IBM eserver

IBM

IBM eserver pSeries 670 and pSeries 690 System Handbook

- Component-based description of the hardware architecture
- How to use e-config to configure pSeries 670 and pSeries 690
- Capacity Upgrade on Demand explained

Keigo Matsubara
Alvin Hua Juay Teng
Bruno Blanchard

Redbooks

ibm.com/redbooks

International Technical Support Organization

IBM @server pSeries 670 and pSeries 690 System Handbook

May 2003

SG24-7040-02

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

Third Edition (May 2003)

This edition applies to IBM @server pSeries 670 Model 671 and IBM @server pSeries 690 for use with the AIX 5L Version 5.1 (program number 5765-E61) and AIX 5L Version 5.2 (program number 5765-E62).

Note: This book is based on a pre-GA version of product IBM Hardware Management Console for pSeries software Release 3.2 and may not apply when the product becomes generally available. We recommend that you consult the product documentation or follow-on versions of this redbook for more current information.

Contents

Figures

Tables

Notices

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

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

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive Armonk, NY 10504-1785 U.S.A.

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

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

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

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

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

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

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

Trademarks

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

AIX®
AIX 5L™
AS/400®
Chipkill™
ESCON®
eServer™
@server™
@server ™
IBM®
ibm.com®
IBMLink™
iSeries™
POWER2™
POWER4™
POWER4+™
POWERparallel®
PowerPC®
pSeries™
Redbooks(logo)™ 
Redbooks™
RS/6000®
S/370™
S/390®
SP™
zSeries®

The following terms are trademarks of other companies:

ActionMedia, LANDesk, MMX, Pentium and ProShare are trademarks of Intel Corporation in the United States, other countries, or both.

Microsoft, Windows, Windows NT, 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 Sun Microsystems, Inc. in the United States, other countries, or both.

C-bus is a trademark of Corollary, Inc. in the United States, other countries, or both.

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

SET, SET Secure Electronic Transaction, and the SET Logo are trademarks owned by SET Secure Electronic Transaction LLC.

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

Preface

This IBM® Redbook explains the IBM @server™ pSeries™ 670 Model 671 and the IBM @server pSeries 690, a new level of UNIX servers providing world-class performance, availability, and flexibility. Capable of data center, application, and high performance computing, these servers include mainframe-inspired self-management and security to meet your most demanding needs.

This redbook includes the following topics:

- Overview of the pSeries 670 and pSeries 690
- Hardware architecture of the pSeries 670 and pSeries 690
- Using the IBM Configurator for e-business
- Capacity Upgrade on Demand
- Reliability, availability, and serviceability

This book is an ideal desk-side reference for IBM professionals, Business Partners, and technical specialists who support the pSeries 670 and pSeries 690, and for those who want to learn more about this radically new server in a clear, single-source handbook. It provides the necessary information to successfully order a pSeries 670 or pSeries 690 for a production environment and then, upon successful installation, configure service support functions, such as Service Focal Point and Inventory Scout.

The team that wrote this redbook

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

Keigo Matsubara is an Advisory IT Specialist at the International Technical Support Organization (ITSO), Austin Center. Before joining the ITSO, he worked in the System and Web Solution Center in Japan as a Field Technical Support Specialist (FTSS) for pSeries. He has been working for IBM for ten years.

Alvin Hua Juay Teng is a Technical Sales Specialist at IBM Singapore. He joined IBM in 2000 and has been working with pSeries ever since. He has experience in providing solutions for the banking industries and government organizations. He holds a master's degree in High Performance Computing for Engineered Systems from Singapore-MIT Alliance (A jointed master's degree from Singapore University of Singapore, Singapore, Nanyang Technological

University, Singapore, and Massachusetts Institute of Technology, US) and a B.Eng in Mechanical Engineering from University of Sheffield, UK.

Bruno Blanchard is a Certified IT Specialist working for IBM France at the IGS Pan-EMEA Infrastructure and Technology Group, in La Gaude. He holds an Engineering degree from Ecole Centrale de Paris and a master's degree from Oregon State University. He has been with IBM since 1983, as a System Engineer for VM and AIX®. He is a certified AIX and SP™ Specialist, and his areas of expertise also include pSeries servers and clusters, as well as network management. He is currently working as an Architect on projects deployed in Europe for the IBM @server cluster 1600 and pSeries servers required for the IBM IT infrastructure.

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

International Technical Support Organization, Austin Center
Scott Veter and Wade Wallace

International Technical Support Organization, Rochester Center
Gabrielle Velez

IBM Austin
Alan Standridge, Andy McLaughlin, Bill Casey, Bob Foster, Bruce Wood, Dave Willoughby, Doug Bossen, Eric Marshall, Iggy Haider, James Donnelly, Joel M. Tendler, John Bissell, John Purcell, Kaena Freitas, Mark Dewalt, Mike Stys, Minh Nguyen, Richard Bolton, Truc Nguyen, Walter Lipp

IBM Endicott
Scott Nettleship and Lenny Nichols

IBM Poughkeepsie
James T Mitchell

IBM Rochester
David L Shaw, Lawny Miller, Matt Spinler, Sandy Shirk-Heath

Become a published author

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

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

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

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

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

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

  ibm.com/redbooks

- Send your comments in an Internet note to:

  redbook@us.ibm.com

- Mail your comments to:

  IBM Corporation, International Technical Support Organization
  Dept. JN9B Building 003 Internal Zip 2834
  11400 Burnet Road
  Austin, Texas 78758-3493

Summary of changes

This section describes the technical changes made in this edition of the book and in previous editions. This edition may also include minor corrections and editorial changes that are not identified.

Summary of Changes
for SG24-7040-02
for IBM @server pSeries 670 and pSeries 690 System Handbook
as created or updated on May 30, 2003.

April 2003, Third Edition

This revision reflects the addition, deletion, or modification of new and changed information described below.

New information

- AIX 5L Version 5.2
- Linux information on the pSeries 670 and pSeries 690
- POWER4™+ processors
- New I/O subsystem
- Capacity Upgrade on Demand for memory
- How to update HMC
- Microcode update service

Changed information

- Using the IBM Configurator for e-business
- Capacity Upgrade on Demand for processors
- Description of On/Off Capacity on Demand

October 2002, Second Edition

The second version of this book, SG24-7040-01 IBM @server pSeries 670 and pSeries 690System Handbook, was written by the following authors:

Keigo Matsubara, Bas Hazelzet, Marc-Eric Kahle

This revision reflects the addition, deletion, or modification of new and changed information described below.

New information

- The IBM @server pSeries 670 Model 671
- Capacity Upgrade on Demand
- Overview of the feature codes

Changed information

- Detailed logical partitioning explanation was moved to *The Complete Partitioning Guide for IBM @server* pSeries Servers, SG24-7039.
- Usage examples about how to use IBM Configurator for e-business were updated to include the IBM @server pSeries 670 Model 671.
- A new explanation was added about service functions, such as Service Focal Point and Inventory Scout.
- The CUoD support for AIX 5L Version 5.1 was added.

The following list shows contributors for the first version of this book, SG24-7040-00 IBM @server pSeries 690 System Handbook:

IBM Austin
Alan Standridge, Andy McLaughlin, Bob Foster, Bruce Wood, Dave Lewis, Dave Willoughby, Doug Bossen, George W Clark, James Donnelly, Jane Chilton, John O'Quin, John Purcell, Larry Amy, Luke Browning, Mark Dewalt, Mike Stys, Minh Nguyen, Sajan Lukose, Siraj Ismail, Susan Caunt, Truc Nguyen, Walter Lipp

IBM Endicott
Scott Nettleship and Lenny Nichols

IBM Germany
Knut Muennich

IBM Hamden
Jack Dobkowski

IBM Poughkeepsie
Anthony Pioli, Juan Parrilla II, Michael Schmidt, Rob Overton

IBM Rochester
Matt Spinler

IBM UK
Dave Williams and Clive Benjamin

March 2002, First Edition

The first version of this book, SG24-7040-00 IBM @server pSeries 690 System Handbook, was written by the following authors:

Keigo Matsubara, Cesar Diniz Maciel, KangJin Lee, Martin Springer

The following list shows contributors for the first version of this book, SG24-7040-00 IBM @server pSeries 690 System Handbook:

IBM Austin
Alan Standridge, Andy McLaughlin, Arthur Ban, Balaram Sinharoy, Bob Foster, Bradley McCredie, Dave Willoughby, David Ruth, Doug Bossen, Duke Paulsen, Eric Marshall, Gerald McBrearty, Hung Le, Hye-Young McCreary, Jan Klockow, Joel M. Tendler, Julie Craft, Kurt Szabo, Luc Smolders, Mike Stys, Pat Buckland, Randy Itskin, Richard Cutler, Robert W West, Steve Dodson, Steve Fields, Steven Hartman, Susan Caunt, Walter Lipp, William Hodges

IBM Japan
Shigeru Kishikawa and Shingo Matsuda

IBM Poughkeepsie
Ron Goering

1

Overview of the pSeries 670 and pSeries 690

The IBM @server pSeries 670 Model 671 (hereafter referred to as pSeries 670) and the IBM @server pSeries 690 (hereafter referred to as pSeries 690) are integral parts of the IBM @server product line, featuring servers that can help lower total cost of ownership, improve efficiency, and speed up e-business transformation.

The pSeries 670 and pSeries 690 are IBM's flagship products in the pSeries line of UNIX servers. The pSeries 670, a 4-way to 16-way server, and the pSeries 690, an 8-way to 32-way server, share the same basic technology, and represent the latest generation of performance leadership and reliability from IBM. They incorporate the latest advances in chip technology from IBM, as well as many leading self-managing system capabilities to enable mission-critical operation.

In addition to unparalleled speed, the pSeries 670 and pSeries 690 systems have the ability to consolidate critical applications on a single, data center class server. As a result, there are fewer servers to manage and maintain. Available capacity can be used more effectively and with greater flexibility by the dynamic logical partitioning function to meet changing business demands.

In this chapter we introduce the pSeries 670 and pSeries 690 servers, highlighting their advantages, main characteristics and components, as well as how they fit into the data center.

1.1 What's new in the pSeries 670 and pSeries 690

This third edition of the *IBM @server pSeries 670 and pSeries 690 System Handbook*, SG24-7040 describes these two models of servers, as available after the products' announcements in May 2003, that brings the following new features[1]:

- Processors using the POWER4+ chip, at 1.5 and 1.7 GHz
- L3 Cache and memory modules running at 500 and 575 MHz
- 64 GBs memory modules
- A faster I/O subsystem, using new I/O books, new RIO cables, new I/O drawer planar
- Support for 133 MHz PCI-X
- Improved partitioning facility, with support up to 32 partitions
- Improved Capacity Upgrade on Demand addressing both the processor and the memory resources
- new upgrades possibilities, including model conversion from pSeries 670 to pSeries 690
- Firmware updates from the HMC

The support of these new features leaves the general architecture of the pSeries 670 and pSeries 690 servers unchanged. However, it results into many important implementation changes, such as the RIO drawer cabling, the PCI adapters supported combinations, or the manner to use the IBM Configurator for e-business to order a new server or an MES (Miscellaneous Equipment Specification).

If you are already familiar with the pSeries 670 and pSeries 690 servers, you can avoid reading their general presentation in the following sections, and directly skip to the following chapters.

If however you are new to the pSeries 670 and pSeries 690 servers, we strongly encourage you to read this introductory chapter in its entirety.

Note: Throughout this book, we use the word "POWER4" to refer to the 1.1, 1.3, 1.5 or 1.7 GHz processors when the technology or speed difference is not of importance, and we specify "POWER4+" to refer exclusively to the 1.5 and 1.7 GHz processors.

[1] Please check the announcement letter for the exact availability date of each new feature.

1.2 pSeries 670 and pSeries 690 characteristics

The pSeries 670 and pSeries 690 family of servers incorporate advanced technologies available from across the IBM @server family, as well as technology enhancements from IBM research divisions. The result are high-performance, high-availability servers that offer enhanced features and benefits.

The pSeries 670 and pSeries 690 servers are based on a modular design. They feature a Central Electronics Complex (CEC) (where memory and processors are installed), power subsystem, and I/O drawers, also known as Remote I/O (RIO) drawers. Optional battery backups can provide energy for an emergency shutdown in case of a power failure.

The CEC provides room for up to 16 POWER4 chips, each one housing two processors and a shared level 2 (L2) cache. The POWER4 chip offers an advanced microprocessor design, with an SMP design in a single silicon substrate.

Advanced multichip module (MCM) packaging places four POWER4 chips with four or eight processors, mounted on a ceramic package. This provides enhanced processor reliability, and faster interconnections between processors and L2 caches. Each MCM connects to two memory card slots through the level 3 (L3) cache modules.

Redundant power supplies and optional battery backups assure that a single power failure does not interrupt system operation. In case of total power failure, an emergency shutdown is initiated whenever the battery is present. The redundant power supplies are connected through separate power cords, in order to prevent a single failure on the external electrical circuit.

Building on IBM @server zSeries® heritage, the pSeries 670 and pSeries 690 deliver true logical partitioning and can be divided respectively into up to 16 or 32 partitions[2], each with its own set of system resources, such as processors, memory, and I/O. Unlike partitioning techniques available on other UNIX servers, logical partitioning provides greater flexibility and finer granularity. Resources can be assigned in any amount or combination for business-critical applications. Each logical partition has its own version of an operating system, being either AIX 5L Version 5.1, AIX 5L Version 5.2, or Linux.

The following sections detail some of the technologies behind the pSeries 670 and pSeries 690.

[2] The number of LPAR that can be instantiated on a system depends on its hardware configuration. Please refer to the *The Complete Partitioning Guide for IBM @server* pSeries Servers, SG24-7039 for explanation of the LPAR configuration rules, and to 2.3.6, “I/O books” on page 47 in this redbook.

1.2.1 Microprocessor technology

The POWER4 chip is a result of advanced research technologies developed by IBM. Numerous technologies are incorporated into the POWER4 to create a high-performance, high-scalability chip design to power future IBM @server pSeries systems. Some of the advanced techniques used in the design and manufacturing processes of the POWER4 include copper interconnects and Silicon-on-Insulator.

Copper interconnects

As chips become smaller and faster, aluminum interconnects, which have been used in chip manufacturing for over 30 years, present increasing difficulties. In 1997, after nearly 15 years of research, IBM scientists announced a new advance in the semiconductor process that involves replacing aluminum with copper. Copper has less resistance than aluminum, which permits the use of smaller circuits with reduced latency that allows for faster propagation of electrical signals. The reduced resistance and heat output make it possible to shrink the electronic devices even further while increasing clock speed and performance without resorting to exotic chip cooling methods.

Silicon-on-Insulator

Silicon-on-Insulator (SOI) refers to the process of implanting oxygen into a silicon wafer to create an insulating layer and using an annealing process until a thin layer of SOI film is formed. The transistors are then built on top of this thin layer of SOI. The SOI layer reduces the capacitance effects that consume energy, generate heat, and hinder performance. SOI improves chip performance by 25 to 35 percent.

Detailed information on the POWER4 chip can be found in 2.3.1, “POWER4 processor and MCM packaging” on page 22.

1.2.2 Memory subsystem

The memory hierarchy in pSeries 670 and pSeries 690 systems is represented as follows:

Internal level 1 (L1) caches	Each microprocessor core has its own L1 instruction and L1 data caches.
Shared L2 cache	Inside a POWER4 chip, both processor cores share an L2 cache.

L3 cache — It operates at a 3:1 ratio with the clock speed, and connects to the MCMs through a dedicated port. Each L3 cache module has a direct connection with one POWER4 chip and one memory slot.

The memory used in the pSeries 670 and pSeries 690 systems is Double Data Rate (DDR), which provides superior bandwidth. The minimum memory in the pSeries 670 system is 4 GB, in the pSeries 690 it is 8 GB, and it can be expanded to up to 256 GB and 512 GB, respectively. The pSeries 670 and pSeries 690 are the first pSeries servers to incorporate level 3 (L3) cache. This contributes to faster data access and faster processing capabilities.

Additional information about memory can be found in 2.3.2, “Memory subsystem for pSeries 690” on page 29, and in 5.4.2, “Memory redundancy mechanisms” on page 164.

1.2.3 I/O drawer subsystem

The standard system comes with one I/O drawer that features 20 Hot-Plug PCI slots, and supports more than 2.3 TB of storage located in 16 drive bays connected to integrated Ultra3 SCSI controllers. All power, thermal control, and communication systems have redundancy to eliminate outages caused by single component failures. Each I/O drawer connects to the system through two or four RIO ports, with up to 8 GB/s of total burst bandwidth per drawer[3].

Up to three drawers can be added to the pSeries 670 to obtain a total of 60 PCI slots and more than 7 TB of internal disk storage. The pSeries 690 can accommodate up to eight drawers, supporting a total of 160 PCI slots and more than 18 TB of internal disk storage. For more information on the I/O subsystem, please refer to 2.4, “I/O subsystem” on page 51.

[3] In the case of an RIO drawer connected to the CEC through two RIO loops.

A picture of the IBM @server pSeries 690 is shown in Figure 1-1.

Figure 1-1 The IBM @server pSeries 690

1.2.4 Hardware technology

The pSeries 690 introduces several advanced technologies into its design, and implements other available technologies with enhancements, truly delivering mainframe-class reliability and an outstanding performance. Some key technologies used are referenced here, as follows:

- Many self-healing, self-optimizing, and fault-avoidance mechanisms.
- zSeries packaging and power distribution, recognized as highly reliable.
- Advanced microprocessor design, with POWER4 dual processors, and multichip module (MCM) packaging.
- Logical partitioning (LPAR) offers true flexibility when selecting resources for partitions, and dynamic logical partitioning (DLPAR) allows for moving resources non-disruptively between logical partitions.

The pSeries 670 and pSeries 690 systems are the most reliable UNIX servers ever built, with the pSeries 690 the most powerful UNIX machine on the market. To achieve the outstanding capabilities that these servers offer, IBM has combined two methods to select the technology: Developing new technology, and enhancements to existing technology.

Developing new technologies is one of the most challenging tasks, especially in the high-end UNIX market. It means increased risks, because of the cost of development and the possibility of missing the product release schedule. As a huge technological step in the high-end UNIX market, IBM developed several new technologies for the pSeries 690, such as the POWER4 processor and distributed switch technology. These technologies are now used across the whole range of the pSeries Family, from the entry model pSeries 610 to the high-end pSeries 690 and the clustered model pSeries 655. Hardware technology is shared with the IBM @server iSeries™.

Enhancing existing technologies is an excellent way to exploit solutions already developed, and add new functions and features. It enables technology reuse, lowering costs and enabling a new product to use the best technology available for each specific requirement. The zSeries power subsystem and packaging and the MCM design are among those technologies that have been enhanced on the pSeries 670 and pSeries 690.

1.3 Logical partitioning

There is a strong demand for high-end systems to provide greater flexibility, in particular the ability to subdivide them into smaller partitions that are capable of running a version of an operating system or a specific set of application workloads.

IBM initially started work on partitioning S/370™ mainframe systems in the 1970s. Since then, logical partitioning on IBM mainframes (now called IBM @server zSeries) has evolved from a predominantly physical partitioning scheme based on hardware boundaries, to one that allows for virtual and shared resources with dynamic load balancing. In 1999 IBM implemented LPAR support on the AS/400® (now called IBM @server iSeries) platform. In 2000 IBM announced the ability to run the Linux operating system in an LPAR or on top of VM on a zSeries server, to create thousands of Linux instances on a single box.

Throughout this publication we refer to the different partitioning mechanisms available on the market. Therefore, it is appropriate to clarify the terms and definitions by which we classify these mechanisms.

- A *building block* is a collection of system resources, such as CPUs, memory, and I/O connections. These may be physically packaged as a self-contained SMP system (rack-mounted or stand-alone) or as boards within a larger multiprocessor system. There is no requirement for the CPUs, memory, and I/O slots to occupy the same physical board within the system, although they often do. Other vendors use the terms *system board*, *cell*, and *Quad Building Block (QBB)* to refer to their building blocks.
- A *physical partition* consists of one or more building blocks linked together by a high-speed interconnect. Generally, the interconnect is used to form a single coherent memory address space. In a system that is only capable of physical partitioning, a partition is a group of one or more building blocks configured to support an operating system image. Other vendors may refer to physical partitions as *domains* or *nPartitions*.
- A *logical partition* is a subset of logical resources that are capable of supporting an operating system. A logical partition consists of CPUs, memory, and I/O slots that are a subset of the pool of available resources within a system.

Note: The major difference between logical partitioning and physical partitioning is the granularity and flexibility available for allocating resources to an operating system image. Logical partitions have finer granularities than physical partitions.

It should be noted that the zSeries LPAR implementation is unique in comparison to the other partitioning implementations available from IBM and other hardware vendors. It is the most mature and dynamic partitioning technology in the industry. IBM experience with physical and logical partitioning over the last 25 years has greatly influenced the design and implementation of logical partitioning on pSeries.

The pSeries 690 were the first pSeries servers to incorporate the ability to be partitioned (This is now also available on pSeries 630, 650, 655, and 670). Their architectural design brings true logical partitioning to the UNIX world, being capable of up to 32 partitions inside a single server, with great flexibility in resource selection. The partitioning implementation on the pSeries 670 and pSeries 690 differs from those of other UNIX system vendors in that the physical resources that can be assigned to a partition are not limited by internal physical system board boundaries.

Processors, memory, and I/O slots can be allocated to any partition, regardless of their locality. For example, two processors on the same POWER4 silicon chip can be in different partitions. PCI slots are assigned individually to partitions, and memory can be allocated in fixed-size increments. The granularity of the resources that can be assigned to partitions is very fine, providing flexibility to create systems with just the desired amount of resources.

The pSeries 670 and pSeries 690 systems can also be shipped with additional capacity, which may be purchased and activated at a certain point in time without affecting normal machine operation. This feature is referred to as Capacity Upgrade on Demand (CUoD), which provides flexibility and fine granularity in processor and memory upgrades. A feature that comes with CUoD is Dynamic Processor Sparing. This is the capability of the system to configure out a failing processor and configure in a non-activated CUoD processor. In this way, the number of activated processors is guaranteed by IBM whenever a processor failure would occur. Details on this new offering can be found in Chapter 4, “Capacity Upgrade on Demand” on page 133.

The pSeries 670 and pSeries 690 are also capable of running Linux inside a partition, and having both AIX 5L Version 5.1, AIX 5L Version 5.2, and Linux running on the system simultaneously. The pSeries 690 is the first high-end UNIX server capable of running Linux, and the only one supporting Linux concurrently with other operating systems.

For more information on logical partitioning technology on the pSeries 690, refer to the *Partitioning for the IBM @server* pSeries 690 *System* white paper, available at:

`http://www.ibm.com/servers/eserver/pseries/hardware/whitepapers/lpar.html`

Also refer to *The Complete Partitioning Guide for IBM @server* pSeries Servers, SG24-7039.

1.4 Dynamic logical partitioning

Starting from AIX 5L™ Version 5.2, IBM supports dynamic logical partitioning (also known as DLPAR) in partitions on several logical-partitioning capable IBM @server pSeries server models, including the pSeries 670 and pSeries 690. The dynamic logical partitioning function allows you to add and remove resources, such as CPUs, memory, and I/O slots, to and from a partition as well as to move resources between two partitions, without operating system reboot (on the fly). However, to use the function, the following conditions must be met, because it is achieved by the combination of these three components:

- The pSeries 670 and pSeries 690 systems must be updated with the 10/2002 system microcode update or later.
- The IBM Hardware Management Console for pSeries (HMC) must be updated with Release 3 or higher, and must be connected through Ethernet to the each partition.
- A partition that requires the dynamic logical partitioning function must be installed with AIX 5L Version 5.2, or must be migrated to AIX 5L Version 5.2 from AIX 5L Version 5.1.

1.5 General overview of pSeries 670 and pSeries 690

The pSeries 670 has some differences from the pSeries 690. Basically it has the same technical advantages and Reliability, Availability, and Serviceability (RAS) features, but it is a single rack server and has several configuration limitations compared to the pSeries 690. Table 1-1 provides a quick overview of the differences between these two models.

Table 1-1 Differences between pSeries 670 and pSeries 690

Component	pSeries 690	pSeries 670
Number of CPUs / MCMs	8-32/1-4	4-16/1-2
4 way 1.1GHz POWER4	No	Yes
8 way 1.1GHz POWER4	Yes	Yes
8 way 1.3GHz POWER4 Turbo	Yes	No
4 way 1.3GHz POWER4 HPC	Yes	No
4 way 1.5GHz POWER4+	No	Yes
8 way 1.5GHz POWER4+	Yes	Yes
8 way 1.7GHz POWER4+	Yes	No
Amount of memory (GB)	8-512	4-256
Number of memory cards	1-8	1-4
Number of I/O drawers	1-8	1/2[a] - 3
Number of PCI slots	20-160	10-60
IBF feature	Yes	Yes

Component	pSeries 690	pSeries 670
Number of racks	1-2	1

a. As an initial order, just one half I/O drawer can be used with 10 PCI slots.

The MCMs used in the pSeries 670 are either 4- or 8-way modules. Since it is only possible to install two MCMs in the pSeries 670, only the inward-facing memory card slots can be used.

Note: It is possible to upgrade a pSeries 670 into a pSeries 690 by model conversion.

1.6 Market positioning

The pSeries 690 is one of the most powerful UNIX servers ever built, with the best performance per processor in many application areas. The same holds for the pSeries 670 in the midrange. From copper-based, Silicon-on-Insulator (SOI) POWER4 microprocessor technology and multichip packaging expertise to the industry's most advanced self-management capabilities, the pSeries 670 and pSeries 690 are designed to provide the ultimate in performance, configuration flexibility, and availability.

The pSeries 670 and pSeries 690 are positioned to fulfill the applications' need for performance and scalability, and the reliability required in mission-critical environments. It is an ideal server on which to run corporate applications such as Online Transaction Processing (OLTP), Enterprise Resource Planning (ERP), Business Intelligence (BI), and high-performance e-business infrastructures. It also provides attractive facilities for server consolidation with its flexible partitioning mechanism, as well as the dynamic logical partitioning function, possibilities, and advanced management functions.

Performance wise, the pSeries 670 and pSeries 690 excel in commercial processing, and the pSeries 690 set many new records in High Performance Computing (HPC) applications. The balanced architecture, together with high performance processors, offer unparalleled price/performance, helping to reduce costs not only on the server itself, but also on power, cooling, and software licenses.

In the case of very resource-consuming applications, or for consolidation of a very high number of servers, it is possible to group several pSeries 670 and pSeries 690 into clusters, (optionally with other pSeries, RS/6000® and SP servers) known as IBM @serverCluster 1600.

Additional clustering information can be found at:

http://www.ibm.com/servers/eservers/clusters/hardware/1600.html

For more information on the AIX 5L operating system and related IBM products, the following link may be of interest:

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

For application availability on the AIX 5L operating system, the following link can be used for alphabetical listing and advanced search options, both for IBM software products and third-party software products.

http://www.ibm.com/servers/aix/products/

1.7 Supported operating systems

The goal of this redbook is to present the hardware features of the pSeries 670 and pSeries 690 servers. It does not have a designated chapter providing information about the operating systems that are supported on this hardware. However, throughout the book, when a feature requires a specific level of software, we will mention the prerequisite.

The following sections provide general information and pointers to the relevant documentation about operating systems you can run on these servers:

- "AIX 5L Version 5.1" on page 12
- "AIX 5L Version 5.2" on page 13
- "Linux - SuSE" on page 13

1.7.1 AIX 5L Version 5.1

When using AIX on pSeries 670 or pSeries 690, AIX 5L Version 5.1 is the minimum level that must be installed. It is the first version of AIX supporting the POWER 4 processor technology. Any prior version of AIX, including AIX 4.3.3, is not supported on pSeries 670 or pSeries 690.

AIX 5.1 provides support for the features that were announced on the first generation of pSeries 690. It therefore does not support some of the features that were announced later. In particular, AIX 5.1 does not support:

- Dynamic LPAR
- Memory Capacity Upgrade on Demand
- Dynamic Processor Sparing
- Dynamic CPU Guard

While models that support greater than 256 GB memory sizes have the potential to define greater than 256 GB logical partition sizes, AIX 5.1 logical partitions have a maximum logical partition memory size of 256 GB. AIX 5.1 partitions defined to have greater than 256 GB would fail to activate with an insufficient real mode memory failure. AIX 5.2 and Linux partitions can be greater than 256 GB in size.

1.7.2 AIX 5L Version 5.2

AIX 5L Version 5.2 is the latest available version of AIX. As a general rule, all features available with the pSeries 670 or pSeries 690 are supported by this level of AIX. Some features may require a specific maintenance level.

1.7.3 Linux - SuSE

The Linux distribution which is supported on the pSeries 670 or pSeries 690 is the SuSE Linux Enterprise Server 8. Other distributions of Linux are not officially supported on pSeries 670 or pSeries 690, although some distributors have announced their intention to port their products to this hardware platform.

To find the latest information about the Linux products supported on pSeries, refer to these Web sites:

http://www.ibm.com/servers/eserver/pseries/linux/
http://www.ibm.com/servers/eserver/pseries/linux/whitepapers/linux_pseries.pdf

Information about SuSE Linux Enterprise Server 8 can be found at:

http://www.suse.com/us/business/products/server/sles/i_pseries.html

Release notes for the SuSE Linux Enterprise Server 8 are available at:

http://www.ibm.com/servers/eserver/pseries/linux/sles8_release_notes.pdf

There are some restrictions that you must be aware of regarding Linux on pSeries 670 or pSeries 690:

- Linux-SuSE can only be installed on partitions and not supported on the Full System Partition.
- Only a subset of the hardware options (disks, adapters, ...) are supported. For the latest list of supported features, refer to:

  http://www.ibm.com/servers/eserver/pseries/hardware/linux_facts.pdf
- We do not recommend using Linux in partition with a large number of processors. Usually, Linux applications perform at their best with less than 4 to 8 processors in the partition.
- CuOD is not available for Linux systems.

Redbooks are available that specifically address the topic of Linux on pSeries. In particular, we recommend that you read *Linux Applications on pSeries*, SA24-6033, for information about installing Linux on pSeries 670 or pSeries 690, or porting Linux applications onto AIX using the AIX toolbox for Linux applications.

1.7.4 Comparison of RAS supported features

The hardware and firmware of pSeries 670 or pSeries 690 provides for many RAS features, which are described in Chapter 5, “Reliability, availability, and serviceability” on page 157.

Table 1-2 Comparison of AIX and Linux support for RAS features

RAS functions	AIX	Linux
Service Focal Point	Yes	No
I/O Errors Reporting	Yes	Yes
EEH Detection	Yes	Yes
EEH Recovery	Yes	No
Concurrent Diagnostics	Yes	No
Hot Plug (I/O, SCSI)	Yes	No
Inventory Scout	Yes	No
Update System Firmware	Yes	Yes
Update IOA Firmware	Yes	No
Platform errors to syslog	Yes	Yes
EPOW Warnings	Yes	No
lscfg	Yes	Partial
Scan Dump	Yes	Yes
DASD Mirroring	Yes	No
Service Agent	Yes	No
Error Log Analysis	Yes	No
Predictive callout	Yes	No
Dynamic CPU Deallocation	Yes	No
snap	Yes	Yes

RAS functions	AIX	Linux
lsvpd	Yes	Yes
lsmcode	Yes	Yes
SRC codes(IPL prog, Machine check, dump state	Yes	No
HACMP Options	Yes	No

The RAS features also need some support by the operating system, so that the system administrator can use them. Table 1-2 on page 14 lists which features are available in AIX or Linux.

Some features may require a specific maintenance level of code to be supported.

For many features that are not supported by Linux, there are statements of the direction for support in future versions of the operating system.

1.7.5 Installation and backup of the operating systems

For a pSeries 670 or pSeries 690, all internal media bays on the front are connected to one I/O slot, while all internal media bays in the back are connected to another I/O slot. Therefore, a maximum of two partitions can be directly connected to a set of internal media bays at any one time. In many cases, because of the performance and flexibility of the systems, more than two partitions are configured on a pSeries 670 or pSeries 690. In this section, we briefly describe the three following options to install and back up multiple partitions on a pSeries 670 or pSeries 690.

- Share over time a set of internal media (CD-ROM and tape drive) between all the partitions. This option is likely to have the lowest cost and works with both AIX and Linux. But note that unless you are running AIX 5L Version 5.2, you would need to shutdown the partitions that are affected when transferring the media bays from one partition to another. Another point to take into account is that at any one time, a maximum of two partitions can be installed or backed up if you have the maximum number of internal media bays.
- Provide a set of media (CD-ROM and tape drive) for each partition. If there is more than two partitions, you need to configure additional external CD-ROM, tape drives and the required adapters on the I/O drawers. This option can be used for both AIX and Linux. The advantage if this option is that all the partitions can be scheduled to be backed up at the same time or different times, but there is an additional cost for the external devices.
- Provide a set of internal media to a partition and use the AIX Network Installation Manager (NIM) to install and manage the partitions. This option is only supported with AIX. NIM provides capabilities, such as installation, fixes

through the network, and also creates system backups using the **mksysb** command. For more details about NIM, refer to *The Complete Partitioning Guide for IBM @server* pSeries Servers, SG24-7039.

2

Hardware architecture of the pSeries 670 and pSeries 690

Both the pSeries 670 and pSeries 690 are based on a modular design, where all components are mounted in one or two 24 inch racks. There are three major subsystems: the Central Electronics Complex (CEC), the power subsystem, and the I/O subsystem. In addition, all these components are managed by an external management workstation, called the Hardware Management Console (HMC). Each of these subsystems and the HMC are explained in detail in this chapter:

- 2.3, “Central Electronics Complex” on page 19
- 2.4, “I/O subsystem” on page 51
- 2.5, “Power subsystem” on page 72
- 2.6, “IBM Hardware Management Console for pSeries” on page 75

The pertinent information about the differences between the pSeries 690 and the pSeries 670 is also provided in this chapter.

The support of the Capacity Upgrade on Demand feature on the pSeries 670 and pSeries 690 servers requires some specific hardware features and introduces additional configuration rules. All CUoD specific aspects are described in Chapter 4, “Capacity Upgrade on Demand” on page 133, and are not addressed in this chapter.

2.1 What's new in the pSeries 670 and pSeries 690

In May 2003, IBM has announced new hardware features on the pSeries 690 and pSeries 670 servers:

- In addition to the 1.1GHz and 1.3 GHz chip using POWER4 technology, the customer can now also choose processors using the POWER4+ technology, with a clock speed of 1.5 or 1.7 GHz. These new processors are described in 2.3.1, "POWER4 processor and MCM packaging" on page 22, as well as their new associated L3 cache modules.
- As for the L3 cache, new memory modules are announced, that support a faster clock speed to be used with the new POWER4+ processors. In addition, a larger 64 GB memory board is now supported. These memory modules and the new configuration possibility they offer are described in 2.3.2, "Memory subsystem for pSeries 690" on page 29 and 2.3.4, "Memory subsystem for pSeries 670" on page 42.
- A completely new I/O subsystem, using a faster technology (called RIO-2), is mandatory to support the new POWER4+ processors, and can also be used in conjunction with the POWER4 processors. It consists of new I/O books, described in 2.3.6, "I/O books" on page 47, new I/O planers for the RIO drawers (see 2.4.1, "I/O drawer" on page 51), and a completely new set of cabling options between the I/O books and the RIO drawers, explained in 2.4.2, "I/O subsystem communication and monitoring" on page 56.

2.2 Modular design of the pSeries 670 and pSeries 690

Both the pSeries 670 and pSeries 690 are rack-based servers, housed on the same 24-inch wide, 42 EIA height rack used by the IBM @server zSeries. Inside this rack all the server components are placed in specific positions. This design and mechanical organization offers advantages in optimization of floor space usage.

Table 2-1 shows the main subsystems of the pSeries 690 and pSeries 670, and the sections where they are described.

Table 2-1 pSeries 690 main subsystems

Subsystem name	Section number
Central Electronics Complex (CEC)	2.3, "Central Electronics Complex" on page 19
I/O subsystem	2.4, "I/O subsystem" on page 51 2.4.4, "Media drawer" on page 70
Power subsystem	2.5, "Power subsystem" on page 72

A representation of these components inside the base rack is shown in Figure 2-1.

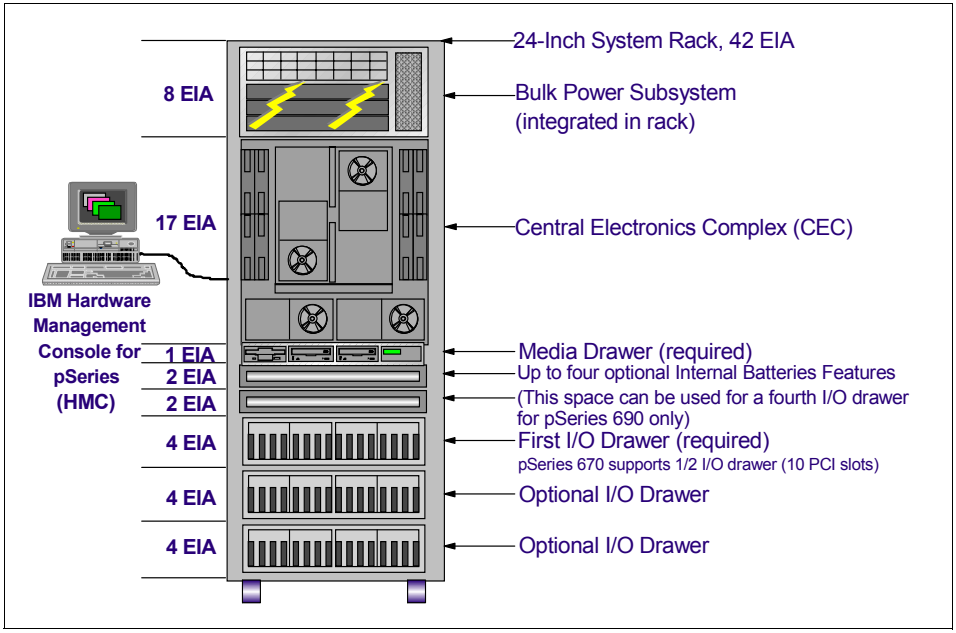


Figure 2-1 The pSeries 670 and pSeries 690 base rack with components

2.3 Central Electronics Complex

The Central Electronics Complex is a 17 EIA height drawer housing the processors and memory of the pSeries 670 and pSeries 690. The CEC contains the following components:

- The CEC backplane, where the components are mounted
- The multichip modules (MCMs), which contain the POWER4 processors
- Memory cards
- L3 cache modules
- I/O books, which provide the Remote I/O (RIO) ports for the I/O drawers, and the service processor function
- Fans and blowers for CEC cooling

Major design efforts have contributed to the development of the pSeries 670 and pSeries 690 to analyze single points of failure within the CEC to either eliminate them or to provide hardening capabilities to significantly reduce their probability of failure.

The front view of CEC is shown in Figure 2-2 on page 20. There are eight memory card slots available for pSeries 690 and four for pSeries 670. For detailed information about memory, see 2.3.2, “Memory subsystem for pSeries 690” on page 29, and 2.3.4, “Memory subsystem for pSeries 670” on page 42.

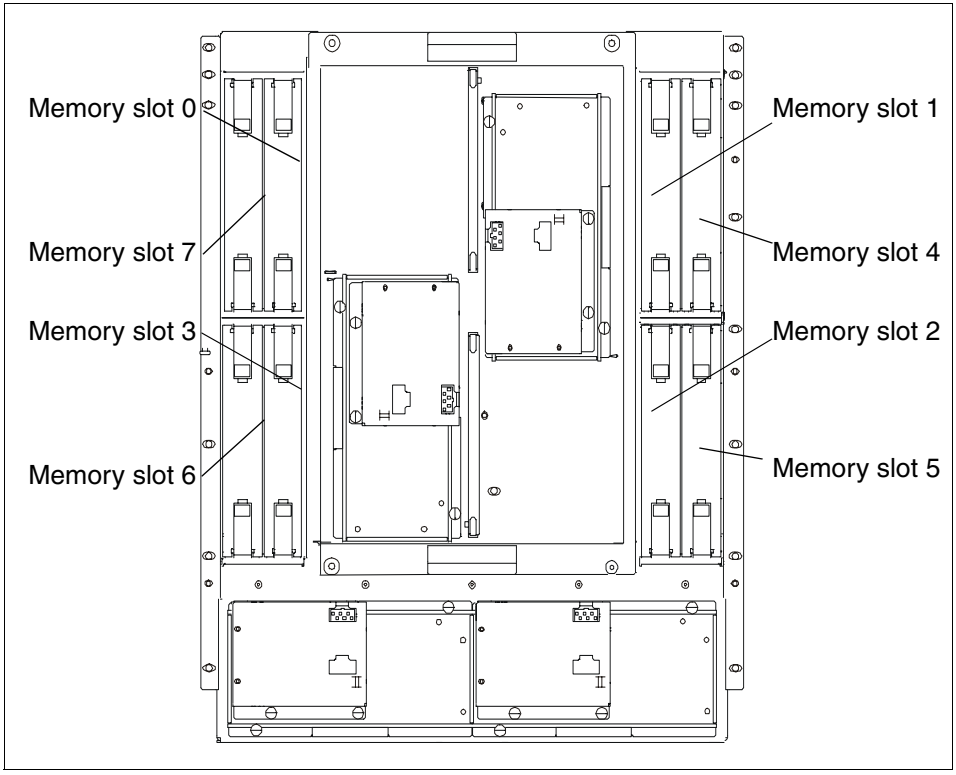


Figure 2-2 CEC front view

The rear view of CEC is shown in Figure 2-3 on page 21. The following slots are provided:

- Slots for distributed converter assembly (DCA) books and two capacitor books

  These slots are populated by up to six DCA books and up to two capacitor books. They supply electricity power to the CEC backplane and convert voltage.

- GX bus slots 0-3

  The GX bus slot 0 is used to insert the primary I/O book. The GX bus slots 1, 2, and 3 are used for the optional secondary I/O books. In Figure 2-3, the slot 2 is populated with an IO book, while slots 1 and 3 are not. The pSeries 670 is restricted to at most one secondary I/O book, while the pSeries 690 can have from zero to three secondary books. For detailed information about the order in which to populate the GX slots, see 2.3.3, “MCMs and GX slots relationship for pSeries 690” on page 39, and for description of the books, see 2.3.6, “I/O books” on page 47.

