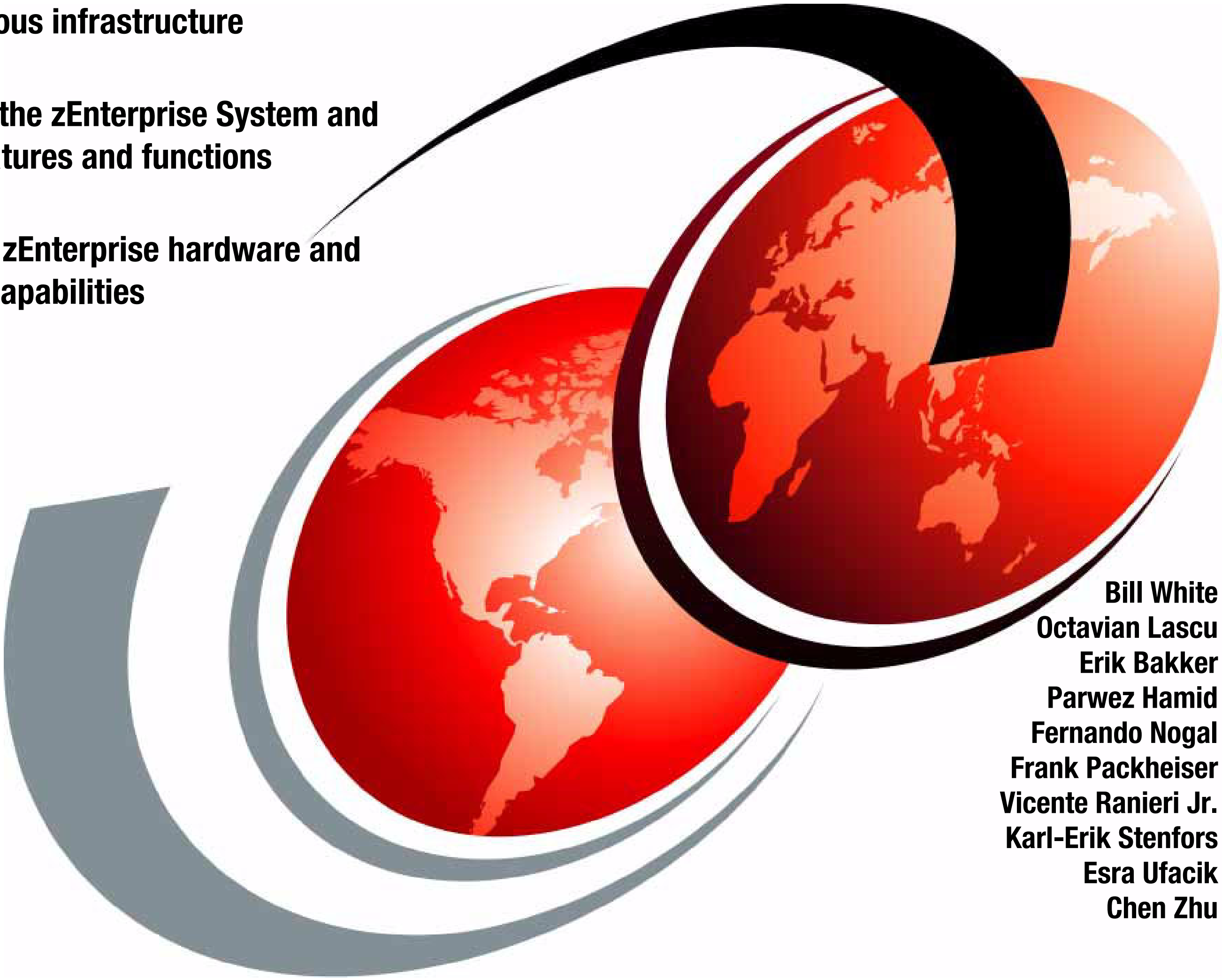
IBM zEnterprise 114 Technical Guide

Explains virtualizing and managing the heterogenous infrastructure

Describes the zEnterprise System and related features and functions

Discusses zEnterprise hardware and software capabilities



Bill White
Octavian Lascu
Erik Bakker
Parwez Hamid
Fernando Nogal
Frank Packheiser
Vicente Ranieri Jr.
Karl-Erik Stenfors
Esra Ufacik
Chen Zhu

Redbooks

ibm.com/redbooks

International Technical Support Organization

IBM zEnterprise 114 Technical Guide

September 2011

SG24-7954-00

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

First Edition (September 2011)

This edition applies to the IBM zEnterprise 114.

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

IBM may use or distribute any of the information you 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®
- AIX®
- BladeCenter®
- CICS®
- DataPower®
- DB2 Connect™
- DB2®
- Distributed Relational Database Architecture™
- Domino®
- DRDA®
- DS8000®
- ECKD™
- ESCON®
- FICON®
- FlashCopy®
- GDPS®
- Geographically Dispersed Parallel Sysplex™
- HACMP™
- HiperSockets™
- IBM Systems Director Active Energy Manager™
- IBM®
- IMS™
- Language Environment®
- Lotus®
- MQSeries®
- Parallel Sysplex®
- Passport Advantage®
- Power Systems™
- POWER6®
- POWER7™
- PowerHA™
- PowerPC®
- PowerVM™
- POWER®
- PR/SM™
- Processor Resource/Systems Manager™
- RACF®
- Redbooks®
- Redbooks (logo) ®
- Resource Link™
- Resource Measurement Facility™
- RETAIN®
- RMF™
- Sysplex Timer®
- System p®
- System Storage®
- System x®
- System z10®
- System z9®
- System z®
- Tivoli®
- WebSphere®
- z/Architecture®
- z/OS®
- z/VM®
- z/VSE™
- z10™
- z9®
- zSeries®

The following terms are trademarks of other companies:

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

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

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

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

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 in an 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 separate 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 114 central processor complex, the IBM zEnterprise Unified Resource Manager, and the IBM zEnterprise BladeCenter® Extension. The z114 is designed with improved scalability, performance, security, resiliency, availability, and virtualization. The z114 provides up to 18% improvement in uniprocessor speed and up to a 12% increase in total system capacity for z/OS®, z/VM®, and Linux on System z over the z10™ Business Class (BC).

The zBX infrastructure works with the z114 to enhance System z virtualization and management through an integrated hardware platform that spans mainframe, POWER7™, and System x technologies. The federated capacity from multiple architectures of the zEnterprise System is managed as a single pool of resources, integrating system and workload management across the environment through the Unified Resource Manager.

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.

Octavian Lascu is a Project Leader for System z hardware and Senior IT Specialist at the International Technical Support Organization, Poughkeepsie Center.

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 clients. For many years, he worked for Global Technology Services as a systems programmer providing implementation and consultancy services at many client sites. He currently provides pre-sales System z technical consultancy in support of large and small System z clients. His areas of expertise include Parallel Sysplex®, z/OS, and System z.

Parwez Hamid is an Executive IT Consultant working for the IBM 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 mainframe clients and spent much of his time introducing new technology. Currently, he provides pre-sales technical support for the IBM System z product portfolio and is the lead System z Technical Specialist for UK and Ireland. Parwez co-authored a number of IBM Redbooks publications. He prepares technical material for the worldwide announcement of System z servers. Parwez works closely with System z product development in Poughkeepsie and provides input and feedback for future product plans. Additionally, Parwez is a member of the IBM IT Specialist profession certification board in the UK and is also a Technical Staff member of the IBM UK Technical Council, which is made of senior technical specialists representing all of the IBM Client, Consulting, Services, and Product groups. Parwez teaches and presents at numerous IBM user group and IBM internal conferences.

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 clients. 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 clients 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 IBM Redbooks publications 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 client engagements and as a speaker at IBM and client events and trade shows.

Frank Packheiser is a Senior zIT Specialist at the Field Technical Sales Support office in Germany. He has 20 years of experience in zEnterprise, System z, zSeries®, and predecessor mainframe servers. He has worked for 10 years for the IBM education center in Germany, developing and providing professional training. He also provides professional services to System z and mainframe clients. He recently supported clients in Middle East/North Africa (MENA) for two years as a zIT Architect.

Vicente Ranieri Jr. is an Executive IT Specialist at the STG Advanced Technical Skills (ATS) team supporting System z in Latin America. He has more than 30 years of experience working for IBM. Ranieri is a member of the zChampions team, a worldwide IBM team to participate in the creation of System z technical roadmap and value proposition materials. Besides co-authoring several IBM Redbooks publications, 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 Ranieri is certified as a Distinguished IT Specialist by the Open Group. Vicente is a member of the Technology Leadership Council – Brazil (TLC-BR), and he is also a member of the IBM Academy of Technology.

Karl-Erik Stenfors is a Senior IT Specialist in the PSSC Customer Center in Montpellier, France. He has more than 42 years of working experience in the Mainframe environment, as a systems programmer, as a consultant with IBM clients, 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.

Esra Ufacik is a System z Client Technical Specialist working for Systems and Technology Group. She holds a B.Sc. degree in Electronics and Telecommunication Engineering from Istanbul Technical University. Her IT career started with a Turkish Bank as a z/OS Systems Programmer. Her responsibilities were maintaining a Parallel Sysplex environment with DB2® data sharing, planning and executing hardware migrations, and installing new operating system releases, middleware releases and system software maintenance, as well as generating performance reports for capacity planning. Esra joined IBM in 2007 as a Software Support Specialist in Integrated Technology Services where she works with mainframe clients within the country, acting as an account advocate. Esra was assigned as the Software Support Team Leader in ITS and got involved in several projects. Since 2010, she has been with STG, where her role covers pre-sales technical support, conducting System z competitive assessment studies, presenting the value of System z to various audiences, and assuring the technical feasibility of proposed solutions. Esra is also a guest lecturer for System z and large scale computing classes, which are given to undergraduate Computer Engineering students within the IBM Academic Initiative.

Chen Zhu is a Senior System Service Representative at the IBM Global Technology Services in Shanghai, China. He joined IBM in 1998 to support and maintain System z products for clients throughout China. Chen has been working in the Technical Support Group (TSG) providing second-level support to System z clients since 2005. 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

Gerard Laumay, Laurent Boudon
IBM Montpellier

IBM 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 email to:

  redbooks@us.ibm.com

- Mail your comments to:

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

Stay connected to IBM Redbooks publications

- Find us on Facebook:

  http://www.facebook.com/IBMRedbooks

- Follow us on Twitter:

  http://twitter.com/ibmredbooks

- Look for us on LinkedIn:

  http://www.linkedin.com/groups?home=&gid=2130806

- Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:

  https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm

- Stay current on recent Redbooks publications with RSS Feeds:

  http://www.redbooks.ibm.com/rss.html

1

Introducing the IBM zEnterprise 114

The IBM zEnterprise 114 is the second member in the zEnterprise family. Similarly to the z196, it was designed to help overcome problems in today's IT infrastructure 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 114 (z114) central processor complex (CPC) has the same newly designed quad-core chip as the z196, operating at a clock speed of 3.8 GHz. The z114 can be configured with up to 10 processors running concurrent production tasks with up to 256 GB (includes 8 GB for the HSA) of memory. It offers hot pluggable PCIe I/O drawers and I/O drawers, and continues the utilization of advanced technologies, such as PCIe and InfiniBand.

The z114 goes beyond previous designs while continuing to enhance the traditional mainframe qualities, delivering unprecedented performance and capacity growth. The z114 has a well-balanced general-purpose design that allows it to be 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 and the IBM WebSphere® DataPower® XI50 for zEnterprise.

Another key zEnterprise element is the zEnterprise Unified Resource Manager firmware. The zEnterprise 114 with or without a zBX attached, but managed by the Unified Resource Manager, constitutes a *node* in a zEnterprise *ensemble*. A 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 a combination of up to eight z114 and/or z196 systems and up to 896 blades (housed in eight zBXs, eight BladeCenters each). The ensemble has dedicated networks for management and data transfer along with Unified Resource Manager functions.

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

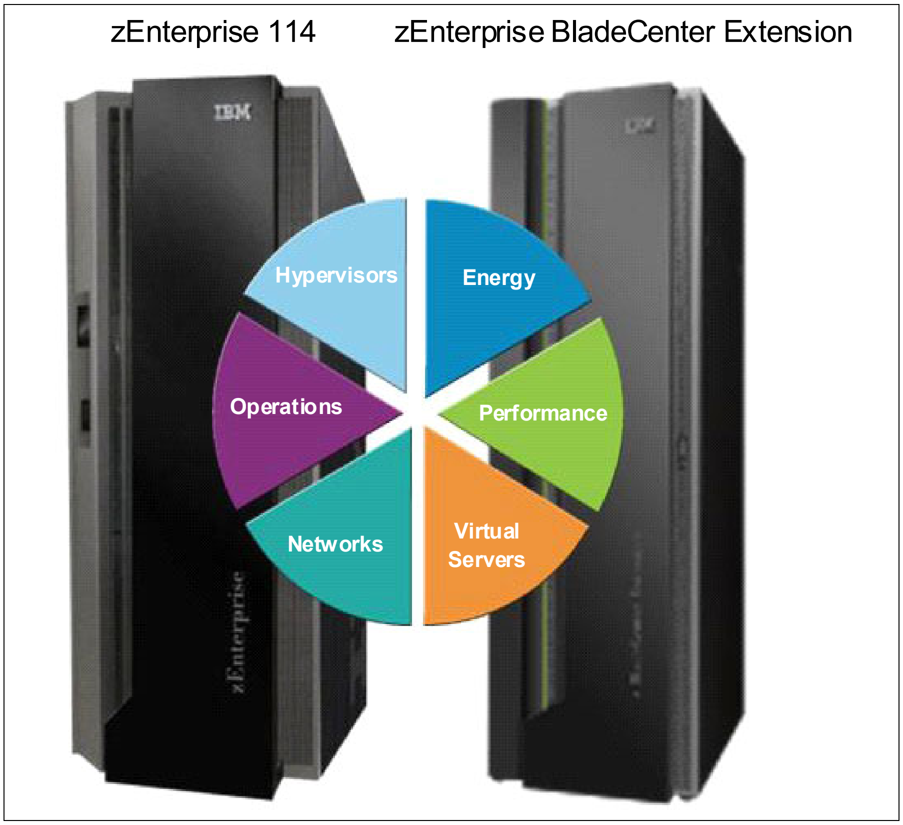


Figure 1-1 Elements of the IBM zEnterprise 114 Node

Recent decades have witnessed an explosion in applications, architectures, and platforms. A lot of experimentation has occurred in the marketplace. With the generalized availability of the Internet and the appearance of commodity hardware and software, various application 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 z114 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 that is needed to provide the high quality 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 management. Additional staff 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 various tiers is required to provide the robustness, scaling, consistent response time, and other characteristics that are 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 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 z114 continues to offer a wide range of subcapacity settings with 26 subcapacity levels for up to five central processors, giving a total of 130 distinct capacity settings in the system, and providing for a range of over 1:8 in processing power. The z114 delivers scalability and granularity to meet the needs of small to medium-sized enterprises, while also satisfying the mission-critical transaction and data processing requirements. The z114 continues to offer all the specialty engines that are available with System z10®.

The IBM holistic approach to System z design 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 114 highlights

The z114 CPC provides a record level of capacity over the previous mid-size System z servers. This capacity 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 114 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 z114. The server is built using an IBM single-chip modules design and processor drawers. Up to two processor drawers are supported per CPC. Each processor drawer contains three Single-Chip Modules (SCM), which host 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 in the marketplace. In addition, the system I/O buses take advantage of the InfiniBand technology, which is also exploited in coupling links, and PCIe technology. 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.

1.1.1 Models

The z114 has two model offerings ranging from one to ten configurable processor units (PUs), with a maximum of five CPs. Model M05 has one processor drawer and the model M10 has two. Each processor drawer houses, besides other components, two physical unit (PU) Single-Chip Modules (SCMs), one with four active cores, the other with three active cores. Model M10 is estimated to provide up to 12% more total system capacity than the largest z10 Business Class model, with a redesigned memory and additional PCIe-based I/O features. 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. z114 provides up to 26 subcapacity settings across a maximum of five CPs, offering 130 subcapacity setting options in total.

Depending on the model, the z114 can support from a minimum of 8 GB to a maximum of 248 GB of usable memory, with up to 128 GB per processor drawer. In addition, a fixed amount of 8 GB is reserved for the Hardware System Area (HSA) and is not part of customer-purchased memory. Memory is implemented as a Redundant Array of Independent Memory (RAIM). To exploit the RAIM function, up to 160 GB are installed per processor drawer, for a system total of 320 GB.

The z114 introduces the new PCIe I/O drawer. In addition, I/O drawers, which were introduced with the IBM z10 BC, are also supported. The I/O cages of previous System z servers are *not* supported. There are up to eight high-performance fanouts for data communications between the server and the peripheral environment. The multiple channel subsystems (CSS) architecture allows up to two CSSs, each with 256 channels. I/O constraint relief, using two 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 symmetric multiprocessor (SMP) system. It enables the configuration and operation of up to 30 logical partitions, which have processors, memory, and I/O resources assigned from the installed processor drawers. 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 the 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 (z114 supports 32 HiperSockets).

The z114 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 z114, further reduction of outages can be attained through improved nondisruptive replace, repair, and upgrade functions for memory, I/O drawers, and I/O adapters, as well as extending the nondisruptive capability to download Licensed Internal Code (LIC) updates.

Enhancements include, on the model M10, two dedicated processor spares, and removing preplanning requirements with the fixed 8 GB HSA. Clients 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 the following functions:

- Maximum configuration of 30 logical partitions (LPARs), two logical channel subsystems (LCSSs), and two managed software systems (MSSs)
- Dynamic addition and removal of a new LPAR to new or existing LCSSs
- Dynamic addition and removal of Crypto Express3 features
- Dynamic I/O enabled as a default
- Add or change the number of logical central processors (CPs), Integrated Facility for Linux (IFL) processors, Internal Coupling Facility (ICF) processors, System z Application Assist Processors (zAAPs), and System z Integrated Information Processors (zIIPs) per partition
- Dynamic LPAR physical unit (PU) assignment optimization for CPs, ICFs, IFLs, zAAPs, and zIIPs

1.1.2 Capacity on Demand

On-demand enhancements enable clients to have more flexibility in managing and administering their temporary capacity requirements. The z114 supports the architectural approach for temporary offerings that was introduced with z10, which has the potential to change the thinking about on-demand capacity. Within the z114, one or more flexible configuration definitions can be available to solve multiple temporary situations, and multiple capacity configurations can be active simultaneously.

Up to 200 staged records can be created for many 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 Capacity on Demand (CoD) either before or after execution.

1.2 zEnterprise 114 models

The z114 has a single frame, which is known as the *A frame*. The frame contains many components, including these components:

- Up to two processor drawers
- PCIe I/O drawers and I/O drawers
- Power supplies
- An optional internal battery feature (IBF)
- Cooling units (air cooling)
- Support elements

The zEnterprise 114 has a machine type of 2818. Two models are offered: M05 and M10. 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 Single-Chip Module (SCM) 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 the 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 two-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 cannot 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 maximum number of CPs for either of the two models is five.

The two-drawer (processor) system design provides an opportunity to increase the capacity of the system in three ways:

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

The following I/O features or channel types are supported:

- Enterprise Systems Connection (ESCON®)
- Fibre Channel connection (FICON) Express8S SX and 10 KM LX
- FICON Express8
- FICON Express4 (four port cards only)
- FICON Express4-2C
- Open Systems Adapter (OSA)-Express4S GbE LR and SR, GbE LX and SX
- OSA-Express3 1000BASE-T
- OSA-Express3-2P 1000BASE-T
- OSA-Express3 10 GbE LR and SR
- OSA-Express3 GbE LX and SX
- OSA-Express3-2P Gbe SX
- OSA-Express2
  (the OSA-Express2 10 GbE LR feature is not supported)
- Crypto Express3
- Crypto Express3-1P
- Coupling Links - peer mode only (InterSystem Channel (ISC)-3)
- Parallel Sysplex InfiniBand coupling link (IFB)

1.2.1 Model upgrade paths

A z114 Model M05 can be upgraded a model M10, and the model M10 can be upgraded to a z196 Model M15. All of these upgrades are disruptive (that is, the machine is unavailable during these upgrades). Any z9® BC or z10 BC model can be upgraded to any z114 model. Figure 1-1 on page 2 presents a diagram of the upgrade paths.

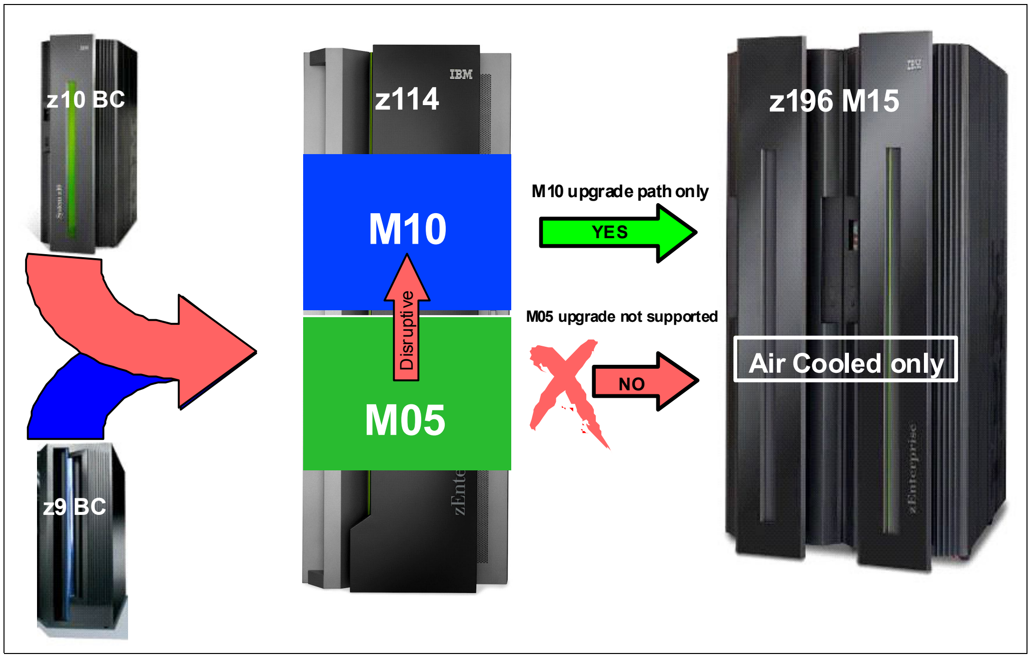


Figure 1-2 zEnterprise z114 upgrades

1.2.2 Concurrent processor unit conversions

The z114 supports concurrent conversion between various PU types, providing flexibility to meet changing business environments. CPs, IFLs, zAAPs, zIIPs, ICFs, or optional SAPs can be converted to CPs, IFLs, zAAPs, zIIPs, ICFs, or optional SAPs.

1.3 System functions and features

The z114 has a single frame. The frame contains the key components.

1.3.1 Overview

In this section, we introduce the functions and features of the system design.

Functions and features

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

- Single processor core sparing
- Large page (1 MB)
- Redundant 100 Mb Ethernet Service Network with virtual LAN (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 z114 provides the following benefits:

- Increased bandwidth between memory and I/O.
- Reduction in the impact of planned and unplanned server outages through these components and functions:
  - Hot-pluggable PCIe I/O drawers and I/O drawers
  - Redundant I/O interconnect
  - Concurrent PCIe fanout and Host Channel Adapter (HCA-O and HCA-C) fanout card hot-plug
  - Enhanced driver maintenance
- Up to two subchannel sets that are designed to allow improved device connectivity for Parallel Access Volumes (PAVs), Peer-to-Peer Remote Copy (PPRC) secondaries, and FlashCopy® devices; the second subchannel set allows the user to extend the amount of addressable external storage. The z114 allows you to IPL from subchannel set 1 (SS1), in addition to subchannel set 0.
- More capacity over native FICON channels for programs that process data sets, which exploit striping and compression (such as DB2, VSAM, Partitioned Data Set Extended (PDSE), Hierarchical File System (HFS), and zSeries File System (zFS)) by reducing channel, director, and control unit overhead when using the Modified Indirect Data Address Word (MIDAW) facility.
- Improved access to data for online transaction processing (OLTP) applications with High Performance FICON for System z (zHPF) on FICON Express8S, 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 Fibre Channel Protocol (FCP).

1.3.2 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 that is purchased.

Processor features

The z114 design employs processor drawers containing Singe-Chip Modules (SCM). Using two drawers provides for better capacity granularity. Each processor drawer houses two SCMs with processor chips and one SCM with a storage control chip. The processor chip has a quad-core design, with either three or four active cores, and operates at 3.8 GHz. One processor chip has four active cores and the other processor chip has three active cores.

The SCMs are interconnected with 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) system.

On the model M10, the PU configuration includes two spare PUs per CPC. Two system assist processors (SAPs) are available, regardless of the model. The remaining PUs can be characterized as central processors (CPs), with a maximum of five, Integrated Facility for Linux (IFL) processors, with a maximum of ten, System z Application Assist Processors (zAAPs), with a maximum of five, System z Integrated Information Processors (zIIPs), with a maximum of five, Internal Coupling Facility (ICF) processors, with a maximum of ten, or additional SAPs, with a maximum of two. 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 coprocessors right on the processor translate to high-speed cryptography for protecting data, 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

The z114 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 later support this capability, which increases flexibility and simplifies systems management. In a single LPAR, z/VM can manage guests that exploit Linux on System z on IFLs, z/VSE™, z/TPF, and z/OS on CPs, execute designated z/OS workloads, such as parts of DB2 Distributed Relational Database Architecture (DRDA®) processing and XML, on zIIPs, and provide an economical Java execution environment under z/OS on zAAPs.

1.3.3 Memory subsystem and topology

The z114 employs the memory technology that was introduced with the z196, which includes a buffered dual inline memory module (DIMM). For this purpose, IBM has developed a chip that controls communication with the PU and drives the address and control from DIMM to DIMM. The DIMM capacities are 4, 8, and 16 GB.

Memory topology provides the following benefits:

- Redundant array of independent memory (RAIM) for protection at the dynamic random access memory (DRAM), DIMM, and memory channel levels
- Maximum of 248 GB of user configurable memory with a maximum of 320 GB of physical memory (with a maximum of 248 GB configurable to a single logical partition)
- One memory port for each PU chip; up to two independent memory ports per processor drawer
- Asymmetrical memory size and DRAM technology across drawers
- 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 (8 GB) fixed-size HSA eliminates having to plan for an HSA

1.3.4 Processor drawer

This section highlights new characteristics in the processor drawer.

SCM technology

The z114 is built on a proven superscalar microprocessor architecture. In each processor drawer, there are three SCMs. Two SCMs have one PU chip each and one SCM has an SC chip. 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. The z114 is an air-cooled system using an evaporator/heat sink and a modular refrigeration unit (MRU) and air backup.

Out-of-order execution

The z114 has a superscalar microprocessor with out-of-order (OOO) execution to achieve faster throughput. With OOO, instructions might not execute in the original program order, although results are presented in the original order. For instance, OOO allows a few instructions to complete while another instruction is waiting. Up to three instructions can be decoded per cycle and up to five instructions can be executed per cycle.

PCIe fanout

The PCIe fanout provides the path for data between memory and the PCIe I/O cards using the PCIe 8 GBps bus. The PCIe fanout is hot-pluggable. In the event of an outage, a PCIe fanout can be concurrently repaired without loss of access to its associated I/O cards, using redundant I/O interconnect. Up to four PCIe fanouts are available per processor drawer.

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 the loss of access to its associated I/O cards, using redundant I/O interconnect. Up to four HCA fanouts are available per drawer.

1.3.5 I/O connectivity, PCIe, and InfiniBand

The z114 offers various improved features and exploits technologies, such as PCIe, InfiniBand, and Ethernet. In this section, we briefly review the most relevant I/O capabilities.

The z114 takes advantage of PCIe Generation 2 to implement the following features:

- An I/O bus, which implements the PCIe infrastructure. This approach is a preferred infrastructure and can be used alongside InfiniBand.
- PCIe fanouts, which provide 8 GBps connections to the PCIe I/O features.

The z114 takes advantage of InfiniBand to implement the following features:

- An I/O bus, which includes the InfiniBand infrastructure.

  This I/O bus replaces the self-timed interconnect bus that is found in System z servers prior to z9.
- Parallel Sysplex coupling using InfiniBand (IFB).

  This link has a bandwidth of 6 GBps between two zEnterprise CPCs, and between zEnterprise CPCs and z10 servers, and a bandwidth of 3 GBps between zEnterprise CPCs or z10 and System z9 servers.
- Host Channel Adapters for InfiniBand (HCA3) are designed to deliver up to 40% faster coupling link service times than HCA2.

1.3.6 I/O subsystems

The I/O subsystem draws on developments from z10, and it also includes a PCIe infrastructure. The I/O subsystem is supported by both a PCIe bus and an I/O bus similar to that of the z10 BC, and it includes the InfiniBand Double Data Rate (IB-DDR) infrastructure (replacing the self-timed interconnect that was found in the prior System z servers). This infrastructure is designed to reduce overhead and latency, and it provides increased throughput. The I/O expansion network uses the InfiniBand Link Layer (IB-2, Double Data Rate).

z114 also offers two I/O infrastructure elements for holding the I/O cards: PCIe I/O drawers, for PCIe cards, and I/O drawers, for non-PCIe cards.

PCIe I/O drawer

The PCIe I/O drawer, together with the PCIe I/O features, offers improved granularity and capacity over previous I/O infrastructures, and can be concurrently added and removed in the field, easing planning. A PCIe I/O drawer occupies one drawer slot, the same as an I/O drawer, yet it offers 32 I/O card slots. Only PCIe cards are supported, in any combination. Up to two PCIe I/O drawers are supported.

I/O drawer

I/O drawers provide increased I/O granularity and capacity flexibility over the I/O cages that were offered by previous System z servers, and I/O drawers can be concurrently added and removed in the field, which is another advantage over I/O cages. This design also eases planning. The z114 CPC can have up to two I/O drawers. I/O drawers were first offered with the z10 BC and can accommodate up to eight I/O features in any combination.

I/O features

The z114 supports the following I/O features, which can be installed in the PCIe I/O drawers:

- FICON Express8S SX and 10 KM LX (Fibre Channel connection)
- OSA-Express4S 10 GbE LR and SR, GbE LX and SX

The z114 also supports the following I/O features, which can be installed in both the I/O drawers and I/O cages:

- ESCON
- FICON Express8[1]
- FICON Express4[1] (four port cards only)
- FICON Express4-2C
- OSA-Express3 10GbE[1] LR and SR, and GbE LX and SX[1] (includes OSA-Express3-2P)
- OSA-Express3 1000BASE-T (includes OSA-Express3-2P)
- OSA-Express2 (except OSA-Express2 10 GbE LR)
- Crypto Express3
- Crypto Express3-1P
- InterSystem Channel (ISC)-3 coupling links

ESCON channels

The high-density ESCON feature (Feature code (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. With the z114, 240 channels are supported.

FICON channels

Up to 64 features with up to 128 FICON Express8S channels are supported. The FICON Express8S features support a link data rate of 2, 4, or 8 Gbps.

Up to 16 features with up to 64 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.

[1] Ordering this feature is determined by the fulfillment process.

The z114 supports FICON, High Performance FICON for System z (zHPF), channel-to-channel (CTC), and Fibre Channel Protocol (FCP).

The z114 continues to support, when carried forward, the FICON-Express4-2C features.

Open Systems Adapter

The z114 allows any mix of the supported Open Systems Adapter (OSA) Ethernet features, for up to 96 ports of LAN connectivity. For example, up to 48 OSA-Express4S features, in any combination, can be installed in the PCIe I/O drawer. For another example, up to 24 features of the OSA-Express3 or OSA-Express2 features are supported, with the exception of the OSA-Express2 10 GbE LR, which is not supported.

Each OSA-Express3 or OSA-Express2 feature that is installed in an I/O drawer reduces by two the number of OSA-Express4S features allowed.

OSM and OSX channel path identifier (CHPID) types

The z114 provides OSA-Express4S and OSA-Express3 CHPID types OSM and OSX for zBX connections:

- OSA-Express for Unified Resource Manager (OSM)

  Connectivity to the intranode management network (INMN). It connects the z114 to the zBX’s Bulk Power Hubs (BPHs) for the use of the Unified Resource Manager functions in the HMC. It uses OSA-Express3 1000BASE-T Ethernet exclusively.

- OSA-Express for zBX (OSX)

  Connectivity to the intraensemble data network (IEDN). It provides a data connection from the z114 to the zBX. It uses, preferably, OSA-Express4S 10 GbE or OSA-Express3 10 GbE.

OSA-Express4S and OSA-Express3 feature highlights

The z114 has four OSA-Express4S and seven OSA-Express3 features. When compared to similar OSA-Express2 features, which they replace, OSA-Express4S and 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 that is present in all OSA-Express4S and OSA-Express3 features. What previously was performed in the firmware, the OSA-Express4S and OSA-Express3 perform in the hardware. Additional logic in the IBM application-specific integrated circuit (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, which avoids a *hop* and is designed to reduce latency and to increase throughput for standard frames (1,492 byte) and jumbo frames (8,992 byte).

The z114 continues to support, when carried forward, the OSA-Express3-2P features.

For more information about the OSA features, refer to 4.8, “Connectivity” on page 122.

HiperSockets

The HiperSockets function, which is also known as internal queued direct input/output (internal QDIO or iQDIO), is an integrated function of the z114 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 z114 server. HiperSockets offers significant value in server consolidation by connecting many virtual servers, and it can be used instead of certain coupling link configurations in a Parallel Sysplex.

1.3.7 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 (FIPS 140-2 Level 4[2]). The crypto cards were enhanced with additional capabilities to add or move crypto coprocessors to logical partitions dynamically without pre-planning.

CP Assist for Cryptographic Function

The z114 implements the Common Cryptographic Architecture (CCA) in its cryptographic features. The CP Assist for Cryptographic Function (CPACF) offers the full complement of the Advanced Encryption Standard (AES) algorithm and Secure Hash Algorithm (SHA) along with the Data Encryption Standard (DES) algorithm. Support for CPACF is available through a group of instructions known as the Message-Security Assist (MSA). The z/OS Integrated Cryptographic Service Facility (ICSF) callable services and z90crypt device driver running at Linux on System z also invoke CPACF functions. ICSF is a base element of z/OS, and it can transparently use the available cryptographic functions, CPACF or Crypto Express3, to balance the workload and help address the bandwidth requirements of your applications.

CPACF must be explicitly enabled, using a no-charge enablement feature (FC 3863), except for the Secure Hash Algorithms (SHA), which are shipped enabled with each server.

The enhancements to CPACF are exclusive to the zEnterprise CPCs and they are 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. Each one can be configured as a coprocessor or as an accelerator:

- Crypto Express3 Coprocessor is for secure key encrypted transactions (default).
- Crypto Express3 Accelerator is for Secure Sockets Layer/Transport Layer Security (SSL/TLS) acceleration.

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

On z114, it is possible to have the Crypto Express3-1P feature, which has only one PCIe adapter. This adapter can also be configured as a coprocessor or as an accelerator.

The following functions have been recently added:

- Elliptic Curve Cryptography (ECC) Digital Signature Algorithm
- Elliptic Curve Diffie-Hellman (ECDH) algorithm
- PIN block decimalization table protection
- ANSI X9.8/ISO 9564 PIN security enhancements
- Enhanced Common Cryptographic Architecture (CCA), which is a key wrapping to comply with ANSI X9.24-1 key bundling requirements
- Expanded key support for AES algorithm
- Enhanced ANSI TR-31 interoperable secure key exchange
- Hardware support for long RSA keys (up to 4096-bit keys)
- Secure Keyed-Hash Message Authentication Code (HMAC)
- Remote loading of initial ATM keys
- PKA RSA Optimal Asymmetric Encryption Padding (OAEP) with SHA-256 algorithm
- Enhanced Driver Maintenance (EDM) and Concurrent Machine Change Level (MCL) apply

The z114 continues to support, when carried forward, the Crypto Express3-1P feature.

The configurable Crypto Express3 and Crypt Express3-1P features are 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 and the TKE 7.1 Licensed Internal Code (LIC) are optional features on the z114. The TKE workstation offers a security-rich solution for local and remote key management. It provides authorized personnel a method for key identification, exchange, separation, update, backup, and a secure hardware-based key loading for operational and master keys. TKE also provides secure management of host cryptographic module and host capabilities. Recent enhancements include support for the AES encryption algorithm operational keys, audit logging, and an infrastructure for payment card industry data security standard (PCIDSS), as well as these functions:

- ECC master key support
- CBC default settings support
- Use of elliptic curve Diffie-Hellman (ECDH) to derive shared secret
- TKE audit record upload configuration utility support
- New access control support for all TKE applications
- Single process for loading an entire key
- Single process for generating multiple key parts of the same type
- Decimalization table support
- Host cryptographic module status support
- Display of active IDs on the TKE console
- USB flash memory drive support
- Stronger PIN strength support
- Stronger password requirements for TKE passphrase user profile support
- Increased number of key parts on smart card

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

During a migration from a lower release of TKE to TKE 7.1 LIC, it will be necessary to add access control points to the existing roles. The new access control points can be added through the new Migrate Roles Utility or by manually updating each role through the Cryptographic Node Management Utility.

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 are protected by a user-defined personal identification number (PIN).

1.3.8 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 z114 does not support Integrated Cluster Bus (ICB)4 connectivity. Parallel Sysplex connectivity supports the following features:

- 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 z114, z196, and z10 servers and up to 3 GBps of bandwidth between z114, z196, or z10 servers and z9 servers for a distance up to 150 m (492 feet). With the introduction of a new type of InfiniBand coupling links (HCA3-O (12xIFB)), improved service times can be obtained.
- 1x InfiniBand up to 5 Gbps connection bandwidth between z114, z196, and z10 servers for a distance up to 10 km (6.2 miles). The new HCA3-O LR (1xIFB) type has doubled the number of links per fanout card, compared to type HCA2-O LR (1xIFB).

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 z114, with the following enhancements:

- Greater than 1,024 CF structures

  The limit has been increased to 2,047 structures. Greater than 1,024 CF structures requires a new version of the CFRM CDS:

  – All systems in the sysplex need to be running z/OS V1R12 or have the coexistence/preconditioning PTF installed.

  – Falling back to a previous level (without the coexistence PTF installed) is *not* supported without sysplex IPL.

- Greater than 32 connectors

  Limits have been increased, depending on the 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 higher than CF Level 17.

- Improved CFCC diagnostics and link diagnostics.
- The number of available CHPIDs 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 LIC 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 can be enabled for STP by installing the STP feature. You must enable STP for each server and CF that you plan to configure in a coordinated timing network (CTN).

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 z114, z196, z10, and z9. With this functionality, the z114, 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 z114, 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 various vendors that offer network timing solutions.

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

A System z114 cannot be connected to a Sysplex Timer. Preferably, migrate to an STP-only Coordinated Time Network (CTN) for existing environments. It is possible to have a System z114 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 (zBX)

The IBM zEnterprise BladeCenter Extension (zBX) is available as an optional machine to work along with the z114 server and consists of the following components:

- Up to four IBM 42U Enterprise racks.
- Up to eight BladeCenter chassis with up to 14 blades each.
- Blades, up to 112[3].
- Intranode management network (INMN) Top of Rack (TOR) switches. The INMN provides connectivity between the z114 Support Elements and the zBX, for management purposes.
- Intraensemble data network (IEDN) Top of Rack (TOR) switches. The IEDN is used for data paths between the z114 and the zBX, and the other ensemble members.

[3] The maximum number of blades varies according to the blade type and blade function.

- 8 Gbps Fibre Channel switch modules for connectivity to a SAN.
- Advanced Management Modules (AMMs) for monitoring and management functions for all the components in the BladeCenter.
- Power Distribution Units (PDUs) and cooling fans.
- Optional acoustic rear door or optional rear door heat exchanger.

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 the IBM judgement of the market's needs, additional specialized or general purpose blades might be introduced.

1.4.1 Blades

There are two types of blades that can be installed and operated in the IBM zEnterprise BladeCenter Extension (zBX):

- Optimizer blades:
  - IBM Smart Analytics Optimizer for DB2 for z/OS, V1.1 blades
  - IBM WebSphere DataPower Integration Appliance XI50 for zEnterprise blades
- IBM blades:
  - A selected subset of IBM POWER7 blades
  - A selected subset of IBM BladeCenter HX5 blades

These blades have been thoroughly tested to ensure compatibility and manageability in the IBM zEnterprise System environment.

IBM POWER7 blades are virtualized by PowerVM™ Enterprise Edition, and the virtual servers run the AIX® operating system. IBM BladeCenter HX5 blades are virtualized using an integrated hypervisor for System x® and the virtual servers run Linux on System x (Red Hat Enterprise Linux (RHEL) and SUSE Linux Enterprise Server (SLES) operating systems).

zEnterprise enablement for the blades is specified with an entitlement feature code to be configured on the zEnterprise CPCs.

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 sizes based on the amount of DB2 data (DB2 tables, number of indexes, and number of AQTs[4]) to be queried.

[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 seven to 56 blades that are housed in one to four dedicated BladeCenter chassis. One or two standard 19-inch IBM 42U Enterprise racks 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.

1.4.3 IBM WebSphere DataPower Integration Appliance XI50 for zEnterprise

The IBM WebSphere DataPower Integration Appliance XI50 for zEnterprise (DataPower XI50z) is a multifunctional appliance that can help provide multiple levels of XML optimization, streamline and secure valuable service-oriented architecture (SOA) applications, and provide drop-in integration for heterogeneous environments by enabling core enterprise service bus (ESB) functionality, including routing, bridging, transformation, and event handling. It can help to simplify, govern, and enhance the network security for XML and web services.

The zEnterprise BladeCenter Extension (zBX) is the new infrastructure for extending tried and true System z qualities of service and management capabilities across a set of integrated compute elements in the zEnterprise System.

When the DataPower XI50z is installed within the zEnterprise environment, Unified Resource Manager will provide integrated management for the appliance to simplify control and operations including the change management, energy monitoring, problem detection, problem reporting, and dispatching of an IBM System z service representative, as needed.

1.5 Unified Resource Manager

The zEnterprise Unified Resource Manager is the integrated management fabric that executes on the Hardware Management Console (HMC) and Support Element (SE). The Unified Resource Manager is comprised of six management areas (see Figure 1-1 on page 2):

- Operational controls (Operations)

  Includes extensive operational controls for various management functions

- Virtual server lifecycle management (Virtual servers)

  Enables directed and dynamic virtual server provisioning across hypervisors from a single uniform point of control

- Hypervisor management (Hypervisors)

  Enables the management of hypervisors and support for application deployment

- Energy management (Energy)

  Provides energy monitoring and management capabilities that can be used to better understand the power and cooling demands of the zEnterprise System

- Network management (Networks)

  Creates and manages virtual networks, including access control, which allows virtual servers to be connected together

- Workload Awareness and platform performance management (Performance)

  Manages CPU resource across virtual servers hosted in the same hypervisor instance to achieve workload performance policy objectives

The Unified Resource Manager provides energy monitoring and management, goal-oriented policy management, increased security, virtual networking, and data management for the physical and logical resources of a given ensemble.

1.6 Hardware Management Consoles and Support Elements

The Hardware Management Consoles (HMCs) and Support Elements (SEs) are appliances, which together provide hardware platform management for System z. The HMC is used to manage, monitor, and operate one or more zEnterprise CPCs, zBXs, 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.

1.7 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 objective 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 LIC and hardware design, failures are eliminated through rigorous design rules; design walk-through; peer reviews; element, subsystem, and system simulation; and extensive engineering and manufacturing testing.

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

1.8 Performance

z114 system resources are powered by up to 14 microprocessors running at 3.8 GHz and provide up to 18% uniprocessor performance improvement and up to 12% improvement in total system capacity for z/OS, z/VM, and Linux workloads on System z, as compared to the z10 BC. Figure 1-3 on page 20 shows the processor capacity settings for z114.

[5] From Version 2.11 and later. See 12.7, "HMC in an ensemble" on page 363.

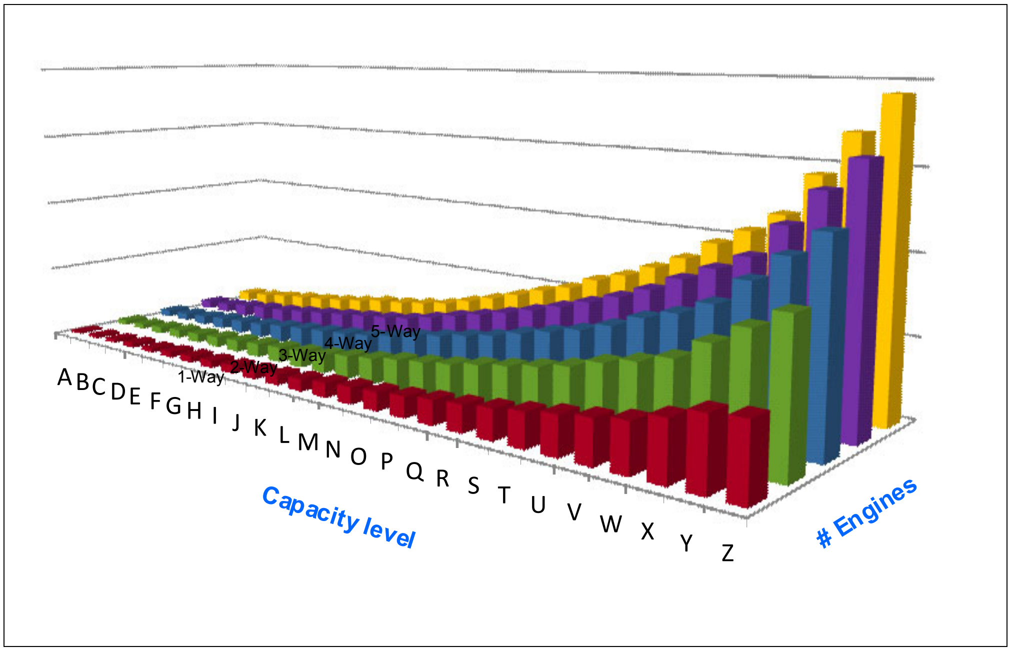


Figure 1-3 z114 processor capacity settings

Consult the Large System Performance Reference (LSPR) when you consider performance on the z114. The range of performance ratings across the individual LSPR workloads is likely to have a large spread. More performance variation of individual LPARs 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 more information, read 1.8.6, "Workload performance variation" on page 25.

For detailed performance information, see the LSPR website:

https://www-304.ibm.com/servers/resourcelink/lib03060.nsf/pages/lsprindex

The MSU ratings are available from the following website:

http://www-03.ibm.com/systems/z/resources/swprice/reference/exhibits/

1.8.1 LSPR workload suite

Historically, LSPR capacity tables, including pure workloads and mixes, have been identified with application names or a *software* characteristic. Examples are CICS®, IMS™, OLTP-T[6], CB-L[7], LoIO-mix[8] and TI-mix[9]. However, capacity performance is more closely associated with how a workload uses and interacts with a particular processor *hardware* design. With the availability of CPU measurement facility (MF) data on z10, the ability to gain insight into the interaction of workload and *hardware design* in production workloads has arrived. CPU MF data helps LSPR to adjust workload capacity curves based on the underlying hardware sensitivities, in particular, the processor access to caches and memory, which is known as *nest activity intensity.* With this nest activity intensity, the LSPR introduces three new workload capacity categories, which replace all prior primitives and mixes.

[6] Traditional online transaction processing workload (formerly known as IMS)
[7] Commercial batch with long-running jobs
[8] Low I/O Content Mix Workload
[9] Transaction Intensive Mix Workload

The LSPR contains the internal throughput rate ratios (ITRRs) for the zEnterprise CPCs 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.

1.8.2 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 factors.

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 vary for each transaction or job depending on the complexity of the tasks 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 might vary based on a number of factors:

- Competition with other tasks in the system for shared resources. As the total number of tasks grows, more instructions are needed to manage the resources.
- The *N*way (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:

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

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

Memory hierarchy and memory nest

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

- 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

Certain caches are “private” to the micro-processor, which means only that specific micro-processor can access them. Other caches are shared by multiple micro-processors. We 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. The best performance occurs when the instructions and data are found in the caches 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 various memory hierarchy designs, performance will vary as the average time to retrieve instructions and data from within the memory hierarchy will vary. Additionally, when on a processor, this component will continue to vary significantly, because the location of a workload’s instructions and data within the memory hierarchy is affected by many factors including, but not limited to these factors:

- Locality of reference
- I/O rate
- Competition from other applications and LPARs

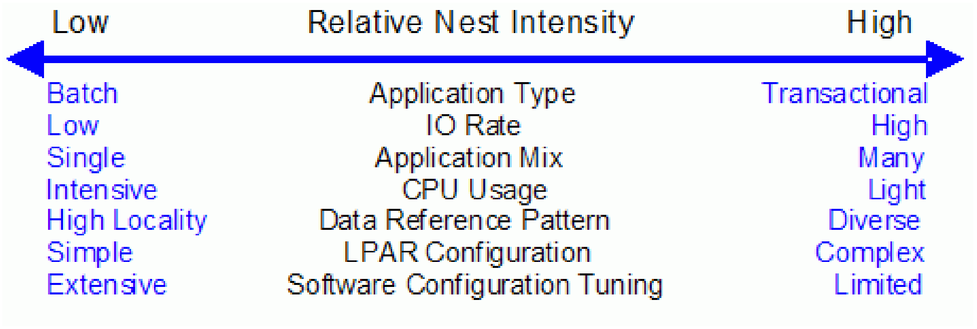
1.8.3 Relative nest intensity

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 can be calculated. The higher the RNI, the deeper into the memory hierarchy the processor must go to retrieve the instructions and data for that workload.

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 results in a net RNI for the workload, which in turn directly relates to the performance of the workload.

We emphasize that these factors are simply tendencies and not absolutes. For example, a workload might have a low I/O 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.

Figure 1-4 on page 23 lists the traditional factors that have been used to categorize workloads in the past along with their RNI tendency.



Low	Relative Nest Intensity	High
Batch	Application Type	Transactional
Low	IO Rate	High
Single	Application Mix	Many
Intensive	CPU Usage	Light
High Locality	Data Reference Pattern	Diverse
Simple	LPAR Configuration	Complex
Extensive	Software Configuration Tuning	Limited

Figure 1-4 The traditional factors that have been used to categorize workloads

Note that you can do little 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. I/O rate can be influenced somewhat through buffer pool tuning.

However, one factor that can be affected, *software 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 application-owning regions (AORs) or batch initiators) that are needed to support a workload. This factor has always existed but its sensitivity is higher with today's high frequency microprocessors. Spreading the same workload over a larger number of address spaces than necessary can raise a workload's RNI, because the working set of instructions and data from each address space increase the competition for the processor caches.

Tuning to reduce the number of simultaneously active address spaces to the proper number that is needed to support a workload can reduce RNI and improve performance. In the LSPR, the number of address spaces for each processor type and Nway configuration is tuned 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. Re-tuning the software configuration of a production workload as it moves to a bigger or faster processor might be needed in order to achieve the published LSPR ratios.

1.8.4 LSPR workload categories based on relative nest intensity

A workload's relative nest intensity is the most influential factor that determines workload performance. Other more traditional factors, such as application type or I/O 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. The LSPR now runs various combinations of former workload primitives, such as CICS, DB2, IMS, OSAM, VSAM, WebSphere, COBOL, and utilities, to produce capacity curves that span the typical range of RNI.

Three new workload categories are represented in the LSPR tables:

- *LOW* (relative nest intensity)

  A workload category representing light use of the memory hierarchy. This category is similar to past high-scaling primitives.

- *AVERAGE* (relative nest intensity)

  A workload category representing an average use of the memory hierarchy. This category is 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 category is similar to the past TI-mix workload.

These categories are based on the relative nest intensity, which is influenced by many variables, such as application type, I/O rate, application mix, CPU usage, data reference patterns, LPAR configuration, and software configuration running, among others. CPU MF data can be collected by z/OS System Measurement Facility on SMF 113 records.

1.8.5 Relating production workloads to LSPR workloads

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

- Application name (a client running CICS can use the CICS LSPR workload)
- Application type (create a mix of the LSPR online and batch workloads)
- I/O rate (low I/O rates used a mix of the low I/O rate LSPR workloads)

The previous LSPR workload suite was made up of the following workloads:

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

The traditional Commercial Batch Short Job Steps (CB-S) workload (formerly CB84) was dropped.

The previous LSPR provided performance ratios for individual workloads and for the default mixed workload, which was composed of equal amounts of four of the workloads described previously (OLTP-T, OLTP-W, WASDB, and CB-L). Guidance in converting LSPR previous categories to the new categories is provided, and built-in support on the zPCR tool[10] is provided.

However, as discussed in 1.8.1, “LSPR workload suite” on page 20, 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 through these hardware characteristics (see 1.8.3, “Relative nest intensity” on page 22 for a discussion about RNI).

The AVERAGE RNI LSPR workload is intended to match the majority of client workloads. When no other data is available, use it for capacity analysis.

DASD I/O rate has been used for many years to separate workloads into two categories: those workloads whose DASD I/O per MSU (adjusted) is less than 30 (or DASD I/O per PCI

[10] IBM Processor Capacity Reference: A no-cost tool that reflects the latest IBM LSPR measurements. Available at http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/PRS1381

less than 5) and those workloads whose DASD I/O per MSU is higher than these values. The majority of production workloads fell into the "low I/O" category and a LoIO-mix workload was used to represent them. Using the same I/O test, these workloads now use the AVERAGE RNI LSPR workload. Workloads with higher I/O rates can use the HIGH RNI workload or the AVG-HIGH RNI workload that is included with zPCR. Low-Average and Average-High categories allow better granularity for workload characterization.

For z10 and newer processors, the CPU MF data can be used to provide an 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 can be classified as LOW, AVERAGE, or HIGH RNI. This classification and the resulting "hint" are automated in the zPCR tool. It is best to use zPCR for capacity sizing.

The LSPR workloads, which have been updated for z114, are considered to reasonably reflect the current and growth workloads of the client. The set contains three generic workload categories based on z/OS R1V11 supporting up to five processors in a single image.

1.8.6 Workload performance variation

One of the major changes in z114 when compared with z10 BC is the introduction of a two processor drawer design. The z10 BC has a single processor drawer. Performance variability from application to application, similar to that seen on the z9 BC and z10 BC, is expected. This variability can be observed in certain ways. The range of performance ratings across the individual workloads is likely to have a spread.

The memory and cache designs affect various workloads in a number of ways. All workloads are improved, with cache-intensive loads benefiting the most. When comparing moving from z9 BC to z10 BC with moving from z10 BC to z114, it is likely that the relative benefits per workload will vary. Those workloads that benefited more than the average when moving from z9 BC to z10 BC will benefit less than the average when moving from z10 BC to z114, and vice-versa. Also, enhancements, such as out-of-order instruction execution, yields significant performance benefit for especially compute-intensive applications while maintaining good performance growth for traditional workloads.

z114 provides 130 available capacity settings. Each subcapacity indicator is defined with the notation A0*x*-Z0*x*, where *x* is the number of installed CPs, from one to five. There are a total of 26 subcapacity levels, designated by the letters A through Z. This extreme granularity is extremely helpful when choosing the right capacity setting for your needs. The client 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.

Experience demonstrates that System z servers can be run at up to 100% utilization levels, sustained, although most clients prefer to leave a bit of white space and run at 90% or slightly under. For any capacity comparison exercise, using only one number, such as the MIPS or MSU metric, is not a valid method. Be careful when deciding on the number of processors and the uniprocessor capacity to keep both the workload characteristics and LPAR configuration in mind. That's why, when planning capacity, we recommend using zPCR and involving IBM technical support.

1.9 Operating systems and software

The z114 is supported by a large set of software, including independent software vendor (ISV) applications. This section lists only the supported operating systems. Exploitation of various features might require the latest releases. Further information is available in Chapter 8, “Software support” on page 211.

The z114 supports any of the following operating systems:

- z/OS Version 1 Release 10 and later releases
- z/OS Version 1 Release 9 with IBM Lifecycle Extension
- z/OS Version 1 Release 8 with IBM Lifecycle Extension
- z/VM Version 5 Release 4 and later
- z/VSE Version 4 Release 2 and later
- z/TPF Version 1 Release 1
- Linux on System z distributions:
  - Novell SUSE: SLES 10 and SLES 11[11]
  - Red Hat: RHEL 5[12]

Operating system support for the zBX blades includes:

- AIX Version 5 Release 3 or later, with PowerVM Enterprise Edition
- Linux on System x (Red Hat Enterprise Linux - RHEL and SUSE Linux Enterprise Server - SLES)
- Microsoft Windows (Statement of Direction)

Finally, a large software portfolio is available to the zEnterprise 114, 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, Java 2 Platform, Enterprise Edition (J2EE), Linux, and Open Standards, the zEnterprise 114 is intended to be a platform of choice for the integration of a new generation of applications with existing applications and data.

[11] SLES is the abbreviation for Novell SUSE Linux Enterprise Server.

[12] RHEL is the abbreviation for Red Hat Enterprise Linux.

2

Central processor complex hardware components

This chapter introduces IBM zEnterprise 114 (z114) hardware components along with significant features and functions with their characteristics and options. Our objective is to explain the z114 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:

- “Frames and drawers” on page 28
- “Processor drawer concept” on page 31
- “Single-chip module” on page 35
- “Processor units and storage control chips” on page 36
- “Memory” on page 41
- “Reliability, availability, and serviceability (RAS)” on page 47
- “Connectivity” on page 48
- “Model configurations” on page 51
- “Power and cooling” on page 57
- “Summary of z114 structure” on page 59

2.1 Frames and drawers

System z frames are enclosures that are built to Electronic Industry Association (EIA) standards. The z114 central processor complex (CPC) has one 42U EIA frame, which is shown in Figure 2-1. The frame has positions for one or two processor drawers and a combination of I/O drawers or PCIe I/O drawers.

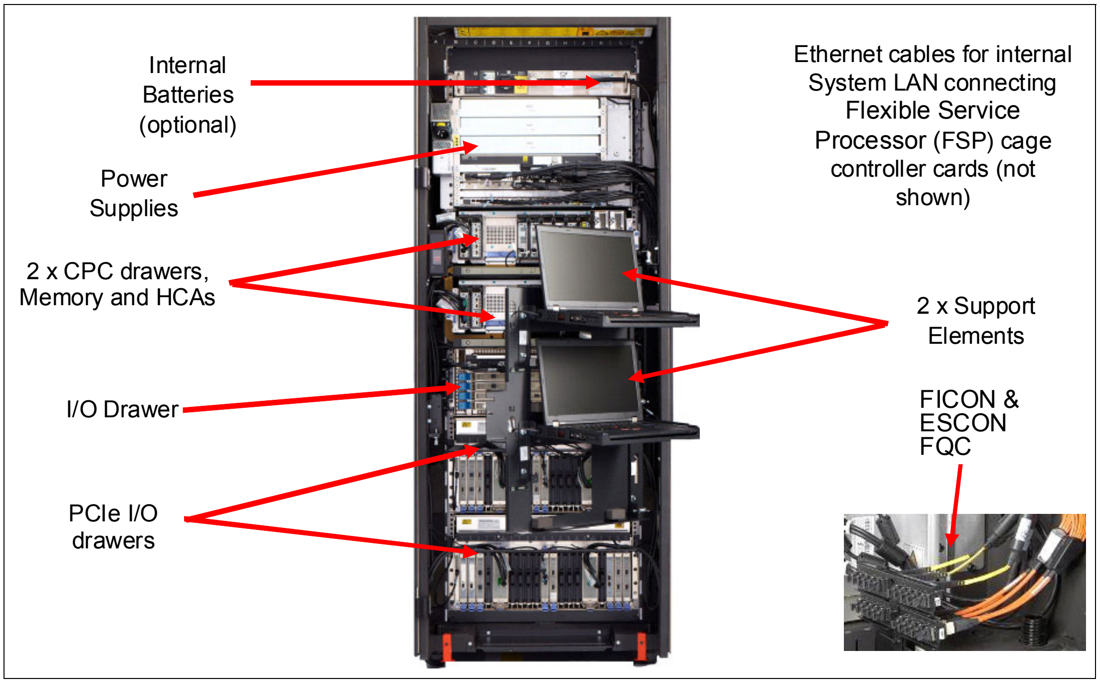

Internal Batteries (optional)
Power Supplies
2 x CPC drawers, Memory and HCAs
I/O Drawer
PCIe I/O drawers
Ethernet cables for internal System LAN connecting Flexible Service Processor (FSP) cage controller cards (not shown)
2 x Support Elements
FICON & ESCON FQC

Figure 2-1 Front view of processor drawers and I/O drawers

Figure 2-1 shows the front view of the z114 with two processor drawers, one I/O drawer, and two PCIe I/O drawers.

2.1.1 The z114 frame

The frame includes the following elements (see Figure 2-1):

- 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 both power feeds from the utility company. The IBF provides battery power to preserve full system function despite the loss of power. 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 25 minutes of full power, depending on the I/O configuration. Table 2-1 on page 29shows the IBF delay or holdup times for various configurations.

Table 2-1 z114 IBF holdup times in minutes

I/O drawers	Model M05	Model M10
No I/O drawers	25	15
1 FC 4000	18	10.5
1 FC 4003	12	8.5
2 FC 4000	12	8.5
1 FC 4000 plus 1 FC 4003	9	6.5
3 FC 4000 (RPQ 8P2533)	9	6.5
2 FC 4003	7	5
2 FC 4000 plus 1 FC 4003	7	5
4 FC 4000 (RPQ 8P2533)	7	N/A
1 FC 4000 plus 2 FC 4003	N/A	4
FC 4000 = I/O Drawer, FC 4003 = PCIe I/O Drawer		

- One or two processor drawers, each of which contains three Single-Chip Modules (SCM) and memory DIMMs
- Up to 4 I/O drawers in various combinations, as shown in Table 2-1
- Power supplies
- Support Elements (SEs)

2.1.2 I/O drawers and PCIe I/O drawers

Each processor drawer has up to four dual port fanouts to support two types of I/O infrastructures for data transfer:

- PCIe I/O infrastructure with bandwidth of 8 GBps
- InfiniBand I/O infrastructure with bandwidth of 6 GBps

PCIe I/O infrastructure uses the PCIe fanout to connect to a PCIe I/O drawer that can contain Fibre Channel (FICON) and Open Systems Adapter (OSA) feature cards:

- FICON Express8S channels (two port cards)
- OSA-Express channels:
  - OSA-Express4S 10 Gb Ethernet Long Reach and Short Reach (one port card, LR and SR)
  - OSA-Express4S Gb Ethernet (two port cards, LX and SX)

InfiniBand I/O infrastructure uses the HCA2-C fanout to connect to an I/O drawer 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 Express4-2C channels (two port cards)
  - FICON Express4 channels (four port cards)
  - FICON Express8 channels (four port cards)

- InterSystem Channel (ICS)-3 links (up to four coupling links, two links per daughter card). Two daughter cards (InterSystem Channel (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-Express3-2P (two port cards, 1000BASE-T and GbE SX)
  - OSA-Express2 Gb Ethernet (two port cards, SX and LX)
  - OSA-Express2 1000BASE-T Ethernet (two port cards)
- Crypto Express3 feature (FC 0864) has 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.

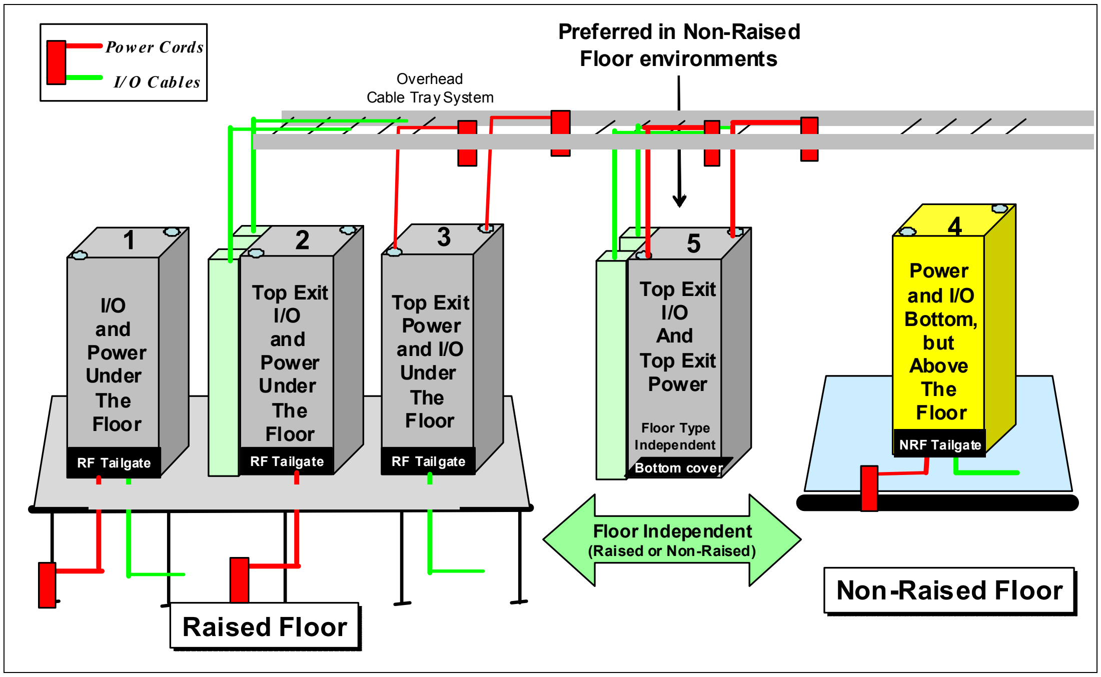
  On z114, it is possible to order the Crypto Express3-1P feature (FC 0871) which has only one PCIe adapter per feature. This adapter can also be configured as a coprocessor or as an accelerator.

InfiniBand coupling to a coupling facility is achieved directly from the HCA2-O (12xIFB) fanout and HCA3-O (12xIFB) fanout to the coupling facility with a bandwidth of 6 GBps.

The HCA2-O LR (1xIFB) fanout and HCA3-O LR (1xIFB) fanout support long distance coupling links for up to 10 km or 100 km when extended by using System z qualified dense wavelength division multiplexing (DWDM) equipment. Supported bandwidths are 5 Gbps (1x IB DDR) and 2.5 Gbps (1x IB SDR), depending on the DWDM equipment that is used.

2.1.3 I/O cabling

On the z114, there are a number of options for installation on a raised floor, as well as on a non-raised floor. Furthermore, there are options for cabling coming into the bottom of the machine or from the top of the machine, as shown in Figure 2-2 on page 31.



Power Cords
I/O Cables
Preferred in Non-Raised Floor environments
Overhead Cable Tray System
1
I/O and Power Under The Floor
RF Tailgate
2
Top Exit I/O and Power Under The Floor
RF Tailgate
3
Top Exit Power and I/O Under The Floor
RF Tailgate
5
Top Exit I/O And Top Exit Power
Floor Type Independent
Bottom cover
4
Power and I/O Bottom, but Above The Floor
NRF Tailgate
Floor Independent
(Raised or Non-Raised)
Raised Floor
Non-Raised Floor

Figure 2-2 z114 cabling options

2.2 Processor drawer concept

The z114 CPC uses a packaging concept for its processors that is based on processor drawers. A processor drawer contains SCMs, memory, and connectors to an I/O drawer or PCIe I/O drawer and other servers. The z114 M05 has one installed processor drawer, and the z114 M10 has two installed processor drawers. A processor drawer and its components are shown in Figure 2-3.

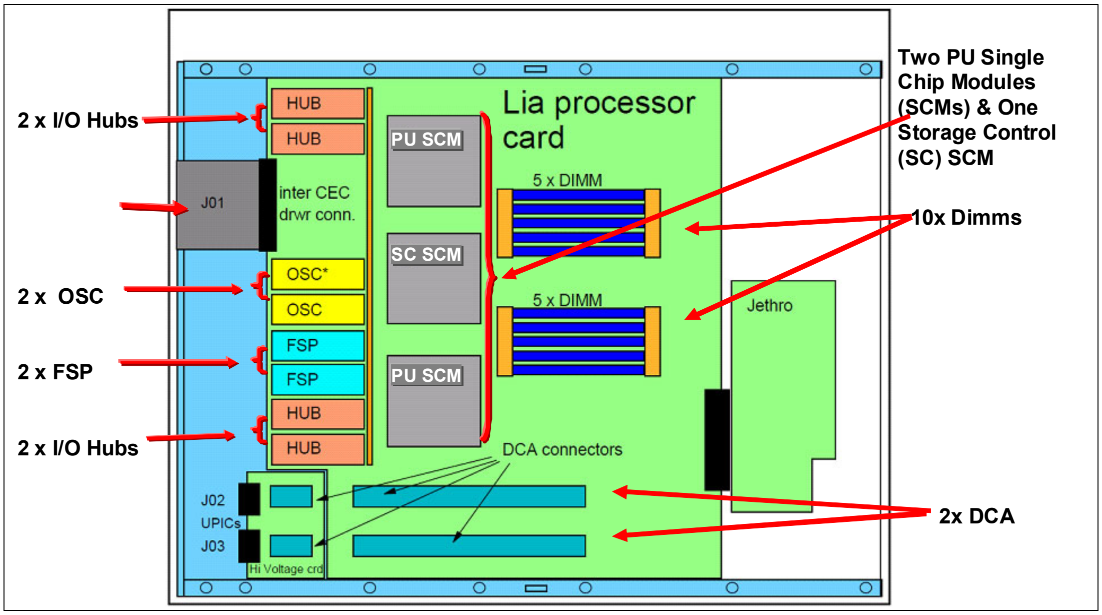


Figure 2-3 Processor drawer structure and components

Each processor drawer contains these components:

- One SC SCM with 96 MB L4 cache.
- Two PU SCMs (each PU SCM is a quad-core chip with three or four active cores).
- Memory DIMMs plugged into 10 available slots, providing up to 160 GB of physical memory installed in a processor drawer.
- A combination of up to four fanout cards. PCIe fanout connects are for links to the PCIe I/O drawers in the server, HCA2-Copper connections are for links to the I/O drawers in the server, and HCA-Optical (HCA2-O (12xIFB), HCA2-O LR (1xIFB), HCA3-O (12xIFB), and HCA3-O LR (1xIFB)) connections are to external servers (coupling links).
- Two distributed converter assemblies (DCAs) that provide power to the processor drawer. The loss of a DCA leaves enough power to satisfy the processor drawer's power requirements (N+1 redundancy). The DCAs can be concurrently maintained.
- Two flexible service processor (FSP) cards for system control.
- Two oscillator (OSC) cards with Pulse Per Second (PPS).

Figure 2-4 displays the processor drawer logical structure, showing its component connections, including the PUs on SCM.

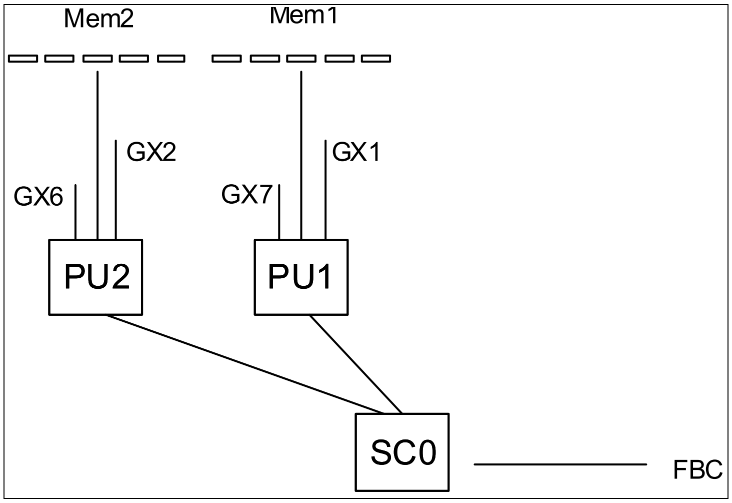


Figure 2-4 Processor drawer logical structure

Memory is connected to SCM through two memory control units (MCUs). GX1, GX2, GX6, and GX7 are the I/O bus interfaces to fanouts, with full store buffering, maximum of 10 GBps 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 processor drawers.

2.2.1 Processor drawer interconnect topology

Figure 2-5 on page 33 shows the point-to-point topology for processor drawer communication. Two processor drawers communicate directly with each other.

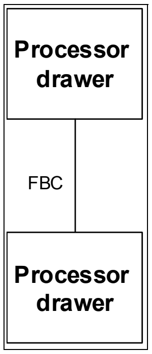


Figure 2-5 Communication between processor drawers

Important: The processor drawer slot locations are important in the sense that in the physical channel ID (PCHID) report, resulting from the IBM configurator tool.

2.2.2 Oscillator

The z114 has two oscillator cards (OSCs), a primary and a backup. Although not part of the processor drawer design, they are found at the front of the processor drawers. If the primary fails, the secondary detects the failure, takes over transparently, and continues to provide the clock signal to the server.

Figure 2-3 on page 31 shows the location of the two OSC cards on the processor drawer.

2.2.3 Pulse per second

The two oscillator cards in the first drawer of the z114 are each equipped with an interface for pulse per second (PPS), providing redundant connection to the network time protocol servers equipped with PPS output. This redundancy allows continued operation even if a single oscillator card fails. The redundant design also allows concurrent maintenance. Figure 2-6 shows the two oscillator cards in processor drawer 1 that are equipped with PPS ports.

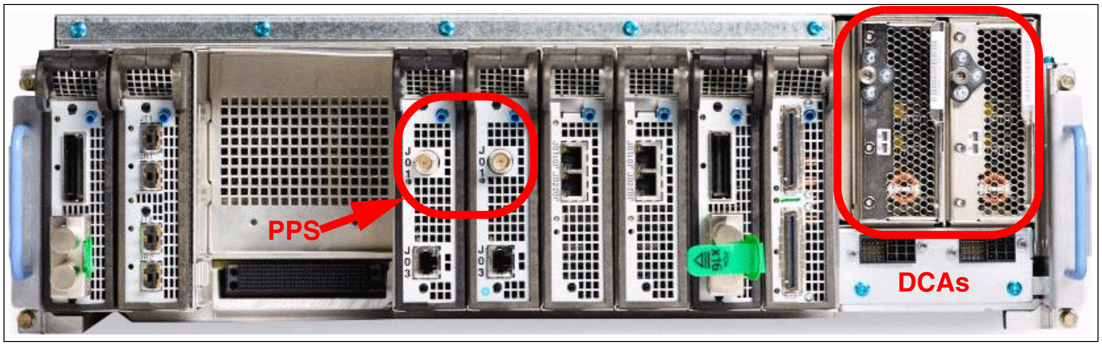


Figure 2-6 Location of PPS ports (first processor drawer)

The Support Element provides the Simple Network Time Protocol (SNTP) client. 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 synchronize to the same time source.

The time accuracy of an STP-only CTN is improved by adding an NTP server with the PPS output signal 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 oscillator card to the PPS output of the NTP server is required when the z114 is using STP and configured in an STP-only CTN using NTP with PPS as the external time source.

STP tracks the highly stable accurate PPS signal from the NTP server and maintains an 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.

STP: 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

2.2.4 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 SEs and provides a subsystem interface (SSI) for controlling components. Figure 2-7 is a conceptual overview of the system control design.

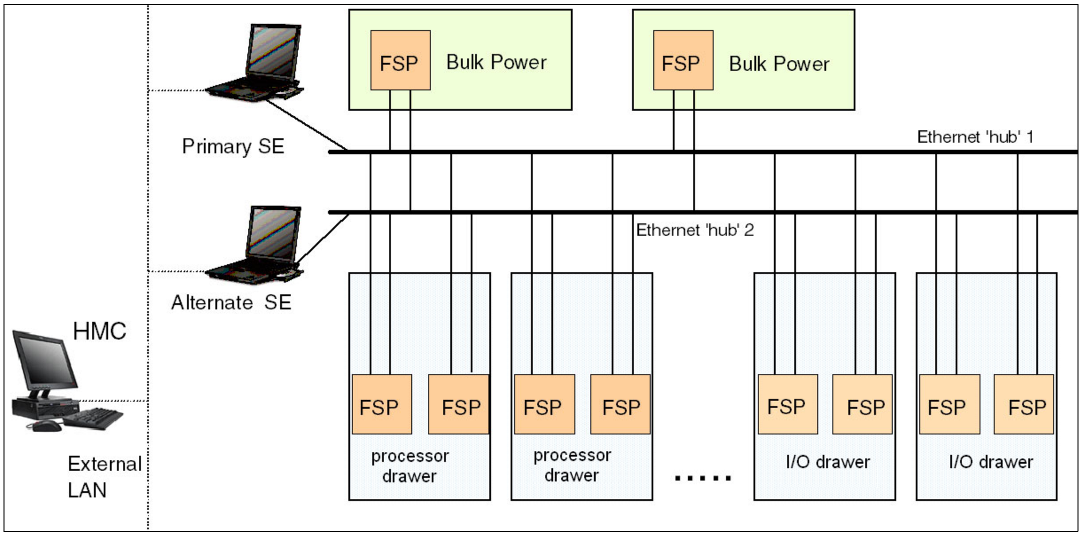


Figure 2-7 Conceptual overview of system control elements

One 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. There are two internal Ethernet LANs and two SEs for redundancy. There is also crossover capability between the LANs, 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 the same external LANs. One or more HMCs can be used, but, in an ensemble, two (a primary and an alternate[1]) are mandatory. Additional HMCs can operate a zEnterprise CPC when it is not a member of an ensemble.

Important: For ensemble configurations, the primary and the alternate HMCs must be connected to the same virtual LAN (VLAN) and have IP addresses belonging to the same subnet to allow the alternate HMC to take over the IP address in case the primary HMC fails.

If the zEnterprise CPC server is not a member of an ensemble, the controlling HMCs are stateless (there is no system status kept on the HMCs), and therefore system operations are not affected if any HMC is disconnected. At that time, the system can be managed from either SE.

However, if the zEnterprise CPC 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 user-replaceable module (URM)-managed nodes (CPCs and zEnterprise BladeCenter Extensions (zBXs)) in the collection (ensemble). In this case, the HMC is no longer simply a console/access point for configuration and policies (otherwise owned by each of the managed CPCs). The HMC of an ensemble also has an active role in ongoing system monitoring and adjustment. This role requires that the HMC is paired with an active backup (alternate) HMC[1].

2.2.5 Processor drawer power

Each processor drawer gets its power from two distributed converter assemblies (DCAs) that reside in the processor drawer (see Figure 2-6 on page 33). The DCAs provide the power for the processor drawer. Loss of one DCA leaves enough power to satisfy processor drawer power requirements. The DCAs can be concurrently maintained and are accessed from the rear of the frame.

2.3 Single-chip module

Two SCM types, which are shown on Figure 2-8 on page 36, are available:

- The microprocessor (PU chip) SCM with three or four active cores
- The system controller (SC chip) SCM

Each processor drawer has two PU SCMs (size is 50 x 50 mm) and one SC SCM (size is 61 x 61 mm).

[1] These HMCs must be running with Version 2.11 or higher. See section 12.7, “HMC in an ensemble” on page 363 for more information.

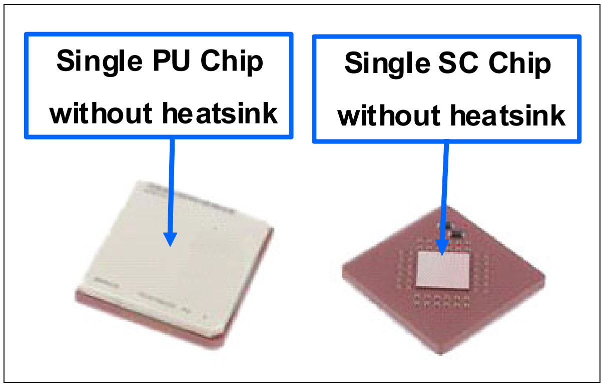


Figure 2-8 z114 Single Chip Module

The SCMs plug into a horizontal planar board (as shown in Figure 2-9) using Land Grid Arrays (LGA) connectors. Each SCM is topped with a heat sink to assure proper cooling.

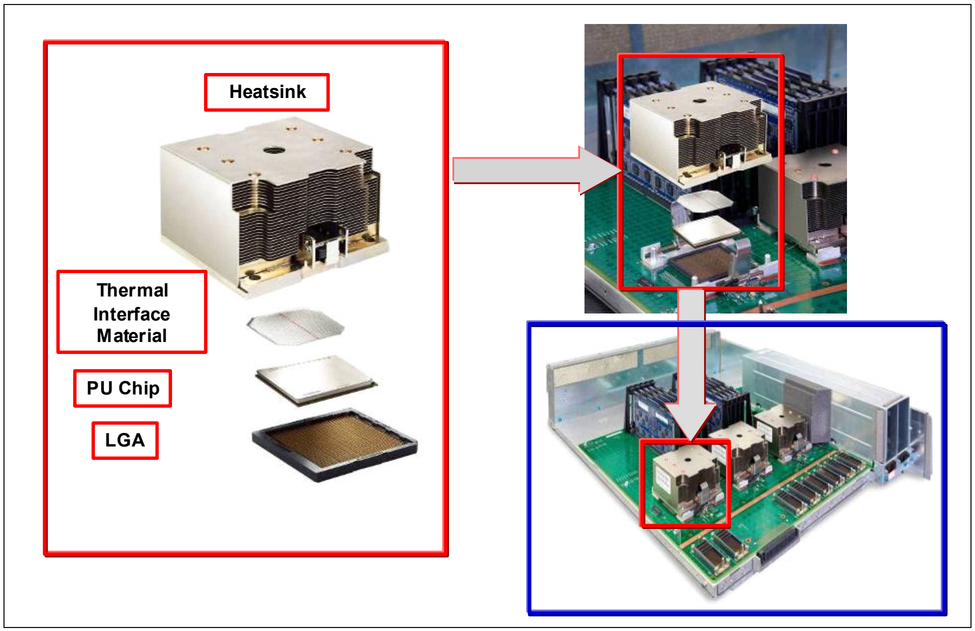


Figure 2-9 SCM and heat sink

2.4 Processor units and storage control chips

Both processor unit (PU) and storage control (SC) chips on the SCM 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 SCM, four serial electrically erasable programmable ROM (SEEPROM) chips, which are rewritable memory chips that hold data without power and are based on the same technology, are used for retaining product data for the SCM and relevant engineering information.

2.4.1 PU chip

The z114 PU chip is an evolution of the System 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 higher clock frequency, larger caches, and compiler enhancements to allow applications the benefit of the new execution units.

Each PU chip has up to four cores running at 3.8 GHz. The PU chips come in two versions, having three active cores or all four active cores. A schematic representation of the PU chip is shown in Figure 2-10.

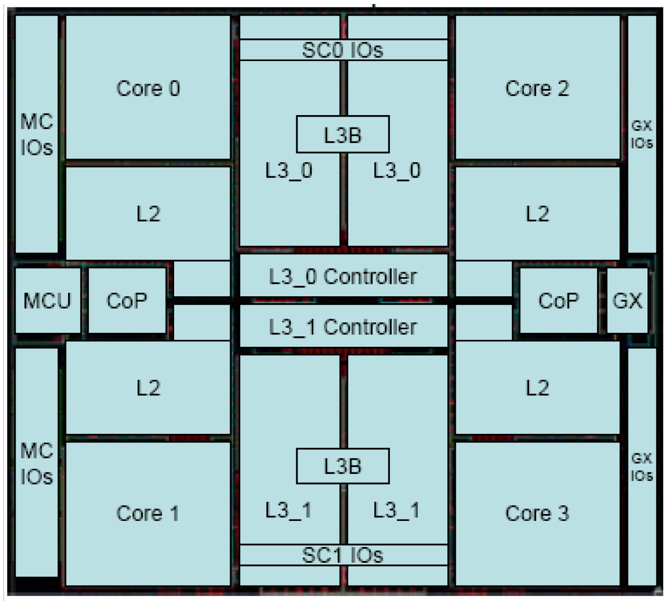


Figure 2-10 PU chip diagram

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 12 MB L3 cache. The L3 cache is a store-in shared cache across all four cores in the PU chip. It has 192 x 512 Kb 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 GBps 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 fanouts 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. For details, see 3.3.3, “Compression and cryptography accelerators on a chip” on page 70.

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 program order processor, having the following six execution units:

- 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 implemented on each core, as shown in Figure 2-11 on page 39:

- 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. The relative size of these sub-units is the result of the elaborate branch prediction design, which is further described in 3.3.2, "Superscalar processor" on page 70.

- Instruction decode unit (IDU)

  The IDU is fed from the IFU buffers and is responsible for the 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.

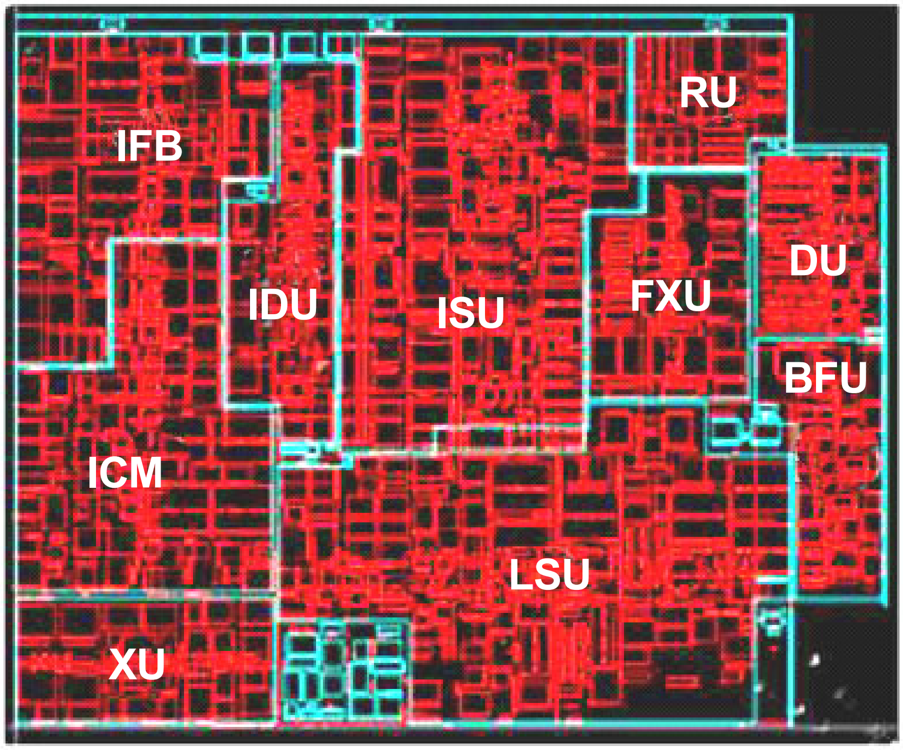


Figure 2-11 Core layout

2.4.3 PU characterization

In each processor drawer, certain PUs can be characterized for client use. The characterized PUs can be used for general purpose to run supported operating systems, such as z/OS, z/VM, and Linux on System z, or specialized to run specific workloads, such as Java, XML services, IPSec, and specific DB2 workloads or functions, such as Coupling Facility Control Code. For more information about PU characterization, see 3.4, “Processor unit functions” on page 75.

