration is a 2-way 402, then providing a CBU configuration of a 4-way of the same capacity setting requires two CBU feature codes. If the required CBU capacity changes the capacity setting of the CPs, then going from model capacity identifier 402 to a CBU configuration of a 4-way 504 would require four CBU feature codes with a capacity setting of 5xx.

If the capacity setting of the CPs is changed, then more CBU features are required, not more physical PUs. This means that your CBU contract requires more CBU features if the capacity setting of the CPs is changed.

Note that CBU can add CPs through LICCC-only, and the z196 must have the proper number of books installed to allow the required upgrade. CBU can change the model capacity identifier to a *higher* value than the base setting, 4xx, 5xx, or 6xx, but does not change the *server* model 2817-Mvv. The CBU feature cannot *decrease* the capacity setting.

A CBU contract must be in place before the special code that enables this capability can be installed on the server. CBU features can be added to an existing z196 non-disruptively. For each machine enabled for CBU, the authorization to use CBU is available for a definite number of years of 1-5 years.

The installation of the CBU code provides an alternate configuration that can be activated in case of an actual emergency. Five CBU tests, lasting up to 10 days each, and one CBU activation, lasting up to 90 days for a real disaster and recovery, are typically allowed in a CBU contract.

The alternate configuration is activated *temporarily* and provides additional capacity greater than the server's original, *permanent* configuration. At activation time, you determine the

capacity required for a given situation, and you can decide to activate only a sub-set of the capacity specified in the CBU contract.

> **Note:** Do not run production work on a server that has an active test CBU. Instead, run only a copy of production.

The base server configuration must have sufficient memory and channels to accommodate the potential requirements of the large CBU target server. Ensure that all required functions and resources are available on the backup servers, including CF LEVELs for coupling facility partitions, memory, and cryptographic functions, as well as connectivity capabilities.

When the emergency is over (or the CBU test is complete), the server must be taken back to its original configuration. The CBU features can be deactivated by the customer at any time before the expiration date. Failure to deactivate the CBU feature before the expiration date can cause the system to degrade gracefully back to its original configuration. The system does *not* deactivate dedicated engines, or the last of in-use shared engines.

> **Note:** CBU for processors provides a concurrent upgrade, resulting in more enabled processors or changed capacity settings available to a server configuration, or both. You decide, at activation time, to activate a sub-set of the CBU features ordered for the system. Thus, additional planning and tasks are required for *nondisruptive* logical upgrades. See "Recommendations to avoid disruptive upgrades" on page 308.

For detailed instructions, see the *System z Capacity on Demand User's Guide*, SC28-6846.

9.7.2 CBU activation and deactivation

The activation and deactivation of the CBU function is a customer responsibility and does not require on-site presence of IBM service personnel. The CBU function is activated/deactivated concurrently from the HMC using the API. On the SE, CBU is be activated either using the Perform Model Conversion task or through the API (API enables task automation).

CBU activation
CBU is activated from the SE by using the Perform Model Conversion task or through automation by using API on the SE or the HMC. In case of real disaster, use the Activate CBU option to activate the 90-day period.

Image upgrades
After the CBU activation, the z196 can have more capacity, more active PUs, or both. The additional resources go into the resource pools and are available to the logical partitions. If the logical partitions have to increase their share of the resources, the logical partition weight can be changed or the number of logical processors can be concurrently increased by configuring reserved processors online. The operating system must have the capability to concurrently configure more processors online. If necessary, additional logical partitions can be created to use the newly added capacity.

CBU deactivation
To deactivate the CBU, the additional resources have to be released from the logical partitions by the operating systems. In some cases, this is a matter of varying the resources offline. In other cases, it can mean shutting down operating systems or deactivating logical partitions. After the resources have been released, the same facility on the SE is used to turn

off CBU. To deactivate CBU, click the **Undo temporary upgrade** option from the Perform Model Conversion task on the SE.

CBU testing

Test CBUs are provided as part of the CBU contract. CBU is activated from the SE by using the Perform Model Conversion task. Select the test option to initiate a 10-day test period. A standard contract allows five tests of this type. However, you may order additional tests in increments of one up to a maximum of 15 for each CBU order. The test CBU has a 10-day limit and must be deactivated in the same way as the real CBU, using the same facility through the SE. Failure to deactivate the CBU feature before the expiration date can cause the system to degrade gracefully back to its original configuration. The system does *not* deactivate dedicated engines, or the last of in-use shared engine. Testing can be accomplished by ordering a diskette, calling the support center, or using the facilities on the SE. The customer has the possibility of purchasing additional tests.

CBU example

An example of a capacity backup operation could be as follows; 12 CBU features are installed on a backup model M32 with model capacity identifier 708. When a production model M15 with model capacity identifier 708 has an unplanned outage, the backup server can be temporarily upgraded from model capacity identifier 708 to 720, so that the capacity can take over the workload from the failed production server.

Furthermore, you may configure systems to back up each other. For example, if you use two models of M15 model capacity identifier 705 for the production environment, each can have five or more features installed. If one server suffers an outage, the other one uses a temporary upgrade to recover approximately the total original capacity.

9.7.3 Automatic CBU enablement for GDPS

The intent of the Geographically Dispersed Parallel Sysplex™ (GDPS) CBU is to enable automatic management of the PUs provided by the CBU feature in the event of a server or site failure. Upon detection of a site failure or planned disaster test, GDPS will concurrently add CPs to the servers in the take-over site to restore processing power for mission-critical production workloads. GDPS automation does the following tasks:

► Performs the analysis required to determine the scope of the failure. This minimizes operator intervention and the potential for errors.

► Automates authentication and activation of the reserved CPs

► Automatically restarts the critical applications after reserved CP activation.

► Reduces the outage time to restart critical workloads from several hours to minutes.

The GDPS service is for z/OS only, or for z/OS in combination with Linux on System z.

9.8 Nondisruptive upgrades

Continuous availability is an increasingly important requirement for most customers, and even planned outages are no longer acceptable. Although Parallel Sysplex clustering technology is the best continuous availability solution for z/OS environments, non-disruptive upgrades within a single server can avoid system outages and are suitable to additional operating system environments.

The z196 allows *concurrent* upgrades, meaning that dynamically adding more capacity to the server is possible. If operating system images running on the upgraded server do not require disruptive tasks in order to use the new capacity, the upgrade is also *non-disruptive.* This means that power-on reset (POR), logical partition deactivation, and IPL do not have to take place.

If the concurrent upgrade is intended to satisfy an *image upgrade* to a logical partition, the operating system running in this partition must also have the capability to concurrently configure more capacity online. z/OS operating systems have this capability. z/VM can concurrently configure new processors and I/O devices online, memory can be dynamically added to z/VM partitions.

If the concurrent upgrade is intended to satisfy the need for more operating system images, additional logical partitions can be created *concurrently* on the z196 server, including all resources needed by such logical partitions. These additional logical partitions can be activated concurrently.

These enhanced configuration options are made available through the separate HSA, which is introduced on the System z196.

Linux operating systems in general do *not* have the capability of adding more resources concurrently. However, Linux, and other types of virtual machines running under z/VM, can benefit from the z/VM capability to non-disruptively configure more resources online (processors and I/O).

With z/VM, Linux guests can manipulate their logical processors through the use of the Linux CPU hotplug daemon. The daemon can start and stop logical processors based on the Linux average load value. The daemon is available in Linux SLES 10 SP2. IBM is working with our Linux distribution partners to have the daemon available in other distributions for the System z servers.

Processors

CPs, ICFs, zAAPs, zIIPs, IFLs, and SAPs can be concurrently added to a z196 if unassigned PUs are available on any installed book. The number of zAAPs cannot exceed the number of CPs plus unassigned CPs. The same holds true for the zIIPs.

Additional books can also be installed concurrently, allowing further processor upgrades.

Concurrent upgrades are not supported with PUs defined as additional SAPs.

If necessary, additional logical partitions can be created concurrently to use the newly added processors.

The Coupling Facility Control Code (CFCC) can also configure more processors online to coupling facility logical partitions by using the CFCC image operations window.

Memory

Memory can be concurrently added up to the physical installed memory limit. Additional books can also be installed concurrently, allowing further memory upgrades by LICCC, enabling memory capacity on the new books.

Using the previously defined reserved memory, z/OS operating system images, and z/VM partitions, can dynamically configure more memory online, allowing nondisruptive memory upgrades. Linux on System z supports Dynamic Storage Reconfiguration.

I/O

I/O cards can be added concurrently if all the required infrastructure (I/O slots and HCAs) is present on the configuration. The plan-ahead process can assure that an initial configuration will have all the infrastructure required for the target configuration.

I/O ports can be concurrently added by LICCC, enabling available ports on ESCON and ISC-3 daughter cards.

Dynamic I/O configurations are supported by certain operating systems (z/OS and z/VM), allowing nondisruptive I/O upgrades. However, having dynamic I/O reconfiguration on a stand-alone coupling facility server is not possible because there is no operating system with this capability running on this server.

Cryptographic adapters

Crypto Express3 features can be added concurrently if all the required infrastructure, I/O slots, and STIs are present on the configuration. The plan-ahead process can assure that an initial configuration will have all the infrastructure required for the target configuration.

Concurrent upgrade considerations

By using MES upgrade, On/Off CoD, CBU, or CPE, a z196 can be concurrently upgraded from one model to another, either temporarily or permanently.

Enabling and using the additional processor capacity is transparent to most applications. However, certain programs depend on processor model-related information, for example, Independent Software Vendor (ISV) products. You should consider the effect on the software running on a z196 when you perform any of these configuration upgrades.

Processor identification

Two instructions are used to obtain processor information:

► Store System Information instruction (STSI)

   STSI reports the processor model and model capacity identifier for the base configuration and for any additional configuration changes through temporary upgrade actions. It fully supports the concurrent upgrade functions and is the preferred way to request processor information.

► Store CPU ID instruction (STIDP)

   STIDP is provided for purposes of backward compatibility.

Store system information (STSI) instruction

Figure 9-14 shows the relevant output from the STSI instruction. The STSI instruction returns the model capacity identifier for the permanent configuration, and the model capacity identifier for any temporary capacity. This is key to the functioning of Capacity on Demand offerings.

Figure 9-14   STSI output on z196

The model capacity identifier contains the base capacity, the On/Off CoD, and the CBU. The model permanent capacity identifier and the Model Permanent Capacity Rating contain the base capacity of the system, and the model temporary capacity identifier and model temporary capacity rating contain the base capacity and the On/Off CoD.

Store CPU ID instruction

The STIDP instruction provides information about the processor type, serial number, and logical partition identifier. See Table 9-4. The logical partition identifier field is a full byte to support greater than 15 logical partitions.

Table 9-4   STIDP output for z196

Description	Version code	CPU identification number		Machine type number	Logical partition 2-digit indicator
Bit position	0 - 7	8 - 15	16 - 31	32 - 48	48 - 63
Value	x'00' [a]	Logical partition ID[b]	6-digit number derived from the CPC serial number	x'2817'	x'8000' [c]

a. The version code for z196 is x00.
b. The logical partition identifier is a two-digit number in the range of 00 - 3F. It is assigned by the user on the image profile through the Support Element or HMC.
c. High order bit on indicates that the logical partition ID value returned in bits 8 - 15 is a two-digit value.

When issued from an operating system running as a guest under z/VM, the result depends on whether the SET CPUID command has been used, as follows:

► Without the use of the SET CPUID command, bits 0 - 7 are set to FF by z/VM, but the remaining bits are unchanged, which means that they are exactly as they would have been without running as a z/VM guest.

► If the SET CPUID command has been issued, bits 0 - 7 are set to FF by z/VM and bits 8 - 31 are set to the value entered in the SET CPUID command. Bits 32 - 63 are the same as they would have been without running as a z/VM guest.

Table 9-5 lists the possible output returned to the issuing program for an operating system running as a guest under z/VM.

Table 9-5   z/VM guest STIDP output for z196

Description	Version code	CPU identification number		Machine type number	Logical partition 2-digit indicator
Bit position	0 - 7	8 - 15	16 - 31	32 - 48	48 - 63
Without SET CPUID command	x'FF'	Logical partition ID	4-digit number derived from the CPC serial number	x'2817'	x'8000'
With SET CPUID command	x'FF'	6-digit number as entered by the command SET CPUID = nnnnnn		x'2817'	x'8000'

Planning for nondisruptive upgrades

Online permanent upgrades, On/Off CoD, CBU, and CPE can be used to concurrently upgrade a z196. However, certain situations require a disruptive task in order to enable the new capacity that was recently added to the server. Some of these situation can be avoided if planning is done in advance. Planning ahead is a key factor for non-disruptive upgrades.

The following list describes main reasons for disruptive upgrades. However, by carefully planning and by reviewing "Recommendations to avoid disruptive upgrades" on page 308, you can minimize the need for these outages:

▶ z/OS logical partition processor upgrades when reserved processors were not previously defined are disruptive to image upgrades.

▶ Logical partition memory upgrades when reserved storage was not previously defined are disruptive to image upgrades. z/OS and z/VM support this function.

▶ Installation of an I/O cage is disruptive.

▶ An I/O upgrade when the operating system cannot use the dynamic I/O configuration function is disruptive. Linux, z/VSE, TPF, z/TPF, and CFCC do not support dynamic I/O configuration.

Recommendations to avoid disruptive upgrades

Based on the previous list of reasons for disruptive upgrades ("Planning for nondisruptive upgrades" on page 307), here are several recommendations for avoiding or at least minimizing these situations, increasing the possibilities for nondisruptive upgrades:

▶ For z/OS logical partitions configure as many reserved processors (CPs, ICFs, zAAPs, and zIIPs) as possible.

Configuring reserved processors for all logical z/OS partitions *before* their activation enables them to be non-disruptively upgraded. The operating system running in the logical partition must have the ability to configure processors online. The total number of defined and reserved CPs cannot exceed the number of CPs supported by the operating system. z/OS V1R10 with PTF's, z/OS V1R11, and z/OS V1R12 support up to 80 processors including CPs, zAAPs, and zIIPs. z/VM supports up to 32 processors.

▶ Configure reserved storage to logical partitions.

Configuring reserved storage for all logical partitions *before* their activation enables them to be non-disruptively 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, even if no other book is already installed. The current partition storage limit is 1TB. z/OS and z/VM support this function.

▶ Consider the flexible and plan-ahead memory options.

Use a convenient entry point for memory capacity and consider the memory options to allow future upgrades within the memory cards already installed on the books. For details about the offerings, see

– 2.5.6, "Flexible memory option" on page 46
– 2.5.7, "Plan-ahead memory" on page 47

▶ Use the plan-ahead concurrent condition for I/O.

Use the plan-ahead concurrent condition process to include in the initial configuration all the I/O cages required by future I/O upgrades, allowing the planned concurrent I/O upgrades.

Considerations when installing additional books

During an upgrade, additional books can be installed concurrently. Depending on the number of additional books in the upgrade and your I/O configuration, an HCA2 rebalancing might be recommended for availability reasons. It will change PCHID numbers, requiring an I/O definition update.

9.9 Summary of Capacity on Demand offerings

The capacity on demand infrastructure and its offerings are major features introduced with the z196 server. The reasons for the introduction of these features are based on numerous customer requirements for more flexibility, granularity, and better business control over the System z infrastructure, operationally and financially.

One major customer requirement is to dismiss the necessity for a customer authorization connection to IBM Resource Link system at the time of activation of any offering. This requirement is being met by the z196. After the offerings have been installed on the z196, they can be activated at any time, completely at the customer's discretion. No intervention through IBM or IBM personnel is necessary. In addition, the activation of the Capacity Backup does not require a password.

The z196 can have up to eight offerings installed at the same time, with the limitation that only one of them can be an On/Off Capacity on Demand offering; the others can be any combination. The installed offerings can be activated fully or partially, and in any sequence and any combination. The offerings can be controlled manually through command interfaces on the HMC, or programmatically through a number of APIs, so that IBM applications, ISV programs, or customer-written applications, can control the usage of the offerings.

Resource consumption (and thus financial exposure) can be controlled by using capacity tokens in On/Off CoD offering records.

The Capacity Provisioning Manager (CPM) is an example of an application that uses the Capacity on Demand APIs to provision On/Off CoD capacity based on the requirements of the executing workload. The CPM cannot control other offerings.

10

RAS

This chapter describes several of the reliability, availability, and serviceability (RAS) features of the zEnterprise System.

The z196 design is focused on providing higher availability by reducing planned and unplanned outages. RAS can be accomplished with improved concurrent replace, repair, and upgrade functions for processors, memory, books, and I/O. RAS also extends to the nondisruptive capability for downloading Licensed Internal Code (LIC) updates. In most cases a capacity upgrade can be concurrent without a system outage. As an extension to the RAS capabilities we will discuss environmental controls implemented in the server to help reduce power consumption and cooling requirements.

The design of the memory on the z196 has taken a major step forward by implementing a fully redundant memory infrastructure - Redundant Array of Independent Memory - a concept similar to the RAID design used in external disk storage systems. The z196 is the only server in the industry offering this level of memory design.

To make the delivery and transmission of microcode (LIC), fixes and restoration/backup files are digitally signed. Any data transmitted to IBM Support is encrypted.

The design goal for the z196 has been to remove all sources of planned outages.

This chapter discusses the following topics:

10.1 z196 Availability characteristics

The following functions include availability characteristics on the z196:

► Enhanced book availability (EBA)

EBA is a *procedure* under which a book in a multi-book machine can be removed and re-installed during an upgrade- or repair action with no impact on the executing workload.

► Concurrent memory upgrade or replacement

Memory can be upgraded concurrently using LICCC if physical memory is available on the books. If the physical memory cards have to be changed in a multibook configuration, which would require the book to be removed, the enhanced book availability function can be useful. It requires the availability of additional resources on other books or reducing the need for resources during this action. To help ensure that the appropriate level of memory is available in a multiple-book configuration, consider the selection of the flexible memory option for providing additional resources to exploit EBA when repairing a book or memory on a book, or when upgrading memory where larger memory cards might be required.

Memory can be upgraded concurrently by using LICCC if physical memory is available as previously explained. The plan-ahead memory function available with the z196 server provides the ability to plan for nondisruptive memory upgrades by having the system pre-plugged based on a target configuration. Pre-plugged memory is enabled when you place an order through LICCC.

► Enhanced driver maintenance (EDM)

One of the greatest contributors to downtime during planned outages is Licensed Internal Code driver updates performed in support of new features and functions. The z196 is designed to support activating a selected new driver level concurrently.

► Concurrent HCA/MBA fanout addition or replacement

A Host Channel Adapter (HCA)/Memory Bus Adapter (MBA) fanout card provides the path for data between memory and I/O using InfiniBand (IFB) cables. With the z196, a hot-pluggable and concurrently upgradable HCA/MBA fanout card is available. Up to eight HCA/MBA fanout cards are available per book for a total of up to 32 HCA/MBA fanout cards when four books are installed. In the event of an outage, an HCA/MBA fanout card, used for I/O, may be concurrently repaired while redundant I/O interconnect ensures that no I/O connectivity is lost.

► Redundant I/O interconnect

Redundant I/O interconnect helps maintain critical connections to devices. The z196 allows a single book, in a multibook server, to be concurrently removed and reinstalled during an upgrade or repair, continuing to provide connectivity to the server I/O resources using a second path from a different book.

► Dynamic oscillator switch-over

The z196 has two oscillator cards, a primary and a backup. In the event of a primary card failure, the backup card is designed to transparently detect the failure, switch-over, and provide the clock signal to the server.

► Cooling improvements

The z196 comes with an improved cooling system including a water cooling option, MCM backup cooling with heat exchanger, and enhanced evaporator design minimizing temperature variations.

10.2 z196 RAS functions

Hardware RAS function improvements focus on addressing all sources of outages. Sources of outages have three classifications:

Unscheduled This outage occurs because of an unrecoverable malfunction in a hardware component of the server.

Scheduled This outage is caused by changes or updates that have to be done to the server in a timely fashion. A scheduled outage can be caused by a disruptive patch that has to be installed, or other changes that have to be done to the system.

Planned This outage is also caused by changes or updates that have to be done to the server. A planned outage can be caused by a capacity upgrade or a driver upgrade. A planned outage is usually requested by the customer and often requires pre-planning. The z196 design phase focused on this pre-planning effort and was able to simplify or eliminate it.

Unscheduled, scheduled, and planned outages have been addressed for the mainframe family of servers for many years.

A fixed size HSA of 16 GB helps eliminate pre-planning requirements for HSA and to provide flexibility to dynamically update the configuration.

Performing the following tasks dynamically is possible:

- ► Add a logical partition.
- ► Add a logical channel subsystem (LCSS).
- ► Add a subchannel set.
- ► Add a logical CP to a logical partition.
- ► Add a cryptographic coprocessor.
- ► Remove a cryptographic coprocessor.
- ► Enable I/O connections.
- ► Swap processor types.

In addition, by addressing the elimination of planned outages, the following tasks are also possible:

- ► Concurrent driver upgrades
- ► Concurrent and flexible customer-initiated upgrades

For a description of the flexible customer-initiated upgrades see Chapter 9, "System upgrades" on page 263.

10.2.1 Scheduled outages

Concurrent hardware upgrades, concurrent parts replacement, concurrent driver upgrade, and concurrent firmware fixes available with the z196, all address elimination of scheduled outages. Furthermore, the following indicators and functions that address scheduled outages are included:

- ► Double memory data bus lane sparing

    This feature reduces the number of repair actions for memory

- ► Single memory clock sparing
- ► Double DRAM chipkill tolerance
- ► Field repair of the cache fabric bus

► Fast bitline delete on L3 and L4 caches

► Zero Address Detect PER to improve software error detection

  This feature helps detect wild branches following a zero value used for branch destination

► Power distribution N+2 design

  This feature is using Voltage Transformation Module (VTMs) in a highly redundant N+2 configuration

► Redundant humidity sensors

► Redundant altimeter sensors

► Unified support for the zBX

  The zBX will be supported like any other feature on the z196

► Dual in-line memory module (DIMM) field replaceable unit (FRU) indicators

  These indicators imply that a memory module is not error free and could fail sometime in the future. This gives IBM a warning, and the possibility and time to concurrently repair the storage module. To do this, first fence-off the book, then remove the book, replace the failing storage module, and then add the book. The flexible memory option might be necessary to maintain sufficient capacity while repairing the storage module.

► Single processor core checkstop and sparing

  This indicator implies that a processor core has malfunctioned and has been *spared*. IBM has to consider what to do and also take into account the history of the server by asking the question: Has this type of incident happened previously on this server?

► Point-to-point fabric for the SMP

  Having fewer components that can fail is an advantage. In a two-book or three-book machine the need to complete a ring has been removed. In addition, a book can always be added concurrently.

► Hot swap InfiniBand (IFB) hub cards

  When properly configured for redundancy, hot swapping (replacing) the IFB (HCA2-O) hub cards is possible, thereby avoiding any kind of interruption when the need for replacing these types of cards occurs.

► Redundant 100 Mbps Ethernet service network with VLAN

  The service network in the machine gives the machine code the capability to monitor each single internal function in the machine. This helps to identify problems, maintain the redundancy, and provides assistance in concurrently replacing a part. Through the implementation of the VLAN to the redundant internal Ethernet service network, these advantages are improving even more, as it makes the service network itself easier to handle and more flexible.

► I/O drawer

  The smaller-form factor I/O drawer originally introduced on the System z10 Business Class server is available on the z196. This drawer can be installed concurrently and I/O cards can be added to the drawer concurrently.

10.2.2 Unscheduled outages

An unscheduled outage occurs because of an unrecoverable malfunction in a hardware component of the server.

The following improvements can minimize unscheduled outages:

► Continued focus on firmware quality

For Licensed Internal Code and hardware design, failures are eliminated through rigorous design rules, design walk-through, peer reviews, element, subsystem and system simulation, and extensive engineering and manufacturing testing.

► Memory subsystem improvements

z196 introduces the Redundant Array of Independent Memory (RAIM) on System z servers, which in the disk industry is known as Redundant Array of Independent Drives (RAID). RAIM design detects and recovers from DRAM, socket, memory channel or DIMM failures.The RAIM design requires the addition of one memory channel that is dedicated for RAS. The parity of the four "data" DIMMs are stored in the DIMMs attached to a fifth memory channel. Any failure in a memory component can be detected and corrected dynamically. This design takes the RAS of the memory subsystem to another level, making it essentially a fully fault tolerant "N+1" design. The memory system on the z196 is implemented with an enhanced version of the Reed-Solomon ECC code known as 90B/64B, as well as protection against memory channel and DIMM failures. A very precise marking of faulty chips help assure timely DRAM replacements. The key cache on the z196 memory is completely mirrored. For a full description of the memory system on the z196, see 2.5, "Memory" on page 40

► Improved thermal- and condensation management

► Soft-switch firmware

The capabilities of soft-switching firmware have been enhanced. Enhanced logic in this function ensures that every affected circuit is powered off during soft-switching of firmware components. For example, if you must upgrade the microcode of a FICON feature, enhancements have been implemented to avoid any unwanted side-effects detected on previous servers.

10.3 z196 Enhanced book availability

Enhanced Book Availability (EBA) is a *procedure* under which a book in a multi-book machine can be removed and re-installed during an upgrade- or repair action with no impact on the executing workload

With the EBA procedure, and with proper planning to ensure that all the resources are still available to run critical applications in a (n-1) book configuration, you might be able to avoid planned outages. Consider, also, the selection of the flexible memory option to provide additional memory resources when replacing a book.

To minimize affecting current workloads, ensure that there are sufficient inactive physical resources on the remaining books to complete a book removal. Also consider non-critical system images, such as test or development logical partitions. After these non-critical logical partitions have been stopped and their resources freed, you might find sufficient inactive resources to contain critical workloads while completing a book replacement.

10.3.1 EBA planning considerations

To take advantage of the enhanced book availability function, configure enough physical memory and engines so that the loss of a single book does not result in any degradation to critical workloads during the following occurrences:

► A degraded restart in the rare event of a book failure
► A book replacement for repair or physical memory upgrade

We recommend the following configurations that enable exploitation of the enhanced book availability function. The PU and models suggested have enough unused capacity so that 100% of the customer-owned PUs can be activated even when one book within a model is fenced.

► A maximum of 15 customer PUs are configured on the M32

► A maximum of 32 customer PUs are configured on the M49

► A maximum of 49 customer PUs are configured on the M66

► A maximum of 60 customer PUs are configured on the M80

► No special feature codes are required for PU and model configuration

► For the four book models, the number of SAPs in a book is not the same for all books. This is a planning consideration. For the exact distribution of SAPs and spares over the four books, see Table 2-3 on page 38.

► Flexible memory option, which delivers physical memory so that 100% of the purchased memory increment can be activated even when one book is fenced.

The main point here is that the server configuration should have sufficient dormant resources on the remaining books in the system for the *evacuation* of the book to be replaced or upgraded. Dormant resources can be:

► Unused PUs or memory are not enabled by LICCC

► Inactive resources that are enabled by LICCC (memory that is not being used by any activated logical partitions)

► Memory purchased with the flexible memory option

► Additional books, as discussed previously

The I/O connectivity must also support book removal. The majority of the path to the I/O has redundant I/O interconnect support in the I/O cages and that enable connection through multiple IFBs.

If sufficient resources are not present on the remaining books, certain non-critical logical partitions might have to be deactivated, and one or more CPs, specialty engines, or storage might have to be configured offline to reach the required level of available resources. Planning that addresses these possibilities can help to reduce operational errors.

> **Note:** Single-book systems cannot make use of the EBA function.

The planning should be included as part of the initial installation and any follow-on upgrade that modifies the operating environment. The eConfig report can be used to determine the number of books, active PUs, memory configuration, and the channel layout.

If the z196 is installed, you may click the **Prepare for Enhanced Book Availability** option in the Perform Model Conversion panel of the EBA process on the HMC. This task helps you

determine the resources required to support the removal of a book with acceptable degradation to the operating system images.

The EBA process determines which resources, including memory, PUs, and I/O paths will have to be freed to allow for the removal of a given book. You may run this preparation on each book to determine which resource changes are necessary; use the results as input to the planning stage to help identify critical resources.

With this planning information, you can examine the logical partition configuration and workload priorities to determine how resources might be reduced and allow for the book to be removed.