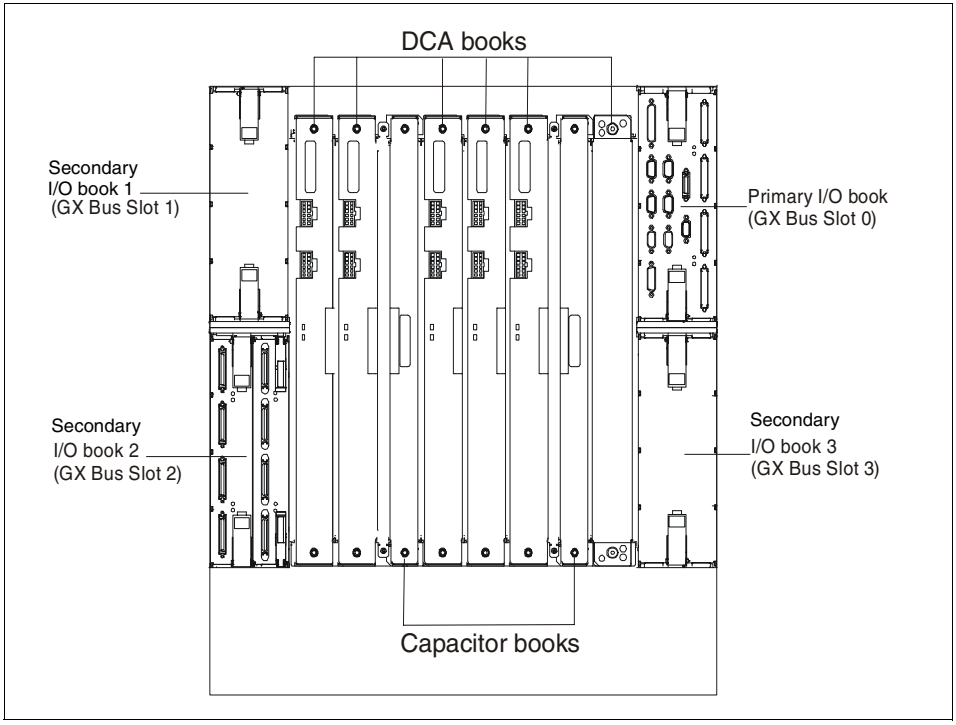


Figure 2-3 CEC rear view

In the center of the CEC, a CEC backplane is positioned vertically. As shown in Figure 2-4 on page 22, it provides mount spaces for up to four multichip modules (MCMs), sixteen level 3 (L3) cache modules, the eight memory cards and the four I/O books. In the center position of the backplane is the clock card. It distributes sixteen paired clock signals to the four MCMs. Please note that the four GX bus slots (shown as “I/O Card Slots Books” in Figure 2-4 on page 22) are located in the rear side of the CEC backplane.

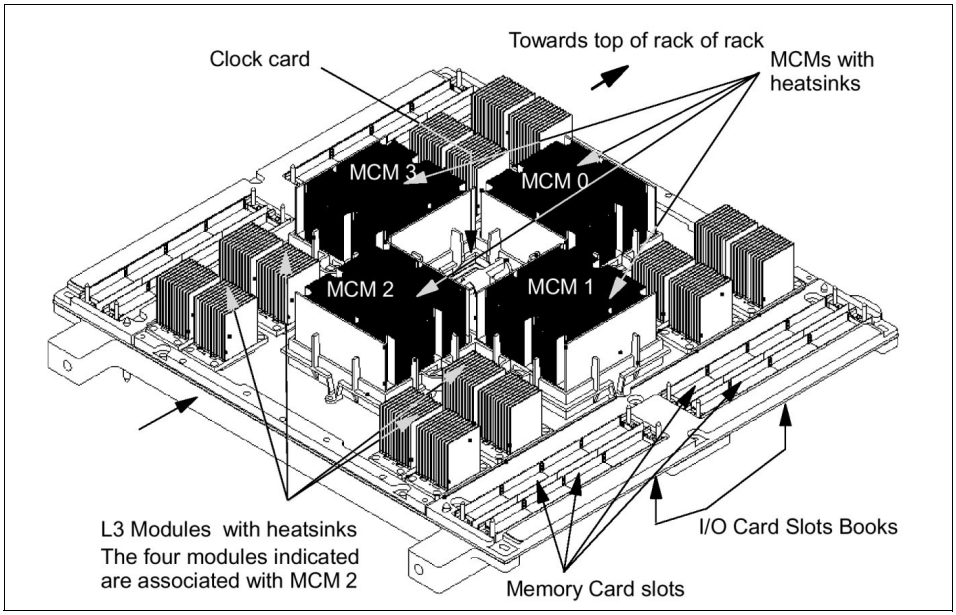


Figure 2-4 CEC backplane orthogonal view

2.3.1 POWER4 processor and MCM packaging

Several years ago, scientists working for microprocessor development set out to design a new microprocessor that would leverage IBM strengths across many different disciplines to deliver a server that would redefine what was meant by the term *server*. POWER4 is the result. It was developed by over 300 engineers in several IBM development laboratories. A new generation of this chip is available since May 2003, which is called POWER4+. It has the same architecture as the initial POWER4 chip, but is able to sustain higher clock frequency.

The pSeries 670 and pSeries 690 are based on these high performance POWER4 chips. Up to 32 processors can be configured, and they are mounted together in multichip modules. This packaging methodology was carefully designed to provide increased levels of reliability by eliminating the tiered packaging levels of separate processor modules mounted on processor cards; these were mounted on a backplane. A pSeries 690 can have up to four MCMs installed.

The pSeries 670 is using the same high performance POWER4 chips, but it only supports up to 16 processors, which is a total of two MCMs installed.

Several versions of the POWER4 chips are available on pSeries 670 and pSeries 690 servers, with either 4 or 8 CPUs, executing at different speeds: 1.1GHz, 1.3

GHz, 1.5 GHz or 1.7 GHz. All processors within pSeries 670 or pSeries 690 server, must be identical and operate at the same speed. The following table describes the supported combinations of processors.

Table 2-2 Supported combinations of processors

Architecture	Clock frequency	Number of Processors per MCM	pSeries 670	pSeries 690	Designation
POWER4	1.1 GHz	4	4-way	N/A	Standard
		8	8-way or 16-way	8-way, 16-way, 24-way or 32-way	Standard
	1.3 GHz	4	N/A	8-way or 16-way	HPC
		8	N/A	8-way, 16-way, 24-way or 32-way	Turbo
POWER4+	1.5 GHz	4	4-way	N/A	
		8	8-way or 16-way	8-way, 16-way, 24-way or 32-way	
	1.7 GHz	8	N/A	8-way, 16-way, 24-way or 32-way	Turbo

The 1.1 GHz and 1.3 GHz 4-way MCMs are slightly different from the others. They use a POWER4 chip with a single core. The 1.3 GHz 4-way MCM (called the HPC feature) is specifically described in “Single core POWER4 processor feature” on page 28. It is a special option for technical applications that demand high memory bandwidth.

The chip has two processor cores and a level 2 (L2) cache, all in the same silicon substrate. The L2 cache is shared between the two cores through a crossbar switch known as the core interface unit (CIU). All these components are part of a single POWER4 chip, along with communication buses and a level 3 (L3) cache directory.

The following sections give a brief explanation about the POWER4 and MCM technologies.

For further detailed information about POWER4 and memory subsystem in the pSeries 690, refer to the following whitepapers and redbook:

- *POWER4 System Microarchitecture*, found at:

  http://www.ibm.com/servers/eserver/pseries/hardware/whitepapers/power4.html

- *IBM @server* pSeries 690 *Configuring for Performance*, found at:

  http://www.ibm.com/servers/eserver/pseries/hardware/whitepapers/p690_config.html

- The *POWER4 Processor Introduction and Tuning Guide*, SG24-7041

POWER4 core

The internal micro-architecture of the POWER4 core is a speculative superscalar Out-of-Order execution design. Up to eight instructions can be issued each cycle, with a sustained completion rate of five instructions. Register rename pools and other Out-of-Order resources coupled with the pipeline structure allow the design to have over 200 instructions in flight at any given time.

In order to exploit instruction level parallelism, there are eight execution units, each capable of being issued an instruction each cycle. Two identical floating-point execution units, each capable of starting a fused multiply and add each cycle are provided. This combination can achieve four floating-point operations (FLOPs) per cycle per core. In order to feed the dual floating-point units, two load/store units, each capable of performing address generation arithmetic, are provided.

The microprocessor cores inside a POWER4 chip are a new design, although they support the same PowerPC® instruction set architecture as prior pSeries processors. Each core has several execution units, along with level 1 (L1) instruction and data caches, and run at clock speeds of up to 1.7 GHz. Each L1 instruction cache is 64 KB and each L1 data cache is 32 KB. All data stored in the L1 data cache is available in the L2 cache, guaranteeing no data loss.

The initial POWER4 manufacturing process, known as *CMOS-8S3SOI*, implements seven-layer copper metallization, and 0.18 µm SOI CMOS technology. The POWER4+ manufacturing process is called *CMOS-9SSOI* with a 0.13 µm SOI CMOS technology. Along with performance improvements, these technologies deliver high-reliability components that IBM considers to be fundamental for high-end, continuous operation servers. All the buses scale with processor speed.

L2 cache subsystem

The L2 cache on the POWER4 chip is composed of three separate cache controllers, connected to the two processor cores through a core interface unit. In POWER4 and POWER4+ chips, each L2 cache contains, respectively, 480 KB and 512KB, for a total of 1.44 MB per POWER4 chip, and 1.5 MB per POWER4+ chip.

Each L2 cache controller can operate concurrently and feed 32 bytes of data per cycle. The CIU connects each of the three L2 controllers to either the L1 data cache or the L1 instruction cache in either of the two processors.

Additionally, the CIU accepts stores from the processors across 8-byte wide buses and sequences them to the L2 controllers. Each processor has a non-cacheable (NC) unit associated with it, responsible for handling instruction serializing functions and performing any non-cacheable operations in the memory hierarchy. In a logical view, the NC units are part of the L2 cache.

The L2 cache on the POWER4 chip is dedicated to delivering data to the two cores as fast as they can process it, maximizing the overall performance. It has multiple ports and is capable of multiple concurrent operations. The L2 cache can provide data at the following peak rates:

- 1.1 GHz Standard system: 105.6 GB/s to the two cores
- 1.3 GHz Turbo feature: 124.8 GB/s shared to the two cores
- 1.3 GHz HPC feature: 83.2 GB/s to the single core
- 1.5 GHz POWER4+ system: 144 GB/s shared to the two cores
- 1.7 GHz POWER4+Turbo feature: 163.2 GB/s shared to the two cores

L3 cache controller and directory

Each POWER4 processor chip, not each core, controls one L3 cache module. The L3 module itself is located outside the chip, but the L3 directory and L3 controller are located on the POWER4 chip. A separate functional unit, referred to as the Fabric Controller, is responsible for controlling data flow between the L2 and L3 controller for the chip and for POWER4 communication.

Each POWER4 chip has its own interface to the off chip L3 across two 16-byte wide buses, operating at one third of the processor frequency. To communicate with I/O devices, two 4-byte wide GX buses, operating at one third processor frequency, are used. Finally, each chip has its own Joint Test Action Group (JTAG) interface to the system service processor.

Also included on the chip are functions called pervasive functions. These include trace and debug facilities used for First Failure Data Capture (FFDC), Built-in Self Test (BIST) facilities, Performance Monitoring Unit, an interface to the service processor used to control the overall system, Power On Reset (POR) Sequencing logic, and error detection and logging circuitry.

Multichip modules

The POWER4 chips are packaged on a single module called multichip modules. Each MCM houses four chips (eight CPU cores) that are connected through chip-to-chip ports.

The chips are mounted on the MCM such that they are all rotated 90 degrees from one another, as shown in Figure 2-5. This arrangement minimizes the interconnect distances, which improves the speed of the inter-chip communication. There are separate communication buses between processors in the same MCM, and processors in different MCMs (see Figure 2-6 on page 27).

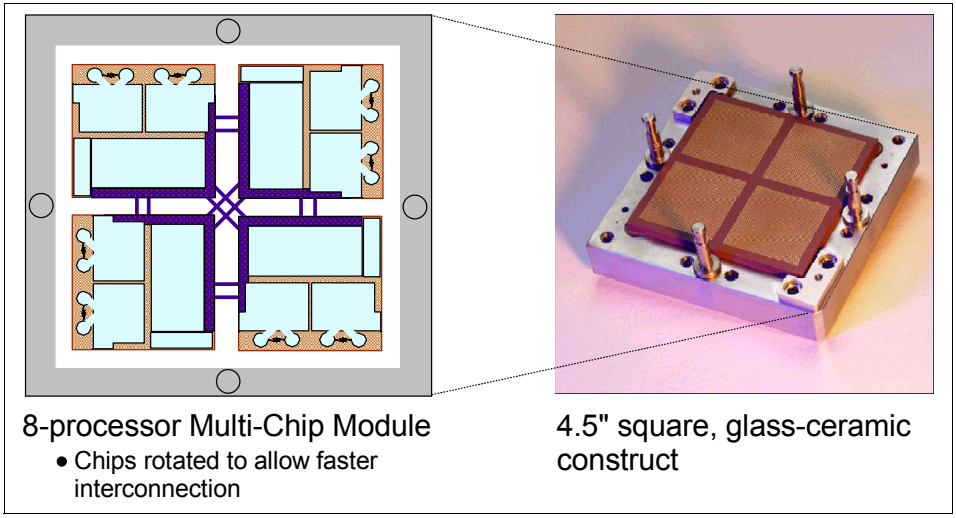


Figure 2-5 POWER4 multichip module

An internal representation of the MCM is shown in Figure 2-6, with four interconnected POWER4 chips. Each installed MCM comes with 128 MB of L3 cache. This provides 32 MB of L3 cache per POWER4 chip. The system bus (L3 cache, GX Bus, memory nest) operates at a 3:1 ratio with the processor frequency. Therefore the L3 cache to MCM connections operate at:

- 375 MHz for 1.1 GHz processors
- 433 MHz for 1.3 GHz processors
- 500 MHz for 1.5 GHz processors
- 567 MHz for 1.7 GHz processors

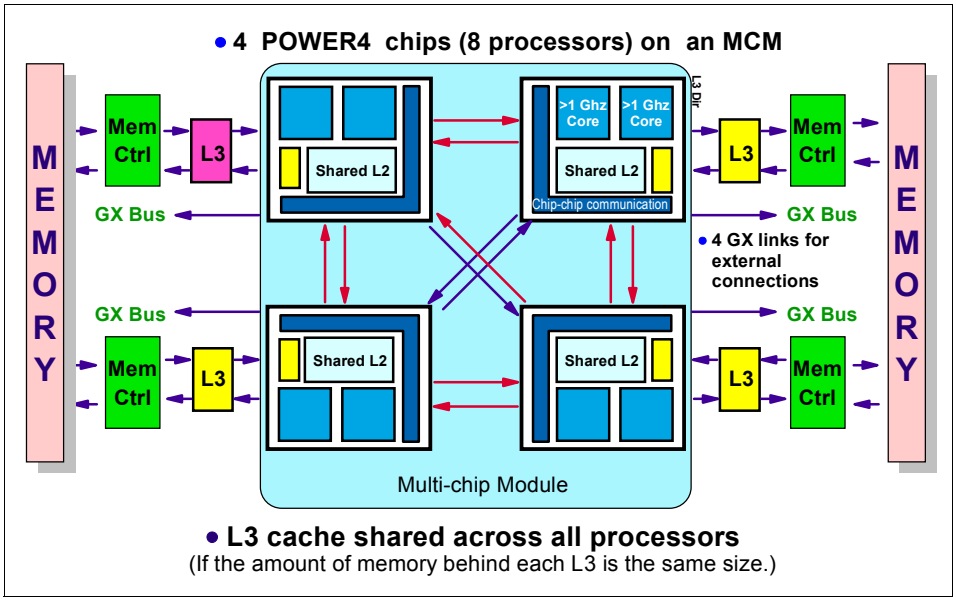


Figure 2-6 Multichip module with L2, L3, and memory

The MCM is a proven technology that IBM has been using for many years in the mainframe systems (now IBM @server zSeries). It offers several benefits in mechanical design, manufacturing, and component reliability. IBM has also used MCM technology in the RS/6000 servers in the past. The IBM RS/6000 Model 580 was based on the POWER2™ processor that has all its processing units and chip-to-chip wiring packaged in an MCM.

Single core POWER4 processor feature

As stated before, some technical applications benefit from very large bandwidth between processors and memory. The POWER4 processor delivers an exceptional bandwidth to the cores inside. For those applications that require extremely high bandwidth, the high performance computing (HPC) feature is an attractive alternative. Instead of 8-way MCMs, you have 4-way MCMs with the same amount of L2 and L3 caches and the same bus interconnection (see Figure 2-7).

This configuration provides twice the amount of L2 and L3 cache per processor and additional memory bandwidth, when compared to the pSeries 690 configured with 8-way processor MCMs. This additional cache and memory bandwidth available for each processor in this configuration may provide significantly higher performance per processor for certain engineering and technical environment applications.

Note: The HPC feature is only available on the pSeries 690 with a POWER4 1.3 GHz processor. It is not offered with the POWER4+ 1.5 or 1.7 GHz processors.

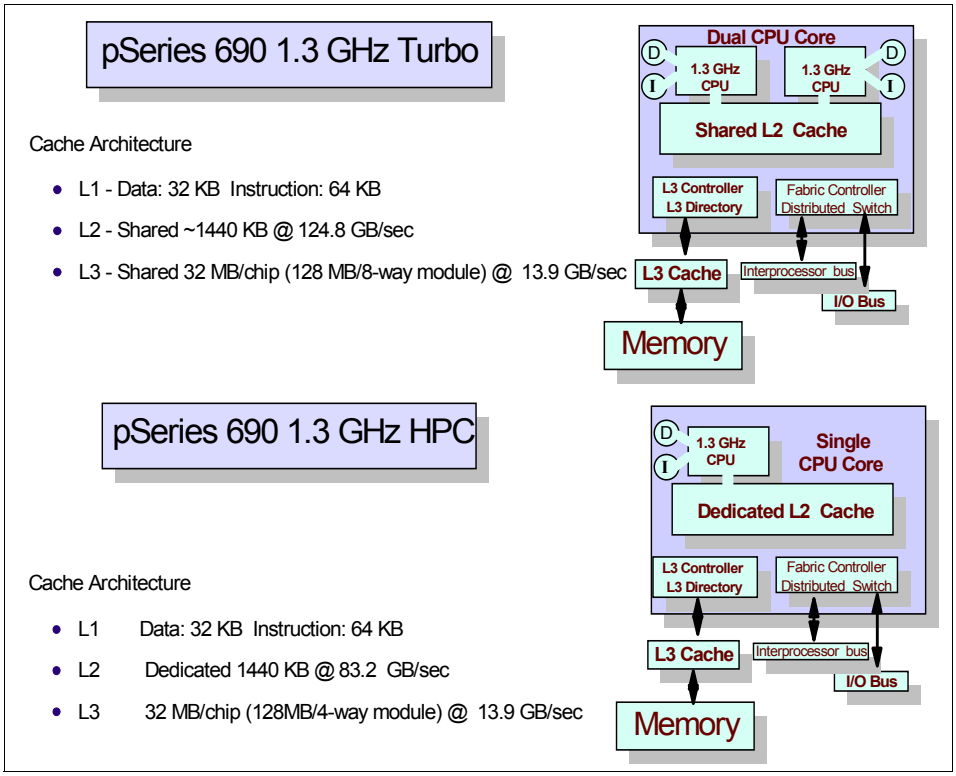


Figure 2-7 pSeries 690 1.3 GHz Turbo and 1.3 GHz HPC feature

2.3.2 Memory subsystem for pSeries 690

The memory subsystem in the pSeries 690 is physically composed of L3 cache modules and up to eight memory cards, both mounted on the CEC backplane. Since each L3 cache module is controlled by an L3 controller that physically resides in a POWER4 processor chip packaged in an MCM, there are tight relationships between MCMs, L3 cache modules, and memory cards.

MCM population order

The pSeries 690 can be configured with up to four MCMs mounted on the backplane, as shown in Figure 2-8. The MCM 0 is always configured. The other optional MCMs are populated in this order: MCM 2, MCM 1, and MCM 3. In a two-MCM configuration (MCM 0 and MCM 2), because the spaces for MCM 1 and MCM 3 are not populated, you have to have a two-processor bus pass-through modules in this space to establish the link between two populated MCMs. In a three-MCM configuration (MCM 0, MCM 1, and MCM 2), since the space for MCM 3 is not populated, you have to have a processor bus pass

through a module in this space. For the configuration rules of multiple MCMs, see 3.2.3, “Processor configuration rules” on page 85.

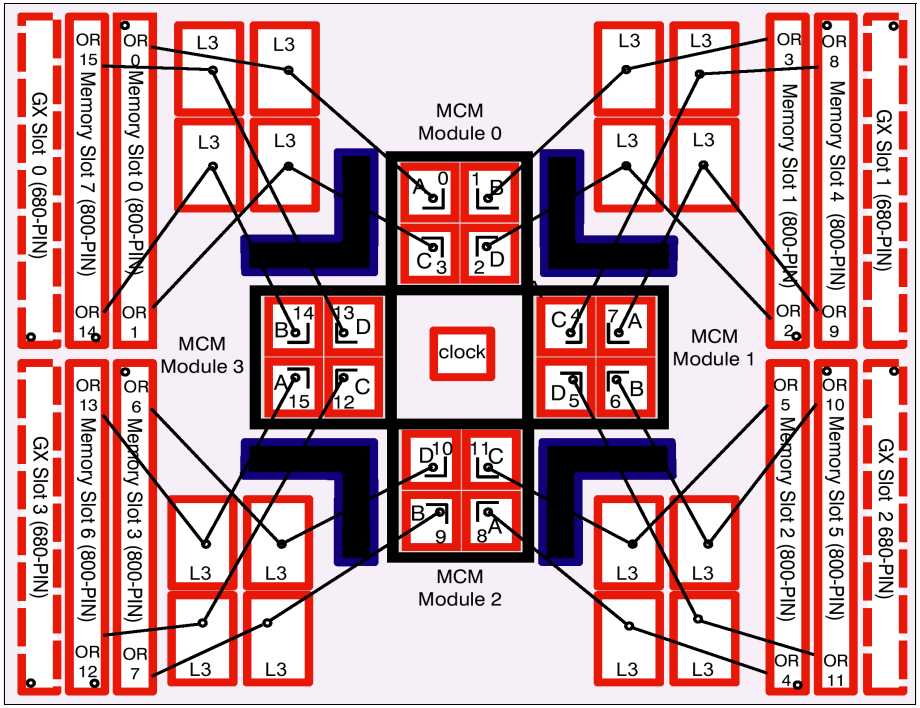


Figure 2-8 MCM, L3 cache, and memory slots relationship on backplane

Figure 2-9 provides a logical view of the relationship between MCMs and memory slots.

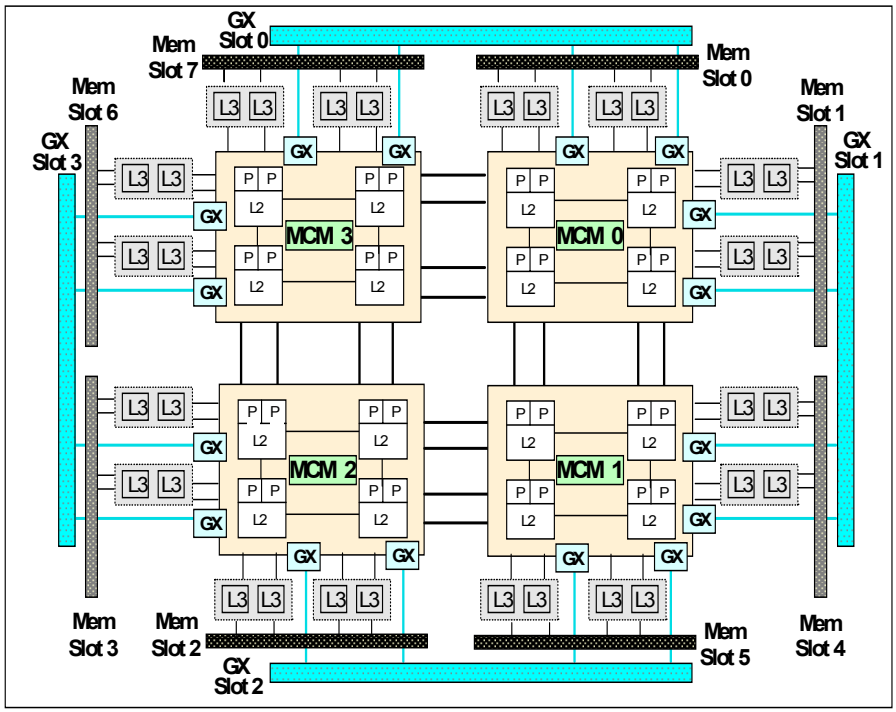


Figure 2-9 Logical relationship between MCMs, memory, and GX Slots

Multiple MCM configuration

In a four-MCM configuration, there is no processor bus pass-through module required, because all the space is populated with MCMs. In this configuration, there are sixteen 32 MB L3 caches, 512 MB total, and four rings interconnecting the MCMs, as shown in Figure 2-10 on page 32.

In two- and three-MCM configurations, unpopulated spaces are replaced with processor bus pass-through modules. If one of the MCMs is replaced with a processor bus pass-through module, MCMs are still connected through a four-ring interconnection. However you lose access to the specific memory slots under the unpopulated MCMs, because a processor bus pass-through module does not provide access to the memory slots directly connected to it. The processor bus modules also do not have access to the GX bus (see 2.3.3, “MCMs and GX slots relationship for pSeries 690” on page 39).

Forming the four rings interconnecting the MCMs, a processor can access all the memory cards.

- If the referenced address is in the memory cards under the directly attached L3 cache to the MCM, then interconnection between processor cores within

MCM is used to reference the memory. This access is explained in “Single MCM configuration” on page 32.

- If the referenced address is not in the memory cards under the directly attached L3 cache to the MCM, then interconnection between MCMs is used for access. This memory access is achieved by a very small overhead.

See Figure 2-10 for more details about the interconnection between the four MCMs.

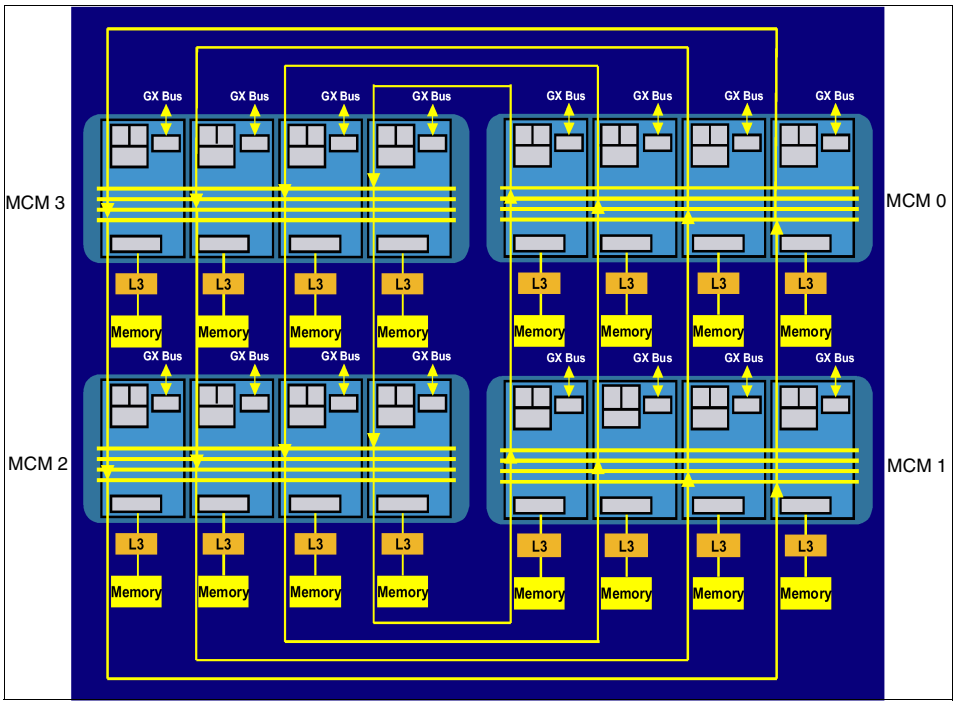


Figure 2-10 Interconnection between four MCMs

Single MCM configuration

In a single MCM configuration (always MCM 0), there is no processor buss pass-through module required, because there are no interconnections between MCMs. In this configuration, a processor can access the memory cards in slot 0 and 1, but cannot access the other memory cards. Within this MCM, each processor can access all the memory addresses through its own L3 cache, or through the interconnection between processors within MCM, as shown in Figure 2-11.

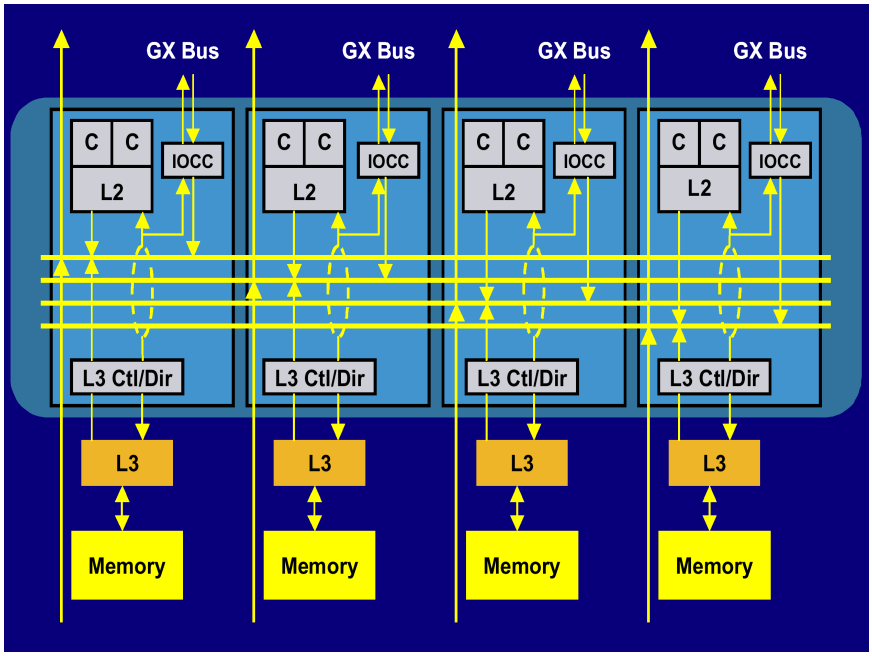


Figure 2-11 Interconnection between processors in an MCM

Access to memory through L3 cache

Each POWER4 processor chip controls a 32 MB L3 cache, and each MCM has access to two memory cards through these four L3 caches, 128 MB total, as shown in Figure 2-11.

The general rule is that the two memory slots attached to an MCM are populated with two identically sized memory cards. Then these four L3 caches act as a single interleaved 128 MB sized L3 cache. Combining L3 cache chips into a group not only increases the L3 cache size, but also scales the available L3 bandwidth. When combined into a group, four L3 caches and the memory behind them are interleaved on 512 byte granularity.

In two special cases (Single 8 GB memory card attached to an 8-way system and single 4 GB memory card in a pSeries 670 system), the two memory slots are populated with two differently sized memory cards: one of the memory slots is unpopulated. Then the four L3 caches act as two separate 64 MB sized L3 caches. For optimal performance, memory cards of equal capacity should be attached to each MCM.

Each memory card has two memory controller chips, which are connected to an L3 cache module and to synchronous memory interface (SMI) chips.

Each memory controller chip can have one[1] or two[2] memory data ports and can support up to 32 GB of memory.

Two sets of memory cards can be found in the pSeries 690 servers:

- The cards in the first set sustain a bus frequency up to 500 MHz. They are exist in several sizes: 4 GB, 8 GB, 16 GB, and 32 GB (the 16 GB, and 32 GB cards will no longer be available after May 2003).
- The cards in the second set sustain a bus frequency up to 567 MHz. They are available in several sizes: 4 GB, 8 GB, 16 GB, 32 GB, and 64 GB.

All models of pSeries 670 and pSeries 690 can use the 500 MHz cards and the 16 GB, 32 GB, and 64 GB at 567 MHz cards. The 4GB and 8GB at 567 MHz memory features are only available with the 1.7 GHz processor configurations.

The memory controller is attached to the L3 modules, with each module having two 16-byte buses to the data interface in the memory controller. These buses operate at one-third processor speed using the Synchronous Wave Pipeline Interface to operate at high frequencies.

Restriction: If a 500 MHz card is used with a 1.7 GHz processor, it is not able to run at a 3:1 ratio of the processor speed. In this case, it will be detected as a slow card, and as a result, all memory buses and all memory cards in the system will operate at a 4:1 ratio of the processor speed: 425 MHz.

The memory modules are Dual Data Rate (DDR) and Synchronous DRAM (SDRAM), soldered on the memory card for improved reliability, and use book packaging.

Memory card configuration for pSeries 690

The pSeries 690 has eight memory slots. Memory is available in two types: Inward-facing and outward-facing. The inward-facing memory slots (0, 1, 2, and 3) are enabled by the MCM 0 and MCM 2, while the outward-facing memory slots (4, 5, 6, and 7) are enabled by the MCM 1 and MCM 3 (see Figure 2-8 on page 30).

As you add more MCMs, you activate more memory ports. It increases the configurable memory size, but also increases the memory bandwidth. Table 2-3 summarizes this relationship.

[1] In 4 GB and 8 GB memory cards.
[2] In 16 GB, 32 GB, and 64 GB memory cards.

Table 2-3 Relationship between MCMs, L3 cache, memory slots, and size

Populated MCMs	L3 cache modules	Usable memory slots	Maximum memory size
1	4	2	128 GB
2	8	4	256 GB
3	12	6	384 GB
4	16	8	512 GB

The minimum sized memory, 8 GB, is available with two 4 GB memory cards on a single MCM or with a single 8 GB memory card. The maximum sized memory, 512 GB, is available in a four-MCM configuration, which supports a full eight memory slots, with each populated by a 64 GB memory card.

Note: The whole memory related to an MCM is available to the system, even if all the processors on this MCM are not activated. Customers with applications requiring very large amounts of memory can therefore use the CUoD feature to increase the maximum memory size supported on their server. For example, a server with one 8-way MCM can have up to 128 GB of memory, while an 8-way server with 2 CUoD MCMs (4 active processors and 4 inactive processors) supports up to 256 GB of memory. See Chapter 4, “Capacity Upgrade on Demand” on page 133 for details on memory configurations supported with the CUoD feature.

The optimal amount of memory for a particular system depends upon many factors, such as the requirements of the key applications targeted for the system. However, it is the size and number of memory cards in the system that determine the maximum bandwidth the system can deliver.

Memory should be installed and balanced across all populated MCM positions for optimal performance. Unbalanced configurations will function properly, but the unbalanced portion of the memory will not utilize the full memory bus bandwidth.

Table 2-4 provides supported memory card placement for various memory requirements. The memory configurations with (*) are valid but not recommended. You have to review application performance implications before ordering this memory configuration.

Table 2-4 Supported memory cards configurations

Number of MCMs	Memory size in GB	Memory slots 0 & 1	Memory slots 2 & 3	Memory slots 4 & 5	Memory slots 6 & 7
1	8	4 + 4			
	8 (*)	8 + 0			
	16	8 + 8			
	32	16 + 16			
	64	32 + 32			
	128	64+64			
2	8	4 + 4	0 + 0		
	16	4 + 4	4 + 4		
	16 (*)	8 + 8	0 + 0		
	24	8 + 8	4 + 4		
	32	8 + 8	8 + 8		
	48	16 + 16	8 + 8		
	64	16 + 16	16 + 16		
	96	32 + 32	16 + 16		
	128	32 + 32	32 + 32		
	192	64 + 64	32 + 32		
	256	64 + 64	64 + 64		

Number of MCMs	Memory size in GB	Memory slots 0 & 1	Memory slots 2 & 3	Memory slots 4 & 5	Memory slots 6 & 7
3	8	4 + 4	0 + 0	0 + 0	
	16	4 + 4	4 + 4	0 + 0	
	16 (*)	8 + 8	0 + 0	0 + 0	
	24	4 + 4	4 + 4	4 + 4	
	24 (*)	8 + 8	4 + 4	0 + 0	
	32	8 + 8	4 + 4	4 + 4	
	32 (*)	8 + 8	8 + 8	0 + 0	
	40	8 + 8	8 + 8	4 + 4	
	48	8 + 8	8 + 8	8 + 8	
	64	16 + 16	8 + 8	8 + 8	
	80	16 + 16	16 + 16	8 + 8	
	96	16 + 16	16 + 16	16 + 16	
	128	32 + 32	16 + 16	16 + 16	
	160	32 + 32	32 + 32	16 + 16	
	192	32 + 32	32 + 32	32 + 32	
	256	64 + 64	32 + 32	32 + 32	
	320	64 + 64	64 + 64	32 + 32	
	384	64 + 64	64 + 64	64 + 64	

Number of MCMs	Memory size in GB	Memory slots 0 & 1	Memory slots 2 & 3	Memory slots 4 & 5	Memory slots 6 & 7
4	8	4 + 4	0 + 0	0 + 0	0 + 0
	16	4 + 4	4 + 4	0 + 0	0 + 0
	16 (*)	8 + 8	0 + 0	0 + 0	0 + 0
	24	4 + 4	4 + 4	4 + 4	0 + 0
	24 (*)	8 + 8	4 + 4	0 + 0	0 + 0
	32	4 + 4	4 + 4	4 + 4	4 + 4
	32 (*)	8 + 8	4 + 4	4 + 4	0 + 0
	32 (*)	8 + 8	8 + 8	0 + 0	0 + 0
	40	8 + 8	4 + 4	4 + 4	4 + 4
	40 (*)	8 + 8	8 + 8	4 + 4	0 + 0
	48	8 + 8	8 + 8	4 + 4	4 + 4
	48 (*)	8 + 8	8 + 8	8 + 8	0 + 0
	56	8 + 8	8 + 8	8 + 8	4 + 4
	64	8 + 8	8 + 8	8 + 8	8 + 8
	80	16 + 16	8 + 8	8 + 8	8 + 8
	96	16 + 16	16 + 16	8 + 8	8 + 8
	112	16 + 16	16 + 16	16 + 16	8 + 8
	128	16 + 16	16 + 16	16 + 16	16 + 16
	160	32 + 32	16 + 16	16 + 16	16 + 16
	192	32 + 32	32 + 32	16 + 16	16 + 16
	224	32 + 32	32 + 32	32 + 32	16 + 16
	256	32 + 32	32 + 32	32 + 32	32 + 32
	320	64 + 64	32 + 32	32 + 32	32 + 32
	384	64 + 64	64 + 64	32 + 32	32 + 32
	448	64 + 64	64 + 64	64 + 64	32 + 32
	512	64 + 64	64 + 64	64 + 64	64 + 64

In the configurations listed previously in Table 2-4, you can notice that the first memory slots are populated with the largest memory boards in the first memory slots. This corresponds to the recommended order for inserting memory boards, starting with the largest boards in the first slots, and then installing the following boards in decreasing memory size order. IBM manufacturing will deliver the initial configuration of pSeries 690 with memory installed according to this rule. However, this order is only a recommendation, and there may be cases where it is not respected. For example, if a customer has a system initially configured with 48 GB memory as follows:

16 + 16	8 + 8		

Then the customer orders an MES to increase the memory to 96 GB with one pair of 16 GB memory board and one pair of 8 GB memory board (assuming the system is populated with 4 MCM), the memory configuration will become:

16 + 16	8 + 8	16 + 16	8 + 8

This is a valid and supported configuration. It would not be possible to reorder the memory boards since the 8 GB memory boards initially installed in slots 2 and 3 are inward facings and could not be moved to slots 4, 5, 6, or 7 when the MES memory boards are installed.

Note: Consideration for system upgrades. When initially ordering a pSeries 690 configuration with the intent to order additional hardware later, you must pay attention to the supported memory configurations available with different numbers of MCM. For example, if you order a two MCM pSeries 690 with 256 GB of memory, and later you want to order two extra MCM, you will have to also order at least 128 GB extra memory.

For more information on physical memory configuration and performance considerations, please refer the *IBM @server* pSeries 690 *Configuring for Performance* white paper, found at:

http://www.ibm.com/servers/eserver/pseries/hardware/whitepapers/p690_config.html

2.3.3 MCMs and GX slots relationship for pSeries 690

The number of installed MCMs also affects the number of I/O drawers supported. An MCM is connected to two GX slots through two GX buses using four ports. The GX slot[3] is a physical connector that is used for inserting an I/O book. For detailed information about I/O books, see 2.3.6, “I/O books” on page 47.

[3] Although the GX bus and the GX slot are physically different components, from here on we use *GX slot* to express both of them, since there is a one-to-one relationship between a GX bus and a GX slot.

If there is direct connection to the specific GX slot between a POWER4 chip and this slot, then this path is used. If the GX slot is being accessed from the other chip within an MCM, then an inter-chip connection link is used. If this GX slot is being accessed from a processor chip in the other MCMs only, then an inter-MCM connection link is used. Figure 2-9 on page 31 provides a logical view of the relationship between MCMs and GX slots.

As explained in 2.3.6, “I/O books” on page 47, the primary I/O book has four Remote I/O (RIO) ports, and each of the three optional secondary I/O books has eight RIO ports each.

The GX slots can be populated with IO books in a predefined order: GX Slot 0, then GX Slot 2, GX Slot 3 and finally GX Slot 1. The primary IO book in GX Slot 0 is always installed. The number of installed MCM defines how many secondary IO books can be installed in the GX slots:

- In a single MCM configuration (only MCM 0 is populated), the POWER4 processor chips 0 and 2 are connected to GX Slot 0, and the processor chips 1 and 3 are connected to GX Slot 1, as shown in Figure 2-12 on page 41. But since the GX Slot 1 cannot be used in a single MCM configuration, only the primary I/O book is accessible in GX Slot 0. This means that up to two I/O loops can be configured.
- In a two-MCM configuration, GX Slots 0, 2 and 3 can be populated with IO books. Therefore, 20 RIO ports are physically available, shown in the column Physically Available RIO ports in Table 2-5. Actually, in a two-MCM configuration (MCM 0 and MCM 2 are populated), GX Slot 2 and 3 are accessed from the MCM 2 only. Therefore, in addition to the four ports on GX slot 0, only four out of eight IO ports on GX Slot 2, and only four RIO ports out of eight RIO ports on GX Slot 3 are accessed, for a total of 12 usable ports: only 6 I/O loops[4] are supported at most in a two-MCM configuration.
- In a three-MCM configuration, the four GX bus can be populated with RIO books. Only four RIO ports on GX Slot 3 can be used. A total of 24 RIO ports are usable, for a maximum of 12 RIO loops.
- In a four-MCM configuration, all 28 RIO ports can be used, providing connectivity for up to 14 I/O loops.

[4] An I/O loop requires two RIO ports to be connected.