The maximum number of characterized PUs depends on the z114 model. Certain PUs are characterized by the system as standard system assist processors (SAPs), to run the I/O processing. On M10, there are two dedicate spare PUs, which are used to assume the function of a failed PU. The remaining installed PUs can be characterized for client use. A z114 model nomenclature includes a number, which represents this maximum number of PUs that can be characterized for client use, as shown on Table 2-2.

Table 2-2 Number of PUs per z114 model

Model	Processor drawer	Installed PUs	Standard SAPs	Spare PUs	Max characterized PUs
M05	1	7	2	0	5
M10	2	14	2	2	10

2.4.4 Storage control 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 processor has one SC chip. The L4 cache on the SC chip has 96 MB.

Figure 2-12 shows a schematic representation of the SC chip with its elements.

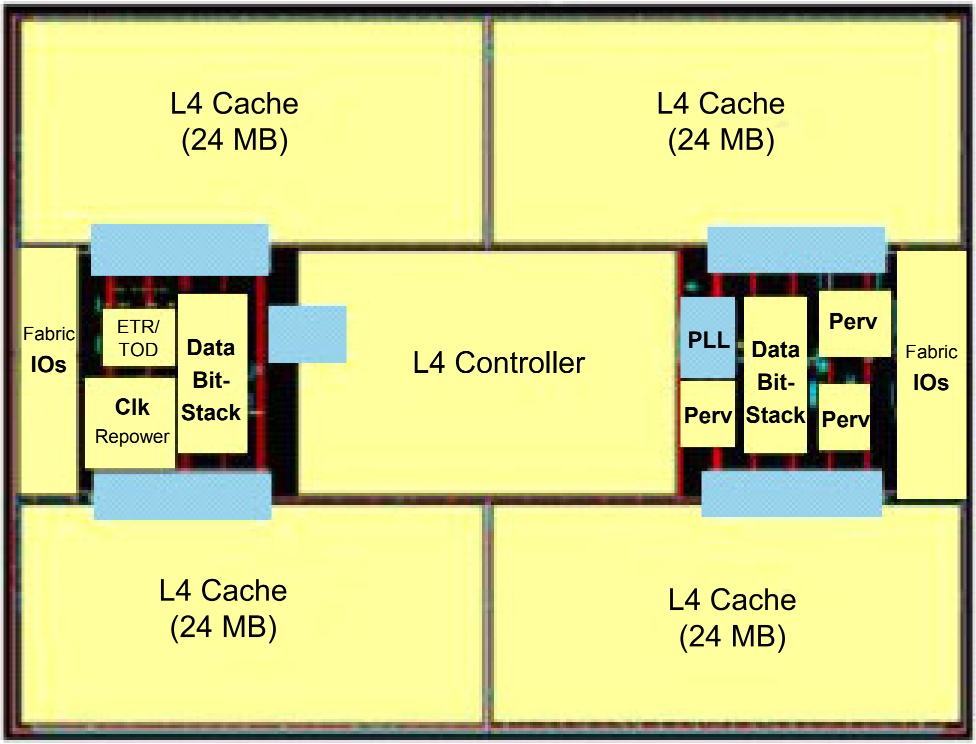


Figure 2-12 SC chip diagram

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.

2.4.5 Cache-level structure

The z114 server implements a four-level cache structure, as shown on Figure 2-13 on page 41.

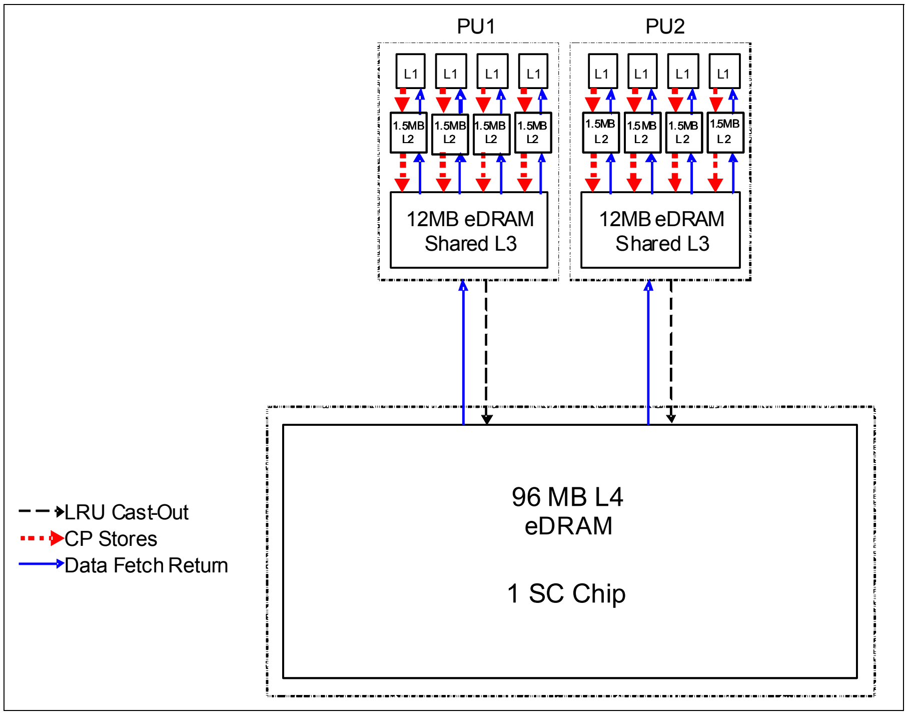


Figure 2-13 z114 processor drawer cache levels structure

Each core has its own 192 KB cache Level 1 (L1), which is 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) that is 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 PU 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 cache Level 4 (L4), which is located on the SC chip. Each SC chip has 96 MB, which is shared by both PUs on the processor drawer. The L4 cache uses a store-in design.

2.5 Memory

Maximum physical memory size is directly related to the number of processor drawer in the system. Each processor drawer can contain up to 160 GB of physical memory, for a total of 320 GB of installed memory per system.

A z114 server has more memory installed than ordered. Part of the physically installed memory is used to implement the redundant array of independent memory (RAIM) design, resulting in up to 128 GB of available memory per processor drawer and up to 256 GB per system with fixed 8 GB hardware system area (HSA). Table 2-3 on page 42 shows the maximum and minimum memory sizes that a client can order for each z114 model with separate increments.

Table 2-3 z114 server memory sizes

Model	Number of processor drawers	Increment (GB)	Customer memory (GB)
M05	1	8	8 - 120
M10	2	8	16 - 120
M10	2	32	152 - 248

On z114 servers, the memory granularity is 8 GB, for customer memory sizes from 8 to 120 GB, and 32 GB, for servers having from 152 GB to 248 GB of customer memory. Memory is physically organized in the following manner:

- A processor drawer always contains a minimum of 40 GB of physically installed memory.
- A processor drawer can have more memory installed than enabled. The excess amount of memory can be enabled by a Licensed Internal Code load when required by the installation.
- Memory upgrades are satisfied from already-installed unused memory capacity until exhausted. When no more unused memory is available from the installed memory cards, either the cards must be upgraded to a higher capacity or the second processor drawer with additional memory must be installed.

2.5.1 Memory subsystem topology

The z114 memory subsystem uses high-speed, differential-ended communications memory channels to link a host memory to the main memory storage devices. Figure 2-14 on page 43 shows an overview of the z114 memory topology.

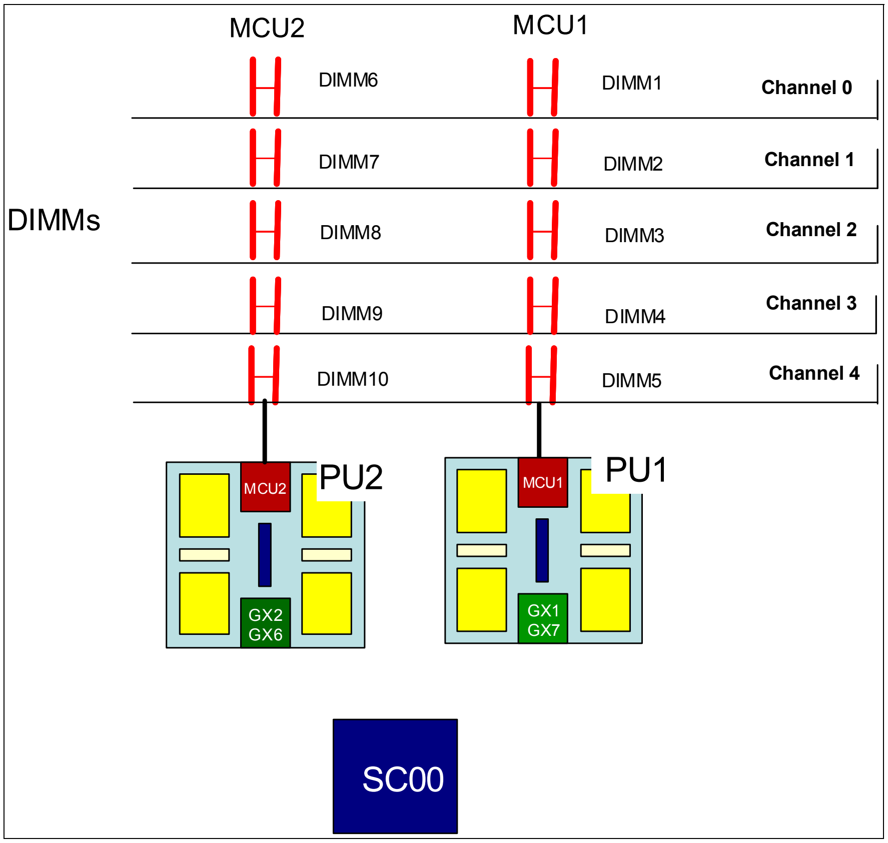


Figure 2-14 z114 memory topology

Each processor drawer has 10 dual in-line memory modules (DIMMs). DIMMs are connected to the L4 cache through two memory control units (MCUs) located on PU1 and PU2. Each MCU uses five channels, one of them for RAIM implementation, on a 4 +1 (parity) design. Each channel has one chained DIMM, so a single MCU can have five DIMMs. Each DIMM has a size of 4 GB, 8 GB, or 16 GB, and there is no mixing of DIMM sizes on a processor drawer.

2.5.2 Redundant array of independent memory (RAIM)

The z114 supports the redundant array of independent memory (RAIM), like z196. 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 2-15 on page 44.

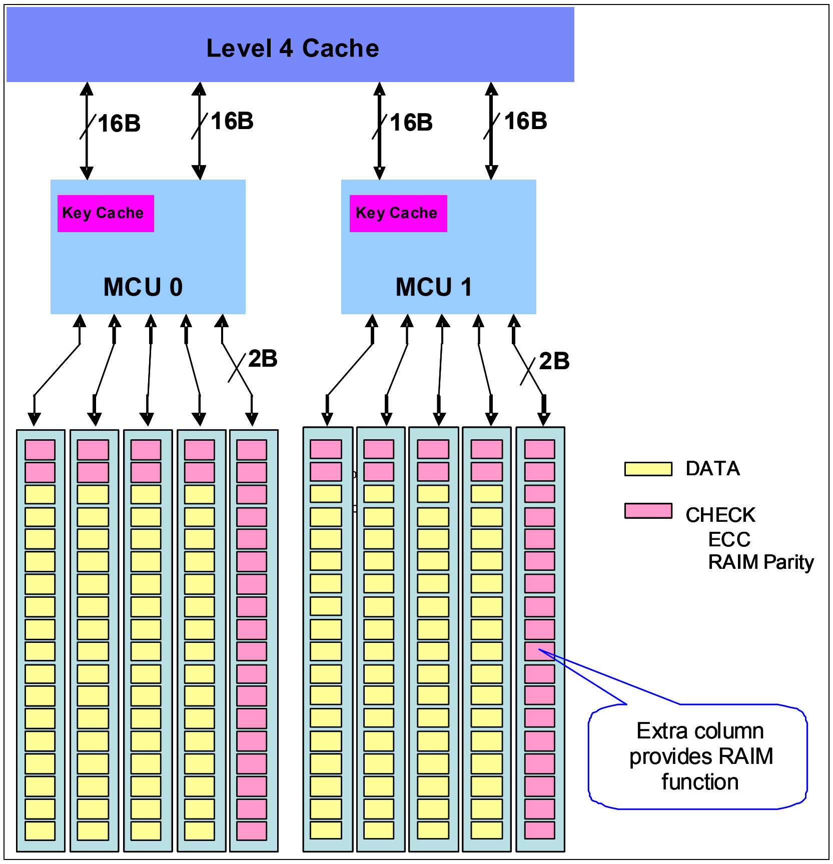


Figure 2-15 z114 RAIM DIMMs

The parity information of the four “data” DIMMs is stored in the DIMMs that are attached to the fifth memory channel. Any failure in a memory component can be detected and corrected dynamically. 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 can be purchased in increments of 8 GB up to a total size of 120 GB for M05 and M10. From 120 GB, the increment size increases to 32 GB up to 248 GB, which is for M10 only. Table 2-4 on page 45 shows all memory configurations as seen from a client and hardware perspective.

Table 2-4 z114 memory offerings

FC	GB	Increment	M05			M10 (2 processor drawers)		
			Dial max	DIMM (GB)	Number of plugged	Dial max	DIMM (GB)	Number of plugged
3509	8	8	24	4	10	N/A	N/A	N/A
3610	16	8		4	10	56	4/4	10/10
3611	24	8		4	10		4/4	10/10
3612	32	8	56	8	10		4/4	10/10
3613	40	8		8	10		4/4	10/10
3614	48	8		8	10		4/4	10/10
3615	56	8		8	10		4/4	10/10
3616	64	8	120	16	10	88	4/8	10/10
3617	72	8		16	10		4/8	10/10
3618	80	8		16	10		4/8	10/10
3619	88	8		16	10		4/8	10/10
3620	96	8		16	10	120	8/8	10/10
3621	104	8		16	10		8/8	10/10
3622	112	8		16	10		8/8	10/10
3623	120	8		16	10		8/8	10/10
3624	152	32	N/A			152	4/16	10/10
3625	184	32				184	8/16	10/10
3626	216	32				248	16/16	10/10
3627	248	32					16/16	10/10

Physically, memory is organized in the following manner:

- A processor drawer always contains 10 DIMMs with 4 GB, 8 GB, or 16 GB each.
- z114 has more memory installed than enabled. The amount of memory that can be enabled by the client is the total physically installed memory minus the RAIM amount and minus the 8 GB HSA memory.
- A processor drawer can have available unused memory, which can be ordered on a memory upgrade.

Figure 2-16 on page 46 illustrates how the physically installed memory is allocated on a z114 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.

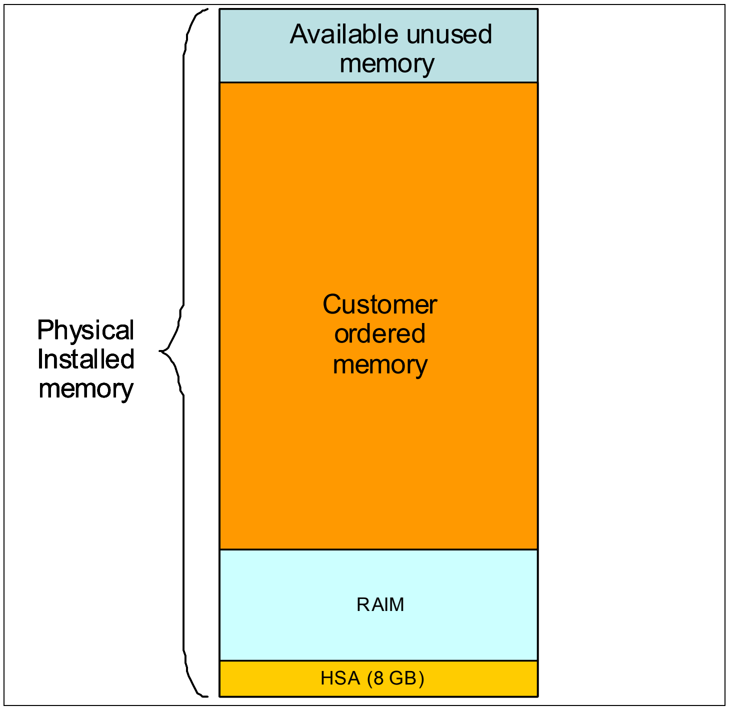


Figure 2-16 z114 Memory allocation diagram

As an example, a z114 server model M10 (two processor drawers) ordered with 216 GB of memory has the following memory sizes (refer to Figure 2-16):

- Physically installed memory is 320 GB: 160 GB on processor drawer 1 and 160 GB on processor drawer 2.
- Processor drawer 1 has the 8 GB HSA memory and up to 120 GB for customer memory, and processor drawer 2 has up to 128 GB for customer memory, resulting in 248 GB of available memory for the client.
- Because the client ordered 216 GB, provided the granularity rules are met, 32 GB (248 - 216 GB) is still available to be used in conjunction with additional memory for future upgrades by LIC.

When activated, a logical partition can use memory resources that are located in either processor drawer. For more information, see 3.6, “Logical partitioning” on page 89.

2.5.4 Memory upgrades

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 processor drawer 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. See 2.8, “Model configurations” on page 51.

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 3.6.3, “Reserved storage” on page 96. Alternately, additional memory can be used by an already defined logical partition that is activated after the memory addition.

2.5.5 Pre-planned memory

Pre-planned memory provides the ability to plan for nondisruptive permanent memory upgrades. When preparing in advance for a future memory upgrade, note that memory can be preplugged in, based on a target capacity. The preplugged memory can be made available through a Licensed Internal Code (LIC) configuration code (LICCC) update. You can order this LICCC through one of these sources:

- 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), which are described in Table 2-5.

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 use. For the exact terms and conditions, contact your IBM representative.

Table 2-5 Feature codes for plan-ahead memory

Memory	z114 feature code
Pre-planned memory Charged when physical memory is installed. Used for tracking the quantity of physical increments of plan-ahead memory capacity.	FC 1993
Pre-planned memory activation Charged when plan-ahead memory is enabled. Used for tracking the quantity of increments of plan-ahead memory being activated.	FC 1903

You install pre-planned memory by ordering FC 1993. The ordered amount of plan-ahead memory is charged with a reduced price compared to the normal price for memory. One FC 1993 is needed for each 8 GB physical increment.

The activation of installed pre-planned memory is achieved by ordering FC 1903, which causes the other portion of the previously contracted charge price to be invoiced. FC 1903 indicates 8 GB (or 32 GB in larger configurations) of LICCC increments of memory capacity.

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

2.6 Reliability, availability, and serviceability (RAS)

IBM System z continues to deliver enterprise RAS with the IBM zEnterprise 114. Patented error correction technology in the memory subsystem provides the most robust IBM 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 application-specific integrated circuit (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 SCM, can be corrected. Upstream and downstream 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 upstream and one downstream). Taken together, this design provides System z's strongest memory subsystem.

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

z114 introduces fully fault-protected N+2 voltage transformation module (VTM) power conversion in the processor drawer. This redundancy protects processor workloads from loss of voltage due to virtual terminal manager (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 drawers and Parallel Sysplex InfiniBand coupling (IFB) are driven from the host channel adapter fanouts that are located on the front of the processor drawer. Connections to PCIe I/O drawers are driven from the PCIe fanouts. Figure 2-17 shows the location of the fanouts and connectors.

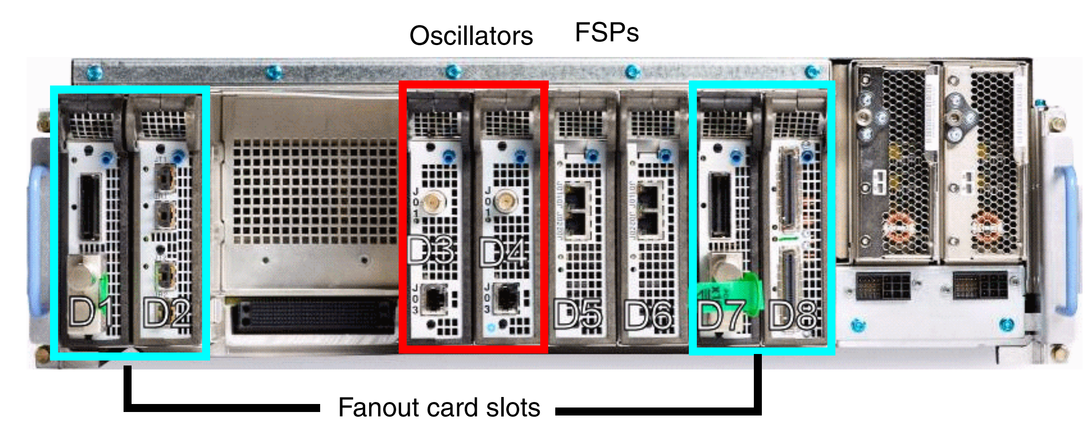


Figure 2-17 Location of the host channel adapter fanouts (first processor drawer shown)

Each processor drawer has up to four fanouts (numbered D1, D2, D7, and D8). The fanout slot sequence for plugging all fanout cards is strictly an outside in, right-to-left sequence, D8, D1, D7, and D2. Slots D3 and D4 are used for OSC, and slots D5 and D6 are used for FSP cards, not fanouts. CP chips are wired to certain fanout slots: one CP to D1 and D2 and the other CP to D7 and D8. There is a possible degrade mode if one CP chip is lost, which is the reason for the fanout plugging order. A fanout can be repaired concurrently with the use of redundant I/O interconnect. See 2.7.1, "Redundant I/O interconnect" on page 49.

Six types of fanouts are available:

- 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 drawers.
- PCIe fanout provides copper connections for PCIe I/O interconnect to all I/O cards in PCIe I/O drawers.

- Host Channel Adapter2-O (HCA2-O (12xIFB)) provides optical connections for 12x InfiniBand for coupling links (IFB). The HCA2-O (12xIFB) provides a point-to-point connection over a distance of up to 150 m (492.17 ft.), using four 12x Multi-Fiber Push-On (MPO) fiber connectors and OM3 fiber optic cables (50/125 µm).

  z114 to z196, z114, or System z10 connections use a 12-lane InfiniBand link at 6 GBps.
- The HCA2-O LR (1xIFB) fanout provides optical connections for 1x InfiniBand and supports IFB Long Reach (IFB LR) coupling links for distances of up to 10 km (6.21 miles) and up to 100 km (62.1 miles) when repeated through a System z-qualified DWDM. This fanout is supported on z196, z114, and System z10 only.

  IFB LR coupling links operate at up to 5.0 Gbps (1x IB-DDR) between two servers or automatically scale down to 2.5 Gbps (1x IB-SDR), depending on the capability of the attached equipment.
- Host Channel Adapter3-O (HCA3-O (12xIFB)) provides optical connections for 12x IFB or 12x IFB3 for coupling links (IFB). For details, refer to “12x IFB and 12x IFB3 protocols” on page 117. The HCA3-O (12xIFB) provides a point-to-point connection over a distance of up to 150 m (492.17 ft.), using four 12x MPO fiber connectors and OM3 fiber optic cables (50/125 µm). This fanout is supported on z196 and z114 only.

  z114 to z196, z114, or System z10 connections use a 12-lane InfiniBand link at 6 GBps.
- The HCA3-O LR (1xIFB) fanout provides optical connections for 1x InfiniBand and supports IFB Long Reach (IFB LR) coupling links for distances of up to 10 km (6.21 miles) and up to 100 km (62.1 miles) when repeated through a System z-qualified DWDM. This fanout is supported on z196 and z114 only.

  IFB LR coupling links operate at up to 5.0 Gbps between two servers or automatically scale down to 2.5 Gbps, depending on the capability of the attached equipment.

Up to 4 fanouts can be installed on the z114 M05. Up to 8 fanouts can be installed on the z114 M10.

2.7.1 Redundant I/O interconnect

Next, we describe redundant I/O interconnect.

InfiniBand I/O connection

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 that is located in the InfiniBand fanout of the processor drawer. IFB-MP cards are interconnected, allowing redundant I/O connection in case the connection coming from a processor drawer ceases to function. A conceptual view of how redundant I/O interconnect is accomplished is shown in Figure 2-18 on page 50.

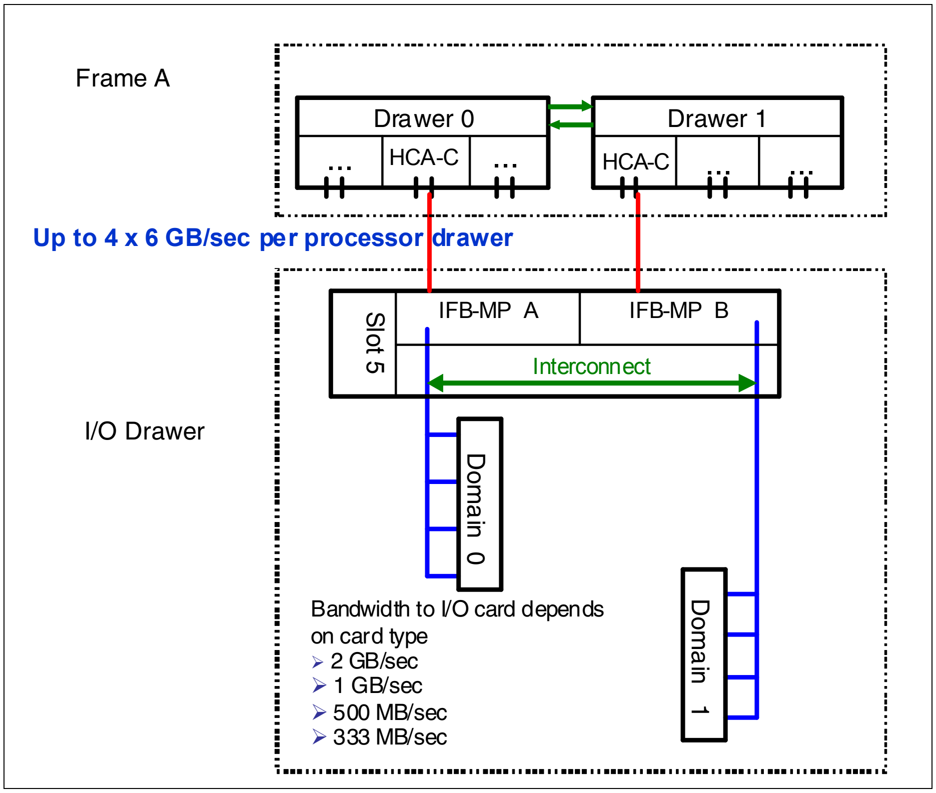


Figure 2-18 Redundant I/O interconnect for I/O drawer

Normally, the HCA2-C fanout in the first processor drawer connects to the IFB-MP (A) card and services domain 0 in an I/O drawer. In the same fashion, another HCA2-C fanout of the processor drawer of the model M05 or of the second processor drawer in case of a model M10 connects to the IFB-MP (B) card and services domain 1 in an I/O drawer. If one of the connections to the IFB-MP card is removed, connectivity to the failing domain is maintained by guiding the I/O to this domain through the interconnect between IFB-MP (A) and IFB-MP (B).

In configuration reports, drawers are identified by their location in the rack. HCA2-C fanouts are numbered from D1, D2, and D7, D8. The jacks are numbered J01 and J02 for each HCA2-C fanout port.

PCIe I/O connection

The PCIe I/O drawer supports up to 32 I/O cards. They are organized in four hardware domains per drawer, as shown on Figure 2-19 on page 51.

Each domain is driven through a PCIe switch card. Two PCIe switch cards always provide a backup path for each other through the passive connection in the PCIe I/O drawer backplane. That way, in case of a PCIe fanout or cable failure, all 16 I/O cards in the two domains can be driven through a single PCIe switch card.

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

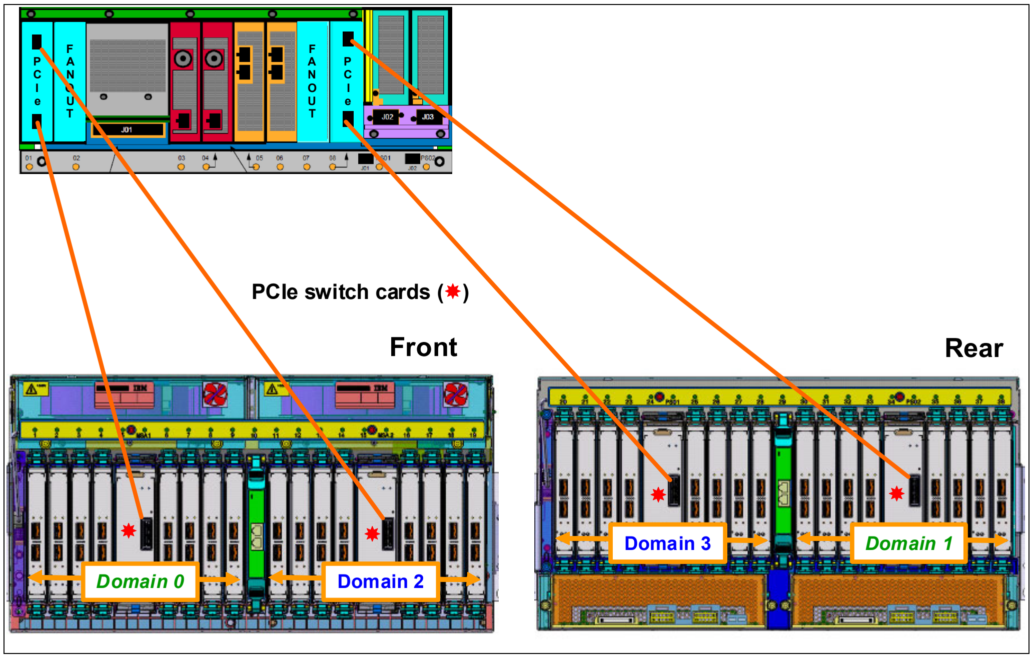


Figure 2-19 Redundant I/O interconnect for PCIe I/O drawer

2.8 Model configurations

When a z114 order is configured, PUs are characterized according to their intended use. 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 unavailable for use until an upgrade for the CP is installed. It does not affect software licenses or maintenance charges.

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

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

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

zAAP A z114 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 z114 Integrated Information Processor (zIIP) purchased and activated to run eligible workloads, such as DB2 DRDA or z/OS[2] Communication Server IPSec.

[2] z/VM V5R4 and higher support zAAP and zIIP processors for guest configurations.

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

A minimum of one PU characterized as a CP, IFL, or ICF is required per system. The maximum number of CPs is five, the maximum number of IFLs is 10, and the maximum number of ICFs is 10. The maximum number of zAAPs is five, but it requires an equal or greater number of characterized CPs. The maximum number of zIIPs is also five, and it 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. Table 2-6 shows these details.

Table 2-6 z114 configurations

Model	Processor drawer	CPs	IFLs/ uIFL	ICFs	zAAPs	zIIPs	Add. SAPs	Std. SAPs	Spares
M05	1	0 - 5	0 - 5	0 - 5	0 - 2	0 - 2	0 - 2	2	0
M10	2	0 - 5	0 - 10	0 - 10	0 - 5	0 - 5	0 - 2	2	2

Not all PUs on a given model are required to be characterized. The z114 model nomenclature is based on the number of PUs available for client use in each configuration.

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 that is intended to be used as CPs in the future. This capacity usually has this status for software-charging reasons. Unused CPs are not a factor when establishing the millions of service units (MSU) value that is used for charging MLC software, or when charged on a per-processor basis.

2.8.1 Upgrades

Concurrent CP, IFL, ICF, zAAP, zIIP, or SAP upgrades are done within a z114. 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. On the model M05, eventual unassigned PUs will be used as spares. A fully configured M05 does not have any spares. The model M10 always has two dedicated spares.

If an upgrade request cannot be accomplished within the given M05 configuration, a hardware upgrade to model M10 is required. The upgrade enables the addition of another processor drawer to accommodate the desired capacity. The upgrade from M05 to M10 is disruptive.

You can upgrade a System z10 Business Class (BC) or a System z9® BC to a z114, preserving the server serial number (S/N). The I/O cards are also moved up (with certain restrictions).

Important: Upgrades from System z10 and System z9 are disruptive.

2.8.2 Concurrent PU conversions

Assigned CPs, assigned IFLs, and unassigned IFLs, ICFs, zAAPs, zIIPs, and SAPs can 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 M05 with five CPs is converted to an all IFL system. In addition, a logical partition might be disrupted if PUs must be freed before they can be converted.

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 characterizations have no effect on software charging. More information about the STSI output is shown in “Processor identification” on page 313.

Capacity identifiers: Within a z114, all CPs have the same capacity identifier. Specialty engines (IFLs, zAAPs, zIIPs, and ICFs) operate at full speed.

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 as shown in Table 2-7.

Table 2-7 Model capacity identifier and MSU values

Model capacity identifier	MSU	Model capacity identifier	MSU	Model capacity identifier	MSU
A01	3	B01	4	C01	5
A02	6	B02	7	C02	9
A03	9	B03	10	C03	12
A04	11	B04	12	C04	16
A05	13	B05	15	C05	19
D01	6	E01	7	F01	7
D02	11	E02	12	F02	13
D03	15	E03	16	F03	19
D04	19	E04	21	F04	25
D05	23	E05	26	F05	30
G01	9	H01	10	I01	11
G02	16	H02	18	I02	20
G03	23	H03	26	I03	29
G04	29	H04	33	I04	37
G05	35	H05	40	I05	44
J01	12	K01	14	L01	16
J02	22	K02	25	L02	30
J03	32	K03	36	L03	42

Model capacity identifier	MSU	Model capacity identifier	MSU	Model capacity identifier	MSU
J04	41	K04	46	L04	54
J05	49	K05	55	L05	65
M01	19	N01	21	O01	24
M02	34	N02	40	O02	44
M03	49	N03	56	O03	62
M04	62	N04	71	O04	80
M05	75	N05	86	O05	97
P01	27	Q01	31	R01	34
P02	49	Q02	56	R02	62
P03	70	Q03	80	R03	88
P04	90	Q04	102	R04	113
P05	108	Q05	123	R05	136
S01	38	T01	42	U01	49
S02	69	T02	77	U02	88
S03	98	T03	110	U03	125
S04	125	T04	140	U04	161
S05	151	T05	170	U05	194
V01	53	W01	60	X01	73
V02	98	W02	109	X02	132
V03	138	W03	155	X03	188
V04	177	W04	198	X04	240
V05	214	W05	239	X05	290
Y01	86	Z01	98		
Y02	156	Z02	177		
Y03	221	Z03	251		
Y04	283	Z04	321		
Y05	343	Z05	388		

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

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.

There are six CBU types:

- 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 on the server. Up to eight records can be active, and they contain several CBU activation variations that apply to the installation.

You can test the CBU. The number of free CBU test activations in each CBU record is now determined by the number of years that are purchased with the CBU record. (For example, a 3-year CBU record has three test activations, and a 1-year CBU record has one test activation.) You can increase the number of tests up to a maximum of 15 for each CBU record. The real activation of CBU lasts up to 90 days with a grace period of two days to prevent sudden deactivation when the 90-day period expires. The contract duration can be set from 1 - 5 years.

The CBU record describes the following properties 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 the IBM configurator output, as 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 D03, and the capacity level does not change, the D03 becomes D05:

  `(D03 + 2 = D05)`

  If the capacity level changes to a D06, the number of additional CPs (3) are added to the 3 CPs of the D03, 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.

On/Off Capacity on Demand: 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 can 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 which also applies to CBU for specialty engines. Note that the CBU record can contain larger numbers of CBUs than can fit in the current model.

Unassigned IFLs are ignored. They are considered spares and are available for use as CBUs. 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 D03 has two CPs added temporarily, it becomes a model capacity identifier D05.

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 configuration with model capacity identifier D03 specifies four temporary CPs through On/Off CoD, the result is a server with model capacity identifier E05. A cross-over does not necessarily mean that the CP count for the additional temporary capacity will increase. The same D03 can temporarily be upgraded to a server with model capacity identifier F03. In this case, the number of CPs does not increase, but 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.

Appendix B, “Valid z114 On/Off Capacity on Demand upgrades” on page 375 shows possible On/Off CoD CP upgrades. For more information about temporary capacity increases, see Chapter 9, “System upgrades” on page 275.

2.9 Power and cooling

As environmental concerns raise the focus on energy consumption, the IBM zEnterprise 114 offers a holistic focus on the environment. New efficiencies and functions, such as power capping, enable a dramatic reduction of energy usage and floor space when consolidating workloads from distributed servers.

The power service specifications for the zEnterprise CPCs are the same as their particular predecessors, but the power consumption is more efficient. A fully loaded z114 CPC maximum consumption is 6.02 kW, which is nearly the same as a z10 BC, but with a maximum performance ratio of 1.64, it has a much higher exploitation on the same footprint.

2.9.1 Power considerations

The zEnterprise CPCs operate with two completely redundant power supplies. Each power supply has an individual line cord for the z114.

For redundancy, the servers have two power feeds. Line cords attach either 50/60 Hz, AC power single-phase 200 to 415 v or 3-phase 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.

There is a Balanced Power Plan Ahead feature available for future growth, also assuring adequate and balanced power with AC line cord selection. With this feature, downtimes for upgrading a server will be eliminated by including the maximum power requirements in terms of Bulk Power Regulators (BPR) and line cords to your installation. For ancillary equipment, such as the HMC, its display, and its modem, additional single-phase outlets are required.

The power requirements depend on the installed cooling facility and on the number of books, or respectively processor drawers in the z114, and the number and kind of I/O units installed. Table 11-1 on page 331 shows the maximum power consumption tables for the various configurations and environments.

The z114 can operate in raised floor and non-raised floor environments. For both kinds of installation, an overhead power cable option for the top exit of the cables is available. In the case of a non-raised floor environment, if the top exit option is chosen for the power cables, the I/O cables must use the top exit option as well.

2.9.2 High-voltage DC power

In data centers today, many businesses pay increasingly expensive electric bills and are running out of power. The zEnterprise CPC High Voltage Direct Current power feature adds nominal 380 to 520 volt DC input power capability to the existing System z, universal 3 phase, 50/60 hertz, totally redundant power capability (nominal 200 - 240VAC or 380 - 415VAC or 480VAC). It allows CPCs to directly use the high voltage DC distribution in new, green data centers. A direct HV DC data center power design can improve data center energy efficiency by removing the need for a DC to AC inversion step.

The zEnterprise CPCs bulk power supplies have been modified to support HV DC so the only difference in shipped hardware to implement the option is the DC line cords. Because HV DC is a new technology, there are multiple proposed standards. The zEnterprise CPC supports both ground-referenced and dual polarity HV DC supplies, such as +/- 190V or +/- 260V, or +380V, and so on. Beyond the data center, UPS, and power distribution energy savings, a zEnterprise CPC run on HV DC power will draw 1 - 3% less input power. HV DC does not change the number of line cords that a system requires.

2.9.3 Internal Battery Feature (IBF)

IBF is an optional feature on the zEnterprise CPC server. See Figure 2-1 on page 28 for the z114 for a pictorial view of the location of this feature. This optional IBF provides the function of a local uninterrupted power source.

The IBF further enhances the robustness of the power design, increasing power line disturbance immunity. It provides battery power to preserve processor data in case of a loss of power on all four AC feeds from the utility company. The IBF can hold power briefly during a brownout, or for orderly shutdown in case of a longer outage. The values for the holdup time depend on the I/O configuration, as shown in Table 11-2 on page 332.

2.9.4 Power capping

The IBM zEnterprise 114 supports power capping, which gives the ability to control the maximum power consumption and reduce cooling requirements (especially with zBX). To use power capping, you must order FC 0020, Automate Firmware Suite. This feature is used to enable the Automate suite of functionality that is associated with the IBM zEnterprise Unified Resource Manager. The Automate suite includes representation of resources in a workload context, goal-oriented monitoring and management of resources, and energy management.

2.9.5 Power estimation tool

The Power estimator tool for the zEnterprise CPCs allows you to enter your precise server configuration to produce an estimate of power consumption. Log in to Resource Link with any user ID. Navigate to **Planning** → **Tools** → **Power Estimation Tools**. Specify the quantity for the features that are installed in your machine. This tool estimates the power consumption for the specified configuration. The tool does not verify that the specified configuration can be physically built.

> **Power consumption:** The exact power consumption for your machine will vary. The object of the tool is produce an estimation of the power requirements to aid you in planning for your machine installation. Actual power consumption after installation can be confirmed on the HMC System Activity Display.

2.9.6 Cooling requirements

The z114 is an air-cooled system. It requires chilled air, ideally coming from under the raised floor, to fulfill the air-cooling requirements. The chilled air is usually provided through perforated floor tiles. The amount of chilled air that is required for a variety of temperatures under the floor of the computer room is indicated in *Installation Manual - Physical Planning (IMPP),* GC28-6907.

2.10 Summary of z114 structure

Table 2-8 on page 60 summarizes all aspects of the z114 structure.

Table 2-8 System structure summary

Description	Model M05	Model M10
Number of PU SCMs	2	4
Number of SC SCMs	1	2
Total number of PUs	8	16
Maximum number of characterized PUs	7	14
Number of CPs	0 - 5	0 - 5
Number of IFLs	0 - 5	0 - 10
Number of ICFs	0 - 5	0 - 10
Number of zAAPs	0 - 2	0 - 5
Number of zIIPs	0 - 2	0 - 5
Standard SAPs	2	2
Additional SAPs	0 - 2	0 - 2
Standard spare PUs	0	2
Enabled memory sizes	8 - 120 GB	16 - 248 GB
L1 cache per PU	64-I/128-D KB	64-I/128-D KB
L2 cache per PU	1.5 MB	1.5 MB
L3 shared cache per PU chip	12 MB	12 MB
L4 shared cache	96 MB	192 MB
Cycle time (ns)	0.26	0.26
Clock frequency	3.8 GHz	3.8 GHz
Maximum number of fanouts	4	8
I/O interface per IFB cable	6 GBps	6 GBps
I/O interface per PCIe cable	8 GBps	8 GBps
Number of Support Elements	2	2
External AC power	3 phase	3 phase
Optional external DC	570V/380V	570V/380V
Internal Battery Feature	Optional	Optional

3

Central processor complex system design

In this chapter, we explain how the z114 central processor complex (CPC) is designed. You can use this information to understand the functions that make the z114 a server that suits a broad mix of workloads for enterprises.

We cover the following topics:

The design of the z114 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 toward new types of applications on which clients are becoming more and more dependent.

The z114 offers high levels of serviceability, availability, reliability, resilience, and security. It fits in the IBM strategy in which mainframes play a central role in creating an intelligent, energy efficient, integrated infrastructure. The z114 is designed so that the server and everything around it (operating systems, middleware, storage, security, and network technologies supporting open standards) is important for the infrastructure and helping clients to achieve their business goals.

The modular I/O drawer and the new introduced PCIe I/O drawer design aim to reduce planned and unplanned outages by offering concurrent repair, replace, and upgrade functions for I/O. The z114 with its ultra-high frequency, large high-speed buffers (caches) and memory, superscalar processor design, out-of-order core execution, and flexible configuration options is the next implementation in the mid-sized server area to address the ever-changing IT environment.

3.1 Design highlights

The physical packaging of the z114 compares to the packaging used for z10 Business Class (BC) systems. Its processor drawer design creates the opportunity to address the ever-increasing costs related to building systems with ever-increasing capacities and offers unprecedented capacity settings to meet consolidation needs in the Mid-size world.

The z114 continues the line of compatibility with previous systems and has 246 complex instructions which are executed by millicode, and another 211 complex instructions executed through multiple operations. It uses 24, 31, and 64-bit addressing modes, multiple arithmetic formats, and multiple address spaces robust inter-process security.

The z114 system design, which are discussed in this and subsequent chapters, has the following major objectives:

- Offers 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.
- Offers state-of-the-art *integration* capability for server consolidation, offering virtualization techniques:
  - Logical partitioning, which allows 30 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 integration capability allows for a logical and virtual server coexistence and maximizes system utilization and efficiency, by sharing hardware resources.
- Offers *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.
- Offers the *high scalability* required by the most demanding applications, both from single-system and clustered-systems points of view compared to z10 BC.
- Offers the capability of *concurrent upgrades* for memory and I/O connectivity, avoiding server outages in planned situations.
- Implements 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.
- Has broad internal and external *connectivity* offerings, supporting open standards, such as Gigabit Ethernet (GbE), and Fibre Channel Protocol (FCP) for the Small Computer System Interface (SCSI).
- Provides *leading* cryptographic *performance* in which every two PUs share a CP Assist for Cryptographic Function (CPACF). You can add optional Crypto Express features with cryptographic coprocessors providing the *highest* standardized *security* certification[1] and Cryptographic Accelerators for Secure Sockets Layer/Transport Layer Security (SSL/TLS) transactions.
- *Self-manages* and *self-optimizes*, adjusting itself on workload changes to achieve the best system throughput, through the Intelligent Resource Director or the Workload Manager (WLM) functions, assisted by HiperDispatch.

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

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

The following sections describe the z114 system structure, showing a logical representation of the data flow from PUs, caches, memory cards, and a variety of interconnect capabilities.

3.2 processor drawer design

The z114 is available in two models: M05 with a single processor drawer, and M10 with two processor drawers, offering additional flexibility for I/O and coupling expansion and increased specialty engine capability. Up to 10 processor units and up to 256 GB of memory, including the HSA, can be characterized. Memory has up to four memory controllers, using a five-channel redundant array of independent memory (RAIM) protection, with DIMM bus cyclic redundancy check (CRC) error retry. The four-level cache hierarchy is implemented with eDRAM (embedded) caches. Up until recently, eDRAM was considered to be too slow for this purpose, but a break-through in IBM technology has demonstrated the opposite. In addition, eDRAM offers higher density, less power utilization, fewer soft errors, and better performance.

With up to 10 configurable cores, the model naming is indicative of how many total processor units are available for user characterization. Table 3-1 shows how the cores can be configured.

Table 3-1 z114 PU characterization

Model	CPs	IFLs unassigned IFLs	zAAPs	zIIPs	ICFs	Standard SAPs	Additional SAPs	Spares
M05	0 - 5	0 - 5	0 - 2	0 - 2	0 - 5	2	0 - 2	0
M10	0 - 5	0 - 10	0 - 5	0 - 5	0 - 10	2	0 - 2	2

3.2.1 Cache levels and memory structure

The z114 memory subsystem focuses on keeping data “closer” to the processor unit, implementing new chip-level shared cache (L3) and much larger shared cache (L4).

Figure 3-1 shows the z114 cache levels and memory hierarchy.

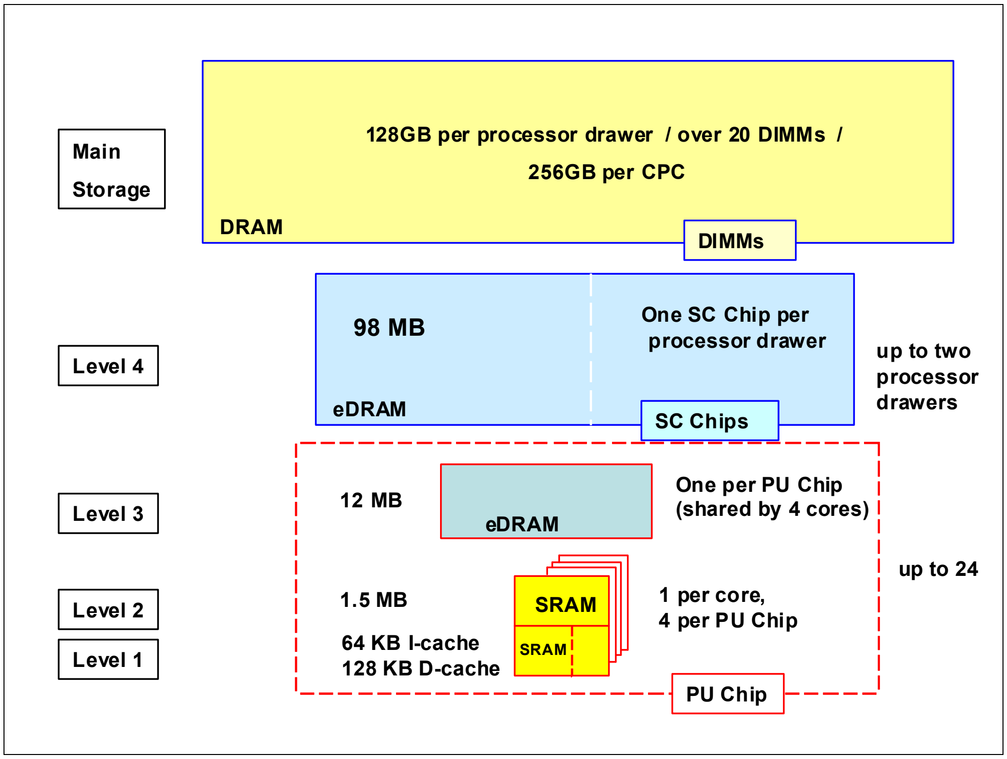


Figure 3-1 z114 cache levels and memory hierarchy

The 4-level cache structure is implemented within the SCM. 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.

> **Models:** z114 M05 has two of them, resulting in 24 MB (12 MB x 2), and z114 M10 has 4 of them, resulting in 48 MB (12 MB x 2 x 2 drawers).

- L4 cache also uses eDRAM and is shared by all PU chips on the SCM. The z114 M05 has 96 MB, and the z114 M10 has 192 MB (2 x 96 MB) of shared L4 cache.
- Main storage: z114 M05 has up to 128 GB using up to 10 DIMMs and z114 M10 has up to 256 GB using up to 20 DIMMs.

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

Also, the distance to remote caches as seen from the microprocessor becomes a significant factor. An example is the L4 cache, which is not on the microprocessor. Although the L4 cache is rather large, the reduced cycle time results in more cycles needed to travel the same distance.

In order to overcome this distance and avoid potential latency, zEnterprise CPC uses two additional cache levels (L2 and L3) within the PU chip, with denser packaging. 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. On z114 M10, the L4 cache straddles up to two processor drawers. Therefore, 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 processor drawer 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 processor drawer as the L4 cache. Have 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 processor drawer 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 z114. The cooperation between z/OS and PR/SM has been bundled in a function called HiperDispatch. HiperDispatch exploits the new z114 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 89.

Figure 3-2 compares the cache structures of the z114 with z10 BC.

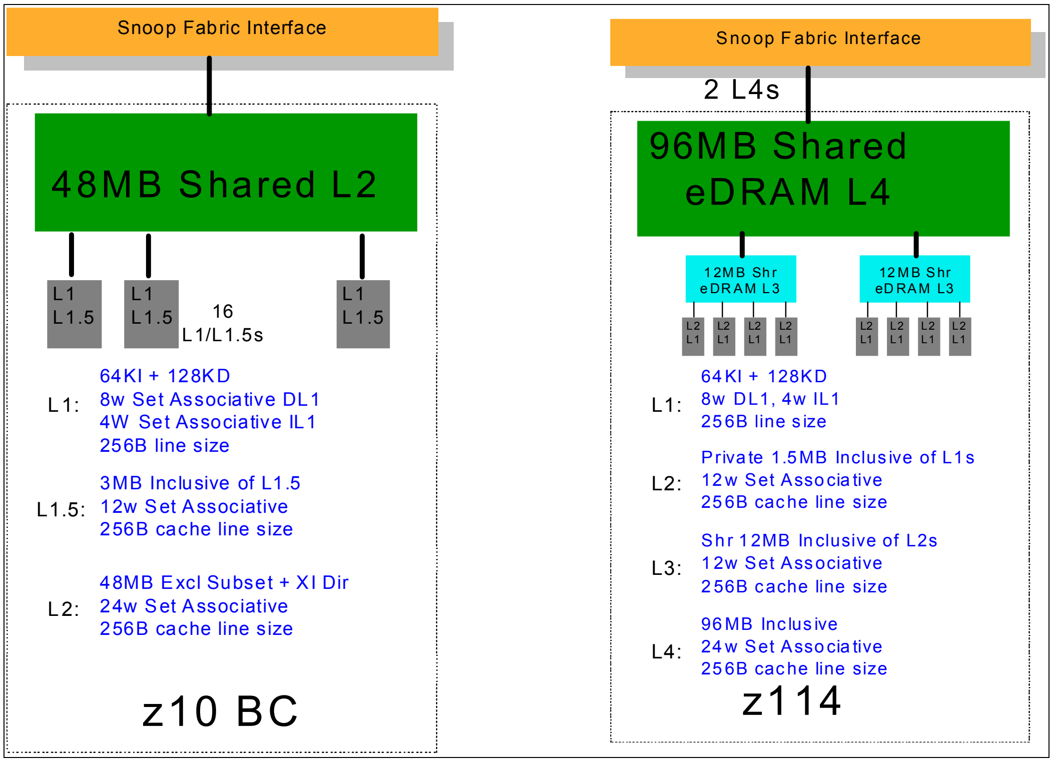


Figure 3-2 z10 BC (z196) and z114 cache level comparison

Compared to the z10 BC, the z114 cache design has one more shared level. The z114 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 Processor drawer interconnect topology

z114 is built from a subset of the z196 design and chip set. Two processor drawers are connected to each other. Figure 3-3 shows a simplified topology for the z114 internal system structure.

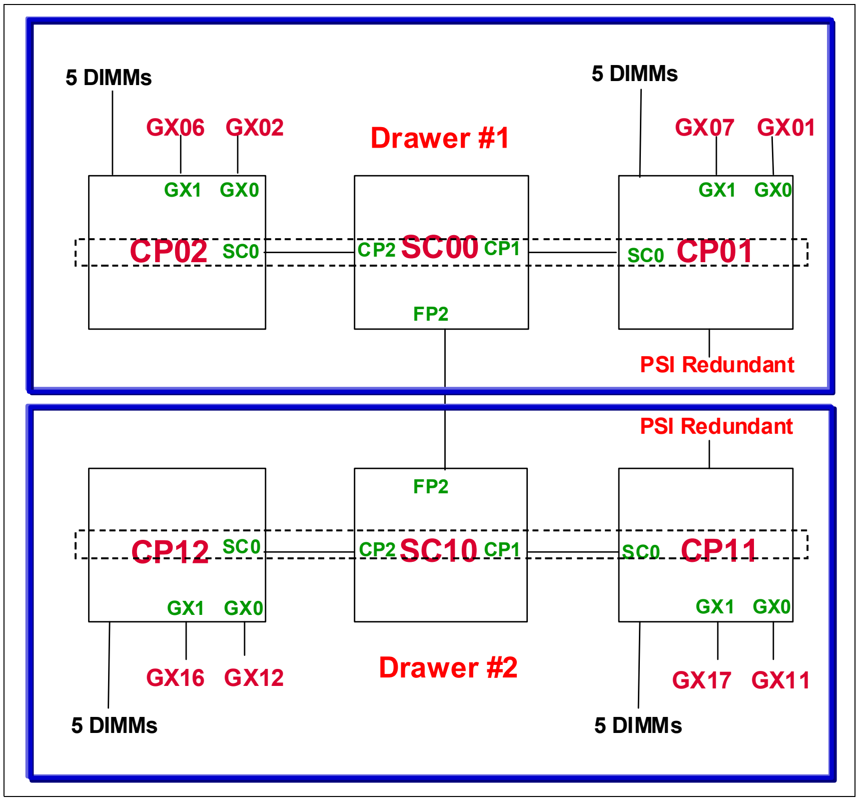


Figure 3-3 z114 internal system structure

The processor drawers' communication takes place at the L4 cache level, which is implemented on SC cache chips in each SCM. The SC function regulates coherent data traffic between the processor drawers.

3.3 Processor unit design

Today, systems design is driven by processor cycle time, although this design does not automatically mean that the performance characteristics of the system improve. Processor cycle time is especially important for CPU-intensive applications. The z114 system resources are powered by up to 14 microprocessors running at 3.8 GHz. The z114 is designed with improved scalability, performance, security, resiliency, availability, and virtualization. The z114 provides up to an 18% improvement in uniprocessor speed and a 12% increase in total system capacity for z/OS, z/VM, and Linux on System z over the z10 BC.

Besides the cycle time, other processor design aspects, such as pipeline, execution order, branch prediction, and high-speed buffers (caches), are also important for the performance of the system. Each z114 processor unit core is a superscalar, out of program order processor, having six execution units. There are 246 complex instructions executed by millicode and another 211 complex instructions broken down into multiple operations. There are 100 new instructions that will help to deliver CPU-centric performance.

The z114 is expected to exhibit up to 25% performance improvement for CPU-intensive C/C++ programs, based on measurements of z196 performance and projections of z114 performance, when the programs are recompiled using the z/OS V1.12 level of XL C/C++. This comparison is against the performance expected for programs that were last compiled using the z/OS V1.8 level, or an earlier level, of C/C++.

z114 introduces architectural extensions, with new instructions to allow reduced processor quiesce effects, reduced cache misses, and reduced pipeline disruption. The z114 new PU architecture includes the following features:

- New truncation with OR inexactness Binary Floating Point rounding mode.
- New Decimal Floating Point quantum exception, which eliminates the need for a test data group for every operation.
- Virtual Architecture Level, which allows the z/VM Live Guest Relocation Facility to make a z114 behave architecturally like a System z10, and facilitates moving work transparently between z114 and System z10 for backup and capacity reasons.
- On a non-quiescing set storage key extended (SSKE) instruction, with significant performance improvement for systems with large number of PUs. It also improves the 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 higher versions, and planned potentially to be exploited by Linux and z/VM in the future).
- Other minor architecture: RRBM, Fast-BCR-Serialization Facility, Fetch-Store-Access Exception Indicator, and CMPSC Enhancement Facility.

The z114 new instruction set architecture (ISA) includes 110 new instructions that have been added to improve compiled code efficiency:

- High-Word Facility (30 new instructions), with independent addressing to high word of 64-bit general purpose registers (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 (separate from 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 counting than prior software implementations
- Integer to/from Floating point converts (39 new instructions)

These new instructions result in 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 z114 supports implementing an out-of-order (OOO) program execution as well. OOO yields significant performance benefit for compute-intensive applications through reordering instruction execution, allowing later (younger) instructions to be executed ahead of a stalled instruction, and reordering storage accesses and parallel storage accesses. OOO maintains good performance growth for traditional applications. Out-of-order (OOO) execution can improve performance in these ways:

- Reordering instruction execution

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

- Reordering storage accesses

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

The 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 z114 PU core is shown in Figure 3-4 on page 69.

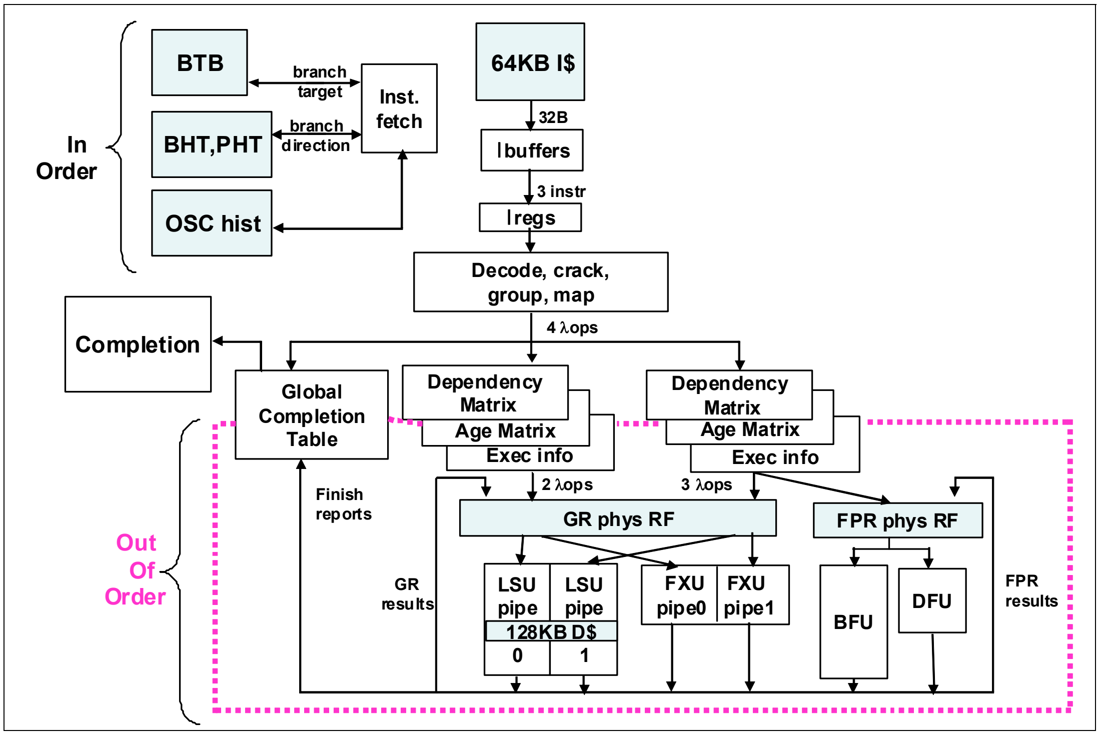


Figure 3-4 z114 PU core logical diagram

Memory address generation and memory accesses can occur out of (program) order. This capability can provide a greater exploitation of the z114 superscalar core and can improve system performance. Figure 3-5 shows how OOO core execution can reduce the execution time of a program.

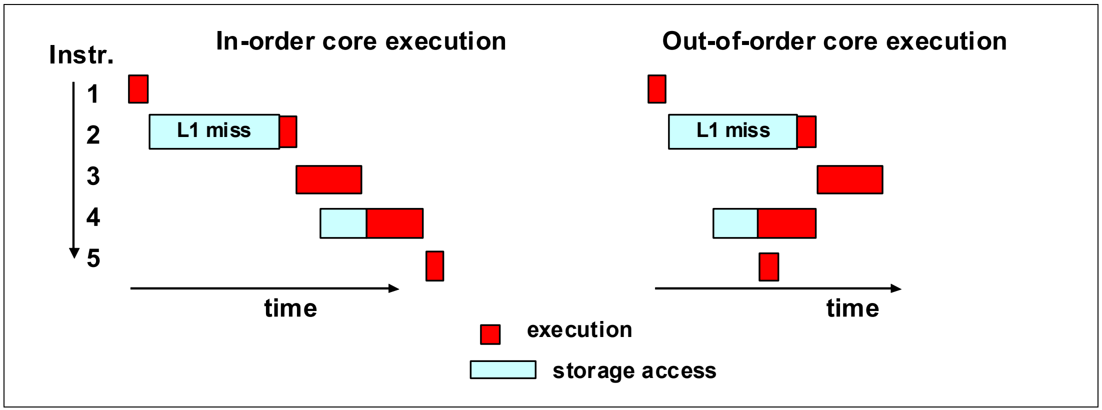


Figure 3-5 In-order and out-of-order core execution

The left side of the example shows an in-order core execution. Instruction 2 has a big delay due to an L1 miss, and the next instructions wait until instruction 2 finishes. In the usual in-order execution, the next instruction waits until the previous one finishes. Using OOO core execution, which is shown on the right side of the example, instruction 4 can start its storage access and execution while instruction 2 is waiting for data, if no dependencies exist between both instructions. So, when the L1 miss is solved, instruction 2 can also start its execution while instruction 4 is executing. Instruction 5 might 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 z114 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 z114, up to three instructions can be decoded per cycle and up to five instructions or 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 wrong branch prediction is more costly in a high-frequency processor design, as we discussed previously. Therefore, the branch prediction techniques used are extremely 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 z114 PU core logical diagram in Figure 3-4 on page 69. 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

In this section, we discuss the compression and cryptography features.

Coprocessor units

There are two coprocessor units for compression and cryptography on each quad-core chip (marked in Figure 3-6 on page 71). 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 64 KB, with 8 K 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.

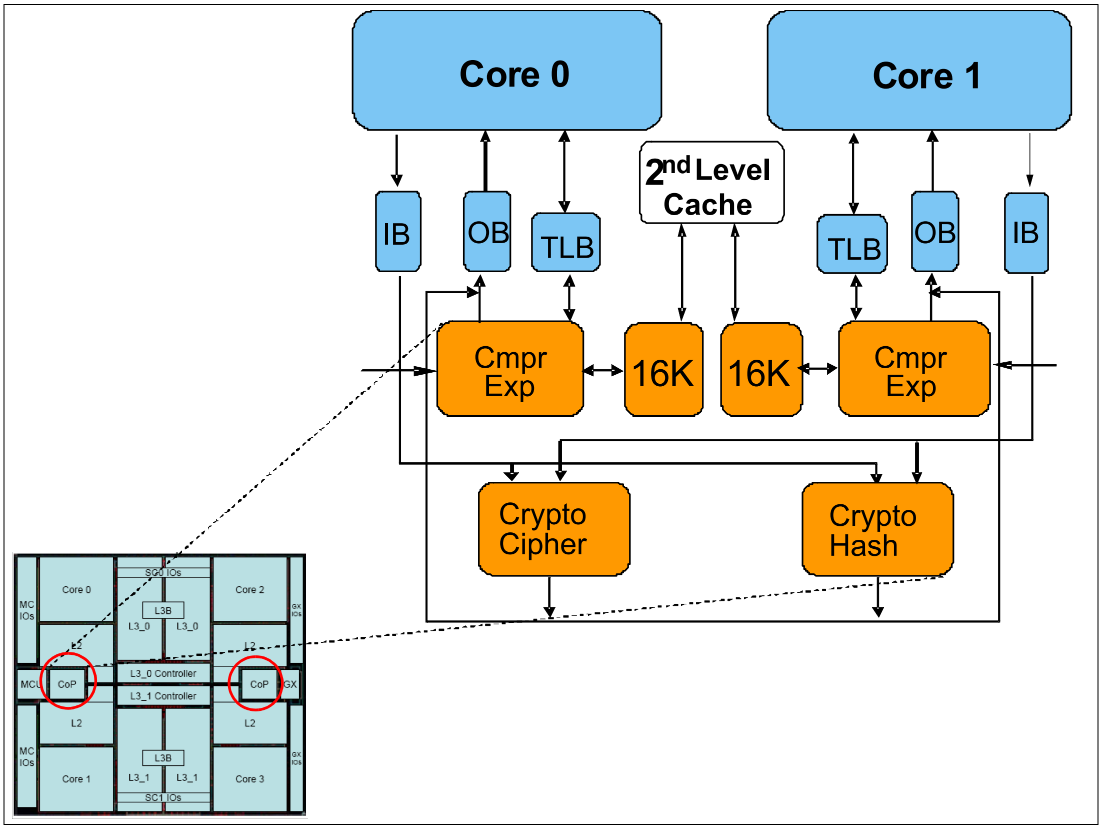


Figure 3-6 Compression and cryptography accelerators on a quad-core chip

CP assist for cryptographic function

The CP assist for cryptographic function (CPACF) accelerates the encrypting and decrypting of SSL/TLS transactions and VPN-encrypted data transfers. The assist function uses a special instruction set for symmetrical clear key cryptographic encryption and decryption, as well as for hash operations. This group of instructions is known as the Message-Security Assist (MSA). For information about the instructions (and micro-programming), see the IBM Resource Link website, which requires registration:

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

For more information about cryptography on z114, see Chapter 6, “Cryptography” on page 159.

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 situation 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 the DFP accelerator

The DFP accelerator offers the following benefits:

- Avoids rounding issues, such as those issues 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 various programming languages:

- 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 a process 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 3-7.

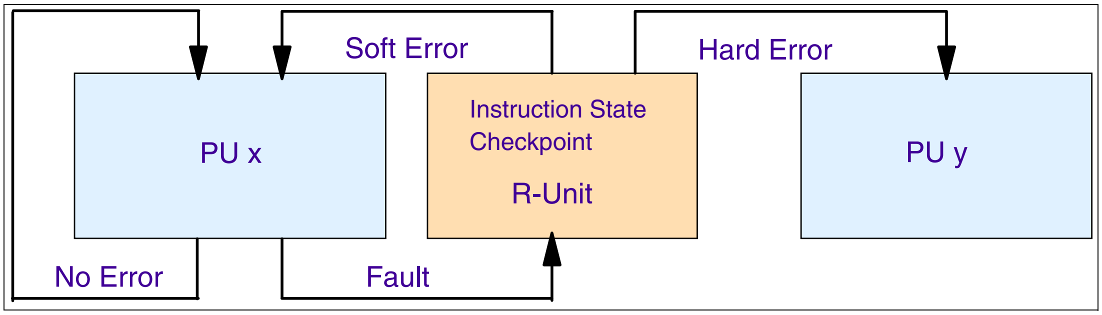


Figure 3-7 PU error detection and recovery

3.3.6 Branch prediction

Because of the ultra high frequency of the PUs, the penalty for a wrongly predicted branch is high. So a multi-pronged strategy for branch prediction, based on gathered branch history combined with 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 z114 uses a variety of techniques to improve the prediction of the correct branch to be executed. The following techniques are included:

- 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 z114. 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 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. Therefore, this approach decreases the many debugging steps that are necessary when determining from where the branch came.

3.3.8 IEEE floating point

Over 130 binary and hexadecimal floating-point instructions are present in z114. 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. Therefore, 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 uses 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 87.

3.3.10 Instruction fetching, decoding, and grouping

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

Instruction fetching

Instruction fetching normally tries to get as far ahead of instruction decoding and execution as possible because of the relatively large instruction buffers that are available. In the microprocessor, smaller instruction buffers are used. The operation code is fetched from the 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 Java virtual machines (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. 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:

- 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 half words used for Unicode).
- Cryptographic instructions, which is known as the Message-Security Assist (MSA), offers the full complement of the AES, SHA, and DES algorithms along with 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

In this section, we describe the processor unit (PU) functions.

3.4.1 Overview

All PUs on a z114 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 that is 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 z114.

This design brings outstanding flexibility to the z114 server, because any PU can assume any available characterization. This design 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.

Also see Chapter 8, “Software support” on page 211 for information about software-level support on functions and features.

Concurrent upgrades

Except on model conversion from M05 to M10, you can perform concurrent upgrades by the Licensed Internal Code (LIC), which assigns a PU function to a previously non-characterized PU. Within the processor drawer boundary or boundary of two processor drawers, no hardware changes are required, and the upgrade can be done concurrently through the following facilities:

- 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 SCMs in the installed processor drawer have no available remaining PUs, an upgrade results in a model upgrade and the installation of an additional processor drawer (up to the limit of two processor drawers). Processor drawer installation is disruptive.

For more information about Capacity on Demand, see Chapter 9, “System upgrades” on page 275.

PU sparing

The PU sparing on z114 M05 is based on prior Business Class (BC) offerings. Because of no dedicated spares, in the rare event of a PU failure, the failed PU’s characterization is dynamically and transparently reassigned to another PU. Because no designated spare PUs are in the z114 M05, an unassigned PU is used as a spare when available. The PUs can be used for sparing any characterization, such as CP, IFL, ICF, zAAP, zIIP, or SAP.

The PU sparing on z114 M10 is based on Enterprise Class (EC) offerings. There are two dedicated spare PUs on a z114 server. PUs that are not characterized on a server configuration are also used as additional spare PUs. More information about PU sparing is provided in 3.4.11, “Sparing rules” on page 85.

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 approach 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 76.

All assigned PUs are grouped together in the PU pool. These PUs are dispatched to online logical PUs. As an example, consider a z114 with five CPs, one zAAP, one IFL, one zIIP, and one ICF. This system has a PU pool of nine PUs, which is called the *pool width*. Subdivision of the PU pool defines these pools:

- A CP pool of five CPs
- An ICF pool of one ICF
- An IFL pool of one IFL
- A zAAP pool of one zAAP
- A zIIP pool of one zIIP

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

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 separate 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. Therefore, 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 the *zEnterprise 196 Processor Resource/Systems Manager Planning Guide*, SB10-7155.

3.4.2 Central processors

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

The z114 can only be initialized in logical partition (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 five logical CPs defined (the sum of active and reserved logical CPs).

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 z114 has 26 capacity levels, which are named from A to Z. Within each capacity level, a one-, two-, three-, four-, and five-way model is offered, each of which is identified by its capacity level indicator (A through Z) followed by an indication of the number of CPs available (01 to 05). This way, the z114 offers 130 capacity settings. All models have a related MSU value that is used to determine the software license charge for MLC software.

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

See 2.8, “Model configurations” on page 51, for more details about granular capacity.

3.4.3 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 five PUs can be characterized as IFLs on M05 and up to 10 PUs can be characterized as IFLs on M10. 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 z114 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 3394).

IFL pool

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 z114. 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 3399). When the system is subsequently upgraded with an additional IFL, the system recognizes that an IFL was already purchased and is present.

3.4.4 Internal coupling facilities

An Internal Coupling Facility (ICF) is a PU that is used to run the coupling facility control code (CFCC) for Parallel Sysplex environments. Up to five ICFs can be characterized on M05 and up to 10 ICFs can be characterized on M10. ICFs are orderable by feature code (FC 3395).

Only CFCC can run on ICFs. ICFs do not change the model capacity identifier of the z114. Software product license charges that are based on the model capacity identifier are not affected by the addition of ICFs.

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

Coupling facility combinations

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 desirable production configuration. It is preferable for a production CF to operate by using dedicated ICFs.

In Figure 3-8, the server on the left has two defined environments (production and test), each having one z/OS and one coupling facility image. The coupling facility images share the same ICF.

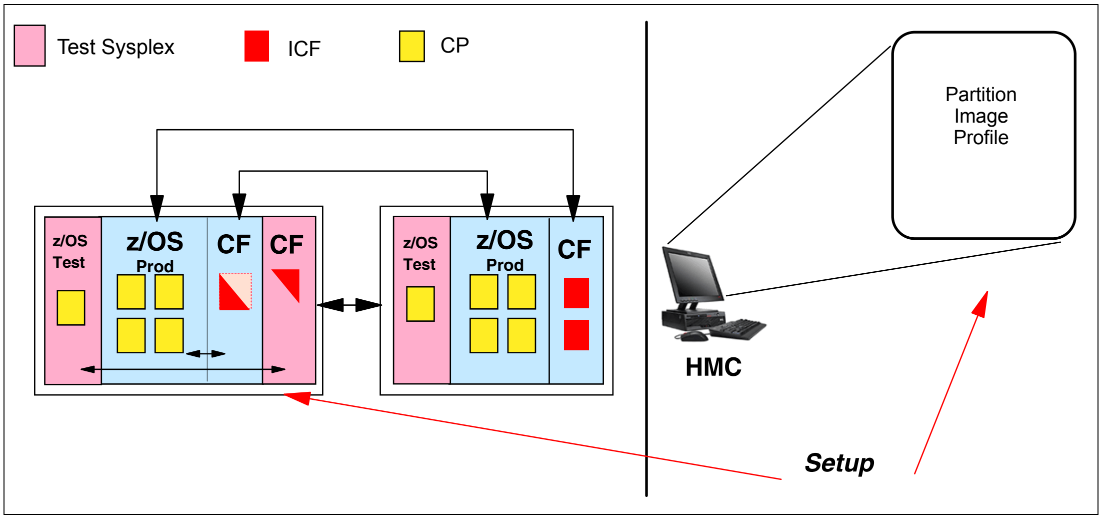


Figure 3-8 ICF options; shared ICFs

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 few processor resources. When the backup CF becomes active, only the resource that is necessary to provide coupling is allocated.

3.4.5 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 Software Developer Kit (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 that is performed in TCB mode is eligible to be executed on the zAAP processors.

zAAPs include the following benefits:

- 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 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. Up to two zAAPs can be characterized on M05 and up to five zAAPs can be characterized on M10.

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

zAAPs are orderable by feature code (FC 3397). 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. 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.

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

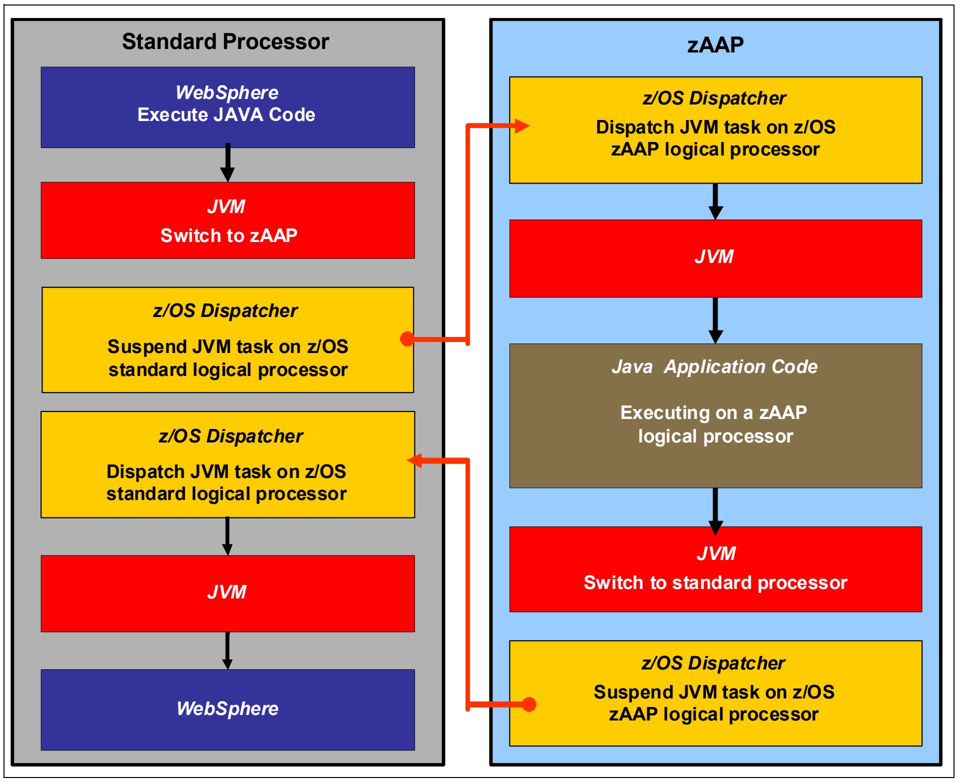
How zAAPs work

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

Figure 3-9 shows the logical flow of Java code running on a z114 that has a zAAP available. When JVM starts the execution of a Java program, it passes control to the z/OS dispatcher that will verify the availability of a zAAP.

Availability is treated in the following manner:

- If a zAAP is available (not busy), the dispatcher suspends the JVM task on the CP and assigns 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 can allow a Java task to run on a standard CP, depending on the option that is used in the OPT statement in the IEAOPTxx member of SYS1.PARMLIB.



Standard Processor
WebSphere
Execute JAVA Code
JVM
Switch to zAAP
z/OS Dispatcher
Suspend JVM task on z/OS standard logical processor
z/OS Dispatcher
Dispatch JVM task on z/OS standard logical processor
JVM
WebSphere
zAAP
z/OS Dispatcher
Dispatch JVM task on z/OS zAAP logical processor
JVM
Java Application Code
Executing on a zAAP logical processor
JVM
Switch to standard processor
z/OS Dispatcher
Suspend JVM task on z/OS zAAP logical processor

Figure 3-9 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 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 so 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 following current options are supported for z/OS V1R8 and later releases:

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

  This option 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 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 rule is that zAAP-eligible work sometimes runs on a CP to resolve resource conflicts, and for 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, 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.

3.4.6 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.

Here, we list several eligible DB2 UDB for z/OS V8 (and later) workloads executing in SRB mode:

- 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, which is mostly used in Business Intelligence (BI) work. A star schema is a relational database schema for representing multidimensional data. It stores data in a central fact table and is surrounded by additional dimension tables holding information about each perspective of the data. A star schema query, for example, joins various dimensions of a star schema data set.
- DB2 utilities that are used for index maintenance, such as LOAD, REORG, and REBUILD. Indexes allow quick access to table rows, but over time as data in large databases is manipulated, 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, separate variables determine how much work is actually redirected to the zIIP.

The z/OS communications server exploits the zIIP for eligible Internet protocol security (IPSec) network encryption workloads. This function 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 the 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 function 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.

The 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 that are based on XML, HTTP, SOAP, Java, and so on, as well as traditional file transfer, can benefit.

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

For more information, see the IBM zIIP website:

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. Up to two zIIPs can be characterized on M05 and up to five zIIPs can be characterized on M10.

zIIPs are orderable by feature code (FC 3398). Up to one zIIP can be ordered for each CP or marked CP configured in the server.

PUs that are 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 the 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.

Logical partition: 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.

3.4.7 zAAP on zIIP capability

zAAPs and zIIPs support separate 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 capability of combining zAAP and zIIP workloads on zIIP processors, provided that no zAAPs are installed on the server. This combination can provide the following benefits:

- The combined eligible workloads can make the zIIP acquisition more cost-effective.
- When zIIPs are already present, the 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. The support is provided by z/OS. See 8.3.2, “zAAP support” on page 225.

When zAAPs are present[1], this capability is not available, because it is not intended as a replacement for zAAPs, which continue to be available, and it is not intended as an overflow possibility for zAAPs. Do not convert zAAPs to zIIPs to take advantage of the zAAP to zIIP capability for the following reasons:

- Having both zAAPs and zIIPs maximizes the system potential for new workloads.
- zAAPs have been available for over 5 years and there might exist applications or middleware with zAAP-specific code dependencies. For example, the code can use the number of installed zAAP engines to optimize multithreading performance.

It is a good idea to plan and test before eliminating all zAAPs, because application code dependencies can exist that might affect performance.

3.4.8 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. Both models have two standard SAPs configured and up to two additional SAPs.

SAP configuration

A standard SAP configuration provides a well-balanced system for most environments. However, there are application environments with high I/O rates (typically various Transaction Processing Facility (TPF) environments). In this case, optional additional SAPs can be ordered. Assignment of additional SAPs can increase the capability of the channel subsystem to perform I/O operations. In z114 servers, the number of SAPs can be greater than the number of CPs.

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

Optional additional orderable SAPs

An available option on all models is additional orderable SAPs (FC 3396). These additional SAPs increase the capacity of the channel subsystem to perform I/O operations, usually suggested for Transaction Processing Facility (TPF) environments.

Optionally assignable SAPs

Assigned CPs can 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 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 available CPs 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 available CPs. 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 available CPs.

3.4.9 Reserved processors

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

Reserved processors can be dynamically configured online by an operating system that supports this function, if enough unassigned PUs are available to satisfy 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. Therefore, you can 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.

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

Do not define more active and reserved processors than the operating system for the logical partition can support. For more information about logical processors and reserved processors and their definition, see 3.6, “Logical partitioning” on page 89.

3.4.10 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 rule is to keep PUs of the same characterization type grouped together as much as possible regarding PU chips to optimize shared cache usage.

The assignment rules follow this order:

- Spares: No dedicated spare PU resides on M05. Two dedicated spare PUs reside on M10, where each processor drawer has one.

- SAPs: Spread across processor drawers and high PU chips. Start with the high PU chip high core, then the next PU chip high core, which prevents all the SAPs from being assigned on one PU chip.
- CPs: Fill PU chip and spill into next chip on the low processor drawer first before spilling over into the high processor drawer.
- ICFs: Fill the high PU chip on the high processor drawer.
- IFLs: Fill the high PU chip on the high processor drawer.
- zAAPs: Attempts are made to align these zAAPs close to the CPs.
- zIIPs: Attempts are made to align these zIIPs close to the CPs.

This implementation is to isolate as much as possible on separate processor drawers (and even on separate PU chips) processors that are used by separate 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 from sharing L3 and L4 caches with z/OS processors.

This initial PU assignment that is done at POR can be dynamically rearranged by LPAR to improve system performance (see 3.6.2, “Storage operations” on page 94).

3.4.11 Sparing rules

On a z114 M05, because of no dedicated spare PU, non-characterized PUs are used for sparing.

On a z114 M10, two dedicated spare PUs are available, one for each processor drawer. The two spare PUs can 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.

Follow these sparing rules:

- 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.

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

If a spare PU is available, 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 client 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 that is used by the operating system and, through operating system recovery services, the task is resumed successfully (in most cases, without client 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, client intervention is not required. This capability eliminates an unplanned outage and permits a service action to be deferred to a more convenient time.

3.4.12 Increased flexibility with z/VM-mode partitions

The z114 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 perform these functions:

- 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

In this section, we describe various considerations regarding the z114 memory design.

3.5.1 Overview

The z114 memory design supports concurrent memory upgrades up to the limit provided by the physically installed capability.

The z114 can have more physically installed memory than the initial available capacity. Memory upgrades within the physically installed capacity can be done concurrently by LIC, and no hardware changes are required. Note that memory upgrades *cannot* be done through Capacity BackUp (CBU) or On/Off CoD.

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 upgrade is disruptive if the physically installed capacity is reached.

Large page support

By default, page frames are allocated with a 4 KB size. The z114 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 desirable to have your 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. Therefore, 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. The reason 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.

Extremely 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 are required to resolve all the memory that it needs. So, an extremely long-running process can benefit somewhat even without frequent memory access. You must weigh the benefits of whether work in this category must use large pages as a result of the system-level costs of tying up real storage. There is a balance between the performance of a process 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.

It appears that increasing the TLB size is a feasible option to deal with TLB-miss situations. However, this approach 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 complex statistical modelling in order to find the optimal trade-off between size and performance.

3.5.2 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 these functions:

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

Central storage can be accessed by all processors, but it 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.

3.5.3 Expanded storage

Expanded storage can optionally be defined on z114. *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.

Storage considerations

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 z114 only runs in LPAR mode. Storage is placed into a single storage pool called the 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 8 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 do not require a power-on reset.

3.5.4 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 8 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 the space for the following functions:

- Two channel subsystems (CSSs)
- Fifteen logical partitions in each CSS for a total of 30 logical partitions
- Subchannel set 0 with 63.75 K devices in each CSS
- Subchannel set 1 with 64 K devices in each CSS

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

3.6 Logical partitioning

In this section, we discuss logical partitioning features.

3.6.1 Overview

Logical partitioning (LPAR) is a function implemented by the Processor Resource/Systems Manager (PR/SM) on all z114 servers. The z114 runs only in LPAR mode. Therefore, all system aspects are controlled by PR/SM functions.

PR/SM is aware of the processor drawer structure on the z114. 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 in 3.4.10, “Processor unit assignment” on page 84, the initial PU assignment is done during POR, using rules to optimize cache usage. This step is the “physical” step, where CPs, zIIPs, zAAPs, IFLs, ICFs, and SAPs are allocated on the processor drawer.

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

Memory allocation is spread across both processor drawers. 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.

PR/SM optimizes chip assignments within the assigned processor drawer, 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 the same processor group, keeping processes running near their cached instructions and data, and minimizing transfers of data ownership among processors.

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

z/OS dispatcher manages multiple queues, which are 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.

Logical partitions

PR/SM enables z114 servers to be initialized for a logically partitioned operation, supporting up to 30 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). Certain 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

  They are called *logical processors*, and 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 z114, which is described in Chapter 9, “System upgrades” on page 275, adds another dimension to the dynamic management of logical partitions.

For the 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:

- 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 under the defined capacity, the soft cap is removed.

For more information regarding WLM, see *System Programmer's Guide to: Workload Manager*, SG24-6472.

> **Weight settings:** 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. Therefore, 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 best to avoid this case. Try to establish a defined capacity that 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 3.6.5, "LPAR dynamic storage reconfiguration" on page 98.