Planning should include the following tasks:

▶ Review of the z196 configuration to determine:

- Number of books installed and the number of PUs enabled. Note the following information:

 • Use the eConfig output or the HMC to determine the model, number and types of PUs (CPs, IFL, ICF, zAAP, and zIIP).

 • Determine the amount of memory, both physically installed and LICCC-enabled.

 • Work with your IBM service personnel to determine the memory card size in each book. The memory card sizes and the number of cards installed for each installed book can be viewed from the SE under the CPC configuration task list, using the view hardware configuration option.

- Channel layouts, IFB to channel connections.

 Use eConfig output to review channel configuration including the IFB path. This is a normal part of the I/O connectivity planning. The alternate paths should be separated as far into the system as possible.

▶ Review the system image configurations to determine the resources for each.

▶ Determine the importance and relative priority of each logical partition.

▶ Identify the logical partition or workloads and the actions to be taken:

- Deactivate the entire logical partition.

- Configure PUs.

- Reconfigure memory, which might require the use of Reconfigurable Storage Units (RSU) value.

- Vary off of the channels.

▶ Review the channel layout and determine whether any changes are necessary to address single paths.

▶ Develop the plan to address the requirements.

When performing the review, document the resources that could be made available if the EBA were to be used. The resources on the books are allocated during a power-on reset (POR) of the system and can change during a POR. Perform a review when changes are made to the z196, such as adding books, CPs, memory, or channels, or when workloads are added or removed. The Prepare for Enhanced Book Availability function can be used ahead of any EBA action to determine whether the system can be conditioned to allow for book removal or what actions are required to support the removal.

10.3.2 Enhanced book availability processing

To use the EBA, certain conditions must be satisfied:

▶ Free the used processors (PUs) on the book that will be removed.

▶ Free the used memory on the book.

▶ For all I/O domains connected to this book, ensure that alternate paths exist, otherwise place the I/O paths offline.

For the EBA process, this is the preparation phase, and is started from the SE, either directly or on the HMC by using the single object operation option in the Perform Model Conversion panel from the CPC configuration task list. See Figure 10-1 on page 319.

Processor availability

Processor resource availability for reallocation or deactivation is affected by the type and quantity of resources in use, as follows:

▶ Total number of PUs that are enabled through LICCC
▶ PU definitions in the profiles and that can be
  – Dedicated and dedicated reserved
  – Shared
▶ Active logical partitions with dedicated resources at the time of book repair or replacement

To maximize the PU availability option, ensure that there are sufficient inactive physical resources on the remaining books to complete a book removal.

Memory availability

Memory resource availability for reallocation or deactivation depends on:

▶ Physically installed storage
▶ Image profile storage allocations
▶ Amount of memory enabled through LICCC
▶ Flexible memory option

See 2.7.2, "Enhanced book availability" on page 51.

HCA-2 to IFB-MP connectivity requirements

The optimum approach is to maintain maximum I/O connectivity during book removal. The redundant I/O interconnect (RII) function provides for redundant IFB connectivity to all installed I/O domains in all I/O cages.

Preparing for enhanced book availability

The Prepare Concurrent Book replacement option validates that enough dormant resources exist for this operation. If enough resources are not available on the remaining books to complete the EBA process, the process identifies the resources and guides you through a series of steps to select and free up resources. The preparation process does not complete until all memory and I/O conditions have been successfully resolved.

> **Note:** The preparation step does not reallocate any resources. It is only used to record customer choices and to produce a configuration file on the SE that will be used by the perform concurrent book replacement operation.

The preparation step can be done in advance. However, if any changes to the configuration occur between the time the preparation is done and when the book is physically removed, you must rerun the preparation phase.

The process can be run multiple times, because it does not move any resources. To view results of the last preparation operation, select **Display Previous Prepare Enhanced Book Availability Results** from the Perform Model Conversion panel (in SE).

The preparation step can be run several times without actually performing a book replacement. It enables you to dynamically adjust the operational configuration for book repair or replacement prior to Customer Engineer (CE) activity. Figure 10-1 shows the Perform Model Conversion panel where you select the **Prepare for Enhanced Book Availability** option.
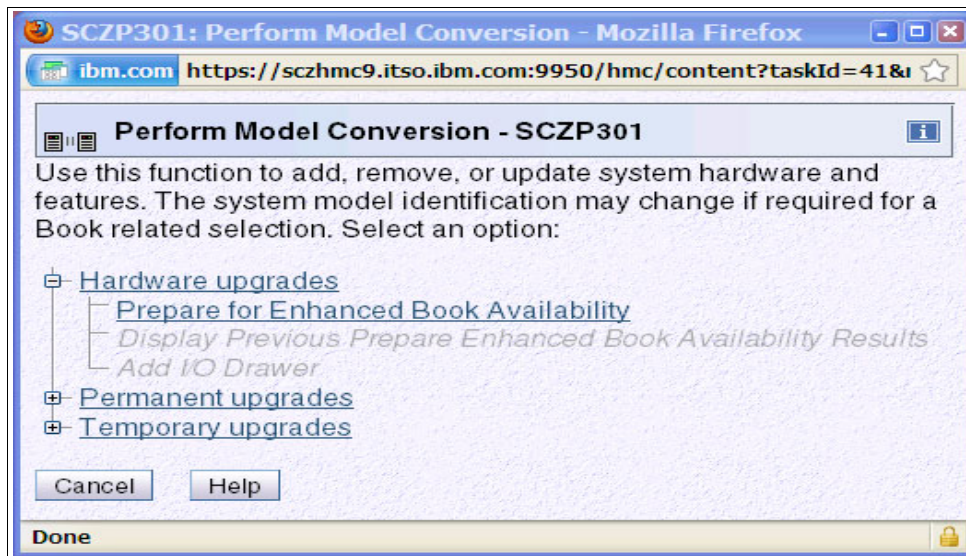
Figure 10-1   Perform Model Conversion; select Prepare for Enhanced Book Availability

After you select **Prepare for Enhanced Book Availability**, the Enhanced Book Availability panel opens. Select the book that is to be repaired or upgraded and click **OK**. See Figure 10-2. Only one target book can be selected each time.
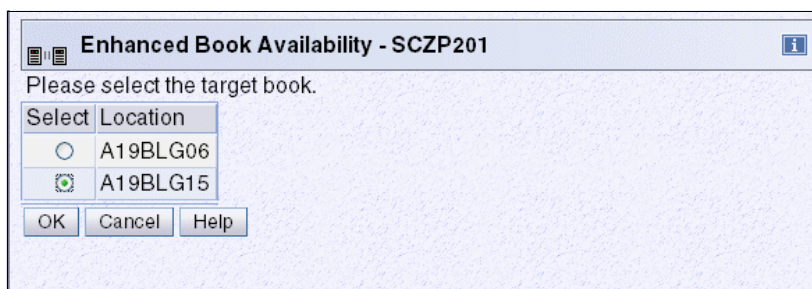
Figure 10-2   Enhanced Book Availability, selecting the target book

The system verifies the resources required for the removal, determines the required actions, and presents the results for review. Depending on the configuration, the task can take approximately 30 minutes.

The preparation step determines the readiness of the system for the removal of the targeted book. The configured processors and the memory that is in are evaluated against unused resources available across the remaining books.

If not enough resources are available, the system identifies the conflicts so you can take action to free other resources. The system analyzes I/O connections associated with the removal of the targeted book for any single path I/O connectivity.

Three states can result from the preparation step:

► The system is ready to perform the enhanced book availability for the targeted book with the original configuration.

► The system is not ready to perform the enhanced book availability because of conditions indicated from the preparation step.

► The system is ready to perform the enhanced book availability for the targeted book. However, to continue with the process, processors are reassigned from the original configuration. Review the results of this reassignment relative to your operation and business requirements. The reassignments can be changed on the final window that is presented. However, before making changes or approving reassignments, ensure that the changes have been reviewed and approved by the correct level of support, based on your organization's business requirements.

Preparation tabs

The results of the preparation are presented for review in a tabbed format. Each tab indicates conditions that prevent the EBA option from being performed. Tabs are for processors, memory, and various single path I/O conditions. See Figure 10-3 on page 321.Possible tab selections are:

► Processors
► Memory
► Single I/O
► Single Domain I/O
► Single Alternate Path I/O

Only the tabs that have conditions that prevent the book from being removed are displayed. Each tab indicates what the specific conditions are and possible options to correct the conditions.

Example panels from the preparation phase

The figures in this section are examples of panels that are displayed, during the preparation phase, when a condition requires further actions to prepare the system for the book removal.

Figure 10-3 shows the results of the preparation phase for removing book 0. The tabs labeled Memory and Single I/O indicate the conditions that were found in preparing the book to be removed. In the figure, the Memory tab indicates the amount of memory in use, the amount of memory available, and the amount of memory that must be made available. The amount of in-use memory is indicated in megabytes for each partition name. After the required amount of memory has been made available, rerun the preparation to verify the changes.
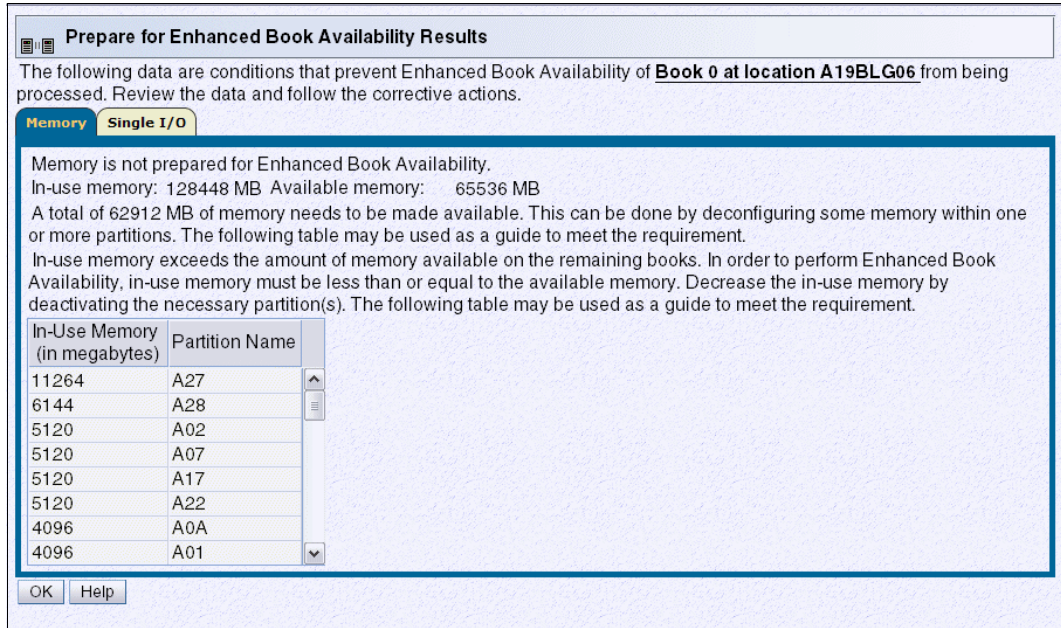
Figure 10-3   Prepare for EBA: Memory conditions

Figure 10-4 shows the Single I/O tab. The preparation has identified single I/O paths associated with the removal of book 0. The paths have to be placed offline to perform the book removal. After addressing the condition, rerun the preparation step to ensure that all the required conditions have been met.
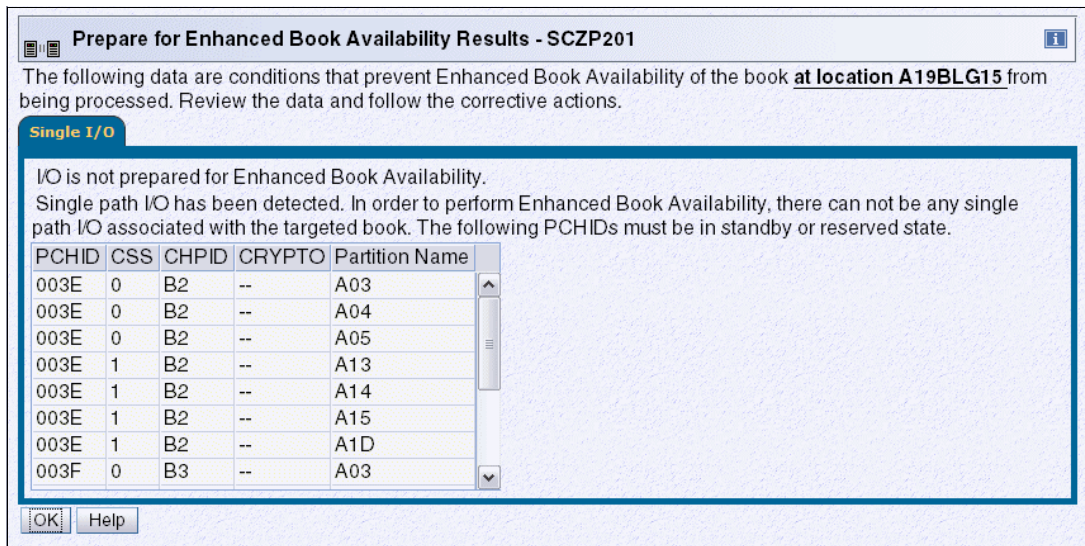
Figure 10-4   Prepare for EBA: Single I/O conditions

Preparing the server to perform enhanced book availability

During the preparation, the system determines the CP configuration that is required to remove the book. Figure 10-5 shows the results and provides the option to change the assignment on non-dedicated processors.

Reassign Non-Dedicated Processors

Accept or reassign non-dedicated processors without exceeding the LICCC count.

Warning: These values should only be reassigned under the direction of the System Programmer. Otherwise accept the

Processor Type	Dedicated Count	Non-Dedicated Count	Processor Totals	LICCC Count
CPU	0	7	7	12
ICF	1	0	1	4
IFL	0	0	0	0
IFA	0	1	1	2
SAP	3	0	3	4
Available to use		0	0	
Remaining Book Totals	4	8	12	

OK   Cancel   Help

Figure 10-5   Reassign Non-dedicated Processors results

Important: Consider the results of these changes relative to the operational environment. Understand the potential impact of making such operational changes. Changes to the PU assignment, although technically correct, can result in constraints for critical system images. In some cases, the solution might be to defer the reassignments to another time that would have less impact on the production system images.

After reviewing the reassignment results, and making adjustments if necessary, click **OK**.

The final results of the reassignment, which include changes made as a result of the review, are displayed. See Figure 10-6. These results will be the assignments when the book removal phase of the EBA is completed.

Reassign Non-Dedicated Processors

The following processor allocation will be made if OK is selected.
Select CANCEL if you wish to not make changes or abort the allocation.

Number of CPUs = 7
Number of ICFs = 0
Number of IFLs = 0
Number of IFAs = 1
Number of zIIPs = 0
PUs not assigned = {5}

ACT37294

OK   Cancel

Figure 10-6   Reassign Non-Dedicated Processors, message ACT37294

Summary of the book removal process steps

This section steps through the process of a concurrent book replacement.

To remove a book, the following resources must be moved to the remaining active books:

► PUs: Enough PUs must be available on the remaining active books, including all types of characterizable PUs (CPs, IFLs, ICFs, zAAPs, zIIPs, and SAPs).

► Memory: Enough installed memory must be available on the remaining active books.

► I/O connectivity: Alternate paths to other books must be available on the remaining active books or the I/O path must be taken offline.

By understanding both the server configuration and the LPAR allocation for memory, PUs, and I/O, you should be able to make the best decision about how to free necessary resources and allow for book removal.

To concurrently replace a book:

1. Run the preparation task to determine the necessary resources.

2. Review the results.

3. Determine the actions to perform in order to meet the required conditions for EBA.

4. When you are ready for the book removal, free the resources that are indicated in the preparation steps.

5. Rerun the step in Figure 10-1 on page 319 (Prepare for Enhanced Book Availability task) to ensure that the required conditions are all satisfied.

6. Upon successful completion (Figure 10-7), the system is ready for the removal of the book.
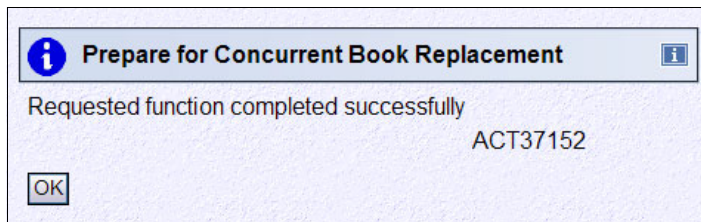
Figure 10-7   Preparation completed successfully, message ACT37152

The preparation process can be run multiple times to ensure that all conditions have been met. It does not reallocate any resources. All it does is to produce a report. The resources are not reallocated until the Perform Book Removal process is invoked.

Rules during EBA

Processor, memory, and single I/O rules during EBA are as follows:

► Processor rules

All processors in any remaining books are available to be used during EBA. This includes the two spare PUs or any available PU that is non-LICCC.

The EBA process also allows conversion of one PU type to another PU type. One example is converting a zAAP to a CP for the duration of the EBA function. The preparation for concurrent book replacement task indicates whether any SAPs have to be moved to the remaining books.

► Memory rules

All physical memory installed in the system, including flexible memory, is available for the duration of the EBA function. Any physical installed memory, whether purchased or not, is available to be used by the EBA function.

► Single I/O rules

Alternate paths to other books must be available or the I/O path must be taken offline.

Review the results. The result of the preparation task is a list of resources that need to be made available before the book replacement can take place.

Free any resources

At this stage, create a plan to free up these resources. The resources and actions to free them are in the following list:

► To free any PUs:

– Vary the CPs offline, reducing the number of CP in the shared CP pool.
– Use any spare CP.
– Deactivate logical partitions.
– Convert the PU. For example, convert a zAAP to a CP.

► To free memory:

– Deactivate a logical partition.

– Vary offline a portion of the reserved (online) memory. For example, in z/OS issue the command:

```
CONFIG_STOR(E=1),<OFFLINE/ONLINE>
```

This command enables a storage element to be taken offline. Note that the size of the storage element depends on the RSU value. In z/OS, the following command enables you to configure offline smaller amounts of storage than what has been set for the storage element:

```
CONFIG_STOR(nnM),<OFFLINE/ONLINE>
```

– A combination of both logical partition deactivation and varying memory offline.

> **Note:** If you plan to use the EBA function with z/OS logical partitions, we recommend setting up reserved storage and setting an RSU value. Use the RSU value to specify the number of storage units that are to be kept free of long-term fixed storage allocations, allowing for storage elements to be varied offline.

10.4 z196 Enhanced driver maintenance

Enhanced Driver Maintenance (EDM) is another step in reducing both the necessity and the eventual duration of a scheduled outage. One of the contributors to planned outages is Licensed Internal Code (LIC) Driver updates that are performed in support of new features and functions.

When properly configured, the z196 supports concurrently activating a selected new LIC Driver level. Concurrent activation of the selected new LIC Driver level is supported only at specifically released sync points. Concurrently activating a selected new LIC Driver level anywhere in the maintenance stream is not possible. There are certain LIC updates where a concurrent update/upgrade may not be possible.

The key points of EDM are:

► The HMC is capable of querying whether a system is ready for a concurrent driver upgrade.

► Previous firmware updates, which require an initial machine load (IML) of z196 to be activated, may block the ability to perform a concurrent driver upgrade.

► An icon on the Support Element (SE) allows you or your IBM support personnel to define the concurrent driver upgrade sync point to be used for an EDM.

► The ability to concurrently install and activate a new Driver can eliminate or reduce a planned outage.

► Concurrent crossover from Driver level N to Driver level N+1, to Driver level N+2 must be done serially; no composite moves are allowed.

► Disruptive upgrades are permitted at any time and allow for a composite upgrade (Driver N to Driver N+2).

► Concurrent back-off to the previous Driver level is not possible. The driver level must move forward to driver level N+1 after EDM is initiated. Catastrophic errors during an update could result in a scheduled outage to recover.

The EDM function does not completely eliminate the need for planned outages for driver-level upgrades. Upgrades may require a system level or a functional element scheduled outage to activate the new LIC. The following circumstances require a scheduled outage:

► Specific complex code changes might dictate a disruptive driver upgrade. You are alerted in advance so you can plan for the following changes:
  – Design data or hardware initialization data fixes
  – CFCC release level change

► Non-QDIO OSA CHPID types, CHPID type OSC, and CHPID type OSE require CHPID Vary OFF/ON in order to activate new code.

► Cryptographic code on a Crypto Express2 and/or Crypto Express3 feature requires a Config OFF/ON in order to activate the new code.

► FICON and FCP code changes requiring a CHPID Config OFF/ON to activate the new code.

10.5 RAS capability for the HMC

The primary HMC for the zEnterprise is where portions of the Unified Resource Manager routines execute. The Unified Resource Manager is an active part of the zEnterprise System infrastructure. The HMC is therefore in a stateful environment that needs high availability features to assure survival of the system in case of failure. So each zEnterprise must be equipped with two HMC workstations—a primary and an alternate. The contents and activities of the primary are synchronously mirrored on the alternate HMC, so that it can automatically take over the activities of the primary if it were to fail. While the primary HMC can do all HMC activities (including Unified Resource Manager activities), the alternate can only be the backup and cannot be used for tasks or activities.

> **Note:** The primary HMC and its alternate must be connected to the same subnetwork to allow the alternate HMC to take over the IP address of the primary HMC during failover processing.

For a full description of the HMC and its capabilities, see 13.3, "Ensemble Physical Resource Management" on page 372

10.6 RAS capability for zBX

The zBX has been built with the traditional System z quality of service to include RAS capabilities. The zBX offering provides extended service capability with the z196 hardware management structure. The HMC/SE functions of the z196 server provide management and control functions for the zBX solution.

Apart from a zBX configuration with one chassis installed, the zBX is configured to provide N + 1 components. All the components are designed to be replaced concurrently. In addition zBX configuration upgrades can be performed concurrently.

The zBX has two Top of Rack Switches (TORs). These switches provide N + 1 connectivity for the private networks between the z196 server and the zBX for monitoring, controlling, and managing the zBX components.

Each BladeCenter has the following:

- ► Up to 14 Blades. Blades can be removed, repaired, and replaced concurrently.
- ► N + 1 PDUs. Provided the PDUs have power inputs from two different sources. In case of a single source failure, the second PDU will take over the total load of its BladeCenter.
- ► N + 1 hot-swap power module with fan. A pair of power modules provide power for seven blades. A fully configured BladeCenter with 14 Blades has a total of four power modules
- ► N + 1 1 GbE switch modules for the PSCN
- ► N + 1 10 GbE High Speed switches for the IEDN
- ► N + 1 1000BaseT switches for the INMN
- ► N + 1 8 Gb Fiber Channel switches for the external disk
- ► Two hot-swap Advanced Management Modules
- ► Two hot-swap fans/blowers

Note: Some BladeCenter configurations do not physically fill up the rack with their components, but they have reached other maximums, such as power usage.

zBX Firmware

The testing, delivery, installation, and management of the zBX firmware is handled exactly the same way as for the z196 server. The same z196 server process and controls are used. Any fixes to the zBX machine are downloaded on to the owning z196 server's SE and are applied to the zBX.

The MCLs for the zBX are designed to be concurrent and their status can be viewed at the z196 server's HMC.

11

Environmental requirements

> *"You can't make a product greener, whether it's a car, a refrigerator, or a city without making it smarter: smarter materials, smarter software, or smarter design."*
>
> .... Thomas Friedman, New York Times

This chapter briefly describes the environmental requirements for the zEnterprise System. We list the dimensions, weights, power, and cooling requirements as an overview of what is needed to plan for the installation of a zEnterprise 196 and zEnterprise BladeCenter Extension.

There are a number of options for the physical installation of the server including air or water cooling, I/O cabling under the raised floor or off the top of the server frames, and the possibility of having high-voltage DC power supply directly into the server in stead of the usual AC power supply.

For comprehensive physical planning information see *System z10 Enterprise Class Installation Manual for Physical Planning,* GC28-6865.

This chapter discusses the following topics:

11.1 z196 power and cooling

The z196 is always a two-frame system. The frames are shipped separately and are bolted together during the installation procedure. Installation is always on a raised floor with power and optional water hoses arriving to the server from underneath the raised floor. I/O cables also exit from the bottom of the server frames unless the Top exit cabling feature code (FC 7942) is installed, so that I/O cables can exit directly from the top corners of the server in to overhead cabling rails.

11.1.1 Power consumption

The system operates with two completely redundant power supplies. Each of the power supplies have their individual line-cords or pair of line-cords depending on the configuration.

For redundancy, the server should have two power feeds. Each power feed is either 1 or 2 line cords. The number of line cords required depends on system configuration. Line cords attach to either 3 phase, 50/60 Hz, 200 to 480 V AC power or 380 to 520 V DC power. There is no impact to system operation with the total loss of one power feed.

For ancillary equipment such as the Hardware Management Console, its display, and its modem, additional single-phase outlets are required.

The power requirements depend on the cooling facility installed and on number of books, as well as the number of I/O units installed. I/O units are values for I/O cages (equals 2 I/O units) and I/O drawers (equals 1 I/O unit).

Input power in kVA is equal to the output power in kW. Heat output expressed in kBTU per hour can be derived from multiplying the table entries by a factor of 3.4.

Table 11-1 lists the absolute maximum power requirements for the air cooled models (typical systems will draw significantly less power than this).

Table 11-1 Power requirements - air cooled models

Power requirement kVA	Number of I/O units						
	0	1	2	3	4	5	6
M15	6.8	7.7	8.7	10.8	12.8	12.9	13.1
M32	11.9	13.2	14.1	16.1	18.0	18.9	20.7
M49	17.3	18.2	19.2	21.1	23.0	23.8	25.8
M66 / M80	22.7	23.6	24.6	26.4	28.4	29.2	30.1
Notes: FC 4000 = 1 I/O unit, FC 4002 = 2 I/O units. For an I/O cage in the A Frame, FC 4016 (M15) or FC 4020 (M32, M49, M66, M80)							

Table 11-2 on page 329 lists the absolute maximum power requirements water cooled models (typical systems will draw significantly less power than this).

Table 11-2   Power requirements - water cooled models

Power requirement kVA	Number of I/O units						
	0	1	2	3	4	5	6
M15	6.1	6.8	7.8	9.7	11.6	11.8	12.0
M32	9.8	10.6	11.6	13.4	15.3	16.2	18.2
M49	13.7	14.4	15.5	17.3	19.2	20.1	22.0
M66 / M80	18.1	18.9	19.8	21.7	23.6	24.5	26.4
Notes: FC 4004 = 1 I/O unit, FC 4005 = 2 I/O units For an I/O cage in the A Frame, FC 4018 (M15, M32, M49, M66, M80)							

Table 11-3 shows the line-cord pair and Bulk Power Regulator (BPR) requirements for books and I/O units. FC 3004 (air-cooled) or FC 3006 (water-cooled) will provide an additional BPR pair.

If your initial configuration needs one line-cord pair, but for growth would need a second pair, you can order the line-cord plan ahead feature (FC 2000), which will install two line-cord pairs at the initial configuration. Also, if Balanced Power Plan Ahead (FC 3003) is ordered, two line-cord pairs are shipped.

Table 11-3   Line-cord requirements

Line-cord requirements - number of pairs/number of BPRs per side	Number of I/O units						
	0	1	2	3	4	5	6
M15	1/1	1/1	1/1	1/2	1/3	1/3	1/3
M32	1/2	1/3	1/3	1/3	1/3	2/4	2/4
M49	1/3	1/3	2/4	2/4	2/4	2/4	2/5
M66 / M80	2/4	2/4	2/5	2/5	2/5	2/5	2/5

Systems that specify two line-cords can be brought up with one line-cord and will continue to run without power redundancy. The larger machines that specify four line-cords can be brought up with two line-cords and will continue to run without power redundancy. Four line-cords offer power redundancy, so that when a line-cord fails, the remaining cords deliver sufficient power to keep the system up and running.

11.1.2 Internal Battery Feature

The optional Internal Battery Feature (IBF) provides sustained system operations for a relatively short period of time, allowing for orderly shutdown. In addition, an external uninterruptible power supply system can be connected, allowing for longer periods of sustained operation.

If the batteries are not older than three years and have been discharged regularly, the IBF is capable of providing emergency power for the periods of time listed in Table 11-4 on page 330.