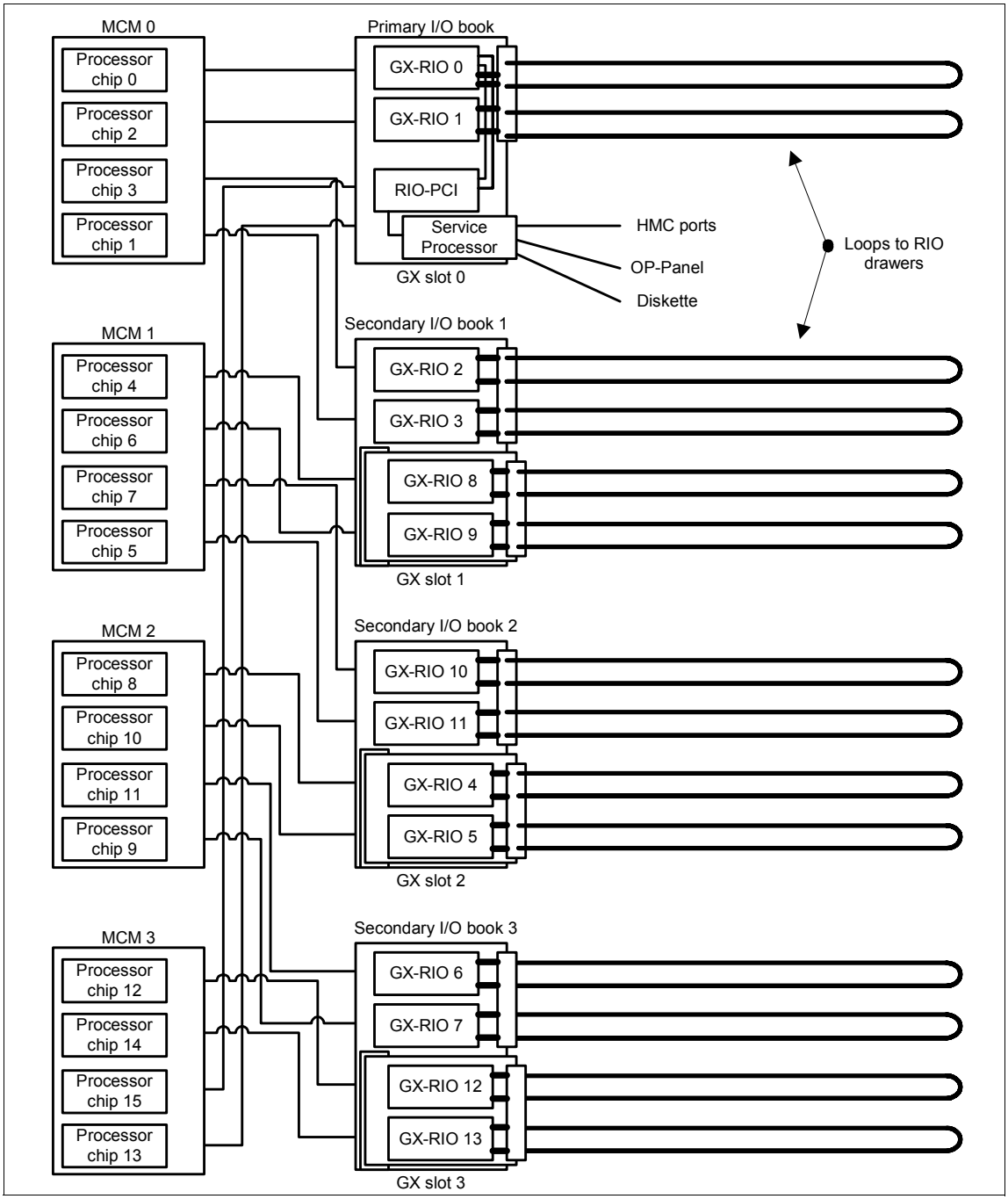


Figure 2-12 MCMs and RIO ports relationship (pSeries 690)

We summarize the supported RIO ports numbers in Table 2-5.

Table 2-5 Relationship between MCMs, RIO ports, and I/O drawers

Populated MCMs	Available GX Slots	Number of Physically existing RIO ports	Number of usable RIO ports
MCM 0	0	4	4
MCM 0, 2	0, 2, 3	20	12
MCM 0, 1, 2	0, 1, 2, 3	28	24
MCM 0, 1, 2, 3	0, 1, 2, 3	28	28

Note: As for memory, the limitations on the number of supported GX slots and IO ports depends on the number of installed MCM, and not the number of activated processors. Therefore, the CUoD feature can be used when a large number of IO drawers are required on a system with “few” processors. For example, a server with one 8-way MCM can have up to two I/O drawers, while a 8-way server with 2 CUoD MCMs (4 active processors and 4 inactive processors) supports up four 4 I/O drawers.

2.3.4 Memory subsystem for pSeries 670

This section only mentions information related to the memory subsystem, which are different for pSeries 670 and for pSeries 690.

Information presented in the these three sections for the pSeries 690 also apply to the pSeries 670:

- “MCM population order” on page 29
- “Multiple MCM configuration” on page 31
- “Single MCM configuration” on page 32

Besides the fact that only the number of MCMs is restricted to two for pSeries 670, the backplane is basically the same as for the pSeries 690. Only two MCMs and four memory slots can be used. MCM positions MCM1 and MCM3 are unused. The minimum required memory with one MCM installed is 4 GB. The maximum memory available with two MCMs is 256 GB. Only the inner memory slots can be used and should be populated in identical pairs. Figure 2-13 on page 43 shows the backplane of pSeries 670.

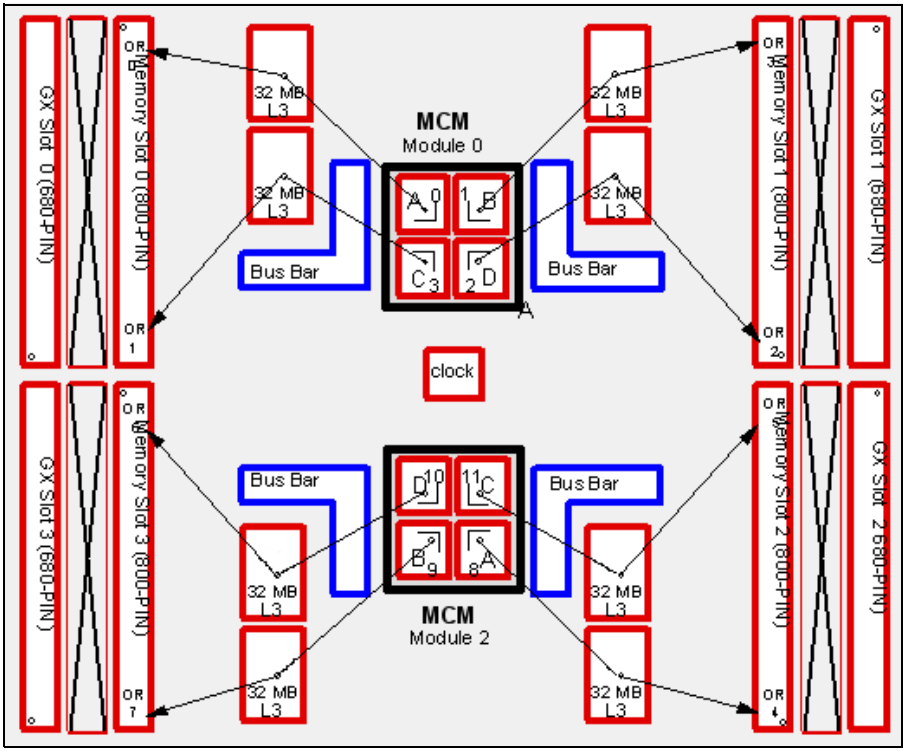


Figure 2-13 MCM, L3 cache, and memory slots relationship on backplane

Memory cards configuration for pSeries 670

The first MCM activates the first and second memory positions. The second MCM activates the remaining third and fourth positions. For best performance you should populate all activated memory card positions with memory cards, and memory sizes should be balanced as closely as possible across all populated MCM locations.

As for pSeries 690, two sets of memory cards can be found in the pSeries 670 servers:

- The 500 MHz set, with memory boards available in sizes: 4 GB, 8 GB, 16 GB, and 32 GB.
- The 567 MHz set, with memory boards available in sizes: 16 GB, 32 GB, and 64 GB.

The memory boards operate at a 1:3 ratio of the processor speed. Table 2-6 shows the supported memory card configurations on pSeries 670.

Table 2-6 Supported memory configurations for pSeries 670

MCMs installed	Memory size in GB	Memory slots 0 & 1	Memory slots 2 & 3
1	4	4 + 0	N/A
	8	4 + 4	
	8[a]	8 + 0	
	16	8 + 8	
	32	16 + 16	
	64	32 + 32	
	128	64+64	
2	8	4 + 4	0 + 0
	16	4 + 4	4 + 4
	16[a]	8 + 8	0 + 0
	24	8 + 8	4 + 4
	32	8 + 8	8 + 8
	48	16 + 16	8 + 8
	64	16 + 16	16 + 16
	96	32 + 32	16 + 16
	128	32 + 32	32 +32
	192	64+64	32 + 32
	256	64+64	64+64

a. This is a valid configuration, but not a recommended configuration.

2.3.5 MCMs and GX slots relationship for pSeries 670

The pSeries 670 shares the same MCM-to-GX slot relationship with pSeries 690, although it only supports up to two MCMs and a maximum of three I/O drawers. Figure 2-14 on page 46 illustrates the relationship between MCMs and GX slots on the pSeries 670. Only two I/O books (GX slot 0 and 2) are supported.

The number of installed MCM defines how many secondary IO books can be installed in the GX slots.

- In a single MCM configuration (only MCM 0 is populated), the POWER4 processor chips 0 and 2 are connected to GX slot 0, and the processor chips 1 and 3 are connected to GX slot 1, as shown in Figure 2-12 on page 41. But since the GX slot 1 cannot be used in a pSeries 670 configuration, only the primary I/O book is accessible in GX slot 0. This means that up to two I/O loops can be configured.
- In a two-MCM configuration, GX slots 0 and 2 can be populated with IO books. Therefore, 12 RIO ports are physically available, shown in the column Physically Available RIO ports in Table 2-5. Actually, in a two-MCM configuration (MCM 0 and MCM 2 are populated), GX slot 2 is accessed from the MCM 2 only. Therefore, in addition to the four ports on GX slot 0, only four out of eight IO ports on GX slot 2 are accessed, for a total of 8 usable ports: only 4 I/O loops[5] are supported at most in a two-MCM configuration.

Table 2-7 summarizes the supported RIO ports numbers.

Table 2-7 Relationship between MCMs, RIO ports, and I/O drawers

Populated MCMs	Available GX Slots	Number of Physically existing RIO ports	Number of usable RIO ports
MCM 0	0	4	4
MCM 0, 2	0, 2	12	8

[5] An I/O loop requires two RIO ports to be connected.

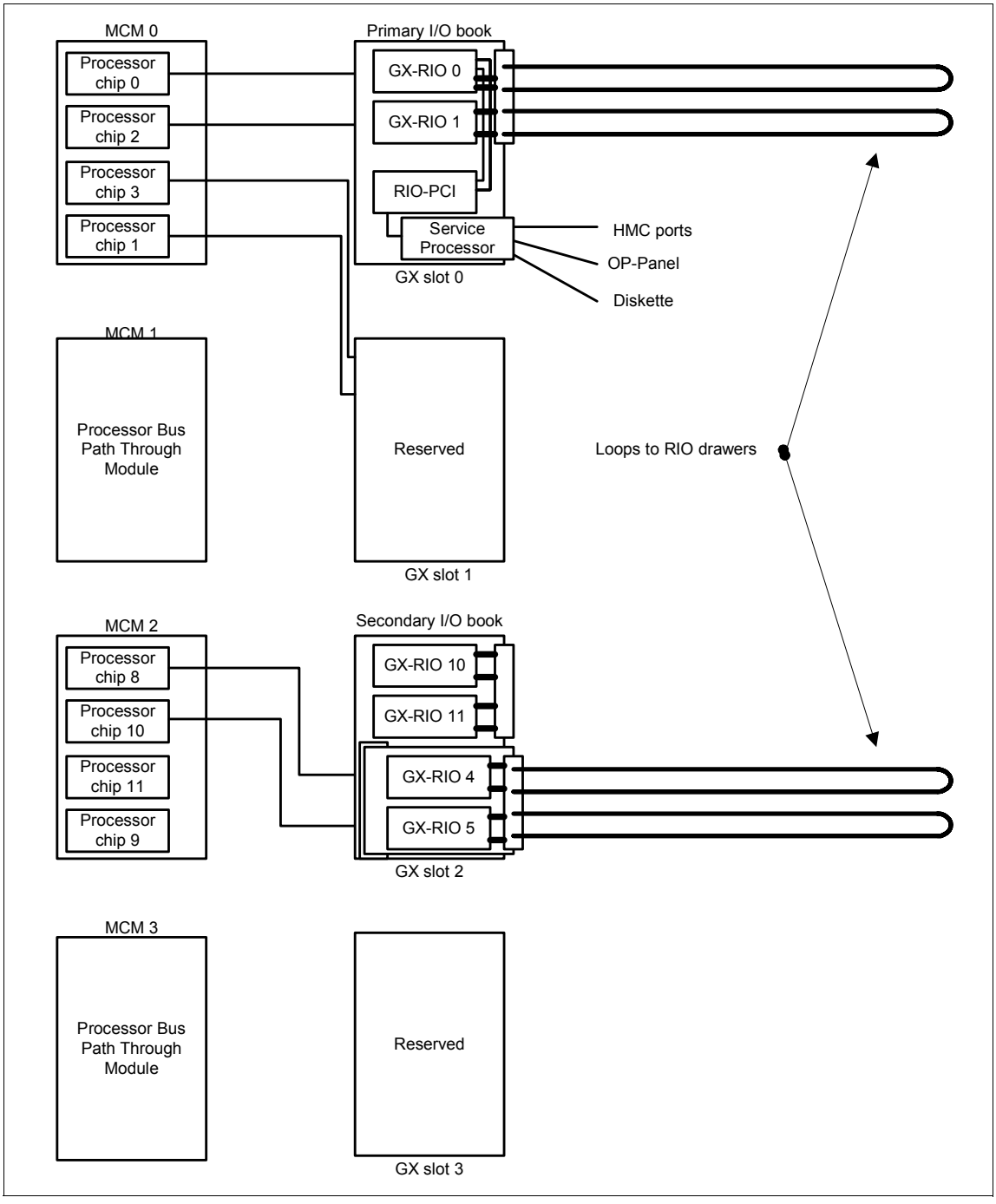


Figure 2-14 MCMs and RIO ports relationship (pSeries 670)

2.3.6 I/O books

The I/O books plug into the GX slots in the CEC backplane and provide the Remote I/O (RIO) ports. The RIO ports are used to connect I/O drawers to the GX bus. Each I/O book contains a base card, a riser, and a daughter card, and is physically packaged using *book packaging* technology, explained in “Book packaging” on page 49.

There are two types of I/O books available, called primary I/O book and secondary I/O book and shown in Figure 2-15 on page 48. The primary I/O book is mandatory in all pSeries 670 and pSeries 690 and there must be exactly one per system. The secondary I/O books are optional. Their number depends on the requirements for external IO drawer attachments. A pSeries 670 can contain at most one secondary I/O book, while a pSeries 690 can contain up to three secondary I/O books.

- **The primary I/O book**
  - Contains the following ports:
    - Two HMC connections ports: HMC1 and HMC2
    - Two native serial ports: S1 and S2

      These two native serial ports are under the control of the service processor and are not available for functions, such as HACMP heartbeat cabling or UPS control, which require fully dedicated ports.
    - An operator panel port
    - A diskette drive port
    - A bulk power controller (BPC) Y-cable connector

      This port is used to connect between two BPCs on the bulk power assembly (BPA) and the service processor using a Y-cable.
  - Contains the service processor.
  - Contains non-volatile random access memory (NVRAM).
  - Contains four RIO ports for the first I/O drawer (mandatory) and an optional second I/O drawer.
  - This I/O book plugs into the GX slot 0 in the CEC backplane.
- **Each secondary I/O book**
  - Contains eight RIO ports.
  - Plugs into the GX slots of the CEC backplane in the order: GX slot 2, GX slot 3 and GX slot 1. In Figure 2-15 on page 48, the RIO Port numbers correspond to the first secondary I/O book. For the second and third

secondary books, the port number would be, respectively, 6, 7, 12, and 13, and 2, 3, 8, and 9.

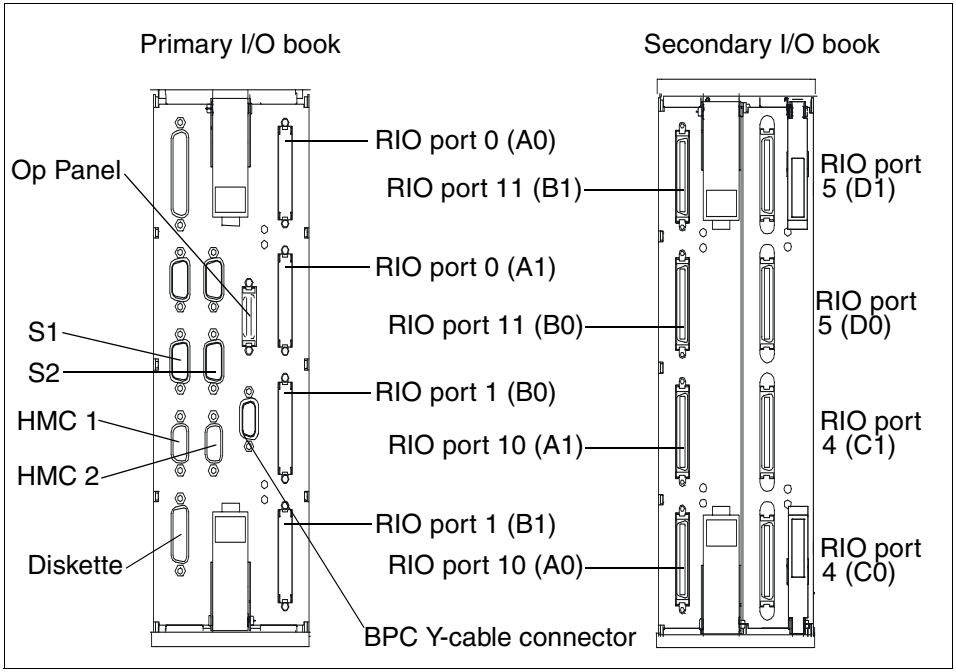


Figure 2-15 Primary and secondary I/O books

For details on connections between the I/O books and the I/O drawers, see 2.4.2, “I/O subsystem communication and monitoring” on page 56.

There are two sets of I/O books, which differ by the technology used in the base card, a riser, and a daughter card installed in the book. The first set is referred to as RIO while the second is called RIO-2.

- The RIO primary and secondary books are:
  - Primary: FC 6404, Service Processor and RIO Loop Attachment, Two Loops
  - Secondary: FC 6410, RIO Loop Adapter, Four Loops
- The RIO-2 books are called:
  - Primary: FC 6418, Service Processor and RIO-2 Loop Attachment, Two Loops
  - Secondary: FC 6419, RIO-2 Loop Adapter, Four Loops

The RIO-2 books provide improvement in I/O performance and throughput required by the new 1.5 and 1.7 GHz processors. The RIO-2 primary book contains a larger NVRAM than the RIO primary book, and provide the increased support from 16 to 32 LPARs. There are restrictions in the use of the books:

- All I/O books must be of the same technology. You cannot mix FC 6404 and 6419 or FC 6418 and FC6410 in the same system.
- The RIO books can only be used with the 1.1 and 1.3 GHz processors.
- The RIO-2 books can be used with the 1.1, 1.3, 1.5, and 1.7 GHz processors.
- When using RIO books, the maximum number of supported LPARs is 16.
- With RIO-2 books, it is possible to instantiate up to 32 LPARs, even with 1.1 GHz and 1.3 GHz systems.

Book packaging

Memory cards, I/O books, and capacitors in the CEC are packaged between metal sheets to form what are called *books*. Book packaging helps protect components from electrostatic discharge and physical damage, and also helps to stabilize electronics and distribute air-flow throughout the CEC for proper temperature control.

Frame guide rails, as shown in Figure 2-16 on page 50, help align the books when connecting to the backplane; guidance pins assure final positioning and two book locks secure the book to the backplane and frame cage. This results in a reduction in pin damage when installing memory and I/O book upgrades, and stabilizes the system for shipment and physical installation.

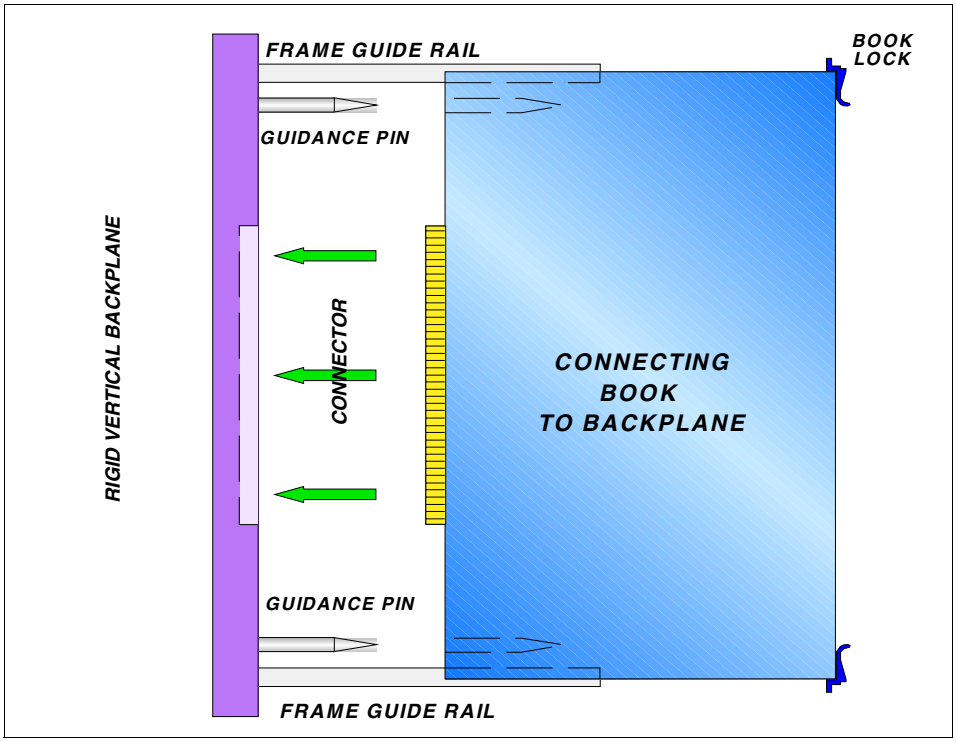


Figure 2-16 Book packaging

2.3.7 Service processor

The service processor is a complete microcomputer, including random access memory and program storage, within a computer system. This auxiliary processor serves two main purposes:

- It initializes and tests the chip logic interconnects that constitute the processor subsystem of the server, configuring them for normal operation.
- It constantly supervises the functional health of the server while the computer system is in operation to detect any failing components as they occur.

The service processor is also responsible for hardware management and control, and executes the required changes in the hardware configuration when creating or modifying a partition. It is the interface between the pSeries 670 and pSeries 690 and the HMC. The service processor is packaged along with a quad RIO port hub adapter in the primary I/O book.

The service processor is responsible for monitoring the entire system, and for storing the machine configuration and partition information in NVRAM.

2.4 I/O subsystem

The I/O drawer subsystem of the pSeries 670 and pSeries 690 uses similar technology to that used in the IBM @server pSeries 660 Model 6M1. The PCI I/O adapters are housed in separate drawers that are connected to the CEC with RIO cables.

Note: The pSeries 670 supports up to three I/O drawers, while the pSeries 690 supports up to eight I/O drawers. The pSeries 670 also supports a half I/O drawer configuration (10 PCI slots).

For removable media, both pSeries 670 and pSeries 690 have a media drawer with configurable devices that can be connected to different logical partitions.

2.4.1 I/O drawer

The I/O drawers provide internal storage and I/O connectivity to the system. Figure 2-17 shows the rear view of an I/O drawer, with the PCI slots and riser cards that connect to the RIO ports in the I/O books.

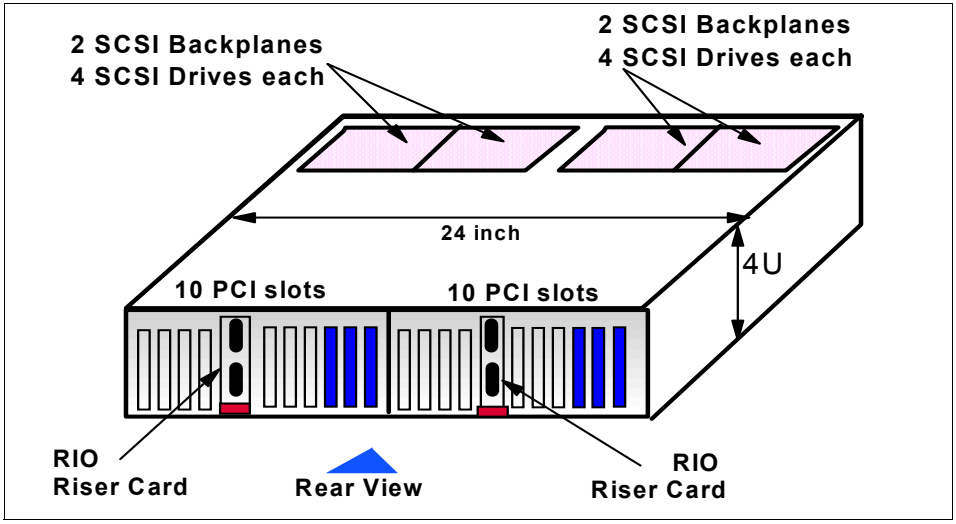


Figure 2-17 I/O drawer rear view

Each drawer is composed of two physically symmetrical I/O planar boards that contain 10 Hot-Plug PCI slots each, and PCI adapters can be inserted in the rear of the I/O drawer. The planar boards also contain two integrated Ultra3 SCSI adapters and SCSI Enclosure Services (SES), connected to a SCSI 4-pack

backplane. The pSeries 670 can be initially configured with only a half I/O drawer with 10 PCI slots and up to eight disk drives installed.

As for I/O books, the I/O drawers exist in two technologies: RIO and RIO-2. I/O drawers of both technologies offer the same number of PCI slots, the same number of disks, and are packaged in the same chassis. The difference is the type of I/O planar that is installed inside the drawer. Externally, the only visible difference between the two technologies is the shape of the cable connectors on the RIO Riser card (see Figure 2-18).

- RIO connectors have a thumbscrew retention physical connector.
- RIO-2 connectors have a bayonet retention physical connector.

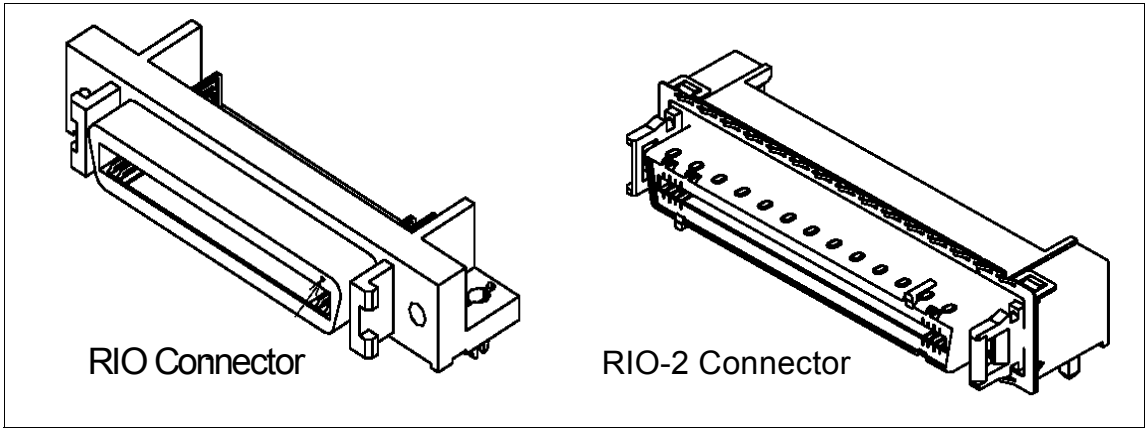


Figure 2-18 Difference between RIO and RIO-2 connectors

The I/O Drawer for pSeries 670 and pSeries 690 has its own product number: 7040-61D, which refers to the two technologies. For ordering purposes, the difference between them is the Feature Code of the configured I/O planar:

- RIO drawer uses the “FC **6563**, I/O Drawer **PCI** Planar, 10 slots, 2 Integrated Ultra3 SCSI Ports”.
- RIO-2 drawer is configured with the “FC **6571**, I/O Drawer **PCI-X** Planar, 10 slots, 2 Integrated Ultra3 SCSI Ports”.

A logical connection view of an RIO I/O drawer is shown in Figure 2-19.

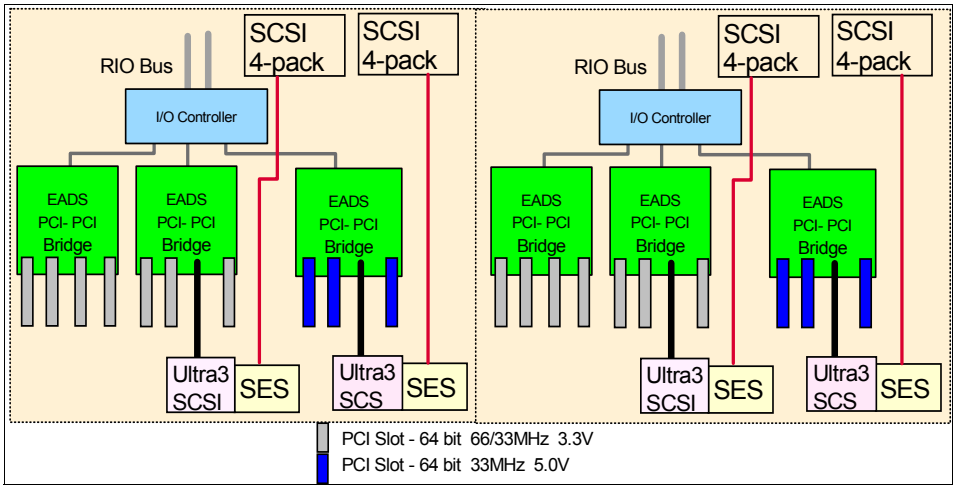


Figure 2-19 Logical view of an RIO drawer

For a RIO-2 I/O drawer, the topology of the logical connection view in identical, as shown in Figure 2-20. The PCI-PCI bridge uses an EADS-X chip instead of an EADS chip. The PCI slots' differences are explained later in this section.

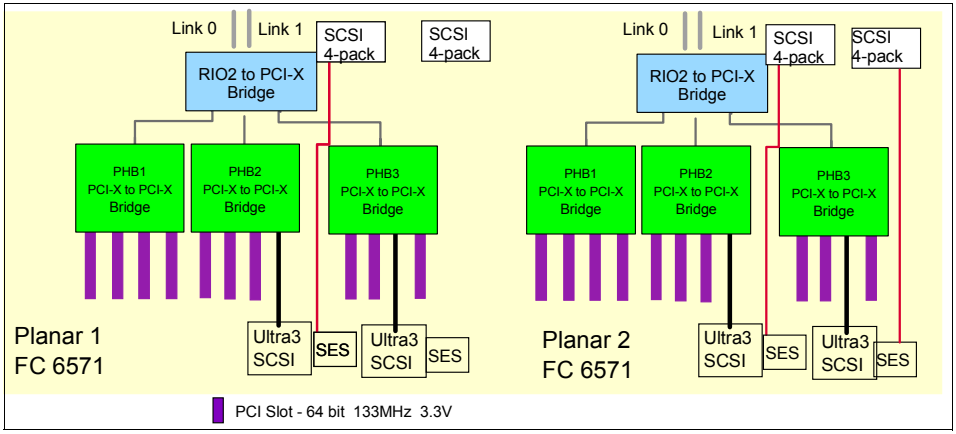


Figure 2-20 Logical view of an RIO -G drawer

Each of the 4-packs supports up to four hot-swappable Ultra3 SCSI disk drives, which can be used for installation of the operating system or storing data.

There are three different disk capacities: 18.2 GB, 36.4 GB, 73.4 GB and 146.8 GB, and they have the following characteristics:

- Form factor: 3.5-inch, 1-inch (25 mm) high
- SCSI interface: SCSI Ultra3 (fast 80) 16 bit
- Rotational speed: 10,000 RPM and 15,000 RPM

Optionally, external disk subsystems can be connected to serve as operating system disks. The pSeries 670 and pSeries 690 and their logical partitions can boot[6] from SCSI, SSA, and Fibre Channel disk subsystems.

The RIO riser cards are connected to the planar boards. The RIO ports of each riser card are connected through IO loops to RIO ports on I/O books in the CEC. The connectivity between the I/O drawer RIO ports and the I/O books RIO ports is described in “Remote I/O loop” on page 57.

The PCI slots have different characteristics in the RIO and RIO-2 drawers:

- **RIO Drawer**

  On each planar board, the first seven PCI slots have a 3.3V PCI bus signaling and operating at 66 MHz or 33 MHz, depending on the adapter. The last three PCI slots have a 5V PCI bus signaling and operating at 33 MHz. All PCI slots are PCI 2.2 compliant and are Hot-Plug enabled, which allows most PCI adapters to be removed, added, or replaced without powering down the system. This function enhances system availability and serviceability.

- **RIO-2 Drawer**

  On each planar board, the ten PCI-X slots have a 3.3V PCI bus signaling and operating at 33 MHz, 66MHz or 133 MHz, depending on the adapter. All PCI slots are PCI 2.2 compliant and are Hot-Plug enabled.

PCI adapters have different bandwidth requirements, and there are functional limitations on the number of adapters of a given type in an I/O drawer or a system. The limitations differ for the RIO and RIO-2 drawers.

The limitations are of several types:

- Maximum number of adapters of one type per IO planar
- Maximum number of adapters of one type per IO drawer
- Maximum number of adapters of one type per LPAR
- Maximum number of adapters of one type per pSeries 690 or per pSeries 670
- Maximum combination of specific adapters

[6] The boot capability from different disk technology depends on the operating system support. AIX supports these types of disk for a boot device.

The complete set of limitations are described in the *PCI Adapter Placement References*, SA38-0538. This book is regularly updated and should be considered as the reference for any questions related to PCI limitations.

Also see the *IBM @server* pSeries 690 *Configuring for Performance* white paper, found at:

http://www.ibm.com/servers/eserver/pseries/hardware/whitepapers/p690_config.html

We only mention two general rules of thumb here:

- All adapters that are supported in the RIO drawers are supported in the RIO-2 drawers, except three of them which use a 5V signaling:
  - FC2751: S/390® ESCON® Channel PCI Adapter
  - FC6206: PCI Single-Ended Ultra SCSI Adapter
  - FC8396: IBM RS/6000 SP System Attachment Adapter
- In an RIO drawers, there is a maximum number of high speed adapters of 5 per planar, and 10 per drawer, while all the I/O slots of an RIO-2 drawers can be populated with high speed adapters (for example, Gigabit Ethernet, Fiber Channel, ATM or Ultra-320 SCSI adapters).

Two special configurations of I/O drawers are supported, which we mention even though they are not recommended:

- A half-RIO drawer with only one I/O planar is available for an entry pSeries 670 with only one I/O drawer. In this case, only 8 SCSI disk slots and 10 I/O slots are usable in the I/O drawer.
- A mixed I/O drawer where one of the I/O planars is an RIO planar, and the other one is an RIO-2 planar. This configuration should only be used in very special cases: when the customer needs one of the three I/O adapters which are not supported with the RIO-2 planar, or when the customer upgrades a half-RIO drawer into a full RIO drawer.

I/O drawer RAS

If there is an RIO failure in a port or cable, an I/O planar board can route data through the other I/O connection and share the remaining RIO cable for I/O.

For power and cooling, each drawer has two redundant DC power supplies and four high reliability fans. The power supplies and fans have redundancy built into them, and the drawer can operate with a failed power supply or a failed fan. The hard drives and the power supplies are hot swappable, and the PCI adapters are hot plug.

All power, thermal, control, and communication systems are redundant in order to eliminate outages due to single-component failures.

2.4.2 I/O subsystem communication and monitoring

There are two main communication subsystems between the CEC and the I/O drawers. The power and RAS infrastructure are responsible for gathering environmental information and controlling power on I/O drawers. The RIO loops are responsible for data transfer to and from I/O devices.

Power and RAS infrastructure

The power cables that connect each I/O drawer and the bulk power assembly (BPA) provide both electricity power distribution and Reliability, Availability, and Serviceability (RAS) infrastructure functions. The BPA is described in 2.5.1, “Bulk power assembly” on page 73.

The primary I/O book has a BPC Y-cable connector that is used to connect between two bulk power controllers (BPCs) on the bulk power assembly (BPA) and the service processor using a Y-cable. Then two BPCs fan out the connection to each of the I/O drawers, media drawers, fans, and Internal Battery Features (IBFs). They compose the logical network (which we refer to as the RAS infrastructure).

The RAS infrastructure includes:

- Powering all system components up or down, when requested. These components include I/O drawers and the CEC.
- Powering down all the system enclosures on critical power faults.
- Verifying power configuration.
- Reporting power and environmental faults, as well as faults in the RAS infrastructure network itself, on operator panels and through the service processor.
- Assigning and writing location information into various VPD elements in the system.

Note: It is the cabling between the RIO drawer and the BPA that defines the numbering of the I/O drawer.

The power and RAS infrastructure monitors power, fans, and thermal conditions in the system for problem conditions. These conditions are reported either through an interrupt mechanism (for critical faults requiring immediate AIX action) or through messages passed from the RAS infrastructure to the service processor to RTAS. For more detailed information on the RAS features of both the pSeries 670 and pSeries 690, see Chapter 5, “Reliability, availability, and serviceability” on page 157.

Remote I/O loop

The communication from the CEC to the I/O drawers is done over the remote I/O link. This link uses a loop interconnect technology to provide redundant paths to I/O drawers. RIO availability features include CRC checking on the RIO bus with packet retry on bus timeouts. In addition, if a RIO link fails, the hardware is designed to automatically initiate an RIO bus reassignment to route the data through the alternate path to its intended destination.

The components involved in the creation of a loop are:

- RIO books, located in the CEC: The primary and secondary books have, respectively, 2 or 4 pairs of RIO ports.
- RIO planars located in the I/O drawer: All drawers have 2 pairs of ports: one on each planar.
- RIO cables: Cables are different to plug into an RIO port or an RIO-2 port.

There are two types of loops, RIO and RIO-2, depending on the technology used in these components.

- RIO Loop connections operate at 500 MHz, and connect to the CEC using the RIO books (FC6404 or 6410).
- RIO-2 Loop connections operate at 1 GHz, and connect to the CEC using the RIO-2 books (FC6418 or 6419).

There are two modes of loop operation:

- Single-loop mode: The two I/O planars of an I/O drawer belongs to the same loop, which is connected to one pair of ports of an I/O book
- Dual-loop mode: Each I/O planar of an I/O drawer belongs to one loop, which connects to one pair of ports of an I/O book. The two loops connected to one I/O drawer must connect to pairs of ports in the same IO book.

RIO books only support single-mode loops, while RIO-2 books support both single-mode and dual mode loops.

Figure 2-21 presents the cabling pattern that is supported for configuration of loops between RIO books and RIO planers.

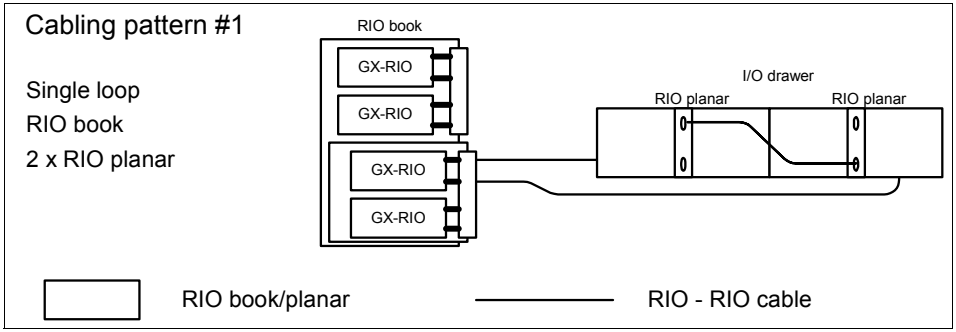


Figure 2-21 RIO loops supported configurations

Figure 2-22 on page 59 presents the four cabling patterns that are supported for configuration of loops between RIO-2 books and RIO or RIO-2 planers.