- 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 preferred, tool used to map CHPIDs onto physical channel identifiers (PCHIDs) that represent the physical location of a port on a card in an PCIe I/O drawer 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 3-2 on page 92 shows the modes of operation, summarizing all available mode combinations: operating modes and their processor types, operating systems, and addressing modes.

Table 3-2 z114 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	z/VSE and Linux on System z (64-bit)	64-bit
	CP	Linux on System z (31-bit)	31-bit
ESA/390 TPF mode	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, and 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 image's 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 211.

Logically partitioned mode

The z114 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 these operating systems:
  - 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

> **Important:** zAAPs and zIIPs can be defined to an ESA/390 mode or z/VM-mode image (see Table 3-2 on page 92). 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 V5R4 and later can provide zAAPs or zIIPs to a guest z/OS.

- ESA/390 TPF mode, to run the 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 3-3 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 3-3 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 (V5R4 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	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.

Partition profile: Cryptographic coprocessors are not tied to partition numbers or MIF IDs. They are set up with Adjunct Processor (AP) numbers and domain indexes, which are assigned to a partition profile of a given name. The client assigns these AP numbers and domains to the partitions and continues to have the responsibility to clear them out when their profiles change.

Adding the 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 Indexes in advance of installation. By using the Change LPAR Cryptographic Controls task, it is possible to add crypto adapters dynamically to a logical partition without an outage of the LPAR. 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 z114 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 feature 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 feature, 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.

3.6.2 Storage operations

In z114 servers, memory can be assigned as a combination of central storage and expanded storage, supporting up to 30 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.

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 88.

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 function 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 3.6.5, “LPAR dynamic storage reconfiguration” on page 98.

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 3-4 on page 95 shows the z114 storage *allocation* and *usage* possibilities, depending on the image mode.

Table 3-4 Storage definition and usage possibilities

Image mode	Architecture mode (addressability)	Maximum central storage		Expanded storage	
		Architecture	**z114 definition**	**z196 definable**	**Operating system usage**[1]
ESA/390	z/Architecture (64-bit)	16 EB	248 GB	Yes	Yes
	ESA/390 (31-bit)	2 GB	128 GB	Yes	Yes
z/VM[2]	z/Architecture (64-bit)	16 EB	248 GB	Yes	Yes
ESA/390 TPF	ESA/390 (31-bit)	2 GB	2 GB	Yes	No
Coupling facility	CFCC (64-bit)	1.5 TB	248 GB	No	No
Linux only	z/Architecture (64-bit)	16 EB	248 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 bits 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.

Certain 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 following modes are provided:

- 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 z114 is 248 GB 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 higher than 2 GB can 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.

Storage resources: Either a z/Architecture mode or an ESA/390 architecture mode operating system can run in an ESA/390 image on a z114. 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, certain types of processor units can be defined within one LPAR. This capability 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 z114 definition limit for logical partitions is 248 GB of storage.

CFCC Level 17, which is available for the z114, allows the following capabilities:

- Greater than 1023 CF structures: New limit is 2047
- Greater than 32 connectors: New limits are 255 cache, 247 lock, or 127 serialized list
- Improved CFCC diagnostics and link diagnostics
- An increase from 64 to 128 CHPIDs

For details, see 3.8.1, “Coupling facility control code” on page 101. 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.

3.6.3 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 storage and expanded storage can have two storage sizes defined:

- The initial value is the storage size that is allocated to the partition when it is activated.
- The reserved value is an additional storage capacity beyond its initial storage size that 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 the following actions:

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 these sources:

- Any unused available storage
- Another partition that has released 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.

3.6.4 Logical partition storage granularity

Granularity of central storage for a logical partition depends on the largest central storage amount that is defined for either initial or reserved central storage, as shown in Table 3-5.

Table 3-5 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

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 the z/OS Reconfigurable Storage Units definition. For z/VM V5R4 and later, the limitation is 256 GB.

3.6.5 LPAR dynamic storage reconfiguration

Dynamic storage reconfiguration on z114 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 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 capability gives WLM the ability to manage resources, both processor and I/O, not only in a single image, but across the entire cluster of system images.

Figure 3-10 on page 99 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.

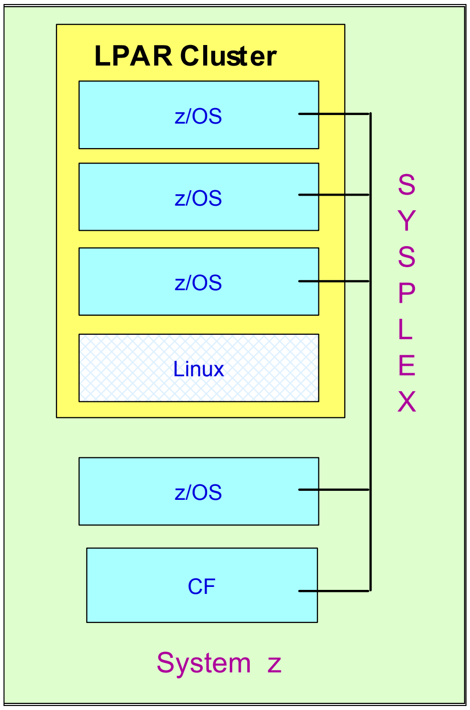


Figure 3-10 IRD LPAR cluster example

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 the WLM goal mode policy.

  We introduced HiperDispatch in 3.6, “Logical partitioning” on page 89. 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. Performing this mapping 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 z114 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 that is used with z114 servers. Figure 3-11 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.

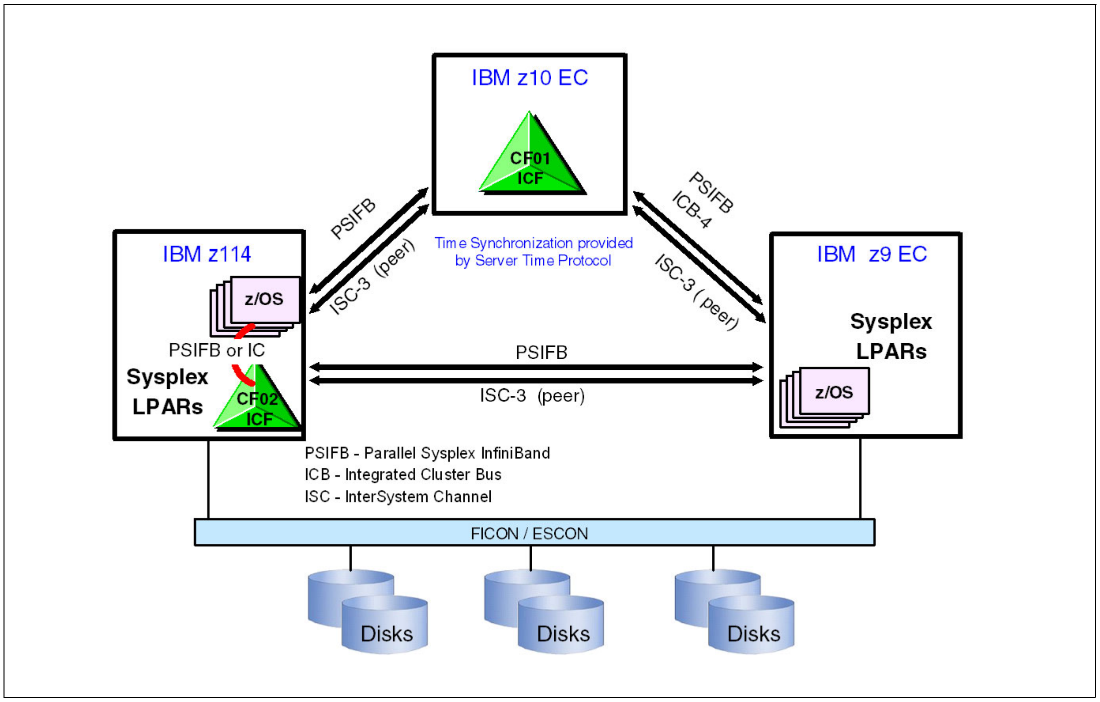


Figure 3-11 Sysplex hardware overview

Figure 3-11 shows a z114 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. Server Time Protocol (STP) over coupling links provides time synchronization to all servers. CF link technology (InfiniBand (IFB), Integrated Cluster Bus (ICB)-4, and InterSystem Channel-3) selection depends on server configuration. We describe link technologies in 4.9.1, “Coupling links” on page 139.

Parallel Sysplex technology is an enabling technology, allowing highly reliable, redundant, and robust System z technology to achieve near-continuous availability. A Parallel Sysplex consists of 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:

- 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 client requirements for continuous availability, and provides techniques for achieving simplified systems management consistent with this requirement.

- Resource sharing

  A number of base (z/OS) components 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.

3.8.1 Coupling facility control code

Coupling facility control code (CFCC) Level 17 is made available on the z114.

CFCC Level 17 allows an increase in the number of CHPIDs from 64 to 128. (This increase applies to IC, 12x IFB, 1x IFB, and active InterSystem Channel (ISC)-3 links.) This constraint relief can help support better CF link throughput, because most[1] coupling CHPIDs carries with them only seven 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 2047 structures. When sysplex was first implemented, only 64 structures were supported in the sysplex. Before long, sysplex exploitation took off, and clients levied requirements up to 1023 structures with CFCC Level 16. New exploiters demanded more structures, for example:

- Logical groupings, such as DB2, IMS, and MQ datasharing groups, for which multiple group instances can exist in the same sysplex (each potentially with many structures)
- “Service provider” clients that provide IT services for many customers, and define large numbers of individual small datasharing groups, one per customer
- Customer mergers, acquisitions, and sysplex consolidations, which often grow the requirements in quantum leaps rather than slow and 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, which 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 (for example, threading constraints, virtual storage constraints, or 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.

[1] Up to 32 subchannels per CHPID for 1x InfiniBand coupling links. For more information, see the description of IFB LR coupling links in 4.9, “Parallel Sysplex connectivity” on page 139.

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 technique 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), using all processor capacity (cycles) is not a problem. However, this technique can be an issue if the LPAR that is running the CFCC also has shared processors. Therefore, it is best to enable dynamic dispatching on the CF LPAR.

3.8.2 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 3-12 shows the dynamic CF dispatching.

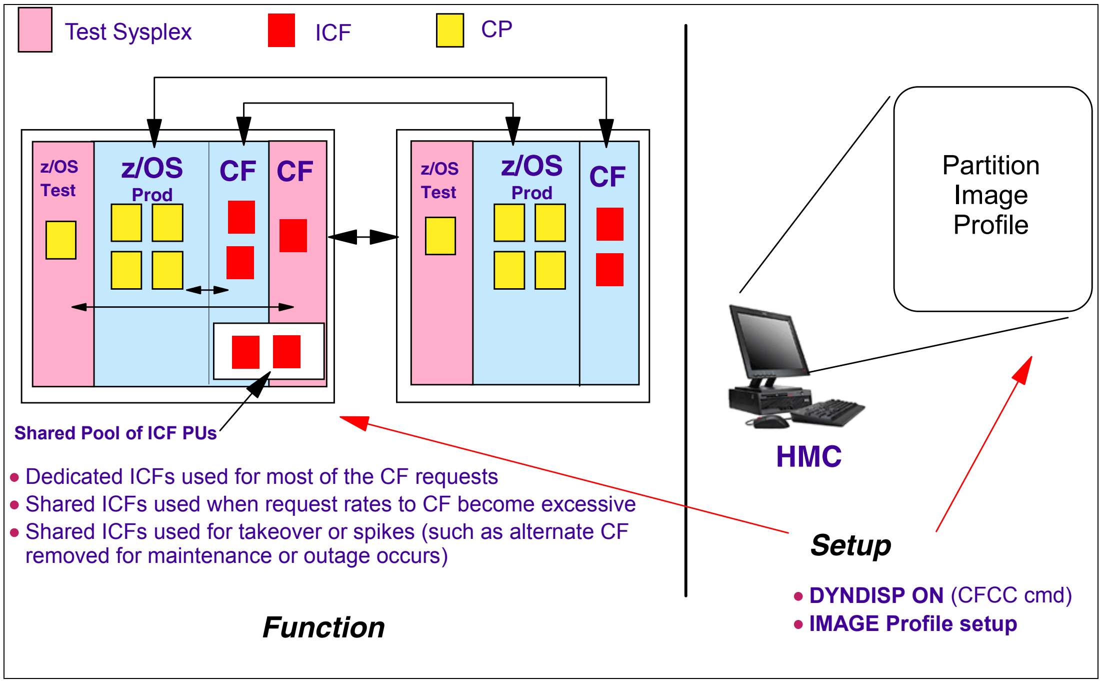

Test Sysplex
ICF
CP
z/OS
Test
z/OS
Prod
CF
CF
z/OS
Test
z/OS
Prod
CF
Shared Pool of ICF PUs
Partition
Image
Profile
HMC
• Dedicated ICFs used for most of the CF requests
• Shared ICFs used when request rates to CF become excessive
• Shared ICFs used for takeover or spikes (such as alternate CF removed for maintenance or outage occurs)
Function
Setup
• DYNDISP ON (CFCC cmd)
• IMAGE Profile setup

Figure 3-12 Dynamic CF dispatching (shared CPs or shared ICF PUs)

For additional details regarding CF configurations, see *Coupling Facility Configuration Options,* GF22-5042, which is also available from the Parallel Sysplex website:

http://www.ibm.com/systems/z/advantages/pso/index.html

4

Central processor complex I/O system structure

In this chapter, we describe the I/O system structure and the connectivity options that are available on the IBM zEnterprise 114 (z114).

We cover the following topics:

4.1 Introduction

The z114 supports two internal I/O infrastructures:

- InfiniBand-based infrastructure for I/O drawers
- PCIe-based infrastructure for PCIe I/O drawers with a new form factor drawer and I/O features

InfiniBand I/O infrastructure

The InfiniBand I/O infrastructure was first made available on System z10 and consists of these components:

- InfiniBand fanouts supporting the current 6 GBps InfiniBand I/O interconnect
- InfiniBand I/O card domain multiplexers with Redundant I/O Interconnect in the 5U, 8-slot, 2-domain IO drawer

PCIe I/O infrastructure

IBM extends the use of industry standards on the System z platform by offering a Peripheral Component Interconnect Express Generation 2 (PCIe Gen2) I/O infrastructure. The PCIe I/O infrastructure that is provided by the zEnterprise CPCs improves I/O capability and flexibility, while allowing for the future integration of PCIe adapters and accelerators.

The z114 PCIe I/O infrastructure consists of these components:

- PCIe fanouts supporting 8 GBps I/O bus interconnections for processor drawer connectivity to the PCIe I/O drawer
- The 7U, 32-slot, 4-domain PCIe IO drawer for PCIe I/O features

The zEnterprise PCIe I/O infrastructure offers these benefits:

- Increased bandwidth from the processor book or drawer to the I/O domain in the PCIe I/O drawer via an 8 gigabytes per second (GBps) bus
- Better granularity for the storage area network (SAN) and the local area network (LAN) For the FICON, zHPF and FCP storage area networks, the FICON Express8S has two channels per feature. The OSA-Express4S GbE features have two ports each and the OSA-Express4S 10 GbE features have one port each for LAN connectivity.

4.1.1 InfiniBand

The InfiniBand specification defines the raw bandwidth of a one lane (which is referred to as 1x) connection at 2.5 Gbps. Two additional lane widths are specified, which are referred to as 4x and 12x, as multipliers of the base link width.

Similar to Fibre Channel, PCI Express, Serial ATA, and many other contemporary interconnects, InfiniBand is a point-to-point, bidirectional serial link that is intended for the connection of processors with high-speed peripherals, such as disks. InfiniBand supports various 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, per physical connection. InfiniBand also supports 5 Gbps or 10 Gbps signaling rates.

Data signalling and link rates

Links use 8b/10b encoding (every ten bits sent carry 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, links carry 2, 4, or 8 Gbps of useful data.

Links can be aggregated in units of 4 or 12, indicated as 4x[1] or 12x. A 12x 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 z114, and for inter-switch connections.

For details and the standards for InfiniBand, see the InfiniBand website:

http://www.infinibandta.org

z114 and InfiniBand: Not all properties and functions offered by InfiniBand are implemented on the z114. Only a subset is used to fulfill the interconnect requirements that have been defined for z114.

4.1.2 PCIe

PCIe is a serial bus with an embedded clock and uses 8b/10b encoding, where every 8 bits are encoded into a 10-bit symbol that is then decoded at the receiver. Thus, the bus needs to transfer 10 bits to send 8 bits of actual usable data. A PCIe bus generation 2 single lane can transfer 5 Gbps of raw data (duplex connection), which is 10 Gbps of raw data in total. From these 10 Gbps, only 8 Gbps are actual data. Therefore, an x16 (16 lanes) PCIe gen2 bus transfers 160 Gbps encoded, which is 128 Gbps of unencoded data. This example is 20 GBps raw data and 16 GBps of encoded data. Now, the new measuring unit GT/s which means Giga Transfers per second refers to the raw data even though only 80% of this transfer is actual data. The translation between GT/s to GBps is 5 GT/s equals 20 GBps or 1 GT/s equals 4 GBps.

The 16 lanes of the PCIe bus are virtual lanes, always consisting of one transmit and one receive lane. Each of these lanes consists of two physical copper wires, because the physical method used to transmit signals is a differential bus, which means that the signal is encoded into the various voltage levels between two wires (as opposed to one voltage level on one wire in comparison to the ground signal). Therefore, each of the 16 PCIe lanes actually uses four copper wires for the signal transmissions.

4.2 I/O system overview

This section lists the z114 I/O subsystem characteristics and a summary of the supported features.

4.2.1 Characteristics

The z114 I/O subsystem design provides great flexibility, high availability, and excellent performance characteristics:

- High bandwidth

  The z114 uses PCIe as new internal interconnect protocol to drive PCIe I/O drawers. The I/O bus infrastructure data rate increases up to 8 GBps.

[1] z114 does not support this data rate.

The z114 uses InfiniBand as the internal interconnect protocol to drive I/O drawers and CPC to CPC connection. InfiniBand supports I/O bus infrastructure data rates up to 6 GBps.

- Connectivity options:
  - The z114 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 LAN connectivity.
  - For CPC to CPC connection, z114 uses Parallel Sysplex InfiniBand (IFB) or InterSystem Channel (ISC)-3 coupling links.
- Concurrent I/O upgrade

  You can concurrently add I/O cards to the server if an unused I/O slot position is available.
- Concurrent I/O drawer upgrade

  Additional I/O drawers can be installed concurrently without preplanning.
- Concurrent PCIe I/O drawer upgrade

  Additional PCIe 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 unit, and I/O devices without a planned outage.
- Pluggable optics

  The FICON Express8, FICON Express8S, 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

  Every I/O card plugged in an I/O drawer or PCIe I/O drawer supports concurrent card replacement in the case of a repair action.

4.2.2 Summary of supported I/O features

The following I/O features are supported (a few of them are carried forward on upgrade only):

- Up to 240 ESCON channels
- Up to 64 FICON Express4 channels
- Up to 32 FICON Express4-2C channels
- Up to 64 FICON Express8 channels
- Up to 128 FICON Express8S channels
- Up to 32 OSA-Express2 ports
- Up to 64 OSA-Express3 ports
- Up to 32 OSA-Express3-2P ports
- Up to 96 OSA-Express4S ports
- Up to 48 ISC-3 coupling links
- Up to 8 InfiniBand fanouts using one of these links:
  - Up to 16 12x InfiniBand coupling links
  - Up to 16 1x InfiniBand coupling links with HCA2-O LR (1xIFB) fanout
  - Up to 24 1x InfiniBand coupling links with HCA3-O LR (1xIFB) fanout

> **Coupling links:** The maximum number of external coupling links combined (ISC-3 and IFB coupling links) cannot exceed 72 for the z114 M10 CPC and 56 for the z114 M05 CPC.

4.3 I/O drawers

The I/O drawer is five EIA units high and supports up to eight I/O feature cards. Each I/O 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 processor drawer.

The link between the HCA in the processor drawer 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. Figure 4-1 shows the locations of the DCAs, I/O feature cards, and IFB-MP cards in the I/O drawer.

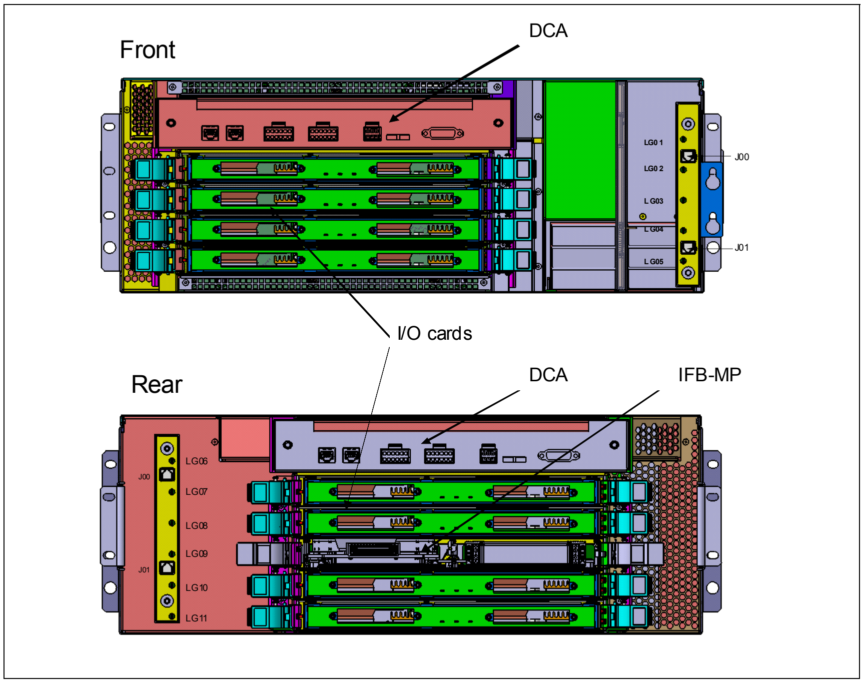


Figure 4-1 I/O drawer

The I/O structure in a z114 server is illustrated in Figure 4-2 on page 108. 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 (RII). RII 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.

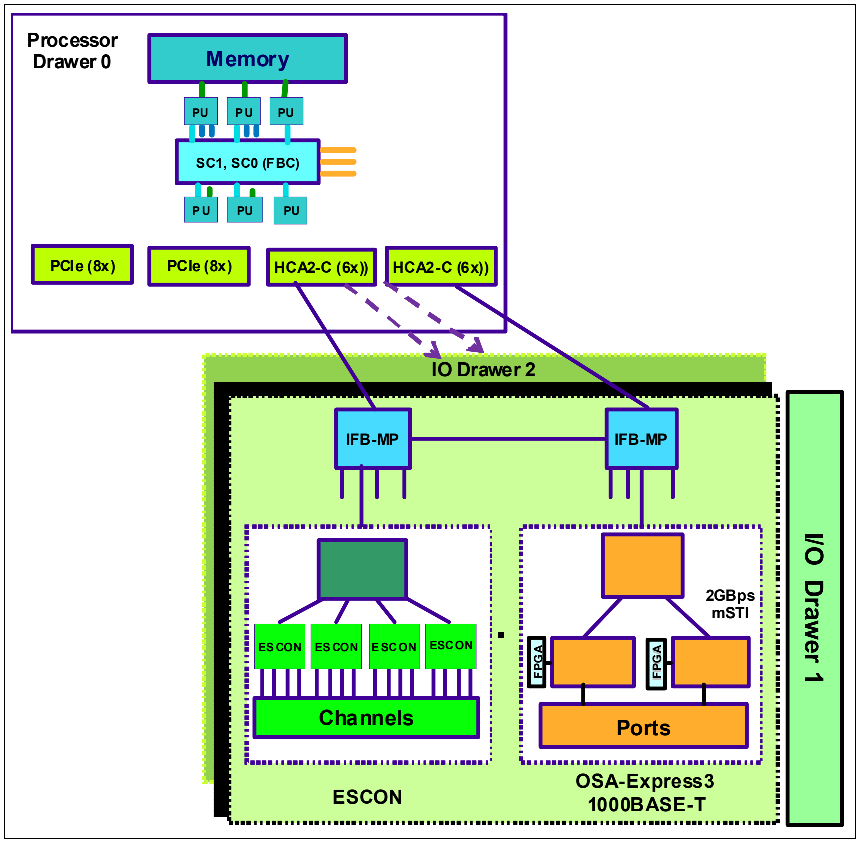


Figure 4-2 z114 I/O structure when using I/O drawers

The I/O drawer domains and their related I/O slots are shown in Figure 4-3. 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 (FICON or OSA). The I/O cards are connected to the IFB-MP card through the backplane board.

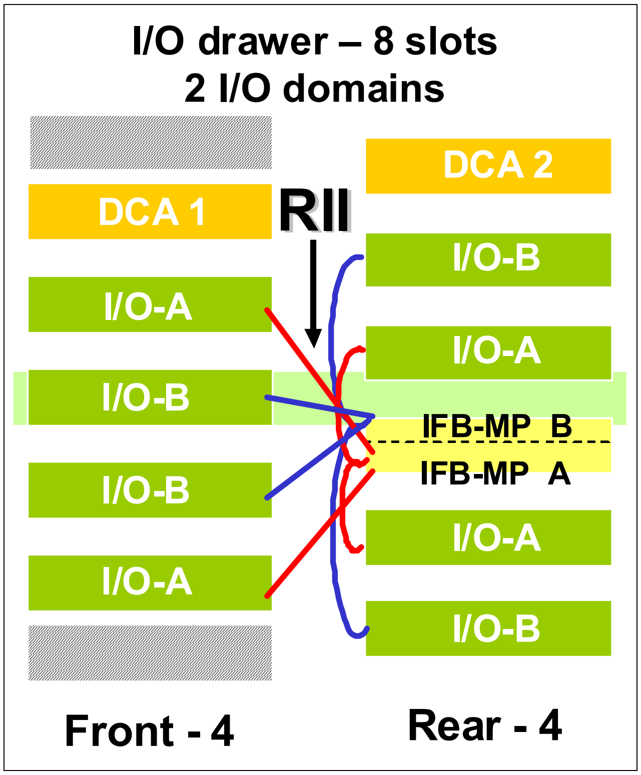


Figure 4-3 I/O domains of I/O drawer

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 4-1 lists the I/O domains and their related I/O slots.

Table 4-1 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.

Power Sequence Controller: It is intended that the z196 and z114 are the last System z servers to support the Power Sequence Controller feature.

4.4 PCIe I/O drawers

The PCIe I/O drawer attaches to the processor node via a PCIe bus and uses PCIe as the infrastructure bus within the drawer. The PCIe I/O bus infrastructure data rate is up to 8 GBps. PCIe switch Application-Specific Integrated Circuits (ASICs) are used to fan out the host bus from the processor node to the individual I/O cards.

The PCIe drawer is a two-sided drawer (I/O cards on both sides) that is 7U high (one half of the I/O cage) and fits into a 7.3 m (24 in.) System z frame. The drawer contains 32 I/O card slots, four switch cards (two in front, and two in rear), two DCAs to provide the redundant power, and two air moving devices (AMDs) for redundant cooling. The locations of the DCAs, AMDs, PCIe switch cards, and I/O feature cards in the PCIe I/O drawer are shown in Figure 4-4.

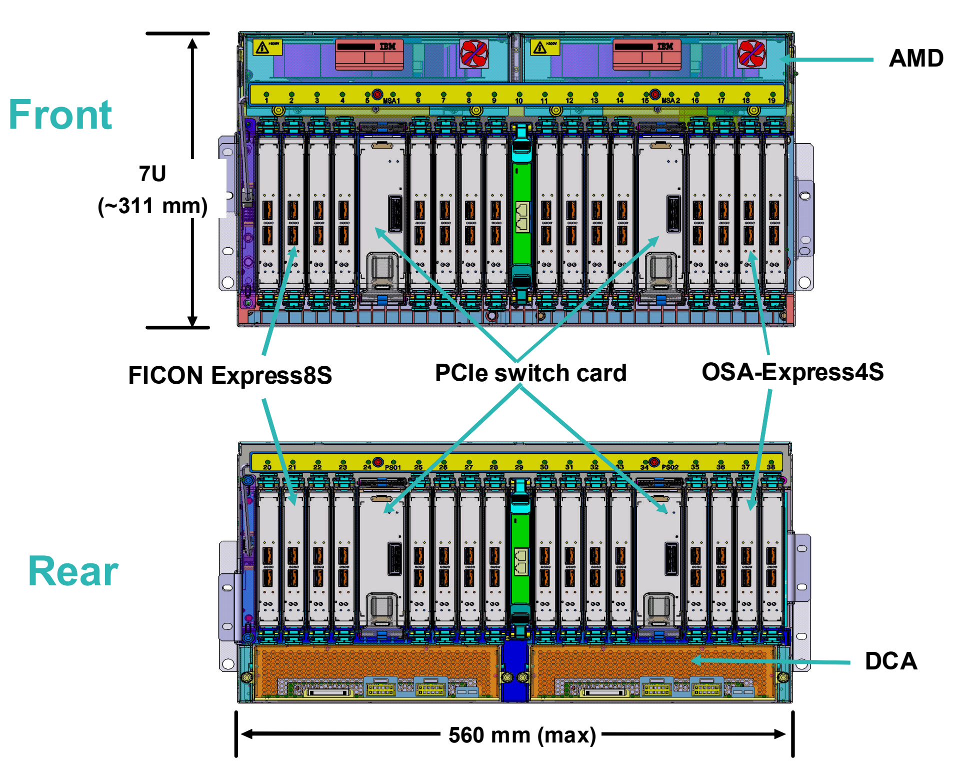


Figure 4-4 PCIe I/O drawer

The I/O structure in a z114 server is illustrated in Figure 4-5. The PCIe switch card provides the fanout from the high speed x16 PCIe host bus to eight individual card slots. The PCIe switch card is connected to the processor nest via a single x16 PCIe Gen 2 bus from a PCIe fanout card, which converts the processor drawer internal bus into two PCIe buses.

A switch card in the front connects to a switch card in the rear through the PCIe I/O drawer board to provide a failover capability in case of a PCIe fanout card failure. In the PCIe I/O drawer, the eight I/O cards that directly attach to the switch card constitute an I/O domain. The PCIe I/O drawer supports concurrent add and delete to enable a client to increase I/O capability as needed without planning ahead.

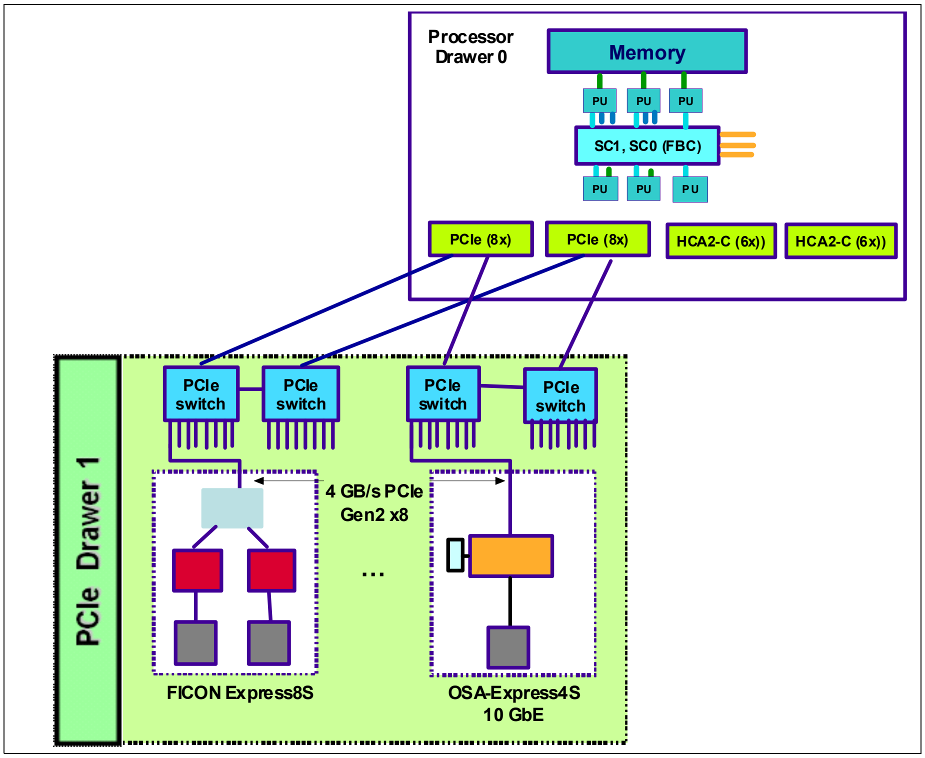


Figure 4-5 z114 I/O structure when using PCIe I/O drawer

The PCIe I/O drawer supports up to 32 I/O cards. They are organized in four hardware domains per drawer. Each domain is driven through a PCIe switch card. Two PCIe switch cards always provide a backup path for each other through the passive connection in the PCIe I/O drawer backplane, so that in the case of a PCIe fanout card or cable failure, all 16 I/O cards in the two domains can be driven through a single PCIe switch card.

To support redundant I/O interconnect (RII) between the front to back domain pairs 0 - 1 and 2 - 3, the two interconnects to each pair must be from two separate PCIe fanouts (all four domains in one of these cages can be activated with two fanouts). The flexible service processors (FSPs) are used for system control.

The PCIe I/O domains and their related I/O slots are shown in Figure 4-6 on page 112.

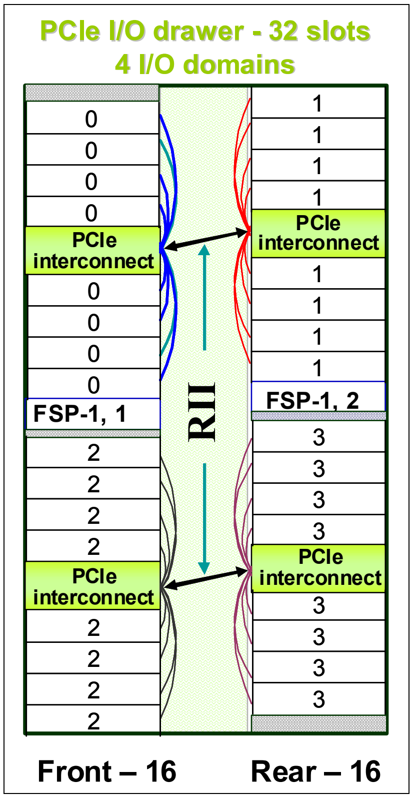


Figure 4-6 I/O domains of PCIe I/O drawer

> **Power Sequence Controller:** The PCIe I/O drawer does not support the Power Sequence Controller (PSC) feature.

4.5 I/O drawer and PCIe I/O drawer offerings

The I/O drawers for z114 cannot be ordered. I/O feature types determine the appropriate mix of I/O drawers and PCIe I/O drawers.

All new system builds, migration offerings, and exchange programs include FICON Express8S and OSA-Express4S features. Crypto Express3, ESCON, ISC-3, OSA-Express3 1000BASE-T, and PSC features cannot be used in the PCIe I/O drawer.

4.6 Fanouts

The z114 server uses fanout cards to connect the I/O hardware subsystem to the CEC drawers and to provide the InfiniBand coupling links for Parallel Sysplex, as well. All fanout cards support concurrent add, delete, and move.

z114 supports two separate internal I/O infrastructures for the internal connection. The z114 uses InfiniBand-based infrastructure for the internal connection to I/O drawers. The z114 uses PCIe-based infrastructure for PCIe I/O drawers in which the cards for the connection to peripheral devices and networks reside.

The InfiniBand and PCIe fanouts are located in the front of the processor drawer. Each processor drawer has four fanout slots. Six types of fanout cards are supported by z114. Each slot holds one of the following six fanouts:

- Host Channel Adapter (HCA2-C)

  This copper fanout provides connectivity to the IFB-MP card in the I/O drawer.
- PCIe Fanout

  This copper fanout provides connectivity to the PCIe switch card in the PCIe I/O drawer.
- Host Channel Adapter (HCA2-O (12xIFB))

  This optical fanout provides 12x InfiniBand coupling link connectivity up to 150 m (492.1 ft.) distance to a z196, z114, System z10, and System z9.
- Host Channel Adapter (HCA2-O LR (1xIFB))

  This optical long-range fanout provides 1x InfiniBand coupling link connectivity up to 10 km (6.2 miles) unrepeated distance to a z196, z114, or System z10 server.
- Host Channel Adapter (HCA3-O (12xIFB))

  This optical fanout provides 12x InfiniBand coupling link connectivity up to 150 m (492.1 ft.) distance to a z196, z114, and System z10. It cannot communicate with an HCA1-O fanout on z9.
- Host Channel Adapter (HCA3-O LR (1xIFB))

  This optical long-range fanout provides 1x InfiniBand coupling link connectivity up to 10 km (6.2 miles) unrepeated distance to a z196, z114, or System z10 server.

The HCA3-O LR (1xIFB) fanout ships with four ports, and the other fanouts ship with two ports.

Figure 4-7 on page 114 illustrates the z114 coupling links.

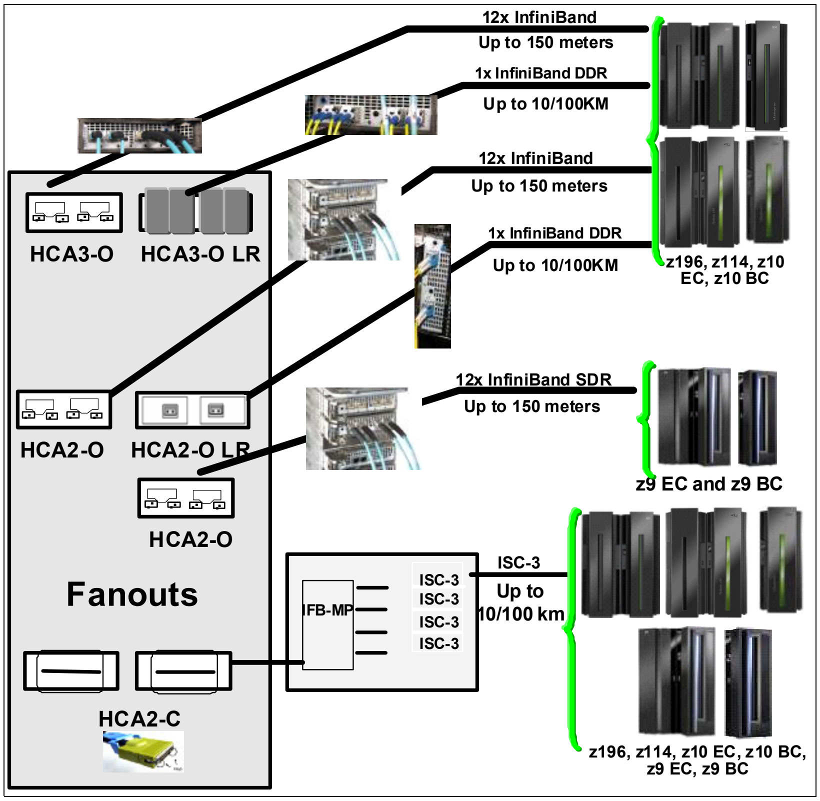


Figure 4-7 z114 coupling links

4.6.1 HCA2-C fanout

The HCA2-C fanout that is shown on Figure 4-2 on page 108 is used to connect to 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 copper cable of 1.5 m (4.11 ft.) to 3.5 m (11.5 ft.) is used for connection to the IFB-MP card in the I/O drawer.

HCA2-C fanout: The HCA2-C fanout is used exclusively for I/O and cannot be shared for any other purpose.

4.6.2 PCIe copper fanout

The PCIe fanout card shown on Figure 4-5 on page 111 supports the PCIe Gen2 bus and is used exclusively to connect to the PCIe I/O drawer. PCIe fanout cards are always plugged in pairs. The bandwidth of each port on the PCIe fanout supports a link rate of up to 8 GBps.

The PCIe fanout supports FICON Express8S and OSA Express4S in the PCIe I/O drawer.

PCIe fanout: The PCIe fanout is used exclusively for I/O and cannot be shared for any other purpose.

4.6.3 HCA2-O (12xIFB) fanout

The HCA2-O fanout for 12x InfiniBand provides an optical interface used for coupling links. The two ports on the fanout are dedicated to coupling links to connect to z196, z114, System z10, and System z9 servers, or to connect to a coupling port in the same server by using a fiber cable. Each fanout has an optical transmitter and receiver module and allows dual simplex operation. Up to 8 HCA2-O (12xIFB) fanouts are supported by z114 and provide up to 16 ports for coupling links.

The HCA2-O (12xIFB) fanout supports InfiniBand double data rate (12x IFB-DDR) and InfiniBand single data rate (12x IFB-SDR) optical links that offer longer distance, configuration flexibility, and high bandwidth for the 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 (492.1 ft.). There are 12 pairs of fibers: 12 fibers for transmitting and 12 fibers for receiving.

Each fiber supports a link rate of 6 GBps if connected to a z196, z114, or System z10 server, and 3 GBps when connected to a System z9 server. The link rate is auto-negotiated to the highest common rate.

HCA2-O (12xIFB) fanout: Ports on the HCA2-O (12xIFB) 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 as channel type CIB in the input/output configuration data set (IOCDS). The coupling links can be defined as shared between images within a channel subsystem and they can also be spanned across multiple CSSs in a server.

Each HCA2-O (12xIFB) fanout used for coupling links has an assigned adapter ID (AID) number that must be used for definitions in IOCDS to create a relationship between the physical fanout location and the CHPID number. For details about AID numbering, see “Adapter ID number assignment” on page 118.

For detailed information about how the AID is used and referenced in HCD, see *Getting Started with InfiniBand on System z10 and System z9,* SG24-7539.

When Server Time Protocol (STP) is enabled, IFB coupling links can be defined as timing-only links to other z196, z114, and System z10 servers.

4.6.4 HCA2-O LR (1xIFB) fanout

HCA2-O LR (1xIFB) is only available on z114 when carried forward (that is, during an upgrade) from a z10 BC. You cannot order HCA2-O LR (1xIFB) on an initial order of a z114.

The HCA2-O LR (1xIFB) fanout for 1x InfiniBand provides an optical interface that is used for coupling links. The two ports on the fanout are dedicated to coupling links to connect to z196, z114, and System z10 servers. IFB LR coupling link connectivity to other servers is not supported. Up to eight HCA2-O LR fanouts are supported by z114 and provide 16 ports for the coupling link.

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 a longer distance of coupling links. The cable has one lane containing two fibers: one fiber is used for transmitting data and one fiber is used for receiving data.

Each fiber supports a link rate of 5 Gbps if connected to a z196, z114, System z10 server, System z qualified dense wavelength division multiplexer (DWDM) supporting IB-DDR, or a data link rate of 2.5 Gbps when connected to a System z qualified DWDM that supports IB-SDR. The link rate is auto-negotiated to the highest common rate.

HCA2-O LR (1xIFB) fanout: Ports on the HCA2-O LR (1xIFB) 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 (62.1 miles) with 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 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 (1xIFB) fanout can be used for link definitions to another server or a link from one port to a port in another fanout on the same server.

Definitions of the source and target operating system image, CF image, and the CHPIDs that are used on both ports in both servers, are defined in IOCDS.

Each HCA2-O LR (1xIFB) fanout used for coupling links has an assigned adapter ID (AID) number that must be used for definitions in IOCDS to create a relationship between the physical fanout location and the CHPID number. See “Adapter ID number assignment” on page 118 for details about AID numbering.

When STP is enabled, IFB LR coupling links can be defined as timing-only links to other z196, z114, and System z10 servers.

4.6.5 HCA3-O (12xIFB) fanout

The HCA3-O fanout for 12x InfiniBand provides an optical interface used for coupling links. The two ports on the fanout are dedicated to coupling links to connect to z196, z114, or System z10, or to connect to a coupling port in the same server by using a fiber cable. The HCA3-O (12xIFB) fanout cannot communicate with an HCA1-O fanout on System z9. Up to eight HCA3-O (12xIFB) fanouts are supported and provide up to 16 ports for coupling links.

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 m (492.1 ft.). There are 12 pairs of fibers: 12 fibers for transmitting and 12 fibers for receiving. The HCA3-O (12xIFB) fanout supports a link data rate of 6 GBps.

HCA3-O (12xIFB) fanout: Ports on the HCA3-O (12xIFB) 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 the four ports. These CHPIDs are defined as channel type CIB in the IOCDS. The coupling links can be defined as shared between images within a channel subsystem and they can be also be spanned across multiple CSSs in a server.

Each HCA3-O (12xIFB) fanout used for coupling links has an assigned adapter ID (AID) number that must be used for definitions in IOCDS to create a relationship between the physical fanout location and the CHPID number. For details about AID numbering, see "Adapter ID number assignment" on page 118.

When STP is enabled, IFB coupling links can be defined as timing-only links to other z196, z114, and System z10 servers.

12x IFB and 12x IFB3 protocols

There are two protocols supported by the HCA3-O for 12x IFB feature:

- 12x IFB3 protocol

  When HCA3-O (12xIFB) fanouts communicate with HCA3-O (12xIFB) fanouts and have been defined with four or fewer CHPIDs per port, the 12x IFB3 protocol is used.

- 12x IFB protocol

  If more than four CHPIDs are defined per HCA3-O (12xIFB) port, or HCA3-O (12xIFB) features communicate with HCA2-O features on zEnterprise or System z10 servers, links will run with the 12x IFB protocol.

The HCA3-O feature supporting 12x InfiniBand coupling links has been designed to deliver improved service times. When no more than four CHPIDs are defined per HCA3-O (12xIFB) port, the 12x IFB3 protocol is used. When using the 12x IFB3 protocol, synchronous service times are designed to be 40% faster than when using the 12x IFB protocol.

4.6.6 HCA3-O LR (1xIFB) fanout

The HCA3-O LR (1xIFB) fanout for 1x InfiniBand provides an optical interface that is used for coupling links. The four ports on the fanout are dedicated to coupling links to connect to z196, z114, or System z10, or to connect to a coupling port in the same server by using a fiber cable. The HCA3-O LR (1xIFB) fanout cannot communicate with an HCA1-O LR fanout on System z9. Up to eight HCA3-O LR (1xIFB) fanouts are supported by the z114 and provide up to 32 ports for coupling links.

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 a longer distance of coupling links. The cable has one lane containing two fibers: one fiber is used for transmitting and one fiber is used for receiving data.

Each fiber supports a link rate of 5 Gbps if connected to a z196, z114, z10 server, or a System z qualified DWDM supporting IB-DDR, and a data link rate of 2.5 Gbps when connected to a repeater (System z qualified DWDM) that supports IB-SDR. The link rate is auto-negotiated to the highest common rate.

HCA3-O LR (1xIFB) fanout: Ports on the HCA3-O LR (1xIFB) 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 (62.1 miles) with System z qualified DWDM.

A fanout has four ports for optical link connections and supports up to 16 CHPIDs across both ports. These CHPIDs are defined as channel type CIB in the IOCDS. The coupling links can be defined as shared between images within a channel subsystem and they can also be spanned across multiple CSSs in a server. The fanout is compatible with the HCA2-O LR (1xIFB) fanout, which has two ports.

Each HCA3-O LR (1xIFB) fanout can be used for link definitions to another server or a link from one port to a port in another fanout on the same server.

Definitions of the source and target operating system image, CF image, and the CHPIDs that are used on both ports in both servers are defined in IOCDS.

Each HCA3-O LR (1xIFB) fanout that is used for coupling links has an assigned adapter ID (AID) number that must be used for definitions in IOCDS to create a relationship between the physical fanout location and the CHPID number. See “Adapter ID number assignment” on page 118 for details about AID numbering.

When STP is enabled, IFB LR coupling links can be defined as timing-only links to other z196, z114, and System z10 servers.

4.6.7 Fanout considerations

Because fanout slots in each processor drawer can be used to plug separate fanouts, where each fanout is designed for a special purpose, certain restrictions might apply to the number of available channels located in the I/O drawer and PCIe I/O drawer.

Adapter ID number assignment

Unlike channels installed in an I/O drawer, which are identified by a PCHID number related to their physical location, IFB 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 4-2 shows the assigned AID numbers for a newly built z114.

Table 4-2 AID number assignment for z114

Fanout location	Processor drawer	
	Processor drawer 1	Processor drawer 2
D1	08	00
D2	09	01
D7	0A	02
D8	0B	03

Important: The AID numbers in Table 4-2 are valid only for a newly built server or for a newly added processor drawer. If a fanout is moved, the AID follows the fanout to its new physical location.

The AID that is assigned to a fanout is found in the PCHID REPORT that is provided for each new server or for an MES upgrade on existing servers.

Example 4-1 shows part of a report, which is named PCHID REPORT, for a model M10. In this example, one fanout is installed in processor drawer 2 (A26B) and one fanout is installed in processor drawer slot D1. The assigned AID for the fanout is 00.

Example 4-1 AID assignment in PCHID report

```
CHPIDSTART
  14801905                      PCHID REPORT
  Machine: 2818-M10  NEW1
  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
  Source          Cage  Slot  F/C    PCHID/Ports or AID             Comment
  A26/D1          A26B  D1    0171   AID=00
```

4.6.8 Fanout summary

Table 4-3 shows the fanout features that are supported by the z114 server. The table provides the feature type, feature code, and information about the link supported by the fanout feature.

Table 4-3 Fanout summary

Fanout feature	Feature code	Use	Cable type	Connector type	Maximum distance	Link data rate
HCA2-C	0162	Connect to I/O drawer	Copper	N/A	3.5 m (11.5 ft.)	6 GBps
HCA2-O (12xIFB)	0163	Coupling link	50 µm MM OM3 (2000 MHz-km)	MPO	150 m (492.1 ft.)	6 GBps[a]
HCA2-O LR (1xIFB)	0168	Coupling link	9 µm SM	LC Duplex	10 km[b] (6.2 miles)	5.0 Gbps 2.5 Gbps[c]
PCIe fanout	1069	Connect to PCIe I/O drawer	Copper	N/A	3 m (9.10 ft.)	8 GBps
HCA3-O (12xIFB)	0171	Coupling link	50 µm MM OM3 (2000 MHz-km)	MPO	150 m (492.1 ft.)	6 GBps[d]
HCA3-O LR (1xIFB)	0170	Coupling link	9 µm SM	LC Duplex	10 km[b] (6.2 miles)	5.0 Gbps 2.5 Gbps[c]

[a] 3 GBps link data rate if connected to a System z9 server

[b] Up to 100 km (62.1 miles) with repeaters (System z qualified DWDM)

[c] Auto-negotiated, depending on DWDM equipment

[d] When using the 12x IFB3 protocol, synchronous service times are 40% faster than when using the 12x IFB protocol.

4.7 I/O feature cards

I/O cards have ports to connect the z114 to external devices, networks, or other servers. I/O cards are plugged into the PCIe I/O drawer and I/O drawer based on the configuration rules for the server. Various types of I/O cards are available: one for each channel or link type. I/O cards can be installed or replaced concurrently.

4.7.1 I/O feature card types ordering information

Table 4-4 lists the I/O features that are supported by z114 and their ordering information.

Table 4-4 I/O features and ordering information

Channel feature	Feature code	New build	Carry forward
16-port ESCON	2323	Y	Y
FICON Express4 4KM LX	3324	N	Y
FICON Express4-2C 4KM LX	3323	N	Y
FICON Express4 10KM LX	3321	N	Y
FICON Express8 10KM LX	3325	N[a]	Y
FICON Express8S 10KM LX	0409	Y	N/A
FICON Express4 SX	3322	N	Y
FICON Express4-2C SX	3318	N	Y
FICON Express8 SX	3326	N	Y
FICON Express8S SX	0410	Y	N/A
OSA-Express2 GbE LX	3364	N	Y
OSA-Express3 GbE LX	3362	N[a]	Y
OSA-Express4S GbE LX	0404	Y	N/A
OSA-Express2 GbE SX	3365	N	Y
OSA-Express3 GbE SX	3363	N[a]	Y
OSA-Express3-2P GbE SX	3373	N[a]	Y
OSA-Express4S GbE SX	0405	Y	N/A
OSA-Express2 1000BASE-T Ethernet	3366	N	Y
OSA-Express3 1000BASE-T Ethernet	3367	Y	Y
OSA-Express3-2P 1000BASE-T Ethernet	3369	Y	Y
OSA-Express3 10 GbE LR	3370	N[a]	Y
OSA-Express4S 10 GbE LR	0406	Y	N/A
OSA-Express3 10 GbE SR	3371	N[a]	Y
OSA-Express4S 10 GbE SR	0407	Y	N/A
ISC-3	0217 (ISC-M) 0218 (ISC-D)	Y	Y

Channel feature	Feature code	New build	Carry forward
ISC-3 up to 20 km[b] (12.4 miles)	RPQ 8P2197 (ISC-D)	Y	Y
HCA2-O (12xIFB)	0163	Y	Y
HCA2-O LR (1xIFB)	0168	N	Y
HCA3-O (12xIFB)	0171	Y	N/A
HCA3-O LR (1xIFB)	0170	Y	N/A
Crypto Express3	0864	Y	Y
Crypto Express3-1P	0871	Y	Y

[a] Ordering this feature is determined by the fulfillment process.

[b] RPQ 8P2197 enables the ordering of a daughter card supporting 20 km (12.4 miles) 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.

4.7.2 PCHID report

A Physical Channel ID (PCHID) reflects the physical location of a channel-type interface. A PCHID number is based on the I/O drawer and PCIe I/O drawer location, the channel feature slot number, and the port number of the channel feature. A CHPID does not directly correspond to a hardware channel port, but it is assigned to a PCHID in HCD or IOCP.

A PCHID report is created for each newly built server and for upgrades on existing servers. The report lists all I/O features installed, the physical slot location, and the assigned PCHID. Example 4-2 shows a portion of a sample PCHID report.

The AID numbering rules for InfiniBand coupling links are described in "Adapter ID number assignment" on page 118.

Example 4-2 PCHID report

```
CHPIDSTART
  16009308                      PCHID REPORT                       Jun 09,2011
  Machine: 2818-M10  NEW1
  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
  Source           Cage  Slot  F/C    PCHID/Ports or AID            Comment
  A26/D1/J01       A02B  09    0409   11C/J01 11D/J02

  A21/D1/J01       A02B  38    0404   17C/J01J02

  A26/D8/J01       A16B  10    2323   260/J00 261/J01


Legend:
  Source  CEC Drawer/Fanout Slot/Jack
  A21B    CEC Drawer 1 in A frame
  A26B    CEC Drawer 2 in A frame
  A02B    PCIe Drawer 1 in A frame
  A16B    I/O Drawer 1 in A frame
```

```
2323    ESCON Channel 16 Ports
0409    FICON Express8S 10KM LX
0404    OSA Express4S GbE LX
```

The following list explains the content of the sample PCHID REPORT:

- Feature code 0409 (FICON Express8S 10KM LX) is installed in the PCIe I/O drawer 1 (A02B, slot 09) and has PCHID 11C and 11D assigned.
- Feature code 0404 (OSA Express4S GbE LX) is installed in the PCIe I/O drawer 1 (A02B, slot 38) and has PCHID 17C assigned and shared by port J01 and J02.
- Feature code 2323 (ESCON Channel 16 ports) is installed in the I/O drawer 1 (A16B, slot 10) and has PCHIDs 260 and 261 assigned.

The pre-assigned PCHID number of each I/O port relates directly to its physical location (jack location in a specific slot). For PCHID numbers and their locations, see Table 4-4 on page 120 and Table 4-5.

4.8 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 LANs is provided by the OSA-Express2, OSA-Express3, and OSA-Express4S features.

Connectivity to I/O subsystems to exchange data is provided by ESCON and FICON channels.

4.8.1 I/O feature support and configuration rules

Table 4-5 lists the I/O features supported. The table shows the 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 4-5 z114 Supported I/O features

I/O feature	Number of ports per card	Number of port increments	Max. number of ports	Max. number of I/O slots	PCHID	CHPID definition
ESCON	16 (1 spare)	4 (LICCC)	240	16	Yes	CNC, CVC, CTC, CBY
FICON Express4-2C LX/SX	2	2	32	16	Yes	FC, FCP
FICON Express4 LX/SX	4	4	64	16	Yes	FC, FCP

I/O feature	Number of ports per card	Number of port increments	Max. number of ports	Max. number of I/O slots	PCHID	CHPID definition
FICON Express8 LX/SX	4	4	64	16	Yes	FC, FCP
FICON Express8S LX/SX	2	2	128	64	Yes	FC, FCP
OSA- Express2 GbE LX/SX	2	2	32	16	Yes	OSD, OSN
OSA- Express2 1000BASE-T	2	2	32	16	Yes	OSE, OSD, OSC, OSN
OSA- Express3 10 GbE LR/SR	2	2	32	16	Yes	OSD, OSX
OSA-Express3 GbE LX/SX	4	4	64	16	Yes	OSD, OSN
OSA-Express3 1000BASE-T	4	4	64	16	Yes	OSE, OSD, OSC, OSN, OSM
OSA-Express3-2P GbE SX	2	2	32	16	Yes	OSD, OSN
OSA-Express3-2P 1000BASE-T	2	2	32	16	Yes	OSE, OSD, OSC, OSN
OSA-Express4S GbE LX/SX	2	2	96	48	Yes	OSD
OSA- Express4S 10 GbE LR/SR	1	1	48	48	Yes	OSD, OSX
ISC-3 2 Gbps (10 km (6.2 miles))	2/ISC-D	1	48	12	Yes	CFP
ISC-3 1 Gbps (20 km (12.4 miles))	2/ISC-D	2	48	12	Yes	CFP
HCA2-O for 12x IFB	2	2	M10 - 16 M05 - 8	8	No	CIB
HCA3-O for 12x IFB and 12x IFB3	2	2	M10 - 16 M05 - 8	8	No	CIB
HCA2-O LR for 1x IFB	2	2	M10 - 12 M05 - 8	8	No	CIB
HCA3-O LR for 1x IFB	4	4	M10 - 32 M05 - 16	8	No	CIB

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 are 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:

- 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 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
- OSA-Express4S defined as OSD and OSX
- Coupling links defined as CFP, ICP, or CIB
- HiperSockets defined as IQD

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 Crypto Express3 features do not have a CHPID type, but logical partitions in all CSSs have access to the features. The Crypto Express3 feature (FC 0864) has two PCIe adapters. On z114, it is possible to order the Crypto Express3-1P feature (FC 0871), which has only one PCIe adapter. Each adapter on a Crypto Express3 feature can be defined to up to 16 logical partitions.

I/O feature cables and connectors

Cables: All fiber optic cables, cable planning, labeling, and installation are client responsibilities for new z114 installations and upgrades. Fiber optic conversion kits and mode conditioning patch (MCP) cables are not orderable as features on z114 servers. All other cables have to be sourced separately.

The IBM Facilities Cabling Services - fiber transport system offers a total cable solution service to help with cable ordering requirements and is highly desirable. These services consider the requirements for all of the supported protocols and media types (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 z114 for *quick* connection, which is offered as a feature on z114 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 4-6 on page 125 lists the required connector and cable type for each I/O feature on the z114.

Table 4-6 I/O feature connector and cable types

Feature code	Feature name	Connector type	Cable type
0163	InfiniBand coupling (IFB)	MPO	50 µm MM[a] OM3 (2000 MHz-km)
0168	InfiniBand coupling (IFB LR)	LC Duplex	9 µm SM[b]
0219	ISC-3	LC Duplex	9 µm SM
2324	ESCON	MT-RJ	62.5 µm MM
3323	FICON Express4-2C LX 4 km	LC Duplex	9 µm SM
3318	FICON Express4-2C SX	LC Duplex	50, 62.5 µm MM
3321	FICON Express4 LX 10 km	LC Duplex	9 µm SM
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
0409	FICON Express8S LX 10 km	LC Duplex	9 µm SM
0410	FICON Express8S 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]
3369	OSA-Express3-2P 1000BASE-T	RJ-45	Category 5 UTP[c]
3373	OSA-Express3-2P GbE SX	LC Duplex	50, 62.5 µm MM
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]
0404	OSA-Express4S GbE LX	LC Duplex	9 µm SM
0405	OSA-Express4S GbE SX	LC Duplex	50, 62.5 µm MM

[a] MM is multimode fiber.
[b] SM is single-mode fiber.
[c] UTP is unshielded twisted pair. Consider using Category 6 UTP for 1000 Mbps connections.

4.8.2 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. Each port on the feature uses a 1300 nanometer (nm) optical transceiver, which is 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.

Considerations:

- 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. You install an MT-RJ/ESCON conversion kit between the 16-port ESCON feature MT-RJ port and the ESCON Duplex jumper cable. This approach 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 client responsibilities for new installations and upgrades.
- IBM Facilities Cabling Services: The fiber transport system offers a total cable solution service to help with cable ordering needs and is highly desirable.

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 a few physically installed channels are typically inactive (LICCC-protected). These inactive channel ports are available to satisfy future channel adds.

ESCON to FICON: At this time, the zEnterprise 196 and zEnterprise 114 are intended to be the last System z servers to offer ordering of ESCON channels on new builds, migration offerings, upgrades, and System z exchange programs. Enterprises need to 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 this website:

`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 this website:

`http://www.opticatech.com`

Vendor inquiries: IBM cannot confirm the accuracy of compatibility, performance, or any other claims by vendors for products that have not been System z qualified. Address any questions regarding these capabilities and device support to the suppliers of these products.

4.8.3 FICON channels

The FICON Express8S, 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, and printers) in a Storage Area Network (SAN).

Important: FICON Express and FICON Express2 features installed in previous servers are *not* supported on a z114 and cannot be carried forward on an upgrade.

Each FICON Express8 or FICON Express4 feature occupies one I/O slot in the I/O drawer. Each feature has four ports, each supporting an LC Duplex connector, with one PCHID and one CHPID associated with each port.

Each FICON Express8S feature occupies one I/O slot in the PCIe I/O drawer. Each feature has two ports, each supporting an LC Duplex connector, with one PCHID and one CHPID associated with each port.

All FICON Express8S, 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 the replacement of a complete feature.

All FICON Express8S, FICON Express8, and FICON Express4 features also support cascading (the connection of two FICON Directors in succession) to minimize the number of cross-site connections and help reduce the implementation costs for disaster recovery applications, GDPS®, and remote copy.

Each FICON Express8S, FICON Express8, and FICON Express4 channel can be defined independently, for connectivity to servers, switches, directors, disks, tapes, and printers:

- CHPID type FC

  FICON, High Performance FICON for System z (zHPF), and FICON Channel-to-Channel (CTC). FICON, FICON CTC, and zHPF protocols are supported simultaneously.

- CHPID type FCP

  Fibre Channel Protocol, which supports attachment to SCSI devices directly or through Fibre Channel switches or directors.

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.

FICON Express8S

The FICON Express8S feature resides exclusively in the PCIe I/O drawer. Each of the two independent ports is capable of 2 gigabits per second (Gbps), 4 Gbps, or 8 Gbps depending on the capability of the attached switch or device. The link speed is auto-negotiated, point-to-point, and transparent to users and applications.

The two types of supported FICON Express8S optical transceivers are the long wavelength (LX) and the short wavelength (SX) transceivers:

- FICON Express8S 10km LX feature FC 0409, with two ports per feature, supporting LC Duplex connectors
- FICON Express8S SX feature FC 0410, with two ports per feature, supporting LC Duplex connectors

Each port of the FICON Express8S 10 km LX feature uses a 1300 nanometer (nm) optical transceiver and supports an unrepeated distance of 10 km (6.2 miles) using 9 µm single-mode fiber.

Each port of the FICON Express8S SX feature uses an 850 nanometer (nm) optical transceiver and supports varying distances depending on the fiber used (50 or 62.5 µm multimode fiber).

Auto-negotiation: FICON Express8S features do not support auto-negotiation to a data link rate of 1 Gbps.

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 website:

http://www-03.ibm.com/systems/z/hardware/connectivity/ficon_performance.html

The two types of supported FICON Express8 optical transceivers are the long wavelength (LX) and the short wavelength (SX) transceivers:

- 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 and supports an unrepeated distance of 10 km (6.2 miles) using 9 µm single-mode fiber.

Each port of FICON Express8 SX feature uses an 850 nanometer (nm) optical transceiver and supports varying distances depending on the fiber used (50 or 62.5 µm multimode fiber).

Auto-negotiation: FICON Express8 features do not support auto-negotiation to a data link rate of 1 Gbps.

FICON Express4

The three types of supported FICON Express4 optical transceivers are the two long wavelength (LX) and one short wavelength (SX) transceivers:

- 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 Express4-2C 4km LX feature FC 3323, with two ports per feature, supporting LC Duplex connectors
- FICON Express4 SX feature FC 3322, with four ports per feature, supporting LC Duplex connectors
- FICON Express4-2C SX feature FC 3318, with two ports per feature, supporting LC Duplex connectors

FICON Express4: It is intended that the z196 and z114 are the last servers to support FICON Express4 features. Clients need to review the usage of their installed FICON Express4 channels and, where possible, migrate to FICON Express8S channels.

Both FICON Express4 LX features use 1300 nanometer (nm) optical transceivers. One transceiver supports an unrepeated distance of 10 km (6.2 miles), and the other transceiver supports an unrepeated distance of 4 km (2.48 miles), using 9 µm single-mode fiber. Use of MCP cables limits the link speed to 1 Gbps and the unrepeated distance to 550 m (1804.6 ft.).

The FICON Express4 SX feature use 850 nanometer (nm) optical transceivers and supports varying distances depending on the fiber used (50 or 62.5 µm multimode fiber).

Link speed: FICON Express4 is the last FICON family that is able to negotiate link speed down to 1 Gbps.

FICON feature summary

Table 4-7 shows the FICON card feature codes, cable type, maximum unrepeated distance, and the link data rate on a z114. All FICON features use LC Duplex connectors. For long wave FICON features that can utilize a data rate of 1 Gbps, mode conditioning patch (MCP) cables (50 or 62.5 MM) can be used. The maximum distance for this connection is reduced to 550 m (1804.6 ft.) at a link data rate of 1 Gbps. Details for each feature follow the table.

Table 4-7 z114 channel feature support

Channel feature	Feature codes	Bit rate	Cable type	Maximum unrepeated distance[a]
FICON Express4 10KM LX	3321	1, 2, or 4 Gbps	SM 9 µm	10 km/20 km[b]
FICON Express4 4KM LX	3324	4 Gbps	SM 9 µm	4 km (2.48 miles)
FICON Express4-2C 4KM LX	3323	4 Gbps	SM 9 µm	4 km (2.48 miles)
FICON Express4 SX	3322	4 Gbps	MM 62.5 µm MM 50 µm	70 m (200) 150 m (500) 380 m (2000)
		2 Gbps	MM 62.5 µm MM 50 µm	150 m (200) 300 m (500) 500 m (2000)
		1 Gbps	MM 62.5 µm MM 50 µm	300 m (200) 500 m (500) 860 m (2000)
FICON Express4-2C SX	3318	4 Gbps	MM 62.5 µm MM 50 µm	70 m (200) 150 m (500) 380 m (2000)
		2 Gbps	MM 62.5 µm MM 50 µm	150 m (200) 300 m (500) 500 m (2000)
		1 Gbps	MM 62.5 µm MM 50 µm	300 m (200) 500 m (500) 860 m (2000)
FICON Express8 10KM LX	3325	2, 4, or 8 Gbps	SM 9 µm	10 km (6.2 miles)
FICON Express8 SX	3326	8 Gbps	MM 62.5 µm MM 50 µm	21 m (200) 50 m (500) 150 m (2000)
		4 Gbps	MM 62.5 µm MM 50 µm	70 m (200) 150 m (500) 380 m (2000)
		2 Gbps	MM 62.5 µm MM 50 µm	150 m (200) 300 m (500) 500 m (2000)
FICON Express8S 10KM LX	0409	2, 4, or 8 Gbps	SM 9 µm	10 km (6.2 miles)

Channel feature	Feature codes	Bit rate	Cable type	Maximum unrepeated distance[a]
FICON Express8S SX	0410	8 Gbps	MM 62.5 µm MM 50 µm	21 m (200) 50 m (500) 150 m (2000)
		4 Gbps	MM 62.5 µm MM 50 µm	70 m (200) 150 m (500) 380 m (2000)
		2 Gbps	MM 62.5 µm MM 50 µm	150 m (200) 300 m (500) 500 m (2000)

a. Minimum fiber bandwidths in MHz/km for multimode fiber optic links are included in parentheses where applicable.

b. Under certain conditions, RPQ 8P2263 might be required in conjunction with RPQ 8P2197. For the z10 BC, RPQ 8P2340 might be required. Check with your IBM representative.

4.8.4 OSA-Express4S

The OSA-Express4S feature resides exclusively in the PCIe I/O drawer. The following OSA-Express4S features can be installed on z114 servers:

- OSA-Express4S Gigabit Ethernet LX, feature code 0404
- OSA-Express4S Gigabit Ethernet SX, feature code 0405
- OSA-Express4S 10 Gigabit Ethernet LR, feature code 0406
- OSA-Express4S 10 Gigabit Ethernet SR, feature code 0407

Table 4-8 lists the OSA-Express4S features.

Table 4-8 OSA-Express4S features

I/O feature	Feature code	Number of ports per feature	Port increment	Maximum number of ports	Maximum number of features	CHPID type
OSA-Express4S 10 GbE LR	0406	1	1	48	48	OSD, OSX
OSA-Express4S 10 GbE SR	0407	1	1	48	48	OSD, OSX
OSA-Express4S GbE LX	0404	2	2	96	48	OSD
OSA-Express4S GbE SX	0405	2	2	96	48	OSD

OSA-Express4S Gigabit Ethernet LX (FC 0404)

The OSA-Express4S Gigabit Ethernet (GbE) long wavelength (LX) feature has one PCIe adapter and two ports. The two ports share a channel path identifier (CHPID type OSD exclusively). The ports support attachment to a 1Gbps Ethernet LAN. Each port can be defined as a spanned channel and can be shared among logical partitions and across logical channel subsystems.

The OSA-Express4S GbE LX feature supports the use of an LC Duplex connector. Ensure that the attaching or downstream device has a long-wavelength (LX) transceiver. The sending and receiving transceivers must be the same (LX to LX).

A 9 µm single-mode fiber optic cable terminated with an LC Duplex connector is required for connecting each port on this feature to the selected device. If multimode fiber optic cables are being reused, a pair of Mode Conditioning Patch cables is required, one cable for each end of the link.

OSA-Express4S Gigabit Ethernet SX (FC 0405)

The OSA-Express4S Gigabit Ethernet (GbE) short-wavelength (SX) feature has one PCIe adapter and two ports. The two ports share a channel path identifier (CHPID type OSD exclusively). The ports support attachment to a 1Gbps Ethernet LAN. Each port can be defined as a spanned channel and can be shared among logical partitions and across logical channel subsystems.

The OSA-Express4S GbE SX feature supports the use of an LC Duplex connector. Ensure that the attaching or downstream device has a short-wavelength (SX) transceiver. The sending and receiving transceivers must be the same (SX to SX).

A 50 or 62.5 µm multimode fiber optic cable terminated with an LC Duplex connector is required for connecting each port on this feature to the selected device.

OSA-Express4S 10 Gigabit Ethernet LR (FC 0406)

The OSA-Express4S 10 Gigabit Ethernet (GbE) long reach (LR) feature has one PCIe adapter and one port per feature. The port supports channel path identifier (CHPID) types OSD and OSX. When defined as CHPID type OSX, the 10 GbE port provides connectivity and access control to the intraensemble data network (IEDN) from z114 to zEnterprise BladeCenter Extension (zBX). The 10 GbE feature is designed to support attachment to a single-mode fiber 10 Gbps Ethernet LAN or Ethernet switch that is capable of 10 Gbps. The port can be defined as a spanned channel and can be shared among logical partitions within and across logical channel subsystems.

The OSA-Express4S 10 GbE LR feature supports the use of an industry standard small form factor LC Duplex connector. Ensure that the attaching or downstream device has a long-reach (LR) transceiver. The sending and receiving transceivers must be the same (LR to LR which might also be referred to as LW or LX).

A 9 µm single-mode fiber optic cable terminated with an LC Duplex connector is required for connecting this feature to the selected device.

OSA-Express4S 10 Gigabit Ethernet SR (FC 0407)

The OSA-Express4S 10 Gigabit Ethernet (GbE) Short Reach (SR) feature has one PCIe adapter and one port per feature. The port supports channel path identifier (CHPID) types OSD and OSX. When defined as CHPID type OSX, the 10 GbE port provides connectivity and access control to the intraensemble data network (IEDN) from z114 to zEnterprise BladeCenter Extension (zBX). The 10 GbE feature is designed to support attachment to a multimode fiber 10 Gbps Ethernet LAN or Ethernet switch that is capable of 10 Gbps. The port can be defined as a spanned channel and can be shared among logical partitions within and across logical channel subsystems.

The OSA-Express4S 10 GbE SR feature supports the use of an industry standard small form factor LC Duplex connector. Ensure that the attaching or downstream device has a Short Reach (SR) transceiver. The sending and receiving transceivers must be the same (SR to SR).

A 50 or a 62.5 µm multimode fiber optic cable terminated with an LC Duplex connector is required for connecting each port on this feature to the selected device.

4.8.5 OSA-Express3

This section discusses the connectivity options that are 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 z114 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
- OSA-Express3-2P 1000BASE-T Ethernet, feature code 3369
- OSA-Express3-2P GbE SX, feature code 3373

Table 4-9 lists the OSA-Express3 features.

Table 4-9 OSA-Express3 features

I/O feature	Feature code	Number of ports per feature	Port increment	Maximum number of ports	Maximum number of features	CHPID type
OSA-Express3 10 GbE LR	3370	2	2	32	16	OSD,OSX
OSA-Express3 10 GbE SR	3371	2	2	32	16	OSD,OSX
OSA-Express3 GbE LX	3362	4	4	64	16	OSD, OSN
OSA-Express3 GbE SX	3363	4	4	64	16	OSD, OSN
OSA-Express3 1000BASE-T	3367	4	4	64	16	OSC,OSD, OSE,OSN, OSM
OSA-Express3-2P 1000BASE-T	3369	2	2	32	16	OSC,OSD, OSE,OSN, OSM
OSA-Express3-2P GbE SX	3373	2	2	32	16	OSD, OSN

OSA-Express3 data router

OSA-Express3 features help reduce latency and improve throughput by providing a data router. Functions that were previously performed in firmware (packet construction, inspection, and routing) are 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 (6.2 miles).

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 client-managed external network connections.

CHPID type OSX is dedicated for connecting the z196 and z114 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 m (108 ft.) on a 62.5 µm multimode (200 MHz) fiber optic cable, 82 m (269 ft.) on a 50 µm multi mode (500 MHz) fiber optic cable, and 300 m (984 ft.) 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 client-managed external network connections.

CHPID type OSX is dedicated for connecting the z196 and z114 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 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 feature.

MCP: The use of these multimode cable types requires a mode conditioning patch (MCP) cable at each end of the fiber optic link. The use of the single-mode to multimode MCP cables reduces the supported distance of the link to a maximum of 550 m (1804.6 ft.).

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 212.

OSA-Express3 GbE SX (FC 3363)

Feature code 3363 occupies one slot in the 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 8.2, “Support by operating system” on page 212.

OSA-Express3 1000BASE-T Ethernet feature (FC 3367)

Feature code 3367 occupies one slot in the 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 or Category 6 unshielded twisted pair (UTP) cable with a maximum length of 100 m (328 ft.).

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 that 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 specified speed and duplex mode does not match the speed and duplex mode of the signal on the cable, the OSA-Express port will not connect.

4.8.6 OSA-Express2

This section discusses the connectivity options that are 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 z114 and cannot be carried forward on an upgrade.

Table 4-10 lists the OSA-Express2 features.

Table 4-10 OSA-Express2 features

I/O feature	Feature code	Number of ports per feature	Number of port increments	Max. number of ports	Max. number of I/O slots	CHPID type
OSA-Express2 GbE LX/SX	3364 3365	2	2	48	24	OSD, OSN
OSA-Express2 1000BASE-T	3366	2	2	48	24	OSE, OSD, OSC, OSN

OSA-Express2: It is intended that the z196 and z114 are the last servers to support OSA-Express2 features. Review the usage of installed OSA-Express2 features and, where possible, migrate to OSA-Express4S features.

OSA-Express2 GbE LX (FC 3364)

The OSA-Express2 Gigabit (GbE) Long Wavelength (LX) feature occupies one slot in an 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 can 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. The 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 m (1804 ft.).

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 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 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 or Category 6 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 on the LAN.

The OSA-Express2 1000BASE-T Ethernet feature supports CHPID types OSC, OSD, OSE, and OSN.

You can 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 (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.

4.8.7 OSA-Express for ensemble connectivity

The following OSA-Express features are used to connect the zEnterprise CPC to its attached IBM zEnterprise BladeCenter Extension (zBX) and other ensemble nodes:

- OSA-Express3 1000BASE-T Ethernet, feature code 3367
- OSA-Express3-2P 1000BASE-T Ethernet, feature code 3369
- 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-Express4S 10 Gigabit Ethernet (GbE) Long Range (LR), feature code 0406
- OSA-Express4S 10 Gigabit Ethernet (GbE) Short Reach (SR), feature code 0407

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 client-managed switches or routers are required. The IEDN is managed by a primary HMC[2].

IEDN requires two OSA-Express3 10 GbE ports (one port from two OSA-Express3 10 GbE features) or preferably two OSA-Express4S 10 GbE configured as CHPID type OSX. The connection is from the zEnterprise CPC to the IEDN Top of Rack (TOR) switches on zBX. Or with a stand-alone zEnterprise CPC node (no-zBX), interconnect pairs of OSX ports via LC DUPLEX directly connected cables and *not* wrap cables as has previously been recommended.

Intranode management network (INMN)

The INMN is a private and physically isolated 1000BASE-T Ethernet internal management network, operating at 1 Gbps. It connects all resources (zEnterprise CPC and zBX components) of an ensemble node for management purposes. It is prewired, internally switched, configured, and managed with full redundancy for high availability.

The INMN requires two ports (CHPID port 0 from two OSA-Express3 1000BASE-T features. CHPID port 1 is not used at all in this case) configured as CHPID type OSM. The connection is via port J07 of the bulk power hubs (BPHs) in the zEnterprise CPC. The INMN Top of Rack (TOR) switches on zBX also connect to the BPHs.

For detailed information about OSA-Express in an ensemble network, see 7.4, “zBX connectivity” on page 194.

4.8.8 HiperSockets

The HiperSockets function of zEnterprise CPCs is improved to provide up to 32 high-speed virtual LAN 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 zEnterprise CPC. 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 zEnterprise CPCs support two transport modes:

- Layer 2 (link layer)
- Layer 3 (network or IP layer)

Traffic can be IPv4 or IPv6, or non-IP, such as AppleTalk, DECnet, IPX, NetBIOS, or SNA.

HiperSockets devices are protocol-independent and Layer 3-independent. Each HiperSockets device (Layer 2 and Layer 3 mode) has its own MAC address that is 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, reduce complexity, simplify network configuration, and allow LAN administrators to maintain the mainframe network environment in a similar manner to non-mainframe environments.

[2] This HMC must be running with Version 2.11 or higher with feature codes 0090, 0025, 0019, and optionally 0020.

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 configuration 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 separate servers.

HiperSockets Layer 2 on zEnterprise CPCs is supported by Linux on System z, and by z/VM for Linux guest exploitation.

Statements of Direction:

HiperSockets Completion Queue: At this time, IBM plans to support transferring HiperSockets messages asynchronously, in addition to the current synchronous manner, on z196 and z114. This capability might be especially helpful in burst situations. The Completion Queue function is designed to allow HiperSockets to transfer data synchronously if possible and asynchronously if necessary, thus combining ultra-low latency with more tolerance for traffic peaks. HiperSockets Completion Queue is planned to be supported in the z/VM and z/VSE environments.

HiperSockets integration with the IEDN: Within a zEnterprise environment, it is intended for HiperSockets to be integrated with the intraensemble data network (IEDN), extending the reach of the HiperSockets network outside of the central processor complex (CPC) to the entire ensemble, appearing as a single Layer 2 network. HiperSockets integration with the IEDN is planned to be supported in z/OS V1.13 and z/VM in a future deliverable.

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

4.9 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 consists of one or more z/OS operating system images coupled through one or more coupling facilities.

4.9.1 Coupling links

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 following types of links are available to connect an operating system logical partition to a coupling facility:

- 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, z114, or System z10. They are optic fiber links that support a maximum distance of 10 km (6.2 miles), 20 km (12.4 miles) with RPQ 8P2197, and 100 km (62.1 miles) 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 (6.2 miles), 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 z114 is 48. ISC-3 supports transmission of Server Time Protocol (STP) messages.

  > **ISC-3:** It is intended that the z196 and z114 are the last systems for ordering of ISC-3 coupling links. Clients need to review the usage of their installed ISC-3 coupling links and, where possible, migrate to IFB (FC 0163 and FC 0171) or IFB LR (FC 0168 and FC 0170) coupling links.

- IFB

  Parallel Sysplex using Infinband (IFB) connects a z114 to a z196, z114, System z10, or System z9. 12x InfiniBand coupling links are fiber optic connections that support a maximum distance of up to 150 m (492 ft.). IFB coupling links are defined as CHPID type CIB. IFB supports transmission of STP messages.

  z114 supports two types of 12x InfiniBand coupling links (FC 0171 HCA3-O (12xIFB) fanout and FC 0163 HCA2-O (12xIFB) fanout).

- IFB LR

  IFB LR (Long Reach) connects a z114 to another z196 or System z10 server. 1x InfiniBand coupling links are fiber optic connections that support a maximum unrepeated distance of up to 10 km (6.2 miles) and up to 100 km (62.1 miles) with a System z qualified dense wave division multiplexor (DWDM). IFB LR coupling links are defined as CHPID type CIB. IFB LR supports transmission of STP messages.

  z114 supports two types of 1x InfiniBand coupling links (FC 0170 HCA3-O LR (1xIFB) fanout and FC 0168 HCA2-O LR (1xIFB) fanout). IFB LR supports 7 or 32 subchannels per CHPID.

- IC

  CHPIDs (type ICP) that are defined for internal coupling can connect a CF to a z/OS logical partition in the same z114. 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 4-11 shows the coupling link options.

Table 4-11 Coupling link options

Type	Description	Use	Link rate	Distance	z114-M05 maximum	z114-M10 maximum
ISC-3	InterSystem Channel-3	z114 to z114, z196, z10, z9	2 Gbps	10 km (6.2 miles) unrepeated 100 km (62.1 miles) repeated	48	48
IFB	12x IB-DDR InfiniBand (HCA3-O)[a]	z114 to z114, z196, z10	6 GBps	150 meters (492 feet)	8[b]	16[b]
	12x IB-DDR InfiniBand (HCA2-O)	z114 to z114, z196, z10	6 GBps	150 meters (492 feet)	8[b]	16[b]
	12x IB-SDR InfiniBand (HCA2-O)	z114 to z9	3 GBps	150 meters (492 feet)	8[b]	16[b]
IFB LR	1x IFB (HCA3-O LR)	z114 to z114, z196, z10	2.5 Gbps 5.0 Gbps	10 km (6.2 miles) unrepeated 100 km (62.1 miles) repeated	16[b]	32[b]
	1x IFB (HCA2-O LR)	z114 to z114, z196, z10	2.5 Gbps 5.0 Gbps	10 km (6.2 miles) unrepeated 100 km (62.1 miles) repeated	8[b]	12[b]
IC	Internal coupling channel	Internal communication	Internal speeds	N/A	32	32

a. 12x IFB3 protocol: Maximum 4 CHPIDs and connects to other HCA3-O (12xIFB) port, else 12x IFB protocol. Auto-configured when conditions are met for IFB3. See 4.6.5, “HCA3-O (12xIFB) fanout” on page 116.

b. Uses all available fanout slots. Allows no other I/O or coupling.

The maximum IFB links is 32. The maximum number of combined external coupling links (active ISC-3 links, IFB, and IFB LR) cannot exceed 56[3] for z114 M05 and 72[4] for z114 M10. There is a maximum of 128 coupling CHPIDs limitation, including ICP for IC, CIB for IFB/IFB LR, and CFP for ISC-3.

The z114 supports various connectivity options depending on the connected z196, z114, System z10, or System z9 server. Figure 4-8 on page 142 shows z114 coupling link support for z196, z114, System z10, and System z9 servers.

When defining IFB coupling links (CHPID type CIB), HCD now defaults to 32 subchannels. Thirty-two subchannels are only supported on HCA2-O LR (1xIFB) and HCA3-O LR (1xIFB) on z196 and z114 and when both sides of the connection use 32 subchannels. Otherwise, you need to change the default value from 32 to 7 subchannels on each CIB definition.

[3] M05: Eight 1x IFB and 48 ISC-3, with no 12x IFB links. Uses all available fanout slots.

[4] M10: Twenty-four 1x IFB and 48 ISC-3, with no 12x IFB links. Uses all available fanout slots.

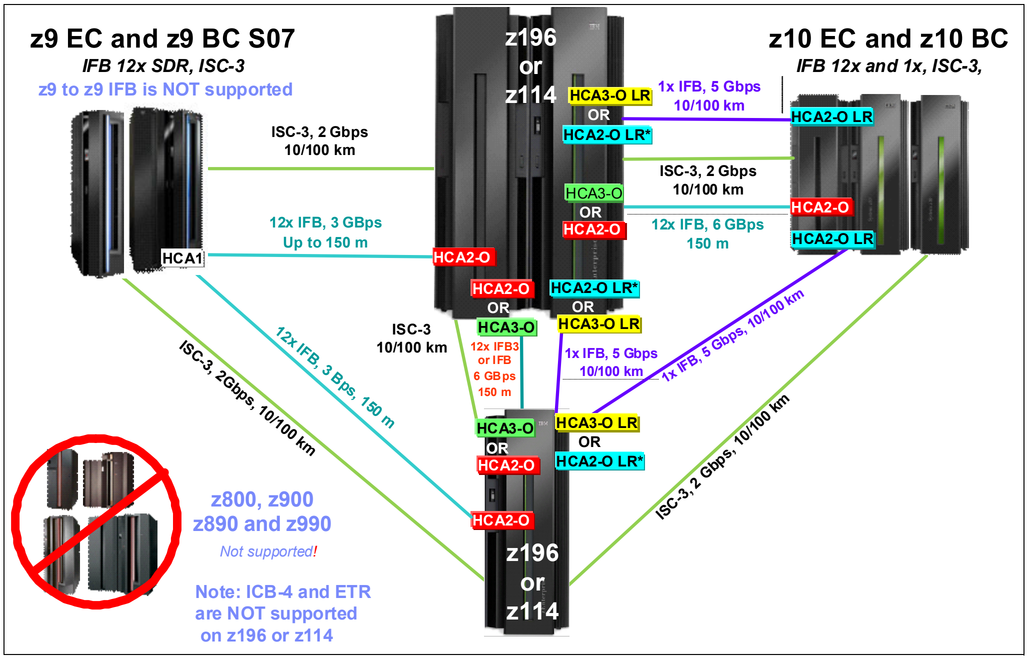


Figure 4-8 zEnterprise CPCs Parallel Sysplex coupling connectivity

z/OS and coupling facility images can 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, have at least these links and facility images:

- Two coupling links between the z/OS and coupling facility images
- Two coupling facility images not running on the same server
- One stand-alone coupling facility. If using system-managed CF structure duplexing or running with *resource sharing* only, a stand-alone coupling facility is not mandatory.