Table 11-4   Battery hold-up times

Internal battery hold-up times in minutes	Number of I/O units						
	0	**1**	**2**	**3**	**4**	**5**	**6**
M15	6	5	4	9	7	6.8	6.8
M32	7.5	7	6	9.5	7.5	7	6
M49	8	7.4	7	6	5	4.6	4.1
M66 / M80	5	4.8	4.5	4	3.6	3.2	3
Notes: The hold-up times in this table are estimates and are valid for batteries 3 years old or less that have seen normal service life (2 or less complete discharges per year) with system input power at N+1 operation.							

11.1.3 Emergency power-off

On the front of frame A is an emergency power-off switch that, when activated, immediately disconnects utility and battery power from the server. This causes all volatile data in the server to be lost.

If the server is connected to a machine-room's emergency power-off switch, and the Internal Battery Feature is installed, the batteries take over if the switch is engaged.

To avoid take-over, connect the machine-room emergency power-off switch to the server power-off switch. Then, when the machine-room emergency power-off switch is engaged, all power will be disconnected from the line-cords and the Internal Battery Features. However, all volatile data in the server will be lost.

11.1.4 Cooling requirements

The z196 can be air cooled or water cooled. The air cooled server requires chilled air to fulfill the air-cooling requirements, ideally coming from under the raised floor. The chilled air is usually provided through perforated floor tiles. The water cooled server requires building water to be supplied to the server. The requirements for both cooling options are indicated in *System z10 Enterprise Class Installation Manual for Physical Planning,* GC28-6865.

11.2 z196 physical specifications

This section describes weights and dimensions of the z196.

11.2.1 Weights and dimensions

Installation on a raised floor is mandatory. In the *System z10 Enterprise Class Installation Manual for Physical Planning,* GC28-6865, 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.

Table 11-5 indicates the maximum system dimension and weights for the M80 models. The weight ranges are based on configuration models with one I/O frame and three I/O cages, and with and without IBF.

Table 11-5   System dimensions and weights

Maximum	A and Z frames without Internal Battery Feature (3212) - Model M80	A and Z frames with Internal Battery Feature (3212) - Model M80
Air cooled servers		
Weight kg (lbs)	1894 (4175)	2177 (4799)
Width mm (in)	1534 (60.7)	1534 (60.7)
Depth mm (in)	1273 (50.1)	1273 (50.1)
Height mm (in)	2012 (79.2)	2012 (79.2)
Water cooled servers		
Weight kg (lbs)	1902 (4193)	2185 (4817)
Width mm (in)	1534 (60.7)	1534 (60.7)
Depth mm (in)	1375 (54.1)	1375 (54.1)
Height mm (in)	2012 (79.2)	2012 (79.2)

Notes:
1. Weight includes covers. Width, depth, and height are indicated without covers.
2. Weight is based on maximum system configuration, not the addition of the maximum weight of each frame.
3. Width increases to 1846 mm (72.7 in) if the top exit for I/O cables feature (FC 7942) is installed.
4. Weight increases by 43.1 kg (95 lbs) if the I/O top exit feature is installed.
5. If Feature code 7942 is specified the weight increases by and the width increases by

Weight distribution plate

The weight distribution plate is designed to distribute the weight of a frame onto two floor panels in a raised-floor installation. As Table 11-5 shows, the weight of a frame can be substantial, causing a concentrated load on a caster or leveling foot to be half of the total frame weight. In a multiple system installation, one floor panel can have two casters from two adjacent systems on it, potentially inducing a highly concentrated load on a single floor panel. The weight distribution plate distributes the weight over two floor panels. The weight distribution kit is ordered and delivered through feature code 9970.

Always consult the floor tile manufacturer to determine the load rating of the tile and pedestal structure. Additional panel support might be required to improve the structural integrity, because cable cutouts significantly reduce the floor tile rating.

3-in-1 bolt down kit for raised floor

A bolt-down kit for raised floor environments can be ordered for the z196 frames. The kit provides hardware to enhance the ruggedness of the frames and to tie down the frames to a concrete floor beneath a raised floor of 228-912 mm (9–36 inches). The kit is offered in the following configurations:

► The Bolt-Down Kit for an air cooled machine (FC 8008) provides frame stabilization and bolt-down hardware for securing a frame to a concrete floor beneath the raised floor.

> ► The Bolt-Down Kit for a water cooled machine (FC 8009) provides frame stabilization and bolt-down hardware for securing a frame to a concrete floor beneath the raised floor.

Each server would need two features one for each of the frames. The kits help secure the frames and their contents from damage when exposed to shocks and vibrations such as those generated by a seismic event. The frame tie-downs are intended for securing a frame weighing less than 1632 kg (3600 lbs) per frame.

11.3 Power estimation tool

Several aids are available to monitor the power consumption and heat dissipation of the z196. This section summarizes the tools that are available to estimate the energy consumption of the z196. The following tools are available:

> ► Power estimation tool
> ► System activity display
> ► IBM Systems Director Active Energy Manager™

Power estimation tool

The power estimation tool for z196 is available through the IBM Resource Link Web site:

http://www.ibm.com/servers/resourcelink

The tool provides an estimate of the anticipated power consumption of a particular machine model given its configuration. For the z196, you input the machine model, memory size, number of I/O cages, and quantity of each type of I/O feature card. The tool outputs an estimate of the power requirements for your configuration.

The tool helps with power and cooling planning for installed/planned z196 servers.

11.4 Energy management

The energy management strategy for the server is not to go after fine-grained power savings by dynamically and automatically vary the performance of the system. The objective is to deliver energy efficient features, energy management functions, and very advanced management solutions that clearly demonstrates value for System z.

In this section we will discuss the elements of energy management in areas of tooling to help understand the requirement for power and cooling, monitoring and trending, and reducing power consumption. The energy management structure for the server is shown in Figure 11-1 on page 333.
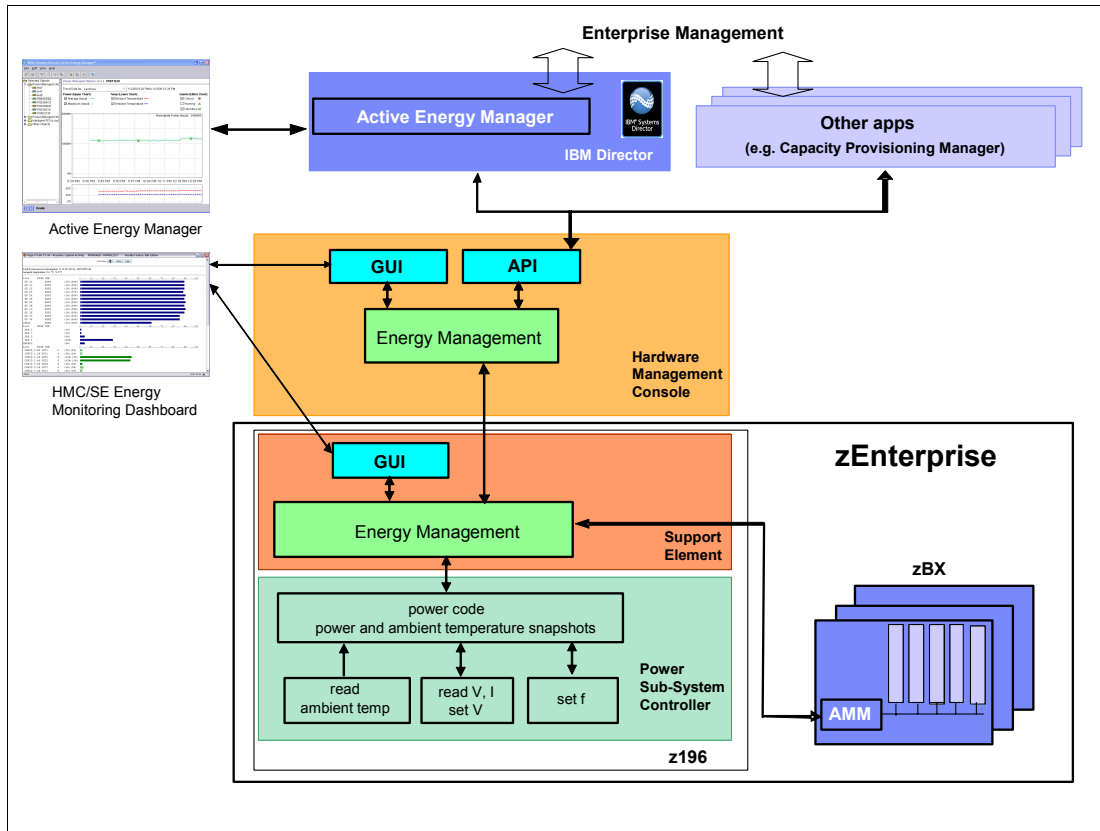
Figure 11-1   Energy management structure

11.4.1 Energy management tooling

Actual power consumption of the system can be seen on a system activity display (SAD) panel of HMC, shown in Figure 11-2 on page 334.
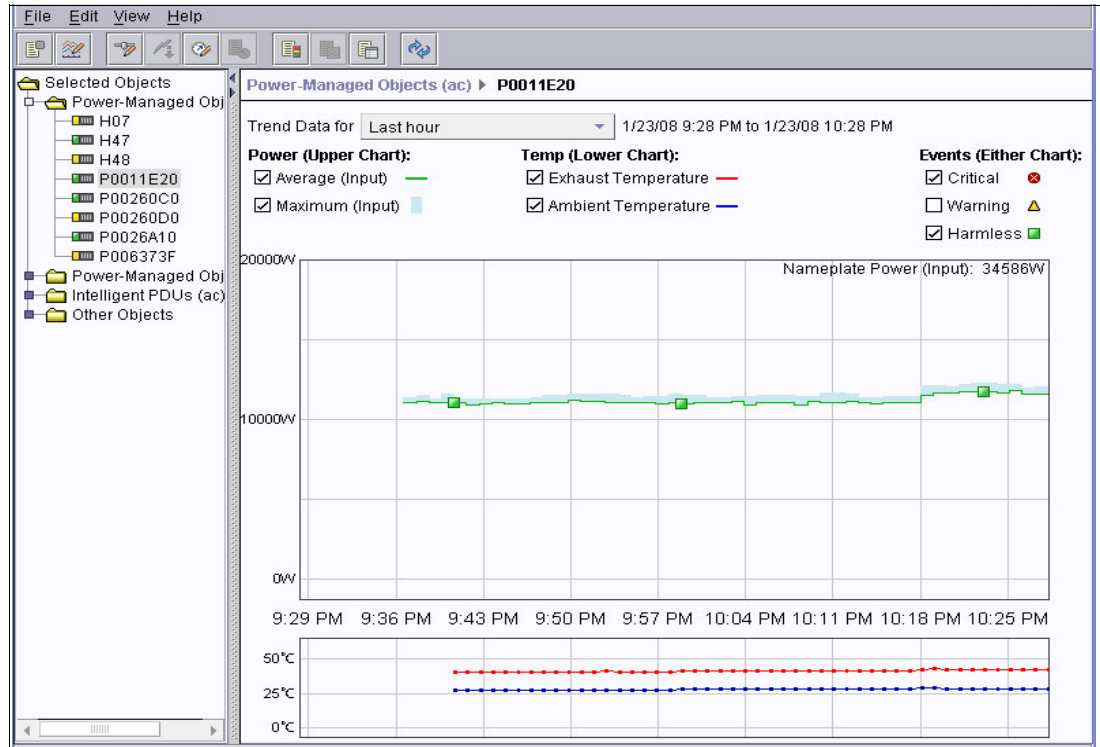
Figure 11-2   SAD frame power consumption example

IBM Systems Director Active Energy Manager

IBM Systems Director Active Energy Manager is an energy management solution building-block that returns true control of energy costs to the customer. Active Energy Manager is an industry-leading cornerstone of the IBM energy management framework and is part of the IBM Cool Blue™ portfolio.

Active Energy Manager Version 4.1.1 is a plug-in to IBM Systems Director Version 6.1 and is available for installation on Linux on System z. It can also run on Windows, Linux on IBM System x, and Linux on IBM Power Systems™. For more specific information see *Implementing IBM Systems Director Active Energy Manager 4.1.1*, SG24-7780.

Active Energy Manger is a management software tool that can provide a single view of the actual power usage across multiple platforms as opposed to the benchmarked or rated power consumption. It can effectively monitor and control power in the data center at the system, chassis, or rack level. By enabling these power management technologies, data center managers can more effectively power manage their systems while lowering the cost of computing.

The following power management functions are available with Active Energy Manager:

► Power trending

  Power trending allows you to monitor, in real time, the consumption of power by a supported power managed object. You use this data to track the actual power consumption of monitored devices and to determine the maximum value over time. The data can be presented either graphically or in tabular form.

► Thermal trending

  Thermal trending allows you to monitor, in real-time, the heat output and ambient temperature of a supported power managed object. Use this data to help avoid situations

where overheating might cause damage to computing assets, and to study how the thermal signature of various monitored devices varies with power consumption. The data can be presented either graphically or in tabular form.

The following data is available from System z HMC:

► System name, machine type, model, serial number, firmware level

► Ambient temperature

► Exhaust temperature

► Average power usage over a one minute period

► Peak power usage over a one minute period

► Limited status and configuration information. This information helps explain changes to the power consumption, called Events, which can be:

  – Changes in fan speed
  – MRU failures
  – Changes between power-off, power-on, and IML-complete states
  – Number of books and I/O cages
  – CBU records expiration(s)

Figure 11-3 on page 336 shows a sample chart of the data that is available from Active Energy Manager and System z196.

IBM Systems Director Active Energy Manager is the first solution on the market that provides customers with the intelligence necessary to effectively manage power consumption in the data center. Active Energy Manager, which is an extension to IBM Director systems management software, enables you to *meter* actual power usage and trend data for any single physical system or group of systems. Active Energy Manager uses monitoring circuitry, developed by IBM, to help identify how much actual power is being used and the temperature of the system.
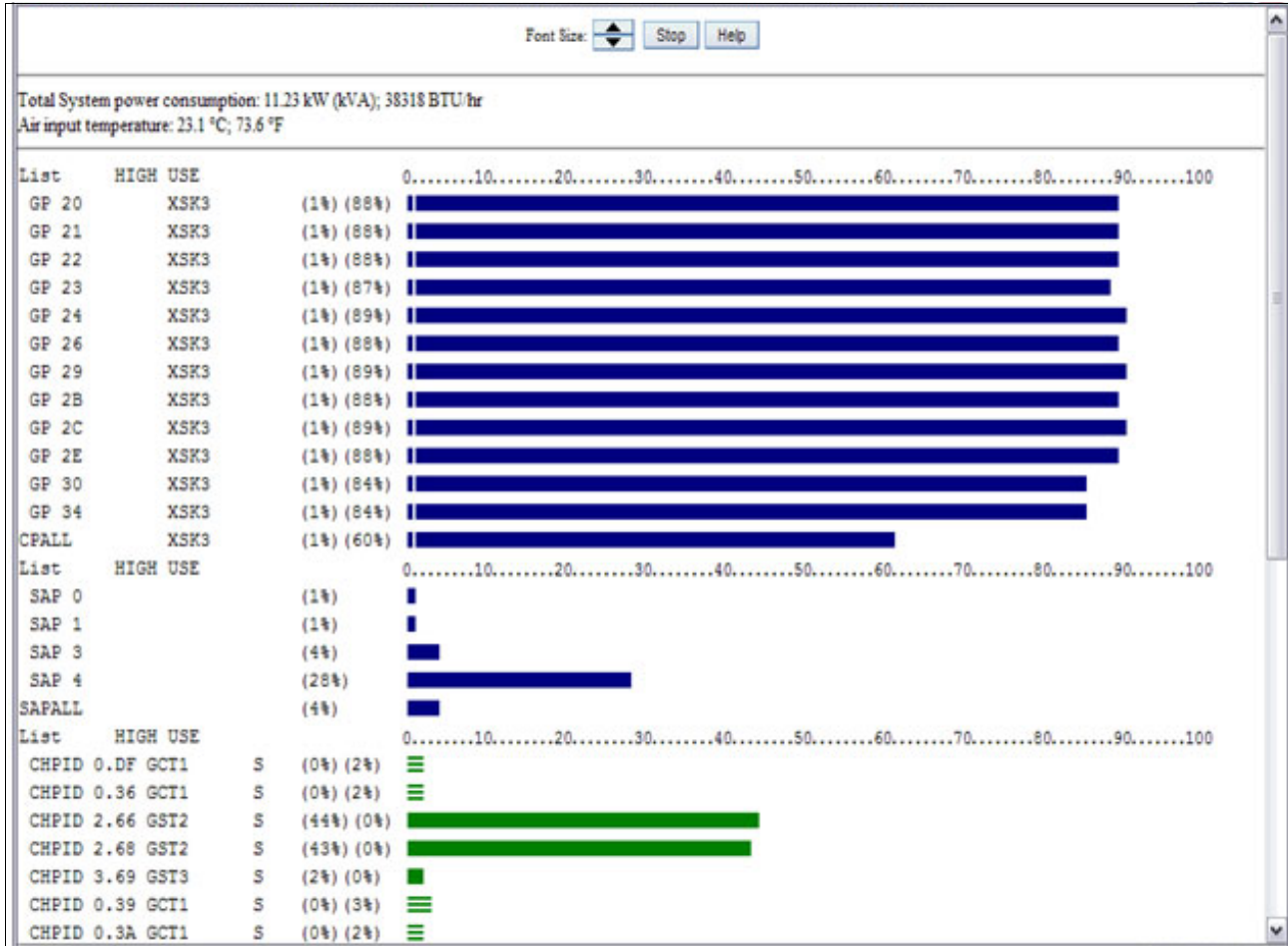
Figure 11-3   Active Energy Manager - dashboard example

11.4.2 Static power saving mode

The server has a mechanism to vary frequency and voltage, originally developed to help avoid interruptions due to cooling failures. The mechanism can be used to reduce the energy consumption of the system in periods of low utilization and to partially power off systems designed mainly for disaster recovery - CBU-systems. The mechanism is under full customer control, there is no autonomous function to perform changes under the covers. The customer controls are implemented in the HMC, in the SE, and in the Active Energy Manager with one power saving mode. The expectation is that the frequency reduction is 17%,the voltage reduction 9%, and the total power savings is from 10 to 20% depending on the configuration.

Figure 11-4 on page 337 shows the Set Power Saving panel.
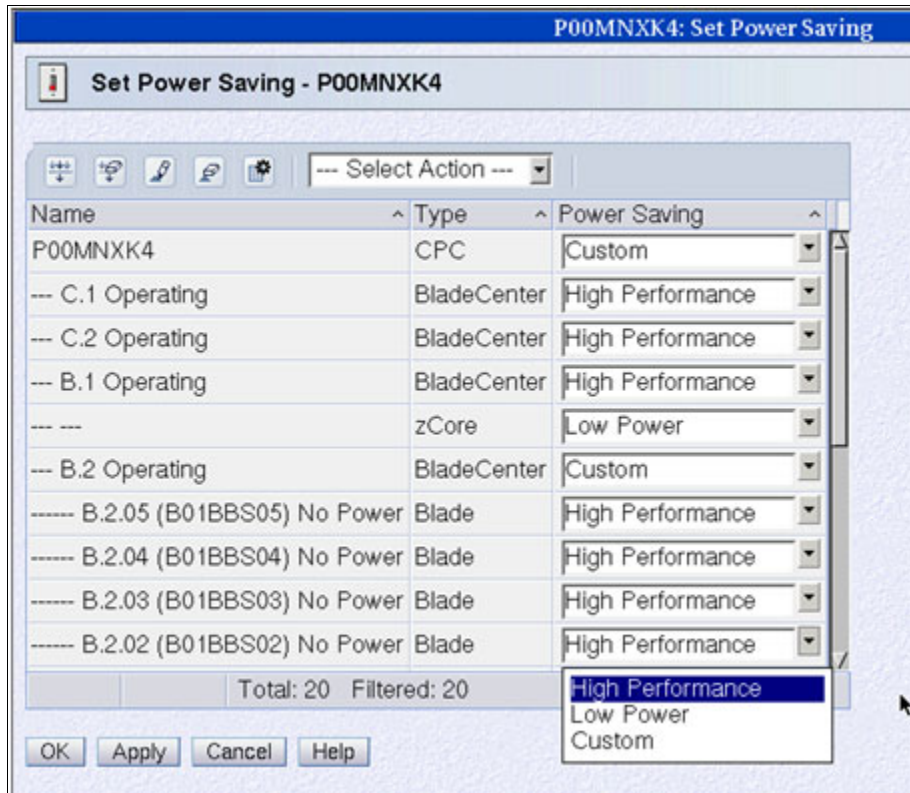
Figure 11-4   Set Power Saving panel

11.4.3 Query maximum potential power

This function is implemented in the SE using an algorithm to calculate the maximum potential power draw of the system based on the configuration, the altitude of the computer room, the room temperature, and the highest single fault service scenario power condition for the configuration applying reasonable tolerances. The functions looks like power capping to higher level management tools and allows for reducing the power allocation to the system based on the knowledge of the maximum possible draw. The functions facilitates operations personnel with no System z knowledge to query the maximum possible power draw of the system. The implementation helps avoid capping enforcement through dynamic capacity reduction. The customer controls are implemented in the HMC, the SE, and in the Active Energy Manager. We recommended that this function be used in conjunction with the Power Estimation tool that allows for pre-planning for power and cooling requirements. See "Power estimation tool" on page 332

An example of the Set Power Cap panel is shown in Figure 11-5 on page 338.
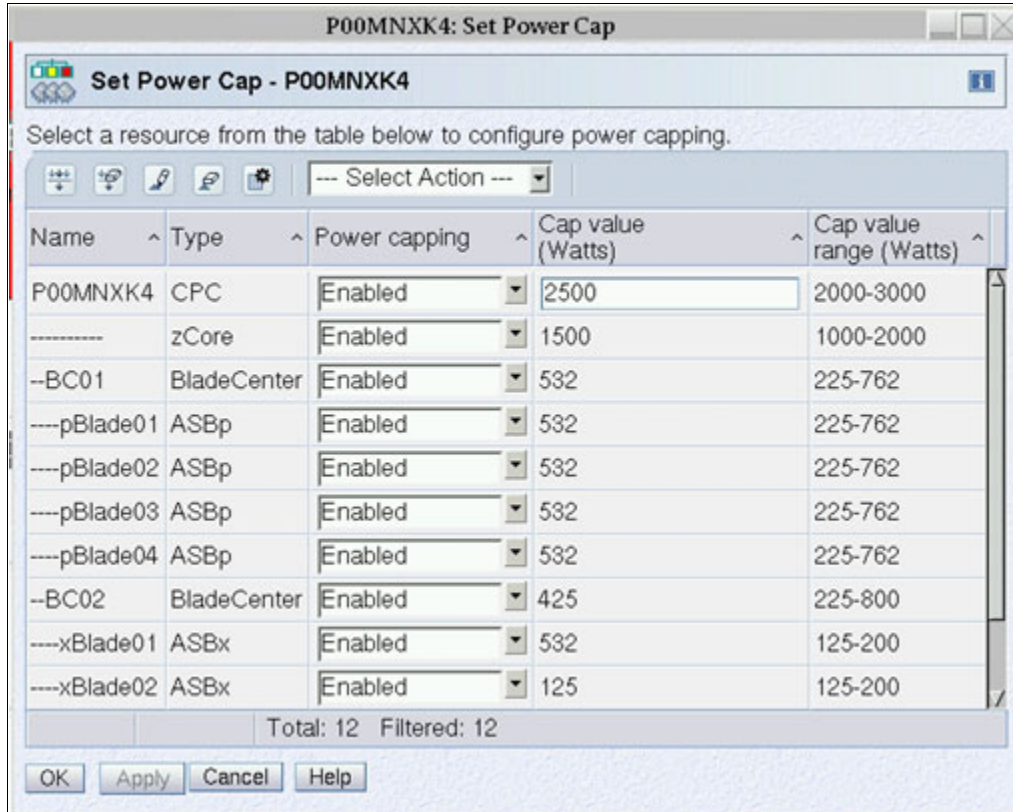
Figure 11-5   Set Power Cap panel

11.5 zBX environmentals

The following sections discuss the environmentals in summary for zEnterprise BladeCenter
Extension (zBX). For a full description of the environmentals for the zBX see *zBX Model 002
Installation Manual - Physical Planning,* GC27-2611-00

11.5.1 zBX configurations

The zBX can have from one to four racks. The racks are shipped separately and are bolted
together at installation time. Each rack can contain up to two BladeCenter chassis, and each
chassis can contain up to fourteen single-wide blades. The number of blades required
determines the actual components required for each configuration. The number of
BladeCenters and racks are generated by the quantity of blades (see Table 11-6).

Table 11-6   zBX configurations

Number of blades	Number of BladeCenters	Number of racks
7	1	1
14	1	1
28	2	1
42	3	2
56	4	2

Number of blades	Number of BladeCenters	Number of racks
70	5	3
84	6	3
98	7	4
112	8	4

zBX power

The zBX has its own power supplies and cords, independent of the z196 server power. Depending on the configuration of the zBX, up to 16 customer supplied power feeds may be required. A fully configured four-rack zBX has 16 power distribution units (PDUs). The zBX operates with:

► 50/60Hz AC power
► Voltage (240V)
► Both single-phase and three-phase wiring

The Power Distribution Units (PDU) options available for the zBX are as follows:

► FC 0520 - 7176 Model 3NU with attached Line-cord (US)
► FC 0521 - 7176 Model 2NX (WW)

The power cord options available for the zBX are as follows:

► FC 0531 - 4.3 meter, 60A/208V, US Line-cord, Single Phase
► FC 0532 - 4.3 meter, 63A/230V, non-US Line-cord, Single Phase
► FC 0533 - 4.3 meter, 32A/380V-415V, non-US Line-cord, Three Phase. Note that 32A WYE 380V or 415V gives you 220V or 240V line to neutral, respectively. This ensures that the BladeCenter maximum of 240V is not exceeded.

Power installation considerations

Each zBX BladeCenter operates from two fully-redundant power distribution units (PDUs) installed in the rack with the BladeCenter. These PDUs each have their own line-cords (see Table 11-7), allowing the system to survive the loss of customer power to either line-cord. If power is interrupted to one of the PDUs, the other PDU will pick up the entire load and the BladeCenter will continue to operate without interruption.

Table 11-7   Number of BladeCenter power cords

Number of BladeCenters	Number of power cords
1	2
2	4
3	6
4	8
5	10
6	12
7	14
8	16

For the maximum availability, the line-cords on each side of the racks should be powered from different building power distribution units.

Actual power consumption is dependent on the zBX configuration in terms of the number of BladeCenters and blades installed.

Input power in kVA is equal to the out power in kW. Heat output expressed in kBTU per hour is derived by multiplying the table entries by a factor of 3.4. For 3-phase installations, phase balancing is accomplished with the power cable connectors between the BladeCenters and the PDUs.

zBX cooling

The individual BladeCenter configuration is air cooled with two hot swap blower modules. The blower speeds vary depending on the ambient air temperature at the front of the BladeCenter unit and the temperature of internal BladeCenter components.

► If the ambient temperature is 25°C (77°F) or below, the BladeCenter unit blowers will run at their minimum rotational speed, increasing their speed as required to control internal BladeCenter temperature.
► If the ambient temperature is above 25°C (77°F), the blowers will run faster, increasing their speed as required to control internal BladeCenter unit temperature.
► If a blower fails, the remaining blower will run full speed and continues to cool the BladeCenter unit and blade servers.

The typical heat released by the different zBX solution configurations is in Table 11-8.

Table 11-8   zBX power consumption and heat output

Number of blades	Max utility power (kW)	Heat output (kBTU/hour)
7	7.3	24.82
14	12.1	41.14
28	21.7	73.78
42	31.3	106.42
56	40.9	139.06
70	50.5	171.70
84	60.1	204.34
98	69.7	236.98
112	79.3	269.62

Optional Rear Door Heat eXchanger - FC 0540

For data centers that have limited cooling capacity, using the Rear Door Heat eXchanger (see Figure 11-6 on page 341) is a more cost-effective solution than adding another air conditioning unit.

> **Note:** The Rear Door Heat eXchanger is not a requirement for BladeCenter cooling. It is a solution for clients that cannot upgrade a data center's air conditioning units due to space, budget, or other constraints.