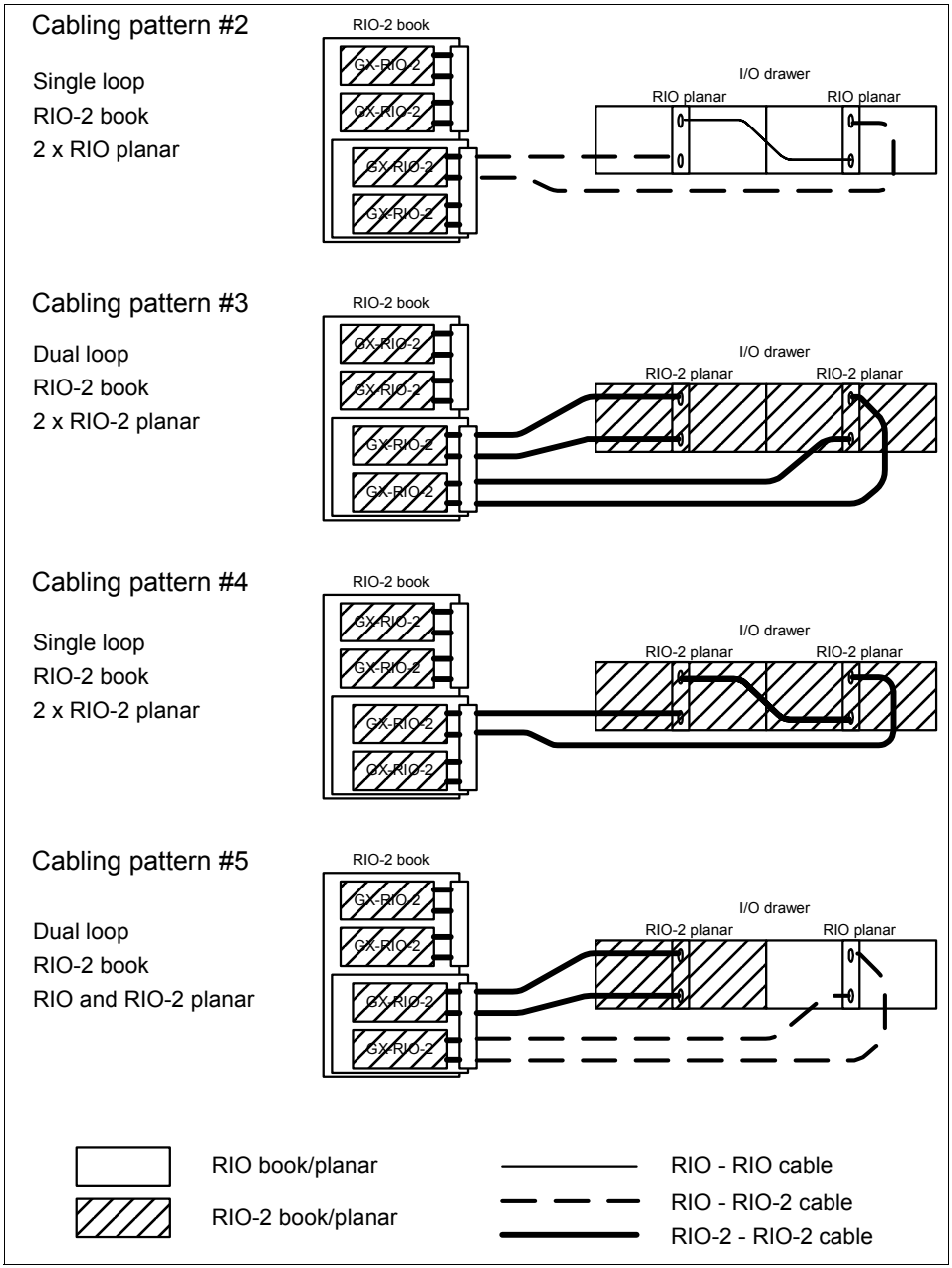


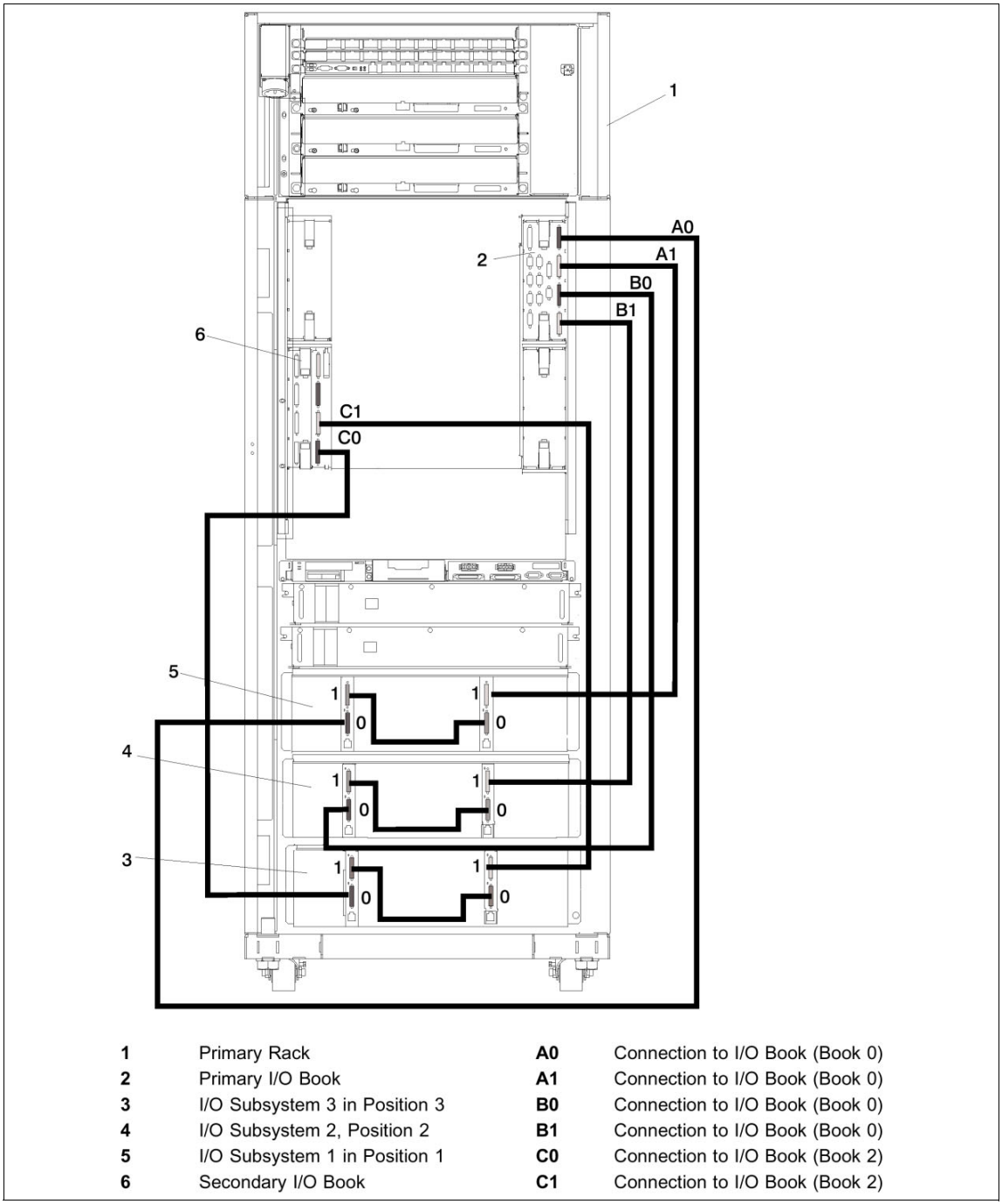
Figure 2-22 RIO-2 loops supported configurations

Table 2-8 lists the maximum number of RIO drawers and the maximum data rate that can be reached for each of the cabling pattern presented in Figures 2-21 and 2-22.

Table 2-8 Maximum configuration for each cabling pattern

Cabling Pattern	Max # of RIO Drawers for pSeries 690 Configuration	Max # of RIO Drawers for pSeries 670 Configuration	Maximum I/O burst rate per RIO Drawer	Maximum I/O sustained rate per RIO Drawer
#1	8	3	2 GB/s	800 MB/s
#2	8	3	2 GB/s	800 MB/s
#3	7	2[1]	8 GB/s	2.6 GB/s
#4	8	3	4 GB/s	1.6 GB/s
#5	8	2[1]	3 GB/s	2 GB/s
1. Only two pairs out of four can be used on the Secondary I/O book of a pSeries 670.				

An example of RIO cabling for the first three I/O drawers with IBF configuration is shown in Figure 2-23 on page 61. For a complete set of RIO cabling diagrams, refer to Appendix G, “Subsystem Positioning and Cabling” of *IBM @server pSeries 690 User’s Guide*, SA38-0588 and *IBM @server pSeries 670 Installation Guide*, SA38-0613.



1	Primary Rack	**A0**	Connection to I/O Book (Book 0)
2	Primary I/O Book	**A1**	Connection to I/O Book (Book 0)
3	I/O Subsystem 3 in Position 3	**B0**	Connection to I/O Book (Book 0)
4	I/O Subsystem 2, Position 2	**B1**	Connection to I/O Book (Book 0)
5	I/O Subsystem 1 in Position 1	**C0**	Connection to I/O Book (Book 2)
6	Secondary I/O Book	**C1**	Connection to I/O Book (Book 2)

Figure 2-23 RIO connections for three I/O drawers with IBF configuration

The pairs of RIO book ports in the system are numbered from 0 to 13 as shown in Figure 2-12 on page 41 (called GX-RIO in figure). In each book, the RIO port pairs are named A0&A1, B0&B1, C0&C1, D0&D1, as shown in Figure 2-15 on page 48. The relation between the pair number and the physical port pair is given in Table 2-9.

Table 2-9 Physical I/O book ports and system logical port numbers

I/O Book	Pair Name	Pair Number	I/O Book	Pair Name	Pair Number
Primary - 0	A0&A1	0	Secondary - 2	A0&A1	10
	B0&B1	1		B0&B1	11
	N/A			C0&C1	4
	N/A			D0&D1	5
Secondary - 1	A0&A1	2	Secondary - 3	A0&A1	6
	B0&B1	3		B0&B1	7
	C0&C1	8		C0&C1	12
	D0&D1	9		D0&D1	13

Of course, if I/O books are not present, the corresponding pair numbers do not appear in the system. For example, a system with only the Primary I/O book and the Secondary I/O book 2 contains only the pairs numbered 0, 1, 4, 5, 10, 11.

The cabling rule for the I/O book port pairs is to connect the I/O loop pairs in the following order: 0, 1, 4, 5, 10, 11, 12, 13, 6, 7, 2, 3, 8, and 9, without skipping any installed pair.

There is no mandatory rule for the order to connect the I/O drawers. However, pairs 0, 1, 4, 5, 10, 11, 12, 13 support both single-loop and dual-loop modes, while pairs 6, 7, 2, 3, 8, and 9 support only dual-loop mode. Therefore, we recommend you first connect to the CEC the drawers configured in single-loop mode. For initial order, we also recommend that if the configuration contains single-loops, you assign them to the first I/O drawers, to connect the first I/O drawers to the first pairs of I/O book ports. Another consequence of the pairs 2, 3, 8, and 9 supporting only dual-loop mode is that the last GX bus position (secondary book 1) is never used in a configuration containing only RIO drawers.

Table 2-10 presents for each combination of installed MCM and I/O books, the lists of I/O ports that are usable on the CEC. Please note that this table is cumulative top to bottom, and left to right: a system in row marked “MCM1” also contains the MCM0 and MCM 2, and a system in column marked “Secondary-3” also contains the primary book and the secondary-2 book. The entry for each

combination contains two lines: the first line indicates the usable ports supporting both single-loop and dual-loop modes, while the second line lists the port that only supports dual-loop mode. Ports in the second lines cannot be used to connect RIO drawers, but only RIO-2 drawers.

For example, with two MCMs and three I/O books, ports 0, 1, 4, and 5 are usable to connect single-loops, and ports 6 and 7 are usable to connect dual-loops.

Table 2-10 Available I/O ports versus installed MCM and Books

MCM position	Primary - 0	Secondary- 2	Secondary - 3	Secondary 1
MCM 0	0, 1	N/A	N/A	N/A
	N/A	N/A	N/A	N/A
MCM 2	0, 1	0, 1, 4, 5	0, 1, 4, 5	N/A
	N/A	N/A	6, 7	N/A
MCM 1	0, 1	0, 1, 4, 5, 10, 11	0, 1, 4, 5, 10, 11	0, 1, 4, 5, 10, 11
	N/A	N/A	6, 7	6, 7, 2, 3, 8, 9
MCM 3	0, 1	0, 1, 4, 5, 10, 11	0, 1, 4, 5, 10, 11, 12, 13	0, 1, 4, 5, 10, 11, 12, 13
	N/A	N/A	6, 7	6, 7, 2, 3, 8, 9

Figure 2-24 graphically presents the same information. It depicts four simplified rear views of the CEC, when 1, 2, 3, or 4 MCMs are installed, and shows in each case which pairs of ports can be used.

When installing additional MCM through MES to an existing system, the set of available port pairs will be modified. For example, a 2-MCM and 3-I/O Book systems can use pairs 1, 2, 4, 5, 6, and 7. After installation of a third MCM, it can use pairs 0, 1, 4, 5, 10, 11, 6 and 7. If all pairs were used in the initial system, there is no need to re-cable the I/O loops after installation of the third MCM, even though the newly available pairs 10 an 11 have a higher cabling priority than pairs 6 and 7.

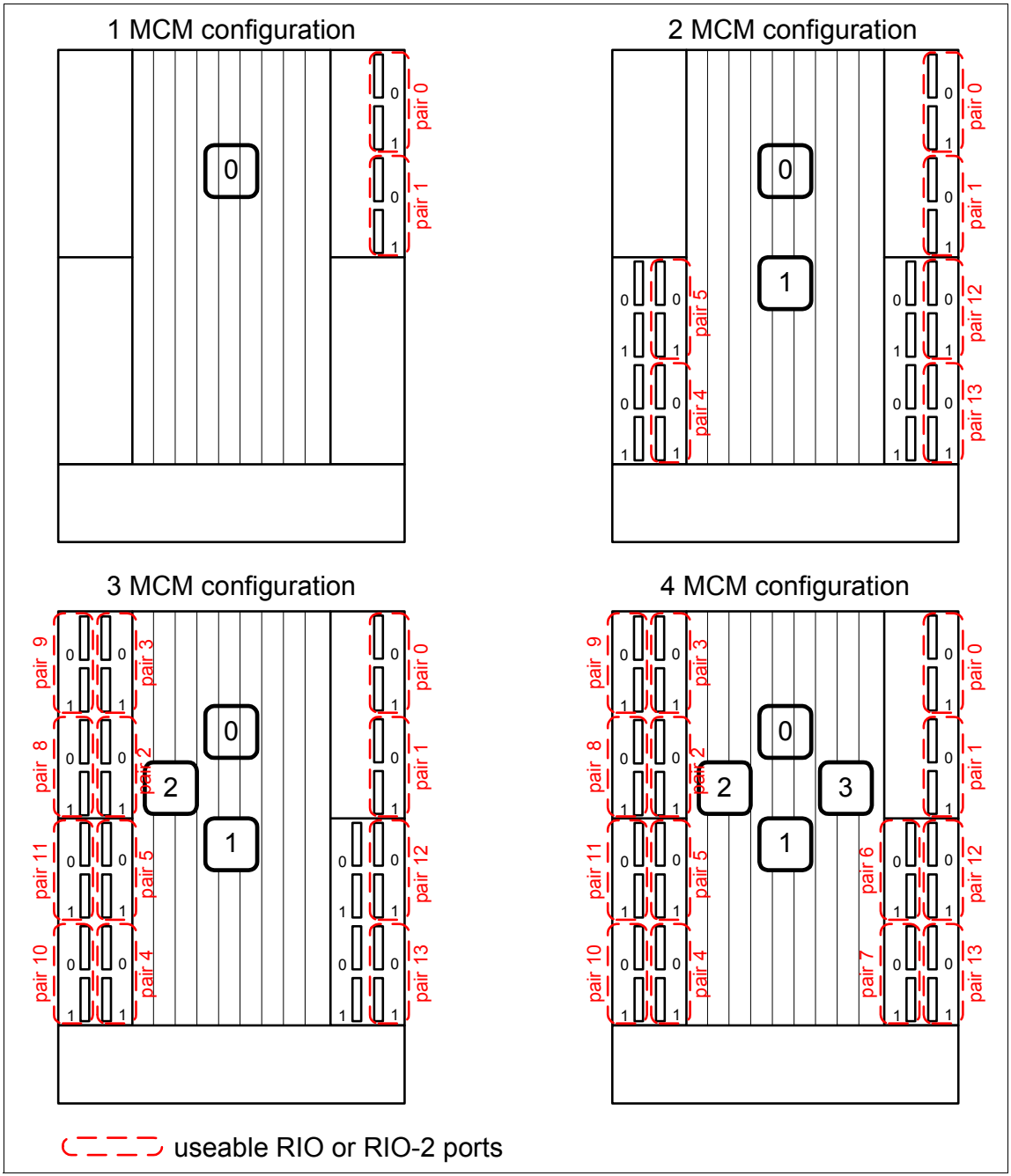


Figure 2-24 Number of usable RIO or RIO-2 ports

I/O cabling conclusion

An important change that is brought along with the new features announced for pSeries 670 and pSeries 690 in May 2003 relates to the I/O loop cabling rules. Before this date, the I/O loop had to be cabled according to the rules specified in the *IBM @server pSeries 670 Installation Guide*, SA38-0613 and the *IBM @server pSeries 690 Installation Guide*, SA38-0587. Now, these rules are only guidelines, and the I/O loops only need to be cabled in the next available I/O port pair on the I/O books.

However, to simplify the management of the server we strongly recommend that I/O loops be configured as described in the pSeries installation guides, and to only follow a different order when absolutely necessary, for example, when required by an MES installation or when using a mix of RIO and RIO-2 drawers.

In any case, it becomes extremely important for the management of the system to keep an up-to-date cabling documentation of your systems, because it may be different from the cabling diagrams of the pSeries installation guides.

Examples of RIO loop cabling and associated performance figures are given in Appendix B, “I/O loop cabling and performance” on page 217.

2.4.3 I/O drawer physical placement order

The I/O drawer physical placement order in the pSeries 690 depends on whether the system is equipped with the Internal Battery Feature (IBF).

The first three I/O drawers have a fixed placement order, regardless of the system configuration in the base rack. They occupy the drawer positions 1, 2, and 3, as shown in Figure 2-25.

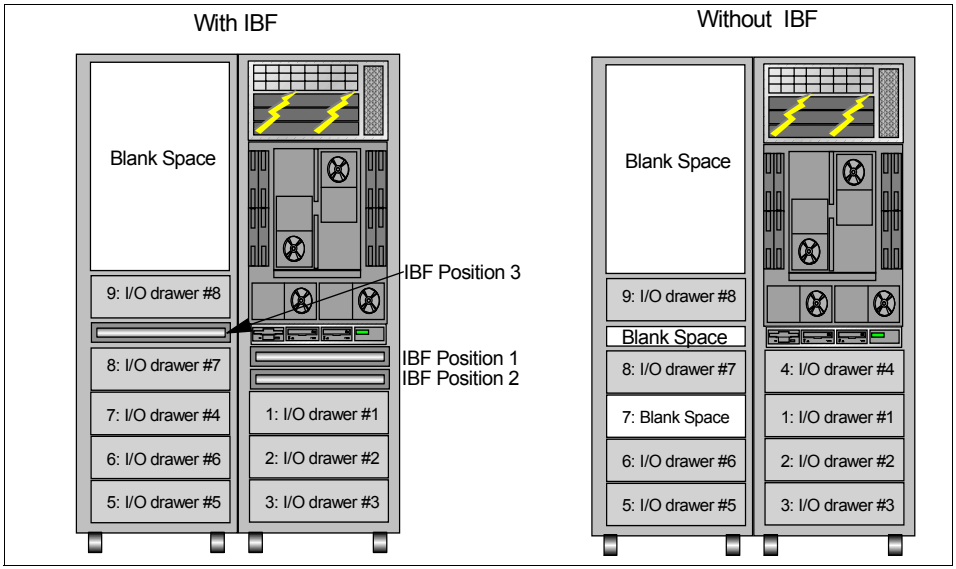


Figure 2-25 I/O drawer and IBF placement positions (pSeries 690)

With IBF configuration, the fourth I/O drawer occupies drawer position 7 in the expansion rack, as shown in the left side of Figure 2-25. Additional I/O drawers are placed on the expansion rack in sequence from the drawer position 5 to 9, but the drawer position 7 is skipped.

If the IBF is ordered, it is placed on the base rack between the media drawer and the first I/O drawer (shown as IBF position 1 and IBF position 2; see Figure 2-25). Up to four batteries can be placed on the base rack. You can also install two additional batteries on the expansion rack between drawer positions 8 and 9 (shown as IBF position 3; see Figure 2-25).

Without IBF configuration, the fourth I/O drawer can be placed on the base rack. It is placed between the media drawer and the first I/O drawer, as shown in the right side of Figure 2-25. In this case, drawer position 7 in the expansion rack is never used, as shown in the right side of Figure 2-25.

As already mentioned, the pSeries 670 only offers a maximum of three I/O drawers and the IBF feature. Therefore only one frame is needed. Figure 2-26 shows the maximum I/O drawer configuration of pSeries 670.

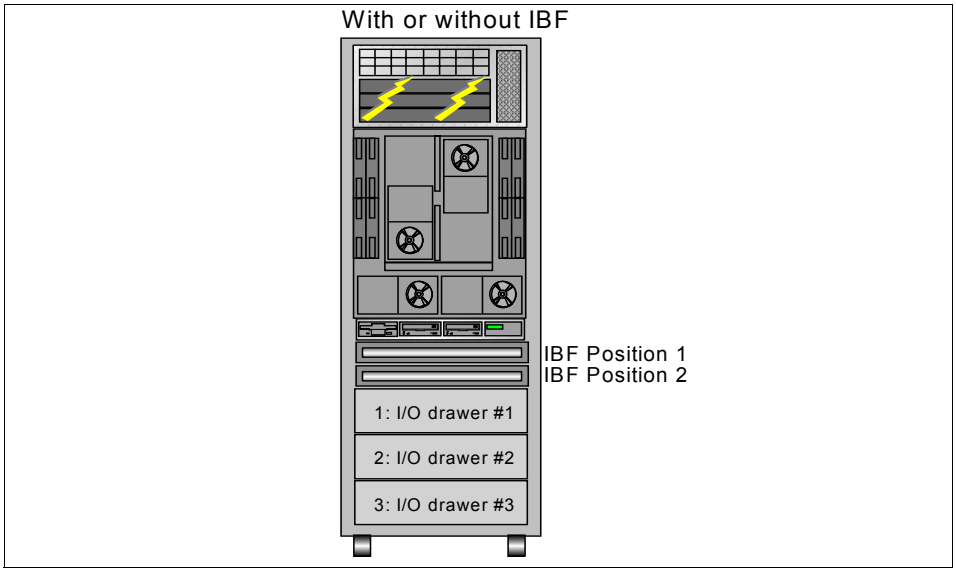


Figure 2-26 I/O drawer and IBF placement positions (pSeries 670)

The service processor is aware of the presence of the IBF, using the same mechanism explained in “Power and RAS infrastructure” on page 56.

Note: Do not misunderstand the difference between a drawer position number and an I/O drawer number. The drawer position number represents a physically fixed position in the base or expansion racks. An I/O drawer number represents the order of I/O drawers configured on the system.

All components of a pSeries 670 or a pSeries 690 can be identified by their physical location code. The format of this code is:

```
U<rack_number>.<rack_position>[<optional location info>]
```

Where:

<rack_number> is 1 for the first rack, 2 for the second rack.

<rack_position> is the lowest position occupied by the device in the rack, from 1 at the bottom of the rack to 42 at the top of the rack.

<optional location info> indicates a position within the device.

Figure 2-27 and Figure 2-28 present all physical location codes existing in pSeries 670 or a pSeries 690 frames. For detailed information, refer to Chapter 1 of *IBM @server pSeries 670 Service Guide*, SA38-0615 and *IBM @server pSeries 690 Service Guide*, SA38-0589.

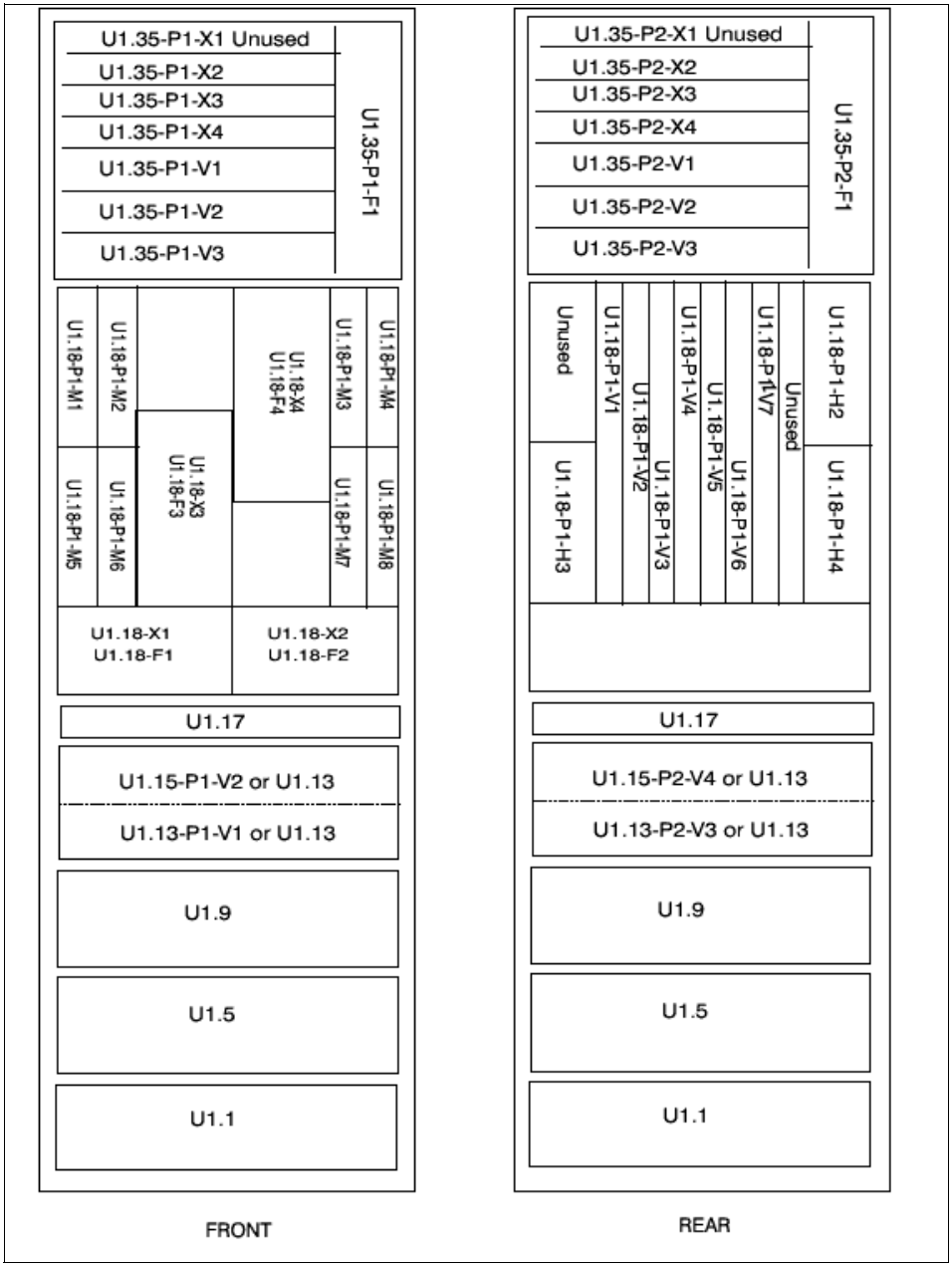


Figure 2-27 Physical location in frame 1 of pSeries 670 and pSeries 690

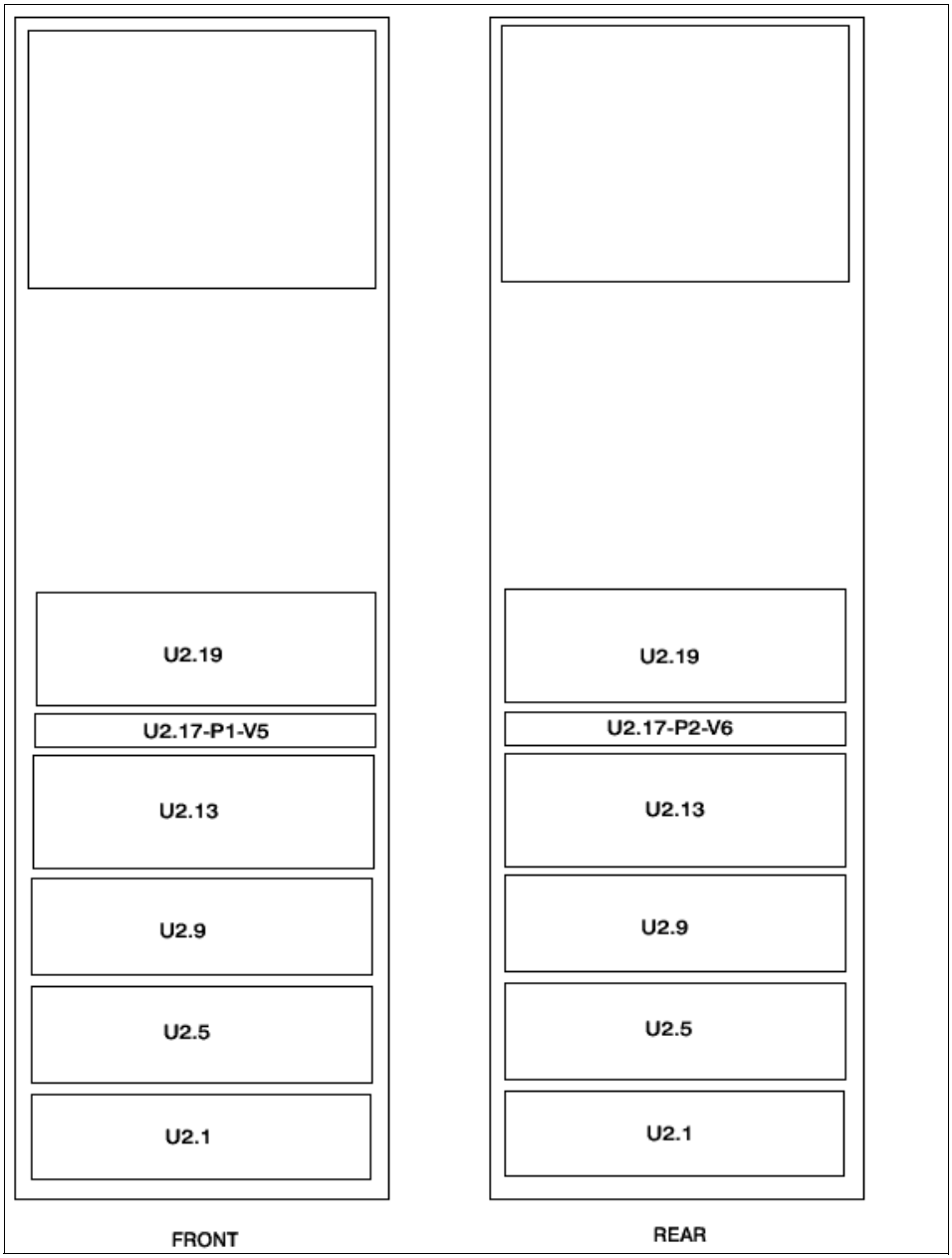


Figure 2-28 Physical location in frame 2 of pSeries 690

Table 2-11 shows the relationship between I/O drawer positions and physical location codes.

Table 2-11 Physical location code of drawer position number

Drawer position number	Physical location code	pSeries 670	pSeries 690
1	U1.9	Yes	Yes
2	U1.5	Yes	Yes
3	U1.1	Yes	Yes
4	U1.13	No	Yes
5	U2.1	No	Yes
6	U2.5	No	Yes
7	U2.9	No	Yes
8	U2.13	No	Yes
9	U2.19	No	Yes

2.4.4 Media drawer

The media drawer provided in the base configuration contains an operator panel, a 1.44 MB diskette drive and two sets of two removable media bays. The operator panel and the diskette drive are each connected to the service processor book in the CEC using point-to-point cables, and are powered from the CEC standby power. Each set of media devices is powered and logically driven by an I/O planar in the first I/O drawer, as shown in Figure 2-29.

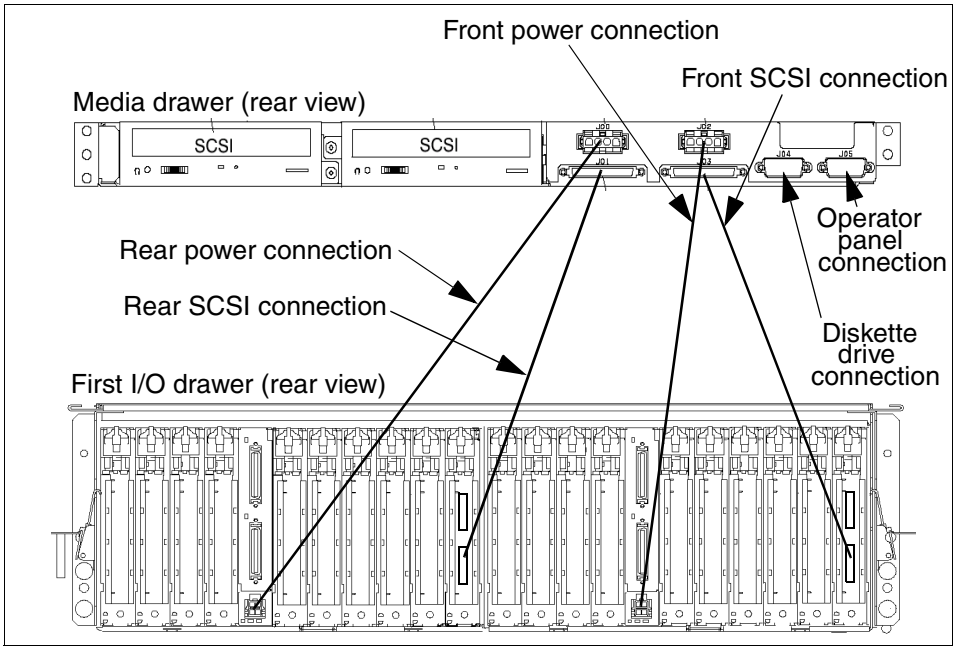


Figure 2-29 Media drawer power and SCSI connection

The media drawer is split into two separate sections, as shown in Figure 2-30. The front section houses the system operator panel, a diskette drive, and two media bays. The rear section provides space for two additional media devices. The front and rear sections of the media drawer are in two separate SCSI buses, and must be connected to separate SCSI PCI adapters on the I/O drawers. Therefore, these devices configured in the front and rear bays of the media drawer can either be in the same partition or in different partitions.

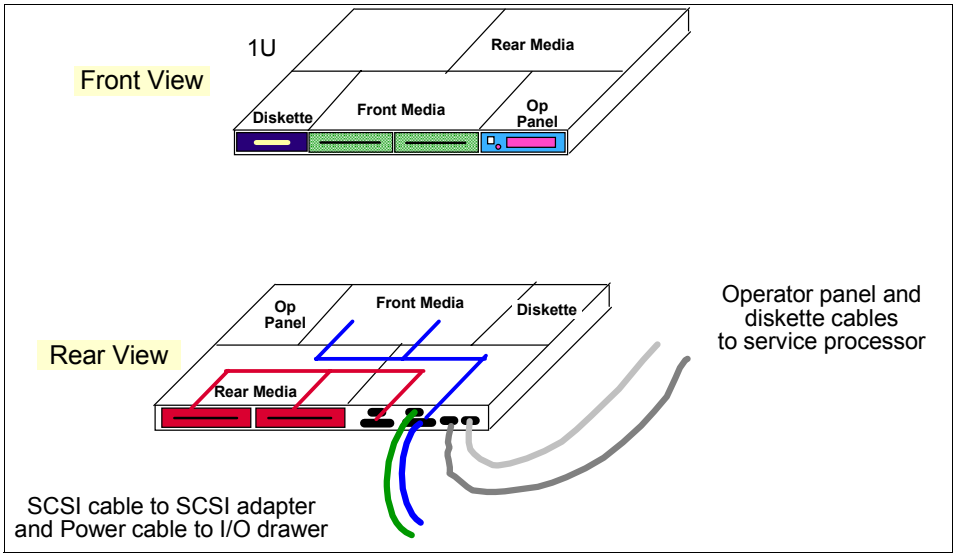


Figure 2-30 Media drawer

In order to use an IDE DVD-ROM drive (FC 2634) in the media drawer, the IDE media to LVD SCSI interface bridge card (FC 4253) must be attached to each IDE DVD-ROM drive. The IDE DVD-ROM drive is treated as a SCSI CD-ROM drive on AIX.

Note: The IDE DVD-ROM drive (FC 2634) is supported on AIX 5L Version 5.2 and later.

2.5 Power subsystem

The power subsystem on the pSeries 670 and pSeries 690 are identical and provide full redundancy. The power subsystem consists of redundant bulk power assemblies, bulk power regulators, power controllers, power distribution assemblies, DC power converters, and associated cabling.

2.5.1 Bulk power assembly

The bulk power assembly (BPA) is the main power distribution unit for the pSeries 670 and pSeries 690. The redundant bulk power assembly converts AC input to DC and distributes power at 350 V to each drawer where conversion is made to the required chip level. It is composed of several power converters and regulators. Two BPAs are mounted in front and rear positions and occupy the top of the rack as an eight EIA height drawer unit (see Figure 2-31). They are powered from separate power sources via separate line cords.

Each BPA consist of a stack of components (from bottom to top):

- Three Bulk Power Regulator (BPR)
- One Bulk Power Controller (BPC)
- Three Bulk Power Distributor (BPD). The first BPD must be installed on all systems. The second and third are optional, depending on the number of I/O drawers installed in the system.

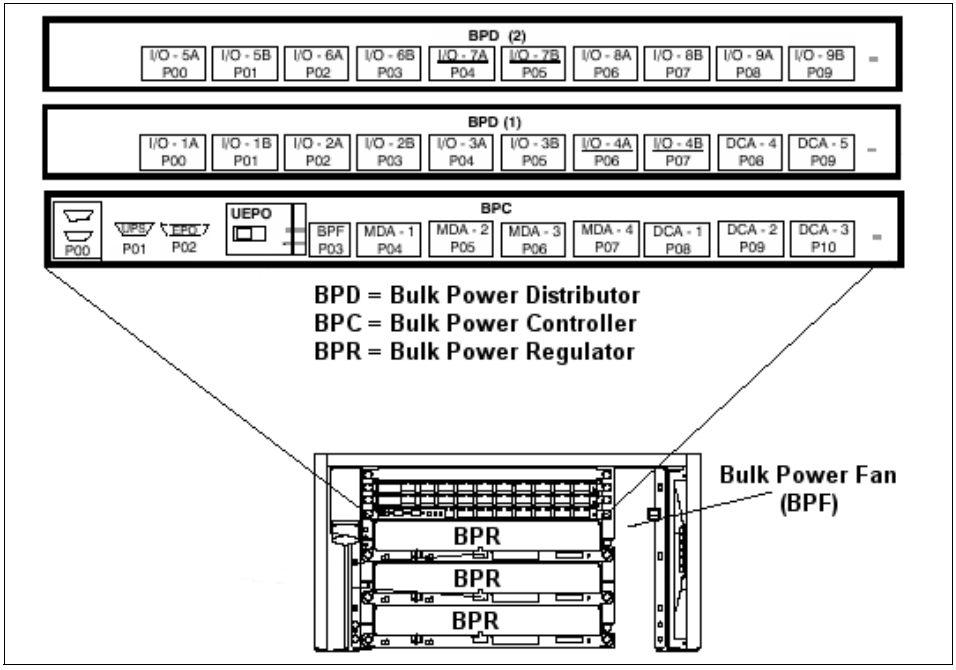


Figure 2-31 Power subsystem locations in BPA

The BPA distributes power to the I/O drawers and the DCA in the CEC via appropriate power cables, plugged into the BPC and BPD connectors (The

number of the DCA in the CEC depends on the number of installed MCMs and I/O books).

Constant power monitoring hardware assists in the detection of failures in the power subsystem that require maintenance. The power supplies, power converters, and cables can be serviced without system disruption. See Figure 2-31.

The BPA also provides a major function in the RAS infrastructure, as explained in “Power and RAS infrastructure” on page 56.

2.5.2 Internal battery feature

The internal battery feature (IBF) provides emergency power in case of a power outage. The pSeries 670 and pSeries 690 incorporate an Early Power Off Warning (EPOW) capability that assists in performing an orderly system shutdown in the event of a sudden power loss. The IBF protects against power line disturbances and provides sufficient power to enable the system to shut down gracefully in the event that the power sources fail.

Up to two IBF enclosures can be installed in the base rack configuration. One additional IBF enclosure can be installed in an expansion rack (see Figure 2-25 on page 66).

The hold-up time of the IBF will vary considerably with conditions present at the time of the power loss. Ambient temperature, age, and past use of the battery can affect the discharge time.

For more detailed information about the IBF hold-up times for the pSeries 670 and pSeries 690, see the 7040 sections in Chapter 1, “Physical Characteristics of Systems” of the publication *Site and Hardware Planning Information*, SA38-0508.

Attention: The document *Site and Hardware Planning Information*, SA38-0508 is updated very often. The version of this document which includes information related to the related to the pSeries 670 and pSeries 690 models announced in May 2003 is version 18. If you have an old version of this document, please download the latest version from the IBM Web page:

http://publib16.boulder.ibm.com/pseries/en_US/infocenter/base/HW_site_hardware.htm

2.5.3 Cooling

Several fans and blowers provide air cooling for the pSeries 670 and pSeries 690. The power supplies, the CEC, and each I/O drawer have a group of redundant cooling devices. They operate at normal speed when all the cooling devices are operational, and increase their speed if one device stops working in order to maintain the air flow constant and provide proper cooling.

The CEC is cooled by the following devices:

- Four high-pressure, high-flow blowers in the base of the CPU cage that pull air down the front of the cage through the memory cards and exhaust air up the rear of the cage through the power converters and RIO books.
- Two additional identical blowers in the front of the cage blow air directly down on the MCMs and L3 modules, which is then exhausted out the top rear of the cage.

Each fan or blower is referred to as an Air Moving Device, which consists of a Motor/Scroll Assembly (MSA) and a Motor Drive Assembly (MDA).

- The MSA consists only of a motor, wheel, and metal housing.
- The MDA has two 350 V and communication inputs (one from each BPA) and a three-phase output designed to drive a brush-less DC motor using back EMF commutation control. The power and RAS infrastructure is used to control and activate the Air Moving Devices through the communication ports (see 2.4.2, “I/O subsystem communication and monitoring” on page 56).

A micro controller within the MDA monitors and controls the motor drive and provides precise speed control capability. Each Air Moving Device can draw 15-300 watts of power, depending on speed setting and back pressure, with the maximum power for a set of two blowers being 500 watts. The Air Moving Device design is identical to that of the Air Moving Device used in the zSeries servers. All Air Moving Devices from the CEC can be removed and replaced without powering down the system.

Please note that each I/O drawer has the cooling devices that have the same functionality within the CEC, and also offers N+1 redundancy.

2.6 IBM Hardware Management Console for pSeries

The IBM Hardware Management Console for pSeries (HMC) provides a standard user interface for configuring and operating partitioning-capable pSeries servers. The HMC provides a graphical user interface for configuring and operating single or multiple partitioning-capable pSeries servers (also called managed systems).

Figure 2-32 shows an example of the graphical user interface on the HMC.

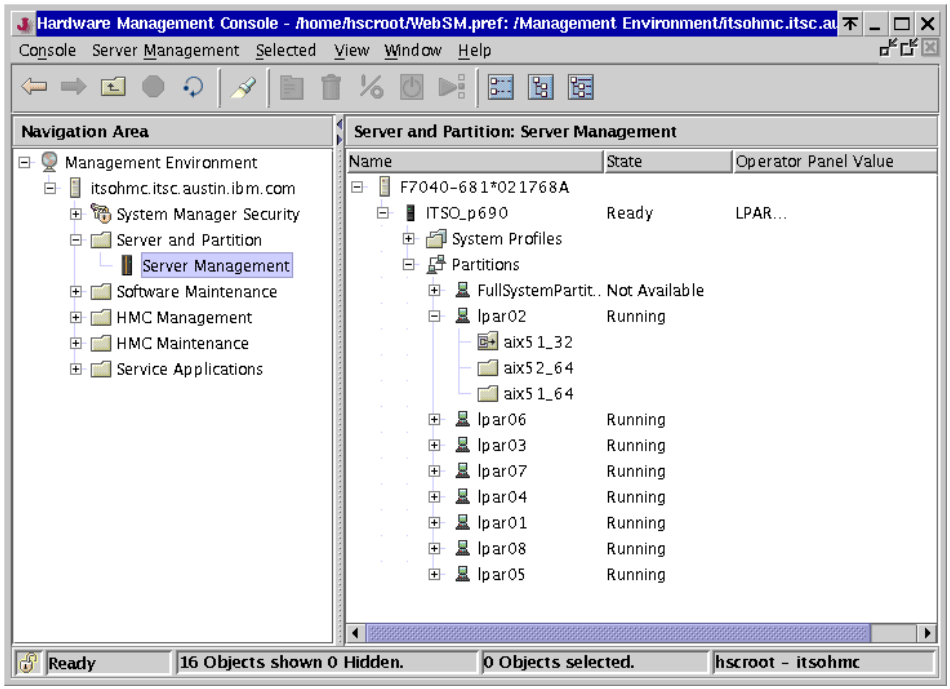


Figure 2-32 Graphical user interface on the HMC

The HMC consists of a 32-bit Intel-based desktop PC with a DVD-RAM drive running the Linux operating system. The application environment, with a set of hardware management applications for configuration and partitioning, is written in Java. The applications are based on the object-oriented schema using the Common Information Model (CIM), an industry standard sponsored by the Distributed Management Task Force (DMTF). A CIM object manager acts as a repository and database look-up for all managed objects.

The pSeries 670 and pSeries 690 provide two special asynchronous ports dedicated for HMC attachment, HMC1 and HMC2, shown in Figure 2-15 on page 48. The connectors on these ports are differentiated from standard asynchronous ports by providing a 9-pin female connector versus the standard 9-pin male serial connector. The HMC and managed systems are connected by a serial cable available in 6 meter and 15 meter lengths[7].