Coupling link features

The z114 supports five types of coupling link options:

- InterSystem Channel-3 (ISC-3) FC 0217, FC 0218, and FC 0219
- HCA2-O fanout for 12x InfiniBand, FC 0163
- HCA2-O LR fanout for 1x InfiniBand, FC 0168
- HCA3-O fanout for 12x InfiniBand, FC 0171
- HCA3-O LR fanout for 1x InfiniBand, FC 0170

The coupling link features that are available on the z114 connect z114 servers to the identified System z servers by various link options:

- ISC-3 at 2 Gbps to z196, z114, System z10, and System z9
- 12x InfiniBand using HCA2-O fanout card at 6 GBps to z196, z114, and System z10, or 3 GBps to z9 EC and z9 BC
- 12x InfiniBand using HCA3-O fanout card at 6 GBps to z196, z114, and System z10
- 1x InfiniBand using both HCA3-O LR (1xIFB) and HCA2-O LR (1xIFB) at 5.0 or 2.5 Gbps to z196, z114, and System z10 servers

ISC-3 coupling links

Three feature codes are available to implement ISC-3 coupling links:

- FC 0217, ISC-3 mother card
- FC 0218, ISC-3 daughter card
- FC 0219, ISC-3 port

The ISC mother card (FC 0217) occupies one slot in the I/O cage or I/O drawer and supports up to two daughter cards. The ISC daughter card (FC 0218) provides two independent ports with one CHPID that is 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.

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 (6.2 miles). With repeaters, the maximum distance extends to 100 km (62.1 miles). 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 that 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 (12.4 miles). 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)

For detailed information, see 4.6.3, “HCA2-O (12xIFB) fanout” on page 115.

InfiniBand coupling links LR (FC 0168)

For detailed information, see 4.6.4, “HCA2-O LR (1xIFB) fanout” on page 115.

HCA3-O fanout for 12x InfiniBand (FC 0171)

For detailed information, see 4.6.5, “HCA3-O (12xIFB) fanout” on page 116.

HCA3-O LR fanout for 1x InfiniBand (FC 0170)

For detailed information, see 4.6.6, “HCA3-O LR (1xIFB) fanout” on page 117.

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 they 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, 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 that is associated with Parallel Sysplex technology.
- Can be used in test or production configurations, and reduce the cost of moving into Parallel Sysplex technology while enhancing performance and reliability.
- Can be defined as spanned channels across multiple CSSs.
- Are no 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

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:

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 IFB links, can meet more stringent synchronization requirements than servers exchanging messages over long ISC-3 links (distances up to 100 km (62.1 miles)). This advantage is an enhancement over the IBM Sysplex Timer implementation, which does not scale with distance.
- Coupling links also provide the necessary connectivity in a Parallel Sysplex. Therefore, there is a potential benefit of minimizing the number of required cross-site links in a multiple site Parallel Sysplex.

Between any two servers that are intended to exchange STP messages, it is best 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 and z114 do not support attachment to the IBM Sysplex Timer. A z114 can be added into a Mixed Coordinated Timing Network (CTN) only when there is a z10 or z9 attached to the Sysplex Timer operating as the Stratum 1 server. Connections to two Stratum 1 servers are preferable to provide redundancy and avoid a single point of failure.

Important: A Parallel Sysplex in an ETR network *must* migrate to a Mixed CTN or STP-only CTN *before* introducing a z114.

STP recovery enhancement

The new generation of host channel adapters (HCA3-O (12xIFB) or HCA3-O LR (1xIFB)) that have been introduced for coupling has been designed to send a reliable unambiguous "going away signal" to indicate that the server on which the HCA3 is running is about to enter a failed (check stopped) state. When the "going away signal" sent by the Current Time Server (CTS) in an STP-only Coordinated Timing Network (CTN) is received by the Backup Time Server (BTS), the BTS can safely take over as the CTS without relying on the previous Offline Signal (OLS) in a two-server CTN or as the Arbiter in a CTN with three or more servers.

This enhancement is exclusive to z196 and z114 and is available only if you have an HCA3-O (12xIFB) or HCA3-O LR (1xIFB) on the CTS communicating with an HCA3-O (12xIFB) or HCA3-O LR (1xIFB) on the BTS. Note that the already available STP recovery design is still available for the cases when a GOSIG is not received or for other failures besides a server failure.

For more information: See the following white paper if you have configured a Server Time Protocol (STP) Coordinated Timing Network (CTN) with three or more servers:

http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP101833

The suggestions, if not followed, might result in all the servers in the CTN becoming unsynchronized, a condition that results in a sysplex-wide outage.

For STP configuration information, see the *Server Time Protocol Planning Guide*, SG24-7280, and *Server Time Protocol Implementation Guide*, SG24-7281.

4.9.2 Pulse Per Second (PPS)

The two oscillator cards in the first processor drawer of the z114 provide Pulse Per Second (PPS) connections to external time sources (ETSs). Two oscillator cards are installed to provide redundancy for continued operation and concurrent maintenance when a single oscillator card fails. Each oscillator card has a Bayonet Neill-Concelman (BNC) connector for PPS connection support, attaching to two separate ETSs.

The time accuracy of an STP-only CTN is improved by adding a Network Time Protocol (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 an 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 various vendors that offer network timing solutions.

PPS connection from a NTP server with PPS output to the ECF card to is required when the z114 is configured in an STP-only CTN using NTP with pulse per second as the external time source. Two PPS connections from two separate NTP servers are preferable for redundancy.

4.10 Cryptographic functions

Cryptographic functions are provided by CP Assist for Cryptographic Function (CPACF) and the Crypto Express3 feature.

CPACF functions (FC 3863)

Feature code 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 is installed; thereafter, the number of features increases by one at a time up to a maximum of eight features. Each Crypto Express3 feature holds two PCI Express cryptographic adapters. Either of the adapters can be configured by the installation as a coprocessor or accelerator.

Crypto Express3-1P feature (FC 0871)

On z114, you have the capability to order the Crypto Express3-1P feature, which has just one PCIe adapter. It is also an optional feature. On the initial order, the minimum of two features is installed. Thereafter, the number of features increases by one at a time up to a maximum of eight features. The adapter can be configured by the installation as a coprocessor or accelerator.

Each Crypto Express3 feature occupies one I/O slot in the I/O drawer, and it has no CHPIDs assigned, but it uses two PCHIDS. We describe cryptographic functions in Chapter 6, “Cryptography” on page 159.

5

Central processor complex channel subsystem

In this chapter, we describe the concepts of the System z channel subsystem, including multiple channel subsystems. We also discuss the technology, terminology, and implementation aspects of the channel subsystem.

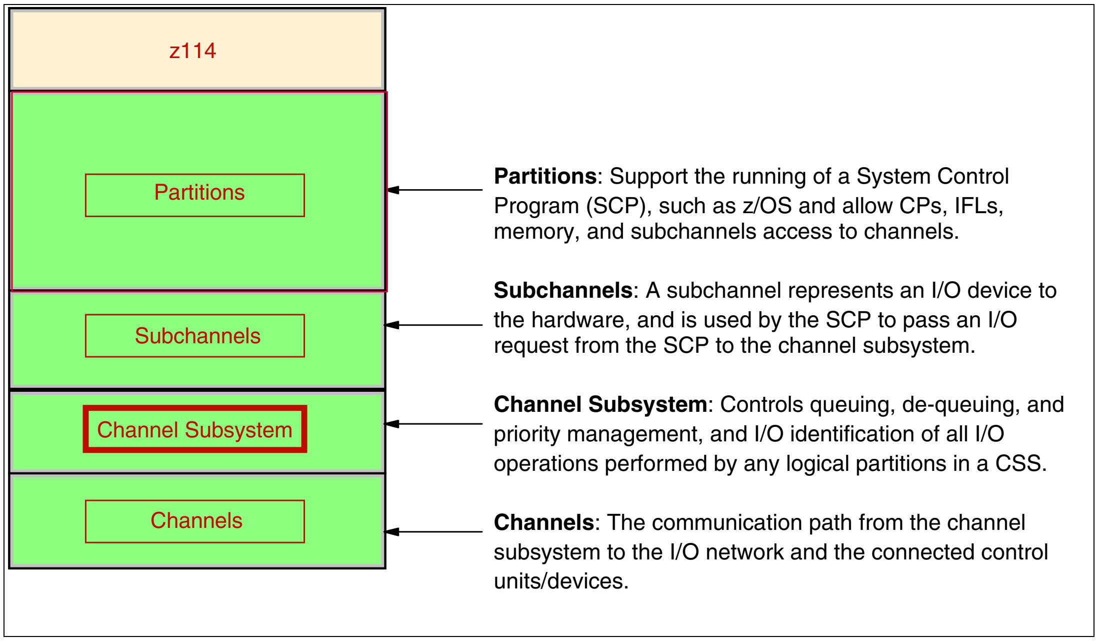
We cover the following topics:

5.1 Channel subsystem

The role of the channel subsystem (CSS) is to control the 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 the 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.

Figure 5-1 shows the building blocks that make up a channel subsystem.



Partitions: Support the running of a System Control Program (SCP), such as z/OS and allow CPs, IFLs, memory, and subchannels access to channels.

Subchannels: A subchannel represents an I/O device to the hardware, and is used by the SCP to pass an I/O request from the SCP to the channel subsystem.

Channel Subsystem: Controls queuing, de-queuing, and priority management, and I/O identification of all I/O operations performed by any logical partitions in a CSS.

Channels: The communication path from the channel subsystem to the I/O network and the connected control units/devices.

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, but on z114, two channel subsystems are supported. 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 can have from 1 to 256 channels and be configured with 1 to 15 logical partitions. Therefore, two CSSs support a maximum of 30 logical partitions. CSSs are numbered 0 to 1 and are sometimes referred to as the *CSS image ID* (CSSID 0 and 1).

5.1.2 CSS elements

We describe the elements that encompass the CSS 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 149, are available to increase addressability. Two subchannel sets per CSS are supported on z114. Subchannel set 0 can have up to 63.75 K subchannels, and subchannel set 1 can have up to 64 K subchannels.

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 256 are supported per z114.

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 client, 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 can be housed separately, or it can 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

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, which is not technically correct. We might 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. Subchannel set 1 provides the full range of 64 K subchannels. In principle, subchannels in either set can be used for any device-addressing purpose. These subchannels are referred to as special devices in the following topics.

Figure 5-2 summarizes the multiple channel subsystems and multiple subchannel sets.

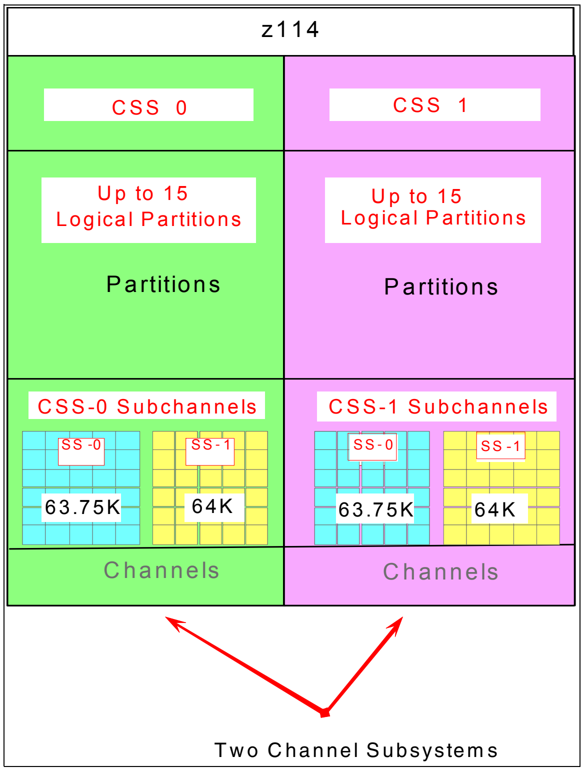


Figure 5-2 Multiple channel subsystems and multiple subchannel sets

The additional subchannel set, in effect, adds an extra high-order digit (either 0 or 1) to existing device numbers. For example, consider an address of 08000 (subchannel set 0) or 18000 (subchannel set 1). 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, certain messages contain the subchannel set number, and you can mentally use that as a high-order digit for device numbers. Only a few requirements refer to the subchannel set 1, because subchannel sets 1 is only used for these special devices. JCL, messages, and programs rarely refer directly these special devices.

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

[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. The informal name, 63.75 K subchannel, represents the following equation: (63 x 1024) + (0.75 x 1024) = 65280

definitions or in the HCD definitions that produce the input/output configuration data set (IOCDS). The subchannel set number defaults to zero.

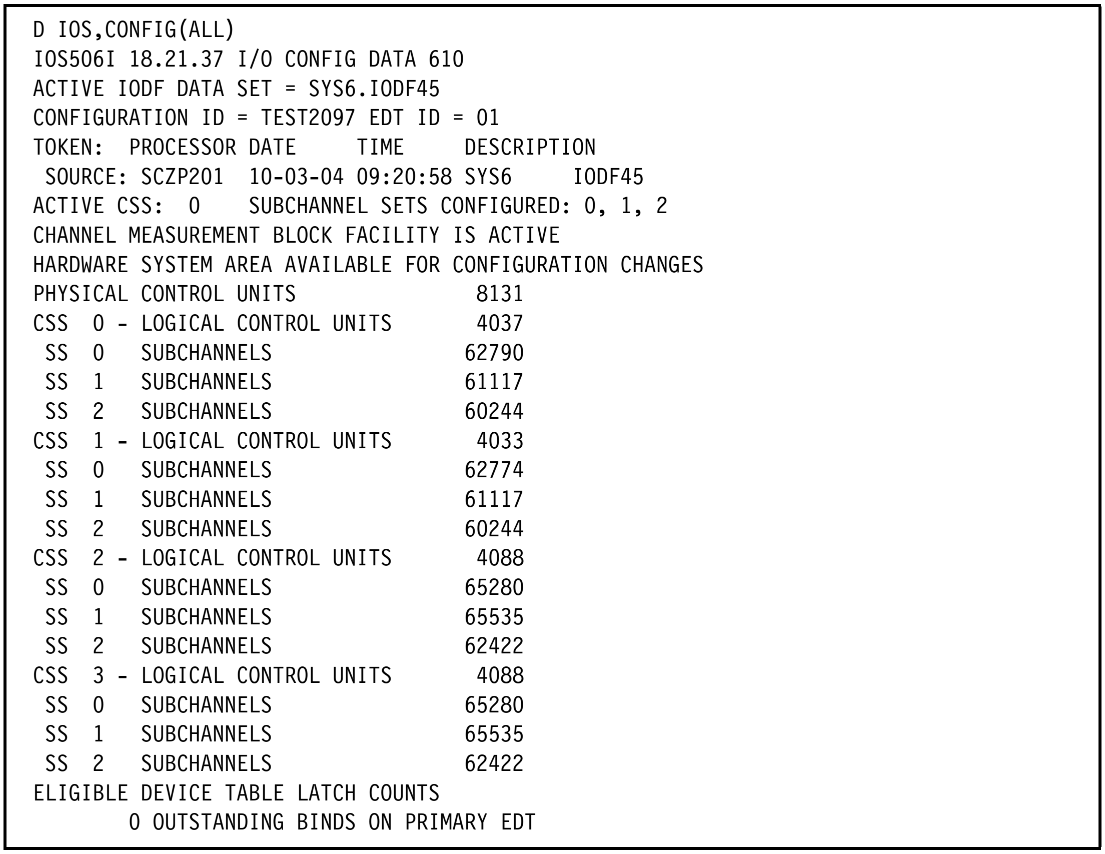
IPL from an alternate subchannel set

z114 supports IPL from subchannel set 1 (SS1), in addition to subchannel set 0. Devices used early during IPL processing can now be accessed using subchannel set 1. This capability allows the users of Metro Mirror (PPRC) secondary devices defined using the same device number and a new device type in an alternate subchannel set to be used for IPL, IODF, and stand-alone dump volumes when needed.

IPL from an alternate subchannel set is exclusive to z196 and z114, and is supported by z/OS V1.13, as well as V1.12 and V1.11 with PTFs, and applies to the FICON and zHPF protocols.

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 Workload Manager (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 separate HyperPAV hosts to use one alias to access separate 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 extremely large disk volumes using virtualization technology. By being able to extend the disk volume size, a client might 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 might 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.

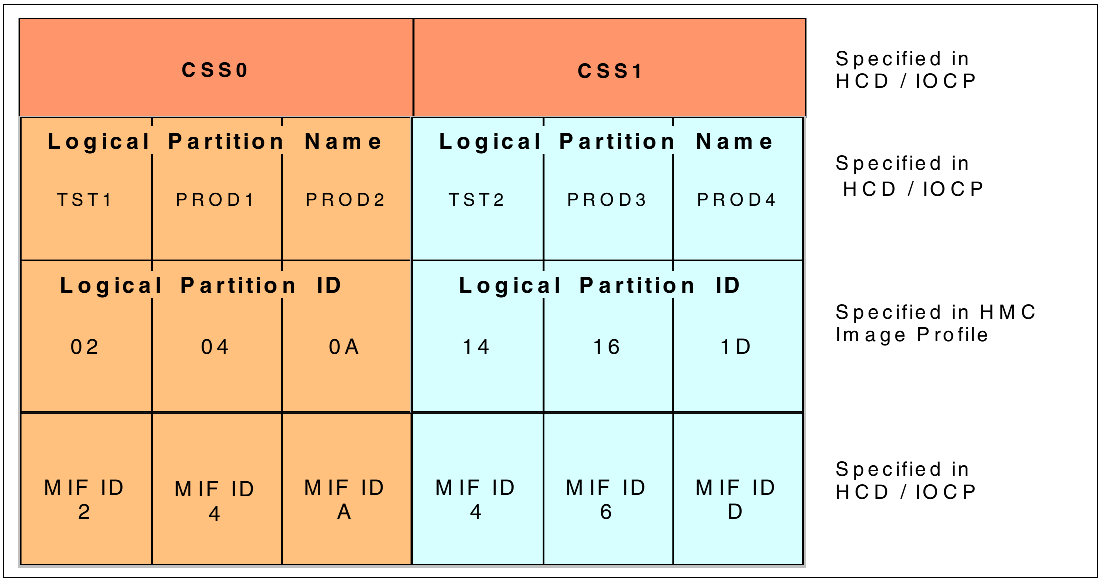
The multiple image facility enables resource sharing across logical partitions within a single CSS or across the multiple CSSs. When a channel resource is shared across logical partitions in multiple CSSs, this design is known as *spanning*. Multiple CSSs can specify the same MIF image ID. However, the 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

It is good practice to establish a naming convention for the logical partition identifiers. As shown in Figure 5-4 on page 153, which summarizes the identifiers and how they are defined, you can use the CSS number concatenated to the MIF ID, which means that logical partition ID 1D is in CSS 1 with MIF ID D. This method fits within the allowed range of logical partition IDs and conveys helpful information to the user.



CSS0			CSS1			Specified in HCD / IOCP
Logical Partition Name			Logical Partition Name			Specified in HCD / IOCP
TST1	PROD1	PROD2	TST2	PROD3	PROD4	
Logical Partition ID			Logical Partition ID			Specified in HMC Image Profile
02	04	0A	14	16	1D	
MIF ID 2	MIF ID 4	MIF ID A	MIF ID 4	MIF ID 6	MIF ID D	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 PCIe I/O drawer location, the channel feature slot number, and the port number of the channel feature. A hardware channel is identified by a PCHID. The physical channel, which uniquely identifies a connector jack on a channel feature, is known by its PCHID number. For further information, see 4.7.2, “PCHID report” on page 121.

Do not confuse PCHIDs with CHPIDs. A CHPID does not directly correspond to a hardware channel port, and it can be arbitrarily assigned. Within a single channel subsystem, 256 CHPIDs can be addressed, which gives a maximum of 512 CHPIDs when two 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 separate 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 approach applies, but with no PCHID association. The internal channels 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.

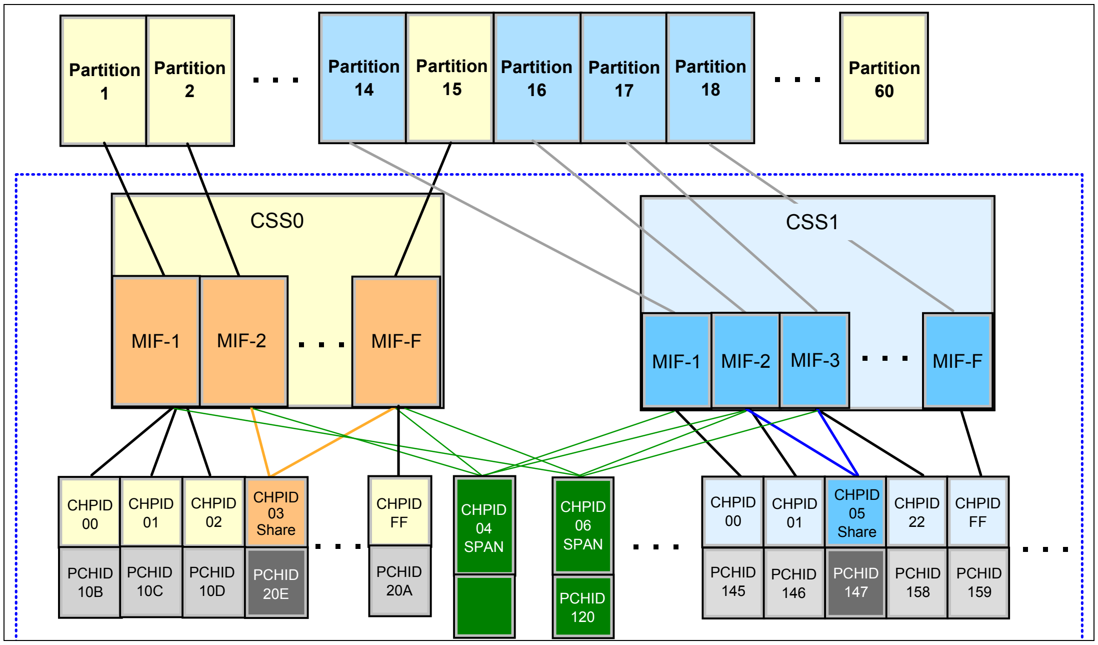


Figure 5-5 z114 CSS: Two channel subsystems with channel spanning

CHPIDs that span CSSs reduce the total number of available channels. The total is reduced, because no CSS can have more than 256 CHPIDs. For a z114 with two CSSs defined, a total of 512 CHPIDs is supported. If all CHPIDs are spanned across the two CSSs, only 256 channels are supported.

Channel spanning is supported for internal links (HiperSockets and Internal Coupling (IC) links) and for certain external links (FICON Express8S, FICON Express8, and FICON Express4 channels, OSA-Express2, OSA-Express3, OSA-Express4S, and Coupling Links).

ESCON channels: Spanning of ESCON channels is not supported.

5.1.8 Multiple CSS construct

Figure 5-6 on page 155 is a pictorial view of a z114 with multiple CSSs defined. In this example, two channel subsystems are defined (CSS0 and CSS1). Each CSS has three logical partitions with their associated MIF image identifiers.

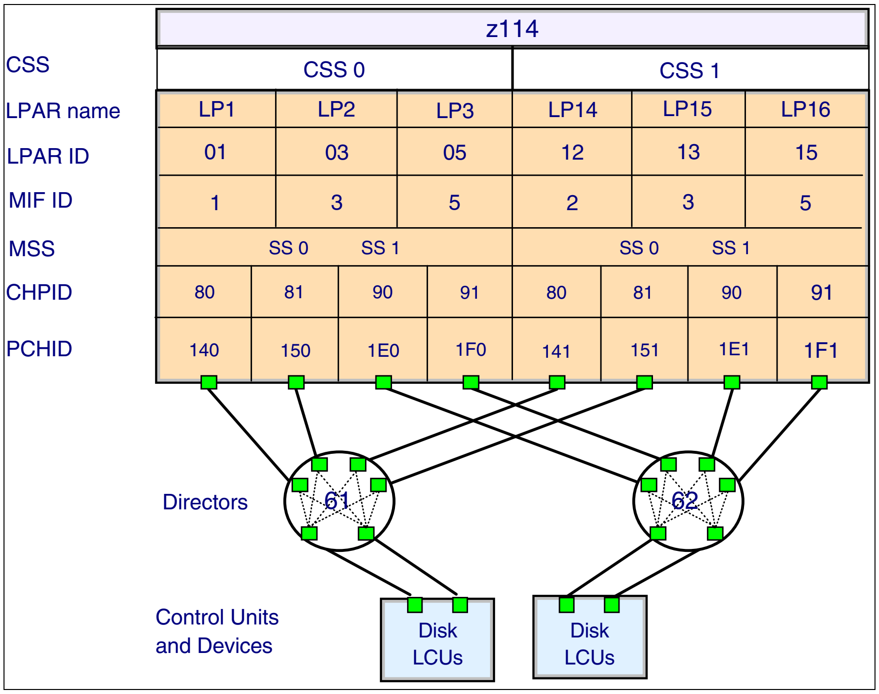


Figure 5-6 z114 CSS connectivity

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 (IFB) coupling link port, an adapter ID (AID) number is required.

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.

For detailed information, see “Adapter ID number assignment” on page 118.

5.2 I/O configuration management

For ease of management, it is preferable to use HCD 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 the 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). It is good practice to use HCD or HCM 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 you to manage both the physical and the logical 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 z114. It provides the best availability choices 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

Table 5-1 shows z114 CSS-related information in terms of maximum values for devices, subchannels, logical partitions, and CHPIDs.

Table 5-1 z114 CSS overview

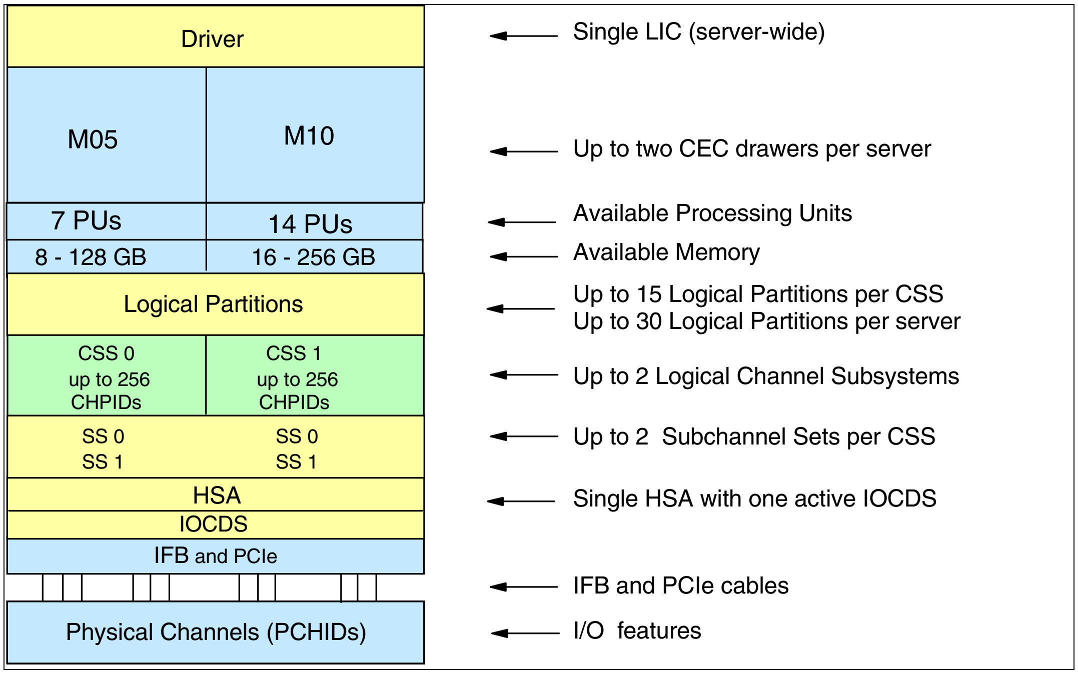
Setting	z114
Maximum number of CSSs	2
Maximum number of CHPIDs	512
Maximum number of LPARs supported per CSS	15
Maximum number of LPARs supported per system	30
Maximum number of HSA subchannels	3832.5 K (127.75 K per partition x 30 partitions)
Maximum number of devices	127.75 K (2 CSSs x 63.75 K devices)
Maximum number of CHPIDs per CSS	256
Maximum number of CHPIDs per logical partition	256
Maximum number of subchannels per logical partition	127.75 K (63.75 K + 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 8 GB for z114. This pre-allocation eliminates planning for HSA and pre-planning for HSA expansion, because HCD/IOCP always reserves the following items by the IOCDS process:

- Two CSSs
- Fifteen LPARs in each CSS
- Subchannel set 0 with 63.75 K devices in each CSS
- Subchannel set 1 with 64 K devices in each CSS

All these items 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 30 logical partitions are supported.



Driver
M05
M10
7 PUs
14 PUs
8 - 128 GB
16 - 256 GB
Logical Partitions
CSS 0 up to 256 CHPIDs
CSS 1 up to 256 CHPIDs
SS 0
SS 1
SS 0
SS 1
HSA
IOCDS
IFB and PCIe
Physical Channels (PCHIDs)
Single LIC (server-wide)
Up to two CEC drawers per server
Available Processing Units
Available Memory
Up to 15 Logical Partitions per CSS
Up to 30 Logical Partitions per server
Up to 2 Logical Channel Subsystems
Up to 2 Subchannel Sets per CSS
Single HSA with one active IOCDS
IFB and PCIe cables
I/O features

Figure 5-7 Logical view of z114 models, CSSs, IOCDS, and HSA

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 InterSystem Channel (ISC)-3 link is shared across logical partitions on a z114 server. When an I/O card is to be replaced for a repair, it usually has a few failed channels and others 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 requirement means that the IBM service support representative (SSR) 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 SE 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 SE sends another signal to the channel subsystem for each channel. This time, the channel subsystem alerts the operating system that the channel has to 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. Multipath IPL allows the 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

In this chapter, we describe the hardware cryptographic functions available on the IBM zEnterprise 114. The CP Assist for Cryptographic Function (CPACF) along with the PCIe Cryptographic Coprocessors offer a balanced use of resources and unmatched scalability.

The zEnterprise CPCs include 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-sensing and tamper-responding programmable 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 z114 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.

We cover 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 client-written programs, CPACF functions can be invoked by instructions described in the *z/Architecture Principles of Operation*, SA22-7832-08. As a group, these instructions are known as the Message-Security Assist (MSA). z/OS Integrated Cryptographic Service Facility (ICSF) callable services on z/OS, as well as in-kernel crypto APIs and the libica cryptographic functions library running at Linux on System z, also invoke CPACF synchronous functions.

The z114 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 following list shows the synchronous functions:

- Data encryption and decryption algorithms for data privacy and confidentially

  Data Encryption Standard (DES):

  - Single-length key DES
  - Double-length key DES
  - Triple-length key DES (also known as Triple-DES)

  Advanced Encryption Standard (AES) for 128-bit, 192-bit, and 256-bit keys
- Hashing algorithms for data integrity, such as SHA-1, and SHA-2 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

> **Keys:** The 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, zTPF, 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):

  - 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) functions

  Supported callable services intended for application programs that use PKA include these services:

  - Importing RSA public-private key pairs in clear and encrypted forms
  - Rivest-Shamir-Adelman (RSA):
    - Key generation, up to 4,096 bit
    - Signature generation and 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 operations 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 can be written to comply with the EMV 2000 standard for financial transactions between heterogeneous hardware and software. Support for EMV 2000 requires the PCIe feature at the zEnterprise CPC.

The Crypto Express3 card, a PCI Express cryptographic adapter, offers SHA-2 functions similar to those functions offered in the CPACF. This card is in addition to the functions mentioned.

6.2.2 CPACF protected key

The zEnterprise CPCs support 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 coprocessors (CEX3C) and the performance characteristics of the CPACF, running closer to the speed of clear keys.

An enhancement to CPACF facilitates the continued privacy of cryptographic key material when used for data encryption. In Crypto Express3 coprocessors, a secure key is encrypted under a master key, whereas a protected key is encrypted under a wrapping key that is unique to each LPAR. After 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 the hardware system area (HSA). The wrapping key is accessible only by firmware. It cannot be accessed by operating systems or applications. DES/T-DES and AES algorithms were implemented in CPACF code with the support of the hardware assist functions. Two variations of wrapping keys are generated: one version for DES/T-DES keys and another version for AES keys.

Wrapping keys are generated during the clear reset each time that an LPAR is activated or reset. No customizable option is available at the SE or HMC that permits or avoids the wrapping key generation. Figure 6-1 shows this function.

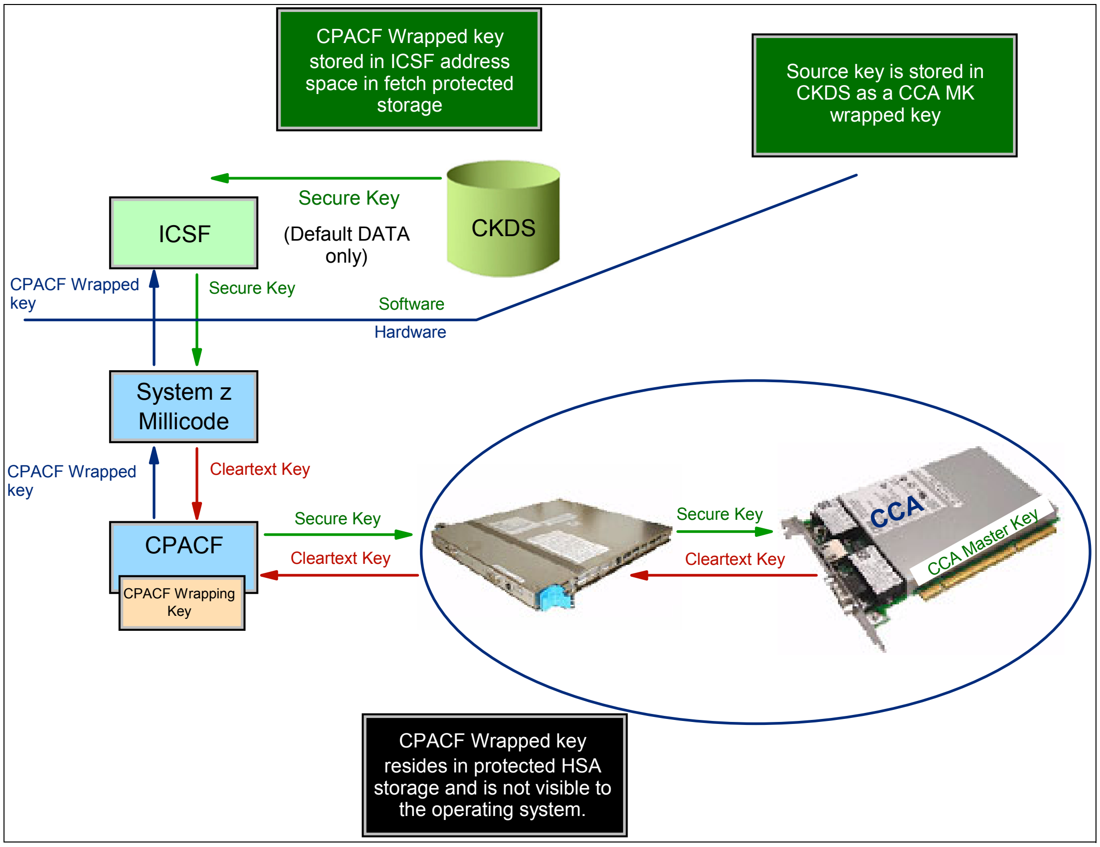

CPACF Wrapped key stored in ICSF address space in fetch protected storage
Source key is stored in CKDS as a CCA MK wrapped key
Secure Key
(Default DATA only)
CKDS
ICSF
CPACF Wrapped key
Secure Key
Software
Hardware
System z Millicode
CPACF Wrapped key
Cleartext Key
Secure Key
Cleartext Key
CPACF
CPACF Wrapping Key
Secure Key
Cleartext Key
CCA
CCA Master Key
CPACF Wrapped key resides in protected HSA storage and is not visible to the operating system.

Figure 6-1 CPACF key wrapping

If a CEX3 coprocessor is available, a protected key can begin its life as a secure key. Otherwise, an application is responsible for creating or loading a clear key value and then using the new PCKMO instruction to wrap the key. ICSF is not called by the application if CEX3C is not available.

A new segment in profiles at the CSFKEYS class in RACF® restricts which secure keys can be used as protected keys. By default, all secure keys are considered ineligible to be used as protected keys. The process that is described in Figure 6-1 considers a secure key as the source of a protected key.

In 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 then it is returned to ICSF. After the key is wrapped, ICSF can keep the protected value in memory, passing it to the CPACF, where the key will be unwrapped for each encryption/decryption operation.

The protected key is designed to provide substantial throughput improvements for a large volume of data encryption as well as low latency for encryption of small blocks of data. A high performance secure key solution, which is also known as a protected key solution, requires ICSF HCR7770, and it is highly desirable to use a 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. A new standard ANSI X9.24-2 defines the acceptable methods of doing this loading using public key cryptographic techniques. 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 the remote key loading process.

- 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 they can use keys that have no associated control information. Enhancements to key exchange functions added to CCA the capability to exchange keys between CCA systems and systems that do not use control vectors. The capability 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.

- Elliptic Curve Cryptography (ECC) Digital Signature Algorithm support

  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 that are compliant with the ECDSA method that is 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 elliptical curves. This Crypto function is supported by z/OS, z/VM, and Linux on System z.

Licensing: Elliptical Curve Cryptography technology (ECC) is delivered through the machine's Machine Code (also called Licensed Internal Code (LIC)) and requires licensing 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 zEnterprise CPC order, or when a cryptography feature is carried forward as part of an MES order into zEnterprise CPC.

- Elliptic Curve Diffie-Hellman (ECDH) algorithm support

  The Common Cryptographic Architecture has been extended to include the Elliptic Curve Diffie-Hellman (ECDH) algorithm.

  Elliptic Curve Diffie-Hellman (ECDH) is a key agreement protocol that allows two parties, each having an elliptic curve public-private key pair, to establish a shared secret over an insecure channel. This shared secret can be used directly as a key or to derive another key, which can then be used to encrypt subsequent communications using a symmetric key cipher, such as AES key encrypting keys (KEK). This list shows the enhancements:

  - Key management function to support AES KEK
  - Generation of an ECC private key wrapped with an AES KEK
  - Import and export of an ECC private key wrapped with an AES KEK
  - Support for ECDH with a new service

- PKA RSA OAEP with SHA-256 algorithm

  RSA Encryption Scheme - Optimal Asymmetric Encryption Padding (RSA OAEP) is a public-key encryption scheme or method of encoding messages and data in combination with the RSA algorithm and a hash algorithm.

  Currently, the Common Cryptographic Architecture and z/OS Integrated Cryptographic Service Facility (ICSF) provide key management services supporting the RSA OAEP method using the SHA-1 hash algorithm, as defined by the Public Key Cryptographic standards (PKCS) #1 V2.0 standard. These services can be used to exchange AES or DES/TDES key values securely between financial institutions and systems. However, PKCS#1 V2.1 extends the OAEP method to include the use of the SHA-256 hashing algorithm to increase the strength of the key wrapping and unwrapping mechanism. The CCA key management services have been enhanced so that they can use RSA OAEP with SHA-256 in addition to RSA OAEP with SHA-1. This enhancement provides support for PKCS that is mandated by certain countries for interbank transactions and communication systems.

- User-Defined Extensions (UDX) support

  UDX allows the user to add customized operations to a cryptographic coprocessor. User-Defined Extensions to the Common Cryptographic Architecture (CCA) support customized operations that execute within the Crypto Express features when defined as a coprocessor.

  UDX is supported under a special contract through an IBM or approved third-party service offering. The CryptoCards website directs your request to an IBM Global Services location that is appropriate for your geographic location. A special contract is negotiated between you and IBM Global Services. The contract is for the 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.

A UDX toolkit for System z is available for the Crypto Express3 feature. In addition, there is a migration path for clients with UDX on a previous feature to migrate their code to the Crypto Express3 feature. A UDX migration is no more disruptive than a normal Machine Change Level (MCL) or ICSF release migration.

For more information, see the IBM CryptoCards website:

http://www.ibm.com/security/cryptocards

6.2.4 Cryptographic feature codes

Table 6-1 lists the available cryptographic features.

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 can be ordered. Each feature contains two PCI Express cryptographic adapters (adjunct processors).
0871	Crypto Express3-1P feature A maximum of eight features can be ordered. Each feature contains one PCI Express cryptographic adapter (adjunct processor).
0841	Trusted Key Entry (TKE) workstation This feature is optional. TKE provides simple key management (key identification, exchange, separation, update, and 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 10 features per zEnterprise CPC can be installed.
0867	TKE 7.1 Licensed Internal Code (TKE 7.1 LIC) The 7.1 LIC requires Trusted Key Entry workstation feature code 0841. It is required to support zEnterprise CPC. The 7.1 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 feature code includes two Smart Card Readers, two cables to connect to the TKE 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 7.1 LIC or later is required. See 6.5, “TKE workstation feature” on page 173 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 client'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 with the zEnterprise CPC:

- Compute intermediate Message Digest (KIMD) adds KIMD-GHASH

These functions are provided as problem-state z/Architecture instructions, which are directly available to application programs. These instructions are known as Message-Security Assist (MSA). When enabled, the CPACF runs at processor speed for every CP, IFL, 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 client's site as an MES installation, except for SHA-1, and SHA-2 support for SHA-224, SHA-256, SHA-384, and SHA-512, which are always enabled.

[2] Federal Information Processing Standard

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 that is 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. There is a separate service processor that is used to manage self-test and firmware updates, RAM, flash memory, and battery-powered memory, cryptographic-quality 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.

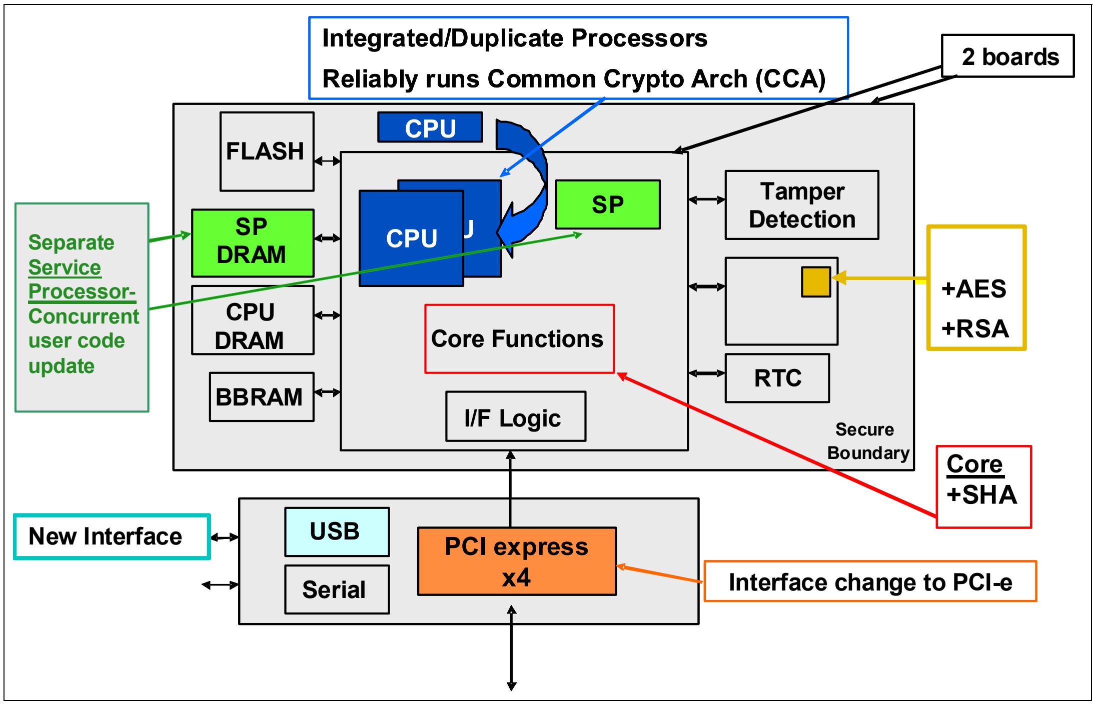


Figure 6-2 Crypto Express3 feature layout

The Crypto Express3 feature does not have external ports and does not use fiber optic or other cables. It does not use CHPIDs, but it requires one slot in the I/O drawer and one PCHID for each PCI-e cryptographic adapter. The removal of the feature or card "*zeros out*" the content.

The z114 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.

Adapter: Though 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 drawer of the z114, continues to support all of the cryptographic functions that are 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, which were introduced at z114, are supported:

- Expanded key support for AES algorithm

  CCA currently supports the Advanced Encryption Standard (AES) algorithm to allow the use of AES keys to encrypt data. Expanded key support for AES adds a framework to support a much broader range of application areas, and it lays the groundwork for future use of AES in areas where standards and client applications are expected to evolve.

  As stronger algorithms and longer keys become increasingly common, security requirements dictate that these keys must be wrapped using key encrypting keys (KEKs) of sufficient strength. This feature adds support for AES key encrypting keys. These AES wrapping keys have adequate strength to protect other AES keys for transport or storage. The new AES key types use the variable length key token. The supported key types are EXPORTER, IMPORTER, and for use in the encryption and decryption services, CIPHER.

  New APIs have been added or modified to manage and use these new keys.

  The following new or modified CCA API functions are also supported:

  - Key Token Build2: Builds skeleton variable length key tokens
  - Key Generate2: Generates keys using random key data
  - Key Part Import2: Creates keys from key part information
  - Key Test2: Verifies the value of a key or key part
  - Key Translate2:
    - Translates a key: Changes the key encrypting key (KEK) that is used to wrap a key
    - Reformats a key: Converts keys between the previous token format and the newer variable length token format
  - Symmetric Key Export: Modified to also export AES keys
  - Symmetric Key Import2: Imports a key that has been wrapped in the new token format
  - Secure Key Import2 (System z-only): Wraps key material under the master key or an AES KEK
  - Restrict Key Attribute: Changes the attributes of a key token
  - Key Token Parse2: Parses key attributes in the new key token
  - Symmetric Algorithm Encipher: Enciphers data
  - Symmetric Algorithm Decipher: Deciphers data

- Enhanced ANSI TR-31 interoperable secure key exchange

  ANSI TR-31 defines a method of cryptographically protecting Triple Data Encryption Standard (TDES) cryptographic keys and their associated usage attributes. The TR-31 method complies with the security requirements of the ANSI X9.24 Part 1 standard, although use of TR-31 is not required in order to comply with that standard. CCA has added functions that can be used to import and export CCA TDES keys in TR-31 formats.

These functions are designed primarily as a secure method of wrapping TDES keys for improved and more secure key interchange between CCA and non-CCA devices and systems.

- PIN block decimalization table protection

  To help avoid a decimalization table attack to learn a personal identification number (PIN), a solution is now available in the CCA to thwart this attack by protecting the decimalization table from manipulation. PINs are most often used for automated teller machines (ATMs) but are increasingly used at the point-of-sale, for debit and credit cards.

- ANSI X9.8 PIN security

  This enhancement 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 (PINs).

- Enhanced 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 keys 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.

- Enhanced Driver Maintenance (EDM) and Concurrent Machine Change Level (MCL) apply

  This enhancement is a process to eliminate or reduce cryptographic coprocessor card outages for new Cryptographic function releases. With Enhanced Driver Maintenance and Concurrent MCL applied, new cryptographic functions can be applied without configuring the Cryptographic coprocessor card off and on. It is now possible to upgrade Common Cryptographic Architecture (CCA), segment 3, LIC without any performance impact during the upgrade. However, certain levels of Common Cryptographic Architecture (CCA) or hardware changes will still require cryptographic coprocessor card vary off and on. This Crypto function is exclusive to the zEnterprise CPC.

The enhancements include the following additional key features of Crypto Express3:

- 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) having access to 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 PCIe 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 separate 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 that is 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 the FIPS 140-2 Level 4 compliance rating for secure cryptographic hardware modules. Unauthorized removal of the adapter or feature *“zeros out”* its content.

The Crypto Express3 coprocessor enables the user to perform the following tasks:

- 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).
- Use CEX3C, which supports 13-digit 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 lower performance than the Crypto Express3 accelerator.

Three methods of master key entry are provided by Integrated Cryptographic Service Facility (ICSF) for the Crypto Express3 feature coprocessor:

- A passphrase 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 the 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. It operates 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 on a 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.

Coprocessors: Cryptographic coprocessors are not tied to logical partition numbers or MIF IDs. They are set up with PCI-e adapter numbers and domain indexes that are defined in the partition image profile. The client 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 a coprocessor 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):

- 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 zEnterprise CPC supports up to eight Crypto Express3 features, which equals up to a maximum of 16 PCI-e cryptographic adapters. Table 6-2 summarizes configuration information for Crypto Express3.

Table 6-2 Crypto Express3 feature

Feature	Number of adapters
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)[b]	2
Maximum number of PCI-e adapters for each server	16
Number of cryptographic domains for each PCI-e adapter[c]	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] If running Crypto Express3-1P, we have only one PCI-e adapter per feature.

[c] More than one partition, defined to the same CSS or to separate CSSs, can use the same domain number when assigned to separate PCI-e cryptographic adapters.

The concept of *dedicated processor* does not apply to the PCI-e cryptographic adapter. Whether configured as a 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 needs to be planned ahead to allow for nondisruptive changes:

- 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 are dynamic.
- The same usage domain index can 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 can 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 z114 allows for up to 30 logical partitions to be active concurrently. Each PCI Express supports 16 domains, whether it is configured as a Crypto Express3 accelerator or as 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 30 logical partitions require concurrent access to cryptographic functions. More Crypto Express3 features might be needed to satisfy application performance and availability requirements.

6.5 TKE workstation feature

The Trusted Key Entry (TKE) 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 writable 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.

Adapters: 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. You use the TKE to manage the cryptographic keys of a z114 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 Coprocessors. 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. You can use the TKE feature code 0841 to control the z114 and also z196, 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 that is 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 that is used for the passphrase logon to the TKE workstation crypto adapter. The passphrase must meet the following requirements:

  – Must be 8 to 64 characters in length
  – Contains at least two numeric and two non-numeric characters
  – Does 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 separate 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 z114. This wizard offers the following benefits:

  – Reduces migration steps, thereby minimizing user errors
  – Minimizes the number of user clicks
  – Significantly reduces migration task duration

The TKE workstation feature code 0841 along with LIC 7.1 offers additional enhancements:

- New access control support for all TKE applications

  Every TKE application and the ability to create and manage the crypto module and domain groups now require the TKE local cryptographic adapter profile to have explicit access to the TKE application or function that the user wants to run. This enhancement provides more control of the functions that TKE users are allowed to perform.

- New migration utility

  During a migration from a lower release of TKE to TKE 7.1 LIC, it will be necessary to add access control points to the existing roles. The new access control points can be added through the new Migrate Roles Utility or by manually updating each role through the Cryptographic Node Management Utility. The IBM-supplied roles created for TKE 7.1 LIC have all of the access control points that are needed to perform the functions they were permitted to use in TKE releases prior to TKE 7.1 LIC.

- Single process for loading an entire key

  The TKE now has a wizard-like feature that takes users through the entire key loading procedure for a master or operational key. The feature preserves all of the existing separation of duties and authority requirements for clearing, loading key parts, and completing a key. The procedure saves time, by walking users through the key loading procedure. However, this feature does not reduce the number of people that it takes to perform the key load procedure.

- Single process for generating multiple key parts of the same type

  The TKE now has a wizard-like feature that allows a user to generate more than one key part at a time. The procedure saves time because the user has to start the process only one time, and the TKE efficiently generates the desired number of key parts.

- AES operational key support

  CCA V4.2 for the Crypto Express3 feature includes three new AES operational key types. From the TKE, users can load and manage the new AES EXPORTER, IMPORTER, and CIPHER operational keys from the TKE workstation crypto module notebook.

- Decimalization table support

  CCA V4.2 for the Crypto Express3 feature includes support for 100 decimalization tables for each domain on a Crypto Express3 feature. From the TKE, users can manage the decimalization tables on the Crypto Express3 feature from the TKE workstation crypto module notebook. Users can manage the tables for a specific domain or manage the tables of a set of domains if they use the TKE workstation Domain Grouping function.

- Host cryptographic module status support

  From the TKE workstation crypto module notebook, users will be able to display the current status of the host cryptographic module that is being managed. If they view the Crypto Express3 feature module information from a crypto module group or a domain group, they will see only the status of the group’s master module.

- Display of active IDs on the TKE console

  A user can be logged onto the TKE workstation in privileged access mode. In addition, the user can be signed onto the TKE workstation’s local cryptographic adapter. If a user is signed on in privileged access mode, that ID is shown on the TKE console. With this new support, both the privileged access mode ID and the TKE local cryptographic adapter ID will be displayed on the TKE console.

- Increased number of key parts on a smart card

  If a TKE smart card is initialized on a TKE workstation with a 7.1 level of LIC, it will be able to hold up to 50 key parts. Previously, TKE smart cards held only 10 key parts.

- Use of ECDH to derive a shared secret

  When the TKE workstation with a 7.1 level of LIC exchanges encrypted material with a Crypto Express3 at CCA Level V4.2, Elliptic Curve Diffie-Hellman (ECDH) is used to derive the shared secret. This function increases the strength of the transport key that is used to encrypt the material.

6.5.1 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 that is assigned to the logical partitions that are defined to the TKE workstation.

Each logical partition in the same system that uses a domain that is managed through a TKE workstation connection is either a TKE host or a TKE target. A logical partition with a TCP/IP connection to the TKE is referred to as a TKE host. All other partitions are TKE targets.

The cryptographic control setting for a logical partition through the Support Element determines whether the workstation is a TKE host or TKE target.

6.5.2 Optional smart card reader

Adding an optional smart card reader (FC 0885) to the TKE workstation is possible. One 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. The access to and the 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. One feature code is ordered, and a quantity of ten smart cards is shipped. The order increment is one up to 99 (990 blank smart cards).

6.6 Cryptographic functions comparison

Table 6-3 lists the functions or attributes on z114 of the three cryptographic hardware features.

Table 6-3 Cryptographic functions on z114

Functions or attributes	CPACF	Crypto Express3 Coprocessor	Crypto Express3 Accelerator
Supports z/OS applications using ICSF	Yes	Yes	Yes
Supports Linux on System z CCA applications	Yes	Yes	Yes
Encryption and decryption using secret-key algorithm	No	Yes	No
Provides highest SSL/TLS handshake performance	No	No	Yes[a]
Provides highest symmetric (clear key) encryption performance	Yes	No	No
Provides highest asymmetric (clear key) encryption performance	No	No	Yes
Provides highest asymmetric (encrypted key) encryption performance	No	Yes	No
Disruptive process to enable	No	Note[b]	Note[b]
Requires IOCDS definition	No	No	No
Uses CHPID numbers	No	No	No
Uses PCHIDs		Yes [c]	Yes[c]
Requires CPACF enablement (FC 3863)	Yes[d]	Yes[d]	Yes[d]
Requires ICSF to be active	No	Yes	Yes
Offers user programming function (UDX)	No	Yes	No
Usable for data privacy: encryption and decryption processing	Yes	Yes	No
Usable for data integrity: hashing and message authentication	Yes	Yes	No

Functions or attributes	CPACF	Crypto Express3 Coprocessor	Crypto Express3 Accelerator
Usable for financial processes and key management operations	No	Yes	No
Crypto performance RMF™ monitoring	No	Yes	Yes
Requires system master keys to be loaded	No	Yes	No
System (master) key storage	No	Yes	No
Retained key storage	No	Yes	No
Tamper-resistant hardware packaging	No	Yes	Yes[e]
Designed for FIPS 140-2 Level 4 certification	No	Yes	No
Supports SSL functions	Yes	Yes	Yes
Supports Linux applications doing SSL handshakes	No	No	Yes
RSA functions	No	Yes	Yes
High performance SHA-1 and SHA2	Yes	Yes	No
Clear key DES or triple DES	Yes	No	No
Advanced Encryption Standard (AES) for 128-bit, 192-bit, and 256-bit keys	Yes	Yes	No
Pseudorandom number generator (PRNG)	Yes	No	No
Clear key RSA	No	No	Yes
Europay Mastercard VISA (EMV) support	No	Yes	No
Public Key Decrypt (PKD) support for Zero-Pad option for clear RSA private keys	No	Yes	Yes
Public Key Encrypt (PKE) support for MRP function	No	Yes	Yes
Remote loading of initial keys in ATM	No	Yes	No
Improved key exchange with non-CCA system	No	Yes	No
ISO 16609 CBC-mode triple DES MAC support	No	Yes	No

a. This function 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. CPACF 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 function is physically present but is not used when configured as an accelerator (clear key only).

6.7 Software support

We list the software support levels in 8.4, “Cryptographic support” on page 252.

7

zEnterprise BladeCenter Extension Model 002

IBM has extended the role of the mainframe by adding new infrastructure that is based on the IBM BladeCenter. It is called the zEnterprise BladeCenter Extension (zBX) Model 002.

The zBX brings the computing capacity of systems in a blade form factor to the zEnterprise System. It is designed to provide a redundant hardware infrastructure that supports the multiple platform environment of the zEnterprise System in a seamlessly integrated way.

Key to the zEnterprise System is also the Unified Resource Manager (URM), which helps deliver end-to-end virtualization and management, as well as the ability to optimize multiple platform technology deployment according to applications' requirements. For more information about Unified Resource Manager, refer to Chapter 12, "Hardware Management Console" on page 345 and *IBM zEnterprise Unified Resource Manager,* SG24-7921.

In this chapter, we introduce the zBX Model 002 and describe its hardware components. We also explain the concepts and building blocks for zBX connectivity.

You can use the information in this chapter for planning purposes and to help define the configurations that best fit your requirements.

We cover the following topics:

7.1 zBX concepts

The IBM zEnterprise System represents a new height for mainframe functionality and qualities of service (QoS). It has been rightly portrayed as a cornerstone for the IT infrastructure, especially when flexibility for rapidly changing environments is required.

IBM zEnterprise System characteristics make it especially valuable for mission-critical workloads. Today, most of these applications have multi-tiered architectures that span various hardware and software platforms. However, there are differences in the QoS 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 platforms require personnel with distinct skill sets, various 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 affect the bottom line.

IBM mainframe systems have been providing specialized hardware and fit-for-purpose (tuned to the task) computing capabilities for a long time. In addition to the machine instruction assists, another example was 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 are not 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. zEnterprise with its zBX infrastructure offers the capability of running an application that spans z/OS, z/VM, zVSE, Linux on System z, AIX on POWER7, and Linux on System x, yet has it under a single management umbrella. Also, zBX can host and integrate special-purpose workload optimizers, such as the IBM Smart Analytics Optimizer and WebSphere DataPower Integration Appliance XI50 for zEnterprise (DataPower XI50z).

7.2 zBX hardware description

The zBX has a machine type of 2458-002 and is exclusive to the zEnterprise central processor complexes (CPCs). 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 IBM 42U zEnterprise racks with the required network and power infrastructure
- One to eight BladeCenter chassis with a combination of up to 112[1] separate blades
- Redundant infrastructure for fault tolerance and higher availability
- Management support through the z114 Hardware Management Console (HMC) and Support Element (SE)

You can read more about zBX reliability, availability, and serviceability (RAS) in 10.5, "RAS capability for zBX" on page 324.

You can order the zBX with a new z114 or as an MES to an existing z114. Either way, the zBX is treated as an extension to a z114 and cannot be ordered as a stand-alone feature.

Figure 7-1 shows a z114 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.

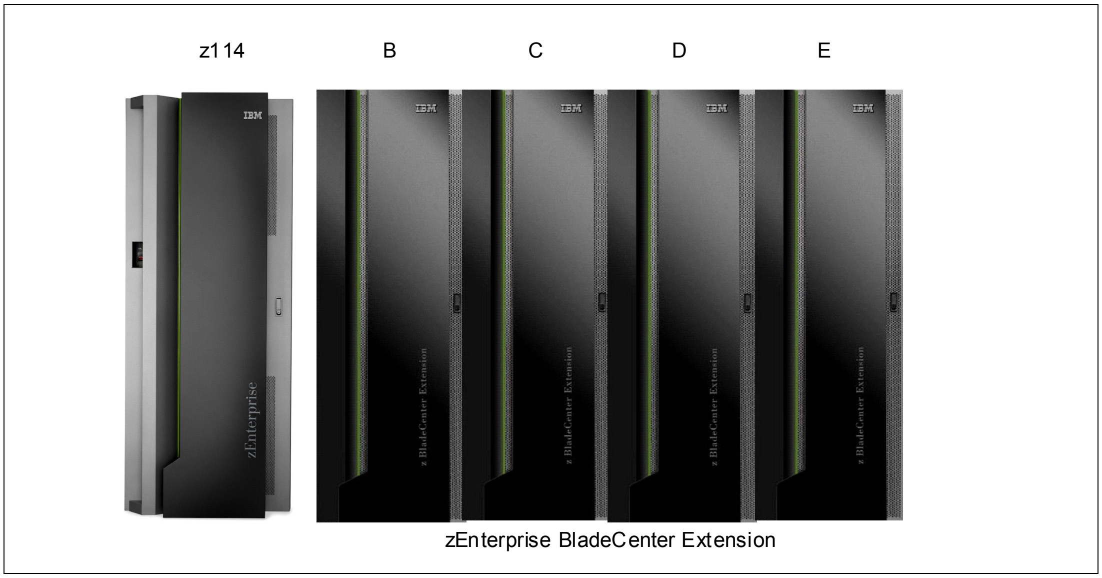


Figure 7-1 z114 with a maximum zBX configuration

7.2.1 zBX racks

The zBX Model 002 (2458-002) hardware is housed in up to four IBM zEnterprise racks. Each rack is an industry-standard 19-inch, 42U-high rack, and has four sidewall compartments to support the installation of power distribution units (PDUs) and switches, with additional space for cable management.

[1] The number of chassis and blades varies depending on the type of blades configured within the zBX. See 7.2.4, "zBX blades" on page 187 for more information.

Figure 7-2 shows the rear view of a two-rack zBX configuration:

- Two Top of Rack (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:
  - Up to 14 blade server slots per chassis
  - Advanced management modules (AMMs)
  - Ethernet switch modules (ESMs)
  - High-speed switch (HSS) modules
  - 8 Gbps Fibre Channel (FC) switches for connectivity to the SAN environment[1]
  - Blower modules
- Power distribution units (PDUs)

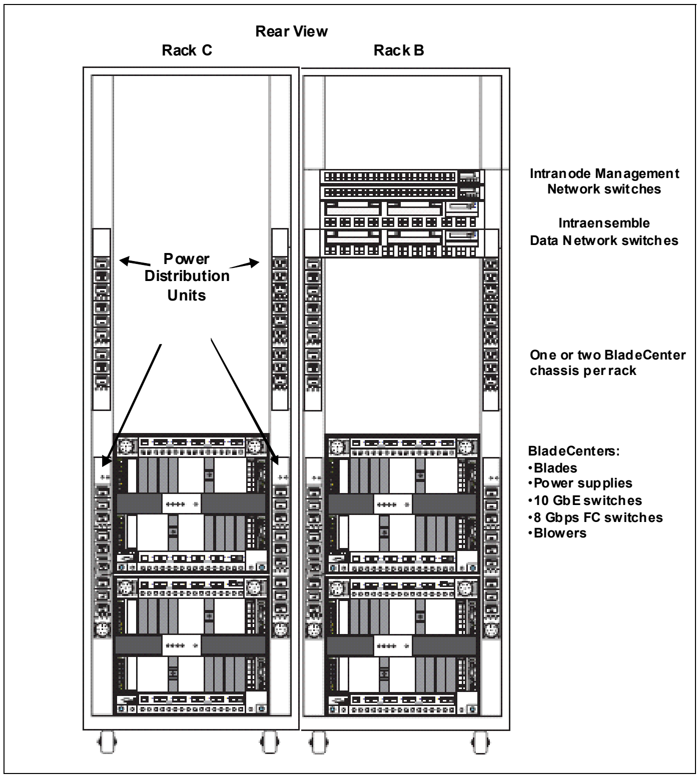


Figure 7-2 zBX racks: Rear view with BladeCenter chassis

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.

The zBX racks ship with lockable standard non-acoustic doors and side panels. The following optional features are also available:

- IBM rear door heat eXchanger (Feature Code (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 heat of the zBX to chilled water (customer-supplied data center

[1] Client supplied FC switches are required and must support N_Port ID Virtualization (NPIV). Certain FC switch vendors also require “interop” mode. Check the interoperability matrix for the latest details: http://www-03.ibm.com/systems/support/storage/ssic/interoperability.wss

infrastructure). The rear door heat eXchanger requires external conditioning units for its use.

- IBM acoustic door (FC 0543) can be used to reduce the acoustical noise from the zBX.
- Height reduction (FC 0570) reduces the rack height to 36U high and accommodates doorway openings as low as 1,832 mm (72.1 in). Order this choice if you have doorways with openings less than 1,941 mm (76.4 in) 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) does 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 the top down, there 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 z114 through the INMN connections. Each VLAN-capable 1000BASE-T switch has 48 ports. The switch ports are reserved in the following manner:

- One port for each of the two bulk power hubs (BPH) on the controlling z114
- One port for each of the advanced management modules (AMM) and Ethernet switch modules (ESM), in each zBX BladeCenter chassis
- One port for each of the two IEDN 10 GbE TOR switches
- Two ports each for interconnecting the two switches

Both switches have the same connections to the corresponding redundant components (BPH, AMM, ESM, and IEDN TOR switches) to avoid any single point of failure. Table 7-5 on page 196 shows port assignments for the 1000BASE-T TOR switches.

Important: Although IBM provides a 26 m (85.3 ft) cable for the INMN connection, it is best to have the zBX installed next to or near the *controlling* z114 server, for easy access to the zBX for service-related activities or tasks.

Each virtual LAN (VLAN)-capable 10 GbE TOR switch has 40 ports that are dedicated to the IEDN. The switch ports have the following connections:

- Up to eight ports are used for connections to an HSS module (SM07 and SM09) of each BladeCenter chassis in the same zBX (as part of IEDN), to provide data paths to blades.
- Up to eight ports are used for OSA-Express4S 10 GbE or OSA-Express3 10 GbE (long reach (LR) or short reach (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 are used for zBX to zBX connections within the same ensemble (as part of the IEDN).
- Up to nine ports are used for the customer-managed data network. These connections are not part of IEDN, and they cannot be managed or provisioned by the Unified Resource Manager. The Unified Resource Manager will recognize these connections as migration connections and provide access control for their connection to the 10 GbE TOR switches.
- One port is the management port that connects to the INMN 1000BASE-T TOR switch.
- Two ports are used for interconnections between two switches (as a failover path), using two direct-attached cables (DAC) to interconnect both switches.

Table 7-7 on page 200 shows port assignments for the 10GbE TOR switches. See Figure 7-3 for a graphical illustration of zEnterprise network connections. For more information about the connectivity options for the INMN and the IEDN, as well as the connectivity rules, see 7.4, “zBX connectivity” on page 194.

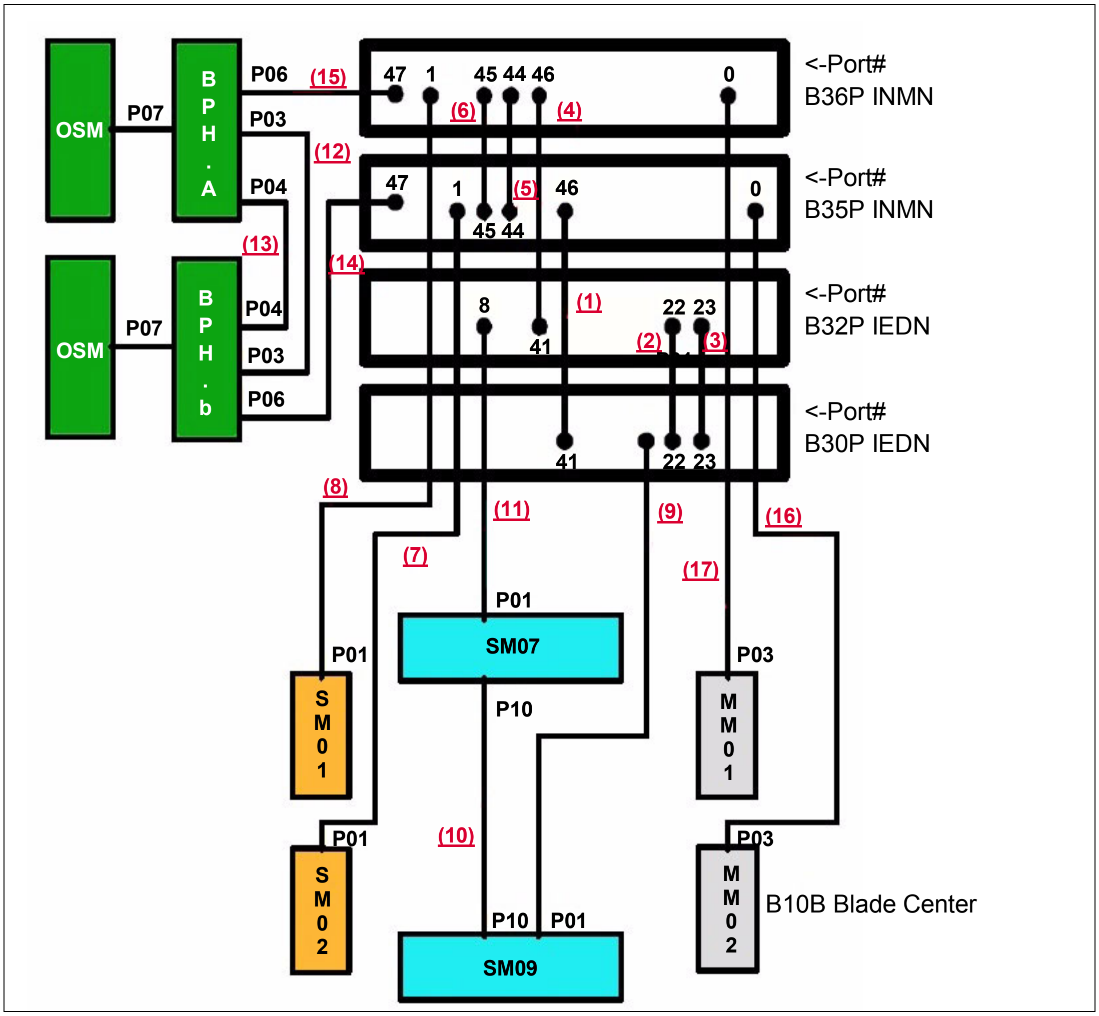


Figure 7-3 Graphical illustration of zEnterprise network connections

7.2.3 zBX BladeCenter chassis

Each zBX BladeCenter chassis is designed with additional components that are installed for high levels of resiliency.

The front of a zBX BladeCenter chassis has the following components:

- Blade server slots:

  There are 14 available blade server slots (BS01 to BS14) in a zBX BladeCenter chassis. Each slot is capable of housing any zBX-supported blades, with the following restrictions:

  - Slot 14 cannot hold a double-wide blade.
  - The DataPower XI50z blades are double-wide. Each feature takes two adjacent BladeCenter slots, so the maximum number of DataPower blades per BladeCenter is seven. The maximum number of DataPower blades per zBX is 28.

Placement of blades: Blades must be sequentially plugged into each BladeCenter (plug a blade into the first slot and go linearly out; slots must not be skipped).

- Power module:

  The power module (PM) includes a power supply and a package of three fans. Two of the three fans are needed for a power module operation. Power modules 1 and 2 (PM01 and PM02) are installed as a pair to supply power for the seven blade server slots from BS01 to BS07. Power modules 3 and 4 (PM03 and PM04) support slots BS08 to BS14.

  The two power connectors (marked with “1” and “2” in Figure 7-4) provide power connectivity for the power modules (PM) and blade slots. PM01 and PM04 are connected to power connector 1. PM02 and PM03 are connected to power connector 2. Thus, each slot has fully redundant power from a different power module that is connected to a different power connector.

Figure 7-4 shows the rear view of a zBX BladeCenter chassis.

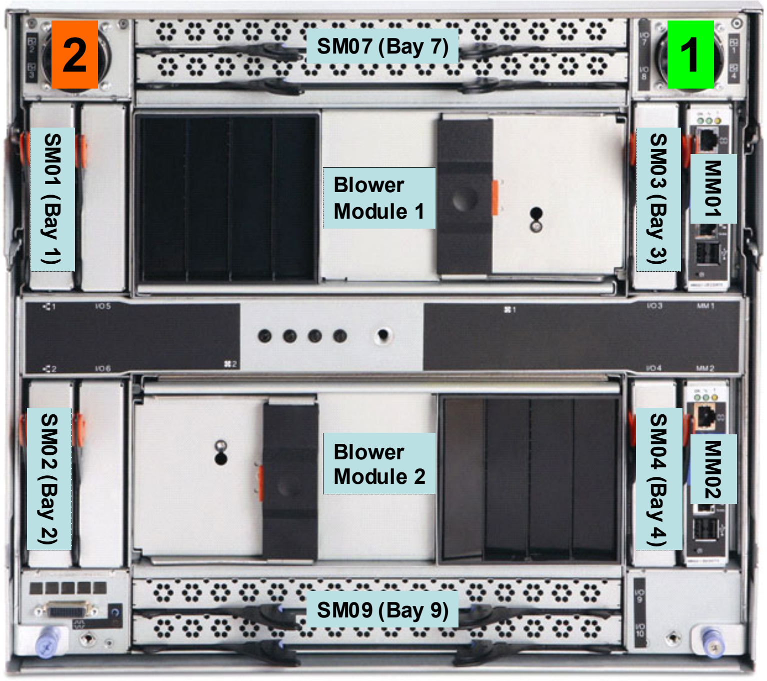


Figure 7-4 zBX BladeCenter chassis rear view

The rear of a zBX BladeCenter chassis has the following components:

- Advanced management module:

  The advanced 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 of the controlling z114, with all the service management and reporting function of AMMs.

Two AMMs (MM01 and MM02) are installed in the zBX BladeCenter chassis. Only one AMM has primary control of the chassis (it is active); the second module is in passive (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 (ESMs) - SM01 and SM02 - 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 01 is connected to one of the 1000BASE-T TOR switches. As part of the INMN, configuration and firmware of ESM is controlled by the controlling z114 Support Element (SE).

- High-speed switch module:

  Two high-speed switch (HSS) modules (SM07 and SM09) are installed to the switch bays 7 and 9. The HSS modules provide 10 GbE uplinks to 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 switches. Port 10 is used to interconnect HSS in bays 7 and 9 as a failover path.

- Eight Gbps Fibre Channel switch module:

  Two 8 Gbps Fibre Channel (FC) switches (SM03 and SM04) 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 FC ports to provide connectivity to the SAN environment.

- 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 the internal BladeCenter components. If a blower fails, the remaining blowers run full speed.

- BladeCenter mid-plane fabric connections:

  The BladeCenter mid-plane provides redundant power, control, and data connections to a blade server by internally routed chassis components (power modules, AMMs, switch modules, and 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 MM01, SM01, and SM03.

  – Power connector from power module 1 (blade server slots 1 to 7) or power module 3 (blade server slots 8 to 14).

  – Top 4X fabric connects blade to SM07.

  – Bottom 4X fabric connects blade to SM09.

  – Bottom 1X fabric connects blade to MM02, SM02, and SM04.

  – Power connector from power module 2 (blade server slots 1 to 7) or power module 4 (blade server slots 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 supports the following blade types:

- POWER7 blades:

  Three configurations of POWER® blades are supported, depending on their memory sizes (see Table 7-2 on page 191). The number of blades can be from one to 112.

- IBM WebSphere DataPower XI50 for zEnterprise blades:

  Up to 28 IBM WebSphere DataPower XI50 for zEnterprise (DataPower XI50z) blades are supported. These blades are double-wide (each blade occupies two blade server slots).

- IBM HX5 System x blades:

  Up to 28 IBM System x HX5 blades are 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 through the mid-plane to the high speed switches and switch modules in the chassis.

Depending on the blade type, 10 GbE CFFh expansion cards and 8 Gbps Fibre Channel CIOv expansion cards provide I/O connectivity to the IEDN, INMN, or client-supplied SAN-attached storage disks.

POWER7 blade

The POWER7 blade (Table 7-1) is a single-width blade, which includes a POWER7 processor, up to 16 dual inline memory modules (DIMMs), and an HDD. The POWER7 blade supports 10 GbE connections to IEDN, and 8 Gbps FC connections to client-provided Fibre Channel storage through the Fibre Channel (FC) switches (SM03 and SM04) in the chassis.

The POWER7 blade is loosely integrated to a zBX, so that you can acquire supported blades through existing channels from IBM. The primary HMC and SE of the controlling z114 perform entitlement management for installed POWER7 blades on a one-blade basis.

Table 7-1 Supported configurations of POWER7 blades

Feature	Feature code	Configuration 1 quantity	Configuration 2 quantity	Configuration 3 quantity
Processor (3.0GHz@150W)	8412	8	8	8
8 GB Memory	8208	4	8	0
16 GB Memory	8209	0	0	8
Internal HDD (300 GB)	8274	1	1	1
CFFh 10 GbE expansion	8275	1	1	1
CIOv 8 Gb FC expansion	8242	1	1	1

Feature	Feature code	Configuration 1 quantity	Configuration 2 quantity	Configuration 3 quantity
PowerVM Enterprise Edition	5228	8	8	8

DataPower XI50z blades

The IBM WebSphere DataPower Integration Appliance XI50 for zEnterprise (DataPower XI50z), which is integrated into the zEnterprise System, is a high-performance hardware appliance that offers these functions:

- Provides fast and flexible integration with any-to-any transformation between diverse message formats with integrated message-level security and superior performance.
- Provides web services enablement for core System z applications to enable web-based workloads. As a multifunctional appliance, DataPower XI50z can help provide multiple levels of XML optimization and streamline and secure valuable service-oriented architecture (SOA) applications.
- Enables SOA and XML applications with System z web services for seamless integration of distributed and System z platforms. It can help to simplify, govern, and enhance the network security for XML and web services.
- Provides drop-in integration for heterogeneous environments by enabling core enterprise service bus (ESB) functionality, including routing, bridging, transformation, and event handling.
- Offers standards-based, centralized System z governance, and extreme reliability through integrated operational controls, “call home” capability, and integration with RACF security through a secured private network.

The zBX provides additional benefits to the DataPower appliance environment in these areas:

- Blade hardware management:
  - Improved cooling and power management controls, includes cooling of the frame and energy monitoring and management of the DataPower blades.
  - Virtual network provisioning.
  - Call home capability for current and expected problems.
- Hardware Management Console integration:
  - Single view showing the System z environment together with the DataPower blades in an overall hardware operational perspective.
  - Group GUI operations for functions that are supported on the HMC, such as activating or deactivating blades.
- Improved availability:
  - Guided placement of blades to optimize built-in redundancy in all components at the rack, BladeCenter, and HMC levels, including the TOR switch, ESM switches, and physical network.
  - Detection and reporting by the HMC/SE on appliance failures. The HMC/SE can also be used to recycle the DataPower appliance.
- Networking:
  - Virtual network provisioning.
  - Enforced isolation of network traffic via VLAN support.
  - 10 Gb end-to-end network infrastructure.
  - Built-in network redundancy.

  - Network protection via IEDN, possibly obviating any perceived need for the encryption of flows between the DataPower and the target back-end System z server.
- Monitoring and reporting:
  - Monitoring and reporting the DataPower hardware health and degraded operations via the HMC.
  - Monitoring all hardware, call home, and automated parts replacement.
  - Consolidating and integrating the DataPower hardware problem reporting with other problems that are reported in the zBX.
- System z value:
  - Simplified ordering of the DataPower appliance via System z allows the proper blade infrastructure to be transparently ordered.
  - Simplified upgrades keep MES history so that the upgrades flow based on the installed components.
  - System z service on the zBX and the DataPower blade with a single point of service.
  - The DataPower appliance becomes part of the data center and comes under data center control.

In addition, although not specific to the zBX environment, dynamic load balancing to the DataPower appliances is available using the z/OS Communications Server Sysplex Distributor.

Configuration

The DataPower XI50z is a double-wide IBM HS22 blade. Each DataPower XI50z takes two slots, so the maximum number of DataPower blades per BladeCenter is seven. The maximum number of DataPower blades per zBX is 28. The DataPower XI50z can coexist with POWER7 blades and with IBM BladeCenter HX5 blades in the same zBX BladeCenter. DataPower XI50z blades are configured and ordered as zBX (machine type 2458-002) features, but they have their own machine type (2462-4BX).

The DataPower XI50z with DataPower expansion unit has the following specifications:

- 2.13 GHz.
- 2x quad core processors.
- 8 M cache.
- 3 x 4 Gb DIMMs (12 Gb memory).
- 4 Gb USB Flash Key that contains the DataPower XI50z firmware load.
- 2 x 300 GB HDDs that are used by the client for logging, storing style sheets, and storing XML files. The hard disk array consists of two hard disk drives in a RAID-1 (mirrored) configuration.
- Broadcom BCM5709S x2 with TOE (integrated on planar).
- BPE4 Expansion Unit. It is a sealed field-replaceable unit (FRU) with one-way tamperproof screws (it contains the crypto for secure SOA applications).
- XG5 accelerator PCI-e card.
- CN1620 Cavium crypto PCI-e card.
- Dual 10Gb Ethernet card (Cobia).

2462 Model 4BX (DataPower XI50z)

The 2462 Model 4BX is designed to work together with the IBM 2458 Model 002 (zBX). It is functionally equivalent to an IBM 4195-4BX with similar feature codes. The IBM 2462 Model 4BX is ordered through certain feature codes for the 2458-002.

When you configure the IBM 2458 Model 002 with Feature Code 0611 (DataPower XI50z), you order a machine type IBM 2462 Model 4BX for each configured feature code. It requires software PID 5765-G84.

The Software Maintenance Agreement (SWMA) must be active for the IBM software that runs on the DataPower XI50z in order to obtain service or other support for the IBM software. Failure to maintain the SWMA will result in the client not being able to obtain service for the IBM software, even if the DataPower XI50z is under warranty or a post-warranty IBM hardware maintenance service contract.

DataPower XI50z has these license entitlements:

- DataPower Basic Enablement (Feature Code 0650)
- Tivoli® Access Manager (Feature Code 0651)
- TIBCO (Feature Code 0652)
- Database Connectivity (DTB) (Feature Code 0653)
- Application Optimization (AO) (Feature Code 0654)
- Month Indicator (Feature Code 0660)
- Day Indicator (Feature Code 0661)
- Hour Indicator (Feature Code 0662)
- Minute Indicator (Feature Code 0663)

5765-G84 IBM WebSphere DataPower Integration Blade XI50B has these feature codes and descriptions:

- FC 0001: License with one year SWMA
- FC 0002: Option for TIBCO
- FC 0003: Option for Application Optimization
- FC 0004: Option for Database Connectivity
- FC 0005: Option for Tivoli Access Manager

Every IBM 2462 Model 4BX includes feature codes 0001, 0003, and 0005 (they are optional on DataPower XI50B). Optional Software feature codes 0002 and 0004 are required if you order FC 0652 TIBCO or FC 0653 Database Connectivity.

The TIBCO option (FC 0002) lets you extend the DataPower XI50z so that you can send and receive messages from TIBCO Enterprise Message Service (EMS).

The Option for Database Connectivity (FC 0004) lets you extend the DataPower XI50z to read data from and write data to relational databases, such as IBM DB2, Oracle, Sybase, and Microsoft SQL Server.

For software PID number 5765-G85 (registration and renewal), every IBM 2462 Model 4BX includes Feature Code 0001. Feature code 0003 is available at the end of the first year to renew the software maintenance for one more year.

For software PID number 5765-G86 (maintenance reinstatement - 12 months), Feature Code 0001 is available if software PID 5765-G85 Feature Code 0003 was not ordered before the first year expired.

For software PID number 5765-G87 (three-year registration), Feature Code 0001 can be ordered instead of software PID 5765-G85 Feature Code 0003 to make the initial period three years, instead of one year.

For software PID number 5765-G88 (three-year renewal), Feature Code 0001 can be used as an alternative to software PID 5765-G85 Feature Code 0003 if a three-year renewal is desired. The maximum duration is five years.

For software PID number 5765-G89 (three years after license), Feature Code 0001 is available if software PID 5765-G85 Feature Code 0003 was not ordered before the first year expired if a three-year renewal is desired.

IBM BladeCenter HX5 blades

The IBM BladeCenter HX5 is a scalable blade server that is designed to provide new levels of utilization, performance, and reliability for compute-intensive and memory-intensive workloads, such as database, virtualization, business intelligence, modeling and simulation, and other enterprise applications.

Select System x blades running Linux on System x are supported in the zBX, utilizing the zBX integrated hypervisor for IBM System x blades (using a kernel-based virtual machine) and providing logical device integration between the System z and System x blades for multi-tiered applications. System x blades are licensed separately and are enabled and managed as part of the ensemble by Unified Resource Manager.

The support of select IBM System x blades in the zBX allows the zEnterprise to access a whole new application portfolio. Front-end applications that need access to centralized data serving are a good fit for running on the blades, as well as applications that are a front end to core CICS or IMS transaction processing, such as IBM WebSphere. BladeCenter HX5 blades are client-acquired through existing channels or through IBM. POWER7, DataPower XI50z, and System x blades can be in the same BladeCenter chassis. Table 7-2 on page 191 lists the supported configuration options.

IBM BladeCenter HX5 7873 is a dual-socket 16-core blade with following features:

- Intel 8 core processor
- Two processor sockets
- 2.13 GHz 105W
- Maximum of 14 A16Ms per BC-H
- Memory: Up to 16 DIMM DDR-3 with 6.4 GTs
- 100 GB SSD internal disk

Table 7-2 Supported configurations of System x blades

System x blades	Part number	Feature code	Config 0	Config 1	Config 2	Config 3
Blades base - HX5	MT 7873	A16M	1	1	1	1
Processor 2.13 GHz 105W	69Y3071 69Y3072	A16S A179	1 1	1 1	1 1	1 1
Intel processors			2	2	2	2
Blade width			Single	Single	Single	Single
Total cores			16	16	16	16
Memory kits: 8 GB 1333 MHz 16 GB 1333 MHz	 46C0558 49Y1527	 A17Q 2422	 8 0	 16 0	 8 8	 0 16
GB/core			4	8	12	16
Speed Burst	46M6843	1741	1	1	1	1

System x blades	Part number	Feature code	Config 0	Config 1	Config 2	Config 3
SSD Exp Card 50 GB SSD MLC No Internal RAID	46M6906 43W7727	5765 5428 9012	1 2 1	1 2 1	1 2 1	1 2 1
CFFh 10GbE	46M6170	0099	1	1	1	1
CIOv 8 Gb FC	44X1946	1462	1	1	1	1

7.2.5 Power distribution units

The power distribution units (PDUs) provide connection to the main power source for intranode management network and intraensemble data network TOR switches, and the BladeCenter. The number of necessary power connections 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 controlling z114 server will have the entitlements feature for the configured blades. The entitlements are similar to a high-water mark or maximum-purchased flag, and only a blade quantity that is equal to or less than the quantity that is installed in the zBX can communicate with the CPC.