The Rear Door Heat eXchanger has the following features:

► A water-cooled heat exchanger door is designed to dissipate heat generated from the back of the computer systems before it enters the room.
► An easy-to-mount rear door design attaches to client-supplied water, using industry standard fittings and couplings.
► Up to 50,000 BTUs (or approximately 15 kW) of heat can be removed from air exiting the back of a zBX rack.

The IBM Rear Door Heat eXchanger is an effective way to assist your air conditioning system in keeping your data center cool. It removes heat from the rack before the heat enters the room, allowing your air conditioning unit to handle the increasingly dense system deployment your organization requires to meet its growing computing needs (see Figure 11-6).
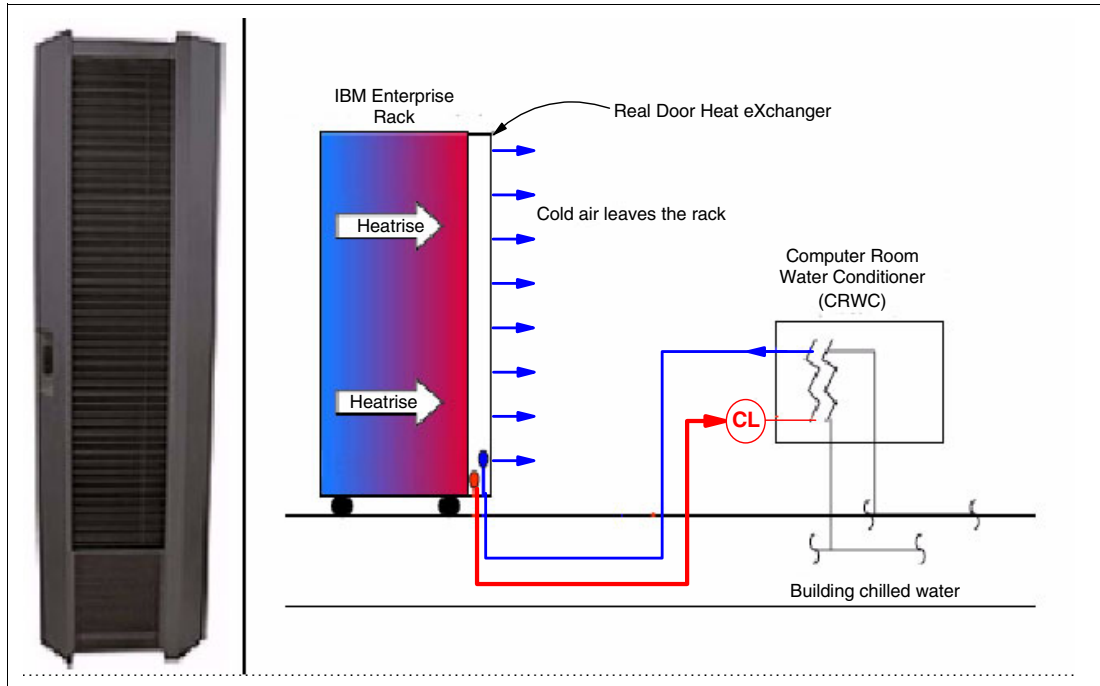
Figure 11-6   Rear Door Heat eXchanger (left) and functional diagram

The IBM Rear Door Heat eXchanger also offers a convenient way to handle dangerous "hot spots" in your data center.

zBX physical specifications

The zBX solution is delivered either with one (Rack B) or four racks (Rack B, C, D, and E). The physical dimensions of the zBX minimum and maximum solutions are in Table 9 on page 341.

Table 9   Dimensions of zBX racks

Racks with covers	Width mm (in)	Depth mm (in)	Height mm (in)
B	648 (25.5)	1105 (43.5)	2020 (79.5)
B+C	1296 (51.0)	1105 (43.5)	2020 (79.5)
B+C+D	1994 (76.5)	1105 (43.5)	2020 (79.5)
B+C+D+E	2592 (102)	1105 (43.5)	2020 (79.5)

Height Reduction FC 0570

This feature is required if it is necessary to reduce the shipping height for the zBX. This feature should be selected when it has been deemed necessary for delivery clearance purposes. This feature should be ordered if you have doorways with openings less than 1941 mm (76.4 inches) high. This feature accommodates doorway openings as low as 1832 mm (72.1 inches).

zBX weight

Table 11-10 shows the maximum weights of fully populated zBX racks and BladeCenters.

Table 11-10   Weights of zBX racks

Rack Description	Weight kg (lbs.)
B with 28 blades	740 (1630)
B + C full	1234 (2720)
B + C + D full	1728 (3810)
B + C + D + E full	2222 (4900)

Note: A fully configured Rack B is heavier than a fully configured Rack C, D, or E because Rack B has the two TORs installed.

For a complete view of the physical requirements see *zBX Model 002 Installation Manual - Physical Planning,* GC27-2611-00

12

Hardware Management Console

The Hardware Management Console (HMC) supports many functions and tasks to extend the management capabilities of z196. The HMC is important in the overall management of the data center infrastructure. When tasks are performed on the HMC, the commands are sent to one or more SEs, which then issue commands to their CPCs.

This chapter describes the HMC and Support Element (SE). HMCs that manage ensembles are discussed in Chapter 13, "Unified Resource Manager" on page 363.

This chapter discusses the following topics:

12.1 HMC and SE introduction

The Hardware Management Console (HMC) is a combination of a stand-alone computer and a set of management applications. The HMC is a closed system, which means that no other applications can be installed on it.

The HMC is used to set up, manage, monitor, and operate one or more IBM System z servers. It manages System z hardware, its logical partitions, and provides support applications. At least one HMC is required to operate a zEnterprise 196 server. If the zEnterprise 196 server is defined as a member of an ensemble, a pair of HMCs are required (a primary and an alternate). See 13.3, "Ensemble Physical Resource Management" on page 372 for a description and prerequisites.

> **Note:** The primary HMC and its alternate must be connected to the same subnetwork to allow the alternate HMC to take over the IP address of the primary HMC during failover processing.

An HMC can manage multiple System z servers and can be located at a local or a remote site. However, when a zEnterprise 196 server is defined as a member of an ensemble, some restrictions apply; see Chapter 13, "Unified Resource Manager" on page 363.

The Support Elements (SEs) are two integrated ThinkPads that are supplied with the System z server. One is always the active SE and the other is a strictly alternate element. The SEs are closed systems and no other applications can be installed on them.

When tasks are performed at the HMC, the commands are routed to the active SE of the System z server. One HMC can control up to 100 SEs and one SE can be controlled by up to 32 HMCs.

At the time of this writing, the zEnterprise System is shipped with HMC version 2.11.0, which is capable of supporting different System z server types. Many functions that are available on Version 2.11.0 are only supported when connected to a zEnterprise System server. HMC Version 2.11.0 supports the servers and SE versions shown in Table 12-1.

Table 12-1   System z196 HMC server support summary

Server	Machine type	Minimum firmware driver	Minimum SE version
z196	2817	86	2.11.0
z10 BC	2098	76	2.10.1
z10 EC	2097	73	2.10.0
z9 BC	2096	67	2.9.2
z9 EC	2094	67	2.9.2
z890	2086	55	1.8.2
z990	2084	55	1.8.2

12.2 HMC and SE connectivity

The HMC has two Ethernet adapters. Each SE has one Ethernet adapter and both SEs are connected to the same Ethernet switch. The Ethernet switch (FC 0089) can be supplied with every system order. Additional Ethernet switches (up to a total of ten) may be added.

The switch is a standalone unit located outside the frame; it operates on building AC power. A client-supplied switch may be used if it matches IBM specifications.

The internal LAN for the SEs on the zEnterprise 196 server (z196) connects to the Bulk Power Hub. The HMC must be connected to the Ethernet switch through one of its Ethernet ports. Only the switch may be connected to the client ports J01 and J02 on the Bulk Power Hub. With respect to the zEnterprise System network topology architecture, as discussed in 13.4.3, "Network Virtualization Management" on page 377, this network is referred to as the *Customer-managed Management Network*. Other server's SEs may also be connected to the switches. To provide redundancy for the HMCs, two switches are recommended, as shown in Figure 12-1.

For more information, see *zEnterprise 196 Installation Manual for Physical Planning, GC28-6897*.
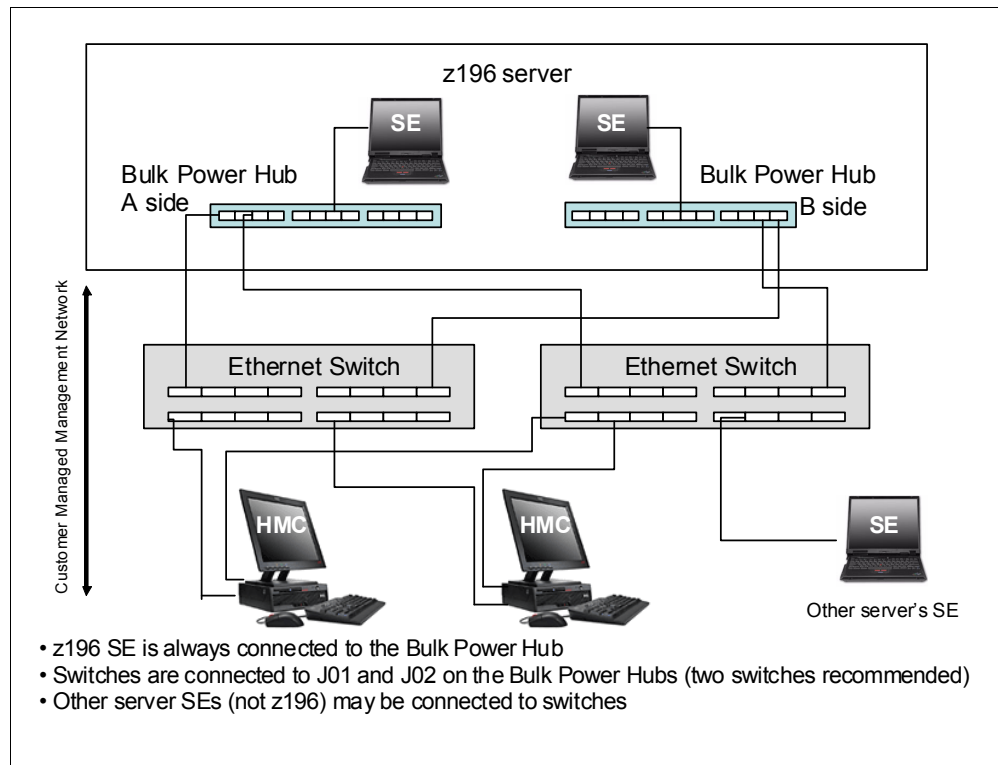
Figure 12-1   HMC to SE connectivity

The HMC and SE have several exploiters that either require or can take advantage of broadband connectivity to the Internet and your corporate intranet.

Several methods are available for setting up the network to allow access to the HMC from your corporate intranet or to allow the HMC to access the Internet. The method you select depends on your connectivity and security requirements.

One example is to connect the second Ethernet port of the HMC to a separate switch that has access to the intranet or Internet, as shown in Figure 12-2.

Also, the HMC has built-in firewall capabilities to protect the HMC and SE environment. The HMC firewall can be set up to allow certain types of TCP/IP traffic between the HMC and permitted destinations in your corporate intranet or the Internet.

> **Note:** Configuration of network components, such as routers or firewall rules, is beyond the scope of this document. Anytime networks are interconnected, security exposures can exist. Network security is a client's responsibility.
>
> The document "IBM System z HMC Security" provides information about HMC security. It is available on IBM Resource Link:
>
>   https://www-304.ibm.com/servers/resourcelink/lib03011.nsf/pages/zHmcSecurity
>   /$file/zHMCSecurity.pdf
>
> Registration is required to access IBM Resource Link.
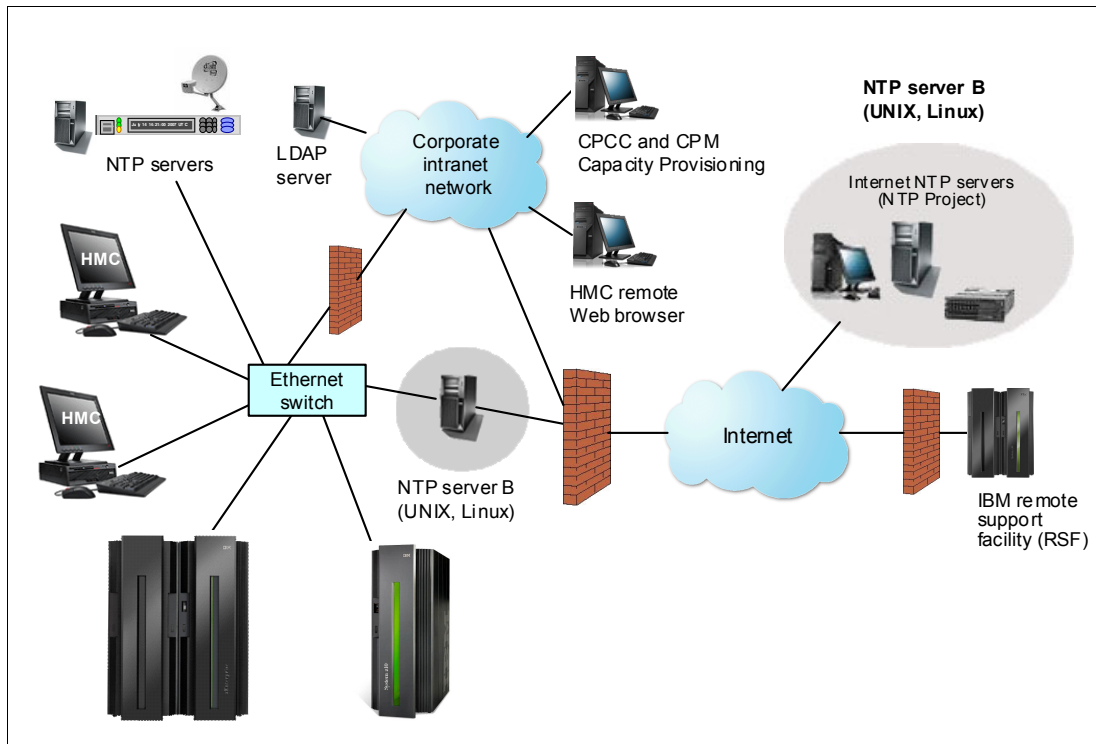
Figure 12-2   HMC connectivity

The HMC and SE network connectivity should be planned carefully to allow for current and future use. Many of the System z capabilities benefit from the various network connectivity options available. For example, functions available to the HMC that depend on the HMC connectivity include:

► LDAP support, that can be used for HMC user authentication.

► STP and NTP client/server support.

► RSF, available through the HMC with an Internet-based connection, providing increased performance as compared to dial-up.

▶ Enablement of the SNMP and CIM APIs to support automation or management applications such as Capacity Provisioning Manager and Active Energy Manager (AEM).

TCP/IP Version 6 on HMC and SE

The HMC and SE can communicate using IPv4, IPv6, or both. Assigning a static IP address to an SE is unnecessary if the SE only has to communicate with HMCs on the same subnet. An HMC and SE can use IPv6 link-local addresses to communicate with each other.

IPv6 link-local address characteristics are:

▶ Every IPv6 network interface is assigned a link-local IP address.

▶ A link-local address is for use on a single link (subnet) and is never routed.

▶ Two IPv6-capable hosts on a subnet can communicate by using link-local addresses, without having any other IP addresses assigned.

Assigning addresses to HMC and SE

An HMC can have the following IP addresses:

▶ Statically assigned IPv4 or statically assigned IPv6.

▶ DHCP assigned IPv4 or DHCP assigned IPv6.

▶ Autoconfigured IPv6:

– Link-local is assigned to every network interface.

– Router-advertised, which is broadcast from the router, can be combined with a MAC address to create a totally unique address.

– Privacy extensions can be enabled for these addresses as a way to avoid using MAC address as part of address to ensure uniqueness.

An SE can have the following IP addresses:

▶ Statically assigned IPv4 or statically assigned IPv6.

▶ Autoconfigured IPv6 as link-local or router-advertised.

IP addresses on the SE cannot be dynamically assigned through DHCP to ensure repeatable address assignments. Privacy extensions are not used.

The HMC uses IPv4 and IPv6 multicasting to automatically discover SEs. The HMC Network Diagnostic Information task may be used to identify the IP addresses (IPv4 and IPv6) that are being used by the HMC to communicate to the CPC SEs.

IPv6 addresses are easily identified. A fully qualified IPV6 address has 16 bytes, written as eight 16-bit hex blocks separated by colons, as shown in the following example:

`2001:0db8:0000:0000:0202:B3FF:fe1e:8329`

Because many IPv6 addresses are not fully qualified, shorthand notation can be used. This is where the leading zeros can be omitted and a series of consecutive zeros can be replaced with a double colon. The address in the previous example can also be written as:

`2001:db8::202:B3FF:fe1e:8329`

For remote operations using a Web browser, if an IPv6 address is assigned to the HMC, navigate to it by specifying that address. The address must be surrounded with square brackets in the browser's address field:

`https://[fdab:1b89:fc07:1:201:6cff:fe72:ba7c]`

Using link-local addresses must be supported by browsers.

12.3 Remote Support Facility

The HMC Remote Support Facility (RSF) provides communication to a centralized IBM support network for hardware problem reporting and service. The types of communication provided include:

- ► Problem reporting and repair data.
- ► Fix and firmware delivery to the service processor and HMC.
- ► Hardware inventory data.
- ► On-demand enablement.

The HMC can be configured to send hardware service related information to IBM by using a dialup connection over a modem or using an Internet connection. The advantages of using an Internet connection include:

- ► Significantly faster transmission speed.

- ► Ability to send more data on an initial problem request, potentially resulting in more rapid problem resolution.

- ► Reduced customer expense (for example, the cost of a dedicated analog telephone line)

- ► Greater reliability.

Unless the enterprise's security policy prohibits any connectivity from the HMC over the Internet, an Internet connection is recommended. With z196 there will be limitations to some firmware components such they can only be updated via an Internet connection or media, not via a dialup connection.

If both types of connections are configured, the Internet will be tried first and, if it fails, then the modem is used.

The following security characteristics are in effect regardless of the connectivity method chosen:

- ► Remote Support Facility requests are always initiated from the HMC to IBM. An inbound connection is never initiated from the IBM Service Support System.

- ► All data transferred between the HMC and the IBM Service Support System is encrypted in a high-grade Secure Sockets Layer (SSL) encryption.

- ► When initializing the SSL encrypted connection the HMC validates the trusted host by its digital signature issued for the IBM Service Support System.

- ► Data sent to the IBM Service Support System consists solely of hardware problems and configuration data. No application or customer data is transmitted to IBM.

12.4 HMC remote operations

The z196 HMC application simultaneously supports one local user and any number of remote users. Remote operations provide the same interface used by a local HMC operator. The two ways to perform remote manual operations are:

- ► Using a Remote HMC

   A remote HMC is an HMC that is on a different subnet from the SE, therefore the SE cannot be automatically discovered with IP multicast.

▶ Using a Web browser to connect to an HMC

The choice between a remote HMC and a Web browser connected to a local HMC is determined by the scope of control needed. A remote HMC can control only a specific set of objects, but a Web browser connected to a local HMC controls the same set of objects as the local HMC.

In addition, consider communications connectivity and speed. LAN connectivity provides acceptable communications for either a remote HMC or Web browser control of a local HMC, but dialup connectivity is only acceptable for occasional Web browser control.

Using a remote HMC

Although a remote HMC offers the same functionality as a local HMC, its connection configuration differs from a local HMC. The remote HMC requires the same setup and maintenance as other HMCs (see Figure 12-2 on page 346).

A remote HMC requires TCP/IP connectivity to each SE to be managed. Therefore, any existing customer-installed firewall between the remote HMC and its managed objects must permit communications between the HMC and SE. For service and support, the remote HMC also requires connectivity to IBM, or to another HMC with connectivity to IBM.

Using a Web browser

Each HMC contains a Web server that can be configured to allow remote access for a specified set of users. When properly configured, an HMC can provide a remote user with access to all the functions of a local HMC except those that require physical access to the diskette or DVD media. The user interface in the browser is the same as the local HMC and has the same functionality as the local HMC.

The Web browser can be connected to the local HMC by using either a LAN TCP/IP connection or a switched, dial-up, or network PPP TCP/IP connection. Both connection types use only encrypted (HTTPS) protocols, as configured in the local HMC. If a PPP connection is used, the PPP password must be configured in the local HMC and in the remote browser system. Logon security for a Web browser is provided by the local HMC user logon procedures. Certificates for secure communications are provided, and can be changed by the user.

A remote browser session to the primary HMC that is managing an ensemble allows a user to perform ensemble-related actions.

12.5 z196 HMC and SE key capabilities

The z196 comes with the HMC application Version 2.11.0. For a complete list of traditional HMC functions, see *System z HMC Operations Guide Version 2.11.0, SC28-6895*.

12.5.1 CPC management

The HMC is the primary place for central processor complex (CPC) control. For example, to define hardware to the z196, the I/O configuration data set (IOCDS) must be defined. The IOCDS contains definitions of logical partitions, channel subsystems, control units and devices and their accessibility from logical partitions. IOCDS can be created and put into production from the HMC.

The z196 server is powered on and off from the HMC. The HMC is used to initiate power-on reset (POR) of the server. During the POR, among other things, PUs are characterized and placed into their respective pools, memory is put into a single main storage pool and the IOCDS is loaded and initialized into the hardware system area.

The Hardware messages task displays hardware-related messages at CPC level, at logical partition level, SE level, or hardware messages related to the HMC itself.

12.5.2 LPAR management

Use the HMC to define logical partition properties, such as how many processors of each type, how many are reserved, or how much memory is assigned to it. These parameters are defined in logical partition profiles and they are stored on the SE.

Because PR/SM has to manage logical partition access to processors and initial weights of each partition, weights are used to prioritize partition access to processors.

A "Load" task on the HMC enables you to IPL an operating system. It causes a program to be read from a designated device and initiates that program. The operating system can be IPLed from disk, the HMC CD-ROM/DVD, or an FTP server.

When a logical partition is active and an operating system is running in it, you may use the HMC to dynamically change certain logical partition parameters. The HMC also provides an interface to change partition weights, add logical processors to partitions, and add memory.

LPAR weights can be also changed through a scheduled operation. Use the HMCs "Customize Scheduled Operations" task to define the weights that will be set to logical partitions at the scheduled time.

Channel paths can be dynamically configured on and off, as needed for each partition, from an HMC.

The "Change LPAR Controls" task for z196 has the ability to export the "Change LPAR Controls" table data to a .csv formatted file. This support is available to a user when connected to the HMC remotely via a web browser.

Partition capping values can be scheduled and are specified on the "Change LPAR Controls scheduled operation" support. Viewing of Details about an existing "Change LPAR Controls schedule operation" is available on the SE.

The "Change LPAR Group Controls" task provides the ability to modify the group members and group capacity setting. These updates can be applied dynamically to the running system or saved to the Group and corresponding Image profiles. In z196 the SNMP and CIM API allow dynamic changes to both the group members and group capacity setting.

12.5.3 Operating system communication

The Operating system messages task displays messages from a logical partition. You may also enter operating system commands and interact with the system.

The HMC also provides integrated 3270 and ASCII consoles so you can access an operating system without requiring other network or network devices (such as TCP/IP or control units).

12.5.4 SE access

Being physically close to an SE is not necessary to use it. The HMC can be used to remotely access the SE; the same interface as in the SE provided.

The HMC enables you to:

► Synchronize content of the primary SE to the alternate SE.
► Determine whether a switch from primary to the alternate can be performed.
► Switch between primary and alternate SEs.

12.5.5 Monitoring

Monitor Task Group

The task group ("Monitor") holds "monitoring" related tasks (for both HMC and SE). In previous versions some tasks ("Activity") were located elsewhere.
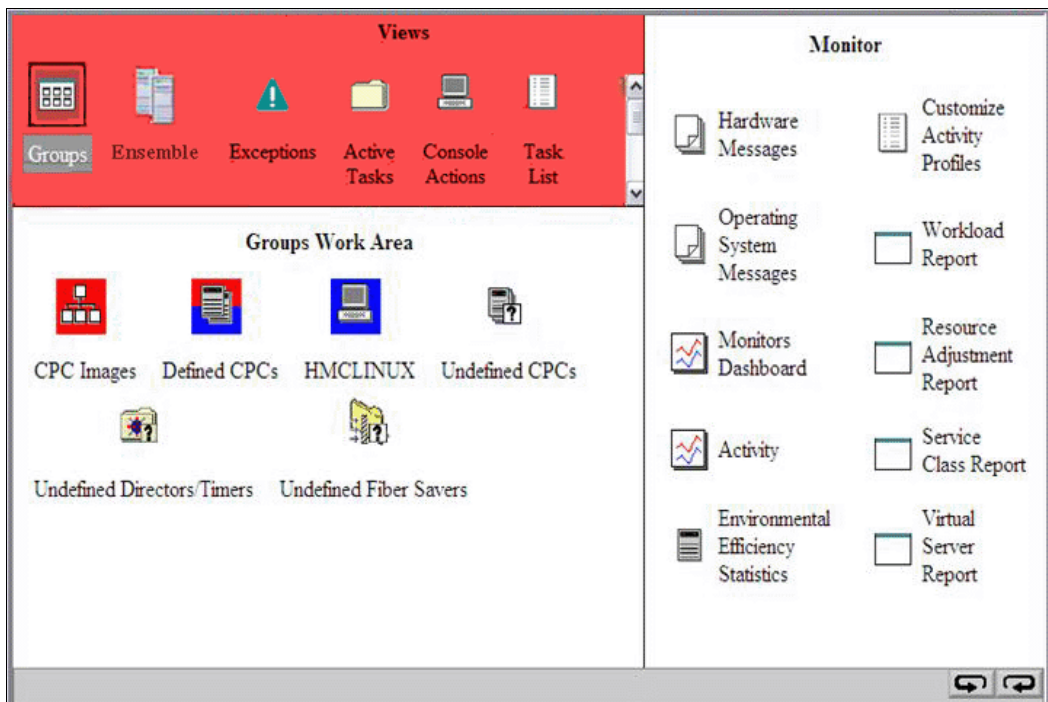
Figure 12-3   HMC Monitor Task Group

Use the System Activity Display (SAD) task on the HMC to monitor the activity of one or more CPCs. The task monitors processor and channel usage. You may define multiple activity profiles. The task also includes power monitoring information, the power being consumed, and the air input temperature for the server.

For HMC users with Service authority, SAD shows information about each power cord. Power cord information should only be used by those with extensive knowledge about System z196 internals and three-phase electrical circuits. Weekly call-home data includes power information for each power cord.