The HMC allows a system administrator to do the following management tasks on managed systems:

[7] The 128-port asynchronous adapter (FC 2944) provides a distance solution, as explained in 3.2.11, “HMC configuration rules” on page 93.

- Creating and maintaining a multiple-partitioned environment
- Displaying a Virtual Terminal window for each operating system session
- Displaying operator panel values for each partition
- Detecting, reporting, and storing changes in the hardware conditions
- Controlling the powering on and off of the managed system
- Acting as the Service Focal Point for service representatives to determine an appropriate service strategy and enable the Service Agent *Call-Home* capability
- Activating additional processor resources on demand (Capacity Upgrade on Demand)

The graphical user interface on the HMC is based on the AIX 5L Version 5.2 Web-based System Manager, which allows the management integration of other HMCs or pSeries systems running AIX 5L Version 5.1 and 5.2. Except for IBM customer support representatives, the native Linux interfaces are hidden from the user and are not accessible. No Linux skills are required to operate the HMC.

For further information about how to use and manage HMC, refer to the following publications:

- *IBM Hardware Management Console for pSeries Installation and Operations Guide*, SA38-0590
- *IBM Hardware Management Console for pSeries Maintenance Guide*, SA38-0603

Note: The HMC is a mandatory feature of the pSeries 670 and pSeries 690.

3

Using the IBM Configurator for e-business

This chapter is intended to help IBM employees and IBM business partners to prepare a pSeries 670 and pSeries 690 configuration for price quotation using the IBM Configurator for e-business (also known as e-config).

- Section 3.1, “What’s new with e-config” summarizes the changes in the use of the configurator due to the May 2003 announcements.
- Section 3.2, “Configuration rules for pSeries 670 and pSeries 690” details information you should know before you start using the configurator.
- Section 3.3, “IBM Configurator for e-business (e-config)” presents the steps performed in the configurator to create an initial order.
- Section 3.4, “Configuration examples” contains examples of initial orders, feature conversions, and model conversions.

We assume that you are familiar with installation and basic use of e-config. More information on the IBM Configurator for e-business is available at:

http://ftp.ibmlink.ibm.com/econfig/announce/index.htm

3.1 What's new with e-config

To support the announcements of the new pSeries 670 and pSeries 690 features in May 2003, a new Version 4.1.1 of e-config has been released.

Note: At the time of writing of this book, an internal test of Version 4.1.1 of the IBM Configurator for e-business was used as shown in Figure 3-1. Changes or enhancements to e-config may occur so that information in this chapter might not be completely accurate over time. The version of e-config is the number displayed in the Base field of the ECheck tool provided with e-config (see Figure 3-1). It is not the version displayed if you click **Help** -> **About SCPortofolio** in the configurator.

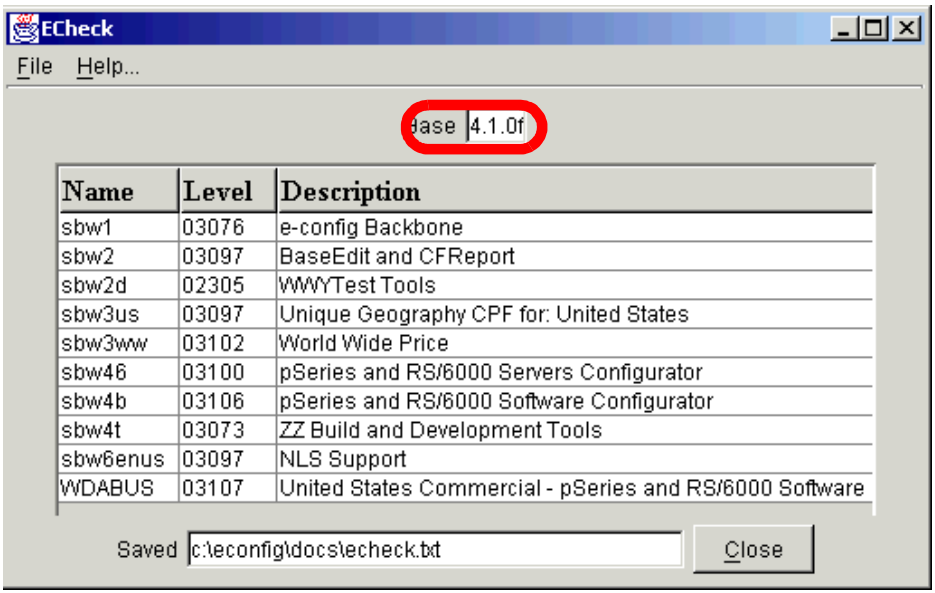


Figure 3-1 IBM Configurator for e-business version

There are a few changes in the way you will use this new version to create new configurations or to prepare MES to an existing system, and especially:

- The I/O books on the CEC and the I/O planars on the I/O drawers are available in two technologies (RIO and RIO-2). As a result, you must now specify which kind of these devices you want, while in the previous version, e-config would automatically configure the I/O books and planars for you.
- Because the I/O drawers can be connected using single-loop or dual-loop mode, there is no longer a unique relationship between the number of

installed I/O drawers and the number of required I/O books. You need to specify your requirements and cannot only rely on default values provided by the configurator.

Furthermore, e-config is not able to verify the consistency between requirements and configured features, like the number of LPARs and number of disks or Ethernet adapters. For example, you may want to use external disks in a 7133 SSA drawer as boot devices so a config with four internal disks and eight LPARs is perfectly valid.

Be sure you understand the architectural constraints presented in Chapter 2, “Hardware architecture of the pSeries 670 and pSeries 690” on page 17 and the configuration rules presented in 3.2, “Configuration rules for pSeries 670 and pSeries 690” before you start using e-config.

3.2 Configuration rules for pSeries 670 and pSeries 690

In order to properly configure a pSeries 670 or pSeries 690, some requirements and constraints must be satisfied. For example, there are dependencies between some hardware components: installation of a 1.7 GHz MCM implies that the L3 cache feature FC 4199 is also installed. This section explains the requirements and options needed to configure the minimum configuration, as well as additional components configuration.

Besides the correctness of the configuration, it is important to configure the system that best meets the needs of the application. Requirements like availability, performance, and flexibility may require additional components or devices. For more information on how to best configure a pSeries 690, see the *IBM @server pSeries 690 Availability Best Practices* white paper and the *IBM @server pSeries 690 Configuring for Performance* white paper, found at the following URLs:

http://www.ibm.com/servers/eserver/pseries/hardware/whitepapers/p690_avail.html
http://www.ibm.com/servers/eserver/pseries/hardware/whitepapers/p690_config.html

The complete and up to date list of configuration guidelines, dependencies and requirements can be found in the “Limitations” section of the pSeries 670 and pSeries 690 sales manual, and in the guide *PCI Adapter Placement References*, SA38-0538.

Some of the requirements do not leave a choice to the user of the configurator, for example, all systems need one and only one clock (FC 5251). In this case, these mandatory features are automatically added by the configurator, and we will not detail the related configuration rules in the following sections.

A pSeries 670 or pSeries 690 server consists of several components which all have a machine type. When using e-config, each of these components is customized using a different wizard.

- One or two system rack: The 7040-61R wizard configures the options of the rack itself (doors, …) but also of the Bulk Power Assembly and of the Integrated Battery Feature.
- One Central Electronics Complex: The 7040-671 or 7040-681 wizard configures the CEC options as well as the options in the Media Drawer.
- At least one and up to eight I/O drawer: The 7040-61D wizard configures the adapters and internal disks, as well as the displays, mouse and keyboard that can be attached to the LPARs.
- One or two HMC: The 7315-C01 or 7315-C02 wizard now configures the HMC as a separate device.

Note: In the first versions of the pSeries 690, the HMC was configured as a feature of the CEC (FC 7315 and FC7316). These features are no longer available. On new systems, the HMC must be ordered as machine type 7315-C01 or 7315-C02. However, for compatibility issues with older configuration, the CEC wizard still contains options for the HMC: keyboard, mouse, power cables, … You should no longer use these options of the CEC wizard. All options of the currently available HMC must be configured through the HMC wizard.

The automatically configured features are not visible in the wizards. You will only find them in the reports generated by e-config.

There are many dependencies between the different components of pSeries 670 and pSeries 690. In many cases, the number of LPARs and the set of adapters to include in each LPAR will define the number of IO drawers to configure. The type of these adapters (especially if they are high-speed adapters), and the constraints on the supported number of concurrent adapters is likely to be the first factor to decide on the number of I/O required drawers. This will also influence the model of the I/O drawers to use and the type of I/O loops to configure. The number of I/O drawers will decide on the number of racks to install. The type of drawers and loops will decide on the type and number of I/O books to install in the CEC. We therefore recommend that you start by configuring all I/O drawers options, then the racks option, and finish by the CEC wizard. The HMC can be configured independently from the other components. This will minimize the number of times you need to revisit your configurations.

3.2.1 Minimum configuration for the pSeries 670 and pSeries 690

The sales manual for the pSeries 670 and pSeries 690 describes a minimum configuration including the following items:

- One 7040-671 Central Electronics Complex with 1 MCM and 4 GB memory for the pSeries 670. The minimum processor configuration is one 4-way 1.1 GHz MCM.
- One 7040-681 Central Electronics Complex with 1 MCM and 8 GB memory for the pSeries 690. The minimum processor configuration is one 8-way 1.1 GHz MCM.
- One 7040-61D I/O drawer with at least one network adapter per partition. The minimum configuration for the pSeries 670 is an I/O drawer with only 1 planar (10 PCI slots). The minimum configuration for the pSeries 690 is an I/O drawer with both planars (20 PCI slots).
- One 7040-61R system rack with front and rear doors.
- A media drawer with one CD-ROM, DVD-RAM, or DVD-ROM drive.
- The bulk power assembly (BPA) and associated converters and cables.

The sales manual minimum configuration corresponds to the absolute lowest price of an orderable server.

The IBM Configurator for e-business has a concept of default configuration. This is the configuration you obtain when you create a server configuration without selecting any options.

Note: In general, the Sales manual minimum configuration and the e-config default configuration are different, even though they do not differ by many features.

There are several possible reasons for the differences between the sales manual minimum configuration and the e-config default configuration:

- The default configuration may recommend a feature configuration that gives better performance than the cheaper alternative indicated in the sales manual.
- The default configurations may contain feature options expected by most customers that are not part of the cheapest configurations.

In the case of the pSeries 670 and pSeries 690:

- The default configuration is not a valid orderable configuration. It does not contain the mandatory rack rear door, and you must specify if you need a slim

or an acoustic rear door. With the addition of this single option, the configuration is valid.

- The default configuration contains language and power cable options compatible with your e-config country (or region) setting, while the sales manuals only contains generic options.
- The pSeries 670 minimum configuration is a server with one half RIO drawer (only one RIO planar), while the default configuration contains one full RIO drawer.

Appendix A, “Minimum and default configurations” on page 209 contains the sales manual minimum configuration and the e-config default configuration of pSeries 670 and pSeries 690.

The minimum pSeries 670 and pSeries 690 configurations require AIX 5L Version 5.1 (5100-01 Recommended Maintenance Level with APAR IY39794) or higher.

For configurations other than the minimal, the set of rules described in the following section applies. For an overview of all available feature codes of the pSeries 670 and pSeries 690, refer to the IBM Sales Manual.

3.2.2 LPAR considerations

Before defining the set of hardware devices that you want to be present in your pSeries 670 or pSeries 690 system, you must pay attention to the constraints implied by the use of logical partitions. If you want to use several LPARs, then you must remember that:

- For up to 16 LPARs, you can use either RIO or RIO-2 Books in the CEC (FC 6404, 6410, 6418 or 6419). For instantiating from 17 to 32 LPARs, you must only use RIO-2 Books (FC 6418 and 6419).
- Each LPAR must contain at least one processor, one bootable device, one network adapter, and a minimum of memory:
  - 1 GB per LPAR with a system firmware and HMC software release at a level earlier than the 10/2002 level
  - 256 MB per LPAR with a system firmware and HMC software release at 10/2002 level or later
- Each PCI adapter and each built-in SCSI adapter of the RIO drawers belongs to at most one LPAR. There is no sharing of these resources between the LPARs. As a consequence:
  - If you use disks internal to the RIO drawers as boot devices, you need at least one RIO drawer for each set of four LPARs.

– The four disk of each four disk pack belongs to the same LPAR.

– All storage devices in the front of the media drawer belong to the same LPAR. All storage devices in the back of the media drawer belong to the same partition. At most, two partitions can use the internal CD-ROM, DVD-ROM, DVD-RAM, and tapes devices.

Note: Even though you enter in e-config the number of LPAR you want to instantiate on your server, the configurator does not check that the number of configured hardware features will be sufficient to support the requested LPARs. You have to it check by yourself.

A TCP/IP connection between the HMC and the LPARs is mandatory to use the DLPAR feature, and recommended in all cases. We strongly recommend that you provide this TCP/IP connectivity using a dedicated Ethernet network. Each LPAR should contain dedicated Ethernet adapters for this connection.

3.2.3 Processor configuration rules

There are some rules that must be followed when configuring processors in a pSeries 670 or pSeries 690 system, as part of the options of the 7040-671 or 7040-681 wizard. These include:

- All processors in a pSeries 670 or pSeries 690 must operate at the same speed.
- Each installed MCM is supported with 128 MB of Level 3 cache. There is a specific L3 cache FC for each MCM FC. e-config automatically inserts the right L3 cache option for you.
- The mandatory programmable processor clock card (FC 5251) is automatically inserted by e-config.
- If needed, the Processor Bus Pass Through Modules (FC 5257) are automatically configured by e-config. pSeries 690 servers configured with two or three MCMs must have the empty processor positions populated with Processor Bus Pass Through Modules (FC 5257). No Processor Bus Pass Through Module are required on pSeries 690 with one MCM or four MCMs, or on pSeries 670.
- MCMs require one or several capacitor books (FC 6198) to operate. The number of capacity books depends on the number of installed MCM, and these features are automatically configured by e-config.

3.2.4 Memory configuration rules

There are some rules when configuring memory cards in the pSeries 670 and pSeries 690 servers, using the CEC wizard. These include:

- The pSeries 690 has eight memory slots. Four memory slots utilize inward-facing memory cards and four utilize outward-facing memory cards. The inward-facing memory slots are utilized by the first and second MCM positions, while the outward-facing memory slots support the third and fourth MCM positions. The pSeries 670 only uses the four inward-facing memory slots, because it only uses the first and second MCM position.
- The combinations of memory boards supported on pSeries 670 and pSeries 690 are detailed in 2.3.2, “Memory subsystem for pSeries 690” on page 29, 2.3.4, “Memory subsystem for pSeries 670” on page 42, and 4.2.4, “Supported CUoD Memory configurations” on page 141.
- Memory boards are available in sizes: 4 GB, 8 GB, 16 GB, 32 GB, and 64 GB.
- The 4 GB and 8 GB at 575 MHz memory features are only available for the 1.7 GHz processor configurations.
- All the available memory slots should be populated with memory boards which size are as closely balanced as possible. Only one memory increment difference is allowed between two memory boards. Therefore, if you mix different sized memory cards, the following combinations are allowed: 4 GB and 8 GB, 8 GB and 16 GB, 16 GB and 32 GB, and 32 and 64 GB.
- Memory slots must be populated in pairs of equal size memory boards. The only exception is for entry server pSeries 670 or pSeries 690 system with only one MCM, which can be equipped with only one memory board in slot 0. The pSeries 670 can have either one 4 GB board or one 8 GB board in position 0. The pSeries 690 can have one 8 GB board in position 0.
- Servers using the 1.7 GHz processor should use 567 MHZ memory boards to obtain best performance. Using 500 MHz memory on these servers results in degraded memory subsystem performance.
- When the system is running in a partitioned environment, effective memory is reduced due to partition page tables and translation control entry (TCE) tables.

Table 3-1 on page 87 details the approximate reserved memory for page and TCE tables, effective usable partition memory, and the maximum number of partitions for various configurations. The configuration number in the header row of columns 5 to 8 refer to:

Conf.A Partitions with any version of AIX or Linux, with firmware and HMC release levels earlier than the 10/2002 release level.

Conf.B AIX 5L Version 5.1, Post-10/2002 firmware

Conf.C AIX 5L Version 5.2 or Linux, Post-10/2002 firmware

Conf.D AIX 5L Version 5.1, Post-5/2003 firmware

Conf.E AIX 5L Version 5.2 or Linux, Post-5/2003 firmware

Table 3-1 Physical memory size and number of allocatable partitions

Total Memory (in GB)	Approx. Memory Overhead (in GB)	Approx. Usable Partition Memory (in GB)	Conf.A Maximum Partition Number (see 1, 2 and 3)	Conf.B Maximum Partition Number (see 1, 2 and 4)	Conf.C Maximum Partition Number (see 2, and 5)	Conf.D Maximum Partition Number (see 1, 2, 6 and 7)	Conf.E Maximum Partition Number (see 1, 2, and 7)
2G	.75 to 1	1 to 1.25	0 and 0	4 and 0	4	4 and 0	4
4GB	.75 to 1	3 to 3.25	2 and 0	12 and 0	12	12 and 0	12
8GB	.75 to 1	7 to 7.25	6 and 0	16 and 0	16	28 and 0	28
16GB	.75 to 1	15 to 15.25	14 and 0	16 and 0	16	32 and 0	32
24GB	1 to 1.25	22.75 to 23	16 and 0	16 and 0	16	32 and 0	32
32GB	1 to 1.5	30.5 to 31	16 and 0	16 and 0	16	32 and 0	32
48GB	1.5 to 2	46 to 46.5	16 and 1	16 and 1	16	32 and 1	32
64GB	1.5 to 2.25	61.75 to 62.5	16 and 2	16 and 2	16	32 and 2	32
96GB	2 to 3.5	92.75 to 94	16 and 4	16 and 4	16	32 and 4	32
128GB	2.5 to 4	124 to 125.5	16 and 6	16 and 6	16	32 and 6	32
192GB	3.5 to 5.75	186.25 to 188.5	16 and 10	16 and 10	16	32 and 10	32
256GB	4.5 to 7.5	248.5 to 251.5	16 and 14	16 and 14	16	32 and 14	32
320 GB	5.5 to 9.25	310.75 to 314.5			16	32 and 18	32
384 GB	6.5 to 11	373 to 377.5			16	32 and 22	32

Total Memory (in GB)	Approx. Memory Overhead (in GB)	Approx. Usable Partition Memory (in GB)	Conf.A Maximum Partition Number (see 1, 2 and 3)	Conf.B Maximum Partition Number (see 1, 2 and 4)	Conf.C Maximum Partition Number (see 2, and 5)	Conf.D Maximum Partition Number (see 1, 2, 6 and 7)	Conf.E Maximum Partition Number (see 1, 2, and 7)
448 GB	7.5 to 12.75	435.25 to 440.5			16	32 and 26	32
512 GB	8.5 to 14.5	497.5 to 503.5			16	32 and 30	32

1. These columns contain two numbers. The first number corresponds to the maximum number of partitions with memory less or equal to 16 GB and the second number corresponds to the maximum number of partitions with memory greater than 16 GB.
2. All partition maximums are subject to availability of sufficient processor, memory, and I/O resources to support that number of partitions. For example, a system with only eight processors can only support a maximum of eight partitions.
3. These rules apply to systems running partitions with any version of AIX or Linux, if the firmware and HMC release levels are earlier than the 10/2002 release level. The overall number of partitions must be less than or equal to 16.
4. These rules apply to systems running partitions with AIX 5L Version 5.1 if the 10/2002 system microcode update or later and HMC Release 3 or higher are used. The HMC partition profile option for the Small Real Mode Address Region option should not be selected for AIX 5L Version 5.1 partitions. These numbers reflect the maximum when running only AIX 5L Version 5.1 partitions, but AIX 5L Version 5.1 and AIX 5L Version 5.2 partitions can be mixed, and may allow for additional partitions to be run (up to the maximum of 16).
5. These rules apply to systems running partitions with AIX 5L Version 5.2 (or later) or Linux, if the firmware and HMC release levels are at the 10/2002 release level or later. The HMC partition profile option for Small Real Mode Address Region should be selected for these partitions.
6. These rules apply to systems running partitions with AIX 5L Version 5.1 if the 5/2003 system microcode update or later and HMC Release 3 or higher are used. The overall number of partitions must be less than or equal to 32.
7. The I/O book FC6418 is required on pSeries 690 to support more than 16 partitions.

3.2.5 I/O books

You have to follow some rules when configuring the I/O drawers in a pSeries 670 or pSeries 690 with the 7040-681 or 7040-671 CEC wizard:

- All I/O books must use the same RIO or RIO-2 technology. Therefore, you can configure either:
  - One FC 6404, and zero to three optional FC6410, (one max for pSeries 670)

  - One FC 6418 and zero to three optional FC 6419 (one max for pSeries 670)
- By default, e-config will include one FC 6404 RIO book, that you can replace with a FC 6418 RIO-2 book.
- If you configure POWER4+ MCM, you must manually configure FC 6418 and FC 6419 RIO-2 book.
- If you need more I/O loops than what can be supported by the Primary I/O book, you need to manually add the required number of secondary books.
- The rules explaining the constraints between the installed MCM and the I/O books are described in detail in 2.3.3, “MCMs and GX slots relationship for pSeries 690” on page 39 and 2.3.5, “MCMs and GX slots relationship for pSeries 670” on page 44. The rules defining the relation between the number of required I/O books and the number of required I/O loops are described in 2.4.2, “I/O subsystem communication and monitoring” on page 56.

3.2.6 Media drawer configuration rules

The media drawer is configured within the CEC wizard. It consists of two sections and contains four media slots available for optional storage devices: The front section contains two media bays, and the rear section contains the two other media bays.

- One media drawer (FC 8692) is required for each pSeries 670 or pSeries 690 server and is automatically configured by default with one diskette drive and one optional CD-ROM driver (FC-2624).
- One media device capable of reading CD-ROM media (either CD-ROM FC 2624, DVD-ROM FC 2634, or DVD-RAM FC 2623) is required for each pSeries 670 or pSeries 690 system. You can replace the CD-ROM drive configured by default with a DVD-ROM or DVD-RAM drive.

Note: In order to use IDE DVD-ROM drive (FC 2634) in the media drawer, one SCSI to IDE Interface Bridge card (FC 4253) must be attached to each IDE DVD-ROM drive.

- The 4 mm tape drive (FC 6158) is allowed in the rear bays of the media drawer only if the operating environment is maintained at 24 degrees C (75.2 degrees F) or below and only on systems installed at 2134 meters (7,000 ft) altitude or below.
- The DVD-RAM drive, CD-ROM drive, and DVD-ROM drive are limited to systems without Primary Integrated Battery Backup (FC 6200) or Redundant Integrated Battery Backup FC #6201) when installed in the Media Drawer (FC 8692) rear bay locations.

- The IBM 80/160 GB Internal Tape Drive with VXA Technology (FC 6120) is limited to systems without Primary Integrated Battery Backup (FC 6200) or Redundant Integrated Battery Backup (FC 6201) installed, Media Drawer (FC 8692) front bay locations only, and a maximum system ambient operating temperature of 28 C (82.4 F) at a maximum operating altitude of 2134 m (7000 ft). Lower altitude have higher maximum ambient operating temperatures. Refer to the 7040 sections of *Site and Hardware Planning Information*, SA38-0508, for additional details.

The power to the media drawer is provided by the first RIO drawer. The connection of the media drawer storage options (CD, DVD Diskette) to the CEC is provided by SCSI adapters in the first RIO drawer. The cables to provide these connections are configured using the 7040-61D wizard:

- The connectivity between the front section of the media drawer and the first media drawer is configured by default by e-config, which will add in the first RIO drawer one power cable FC6179, one SCSI cable FC 2122 and one Single Ended SCSI adapter FC 6206.

  Note: The Single Ended SCSI adapter FC 6206 is a 5V adapter that requires an RIO planar. If you want to only use RIO-2 drawers, you need to change the default configuration to use a different SCSI adapter to attach the front section of the media drawer.

- If you install more than two optional storage devices in the media drawer, the rear section of the media will be used and therefore needs also to be connected to the first I/O drawer. You have to specify a second SCSI adapter. The second power cable FC6179 is automatically configured. If you add another Single Ended SCSI adapter FC 6206, one SCSI cable FC 2122 is automatically added to the configuration.
- You can also replace the default SCSI adapter by a PCI Dual Channel Ultra3 SCSI adapter (FC 6203).
- Each section of the media drawer needs to be connected to a different adapter, so if you use the rear media slots, and you do not want to use the Single Ended SCSI adapter, you must configure two PCI Dual Channel Ultra3 SCSI adapters. You cannot connect the two sides of the media drawer to the two ports of one Dual Channel Ultra3 SCSI adapter. If you only configure one SCSI adapter in I/O drawer 1, e-config will add a Single Ended SCSI adapter. For each PCI Dual Channel Ultra3 SCSI adapter you specify, one SCSI cable FC 2122 and one Converter cable VHDCI to P FC 2118 are automatically added to the I/O drawer configuration.

3.2.7 I/O drawer configuration rules

You have to follow some rules when configuring the I/O drawers in a pSeries 670 or pSeries 690 with the 7040-61D wizard, regarding the type of I/O planar to chose depending on the I/O books installed in the CEC and the type of adapters you need to install in the drawers. These rules have been described in detail in 2.4, “I/O subsystem” on page 51.

In addition, there are constraints that apply to the type, number and position in the drawer of I/O adapters. These contraints are described in the publication *PCI Adapter Placement References*, SA38-0538, and the pSeries 670 or pSeries 690 sales manual.

By default, e-config instantiates the first I/O drawer automatically for each new pSeries 670 or pSeries 690 configuration. You then need to manually add any extra drawer.

By default, each drawer contains four 4-pack Hot swap backplane (FC 6564) to receive internal SCSI disks, the first drawer is populated with 2 disks of 36 GB (FC 3158), while the other drawers do not contain any disks.

The internal SCSI adapters supporting the internal disk in the hot swap backplane are configured by default. They are part of the I/O planars.

3.2.8 I/O loops and cabling

The RIO and RIO-2 cables are automatically added to the configuration of each I/O drawer:

- Their type (RIO, RIO-2 or RIO-to-RIO-2) is defined by the types of I/O books configured in the CEC, and the type of I/O planar configured in the I/O drawers.
- Their number depends on the type of loop you choose: single-loop mode or dual-loop mode.

The granularity of loop-mode is on a per drawer basis. For each I/O drawer you configure, you find in the **Storage** tab of the wizard a checkbox called **Prefer Dual Loop**. If you check this box, two pairs of RIO cables are configured for this drawer.

The defaults settings are to use RIO books, drawers, and cables in single-loop mode.

3.2.9 Graphics console configuration rules

The pSeries 670 and pSeries 690 can have dedicated graphics consoles for each partition, including the Full System Partition. The configuration of the optional graphics console is performed through the 7040-61D wizard.

Up to eight graphics consoles can be configured in a system, and each console requires the following features:

- One POWER GXT135P graphics accelerator (FC 2848)
- One quiet touch keyboard: USB (FC 8800 or 8840)
- One mouse: Stealth black with keyboard attachment cable (FC 8841)
- One keyboard/mouse attachment card: PCI (FC 2737)

You can have a maximum of two of the GXT135P graphics accelerators and keyboard/mouse attachment cards per PCI planar, and a maximum of eight per system. Only one USB port is used to connect to the USB keyboard out of the available four USB ports on the keyboard/mouse attachment card. The USB mouse is daisy-chained from the USB keyboard.

Note: You cannot use this graphics console for the operating system installation.

3.2.10 Rack and power units configuration rules

The racks, Bulk Power Assembly and optional Integrated Battery Features are configured using the 7040-61R wizard.

When configuring the system rack, you must consider the following rules:

- All 7040-61R racks and expansion feature racks must have door assemblies installed. The following door assemblies are available:
  - A sculptured black front door with copper accent is automatically configured for the primary Model 61R rack (FC 6070) and the expansion rack (FC 6071), if installed.
  - An acoustic rear door (FC 6075) is available for the primary rack or the expansion feature rack. This door should be utilized for installations requiring a quieter environment.

Note: With the acoustic rear door, the hot air flowing out of the rack is blown upward toward the ceiling, while with the slim line rear door, the hot air is blown to the back of the rack. If your pSeries 670 or pSeries 690 servers are located in front of other machines requiring cooling, you may need to order the rack with an acoustic rear door. Please refer to the *Site and Hardware Planning Information*, SA38-0508, for environmental characteristics.

– A slim line rack door (FC 6074) is available for the primary or expansion rack feature. This door should be utilized for installations where system footprint is the primary consideration.

Note: The racks are 2.02 meters high, and may not pass through doors or fit in elevators during the delivery to a customer site. It is possible to order the rack with its top shipped separately, so that it is only 1.65 meters high. This feature is not configurable though e-config. You must create a valid rack configuration, and add an RPQ 8A1173 to the order. The RPQ is only orderable with the initial factory order.

The power assembly subsystem and all its options (Bulk power regulators FC 6186, Power controller features FC 6187, Power Distribution Assembly FC6188) are automatically configured with the right number of optional features to support the hardware features configured in the CEC and I/O drawers. The only feature of the power subsystem you need to specify is the line cord use to connect the frame to your site power distribution system. A default power cord is configured according to the country (or region) settings of the configurator.

The battery backup features (if ordered) supply power to the bulk power regulators (FC 6186). The primary battery backup features (FC 6200) function with the bulk power regulators located in the bulk power assembly in the front of the rack. The redundant battery backup features (FC 6201) function with the bulk power regulators in the bulk power assembly in the rear portion of the rack. If you decide to order the battery option, you need to order as many batteries as there are Bulk Power Regulators (FC 6186) configured in the racks. The cabling for the battery is automatically configured.

3.2.11 HMC configuration rules

The HMC models which are now available, and all optional features of the HMC must be configured using the 7315-C01 or 7315-C02 wizard. The HMC options that are visible in the CEC wizard are only for compatibility with former models of HMC which are no longer available.

The IBM Configurator for e-business includes by default with each HMC a mouse, a keyboard, a graphical display, the attachment cable to one pSeries 670 and pSeries 690 host, and the required software. They do not need to be ordered separately unless you want a different model than the default ones.

We strongly recommend that you order one optional Ethernet adapter. An Ethernet connection between the HMC and each active partition on the partitioning-capable pSeries server is required. This connection is utilized to provide several system management tasks, such as dynamic logical partitioning to each individual partition, and collection and passing of hardware service events to the HMC from the partition for automatic notification of error conditions to IBM.

For a serial connection between the HMC and the managed systems, the following rules apply:

- The HMC provides two integrated serial ports. One serial port is required for each pSeries server controlled by the HMC. An additional serial port is required for modem attachment if the Service Agent call-home function is implemented.
- If more than two serial ports are required on the HMC, an 8-port or 128-port asynchronous adapter (FC 2943 or 2944) should be utilized.
- One HMC[1] can manage up to twelve pSeries 670 or pSeries 690 systems, with a maximum total of 64 logical partitions on these systems combined.
- If a redundant HMC function is desired, the pSeries 670 or pSeries 690 can be attached to two separate HMCs.
- The 128-port adapter (FC 2944), in combination with a Remote Asynchronous Node (FC 8137), can be used for a long distance solution between HMC and a managed system. This will provide distances up to 1100 feet or 330 meters, while normal RS-232 connections allow up to 15 meters.
- Up to two 8-port and 128-port Asynchronous Adapters are allowed per HMC.

For use of an HMC in a cluster environment or with other pSeries models (630, 650, or 655), refer to the Sales Manual for exact limitation of the number of supported server per HMC.

3.2.12 SP Cluster 1600 considerations

pSeries 670 or pSeries 690 can be configured as part of a Cluster 1600 (7018-160) using e-config, by selecting the RS Cluster 1600 option in the

[1] The previous model of HMC (FC 7315 and 7316) only supports respectively up to four or eight managed systems.

"Product Line" section of the "Global Settings" panel when you start using the configurator (see 3.3.1, "Initial order" on page 98 for details).

In addition to the wizards mentioned in the previous sections, you will have to use the "Cluster Model 1600" wizard to specify the cluster options, which are the number of nodes of each type you will include in your cluster. e-config uses this number to check that your cluster complies with the limitations on the number of Cluster 1600 supported nodes, that are fully documented in the publication IBM @server *Cluster 1600 Harware Planning, Installation and Service*, GA22-7863.

Note: All limitations regarding maximum number of nodes are not checked by e-config. If you configure a large or complex cluster, we recommend that you manually check compliance against the information in IBM @server *Cluster 1600 Harware Planning, Installation and Service*, GA22-7863.

A cluster must contain at least two nodes. If you plan to grow your cluster over time, you may want to start with only one pSeries 670 or pSeries 690, and add extra servers into the cluster later on. This is possible as long as you create at least two LPARs in your server. The "Cluster Model 1600" wizard checks that there are at least two servers in the cluster. Select either two features 7018-0008 7040 type server or 7018-0009 LPAR (Switched).

After you have configured the cluster, you can configure pSeries 670 or pSeries 690 system with the same wizards than for a standalone server.

Cluster 1600 has limitations that are documented in the publication *RS/6000 SP Planning Vol.1, Harware and Physical Environment*, GA22-7863. For example, you cannot attach a graphical display to an LPAR in a clustered environment.

Note: The e-config wizards for the pSeries 670 or pSeries 690 system components are not aware that the servers are built as part of a cluster, therefore, they do not check for any cluster related restrictions. You have to check these restrictions manually.

If you were to configure a graphical display attached to an LPAR, e-config will not flag an error or warning. It will consider the configuration as valid, even though it is not a supported configuration.

Depending on whether you plan to manage your cluster with PSSP or CSM, you will also need to configure a Control Workstation or a Management station. These servers are not configured within the cluster. You need to add another "Initial Order" to your configuration, specifying "pSeries and RS/6000" as the product line, to chose the server you need.

The SP Switch (FC 8396) or SP Switch2 (FC 8397) adapters are “double-wide” PCI adapters. When you configure then, e-config will automatically add the right number of blind swap cassette for double-wide adapters (FC 4598).

The SP Switch (FC 8396) is a 5V PCI adapters that can only be installed in an RIO drawer (with I/O planar FC 6563).

You must think of configuring Ethernet adapters in each LPAR for the SP Management Ethernet network. If you use switched configurations, there are limitations on the slots where these adapters can be installed: Slots 8, 9, 18 or 19 for an SP Switch2 configuration, and slots 1 though 7 or 11 through 17 for an SP Switch configuration.

3.2.13 Upgrade considerations

If you have an installed pSeries 670 or pSeries 690 system, you may want to increase its processing power. As of May 2003, you have three possibilities for upgrading the system:

- MES: To install addition features, for example, to add an MCM or install RIO-2 books. This is the solution to use when your system limitations lies with the I/O subsystem, while the CPU and memory are able to cope with your applications workload, by replacing all RIO component with RIO-2 components.
- Feature conversion: To replace an installed component by a more powerful one. This is available for MCM and memory replacements to go, for example, from a 1.1 GHz system to a 1.5 GHz system.
- Model conversion, to convert a pSeries 670 system into a pSeries 690 system.

The complete list of all possibilities for any pSeries 670 and pSeries 690 component can be found in the sales manual. Here are a few rules related to these upgrades:

- Feature conversions are implemented on a “quantity of one for quantity of one” basis.

  In a pSeries 670, you can replace:

  – One 1.1 GHz 4-way MCM with one 1.1 GHz 8-way MCM
  – One 1.1 GHz 4-way MCM with one 1.5 GHz 8-way MCM
  – One 1.1 GHz 8-way MCM with one 1.5 GHz 8-way MCM

  You cannot replace one 1.1 GHz 4-way MCM with one 1.5 GHz 4-way MCM.

  In a pSeries 690, you can replace:

  – One 1.3 GHz 4-way HPC MCM with one 1.3, 1.5 or 1.7 GHz 8-way MCM

  - One 1.1 GHz 8-way MCM with one 1.3, 1.5 or 1.7 GHz 8-way MCM
  - One 1.3 GHz 8-way MCM with one 1.5 or 1.7 GHz 8-way MCM
  - One 1.5 GHz 8-way MCM with one 1.7 GHz 8-way MCM

  This is to say you can go from any pSeries 690 MCM configuration to a configuration with the same number of any more powerful MCM type.

- Model conversion from pSeries 670 system into a pSeries 690 is only available if the pSeries 670 has already two 8-way MCM installed. The conversion allows transition from a 1.1 GHz pSeries 670 to a 1.1 GHz pSeries 690, or from a 1.5 GHz pSeries 670 to a 1.5 GHz pSeries 690.

  During the conversion, some parts are replaced like the system planar (FC 9670), others are added like the Pass-through modules, and others are carried over, like the MCM, caches, I/O drawers, …

- During a feature conversion replacing POWER 4 MCM with POWER4+ MCM, the RIO books must be replaced with RIO-2 books.

  For the features that are prerequisite to the processor upgrade from POWER 4 to POWER4+, e-config will either automatically replace them, or flag an incompatibility and ask the user to change them.

3.3 IBM Configurator for e-business (e-config)

The IBM Configurator for e-business is an application that provides configuration support for hardware, software, and peripherals associated with IBM @server product lines that are available for marketing. Functions provided by e-config include:

- The ability to create, save, restore, and print hardware, software, and peripherals configurations.
- Hardware and software interaction for identifying prerequisite and incompatibility conditions.
- Interactive support for re-entering product categories and continuous modification and adjustments to the configuration.
- The ability to modify new or existing initial order, MES, or upgrade configurations.
- The ability to modify an installed base prior to beginning MES or upgrade configuration.
- Support for feature exchange and feature conversion.
- The ability to download and upload saved files to the Host/IBMLink™.

Note: Effective July 5th, 2002, the IBM Portable Configurator for RS/6000 (PCRS6000) is no longer supported. It has been replaced by e-config.

The e-config allows you to make valid configurations for pSeries 670 and pSeries 690 servers. The e-config checks for prerequisites and incompatibilities, and often will correct them automatically. However, you will save time if, before using e-config, you have an understanding of the constraints implied by the pSeries 670 and pSeries 690 server architecture, that are listed in:

- The IBM sales manuals for the latest information
- Section 3.2, “Configuration rules for pSeries 670 and pSeries 690” on page 81

This section assumes that you have e-config already installed on your workstation. It will be easier for you to understand this section if you can exercise e-config on your display while you read this chapter.

In this section, we only address standalone pSeries 670and pSeries 690 servers. We do not cover configuration of clustered servers.

For information on how to install and start e-config, please check the e-config home page:

`http://ftp.ibmlink.ibm.com/econfig/announce/index.htm`

3.3.1 Initial order

Start e-config and select a blank portfolio. In the main menu, select **Portfolio -> Add Initial Order** or click the **Add Initial Order** icon.

Figure 3-2 Add Initial Order icon

Enter the name and description of the configuration, select **pSeries and RS/6000 Systems**, and click **OK**. A menu will appear with options to select the geographic region and/or country, and customer requested arrival date (CRAD). You can also specify pricing options on the second tab.

On the third tab, select **pSeries** and **RS/6000 for Product Line**. Select **Initial** for Order Type and select either AIX pre-installed, AIX not required, or AIX but not pre-installed in the AIX options box. Click **Configure**. A panel similar to Figure 3-3 on page 99 appears.

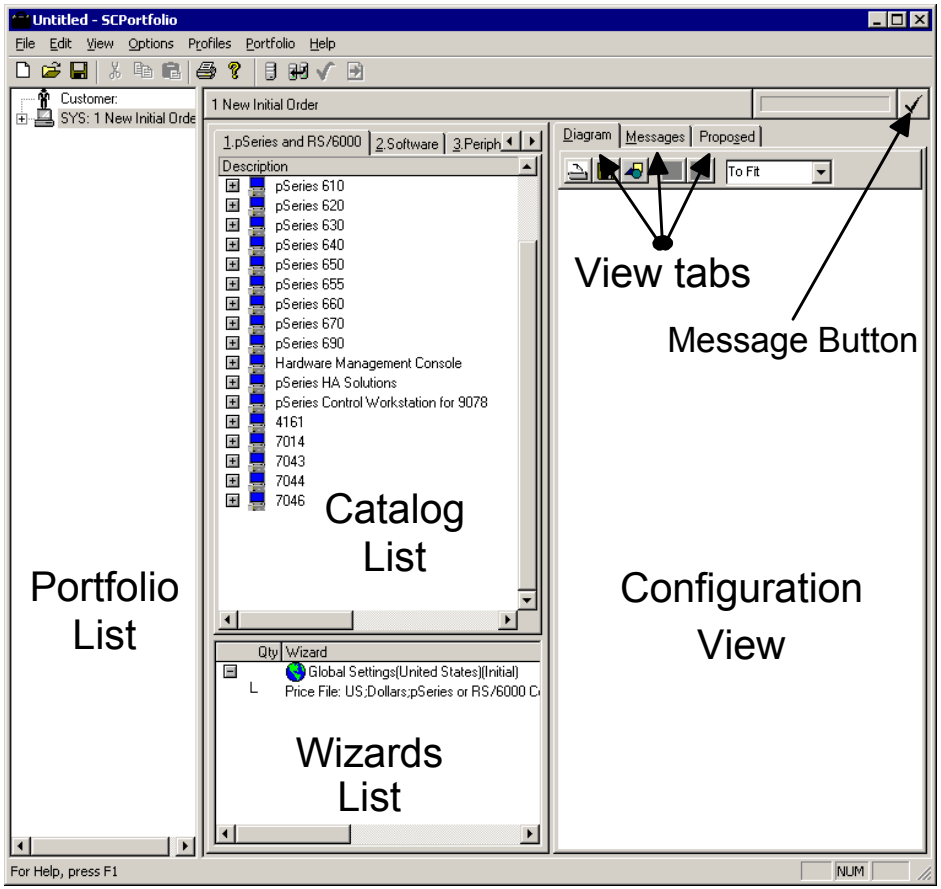


Figure 3-3 The e-config main panel

On the Catalog list, search for pSeries 690 and click it to open the menu tree. Double-click **Server**. You will see that the Wizards list has been updated with the pSeries 690 basic components: The rack, the I/O drawer, and the server (CEC) itself. Also, if you have selected the option to include AIX, the Software Global Options and Software Products for the specified server will be present.
Figure 3-4 shows the physical placement of these components inside the rack. In the Configuration view there is a graphical representation of the pSeries 690 components.