Also, Unified Resource Manager has two management suites: Manage suite (FC 0019) and Automate/Advance Management Firmware suite (FC 0020).

If the controlling z114 has Manage suite (FC 0019), the same quantity that is entered for any blade enablement feature code (FC 0611, FC 0612, or FC 0613) will be used for the Manage Firmware (FC 0040, FC 0041, or FC 0042) of the corresponding blades.

If the controlling z114 has Automate/Advance Management Firmware suite (FC 0020), the same quantity that is entered for the Blade Enablement feature codes (FC 0611, FC 0612, or FC 0613) will be used for the Manage Firmware (FC 0040, FC 0041, or FC 0042) and Automate Firmware (FC 0044, FC 0045, or FC 0046) of the corresponding blades.

Table 7-3 lists these features. Table 7-4 shows the minimum and maximum quantities for the feature codes that are listed in Table 7-3.

Table 7-3 Feature codes for blade enablement and Unified Resource Manager suites

	Blade Enablement	Manage FC per connection	Advanced Management FC per connection	Automate FC per connection
z/OS only	N/A	FC 0019	N/A	FC 0020
IFL	N/A	N/C	N/A	FC 0052
DataPower XI50z	FC 0611	FC 0040	N/A	FC 0044
POWER7 Blade	FC 0612	FC 0041	N/A	FC 0045
IBM System x HX5 Blade	FC 0613	FC 0042	FC 0046	N/A

Blades: If any attempt is made to install additional blades that exceed the FC 0611, FC 0612, or FC 0613 count, those blades will be not be powered on by the system. The blades will also be checked for minimum hardware requirements.

Table 7-4 Minimum and maximum quantities for Unified Resource Manager feature codes

Feature code	Minimum quantity	Maximum quantity
FC 0040	1	28
FC 0044	1	28
FC 0041	1	112
FC 0045	1	112
FC 0042	1	28
FC 0046	1	28

Note that FC 0040, FC 0041, FC 0042, FC 0044, FC 0045, FC 0046, and FC 0052 are priced features. To get ensemble member management and cables, you also need to order FC 0025 on the z114.

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 z114 HMC/SE:

- View firmware information for the BladeCenter and blades
- Retrieve the firmware changes
- Change the firmware level
- Back up and restore critical data:
  - zBX configuration data is backed up as part of the System z114 SE backup.
  - It is restored on the replacement of a blade.

For more details, see *IBM zEnterprise Unified Resource Manager,* SG24-7921.

7.3.2 zBX firmware

The firmware for the zBX is managed, controlled, and delivered in the same way as the firmware for the z114 server. It is packaged and tested with System z microcode, and changes are supplied and applied with Machine Change Level (MCL) bundle releases.

Here, we summarize the benefits of the zBX firmware that is packaged with the System z microcode:

- Tested together with the System z driver code and MCL bundle releases.
- Retrieval of code is the same integrated process for System z (IBM RETAIN® or media).
- No need to use separate tools and connect to websites to obtain code.

- New upcoming System z firmware features, such as Digitally Signed Firmware, are used.
- 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 application failures.
- Optimizer firmware.

7.4 zBX connectivity

There are three types of LANs (each with redundant connections) that attach to the zBX: the intranode management network (INMN), the intraensemble data network (IEDN), and the customer managed data network. The INMN is fully isolated and only established between the controlling z114 and the zBX. The IEDN connects the zBX to a maximum of eight z114s.

Each z114 must have a minimum of two connections to the zBX. The IEDN is also 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.

Example 7-5 (high-level diagram) shows the connectivity that is required for the zBX environment. The z114 connects through two OSA-Express3 1000BASE-T features (CHPID type OSM) to the INMN TOR switches. The OSA-Express4S 10 GbE or OSA-Express3 10 GbE features (CHPID type OSX) connect to the two IEDN TOR switches. Depending on the requirements, any OSA-Express4S, OSA-Express3, or OSA-Express2[1] features (CHPID type OSD) can connect to the customer managed data network.

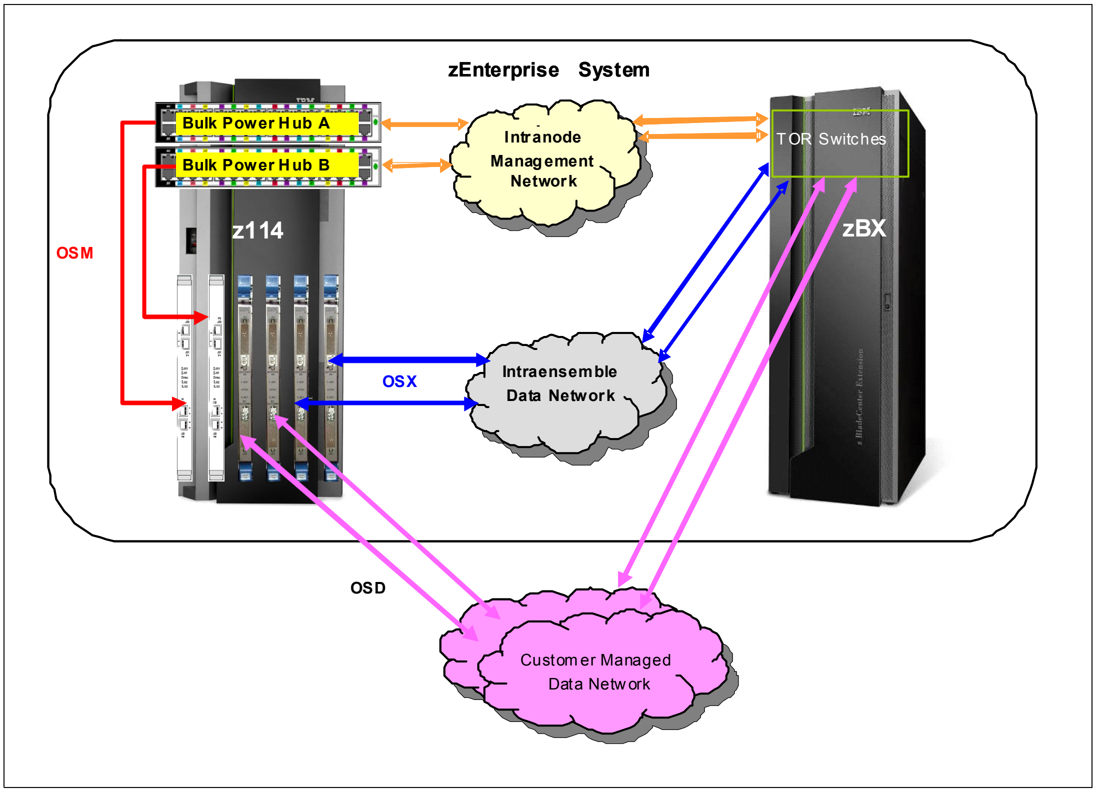

zEnterprise System
Bulk Power Hub A
Bulk Power Hub B
Intranode Management Network
TOR Switches
z114
zBX
OSM
OSX
Intraensemble Data Network
OSD
Customer Managed Data Network

Figure 7-5 INMN, IEDN, and customer managed local area network

[1] Carry forward only for zEnterprise CPCs when upgrading from earlier generations.

The IEDN provides private and secure 10 GbE high-speed data paths between all elements of a zEnterprise ensemble (up to eight zEnterprise CPCs with optional zBXs).

The zBX is managed by the HMC through the physically isolated INMN, which interconnects all resources of the zEnterprise System (zEnterprise CPC 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 z114 and its optional zBX. INMNs in separate nodes do not connect to each other. The INMN connects the SE of the z114 to the hypervisor, optimizer, and guest management agents within the node.

INMN communication

Communication across the INMN is exclusively for the purpose of enabling the Unified Resource Manager of the HMC to perform its various management disciplines (virtual server, performance, network virtualization, storage, energy management, and so on) for the node. The z114 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

Consider these key points for an INMN:

- Each z114 must have two OSA-Express3 1000BASE-T ports connected to the Bulk Power Hub in the same z114:
  - The two ports provide a redundant configuration for failover purposes in case one link fails.
  - For availability, each connection must be from two separate OSA-Express3 1000BASE-T features within the same z114 server.

  Figure 7-6 shows the OSA-Express3 1000BASE-T feature and the required cable type.

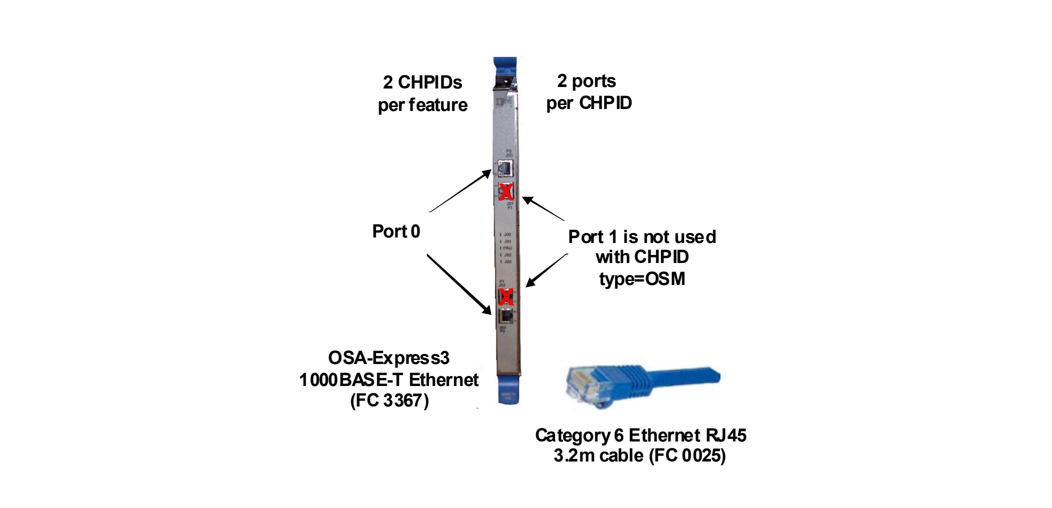


Figure 7-6 OSA-Express3 1000BASE-T feature and cable type

- OSA-Express3 1000BASE-T ports can be defined in the input/output configuration data set (IOCDS) as SPANNED, SHARED, or DEDICATED:
  - DEDICATED restricts the OSA-Express3 1000BASE-T port to a single logical partition (LPAR).
  - SHARED allows the OSA-Express3 1000BASE-T port to be used by all or selected LPARs in the same z114.
  - SPANNED allows the OSA-Express3 1000BASE-T port to be used by all or selected LPARs across multiple channel subsystems in the same z114.
  - SPANNED and SHARED ports can be restricted by the PARTITION keyword in the CHPID statement to allow only a subset of LPARs in the z114 to use the OSA-Express3 1000BASE-T port.
  - SPANNED, SHARED, and DEDICATED link pairs can be defined within the maximum of 16 links that is supported by the zBX.
- z/OS Communication server TCP/IP stack must be enabled for IPv6. The CHPID type OSM-related definitions will be created dynamically.

  No IPv4 address is needed. An IPv6 link local address will be applied dynamically.
- z/VM virtual switch types provide INMN access:
  - The uplink can be a virtual machine network interface card (NIC).
  - Ensemble membership conveys the Universally Unique IDentifier (UUID) and the Media Access Control (MAC) prefix.
- Two 1000BASE-T TOR switches in the zBX (Rack B) are used for the INMN; no additional 1000BASE-T Ethernet switches are required. Figure 7-7 shows the 1000BASE-T TOR switches.

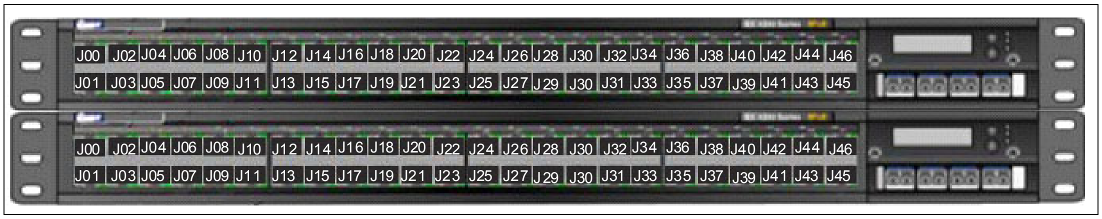

J00 J02 J04 J06 J08 J10 J12 J14 J16 J18 J20 J22 J24 J26 J28 J30 J32 J34 J36 J38 J40 J42 J44 J46
J01 J03 J05 J07 J09 J11 J13 J15 J17 J19 J21 J23 J25 J27 J29 J30 J31 J33 J35 J37 J39 J41 J43 J45
J00 J02 J04 J06 J08 J10 J12 J14 J16 J18 J20 J22 J24 J26 J28 J30 J32 J34 J36 J38 J40 J42 J44 J46
J01 J03 J05 J07 J09 J11 J13 J15 J17 J19 J21 J23 J25 J27 J29 J30 J31 J33 J35 J37 J39 J41 J43 J45

Figure 7-7 Two 1000BASE-T TOR switches

Table 7-5 listed the port assignments for both 1000BASE-T TOR switches.

Table 7-5 Port assignments for the 1000BASE-T TOR switches

Ports	Description
J00 - J03	Management for the BladeCenters that are located in zBX Rack-B
J04 - J07	Management for the BladeCenters that are located in zBX Rack-C
J08 - J11	Management for the BladeCenters that are located in zBX Rack-D
J12 - J15	Management for the BladeCenters that are 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 z114 BPH-A port J06/INMN-B to z114 BPH-B port J06

- 1000BASE-T supported cable:
  - The 3.2 m (10.59 ft) Category 6 Ethernet cables are shipped with the z114 ensemble management flag feature (FC 0025). These cables connect the OSA-Express3 1000BASE-T ports to the Bulk Power Hubs (port 7).
  - The 26 m (85.3 ft) Category 5 Ethernet cables ship with the zBX. These cables are used to connect the z114 Bulk Power Hubs (port 6) and the zBX TOR 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. The alternate HMC 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. For more information about the HMC network configuration, see Chapter 12, “Hardware Management Console” on page 345.

Figure 7-8 shows the primary and alternate HMC configuration connecting into the two bulk power hubs (BPHs) in the z114 via a customer-managed management network. The 1000BASE-T TOR switches in the zBX are also connected to the BPHs in the z114.

Ports: All ports on the z114 BPH are reserved for specific connections. Any deviations or incorrect cabling will affect the operation of the z114 system.

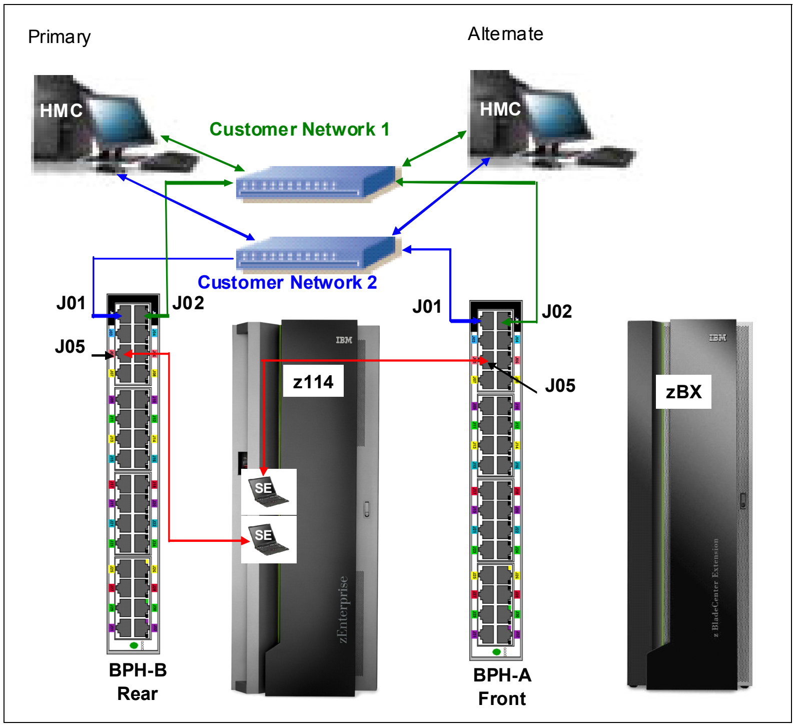


Figure 7-8 HMC configuration in an ensemble node

Table 7-6 on page 198 shows the port assignments for both bulk power hubs (BPHs).

Table 7-6 Port assignments for the BPHs

BPH A		BPH B	
Port number	**Connects to**	**Port number**	**Connects to**
J01	HMC to SE Customer Network2 (VLAN 0.40)	J01	HMC to SE Customer Network2 (VLAN 0.40)
J02	HMC to SE Customer Network1 (VLAN 0.30)	J02	HMC to SE Customer Network1 (VLAN 0.30)
J03	BPH B J03	J03	BPH A J03
J04	BPH B J04	J04	BPH A J04
J05	SE A-Side (Top SE)	J05	SE B-Side (Bottom SE)
J06	zBX TOR Switch B36P, Port 47 (INMN-A)	J06	zBX TOR Switch B35P, Port 47 (INMA-B)
J07	OSA-Express3 1000BASE-T (CHPID type OSM)	J07	OSA-Express3 1000BASE-T (CHPID type OSM)
J08	Not used	J08	Not used
J09 - J32	Used for internal z114 components	J09 - J32	Used for internal z114 components

For more information, see Chapter 12, “Hardware Management Console” on page 345.

7.4.3 Intraensemble data network

The intraensemble data network (IEDN) is the major application data path that is provisioned and managed by the Unified Resource Manager of the controlling z114. The data communications for ensemble-defined workloads flow over the IEDN between the 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 z114 through the OSA-Express4S 10 GbE or OSA-Express3 10 GbE ports when defined as CHPID type OSX. The minimum number of OSA10 GbE features is two per z114. Similarly, a 10 GbE networking infrastructure within the zBX is used for IEDN access.

Terminology: If not specifically stated otherwise, the term OSA10 GbE applies to the OSA-Express4S 10 GbE and OSA-Express3 10GbE features throughout 7.4.3, “Intraensemble data network” on page 198.

IEDN configuration

You can configure the IEDN connections in a number of ways. Consider these key points for an IEDN:

- Each z114 must have a minimum of two OSA 10 GbE ports that are connected to the zBX through the IEDN:
  - The two ports provide a redundant configuration for failover purposes in case one link fails.
  - For availability, each connection must be from two separate OSA 10 GbE features within the same z114.

  - The zBX can have a maximum of 16 IEDN connections (eight pairs of OSA 10 GbE ports).

Figure 7-9 shows the OSA 10 GbE feature (long reach or short reach) and the required fiber optic cable types.

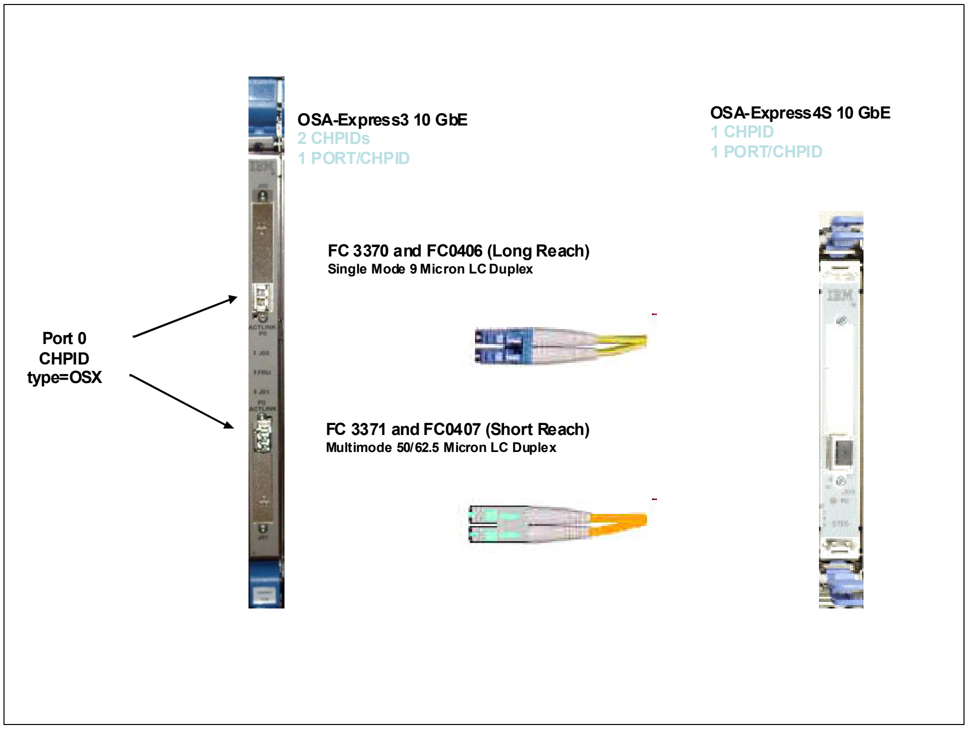

OSA-Express3 10 GbE
2 CHPIDs
1 PORT/CHPID
OSA-Express4S 10 GbE
1 CHPID
1 PORT/CHPID
FC 3370 and FC0406 (Long Reach)
Single Mode 9 Micron LC Duplex
Port 0
CHPID
type=OSX
FC 3371 and FC0407 (Short Reach)
Multimode 50/62.5 Micron LC Duplex

Figure 7-9 OSA-Express4S 10 GbE and OSA-Express3 10 GbE features and cables

- OSA 10 GbE ports can be defined in the IOCDS as SPANNED, SHARED, or DEDICATED:
  - DEDICATED restricts the OSA 10 GbE port to a single LPAR.
  - SHARED allows the OSA 10 GbE port to be used by all or selected LPARs in the same z114.
  - SPANNED allows the OSA 10 GbE port to be used by all or selected LPARs across multiple channel subsystems (CSSs) in the same z114.
  - SHARED and SPANNED ports can be restricted by the PARTITION keyword in the CHPID statement to allow only a subset of LPARs on the z114 to use the OSA 10 GbE port.
  - SPANNED, SHARED, and DEDICATED link pairs can be defined within the maximum of 16 links that is supported by the zBX.
- z/OS Communication Server requires minimal configuration:
  - IPv4 or IPv6 addresses are used.
  - VLAN must be configured to match the HMC (Unified Resource Manager) configuration.
- z/VM virtual switch types provide IEDN access:
  - The uplink can be a virtual machine NIC.
  - Ensemble membership conveys the ensemble UUID and MAC prefix.
- The IEDN network definitions are completed from the primary HMC “Manage Virtual Network” task.

- Two 10 GbE TOR switches in the zBX (Rack B) are used for the IEDN. No additional Ethernet switches are required. Figure 7-10 shows the 10 GbE TOR switches.

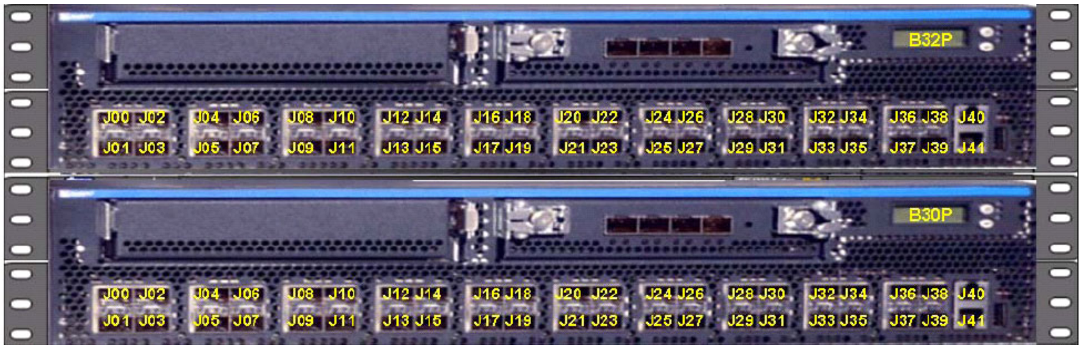

B32P
J00 J02 J04 J06 J08 J10 J12 J14 J16 J18 J20 J22 J24 J26 J28 J30 J32 J34 J36 J38 J40
J01 J03 J05 J07 J09 J11 J13 J15 J17 J19 J21 J23 J25 J27 J29 J31 J33 J35 J37 J39 J41
B30P
J00 J02 J04 J06 J08 J10 J12 J14 J16 J18 J20 J22 J24 J26 J28 J30 J32 J34 J36 J38 J40
J01 J03 J05 J07 J09 J11 J13 J15 J17 J19 J21 J23 J25 J27 J29 J31 J33 J35 J37 J39 J41

Figure 7-10 Two 10 GbE TOR switches

Statement of Direction: HiperSockets integration with the IEDN: Within a zEnterprise environment, it is planned for HiperSockets to be integrated with the intraensemble data network (IEDN), extending the reach of the HiperSockets network outside of the central processor complex (CPC) to the entire ensemble, appearing as a single Layer 2 network. HiperSockets integration with the IEDN is planned to be supported in z/OS V1.13 and z/VM in a future deliverable.

Port assignments

Table 7-7 lists the port assignments for both 10 GbE TOR switches.

Table 7-7 Port assignments for the 10 GbE TOR switches

Ports	Description
J00 - J07	Small Form Factor Pluggable (SFP) and reserved for z114 (OSX) IEDN connections
J08 - J21[1]	Direct-attached cables (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 - J37	SFP reserved for client IEDN connections
J38, J39	Reserved for future use
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 currently used.

The following considerations apply:

- All IEDN connections must be point-to-point to the 10 GbE switch:
  - The IEDN connection uses the MAC address, not the IP address (Layer 2 connection).
  - No additional switches or routers are needed.
  - The point-to-point connections limit the distance that the CPCs can be from the 10 GbE switches in an ensemble.

- The 10 GbE TOR switches use small form-factor pluggable (SFP) optics for the external connections and direct attach cables (DAC) for connections.

  Ports J00 - J07 are reserved for the z114 OSX IEDN connections. These ports use SFPs that are plugged according to the zBX order:

  - FC 0632 LR SFP to FC 0406 OSA-Express4S 10 GbE LR
  - FC 0633 SR SFP to FC 0407 OSA-Express4S 10 GbE SR
  - FC 0632 LR SFP to FC 3370 OSA-Express3 10 GbE LR
  - FC 0633 SR SFP to FC 3371 OSA-Express3 10 GbE SR

  Ports J08 - J23 are reserved for IEDN to BladeCenter attachment. The cables that are used are direct-attached cables (DAC) and are included with the zBX. These cables are hard-wired 10 GbE SFP cables. The feature codes indicate the length of the cable:

  - (FC 0626): One meter (3.33 ft) for rack B BladeCenters and IEDN to IEDN
  - (FC 0627): Five m (16.4 ft) for rack C BladeCenter
  - (FC 0628): Seven m (22.11 ft) for racks D and E BladeCenters

- The 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 m (984.3 ft)
    - 50 micron fiber at 500 MHz-km: 82 m (269 ft)
    - 62.5 micron fiber at 200 MHz-km: 33 m (108.3 ft)
  - Single-mode fiber:
    - 10 km (6.2 miles)

7.4.4 Network connectivity rules with zBX

Follow these network connectivity rules for interconnecting a zBX:

- Only one zBX is allowed per controlling z114.
- The zBX can be installed next to the controlling z114 or within the limitation of the 26 m (85.3 ft) 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).
- Customer-managed data networks are outside the ensemble. A customer-managed data network connects with these components:

  - CHPID type OSD from the z114
  - IEDN TOR switch ports J31 - J37 from the zBX

7.4.5 Network security considerations with zBX

The private networks that are involved in connecting the z114 to the zBX are constructed with extreme security in mind, for example:

- The INMN is entirely private and can be accessed only by the SE (standard HMC security 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. Extremely strict authorizations for users and programs control who is allowed to take advantage of the INMN.
- The INMN network uses “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, because it is a “flat network” with all virtual servers residing on the same IPv6 network.

The Unified Resource Manager communicates with the virtual servers through the SE over the INMN. The virtual servers cannot communicate with each other directly through INMN; they can only communicate with the SE.

- Only authorized programs or agents can take advantage of the INMN; currently, the Performance Agent can take advantage of the INMN. However, there can be other platform management applications in the future. These applications must be authorized to access the INMN.
- The IEDN is built on a flat network design (same IPv4 or IPv6 network). Each server accessing the IEDN must be an authorized virtual server and must belong to an authorized VLAN within the physical IEDN. VLAN enforcement resides 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 (VMACs) that are assigned to the connections from the virtual servers are tightly controlled through the Unified Resource Manager. Thus, there is no chance of either MAC or VLAN spoofing for any of the servers on the IEDN. If you decide to attach your network to the TOR switches of the zBX in order to communicate with the virtual servers on the zBX blades, access must be authorized in the TOR switches (MAC-based or VLAN-based).

  Although the TOR switches will enforce the VMACs and VLAN IDs here, you must take the usual network security measures to ensure that the 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. Therefore, whenever you decide to interconnect the external network to the secured IEDN, the security of that external network must 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 z/VM. Only certain virtual servers on the z114 can utilize these networks. INMN is not accessible at all from within the virtual servers.
- Although we think it is unnecessary to implement firewalls, IP filtering, or encryption for data flowing over the IEDN, if your company security policy mandates these measures, they are supported. You can implement any of the available security technologies, for example, Secure Sockets Layer (SSL)/Transport Layer Security (TLS), or IP filtering.
- The centralized and internal network design of both the INMN and the IEDN limit vulnerability to security breaches. Both networks reduce the amount of network equipment and administration tasks, as well as routing hops that are under the control of multiple individuals and subject to security threats. Both use IBM-only equipment (switches and blades) that have been tested previously and, in certain cases, pre-installed.

In summary, many more technologies than in the past have been architected in a more robust, secure fashion to integrate into the client network. These more secure technologies have been achieved with the help of either the Unified Resource Manager, or additional System Authorization Facility (SAF) controls that are specific to the zEnterprise System and the ensemble:

- 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 Multi-Level 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

The Fibre Channel (FC) connections can be established between the zBX and a SAN environment. Client-supplied FC switches are required and must support N_Port ID Virtualization (NPIV). Certain FC switch vendors also require “interop” mode. Check the interoperability matrix for the latest details:

http://www-03.ibm.com/systems/support/storage/ssic/interoperability.wss

Cables: 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 SAN environment.

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 7-11 shows an image of the external ports.

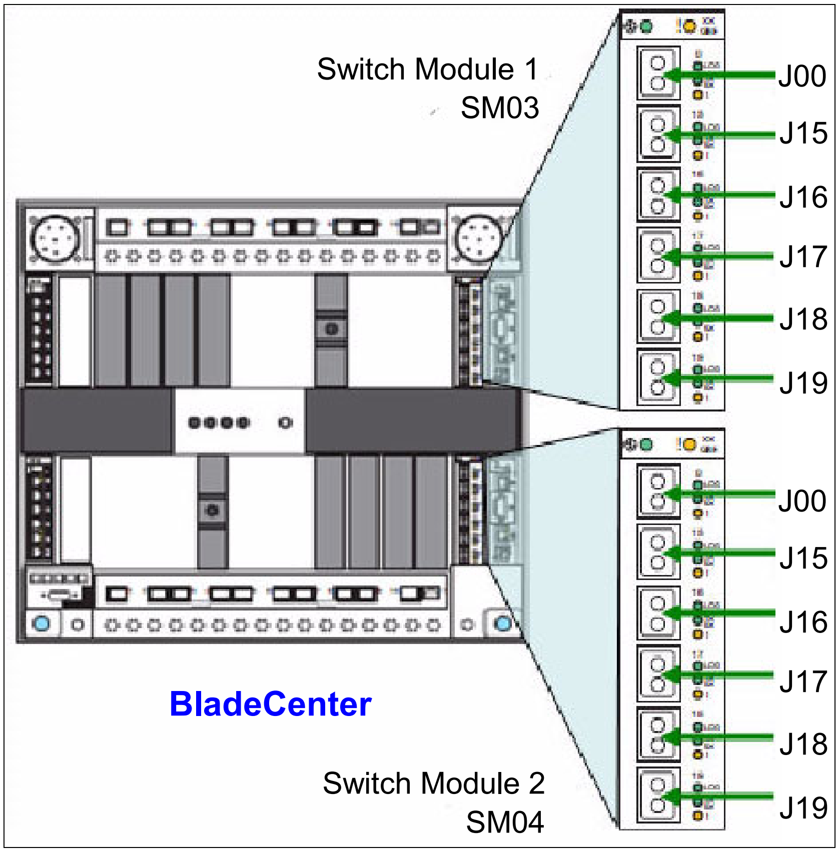


Figure 7-11 8 Gb FC switch external ports

Client-provided multimode LC duplex cables are used for FC disk connections to support speeds of 8 Gbps, 4 Gbps, or 2 Gbps. (A speed of 1 Gbps is not supported.) The maximum distance depends on the speed and fiber type.

Cabling specifications are defined by the Fibre Channel - Physical Interface - 4 (FC-PI-4) standard. Table 7-8 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 that is defined by the FC-PI-4 standard (Revision 8.00).

Table 7-8 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	Link loss budget (dB)	Distance in meters	Link loss budget (dB)	Distance in meters	Link loss budget (dB)
50 µm MM[1] (SX laser)	500 (1,641 ft)	3.31	380 (1,247 ft)	2.88	150 (492 ft)	2.04
50 µm MM[2] (SX laser)	300 (984 ft)	2.62	150 (492 ft)	2.06	50 (164 ft)	1.68
62.5 µm MM[3] (SX laser)	150 (492 ft)	2.1	70 (229.7 ft)	1.78	21 (68.10 ft)	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 MHz-km 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 MHz-km at 850 nm and 500 MHz-km at 1300 nm in accordance with IEC 60793-2-10 Type A1b fiber

Cabling: 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 in both FC switch modules of the BladeCenter chassis and must connect through an FC switch to FC disk storage (see Figure 7-12).

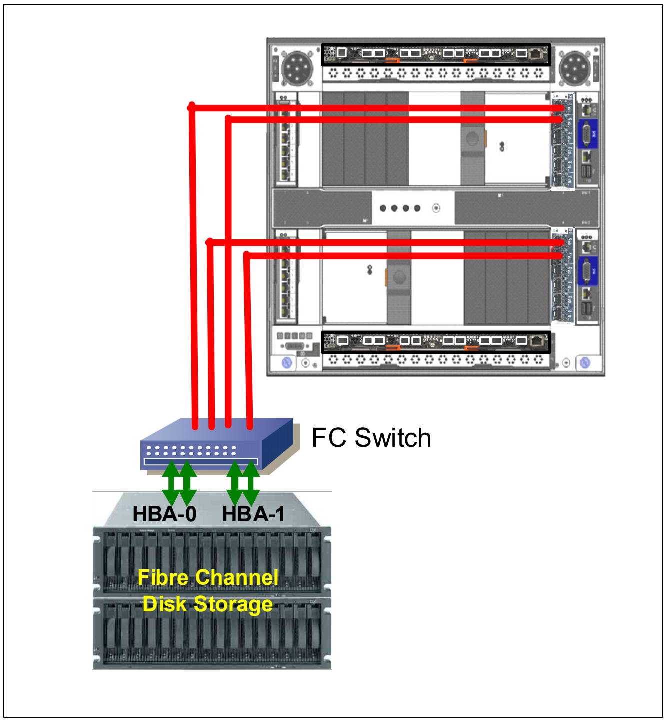


Figure 7-12 BladeCenter chassis storage connectivity

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.

Supported FC disk storage

Supported FC disk types and vendors with IBM blades are listed on the IBM System Storage Interoperation Center (SSIC) website:

http://www-03.ibm.com/systems/support/storage/config/ssic/displayesssearchwithoutjs.wss?start_over=yes

7.5 zBX connectivity examples

This section illustrates various ensemble configuration examples containing a zBX and the necessary connectivity for operation. For simplicity, we do not show the redundant connections in the configuration examples.

Subsequent configuration diagrams build on the previous configuration and only additional connections will be noted.

7.5.1 A single node ensemble with a zBX

Figure 7-13 shows a single node ensemble with a zBX. The necessary components include the controlling z114 CPC1 and the attached zBX, switches, and FC disk storage.

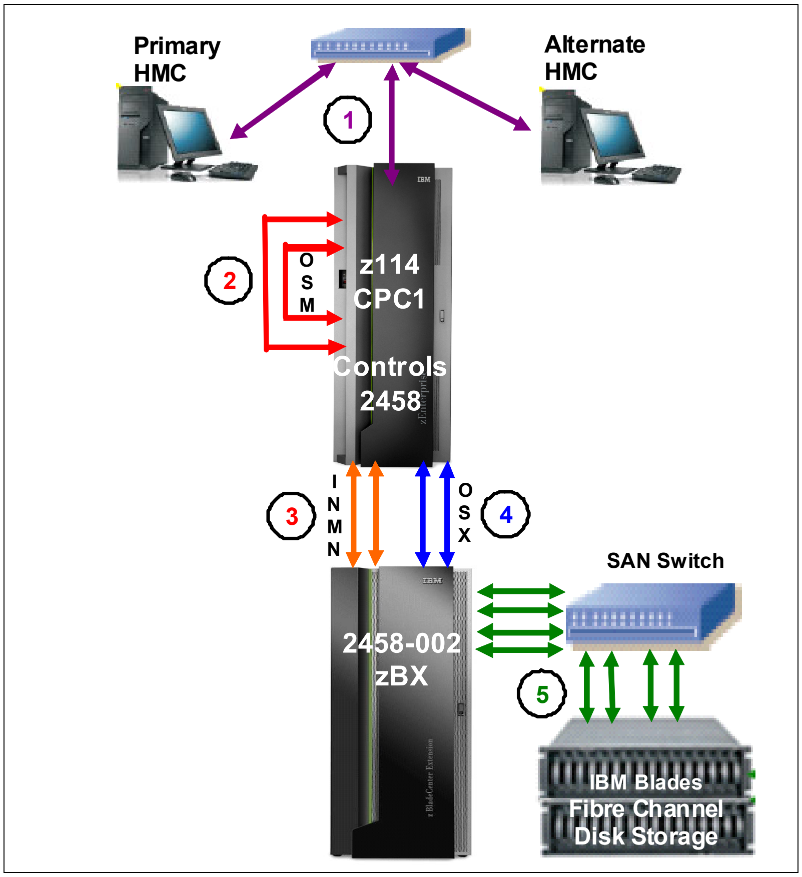


Figure 7-13 Single node ensemble with zBX

The diagram shows the following components:

1. Client-provided management network:
   - IBM supplies a 15 m (49.2 ft) Ethernet RJ-45 cable with the 1000BASE-T (1GbE) switch (FC 0070).
   - The 1000BASE-T switch (FC 0070) connects to the reserved client network ports of the Bulk Power Hubs in z196 - Z29BPS11 (on the A side) and Z29BPS31 (on the B side) - port J02. A second switch connects to Z29BPS11 and Z29BPS31 on port J01.
2. Intranode management network:
   - Two CHPIDs from two separate OSA-Express3 1000BASE-T features are configured as CHPID type OSM.
   - IBM supplies two 3.2 m (10.5 ft) Ethernet Category 6 cables from the OSM CHPIDs (ports) to both Z29BPS11 and Z29BPS31 on port J07. (This connection is a z196 internal connection that is supplied with FC 0025.)
3. Intranode management network - extension:
   - IBM supplies two 26 m (85.3 ft) Category 5 Ethernet cables (chrome gray plenum rated cables) from zBX Rack B INMN-A/B switches port J47 to Z29BPS11 and Z29BPS31 on port J06.

4. Intraensemble data network:
   - Two ports from two separate OSA-Express4S 10 GbE or OSA-Express3 10 GbE (Short Reach (SR) or Long Reach (LR)) features are configured as CHPID type OSX.
   - The client supplies the fiber optic cables (single mode or multimode).
5. 8 Gbps Fibre Channel switch:
   - The client supplies all Fibre Channel cables (multimode) from the zBX to the attached FC switch.
   - The client is responsible for the configuration and management of the FC switch.

7.5.2 A dual-node ensemble with a single zBX

A second z114 CPC2 (node) is introduced in Figure 7-14, showing the additional hardware. Up to eight additional nodes (z114 servers) can be added in the same fashion.

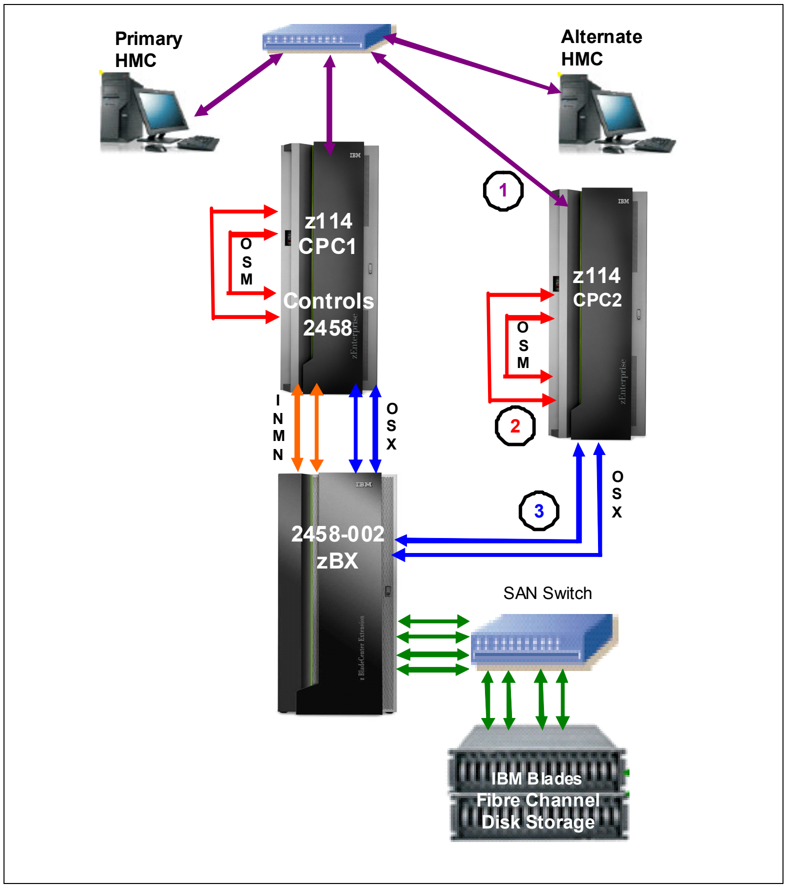


Figure 7-14 Dual-node ensemble with a single zBX

The diagram shows the following components:

1. Client provided management network:
   - The client supplies an Ethernet RJ-45 cable.
   - The 1000BASE-T switch (FC 0070) connects to the reserved client network ports of Z29BPS11 and Z29BPS31 - J02. A second switch connects to Z29BPS11 and Z29BPS31 on port J01.

2. Intranode management network:
   - Two ports from two separate OSA-Express3 1000BASE-T features are configured as CHPID type OSM.
   - IBM supplies two 3.2 m (10.5 ft) Ethernet Category 6 cables from the OSM CHPIDs (ports) to both Z29BPS11 and Z29BPS31 on port J07. (This connection is a z196 internal connection that is supplied with FC 0025.)
3. Intraensemble data network:
   - Two ports from two separate OSA-Express4S 10 GbE or OSA-Express3 10 GbE (Short Reach (SR) or Long Reach (LR)) features are configured as CHPID type OSX.
   - The 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 that has been added to the original configuration. The two zBXs are interconnected through fiber optic cables to SFPs in the IEDN switches for isolated communication (SR or LR) over the IEDN network.

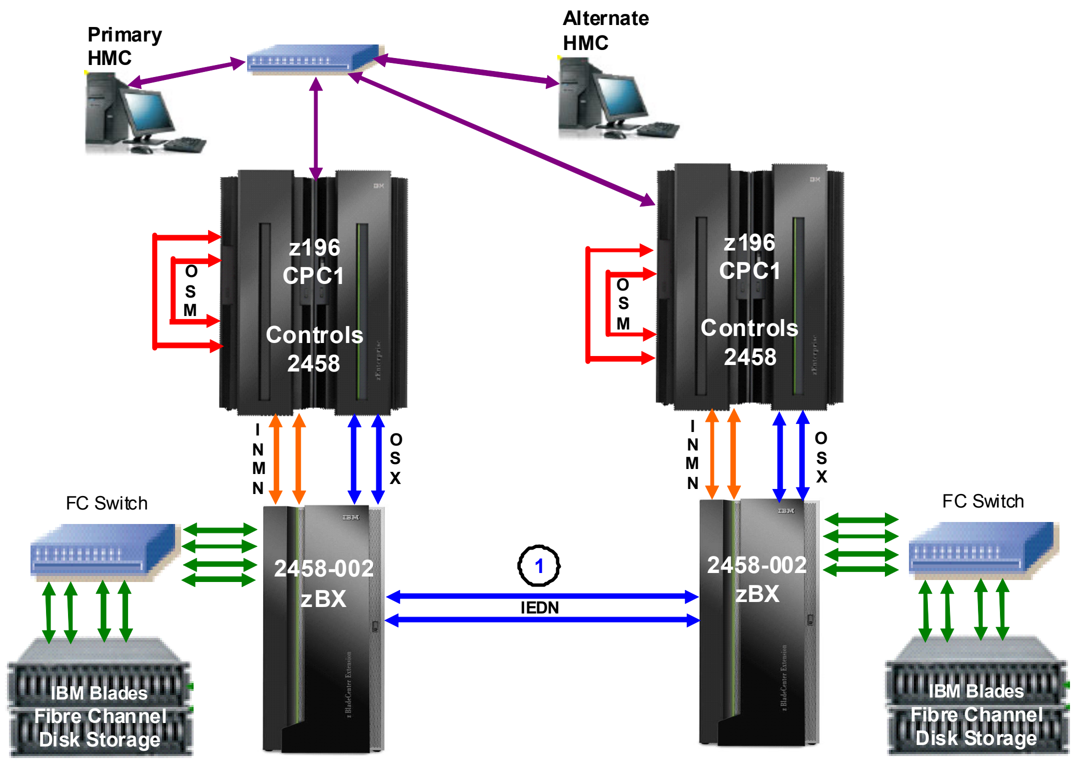


Figure 7-15 Dual-node ensemble

The diagram shows the following components:

1. Intraensemble data network:

   Two 10 GbE ports in the TORs are used to connect the two zBXs (10 GbE TOR switch to 10 GbE TOR switch).

Up to eight z196s CPCs can be connected to a zBX using the IEDN. Additional z196 CPCs added and connected to the zBX through the OSA-Express4S 10 GbE or OSA-Express3 10 GbE (Short Reach (SR) or Long Reach (LR)) features are configured as CHPID type OSX.

Ensembles: z10 servers cannot participate in an ensemble; however, they can utilize the zBX environment (applications).

zEnterprise CPC and z10 servers that are not part of an ensemble can connect to a zBX Model 002 through OSA-Express4S, OSA-Express3, or OSA-Express2 features when defined as CHPID type OSD. The OSA ports can be connected either directly to the IEDN or through client-supplied Ethernet switches that are connected to the IEDN.

7.6 References

You can obtain installation details in *IBM zEnterprise BladeCenter Extension Model 002 Installation Manual for Physical Planning*, GC27-2611, and the *IBM zEnterprise BladeCenter Extension Model 002 Installation Manual*, GC27-2610.

For details about the BladeCenter components, see *IBM BladeCenter Products and Technology,* SG24-7523.

For further discussion about the benefits and usage of the IBM Smart Analytics Optimizer solution, see *Using IBM System z As the Foundation for Your Information Management Architecture*, REDP-4606.

Go to this website for more information about DataPower XI50z blades:

http://www-01.ibm.com/software/integration/datapower/xi50z

Additional documentation is available on IBM Resource Link:

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

8

Software support

This chapter lists the minimum operating system requirements and support considerations for the z114 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 2818DEVICE for the most current information. We also discuss the generic software support for zEnterprise BladeCenter Extension.

Support of IBM zEnterprise 114 functions depends 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 that are required on the z114. For zEnterprise BladeCenter Extension (zBX), see 8.11, “zEnterprise BladeCenter Extension software support” on page 268.

Note that operating system levels that are no longer in service are not covered in this publication. These older levels might provide support for certain features.

Table 8-1 z114 minimum operating system requirements

Operating systems	ESA/390 (31-bit mode)	z/Architecture (64-bit mode)	Notes
z/OS V1R8[a]	No	Yes	Service is required. See the following shaded Updates 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 214.	See Table 8-2 on page 214.	Novell SUSE SLES 10 Red Hat RHEL 5

[a] Regular service support for z/OS V1R8 ended in September 2009. However, by ordering the IBM Lifecycle Extension for z/OS V1.8 product, fee-based corrective service can be obtained for up to two years after the withdrawal of service (September 2011). Similarly, for z/OS V1R9 and V1R10, there are IBM Lifecycle Extension products, which provide fee-based corrective service up to September 2012 and September 2013.

[b] z/VM V5R4 provides compatibility support only. z/VM V6R1 provides both compatibility and exploitation items.

[c] z/VM supports both 31-bit and 64-bit mode guests.

Updates: Exploitation of certain features depends on a particular operating system. In all cases, PTFs might be required with the operating system level that is indicated. Check the z/OS, z/VM, z/VSE, and z/TPF subsets of the 2818DEVICE 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.

For Linux on System z distributions, consult the distributor’s support information.

8.2 Support by operating system

System z196 introduced several new functions and z114 employs the same technologies. In this section, we discuss the support of those functions by the current operating systems. Also, we included several of the functions that were introduced in previous System z servers and that have been carried forward or enhanced in the z114. The features and functions that were available on previous servers but that are no longer supported by z114 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 215. For z/VSE, z/TPF, and Linux on System z, see Table 8-4 on page 220.

The tabular format is intended to help you determine, by a quick scan, which functions are supported and the minimum operating system level that is required.

8.2.1 z/OS

z/OS Version 1 Release 10 is the earliest in-service release supporting the z114. After September 2011, a fee-based extension for defect support (for up to two years) can be obtained by ordering the IBM Lifecycle Extension for z/OS V1.10. Although service support for z/OS Version 1 Release 9 ended in September of 2010, a fee-based extension for defect support (for up to two years) can be obtained by ordering the IBM Lifecycle Extension for z/OS V1.9. Similarly, IBM Lifecycle Extension for z/OS V1.8 provides fee-based support for z/OS Version 1 Release 8 until September 2011. Extended support for z/OS Version 1 Release 7 ended on 30 September 2010. Also, note that z/OS.e is not supported on z114 and that z/OS.e Version 1 Release 8 was the last release of z/OS.e.

See Table 8-3 on page 215 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 215 for a list of the supported functions and their minimum required support levels.

Important: We recommend that the capacity of any z/VM logical partitions, and any z/VM guests, in terms of the number of Integrated Facility for Linux (IFL) processors and logical central processors (CPs), real or virtual, be adjusted to accommodate the physical unit (PU) capacity of the z114.

8.2.3 z/VSE

Support is provided by z/VSE V4. Note these z/VSE characteristics:

- Executes in z/Architecture mode only.
- Exploits 64-bit real memory addressing.
- Support for 64-bit virtual addressing will be provided by z/VSE V5R1, when available.
- z/VSE V5R1 requires an architectural level set that is specific to the IBM System z9.

See Table 8-4 on page 220 for a list of the supported functions and their minimum required support levels.

8.2.4 z/TPF

See Table 8-4 on page 220 for a list of the supported functions and their minimum required support levels.

8.2.5 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 z114 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
Red Hat RHEL 6	Yes

IBM is working with its Linux distribution partners to provide further exploitation of selected z114 functions in future Linux on System z distribution releases.

We recommend these steps:

- Use Novell SUSE SLES 11 or Red Hat RHEL 6 in any new projects for the z114.
- Update any Linux distributions to their latest service level before the migration to the z114.
- Adjust the capacity of any z/VM and Linux on System z logical partition guests, as well as z/VM guests, in terms of the number of IFLs and CPs, real or virtual, according to the PU capacity of the z114.

8.2.6 z114 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 2818DEVICE.

The following two tables summarize the z114 functions and their minimum required operating system support levels:

- Table 8-3 on page 215 is for z/OS and z/VM.
- Table 8-4 on page 220 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.
- N/A

  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 z114 functions minimum support requirements summary, part 1

Function	z/OS V1 R13	z/OS V1 R12	z/OS V1 R11	z/OS V1 R10	z/OS V1 R9	z/OS V1 R8	z/VM V6 R1	z/VM V5 R4
z114	Y	Y[j]	Y[j]	Y[j]	Y[j]	Y[j]	Y[j]	Y[j]
Support of Unified Resource Manager	Y	Y[j]	Y[j]	Y[j]	Y[a j]	Y[a j]	Y	N
System z Integrated Information Processors (zIIPs)	Y	Y	Y	Y	Y	Y	Y[b]	Y[b]
System z Application Assist Processors (zAAPs)	Y	Y	Y	Y	Y	Y	Y[b]	Y[b]
zAAP on zIIP	Y	Y	Y	Y[j]	Y[j]	N	Y[c]	Y[c]
Large memory (> 128 GB)	Y	Y	Y	Y	Y	Y	Y[d]	Y[d]
Large page support	Y	Y	Y	Y	Y	N	N[e]	N[e]
Out-of-order execution	Y	Y	Y	Y	Y	Y	Y	Y
Guest support for execute-extensions facility	N/A	N/A	N/A	N/A	N/A	N/A	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	Y	N	N	N	N	N	N
Thirty logical partitions	Y	Y	Y	Y	Y	Y	Y	Y
Logical partition (LPAR) group capacity limit	Y	Y	Y	Y	Y	Y	N/A	N/A
CPU measurement facility	Y	Y	Y	Y[j]	Y[j]	Y[j]	Y[bj]	Y[bj]
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	Y[j]	N	N[e]	N[e]
Enhanced flexibility for Capacity on Demand (CoD)	Y	Y[f]	Y[f]	Y[f]	Y[f]	Y[f]	Y[f]	Y[f]
HiperDispatch	Y	Y	Y	Y	Y	Y	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[j]	Y[j]	Y[j]	N	N	N[e]	N[e]
Multiple subchannel sets	Y	Y	Y	Y	Y	Y	N[e]	N[e]
IPL from alternate subchannel set	Y[j]	Y[j]	Y[j]	N	N	N	N[e]	N[e]
Modified Indirect Data Address Word (MIDAW) facility	Y	Y	Y	Y	Y	Y	Y[b]	Y[b]
Cryptography								
CP Assist for Cryptographic Function (CPACF)	Y	Y	Y	Y	Y	Y	Y[b]	Y[b]
CPACF AES-128, AES-192, and AES-256	Y	Y	Y	Y	Y	Y	Y[b]	Y[b]
CPACF SHA-1, SHA-224, SHA-256, SHA-384, and SHA-512	Y	Y	Y	Y	Y	Y	Y[b]	Y[b]

Function	z/OS V1 R13	z/OS V1 R12	z/OS V1 R11	z/OS V1 R10	z/OS V1 R9	z/OS V1 R8	z/VM V6 R1	z/VM V5 R4
CPACF protected key	Y	Y	Y[g]	Y[g]	Y[g]	N	Y[bj]	Y[bj]
CPACF recent enhancements (z114)	Y	Y[g]	Y[g]	Y[g]	N	N	Y[bj]	Y[bj]
Crypto Express3 [f]	Y	Y	Y[g]	Y[g]	Y[g]	Y[agj]	Y[bj]	Y[bj]
Crypto Express3-1P [f]	Y	Y	Y[g]	Y[g]	Y[g]	Y[agj]	Y[bj]	Y[bj]
Crypto Express3 enhancements [f]	Y[gj]	Y[gj]	Y[gj]	Y[gj]	N	N	Y[bj]	Y[bj]
Elliptic Curve Cryptography (ECC)	Y	Y[g]	Y[g]	Y[g]	N	N	Y[bj]	Y[bj]
HiperSockets								
32 Hipersockets	Y	Y[j]	Y[j]	Y[j]	N	N	Y[j]	Y[j]
HiperSockets integration with IEDN [h]	Y[j]	N	N	N	N	N	N	N
HiperSockets Completion Queue [i]	Y	N	N	N	N	N	N	N
HiperSockets Network Traffic Analyzer	N	N	N	N	N	N	Y[b j]	Y[b j]
HiperSockets Multiple Write Facility	Y	Y	Y	Y	Y[j]	N	N[e]	N[e]
HiperSockets support of IPV6	Y	Y	Y	Y	Y	Y	Y	Y
HiperSockets Layer 2 support	Y	Y	N	N	N	N	Y[b]	Y[b]
HiperSockets	Y	Y	Y	Y	Y	Y	Y	Y
Enterprise Systems Connection (ESCON)								
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	Y[j]	N	N	N	N	N	N
zHPF enhanced multitrack support	Y	Y	Y[j]	Y[j]	N	N	N	N
High Performance FICON for System z (zHPF)	Y	Y	Y	Y[j]	Y[j]	Y[j]	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 and 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	N	N	N	N	N	N	Y	Y

Function	z/OS V1 R13	z/OS V1 R12	z/OS V1 R11	z/OS V1 R10	z/OS V1 R9	z/OS V1 R8	z/VM V6 R1	z/VM V5 R4
FICON Express8S, FICON Express8, FICON Express4, and FICON Express-2C support of SCSI disks CHPID type FCP	N	N	N	N	N	N	Y[j]	Y[j]
FICON Express8S CHPID type FC	Y	Y[j]	Y[j]	Y[j]	Y[j]	Y[j]	Y	Y
FICON Express8 CHPID type FC	Y	Y	Y[k]	Y[k]	Y[k]	Y[k]	Y[k]	Y[k]
FICON Express4-2C CHPID type FC	Y	Y	Y	Y	Y	Y	Y	Y
FICON Express4 [l] CHPID type FC	Y	Y	Y	Y	Y	Y	Y	Y
Open Systems Adapter (OSA)								
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	N/A	N/A	N/A	N/A	N/A	N/A	Y	Y[j]
OSA Layer 3 Virtual MAC	Y	Y	Y	Y	Y	Y	Y[b]	Y[b]
OSA Dynamic LAN idle	Y	Y	Y	Y	Y	Y	Y[b]	Y[b]
OSA/SF enhancements for IP, MAC addressing (CHPID type OSD)	Y	Y	Y	Y	Y	Y	Y	Y
Queued direct I/O (QDIO) diagnostic synchronization	Y	Y	Y	Y	Y	Y	Y[b]	Y[b]
OSA-Express2 Network Traffic Analyzer	Y	Y	Y	Y	Y	Y	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[m]	Y[m]
OSA-Express4S and OSA-Express3 inbound workload queueing for Enterprise Extender	Y	Y	N	N	N	N	Y[b j]	Y[b j]
OSA-Express4S 10 Gigabit Ethernet LR and SR CHPID type OSD	Y	Y	Y	Y	Y	Y	Y	Y
OSA-Express4S 10 Gigabit Ethernet LR and SR CHPID type OSX	Y	Y[j]	Y[j]	Y[j]	N	N	Y[j]	Y[j]
OSA-Express4S Gigabit Ethernet LX and SX CHPID type OSD (using two ports per CHPID)	Y	Y	Y	Y	Y[j]	Y[j]	Y	Y
OSA-Express4S Gigabit Ethernet LX and SX CHPID type OSD (using one port per CHPID)	Y	Y	Y	Y	Y	Y	Y	Y
OSA-Express3 10 Gigabit Ethernet LR and SR CHPID type OSD	Y	Y	Y	Y	Y	Y	Y	Y
OSA-Express3 10 Gigabit Ethernet LR and SR CHPID type OSX	Y	Y[j]	Y[j]	Y[j]	N	N	Y[j]	Y[j q]

Function	z/OS V1 R13	z/OS V1 R12	z/OS V1 R11	z/OS V1 R10	z/OS V1 R9	z/OS V1 R8	z/VM V6 R1	z/VM V5 R4
OSA-Express3 Gigabit Ethernet LX and SX CHPID types OSD, OSN [n] (using two ports per CHPID)	Y	Y	Y	Y	Y[j]	Y[j]	Y	Y[j]
OSA-Express3 Gigabit Ethernet LX and SX CHPID types OSD, OSN [n] (using one port per CHPID)	Y	Y	Y	Y	Y	Y	Y	Y
OSA-Express3-2P Gigabit Ethernet SX CHPID types OSD and OSN [n]	Y	Y	Y	Y	Y	Y	Y	Y
OSA-Express3 1000BASE-T CHPID type OSC (using two ports per CHPID)	Y	Y	Y	Y	Y[j]	Y[j]	Y	Y
OSA-Express3 1000BASE-T CHPID type OSD (using two ports per CHPID)	Y	Y	Y	Y	Y[j]	Y[j]	Y	Y[j]
OSA-Express3 1000BASE-T CHPID types OSC and OSD (using one port per CHPID)	Y	Y	Y	Y	Y	Y	Y	Y
OSA-Express3 1000BASE-T CHPID type OSE (using one or two ports per CHPID)	Y	Y	Y	Y	Y	Y	Y	Y
OSA-Express3 1000BASE-T CHPID type OSM (using one port per CHPID)	Y	Y[j]	Y[j]	Y[j]	N	N	Y[j]	Y[j] [q]
OSA-Express3 1000BASE-T CHPID type OSN [n]	Y	Y	Y	Y	Y	Y	Y	Y
OSA-Express3-2P 1000BASE-T Ethernet CHPID types OSC, OSD, OSE, and OSN [o]	Y	Y	Y	Y	Y	Y	Y	Y
OSA-Express3-2P 1000BASE-T Ethernet CHPID type OSM [p]	Y	Y	Y	Y	N	N	Y	Y[q]
OSA-Express2 Gigabit Ethernet LX and SX [p] CHPID type OSD	Y	Y	Y	Y	Y	Y	Y	Y
OSA-Express2 Gigabit Ethernet LX and SX [p] 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 [n]		Y	Y	Y	Y	Y	Y	Y

Function	z/OS V1 R13	z/OS V1 R12	z/OS V1 R11	z/OS V1 R10	z/OS V1 R9	z/OS V1 R8	z/VM V6 R1	z/VM V5 R4
Parallel Sysplex and other								
z/VM integrated systems management	N/A	N/A	N/A	N/A	N/A	N/A	Y	Y
System-initiated CHPID reconfiguration	Y	Y	Y	Y	Y	Y	N/A	N/A
Program-directed re-IPL	N/A	N/A	N/A	N/A	N/A	N/A	Y	Y
Multipath IPL	Y	Y	Y	Y	Y	Y	N	N
Server Time Protocol (STP) enhancements	Y	Y	Y	Y	Y	Y	N/A	N/A
Server Time Protocol	Y	Y	Y	Y	Y	Y	N/A	N/A
Coupling over InfiniBand CHPID type CIB	Y	Y	Y	Y	Y	Y	Y[q]	Y[q]
InfiniBand coupling links (12x IB-SDR or 12x IB-DDR) at a distance of 150 m (492.1 ft.)	Y	Y	Y	Y	Y[j]	Y[j]	Y[q]	Y[q]
InfiniBand coupling links (1x IB-SDR or 1xIB DDR) at an unrepeated distance of 10 km (32.9 ft.)	Y	Y	Y	Y	Y[j]	Y[j]	Y[q]	Y[q]
Dynamic I/O support for InfiniBand CHPIDs	N/A	N/A	N/A	N/A	N/A	N/A	Y[q]	Y[q]
CFCC Level 17	Y	Y[j]	Y[j]	Y[j]	N	N	Y[b]	Y[b]

a. Toleration support only.
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. FMIDs are shipped in a web deliverable.
h. z/VM plans to provide HiperSockets intraensemble data network (IEDN) bridging support in a future deliverable.
i. z/VM plans to provide guest support for HiperSockets Completion Queue in a future deliverable.
j. Service is required.
k. Support varies with operating system and level. For details see 8.3.33, “FCP provides increased performance” on page 239.
l. FICON Express4 10KM LX, 4KM LX, and SX features are withdrawn from marketing.
m. Supported for dedicated devices only.
n. CHPID type OSN does not use ports. All communication is LPAR to LPAR.
o. One port is configured for OSM. The other port in the pair is unavailable.
p. Withdrawn from marketing.
q. Support is for dynamic I/O configuration only.

Table 8-4 z114 functions minimum support requirements summary, part 2

Function	z/VSE V5R1[a]	z/VSE V4R3[b]	z/VSE V4R2[b]	z/TPF V1R1	Linux on System z
z114	Y	Y	Y	Y	Y
Support of Unified Resource Manager	N	N	N	N	N
zIIP	N/A	N/A	N/A	N/A	N/A
zAAP	N/A	N/A	N/A	N/A	N/A
zAAP on zIIP	N/A	N/A	N/A	N/A	N/A
Large memory (> 128 GB)	N	N	N	Y	Y
Large page support	Y	Y	N	N	Y
Out-of-order execution	Y	Y	Y	Y	Y
Guest support for Execute-extensions facility	N/A	N/A	N/A	N/A	N/A
Hardware decimal floating point [c]	N	N	N	N	Y[d]
Zero address detection	N	N	N	N	N
30 logical partitions	Y	Y	Y	Y	Y
CPU measurement facility	N	N	N	N	N
LPAR group capacity limit	N/A	N/A	N/A	N/A	N/A
Separate LPAR management of PUs	Y	Y	Y	Y	Y
Dynamic add/delete logical partition name	N	N	N	N	Y
Capacity provisioning	N/A	N/A	N/A	N	N/A
Enhanced flexibility for CoD	N/A	N/A	N/A	N	N/A
HiperDispatch	N	N	N	N	N
63.75 K subchannels	N	N	N	N	Y
Four logical channel subsystems (LCSSs)	Y	Y	Y	N	Y
Dynamic I/O support for multiple LCSSs	N	N	N	N	Y
Third subchannel set	N	N	N	Y	N
Multiple subchannel sets	N	N	N	N	Y
IPL from alternate subchannel set	N	N	N	N	N
MIDAW facility	N	N	N	N	N
Cryptography					
CPACF	Y	Y	Y	Y[g]	Y
CPACF AES-128, AES-192, and AES-256	Y	Y	Y[g]	Y[ge]	Y
CPACF SHA-1, SHA-224, SHA-256, SHA-384, and SHA-512	Y	Y	Y[g]	Y[gf]	Y
CPACF protected key	N	N	N	N	N
CPACF recent enhancements (z114)	N	N	N	Y	N[k]

Function	z/VSE V5R1[a]	z/VSE V4R3[b]	z/VSE V4R2[b]	z/TPF V1R1	Linux on System z
Crypto Express3 [c]	Y	Y	Y[g]	Y[gh]	Y
Crypto Express3-1P [c]	Y	Y	Y[g]	Y[gh]	Y
Crypto Express3 enhancements [c]	N	N	N	N	N[k]
Elliptic Curve Cryptography (ECC)	N	N	N	N	N[k]
HiperSockets					
32 Hipersockets	Y	Y	Y	Y	Y
HiperSockets integration with IEDN	Y[i]	N	N	N	N
HiperSockets Completion Queue	Y[j]	N	N	N	N
HiperSockets Network Traffic Analyzer	N	N	N	N	Y[k]
HiperSockets Multiple Write Facility	N	N	N	N	N
HiperSockets support of IPV6	N	N	N	N	Y
HiperSockets Layer 2 support	N	N	N	N	Y
HiperSockets	Y	Y	Y	N	Y
Enterprise System Connection (ESCON)					
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	N	Y
Request node identification data	N/A	N/A	N/A	N/A	N/A
FICON link incident reporting	N	N	N	N	N
N_Port ID Virtualization for FICON (NPIV) CHPID type FCP	Y	Y	Y	N	Y
FCP point-to-point attachments	Y	Y	Y	N	Y
FICON SAN platform and name registration	Y	Y	Y	Y	Y
FCP SAN management	N	N	N	N	Y
SCSI IPL for FCP	Y	Y	Y	N	Y
Cascaded FICON Directors CHPID type FC	Y	Y	Y	Y	Y
Cascaded FICON Directors CHPID type FCP	Y	Y	Y	N	Y
FICON Express 8S, FICON Express8, FICON Express4, and FICON Express4-2C support of SCSI disks CHPID type FCP	Y	Y	Y	N	Y

Function	z/VSE V5R1[a]	z/VSE V4R3[b]	z/VSE V4R2[b]	z/TPF V1R1	Linux on System z
FICON Express8S [c] CHPID type FC	Y	Y	Y	Y	Y
FICON Express8 [c] CHPID type FC	Y	Y[l]	Y[l]	Y[l]	Y[l]
FICON Express4-2C [c] CHPID type FC	Y	Y	Y	Y	Y
FICON Express4 [c] [m] CHPID type FC	Y	Y	Y	Y	Y
Open Systems Adapter (OSA)					
VLAN management	N	N	N	N	N
VLAN (IEE 802.1q) support	N	N	N	N	Y
QDIO data connection isolation for z/VM virtualized environments	N/A	N/A	N/A	N/A	N/A
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
QDIO diagnostic synchronization	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	N	Y
Checksum offload for IPv4 packets	N	N	N	N	Y
OSA-Express4S 10 Gigabit Ethernet LR and SR CHPID type OSD	Y	Y	Y	Y	Y
OSA-Express4S 10 Gigabit Ethernet LR and SR CHPID type OSX	Y	N	N	Y	Y[n]
OSA-Express4S Gigabit Ethernet LX and SX CHPID type OSD (using two ports per CHPID)	Y	Y	Y	Y	Y[k]
OSA-Express4S Gigabit Ethernet LX and SX CHPID type OSD (using one port per CHPID)	Y	Y	Y	Y	Y
OSA-Express3 10 Gigabit Ethernet LR and SR CHPID type OSD	Y	Y	Y	Y	Y
OSA-Express3 10 Gigabit Ethernet LR and SR CHPID type OSX	Y	N	N	N	Y[n]
OSA-Express3 Gigabit Ethernet LX and SX CHPID types OSD, OSN [o] (using two ports per CHPID)	Y	Y	Y	Y[n]	Y
OSA-Express3 Gigabit Ethernet LX and SX CHPID types OSD, OSN [o] (using one port per CHPID)	Y	Y	Y	Y[n]	Y
OSA-Express3-2P Gigabit Ethernet SX CHPID types OSD and OSN [o]	Y	Y	Y	Y	Y

Function	z/VSE V5R1[a]	z/VSE V4R3[b]	z/VSE V4R2[b]	z/TPF V1R1	Linux on System z
OSA-Express3 1000BASE-T CHPID type OSC (using four ports)	Y	Y	Y	N	N/A
OSA-Express3 1000BASE-T (using two ports per CHPID) CHPID type OSD	Y	Y	Y	Y[n]	Y
OSA-Express3 1000BASE-T (using one port per CHPID) CHPID type OSD	Y	Y	Y	Y	Y
OSA-Express3 1000BASE-T (using one or two ports per CHPID) CHPID type OSE	Y	Y	Y	N	N
OSA-Express3 1000BASE-T Ethernet CHPID type OSN [o]	Y	Y	Y	Y	Y
OSA-Express3 1000BASE-T CHPID type OSM (using two ports)	N	N	N	N	N
OSA-Express3-2P 1000BASE-T Ethernet CHPID types OSC, OSD, OSE, and OSN [p]	Y	Y	Y	Y[p]	Y[q]
OSA-Express3-2P 1000BASE-T Ethernet CHPID type OSM [r]	N	N	N	N	N
OSA-Express2 Gigabit Ethernet LX and SX [s] CHPID type OSD	Y	Y	Y	Y	Y
OSA-Express2 Gigabit Ethernet LX and SX 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

Function	z/VSE V5R1[a]	z/VSE V4R3[b]	z/VSE V4R2[b]	z/TPF V1R1	Linux on System z
OSA-Express2 1000BASE-T Ethernet CHPID type OSN [o]	Y	Y	Y	Y	Y
Parallel Sysplex and other					
z/VM integrated systems management	N/A	N/A	N/A	N/A	N/A
System-initiated CHPID reconfiguration	N/A	N/A	N/A	N/A	Y
Program-directed re-IPL [s]	Y	Y	Y	N/A	Y
Multipath IPL	N/A	N/A	N/A	N/A	N/A
STP enhancements	N/A	N/A	N/A	N/A	N/A
Server Time Protocol	N/A	N/A	N/A	N/A	N/A
Coupling over InfiniBand CHPID type CIB	N/A	N/A	N/A	Y	N/A
InfiniBand coupling links (1x IB-SDR or IB-DDR) at unrepeated distance of 10 km (32.9 ft.)	N/A	N/A	N/A	N/A	N/A
Dynamic I/O support for InfiniBand CHPIDs	N/A	N/A	N/A	N/A	N/A
CFCC Level 17	N/A	N/A	N/A	Y	N/A

a. z/VSE V5R1 is designed to exploit z/Architecture, specifically 64-bit real and virtual-memory addressing. z/VSE V5R1 requires an architectural level set that is available with IBM System z9 or later.
b. z/VSE V4 is designed to exploit z/Architecture, specifically 64-bit real-memory addressing, but does not support 64-bit virtual-memory addressing.
c. Support varies with operating system and level.
d. Supported by Novell SUSE SLES 11.
e. z/TPF supports only AES-128 and AES-256.
f. z/TPF supports only SHA-1 and SHA-256.
g. Service is required.
h. Supported only running in accelerator mode (CEX3A).
i. z/VSE plans to provide HiperSockets IEDN bridging support in a future deliverable.
j. z/VSE plans to provide guest support for HiperSockets Completion Queue in a future deliverable.
k. IBM is working with its Linux distribution partners to include support in future Linux on System z distribution releases.
l. For details, see 8.3.33, “FCP provides increased performance” on page 239.
m. FICON Express4 10KM LX, 4KM LX, and SX features are withdrawn from marketing.
n. Requires PUT4 with PTFs.
o. CHPID type OSN does not use ports. All communication is LPAR to LPAR.
p. CHPID type OSE is not supported on z/TPF.
q. CHPID types OSC and OSE are not supported on Linux for System z.
r. One port is configured for OSM. The other port is unavailable.
s. This function 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. On the z114, the image size is restricted by the number of PUs available: maximum five CPs, zAAPs, or zIIPs, and maximum ten IFLs or ICFs.

Table 8-5 Single system image software support

Operating system	Maximum number of (CPs+zIIPs+zAAPs)[a] or IFLs per system image[b]
z/OS V1R11 and later	80
z/OS V1R10	80[c]
z/OS V1R9	64
z/OS V1R8	32
z/VM V6R1	32
z/VM V5R4	32
z/VSE V4R2 and later	z/VSE Turbo Dispatcher can exploit up to 4 CPs and tolerates up to 10-way LPARs
z/TPF V1R1	84 CPs
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

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. On a z114, a maximum of five CPs, two ZAAPs, two ZIIPs, and ten IFLs can be configured.

c. Service is required.

The z/VM-mode logical partition

The z114 supports a logical partition (LPAR) mode, named z/VM-mode, which is exclusive for running z/VM. The z/VM-mode requires z/VM V5R4 or later and allows z/VM to utilize a wider variety of specialty processors in a single LPAR. For instance, in a z/VM-mode LPAR, z/VM can manage Linux on System z guests running on IFL processors while also managing z/VSE and z/OS on central processors (CPs), and allowing z/OS to fully exploit IBM System z Integrated Information Processors (zIIPs) and IBM System z Application Assist Processors (zAAPs).

8.3.2 zAAP support

zAAPs do not change the model capacity identifier of the z114. IBM software product license charges based on the model capacity identifier are not affected by the addition of zAAPs. On a z114, z/OS Version 1 Release 8 is the minimum level for supporting zAAPs, together with IBM Software Developer Kit (SDK) for z/OS Java 2 Technology Edition V1.4.1.

The following applications exploit zAAPs:

- Any Java application that uses the current IBM SDK.

- WebSphere Application Server V5R1 and later, and products based on it, such as WebSphere Portal, WebSphere Enterprise Service Bus (WebSphere ESB), WebSphere Business Integration (WBI) for z/OS, and so on.
- CICS/TS V2R3 and later.
- DB2 UDB for z/OS Version 8 and later.
- IMS Version 8 and later.
- All z/OS XML System Services validation and parsing that execute in task control block (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 the rollback of z/OS XML System Services.

The functioning of a zAAP is transparent to all Java programming on Java virtual machine (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 Workload Manager (WLM) service class indicates the percentage of a processor that is zAAP eligible. Because of zAAP's lower prices, as compared to CPs, an utilization as low as 10% might provide benefit.

8.3.3 zIIP support

zIIPs do not change the model capacity identifier of the z114. IBM software product license charges based on the model capacity identifier are not affected by the addition of zIIPs. On a z114, z/OS Version 1 Release 8 is the minimum level for supporting zIIPs.

No changes to applications are required to exploit zIIPs. The following products and functions exploit zIIPs:

- DB2 V8 and later for z/OS data serving, for applications using data Distributed Relational Database Architecture (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 Global Business Services (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 Workload Manager service class indicates the percentage of a processor that is zIIP eligible. Because of zIIP's lower prices, as compared to CPs, an utilization as low as 10% might provide benefit.

8.3.4 zAAP on zIIP capability

This capability, which was first made available on System z9 servers under defined circumstances, enables workloads that are eligible to run on Application Assist Processors (zAAPs) to run on Integrated Information Processors (zIIPs). 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 it 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 might 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 later. 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 lists the maximum amount of main storage that is supported by the current operating systems. Expanded storage, although part of the z/Architecture, is currently exploited only by z/VM. A maximum of 248 GB of main storage can be defined for a logical partition on a z114.

Table 8-6 Maximum memory supported by operating system

Operating system	Maximum supported main storage
z/OS	z/OS V1R8 and higher support 4 TB and up to 3 TB per server[a]
z/VM	z/VM V5R4 and higher support 256 GB
z/VSE	z/VSE V4R2 and higher support 32 GB
z/TPF	z/TPF supports 4 TB[a]
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

a. System z114 restricts the maximum LPAR memory size to 248 GB.

8.3.6 Large page support

In addition to the existing 4 KB pages and page frames, z114 supports large pages and large page frames that are 1 MB in size, as described in “Large page support” on page 87. 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 the execute-extensions facility

Operating system	Support requirement
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 z114 has inherited the decimal floating point accelerator feature that was introduced with the z10 EC and described in 3.3.4, “Decimal floating point accelerator” on page 71.

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 261.

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/VM	z/VM V5R4: Support is for guest use.
Linux on System z	Novell SUSE SLES 11.

8.3.9 Up to 30 logical partitions

This feature, which was first made available in the z10 BC, allows the system to be configured with up to 30 logical partitions. Because channel subsystems can be shared by up to 15 logical partitions, it is necessary to configure two channel subsystems to reach 30 logical partitions. Table 8-10 lists the minimum operating system levels for supporting 30 logical partitions.

Table 8-10 Minimum support requirements for 60 logical partitions

Operating system	Support requirements
z/OS	z/OS V1R8
z/VM	z/VM V5R4
z/VSE	z/VSE V4R2
z/TPF	z/TPF V1R1
Linux on System z	Novell SUSE SLES 10 Red Hat RHEL 5

8.3.10 Separate LPAR management of PUs

The z114 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 230 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 V1R8
z/VM	z/VM V5R4
z/VSE	z/VSE V4R2
z/TPF	z/TPF V1R1
Linux on System z	Novell SUSE SLES 10 Red Hat RHEL 5

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, which is described in 9.8, "Nondisruptive upgrades" on page 312, enables you to better control the configuration and activation of On/Off 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, which is a function that was 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. Table 8-12 lists the support requirements.

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 nondisruptively bringing online the additional computing resources when needed. Starting with z/OS V1R10, z/VM V5R4, and z/VSE V4R3, you can use an enhanced capability, which is the ability to dynamically define and change the number and type of reserved PUs in an LPAR profile, 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 z114, z196, and System z10, represents a cooperative effort between the z/OS operating system and the z114 hardware. It improves efficiencies in both the hardware and the software in the following ways:

- Work can be dispatched across fewer logical processors, therefore reducing the multiprocessor (MP) effects and lowering the interference among multiple partitions.
- Specific z/OS tasks can be dispatched to a small subset of logical processors that Processor Resource/Systems Manager (PR/SM) will tie to the same physical processors. This action improves the hardware cache reuse and the locality of reference characteristics, such as reducing the rate of cross-book communication.

For more information, see 3.6, “Logical partitioning” on page 89. Table 8-13 lists the HiperDispatch support requirements.

Table 8-13 Minimum support requirements for HiperDispatch

Operating system	Support requirements
z/OS	z/OS V1R8 and later with PTFs
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 available subchannels. Reserved subchannels exist only in subchannel set 0. No subchannels are reserved in subchannel set 1[3]. The informal name, *63.75 K subchannels*, represents 65280 subchannels, as shown in the following equation:

```
63 x 1024 + 0.75 x 1024 = 65280
```

[3] Subchannel set 2 is not available on the z114.

Table 8-14 lists the minimum operating system level that was required on the z114.

Table 8-14 Minimum support requirements for 63.75 K subchannels

Operating system	Support requirements
z/OS	z/OS V1R8
z/VM	z/VM V5R4
Linux on System z	Novell SUSE SLES 10 Red Hat RHEL 5

8.3.16 Multiple subchannel sets

Multiple subchannel sets (MSS), which were 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 type FC. z114 has a second subchannel set (SS1). The third subchannel set, which was introduced with the z196, is not available on the z114.

Multiple subchannel sets are not supported for z/OS when running as a guest of z/VM.

Table 8-15 lists the minimum operating systems level required on the z114.

Table 8-15 Minimum software requirement for MSS

Operating system	Support requirements
z/OS	z/OS V1R8
Linux on System z	Novell SUSE SLES 10 Red Hat RHEL 5

8.3.17 IPL from an alternate subchannel set

z114 supports IPL from subchannel set 1 (SS1), in addition to subchannel set 0. Devices used early during IPL processing can now be accessed using subchannel set 1. This capability allows the users of Metro Mirror (PPRC) secondary devices that were defined using the same device number and a new device type in an alternate subchannel set to be used for IPL, input/output definition file (IODF), and stand-alone dump volumes when needed.

IPL from an alternate subchannel set is exclusive to z196 and z114, and it is supported by z/OS V1R13, as well as V1R12 and V1R11 with PTFs, and applies to the FICON and zHPF protocols.

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 263.

Table 8-16 on page 233 lists the minimum support requirements for MIDAW.

Table 8-16 Minimum support requirements for MIDAW

Operating system	Support requirements
z/OS	z/OS V1R8
z/VM	z/VM V5R4 for guest exploitation

8.3.19 Enhanced CPACF

We describe cryptographic functions in 8.4, “Cryptographic support” on page 252.

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 capability is transparent to the operating system in the receiving partition. HiperSockets Multiple Write Facility with fewer I/O interrupts is designed to reduce the CPU utilization of the sending and receiving partitions.

Support for this function is required by the sending operating system. See 4.8.8, “HiperSockets” on page 138.

Table 8-17 Minimum support requirements for HiperSockets multiple write facility

Operating system	Support requirement
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 8-18 lists the minimum support requirements for HiperSockets IPv6 (CHPID type IQD).

Table 8-18 Minimum support requirements for HiperSockets IPv6 (CHPID type IQD)

Operating system	Support requirements
z/OS	z/OS V1R8
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 z114 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 or 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 in the same way that they configure and maintain a non-mainframe environment.

Table 8-19 show the requirements for HiperSockets Layer 2 support.

Table 8-19 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 z114, with support to trace Layer2 and Layer3 HiperSockets network traffic in Linux on System z. This enhancement 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).

You can use Linux on System z tools to format, edit, and process the trace records for analysis by system programmers and network administrators.

8.3.24 HiperSockets statements of direction

BM has issued the following Statement of General Direction.

Statements of General Direction about HiperSockets Completion Queue: IBM plans to support transferring HiperSockets messages asynchronously, in addition to the current synchronous manner on z196 and z114. This capability can be especially helpful in burst situations. The Completion Queue function is designed to allow HiperSockets to transfer data synchronously if possible and asynchronously if necessary, thus combining ultra-low latency with more tolerance for traffic peaks. HiperSockets Completion Queue is planned to be supported in the z/VM and z/VSE environments.

HiperSockets integration with the IEDN: Within a zEnterprise environment, it is planned for HiperSockets to be integrated with the intraensemble data network (IEDN), extending the reach of the HiperSockets network outside of the central processor complex (CPC) to the entire ensemble, appearing as a single Layer 2 network. HiperSockets integration with the IEDN is planned to be supported in z/OS V1.13 and z/VM in a future deliverable.

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

8.3.25 FICON Express8S

The FICON Express8S feature resides exclusively in the PCIe I/O drawer and provides a link rate of 8 Gbps, with autonegotiation to 4 Gbps or 2 Gbps, for compatibility with previous devices and investment protection. Both 10KM LX and SX connections are offered (in a specific feature, all connections must have the same type).

With FICON Express 8S, clients might be able to consolidate existing FICON, FICON Express2, and FICON Express4 channels, while maintaining and enhancing performance.

Table 8-20 lists the minimum support requirements for FICON Express8S. The tables in this section use the following convention:

- N/A is not applicable.
- NA is not available.

Table 8-20 Minimum support requirements for FICON Express8S

Operating system	z/OS	z/VM	z/VSE	z/TPF	Linux on System z
Native FICON and Channel-to-Channel (CTC) CHPID type FC	V1R8	V5R4	V4R2	V1R1	Novell SUSE SLES 10 Red Hat RHEL 5
zHPF single track operations CHPID type FC	V1R8[a]	NA	NA	NA	Novell SUSE SLES 11 SP1 Red Hat RHEL 6
zHPF multitrack operations CHPID type FC	V1R9[a]	NA	NA	NA	NA
Support of SCSI devices CHPID type FCP	N/A	V5R4[a]	V4R2	N/A	Novell SUSE SLES 10 Red Hat RHEL 5

[a] PTFs required

8.3.26 FICON Express8

The FICON Express8 features provide a link rate of 8 Gbps, with autonegotiation to 4 Gbps or 2 Gbps, for compatibility with previous devices and investment protection. Both 10KM LX and SX connections are offered (in a specific feature, all connections must have the same type).

With FICON Express 8, client might be able to consolidate existing FICON, FICON Express2, and FICON Express4 channels, while maintaining and enhancing performance.

Table 8-21 on page 236 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	z/TPF	Linux on System z
Native FICON and Channel-to-Channel (CTC) CHPID type FC	V1R8	V5R4	V4R2	V1R1	Novell SUSE SLES 10 Red Hat RHEL 5
zHPF single track operations CHPID type FC	V1R8[a]	NA	NA	NA	NA
zHPF multitrack operations CHPID type FC	V1R9[a]	NA	NA	NA	NA
Support of SCSI devices CHPID type FCP	N/A	V5R4[a]	V4R2	N/A	Novell SUSE SLES 10 Red Hat RHEL 5

[a] PTFs required

8.3.27 z/OS discovery and autoconfiguration (zDAC)

The z/OS discovery and autoconfiguration for FICON channels (zDAC) function is designed to automatically perform a number of I/O configuration definition tasks for new and changed disk and tape controllers that are connected to a switch or director, when attached to a FICON channel.