Monitors Dashboard Task

In z196 the "Monitors Dashboard" task in the Monitor task group provides a tree-based view of resources and allows an aggregated activity view when looking at large configurations. It

also allows for details for objects with smaller scope. Multiple graphical ways of displaying data are available, such as history charts.
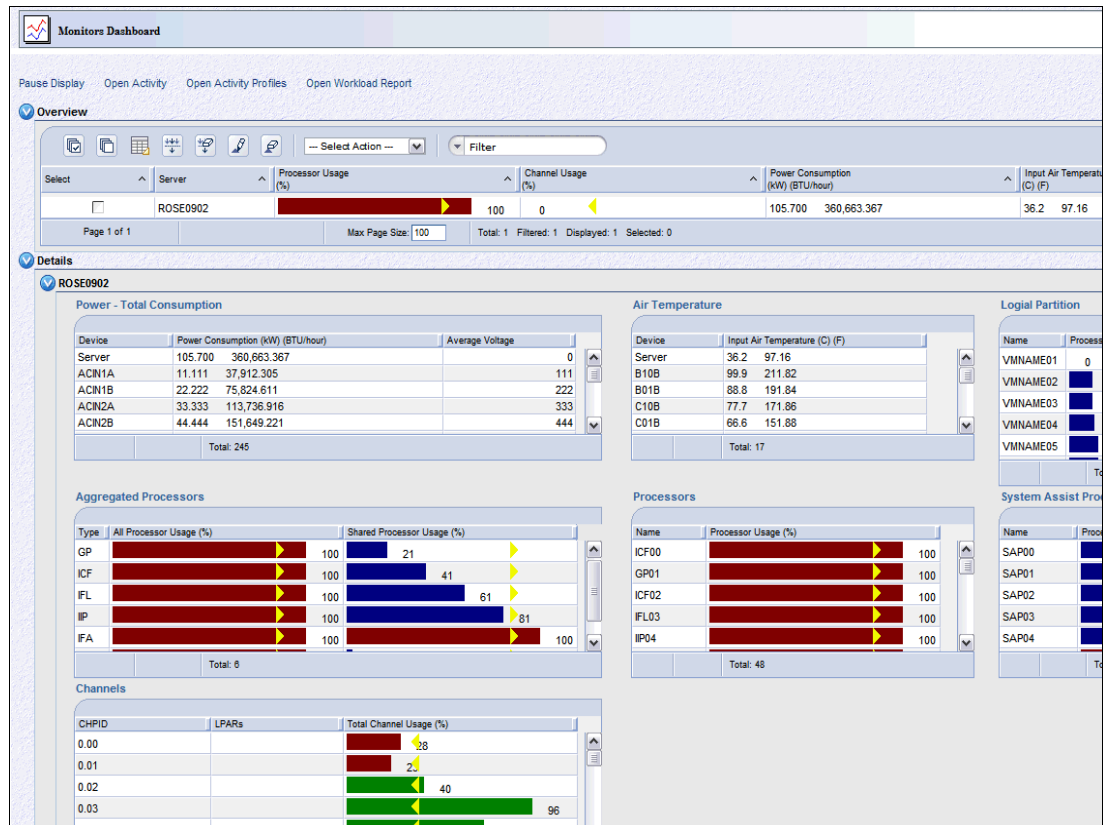
Figure 12-4   Monitors Dashboard

IBM Systems Director Active Energy Manager

As discussed in "Static power saving mode" on page 336, the Active Energy Manager is an energy management solution building-block that returns true control of energy costs to the customer. It is a software tool that provides a single view of the actual power usage across multiple platforms and helps to increase energy efficiency by controlling power use across the data center.

Active Energy Manager runs on Windows, Linux on IBM System x, Linux on IBM System p, and Linux on IBM System z.

How Active Energy Manager works

Active Energy Manager interacts with systems as follows:

► Hardware, firmware, and systems management software in servers and blades provide information to Active Energy Manager.

► Active Energy Manager calculates the power consumption for each component and tracks power usage over time.

► When power is constrained, Active Energy Manager allows power to be allocated on a server-by-server basis.

► Active Energy Manager ensures that limiting the power consumption does not affect performance.

► Sensors and alerts warn the user if limiting power to a particular server can affect performance.

Data available from z196 HMC

The following data is available from the z196 HMC:

► System name, machine type, model, serial number, firmware level.

► Ambient temperature.

► Exhaust temperature.

► Average power (over a one-minute period).

► Peak power (over a one-minute period).

► Limited status and configuration information. This information helps explain changes to the power consumption, called Events, which can be:
  – Changes in fan speed.
  – Changes between power-off, power-on, and IML-complete states.
  – Number of I/O drawers.
  – CBU records expiration(s).

Environmental Efficiency Statistic Task

The Environmental Efficiency Statistic Task is part of the "Monitor" task group.

The Active Energy Manager (AEM) plug-in for the IBM Director provides historical power consumption and thermal information for z196 is available on the HMC. This task provides similar data along with a historical summary of processor and channel utilization.

The data is presented in table form, graphical ("histogram") form and it can also be exported to a .csv formatted file so that it can be imported into tools such as Excel or Lotus 1-2-3®.

Environmental Efficiency Statistics - P0LXSM13

Starting Date: 10/01/09

Amount of Data: One day

Chart Type: Power Consumption (watts)

To display new data, enter the start date and/or the amount of data, and select Refresh.

Sample Date/Time	Power Consumption (watts)	Power Consumption (BTUs)	Temperature (C)	Temperature (F)	CP Utilization (%)	Blade CP
Oct 1, 2009 12:00:00 AM	4987	17016	24.3	75.7	0	
Oct 1, 2009 1:00:00 AM	4992	17033	24.3	75.7	0	
Oct 1, 2009 2:00:00 AM	4990	17027	24.4	75.9	0	
Oct 1, 2009 3:00:00 AM	4990	17027	24.3	75.7	0	
Oct 1, 2009 4:00:00 AM	4991	17030	24.2	75.6	0	
Oct 1, 2009 5:00:00 AM	4990	17027	24.2	75.6	0	

Total: 25

Figure 12-5   Environmental Efficiency Statistics

12.5.6 Capacity on Demand support

All Capacity on Demand upgrades are performed from the SE "Perform a model conversion" task. Use the task to retrieve and activate a permanent upgrade; and to retrieve, install, activate and deactivate a temporary upgrade. The task shows all installed or staged LICCC records to help you manage them. It also shows a history of record activities.

HMC for System z196 CoD capabilities include:

- ► SNMP API support:
  - – API interfaces for granular activation and deactivation.
  - – API interfaces for enhanced Capacity On Demand query information.
  - – API Event notification for any Capacity On Demand change activity on the system.
  - – Capacity On Demand API interfaces (such as On/Off CoD and CBU).
- ► SE panel features (accessed through HMC Single Object Operations):
  - – Panel controls for granular activation and deactivation.
  - – History panel for all Capacity On Demand actions.
  - – Descriptions editing of Capacity On Demand records.

HMC/SE Version 2.11.0 provides CoD information such as:

- ► MSU and processor tokens shown on panels.
- ► Last activation time shown on panels.
- ► Pending resources are shown by processor type instead of just a total count.
- ► Option to show details of installed and staged permanent records.
- ► More details for the *Attention!* state on panels (by providing seven additional flags).

HMC and SE are an integral part for the z/OS Capacity Provisioning environment. The Capacity Provisioning Manager (CPM) communicates with the HMC through System z APIs and enters CoD requests. For this reason, SNMP must be configured and enabled on the HMC.

For additional information about using and setting up CPM, see the publication:

- ► *z/OS MVS Capacity Provisioning User's Guide*, SA33-8299

12.5.7 Server Time Protocol support

Server Time Protocol (STP) is supported on System z servers. With the STP functions, the role of the HMC has been extended to provide the user interface for managing the Coordinated Timing Network (CTN).

z196 relies solely on STP for time synchronization, but continues to provide support of a Pulse per Second (PPS) port. The System (Sysplex) Time task does not contain the ETR Status and ETR Configuration tabs when the target is a z196. An ETR ID can be entered on the STP Configuration tab when system is z196 to support participation in a mixed CTN.

In a mixed CTN (one containing both STP and Sysplex Timer), the HMC can be used to:

- ► Initialize or modify the CTN ID and ETR port states.
- ► Monitor the status of the CTN.
- ► Monitor the status of the coupling links initialized for STP message exchanges.

In an STP-only CTN, the HMC can be used to:

- ► Initialize or modify the CTN ID.

- ► Initialize the time, manually or by dialing out to a time service, so that the Coordinated Server Time (CST) can be set to within 100 ms of an international time standard, such as UTC.

- ► Initialize the time zone offset, daylight saving time offset, and leap second offset.

- ► Schedule periodic dial-outs to a time service so that CST can be steered to the international time standard.

- ► Assign the roles of preferred, backup, and current time servers, as well as arbiter.

- ► Adjust time by up to plus or minus 60 seconds.

- ► Schedule changes to the offsets listed. STP can automatically schedule daylight saving time, based on the selected time zone.

- ► Monitor the status of the CTN.

- ► Monitor the status of the coupling links initialized for STP message exchanges.

For diagnostic purposes the Pulse per Second port state on a z196 can be displayed and fenced ports can be reset individually.

For additional planning and setup information, see the following publications:

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

12.5.8 NTP client/server support on HMC

The Network Time Protocol (NTP) client support allows an STP-only Coordinated Timing Network (CTN) to use an NTP server as an External Time Source (ETS). This capability addresses the requirements for:

- ► Customers who want time accuracy for the STP-only CTN.
- ► Using a common time reference across heterogeneous platforms.

NTP client allows the same accurate time across an enterprise comprised of heterogeneous platforms.

NTP server becomes the single time source, ETS for STP, as well as other servers that are not System z (such as UNIX, Windows NT®, and others) that have NTP clients.

The HMC can act as an NTP server. With this support, z196 can get time from the HMC without accessing other than the HMC/SE network.

When the HMC is used as an NTP server, it can be configured to get the NTP source from the Internet. For this type of configuration, a separate LAN is recommended from the HMC/SE LAN.

The NTP client support can be used to connect to other NTP servers that can potentially receive NTP through the Internet. When using another NTP server, then the NTP server becomes the single time source, ETS for STP, and other servers that are not System z servers (such as UNIX, Windows NT, and others) that have NTP clients.

When the HMC is configured to have an NTP client running, the HMC time will be continuously synchronized to an NTP server instead of synchronizing to a support element.

For additional planning and setup information for STP and NTP check the following manuals:

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

12.5.9 Security and User ID Management

HMC/SE Security Audit Improvements

With "Audit & Log Management" task audit reports can be generated, viewed, saved, and offloaded. The "Customize Scheduled Operations" task allows for scheduling of audit report generation, saving, and offloading. The "Monitor System Events" task allows for Security Logs to result in e-mail notifications using the same type of filters and rules that are used for both hardware and operating system messages.

In z196 the ability to offload the following HMC and SE log files for Customer Audit was added:

► Console Event Log.
► Console Service History.
► Tasks Performed Log.
► Security Logs.
► System Log.

Full log offload as well as delta log offload (since last offload request) is provided. Offloading to removable media as well as to remote locations via FTP is available. The offloading can be manually initiated via the new "Audit & Log Management" task or scheduled via the Scheduled Operations task. The data can be offloaded in the HTML and XML formats.

HMC User ID Templates and LDAP User Authentication

LDAP User Authentication and HMC User ID templates enable adding/removing HMC users according to your own corporate security environment, by using an LDAP server as the central authority. Each HMC User ID template defines the specific levels of authorization levels for the tasks/objects for the user mapped to that template. The HMC User is mapped to a specific User ID template by User ID pattern matching and/or obtaining the name of the User ID template from content in the LDAP Server schema data.

View Only User IDs/Access for HMC/SE

With HMC and SE User ID support users can be created who have View Only access to selected tasks. Support for View Only user IDs is available for:

► Hardware Messages.
► Operating System Messages.
► Customize/Delete Activation Profiles.
► Advanced Facilities.
► Configure On/Off.

12.5.10 System Input/Output Configuration Analyzer on the SE/HMC

A System Input/Output Configuration Analyzer task is provided that supports the system I/O configuration function.

The information necessary to manage a system's I/O configuration has to be obtained from many separate applications. A System Input/Output Configuration Analyzer task enables the system hardware administrator to access, from one location, the information from these many sources. Managing I/O configurations then becomes easier, particularly across multiple servers.

The System Input/Output Configuration Analyzer task performs the following functions:

► Analyzes the current active IOCDS on the SE.

► Extracts information about the defined channel, partitions, link addresses, and control units.

► Requests the channels node ID information. The FICON channels support remote node ID information, which is also collected.

The System Input/Output Configuration Analyzer is a view-only tool. It does not offer any options other than viewing options. With the tool, data is formatted and displayed in five different views, various sort options are available, and data can be exported to a USB flash drive for a later viewing.

The five views are:

► PCHID Control Unit View, which shows PCHIDs, CSS, CHPIDs and their control units.

► PCHID Partition View, which shows PCHIDS, CSS, CHPIDs and the partitions they are in.

► Control Unit View, which shows the control units, their PCHIDs, and their link addresses in each CSS.

► Link Load View, which shows the Link address and the PCHIDs that use it.

► Node ID View, which shows the Node ID data under the PCHIDs.

12.5.11 Test Support Element Communications

The test "Support Element Communications", available on the "Network Diagnostic Information" task, tests that communication between the HMC and SE is available.

The tool performs five tests:

1. HMC pings SE.
2. HMC connects to SE and also verifies the SE is at the correct level.
3. HMC sends a message to SE and receives a response.
4. SE connects back to HMC.
5. SE sends a message to HMC and receives a response.

12.5.12 Automated operations

As an alternative to manual operations, a computer can interact with the consoles through an application programming interface (API). The interface allows a program to monitor and control the hardware components of the system in the same way a human can monitor and control the system. The HMC APIs provide monitoring and control functions through TCP/IP SNMP and CIM to an HMC. These APIs provide the ability to get and set a managed object's attributes, issue commands, receive asynchronous notifications, and generate SNMP traps.

The HMC supports the Common Information Model (CIM) as an additional systems management API. The focus is on attribute query and operational management functions for System z, such as CPCs, images, activation profiles. The System z196 contains a number of enhancements to the CIM systems management API. The function is similar to that provided by the SNMP API.

For additional information about APIs, see the *System z Application Programming Interfaces,* SB10-7030.

12.5.13 Cryptographic support

Cryptographic hardware

The z196 includes both standard cryptographic hardware and optional cryptographic features for flexibility and growth capability.

The HMC/SE interface provides the capability to:

► Define the cryptographic controls.
► Dynamically add a Crypto feature to a partition for the first time.
► Dynamically add a Crypto feature to a partition already using Crypto.
► Dynamically remove Crypto feature from a partition.

A "Usage Domain Zeroize" task is provided to clear the appropriate partition crypto keys for a given usage domain when removing a crypto card from a partition. For detailed set-up information, see *IBM System z10 Enterprise Class Configuration Setup*, SG24-7571.

Digitally signed firmware

One critical issue with firmware upgrades is security and data integrity. Procedures are in place to use a process to digitally sign the firmware update files sent to the HMC, the SE, and the TKE. Using a hash-algorithm, a message digest is generated that is then encrypted with a private key to produce a digital signature. This operation ensures that any changes made to the data will be detected during the upgrade process. It helps ensure that no malware can be installed on System z products during firmware updates. It enables, with other existing security functions, System z196 CPACF functions to comply with Federal Information Processing Standard (FIPS) 140-2 Level 1 for Cryptographic Licensed Internal Code (LIC) changes. The enhancement follows the System z focus of security for the HMC and the SE.

12.5.14 z/VM virtual machine management

The HMC can be used for basic management of z/VM and its virtual machines. The HMC exploits the z/VM Systems Management Application Programming Interface (SMAPI) and provides a graphical user interface (GUI)-based alternative to the 3270 interface.

Monitoring the status information and changing the settings of z/VM and its virtual machines are possible. From the HMC interface, virtual machines can be activated, monitored, and deactivated.

Authorized HMC users can obtain various status information, such as:

► Configuration of the particular z/VM virtual machine
► z/VM image-wide information about virtual switches and guest LANs
► Virtual Machine Resource Manager (VMRM) configuration and measurement data

The activation and deactivation of z/VM virtual machines is integrated into the HMC interface. You can select the Activate and Deactivate tasks on CPC and CPC image objects, and for virtual machines management.

An event monitor is a trigger that is listening for events from objects managed by HMC. When z/VM virtual machines change their status, they generate such events. You can create event monitors to handle the events coming from z/VM virtual machines. For example, selected users can be notified by an e-mail message if the virtual machine changes status from Operating to Exception, or any other state.

In addition, in z/VM V5R4, the APIs can perform the following functions:

- ► Create, delete, replace, query, lock, and unlock directory profiles.

- ► Manage and query LAN access lists (granting and revoking access to specific user IDs).

- ► Define, delete, and query virtual CPUs, within an active virtual image and in a virtual image's directory entry.

- ► Set the maximum number of virtual processors that can be defined in a virtual image's directory entry.

12.5.15 Installation support for z/VM using the HMC

The traditional way of installing Linux on System z in the z/VM virtual machine requires a network connection to a file server that is hosting the installation files of the Linux distribution.

Starting with z/VM V5R4 and System z10, Linux on System z can be installed in a z/VM virtual machine from the HMC workstation DVD drive. This Linux on System z installation can exploit the existing communication path between the HMC and the SE, where *no external network and no additional network setup is necessary* for the installation. This simplification can eliminate potential customer concerns and additional configuration efforts.

12.5.16 Power Saving Mode

The z196 power saving feature is built upon a mechanism for cycle and voltage steering and can be enabled/disabled at the HMC /SE. This mode reduces processor cycle time for all System z processors in the system. Memory and IO cycle times are not affected. Power Saving mode can be specified as "Custom Energy Management", which will use values specified on the "Set Power Saving" panel, or "Emergency High Performance", where all objects are placed into High Performance mode.
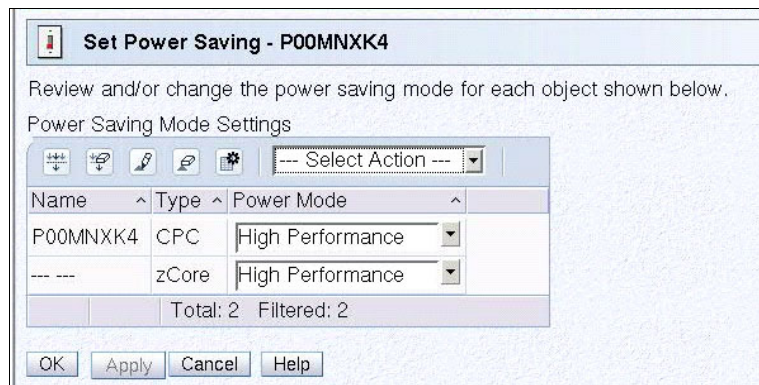
Figure 12-6　Set Power Saving Mode

12.6 HMC in an ensemble

An ensemble is a platform systems management domain consisting of one or more z196 nodes with optional zBXs. Through the ensemble, System z and non-System z resources are effectively integrated into a single platform. The ensemble provides an integrated way to manage virtual server resources and the workloads that can be deployed on those resources.

The ensemble is provisioned and managed through the Unified Resource Manager residing in the HMC. The Unified Resource Manager is a large set of functions for system management that can be grouped as follows:

► Defining and managing virtual environments. This includes the automatic discovery as well as the definition of I/O and other hardware components across z196 and zBX, and the definition and management of LPARs, virtual machines, and virtualized LANs.
► Defining and managing workloads and workload policies.
► Receiving and applying corrections and upgrades to the Licensed Internal Code.
► Performing temporary and permanent z196 capacity upgrades.

The following feature codes must be ordered to equip an HMC with the Unified Resource Manager:

► 0025—Ensemble Membership Flag
► 0019—Manage Firmware Suite
► 0020—Automate Firmware Suite (optional)

The system management functions that pertain to an *ensemble*, exploit the virtual server resources and the intraensemble management network (IEDN). They are provided by the HMC/SE via the internode management network (INMN).

Figure 11-10 depicts an ensemble with two z196s and a zBX that are managed by the Unified Resource Manager residing in the primary and alternate HMCs. CPC1 (z196) owns the zBX, while CPC2 is a stand-alone z196.
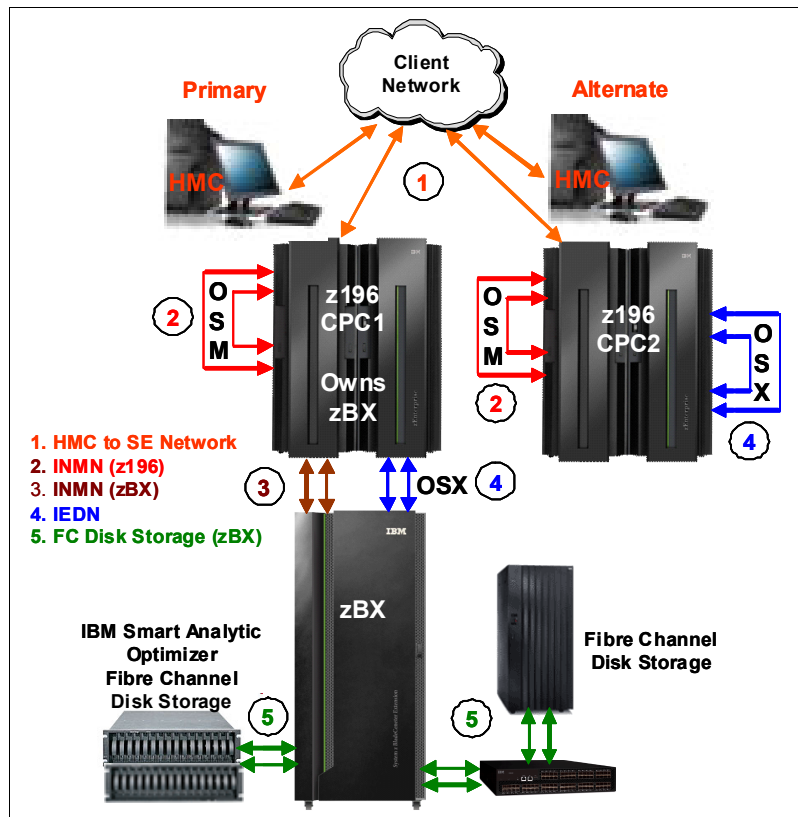
Figure 12-7   Ensemble example - with primary and alternate HMCs

For the stand-alone z196 ensemble node (CPC2), two OSA Express-3 1000BASE-T ports (CHPID type OSM) connect to the Bulk Power Hubs (port J07) with 3.2 meter Category 5 Ethernet cables. The HMCs also communicate with all the components of the ensemble via the BPHs in the z196.

The OSA Express-3 10 GbE ports (CHPID type OSX) are connected to one another in the z196 with client provided 10 GbE loop back cables (either SR or LR, depending on the OSA feature).

Details for ensemble connectivity for a z196 with a zBX can be found in "zBX connectivity" on page 190.

12.6.1 HMC/SE ensemble options

The ensemble starts with a pair of HMCs that are designated as the primary and alternate HMCs, and are assigned an ensemble identity. The z196s and zBXs are then added to the ensemble through an explicit action at the primary HMC.

Feature code 0025 (Ensemble Membership Flag) is associated with an HMC when a z196 is ordered. This feature code is required on the *owning* z196 to be able to attach a zBX.

A new task called *Create Ensemble* will allow the Access Administrator to create an ensemble that contains CPCs, Images, workloads, virtual networks and storage pools, either with or without an optional zBX.

Table 12-2 lists the key management functions in the primary HMC that support an ensemble.

Table 12-2   Key management functions of the primary HMC

Application	Description
Platform Performance Management	Manage Ensemble Image workloads, performance policies, and monitor goals
Virtual Network Management	Manage Ensemble server network connectivity on the Intraensemble data network (IEDN)
Virtual Server Management	Manage Ensemble Virtual server workloads, performance and monitor goals
Hypervisor Management	Manage Ensemble Hypervisor workloads, performance and monitor goals
z/VM Guest Management	HMC access to the services and APIs needed to manage Ensemble z/VM guests
Entitlement Management	Manage Ensemble Entitlement of different levels of Platform Performance Management
Energy Management	Manage Active Energy Manager, to access the Ensemble power/thermal data

If a z196 has been entered into an ensemble, then the CPC Details task on the SE and HMC will reflect the ensemble name.

Unified Resource Manager actions for the ensemble are conducted from a single primary HMC. All other HMCs connected to the ensemble will be able to perform system management tasks (but not ensemble management tasks) for any CPC within the ensemble. The primary HMC can also be used to perform system management tasks on CPCs that are not part of the ensemble, such as Load, Activate, and so on.

Unified Resource Manager considerations:

► All HMCs at the supported code level are eligible to create an ensemble. Only HMCs FC0090 and FC0091 are capable of being primary or alternate HMCs.

► There is a single HMC pair managing the ensemble: primary HMC and alternate HMC.

► Only one primary HMC manages an ensemble, which can consists of a maximum of eight CPCs.

► The HMC that performed the "Create Ensemble" wizard becomes the primary HMC. An alternate HMC is elected and paired with the primary.

► **Primary Hardware Management Console (Version 2.11.0)** and **Alternate Hardware Management Console (Version 2.11.0)** will appear on the HMC banner. When the ensemble is deleted, the titles will resort to default.

► A primary HMC is the only HMC that can perform ensemble-related management tasks (create virtual server, manage virtual networks, create workload, and so on)

► Any HMC can manage up to 100 CPCs. The primary HMC can perform all non-ensemble HMC functions on CPCs that are not members of the ensemble.

► The primary and alternate HMCs *must be on the same LAN segment* (Managed SEs not required)

► The alternate HMC's role is to mirror ensemble configuration and policy information from the primary HMC.

► When failover happens, the alternate HMC will become the primary HMC. This behavior is the same as the current primary and alternate Support Elements

► The alternate HMC has a limited set of tasks. All ensemble functions need to be performed from the primary.

For more information regarding the Unified Resource Manager refer to Chapter 13, "Unified Resource Manager" on page 363.

HMC browser session to a primary HMC

A remote HMC browser session to the primary HMC that is managing an ensemble allows a user currently logged onto another HMC or a workstation to perform ensemble-related actions.

13

Unified Resource Manager

This chapter describes the objectives and composition of the IBM zEnterprise Unified Resource Manager as well as some detail of its implementation as seen from the user interfaces.

The zEnterprise Unified Resource Manager enables management of an **ensemble;** a collection of one or more zEnterprise **nodes** in which each node comprises a zEnterprise CPC (z196) and its optional attached IBM zEnterprise BladeCenter Extension (zBX).

The IBM zEnterprise System (zEnterprise) is a workload optimized technology system that delivers a multi-platform, integrated hardware system; spanning the mainframe, UNIX, and in the future, x86 technologies.

A heterogeneous workload may span multiple platform infrastructures. If the Unified Resource Manager has been assigned the resources required, it has the capabilities to fulfill workload policy objectives. The most cost effective way to fulfill processing requirements is to provide virtualized resources. Through virtualization, the physical resources can be shared among multiple workloads as they likely have different policies with different objectives. The Unified Resource Manager's goal is to fulfill the objectives of the policies in the most optimal and efficient way.

The objective of the Unified Resource Manager is to ensure the workloads executing on the zEnterprise System are treated according to specific workload objectives as expressed in workload policies.

The chapter contains the following sections:

13.1 Unified Resource Manager overview

The IBM zEnterprise System Unified Resource Manager can control an ensemble with up to eight nodes. A node is a z196 with an optionally attached zEnterprise BladeCenter Extension (zBX). Each zBX can have up to four Blade Center racks. Each rack can contain up to two BladeCenter chassis, each containing up to 14 blades.
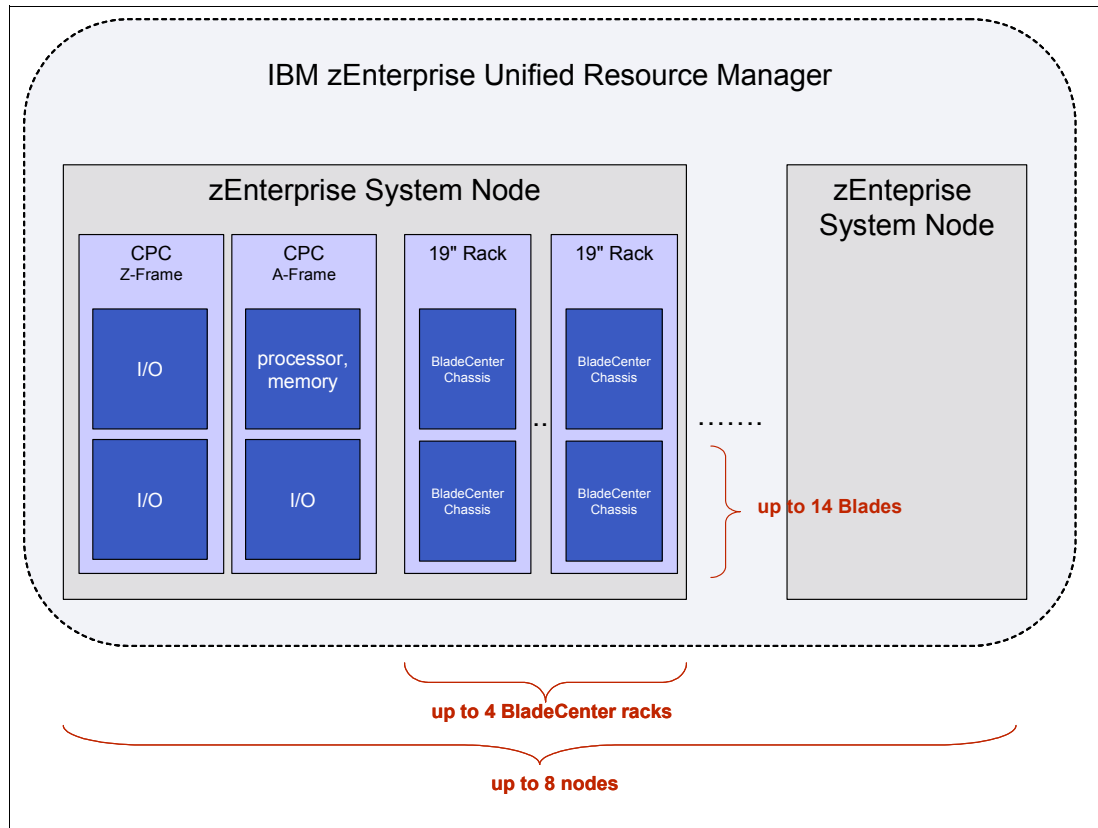
Figure 13-1   Unified Resource Manager topology

As an example, an zEnterprise System node is a complex of a z196 plus 112 blades under control of the Unified Resource Manager.