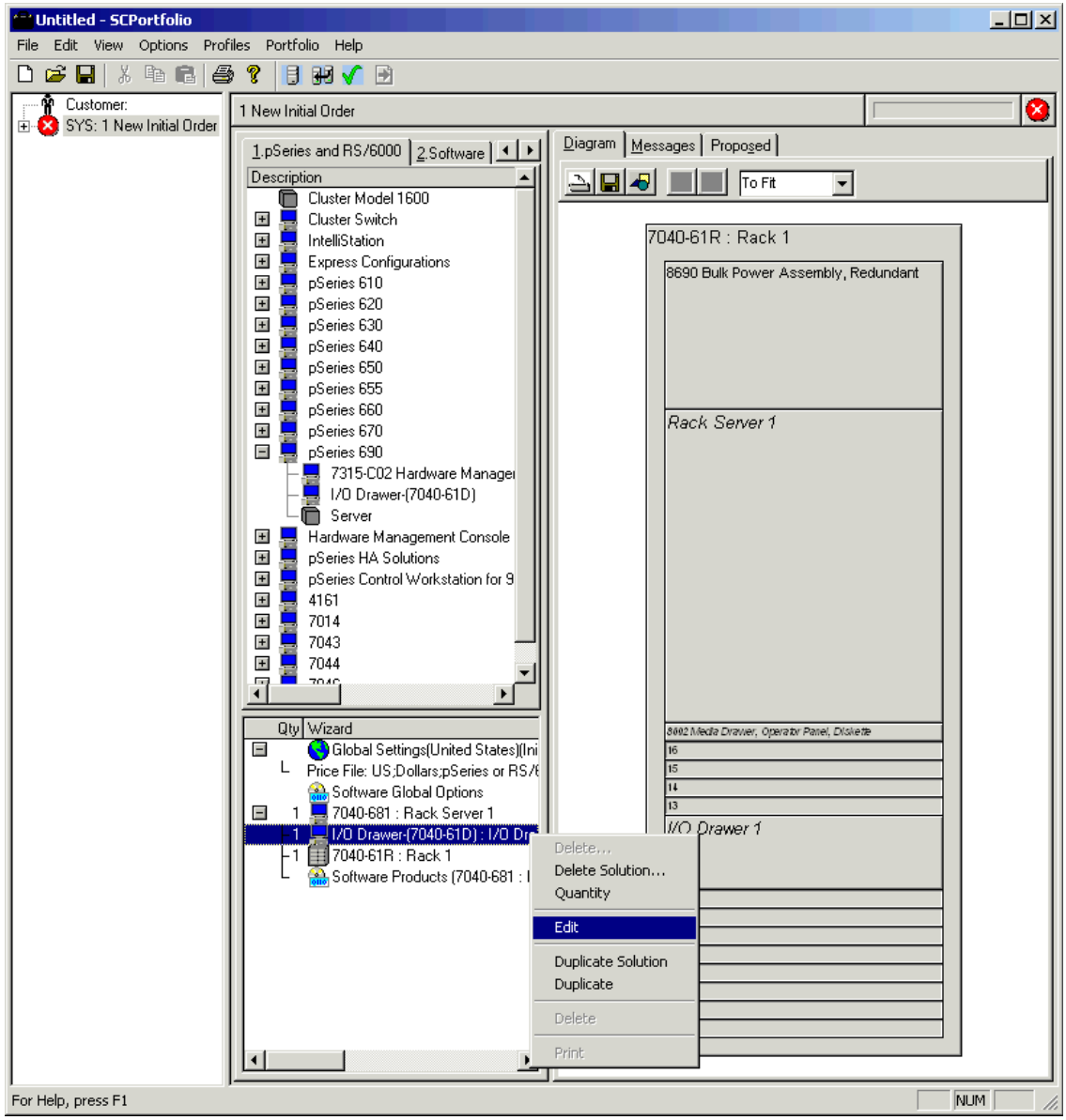


Figure 3-4 Graphical representation of pSeries 690 in configuration view

To edit the configuration and add or remove features, double-click the desired component on the graphic in the Configuration view or in the Wizards list. Or right-click the item in the Wizards list and select **Edit**, as shown in Figure 3-4. To add more items of the same type, select **Duplicate**. A wizard is presented with

the features available for that component. To duplicate the whole pSeries 670 or pSeries 690 configuration, **Duplicate Solution** can be selected. Software options, including AIX-enhanced software subscription options may be altered by double-clicking **Software Products: (7040-681: Rack Server 1)**. Follow the dialog boxes from there and select the desired software components together with the desired period of enhanced software subscription or software maintenance.

CEC options

In the Wizards list, double-click **7040-681: Rack Server 1**. The first panel of the wizard (**Products**) enables you to change the name of the server; select the operating system and specify the number of LPARs that the system will have.

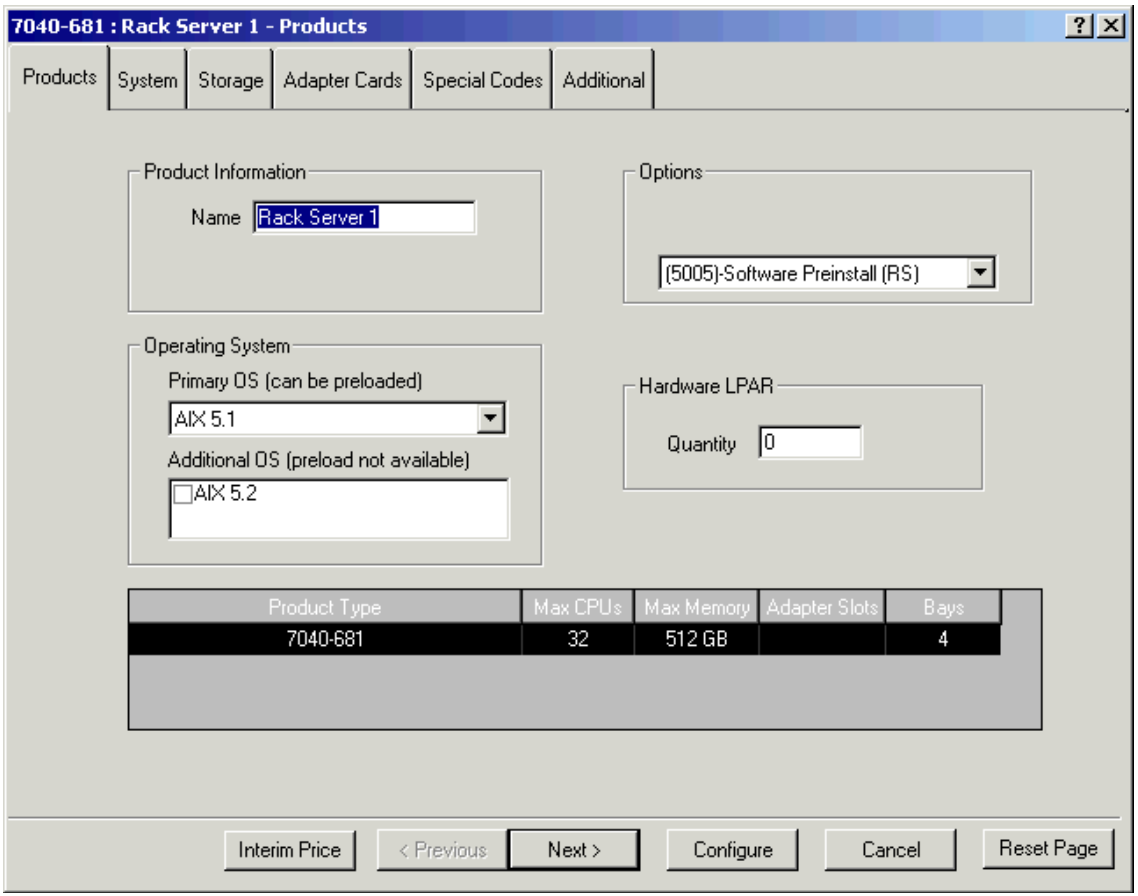


Figure 3-5 pSeries 690 CEC wizard after double-clicking Rack Server 1

Note: The number of LPARs is for information purposes only. It does not change features in the system, and does not prevent the customer from altering the number specified originally.

You can configure the number and type of MCMs, I/O books and memory cards in the **System** tab (see Figure 3-6).

Unless you order CuOD MCM, you should only enter a figure in one line of the processor option list to select one type of MCM. For the 1.1 GHz and 1.3 GHz systems, where the feature code is different for the first MCM and the others, e-config will automatically pick the right feature codes to satisfy the number of MCM you specify.

On the same tab, in the **System Options** list, you choose the RIO books to connect I/O drawers. When configuring one or two RIO drawers in single-loop mode connected to an RIO book, no extra RIO book is needed. For other configurations, refer to 2.4.2, “I/O subsystem communication and monitoring” on page 56 to define the type and number of I/O books needed on your system, then select the required FC 6404, 6410, 6418, and 6419.

Then configure the memory books you need. In systems with one or two MCMs, you should only configure inward-facing memory cards. In systems with more than two MCMs, you can configure both types.

Then press the **Next >** button. If there are inconsistencies between the various choices you made, you will receive a warning or error message.

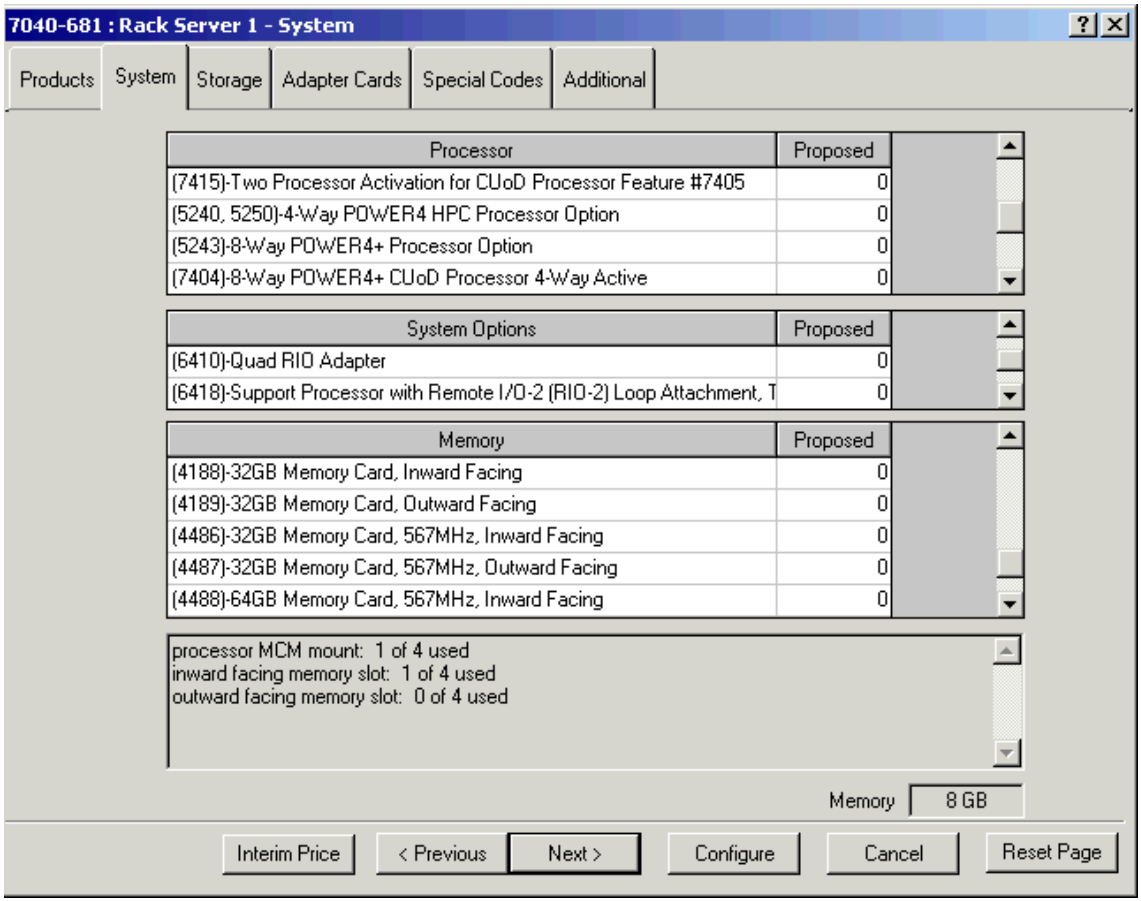


Figure 3-6 System tab for the pSeries 690

After clicking **Next**, the **Storage** tab allows you to configure the media options for the media drawer. You can select between the available media devices; up to four devices can be configured into the media drawer. Remember the media drawer is split into two separate sections. The front section houses the system operator panel, a diskette drive, and two media bays. Utilization of the rear section is optional to provide space for two additional media devices if desired.

The front and rear sections of the media drawer must be connected to separate SCSI adapters in the first 7040-61D I/O drawer. If you want to use the rear section of the media drawer (for example, to assigned media devices to two different partitions), you must configure two SCSI adapters in the first 7040-61D I/O drawer (see “I/O drawer options” on page 106).

The **Adapter Cards** tab presents the cabling options to the HMC for a simple one to one connection. If you want to manage several pSeries from the same HMC, you have to select 8-Port or 128-Port Asynchronous adapters as options of the HMC.

On the **Additional** tab, you can specify the language group for your system. You should not use the entries in the Line Cords options. These features are only left in this panel for consistency with the first generations of pSeries 690, to specify the appropriate power cords for the HMC (FC 7316 and 7315) and its associated displays. The HMC, its display and their power cords are now configured in a different wizard. Power cabling for the 7040-681 Central Electronics Complex is provided from the bulk power assemblies in the 7040-61R rack and is not configured in the CEC wizard.

On the **Additional** tab, you can also specify orders for On/Off CoD.

After you select the options from the tabs, click **Configure** to update your configuration.

If there are inconsistencies between the various choices you made, you receive a warning or error message, and after clicking the **OK** button you are presented again with the wizard panel. If the configuration is valid, the wizard panel closes and you return to the main panel.

You can see the physical location of the components in the CEC. On the diagram in the Configuration view, right-click the area corresponding to the CEC (the area named Rack Server 1), and select **Detailed Diagram**. A figure similar to Figure 3-7 is presented.

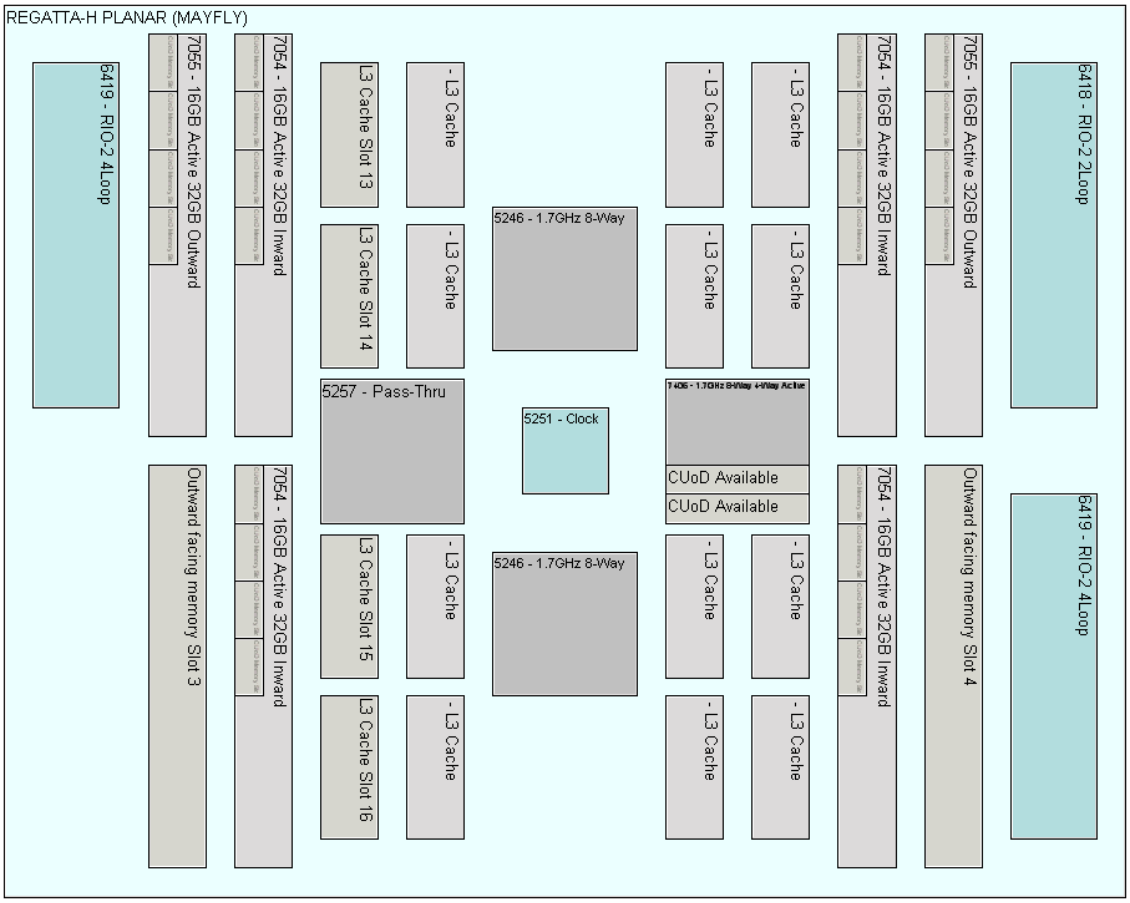


Figure 3-7 Detailed diagram for the pSeries 690 CEC

Note that the unoccupied MCM slots are covered by processor bus pass-through modules, when the machine is configured with two or three MCMs. The e-config automatically places the modules when necessary.

Similarly, you can see the physical location of the components in the media drawer using the same procedure. Figure 3-8 shows a detailed diagram for the media drawer.

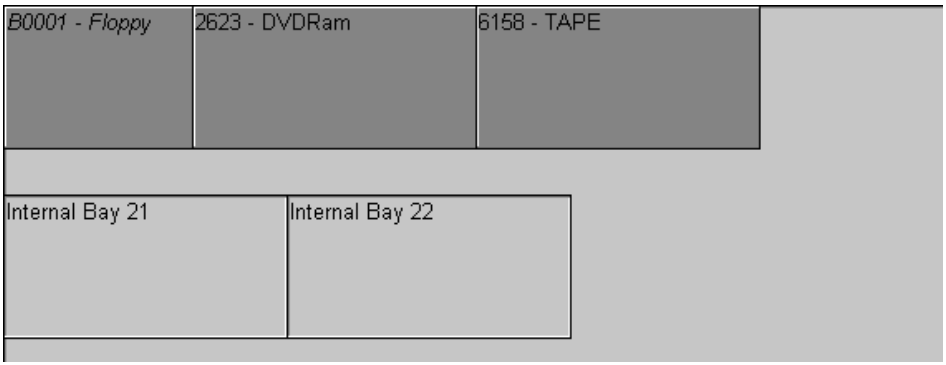


Figure 3-8 Detailed diagram for the media drawer

I/O drawer options

Double-click the I/O Drawer 1 in the Configuration view. The first panel of the wizard enables you to change the name of the drawer. The placement is done automatically. The Target Host drop-down menu lets you select the server to connect the I/O drawer. Currently, you should have a single server in the selection list. Then, click the **Next >** button to display the Storage options, as shown in Figure 3-9.

The **Storage** tab allows you to select the internal disks for the I/O drawer. You can select up to a total of 16 disks. The 4-pack hot swap backplanes come configured by default. Disks will be automatically placed by e-config in an optimum and valid configuration; this cannot be changed in the e-config. Actual physical installation will allow for exact placing of the disks.

In the **Hardware** list of the **Storage** tab, you can choose between RIO, RIO-2 or mixed drawers, by selecting the I/O planar feature codes:

- Two FC 6563 for an RIO drawer,
- Two FC 6571 for an RIO-2 drawer, or
- One FC 6563 and one FC 6571 for a mixed drawer

In the first I/O drawer, you can also select the number of cables needed to connect to the Media drawer (FC 2122). If you use have configured three or four storage devices in the media drawer, you need to manually configure a second SCSI drawer, and the two FC 2122 are automatically configured, or you can specify two FC 2122 cables in the **storage** tab, and the two SCSI adapters are automatically configured. However, if you select only two storage devices in the media drawer, and you want to install one in front, and one in the back section, you must manually specify a second FC2122 in the **storage** tab, as well as a

second SCSI adapter in the **Adapter Cards** tab and a second Power cable (FC 6179) in the **Additional** tab.

Left of the **Hardware** list, you can see the **Prefer Dual Loop** checkbox (highlighted by a circle in Figure 3-9). If you click this box and you have configured a RIO-2 or mixed drawer, e-config will automatically configure two pairs of IO cables of the right type. If you have configured an RIO drawer, clicking this checkbox has no effect, since RIO drawers are not supported in dual-loop mode.

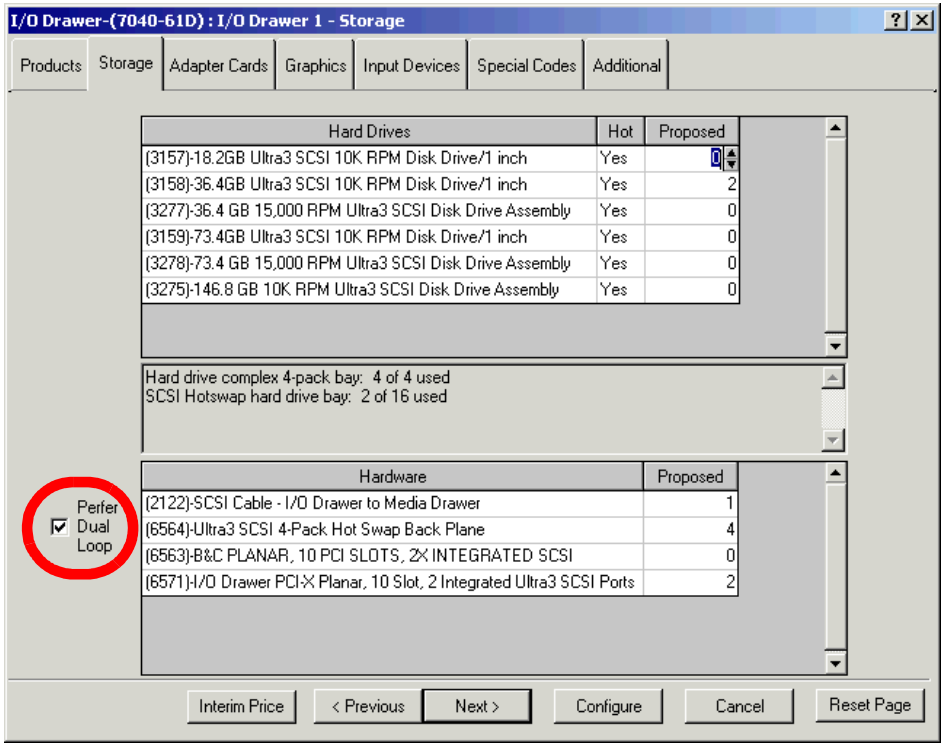


Figure 3-9 Selection of I/O drawer storage and RIO cabling options

The **Adapter Cards** tab provides menus to configure the different adapters on the I/O drawer. Figure 3-10 on page 108 shows the tab with the menu options. Adapters will be automatically placed by e-config in an optimum and valid configuration; this cannot be changed in the e-config. Actual physical installation will allow for exact placing of the adapters.

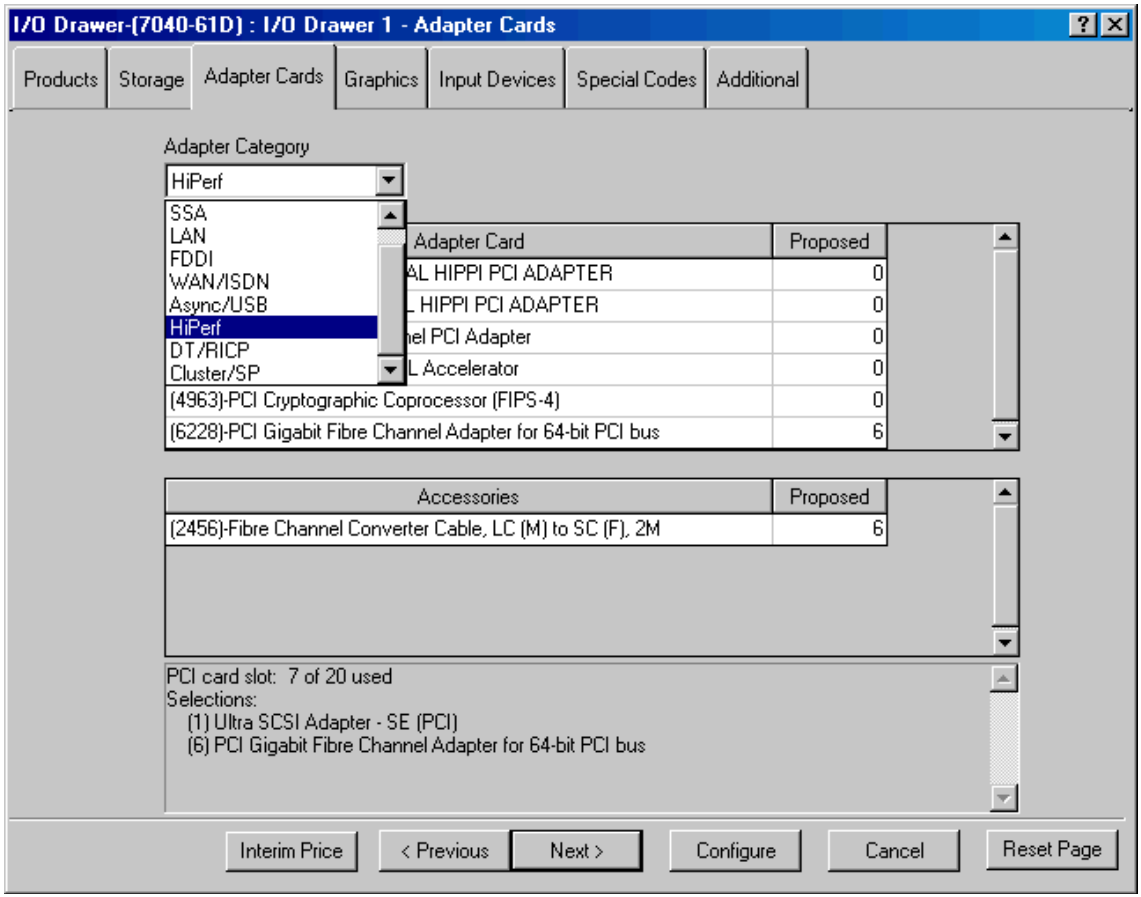


Figure 3-10 Adapter tab for the I/O drawer

Upon selection of one adapter category, the adapters that belong to that category are shown and you can specify the number of adapters on the I/O drawer. The configurator places the adapters automatically into the I/O drawer in a balanced placement whenever possible. Figure 3-11 on page 109 shows the options when selecting the LAN adapter category.

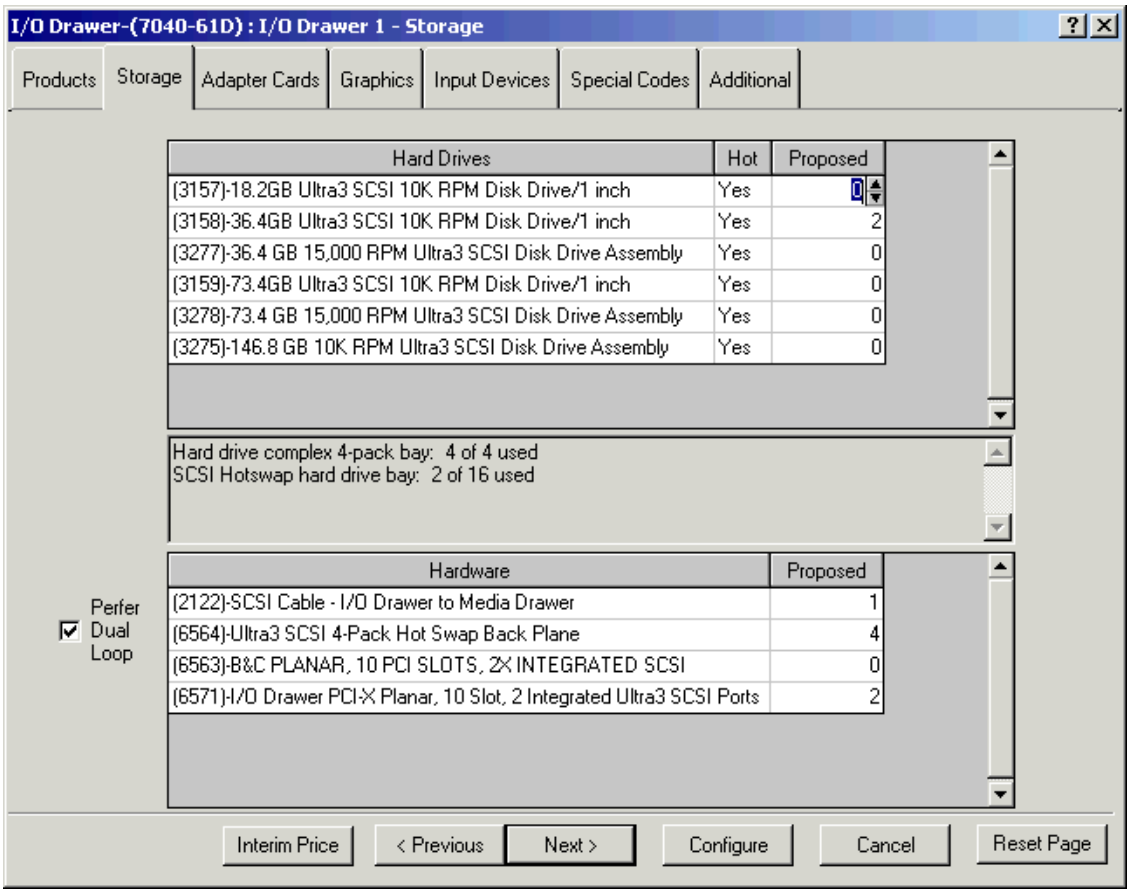


Figure 3-11 Adapter tab with LAN adapter category selected

On the **Graphics** tab, you can select graphics adapters and monitors for the pSeries 670 and pSeries 690 systems. The maximum is one per logical partition and eight per system. Be sure that with every graphics adapter (FC 2848) the following devices are also configured:

- One quiet touch keyboard: USB (FC 8800 to 8840), and one mouse: Stealth black with keyboard attachment cable (FC 8841) from the **Input Devices** tab

  And

- One keyboard/mouse attachment card: PCI (FC 2737, found on the **Adapter Cards** tab, Async/USB)

On the **Additional** tab, you can specify in the first list the language group for the publications provided with the hardware.

The **Line Cords** list of the **Additional** tab presents the optional power cords needed if you configure a graphical display in the **Graphics** tab. When you add a display, a power cord is automatically configured according to the e-config country (or region) settings. You can then override this choice in the **Line Cords** list.

In the **Additional** tab, you can also specify the number of power cables for the media drawer. If you add more than two devices on the media drawer, then you have to add another power cable. If you do not add the cables, e-config will automatically add them for you when necessary. The power cable to the I/O drawer (FC 6172) is automatically configured and should not be modified.

The **Miscellaneous Equipment** list of the **Additional** tab allows you to specify Double Wide Adapters cassette for holding large PCI adapters. The number of cassettes is automatically configured depending on the number of adapters you selected in the **Adapter Cards** tab. You only need to change this number if you want extra cassettes, for example to hold adapters you already own.

After you select the options from the tabs, click **Configure** to update your configuration. The notebook will close and return to the main panel.

Similar to the CEC, you can see the physical location of the components in the I/O drawer by right-clicking the area corresponding to the I/O drawer (the area named pSeries 690 (7040-61D) Additional I/O Drawer: I/O Drawer 1) and selecting **Detailed Diagram**. Figure 3-12 shows the diagram view. PCI adapters and disks are placed in optimal configuration automatically by e-config. This cannot be changed in the configurator. Exact placement of disks and adapters can be done at physical installation time.

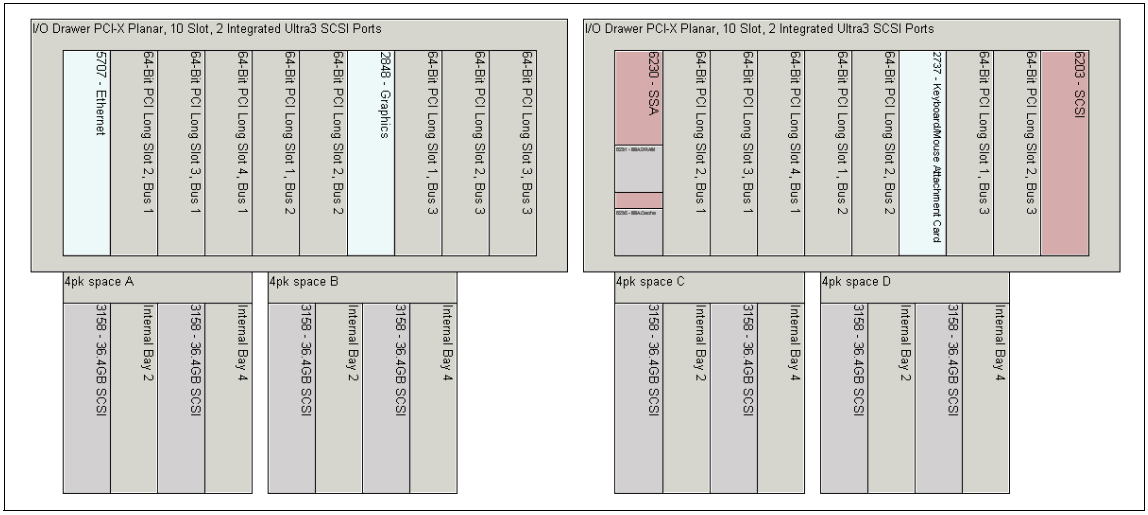


Figure 3-12 Detailed diagram of the I/O drawer

You can now configure additional I/O drawers. If you have a need for several I/O drawers with the same or similar configuration, the easiest solution is to right-click on an already configured I/O drawer in the Wizards list (see Figure 3-13), and then edit the new I/O drawer to modify its features.

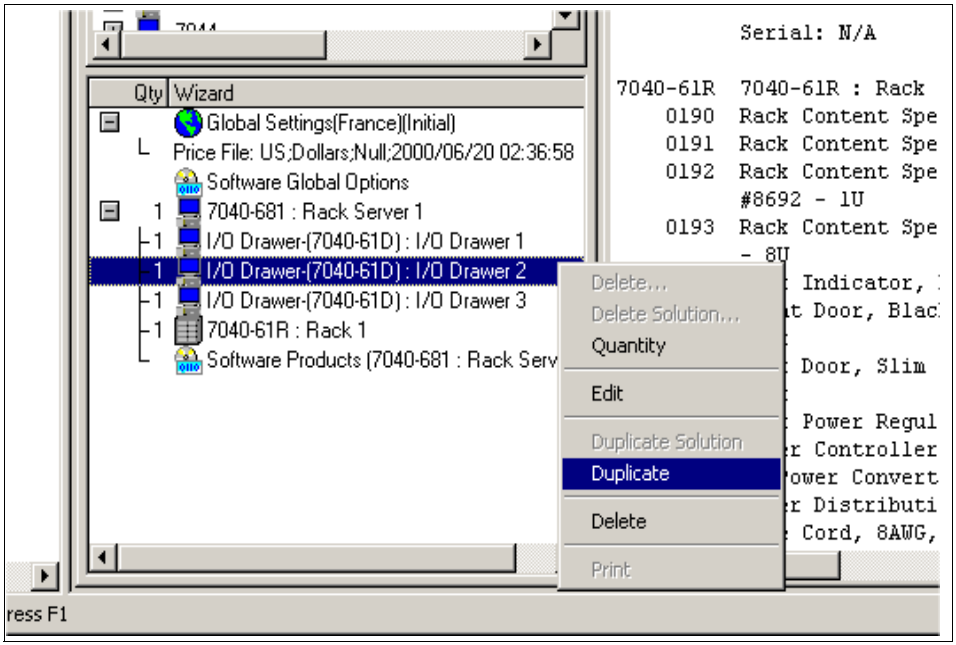


Figure 3-13 Duplicating an existing drawer

If you want a totally different drawer configuration, you should instead double-click on the **I/O Drawer - (7040-6D1)** entry in the Catalogs list, under the **pSeries 690** main entry:

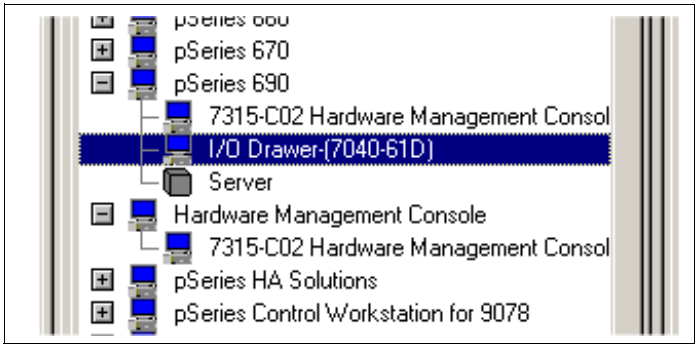


Figure 3-14 Selecting a new I/O drawer

Rack options

Double-click the **7040-61R** graphic in the Configuration view or right-click **7040-61R: Rack 1** in the Wizards list and select **Edit**. The first panel of the wizard enables you to specify front and rear doors for the rack, and add an additional 42U expansion rack. The acoustic rear door (FC 6075) should be utilized for installations requiring a quieter environment.

When more than four I/O drawers are ordered, or four I/O drawers with integrated battery backup (FC 6200), an expansion rack is automatically selected by e-config. There is only one wizard that contains the features for both racks. The expansion rack is indicated by the number “1” in the line **(8691) - 42U Expansion** in the O**ptions** tab. The doors for both racks can be configured by double-clicking any of the two racks. You have to specify two rear doors, either acoustic (FC 6075) or slim line (FC 6074).

The **Additional** tab lets you select the power cord options for the rack. A default selection based on the country (or region) includes two cables. You can change these cables and replace with them others. You must always specify two cables.

Optionally, you can select the optional battery features. Select the integrated battery backup, primary (FC 6200). As an option for the primary battery backup, you can select the redundant features (FC 6201). You can add up to three of each battery. For each redundant battery, you must specify a primary battery.

Note: When primary integrated battery backup feature(s) are ordered, one is required for each bulk power regulator (FC 6186) in the front bulk power assembly. The e-config will notify you when more battery backup features are necessary and select them automatically.

Hardware Management Console

The pSeries 670 and pSeries 690 servers need to be connected to an HMC. If you already have an HMC installed that can be used to manage the server being configured, you can skip this step. Otherwise, you need to configure an HMC. To do this, click on the button on the left of the **Hardware Management Console** entry in the Catalogs list. Then double click on the **7315-C01** or **7315-C02 Hardware Management Console** entry. You are then presented with the HMC wizard window, as presented in Figure 3-15.

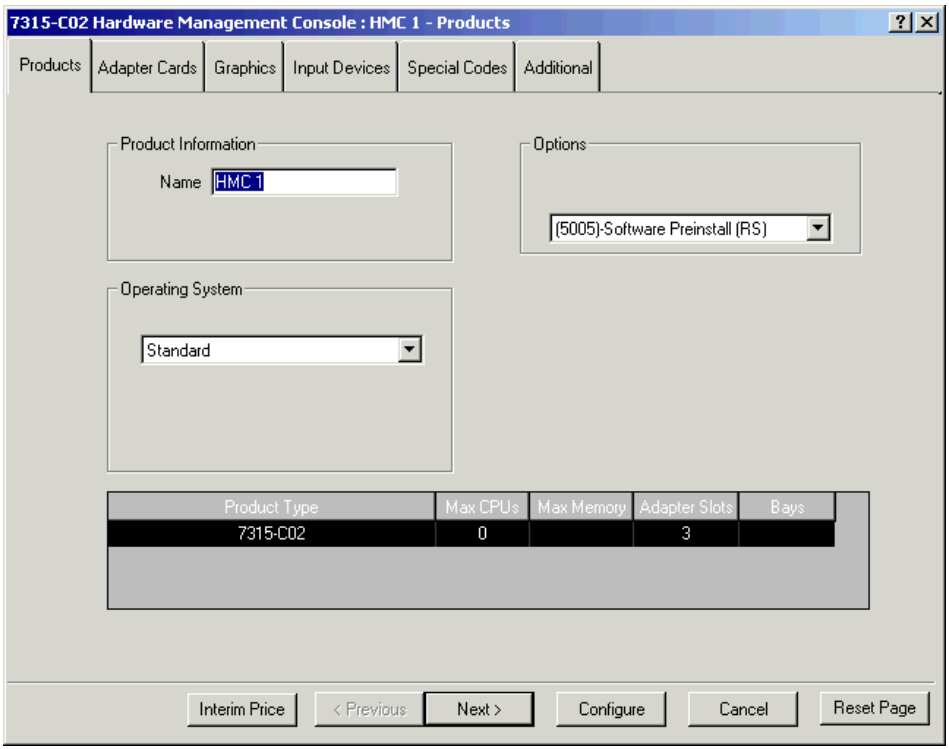


Figure 3-15 HMC wizard

In the **Products** tab, you can change the HMC name.

Then skip to the **Adapter Cards** where you are presented by default with the possibility of ordering extra Ethernet adapters. Click on the Adapter Category pull-down menu to select Async/USB, as shown on Figure 3-16. You can then select the attachment cables and optional Asynchronous adapters to connect the HMC to the pSeries 670 and pSeries 690 servers.

The **Graphics** and **Input Devices** tabs allow you to change the display and keyboard configured by default.

The **Additional** tab contains the default language and power cords corresponding to the country (or region) settings, that you can override by selecting other features.

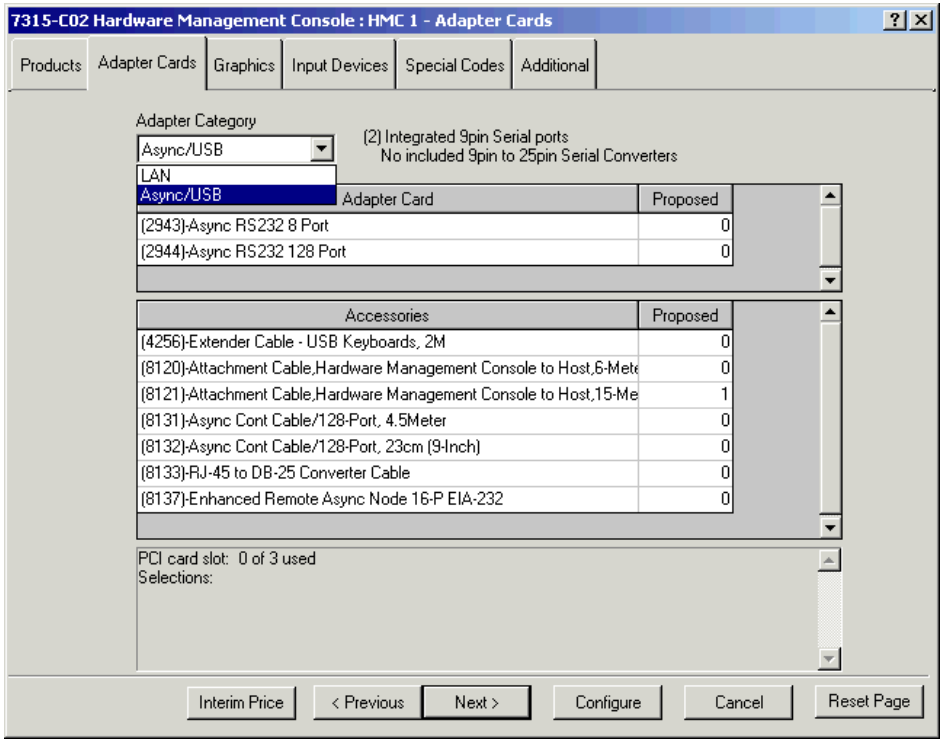


Figure 3-16 HMC adapter options

You then click on the **Configure** button to add the HMC to the configuration.