The zDAC function is integrated into the existing Hardware Configuration Definition (HCD). Clients can define a policy that can 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 that are based on that policy.

zDAC provides real-time discovery for the FICON fabric, subsystem, and I/O device resource changes from z/OS. By exploring the discovered control units for defined logical control units (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 newly 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 based on the defined policy. zDAC uses the channel path chosen algorithm to minimize single point of failure. The zDAC proposed configurations are created as work I/O definition files (IODFs) that can be converted to production IODFs and activated.

zDAC is designed to perform discovery for all systems in a sysplex that support the function. Thus, zDAC helps you to simplify the I/O configuration on z114 systems running z/OS and reduces the complexity and setup time.

zDAC applies to all FICON features that are supported on z114 when configured as CHPID type FC. Table 8-22 on page 237 lists the minimum support requirements for zDAC.

Table 8-22 Minimum support requirements for zDAC

Operating system	Support requirement
z/OS	z/OS V1R12

8.3.28 High performance FICON (zHPF)

High performance FICON (zHPF), which was 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 online 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 (LIC) 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 these standards:

- 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, partitioned data set extended (PDSE), and zFS.

At its announcement, zHPF supported the transfer of small blocks of fixed size data (4 K). This support has been extended on z10 EC to multitrack operations (limited to 64 k bytes), and z114 has removed the 64k byte data transfer limit on multitrack operations. This improvement allows the channel to fully exploit the bandwidth of FICON channels and results in higher throughputs and lower response times.

The multitrack operations extension applies exclusively to the FICON Express8S, FICON Express8, and FICON Express4 on z114 and z196, when configured as CHPID type FC, and when 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.

> **CHPIDs:** All FICON channel paths (CHPIDs) that are 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 that is used for an I/O operation depends on the control unit supporting zHPF and the 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 differs significantly 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 z114, z196, and System z10. The FICON Express8S, FICON Express8, and FICON Express4[4] (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.

Table 8-23 Minimum support requirements for zHPF

Operating system	Support requirements
z/OS	Single track operations: z/OS V1R8 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 z114 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 website:

http://www-03.ibm.com/systems/z/hardware/connectivity/ficon_performance.html

8.3.29 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 requirement
z/OS	z/OS V1R8

8.3.30 Extended distance FICON

An enhancement to the industry standard FICON architecture (FC-SB-3) helps avoid the 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 Express8S, 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. The IBM System Storage DS8000 series supports extended distance FICON for IBM System z environments.

[4] FICON Express4 10KM LX, 4KM LX, and SX features are withdrawn from marketing. All FICON Express2 and FICON features are withdrawn from marketing.

The channel defaults to the current pacing values when it operates with control units that cannot exploit the extended distance FICON.

8.3.31 Platform and name server registration in the FICON channel

The FICON Express8S, FICON Express8, and FICON Express4 features on the z114 servers support platform and name server registration to the fabric for both CHPID types FC and FCP.

Information about the channels that are connected to a fabric, if registered, allows other nodes or storage area network (SAN) managers to query the name server to determine what is connected to the fabric.

The following attributes are registered for the z114 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 the name server registration service are defined in the Fibre Channel - Generic Services 4 (FC-GS-4) standard.

8.3.32 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 requirement
z/OS	z/OS V1R8

8.3.33 FCP provides increased performance

The Fibre Channel Protocol (FCP) LIC 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 website:

http://www-03.ibm.com/systems/z/hardware/connectivity/fcp_performance.html

8.3.34 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 system image was the sole user of the channel. You can use this feature, which was first introduced with z9 EC, with earlier FICON features that have been carried forward from earlier servers.

Table 8-26 on page 240 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 V4R2.
Linux on System z	Novell SUSE SLES 10 SP3. Red Hat RHEL 5.4.

8.3.35 OSA-Express4S 10 Gigabit Ethernet LR and SR

The OSA-Express4S 10 Gigabit Ethernet feature resides exclusively in PCIe I/O drawer. Each feature has one port, which is 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 z114 supports the OSX CHPID type that was introduced with the z196; see 8.3.43, "Intraensemble data network (IEDN)" on page 244.

The OSA-Express4S features have half the number of ports per feature, when compared to the OSA-Express3, and half the size as well. This design actually results in an increased number of installable features while facilitating the purchase of the correct number of ports to help satisfy your application requirements and to better optimize for redundancy. Table 8-27 lists the minimum support requirements for OSA-Express4S 10 Gigabit Ethernet LR and SR features.

Table 8-27 Minimum support requirements for OSA-Express4S 10 Gigabit Ethernet LR and SR

Operating system	Support requirements
z/OS	OSD: z/OS V1R8 OSX: z/OS V1R10[a]
z/VM	OSD: z/VM V5R4 OSX: z/VM V5R4[a] for dynamic I/O only
z/VSE	OSD: z/VSE V4R2 OSX: z/VSE V5R1
z/TPF	OSD: z/TPF V1R1 OSX: z/TPF V1R1 PUT4[a]
Linux on System z	OSD: Novell SUSE SLES 10, Red Hat RHEL 5 OSX: Novell SUSE SLES 10 SP4, Red Hat RHEL 5.6

[a] PTFs required

8.3.36 OSA-Express4S Gigabit Ethernet LX and SX

The OSA-Express4S Gigabit Ethernet feature resides exclusively in the PCIe I/O drawer. Each feature has one PCIe adapter and two ports. The two ports share a channel path identifier (CHPID type OSD exclusively). Each port supports attachment to a 1 Gigabit per second (Gbps) Ethernet LAN. The ports can be defined as a spanned channel and can be shared among logical partitions and across logical channel subsystems.

Operating system support is required in order to recognize and use the second port on the OSA-Express4S Gigabit Ethernet feature. Table 8-28 on page 241 lists the minimum support requirements for the OSA-Express4S Gigabit Ethernet LX and SX features.

Table 8-28 Minimum support requirements for OSA-Express4S Gigabit Ethernet LX and SX

Operating system	Support requirements exploiting two ports per CHPID	Support requirements exploiting one port per CHPID
z/OS	OSD: z/OS V1R8[a]	OSD: z/OS V1R8
z/VM	OSD: z/VM V5R4[a]	OSD: z/VM V5R4
z/VSE	OSD: z/VSE V4R2	OSD: z/VSE V4R2
z/TPF	OSD: z/TPF V1R1 PUT 4[a]	OSD: z/TPF V1R1 PUT 4[a]
Linux on System z	OSD: Novell SUSE SLES 10 SP2, Red Hat RHEL 5.2	OSD: Novell SUSE SLES 10, Red Hat RHEL 5

[a] PTFs required

8.3.37 OSA-Express3 10 Gigabit Ethernet LR and SR

The OSA-Express3 10 Gigabit Ethernet features offer two ports, which are 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 z114 supports the OSX CHPID type that was introduced with the z196; see 8.3.43, “Intraensemble data network (IEDN)” on page 244.

Table 8-29 lists the minimum support requirements for OSA-Express3 10 Gigabit Ethernet LR and SR features.

Table 8-29 Minimum support requirements for OSA-Express3 10 Gigabit Ethernet LR and SR

Operating system	Support requirements
z/OS	OSD: z/OS V1R8 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 V4R2; service required
z/TPF	OSD: z/TPF V1R1
Linux on System z	OSD: Novell SUSE SLES 10 OSD: Red Hat RHEL 5

8.3.38 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, which is 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 8-30 on page 242 (four ports) and Table 8-31 on page 242 (two ports).

Table 8-30 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 V4R2; service required
z/TPF	z/TPF V1R1; service required
Linux on System z	Novell SUSE SLES 10 SP2 Red Hat RHEL 5.2

Table 8-31 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 V1R8
z/VM	z/VM V5R4
z/VSE	z/VSE V4R2; service required
z/TPF	z/TPF V1R1
Linux on System z	Novell SUSE SLES 10 Red Hat RHEL 5

8.3.39 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, which is 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 z114; see 8.3.42, "Intranode management network (INMN)" on page 244.

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. The minimum support requirements for OSA-Express3 1000BASE-T Ethernet feature are listed in Table 8-32 on page 243 (four ports) and Table 8-33 on page 243 (two ports).

Table 8-32 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 V1R8 OSM: z/OS V1R12; z/OS V1R10 and z/OS V1R11, with service OSN[b]: z/OS V1R8
z/VM	OSD: z/VM V5R4; service required OSE: z/VM V5R4 OSM: z/VM service required; V5R4 for dynamic I/O only OSN[b]: z/VM V5R4
z/VSE	OSD: z/VSE V4R2; service required OSE: z/VSE V4R2 OSN[b]: z/VSE V4R2; 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 8-33 on page 243.

b. Although CHPID type OSN does not use any ports (because all communication is LPAR to LPAR), it is listed here for completeness.

Table 8-33 lists the minimum support requirements for OSA-Express3 1000BASE-T Ethernet (two ports).

Table 8-33 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; V1R8
z/VM	OSD, OSE, OSM, and OSN: V5R4
z/VSE	V4R2
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

8.3.40 OSA-Express2 1000BASE-T Ethernet

The OSA-Express2 1000BASE-T Ethernet adapter can be configured in one of these modes:

- QDIO mode with CHPID type OSD or OSN
- Non-QDIO mode with CHPID type OSE
- Local 3270 emulation mode with CHPID type OSC

Table 8-34 lists the support for OSA-Express2 1000BASE-T.

Table 8-34 Minimum support requirements for OSA-Express2 1000BASE-T

Operating system	CHPID type OSC	CHPID type OSD	CHPID type OSE
z/OS V1R8	Supported	Supported	Supported
z/VM V5R4	Supported	Supported	Supported
z/VSE V4R2	Supported	Supported	Supported
z/TPF V1R1	Supported	Supported	Not supported
Linux on System z	Not supported	Supported	Not supported

8.3.41 Open System Adapter for Ensemble

Three separate OSA-Express3 features are used to connect the z114 central processor complex (CPC) to its attached IBM System z BladeCenter Extension (zBX) and other ensemble nodes. These connections are part of the ensemble's two private and secure internal networks.

For the intranode management network (INMN):

- OSA Express3 1000BASE-T Gigabit Ethernet (GbE), feature code 3367

For the intraensemble data network (IEDN):

- OSA-Express3 10 Gigabit Ethernet (GbE) Long Range (LR), feature code 3370
- OSA-Express3 10 Gigabit Ethernet (GbE) Short Reach (SR), feature code 3371

For detailed information about OSA-Express3 in an ensemble network, see 7.4, "zBX connectivity" on page 194.

8.3.42 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 connects all resources (CPC and zBX components) of a zEnterprise ensemble node, for management purposes. It is prewired, internally switched, configured, and managed with full redundancy for high availability.

The z114 exploits the OSA-Express3 OSA Direct-Express Management (OSM) CHPID type, which was introduced with the z196. INMN requires two OSA Express3 1000BASE-T ports, from two separate OSA-Express3 1000Base-T features, which are configured as CHPID type OSM. The OSA connection is via the Bulk Power Hub (BPH) port J07 on the z114 to the Top of the Rack (TOR) switches on the zBX.

8.3.43 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 zEnterprise ensemble and is access-controlled using integrated virtual LAN (VLAN) provisioning. No client-managed switches or routers are required. IEDN is managed by the primary HMC that controls the ensemble, helping to reduce the need of a firewall and encryption, and simplifying network configuration and management, with full redundancy for high availability.

The z114 exploits the OSA-Express3 OSA Direct-Express zBX (OSX) CHPID type that was introduced with the z196. The OSA connection is from the z114 to the Top of the Rack (TOR) switches on the zBX. IEDN requires two OSA Express3 10 GbE ports, which are configured as CHPID type OSX.

8.3.44 OSA-Express3 and OSA-Express2 NCP support (OSN)

OSA-Express3 GbE, OSA-Express3 1000BASE-T Ethernet, OSA-Express2 GbE, and OSA-Express2 1000BASE-T Ethernet features can provide channel connectivity from an operating system in a z114 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 LPAR to the LPAR 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 this website:

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/network control program (NCP) operations. The port that is 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[5] and 1000BASE-T features.

Table 8-35 lists the minimum support requirements for OSN.

Table 8-35 Minimum support requirements for OSA-Express3 and OSA-Express2 OSN

Operating system	OSA-Express3 and OSA-Express2 OSN
z/OS	z/OS V1R8
z/VM	z/VM V5R4
z/VSE	z/VSE V4R2
z/TPF	z/TPF V1R1
Linux on System z	Novell SUSE SLES 10 SP3 Red Hat RHEL 5.4

[5] OSA Express2 GbE is withdrawn from marketing.

8.3.45 Integrated Console Controller

The 1000BASE-T Ethernet features provide the Integrated Console Controller (OSA-ICC) function, which supports TN3270E (RFC 2355) and non-SNA distributed function terminal (DFT) 3270 emulation. The OSA-ICC function uses a definition of CHIPD type OSC and the console controller, and has multiple LPAR support, both as shared or spanned channels.

With the OSA-ICC function, 3270 emulation for console session connections is integrated in the z114 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).

8.3.46 VLAN management enhancements

Table 8-36 lists the minimum support requirements for VLAN management enhancements for the OSA-Express3, OSA-Express2, and OSA-Express features (CHPID type OSD).

Table 8-36 Minimum support requirement for VLAN management enhancements

Operating system	Support requirement
z/OS	z/OS V1R8
z/VM	z/VM V5R4. Support of guests is transparent to z/VM if the device is directly connected to the guest (pass through).

8.3.47 GARP VLAN Registration Protocol

All OSA-Express3 and OSA-Express2 features support VLAN prioritization, which is a component of the IEEE 802.1 standard. GARP[6] 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 capability simplifies the network administration and management of VLANs, because manually entering VLAN IDs at the switch is no longer necessary. Minimum support requirements are listed in Table 8-37.

Table 8-37 Minimum support requirements for GVRP

Operating system	Support requirements
z/OS	z/OS V1R8
z/VM	z/VM V5R4

8.3.48 Inbound workload queueing (IWQ) for OSA-Express4S and OSA-Express3

OSA-Express-3 introduced inbound workload queueing (IWQ), which creates multiple input queues and allows OSA to differentiate workloads “off the wire” and then assign work to a

[6] Generic Attribute Registration Protocol

specific input queue (per device) to z/OS. The support is also available with OSA-Express4S. CHPID types OSD and OSX are supported.

With each input queue representing a unique type of workload and 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-Express4S and 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 the contention that is created by other types of workloads. Two types of z/OS workloads are identified and assigned to unique input queues:

- 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-Express4S and OSA-Express3 CHPID types OSD and OSX and the z/OS operating system. IWQ applies to z114, z196, and System z10. The minimum support requirements are listed in Table 8-38.

Table 8-38 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

8.3.49 Inbound workload queueing (IWQ) for Enterprise Extender

Inbound workload queuing (IWQ) for the OSA-Express features has been enhanced to differentiate and separate inbound Enterprise Extender traffic to a new input queue.

IWQ for Enterprise Extender is exclusive to OSA-Express4S and OSA-Express3 CHPID types OSD and OSX and the z/OS operating system. IWQ applies to z114 and z196. The minimum support requirements are listed in Table 8-38.

Table 8-39 Minimum support requirements for IWQ

Operating system	Support requirements
z/OS	z/OS V1R13
z/VM	z/VM V5R4 for guest exploitation only; service required

8.3.50 Query and display OSA configuration

OSA-Express3 introduced the capability for the operating system to directly query and display the current OSA configuration information (similar to OSA/SF). z/OS exploits this OSA capability by introducing a TCP/IP operator command called **`display OSAINFO`**.

Using **`display OSAINFO`** allows the operator to monitor and verify the current OSA configuration, which will help to improve the overall management, serviceability, and usability of OSA-Express4S and OSA-Express3.

The **`display OSAINFO`** command is exclusive to z/OS and applies to OSA-Express4S and OSA-Express3 CHPID types OSD, OSM, and OSX.

8.3.51 Link aggregation support for z/VM

Link aggregation (IEEE 802.3ad), which is controlled by the z/VM Virtual Switch (VSWITCH), allows the dedication of an OSA-Express4S, OSA-Express3, or OSA-Express2 port to the z/VM operating system, when the port is participating in an aggregated group that is configured in Layer 2 mode. Link aggregation (trunking) is designed to allow combining multiple physical OSA-Express4S, 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-Express4S, OSA-Express3, and OSA-Express2 features when configured as CHPID type OSD (QDIO). Link aggregation is supported by z/VM V5R4.

8.3.52 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 connected QDIO 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-Express4S, OSA-Express3, and OSA-Express2 features on z114.

8.3.53 QDIO interface isolation for z/OS

Specific environments require strict controls for routing data traffic between servers or nodes. In certain cases, the LPAR-to-LPAR capability of a shared OSA connection can prevent such controls from being enforced. With interface isolation, internal routing can be controlled on an LPAR basis. When interface isolation is enabled, the OSA will discard any packets destined for a z/OS LPAR that is registered in the OAT as isolated.

QDIO interface isolation is supported by Communications Server for z/OS V1R11 and all OSA-Express4S, OSA-Express3, and OSA-Express2 features on z114.

8.3.54 QDIO optimized latency mode

QDIO optimized latency mode (OLM) can help improve performance for applications that have a critical requirement to minimize response times for inbound and outbound data.

OLM optimizes the interrupt processing in this manner:

- 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-Express4S or OSA-Express3 without requiring additional program controlled interrupts (PCIs).
- For outbound processing, the OSA-Express4S or OSA-Express3 also look 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.

8.3.55 OSA-Express4S checksum offload

OSA-Express4S features, when configured as CHPID type OSD, provide checksum offload for several types of traffic, as indicated on Table 8-40.

Table 8-40 Minimum support requirements for OSA-Express4S checksum offload

Traffic	Support requirements
LPAR-to-LPAR	z/OS V1R12[a] z/VM V5R4 for guest exploitation[b]
IPv6	z/OS V1R13
LPAR-to-LPAR traffic for IPv4 and IPv6	z/OS V1R13

[a] PTFs are required.

[b] Device is directly attached to guest; PTFs are required.

8.3.56 Checksum offload for IPv4 packets when in QDIO mode

A function that is referred to as *checksum offload* supports z/OS and Linux on System z environments. It is offered on the OSA-Express4S GbE, 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 0404, FC 0405, 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.

8.3.57 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, which were 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-Express4S, OSA-Express3, and OSA-Express2 when using CHPID type OSD).

In extending the use of adapter interruptions to OSD (QDIO) channels, the programming overhead to process a traditional I/O interruption is reduced. This overhead reduction benefits OSA-Express TCP/IP support in z/VM, z/VSE, and Linux on System z.

Adapter interruptions apply to all of the OSA-Express4S, OSA-Express3, and OSA-Express2 features on z114 when in QDIO mode (CHPID type OSD).

8.3.58 OSA Dynamic LAN idle

The OSA Dynamic LAN idle parameter change helps reduce latency and improve performance by dynamically adjusting the inbound blocking algorithm. System administrators can authorize the TCP/IP stack to enable a dynamic setting, 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-Express4S, 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-Express4S, OSA-Express3, and OSA-Express2 features on z114 when in QDIO mode (CHPID type OSD). It is exploited by z/OS V1R8 (or higher) with program temporary fixes (PTFs).

8.3.59 OSA Layer 3 Virtual MAC for z/OS environments

To help simplify the infrastructure and to facilitate load balancing when a logical partition shares 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 that are associated with a TCP/IP stack are accessible by using their own VMAC addresses, 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-Express4S, OSA-Express3, and OSA-Express2 features on z114 when in QDIO mode (CHPID type OSD). OSA Layer 3 VMAC is exploited by z/OS V1R8 and later.

8.3.60 QDIO Diagnostic Synchronization

QDIO Diagnostic Synchronization enables system programmers and network administrators to coordinate and simultaneously capture both software and hardware traces. It allows z/OS to signal OSA-Express4S, OSA-Express3, and OSA-Express2 features (by using a diagnostic assist function) to stop traces and capture the current trace records.

QDIO Diagnostic Synchronization is supported by the OSA-Express4S, OSA-Express3, and OSA-Express2 features on z114 when in QDIO mode (CHPID type OSD). QDIO Diagnostic Synchronization is exploited by z/OS V1R8 and later.

8.3.61 Network Traffic Analyzer

With the large volume and complexity of today's network traffic, the z114 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-Express4S, OSA-Express3, and OSA-Express2 features on z114 when in QDIO mode (CHPID type OSD), and it is exploited by z/OS V1R8 and later.

8.3.62 Program directed re-IPL

First available on the System z9, program directed re-IPL allows an operating system on a z114 to re-IPL without operator intervention. This function is supported for both Small Computer System Interface (SCSI) and extended count key data (ECKD™) devices. Table 8-41 lists the minimum support requirements for program directed re-IPL.

Table 8-41 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	V4R2 on SCSI disks

8.3.63 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 4.9, “Parallel Sysplex connectivity” on page 139.

InfiniBand coupling links

Table 8-42 lists the minimum support requirements for coupling links over InfiniBand.

Table 8-42 Minimum support requirements for coupling links over InfiniBand

Operating system	Support requirements
z/OS	z/OS V1R8
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 (6.2 miles)

Support for HCA2-O LR (1xIFB) fanout supporting InfiniBand coupling links at an unrepeated distance of 10 km (6.2 miles) is listed on Table 8-43.

Table 8-43 Minimum support requirements for coupling links over InfiniBand at 10 km (6.2 miles)

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)

8.3.64 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 8-44 lists the minimum support requirements of dynamic I/O support for InfiniBand CHPIDs.

Table 8-44 Minimum support requirements for dynamic I/O support for InfiniBand CHPIDs

Operating system	Support requirement
z/VM	z/VM V5R4

8.4 Cryptographic support

zEnterprise CPC provides two major groups of cryptographic functions:

- Synchronous cryptographic functions, which are provided by the CP Assist for Cryptographic Function (CPACF)
- Asynchronous cryptographic functions, which are 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 the z114, the CP Assist for Cryptographic Function (CPACF) supports the full complement of the 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 166. Support for these functions is provided through a web deliverable for z/OS V1R7 until z/OS V1R9. The support is already included for z/OS V1R10 and higher. Table 8-45 lists the support requirements for recent CPACF enhancements on the z114.

Table 8-45 Support requirements for enhanced CPACF

Operating system	Support requirements
z/OS[a]	z/OS V1R10 and later with the Cryptographic Support for z/OS V1R10-V1R12 web deliverable.
z/VM	z/VM V5R4 with PTFs and higher: Supported for guest use.
z/VSE	z/VSE V4R2 and later: Supports the CPACF features with the functionality supported on IBM System z10.
z/TPF	z/TPF V1R1
Linux on System z	Novell SUSE SLES 11 SP1 Red Hat RHEL 6.1 For Message-Security-Assist-Extension 4 exploitation, IBM is working with its Linux distribution partners to include support in future Linux on System z distribution releases.

[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 and Crypto Express3-1P

Support of Crypto Express3 and Crypto Express3-1P functions varies by operating system and release. Table 8-46 lists the minimum software requirements for the Crypto Express3 and Crypto Express3-1P features when configured as a coprocessor or an accelerator. For a full description, see 6.4, “Crypto Express3” on page 167.

Table 8-46 Crypto Express3 support on z114

Operating system	Crypto Express3
z/OS	z/OS V1R12 (ICSF FMID HCR7770) and higher z/OS V1R9, z/OS V1R10, or z/OS V1R11 with the Cryptographic Support for z/OS V1R9-V1R11 web deliverable. Z/OS V1R8 with toleration APAR OA29329: Crypto Express3 features handled as Crypto Express2 features.
z/VM	z/VM V5R4: Service required; supported for guest use only.
z/VSE	z/VSE V4R2 and IBM TCP/IP for VSE/ESA V1R5 with PTFs.
z/TPF V1R1	Service required (accelerator mode only).
Linux on System z	For toleration: ▸ Novell SUSE SLES10 SP3 and SLES 11. ▸ Red Hat RHEL 5.4 and RHEL 6.0. For exploitation: ▸ Novell SUSE SLES11 SP1. ▸ Red Hat RHEL 6.1.

8.4.3 Web deliverables

For web-deliverable 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 website:

http://www.ibm.com/developerworks/linux/linux390/

8.4.4 z/OS ICSF FMIDs

Integrated Cryptographic Service Facility (ICSF) is a base component of z/OS. It is designed to transparently use the available cryptographic functions, whether CPACF or Crypto Express3 (and Crypto express3-1P), to balance the workload and help address the bandwidth requirements of the applications.

Despite its being a z/OS base component, ICSF’s new functions are generally made available through web deliverable support a couple of months after a new z/OS release is launched. The new functions must relate to an ICSF FMID instead of a z/OS version.

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 8-47 lists the ICSF FMIDs and web deliverable codes for z/OS V1R7 through V1R13. Later FMIDs include the functions of previous FMIDs.

Table 8-47 z/OS ICSF FMIDs

z/OS	ICSF FMID	Web deliverable name	Supported function
V1R7	HCR7720	Included as a z/OS base element	▸ Crypto Express2 support ▸ Support for 64-bit callers ▸ Support for clear DES/TDES key tokens ▸ 2048-bit key RSA operations ▸ 19-digit Personal Account Numbers (PANs)
	HCR7730	Cryptographic Support for z/OS V1R6/R7 and z/OS.e V1R6/R7[a]	▸ Crypto Express2 Accelerator ▸ CPACF AES-128 and SHA-256 support ▸ Support for clear AES key tokens ▸ CKDS Sysplex-wide consistency
	HCR7731	Enhancements to Cryptographic Support for z/OS and z/OS.e V1R6/R7[a]	▸ ATM Remote Key Loading ▸ PKDS Key Management ▸ ISO 16609 CBC Mode TDES MAC ▸ Enhanced PIN Security
	HCR7750	Cryptographic support for z/OS V1R7-V1R9 and z/OS.e V1R7-V1R8[b]	▸ CPACF AES-192 and AES-256 ▸ CPACF SHA-224, SHA-384, and SHA-512 ▸ 4096-bit RSA keys ▸ ISO-3 PIN block format

z/OS	ICSF FMID	Web deliverable name	Supported function
V1R8	HCR7731	Included as a z/OS base element	► ATM Remote Key Loading ► PKDS Key Management ► ISO 16609 CBC Mode TDES MAC ► Enhanced PIN Security
	HCR7750	Cryptographic support for z/OS V1R7-V1R9 and z/OS.e V1R7-V1R8[b]	► CPACF AES-192 and AES-256 ► CPACF SHA-224, SHA-384, and SHA-512 ► 4096-bit RSA keys ► ISO-3 PIN block format
	HCR7751	Cryptographic Support for z/OS V1R8-V1R10 and z/OS.e V1R8[a]	► IBM System z10 BC support ► Secure key AES ► Key store policy ► PKDS Sysplex-wide consistency ► In-storage copy of the PKDS ► 13-digit through 19-digit Personal Account Numbers (PANs) ► Crypto Query service ► Enhanced SAF checking
V1R9	HCR7740	Included as a z/OS base element	► PKCS#11 support ► CFB and PKCS#7 padding
	HCR7750	Cryptographic support for z/OS V1R7-V1R9 and z/OS.e V1R7-V1R8[b]	► CPACF AES-192 and AES-256 ► CPACF SHA-224, SHA-384, and SHA-512 ► 4096-bit RSA keys ► ISO-3 PIN block format
	HCR7751	Cryptographic Support for z/OS V1R8-V1R10 and z/OS.e V1R8[a]	► IBM System z10 BC support ► Secure key AES ► Key store policy ► PKDS Sysplex-wide consistency ► In-storage copy of the PKDS ► 13-digit through 19-digit Personal Account Numbers (PANs) ► Crypto Query service ► Enhanced SAF checking
	HCR7770	Cryptographic support for z/OS V1R9-V1R11	► Crypto Express3 and Crypto Express3-1P support ► PKA Key Management Extensions ► CPACF Protected Key ► Extended PKCS#11 ► ICSF Restructure (Performance, RAS, ICSF-CICS Attach Facility)

z/OS	ICSF FMID	Web deliverable name	Supported function
V1R10	HCR7750	Included as a z/OS base element	► 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[a]	► IBM System z10 BC support ► Secure key AES ► Key store policy ► PKDS Sysplex-wide consistency ► In-storage copy of the PKDS ► 13-digit through 19-digit Personal Account Numbers (PANs) ► Crypto Query service ► Enhanced SAF checking
	HCR7770	Cryptographic support for z/OS V1R9-V1R11	► Crypto Express3 and Crypto Express3-1P support ► PKA Key Management Extensions ► CPACF Protected Key ► Extended PKCS#11 ► ICSF Restructure (Performance, RAS, ICSF-CICS Attach Facility)
	HCR7780	Cryptographic support for z/OS V1R10-V1R12	► IBM zEnterprise 196 support ► Elliptic Curve Cryptography ► Message-Security-Assist-4 ► HMAC Support ► ANSI X9.8 Pin ► ANSI X9.24 (CBC Key Wrapping) ► CKDS constraint relief ► PCI Audit ► All callable services AMODE(64) ► PKA RSA OAEP with SHA-256 algorithm[c]

z/OS	ICSF FMID	Web deliverable name	Supported function
V1R11	HCR7751	Included as a z/OS base element	- IBM System z10 BC support - Secure key AES - Key store policy - PKDS Sysplex-wide consistency - In-storage copy of the PKDS - 13-digit through 19-digit Personal Account Numbers (PANs) - Crypto Query service - Enhanced SAF checking
	HCR7770	Cryptographic support for z/OS V1R9-V1R11	- Crypto Express3 and Crypto Express3-1P support - PKA Key Management Extensions - CPACF Protected Key - Extended PKCS#11 - ICSF Restructure (Performance, RAS, ICSF-CICS Attach Facility)
	HCR7780	Cryptographic support for z/OS V1R10-V1R12	- IBM zEnterprise 196 support - Elliptic Curve Cryptography - Message-Security-Assist-4 - HMAC Support - ANSI X9.8 Pin - ANSI X9.24 (CBC Key Wrapping) - CKDS constraint relief - PCI Audit - All callable services AMODE(64) - PKA RSA OAEP with SHA-256 algorithm[c]
	HCR7790	Cryptographic Support for z/OS V1R11-V1R13[d]	- Expanded key support for AES algorithm - Enhanced ANSI TR-31 - PIN block decimalization table protection - Elliptic Curve Diffie-Hellman (ECDH) algorithm - RSA in the Modulus Exponent (ME) and Chinese Remainder Theorem (CRT) formats

z/OS	ICSF FMID	Web deliverable name	Supported function
V1R12	HCR7770	Included as a z/OS base element	► Crypto Express3 and Crypto Express3-1P support ► PKA Key Management Extensions ► CPACF Protected Key ► Extended PKCS#11 ► ICSF Restructure (Performance, RAS, ICSF-CICS Attach Facility)
	HCR7780	Cryptographic support for z/OS V1R10-V1R12	► IBM zEnterprise 114 support ► Elliptic Curve Cryptography ► Message-Security-Assist-4 ► HMAC Support ► ANSI X9.8 Pin ► ANSI X9.24 (CBC Key Wrapping) ► CKDS constraint relief ► PCI Audit ► All callable services AMODE(64) ► PKA RSA OAEP with SHA-256 algorithm[c]
	HCR7790	Cryptographic Support for z/OS V1R11-V1R13[d]	► Expanded key support for AES algorithm ► Enhanced ANSI TR-31 ► PIN block decimalization table protection ► Elliptic Curve Diffie-Hellman (ECDH) algorithm ► RSA in the Modulus Exponent (ME) and Chinese Remainder Theorem (CRT) formats
V1R13	HCR7780	Included as a z/OS base element	► IBM zEnterprise 196 support ► Elliptic Curve Cryptography ► Message-Security-Assist-4 ► HMAC Support ► ANSI X9.8 Pin ► ANSI X9.24 (CBC Key Wrapping) ► CKDS constraint relief ► PCI Audit ► All callable services AMODE(64) ► PKA RSA OAEP with SHA-256 algorithm[c]
	HCR7790	Cryptographic Support for z/OS V1R11-V1R13[d]	► Expanded key support for AES algorithm ► Enhanced ANSI TR-31 ► PIN block decimalization table protection ► Elliptic Curve Diffie-Hellman (ECDH) algorithm ► RSA in the Modulus Exponent (ME) and Chinese Remainder Theorem (CRT) formats

[a] Download is no longer available and has been replaced by the Cryptographic Support for z/OS V1R10-V1R12 web deliverable.

[b] This download is installable on V1R6 and V1R7, but not supported. Clients running z/OS V1R9 or later need to install the Cryptographic Support for z/OS V1R10-V1R12 web deliverable.

c. Service required
d. Download is planned to be available during 2H2011.

8.4.5 ICSF migration considerations

Consider the following points about the web-deliverable ICSF code:

- A new PKDS file is required in order to allow 4096-bit RSA keys to be stored. This support requires a new PKDS to be created to allow for the larger keys. Check the new define cluster parameters in the *z/OS ICSF System Programmer's Guide*.

  The existing PKDS needs to be copied into the newly allocated data set. A toleration APAR OA21807 is required for the 4096-bit support on ICSF below HCR7750 to support mixed environments.
- 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 z114 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 114 introduces the latest System z technology. Although support is provided by z/OS starting with z/OS V1R8, the exploitation of the z114 depends on the z/OS release. The z/OS.e is *not* supported on the z114.

In general, we have the following recommendations:

- Do not migrate software releases and hardware at the same time.
- Keep the members of the sysplex at the same software level, except during brief migration periods.
- Migrate to a Server Time Protocol (STP) or Mixed-Coordinated Timing Network (CTN) network prior to introducing a z114 or z196 into a sysplex.
- Review z114 restrictions and migration considerations prior to creating an upgrade plan.

8.5.2 HCD

On z/OS V1R8 and higher, the HCD or the Hardware Configuration Manager (HCM) assist in defining a configuration for z114.