Through the Unified Resource Manager, the zEnterprise System is managed as an ensemble, a single pool of resources, integrating system and workload management across the multi-system, multi-tier, multi-architecture environment. Resource management actions for the entire ensemble are initiated and controlled through a designated Hardware Management Console (HMC) that is configured to be the primary HMC. This HMC must be configured as described in 13.3, "Ensemble Physical Resource Management".

13.1.1 Unified Resource Manager suites

zEnterprise System introduces new platform management functions with the Unified Resource Manager, operating from the Hardware Management Console. The functions are implemented in several operational *suites*, to provide:

► Integrated hardware management across all elements of the system, the z196 and the zBX, deliver:

– Licensed internal code (LIC) inventory, update, and service

- – Hardware and LIC problem detection, reporting, and call home
- – Field-guided repair and verification
- – Physical hardware configuration, backup, and restore
- – Primary/alternate replication and recovery for the HMC

► Fully automatic and coherent integrated resource discovery and inventory for all elements of the system without requiring user configuration, deployment of libraries or sensors, or user scheduling.

► Hypervisors are shipped, serviced, and deployed as System z LIC; booted automatically at power on reset and isolated on the internal platform management network. This way of packaging provides intrusion prevention, integrity, secure virtual switches with integrated configuration, monitoring, and problem management and reporting

► Virtual server lifecycle management, enabling directed and dynamic virtual server provisioning across all hypervisors from a single, uniform, point of control. This includes integrated storage and network configuration, and ensemble membership integrity.

► Representation of the physical and virtual resources that are used in the context of a deployed business function as a named workload. This provides a basis for defining a performance policy and enables performance monitoring, reporting, and resource optimization aligned with business-defined workload service levels. This function allows for automatic adjustment of processor resources across all hypervisors, and for workload balancing recommendations to be sent to network routers

Unified Resource Manager, working with the z196, the zBX infrastructure, and the attached blades provides for an end-to-end virtualized and managed environment. In addition it offers the ability to optimize technology deployment according to individual workload requirements. The Unified Resource Manager is delivered in two suites of tiered functionality - the Manage suite and the Automate suite.

The Manage suite

The Manage suite provides the following functionality:

► Operational Controls

► Virtual Server provisioning and management

- – Lifecycle Operations of Virtual Machines (create, modify, delete, query)
- – Run control (start, stop, suspend, resume)
- – Remote Console to support OS installation
- – Virtual Server Provisioning – CPU, Memory, I/O

► Virtual Network Management

- – Virtual Network Provisioning – Guest LANs, Virtual Switches, VLAN

► Hypervisor Management

- – Hypervisor firmware lifecycle management
- – Start, Stop, Backup/Restore

► Storage Virtualization Management

► Energy Controls

- – Automatically power off unused HW components (Reserved, not entitled, not supported, unused)

► Energy Monitoring

- – Average and maximum power, Input and exhaust air temperature
- – Humidity and heat load to water vs. air, air pressure
- – Query Max Potential Power

- – Enterprise-wide integration

► System Activity Display

- – Display CPU and I/O activity

► Default Workload Performance Context, Monitoring, and Reporting

- – Associate virtual servers with default workload, display default workload
- – Author a performance policy for default Workload
- – Display reporting/monitoring performance data for default workload
- – Monitor platform resources used to support default workload
- – View overall default workload performance health from a platform perspective
- – Display whether objectives defined in default workload performance policy are being achieved
- – Drill-down capability to help identify which virtual servers are contributing to performance problems

The Manage Suite provides for definition of a performance policy for the default workload, and the capability to monitor it against your defined policy objectives. While this might be sufficient for a single workload, it is unlikely that all workloads can be managed with only that level of capability. Therefore, the Automate Suite should be considered to differentiate between multiple workloads and mange them to separate business objectives.

The Automate suite

The Automate suite delivers the following functionality:

► Energy Management

- – Power Savings Mode
- – Power Capping and group capping
- – Hybrid-node-wide power capping

► Workload Performance Context, Monitoring, and Reporting

- – Define workloads, associate virtual servers with workloads, display workloads
- – Author a performance policy for a Workload
- – Display reporting/monitoring performance data for a workload
- – Monitor platform resources used to support a workload
- – View overall workload performance health from a platform perspective
- – Display whether objectives defined in workload performance policy are being achieved
- – Drill-down capability to help identify which virtual servers are contributing to performance problems

► Performance Management

- – Management of CPU resource across virtual servers hosted in the same hypervisor instance to achieve workload performance policy objectives

The Automate Suite expands on the capabilities of workloads and performance management. With the Automate Suite you can define your own custom workloads (by name) and the performance management capabilities are improved as well. By creating your own named workload definitions you can differentiate between multiple workloads in an ensemble.

13.2 How can I tell if my business will benefit?

When distilled to single central theme, the zEnterprise System ensemble has been created simply to optimize and better manage heterogeneous workloads. The ensemble components all contribute toward making it possible to define a workload and identify all the resources

needed by the workload to meet its performance objectives. By building a virtualized infrastructure across the heterogeneous platforms, it is possible to define uniform interfaces to monitor and manage the workload, bringing all the resources of the ensemble together to get the work done.

Workload in this context can be defined as a business function delivered through a collection of IT resources. These IT resources are often deployed across multiple servers based on multiple architectures, but communicating and cooperating to collect or accept requests and deliver business-value results. As various architectures have developed particular strengths, or as application developers have built particular strengths, components of the business function have been deployed over the various architectures. The nature of the infrastructure has thus become more heterogeneous, and each different platform has developed its own monitoring and management processes and policies. These monitoring and management processes often do not integrate seamlessly. As a result of that the workload is managed in a piecemeal manner with no end-to-end visibility of the workload as a single entity.

As an example, consider a common three tier architecture. This is often implemented for World Wide Web based applications, with a pool of http servers, forwarding requests to a pool of application servers, which then rely on a database "back-end" to store and retrieve information. The performance of the http server pool is often unknown to the application servers, even though the application servers have well defined operational objectives and monitoring functions. The database is probably carefully measured and monitored, but only as to how well it receives and responds to requests. The result is three carefully tuned and monitored islands of computing, but no management across the full breadth of the business function.

The zEnterprise System ensemble brings new focus to understanding the business function as a workload, across the islands that comprise it. Integrating heterogeneous platforms under a common management structure allows workloads to be defined by its components across multiple platforms. The ensemble defines all its components in a virtualized structure which allows the creation of consistent functions that monitor and manage the contributions of each component across the environment. With consistent functions across all the architectures you can now define your business objectives in terms of a policy that apply across all the platforms that comprise your business function. Application of performance policies across the ensemble then provides the opportunity for self-tuning of the workload based on the business objectives. When multiple workloads are running across the ensemble, each can be described with its own business objectives and importance allowing IT resources to be intelligently shared to accomplish the business objectives.

Creating this consistent view of workloads and then managing it is the fundamental principle behind the platform management functions of the zEnterprise System ensemble. In the subsequent sections we'll describe each of these functions in more detail.

As mentioned, multi-tier application architectures and their deployment on heterogeneous infrastructures are common today. The zEnterprise System presents a highly optimized environment for deploying these applications with improved monitoring and management. However, these architectures and deployments do not appear as a single pattern in business solutions today. We introduce a number of workload patterns to describe some of the common patterns, and maybe some that are a little less common. We expect that you will recognize some of these patterns in your own applications and see how they can benefit. Maybe in one of these you'll even find a great idea you can apply to your own business.

13.2.1 Mainframe workloads

It is impossible to discuss workloads that can benefit from the zEnterprise System without including the traditional mainframe workloads. The innovations in the z196 continue IBM's long standing history of continuous improvements in mainframe processing. IBM's well known transaction processing systems, such as CICS and IMS, can handle even greater volumes with the additional capabilities the z196 provides.

There are software solutions that leverage z/OS capabilities to enhance SOA architectures and provide a significant improvements for JAVA programs. With z/OS, z/TPF, z/VSE, z/VM, and Linux on System z there are plenty of opportunities to create powerful application solutions.

When you consider how to deploy multi-tier applications, do not forget that they can leverage the unmatched reliability and security of the zEnterprise 196. The z196 is ideal for data and transaction serving for mission critical applications.

13.2.2 Heterogeneous platform deployments

Although the zEnterprise System builds improvements on the traditional System z environment, that is only a part of the story. The introduction of the zBX means that new kinds of workloads will be running on the zEnterprise System. The following sections outline a few types of applications that this new environment is expected to benefit.

World Wide Web application - a three tier architecture

One deployment model that has become common with the explosive growth of the world Wide Web is a three tier web server application. This commonly involves an http server to present web pages to the user and accept their input. The http server identifies the application function requested by the user and sends the request to the next tier, the application server. The application server will accept the request from the http server, select the appropriate business logic, and begin processing the request. When the application server needs to retrieve data to supply the users request, or when it needs to store the information provided by the user, it calls on a data base server. The database server maintains the organized data structures required by the application and invokes the necessary storage (disk) requests to retrieve or store data as request by the application server. An example of such an architecture is pictured in Figure 13-2 on page 369.

This three tier architecture is often deployed with the http server outside a firewall to protect the customer's network from unwanted intrusions from the public internet. The http server may be implemented as a cluster of servers, to insure sufficient capacity for a large volume of requests and to maintain availability while components of the cluster are serviced. The http server then communicates through the firewall to the application servers. The application servers, also likely implemented as a cluster, communicate with the database server and that data flow might be encrypted or flow through another firewall, depending on data sensitivity and industry or government privacy regulations.
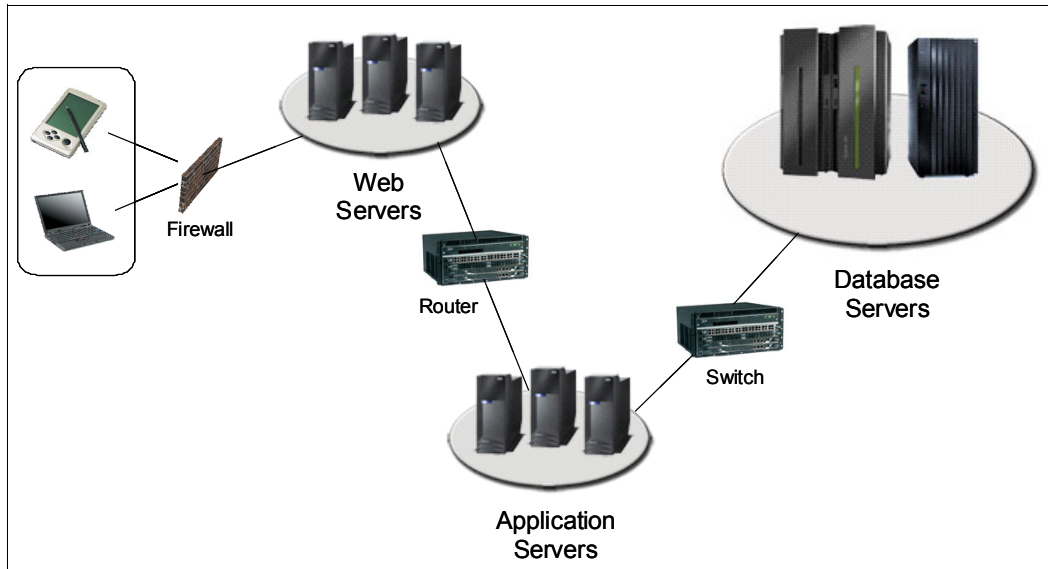
Figure 13-2   An example three-tier application architecture implementation

This application architecture may benefit significantly if deployed across the zEnterprise System architecture. For example, the http server can be virtualized and deployed across several blades in the zBX. The public "internet" communications could be isolated across one VLAN in the intraensemble data network (IEDN) (described in "Network Virtualization Management" on page 377) to which no other virtual servers are allowed access. VLAN isolation is considered to be as secure as physical isolation by many networking industry groups. The internal or "intranet" communications could then be directed to the application server cluster, again deployed in the zBX blades, via a separate VLAN. Lastly, the application server's communications with the database server, which might be a DB2 for z/OS running in the zEnterprise 196, also flow over the IEDN. Since the IEDN is privately managed in the zEnterprise ensemble and with no physical connections intervening between the servers that might be compromised, the application server and database communications might no longer need firewall or encryption protection measures.

All together, this application deployment may benefit from:

► High speed and bandwidth of 10 GbE IEDN
► Reduced number of network "hops" between servers
► Reduction in the number of external routers, switches, and firewalls by leveraging IEDN and VLANs
► Improved response by elimination of encryption (and decryption) overhead
► Monitoring and management of capacity requirement as workload fluctuates

Figure 13-3 on page 370 illustrates how this could be implemented with the zEnterprise System.
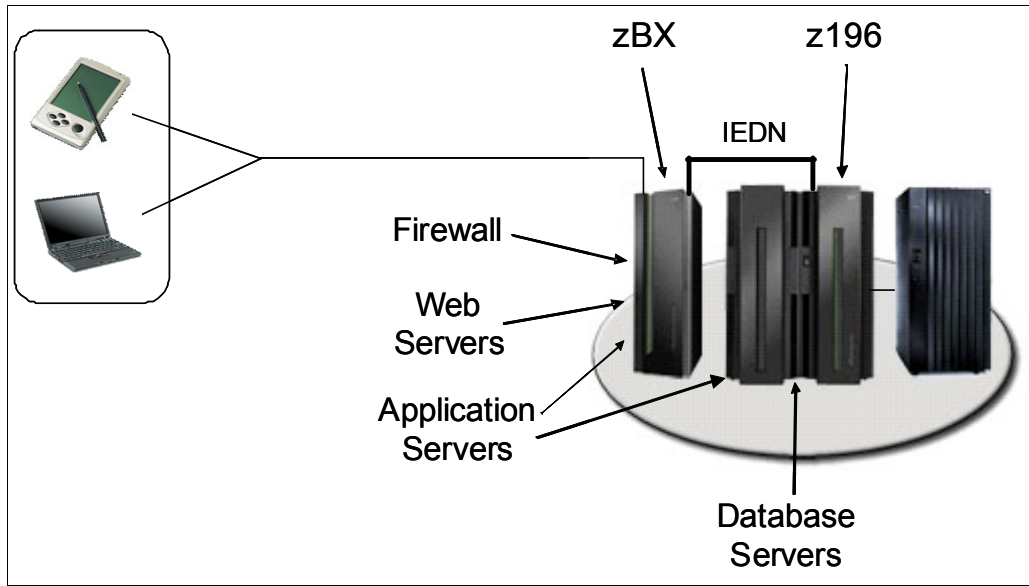
Figure 13-3   Three-tier application architecture on zEnterprise System

Optimizers—leveraging specialized processing capabilities

The evolution of information processing solutions has grown to include special purpose environments, tailored to perform some task very efficiently. Integrating these solutions into a broader application architecture may provide significant performance enhancements that allow business products that were not feasible in the past. Complex analysis tasks that took previous solutions far too long to complete on general purpose processing platforms might be achievable with special cooperative processing capabilities.

Another possibility for leveraging specialized processors is the IBM Smart Analytics Optimizer. It is delivered as a packaged solution that installs directly in the zBX. It provides a number of blades that are managed as an optimizer for DB2 for z/OS. With appropriate software features installed, DB2 can leverage the Smart Analytics Optimizer to build large data structures that can be queried very quickly, in parallel, to produce results much faster than traditional DB2 queries might perform. The Smart Analytics Optimizer does not require the user to build and install operating systems or software products on the blades, as the necessary operating system software is delivered as part of the zEnterprise System firmware. The DB2 specific software is installed and delivered as a feature of the mainframe DB2 and loaded to the blades directly by DB2. You can then select the appropriate data tables and constructs necessary to accelerate the performance of the queries that you desire using the Data Studio software. DB2 will then load the necessary data to the blades where it will compressed, indexed, and prepared for use. Then DB2, using its knowledge of the data loaded into the Smart Analytics Optimizer, will choose queries that will benefit from the performance of the optimizer and send those queries over for processing. When the results are complete they are received back into DB2 and returned to the requester. All of this is done with no change to the Structured Query Language (SQL) used already by your users and applications. If for some reason the Smart Analytics Optimizer is not available when DB2 receives an eligible request, it is handled internally by DB2 as a traditional SQL request. The IBM Smart Analytics Optimizer solution can make short work of many kinds of queries with almost no effort on your part.

Another example of such an environment might be a high performance computing grid. In a grid implementation, numerous independent but coordinated processors are given a part of a complex task and allowed to proceed through the necessary calculations independently. When all the components have completed, the coordinator assembles the final result and

returns it to the requestor. This application is actually very similar to the IBM Smart Analytics Optimizer design, but it can be applied more generally to solve other business problems. Application of this technology might involve investment portfolio analysis that provides recommendations for future investments based on recent market trends. Another might be biomedical research where the search for possible medicines and vaccines involves the testing of so many combinations that simply linear processing can not complete in a feasible timeframe. Such grid implementations can sometimes take advantage of the computing capacity that is built for peak periods and apply it to streamlining other processes during off-peak times. One example might be to assist overnight batch processing if your peak workload occurs during the daytime hours.

Both of these configurations, grid and Smart Analytics Optimizer, can gain from the zEnterprise System. A grid configuration might benefit significantly from:

- ► Fully virtualized environment simplifying grid member deployment
- ► Performance monitoring across the grid workload
- ► High speed and bandwidth of 10 GbE IEDN
- ► Reduced number of network "hops" between servers
- ► Reduction in the number of external routers, switches, and firewalls by leveraging IEDN and VLANs

When IBM Smart Analytics Optimizer is installed in the zBX Model 002 for attachment to the z196, it fits in quite well and gains from the additional capabilities of the zEnterprise System, such as the IEDN.

Virtualization and server consolidation

The zEnterprise System differentiates itself by including monitoring and management functions that are designed to get the most from the virtualized environment.

Because the zEnterprise System is built with virtualization as a central theme it presents an attractive target for server consolidation. This is a topic of current interest in many IT shops today. The past years of adding multiple distributed server platforms for every new application has created a kind of "server sprawl" which has contributed to power, cooling, and floorspace problems in many data centers. With so many separate servers running at very low average utilization rates, there exists a tremendous opportunity to consolidate servers onto fewer hardware instances through virtualization. Now, virtualization is certainly not a new idea, and there are several types of virtualization platforms available for implementation. The zEnterprise System is a versatile virtualization platform.

As has been stated in other parts of this book, the zEnterprise System is "all about the workload." But that doesn't mean a *single* workload. The features of the zEnterprise System are designed to take full advantage of the resources available to meet your business needs. System z operating systems have a history of managing disparate work within a single system using a set of business objectives, based on response times or throughput for example. This workload management is the inspiration for the similar functions included in the Unified Resource Manager suites. The Performance Management component includes the same kind of workload classification rules to define class of work by hostname, the virtual server's name, or other criteria. The business importance of the work and the performance expectations of the work can be specified in one of five categories: Highest, High, Medium, Low, and Lowest. The performance objectives can be specified in one of five categories: Fastest, Fast, Moderate, Slow, and Slowest. In addition, work can be classified as *discretionary* which means if there is any competition for resources then this work can be safely delayed.

By applying these classification rules to different workloads running under the same hypervisor, the Performance Management component can help allocate resources. When a

more important workload needs more processor resource the entitlement to CPU resources can be altered to move the available resources to the most important workload. Thus a hypervisor can change the allocation or processor resources when a workload is not meeting its objectives. This is all performed under the specifications of your policy, performed automatically and continuously by communication between the hypervisors and the management software. If workload importance changes based on time of day or day of week requirements, a scheduled operation can change the assigned policy to change the performance policy at your desired interval. You can see more details about these management functions in , "These operational advantages help to improve service and can also reduce cost by reducing management overhead through increased automation. This automation can be policy based to ensure the prioritization of resources according to preset business policies. Many of the security and resiliency strengths of the System z are applicable to parts of the workload which currently run independent of the mainframe."

Server consolidation with the zEnterprise System might benefit significantly from:

► Fully virtualized environment simplifying deployment
► Performance monitoring across the grid workload
► Performance policy based processor resource allocation
► High speed and bandwidth of 10Gb IEDN
► Reduced number of network "hops" between servers
► Reduction in the number of external routers, switches, and firewalls by leveraging IEDN and VLANs

As stated earlier, these capabilities were inspired by the robust workload management capabilities of z/OS. The implementations of these strategies to manage and balance workload based on work classification and business importance have a proven history in System z. Just as the capabilities for workload management and performance are continuously being improved to extend their capability in z/OS, you can expect that the zEnterprise System workload management capabilities will be extended over time to increase their value to your business. With this kind of management capability it makes server consolidation a strategy that can improve more than just the hardware costs of your servers -- it can actually make your workload perform better with less resources!

Operational advantages and Quality of Service

With the introduction of IBM zEnterprise System the qualities for which the System z is renowned, are extended to other components that are part of the ensemble. This provides support for mission critical workloads running on the heterogenous infrastructure of the ensemble. Improved Quality of Service (QoS) for the ensemble infrastructure is provided through reducing the number of hardware components and centralizing management. In addition to this simplification, all components of the zBX are duplicated for redundancy purposes and are part of the same (well known) servicing regime as other System z hardware components (see Chapter 13.3.2, "Serviceability" on page 375).

These operational advantages help to improve service and can also reduce cost by reducing management overhead through increased automation. This automation can be policy based to ensure the prioritization of resources according to preset business policies. Many of the security and resiliency strengths of the System z are applicable to parts of the workload which currently run independent of the mainframe.

13.3 Ensemble Physical Resource Management

In an ensemble, the Hardware Management Console (HMC) and Support Elements (SE) are appliances which together provide hardware platform management for System z. Hardware

platform management covers a complex set of setup, configuration, operation, monitoring, service management tasks and services that are essential to the use of the hardware platform product.

The HMC allows viewing and managing multi-node servers with virtualization, I/O networks, service networks, power subsystems, cluster connectivity infrastructure, and storage subsystems. The HMC has a global ensemble management function, whereas the SE has node management responsibility. When tasks are performed on the HMC, the commands are sent to one or more SEs, which then issue commands to their CPCs and zBXs.

13.3.1 HMC

An HMC must be at Version 2.11 to manage an ensemble and the workstations for these HMCs must be equipped with feature code 0090. The Unified Resource Manager suite ordered (Manage and Automate) determines the functions available on an HMC for a zEnterprise System. An ensemble is managed by a primary/alternate HMC pair. The alternate HMC cannot perform any HMC functions, since it is mirroring the primary HMC and must be ready to take over its functions in case of failure. These Unified Resource Manager feature codes must be ordered to equip an HMC to manage an ensemble:

▶ 0025—Ensemble Membership Flag
▶ 0019—Manage Firmware Suite
▶ 0020—Automate Firmware Suite (optional)

Figure 13-4 shows an HMC overview of an ensemble and defined workloads for this ensemble. The ensemble contains one member, PZBONZAI, consisting of one z196 CPC with two BladeCenter racks, each rack containing two BladeCenter chassis.

Figure 13-4   HMC ensemble table view

The HMC presents a highly interactive and dynamic web-based user interfaces. The HMC user interface views, management, and monitoring tasks provide everything needed for complete management of the Virtual Machine life cycle across the PR/SM, z/VM, and PowerVM hypervisors. From its inception all the way through monitoring, migration, and policy based administration during its deployment.

The following management activities are performed through the HMC interface:

► Ensemble Membership Management
► zBX Management
► Virtualization Management
► Workloads and Performance Management
► Energy Management
► Network Management
► Storage Management

The ensemble-specific managed objects include:

► Ensemble
► Members
► Blades
► BladeCenters
► Hypervisors
► Storage Resources
► Virtual Servers
► Workloads

In addition to the primary HMC, you have the option of using other HMCs as consoles attached to a zEnterprise node in an ensemble. When another HMC accesses a zEnterprise node in an ensemble, the HMC can do the same tasks as if the zEnterprise were not a part of an ensemble. Some of those tasks have been extended to allow you to configure some ensemble-specific properties (such as setting the virtual network associated with OSAs for a LPAR). Showing ensemble-related data in some tasks is allowed. Generally, if the data affects the operation of the ensemble, then the data is read-only on another HMC. The tasks that show ensemble-related data on another HMC are:

► Scheduled operations—displays ensemble introduced scheduled operations, but you can only view these scheduled operations.
► User role—shows ensemble tasks and you can modify and delete those roles.
► Event monitoring—displays ensemble-related events, but you cannot change or delete the event.

HMC availability

The HMC is attached to the same LAN as the server's support element (SE). This LAN is referred to as the *Customer Managed Management Network*. The HMC communicates with each Central Processor Complex (CPC), and optionally to one or more zEnterprise BladeCenter Extensions (zBXs), through the SE.

If the zEnterprise System server is not a member of an ensemble, it is operated and managed from one or more HMCs (just as any previous generation System z server). These HMCs are stateless (they do not keep any system status) and are therefore not affecting system operations when, if necessary, they are disconnected from the system. The system can (however not recommended) be managed from either SE.

However, if the zEnterprise System node is defined as a member of an ensemble, the primary HMC is the authoritative owning (stateful) component for Unified Resource Manager configuration and policies that have a scope that spans all of the managed CPCs/SEs in the

ensemble. It will no longer simply be a console/access point for configuration and policies that is owned by each of the managed CPC's SEs. The managing HMC has an active role in ongoing system monitoring and adjustment. This requires the HMC to be configured in an primary/alternate configuration and cannot be disconnected from the managed ensemble members.

> **Note:** The primary HMC and its alternate must be connected to the same subnetwork to allow the alternate HMC to take over the IP address of the primary HMC during failover processing.

Considerations for multiple HMCs

Customers often deployed multiple HMC instances to manage an overlapping collection of systems. Until the zEnterprise, all of the HMCs were peer consoles to the managed systems and all management actions are possible to any of the reachable systems while logged into a session on any of the HMCs (subject to access control). With the zEnterprise System Unified Resource Manager, this paradigm has changed. Only one primary alternate pair of HMCs (configured as described in 13.3.1) can manage ensembles. In this environment, if a zEnterprise System node has been added to an ensemble, management actions targeting that system can only be done from the managing (primary) HMC for that ensemble.

Remote HMC access

A remote HMC browser session to the HMC that is the ensemble-managing HMC for an ensemble allows a user currently logged onto another HMC or a workstation to perform ensemble-related actions.

13.3.2 Serviceability

The serviceability function for the components of the ensemble is delivered through the traditional HMC/SE constructs as for earlier System z servers. From a serviceability point of view all the components of the ensemble, including the zBX, are treated as System z features, similar to the treatment of I/O cards and other traditional System z features.

All the traditional functions for management of the components are delivered, including management of change, configuration, operations, and performance management. The zBX receives all of its serviceability and problem management through the HMC/SE infrastructure, and all service reporting, including call-home functions will be delivered in a similar fashion.

All physical zBX components are duplicated for redundancy purposes as dictated by System z QoS. The blades are standard blades provided by the customer, or by a solution, depending on the configuration.

There are several possibilities for blade deployment. Blades can be deployed as part of a solution, delivered by IBM, for example the IBM Smart Analytics Optimizer. In addition, blades can be acquired and deployed by the customer. For the latter solution, IBM provides a list of blade products that can participate in an ensemble.

For blades deployed in a solution configuration, the solution will handle the complete end-to-end management for these blades and their operating systems, middleware, and applications.