Validating the configuration

Before validating the configuration, you should look for any information message, warning or error reported by the configurator. Click the **Message** tab on the **Views tabs** area, or click the **Message** button (see Figure 3-3 on page 99). A panel similar to Figure 3-17 appears.

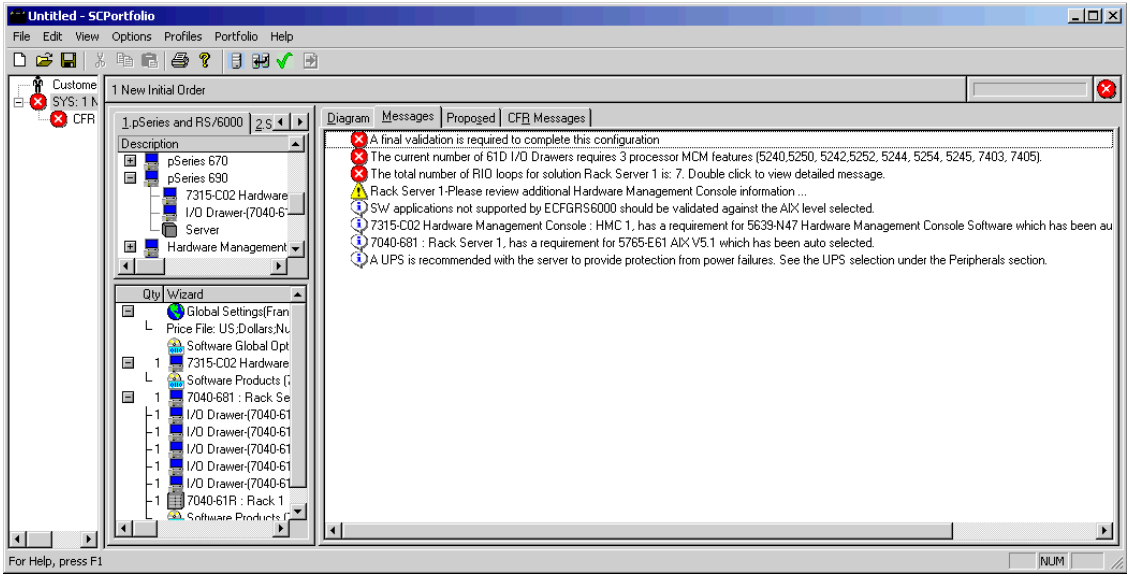


Figure 3-17 The error, warning, and information output of the configuration

If you right-click a message, a context-sensitive menu appears and gives you the ability to execute actions on that message by selecting **Jump to Wizard**. Figure 3-18 shows the possible actions for an error message.

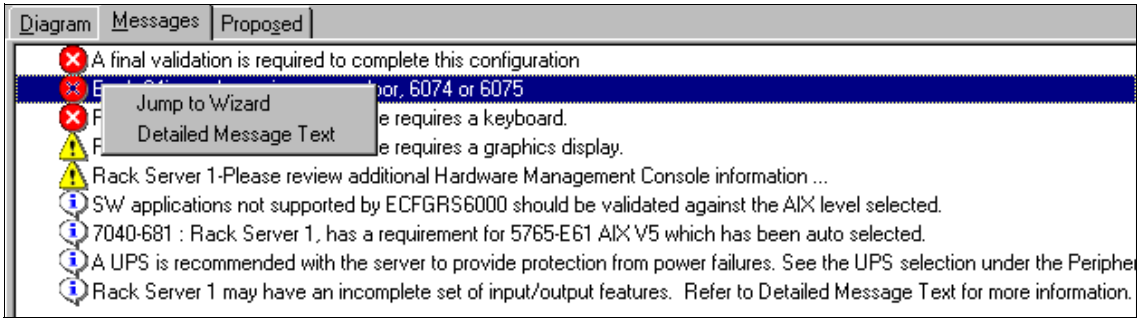


Figure 3-18 Right-click the error message text and it opens a menu

If you select the option **Detailed Message Text**, a panel opens with more information about the message, as shown in Figure 3-19. Click **OK** to return to the messages panel.

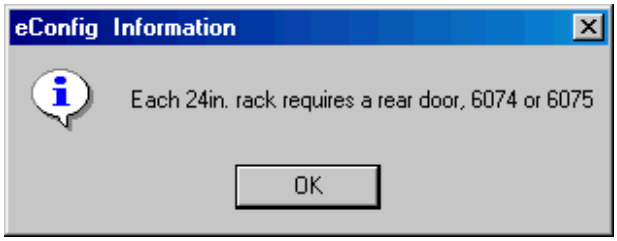


Figure 3-19 Detailed error message

Another option is to select the **Jump to Wizard** action. If you select this, the wizard panels related to that message will open, and you can perform the required changes.

After correcting the configuration errors, right-click the message:

```
A final validation is required to complete this configuration
```

Click the **Perform Final Validation** icon (Figure 3-20).

Figure 3-20 Perform Final Validation icon

If no other errors are encountered, the system is validated and you can save the output files and place the order.

On the **Proposed** tab, you can find a text description of the configuration, with all the feature codes and prices for the hardware and/or software components.

To generate a price proposal and place an order, you have to save the files. Select **File -> Save** or **File -> Save As**, or click the diskette icon in the upper left corner of the panel.

3.3.2 Performing an upgrade

The configuration steps to perform an upgrade on a pSeries 670 or pSeries 690 system are the same steps used to configure a new one. The only thing that changes is that instead of specifying an initial order you should specify Upgrade/MES.

From the main menu, select **Portfolio -> Add Upgrade/MES or Restore CFReport**, or click the **Add Upgrade/MES or Restore CFR** icon.

Figure 3-21 Add Upgrade/MES or Restore CFR icon

If importing from a previously saved CFReport file, select **Import local file**. Select the CFReport file to be imported, and after a moment the base configuration is presented. You can then change components or add new components to the configuration and proceed with the upgrade.

3.4 Configuration examples

In this section we provide the steps needed to make a complete configuration of a pSeries 670 and a pSeries 690 system. We present several scenarios, going from an initial order of a simple basic configuration to the update of a large configuration with multiple I/O drawers and an expansion rack:

- Section 3.4.1 presents the initial order of a pSeries 670.
- Section 3.4.2 describes the initial order of a 2-frame pSeries 690.
- Section 3.4.3 explains how to perform a model conversion from a pSeries 670 into a pSeries 690 system.
- Section 3.4.4 presents feature conversion from a 1.3 GHz pSeries 690 to a 1.7 GHz pSeries 690.

In these sections we assume that the user is familiar with e-config and how to create a new configuration. If any difficulties arise, we recommend that you see 3.3, “IBM Configurator for e-business (e-config)” on page 97.

3.4.1 Configuration example 1: pSeries 670 (16-way 1.1 GHz)

In this section we will prepare an order for a machine with the following configuration:

- 16-way 1.1 GHz processors
- 64 GB of memory
- Eight 36 GB internal disks
- Four logical partitions
- Four Gigabit Ethernet adapters and four Fibre Channel adapters
- One CD-ROM and one 4 mm DAT tape drive
- One I/O drawer (RIO loops)

- One HMC with a 6 meter cable and 21 inch monitor, US keyboard, and mouse
- US Language and US power cords
- AIX 5L Version 5.1 with three years enhanced software subscription

To prepare the configuration reports, follow these steps:

1. In the Catalog list area, double-click pSeries 670 and double-click **Server**. You will find a graphical representation of the pSeries 690 in your Configuration view.
2. Double-click **Rack Server 1**. Select the AIX version and fill in the number of required LPARs. Click **Next**. Go to the **System** tab and change the quantity of the 4-way POWER4 processor from one to zero, and the quantity of the 8-way POWER4 processor (FC 5256) to two. On the same tab, change the quantity 4 GB memory card, inward facing (FC 4196), from one to zero, then add four 16 GB memory cards, inward facing (FC 4183).
3. Go to the **Storage** tab and add a 20/40 GB 4 mm tape drive (FC 6158).
4. Click the **Next** button to move to the **Adapter Cards** tab. Change the quantity of the console/host serial attachment cable, 15 M (FC 8121), from one to zero, and change the quantity of the console/host serial attachment cable, 6 M (FC 8120), from zero to one. Click the **Configure** button to return to the main panel.
5. Double-click **I/O Drawer 1** in the Configuration view to open its wizard. Click the **Storage** tab. Change the quantity of the 36.4 GB Ultra3 SCSI 10 K RPM disk drive/1 inch - (FC 3158) from two to eight.
6. Click the **Next** button to move to the **Adapter Cards** tab and from the **Adapter Category** menu; select **LAN**. Add four of the Gigabit Ethernet–SX PCI-X adapters (FC 5700). Select the **HiPerf** category on the menu and add four of the PCI Gigabit Fibre Channel adapters for 64-bit PCI bus (FC 6228).

   Note: The FC 6228 card uses an LC connector. Most of the Fibre Channel devices still use the SC connector. Therefore, we recommend that you add the Fibre Channel Converter Cable, LC (M), to SC (F), 2M (FC 2456) for each Fibre Channel adapter being configured.

7. Click the **Configure** button to return to the main panel.
8. Double-click **7040-61R** in the Configuration view to open its wizard. Add one rear door, acoustic, primary or secondary rack (FC 6075).
9. Go to the **Additional** tab and select a language group: US English (FC 9300). Click the **Configure** button to return to the main panel.

10. In the Catalog list area, double-click **Hardware Management Console,** then double-click **7015-C02 Hardware Management Console**. The HMC wizard appears on the screen.
11. Change to the **Graphics** tab and add replace the IBM P76/P77 display with one IBM P260/P275 color monitor, stealth black, and cable (FC 3628).
12. Go to the **Input Devices** tab and add one quiet touch keyboard: Stealth black US English, #103P (FC 8800).
13. Click the **Configure** button to return to the main panel.
14. Click the **Messages** tab on the main panel and right-click the message. A final validation is required to complete the configuration, so click the **Perform Final Validation** icon (Figure 3-22).

Figure 3-22 Perform Final Validation icon

15. If you follow all these steps, your configuration will be valid, and you can save the output file for processing. If you encounter any problems, check the steps again and refer to the documentation.

3.4.2 Configuration example 2: pSeries 690 (24-way 1.3 GHz)

In this section we will configure a machine with the following resources:

- 24-way1.3 GHz processors
- 96 GB of memory
- Four I/O drawers, with RIO loops
- Forty 36 GB internal disks (ten in each I/O drawer)
- 16 Gigabit Ethernet adapters 16 Fibre Channel adapters (four in each I/O drawer) and two PCI ESCON Channel adapters (in the first I/O drawer)
- One DVD-RAM drive and one 4 mm DAT tape drive
- Two HMCs with 15 meter cable and 21 inch monitor, US keyboard, and mouse
- Two Integrated Battery Feature options, one being principal and the other redundant
- US Language and US power cords
- AIX 5L Version 5.1 and PCI ESCON Control Unit Connectivity software Version 2.1

Follow these steps to perform the configuration:

1. In the Catalog list area, double-click pSeries 690 and double-click **Server**. A graphical representation of the pSeries 690 appears in your Configuration view.
2. Double-click **Rack Server 1**. Go to the **System** tab and change the quantity of the 8-way POWER4 processor (FC 5242, 5252) from one to zero. Add three of the 8-way POWER4 Turbo processors (FC 5244, 5254). On the same tab, change the 8 GB card from one to zero, then add four 16 GB memory cards, inward facing (FC 4183), and two 16 GB memory cards, outward facing (FC 4184).

   There are several ways to configure memory on the pSeries 670 and pSeries 690. The suggested configuration is recommended for best performance. For more information on how to configure memory on the pSeries 670 and pSeries 690, see Table 2-4 on page 36 and Table 2-6 on page 44, respectively, or consult with the *IBM @server pSeries 690 Configuring for Performance* white paper, found at:

   http://www.ibm.com/servers/eserver/pseries/hardware/whitepapers/p690_config.html
3. Go to the **Storage** tab and change the quantity of the CD-ROM drive–32X (Max) SCSI-2 (FC 2624) from one to zero. Add a 20/40 GB 4 mm tape drive (FC 6158) and one DVD-RAM SCSI re-writable 4.7 GB black bezel (FC 2623).
4. Click the **Next** button to move to the **Adapter Cards** tab. Add another console/host serial attachment cable, 15M (FC 8121).
5. Click the **Configure** button to return to the main panel.
6. Double-click **I/O Drawer 1** to open its wizard.
7. Click **Next** to open the **Storage** tab. Change the quantity of the 36.4 GB Ultra3 SCSI 10 K RPM disk drives/1 inch - (FC 3158) from two to 10.
8. Click the **Next** button to move to the **Adapter Cards** tab, and then from the **Adapter Category** menu, select **LAN**. Add four of the Gigabit Ethernet–SX PCI-X adapters (FC 5700). Select the **HiPerf** category on the menu and add four of the PCI Gigabit Fibre Channel adapters for 64-bit PCI bus (FC 6228). Add four Fibre Channel converter cables, LC (M) to SC (F), 2M (FC 2456), and two S/390 ESCON Channel PCI adapters (FC 2751). Click the **Configure** button to return to the main panel.
9. Right-click the **I/O Drawer (7040-61D): I/O Drawer 1** in the components list and select **Duplicate**. Since the configuration of I/O drawers 2, 3, and 4 is identical, click the **Configure** button to return to the main panel. Repeat this action to add the third and fourth I/O drawer.

Note: Even if the I/O drawer configuration is different in terms of number and type of adapters, you may want to use this Duplicate function. You just need to change the adapters to the desired configuration after duplication. Make sure the maximum number of adapters per system is not surpassed, otherwise you will get an error message when you click **Configure**.

10. In order to accommodate the four I/O drawers and the batteries, we need an additional rack. Double-click **7040-61R** to open its wizard. Add two rear doors, acoustic, primary or secondary rack (FC 6075). Add one front door (black) for 24 in 2M (42U) racks (FC 6071) and one 42U expansion (FC 8691).
11. Go to the **Additional** tab. On the Power Options list, select one of the integrated battery backups, primary (FC 6200) and one of the integrated battery backup, redundant (FC 6201) features. Click the **Configure** button to return to the main panel. The e-config finds out that more battery backup features are necessary and automatically selects them (one is required for each bulk power regulator in the *front* bulk power assembly). Note that the e-config will automatically configure the expansion rack when necessary.
12. Click the **Messages** tab on the main panel. An error message appears because of a needed feature:

    ```
    Rack Server 1: There must be exactly 1 Remote I/O Loop Adapter (6410) for
    the current configuration
    ```

13. Right-click the message and select **Jump to Wizard**.
14. Go to the **System** tab and add one quad RIO adapter (FC 6410). Click the **Configure** button to return to the main panel.
15. You should now add the required software for the ESCON adapter. Double-click **Software Products (pSeries 690 (7040-681): Rack Server 1)** and select the desired software. Configure the prerequisite software, such as the PCI ESCON Channel Control Unit. You can use this panel to add additional software. Select the software you want and click **Next** several times, and finally **Configure**.
16. In the Catalog list area, double-click **Hardware Management Console,** then double-click **7015-C02 Hardware Management Console**. The HMC wizard appears on the screen.
17. Change to the **Graphics** tab and add or replace the IBM P76/P77 display with one IBM P260/P275 color monitor, stealth black, and cable (FC 3628).
18. Go to the **Input Devices** tab and add one quiet touch keyboard: Stealth black US English, #103P (FC 8800).
19. Click the **Configure** button to return to the main panel.

20. Right-click in the Wizards list on the HMC entry, and select **Duplicate** to add a second identical HMC.
21. Click the **Messages** tab on the main panel and right-click the message:

    ```
    A final validation is required to complete the configuration
    ```

    Then select **Perform Final Validation**.
22. The final graphical representation for this configuration should be similar to Figure 3-23 on page 123.

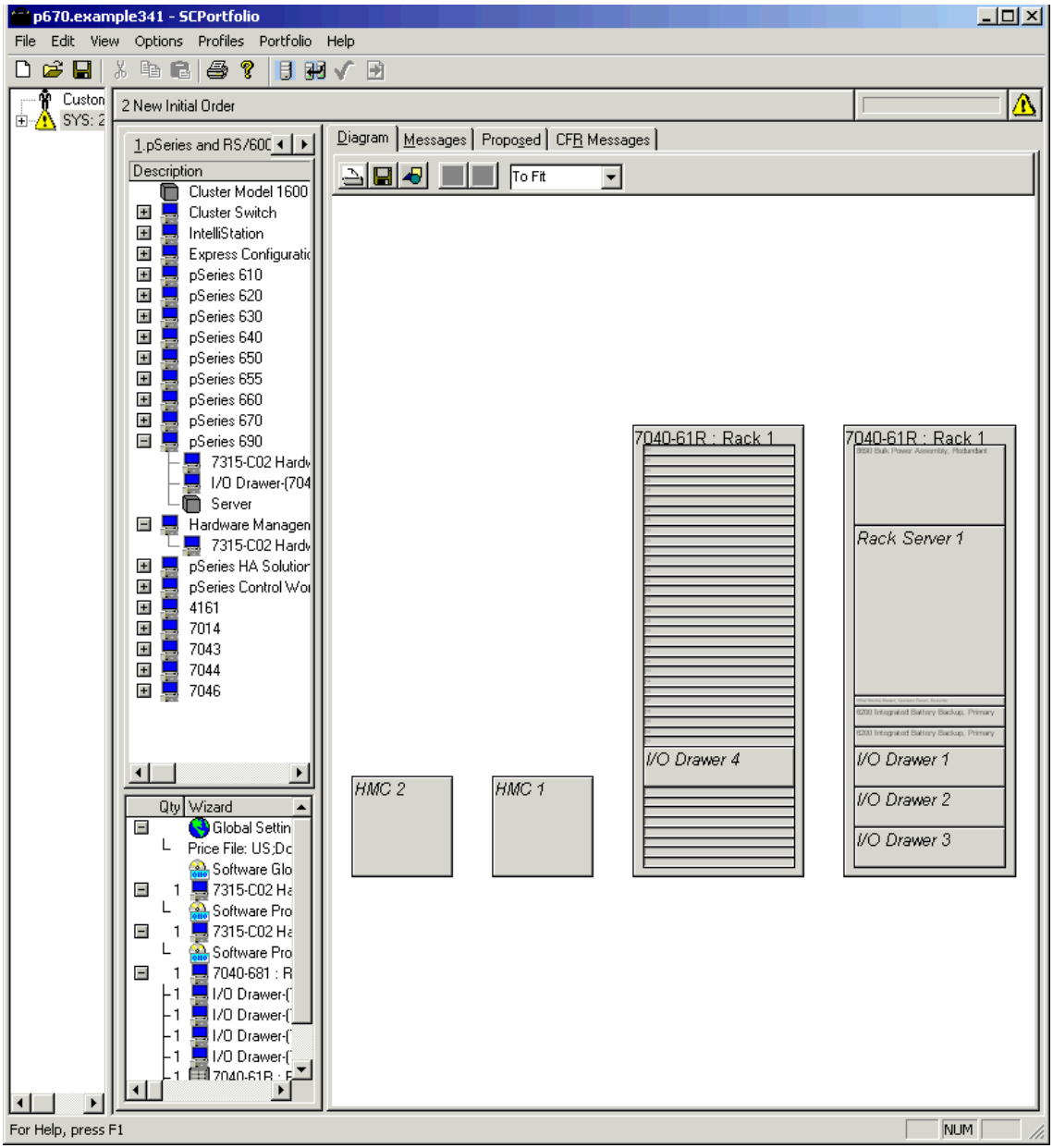


Figure 3-23 A graphical representation of the final configuration

3.4.3 Model conversion from pSeries 670 to pSeries 690

The starting point for this example is the configuration prepared in 3.4.1, “Configuration example 1: pSeries 670 (16-way 1.1 GHz)” .

Rather than using the method described in 3.3.2, “Performing an upgrade” , we assume that the configuration is already loaded in e-config.

The steps involved in the model conversion are:

1. Start upgrading the configuration already loaded in e-config by right-clicking the system in the Portfolios list, and select **Start Upgrade**. A new system is created in the Portfolios list, and automatically displayed in the Configuration view.

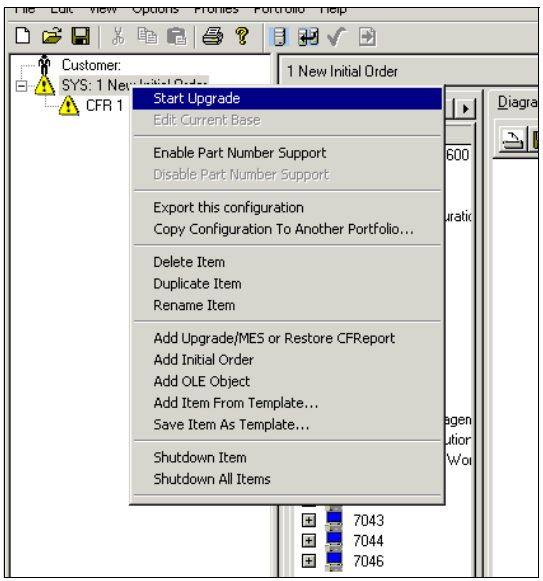


Figure 3-24 Starting to upgrade an existing configuration

2. Right-click the CEC, called Rack Server 1 in the diagram, and select **Upgrade**.

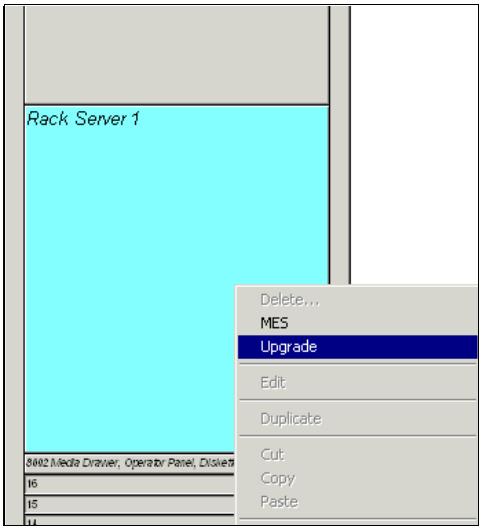


Figure 3-25 Start upgrading the CEC.

The CEC wizard windows pops up, and the **Products** tabs contains a **Product Type** list, where the current model, **7040-671**, is highlighted. Select the **7040-681** type (see Figure 3-26), and click the **Next >** button to change the model.

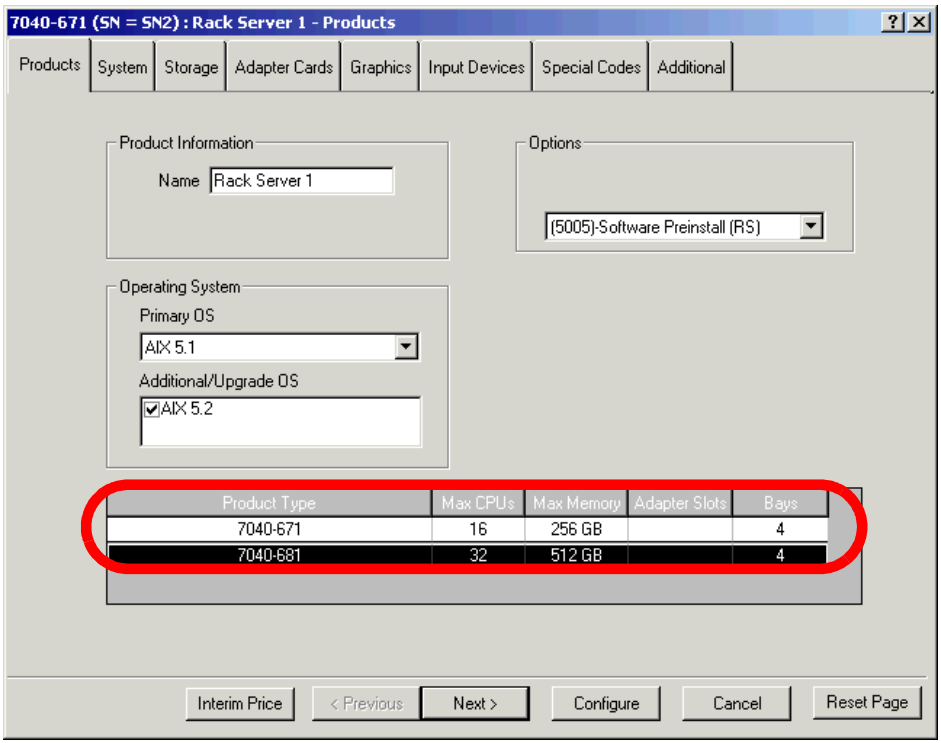


Figure 3-26 Changing model to pSeries 690

3. Click on Configure to return to the Configuration view.
4. Switch to the **Message** view in the **Configuration** window. It contains errors that need to be fixed.

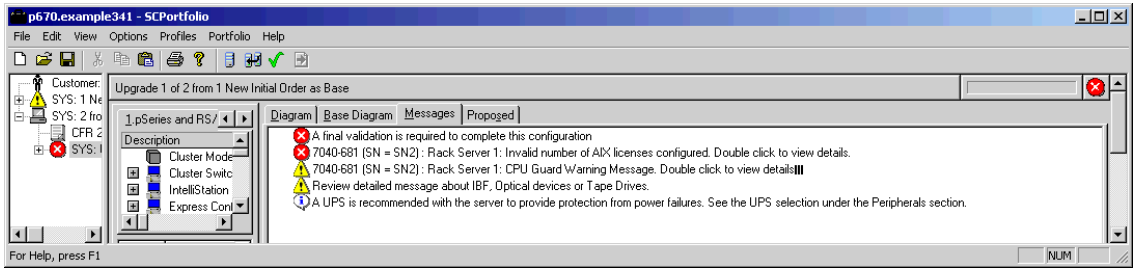


Figure 3-27 Errors in the Messages tab

5. Double-click the error:

   `Each Server 1: Invalid number of AIX licenses configured.`

The detailed message indicates that the number of licenses originally ordered for a 16-way server is no longer valid with a 32-way server. In the Wizards list, right-click the **Software Products** wizard, and select **Edit**.

6. In the Software wizard, click on the **Next >** button, until you are presented with the panel specifying the number of AIX licenses (see Figure 3-28). Add 16 licenses, and click **Configure**. The related error disappears from the Message view.

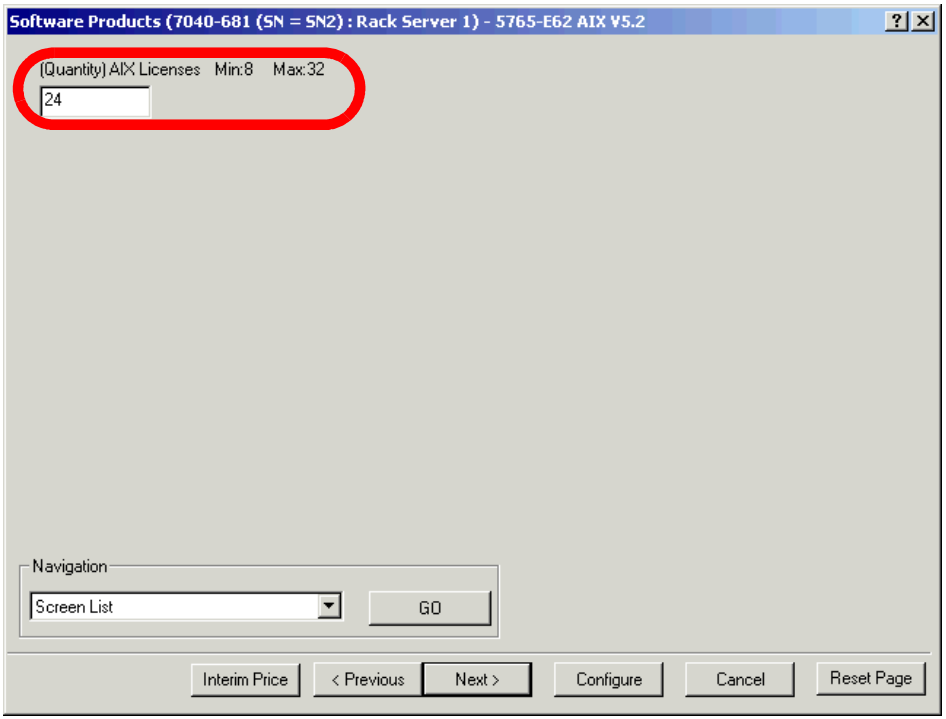


Figure 3-28 Changing the number of AIX licenses

7. You can now perform the final validation and generate the configuration reports.

3.4.4 Feature conversion from POWER4 to POWER4+

The starting point for this example is the configuration prepared in 3.4.2, “Configuration example 2: pSeries 690 (24-way 1.3 GHz)” .

The desired configuration is a pSeries 690 with:

- Four 1.7 GHz MCM

- 128 GB memory
- One RIO drawers and 4 RIO-2 drawers, with contents identical to the content of the drawers in the base pSeries 690 configuration. The RIO drawer is configured in single-loop mode, while the RIO-2 drawers are configured in dual-loop mode.

This example shows that you can include in one configuration report feature conversions (to replace existing processors for example) and MES to add new features.

We assume that the base configuration is already loaded in e-config.

The steps involved in the model conversion are:

1. Start upgrading the base configuration right-clicking the system in the Portfolios list, and select **Start Upgrade** (see the first step in 3.4.3, “Model conversion from pSeries 670 to pSeries 690” for details). A new system is created in the Portfolios list, and automatically displayed in the Configuration view.
2. Right-click the CEC, called Rack Server 1 in the diagram, and select **MES.** The CEC wizard windows pops up, opened on the **Products** tabs.
3. Switch to the **System** tab: In the **Processor** list, change the number of features FC 5245 and FC 5254 to zero, scroll down, and change the number of 1.7 GHz processor (FC 5246) to 4.
4. In the **System Options** list, reset the number of RIO books (FC 6404 and FC 6410) to 0, and select RIO-2 books (FC 6418 and FC 6419): two secondary books are sufficient to connect the 9 I/O loops.
5. In the **memory** list, reset the number of 16 GHz memory card to zero (FC 4183 and FC4184), and select eight 567 MHz memory cards (FC 4484 and FC 4485).
6. Click **Configure** to return to the Configuration view.
7. I/O Drawer 1 must remain in RIO mode. It cannot be migrated to RIO-2, because it contains a SCSI adapter and two ESCON adapters which are not supported by the RIO-2 planar.
8. Right-click on the I/O Drawer 2, and select **MES**. The Drawer wizard windows pops-up.
9. Switch to the **Storage** tab, check the **Prefer Dual Loop** checkbox, reset the RIO planar (FC 6563) to zero, and select Two RIO-2 planar (FC 6571), and click **Configure**.
10. Repeat Step 9 with I/O drawers 3 and 4.

11. In the Wizards window, right-click on the I/O drawer 4, and select duplicate. There are two entries for I/O drawer 4: one with a 1 on its left, one with a 0. Select the one preceded by a 1, which corresponds to the updated drawer.

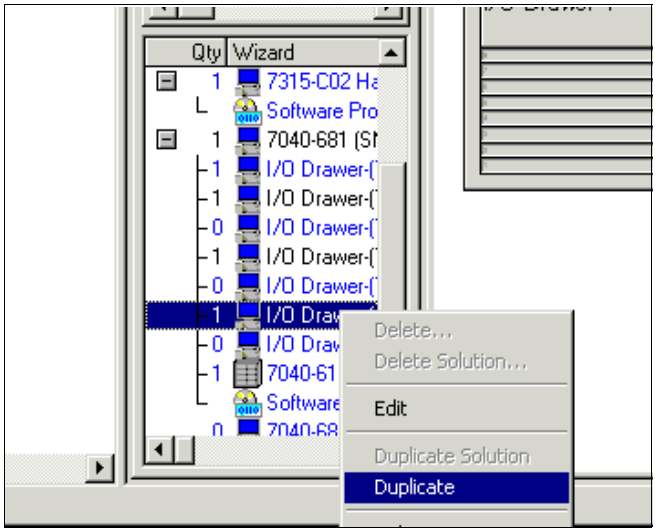


Figure 3-29 Duplicating I/O drawer 4

12. The I/O drawer wizard presents the configuration for I/O drawer 5. Click Configure to create a new drawer identical to drawer 4.

13. If you display the message view in the Configuration area, you may notice that e-config has configured features conversions for the replacement of the MCM and Memory cards.

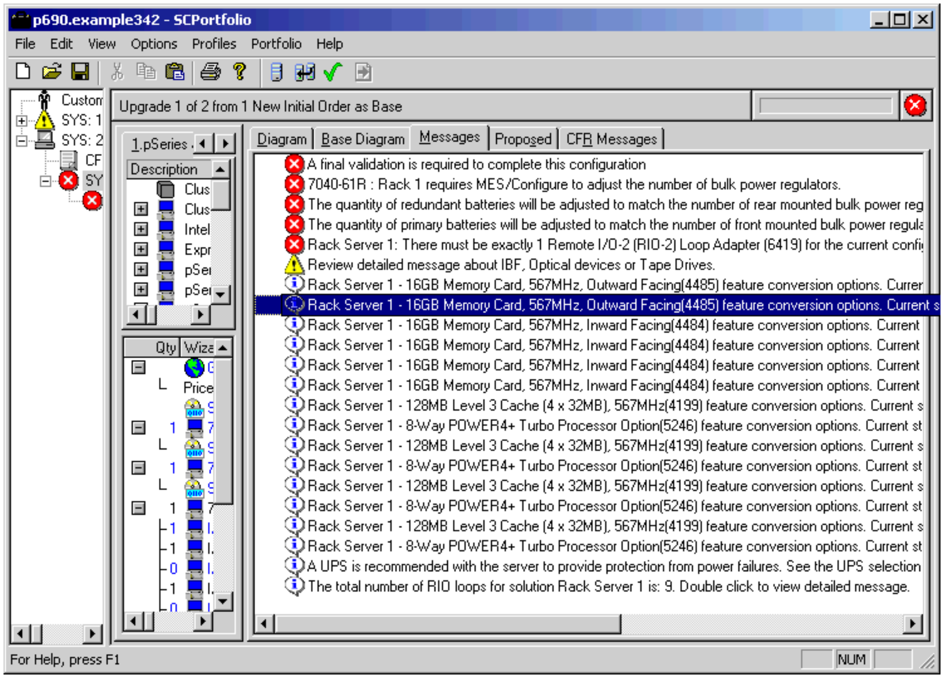


Figure 3-30   Messages view during feature upgrade

14. The message window also contains errors that need to be fixed before you can generate a configuration report.

15. Right-click the error:

    ```
    Rack 1 requires MES/Configure to adjust the number of bulk power
    regulators.
    ```

    Select **MES and jump to Wizard**, and click **Configure**. e-config adjusts the number of Bulk Power features and the number of IBF features. The related error message disappears from the message view.

16. You can now perform the final validation and generate the configuration reports.

17. If you select the Proposal view, and scroll down to the list of replaced, removed and added features, you may notice that e-config has taken care of the L3 cache, Pass-though modules, Bulk power and cables features of the CEC and I/O drawers (see lines in **bold** in Example 3-1).

Example 3-1 Resulting configuration changes for CEC features

```
Removals
    5257  PROC.BUS PASS THROUGH MOD.                      -1
    6404  SUPP.PROC.REM. I/O LOOP ATT.,2                  -1
    6410  REMOTE I/O LOOP AD,4 LOOP                       -1

Additions
    4485  16GB Memory Card, 567MHz, Outward Facing         2
    5246  8-Way POWER4+ Turbo Processor Option             1
    6184  POWER C.GROUP,4TH PROC.MOD.                      1
    6189  DC POWER CONV.,CEC,ADD                           1
    6202  Power Cable Group, CEC to Power Controller,      1
    6418  Support Processor with Remote I/O-2 (RIO-2)      1
    6419  Support Processor with Remote I/O-2 (RIO-2)      2
    4199  128MB Level 3 Cache (4 x 32MB), 567MHz          1

............

Removals
    3149  REMOTE I/O CABLE, 2 M                           -2
    6563  I/O DRAWER PCI PLANAR,10 SLOT                   -2

Additions
    3156  RIO-2 (Remote I/O-2) Cable, 1.75M                4
    6571  I/O Drawer PCI-X Planar, 10 Slot, 2 Integrat     2
```

4

Capacity Upgrade on Demand

The pSeries 670 and pSeries 690 systems can be shipped with non-activated resources (processors and/or memory), which may be purchased and activated at a future point in time without affecting normal machine operation. This ability is called Capacity Upgrade on Demand (CUoD) and provides flexibility and improved granularity in processor and memory upgrades.

This chapter includes the following topics:

Note: This chapter covers the CUoD functions available as of May 2003. Enhancements and changes may appear in the future.

4.1 What's new in CUoD

With the May 2003 announcement, there are several new CUoD enhancements to the CUoD systems:

- In addition to the processor CUoD capability for the 1.1GHz pSeries 670, 1.1 and 1.3 GHz pSeries 690 — the processor CUoD capability for new 1.5 GHz pSeries 670, 1.5 and 1.7 GHz pSeries 690 is introduced and described in 4.2.3, "Supported CUoD Processor configurations" on page 138.
- Memory CUoD capability is introduced to all the pSeries 670 and pSeries 690 systems, providing flexibility and fine-granularity for memory for future growth. The supported configuration is described in 4.2.4, "Supported CUoD Memory configurations" on page 141.
- To improve the response time to unpredictable increased workload, the Trial CoD feature is introduced to the CUoD systems for processors and memory. This feature allows immediate activation of the non-activated resources on CUoD systems, which is described in 4.2.1, "Trial CoD function" on page 135. Section 4.3.2, "Trial CoD processor and memory" on page 151 also describes how to use this feature from the HMC.
- To cater for periodical peak workload, such as monthly batch jobs, which the systems are likely required for additional resources for a short period, there is a new feature called On/Off Capacity on Demand (On/Off CoD), which is described in 4.5, "On/Off Capacity on Demand" on page 155.

Attention: On/Off Capacity on Demand for pSeries is not announced as an orderable product when this redbook was published. The contents in 4.5, "On/Off Capacity on Demand" on page 155 are subject to change.

4.2 Description of CUoD

This section describes the functionality of Capacity Upgrade on Demand (CUoD) for processors and memory on the pSeries 670 and pSeries 690 systems. CUoD systems can be configured with non-activated resources (processors and/or memory). These non-activated resources can be enabled dynamically and non-disruptively to the systems without the need to schedule downtime for the engineer to install additional resources through purchased activation codes at a certain point in time. It provides flexibility and improved granularity in processor upgrades and is very useful to anticipate future growth.

The design is based on the use of a CUoD capacity card that has a smart chip to provide the required high level of asset protection for the CUoD function. This capacity card is placed in a slot in the primary RIO book (FC 6404) or primary

RIO-2 book (FC 6418). This CUoD capacity card tracks activated CUoD resources.

A new feature called *Trial CoD* is introduced to provide immediate activation of the CUoD resources temporarily 30 days without the need for activation code. This is to allow customers to use the resources while they are ordering the activation codes from IBM. The additional temporary resources can either be configured as permanent resources later with the activation code or returned back to the system as CUoD resources.

We also describe a feature associated with CUoD, called Dynamic Processor Sparing in 4.4, “Dynamic Processor Sparing” on page 154. It provides protection against a possible processor failure, by activating a processor from the pool of non-activated CUoD processors.

The information at the following URL briefly explains the CuOD process:

http://www.ibm.com/servers/eserver/pseries/cuod/

4.2.1 Trial CoD function

Trial CoD is a new feature for CUoD systems. With other CUoD offerings, customers with CUoD systems must purchase the activation codes from IBM before the non-activated CUoD resources can be activated to meet the increase in workload. With this Trial CoD feature, customers can now activate the required non-activated CUoD resources immediately and after that, proceed to purchase those resources from IBM.

The following basic rules apply for a pSeries 670 or pSeries 690 system with Trial CoD feature:

- After the CUoD resources are activated through Trial CoD function, a customer must either buy part or all of the activated CUoD resources from IBM as described in 4.3.1, “CUoD resources activation and order process” on page 147 or return the activated CUoD resources back to the system within 30 days.
- The Trial CoD function for processors and memory are activated separately.
- The customer is allowed one use of Trial CoD:
  - When the system first boots up, or
  - After the customer has purchased the activated CUoD resources which were previously activated through Trial CoD function

There are several advantages of using Trial CoD:

- It improves the response time to meet unpredictable increase in workload.

- A customer can evaluate the performance of the system after activating the CUoD resources before placing the order for the activation codes.

4.2.2 Overview of CUoD configurations

The following basic rules apply for a pSeries 670 or pSeries 690 system with the CUoD feature:

- A CUoD system requires a CUoD capacity card to be placed in a slot in the primary RIO book (FC 6404) or primary RIO-2 book (FC 6418) for 1.1 and 1.3 GHz systems.
- A CUoD system requires a CUoD capacity card to be placed in a slot in the primary RIO-2 book (FC 6418) for 1.5 and 1.7 GHz systems.
- A system without CUoD resources does not require a CUoD capacity card.
- For Full System Partition systems, activation of processors and/or memory will require a reboot.
- MCMs can be either fully activated MCMs or four out of eight-way active CUoD MCMs (see 4.2.6, “Logical and physical entities” on page 143 for further information about how the system activates processors). Additional two-way activations may be purchased from IBM and activated on these CUoD MCMs by entering a activation key on the HMC.
- Affinity partitioning is not supported on CUoD systems.
- CUoD resources are not supported on Linux-only systems at this point.
- CUoD is not supported on HPC systems.
- On partitions that support DLPAR, activation of CUoD resources is supported without the need for a reboot of the partitions.
- A CUoD capacity card is required to be present and functional to boot the system if any CUoD processors are present in the system. If a CUoD memory resource is present without a CUoD capacity card, the system will deconfigure the CUoD memory card at boot time, but will allow the system to boot if there is non-CUoD memory cards.
- A CUoD capacity card from another system will allow the system to be rebooted, but none of the CUoD functions (screens, activations) can be used. An error message will be sent to the HMC to indicate that the system CUoD functions and supporting screens are not available.
- CUoD resources can be moved from one CUoD system to another CUoD system; however, activation entitlement stays with the serial number of the purchase.
- Dynamic Processor Sparing is available only on systems running on dynamic logical partitioning (DLPAR).

- To activate CUoD resources, a key called the *activation code* has to be entered on the HMC. The maximum number of invalid activation code entries is five. If more than five activation codes are entered, the system must be rebooted before another activation code can be entered. Some types of failures are not counted against the five invalid attempts, including:
  - User mis-typed the activation code. (There are unique features in the activation codes that are checked.)
  - Previous activation code for the same system.
  - Activation code from a different system that contains a different processor (speed) type.
  - Activation code for memory from a different system is entered and this system does not contain any CUoD memory cards.
- For activation codes for CUoD processors, the following microcode levels, HMC software levels and operating system versions (with fixes) are required (including Full System Partition):

  Note: This applies only for the POWER4 1.1 and 1.3 GHz processor systems.

  - 7040 - 671/681 at 10/2002 system microcode update or later
  - HMC for pSeries with V1.3 or later
  - AIX 5L Version 5.1 with 5100-03 Recommended Maintenance Level with APAR IY36013 or later

    Note: The AIX 5L Version 5.1 partition (including Full System Partition) would have to be rebooted to take advantage of the additional processors.