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:

```
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 V1R8 (requires PTFs for zIIP support) through z/OS V1R13.

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 might 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, readjust the new policies, and repeat the cycle, as needed. To change WLM policies, turning HiperDispatch off and then on is not necessary.

A health check is provided to verify whether HiperDispatch is enabled on a system image that is running on the z114.

8.5.6 Capacity Provisioning Manager

Installation of the capacity provision function on z/OS requires these tasks:

- Setting up and customizing z/OS Resource Measurement Facility (RMF), including the Distributed Data Server (DDS)
- Setting up the z/OS Common Information Model (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

The following requirements are necessary to exploit the capacity provisioning function:

- 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 systems 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 312.
- Runtime systems. These systems 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 that are provided by the compiler require a minimum architecture level. ARCH(8) and ARCH(9) exploit instructions available respectively on the z10 EC and z114.
- The TUNE option, which allows optimization of the application for a specific machine architecture, within the constraints that are imposed by the ARCHITECTURE option. The TUNE level must not be lower than the setting in the ARCHITECTURE option.

For more information about the ARCHITECTURE and TUNE compiler options, see the *z/OS V1R9.0 XL C/C++ User’s Guide*, SC09-4767.

> **ARCHITECTURE and TUNE options:** Use the ARCHITECTURE or TUNE option for C++ programs if the same applications need to run on both the z114 as well as on previous System z servers. However, if C++ applications will only run on z114 servers, use the latest ARCHITECTURE and TUNE options to assure 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 z114 is supported on the z196, z10EC, z10 BC, z9 EC, z9 BC, or another z114. The LPAR running the Coupling Facility Control Code (CFCC) can reside on any of these supported servers. See Table 8-48 on page 263 for Coupling Facility Control Code requirements for supported servers.

> **Coupling link connectivity:** Coupling link connectivity is *not* supported to z890, z990, and previous servers, which might affect the introduction of the z114 into existing installations and require additional planning. Also, consider the level of CFCC. For more information, see “Coupling link migration considerations” on page 144.

The initial support of the CFCC on the z114 is Level 17. CFCC Level 17 is available and is exclusive to z114 and z196. CFCC Level 17 offers the following enhancements:

- Availability improvements with nondisruptive CFCC dumps

  The coupling facility is now designed to collect a serialized, time-consistent dump without disrupting CFCC operations, which 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 defined and used in a sysplex permits more discrete data sharing groups to operate concurrently. Allowing more CF structures to be defined and used in a sysplex can help environments requiring the definitions of many structures, such as to support SAP or service providers.

  > **CFRM CDS:** Having more than 1024 structures requires a new version of the Coupling Facility Resource Manager (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 a sysplex IPL.*

- Increased number of lock and list structure connectors

  The z114 supports 247 connectors to a lock structure and 127 connectors to a list structure, up from 32. This increase 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 connectors to a cache structure for several years.

Enhancements that were available with the 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 the CF processing of the duplexed structure request. CFCC Level 16 allows one of these signals to be asynchronous to CF processing. This change 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 connector to respond reads the new message, but when the other connectors read, they 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, the other connectors are notified as they are prior to CFCC Level 16.

CFCC Level 17 requires an LPAR with 512 MB. As storage requirements might increase when moving to CFCC Level 17, we strongly recommend using the CFSizer Tool, which is located on this website:

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

z114 servers with CFCC Level 17 require z/OS V1R8 or later, and z/VM V5R4 or later for guest virtual coupling.

The current CFCC Level for z114 servers is CFCC Level 17. See Table 8-48 on page 263. To support migration from one CFCC level to the next, separate levels of CFCC can be run concurrently while the coupling facility logical partitions are running on separate servers (CF logical partitions running on the same server share the same CFCC level).

Table 8-48 System z CFCC code level considerations

z114 and 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, the installation of compatibility and 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 website:

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 that is 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.

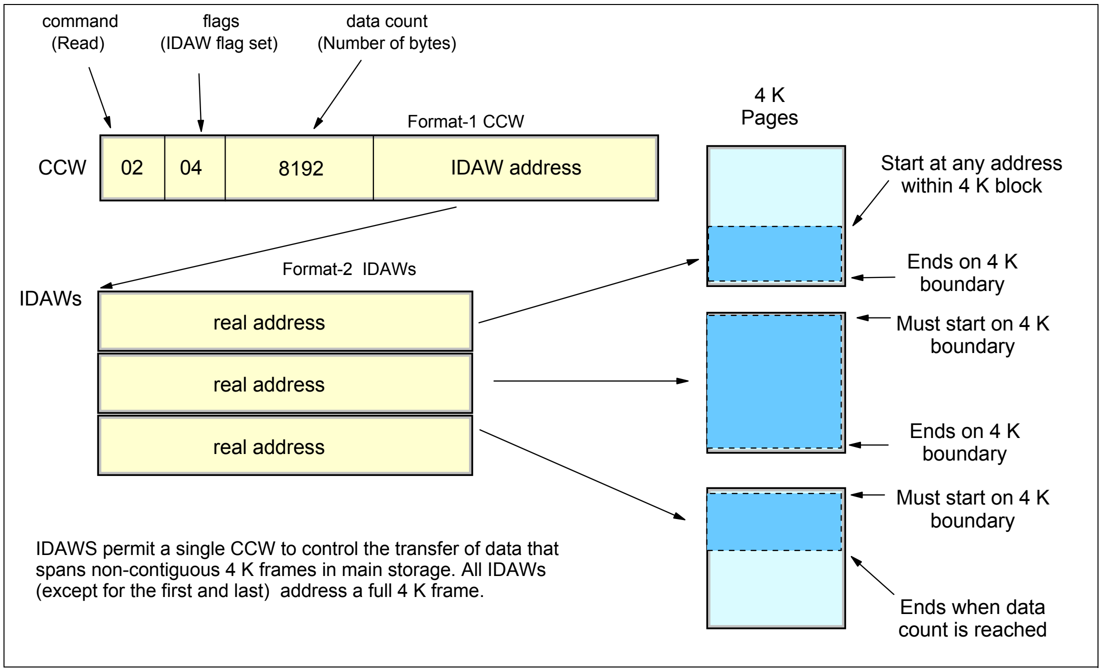
IBM laboratory tests indicate that applications using extended format (EF) data sets, such as DB2, or long chains of small blocks can gain significant performance benefits by using the MIDAW facility.

MIDAW is supported on ESCON channels that are configured as CHPID type CNS and on FICON channels that are 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.[7] 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 264 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.

[7] 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.



IDAWS permit a single CCW to control the transfer of data that spans non-contiguous 4 K frames in main storage. All IDAWs (except for the first and last) address a full 4 K frame.

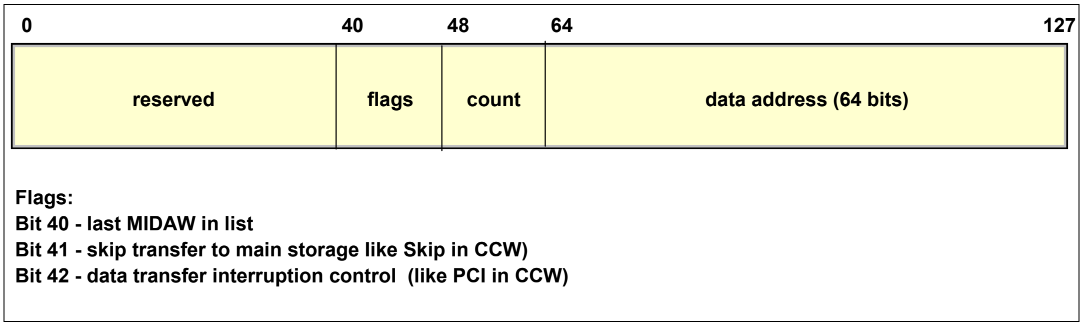
Figure 8-1 IDAW usage

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* can be used to process I/O data blocks that have a more complex internal structure, in which portions of the data block are directed into separate buffer areas (which 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.



Flags:
Bit 40 - last MIDAW in list
Bit 41 - skip transfer to main storage like Skip in CCW)
Bit 42 - data transfer interruption control (like PCI in CCW)

Figure 8-2 MIDAW format

An example of MIDAW usage is shown in Figure 8-3.

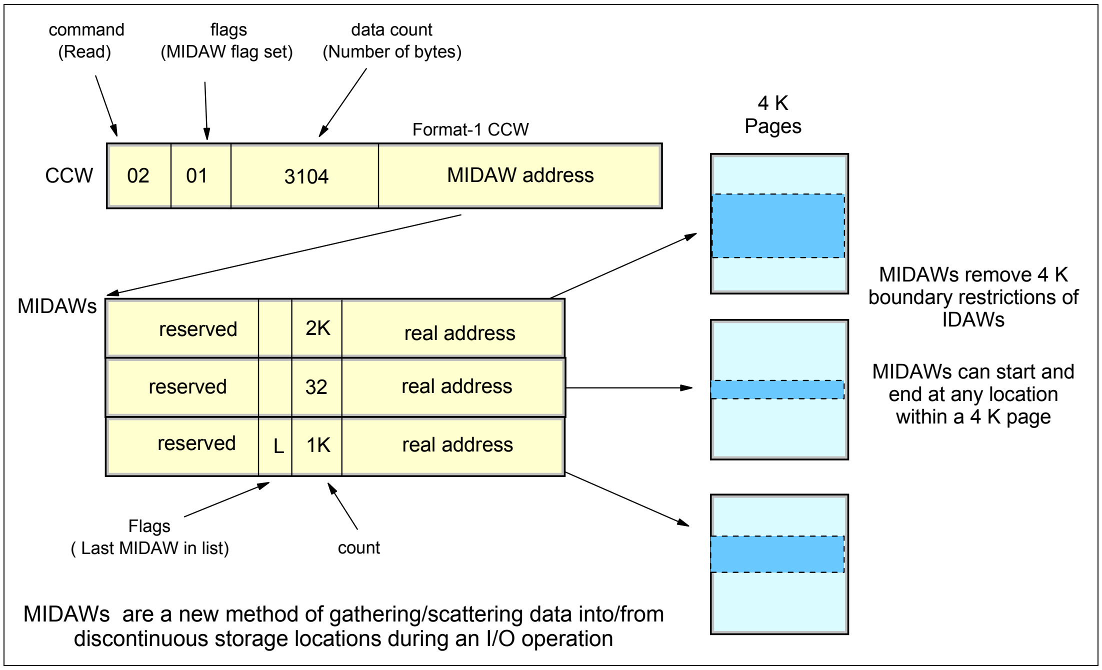

command
(Read)
flags
(MIDAW flag set)
data count
(Number of bytes)
Format-1 CCW
4 K
Pages
CCW
02
01
3104
MIDAW address
MIDAWs
reserved
2K
real address
reserved
32
real address
reserved
L
1K
real address
MIDAWs remove 4 K boundary restrictions of IDAWs
MIDAWs can start and end at any location within a 4 K page
Flags
(Last MIDAW in list)
count
MIDAWs are a new method of gathering/scattering data into/from discontinuous storage locations during an I/O operation

Figure 8-3 MIDAW usage

The use of MIDAWs is indicated by the MIDAW bit in the CCW. If this bit is set, the *skip flag* cannot be set in the CCW. The skip flag in the MIDAW can be used instead. The data count in the CCW must equal the sum of the data counts in the MIDAWs. The CCW operation ends when the CCW count goes to zero or the last MIDAW (with the *last* flag) ends. The combination of the address and count in a MIDAW cannot cross a page boundary. Therefore, the largest possible count is 4 K. The maximum data count of all the MIDAWs in a list cannot exceed 64 K, which is the maximum count of the associated CCW.

The scatter-read or scatter-write effect of the MIDAWs makes it possible to efficiently send small control blocks embedded in a disk record to separate buffers from those buffers that are used for larger data areas within the record. MIDAW operations are on a single I/O block, in the manner of data chaining. Do not confuse this operation with CCW *command* chaining.

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. Therefore, CCW data chaining is required, which produces less than optimal I/O performance. Because the most significant performance benefit of MIDAWs is achieved with extended format (EF) data sets, we include a brief review of the EF data sets 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 that are 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 normally requires at least two CCWs with data chaining. One CCW is used for the 32-byte suffix of the EF data set. With MIDAW, the additional CCW for the EF data set suffix 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 chains 12 CCWs together using data chaining. To read twelve 4 K records from an EF data set, 24 CCWs are chained (two CCWs per 4 K record). Using Media Manager track-level command operations and MIDAWs, an entire track can be transferred using a single CCW.

8.7.3 Performance benefits

z/OS Media Manager has the I/O channel program's support for implementing EF 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 *can* construct channel programs containing MIDAWs, provided that they construct an IOBE with the IOBEMIDA bit set. Users of the EXCP instruction *cannot* 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 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 with 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 website 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 the I/O configuration program (IOCP) for z114 is V2R1L0 (IOCP 2.1.0) or later.

8.9 Worldwide port name (WWPN) tool

A part of the installation of your z114 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 the worldwide port name (WWPN) 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 after the server is installed. In addition, the SAN configuration can be retained instead of altered by assigning the WWPN to physical FCP ports when a FICON feature is replaced.

The WWPN tool takes a `.csv` file 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 it can be exported from the Hardware Configuration Definition/Hardware Configuration Manager (HCD/HCM).

The WWPN tool on z114 (CHPID type FCP) requires these levels:

- z/OS V1R8, V1R9, V1R10, V1R11, with PTFs, or V1R12
- z/VM V5R4 or V6R1, with PTFs

The WWPN tool is available for download at Resource Link and is applicable to all FICON channels that are defined as CHPID type FCP (for communication with SCSI devices) on z114:

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 z114 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 z114.

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 and solution specific.

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 z114 environment.

The blades are virtualized by the PowerVM Enterprise Edition, and their LPARs run either AIX Version 5 Release 3 TL12 (POWER6® mode), AIX Version 6 Release 1 TL5 (POWER7 mode), AIX Version 7 Release 1, or subsequent releases. Applications that are supported on AIX can be deployed to blades.

Also offered are selected IBM System z HX5 blades. Virtualization is provided by an integrated hypervisor, using Kernel-based virtual machines, and supporting Linux on System x.

Statement of General Direction: IBM intends to support running the Microsoft Windows operating system on select IBM BladeCenter HX5 blades that are installed in the IBM zEnterprise BladeCenter Extension (zBX) Model 002.

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

IBM WebSphere DataPower Integration Appliance XI50 for zEnterprise

The IBM WebSphere DataPower Integration Appliance XI50 for zEnterprise (DataPower XI50z) is the latest addition to the zEnterprise integrated infrastructure. The DataPower XI50z is a double-wide blade.

The DataPower XI50z is a multifunctional appliance that can help provide multiple levels of XML optimization, streamline and secure valuable service-oriented architecture (SOA) applications, and provide drop-in integration for heterogeneous environments by enabling core enterprise service bus (ESB) functionality, including routing, bridging, transformation, and event handling. It can help to simplify, govern, and enhance the network security for XML and web services.

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 queries 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 consists 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 z114 Service Element, using the product DVD. When 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 the z114 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 website:

http://www.ibm.com/developerworks/spaces/optim?pageid=649&S_TACT=105AGX01&S_CMP

A client-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 zEnterprise CPC’s IBM software portfolio includes operating system software[8] (that is, 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 metric groups for software licensing are available from IBM, depending on the software product:

- Monthly License Charge (MLC)
- International Program License Agreement (IPLA)

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, in turn, include a variety of offerings.

IPLA metrics have a single, up-front charge for an entitlement to use the product. An optional and separate annual charge called *subscription and support* entitles clients to access IBM product support during the support period and also receive future releases and versions at no additional charge.

[8] Linux on System z distributions are not IBM products.

For details, consult these resources:

- *IBM System z Software Pricing Reference Guide* web page:

  `http://public.dhe.ibm.com/common/ssi/ecm/en/zso01378usen/ZSO01378USEN.PDF`

- IBM System z Software Pricing web pages:

  `http://www.ibm.com/systems/z/resources/swprice/mlc/index.html`

8.12.1 MLC pricing metrics

MLC pricing applies to z/OS, z/VSE, or z/TPF operating systems. Any mix of z/OS, z/VM, Linux, z/VSE, and z/TPF images is allowed. Charges are based on processor capacity, which is measured in Millions of Service Units (MSU) per hour.

A variety of WLC pricing structures supports two charge models:

- Variable charges (several pricing metrics)

  Variable charges apply to products, such as z/OS, z/VSE, z/TPF, DB2, IMS, CICS, MQSeries®, and Lotus® Domino®. There are several pricing metrics employing the following charge types:

  – Full-capacity

    The CPC's total number of MSUs is used for charging. Full-capacity is applicable when the client's CPC is not eligible for sub-capacity.

  – Sub-capacity

    Software charges are based on the utilization of the LPARs where the product is running.

- Flat charges

  Software products that are licensed under flat charges are not eligible for sub-capacity pricing. There is a single charge per CPC on the z196. On the z114, the price is based on the CPC size.

Sub-capacity

Sub-capacity allows, for eligible programs, software charges that are based on the utilization of LPARs instead of the CPC's total number of MSUs. Sub-capacity removes the dependency between software charges and CPC (hardware) installed capacity.

Sub-capacity is based on the highest observed rolling 4-hour average utilization for the LPARs. It is not based on the utilization of each product, but on the utilization of the LPARs where the product runs (with the exception of products licensed using the select application license charge (SALC) pricing metric).

The sub-capacity licensed products are charged monthly based on the highest observed rolling 4-hour average utilization of the LPARs in which the product runs. This type of charge requires measuring the utilization and reporting it to IBM.

The LPAR's rolling 4-hour average utilization can be limited by a defined capacity value on the partition's image profile. This defined capacity value activates the soft capping function of PR/SM, limiting the rolling 4-hour average partition utilization to the defined capacity value. Soft capping controls the maximum rolling 4-hour average usage (the last 4-hour average value at every 5-minute interval), but it does not control the maximum instantaneous partition use.

Also available is an LPAR group capacity limit, which allows you to set soft capping by PR/SM for a group of logical partitions running z/OS.

Even using the soft capping option, the partition's use can reach up to 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.

Certain pricing metrics apply to stand-alone System z servers, and others apply to the aggregation of multiple System z servers' workloads within the same Parallel Sysplex.

For further information about WLC and details about how to combine logical partitions' utilization, see the publication *z/OS Planning for Workload License Charges*, SA22-7506, which is available from the following web page:

http://www-03.ibm.com/systems/z/os/zos/bkserv/find_books.html

The following metrics apply to a stand-alone IBM zEnterprise 114:

- Advanced Entry Workload License Charges (AEWLC)
- System z New Application License Charges (zNALC)
- Parallel Sysplex License Charges (PSLC)

The following metrics apply to an IBM zEnterprise 114 on an actively coupled Parallel Sysplex:

- Advanced Workload License Charges (AWLC), when all CPCs are z114 or z196
  - Variable Workload License Charges (VWLC) are only allowed under the AWLC Transition Charges for Sysplexes when not all CPCs are z196 or z114
- System z New Application License Charges (zNALC)
- Parallel Sysplex License Charges (PSLC)

8.12.2 Advanced Workload License Charges (AWLC)

Advanced Workload License Charges were introduced with the IBM zEnterprise 196. They use the measuring and reporting mechanisms, as well as the existing MSU tiers, from VWLC.

As compared to VWLC, the prices per tier have been lowered and prices for tiers 4, 5, and 6 e differ, allowing for lower costs for charges higher than 875 MSU. AWLC offers improved price performance as compared to PSLC.

Similarly to Workload Licence Charges, AWLC can be implemented in full-capacity or sub-capacity mode. AWLC applies to z/OS and z/TPF and their associated middleware products, such as DB2, IMS, CICS, MQSeries, and Lotus Domino, when running on a z114.

For additional information, see the AWLC web page:

http://www-03.ibm.com/systems/z/resources/swprice/mlc/awlc.html

8.12.3 Advanced Entry Workload License Charges (AEWLC)

Advanced Entry Workload License Charges were introduced with the IBM zEnterprise 196. They use the measuring and reporting mechanisms, as well as the existing MSU tiers, from the Entry Workload License Charges (EWLC) pricing metric and the Midrange Workload License Charges (MWLC) pricing metric.

As compared to EWLC and MWLC, the software price performance has been extended. AEWLC also offers improved price performance as compared to PSLC.

Similarly to Workload License Charges, AEWLC can be implemented in full-capacity or sub-capacity mode. AEWLC applies to z/OS and z/TPF and their associated middleware products, such as DB2, IMS, CICS, MQSeries, and Lotus Domino, when running on a z114.

For additional information, see the AEWLC web page:

http://www-03.ibm.com/systems/z/resources/swprice/mlc/aewlc.html

8.12.4 System z New Application License Charges (zNALC)

System z New Application License Charges offers a reduced price for the z/OS operating system on LPARs 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 that is 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 page:

http://www-03.ibm.com/systems/z/resources/swprice/mlc/znalc.html

8.12.5 Select Application License Charges (SALC)

Select Application License Charges applies to WebSphere MQ for System z only. It allows a WLC client to license MQ under product use rather than the sub-capacity pricing that is provided under WLC.

WebSphere MQ is typically a low-usage product that runs pervasively throughout the environment. Clients who run WebSphere MQ at a low usage can benefit from SALC. Alternatively, you can still choose to license WebSphere MQ under the same metric as the z/OS software stack.

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. The program determines the highest daily usage of a program family, which is the highest of 24 hourly measurements recorded each day. The program refers to all active versions of MQ.
2. The program determines the monthly usage of a program family, which is the fourth highest daily measurement recorded for a month.
3. The program uses the highest monthly usage determined for the next billing period.

For additional information about SALC, see the Other MLC Metrics web page:

http://www.ibm.com/systems/z/resources/swprice/mlc/other.html

8.12.6 Midrange Workload License Charges (MWLC)

Midrange Workload License Charges (MWLC) applies to z/VSE V4 when running on z196, System z10, and z9 servers. The exceptions are the z10 BC and z9 BC servers at capacity setting A01, to which zELC applies.

Similarly to Workload License Charges, MWLC can be implemented in full-capacity or sub-capacity 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.

To report usage, the sub-capacity report tool is used. One SCRT report per server is required.

For additional information, see the MWLC web page:

http://www.ibm.com/systems/z/resources/swprice/mlc/mwlc.html

8.12.7 Parallel Sysplex License Charges (PWLC)

Parallel Sysplex License Charges (PSLC) applies to a large range of mainframe servers. The list can be obtained from this web page:

http://www-03.ibm.com/systems/z/resources/swprice/reference/exhibits/hardware.html

Although it can be applied to stand-alone CPCs, the metric only provides aggregation benefits when applied to a group of CPCs in an actively coupled Parallel Sysplex cluster according to IBM’s terms and conditions.

Aggregation allows charging a product based on the total MSU value of the machines where the product executes (as opposed to all the machines in the cluster). In an uncoupled environment, software charges are based on the MSU capacity of the machine.

For additional information, see the PSLC web page:

http://www.ibm.com/systems/z/resources/swprice/mlc/pslc.html

8.12.8 System z International Program License Agreement (IPLA)

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 z/OS products
- Linux middleware products
- z/VM Versions 5 and 6

Generally, three pricing metrics apply to IPLA products for System z:

- Value unit (VU) pricing applies to the IPLA products that run on z/OS. Value unit pricing is typically based on the number of MSUs and allows for a lower cost of incremental growth. Examples of eligible products are IMS tools, CICS tools, DB2 tools, application development tools, and WebSphere products for z/OS.

- Engine-based value unit (EBVU) pricing enables a lower cost of incremental growth with additional engine-based licenses purchased. Examples of eligible products include z/VM V5 and v6, and certain z/VM middleware products, which are priced based on the number of engines.
- Processor value units (PVU). The number of engines is converted into processor value units under the Passport Advantage® terms and conditions. Most Linux middleware is also priced based on the number of engines.

For additional information, see the System z IPLA web page:

http://www.ibm.com/systems/z/resources/swprice/zipla/index.html

8.13 References

For the most current planning information, see the support website 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/VSE

  http://www.ibm.com/servers/eserver/zseries/zvse/support/preventive.html

- z/TPF

  http://www.ibm.com/software/htp/tpf/pages/maint.htm

- Linux on System z

  http://www.ibm.com/systems/z/os/linux/

9

System upgrades

This chapter provides an overview of zEnterprise 114 upgrade capabilities and procedures, with an emphasis on Capacity on Demand offerings.

The upgrade offerings to the z114 CPC have been developed from previous IBM System z servers. In response to client demands and changes in market requirements, a number of features have been added. The changes and additions are designed to provide increased client control over the capacity upgrade offerings with decreased administrative work and with enhanced flexibility. The provisioning environment gives the client an unprecedented flexibility and a finer control over cost and value.

For detailed tutorials on all aspects of system upgrades, see Resource Link - Customer Initiated Upgrade Information, then select Education, and a list of available servers will help you select your particular product:

`https://www-304.ibm.com/servers/resourcelink/hom03010.nsf/pages/CIUInformation?OpenDocument`

Registration is required to access Resource Link.

Given today's business environment, the benefits of the growth capabilities that are provided by the z114 are plentiful, and include, but are not limited to these benefits:

- 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 lockdown periods
- Enabling planned downtime changes without availability impacts

This chapter discusses the following topics:

- “Capacity Backup” on page 308
- “Nondisruptive upgrades” on page 312
- “Summary of Capacity on Demand offerings” on page 317

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

9.1 Upgrade types

We summarize the types of upgrades for the z114 in this section.

Permanent and temporary upgrades

In various situations, separate types of upgrades are needed. After a certain amount of 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 lost capacity on a server that is down during a disaster or data center maintenance. The z114 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 the installation of Licensed Install Code 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 billable capacity offering is On/Off Capacity on Demand (On/Off CoD). The two replacement capacity offerings that are available are Capacity Backup (CBU) and Capacity for Planned Event (CPE).

For descriptions, see 9.1.1, “Terminology related to CoD for the System z114 CPC” on page 278.

> **MES:** The MES provides a system upgrade that can result in more enabled processors and a separate CP capacity level, but also in a second processor drawer, 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. You order an MES through your IBM representative, and the MES is delivered by IBM service personnel.

Concurrent and nondisruptive upgrades

Depending on the impact on system and application availability, upgrades can be classified in one of these ways:

- 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 281.
- Non-concurrent

  This type of upgrade requires stopping the hardware system. Examples of these upgrades include a model upgrade from an M05 model to the M10 model, and physical memory capacity upgrades.

- Disruptive

  An upgrade is disruptive when the resources that are added to an operating system image require that the operating system is recycled to configure the newly added resources.

- Nondisruptive

  Nondisruptive upgrades do not require that you restart the software that is running or the operating system for the upgrade to take 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 312.

9.1.1 Terminology related to CoD for the System z114 CPC

Table 9-1 briefly describes the most frequently used terms that relate to Capacity on Demand for z114 CPCs.

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.
Capacity	Hardware resources (processor and memory) that are able to process the workload that 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 activates 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 z114 CPC, capacity levels for the CP engines are from A to Z: - Full-capacity CP engine is indicated by Z. - Subcapacity CP engines are indicated by A to Y.
Capacity setting	Derived from the capacity level and the number of processors. For the z114 CPC, the capacity levels are from A01 to Z05, where the last digit indicates the number of active CPs, and the letter from A to Z indicates the processor capacity.
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 can request processor and memory upgrades by using the IBM Resource Link and the system’s Remote Support Facility (RSF) connection.
Capacity on Demand (CoD)	The ability of a computing system to increase or decrease its performance capacity as needed to meet fluctuations in demand.

Term	Description
Capacity Provisioning Manager (CPM)	As a component of z/OS Capacity Provisioning, CPM monitors business-critical workloads that are running on z/OS systems on z114 CPCs.
Customer profile	This information resides on Resource Link and contains client and machine information. A client profile can contain information about more than one machine.
Full capacity CP feature	For z114, capacity settings Z*xx* are full capacity settings.
High water mark	Capacity purchased and owned by the client.
Installed record	The LICCC record has been downloaded, staged to the Support Element (SE), and is now installed on the CPC. A maximum of eight 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 that is 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 a server upgrade without hardware changes by enabling the activation of additional previously installed capacity.
Miscellaneous equipment specification (MES)	An upgrade process initiated through an IBM representative and installed by IBM personnel.
Model capacity identifier (MCI)	Shows the current active capacity on the server, including all replacement and billable capacity. For the z114, the model capacity identifier is in the form of A*xx* to Z*xx* where *xx* indicates the number of active CPs: ▸ *xx* can have a range of 01-05.
Model Permanent Capacity Identifier (MPCI)	Keeps information about capacity settings active before any temporary capacity was activated.
Model Temporary Capacity Identifier (MTCI)	Reflects the permanent capacity with billable capacity only, without replacement capacity. If no billable temporary capacity is active, Model Temporary Capacity Identifier equals Model Permanent Capacity Identifier.
On/Off Capacity on Demand (CoD)	Represents a function that allows a spare capacity in a CPC to be made available to increase the total capacity of a CPC. For example, On/Off CoD can be used to acquire additional capacity for the purpose of handling a workload peak.
Permanent capacity	The capacity that a client 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.
Processor drawer	Packaging technology that contains the SCMs for the PU and SC chips as well as memory and connections to I/O and coupling links.
Purchased capacity	Capacity delivered to and owned by the client. It can be higher than permanent capacity.
Permanent/Temporary entitlement record	The internal representation of a temporary (TER) or permanent (PER) capacity upgrade processed by the CIU facility. An entitlement record contains the encrypted representation of the upgrade configuration with the associated time limit conditions.
Replacement capacity	A temporary capacity that is 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.

Term	Description
Resource Link	IBM Resource Link is a technical support website that is 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 client, 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.
Single Chip Module (SCM)	The packaging technology that is used to hold the processor units (PUs) and SC chips.
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 z114, CP features A01 to Y05 represent subcapacity configurations, and CP features Z01 to Z05 represent full capacity configurations.
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 client.
Vital product data (VPD)	Information that uniquely defines the system, hardware, software, and microcode elements of a processing system.

9.1.2 Permanent upgrades

Permanent upgrades can be ordered through an IBM marketing representative or initiated by the client with the Customer Initiated Upgrade (CIU) on IBM Resource Link.

CIU: 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 perform these tasks:

- Add a processor drawer.
- Add I/O drawers and features.
- Add model capacity.
- Add specialty engines.
- Add memory.
- Activate unassigned model capacity or Integrated Facility for Linux (IFL) processors.
- Deactivate activated model capacity or IFLs.
- Activate channels.
- Activate cryptographic engines.
- Change specialty engine (re-characterization).
- Add zEnterprise BladeCenter Extension (zBX) and zBX features:
  - Chassis
  - Racks
  - Blades
  - Entitlements

Important: Most of the MESs can be concurrently applied without disrupting the existing workload (see 9.2, “Concurrent upgrades” on page 281 for details). However, certain MES changes are disruptive (for example, an upgrade of model M05 to M10, or adding memory).

Permanent upgrades initiated through CIU on IBM Resource Link

Ordering a permanent upgrade by using the CIU application through the IBM Resource Link allows you to add capacity to fit within your existing hardware:

- Add model capacity
- Add specialty engines
- Add memory
- Activate unassigned model capacity or IFLs
- Deactivate activated model capacity or IFLs

9.1.3 Temporary upgrades

System z114 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 the 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 marketing representative. Temporary upgrades capacity changes can be billable or replacement capacity.

Billable capacity

To handle a peak workload, processors can be rented temporarily on a daily basis. You can activate up to twice the purchased capacity of any processor unit (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 the authorized limit.

Two replacement capacity offerings exist:

- Capacity Backup
- Capacity for Planned Event

9.2 Concurrent upgrades

Concurrent upgrades on the z114 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 that are provided by the z114 are plentiful, and include, but are not limited to these benefits:

- Enabling the 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 *lockdown* periods
- Enabling planned-downtime changes without affecting availability

These capabilities are based on the flexibility of the design and structure, which allows concurrent hardware installation and Licensed Internal Code (LIC) control over the configuration.

Subcapacity models provide for a CP capacity increase in two dimensions that can be used together to deliver configuration granularity. The first dimension is by adding CPs to the configuration, and the second dimension is by changing the capacity setting of the CPs currently installed to a higher model capacity identifier. In addition, a capacity increase can be delivered by increasing the CP capacity setting and at the same time decrease the number of active CPs.

The z114 allows the concurrent addition of processors to a running logical partition. As a result, you can have a flexible infrastructure, in which you can add capacity without pre-planning. This function is supported by z/VM. Planning ahead is required for z/OS logical partitions (LPARs). To be able to add processors to a running z/OS, reserved processors must be specified in the LPAR'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 remapped on all SAPs on the z114 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 z114 resources to satisfy the application execution instead of using the zBX infrastructure.

9.2.1 Model upgrades

The z114 has the following machine type and model, and model capacity identifiers:

- Machine type and model is 2818-M*vv*.

  The *vv* can be 05 or 10. The model number indicates how many PUs (*vv*) are available for customer characterization. Model M05 has one processor drawer installed, and model M10 contains two processor drawers.

- Model capacity identifiers (MCI) are A01 to Z05.

  The model capacity identifier describes how many CPs are characterized (01 to 05) and the capacity setting (A to Z) of the CPs.

A hardware configuration upgrade always requires additional physical hardware (processor or I/O drawers, or both). A z114 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 2818-M05, because an additional processor drawer is not added.
  - Can change the model capacity identifier, the capacity setting, or both.
- Hardware installation upgrade:
  - Can change the server model 2818-M05 to M10, if an additional processor drawer is included. This upgrade is non-concurrent.
  - Can change the model capacity identifier, the capacity setting, or both.

The model capacity identifier can be concurrently changed. Concurrent upgrades can be accomplished for both *permanent* and *temporary* upgrades.

Model upgrades: A model upgrade from the M05 to the M10 is disruptive.

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 the following components and situations:

- Processors (logical central processors (CPs), IFL processors, Internal Coupling Facility (ICF) processors, System z Application Assist Processors (zAAPs), System z Integrated Information Processors (zIIPs), and SAPs.
- If unused PUs are available on the installed processor drawers or if the model capacity identifier for the CPs can be increased.
- Memory, when unused capacity is available on the installed memory cards. The plan-ahead memory option is available for clients to gain better control over future memory upgrades. See 2.5.4, “Memory upgrades” on page 46 for more details.
- I/O card ports 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 cards, drawers, and features:

- HCA2, HCA3, or PCI-e fanout cards
- I/O cards, when slots are available in the installed I/O drawers and PCIe I/O drawers
- I/O drawers and PCIe I/O 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.

Concurrent PU conversions (MES-ordered)

The z114 supports concurrent conversion between all PU types, such as any-to-any PUs, including SAPs, to provide flexibility to meet changing business requirements.

LICCC-based PU conversions: 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 for the PU conversion to be done nondisruptively.

PU conversion: Client 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 separate types.
- It might require individual LPAR outage if dedicated PUs are converted.

Unassigned CP capacity is recorded by an MCI. CP feature conversions change (increase or decrease) the MCI.

9.2.2 Customer Initiated Upgrade facility

The Customer Initiated Upgrade (CIU) facility is an IBM online system through which a client can 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 client and IBM, through which the terms and conditions for use of the CIU facility are accepted. The use of the CIU facility for a 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 authorization feature code, FC 9898.

After a client has placed an order through the CIU facility, the client will receive a notice that the order is ready for download. The client can then download and apply the upgrade by using functions that are 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 client.

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 that is required for an upgrade must be previously installed. An additional processor drawer 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 drawers. 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 client must be registered and the system must be set up. After completing the CIU registration, access the CIU application through the IBM Resource Link website:

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

As part of the setup, the client 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 website and invoke the CIU application to upgrade a server for processors, or memory. Requesting a client order approval to conform to customer operation policies is possible. Clients can 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 can generate online permanent upgrade orders to concurrently add processors (CPs, ICFs, zAAPs, zIIPs, IFLs, and SAPs) and memory, or change the MCI, up to the limits of the installed processor drawers on an existing z114 CPC.

Temporary upgrades

The base model z114 describes permanent and dormant capacity (Figure 9-1) using the capacity marker and the number of PU features installed on the CPC. Up to eight temporary offerings can be present. Each offering has its own policies and controls, and each offering 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.

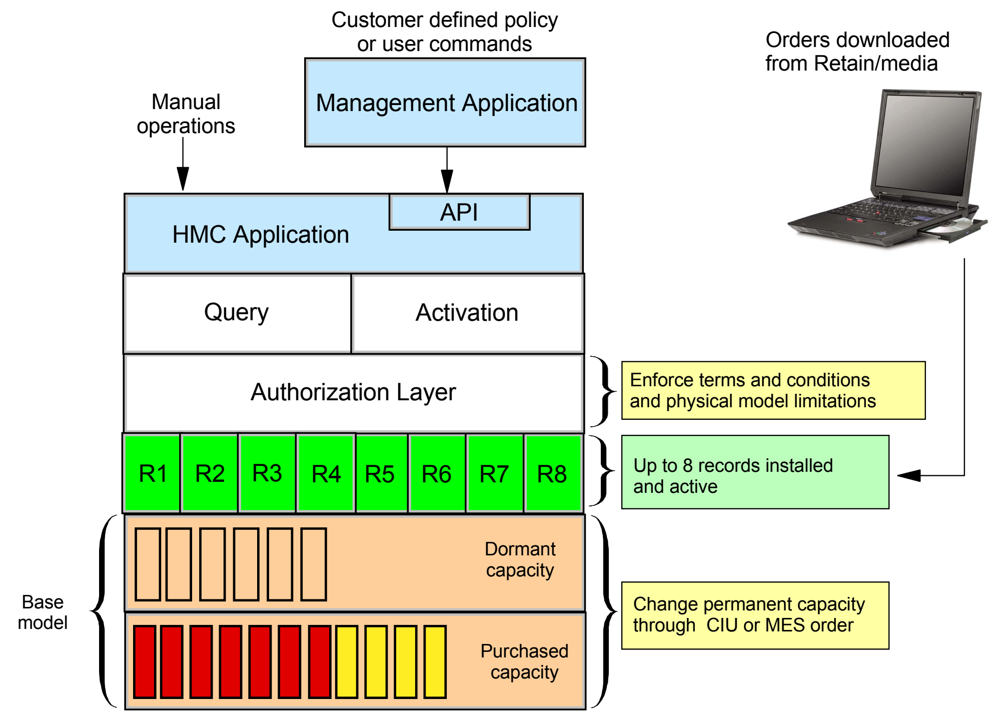

Customer defined policy or user commands
Orders downloaded from Retain/media
Management Application
Manual operations
API
HMC Application
Query
Activation
Authorization Layer
Enforce terms and conditions and physical model limitations
R1
R2
R3
R4
R5
R6
R7
R8
Up to 8 records installed and active
Dormant capacity
Base model
Purchased capacity
Change permanent capacity through CIU or MES order

Figure 9-1 The provisioning architecture

Temporary upgrades are represented in the z114 by a *record*. All temporary upgrade records, downloaded from the remote support facility (RSF) or installed from portable media, are resident on the Support Element (SE) hard drive. At the time of activation, the client 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 the control of an application through a documented application program interface (API).

By using the API approach, you can 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 capability eliminates the need for temporary upgrade specification for all possible scenarios. However, for CPE, the specific 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.

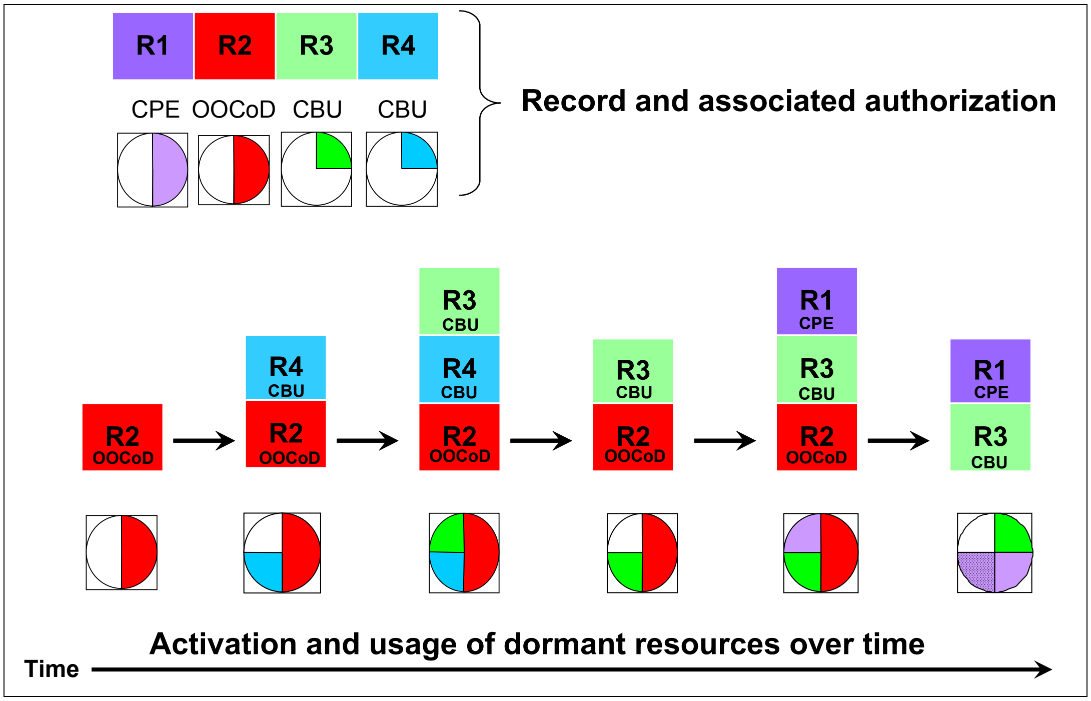

R1
R2
R3
R4
CPE
OOCoD
CBU
CBU
Record and associated authorization
R2 OOCoD
R4 CBU
R3 CBU
R1 CPE
Activation and usage of dormant resources over time
Time

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 can 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 client peak workload requirements, for any length of time, and has a daily hardware and maintenance charge. The software charges can vary according to the license agreement for the individual products. See your IBM Software Group representative for exact details.

  On/Off CoD can concurrently add processors (CPs, ICFs, zAAPs, zIIPs, IFLs, and SAPs), increase the model capacity identifier, or both, up to the limit of the installed processor drawers of an existing server, and is restricted to twice the currently installed capacity. On/Off CoD requires a contract agreement between the client and IBM.

The client decides 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 client 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 the installed processor drawers 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 client's enterprise for planned downtime events, for example, with data center changes. CPE cannot be used for peak load management of the client workload or for a disaster situation.

  The CPE feature requires unused capacity to be available on installed processor drawers of the backup server, either as unused PUs or as a possibility to increase the model capacity identifier, 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 upgrade combinations.

Table 9-2 Concurrent upgrade summary

Type	Name	Upgrade	Process
Permanent	MES	CPs, ICFs, zAAPs, zIIPs, IFLs, SAPs, processor drawer, memory, and I/O	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 On/Off CoD 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. 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 an additional processor drawer, 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 processor drawers.
  - LICCC to adjust the number and types of PUs or to change the capacity setting, or both.
  - Installing an additional processor drawer and LICCC assigning and activating unassigned PUs in the installed drawer.
- 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 processor drawers. Plan-ahead memory features enable you to have better control over future memory upgrades. For details about the memory features, see 2.5.4, “Memory upgrades” on page 46.
  - Installing an additional processor drawer and using LICCC to activate additional memory capacity on installed processor drawers.
- MES upgrades for I/O are done by either of the following methods:
  - Using LICCC to activate additional ports on already installed I/O cards.
  - Installing additional I/O cards and supporting infrastructure if required in I/O drawers or PCIe I/O drawers that are already installed, or installing additional I/O drawers or PCIe I/O drawers to hold the new cards.
- MES upgrades for the zEnterprise BladeCenter Extension can only be performed through your IBM service support representative (SSR).

An MES upgrade requires IBM service personnel for the installation. In most cases, the time that is required for installing the LICCC and completing the upgrade is short. To better exploit the MES upgrade function, we strongly suggest that you 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 power-on resets (PORs) or IPLs. The availability of I/O drawers and PCIe I/O drawers has improved the flexibility to perform unplanned I/O configuration changes concurrently.

The store system information (STSI) instruction gives more useful and detailed information about the base configuration and about temporary upgrades. STSI 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 314 for more details.

Upgrades: The MES provides the physical upgrade, resulting in more enabled processors, separate capacity settings for the CPs, additional memory, and I/O ports. Additional planning tasks are required for nondisruptive logical upgrades (see “Recommendations to avoid disruptive upgrades” on page 316).

9.3.1 MES upgrade for processors

An MES upgrade for processors can concurrently add CPs, ICFs, zAAPs, zIIPs, IFLs, and SAPs to a z114 by assigning available PUs that reside on the processor drawers, through LICCC. Depending on the quantity of the additional processors in the upgrade, an additional processor drawer might be required before the LICCC is enabled. Additional capacity can be provided by adding CPs, by changing the capacity identifier on the current CPs, or by doing both.

Maximums: The sum of CPs, inactive CPs, ICFs, zAAPs, zIIPs, IFLs, unassigned IFLs, and SAPs cannot exceed the maximum limit of PUs available for client use. The number of zAAPs or zIIPs cannot exceed the number of purchased CPs.

9.3.2 MES upgrade for memory

An MES upgrade for memory can concurrently add more memory by enabling, through LICCC, additional capacity up to the limit of the currently installed memory cards. Installing an additional processor drawer and LICCC-enabling the memory capacity on the new drawer are disruptive upgrades.

The Preplanned Memory Feature (FC 1993) is available to allow better control over future memory upgrades. See 2.5.5, “Pre-planned memory” on page 47, for details about plan-ahead memory features.

The M05 model has, as a minimum, ten 4 GB dual inline memory modules (DIMMs), resulting in 40 GB of installed memory in total. The minimum client addressable storage is 8 GB. If you require more than that, a *non-concurrent* upgrade can install up to 120 GB of memory for client use, by changing the existing DIMMs.

The M10 model has, as a minimum, twenty 4 GB DIMMs, resulting in 80 GB of installed memory in total. The minimum client addressable storage is 16 GB. If you require more than that, a *non-concurrent* upgrade can install up to 248 GB of memory for client use, by changing the existing DIMMs.

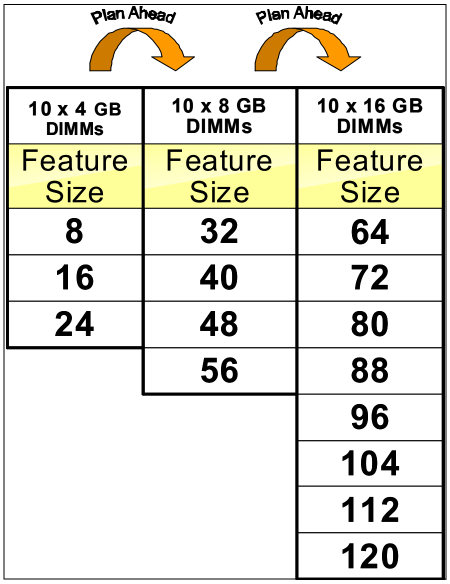
An LPAR can dynamically take advantage of a memory upgrade if reserved storage has been defined to that LPAR. The reserved storage is defined to the LPAR as part of the image profile. Reserved memory can be configured online to the LPAR by using the LPAR dynamic storage reconfiguration (DSR) function. DSR allows a z/OS operating system image, and z/VM partitions, to add reserved storage to their configuration if any unused storage exists. The nondisruptive addition of storage to a z/OS and z/VM partition necessitates that pertinent operating system parameters have been prepared. If reserved storage has not been defined to the LPAR, the LPAR must be deactivated, the image profile changed, and the LPAR reactivated to allow the additional storage resources to be available to the operating system image.

9.3.3 Preplanned Memory Feature

The Preplanned Memory Feature (FC 1993) enables you to install memory for future use:

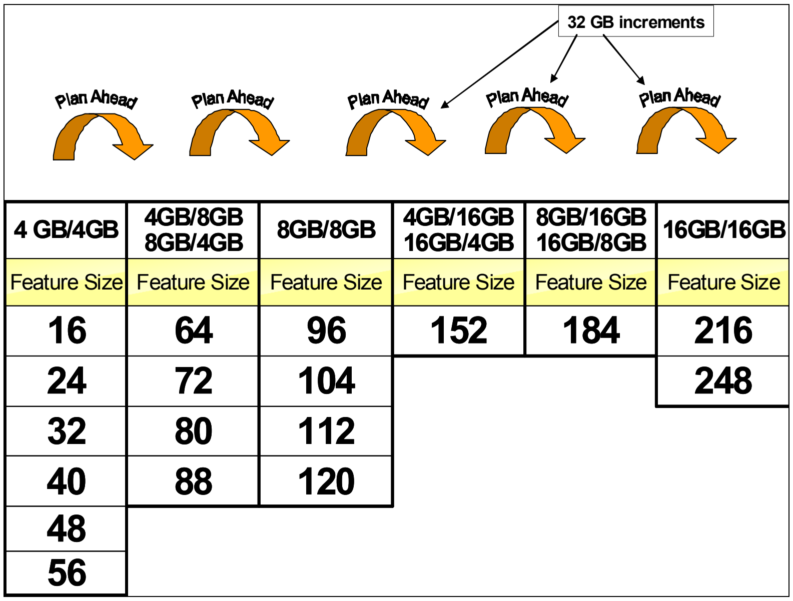
- FC 1993 specifies memory to be installed but not used. Order one feature for each 8 GB that will be usable by the client.
- FC 1903 is used to activate previously installed preplanned memory and can activate all the pre-installed memory or subsets of it. For each additional 8 GB (32 GB for larger memory configurations) of memory to be activated, one FC 1903 must be added, and one FC 1993 must be removed.

See Figure 9-3 and Figure 9-4 for details of memory configurations and upgrades.


Plan Ahead
Plan Ahead

10 x 4 GB DIMMs	10 x 8 GB DIMMs	10 x 16 GB DIMMs
Feature Size	Feature Size	Feature Size
8	32	64
16	40	72
24	48	80
	56	88
		96
		104
		112
		120

Figure 9-3 Memory sizes and upgrades for the M05


32 GB increments
Plan Ahead
Plan Ahead
Plan Ahead
Plan Ahead
Plan Ahead

4 GB/4GB	4GB/8GB 8GB/4GB	8GB/8GB	4GB/16GB 16GB/4GB	8GB/16GB 16GB/8GB	16GB/16GB
Feature Size	Feature Size	Feature Size	Feature Size	Feature Size	Feature Size
16	64	96	152	184	216
24	72	104			248
32	80	112			
40	88	120			
48					
56					

Figure 9-4 Memory sizes and upgrades for the M10

The accurate planning and definition of the target configuration are vital to maximize the value of these features.

9.3.4 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 Enterprise Systems Connection (ESCON) channels and InterSystem Channel (ISC)-3 links, activating ports on the existing 16-port ESCON or ISC-3 daughter (ISC-D) cards.

- Installing additional I/O cards in an existing I/O drawer or PCIe I/O drawer, or adding a new I/O drawer or PCIe I/O drawer to hold the new I/O cards

Figure 9-5 shows a z114 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.

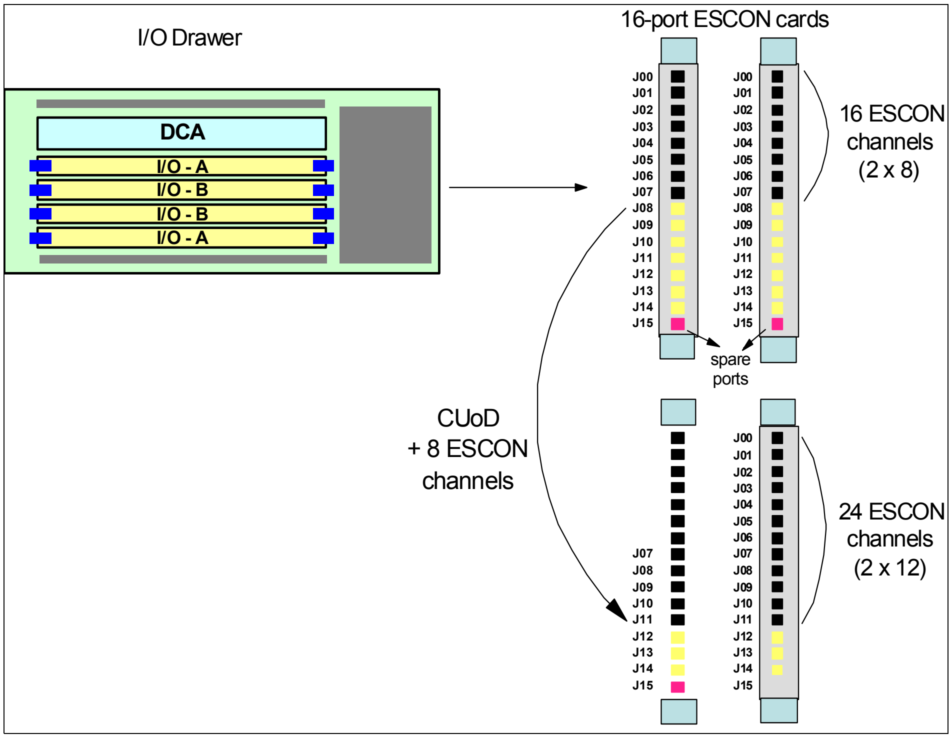


Figure 9-5 MES for I/O LICCC upgrade example

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 must be carefully planned to allow concurrent upgrades up to the target configuration.

9.3.5 MES upgrades for the zBX

The MES upgrades for zBX can concurrently add blades if there are any slots available in the existing blade chassis, add chassis if there are any free spaces in existing racks, add racks up to a maximum of four, and add entitlements for connections to the z114. For the IBM Smart Analytics Optimizer, the solution will continue to support applications using the z114 resources until the zBX has been upgraded and brought back into production status.

9.4 Permanent upgrade through the CIU facility

By using the CIU facility (through the IBM Resource Link on the web), you can 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 client location. You can also unassign previously purchased CPs and IFL processors through the CIU facility.

The capability to add permanent upgrades to a given z114 through the CIU facility requires that the permanent upgrade enablement feature (FC 9898) is installed on the z114. A permanent upgrade might change the MCI if additional CPs are requested or the change capacity identifier as part of the permanent upgrade, but it cannot change the z114 model from an M05 to an M10. If necessary, additional LPARs can be created concurrently to use the newly added processors.

Planning: A permanent upgrade of processors can provide a physical concurrent upgrade, resulting in more enabled processors available to a z114 configuration. Thus, additional planning and tasks are required for *nondisruptive* logical upgrades. See "Recommendations to avoid disruptive upgrades" on page 316 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 the Workload License Charge (WLC) might not be affected by the server upgrade, because their charges are based on LPAR utilization and not based on the server total capacity. See 8.12.2, "Advanced Workload License Charges (AWLC)" on page 271, for more information about the WLC.

Figure 9-6 illustrates the CIU facility process on IBM Resource Link.

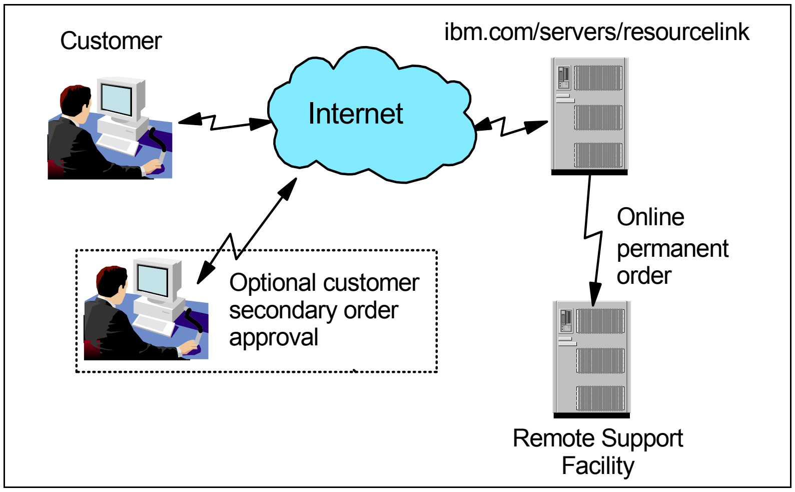

Customer
ibm.com/servers/resourcelink
Internet
Online permanent order
Optional customer secondary order approval
Remote Support Facility

Figure 9-6 Permanent upgrade order example

The following sample sequence on IBM Resource Link initiates an order:

1. Sign on to Resource Link.
2. Select **Customer Initiated Upgrade** from the main Resource Link page. The customer and server details that are 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**. Resource Link limits the options to those options 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 will be available within hours.

Figure 9-7 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.

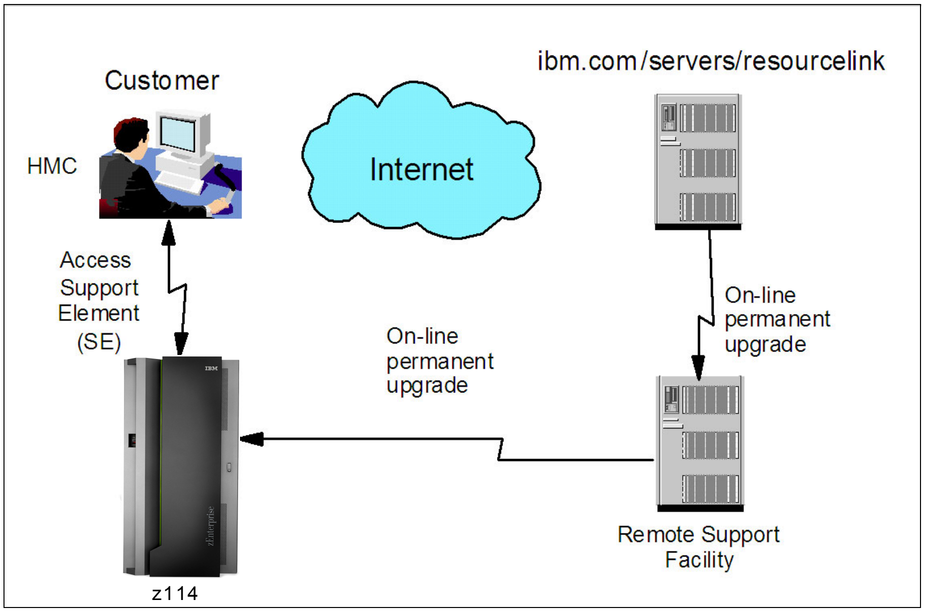


Figure 9-7 CIU-eligible order activation example

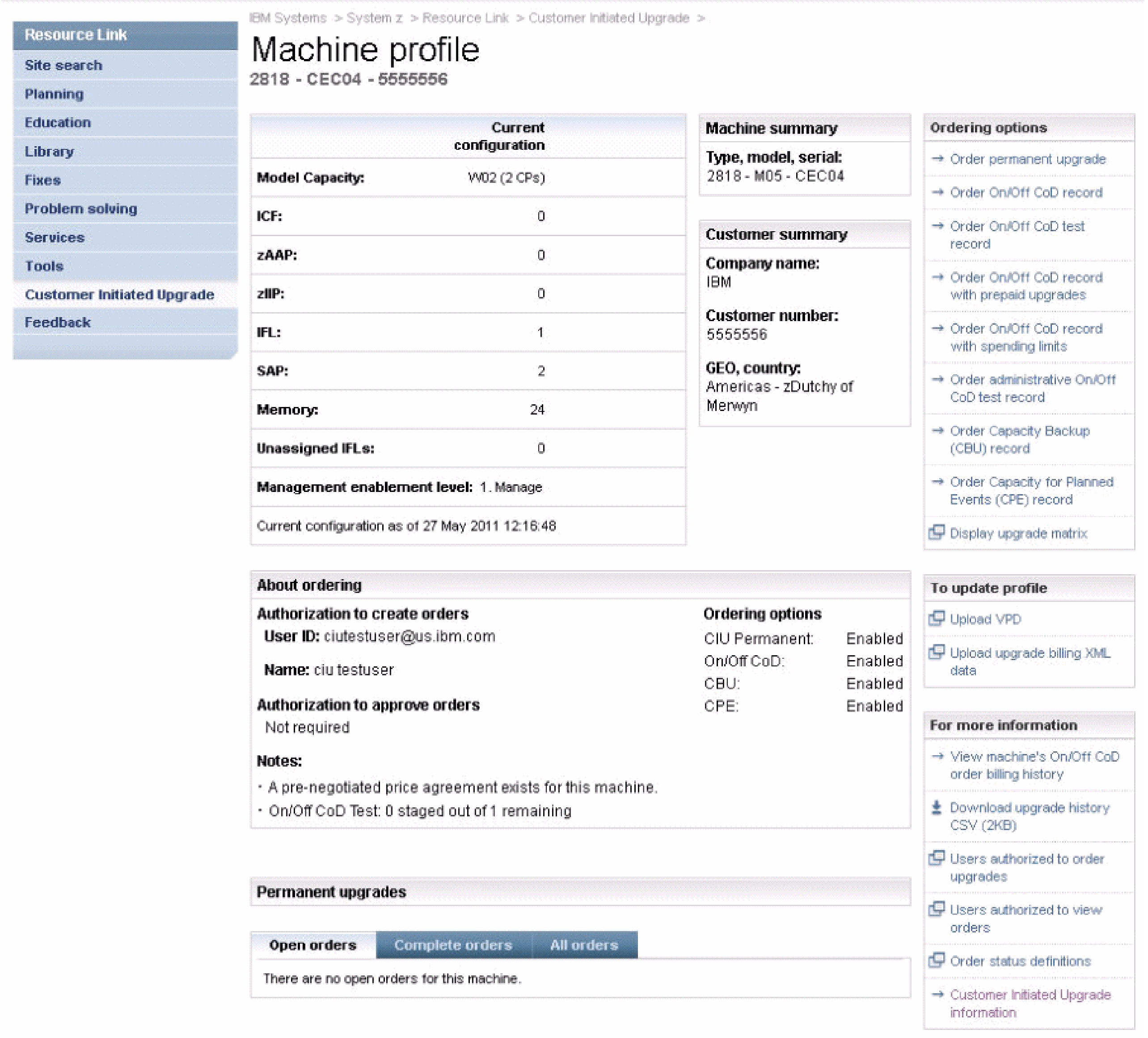
The two major components in the process are *ordering* and *retrieval* (along with activation).

9.4.1 Ordering

Resource Link provides the interface that enables you to order a concurrent upgrade for a server. You can create, cancel, view the order, and view the history of orders that were placed through this interface. Configuration rules enforce that only valid configurations are generated within the limits of the individual 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. For a tutorial, see this website:

https://www-304.ibm.com/servers/resourcelink/hom03010.nsf/pages/CIUInformation?OpenDocument

Figure 9-8 shows the initial view of the machine profile on Resource Link.


Resource Link
Site search
Planning
Education
Library
Fixes
Problem solving
Services
Tools
Customer Initiated Upgrade
Feedback
IBM Systems > System z > Resource Link > Customer Initiated Upgrade >

Machine profile

2818 - CEC04 - 5555556

	Current configuration
Model Capacity:	W02 (2 CPs)
ICF:	0
zAAP:	0
zIIP:	0
IFL:	1
SAP:	2
Memory:	24
Unassigned IFLs:	0
Management enablement level: 1. Manage	
Current configuration as of 27 May 2011 12:16:48	

Machine summary

Type, model, serial: 2818 - M05 - CEC04

Customer summary

Company name: IBM

Customer number: 5555556

GEO, country: Americas - zDutchy of Merwyn

Ordering options

- Order permanent upgrade
- Order On/Off CoD record
- Order On/Off CoD test record
- Order On/Off CoD record with prepaid upgrades
- Order On/Off CoD record with spending limits
- Order administrative On/Off CoD test record
- Order Capacity Backup (CBU) record
- Order Capacity for Planned Events (CPE) record
- Display upgrade matrix

About ordering

Authorization to create orders

User ID: ciutestuser@us.ibm.com

Name: ciu testuser

Authorization to approve orders

Not required

Ordering options	
CIU Permanent:	Enabled
On/Off CoD:	Enabled
CBU:	Enabled
CPE:	Enabled

Notes:

- A pre-negotiated price agreement exists for this machine.
- On/Off CoD Test: 0 staged out of 1 remaining

To update profile

- Upload VPD
- Upload upgrade billing XML data

For more information

- View machine's On/Off CoD order billing history
- Download upgrade history CSV (2KB)
- Users authorized to order upgrades
- Users authorized to view orders
- Order status definitions
- Customer Initiated Upgrade information

Permanent upgrades

There are no open orders for this machine.

Figure 9-8 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.

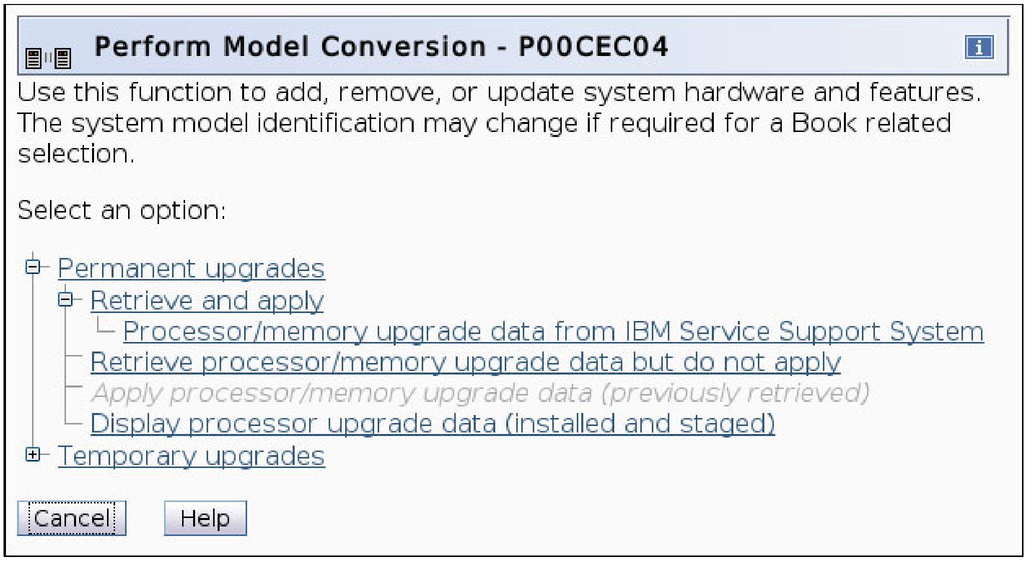
Resource Link retrieves and stores relevant data that is associated with the processor configuration, such as the number of CPs and installed memory cards. It allows you to select only those upgrade options that are deemed valid by the order process. It allows upgrades only within the bounds of the currently installed hardware.

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 can 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 **Permanent upgrades** to start the process. See Figure 9-9.



Perform Model Conversion - P00CEC04

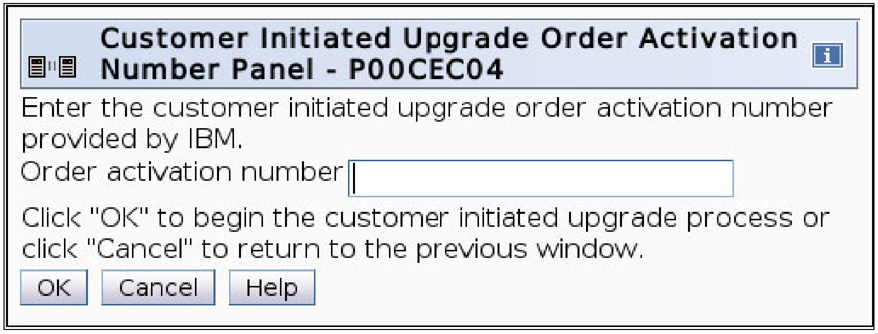
Use this function to add, remove, or update system hardware and features. The system model identification may change if required for a Book related selection.

Select an option:

- Permanent upgrades
  - Retrieve and apply
    - Processor/memory upgrade data from IBM Service Support System
  - Retrieve processor/memory upgrade data but do not apply
  - Apply processor/memory upgrade data (previously retrieved)
  - Display processor upgrade data (installed and staged)
- Temporary upgrades

Figure 9-9 z114 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 9-10.



Customer Initiated Upgrade Order Activation Number Panel - P00CEC04

Enter the customer initiated upgrade order activation number provided by IBM.

Order activation number

Click "OK" to begin the customer initiated upgrade process or click "Cancel" to return to the previous window.

Figure 9-10 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 that are 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 millions of service units (MSUs). The capacity for speciality engines is expressed in the 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 in the following manner:

- 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 must 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 the 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 it represents the time from the first activation of resources in a record until the end of the billing window in which the last resource in a record is deactivated. At the end of each billing window, the tokens are decremented by the highest usage of each resource during the billing window. If any resource in a record does not have enough tokens to cover usage for the next billing window, the entire record will be deactivated.

On/Off CoD requires that the Online CoD Buying feature (FC 9900) be installed on the server that is to be upgraded.

The 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 client is given a window of opportunity to assess the capacity additions to the client's permanent configurations using On/Off CoD. If a purchase is made, the hardware On/Off CoD charges during this window, three days or less, are waived. If no purchase is made, the client is charged for the temporary use.

The resources that are eligible for temporary use are CPs, ICFs, zAAPs, zIIPs, IFLs, and SAPs. The temporary addition of memory and I/O ports is not supported. Unassigned PUs that are on the installed processor drawers can be temporarily and concurrently activated as CPs, ICFs, zAAPs, zIIPs, IFLs, and SAPs through LICCC, up to twice the currently installed CP capacity and up to twice the number of ICFs, zAAPs, zIIPs, or IFLs. Therefore, an On/Off CoD upgrade cannot change the model from an M05 to an M10. The addition of a new processor drawer is not supported. However, the activation of an On/Off CoD upgrade can increase the MCI.

9.5.2 Ordering

Concurrently installing temporary capacity by ordering On/Off CoD is possible:

- 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 two for both models.

On/Off CoD can provide CP temporary capacity in two ways:

- 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 that differs from the installed CPs, the base capacity settings are changed to match.

  On/Off CoD has the following limits associated with its use:

  - The number of CPs cannot be reduced.
  - The target configuration capacity is limited:
    - Twice the currently installed capacity, expressed in MSUs for CPs
    - Twice the number of installed IFLs, ICFs, zAAPs, and zIIPs. The number of additional SAPs that can be activated is two.

    See Appendix B, “Valid z114 On/Off Capacity on Demand upgrades” on page 375 for the valid On/Off CoD configurations for CPs.

On/Off CoD can be ordered as prepaid or postpaid:

- A prepaid On/Off CoD offering record contains the 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, and speciality engines, and it can 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 is 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 z114 with capacity identifier D02, 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 MCI can be changed to a D03, which adds one additional CP (making a 3-way) or to a D04, which adds two additional CPs (making a 4-way).
- The second option is to change to a separate capacity level of the current CPs and change the model capacity identifier to an E02 or to an F02. The capacity level of the CPs is increased, but no additional CPs are added. The D02 can also be temporarily upgraded

to an E03 as indicated in the appendix pointer, thus increasing the capacity level and adding another processor.

We suggest 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 this website:

https://www-304.ibm.com/servers/resourcelink/lib03060.nsf/pages/lsprindex

The On/Off CoD hardware capacity is charged on a 24-hour basis. There is a grace period at the end of the On/Off CoD day. This grace period allows up to an hour after the 24-hour billing period to either change the On/Off CoD configuration for the next 24-hour billing period or deactivate the current On/Off CoD configuration. The times when the capacity is activated and deactivated are maintained in the z114 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 z114 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 LPARs can be activated concurrently to use the newly added processor resources.

Planning: On/Off CoD provides a concurrent *hardware* upgrade, resulting in more enabled processors available to a z114 configuration. Additional planning tasks are required for *nondisruptive* upgrades. See “Recommendations to avoid disruptive upgrades” on page 316.

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 z114. Subsequently, you can concurrently install temporary capacity up to the limits in On/Off CoD and use it for up to 180 days. Monitoring occurs through the z114 call-home facility, and an invoice is generated if the capacity has been enabled during the calendar month. The client will continue to be billed for use of temporary capacity until the z114 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 separate 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.

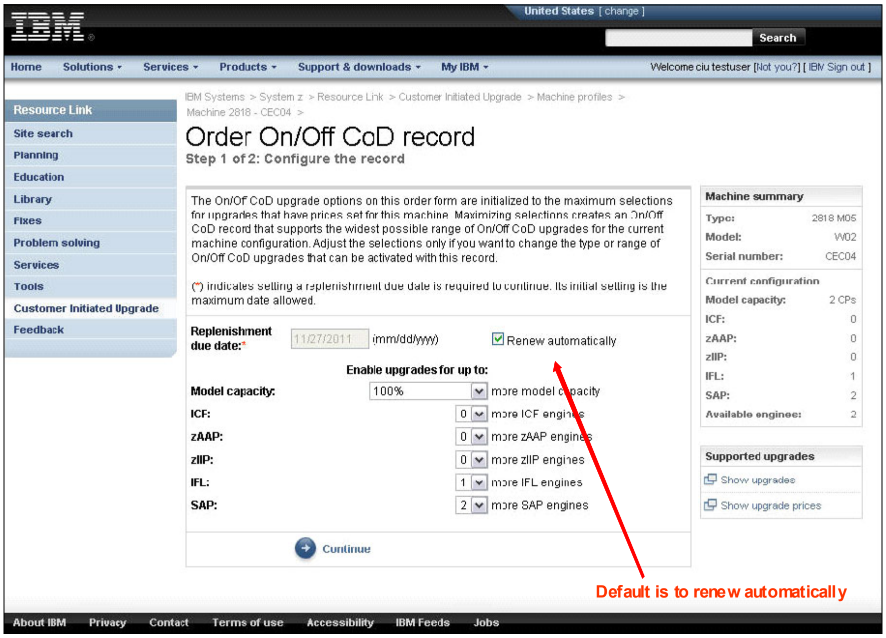
Another possibility is to store unlimited On/Off CoD LICCC records on the Support Element with the same or separate capacities at any given time, giving greater flexibility to quickly enable needed temporary capacity. Each record is easily identified with descriptive names, and you can select from a list of records that can be activated.

Resource Link provides the interface that allows you to order a dynamic upgrade for a specific z114. 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 z114.

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 can 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 various 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 client can use an automatic renewal feature to eliminate the need for a manual replenishment of the On/Off CoD order. This feature is implemented in Resource Link, and the client must enable the feature in the machine profile. The default for the feature is disabled. See Figure 9-11 for more details.


IBM
United States [change]
Search
Home
Solutions
Services
Products
Support & downloads
My IBM
Welcome ciu testuser [Not you?] [IBM Sign out]
Resource Link
Site search
Planning
Education
Library
Fixes
Problem solving
Services
Tools
Customer Initiated Upgrade
Feedback
IBM Systems > System z > Resource Link > Customer Initiated Upgrade > Machine profiles > Machine 2818 - CEC04 >

Order On/Off CoD record

Step 1 of 2: Configure the record

The On/Of CoD upgrade options on this order form are initialized to the maximum selections for upgrades that have prices set for this machine. Maximizing selections creates an On/Off CoD record that supports the widest possible range of On/Off CoD upgrades for the current machine configuration. Adjust the selections only if you want to change the type or range of On/Off CoD upgrades that can be activated with this record.

(*) indicates setting a replenishment due date is required to continue. Its initial setting is the maximum date allowed.

Replenishment due date:* 11/27/2011 (mm/dd/yyyy) ☑ Renew automatically

Enable upgrades for up to:

Model capacity:	100%	more model capacity
ICF:	0	more ICF engines
zAAP:	0	more zAAP engines
zIIP:	0	more zIIP engines
IFL:	1	more IFL engines
SAP:	2	more SAP engines

Machine summary

Type:	2818 M05
Model:	W02
Serial number:	CEC04

Current configuration

Model capacity:	2 CPs
ICF:	0
zAAP:	0
zIIP:	0
IFL:	1
SAP:	2
Available engines:	2

Default is to renew automatically

Figure 9-11 Order On/Off CoD record panel

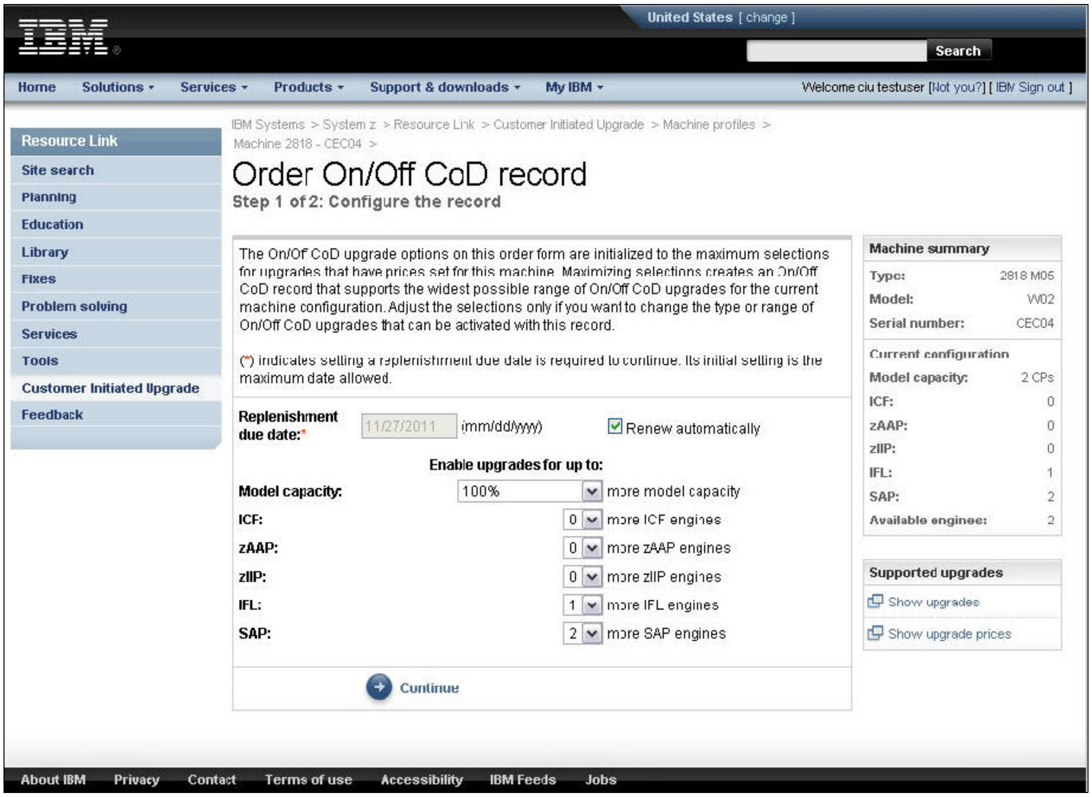
9.5.3 On/Off CoD testing

Each On/Off CoD-enabled z114 is entitled to one no-charge 24-hour test. There will be no IBM charges for the test, including no IBM charges that are 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.

The test can last up to of 24 hours, commencing upon the activation of any capacity resource that is contained in the On/Off CoD record. Activation levels of capacity can change during the 24-hour test period. The On/Off CoD test automatically terminates at the end of the 24-hour period.

In addition, there is a capability to perform administrative testing, through which no additional capacity is added to the z114, but the client can test all the procedures and automation for the management of the On/Off CoD facility.

Figure 9-12 is an example of an On/Off CoD order on the Resource Link web page.



Order On/Off CoD record

Step 1 of 2: Configure the record

The On/Of CoD upgrade options on this order form are initialized to the maximum selections for upgrades that have prices set for this machine. Maximizing selections creates an On/Off CoD record that supports the widest possible range of On/Off CoD upgrades for the current machine configuration. Adjust the selections only if you want to change the type or range of On/Off CoD upgrades that can be activated with this record.

(*) indicates setting a replenishment due date is required to continue. Its initial setting is the maximum date allowed.

Replenishment due date:* 11/27/2011 (mm/dd/yyyy) ☑ Renew automatically

Enable upgrades for up to:

Model capacity:	100%	more model capacity
ICF:	0	more ICF engines
zAAP:	0	more zAAP engines
zIIP:	0	more zIIP engines
IFL:	1	more IFL engines
SAP:	2	more SAP engines

Machine summary

Type:	2818 M05
Model:	W02
Serial number:	CEC04

Current configuration

Model capacity:	2 CPs
ICF:	0
zAAP:	0
zIIP:	0
IFL:	1
SAP:	2
Available engines:	2

Supported upgrades

Show upgrades

Show upgrade prices

Figure 9-12 On/Off CoD order example

The example order in Figure 9-12 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 processor drawers. The maximum number of SAPs is determined by the model number and the number of available PUs on the already installed processor drawers.

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 can activate the order when the capacity is needed, either manually or through automation.

If the On/Off CoD offering record does not contain resource tokens, you must 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 LPARs, you might be required to perform tasks at the LPAR 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 LPARs 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 separate 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 client is contractually obliged to terminate the On/Off CoD right-to-use feature when a transfer in asset ownership occurs. A client can 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 that the CIU enablement feature is removed, the On/Off CoD right-to-use is simultaneously removed. Reactivating the right-to-use feature subjects the client 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 z114. 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 z114 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 the 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) clients are billed at the MSU level that is represented by the combined permanent and temporary capacity. All PSLC products are billed at the peak MSUs enabled during the month, regardless of usage. Clients with WLC licenses are billed by product at the highest four-hour rolling average for the month. In this instance, temporary capacity does not necessarily increase the software bill until that capacity is allocated to LPARs and actually consumed.

Results from the STSI instruction reflect the current permanent and temporary CPs. See “Store system information (STSI) instruction” on page 314 for more details.

9.5.6 z/OS capacity provisioning

The z114 provisioning capability, combined with Capacity Provisioning Manager (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 9-13.

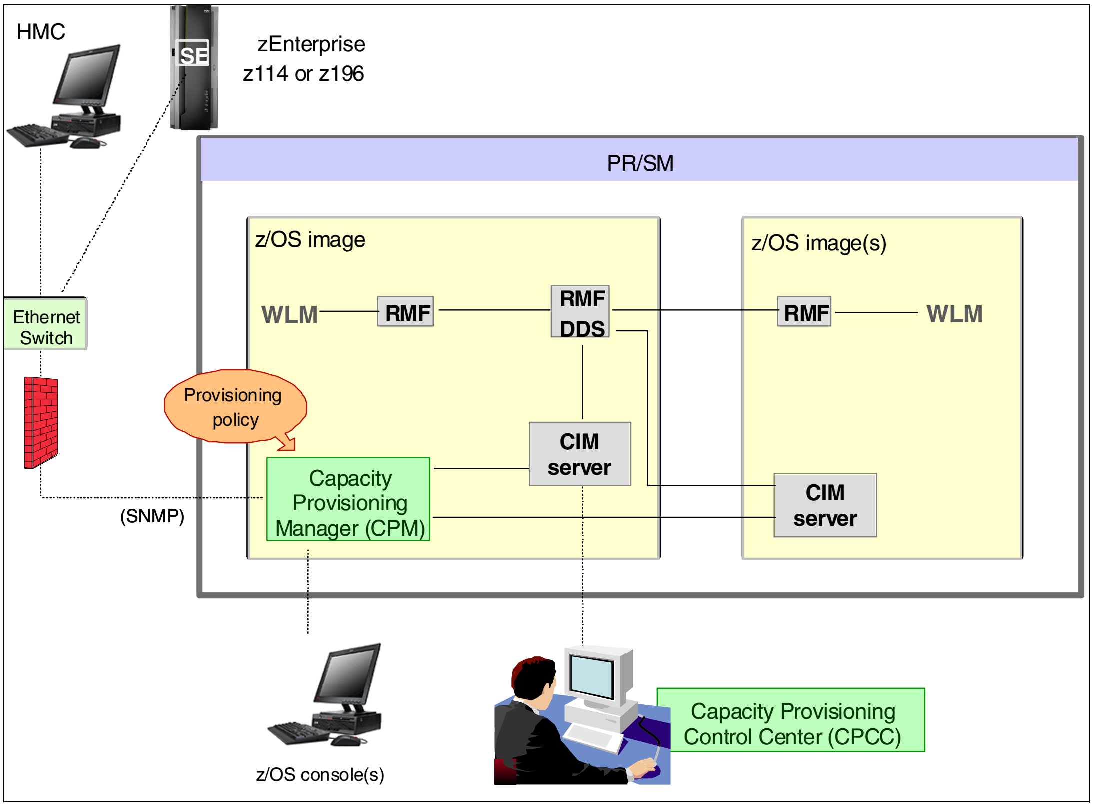


Figure 9-13 The capacity provisioning infrastructure

The z/OS WLM manages the workload by the 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 Simple Network Management Protocol (SNMP).

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 insufficient tokens remain, no resource from the On/Off CoD record is activated.

If a capacity token is completely consumed during an activation that is 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. Warning messages will be issued if capacity tokens are getting close to being fully consumed. The messages will start being generated 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 five days. We suggest that you put operational procedures in place to handle these situations. You can 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 CPCC will over time be moved into the z/OSMF. Parts of the CPCC have been included in z/OSMF V1R13.

The control over the provisioning infrastructure is executed by the CPM through the Capacity Provisioning Domain (CPD), which is controlled by the Capacity Provisioning Policy (CPP). An example of a Capacity Provisioning Domain is shown in Figure 9-14.

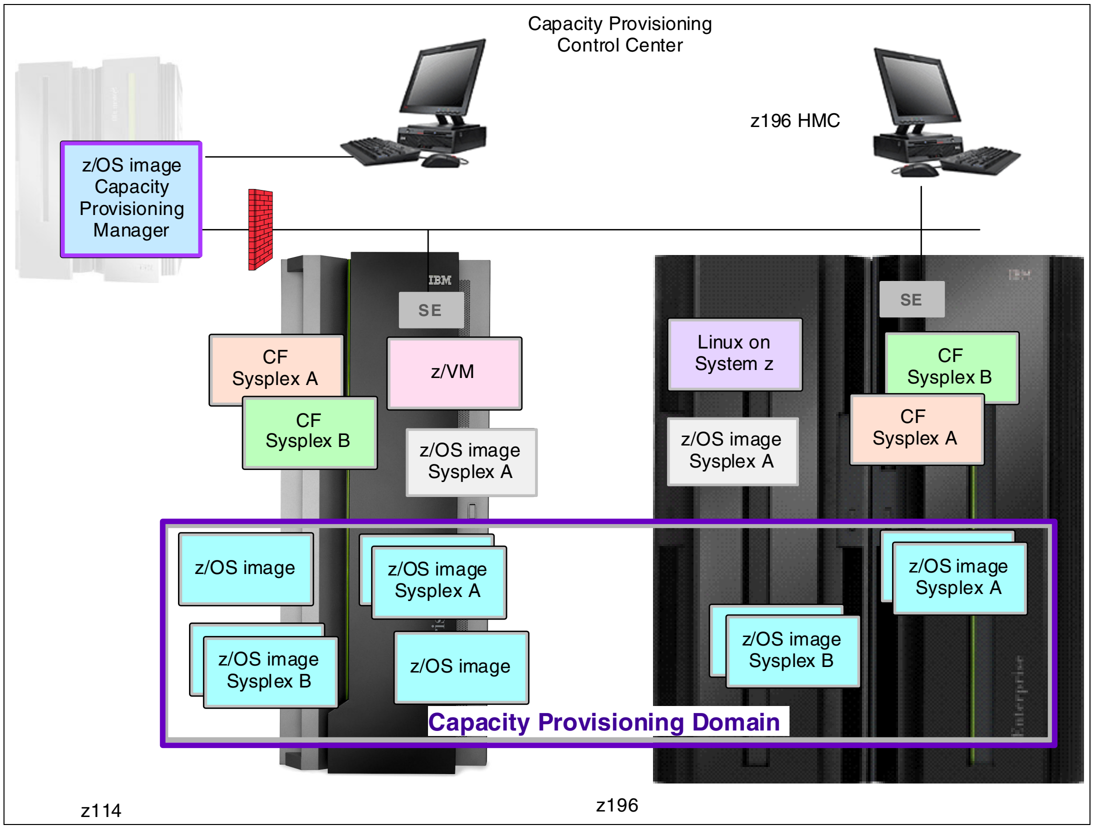


Figure 9-14 A Capacity Provisioning Domain

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 various 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 the policy criteria.
  - The operator determines whether to ignore the information or to manually upgrade or downgrade the system by using the HMC, SE, or available CPM commands.

- Confirmation mode

  In this mode, CPM processes capacity-provisioning policies and interrogates the installed temporary offering records. Every action that is proposed by the CPM 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 the workload, 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 can be provisioned (when, which, and how). The criteria has three elements:

- A *time condition* is when provisioning is allowed:
  - Start time indicates when provisioning can begin.
  - Deadline indicates that the provisioning of the additional capacity is no longer allowed.
  - End time indicates that the deactivation of the additional capacity must begin.
- A *workload condition* is a description of which workload qualifies for provisioning. The parameters include this information:
  - The z/OS systems that might execute the eligible work.
  - The importance filter indicates eligible service class periods, which are identified by Workload Manager (WLM) importance.
  - Performance Index (PI) criteria:
    - Activation threshold: The PI of the service class periods must exceed the activation threshold for a specified duration before the work is considered to be suffering.
    - Deactivation threshold: The PI of the service class periods must fall below the deactivation threshold for a specified duration before the work is considered to no longer be suffering.
  - *Included service classes* are eligible service class periods.
  - *Excluded service classes* are service class periods that must not be considered.

  > **If no workload condition is specified:** The full capacity that is described in the policy will be activated and deactivated at the start and end times that are specified in the policy.

- *Provisioning scope* is how much additional capacity can be activated, and it is expressed in MSUs.

  Specified in MSUs, the number of zAAPs and the 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 can be activated for all the rules in the Capacity Provisioning Domain.

The provisioning rule

The provisioning rule is *in the specified time interval, if the specified workload is behind its objective, up to the defined additional capacity can be activated.*

The rules and conditions are named and stored in the Capacity Provisioning Policy. For more information about z/OS Capacity Provisioning functions, see *z/OS MVS Capacity Provisioning User's Guide,* SA33-8299 .

Planning considerations for using automatic provisioning

Although only one On/Off CoD offering can be active at any one time, several On/Off CoD offerings can be present on the z114. 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 suggest that only one On/Off CoD offering be created on the z114 by specifying the maximum allowable capacity. The Capacity Provisioning Manager can then, at the time that 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 in the offering.

Having an unlimited number of offering records pre-staged on the SE hard disk is possible; changing the content of the offerings if necessary is also possible.

> **Important:** The CPM has control over capacity tokens for the On/Off CoD records. In a situation where a capacity token is completely consumed, the z114 deactivates the corresponding offering record. Therefore, we strongly recommend that you prepare routines for catching the warning messages about capacity tokens being consumed, and have administrative routines in place for these situations. 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 z114 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 it might depend on other work executing with importance level 1. Therefore, considering which workloads to control and which service class periods to specify is extremely important.

9.6 Capacity for Planned Event

Capacity for Planned Event (CPE) is offered with the z114 to provide replacement backup capacity for planned downtime events. For example, if a server room requires an extension or repair work, replacement capacity can be installed temporarily on another z114 in the client's environment.

> **Important:** CPE is for planned replacement capacity only and it *cannot* be used for peak workload management.

CPE has these feature codes:

- FC 6833: Capacity for Planned Event enablement
- FC 0116: 1 CPE Capacity Unit
- FC 0117: 100 CPE Capacity Unit
- FC 0118: 10000 CPE Capacity Unit
- FC 0119: 1 CPE Capacity Unit-IFL
- FC 0120: 100 CPE Capacity Unit-IFL
- FC 0121: 1 CPE Capacity Unit-ICF
- FC 0122: 100 CPE Capacity Unit-ICF
- FC 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 previous list 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 you access to dormant PUs on the server for which you have a contract (as described by the feature codes listed). 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 processor drawer. CPE features can be added to an existing z114 nondisruptively. 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 z114 configuration must have sufficient memory and channels to accommodate the potential requirements of the large CPE configuration. It is important to ensure that all required functions and resources are available on the z114 where CPE is activated, including CF LEVELs for coupling facility partitions, memory, cryptographic functions, and connectivity capabilities.

The CPE configuration is activated temporarily and provides additional PUs in addition to the z114's original, permanent configuration. The number of additional PUs is predetermined by the number and type of feature codes configured as described by the list of the feature codes. The number of PUs that can be activated is limited by the unused capacity that is available on the server. When the planned event is over, the server must be returned to its original configuration. You can deactivate the CPE features at any time before the expiration date.

A CPE contract must be in place before the special code that enables this capability can be installed. CPE features can be added to an existing z114 nondisruptively.

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 z114.

CBU is the quick, temporary activation of PUs and is available in these durations:

- For up to 90 consecutive 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

> **Important:** 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 PUs that you require. CBU has these feature codes:

- FC 9910: CBU enablement
- FC 6805: Additional test activations
- FC 6817: Total CBU years ordered
- FC 6818: CBU records ordered
- FC 6820: Single CBU CP-year
- FC 6821: 25 CBU CP-year
- FC 6822: Single CBU IFL-year
- FC 6823: 25 CBU IFL-year
- FC 6824: Single CBU ICF-year
- FC 6825: 25 CBU ICF-year
- FC 6826: Single CBU zAAP-year
- FC 6827: 25 CBU zAAP-year
- FC 6828: Single CBU zIIP-year
- FC 6829: 25 CBU zIIP-year
- FC 6830: Single CBU SAP-year
- FC 6831: 25 CBU SAP-year
- FC 6832: CBU replenishment

The CBU entitlement record (FC 6818) contains an expiration date that is established at the time of order and depends upon the quantity of CBU years (FC 6817). You have the capability to extend your CBU entitlements through the purchase of additional CBU years. The number of FC 6817 per instance of FC 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 processor drawer. The maximum number of CBU features that can be *ordered* is 10. The number of features that can be *activated* is limited by the number of unused PUs on the server.

However, the ordering system allows for over-configuration in the order itself. You can *order* up to 10 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 subset of the CBU features that are ordered for the system.

Subcapacity makes a difference in the way that the CBU features are done. On the full-capacity models, the CBU features indicate the amount of additional capacity needed. If the amount of necessary CBU capacity is equal to four CPs, the CBU configuration is four CBU CPs.

The number of CBU CPs must be equal to or greater than the number of CPs in the base configuration, and all the CPs in the CBU configuration must have the same capacity setting. For example, if the base configuration is a 2-way D02, providing a CBU configuration of a 4-way of the same capacity setting requires two CBU feature codes. If the required CBU capacity changes the capacity setting of the CPs, going from model capacity identifier D02 to a CBU configuration of a 4-way E04 requires four CBU feature codes with a capacity setting of E*xx*.

If the capacity setting of the CPs is changed, more CBU features are required, not more physical PUs. Therefore, your CBU contract requires more CBU features if the capacity setting of the CPs is changed.

Note that CBU can add CPs through LICCC only, and the z114 must have the proper number of processor drawers installed to allow the required upgrade. CBU can change the model capacity identifier to a *higher* value than the base setting, but it does not change the *z114* model. The CBU feature cannot *decrease* the capacity setting.

A CBU contract must be in place before the special code that enables this capability can be installed on the z114. CBU features can be added to an existing z114 nondisruptively. For each machine enabled for CBU, the authorization to use CBU is available for a definite number of years: one to five 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 subset of the capacity that is specified in the CBU contract.

The base server configuration must have sufficient memory and channels to accommodate the potential requirements of the large CBU target configuration. Ensure that all required functions and resources are available on the backup z114, 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 z114 must be taken back to its original configuration. The CBU features can be deactivated by the client 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.

> **Planning:** CBU for processors provides a concurrent upgrade, resulting in more enabled processors or changed capacity settings available to a z114 configuration, or both. You decide, at activation time, to activate a subset 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 316.

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 client responsibility and does not require the on-site presence of IBM service personnel. The CBU function is activated and deactivated concurrently from the HMC using the API. On the SE, CBU is activated either using the Perform Model Conversion task or through the API (the API enables task automation).

CBU activation

CBU is activated from the SE by using the Perform Model Conversion task or through automation by using the API on the SE or the HMC. In the case of a real disaster, use the Activate CBU option to activate the 90-day period.

Image upgrades

After the CBU activation, the z114 can have more capacity, more active PUs, or both. The additional resources go into the resource pools and are available to the LPARs. If the LPARs have to increase their share of the resources, the LPARs 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 LPARs can be created to use the newly added capacity.

CBU deactivation

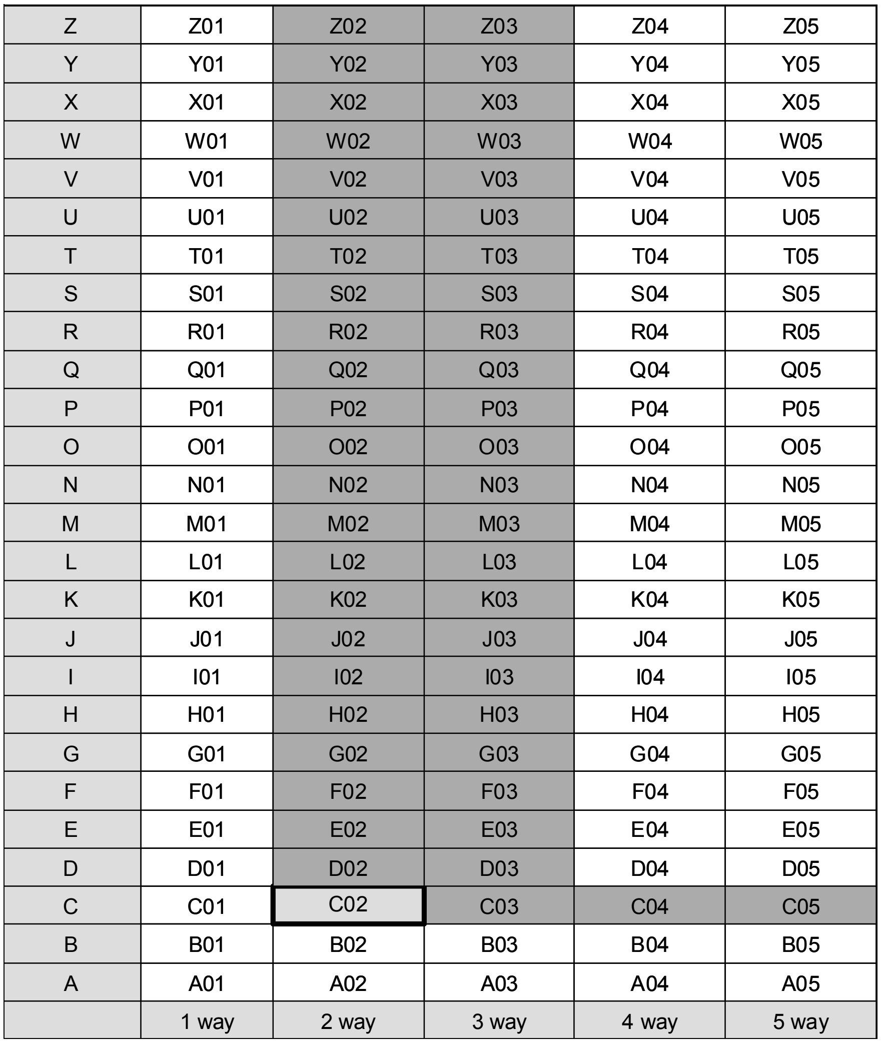
To deactivate the CBU, the additional resources have to be released from the LPARs by the operating systems. In certain cases, releasing resources is a matter of varying the resources offline. In other cases, it can mean shutting down operating systems or deactivating LPARs. After the resources have been released, the same facility on the SE is used to turn off CBU. To deactivate CBU, click **Undo temporary upgrade** 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 can 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 the in-use shared engine. Testing can be accomplished by ordering a diskette, calling the support center, or using the facilities on the SE. The client can purchase additional tests.

CBU example

We describe the following example of a capacity backup operation. The permanent configuration is C02 and a record contains three CP CBU features. During an activation, you can choose among many target configurations. With three CP CBU features, you can add one to three CPs, which allow you to activate C03, C04, or C05. Or, two CP CBU features can be used to change the capacity level of permanent CPs, which means that you can activate D02, E02, and F02 through Z02. Or two CP CBU features can be used to change the capacity level of permanent CPs, and the third CP CBU feature can be used to add a CP, which allows the activation of D03, E03, and F03 through Z03. In this example, you are offered 49 possible configurations at activation time, as shown in Figure 9-15. While CBU is active, you can change the target configuration at any time.



Z	Z01	Z02	Z03	Z04	Z05
Y	Y01	Y02	Y03	Y04	Y05
X	X01	X02	X03	X04	X05
W	W01	W02	W03	W04	W05
V	V01	V02	V03	V04	V05
U	U01	U02	U03	U04	U05
T	T01	T02	T03	T04	T05
S	S01	S02	S03	S04	S05
R	R01	R02	R03	R04	R05
Q	Q01	Q02	Q03	Q04	Q05
P	P01	P02	P03	P04	P05
O	O01	O02	O03	O04	O05
N	N01	N02	N03	N04	N05
M	M01	M02	M03	M04	M05
L	L01	L02	L03	L04	L05
K	K01	K02	K03	K04	K05
J	J01	J02	J03	J04	J05
I	I01	I02	I03	I04	I05
H	H01	H02	H03	H04	H05
G	G01	G02	G03	G04	G05
F	F01	F02	F03	F04	F05
E	E01	E02	E03	E04	E05
D	D01	D02	D03	D04	D05
C	C01	C02	C03	C04	C05
B	B01	B02	B03	B04	B05
A	A01	A02	A03	A04	A05
	1 way	2 way	3 way	4 way	5 way

Figure 9-15 Example of C02 with three CBU features

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 takeover 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 function minimizes operator intervention and the potential for errors.
- Automates the 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 clients, and even planned outages are no longer acceptable. Although Parallel Sysplex clustering technology is the best continuous availability solution for z/OS environments, nondisruptive upgrades within a single server can avoid system outages and are suitable to additional operating system environments.

The z114 allows *concurrent* upgrades, meaning that dynamically adding more capacity to the CPC is possible. If operating system images running on the upgraded CPC do not require disruptive tasks in order to use the new capacity, the upgrade is also *nondisruptive.* Therefore, power-on reset (POR), LPAR deactivation, and IPL do not have to take place.

If the concurrent upgrade is intended to satisfy an *image upgrade* to an LPAR, the operating system running in this partition must also have the capability to concurrently configure more capacity online. z/OS operating systems have this capability. z/VM can concurrently configure new processors and I/O devices online, and 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 LPARs can be created *concurrently* on the z114 CPC, including all resources needed by such LPARs. These additional LPARs can be activated concurrently.

Linux operating systems in general do *not* have the capability of adding more resources concurrently. However, Linux, and other types of virtual machines running under z/VM, can benefit from the z/VM capability to nondisruptively configure more resources online (processors and I/O).

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 z114 if unassigned PUs are available in any installed processor drawer. The number of zAAPs cannot exceed the number of CPs, plus unassigned CPs. The same is true for the zIIPs.

Concurrent upgrades are not supported with PUs defined as additional SAPs.

If necessary, additional LPARs can be created concurrently to use the newly added processors.

The Coupling Facility Control Code (CFCC) can also configure more processors online to coupling facility LPARs by using the CFCC image operations window.

Memory

Memory can be concurrently added up to the physically installed memory limit.

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 fanouts) is present on the 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 is in the configuration.

Concurrent upgrade considerations

By using MES upgrade, On/Off CoD, CBU, or CPE, a z114 can be concurrently upgraded from one capacity identifier 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 need to consider the effect on the software running on a z114 when you perform any of these configuration upgrades.

Processor identification

Two instructions are used to obtain processor information:

- Store System Information instruction (STSI)

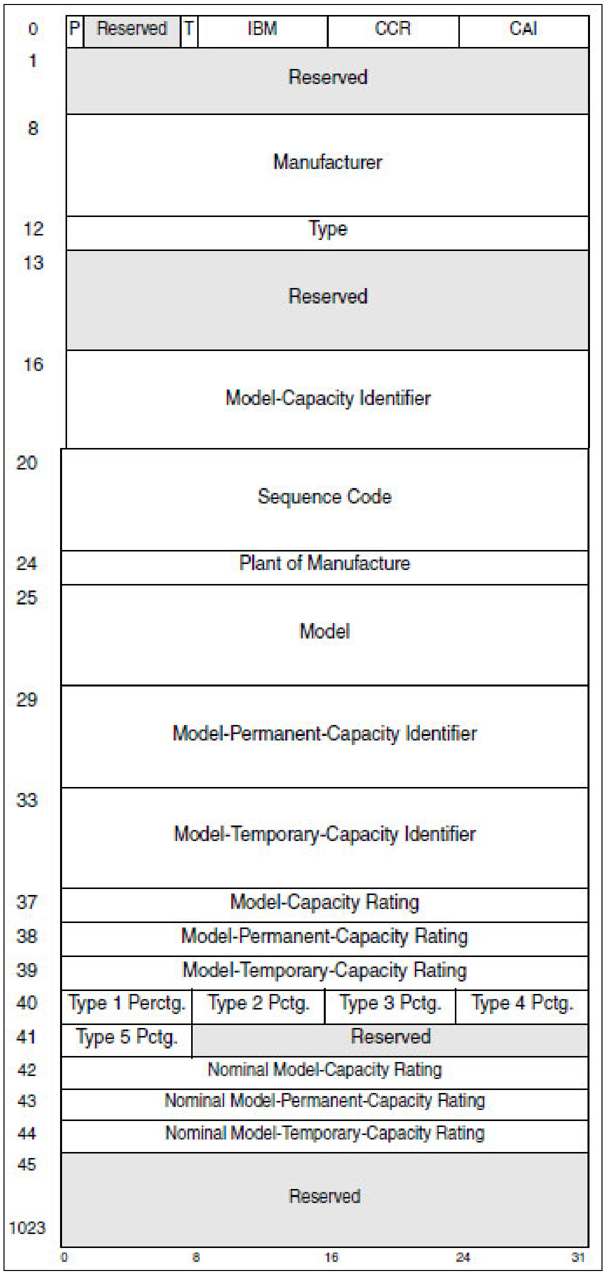
  STSI reports the processor model and model capacity identifier for the base configuration and for any additional configuration changes through temporary upgrade actions. It fully supports the concurrent upgrade functions and is the preferred way to request processor information.

- Store CPU ID instruction (STIDP)

  STIDP is provided for purposes of backward compatibility.

Store system information (STSI) instruction

Figure 9-16 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.



0	P Reserved T	IBM	CCR	CAI
1	Reserved			
8	Manufacturer			
12	Type			
13	Reserved			
16	Model-Capacity Identifier			
20	Sequence Code			
24	Plant of Manufacture			
25	Model			
29	Model-Permanent-Capacity Identifier			
33	Model-Temporary-Capacity Identifier			
37	Model-Capacity Rating			
38	Model-Permanent-Capacity Rating			
39	Model-Temporary-Capacity Rating			
40	Type 1 Perctg.	Type 2 Pctg.	Type 3 Pctg.	Type 4 Pctg.
41	Type 5 Pctg.	Reserved		
42	Nominal Model-Capacity Rating			
43	Nominal Model-Permanent-Capacity Rating			
44	Nominal Model-Temporary-Capacity Rating			
45 1023	Reserved			
	0	8	16	24 31

Figure 9-16 STSI output on z114

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 LPAR identifier. See Table 9-3. The LPAR identifier field is a full byte to support greater than 15 LPARs.

Table 9-3 STIDP output for z114

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'2818'	x'8000' [c]

[a] The version code for z114 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:

- 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 the same as they are 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 exactly the same as they are without running as a z/VM guest.

Table 9-4 lists the possible output returned to the issuing program for an operating system running as a guest under z/VM.

Table 9-4 z/VM guest STIDP output for z114

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'2818'	x'8000'
With SET CPUID command	x'FF'	6-digit number as entered by the command SET CPUID = *nnnnnn*		x'2818'	x'8000'

Planning for nondisruptive upgrades

Online permanent upgrades, On/Off CoD, CBU, and CPE can be used to concurrently upgrade a z114. However, certain situations require a disruptive task to enable the new capacity that was recently added to the CPC. Several of these situations can be avoided if planning is done in advance. Planning ahead is a key factor for nondisruptive upgrades.

The following list describes the major reasons for disruptive upgrades. However, by carefully planning and by reviewing “Recommendations to avoid disruptive upgrades” on page 316, you can minimize the need for these outages:

- z/OS LPAR processor upgrades when reserved processors were not previously defined are disruptive to image upgrades.
- LPAR memory upgrades when reserved storage was not previously defined are disruptive to image upgrades. z/OS and z/VM support this function.
- Installation of an additional processor drawer to upgrade a model M05 to a model M10 is non-concurrent.
- 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 the 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 315), here are several recommendations for avoiding or at least minimizing these situations, increasing the potential for nondisruptive upgrades:

- For z/OS LPARs, configure as many reserved processors (CPs, ICFs, zAAPs, and zIIPs) as possible.

  Configuring reserved processors for all logical z/OS partitions *before* their activation enables them to be nondisruptively upgraded. The operating system running in the LPAR 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 V1R11, z/OS V1R12, and z/OS V1R13 support up to 80 processors, including CPs, zAAPs, and zIIPs. z/VM supports up to 32 processors.

- Configure reserved storage to LPARs.

  Configuring reserved storage for all LPARs *before* their activation enables them to be nondisruptively upgraded. The operating system running in the LPAR must have the ability to configure memory online. The amount of reserved storage can be higher than the processor drawer threshold limit, even if the second processor drawer is not installed. The current partition storage limit is 1 TB. z/OS and z/VM support this function.

- Consider the 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 that are already installed on the processor drawers. For details about the offerings, see 2.5.3, “Memory configurations” on page 44.

Considerations when installing an additional CEC drawer

During an upgrade, an additional CEC drawer can be installed non-concurrently. Depending on your I/O configuration, a fanout rebalancing might be desirable for availability reasons.

9.9 Summary of Capacity on Demand offerings

The capacity on demand infrastructure and its offerings are major features for z114. The introduction of these features was based on numerous client requirements for more flexibility, granularity, and better business control over the System z infrastructure, operationally as well as financially.

One major client requirement is to dismiss the necessity for a client authorization connection to IBM Resource Link system at the time of activation of any offering. This requirement is being met by the z114. After the offerings have been installed on the z114, they can be activated at any time, completely at the client’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 z114 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 client-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

Reliability, availability, and serviceability

In this chapter, we describe several of the reliability, availability, and serviceability (RAS) features of the zEnterprise System.

The z114 design is focused on providing higher availability by reducing planned and unplanned outages. RAS can be accomplished with improved concurrent upgrade functions for processors, memory, and concurrent remove, repair, and add/upgrade for 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 discuss the environmental controls that are implemented in the z114 to help reduce the power consumption and cooling requirements.

The design of the memory on the z114 has taken a major step forward by implementing a fully redundant memory infrastructure, Redundant Array of Independent Memory (RAIM), a concept similar to the RAID design that is used in external disk storage systems. The zEnterprise central processor complexes (CPCs) are the only servers in the industry offering this level of memory design.

To make the delivery and transmission of microcode (LIC) secure, fixes and restoration/backup files are digitally signed. Any data that is transmitted to IBM Support is encrypted.

The design goal for the z114 has been to remove all sources of planned outages.

We cover the following topics:

10.1 z114 availability characteristics

The following functions include the availability characteristics on the z114:

- Concurrent memory upgrade

  Memory can be upgraded concurrently using a Licensed Internal Code (LIC) configuration code (LICCC) update if physical memory is available in the processor drawers. If the physical memory cards have to be changed, the z114 needs to be powered down. To help ensure that the appropriate level of memory is available in a configuration, consider the plan-ahead memory feature.

  The plan-ahead memory feature that is available with the z114 provides the ability to plan for nondisruptive memory upgrades by having the system pre-plugged with memory dual inline memory module (DIMMs) 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 LIC driver updates performed in support of new features and functions. The z114 is designed to support activating a selected new driver level concurrently.

- Concurrent fanout addition or replacement

  A PCIe, Host Channel Adapter (HCA), or Memory Bus Adapter (MBA) fanout card provides the path for data between memory and I/O using InfiniBand (IFB) cables or PCIe cables. With the z114, a hot-pluggable and concurrently upgradable fanout card is available. Up to four fanout cards are available per processor drawer for a total of eight fanout cards when both processor drawers are installed. In the event of an outage, a fanout card, that is used for I/O, can 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 z114 allows a multiplexer card in an I/O drawer or PCIe I/O drawer to be disconnected and replaced, continuing to provide connectivity to the z114's I/O resources using a second connection inside the drawer for the connection while the repair action takes place.

- Dynamic oscillator switchover

  The z114 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 system.

10.2 z114 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 made to the system.

Planned This outage is 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 client and often requires pre-planning. The z114 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 hardware system area (HSA) of 8 GB helps eliminate pre-planning requirements for HSA and 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 client-initiated upgrades

For a description of the flexible client-initiated upgrades, see Chapter 9, “System upgrades” on page 275.

10.2.1 Scheduled outages

Concurrent hardware upgrades, concurrent parts replacement, concurrent driver upgrade, and concurrent firmware fixes that are available with the z114, all address the elimination of scheduled outages. Furthermore, the following indicators and functions that address scheduled outages are included:

- Double memory data bus lane sparing

  This feature reduces the number of repair actions for memory.

- Single memory clock sparing
- Double DRAM chipkill tolerance
- Field repair of the cache fabric bus
- Power distribution N+2 design

  This feature uses Voltage Transformation Module (VTMs) in a highly redundant N+2 configuration.

- Redundant humidity sensors
- Redundant altimeter sensors
- Unified support for the zEnterprise BladeCenter Extension (zBX)

  The zBX is supported like any other feature on the z114.

- 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 z114 by asking the question, “Has this type of incident happened previously on this server?”

- Hot swap InfiniBand (IFB) hub cards

  When properly configured for redundancy, hot swapping (replacing) the IFB (HCA2-O (12xIFB) or HCA3-O (12xIFB)) 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 virtual LAN (VLAN)

  The service network in the machine gives the machine code the capability to monitor each single internal function in the machine. This capability 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, because the VLAN makes the service network itself easier to handle and more flexible.

- I/O drawer and PCIe I/O drawer

  Two types of I/O drawers are available for the z114. Both types can be installed concurrently and I/O cards can be added to the drawers concurrently.

10.2.2 Unscheduled outages

An unscheduled outage occurs because of an unrecoverable malfunction in a hardware component of the z114.

The following improvements are designed to minimize unscheduled outages:

- Continued focus on firmware quality

  For LIC and hardware design, failures are eliminated through rigorous design rules, design walk-throughs, and peer reviews. Failures also are eliminated through element, subsystem, and system simulation, and extensive engineering and manufacturing testing.

- Memory subsystem improvements

  z114 uses the Redundant Array of Independent Memory (RAIM), which is a concept that is known in the disk industry as Redundant Array of Independent Disks (RAID). RAIM design detects and recovers from DRAM, socket, memory channel, or DIMM failures. The RAIM design includes the addition of one memory channel that is dedicated for RAS. The parity of the four “data” DIMMs is stored in the DIMMs that are attached to 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 z114 is implemented with an enhanced version of the Reed-Solomon error correction code (ECC) that is known as 90B/64B, as well as protection against memory channel and DIMM failures. An extremely precise marking of faulty chips helps assure timely DRAM replacements. The key cache on the z114 memory is completely mirrored. For a full description of the memory system on the z114, see 2.5, “Memory” on page 41.

- 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 the soft switching of firmware components. For example, if you must upgrade the microcode of a Fibre Channel connection (FICON) feature, enhancements have been implemented to avoid any unwanted side effects that have been detected on previous servers.

- Server Time Protocol (STP) recovery enhancement

  When HCA3-O (12xIFB) or HCA3-O LR (1xIFB) coupling links are used, an unambiguous “going away signal” will be sent when the server on which the HCA3 is running is about to enter a failed (check stopped) state. When the “going away signal” that is sent by the Current Time Server (CTS) in an STP-only Coordinated Timing Network (CTN) is received by the Backup Time Server (BTS), the BTS can safely take over as the CTS without relying on the previous Offline Signal (OLS) in a two-server CTN or as the Arbiter in a CTN with three or more servers.

10.3 z114 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 LIC driver updates that are performed in support of new features and functions.

When properly configured, the z114 supports concurrently activating a selected new LIC driver level. Concurrent activation of the selected new LIC driver level is supported at specifically released sync points; however, there are certain LIC updates where a concurrent update/upgrade might not be possible.

Consider the following key points of EDM:

- The HMC can query whether a system is ready for a concurrent driver upgrade.
- Previous firmware updates, which require an initial machine load (IML) of z114 to be activated, can 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 the duration of 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 can result in a scheduled outage to recover.

The EDM function does not completely eliminate the need for planned outages for driver-level upgrades. Upgrades might require a system level or a functional element scheduled outage to activate the new LIC. The following circumstances require a scheduled outage:

- Specific complex code changes might dictate a disruptive driver upgrade. You are alerted in advance so that you can plan for the following changes:
  - Design data or hardware initialization data fixes
  - Coupling Facility Control Code (CFCC) release level change
- Non-queued direct input/output (QDIO) Open Systems Adapter (OSA) channel path identifier (CHPID) types, CHPID type OSC, and CHPID type OSE require CHPID Vary OFF/ON in order to activate the new code.
- FICON and Fibre Channel (FC) Protocol (FCP) code changes require a CHPID Config OFF/ON to activate the new code.

10.4 RAS capability for the HMC in an ensemble

The serviceability function for the components of an ensemble is delivered through the traditional HMC/SE constructs as for earlier System z servers. From a serviceability point of view, all 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.

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.

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 the survival of the system in case of failure. Each zEnterprise ensemble must be equipped with two HMC workstations: a primary and an alternate. While the primary HMC can perform all HMC activities (including Unified Resource Manager activities), the alternate can only be the backup and cannot be used for other tasks or activities.

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

For more information, see 12.7, “HMC in an ensemble” on page 363.

10.5 RAS capability for zBX

The zBX has been built with the traditional System z quality of service (QoS) to include RAS capabilities. The zBX offering provides extended service capability with the z114 hardware management structure. The HMC/SE functions of the z114 CPC 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 z114 CPC and the zBX for monitoring, controlling, and managing the zBX components.

BladeCenter components

Each BladeCenter has the following components:

- Up to 14 blade server slots. Blades can be removed, repaired, and replaced concurrently.
- (*N* + 1) PDUs. Provided the Power Distribution Units (PDUs) have power inputs from two separate 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 power system control network (PSCN).
- (*N* + 1) 10 GbE High Speed switches for the intraensemble data network (IEDN).
- (*N* + 1) 1000BaseT switches for the intranode management network (INMN).
- (*N* + 1) 8 Gb FC switches for the external disk.
- Two hot-swap Advanced Management Modules (AMMs).
- Two hot-swap fans/blowers.

> **Maximums:** Certain BladeCenter configurations do not physically fill up the rack with their components, but they might 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 z114 CPC. The same processes and controls are used. All fixes to the zBX are downloaded to the controlling z114's SE and applied to the zBX.

The Machine Change Levels (MCLs) for the zBX are designed to be concurrent and their status can be viewed at the z114's HMC.

zBX RAS and the Unified Resource Manager

The Hypervisor Management function of Unified Resource Manager provides tasks for managing the hypervisor life cycle, managing storage resources, performing RAS and the using the First Failure Data Capture (FFDC) features, and monitoring the supported hypervisors.

For blades that are deployed in a solution configuration, such as the Smart Analytics Optimizer or the DataPower solutions, the solution handles the complete end-to-end management for these blades and their operating systems, middleware, and applications.

For blades that are deployed by the client, the Unified Resource Manager handles the blades:

- The client must have an entitlement for each blade in the configuration.
- When the blade is deployed in the BladeCenter chassis, the Unified Resource Manager will power up the blade, verify that there is an entitlement for the blade, and verify 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 the 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 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 loaded following the activation. For client-deployed blades, all of the application, database, operating system, and network management will be handled by the client's usual system management disciplines.

10.5.1 Considerations for PowerHA in a zBX environment

An application running on AIX can be provided with high availability by the use of the PowerHA™ System Mirror for AIX, which is formerly known as HACMP™[1]. PowerHA is easy to configure because it is menu driven, and it provides high availability for applications that are running on AIX.

PowerHA helps you to define and manage the resources that are required by applications running on AIX, provide service and application continuity through platform resources and application monitoring, and automates actions (start/manage/monitor/restart/move/stop).

Failover: Resource movement and application restart on an alternate server is known as "*failover*".

Automating the failover process speeds up recovery and allows for unattended operations, thus providing improved application availability. In an ideal situation, an application needs to be available 24 hours x 365 days a year, which is also known as 24x7x365. Application availability can be measured as the amount of time that the service is actually available divided by the amount of time in a year, in a percentage.

A PowerHA configuration, which is also known as a "*cluster*", consists of two or more servers[2] (up to 32) that have their resources managed by PowerHA cluster services to provide automated service recovery for the managed applications. Servers can have physical or virtual I/O resources, or a combination of both.

PowerHA performs the following functions at the cluster level:

- Manage and monitor OS and hardware resources
- Manage and monitor application processes
- Manage and monitor network resources
- Automate applications (start/stop/restart/move).

The virtual servers that are defined and managed in zBX use only virtual I/O resources. PowerHA can manage both physical and virtual I/O resources, such as virtual storage and virtual network interface cards.

PowerHA can be configured to perform automated service recovery for the applications that are running in virtual servers that are deployed in zBX. PowerHA automates application failover from one virtual server in a System p® blade to another virtual server in a separate System p blade with a similar configuration.

Failover protects service, because it masks service interruption in case of unplanned or planned (scheduled) service interruptions. During failover, clients might experience a short

[1] High Availability Cluster Multi-Processing

[2] Servers can be also virtual servers; one server equals one instance of the AIX Operating System.

service unavailability, while the resources are being configured by PowerHA on the new virtual server.

The PowerHA configuration for the zBX environment is similar to standard Power environments, with the particularity that it uses only virtual I/O resources. Currently, PowerHA for zBX support is limited to failover inside the same zBX.

PowerHA configuration must cover the following planning, installation, integration, configuration, and testing:

- Network planning (VLAN and IP configuration definition and for server connectivity)
- Storage planning (shared storage must be accessible to all blades that provide resources for a PowerHA cluster)
- Application planning (start/stop/monitoring scripts and OS, CPU, and memory resources)
- PowerHA software installation and cluster configuration
- Application integration (integrating storage, networking, and application scripts)
- PowerHA cluster testing and documentation

Figure 10-1 shows a typical PowerHA cluster.

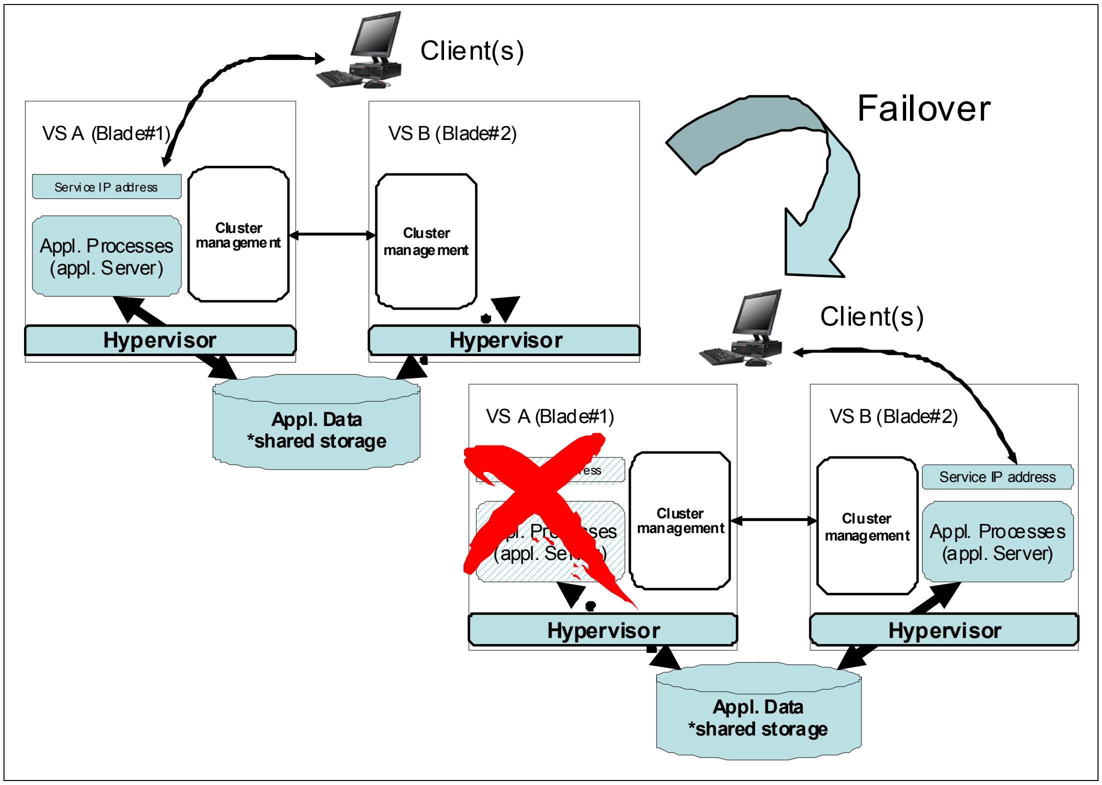


Figure 10-1 Typical PowerHA cluster diagram

For more information about IBM PowerHA System Mirror for AIX, see this website:

http://www-03.ibm.com/systems/power/software/availability/aix/index.html

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.”*
>
> _ “The Green Road Less Traveled” by Thomas L. Friedman, The New York Times, July 15, 2007

In this chapter, we briefly describe several of 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 114 and zEnterprise BladeCenter Extension.

There are a number of options for the physical installation of the z114, including raised floor as well as non-raised floor options, cabling from the bottom of the frame or off the top of the frame, and the option to have a high-voltage DC power supply directly into the z114, instead of the usual AC power supply.

For comprehensive physical planning information, see *zEnterprise 196 Installation Manual for Physical Planning,* GC28-6897.

We cover the following topics:

- “z114 power and cooling” on page 330
- “z114 physical specifications” on page 333
- “zBX environmentals” on page 334
- “Energy management” on page 338

11.1 z114 power and cooling

The z114 is always an air-cooled, one-frame system. Installation can be on a raised floor or non-raised floor with numerous options for top exit or bottom exit for all cabling, both power cords as well as I/O cables, as shown in Figure 11-1.

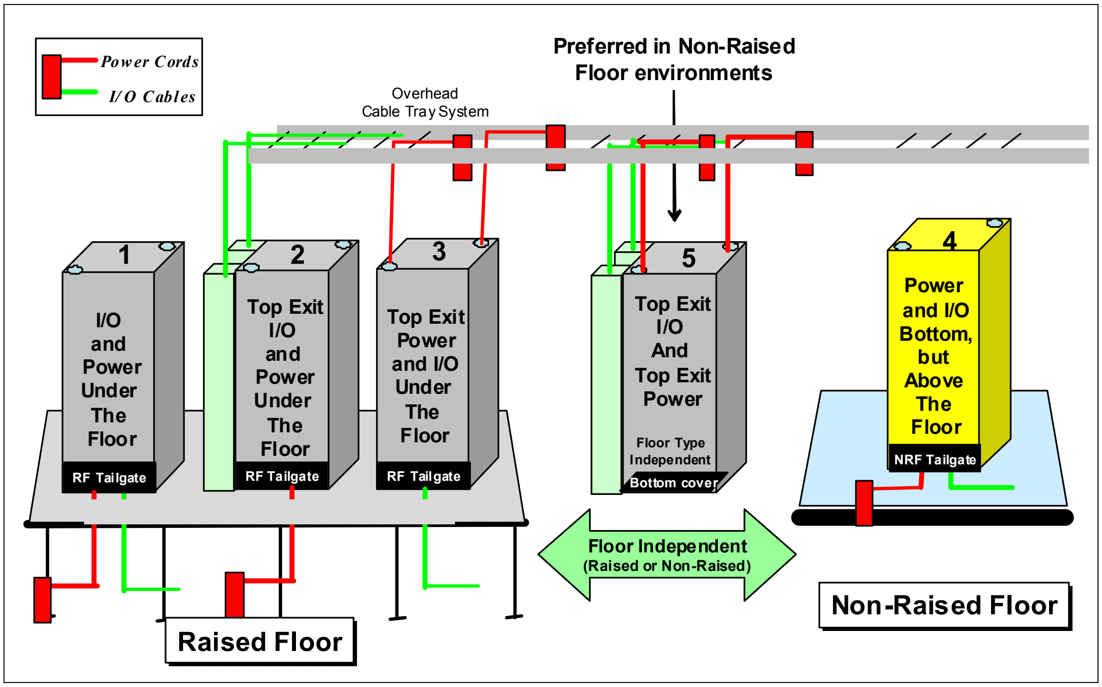

Power Cords
I/O Cables
Preferred in Non-Raised Floor environments
Overhead Cable Tray System
1
I/O and Power Under The Floor
RF Tailgate
2
Top Exit I/O and Power Under The Floor
RF Tailgate
3
Top Exit Power and I/O Under The Floor
RF Tailgate
5
Top Exit I/O And Top Exit Power
Floor Type Independent
Bottom cover
4
Power and I/O Bottom, but Above The Floor
NRF Tailgate
Floor Independent (Raised or Non-Raised)
Raised Floor
Non-Raised Floor

Figure 11-1 z114 cabling options

11.1.1 Power consumption

The system operates with two completely redundant power supplies. Each of the power supplies has its individual line cords. For redundancy, the server must have two power feeds. Each power feed is one line cord. Line cords attach to either 3 phase, 50/60 Hz, 250 or 450 V AC power or 380 to 570 V DC power. Depending on the configuration, single-phase line cords can be used (200 V 30 A). See the shaded area in Table 11-1 on page 331. There is no impact to the system operation with the total loss of one power feed.

For ancillary equipment, such as the Hardware Management Console (HMC), its display, and its modem, additional single-phase outlets are required.

The power requirements depend on the number of processor drawers, as well as the number of I/O drawers, that are installed. Table 11-1 on page 331 lists the maximum power requirements for the z114. These numbers assume the maximum memory configurations, all drawers fully populated, and all fanout cards installed. We strongly suggest that you use the Power Estimation tool to obtain a precise indication for a particular configuration. See 11.4.1, “Power estimation tool” on page 339.

Table 11-1 z114 system power in kilowatts

I/O drawer and PCIe I/O drawer[a]	Model M05			Model M10		
	1	**2**	**3**	**1**	**2**	**3**
No I/O Drawers	1.53	1.86	1.87	2.15	2.77	2.74
1 FC 4000	2.35	2.77	2.92	3.05	3.69	3.80
1 FC 4003	3.05	3.48	3.53	3.73	4.40	4.41
2 FC 4000	3.25	3.69	3.97	3.92	4.62	4.87
1 FC 4000 plus 1 FC 4003	3.92	4.42	4.61	4.62	5.34	5.49
3 FC 4000	4.13	4.62	5.06	4.82	5.55	5.95
2 FC 4003	4.54	5.04	5.12	5.24	5.97	6.02
2 FC 4000 plus 1 FC 4003	4.79	5.30	5.63	5.48	6.23	6.53
4 FC 4000	4.98	5.51	5.84	N/A	N/A	N/A
1 FC 4000 plus 2 FC 4003	N/A	N/A	N/A	6.80	7.58	7.78
FC 4000 = I/O drawer and FC 4003 = PCIe I/O drawer						

Notes:

1. Room ambient temperature below 28 degrees C, altitude below 914.37 m (3000 ft.).
2. Room ambient temperature above 28 degrees C, or altitude above 914.37 m (3000 ft.), but below 1,828.74 m (6000 ft.).
3. Room ambient temperature above 28 degrees C, and altitude above 914.37 m (3000 ft.), but below 1,828.74 m (6000 ft.), or altitude above 1,828.74 m (6000 ft.) at any temperature.

The shaded area of the table indicates configurations that are supported by a single-phase power supply.

[a] Note that I/O drawers cannot be ordered. I/O feature types will determine the appropriate mix of I/O drawers and PCIe I/O drawers.

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 an 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 that are listed in Table 11-2.

Table 11-2 z114 IBF sustained operations in minutes

I/O drawers and PCIe I/O drawers[a]	Model M05	Model M10
No I/O drawers	25	15
1 FC 4000	18	10.5
1 FC 4003	12	8.5
2 FC 4000	12	8.5
1 FC 4000 plus 1 FC 4003	9	6.5
3 FC 4000	9	6.5
2 FC 4003	7	5
2 FC 4000 plus 1 FC 4003	7	5
4 FC 4000	7	N/A
1 FC 4000 plus 2 FC 4003	N/A	4
FC 4000 = I/O drawer and FC 4003 = PCIe I/O drawer		

[a] Note that I/O drawers cannot be ordered. I/O feature types will determine the appropriate mix of I/O drawers and PCIe I/O drawers.

11.1.3 Emergency power-off

On the front of the frame is an emergency power-off switch that, when activated, immediately disconnects utility and battery power from the server. This method causes all volatile data in the z114 to be lost.

If the z114 is connected to a machine room's emergency power-off switch, and the IBF is installed, the batteries take over if the switch is engaged. To avoid this takeover, connect the machine room emergency power-off switch to the z114 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 IBF. However, all volatile data in the z114 will be lost.

11.1.4 Cooling requirements

The z114 is air cooled. The z114 requires chilled air, ideally coming from under a raised floor, to fulfill the air-cooling requirements; however, a non-raised floor option is available. The requirements for cooling are indicated in *zEnterprise 114 Installation Manual for Physical Planning,* GC28-6907.

The front and the rear of z114 dissipate separate amounts of heat. Most of the heat comes from the rear of the system. To calculate the heat output expressed in kBTU per hour for z114 configurations, multiply the table entries from Table 11-1 on page 331 by 3.4. The planning phase must consider the z114 proper placement in relation to the cooling capabilities of the data center.

11.2 z114 physical specifications

This section describes the weights and dimensions of the z114.

11.2.1 Weights and dimensions

Installation can be on a raised floor, as well as a non-raised floor. In the *zEnterprise 196 Installation Manual for Physical Planning,* GC28-6897, you will find more details about the installation requirements for the z114.

Table 11-3 indicates the maximum system dimension and weights for the z114 models.

Table 11-3 z114 physical dimensions

	z114 single frame					
Maximum	**Model M05**			**Model M10**		
	Without IBF	**With IBF**	**With IBF and overhead cabling**	**Without IBF**	**With IBF**	**With IBF and overhead cabling**
Weight in kgs.	872.71	967.97	1011.06	887.68	982.93	1026.03
Weight in lbs.	1924	2134	2229	1957	2167	2262
Height with covers **Width with covers** **Depth with covers**	201.3 cm (79.26 in.) (42 EIA) 78.4 cm (30.87 in.) 157.5 cm (62.0 in.)					
Height reduction **Width reduction**	180.9 cm (71.2 in.) None					
Machine area **Service clearance**	.97 square meters (10.42 square feet) 3.16 square meters (30.38 square feet) (IBF contained within the frame)					
Note: The width increases by 15.2 cm (6 in.) if overhead I/O cabling is configured.						

11.2.2 Three-in-one (3-in-1) bolt-down kit

A bolt-down kit can be ordered for the z114 frame.The kit provides hardware to enhance the ruggedness of the frame and to tie down the frame to a concrete floor. The kit is offered in the following configurations:

- The Bolt-Down Kit for a raised floor installation (FC 8012) provides frame stabilization and bolt-down hardware for securing the frame to a concrete floor beneath the raised floor. The kit will cover raised floor heights from 22.8 cm (9 in.) to 91.4 cm (36 in.).
- The Bolt-Down Kit for a non-raised floor installation (FC 8013) provides frame stabilization and bolt-down hardware.

The kits help to secure the frame and its content from damage when exposed to shocks and vibrations, such as those generated by an earthquake.

11.3 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,* GC28-2611-00.

11.3.1 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 required blades determines the actual components that are required for each configuration. The number of BladeCenters and racks are generated by the quantity of blades (see Table 11-4).

Table 11-4 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
70	5	3
84	6	3
98	7	4
112	8	4

A zBX can be populated by up to 112 Power 701 blades. A maximum of 28 IBM BladeCenter HX5 blades can be installed in a zBX. For DataPower blades, the maximum number is 28. Note that the DataPower blade is a double-wide blade.

11.3.2 zBX power components

The zBX has its own power supplies and cords, which are independent of the z114 server power. Depending on the configuration of the zBX, up to 16 client-supplied power feeds might be required. A fully configured four-rack zBX has 16 power distribution units (PDUs). The zBX has these power specifications:

- 50/60Hz AC power
- Voltage (240V)
- Both single-phase and three-phase wiring

PDUs and power cords

The following PDU options are available for the zBX:

- FC 0520 - 7176: Model 3NU with attached Line-cord (US)
- FC 0521 - 7176: Model 2NX (worldwide (WW))

The following power cord options are available for the zBX:

- FC 0531: 4.3 m (14.1 ft.), 60A/208V, US Line-cord, Single Phase
- FC 0532: 4.3 m (14.1 ft.), 63A/230V, non-US Line-cord, Single Phase
- FC 0533: 4.3 m (14.1 ft.), 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 voltage ensures that the BladeCenter maximum of 240V is not exceeded.

Power installation considerations

Each zBX BladeCenter operates from two fully redundant PDUs that are installed in the rack with the BladeCenter. Each PDU has its own line cords (see Table 11-5), allowing the system to survive the loss of client 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 11-5 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 need to be powered from separate building PDUs.

Actual power consumption depends on the zBX configuration in terms of the number of BladeCenters and blades installed.

Input power in kVA is equal to the outgoing 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.

11.3.3 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 the 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 the internal BladeCenter unit temperature.
- If a blower fails, the remaining blower will run full speed and continue to cool the BladeCenter unit and blade servers.

Heat released by configurations

Table 11-6 shows the typical heat that is released by the various zBX solution configurations.

Table 11-6 zBX power consumption and heat output

Number of blades	Maximum 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 11-2 on page 337) is a more cost-effective solution than adding another air conditioning unit.

Important: 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 that is 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 the air exiting the back of a zBX rack.

Figure 11-2 shows the IBM Rear Door Heat eXchanger details.

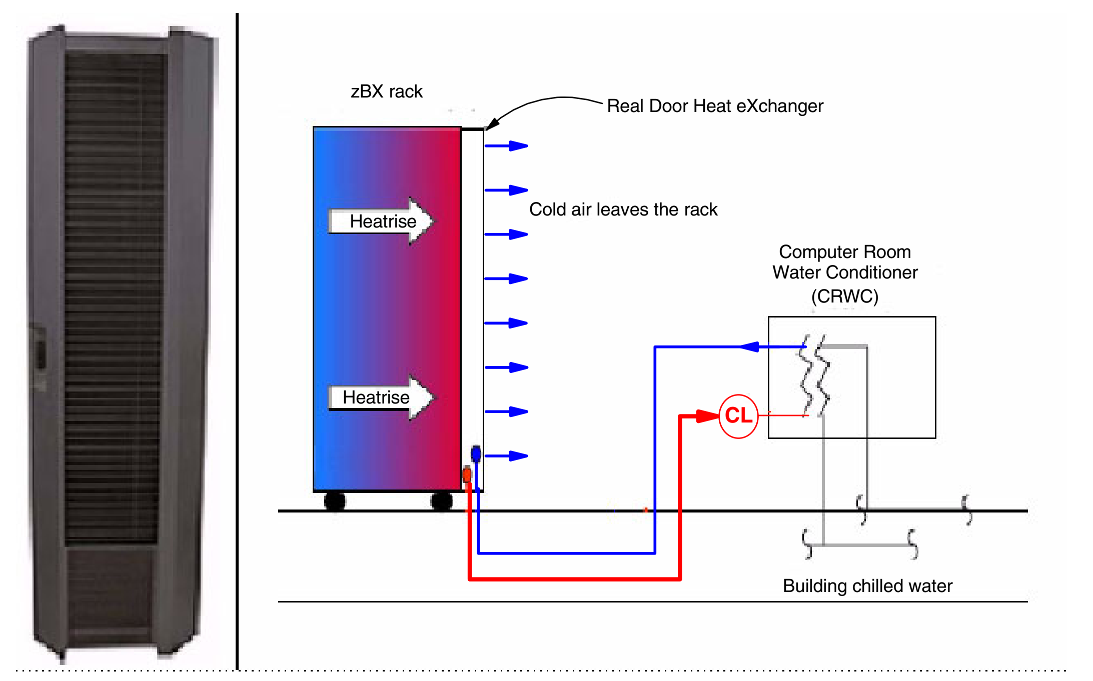


Figure 11-2 Rear Door Heat eXchanger (left) and functional diagram

The IBM Rear Door Heat eXchanger also offers a convenient way to handle hazardous “hot spots”, which might help you lower the total energy cost of your data center.

11.3.4 zBX physical specifications

The zBX solution is delivered either with one rack (Rack B) or four racks (Rack B, C, D, and E). Table 11-7 shows the physical dimensions of the zBX minimum and maximum solutions.

Table 11-7 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. Select this feature when it has been deemed necessary for delivery clearance purposes. Order it if you have doorways with openings less than 1941 mm (76.4 in.) high. It accommodates doorway openings as low as 1832 mm (72.1 in.).

zBX weight

Table 11-8 on page 338 shows the maximum weights of fully populated zBX racks and BladeCenters.

Table 11-8 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)

> **Rack weight:** A fully configured Rack B is heavier than a fully configured Rack C, D, or E, because Rack B has the two Top of Rack switches (TORs) installed.

For a complete view of the physical requirements, see *zBX Model 002 Installation Manual - Physical Planning,* GC28-2611-00.

11.4 Energy management

In this section, we discuss the elements of energy management in areas of tooling to help you understand the requirement for power and cooling, monitoring and trending, and reducing power consumption. Figure 11-3 shows the IBM Rear Door Heat eXchanger details.

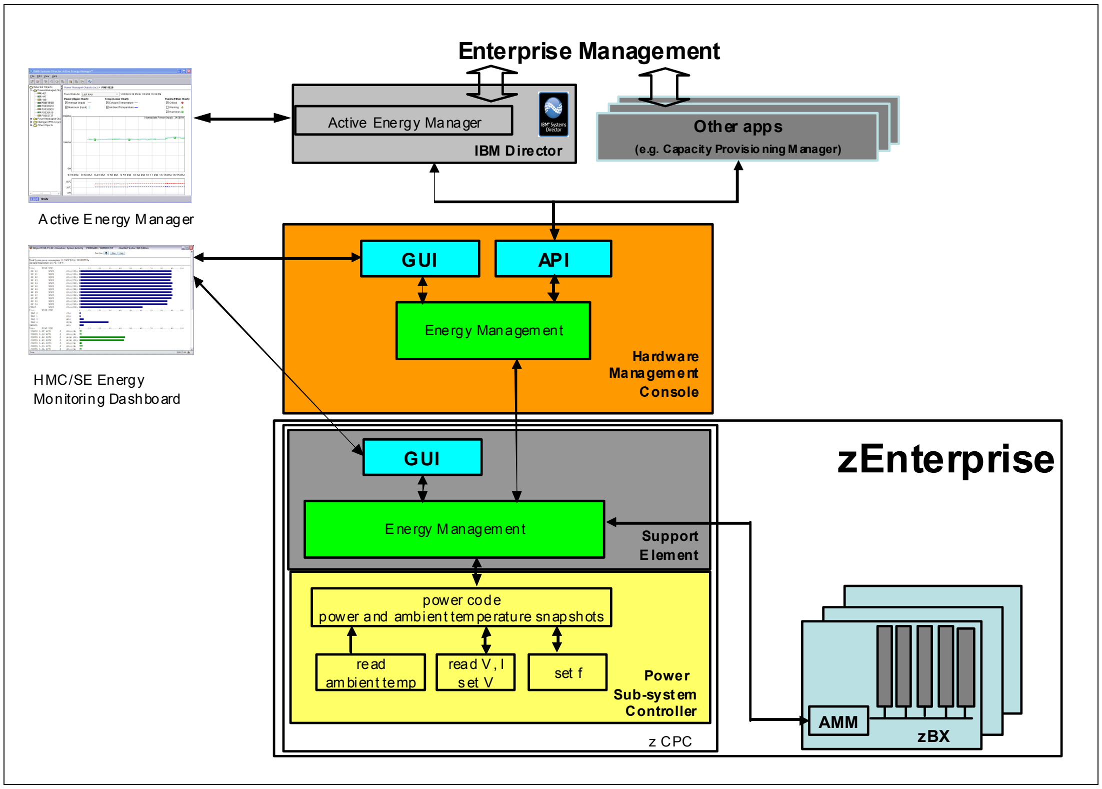


Figure 11-3 zEnterprise Energy Management

The hardware components in the zCPC and the optional zBX are monitored and managed by the Energy Management component in the Support Element (SE) and HMC. The GUI of the SE and the HMC provide views, for instance, with the System Activity Display or Monitors

Dashboard. Through an Simple Network Management Protocol (SNMP) API, energy information is available to, for instance, Active Energy Manager, which is a plug-in of IBM Systems Director. See 11.4.4, “IBM Systems Director Active Energy Manager” on page 341 for more information.

When Unified Resource Manager features are installed (see 12.7.1, “Unified Resource Manager” on page 364), several monitoring and control functions can be used to perform Energy Management. We discuss more details in 11.4.5, “Unified Resource Manager: Energy Management” on page 342.

A few aids are available to plan and monitor the power consumption and heat dissipation of the z114. This section summarizes the tools that are available to plan and monitor the energy consumption of the z114:

- Power estimation tool
- Query maximum potential power
- System Activity Display and Monitors Dashboard
- IBM Systems Director Active Energy Manager™

11.4.1 Power estimation tool

The power estimation tool for System z servers is available through the IBM Resource Link website:

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

The tool provides an estimate of the anticipated power consumption of a machine model, given its configuration. You enter the machine model, memory size, number of I/O cages, I/O drawers, and quantity of each type of I/O feature card. The tool outputs an estimate of the power requirements for that configuration.

If you have a registered machine in Resourcelink, you can access the Power Estimator tool via the machine information page of that particular machine. In the Tools section of Resourcelink, you also can enter the Power Estimator and enter any system configuration for which you want to calculate its power requirements. This tool helps with power and cooling planning for installed and planned System z servers.

11.4.2 Query maximum potential power

The maximum potential power that is used by the system is less than the *label power*, as depicted in the atypical power usage report that is shown in Figure 11-4 on page 340. The *Query maximum potential power* function shows what *your* systems maximum power usage and heat dissipation can be, so that you are able to allocate the proper power and cooling resources.

The output values of this function for *maximum potential power* and *maximum potential heat load* are displayed on the Energy Management tab of the CPC Details view of the HMC.

This function enables operations personnel with no System z knowledge to query the maximum possible power draw of the system. The implementation helps to avoid capping enforcement through dynamic capacity reduction. The client controls are implemented in the HMC, the SE, and the Active Energy Manager. Use this function in conjunction with the Power Estimation tool that allows for pre-planning for power and cooling requirements. See 11.4.1, “Power estimation tool” on page 339.

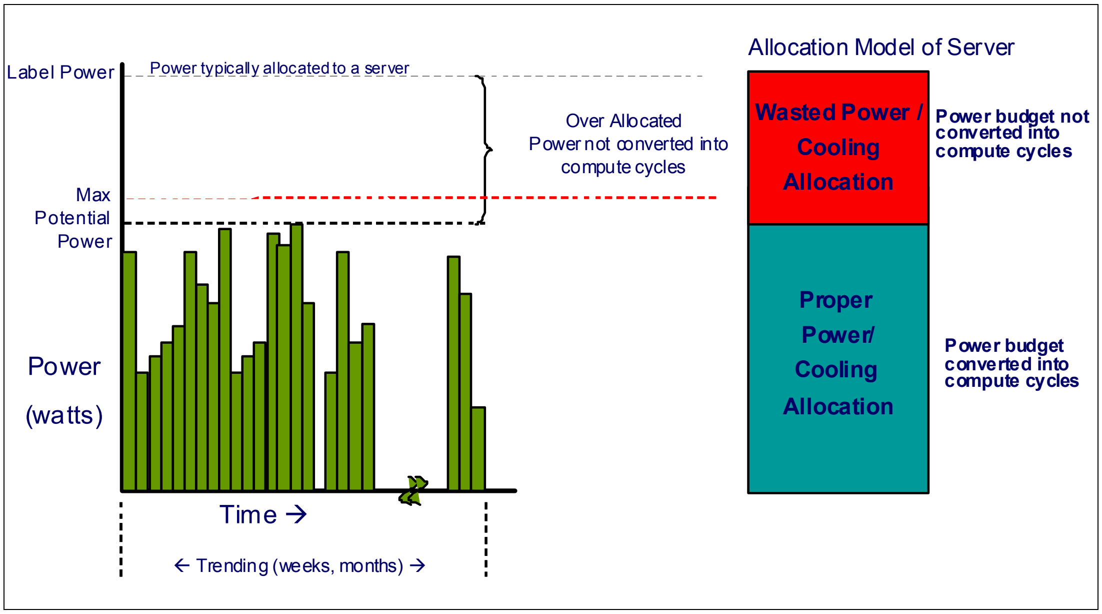
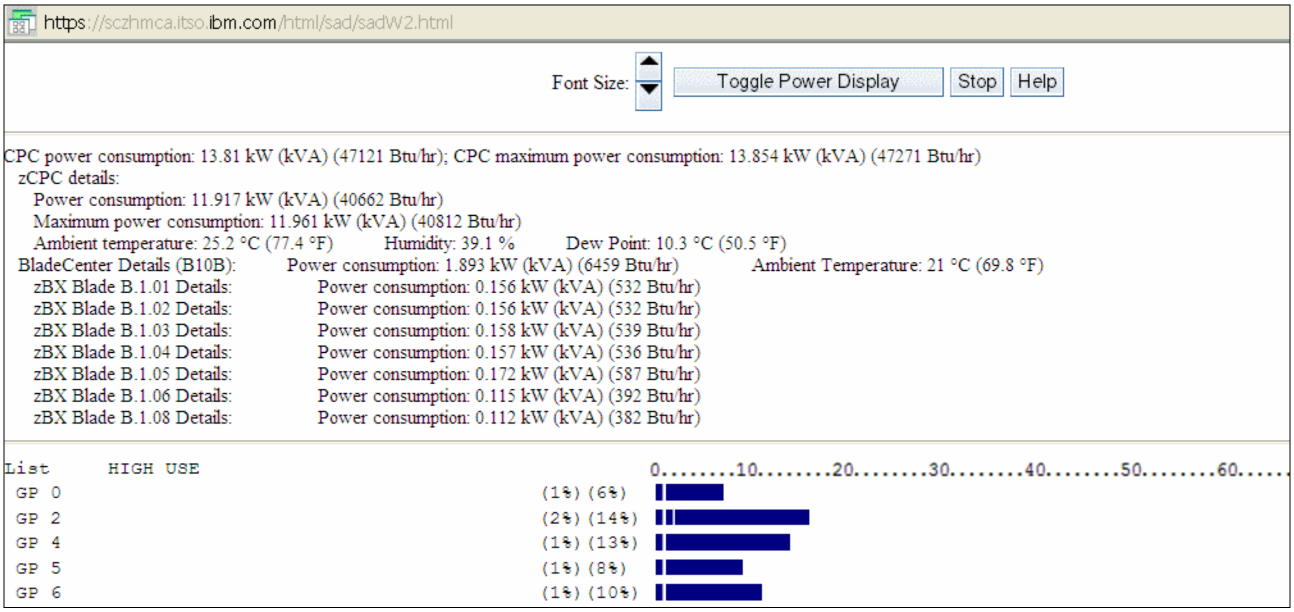


Figure 11-4 Maximum potential power

11.4.3 System Activity Display and Monitors Dashboard

The System Activity Display presents you with the current power usage, among other information, that is shown in Figure 11-5.