For blades deployed by the customer, the Unified Resource Manager will handle the blades as follows;

► The customer must have an entitlement for each blade in the configuration

► When the blade is deployed in the blade center chassis the Unified Resource Manager will power-up the blade, verify that there is an entitlement for the blade, and that the blade can participate in an ensemble. If these two conditions are not met, the Unified Resource Manager powers-down the blade

► The blade will be populated with necessary microcode and firmware

► The appropriate hypervisor will be loaded on the blade

► The management scope will be deployed according to which management enablement level is present in the configuration

► The administrator can now define the blade profile as well as the profiles for virtual servers to execute on the blade through the HMC.

► Based on the profile for individual virtual servers inside the deployed hypervisor, the virtual servers can be activated and an operating system can be IPLed following the activation. For customer deployed blades all of the application, data base, operating system, and network management will be handled by the customers' usual system management disciplines.

13.4 Virtualization management

The purpose of the virtualization management is to create and manage the virtualized resources in the ensemble. Functions to define the virtualization are necessary to create the virtual servers, the virtual network components, and virtual storage volumes. Some of the virtualization is provided by the hypervisors and some by the other parts of the ensemble. Functions to manage the hypervisors and other virtual resources are provided by the Unified Resource Manager firmware through the primary HMC.

13.4.1 Hypervisor management

A hypervisor is a virtual machine monitor program which allows multiple operating systems to execute on the same hardware simultaneously and sharing the hardware resources transparently across the operating systems.

The IBM Unified Resource Manager uses a set of hypervisors (PR/SM, z/VM, and PowerVM) to support deployment of workloads on the different hardware platforms of an ensemble. These hypervisors provides consumable and consistent basic management functions for virtual servers. Hypervisor management tasks are provided by the firmware installed through the **Manage suite** described in "The Manage suite" on page 365.

The available functionality are:

► Deploy and initialize a blade hypervisor

► Start, stop, and query/list hypervisors

► Update and repair a blade hypervisor

► Monitor hypervisors and their resource use via the Monitors Dashboard

– CPU

– Memory consumption

► Create virtual switches

► Allow agents in the operating system of a virtual server to communicate with a manager running in the hypervisor or the hypervisor management stack

13.4.2 Virtual Server management

A virtual server could be described as a container for the operating system required to support a given workload.

The hypervisor provides virtual resources to the server. When provisioning a virtual environment to support a workload, the relevant platform hypervisors will provision the virtual servers and their associated resources defined in "virtual server" definitions.

The functionality to define and manage the virtual server are provided by the Manage and Automate suites described in chapter, "The Automate suite delivers the following functionality:" on page 366. You describe a given workload and its associated resource requirements using the primary Hardware Management Console (HMC). Here you select (or define) the virtual servers required. The resources available to be provisioned to a virtual server are: what hypervisor type to be used, number of processors, size of memory, assign network devices, assign storage devices and "boot" options for the operating system.

Virtual server life-cycle management, enabling directed and dynamic virtual server provisioning across all hypervisors, are done from a single, uniform, point of control and include integrated storage and network configuration and ensemble membership integrity. The available functionality to support the life-cycle management are:

► List server

► Create/Delete server

► Start/Stop server

► View/Modify configuration of server

   – CPU – virtual, shared, dedicated; share (initial/minimum/maximum)

   – Memory – initial/defined

   – Network

   – Console – text/graphical

   – Storage

   – Virtual DVD

A workload policy has to be assigned to a virtual server. The purpose of the policy is to allow performance management routines to monitor performance against a defined policy. Additional capabilities provided by the **Automate suite** allow to provision or withdraw resources from the server in compliance with the workload policy and optimize the use of the physical resources. The creation and management of performance policies are described in "Defining an ensemble, a new Virtual Server and assigning workloads" on page 384.

13.4.3 Network Virtualization Management

Networking is a pervasive component of an ensemble. In fact, each node in an ensemble includes as many as five distinct networks, four of which are depicted in Figure 13-5 on page 378.
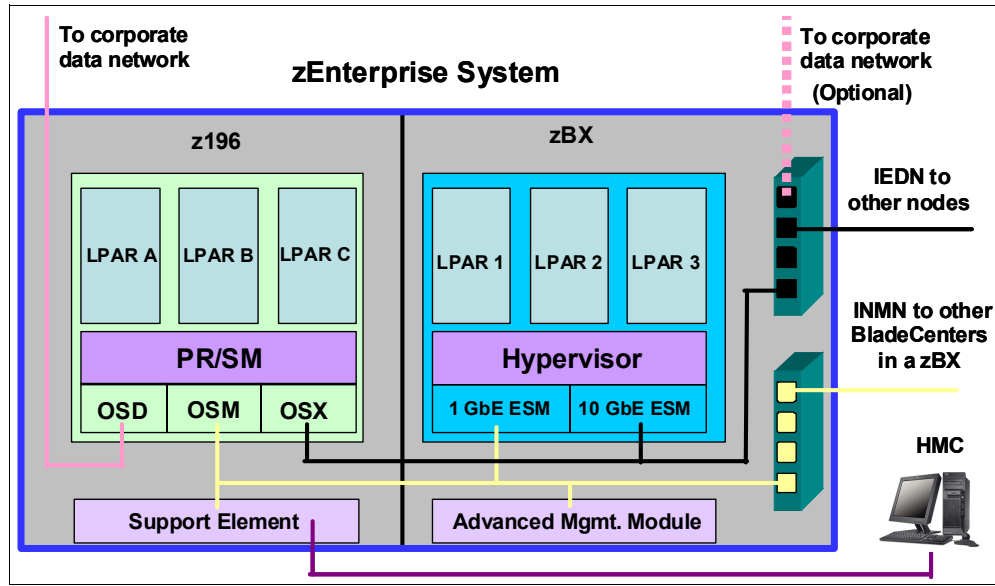
Figure 13-5   Networks contained within an ensemble.

1. Intranode management network (INMN): This private internal network provides the connections necessary to monitor and control components of the node, such as virtual servers or physical switches. The INMN connects to OSA adapters with CHPID type OSM in the z196 CPC, and to the 1Gb Ethernet Electronic Switch Module (ESM) in the zBX BladeCenters. This network requires user definition only for its connection to z/OS or z/VM LPARs.

2. Intraensemble data network (IEDN): This is the network for system and application data communications within the ensemble. It connects to OSA adapters with CHPID type OSX in the z196 CPC, and to the 10 GbE ESM in the zBX BladeCenter. This network connects all nodes, including z196 and zBX frames, together. This is the network that will be virtualized for the use of the virtual servers in the ensemble. Most of the remainder of this chapter relates to customization for the IEDN.

3. Customer management network: Also known as the HMC LAN, this network provides the communication link between Hardware Management Consoles (HMCs) and the nodes of the ensemble. It may connect to other System z machines that are not members of the ensemble. It may connect to support servers such as an Network Time Protocol (NTP) server that provides accurate time to the Server Time Protocol (STP) server in a System z processor. This network will be familiar to previous users of System z, and definitions on this network are unchanged from past System z implementations.

> **Note:** The primary HMC and its alternate must be connected to the same subnetwork to allow the alternate HMC to take over the IP address of the primary HMC during failover processing.

4. Customer managed data network: This network represents the existing enterprise data communication network. This network is attached to Open Systems Adapters (OSAs) such as OSD, in the z196 node, just as it has been attached to previous System z machines. In addition, this network may optionally be connected directly to the IEDN, depending our your configuration requirements.

As noted above, the IEDN requires user customization before it can be used. Network Virtualization Manager (NVM) provides the starting point for that customization. Customizing the IEDN will be the focus of the remainder of this section.

Network virtualization

The IEDN is the network used for application communications within an ensemble. It exists only within an ensemble, although it might also have a connection to the customer data network outside the ensemble. It is implemented as a flat layer-2 network -- which means that all the network interfaces can communicate directly with each other as if they were all connected to a single network switch. No routers are necessary to communicate across the IEDN. While there are physical network switches that are part of the IEDN, the appearance of a single network is maintained through virtualization.

The physical construction of the IEDN contributes to the security and reliability of the ensemble. All the network switches are inside the frames of the z196 and zBX frames and all network cables are point-to-point between the frames. With no intervening switches or routers the opportunity to compromise network integrity is greatly reduced. The switches are managed and configured only from the zEnterprise System firmware.

By virtualizing the network definitions it is possible to isolate the virtual servers from the physical definitions of the network interfaces and devices. This allows the virtual servers to be placed anywhere within the ensemble without changing the network definitions inside the virtual server. In addition, it isolates the virtual servers from "burned in" addresses on physical network interface cards which allows failed cards to be replaced without changing definitions. Finally, the network provisioning is based on the concept of virtuals LANs (VLANs) which provides for multiple logical networks to be defined over the same physical infrastructure. VLANs are a proven method for separating data traffic for multiple applications, as might be required for privacy rules, regulatory requirements, and even separation of production and test communications, all flowing over the same physical network. This virtualization helps you fully utilize the physical network capacity while still meeting your organization's security requirements.

There are four components of the network virtualization of the IEDN:

► VLAN (Virtual LAN) — A logical local area network that flows across the IEDN. A name and a numeric VLAN identifier are required to define a VLAN.

► VSWITCH (Virtual switch) — A virtual switch is a hypervisor component that provides virtualized network resources to a virtual server.

► VNIC (Virtual network interface card) — The VNIC is the network resource that a virtual server uses to access the IEDN. The VNIC is defined in the hypervisor through a VSWITCH.

► VMAC (Virtual media access control) — Virtual MAC addresses are assigned to VNICs. The VMAC replaces the manufacturer's "burned-in" MAC address on a physical network card.

The physical IEDN is connected to all the zBX Blade Centers in the ensemble, as well as all the z196 CPCs. Thus the physical network is shared by all members of the ensemble. A VLAN then provides a logical network on top of the physical IEDN where the virtual servers can connect using a virtual NIC (which has a virtual MAC address). All this virtualization is maintained by the ensemble management firmware cooperating with the hypervisors. The operating systems running in the virtual servers see the VNIC as a real network interface into a real network. They don't need to be aware of the virtualization, but are able to utilize the virtualized resources.

The network virtualization begins with defining a VLAN. Once the ensemble has been defined and the physical resources (z196 CPCs and zBXs) have been added to it, the VLANs can be defined.

As we said at the beginning of this section, networking is pervasive in an ensemble. The various parts of the virtualized network environment are connection points for the various resources of the ensemble. Hypervisors contain the VSWITCH definitions and VNICs that are contained in the VSWITCH. The virtual server and VNICs must be associated to the VLAN where they will connect. As you can see, Network Virtualization is not simply a task in the HMC. It becomes a part of the definition of most of the resources you will define in the ensemble.

> **Note:** Network virtualization manager (NVM) tasks are found on the primary HMC user interface (UI) under **Configuration**. This is an important point! While you will read about NVM and the tasks for which it is an umbrella, NVM is in fact just that, an umbrella term that does not appear anywhere in the primary HMC user interface.

Virtual server IP addressing

Once the virtualized network components have been defined, you will be able to complete networking definitions within the operating systems of the virtual servers. The virtual servers need IP addresses assigned just as they would as real servers on a real network. The IP addressing scheme is not defined in the network virtualization because that is a layer-3 function, which is built on top of the IEDN's layer-2 structure. It is your installation's responsibility to chose an IP addressing scheme appropriate for each of the VLANs on the IEDN. Either IPV4 or IPV6 addressing can be used, depending on the capability of the operating systems in the virtual servers. You should consult with your organization's network engineers for assistance with IP addressing for your environment.

> **Note:** The other networks internal to the z196 node (PSCN and INMN) do not require IP address assignment. They are completely internal to the z196, fully self-defined, and are not exposed to any outside network.

Connection to the existing customer data network

As mentioned above, in some cases in might be appropriate to connect the customer data network to the IEDN. The network configuration tasks allow specific ports on the TOR switches to be configured for attachment to your existing data network, external from the ensemble, and can impose restrictions on the attaching network. In configuring the switches for external connections you must consider whether to extend the VLANs out into the existing data network or keep them internal to the ensemble. If the IEDN switch port chosen for external connection is defined in trunk mode, then the VLAN-tagged data is passed through to the external network. By choosing access mode for the IEDN switch port it can restricted to a single VLAN from the IEDN. These decisions will depend on your network implementation and your network engineers should be consulted and involved before defining the external connection to the IEDN.

Network security

The networks (INMN and IEDN) that connect the z196 to the zBX are constructed with extreme security in mind. For details regarding INMN and IEDN network security refer to "Network security considerations with zBX" on page 198.

13.4.4 Storage Virtualization Management

With the new zEnterprise, additional storage connectivity requirements arise. Traditional FICON connectivity for System z workloads are unchanged; FCP connectivity to storage (SCSI) resources will be discussed in the following sections.

IBM Blades and storage resources

The following terms are used to describe how blades are connected to storage devices:

► A Port is a connection point for a Fibre Channel (FC) cable. A port is uniquely identified by its World-Wide Unique Port-name (WWPN). All suppliers of FC ports adheres to this WW standard.

► A Logical Unit is a virtual SCSI disk drive on a Storage Controller. The logical unit is addressed by WWPN/Logical Unit Number (LUN). It is possible to have multiple WWPN/LUN pairs.

► Access Control to logical units are done through zoning. Zoning is the concept used to define a logical grouping of host and storage controller ports. The definitions are stored in the fabric's name server.

Zoning

Only nodes within a zoning group can see/talk to each other. Zoning is done on a WWPN (or port) basis. Configuring zoning is done through the use of SAN management tools. At runtime, the zoning policies are kept in the switches and are valid for the whole fabric. See Figure 13-6. for a zoning example.
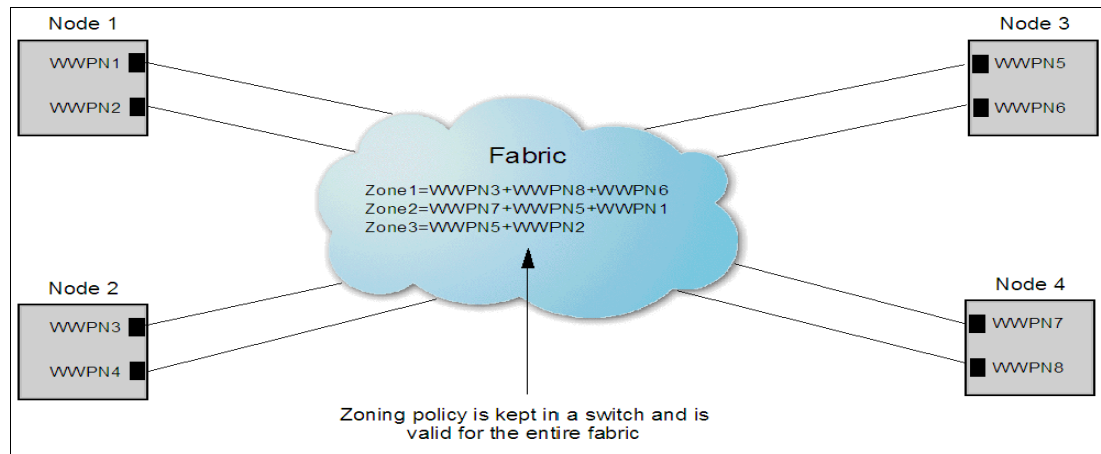
Figure 13-6   Zoning example

LUN masking

Access control to a LUN is enforced by the storage controller which perform LUN masking. LUN masking is done through Storage Controller management tools and is host port WWPN based. See Figure 13-7 on page 382 for details.
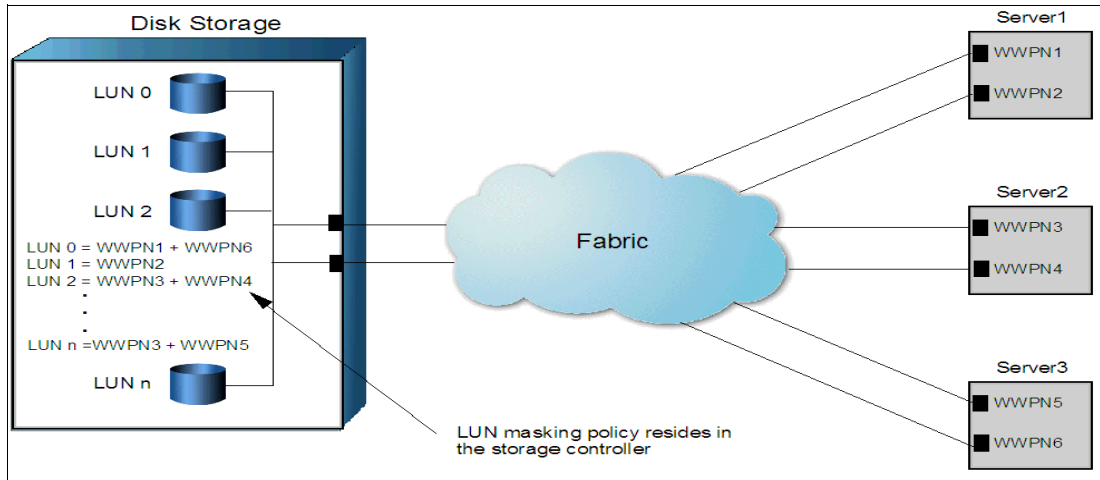
Figure 13-7   LUN masking example.

In this example, only WWPN1 and WWPN6 can access LUN 0. WWPN3 may not access LUN 3 (based on the illustrated rules).

The IBM blades and SAN considerations
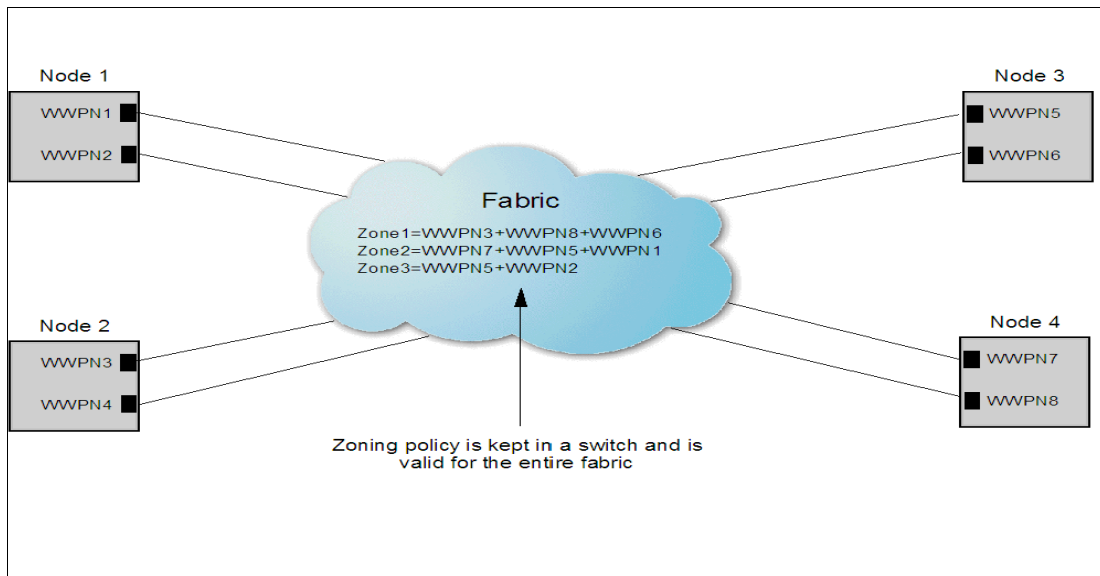
Figure 13-8   Fibre Channel topology overview for blades

Some considerations should be made when connecting the Blade Center (BC) to the SAN. Each blade comes pre-configured from the vendor and has an FC card with two ports. The FC switch in the BC must be set up in "passthru" mode to ensure interoperability with any SAN already installed. If this is not a desired mode, the switch has to be from the same vendor. The use of passthru mode will not add additional hops to customer's SAN. The SAN must therefore operate in "interoperability" mode.

The first SAN switch connected to the zBX must be a N_Port ID Virtualization (NPIV)-capable switch. The reason for this is the pre-NPIV Host Bus Adapter (HBA) sharing problem. To differentiate traffic sent from / to a single N_Port, multiple N_Port IDs would be required.This led to an extension of the Fibre Channel standards, called N_Port ID Virtualization (NPIV). NPIV allows a N_Port to do multiple fabric logins, using different WWPNs. The first login uses

FLOGI command and subsequent logins use FDISC command. WWPNs are created/maintained by firmware. A unique N_Port ID is assigned for each login. Frames sent from the N_Port can use different N_Port IDs.

The Storage Administrator role

The Storage Administrator allocates storage from physical storage pools to support an ensemble of virtual servers. The allocation is done based on input from the Server Administrator role (described below) who provide the number of LUNs/volumes*, host WWPNs, and so on. The Storage Admin define LUNs/volumes*, assigns ports to zones, masks LUNs and provide separate Storage Access Lists (SALs) for each hypervisor required to support the ensembles servers. A SAL contains the accessible storage resources for a Hypervisor. A storage resource may appear in many SALs. The Storage Admin configures the SAN and sets up replication.

*ECKD volumes (applies to z/VM) need to be added to IOCDS

The Server Administrator role

The Server Administrator provided the storage requirements of the ensemble's individual servers to the Storage Administrator.

The Server Administrator receives from the Storage Administrator, a SAL for each hypervisor type required to support the ensemble. The Server Administrator assigns LUNs or volumes[1] to the hypervisor's storage groups.
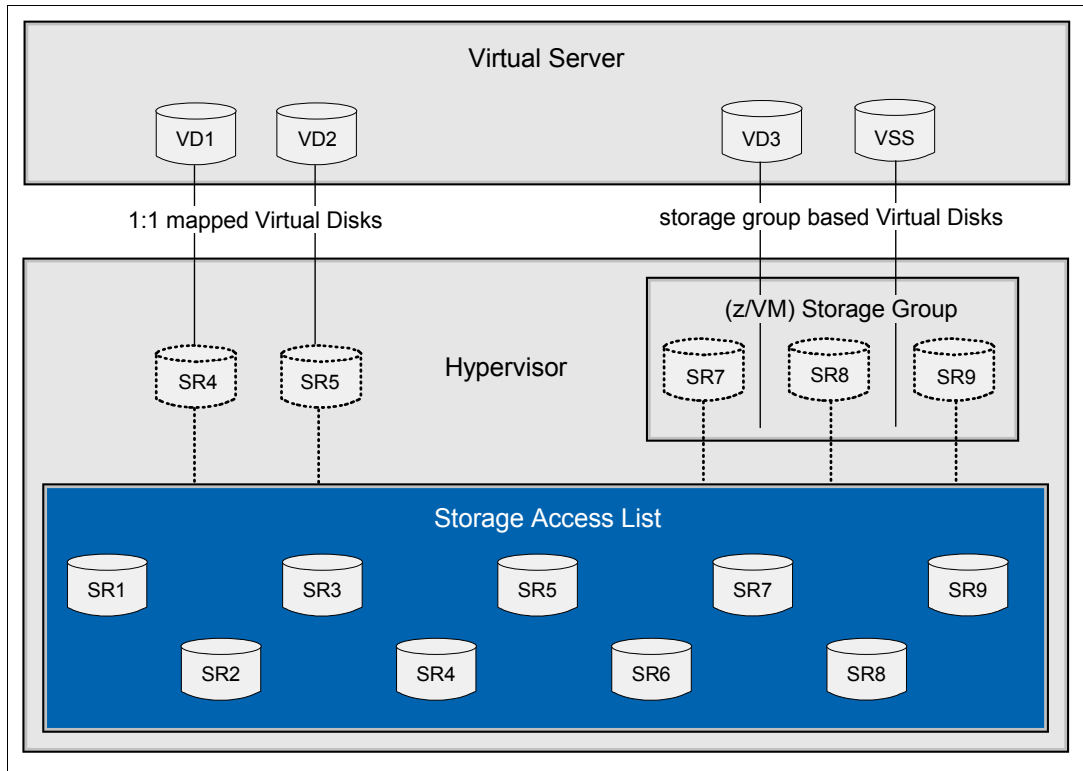
Figure 13-9   Storage virtualization

---

[1] Volumes refers to z/VM storage. Storage management for z/VM has not changed.

Storage Virtualization Manager (SVM)

Virtual servers only have virtual disks. Virtual disks may be either 1:1 mapped storage resources, or (for z/VM only) created from a minidisk storage group. Storage resources are either Fibre Channel Logical Units (to be used by blade and z/VM Hypervisors) or ECKD volumes (for z/VM only).

Configuring virtual disk to virtual servers

The Unified Resource Manager provides an abstraction from the underlying technologies. For SVM this means by providing a simplified storage management interface with common steps across different type of hypervisors. Currently, the SVM functionality is:

► Provide simplified and consistent storage management interface across different type of hypervisors

► Support existing z/VM storage management functions

► Very basic steps due to resource and calendar constraints

► Establish roles between server and storage admin

► Provide interfaces to manage blade storage

► Support for static migration

For OS images running in native LPARs, there are no change from current process (HCD, IOCDS, system specific files). This applies to z/OS, Linux on z, TPF, VSE and z/VM's system disks.

Supported Storage Devices

The IBM System Storage Interconnection Center (SSIC) Web site contains information on supported storage devices.

http://www-03.ibm.com/systems/z/hardware/connectivity/products/index.html

In addition, virtual DVD drives are supported:

► ISO image via file selection dialog box on HMC

  – HMC DVD/USB drive, or local file system

13.5 Ensemble performance management

Defining an ensemble, a new Virtual Server and assigning workloads

The following description is an example of what can be accomplished through the managing HMC functions and capabilities. A new ensemble is instantiated, new virtual servers are defined and assigned to a workload. This requires both the Unified Resource Manager 'Manage' and 'Automate' suites to be available. Those are orderable features of zEnterprise System.

Definition example

The HMC provides an Ensemble Management Guide that assists you with the tasks for setting up an ensemble by providing links to tasks and guidance information. A user ID with the proper role, such as Ensemble Administrator' needs to be used to perform the actions. Creating an ensemble starts with defining the name of the ensemble. The HMC being used to perform this task is the primary HMC. Also, you must assign an alternate HMC, which will make up a primary/alternate pair, for backup. An HMC can manage only one ensemble. The Add Member panel lists the discovered system(s) in the network and will mark the eligible

system(s). A functional ensemble must have at least one member, but it can have up to eight. Figure 13-4 on page 373 shows an example view of the defined ensemble My Ensemble with member PZBONZAI, which also contains four BladeCenters.

You now can use the New Virtual Server Wizard, or select a hypervisor from the Hypervisors tab of the ensemble, to create a new virtual server. After definitions have been made, the summary of such a definition looks like Figure 13-10:
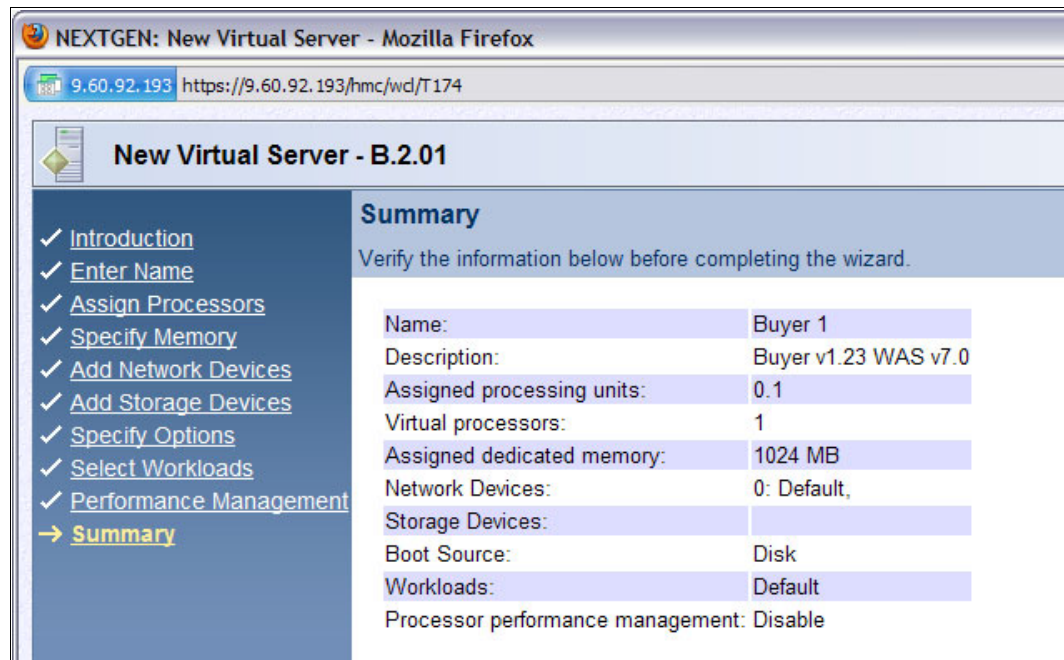
Figure 13-10   Summary of New Virtual Server definition

► The 'New Workload wizard' then guides you through:
  – Naming an categorizing the workload
  – Defining virtual servers which perform work
  – Creating performance policies to specify performance objectives
  – Creating service classes to prioritize and classify work within a policy
  – Activating a performance policy

Management example

With the 'Scheduled Operations' task you can schedule particular performance policy activations, such as policies for day- and night-shifts or for seasonal peaks.

The new 'Monitors Dashboard' (see Figure 12-4 on page 352) contains a link 'Open Workload Report'. This allows you to view performance characteristics, such as the processor usage of entitled blades, or processor usage of Virtual Servers. Workload Reports are available, which provide information such as met or missed objectives, Service Class Performance Index per Workload, or CPU utilization per Workload.

'Event Monitoring' allows you to set up triggers (with the option to send notification e-mails), for example, when a Service Class Performance Index is below a certain threshold.

For more information about setting up and managing ensembles refer to *zEnterprise Ensemble Planning and Configuring Guide GC27-2608.*

13.5.1 zEnterprise System performance management

In the zEnterprise System, the logical partitions (LPARs) running in the z196 are called virtual servers. The z/OS LPARs have the workload performance defined, monitored and adjusted through the z/OS Workload Manager (z/OS WLM), but that is only part of the picture. zEnterprise System performance management functions provide a more comprehensive performance management. It permits setting objectives and tracking the performance of any virtual server in the ensemble:

– The logical partitions (LPARs) running z/OS

– The z/VM guests running Linux

– The virtual servers running AIX on the POWER7 blades

Workloads running on the z/OS operating system are classified into service classes and have objectives defined to express how it should perform. These definitions are used by z/OS WLM to manage the work across all systems of a sysplex environment.

Extending this objective oriented concept to the zEnterprise System, hardware and software resources are also grouped into workloads and performance objectives are defined for them. Orchestration of autonomic management of resources across virtual servers is performed:

– Workload balancing

– Provide Intelligent Resource Director (IRD) like function across the zEnterprise System

– Leveraging of virtual server mobility to achieve performance objectives

zEnterprise System performance management includes the following capabilities:

– Performance monitoring and reporting functions.

– Virtual server CPU management provides the ability to manage CPU resources across virtual servers base on a objective-oriented performance policy.

– Cooperative management enables z/OS WLM to manage the z/OS segments of multi-tiered workloads. It considers the overall workload objectives.

– Load balancing recommendations enable the distribution of incoming work to best achieve the objectives defined in the workload performance policies.

13.5.2 Platform Workload definition

A platform workload is a grouping mechanism of virtual servers and accelerators supporting a business application. Each platform workload definition contains a given name, the associated virtual servers and one or more performance policies. It provides the context within which associate platform resources are presented, monitored, reported and managed. The management policies are associated to platform workload. Among the management policies, we have the performance policy.

Figure 13-11 on page 387 shows an example of distinct performance policies associated to two platform workloads (Payroll and HR).
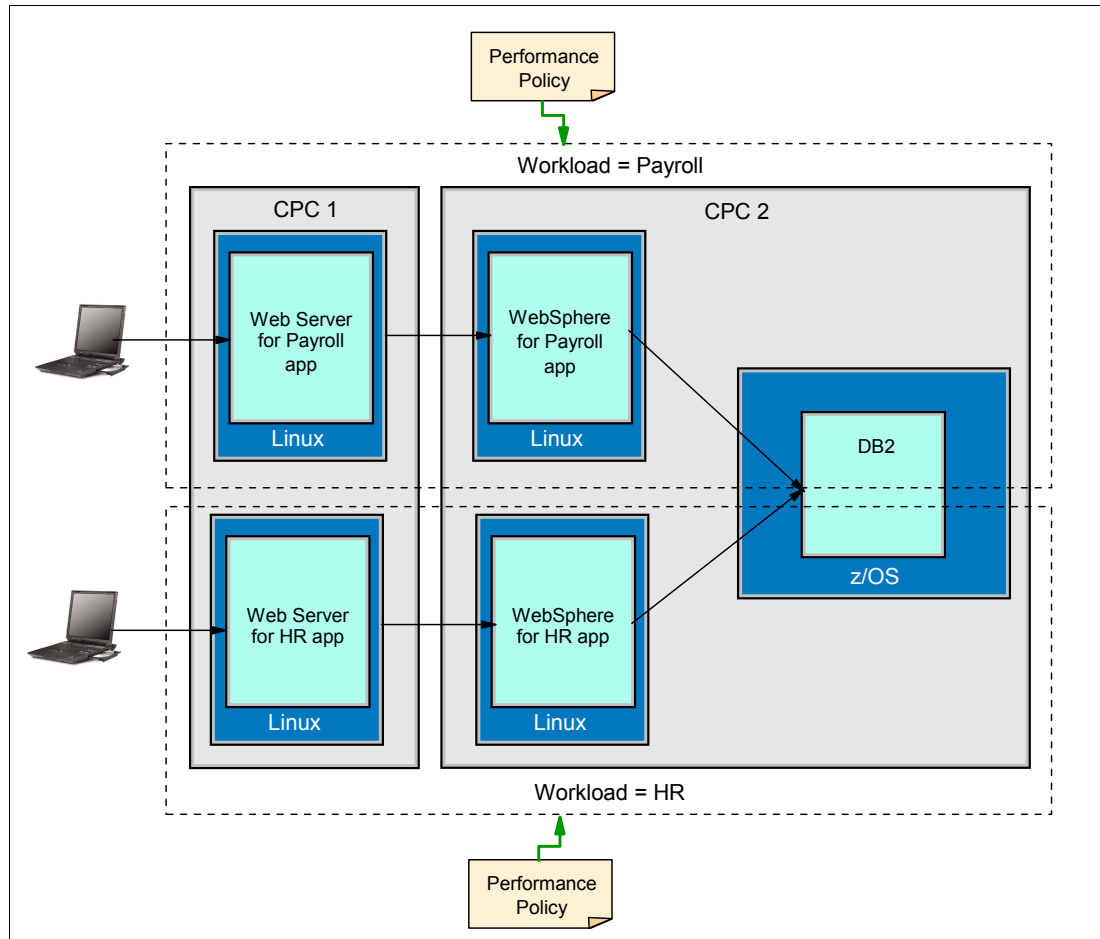
Figure 13-11   Platform Workload Definition

Workload performance policy defines performance objectives for virtual servers in workload. It provides basis for monitoring and management of platform resources used by virtual services in a workload. If different applications have the same performance objectives, it is possible to define only one workload that includes all virtual servers where those applications are going to run. If each application has different performance objectives, more than one workload must be defined.

> **Note:** A workload can have multiple performance policies associated with it that describe the performance objectives and importance, but only one policy can be active at a given time. It is possible to change the active policy dynamically via the user interface (UI) or through a time based schedule

When an ensemble is created, a default workload performance policy applies to all of the virtual servers within that ensemble. The workload performance policy structure contains a set of service classes; classification rules mapping each virtual server within the workload to a service class and the assignment of a service class to a performance objective and an importance value. To provide additional performance objectives besides the default one it is necessary to define one or more workloads for the ensemble.

Policy creation and editing is performed via ensemble Hardware Management Console (HMC). There is a wizard that helps in policy creation and editing. HMC provides a repository

for policies under development and saved policies. It also contains links to workload based performance reporting.

13.5.3 Performance monitoring and reporting

zEnterprise performance manager provides reporting capability that shows usage of platform resources in a workload context with an zEnterprise System scope. Performance monitoring and reporting functions primary objective is to provide data to check whether performance objectives are being met. If these objectives are not being achieved, detailed performance data helps identify the source of problems. Reporting is limited to platform level resources, not replacing tools that report on operating system resources and performance.

Additional customization permits more advanced performance functions, and collect more granular performance data to help identify and resolve performance problems. These advanced capabilities are provided through the use of guest platform management providers that are installed or started on virtual servers. The guest platform management provider (GPMP), is an example of additional customization. GPMP is an optional suite of applications that is installed in specific z/OS, Linux, and AIX operating system images to support platform management functions.

Performance data is available through a variety of reports. The data is collected over a time interval selected by the user. The most recent 36 hours of history is available. Granularity of data kept changes over time. For the most recent hour, 1 minute granularity is kept. After the first hour, granularity changes for 15 minutes. Reports provide different graphical views (topology, trending graphs). The types of performance reports include, but are not limited to:

► Workload report

   Provides a high level performance status and objective achievement information for all the workloads associated with the active performance policy. It contains an indication if a workload contains service class missing objectives. It reports the worst performing service class by performance index. Workload report example is available on Figure 13-12 on page 389.

► Service class report

   Provides a list of all the service classes defined in the workload performance policy. For each service class it includes the active performance policy, objective and importance definitions, indication when service class was part of resource adjustment action and actual performance data and delays. Service class details graphs the service class performance index and service class velocity.

► Virtual server report

   Provides data for virtual servers associated with the workload or a service class.The data includes allocated resources, resource utilization data, and delay statistics.

► Resource adjustment report

   Provides adjustment actions taken over report interval.

► Hypervisor Report

   Provides details about the virtual servers that are running in the same hypervisor instance, and how these virtual servers are competing for shared resources.

> **Note:** Besides setting the time interval for performance data collection, reports also permit setting an alert to notify when a specific type of performance issue arises. Alerts can be set for a workload, for a service class, or for a virtual server.

Workload monitoring report in Figure 13-12 on page 389 presents a graph of Performance Index (PI) of worst performing service class. It also permits to graph other service classes. This example also presents a bar graph of virtual server utilization.
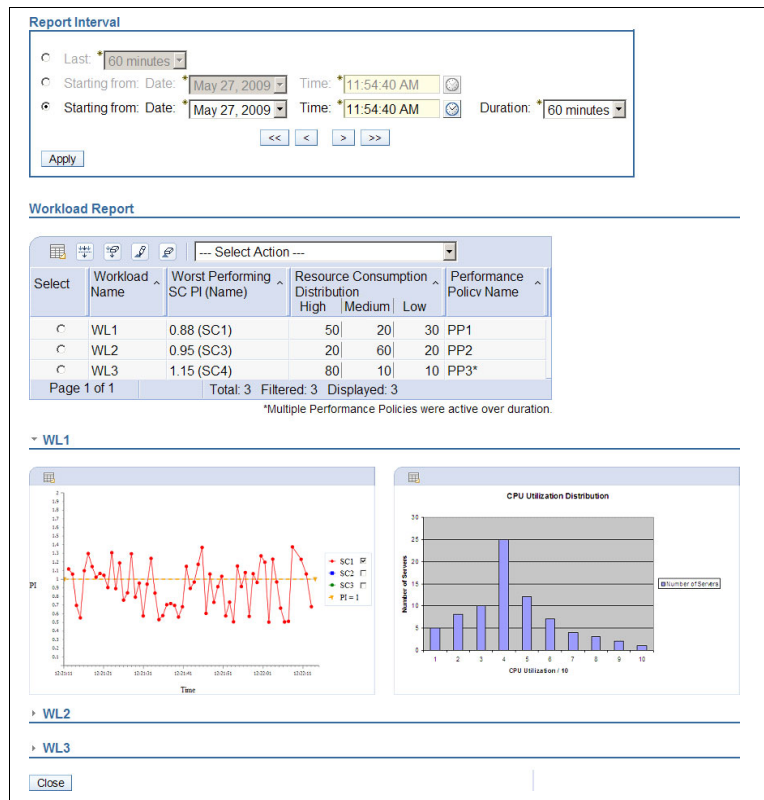
Figure 13-12   Workload Monitoring Report

zEnterprise performance manager provides APIs to allow data to be collected by other monitoring products.

13.5.4 Virtual server CPU management

zEnterprise System performance management functions similar to z/OS Intelligent Resource Director (IRD). On IRD, the scope of management is the set of z/OS LPARs on a single CPC within the same sysplex. zEnterprise System performance management also uses objective-oriented policies to manage CPU resources, but the scope of management is across virtual servers within the ensemble, extending the ability to manage CPU resources to other hypervisors.

Managing z/VM guests

Adjust the CPU allocation across guests with relative, not absolute, CPU shares. For each guest in an ensemble, it is necessary to decide whether z/VM Resource Manager (VMRM) or zEnterprise System performance management will manage the guest's CPU allocation, as both managers perform adjusts to guest CPU shares.

Managing virtual servers on a POWER7 blade

zEnterprise System performance manager raises and lowers entitled LPAR capacity to give or remove CPU resources to an virtual servers running AIX on a POWER7 blade based on objective-oriented policies achievement.

13.6 Energy monitoring and management

Energy monitoring and management can help better understand the power and cooling demand of the zEnterprise System, by providing complete monitoring and trending capabilities.

13.6.1 Multi-system energy monitoring and management

When a workload spans multiple infrastructures, attempting to understand the total energy utilization of all components supporting that workload can be challenging. To address this issue the energy monitoring features of zEnterprise have been extended to cover zBX devices. The IBM Unified Resource Manager has capabilities to monitor power consumption across the ensemble. To monitor is the first step to understand where and what resources are consumed.

13.6.2 The monitor dashboard

Through the Unified Resource Manager GUI on the HMC, the ensemble's power consumption can be closely monitored. At the following levels:

► Ensemble:

– Show Energy Management Information on the Details View

► CPC:

– Show Energy Management Information on the Details View
– Set Power Saving Mode

► BladeCenter:

– Show Energy Management Information on the Details View
– Set Power Saving Mode (for blades supporting this function)

► Blade server:

– Show Energy Management Information on the Details View
– Set Power Saving Mode (for blades supporting this function) Set Power Saving Mode (for blades supporting this function)

See Figure 13-13 on page 391 for details.
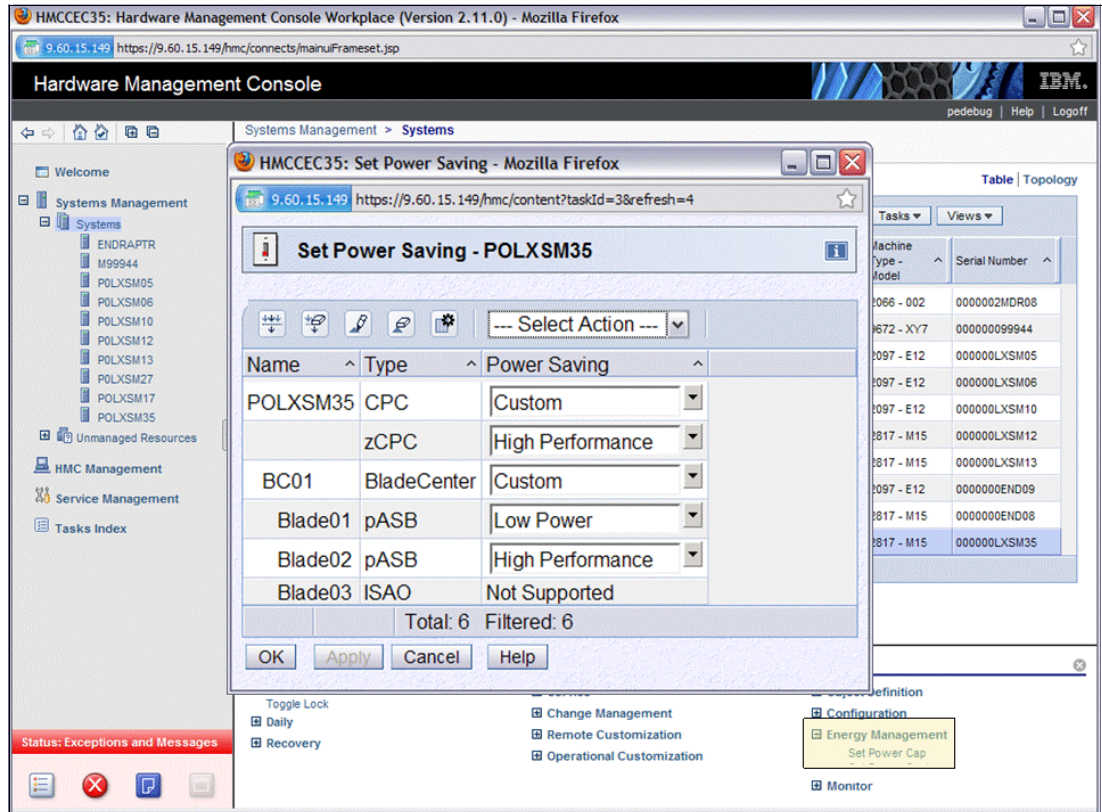
Figure 13-13   Example: Panel for Set Power Setting

A

Channel options

Table A-1 lists the attributes of the channel options supported on z196 servers, the required connector and cable types, the maximum unrepeated distance, and the bit rate.

At least one ESCON, FICON, ISC, or PSIFB feature is required.

Statement of Direction: z196 will be the last high-end server to offer ordering of ESCON channels. IBM intends not to offer ESCON channels on future servers

Table A-1   System z196 channel feature support

Channel feature	Feature codes	Bit rate	Connector	Cable type	Maximum unrepeated distance[a]
Enterprise Systems CONnection (ESCON)					
16-port ESCON	2323	200 Mbps	MT-RJ	MM 62.5 μm	3 km (800)
FIber Connection (FICON)					
FICON Express4 SX[b]	3322	4 Gbps	LC Duplex	MM 62.5 μm MM 50 μm	70 m (230) 380 m (1247) 150 m (492)
		2 Gbps	LC Duplex	MM 62.5 μm MM 50 μm	150 m (492) 500 m (1640) 300 m (984)
		1 Gbps	LC Duplex	MM 62.5 μm MM 50 μm	300 m (984) 860 m (2822) 500 m (1640)
FICON Express4[b] 4KM LX	3324	1, 2, or 4 Gbps	LC Duplex	SM 9 μm	4 km
FICON Express4[b] 10KM LX	3321	1, 2, or 4 Gbps	LC Duplex	SM 9 μm	10 km/20 km[c]

Channel feature	Feature codes	Bit rate	Connector	Cable type	Maximum unrepeated distance[a]
FICON Express8 SX	3326	8 Gbps	LC Duplex	MM 62.5 μm MM 50 μm	21 m (69) 150 m (492) 50 m (164)
		4 Gbps	LC Duplex	MM 62.5 μm MM 50 μm	70 m (230) 380 m (1247) 150 m (492)
		2 Gbps	LC Duplex	MM 62.5 μm MM 50 μm	150 m (492) 500 m (1640) 300 m (984)
FICON Express8 10KM LX	3325	2, 4, or 8 Gbps	LC Duplex	SM 9 μm	10 km
Open Systems Adapter (OSA)					
OSA-Express2 GbE LX[b]	3364	1 Gbps	LC Duplex	SM 9 μm	5 km
				MCP	550 m (500)
OSA-Express2 GbE SX[b]	3365	1 Gbps	LC Duplex	MM 62.5 μm	220 m (166) 275 m (200)
				MM 50 μm	550 m (500)
OSA-Express2 1000BASE-T Ethernet[b]	3366	10/100/1000	RJ45	UTP Cat5	100 m
OSA-Express3 GbE LX	3362	1 Gbps	LC Duplex	SM 9 μm	5 km
				MCP	550 m (500)
OSA-Express3 GbE SX	3363	1 Gbps	LC Duplex	MM 62.5 μm	220 m (166) 275 m (200)
				MM 50 μm	550 m (500)
OSA-Express3 1000BASE-T Ethernet	3367	10/100/1000	RJ45	UTP Cat5	100 m
OSA-Express3 10 GbE LR	3370	10 Gbps	LC Duplex	SM 9 μm	10 km
OSA-Express3 10 GbE SR	3371	10 Gbps	LC Duplex	MM 62.5 μm	33 m (200)
				MM 50 μm	300 m (2000) 82 m (500)
Parallel Sysplex					
IC	n/a		N/A	N/A	N/A
ISC-3 (peer mode)	0217 0218 0219	2 Gbps	LC Duplex	SM 9 μm MCP 50 μm	10 km/20 km 550 m (400)
ISC-3 (RPQ 8P2197 Peer mode at 1 Gbps)[c]		1 Gbps		SM 9 μm	20 km
PSIFB	0163	6 GBps	MPO	OM3 MM 50 μm	150 m
PSIFB LR	0168	5 Gbps	LC Duplex	SM 9 μm	10 km/100 km[d]

Channel feature	Feature codes	Bit rate	Connector	Cable type	Maximum unrepeated distance[a]
Cryptography					
Crypto Express3	0864	N/A	N/A	N/A	N/A

a. Minimum fiber bandwidth in MHz/km for multi-mode fiber optic links are included in parentheses were applicable.

b. Feature is only available if carried forward by an upgrade from a previous server.

c. RPQ 8P2197 enables the ordering of a different daughter card supporting 20 km unrepeated distance for 1 Gbps peer mode. RPQ 8P2262 is a requirement for that option, and other than the normal mode the channel increment is two, that is, both ports (FC 0219) at the card must be activated.

d. Up to 100 km at 2.5 Gbps, with repeater (System z qualified DWDM vendor product that supports 1x IB-SDR)

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this book.

IBM Redbooks

For information about ordering these publications, see "How to get Redbooks" on page 399. Note that some of the documents referenced here may be available in softcopy only.

► IBM zEnterprise System Configuration Setup, SG24-7834

► *Server Time Protocol Planning Guide*, SG24-7280

► *Server Time Protocol Implementation Guide*, SG24-7281

► *System Programmer's Guide to: Workload Manager*, SG24-6472

► *z/OS Intelligent Resource Director*, SG24-5952

► *Parallel Sysplex Application Considerations*, SG24-6523

► *Getting Started with InfiniBand on System z10 and System z9*, SG24-7539

► *IBM System z Connectivity Handbook*, SG24-5444

► *IBM BladeCenter Products and Technology*, SG24-7523

► *IBM Communication Controller for Linux on System z V1.2.1 Implementation Guide*, SG24-7223

► *DS8000 Performance Monitoring and Tuning*, SG24-7146

► *IBM System z10 Enterprise Class Capacity On Demand*, SG24-7504

► *IBM zEnterprise 196 Capacity on Demand User's Guide*, SC28-2605

► *Implementing IBM Systems Director Active Energy Manager 4.1.1*, SG24-7780

Other publications

These publications are also relevant as further information sources:

► *zEnterprise Ensemble Planning and Configuring Guide,* GC27-2608

► *Installation Manual for Physical Planning, 2817 All Models*, GC28-6897

► *zEnterprise 196 Processor Resource/Systems Manager Planning Guide,* SB10-7155

► *Coupling Facility Configuration Options*, GF22-5042

► *z/OS V1R9.0 XL C/C++ User's Guide*, SC09-4767

► *z/OS Planning for Workload License Charges*, SA22-7506

► *z/OS MVS Capacity Provisioning User's Guide*, SA33-8299

► *System z Capacity on Demand User's Guide*, SC28-6846

► *zEnterprise 196 Installation Manual for Physical Planning*, GC28-6897

► *System z HMC Operations Guide Version 2.11.0*, SC28-6895

- ► *Implementing IBM Systems Director Active Energy Manager 4.1.1,* SG24-7780
- ► *zBX Model 002 Installation Manual—Physical Planning*, GC27-2611
- ► *System z Application Programming Interfaces*, SB10-7030

Online resources

These Web sites are also relevant as further information sources:

- ► IBM Resource Link

  http://www.ibm.com/servers/resourcelink/

- ► IBM Communication Controller for Linux on System z

  http://www-01.ibm.com/software/network/ccl//

- ► FICON channel performance

  http://www.ibm.com/systems/z/connectivity/

- ► Rear Door Heat eXchanger

  http://www.rdhxinfo.com

- ► Materialized Query Tables (MQTs)

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

- ► Large Systems Performance Reference measurements

  http://www.ibm.com/developerworks/data/library/techarticle/dm-0509melnyk

- ► IBM zIIP

  http://www-03.ibm.com/systems/z/advantages/ziip/about.html

- ► Parallel Sysplex coupling facility configuration

  http://www.ibm.com/systems/z/advantages/pso/index.html

- ► Parallel Sysplex CFCC code levels

  http://www.ibm.com/systems/z/pso/cftable.html

- ► IBM InfiniBand

  http://www.infinibandta.org

- ► ESCON to FICON migration

  http://www-935.ibm.com/services/us/index.wss/offering/its/c337386u66547p02

- ► Optica Technologies Inc.

  http://www.opticatech.com/

- ► FICON channel performance

  http://www-03.ibm.com/systems/z/hardware/connectivity/ficon_performance.html

- ► z/OS deliverables on the Web

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

- ► LInux on System z

  http://www.ibm.com/developerworks/linux/linux390/

- ► ICSF versions and FMID cross-references

  http://www.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/TD103782

- z/OS

  `http://www.ibm.com/systems/support/z/zos/`

- z/VM

  `http://www.ibm.com/systems/support/z/zvm/`

- z/TPF

  `http://www.ibm.com/software/htp/tpf/pages/maint.htm`

- z/VSE

  `http://www.ibm.com/servers/eserver/zseries/zvse/support/preventive.html`

- Linux on System z

  `http://www.ibm.com/systems/z/os/linux/`

- IBM license charges on System z

  `http://www.ibm.com/servers/eserver/zseries/swprice/znalc.html`
  `http://www.ibm.com/servers/eserver/zseries/swprice/mwlc.html`
  `http://www.ibm.com/servers/eserver/zseries/swprice/zipla/`

How to get Redbooks

You can search for, view, or download Redbooks, Redpapers, Technotes, draft publications and Additional materials, as well as order hardcopy Redbooks publications, at this Web site:

ibm.com/redbooks

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

Index

To determine the spine width of a book, you divide the paper PPI into the number of pages in the book. An example is a 250 page book using Plainfield opaque 50# smooth which has a PPI of 526. Divided 250 by 526 which equals a spine width of .4752". In this case, you would use the .5" spine. Now select the Spine width for the book and hide the others: **Special>Conditional Text>Show/Hide>SpineSize(-->Hide;)>Set** . Move the changed Conditional text settings to all files in your book by opening the book file with the spine.fm still open and **File>Import>Formats** the Conditional Text Settings (ONLY!) to the book files.

IBM

Redbooks

IBM zEnterprise System Technical Guide

(2.0" spine)
2.0" <-> 2.498"
1052 <-> 1314 pages

IBM zEnterprise System Technical Guide

Redbooks

Explains virtualizing and managing the infrastructure for complex workloads

Describes the zEnterprise System and related features and functions

Discusses hardware and software capabilities

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

The zEnterprise System consists of the IBM zEnterprise 196 central processor complex, the IBM zEnterprise Unified Resource Manager, and the IBM zEnterprise BladeCenter Extension. The z196 is designed with improved scalability, performance, security, resiliency, availability, and virtualization. The z196 Model M80 provides up to 1.6 times the total system capacity of the z10 EC Model E64, and all z196 models provide up to twice the available memory of the z10 EC. The zBX infrastructure works with the z196 to enhance System z virtualization and management through an integrated hardware platform that spans mainframe and POWER7 technologies. Through the Unified Resource Manager, the zEnterprise System is managed as a single pool of resources, integrating system and workload management across the environment.

This book provides an overview of the zEnterprise System and its functions, features, and associated software support. Greater detail is offered in areas relevant to technical planning. This book is intended for systems engineers, consultants, planners, and anyone wanting to understand the zEnterprise System functions and plan for their usage. It is not intended as an introduction to mainframes. Readers are expected to be generally familiar with existing IBM System z technology and terminology.

INTERNATIONAL TECHNICAL SUPPORT ORGANIZATION

BUILDING TECHNICAL INFORMATION BASED ON PRACTICAL EXPERIENCE

IBM Redbooks are developed by the IBM International Technical Support Organization. Experts from IBM, Customers and Partners from around the world create timely technical information based on realistic scenarios. Specific recommendations are provided to help you implement IT solutions more effectively in your environment.

For more information:
ibm.com/redbooks

SG24-7833-00   ISBN