```
CPC power consumption: 13.81 kW (kVA) (47121 Btu/hr); CPC maximum power consumption: 13.854 kW (kVA) (47271 Btu/hr)
 zCPC details:
   Power consumption: 11.917 kW (kVA) (40662 Btu/hr)
   Maximum power consumption: 11.961 kW (kVA) (40812 Btu/hr)
   Ambient temperature: 25.2 °C (77.4 °F)        Humidity: 39.1 %        Dew Point: 10.3 °C (50.5 °F)
 BladeCenter Details (B10B):       Power consumption: 1.893 kW (kVA) (6459 Btu/hr)        Ambient Temperature: 21 °C (69.8 °F)
   zBX Blade B.1.01 Details:          Power consumption: 0.156 kW (kVA) (532 Btu/hr)
   zBX Blade B.1.02 Details:          Power consumption: 0.156 kW (kVA) (532 Btu/hr)
   zBX Blade B.1.03 Details:          Power consumption: 0.158 kW (kVA) (539 Btu/hr)
   zBX Blade B.1.04 Details:          Power consumption: 0.157 kW (kVA) (536 Btu/hr)
   zBX Blade B.1.05 Details:          Power consumption: 0.172 kW (kVA) (587 Btu/hr)
   zBX Blade B.1.06 Details:          Power consumption: 0.115 kW (kVA) (392 Btu/hr)
   zBX Blade B.1.08 Details:          Power consumption: 0.112 kW (kVA) (382 Btu/hr)

List     HIGH USE                                        0........10........20........30........40........50........60.....
 GP  0                                         (1%) (6%)
 GP  2                                         (2%) (14%)
 GP  4                                         (1%) (13%)
 GP  5                                         (1%) (8%)
 GP  6                                         (1%) (10%)
```

Figure 11-5 Power usage on the System Activity Display

The Monitors Dashboard of the HMC allows you to display power and other environmental data. It also allows you to start a Dashboard Histogram Display, where you can trend a particular value of interest, such as the power consumption of a blade or the ambient temperature of a zCPC.

11.4.4 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 client. Active Energy Manager is an industry-leading cornerstone of the IBM energy management framework.

Active Energy Manager Version 4.3.1 is a plug-in to IBM Systems Director Version 6.2.1 and is available for installation on Linux on System z. It can also run on Windows, Linux on IBM System x, and AIX and Linux on IBM Power Systems™. For more specific information, see *Implementing IBM Systems Director Active Energy Manager 4.1.1*, SG24-7780. Version 4.3.1 supports IBM zEnterprise System and its optional attached zBX.

Use Active Energy Manager to monitor the power and environmental values of resources, not only System z, but also other IBM products, such as IBM Power Systems, IBM System x, or devices and hardware that are acquired from another vendor. You can view historical trend data for resources, calculate energy costs and savings, view properties and settings for resources, and view active energy-related events.

Active Energy Manager does not directly connect to the System z servers, but it attaches through a LAN connection to the HMC. See Figure 11-3 on page 338 and 12.2, “HMC and SE connectivity” on page 347. Active Energy Manager discovers the HMC managing the server by using a discovery profile that specifies the HMC’s IP address and the SNMP credentials for that System z HMC. As the system is discovered, the System z servers that are managed by the HMC are also discovered.

Active Energy Manager 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 manage the power of their systems while lowering the cost of computing.

The following data is available through Active Energy Manager:

- System name, machine type, model, serial number, and firmware level of the System z servers and optional zBX that is attached to IBM zEnterprise Systems.
- Ambient temperature
- Exhaust temperature
- Average power usage
- Peak power usage
- Limited status and configuration information. This information helps to explain the changes to the power consumption, which are called *events*:
  - Changes in fan speed
  - Changes between power-off, power-on, and IML-complete states
  - CBU records expirations

IBM Systems Director Active Energy Manager provides clients 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, which was developed by IBM, to help identify how much actual power is being used and the temperature of the system.

11.4.5 Unified Resource Manager: Energy Management

The energy management capabilities for Unified Resource Manager that can be used in an ensemble depend on which suite is installed in the ensemble:

- Manage suite (feature code 0019)
- Automate suite (feature code 0020)

Manage suite

For energy management, the manage suite focuses on the monitoring capabilities. Energy monitoring can help you better understand the power and cooling demand of the zEnterprise System. It provides complete monitoring and trending capabilities for the z114 and the zBX using one or multiple of the following options:

- Monitors dashboard
- Environmental Efficiency Statistics
- Details view

Automate suite

The Unified Resource Manager offers multiple energy management tasks as part of the automate suite. These tasks allow you to actually change system behaviors for optimized energy usage and energy saving:

- Power cap
- Group power cap

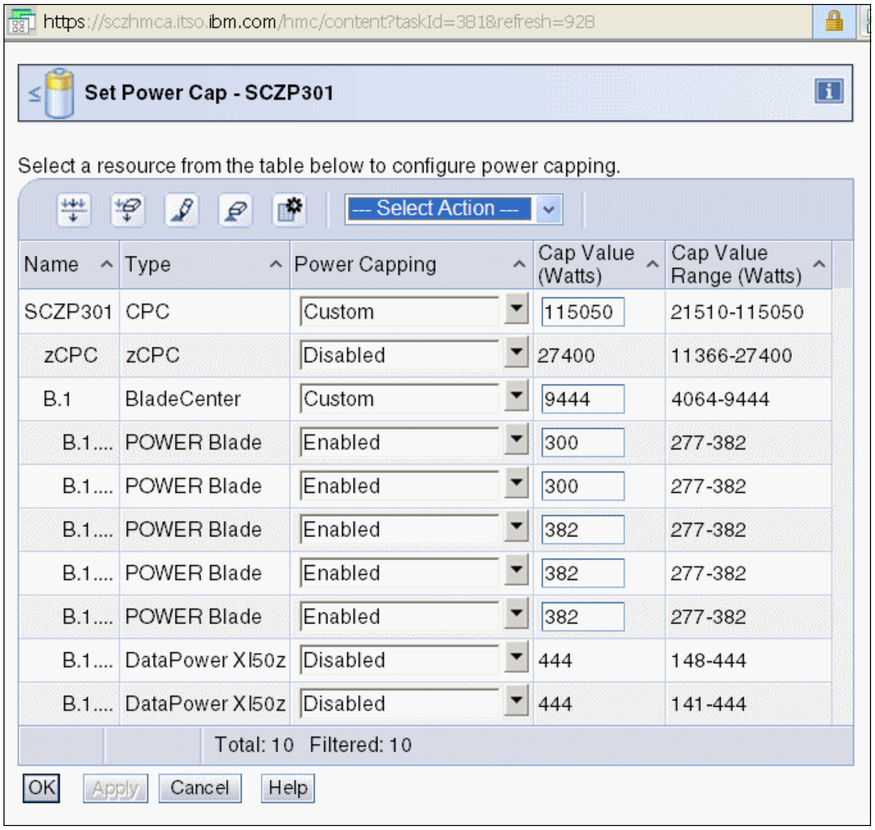
Various options are presented, depending on the scope that is selected inside the Unified Resource Manager GUI.

Set Power Cap

The power cap function can be used to limit the maximum amount of energy that is used by the ensemble. If enabled, it enforces power caps for the hardware by actually throttling the processors in the system.

The Unified Resource Manager shows all components of an ensemble in the Set Power Cap window, as seen in Figure 11-6 on page 343. Because not all components that are used in a specific environment can support power capping, only those components that are marked as “enabled” can actually perform power capping functions.

A zCPC does not support power capping, as opposed to specific blades, which can be power-capped. When capping is enabled for a zCPC, this capping level is used as a threshold for a warning message that informs you that the zCPC went above the set cap level. Being under the limit of the cap level is equal to the maximum potential power value (see 11.4.2, “Query maximum potential power” on page 339).


https://sczhmca.itso.ibm.com/hmc/content?taskId=381&refresh=928

Set Power Cap - SCZP301

Select a resource from the table below to configure power capping.

Name	Type	Power Capping	Cap Value (Watts)	Cap Value Range (Watts)
SCZP301	CPC	Custom	115050	21510-115050
zCPC	zCPC	Disabled	27400	11366-27400
B.1	BladeCenter	Custom	9444	4064-9444
B.1....	POWER Blade	Enabled	300	277-382
B.1....	POWER Blade	Enabled	300	277-382
B.1....	POWER Blade	Enabled	382	277-382
B.1....	POWER Blade	Enabled	382	277-382
B.1....	POWER Blade	Enabled	382	277-382
B.1....	DataPower XI50z	Disabled	444	148-444
B.1....	DataPower XI50z	Disabled	444	141-444

Total: 10 Filtered: 10

Figure 11-6 Set Power Cap panel

More information about energy management with Unified Resource Manager is available in *IBM zEnterprise Unified Resource Manager,* SG24-7921.

12

Hardware Management Console

The Hardware Management Console (HMC) supports many functions and tasks to extend the management capabilities of the IBM zEnterprise System. 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 Support Elements (SE), which then issue commands to their central processor complexes (CPCs).

We first describe the HMC and SE in general. We discuss the HMCs that manage ensembles later in this chapter.

We cover 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 and its logical partitions (LPARs), and provides support applications. At least one HMC is required to operate a zEnterprise CPC. If the zEnterprise CPC is defined as a member of an ensemble, a pair of HMCs is required (a primary and an alternate). See 12.7, “HMC in an ensemble” on page 363 for a description and the prerequisites.

An HMC can manage multiple System z servers, and it can be located at a local or a remote site. However, when a zEnterprise CPC is defined as a member of an ensemble, certain restrictions apply. See 12.7, “HMC in an ensemble” on page 363.

The Support Elements (SEs) are two integrated ThinkPads that are supplied with the System z server. One ThinkPad is always the active SE and the other ThinkPad is a strictly alternate element. The SEs are closed systems, just as the HMCs, 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, a zEnterprise System ships with HMC Version 2.11.1, which is capable of supporting various System z server types. Many functions that are available on Version 2.11.0 and later are only supported when connected to a zEnterprise CPC. HMC Version 2.11.1 supports the CPC and SE versions that are shown in Table 12-1.

Table 12-1 IBM zEnterprise System HMC server support summary

Server	Machine type	Minimum firmware driver	Minimum SE version
z114	2818	93	2.11.1
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

12.1.1 Tree Style User and Classic Style User interfaces

Two user interface styles are provided with an HMC. The Tree Style User Interface (default) uses a hierarchical model that is popular in newer operating systems and features context-based task launching. The Classic Style User Interface uses the drag-and-drop interface style.

Tutorial Note: IBM Resourcelink provides tutorials, which demonstrate how to change from the Classic Style User Interface to the Tree Style User Interface. The tutorial introduces you to the function of the Tree Style Interface on the HMC:

https://www-304.ibm.com/servers/resourcelink/edu03010.nsf/pages/z196Courses?OpenDocument&pathID=

Registration is required to access the IBM Resource Link.

12.2 HMC and SE connectivity

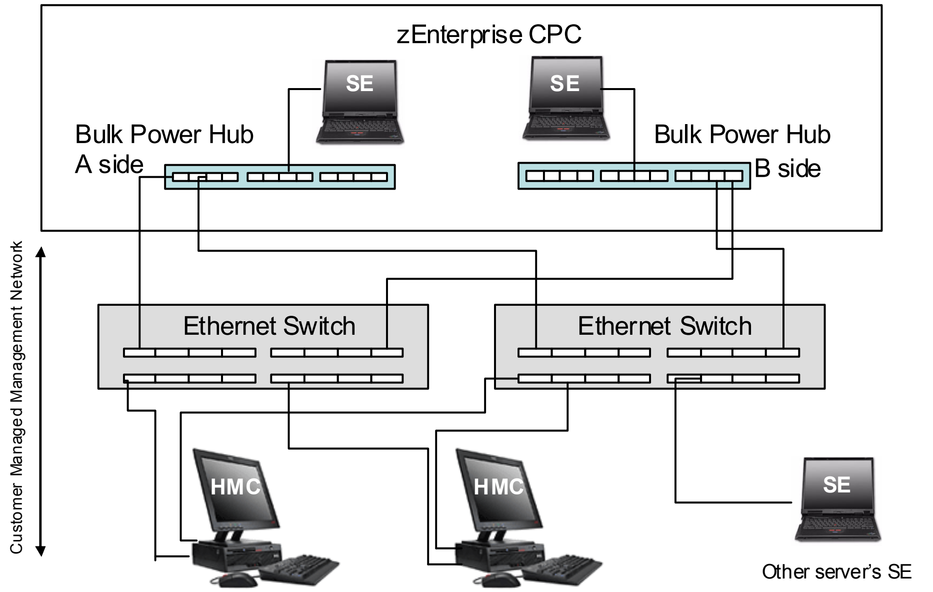
The HMC has two Ethernet adapters. Each SE has one Ethernet adapter and both SEs are connected to the same Ethernet switch. The Ethernet switch (FC 0070[1]) can be supplied with every system order. Additional Ethernet switches (up to a total of ten) can be added. The HMC Version 2.11.1 now supports up to two 10/100/1000 Mbps Ethernet LANs.

The switch is a stand-alone unit that is located outside the frame. It operates on building AC power. A client-supplied switch can be used if it matches the IBM specifications.

The internal LAN for the SEs on the zEnterprise CPC connects to the Bulk Power Hub. The HMC must be connected to the Ethernet switch through one of its Ethernet ports. Only the switch can be connected to the client ports J01 and J02 on the Bulk Power Hub. With respect to the zEnterprise System network topology architecture, this network is referred to as the Customer-Managed Management Network. Other server SEs can also be connected to the switches. To provide redundancy for the HMCs, two switches are preferable, as shown in Figure 12-1 on page 348.

For more information, see *The zEnterprise 114 Installation Manual for Physical Planning,* GC28-6907.

[1] Ethernet switch FC 0070 is not available in certain countries.



- zEnterprise CPC's SE is always connected to the Bulk Power Hub
- Switches are connected to J01 and J02 on the Bulk Power Hubs (two switches recommended)
- Other server's SEs (not zEnterprise) may be connected to switches

Figure 12-1 HMC to SE connectivity

The HMC and SE have exploiters that either require or can take advantage of broadband connectivity to the Internet and your corporate intranet. Various 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 that 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 on page 349. 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.

Security: The configuration of network components, such as routers or firewall rules, is beyond the scope of this document. Any time that networks interconnect, 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 at the IBM Resource Link:

https://www-304.ibm.com/servers/resourcelink/lib03011.nsf/pages/zHmcSecurity/$file/zHMCSecurity.pdf

Registration is required to access the IBM Resource Link.

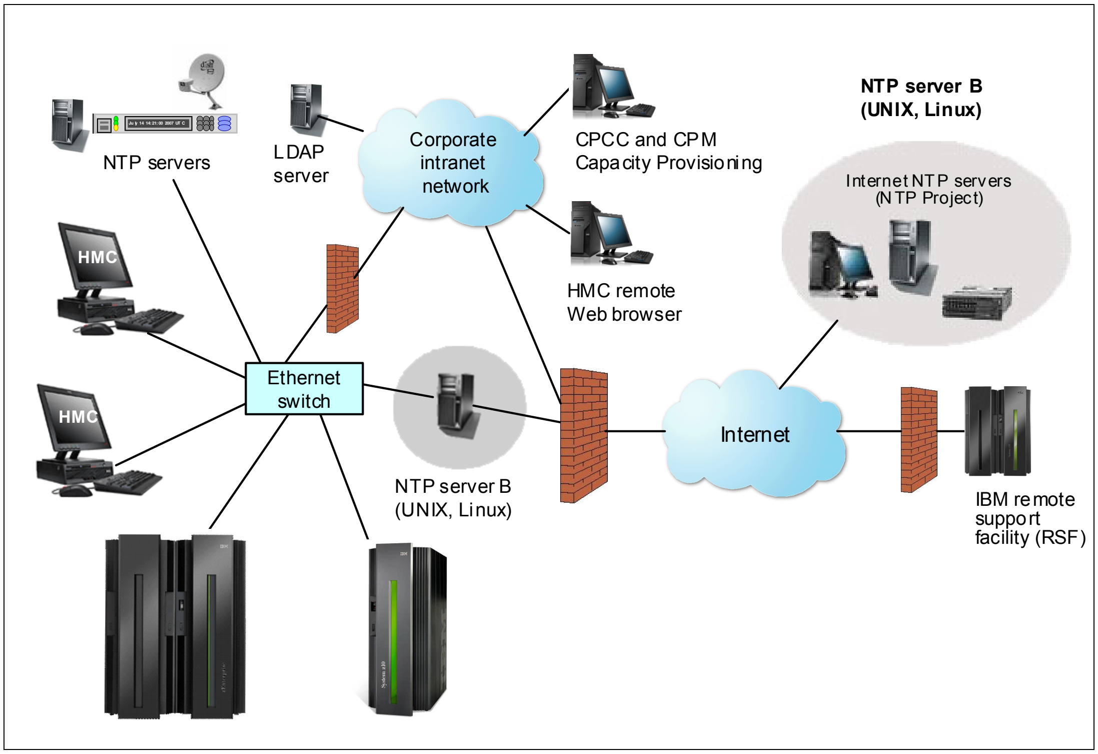

NTP servers
LDAP server
Corporate intranet network
CPCC and CPM Capacity Provisioning
NTP server B (UNIX, Linux)
Internet NTP servers (NTP Project)
HMC
HMC
HMC remote Web browser
Ethernet switch
NTP server B (UNIX, Linux)
Internet
IBM remote support facility (RSF)

Figure 12-2 HMC connectivity

You must plan the HMC and SE network connectivity carefully to allow for current and future use. Many of the System z capabilities benefit from the various network connectivity options that are available. For example, the following functions that are available to the HMC depend on the HMC connectivity:

- Lightweight Directory Access Protocol (LDAP) support, which can be used for HMC user authentication
- Server Time Protocol (STP) and Network Time Protocol (NTP) client/server support
- Remote Support Facility (RSF), which is available through the HMC with an Internet-based connection, providing increased bandwidth as compared to dial-up

> **Statement of Direction for dial-up support:** The IBM zEnterprise 196 and the zEnterprise 114 are the last System z CPCs to support dial-up modems for use with the Remote Support Facility (RSF) and also with the External Time Source (ETS) option of Server Time Protocol (STP).
>
> You can use the currently available Network Time Protocol (NTP) server option for ETS, as well as Internet time services that are available using broadband connections, to provide the same degree of accuracy as dial-up time services.
>
> *Enterprises need to begin migrating from dial-up modems to broadband for RSF connections.*
>
> All statements regarding IBM future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

- Enablement of the Simple Network Management Protocol (SNMP) and Common Information Model (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 addresses have the following characteristics:

- 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 configurations:

- Statically assigned IPv4 addresses or statically assigned IPv6 addresses
- Dynamic Host Configuration Protocol (DHCP)-assigned IPv4 addresses or DHCP-assigned IPv6 addressses
- Autoconfigured IPv6:
  - Link-local is assigned to every network interface.
  - Router-advertised, which is broadcast from the router, can be combined with a Media Access Control (MAC) address to create a unique address.
  - Privacy extensions can be enabled for these addresses as a way to avoid using the MAC address as part of the address to ensure uniqueness.

An SE can have the following IP addresses:

- Statically assigned IPv4 or statically assigned IPv6
- 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 can be used to identify the IP addresses (IPv4 and IPv6) that are being used by the HMC to communicate to the CPC SEs.

IPv6 addresses are easily identified. A fully qualified IPV6 address has 16 bytes, which is written as eight 16-bit hex blocks that are separated by colons, as shown in the following example:

```
2001:0db8:0000:0000:0202:b3ff:fe1e:8329
```

Because many IPv6 addresses are not fully qualified, shorthand notation can be used. Shorthand notation 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 this way:

```
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. RSF provides the following types of communication:

- Problem reporting and repair data
- Fix and firmware delivery to the service processor, HMC, and SE
- Hardware inventory data
- On-demand enablement

The HMC can be configured to send hardware service-related information to IBM by using a dial-up connection over a modem or by using a broadband (Internet) connection. Using a broadband connection offers these advantages:

- Significantly faster transmission speed
- Ability to send more data on an initial problem request, potentially resulting in more rapid problem resolution
- Reduced client expense (for example, the cost of a dedicated analog telephone line)
- Greater reliability
- Availability of Secure Sockets Layer (SSL)/Transport Layer Security (TLS) secured protocol

Broadband connections accelerate support capabilities and normal maintenance time for downloads. IBM intends to withdraw dial-up connection support for future servers. With IBM zEnterprise Systems, there will be limitations to certain firmware components so that they can only be updated by a broadband connection or media, but not by a dial-up connection.

Statement of direction for dial-up support: The IBM zEnterprise 196 and the zEnterprise 114 are the last System z CPCs to support dial-up modems for use with the Remote Support Facility (RSF).

Enterprises need to begin migrating from dial-up modems to broadband for RSF connections.

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

If both types of connections are configured, the Internet will be tried first and, if it fails, the modem is used.

The following security characteristics are in effect regardless of the connectivity method that is chosen:

- RSF requests are always initiated from the HMC to IBM. An inbound connection is never initiated from the IBM Service Support System.
- All data that is transferred between the HMC and the IBM Service Support System is encrypted in a high-grade SSL encryption.
- When initializing the SSL-encrypted connection, the HMC validates the trusted host by its digital signature that is issued for the IBM Service Support System.
- Data that is sent to the IBM Service Support System consists solely of hardware problems and configuration data. No application or client data is transmitted to IBM.

Broadband RSF connection setup:

The following document, which is available on the IBM Resource Link, introduces the benefits of broadband RSF and SSL/TLS secured protocol and includes a sample configuration for the broadband RSF connection:

https://www-304.ibm.com/servers/resourcelink/lib03011.nsf/pages/zHmcBroadbandRsfOverview

Registration is required to access the IBM Resource Link.

12.4 HMC remote operations

The zEnterprise CPC HMC application simultaneously supports one local user and any number of remote users. Remote operations provide the same interface that is used by a local HMC operator. There are two ways to perform remote manual operations:

- Using a Remote HMC

  A remote HMC is an HMC that is on a separate 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 that is connected to a local HMC is determined by the scope of control that is needed. A remote HMC can control only a specific set of objects, but a web browser that is 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 the web browser control of a local HMC, but dial-up 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 349).

A remote HMC requires TCP/IP connectivity to each SE to be managed. Therefore, any existing client-installed firewall between the remote HMC and its managed objects must permit communications between the HMC and the SE. For service and support, the remote HMC also requires connectivity to IBM, or to another HMC with connectivity to IBM.

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 functions that require physical access to the diskette or DVD media. The user interface in the browser is the same interface 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 Point-to-Point Protocol (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 they 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 HMC media support

The HMC provides a CD-ROM drive, and with HMC Version 2.11.0, a USB flash memory drive was introduced as an alternative to DVD-RAM. The tasks that require access to a DVD-RAM now have the ability to access a USB flash memory drive. There can be more than one USB flash memory drive inserted into the console.

You can use USB flash memory for Backup/Restore and also to upload and download data, such as input/output configuration data set (IOCDS) configuration data or to retrieve HMC screen captures, activity, or audit logs. You can use CD-ROM for serviceability tasks (that is, Machine Change Levels (MCLs)) and also for installing z/VM or Linux.

12.6 HMC and SE key capabilities

The zEnterprise CPC comes with the HMC application Version 2.11.1. We encourage you to use the What’s New Wizard on the HMC, to explore the new features that are available for each release. For a complete list of HMC functions, see *System z HMC Operations Guide Version 2.11.1,* SC28-6895.

12.6.1 CPC management

The HMC is the primary place for central processor complex (CPC) control. For example, to define hardware to the zEnterprise CPC, the I/O configuration data set (IOCDS) must be defined. The IOCDS contains the definitions of logical partitions (LPARs), channel subsystems (CSSs), control units, and devices and their accessibility from LPARs. IOCDS can be created and put into production from the HMC.

The zEnterprise CPC is powered on and off from the HMC. The HMC is used to initiate the 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 storage pool, and the IOCDS is loaded and initialized into the hardware system area.

The Hardware messages task displays hardware-related messages at the CPC level, LPAR level, or SE level, or hardware messages that relate to the HMC.

12.6.2 LPAR management

Use the HMC to define logical partition (LPAR) properties, such as how many processors there are of each type, how many processors are reserved, or how much memory is assigned to each processor. These parameters are defined in LPAR profiles, and they are stored on the SE.

Because Processor Resource/Systems Manager (PR/SM) has to manage the LPAR access to the processors and the 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 an LPAR is active and an operating system is running in it, you can use the HMC to dynamically change certain LPAR parameters. The HMC 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 LPARs at the scheduled time.

Channel paths can be dynamically configured on and off, as needed for each partition, from an HMC.

The “Change LPAR Controls” task for the zEnterprise CPC has the ability to export the Change LPAR Controls table data to a `.csv` formatted file. This support is available to a user when the user is connected to the HMC remotely by a web browser.

Partition capping values can be scheduled and are specified on the “Change LPAR Controls scheduled operation” support. Viewing the details about an existing LPAR Controls schedule operation change 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 the zEnterprise CPC, the SNMP and CIM API allow dynamic changes to both the group members and group capacity setting.

12.6.3 Operating system communication

The Operating System Messages task displays messages from an LPAR. You can also enter operating system commands and interact with the system.

The HMC also provides integrated 3270 and ASCII consoles so that you can access an operating system without requiring other network or network devices, such as TCP/IP or control units.

12.6.4 SE access

Being physically close to an SE is not necessary to use it. The HMC can be used to remotely to access the SE in the same interface that is provided in the SE.

The HMC enables you to perform the following tasks:

- Synchronize content of the primary SE to the alternate SE.
- Determine whether a switch from primary to the alternate can be performed.
- Switch between the primary and alternate SEs.

12.6.5 Monitoring

In this section, we discuss monitoring considerations.

Monitor Task Group

The task group, which is called Monitor, holds the monitoring-related tasks for both the HMC and SE. In previous versions, various tasks, such as the Activity tasks, were located elsewhere. See Figure 12-3.

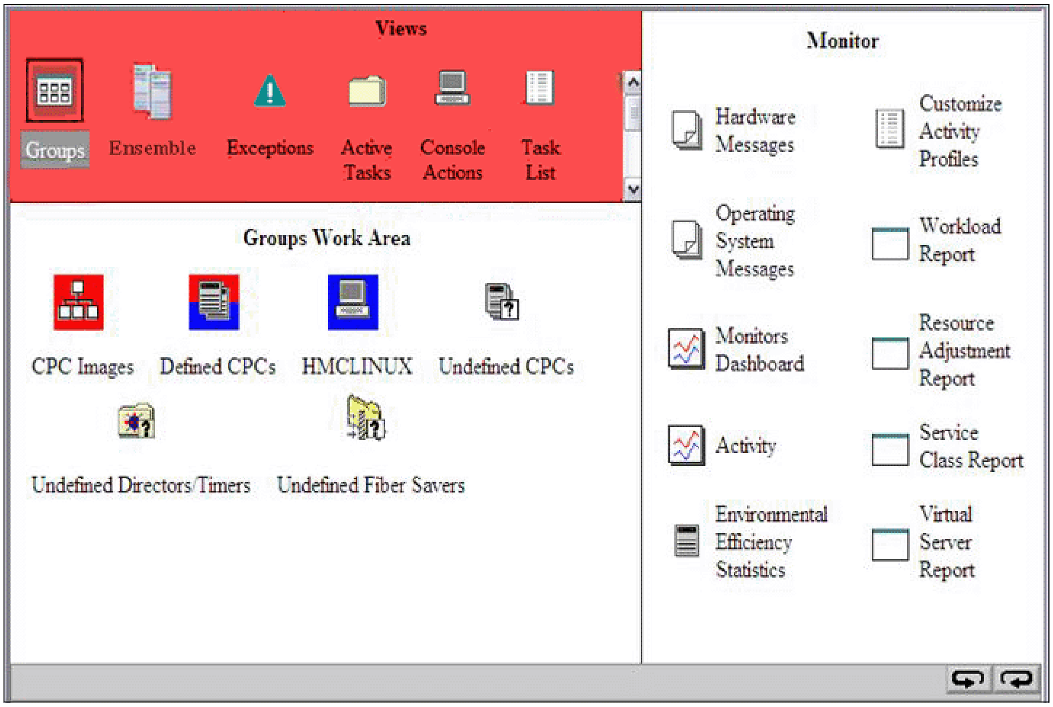


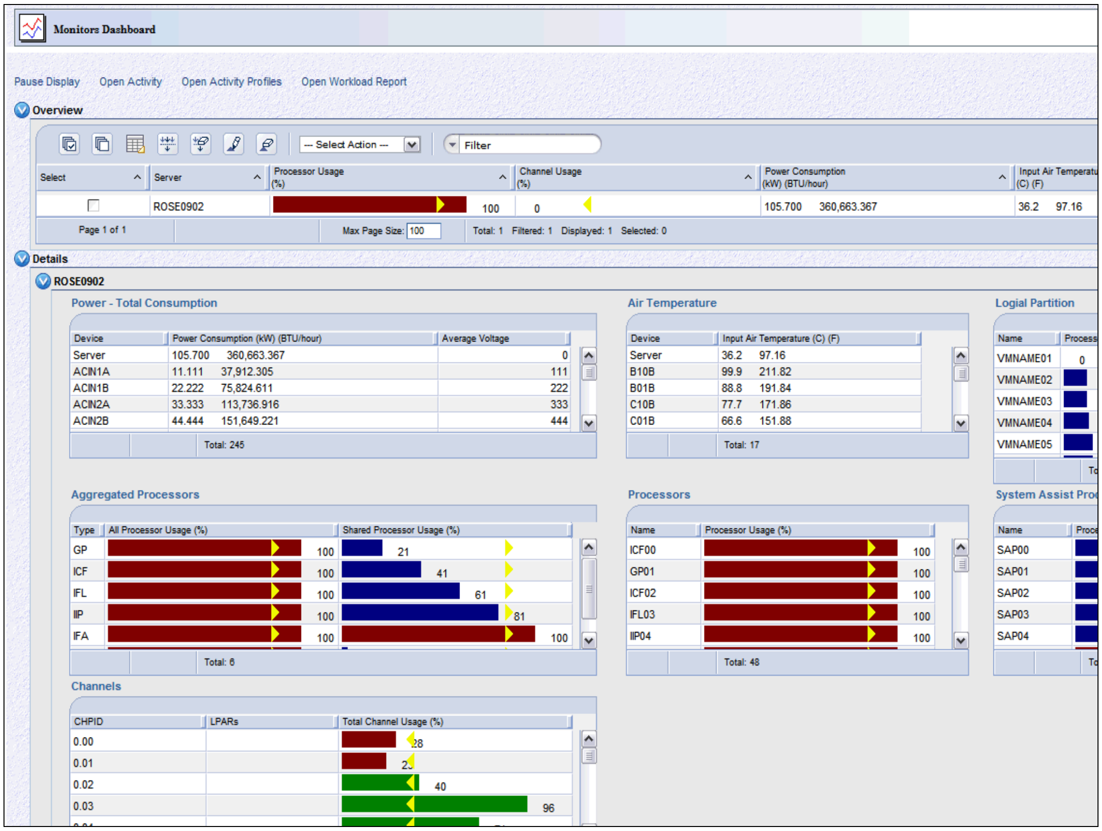
Figure 12-3 HMC Monitor task group

Use the System Activity Display (SAD) task on the HMC to monitor the activity of one or more CPCs. The task monitors the processor and channel usage. You can define multiple activity profiles. The task also includes power monitoring information, the power that is 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 must only be used by those individuals with extensive knowledge about zEnterprise System internals and three-phase electrical circuits. The weekly call-home data includes the power information for each power cord.

Monitors Dashboard task

In zEnterprise CPC, the Monitors Dashboard task in the Monitor task group (Figure 12-3 on page 355) 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 scopes. Multiple graphical ways of displaying data are available, such as history charts (Figure 12-4).



Select	Server	Processor Usage (%)	Channel Usage (%)	Power Consumption (kW) (BTU/hour)	Input Air Temperatu (C) (F)
□	ROSE0902	100	0	105.700 360,663.367	36.2 97.16

Page 1 of 1 Max Page Size: 100 Total: 1 Filtered: 1 Displayed: 1 Selected: 0

Device	Power Consumption (kW) (BTU/hour)	Average Voltage
Server	105.700 360,663.367	0
ACIN1A	11.111 37,912.305	111
ACIN1B	22.222 75,824.611	222
ACIN2A	33.333 113,736.916	333
ACIN2B	44.444 151,649.221	444

Total: 245

Device	Input Air Temperature (C) (F)
Server	36.2 97.16
B10B	99.9 211.82
B01B	88.8 191.84
C10B	77.7 171.86
C01B	66.6 151.88

Total: 17

Name	Process
VMNAME01	0
VMNAME02	
VMNAME03	
VMNAME04	
VMNAME05	

Type	All Processor Usage (%)	Shared Processor Usage (%)
GP	100	21
ICF	100	41
IFL	100	61
IIP	100	81
IFA	100	100

Total: 6

Name	Processor Usage (%)
ICF00	100
GP01	100
ICF02	100
IFL03	100
IIP04	100

Total: 48

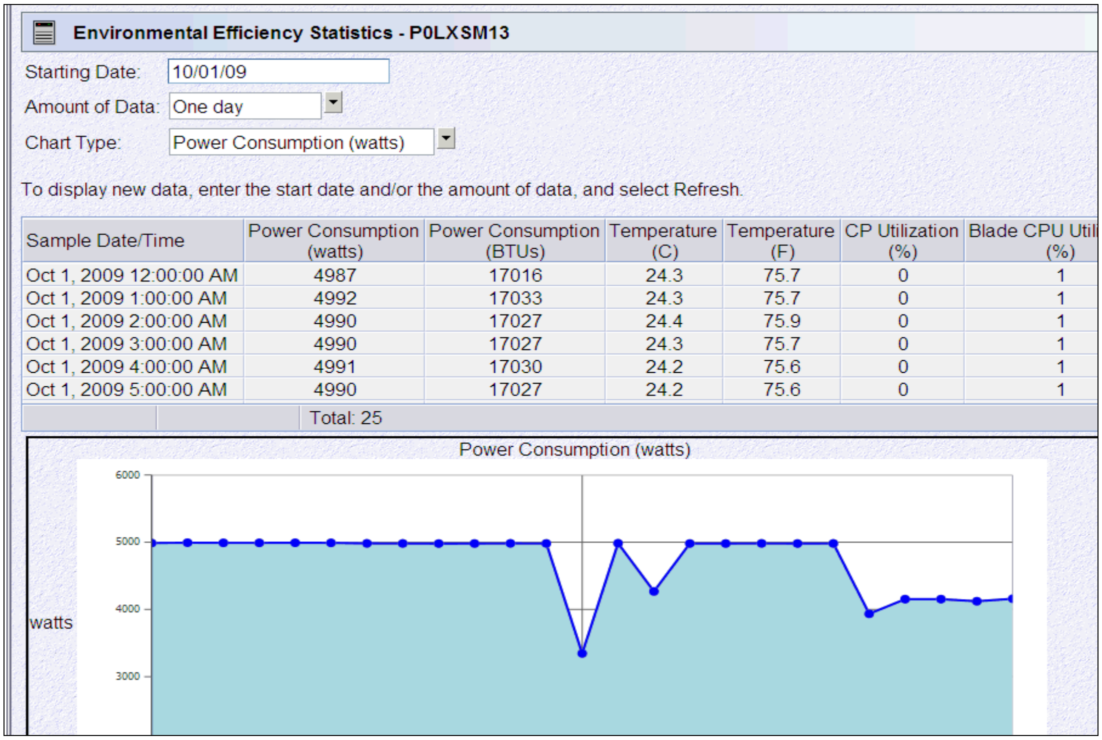
Name	Proce
SAP00	
SAP01	
SAP02	
SAP03	
SAP04	

CHPID	LPARs	Total Channel Usage (%)
0.00		28
0.01		2
0.02		40
0.03		96

Figure 12-4 Monitors Dashboard

Environmental Efficiency Statistics task

The Environmental Efficiency Statistics task (Figure 12-5 on page 357) is part of the Monitor task group and provides historical power consumption and thermal information for the zEnterprise CPC and is available on the HMC. The data is presented in table form and graphical (histogram) form. You can also export the data to a .csv formatted file so that it can be imported into tools, such as Microsoft 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 CPU Util (%)
Oct 1, 2009 12:00:00 AM	4987	17016	24.3	75.7	0	1
Oct 1, 2009 1:00:00 AM	4992	17033	24.3	75.7	0	1
Oct 1, 2009 2:00:00 AM	4990	17027	24.4	75.9	0	1
Oct 1, 2009 3:00:00 AM	4990	17027	24.3	75.7	0	1
Oct 1, 2009 4:00:00 AM	4991	17030	24.2	75.6	0	1
Oct 1, 2009 5:00:00 AM	4990	17027	24.2	75.6	0	1

Total: 25

Figure 12-5 Environmental Efficiency Statistics

12.6.6 Capacity on Demand support

All Capacity on Demand (CoD) upgrades are performed from the SE “Perform a model conversion” task. Use the task to retrieve and activate a permanent upgrade. You also can use the task 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.

The HMC for IBM zEnterprise System CoD includes the following capabilities:

- 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 Capacity Backup (CBU)
- SE panel features, which are accessed through HMC Single Object Operations:
  - Panel controls for granular activation and deactivation
  - History panel for all Capacity On Demand actions
  - Descriptions editing of the Capacity On Demand records

HMC/SE Version 2.11.1 provides the following CoD information:

- MSU and processor tokens shown on panels
- Last activation time shown on panels
- Pending resources shown by processor type instead of merely 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)

The HMC and SE are integral parts of 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, you must configure and enable SNMP on the HMC.

For additional information about using and setting up CPM, see these publications:

- *z/OS MVS Capacity Provisioning User's Guide*, SC33-8299
- *zEnterprise System Capacity on Demand User's Guide*, SC28-2605

12.6.7 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).

The zEnterprise CPC relies solely on STP for time synchronization, but it continues to provide support for 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 zEnterprise CPC. An ETR ID can be entered on the STP Configuration tab when this system is an zEnterprise CPC to support participation in a mixed CTN.

In a mixed CTN (a network containing both STP and Sysplex Timer), you can use the HMC to perform these tasks:

- Initialize or modify the CTN ID and external time reference (ETR) port states.
- Monitor the status of the CTN.
- Monitor the status of the coupling links that are initialized for STP message exchanges.

In an STP-only CTN, you can use the HMC to perform the following tasks:

- 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 that are initialized for STP message exchanges.

For diagnostic purposes, the Pulse per Second port state on a zEnterprise CPC can be displayed, and fenced ports can be reset individually.

STP recovery enhancement

STP recovery has been enhanced for z114 and z196. See “STP recovery enhancement” on page 145.

For additional planning and setup information, see the following publications:

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

12.6.8 NTP client/server support on HMC

> **Statement of direction for dial-up support:** The IBM zEnterprise 196 and the zEnterprise 114 are the last System z CPCs to support dial-up modems for use with the External Time Source (ETS) option of Server Time Protocol (STP).
>
> The currently available Network Time Protocol (NTP) server option for ETS, as well as Internet time services that are available using broadband connections, can be used to provide the same degree of accuracy as dial-up time services.
>
> Enterprises must begin migrating from dial-up modems to broadband for RSF connections.

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 following requirements:

- Clients who want time accuracy for the STP-only CTN
- Clients using a common time reference across heterogeneous platforms

The NTP client allows the same accurate time across an enterprise that is comprised of heterogeneous platforms. The 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, the zEnterprise CPC can get the 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, use a separate LAN 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, 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 an SE.

Time coordination for zBX components

Network Time Protocol (NTP) clients, which are running on blades in the zEnterprise BladeCenter Extension (zBX), can synchronize their time to the NTP server that is provided by the Support Element (SE). Therefore the SE’s Battery Operated Clock (BOC) is synchronized to the server’s Time-of-Day (TOD) clock every hour and allows the SE’s clock to maintain a time accuracy of 100 milliseconds to an NTP server that is configured as the External Time Source in an STP-only CTN.

For additional planning and setup information for STP and NTP, see the following manuals:

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

12.6.9 Security and user ID management

In this section, we discuss security considerations.

HMC and SE security audit improvements

With the Audit & Log Management task, audit reports can be generated, viewed, saved, and offloaded. The Customize Scheduled Operations task allows the scheduling of audit report generation, saving, and offloading. The Monitor System Events task allows for security logs to result in email notifications using the same type of filters and rules that are used for both hardware and operating system messages.

With IBM zEnterprise System, you have the ability to offload the following HMC and SE log files for customer audit:

- Console event log
- Console service history
- Tasks performed log
- Security logs
- System Log

Full log offload, as well as delta log offload (changes since the last offload request), is provided. Offloading to removable media, as well as to remote locations by FTP, is available. The offloading can be manually initiated by the new Audit & Log Management task or scheduled by 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 or 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 or objects for the user that is mapped to that template. The HMC user is mapped to a specific user ID template by user ID pattern matching and by obtaining the name of the user ID template from content in the LDAP server schema data.

View only user IDs and access for the HMC and 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 these objects and tasks:

- Hardware messages
- Operating system messages
- Customize or delete activation profiles
- Advanced facilities
- Configure on and off

HMC and SE secure FTP support

You can use a secure FTP connection from a HMC/SE FTP client to a client FTP server location. It is implemented using the SSH File Transfer Protocol, which is an extension of the Secure Shell Protocol (SSH). The Manage SSH Keys console action, which is available to both the HMC and SE, allows you to import public keys that are associated with a host address.

Secure FTP infrastructure allows HMC/SE applications to query if a public key is associated with a host address, as well as to use the Secure FTP interface with the appropriate public key for a given host. Tasks using FTP now provide a selection for the secure host connection.

When selected, tasks using FTP verify that a public key is associated with the specified host name. If no public key is provided, they put up a message box to point the user to the Manage SSH Keys task to input a public key. The following tasks provide this support:

- Import/Export IOCDS
- Advanced Facilities FTP ICC Load
- Audit and Log Management (Scheduled Operations Only)

12.6.10 System Input/Output Configuration Analyzer on the HMC and SE

A System Input/Output Configuration Analyzer task is provided that supports the system I/O configuration function.

The necessary information 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 the information from these many sources from one location. 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 channel's node ID information. The Fibre Channel connections (FICONs) support the 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 views, various sort options are available, and data can be exported to a USB flash drive for a later viewing.

The following five views are available:

- PCHID Control Unit View, which shows physical channel IDs (PCHIDs), CSS, channel path identifiers (CHPIDs), and their control units.
- PCHID Partition View, which shows PCHIDS, CSS, CHPIDs, and the partitions in which they exist.
- 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.

12.6.11 Test support element communications

The Support Element Communications test, which is available on the Network Diagnostic Information task, tests to see that communication between the HMC and the SE is available.

The tool performs five tests:

1. The HMC pings the SE.
2. The HMC connects to the SE and also verifies that the SE is at the correct level.
3. The HMC sends a message to the SE and receives a response.
4. The SE connects back to the HMC.
5. The SE sends a message to the HMC and receives a response.

12.6.12 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 that a human being 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, and activation profiles. The IBM zEnterprise System 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.

12.6.13 Cryptographic support

In this section, we describe various cryptographic features.

Cryptographic hardware

The IBM zEnterprise System includes both standard cryptographic hardware and optional cryptographic features for flexibility and growth capability.

The HMC/SE interface provides the following capabilities:

- Define the cryptographic controls.
- Dynamically add a Crypto feature to a partition for the first time.
- Dynamically add a Crypto feature to a partition that already uses Crypto.
- Dynamically remove a 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 setup information, see *IBM zEnterprise 196 Configuration Setup*, SG24-7834.

Digitally signed firmware

Security and data integrity are critical issues with firmware upgrades. Procedures are in place to use a process to digitally sign the firmware update files that are sent to the HMC, SE, and the Trusted Key Entry (TKE) workstation. Using a hash algorithm, a message digest is generated that is then encrypted with a private key to produce a digital signature. This operation ensures that any changes that are made to the data will be detected during the upgrade process by verifying the digital signature. It helps to ensure that no malware can be installed on System z products during firmware updates. It enables, with other existing security functions, zEnterprise CPC CP Assist for Cryptographic Function (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.

12.6.14 z/VM virtual machine management

You can use the HMC for simple 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:

- 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 that are managed by the HMC. When z/VM virtual machines change their status, they generate these events. You can create event monitors to handle the events that are coming from z/VM virtual machines. For example, selected users can be notified by an email message if the virtual machine changes status from Operating to Exception, or any other state.

In addition, in z/VM V5R4 (or later releases), 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.

12.6.15 Installation support for z/VM using the HMC

Starting with z/VM V5R4 and System z10, you can install Linux on System z in a z/VM virtual machine from the HMC workstation 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.

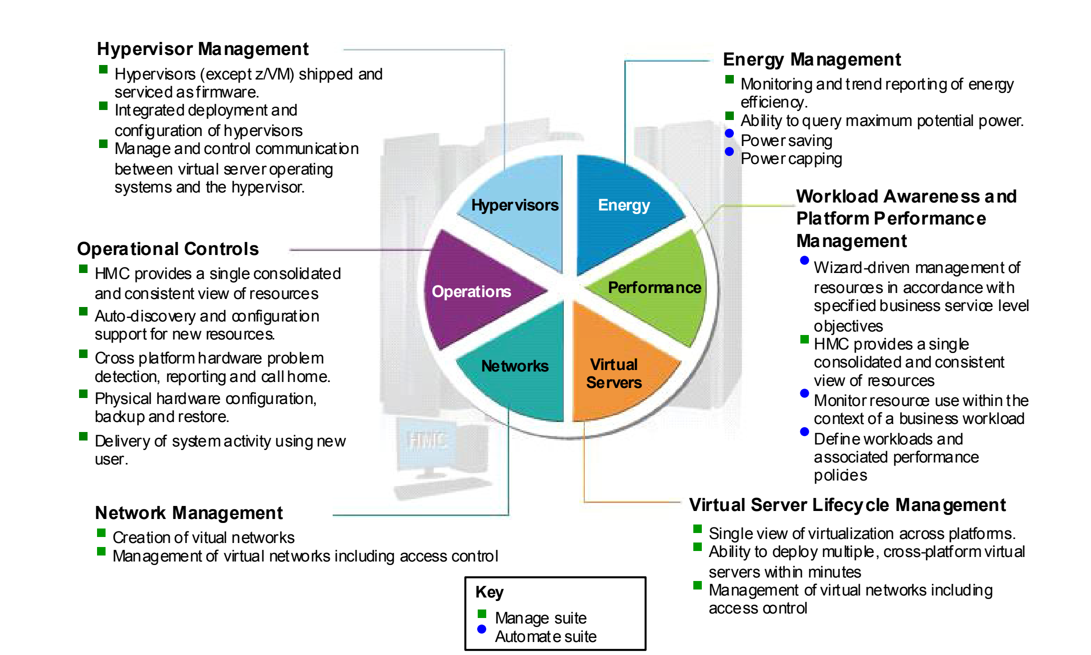
12.7 HMC in an ensemble

An ensemble is a platform systems management domain consisting of one or more zEnterprise *nodes* in which each node comprises a zEnterprise CPC and its optional attached IBM zEnterprise BladeCenter Extension (zBX). The ensemble provides an integrated way to manage virtual server resources and the workloads that can be deployed on those resources.

The IBM zEnterprise System (zEnterprise) is a optimized technology system that delivers a multi-platform, integrated hardware system; spanning System z, System p, and System x server technologies.

12.7.1 Unified Resource Manager

The ensemble is provisioned and managed through the Unified Resource Manager, which resides in the HMC. The Unified Resource Manager is a large set of functions for system management (Figure 12-6).



Hypervisor Management
🟩 Hypervisors (except z/VM) shipped and serviced as firmware.
🟩 Integrated deployment and configuration of hypervisors
🟩 Manage and control communication between virtual server operating systems and the hypervisor.
Operational Controls
🟩 HMC provides a single consolidated and consistent view of resources
🟩 Auto-discovery and configuration support for new resources.
🟩 Cross platform hardware problem detection, reporting and call home.
🟩 Physical hardware configuration, backup and restore.
🟩 Delivery of system activity using new user.
Network Management
🟩 Creation of vitual networks
🟩 Management of virtual networks including access control
Energy Management
🟩 Monitoring and trend reporting of energy efficiency.
🟩 Ability to query maximum potential power.
● Power saving
● Power capping
Workload Awareness and Platform Performance Management
● Wizard-driven management of resources in accordance with specified business service level objectives
🟩 HMC provides a single consolidated and consistent view of resources
● Monitor resource use within the context of a business workload
● Define workloads and associated performance policies
Virtual Server Lifecycle Management
🟩 Single view of virtualization across platforms.
🟩 Ability to deploy multiple, cross-platform virtual servers within minutes
🟩 Management of virtual networks including access control
Hypervisors
Energy
Performance
Virtual Servers
Networks
Operations
HMC
Key
🟩 Manage suite
● Automate suite

Figure 12-6 Unified Resource Manger functions and suites

Unified Resource Manager provides the following functions:

- Hypervisor management

  Provides tasks for managing the hypervisor life cycle, managing storage resources, performing reliability, availability, and serviceability (RAS) and First Failure Data Capture (FFDC) features, and monitoring the supported hypervisors.

- Ensemble membership management

  Provides tasks for creating an ensemble and controlling the membership of the ensemble.

- Storage management

  Provides a common user interface for the allocation and deallocation of physical and virtual storage resources for an ensemble.

- Virtual server management

  Provides life-cycle management to create, delete, activate, deactivate, and modify the definitions of virtual servers.

- Virtual network management

  Provides for the management of networking resources for an ensemble.

- Performance management

  Provides a global performance view of all virtual servers supporting workloads that are deployed in an ensemble. The virtual server workload performance goal is like a simplified z/OS Workload Manager policy:

  – You can define, monitor, report, and manage the performance of virtual servers based on performance policies.

  – Policies are associated to the workload:

    • From the overall Workload performance health report, you can review the contributions of individual virtual servers.

    • You can manage resources across virtual servers within a hypervisor instance.

- Energy management

  – You can monitor energy usage and control power-saving settings, which are accessed through the new Monitors Dashboard.

  – You can monitor virtual server resources for CPU use and delays, with the capability of creating a graphical trend report.

Unified Resource Manager supports various levels of system management. A feature determines the management functions and operational controls that are available for a zEnterprise server and any attached zBX. These named features are Manage and Automate/Advanced Management:

Manage suite Provides Unified Resource Manager's function for core operational controls, installation, and energy monitoring. It is configured by default and activated when an ensemble is created.

Automate/Advanced Management suite
Advanced Management functionality for IBM System x blades delivers workload definition and performance policy monitoring and reporting. The Automate functionality adds, on top of the Advanced Management functionality, goal-oriented resource monitoring management and energy management for the zCPC components, IBM Smart Analytic Optimizer, POWER7 Blade, and DataPower XI50z.

Table 12-2 lists the feature codes that must be ordered for enabling Unified Resource Manager. *To get ensemble membership, make sure that you also order FC 0025 for z196 or z114.*

Table 12-2 Unified Resource Manager feature codes and charge indicators

Unified Resource Manager managed component	**Manage[a] per connection**	**Advanced Management[a] per connection**	**Automate[a] per connection**
Base features	0019[b] - N/C	N/A	0020[c] - N/C
Integrated Facility for Linux (IFL)	N/C	N/A	0052 - Yes
IBM Smart Analytics Optimizer	0039 - Yes	N/A	0043 - N/C
POWER7 Blade	0041 - Yes	N/A	0045 - Yes
DataPower Blade	0040 - Yes	N/A	0044 - N/C
IBM System x Blades	0042 - Yes	0046 - Yes	N/A

[a] Yes = charged feature, N/C = no charge, N/A = not applicable. *All* components are either managed through the Manage suite or the Automate/Advanced Management suite. The Automate/Advanced Management suite contains the functionality of the Managed suite.

[b] Feature code 0019 is a prerequisite for feature codes 0020, 0039, 0040, 0041, and 0042.

[c] Feature code 0020 is a prerequisite for feature codes 0043, 0044, 0045, 0046, and 0052.

APIs for IBM zEnterprise Unified Resource Manager

Statement of direction: IBM intends to offer APIs for IBM zEnterprise Unified Resource Manager. These APIs are designed to provide access to the same underlying functions that support the Unified Resource Manager user interface and can be exploited to enable discovery, monitoring, and provisioning use cases.

Note: These APIs enable management of the Unified Resource Manager from external tools, such IBM Systems Director, IBM Tivoli or independent software vendor (ISV) systems management products. IBM intends to provide, as a priority in the following sequence: discovery, monitoring, and then access to the provisioning functions of Unified Resource Manager.

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

For more information regarding the Unified Resource Manager, see *IBM zEnterprise Unified Resource Manager,* SG24-7921, and *IBM zEnterprise System Introduction to Ensembles,* SC27-2609.

12.7.2 Ensemble definition and management

The ensemble starts with a pair of HMCs. These HMCs are designated as the primary and alternate HMCs. This pair is assigned an ensemble identity. The zEnterprise CPCs 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 zEnterprise CPC is ordered. This feature code is required on the *controlling* zEnterprise CPC to be able to attach a zBX.

A new task called *Create Ensemble* allows the Access Administrator to create an ensemble that contains CPCs, Images, workloads, virtual networks and storage pools, either with or without an optional zBX.

If a zEnterprise CPC has been entered into an ensemble, the CPC Details task on the SE and HMC will reflect the ensemble name.

The Unified Resource Manager actions for the ensemble are conducted from a single primary HMC. All other HMCs that are 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.

The following objects are the ensemble-specific managed objects:

- Ensemble
- Members
- Blades
- BladeCenters
- Hypervisors
- Storage Resources
- Virtual Servers
- Workloads

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. A few of those tasks have been extended to allow you to configure certain ensemble-specific properties, such as setting the virtual network that is associated with Open Systems Adapters (OSAs) for an LPAR. Showing ensemble-related data in certain tasks is allowed. Generally, if the data affects the operation of the ensemble, the data is read-only on another HMC. The following tasks show ensemble-related data on another HMC:

- 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 events.

HMC considerations when used to manage an ensemble

Here, we list considerations for using Unified Resource Manager to manage an ensemble:

- All HMCs at the supported code level are eligible to create an ensemble. Only HMC FC 0090 and FC 0091 are capable of being primary or alternate HMCs.
- The primary and the alternate HMC must be the same machine type/feature code.
- There is a single HMC pair managing the ensemble: primary HMC and alternate HMC.
- Only one primary HMC manages an ensemble, which can consist 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 or later)* and *Alternate Hardware Management Console (Version 2.11.0 or later)* will appear on the HMC banner. When the ensemble is deleted, the titles resort to the defaults.
- 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).

- A zEnterprise ensemble can have a maximum of eight nodes and is managed by one primary HMC and its alternate. Each node comprises a zEnterprise CPC and its optional attached IBM zEnterprise BladeCenter Extension (zBX).
- 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.*
- 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.

12.7.3 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 CPC, and optionally to one or more 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. Although not desirable, the system can 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 controlling (stateful) component for the 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 the 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 HMC is required to be configured in an primary/alternate configuration and cannot be disconnected from the managed ensemble members.

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

12.7.4 Considerations for multiple HMCs

Clients 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 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.

12.7.5 HMC browser session to a primary HMC

A remote HMC browser session to the primary HMC that is managing an ensemble allows a user that is currently logged on to another HMC or a workstation to perform ensemble-related actions.

12.7.6 HMC ensemble topology

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 by the internode management network (INMN).

Figure 12-7 on page 369 depicts an ensemble with two zEnterprise CPCs and a zBX that are managed by the Unified Resource Manager residing in the primary and alternate HMCs. CPC1 controls the zBX, and CPC2 is a stand-alone CPC.

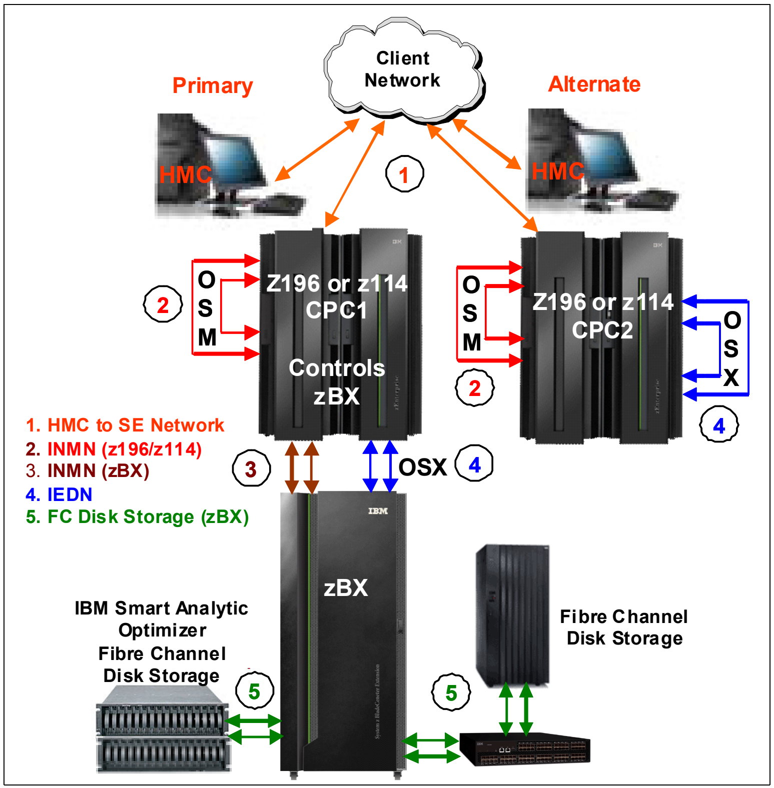


Figure 12-7 Ensemble example with primary and alternate HMCs

For the stand-alone CPC ensemble node (CPC2), two OSA Express-3 1000BASE-T ports (CHPID type OSM) connect to the Bulk Power Hubs (port J07) with 3.2 m (10.5 ft.) Category 6 Ethernet cables. The HMCs also communicate with all the components of the ensemble by the BPHs in the CPC.

The OSA Express-3 10 GbE ports (CHPID type OSX) are plugged in with client-provided 10 GbE cables (either SR or LR, depending on the OSA feature).

You can obtain details for ensemble connectivity for a zEnterprise CPC with a zBX in 7.4, “zBX connectivity” on page 194.

A

Channel options

The following two tables describe all channel attributes, the required cable types, the maximum unrepeated distance, and the bit rate of the zEnterprise CPCs.

For all optical links, the connector type is LC Duplex, except the Enterprise System Connection (ESCON) connector is of an MTRJ type, and the 12xIFB connection is established with a Multi-Fiber Push-On (MPO) connector. The electrical Ethernet cable for the Open Systems Adapter (OSA) connectivity is connected via an RJ45 jack.

Table A-1 lists the attributes of the various channel options that are supported on the IBM zEnterprise 114.

At least one ESCON, Fiber Connection (FICON), InterSystem Channel (ISC), or InfiniBand (IFB) feature is required.

Statement of direction:
ESCON: The z196 and the z114 are the last IBM System z® servers for which you can order ESCON channels. IBM intends not to support ESCON channels on future servers.

ISC-3: The z196 and the z114 are the last IBM System z® servers for which you can order ISC-3 Links.

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

Table A-1 z196 and z114 channel feature support

Channel feature	Feature codes	Bit rate	Cable type	Maximum unrepeated distance[a]	Ordering information	Remark
Enterprise Systems Connection (ESCON)						
16-port ESCON	2323	200 Mbps	MM 62.5 µm	3 km (800)	New build	

Channel feature	Feature codes	Bit rate	Cable type	Maximum unrepeated distance[a]	Ordering information	Remark
Fiber Connection (FICON)						
FICON Express4-2C 4KM LX	3323	1, 2, or 4 Gbps	SM 9 µm	4 km	Carry forward[b]	Only on z114
FICON Express4 4KM LX	3324				Carry forward	
FICON Express4 10KM LX	3321			10 km	Carry forward	
FICON Express8 10KM LX	3325	2, 4, or 8 Gbps	SM 9 µm	10 km	Carry forward[b]	
FICON Express8S 10KM LX	0409				New build	
FICON Express4 SX	3322	1, 2, or 4 Gbps	OM1, OM2, and OM3	See Table A-2 on page 374 because the distance depends on the cable and bit rate.	Carry forward	
FICON Express4-2C SX	3318				Carry forward[b]	Only on z114
FICON Express8 SX	3326	2, 4, or 8 Gbps			Carry forward only	
FICON Express8S SX	0410				New build	
Open Systems Adapter (OSA)						
OSA-Express2 GbE LX	3364	1 Gbps	SM 9 µm MCP 50 µm	5 km 550 m (500)	Carry forward	
OSA-Express3 GbE LX	3362				Carry forward[b]	
OSA-Express4S GbE LX	0404				New build	
OSA-Express2 GbE SX	3365	1 Gbps	MM 62.5 µm MM 50 µm	220 m (166) 275 m (200) 550 m (500)	Carry forward	
OSA-Express3 GbE SX	3363				Carry forward[b]	
OSA-Express3-2P GbE SX	3373				Carry forward[b]	Only on z114
OSA-Express4S GbE SX	0405				New build	

Channel feature	Feature codes	Bit rate	Cable type	Maximum unrepeated distance[a]	Ordering information	Remark
OSA-Express2 1000BASE-T Ethernet	3366	10, 100, or 1000 Mbps	UTP Cat5	100 m	Carry forward	
OSA-Express3 1000BASE-T Ethernet	3367				New build	
OSA-Express3-2P 1000BASE-T Ethernet	3369				New build	Only on z114
OSA-Express3 10 GbE LR	3370	10 Gbps	SM 9 µm	10 km	Carry forward[b]	
OSA-Express4S 10 GbE LR	0406				New build	
OSA-Express3 10 GbE SR	3371	10 Gbps	MM 62.5 µm	33 m (200)	Carry forward[b]	
OSA-Express4S 10 GbE SR	0407		MM 50 µm	300 m (2000) 82 m (500)	New build	
Parallel Sysplex						
IC	N/A		N/A	N/A	N/A	
ISC-3 (peer mode)	0217 0218 0219	2 Gbps	SM 9 µm MCP 50 µm	10 km 550 m (400)	New build	
ISC-3 (RPQ 8P2197 Peer mode at 1 Gbps)[c]		1 Gbps	SM 9 µm	20 km	New build	
HCA2-O 12xIFB	0163	6 GBps	OM3	150 m	New build	
HCA2-O LR 1xIFB	0168	2.5 or 5 Gbps	SM 9 µm	10 km	Carry forward	
HCA3-O 12xIFB	0171	6 GBps	OM3	150 m	New build	
HCA3-O LR 1xIFB	0170	2.5 or 5 Gbps	SM 9 µm	10 km	New build	
Cryptography						
Crypto Express3	0864	N/A	N/A	N/A	New build	
Crypto Express3-1P	0871	N/A	N/A	N/A	New build	Only on z114

a. Minimum fiber bandwidth distance product in MHz·km for multi-mode fiber optic links are included in parentheses where applicable.

b. Ordering of this feature is determined by the fulfillment process.

c. RPQ 8P2197 enables the ordering of a daughter card supporting 20 km (12.4 miles) 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.

Table A-2 shows the maximum unrepeated distance for FICON SX.

Table A-2 Maximum unrepeated distance for FICON SX

Cable type and bit rate	Unit	1 Gbps	2 Gbps	4 Gbps	8 Gbps
OM1 (62,5 µm at 200MHz·km)	meter	300	150	70	21
	foot	984	492	230	69
OM2 (50 µm at 500MHz·km)	meter	500	300	150	50
	foot	1640	984	492	164
OM3 (50 µm at 2000MHz·km)	meter	860	500	380	150
	foot	2822	1640	1247	492

B

Valid z114 On/Off Capacity on Demand upgrades

The tables in this appendix show all valid On/Off Capacity on Demand (OOCoD) upgrade options for the z114. For more information, visit the Resource Link web page for the client-initiated upgrades (CIU) matrix showing the upgrades that are available for a given machine type and model:

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

Table B-1 Valid On/Off Capacity on Demand upgrades for the 1-way z114 capacity identifiers (CIs)

CI	Valid OOCoD upgrade	CI	Valid OOCoD upgrade	CI	Valid OOCoD upgrade
A01	B01, C01, A02, D01	B01	C01, D01, B02, E01, F01	C01	D01, E01, F01, C02, G01, H01
D01	E01, F01, G01, H01, D02, I01, E02, J01	E01	F01, G01, H01, I01, E02, J01, F02, K01	F01	G01, H01, I01, J01, F02, K01
G01	H01, I01, J01, K01, G02, L01, H02	H01	I01, J01, K01, L01, H02, M01, I02	I01	J01, K01, L01, M01, I02, N01, J02
J01	K01, L01, M01, N01, J02, O01	K01	L01, M01, N01, O01, K02, P01	L01	M01, N01, O01, P01, L02, Q01
M01	N01, O01, P01, Q01, M02, R01, S01	N01	O01, P01, Q01, R01, S01, N02, T01	O01	P01, Q01, R01, S01, T01, O02
P01	Q01, R01, S01, T01, P02, U01, V01	Q01	R01, S01, T01, U01, V01, Q02, W01, R02	R01	S01, T01, U01, V01, W01, R02
S01	T01, U01, V01, W01, S02, X01	T01	U01, V01, W01, X01, T02	U01	V01, W01, X01, Y01, U02, V02, Z01
V01	W01, X01, Y01, V02, Z01	W01	X01, Y01, Z01, W02	X01	Y01, Z01, X02
Y01	Z01, Y02	Z01	Z02		

Table B-2 Valid OOCoD upgrades for the 2-way z114 CIs (Capacity Identifiers)

CI	Valid OOCoD upgrade	CI	Valid OOCoD upgrade	CI	Valid OOCoD upgrade
A02	B02, A03, C02, B03, A04, D02, B04, C03, E02	B02	C02, B03, D02, B04, C03, E02, F02	C02	D02, C03, E02, F02, D03, C04, E03, G02, H02
D02	E02, F02, D03, E03, G02, H02, D04, F03, I02, E04, J02	E02	F02, E03, G02, H02, F03, I02, E04, J02, G03	F02	G02, H02, F03, I02, J02, G03, F04, K02, H03
G02	H02, I02, J02, G03, K02, H03, G04, I03, L02, J03	H02	I02, J02, K02, H03, I03, L02, J03, H04, M02, K03	I02	J02, K02, I03, L02, J03, M02, K03, I04, N02
J02	K02, L02, J03, M02, K03, N02, J04, L03, O02	K02	L02, M02, K03, N02, L03, O02, K04, M03, P02	L02	M02, N02, L03, O02, M03, P02, L04, N03, Q02
M02	N02, O02, M03, P02, N03, Q02, M04, O03, R02	N02	O02, P02, N03, Q02, O03, R02, S02, P03, N04, T02, O04, Q03	O02	P02, Q02, O03, R02, S02, P03, T02, O04, Q03, R03, U02
P02	Q02, R02, S02, P03, T02, Q03, R03, U02, P04, S03, V02	Q02	R02, S02, T02, Q03, R03, U02, S03, V02, Q04, W02, T03	R02	S02, T02, R03, U02, S03, V02, W02, T03, R04
S02	T02, U02, S03, V02, W02, T03, S04, U03, X02, V03	T02	U02, V02, W02, T03, U03, X02, V03, T04	U02	V02, W02, U03, X02, V03, W03, Y02, U04
V02	W02, X02, V03, W03, Y02, V04, Z02, X03	W02	X02, W03, Y02, Z02, X03, W04	X02	Y02, Z02, X03, Y03, X04, Z03
Y02	Z02, Y03, Z03, Y04	Z02	Z03, Z04		

Table B-3 Valid OOCoD upgrades for the 3-way z114 CIs (Capacity Identifiers)

CI	Valid OOCoD upgrade	CI	Valid OOCoD upgrade	CI	Valid OOCoD upgrade
A03	B03, A04, B04, C03, A05, B05, D03, C04, E03	B03	B04, C03, B05, D03, C04, E03, C05, D04, F03	C03	D03, C04, E03, C05, D04, F03, E04, D05, G03
D03	E03, D04, F03, E04, D05, G03, F04, E05, H03, G04, I03, F05	E03	F03, E04, G03, F04, E05, H03, G04, I03, F05, J03	F03	G03, F04, H03, G04, I03, F05, J03, H04, G05, K03, I04
G03	H03, G04, I03, J03, H04, G05, K03, I04, H05, J04, L03, I05, K04	H03	I03, J03, H04, K03, I04, H05, J04, L03, I05, K04, J05, M03	I03	J03, K03, I04, J04, L03, I05, K04, J05, M03, L04, K05, N03
J03	K03, J04, L03, K04, J05, M03, L04, K05, N03, M04, O03	K03	L03, K04, M03, L04, K05, N03, M04, O03, L05, P03, N04	L03	M03, L04, N03, M04, O03, L05, P03, N04, M05, O04, Q03
M03	N03, M04, O03, P03, N04, M05, O04, Q03, N05, R03, P04, O05, S03	N03	O03, P03, N04, O04, Q03, N05, R03, P04, O05, S03, Q04, P05, T03	O03	P03, O04, Q03, R03, P04, O05, S03, Q04, P05, T03, R04, Q05
P03	Q03, R03, P04, S03, Q04, P05, T03, R04, Q05, S04, U03, R05, V03, T04	Q03	R03, S03, Q04, T03, R04, Q05, S04, U03, R05, V03, T04, S05, W03	R03	S03, T03, R04, S04, U03, R05, V03, T04, S05, W03, U04, T05
S03	T03, S04, U03, V03, T04, S05, W03, U04, T05, V04, X03, U05	T03	U03, V03, T04, W03, U04, T05, V04, X03, U05, W04, V05	U03	V03, W03, U04, V04, X03, U05, W04, V05, Y03, W05, X04
V03	W03, V04, X03, W04, V05, Y03, W05, X04, Z03	W03	X03, W04, Y03, W05, X04, Z03, Y04, X05	X03	Y03, X04, Z03, Y04, X05, Z04, Y05
Y03	Z03, Y04, Z04, Y05, Z05	Z03	Z04, Z05		

Table B-4 Valid OOCoD upgrades for the 4-way z114 CIs (Capacity Identifiers)

CI	Valid OOCoD upgrade	CI	Valid OOCoD upgrade	CI	Valid OOCoD upgrade
A04	B04, A05, B05, C04, C05, D04, E04	B04	B05, C04, C05, D04, E04, D05	C04	C05, D04, E04, D05, F04, E05, G04, F05
D04	E04, D05, F04, E05, G04, F05, H04, G05, I04	E04	F04, E05, G04, F05, H04, G05, I04, H05, J04	F04	G04, F05, H04, G05, I04, H05, J04, I05, K04, J05
G04	H04, G05, I04, H05, J04, I05, K04, J05, L04, K05	H04	I04, H05, J04, I05, K04, J05, L04, K05, M04, L05	I04	J04, I05, K04, J05, L04, K05, M04, L05, N04
J04	K04, J05, L04, K05, M04, L05, N04, M05, O04	K04	L04, K05, M04, L05, N04, M05, O04, N05, P04	L04	M04, L05, N04, M05, O04, N05, P04, O05, Q04, P05
M04	N04, M05, O04, N05, P04, O05, Q04, P05, R04, Q05	N04	O04, N05, P04, O05, Q04, P05, R04, Q05, S04, R05, T04	O04	P04, O05, Q04, P05, R04, Q05, S04, R05, T04, S05
P04	Q04, P05, R04, Q05, S04, R05, T04, S05, U04, T05, V04	Q04	R04, Q05, S04, R05, T04, S05, U04, T05, V04, U05, W04	R04	S04, R05, T04, S05, U04, T05, V04, U05, W04, V05
S04	T04, S05, U04, T05, V04, U05, W04, V05, W05, X04	T04	U04, T05, V04, U05, W04, V05, W05, X04	U04	V04, U05, W04, V05, W05, X04, Y04, X05, Z04
V04	W04, V05, W05, X04, Y04, X05, Z04, Y05	W04	W05, X04, Y04, X05, Z04, Y05, Z05	X04	Y04, X05, Z04, Y05, Z05
Y04	Z04, Y05, Z05	Z04	Z05		

Table B-5 Valid OOCoD upgrades for the 5-way z114 CIs (Capacity Identifiers)

CI	Valid OOCoD upgrade	CI	Valid OOCoD upgrade	CI	Valid OOCoD upgrade
A05	B05, C05, D05, E05	B05	C05, D05, E05, F05	C05	D05, E05, F05, G05
D05	E05, F05, G05, H05, I05	E05	F05, G05, H05, I05, J05	F05	G05, H05, I05, J05, K05
G05	H05, I05, J05, K05, L05	H05	I05, J05, K05, L05, M05	I05	J05, K05, L05, M05, N05
J05	K05, L05, M05, N05, O05	K05	L05, M05, N05, O05, P05	L05	M05, N05, O05, P05, Q05
M05	N05, O05, P05, Q05, R05	N05	O05, P05, Q05, R05, S05, T05	O05	P05, Q05, R05, S05, T05, U05
P05	Q05, R05, S05, T05, U05, V05	Q05	R05, S05, T05, U05, V05, W05	R05	S05, T05, U05, V05, W05
S05	T05, U05, V05, W05, X05	T05	U05, V05, W05, X05	U05	V05, W05, X05, Y05, Z05
V05	W05, X05, Y05, Z05	W05	X05, Y05, Z05	X05	Y05, Z05
Y05	Z05	Z05	N/A		

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this book.

IBM Redbooks publications

For information about ordering these publications, see “How to get IBM Redbooks publications” on page 381. Note that some of the documents referenced here may be available in softcopy only.

- *IBM zEnterprise 196 Technical Guide*, SG24-7833
- *IBM zEnterprise Unified Resource Manager*, SG24-7921
- *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
- *Using IBM System z As the Foundation for Your Information Management Architecture*, REDP-4606

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 websites 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/
- 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 IBM Redbooks publications

You can search for, view, or download IBM Redbooks publications, Redpaper publications, Web Docs, draft publications, and additional materials, as well as order hardcopy IBM Redbooks publications, at this website:

ibm.com/redbooks

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

Index

IBM zEnterprise 114 Technical Guide

(0.5” spine)
0.475”<->0.873”
250 <-> 459 pages

IBM®

IBM zEnterprise 114 Technical Guide

Redbooks®

Explains virtualizing and managing the heterogenous infrastructure

Describes the zEnterprise System and related features and functions

Discusses zEnterprise hardware and software capabilities

This IBM Redbooks publication discusses the IBM zEnterprise System, an IBM scalable mainframe server. IBM is taking a revolutionary approach by integrating separate 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 114 central processor complex, the IBM zEnterprise Unified Resource Manager, and the IBM zEnterprise BladeCenter Extension. The z114 is designed with improved scalability, performance, security, resiliency, availability, and virtualization. The z114 provides up to 18% improvement in uniprocessor speed and up to a 12% increase in total system capacity for z/OS, z/VM, and Linux on System z over the z10 Business Class (BC).

The zBX infrastructure works with the z114 to enhance System z virtualization and management through an integrated hardware platform that spans mainframe, POWER7, and System x technologies. The federated capacity from multiple architectures of the zEnterprise System is managed as a single pool of resources, integrating system and workload management across the environment through the Unified Resource Manager.

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

SG24-7954-00 ISBN 0738436119