  - AIX 5L Version 5.2
- For Trial CoD of CUoD processors, the following microcode levels, HMC software levels and operating system versions (with fixes) are required (including Full System Partition):
  - 7040 - 671/681 at 05/2003 system microcode update or later
  - HMC for pSeries with V1.3.2 or later
  - AIX 5L Version 5.1 with 5100-04 Recommended Maintenance Level with APAR IY39795 and APAR IY36771 or later
  - AIX 5L Version 5.2 with 5200-01 Recommended Maintenance Level with APAR IY39795 and APAR IY36772 or later

- For activation codes for CUoD memory and Trial CoD of CUoD memory, the following microcode levels, HMC levels and operating system versions (with fixes) are supported (including Full System Partition):
  - 7040-671/681 at 05/2003 system microcode update or later
  - HMC for pSeries with V1.3.2 or later
  - AIX 5L Version 5.2 with 5200-01 Recommended Maintenance Level with APAR IY39795 and APAR IY36772 or later

Note: For HMC for pSeries with V1.3.2, the first number represents version, the second number is the release, and the third number is modification.

Important: The small RMO option must be set to boot AIX 5L Version 5.2 when you create an LPAR from the HMC.

4.2.3 Supported CUoD Processor configurations

Processors are activated in pairs because of the dual core technology of POWER4 and POWER4+ processors and to optimize system performance and maintain consistent, predictable performance. A complex algorithm in the system firmware, based on such things as memory groups, available inactive processors, and possible processor failure or deallocation situations, determines which pair of processors will be activated. In normal circumstances, both processor cores in a single chip will be activated. But there are circumstances in which that may not be possible.

Note: This activation of pairs of processors is also true for On/Off CoD and Trial CoD.

Table 4-1 on page 138 is an overview of all CUoD processor feature codes for the pSeries 670 and pSeries 690.

Table 4-1 CUoD processor feature codes

FC	Description
pSeries 670	
7402	1.1 GHz 4/8-way CUoD MCM + 128 MB (4x 32) L3 cache
7412	1.1 GHz CUoD 2-way Activation for FC 7402
7400	1.5 GHz 4/8-way CUoD MCM + 128 MB (4x32) L3 cache

FC	Description
7410	1.5 GHz CUoD 2-way Activation for FC 7400
pSeries 690	
7403	1.1 GHz 4/8-way CUoD MCM
7413	1.1 GHz CUoD 2-way Activation for FC 7403
7404	1.5 GHz 4/8-way CUoD MCM
7414	1.5 GHz CUoD 2-way Activation for FC 7404
pSeries 690 Turbo	
7405	1.3 GHz 4/8-way CUoD MCM
7415	1.3 GHz CUoD 2-way Activation for FC 7405
7406	1.7 GHz 4/8-way CUoD MCM
7416	1.7 GHz CUoD 2-way Activation for FC 7406

Table 4-2 shows possible CUoD processor combinations for 1.1 and 1.3 GHz systems.

Under the MCM columns:

- The numbers that are not emphasized (for example, 4) indicates the number of processors that are installed and activated on the MCM.
- The numbers that *are emphasized* (for example, *2*) indicates the number of processors that are installed, but can either be activated upon initial installation of the system or activated through the CUoD activation process on the MCM in future.

Table 4-2 Supported 1.1 and 1.3 GHz CUoD processor combinations

Model & System	CUoD Range	MCM 0	MCM 2	MCM 1	MCM 3
pSeries 670 16-way	12 to 16	8	4 / *2* / *2*		
pSeries 690 16-way	12 to 16	8	4 / *2* / *2*		
pSeries 690 24-way	16 to 24	8	4 / *2* / *2*	4 / *2* / *2*	
	20 to 24	8	8	4 / *2* / *2*	

Model & System	CUoD Range	MCM 0	MCM 2	MCM 1	MCM 3
pSeries 690 32-way	20 to 32	8	4 / *2* / *2*	4 / *2* / *2*	4 / *2* / *2*
	24 to 32	8	8	4 / *2* / *2*	4 / *2* / *2*
	28 to 32	8	8	8	4 / *2* / *2*

Table 4-3 shows possible CUoD processor combinations for 1.5 and 1.7 GHz systems.

Table 4-3 Supported 1.5 and 1.7 GHz CUoD processor combinations

Model & System	CUoD Range	MCM 0	MCM 2	MCM 1	MCM 3
pSeries 670 16-way	8 to 16	4 / *2* / *2*	4 / *2* / *2*		
	12 to 16	8	4 / *2* / *2*		
pSeries 690 16-way	8 to 16	4 / *2* / *2*	4 / *2* / *2*		
	12 to 16	8	4 / *2* / *2*		
pSeries 690 24-way	12 to 24	4 / *2* / *2*	4 / *2* / *2*	4 / *2* / *2*	
	16 to 24	8	4 / *2* / *2*	4 / *2* / *2*	
	20 to 24	8	8	4 / *2* / *2*	
pSeries 690 32-way	16 to 32	4 / *2* / *2*	4 / *2* / *2*	4 / *2* / *2*	4 / *2* / *2*
	20 to 32	8	4 / *2* / *2*	4 / *2* / *2*	4 / *2* / *2*
	24 to 32	8	8	4 / *2* / *2*	4 / *2* / *2*
	28 to 32	8	8	8	4 / *2* / *2*

Example 1

We configure a 1.1 GHz 12-way pSeries 670 with twelve processors by configuring one FC 5256 and one FC 7402. This leaves room to order two FC 7412s to reach the maximum processor configuration.

Example 2

We configure a 1.7 GHz 16-way pSeries 690 with sixteen CUoD processors by configuring four FC 7406 (4 out of 8-way operational MCM). This leaves room to

order eight FC 7416 (2-way processor activation) to reach the maximum 32-way configuration.

Note: The two-way CUoD activations may also be configured in the initial order to form additional configurations. For example, a 26 out of 32-way 1.1 GHz pSeries 690 can be configured by selecting one FC 5242, one FC 5252, two FC 7403s, and one FC 7413.

4.2.4 Supported CUoD Memory configurations

Table 4-4 is an overview of all CUoD memory feature codes for the pSeries 670 and pSeries 690. There are some basic rules for these CUoD memory combinations:

- All CUoD memory card must be in identical pairs, that is, 2 x FC 7050 or 2 x 7051.
- All CUoD memory activation must be in identical pairs, that is, 2 x FC 7060 or 2 x FC 7061.
- Configurations consisting of one CUoD memory card and one non-CUoD memory card are not supported.

Table 4-4 CUoD memory feature codes

FC	Description
pSeries 670	
7050	16 GB 567 MHz CUoD memory card, inward facing, 8 GB active
7054	32 GB 567 MHz CUoD memory card, inward facing, 16 GB active
7060	4 GB CUoD activation for FC 7050 and FC 7051
7061	4 GB CUoD activation for FC 7054 and FC 7055
pSeries 690	
7050	16 GB 567 MHz CUoD memory card, inward facing, 8 GB active
7051	16 GB 567 MHz CUoD memory card, outward facing, 8 GB active
7054	32 GB 567 MHz CUoD memory card, inward facing, 16 GB active
7055	32 GB 567 MHz CUoD memory card, outward facing, 16 GB active
7060	4 GB CUoD activation for FC 7050 and FC 7051
7061	4 GB CUoD activation for FC 7054 and FC 7055

The conventional memory configuration table in 2.3.2, “Memory subsystem for pSeries 690” on page 29 applies to CUoD memory cards. Utilize maximum card capacity when applying the configuration rules to CUoD memory cards:

- Partially activated 16 GB CUoD memory cards apply 16 GB configuration rules.
- Partially activated 32 GB CUoD memory cards apply 32 GB configuration rules.

Example 1

We configure a two MCMs pSeries 670 with 32 GB memory by configuring four FC 7050. This leaves room to order eight FC 7060 to reach the maximum memory configuration of 64 GB.

Example 2

We configure a four MCMs pSeries 690 with 160 GB memory by configuring four FC 7054 (16 of 32 GB active inward memory), two FC 7055 (16 of 32 GB active outward memory) and two FC 4487. This leaves room to order sixteen FC 7060 and eight FC 7061 to reach the maximum memory configuration of 256 GB.

Note: The 4 GB CUoD activations may also be configured in the initial order to form additional configurations. For example, a two MCM pSeries 690 with 40 GB out of a maximum of 64 GB can be configured by selecting four FC 7050 and two FC 7060.

4.2.5 CUoD resource sequencing

Once processors and memory from the CUoD resources are activated, an algorithm is used to enable the physical resources in a specific order to ensure the following characteristics:

- Activations are not tied to actual physical CUoD resources.
- All processor activations will be in full chip increments (this means two processors at a time on the same chip).
- Hot sparing may cause performance imbalance within the system since two chips will have only a single core used.
- Processor will be used based upon best performance tuning rather than physical relationship to CUoD resources. The entitled resources used will be selected independently of whether the resource is from a CUoD component or a non-CUoD component.

4.2.6 Logical and physical entities

From an ordering point of view it might seem that a base MCM is always fully operational and a CUoD MCM is initially always four out of eight-way operational. This is the logical view of CUoD. However, whichever specific physical entities will be activated, CUoD is completely determined by the sequencing rules. Since these rules do not consider the type of the resource (a base resource or CUoD resource), there will be situations where specific base resources are not activated, and specific CUoD resources are activated though not purchased.

Also, the technical implementation of CUoD in firmware is implemented such that all theoretical configurations are covered (not only the recommend balanced memory and L3 cache configurations, and dual processor activations). Secondly, the technical implementation is not only meant for the pSeries 670 or pSeries 690, but also for future pSeries machines, possibly with finer CUoD granularity.

On actual shipping of pSeries 670 and pSeries 690 systems, memory guidelines and supported configurations (as described in this book and in Table 2-6 on page 44) are followed, therefore, only balanced memory configurations will exist and actual CUoD activations and sequencing rules will be a subset of those mentioned in 4.2.5, “CUoD resource sequencing” on page 142.

Let us clarify these rules using the following two examples:

In both cases we consider a 16 out 24-way operational pSeries 690 configuration (three MCMs: one base MCM and two CUoD MCMs). We assume that the memory configuration is balanced for all MCMs, as this is recommended for practically all configurations.

Example 1

Total system memory is 40 GB: 8+8, 8+8, and 4+4 for MCM 0, 2, and 1, respectively. Due to sequencing of large balanced memory groups, all processors on MCM 0, and subsequently on MCM 2, are active, and none are active on MCM 1. The active processors will see all the memory, including the memory attached to MCM 1, totalling 40 GB.

Example 2

Total system memory is 16 GB: 4+4, 4+4, and 0+0 for MCM 0, 2, and 1, respectively. Again all processors on MCM 0 and 2 will be active. Now, if an activation code for two CUoD processors is entered, they will become active on MCM 1, which has no memory attached. All 18 processors will use the 16 GB memory of MCM 0 and 2.

4.2.7 CUoD license screen

Upon the first boot of the pSeries 670 or pSeries 690 CUoD systems, the “Click to Accept” message is presented to the system administrator, and it should be accepted for the initial boot of this system. The “Click to Accept” message will be shown on the HMC graphical user interface as shown in Figure 4-1 on page 144.

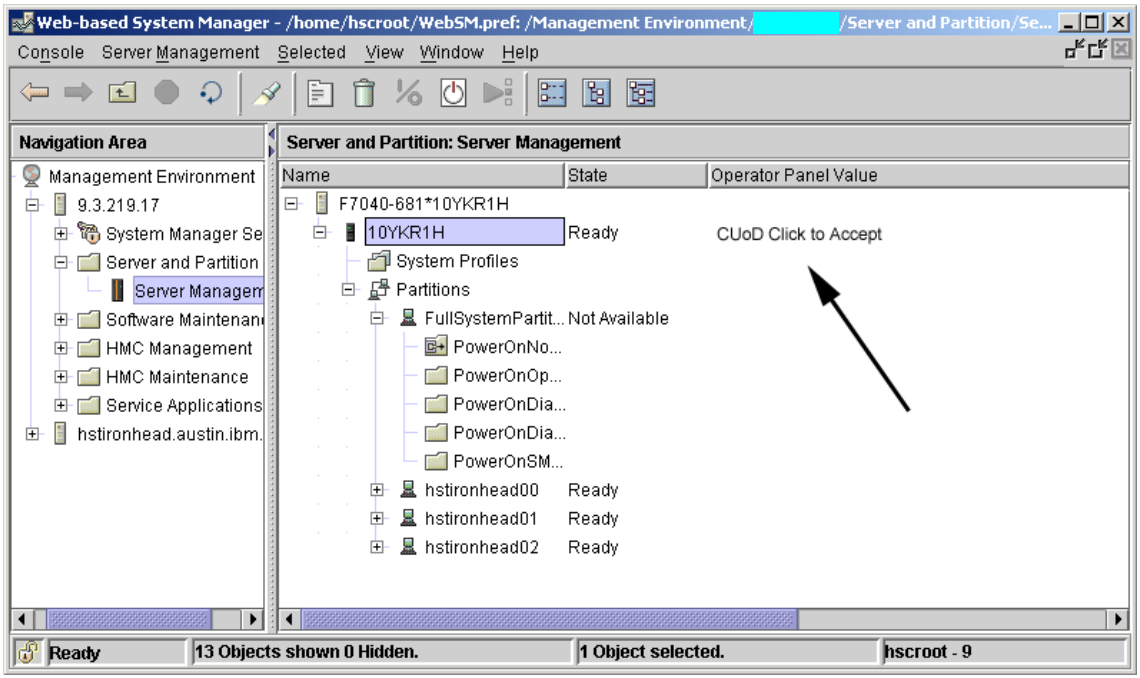


Figure 4-1 Click to Accept on the HMC at initial boot[1]

The CUoD License function works as explained in the following:

1. On every boot, system firmware looks for a Click to Accept (CTA) accepted bit stored in the CUoD capacity card that indicates that the CUoD Click To Accept screen has displayed and the license conditions accepted. If the bit is set, the system will continue the boot. If the CTA accepted bit is not set, `A100C2AC` will be displayed in the operator panel, and firmware will contact the HMC and request that the “Click to Accept” screen be displayed. Firmware will then wait for the HMC to return with the status that the Click To Accept license conditions have been accepted. Once accepted the system will continue the boot.
2. There are two possible responses to accept in the “Capacity Upgrade on Demand License Agreement” window (see Figure 4-2):

[1] The host name of the HMC is hidden in the title bar.

- If **Do not show this information again** is selected and accepted, the boot will continue and the HMC will not display the CUoD Click To Accept screen again.
- If **Do not show this information again** is not selected and accepted, the boot will continue, but the CUoD Click To Accept screen will be displayed on the HMC again on the next reboot.

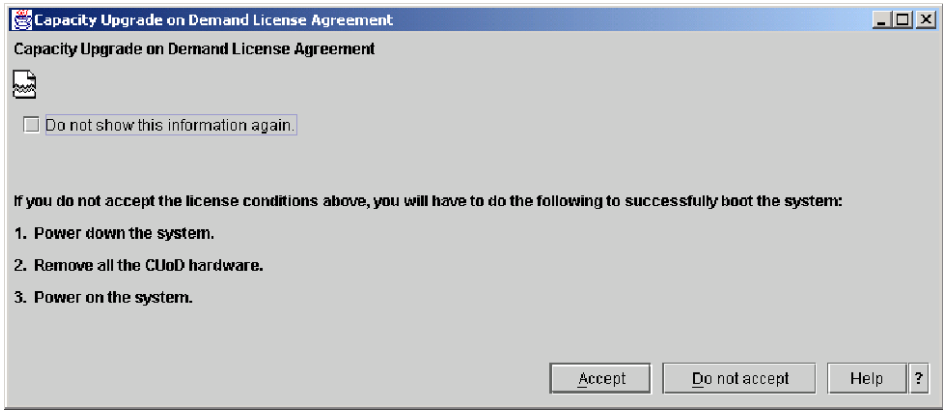


Figure 4-2 Click to Accept window

Note: If you click the **Do not accept** button, the boot process will not continue.

3. At boot time, the system firmware checks if the system power source (wall plug) has been removed as part of a controlled shutdown prior to the boot. If power had been removed, system firmware will present the “Click To Accept” screen on the next boot. This will cover the case that the system may have been sold or transferred and is now being booted by a new customer. The “Click To Accept” screen will not be displayed again if power had been removed due to a power outage.

A service processor menu to reset the CUoD Click To Accept settings is also implemented.

4.2.8 CUoD error messages

The system firmware will detect some error conditions concerning CUoD. The HMC will indicate several errors (as listed in Table 4-5) to the operator in case of CUoD failure situations.

When a user receives out-of-compliance messages from a CUoD system, he can either shutdown one or more partitions containing sufficient processors/memory or to return the exact amount of processors/memory by using dynamic logical partitioning to get the system back in compliance.

Table 4-5 CUoD error codes and messages

Error code	Description
1	Processor capacity activation failed.
2	Memory capacity activation failed.
3	Processor capacity activation code is not valid.
4	Memory capacity activation code is not valid.
5	Processor capacity activation code is not valid, next failed attempt will require a re-boot.
6	Memory capacity activation code is not valid, next failed attempt will require a re-boot.
7	Processor capacity activation code is not valid, a re-boot is required to make another attempt.
8	Memory capacity activation code is not valid, a re-boot is required to make another attempt.
9	Activation Code entered incorrectly, key check error. This is most likely a keying error and the user would be instructed to please re-enter the Activation Code.
10	CUoD capacity card detected from another system or has corrupted data. Contact service provider to replace the card.
11	Request not allowed - Trial CoD function selected had been used.
12	Request not allowed - CUoD capacity card is failing or has corrupted data. Contact your service provider to replace the card.
13	Request not allowed - More processors were selected than are available for immediate activation.
14	Request not allowed - More memory was selected than is available for immediate activation.

4.3 Activating CUoD resources

Here we describe these activation processes on a pSeries 670 or pSeries 690 CUoD system:

4.3.1 CUoD resources activation and order process

To utilize the CUoD resources (processors and memory) on a pSeries 670 or pSeries 690 CUoD machine permanently, the following steps are required:

1. A customer gathers system-specific data from the HMC, as follows:
   a. When you have a CUoD-managed system, the Capacity Upgrade on Demand menu will appear on the HMC when you right-click the managed system as shown in Figure 4-3.

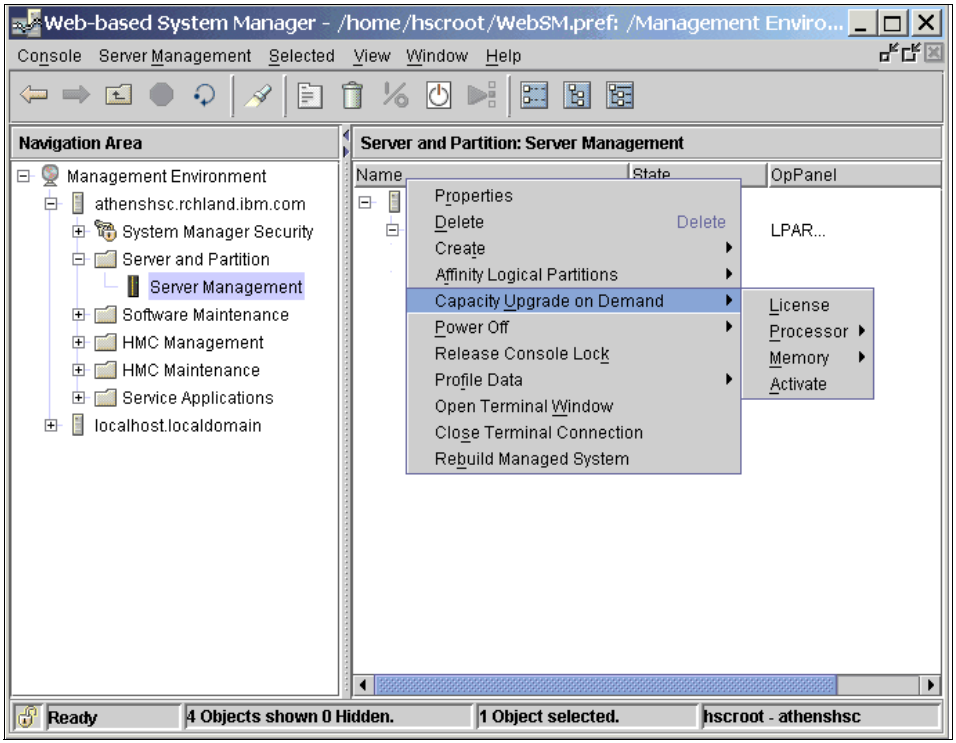


Figure 4-3 CUoD menu on the HMC

   b. On the managed system, select Capacity Upgrade on Demand **->** **Processor** or **Memory** -> **Processor Capacity Settings** or **Memory Capacity Settings** and a menu displaying the resources on the system will appear (see Figure 4-4).

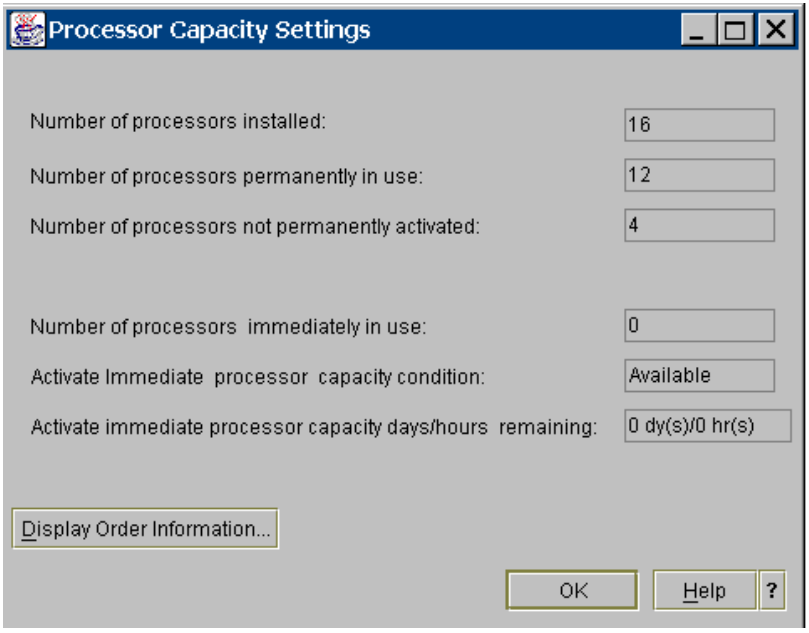


Figure 4-4 Processor Capacity Settings

c. Select **Display Order Information** and the Save Processor or Memory Order Information will appear (see Figure 4-5).

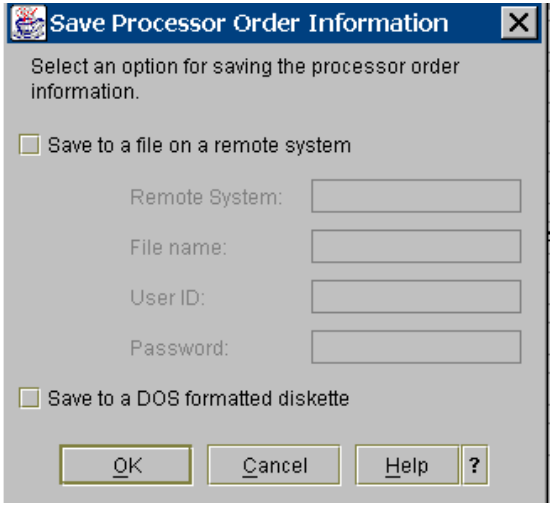


Figure 4-5 Save Processor Order Information

d. Save the processor or memory order information either to a remote system or a diskette.

2. This information can be sent to an IBM Sales Representative using this URL (Figure 4-6):

   http://www.ibm.com/servers/eserver/pseries/cuod/vpd_form.html

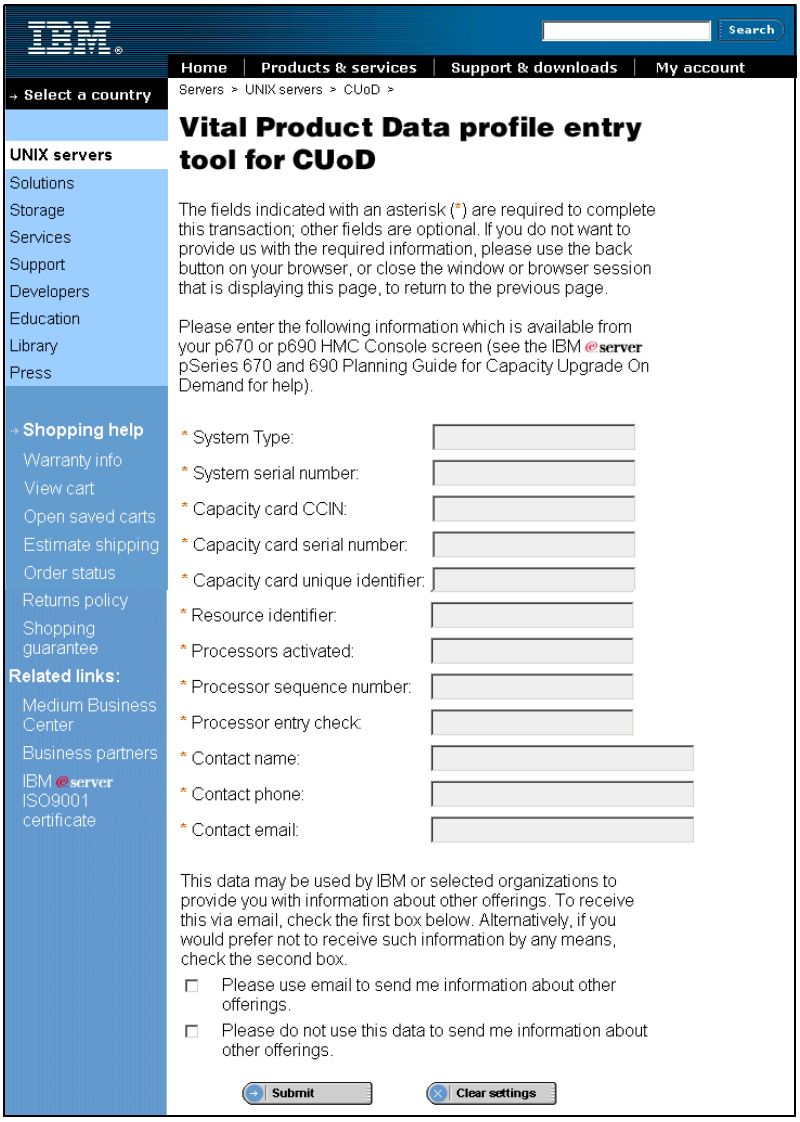


Figure 4-6 Send CUoD and system data

Or, this information can be sent using Service Agent to send CUoD and system data electronically to the IBM Machine Reported Product Data (MRPD) database. Service Agent is the preferred method to gather the data and send it to IBM.

3. The customer places an MES/Upgrade order to a IBM sales representative who will then places the CUoD MES/Upgrade for activation via the e-config. This order process will produce a billing.
4. The CUoD application continually looks for CUoD orders, and when an order comes in it will look for CUoD data input either from the pSeries CUoD URL or from the MRPD database, and will put system CUoD data into the CUoD application. The CUoD application will request the CUoD activation code as ordered from the CUoD activation code tool. The CUoD activation code tool will return the CUoD activation code to the CUoD application.

Note: Step 4 is an IBM internal process.

5. The generated CUoD activation code, a unique code with a system of 34 characters, will be posted to a Web site where the customer can access, and it will also be printed by IBM manufacturing and sent by conventional mail to the customer.
6. When the customer receives the activation code, he or she needs to do the following steps on the HMC to activate the resources:
   a. On the CUoD system that you want to activate the CUoD resources, right-click the managed system and select Capacity Upgrade on Demand -> **Activate**, and a menu will appear (see Figure 4-7).

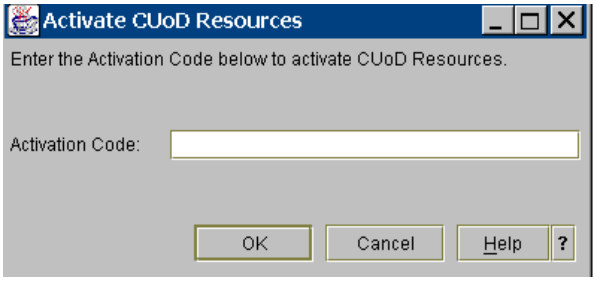


Figure 4-7 Activate CUoD Resources

   b. Enter the CUoD activation code.
7. The activation code is passed to the service processor for authentication. This includes error checking on the key check value and sequence number, decryption of the data, and further error checking to determine that the activation code is for this system.

8. The service processor stores the data in the CUoD capacity card.
9. When the system is running the Full System Partition, a reboot is required to activate the additional resources.
10. When the system is running a partitioned environment, the service processor notifies the system firmware and HMC about the change in licensed status of the additional resources. The customer, using existing HMC screens, decides how to allocate the additional resources to partitions. The normal dynamic logical partitioning (DLPAR) procedure accomplishes the CUoD task as a sequence of basic resource add or remove operations. For partitions that are not running dynamic logical partitioning (DLAPR), allocation of additional resources will require a reboot of the partitions.

The following is a list of the pSeries CUoD data that will be obtained from the HMC and should be faxed by the IBM representative to the CUoD administrator or is electronically sent to the MRPD database for input to the CUoD activation code generation tool:

- System type (four ASCII characters)
- System serial number (eight ASCII characters: pp-sssss)
- CUoD capacity card CCIN (four ASCII characters)
- CUoD capacity card serial number (10 ASCII characters: pp-sssssss)
- CUoD capacity card unique ID (16 ASCII characters)
- CUoD resource identifier (four ASCII characters)
- Activated CUoD function: CUoD increments in use, resources currently paid for (four ASCII characters)
- CUoD sequence number (four hexadecimal characters)
- CUoD activation code entry check (1 byte hex check sum, two ASCII characters—based on items: 1, 2, 3, 4, 5, 6, 7, and 8)

This information is required to generate the activation code (see Figure 4-6 on page 149).

4.3.2 Trial CoD processor and memory

In this section, we describe the necessary steps to activate the CUoD resources (processors and memory) using Trial CoD function on the HMC:

1. On the HMC, right-click the managed system and select Capacity Upgrade on Demand -> **Processor** -> **Processor Activate Immediate** for processors, or select Capacity Upgrade on Demand -> **Memory** -> **Memory Activate Immediate** for memory as shown in Figure 4-8.

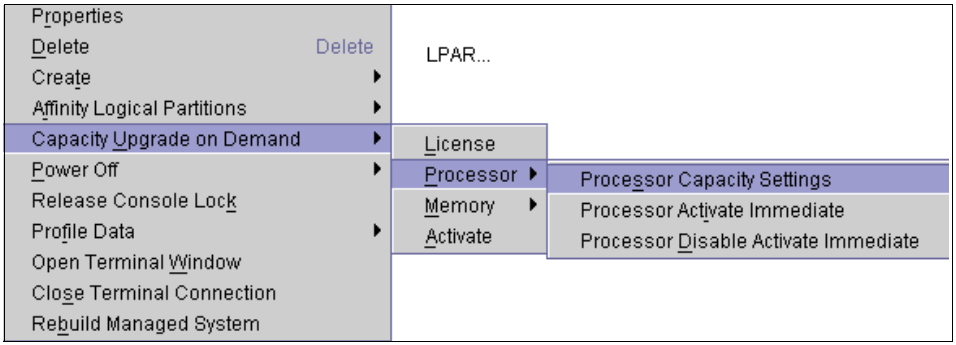


Figure 4-8 To activate the resources using Trial CoD function

2. A menu will appear as shown in Figure 4-9. For processor activation, you can add a minimum of 1 processor and up to the maximum of non-activated CUoD processors available in the system. For memory activation, you can add a minimum of 1 GB and up to a maximum of non-activated CUoD memory available in the system.

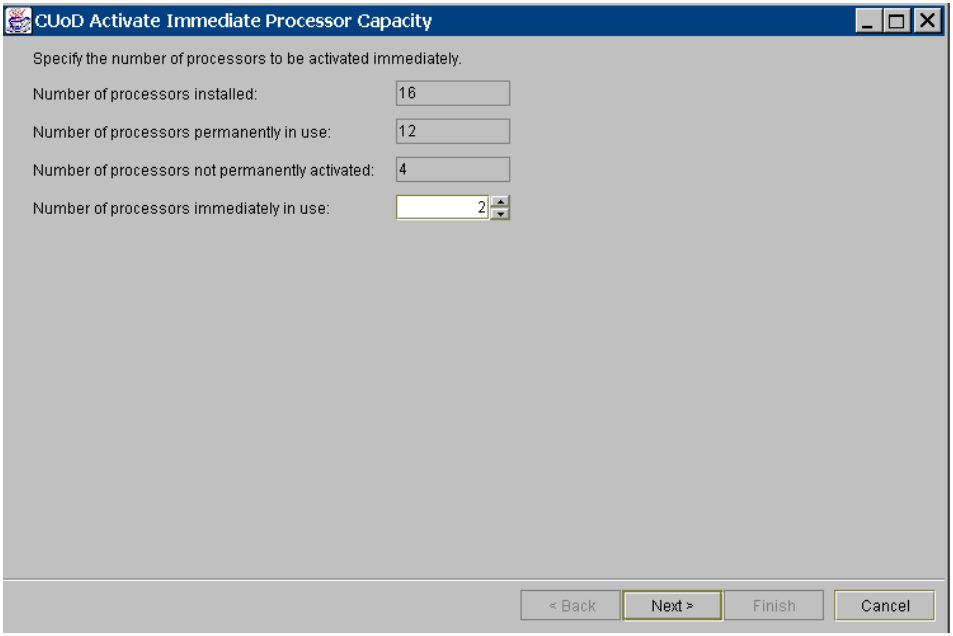


Figure 4-9 CUoD Trial CoD Processor Capacity (Next) screen

Important: All of the screen images for the Trial CoD function were taken before the official product announcement. In the window titles or menus, any text that appears as "Activate Immediate" should be read as "Trial CoD". It will be modified in the future software maintenance update of the HMC software product.

3. After entering the required amount of resources, click **Next** and a summary of the activated and non-activated resources will be displayed (see Figure 4-10).

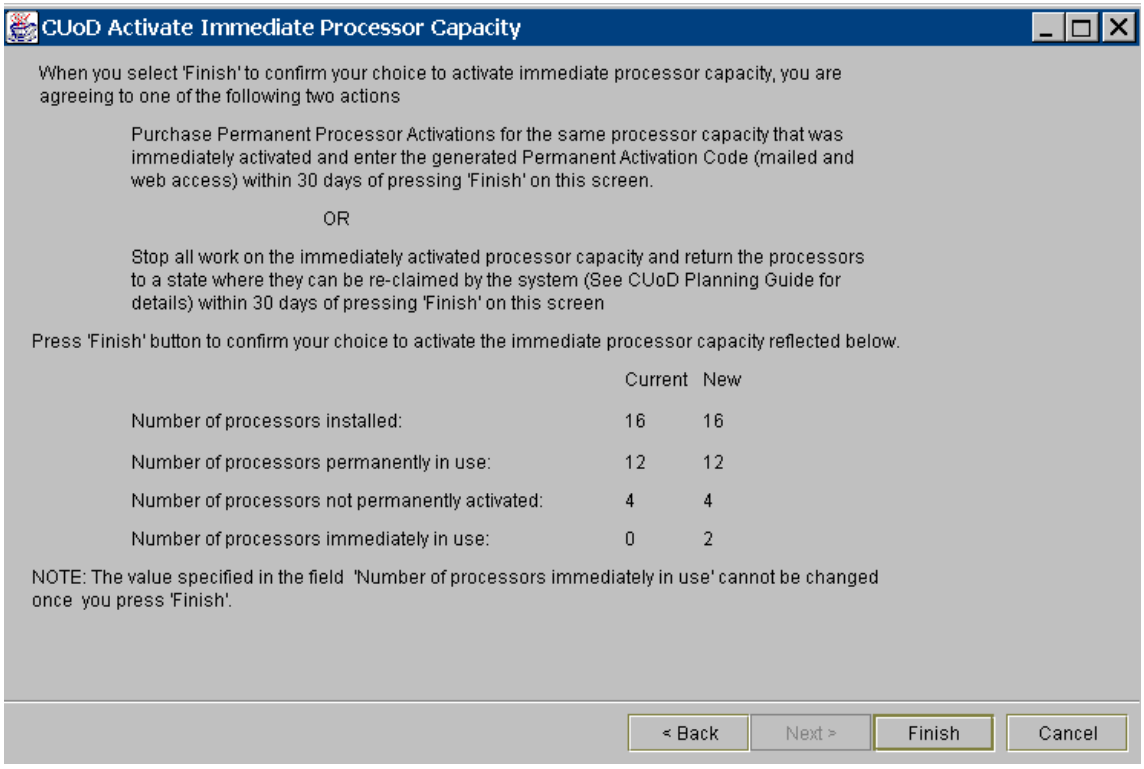


Figure 4-10 CUoD Trial CoD Processor capacity (Finish) screen

4. By clicking **Finish**, the customer has agreed to either purchase part or all of the activated resources or return them back to the system by disabling the Trial CoD function within 30 days.
5. If the customer chooses to purchase the resources, he or she needs to execute the steps as described in 4.3.1, “CUoD resources activation and order process” on page 147.

6. If the customer decides to stop using the resources, he or she will need to disable the Trial CoD function to return the resources back to the system. On the HMC, select Capacity Upgrade on Demand -> **Processor** -> **Processor Disable Activate Immediate** for processors, or select Capacity Upgrade on Demand -> **Memory** -> **Memory Disable Activate Immediate** for memory, and a warning message will appear as shown in Figure 4-11.

Figure 4-11 Warning message to Disable Processor Trial CoD Capacity

7. Click **Yes** to confirm the customer does not require the resources anymore and the resources will be returned to the system as non-activated CUoD resources.

Important: The customer will not be able to use Trial CoD function until the customer has purchased some of the CUoD resources which were activated through Trial CoD function previously.

4.4 Dynamic Processor Sparing

When you have a CUoD system (that is, a system with at least one CUoD MCM), a feature called Dynamic Processor Sparing is automatically provided with it. Dynamic Processor Sparing is the capability of the system to disable a failing processor and enable a non activated CUoD processor without any manual intervation. Non-activated CUoD processors are processors that are physically installed in the system, but not yet activated. They cannot run jobs or tasks unless they become activated via a CUoD processor activation code.

Note: If all the CUoD processors in the system are activated, even if these processors are not allocated to any partition, Dynamic Processor Sparing will not be enabled.

The processor that will be used to replace a failing processor will be the first available CUoD non-activated processor in the CUoD resource sequence, as described in 4.2.5, “CUoD resource sequencing” on page 142.

Both base processors and licensed CUoD processors may be replaced with unlicensed CUoD processors for Dynamic Processor Sparing when failing.

When Dynamic Processor Sparing is performed, the system will be out of CUoD compliance for a period of time between when the spare processor is allocated and the failed processor is deallocated. By design, AIX must request the spare processor, and in doing so AIX is committing to returning the failing processor in a timely fashion. If it fails to return the processor, the CUoD compliance enforcement code will detect the out of compliance condition. System firmware will ignore anything out of compliance during hot sparing for a sufficient time period for AIX to return the failing processor. The out of compliance condition will occur when more processors are active than what the customer is entitled to run (the number of base processors plus the number of CUoD processors by the customer).

Notes:

- Dynamic Processor Sparing requires AIX 5L Version 5.2 or higher, and the system must be run in a partitioned environment. The Dynamic Processor Sparing requires that the CPU guard attribute is set to enable (this is the default value on AIX 5L Version 5.2 or higher), as explained in “Enabling CPU Guard” on page 170.
- There is no “Dynamic Memory Sparing” function. Non activated memory (if it is available in the system) will be sequenced in on the next full system re-boot if there is any faulty memory detected.

4.5 On/Off Capacity on Demand

This section describes the functionality of On/Off Capacity on Demand (On/Off CoD) for processors on the pSeries 670 and pSeries 690 systems with CUoD resources. With On/Off CoD, a customer can enable and disable the non-activated processors dynamically and non-disruptively to the systems depending on their workload at a certain point in time. It provides flexibility to provide additional temporary processing power to the system during peak workload (for example, month-end reporting batch jobs).

The following are the basic features for a pSeries 670 or pSeries 690 system with the On/Off CoD:

- Each On/Off CoD activation entitles the customer to 60 processor days (PDs), defaulting to two processors for 30 days. Processor days (PDs) are actual in use days and not calendar days.
- Multiple On/Off CoD activation features are allowed, each one entitling the customer to an additional 60 PDs from the time it is entered into the system.

- A new customer interface will be introduced for the customer to manage the On/Off CoD processors.
- If the customer deactivates the On/Off CoD-activated processors, the remaining PDs are “banked” for use in the future.

Example 1

When a customer entered two On/Off CoD activation codes. Each activation code by default gives him access to two processors for 30 days, or 60 PDs, in this case a total of 120 PDs. If four non-activated processors are available in the system, there are two possible configurations:

- Activate 4 processors for 30 days, or
- Activate 2 processors for 60 days

5

Reliability, availability, and serviceability

The terms reliability, availability, and serviceability (RAS) are widely used throughout the computer industry as an indication of a product's failure characteristics. RAS refers to a collection of interdependent product attributes that, when taken together, attempt to measure how often a product fails, how quickly it can be repaired, and the overall system disruption caused by a failure.

This chapter describes the various features and mechanisms that are implemented in the pSeries 670 and pSeries 690 servers to minimize failures, and isolate and recover them if possible, by providing the following topics:

All this is developed to keep the impact on server operations for the customer as small as possible.

5.1 What's new in serviceability

All the RAS features that are incorporated in the previous pSeries 670 and pSeries 690 remained with the May 2003 announcement. A new feature to provide the customer the ability to survey and update the microcode levels was introduced. This will help customers ensure that their systems' microcode levels are up-to-date. Three sections have been added to explain the serviceability features:

- Section 5.6.1, "Back up of HMC" on page 180 provides an overview of two types of back up available in the HMC and the usage.
- Section 5.6.2, "Upgrading HMC" on page 181 describes the steps involved in upgrading the HMC to the level that the microcode updates feature can be used.
- Section 5.6.3, "Microcode Updates function" on page 184 describes the steps to survey and update the microcode from the HMC.

5.2 RAS features

Both the pSeries 670 and pSeries 690 bring new levels of availability and serviceability to the enterprise-level UNIX servers. Advanced functions presented on the previous models were enhanced, and new technologies have been added, to reduce faults and minimize their impacts. High-quality components and rigorous verification and testing during the manufacturing processes contribute to the reduction of failures in the system. The diagnostic goal of these RAS features is to isolate the Field Replaceable Units (FRU) challot to 95 percent to a single FRU; of course there will still be a chance that more than one FRU is failing, but the implemented features in the pSeries 670 and pSeries 690 will keep the customer impact as small as possible.

The pSeries 670 and pSeries 690 RAS design enhancements can be grouped into four main areas:

- **Predictive functions**: These are targeted to monitor the system for possible failures, and take proactive measures to avoid the failures.
- **Redundancy in components**: Duplicate components and data paths to prevent single points of failure.
- **Fault recovery**: Provide mechanisms to dynamically recover from failures, without system outage. This includes dynamic deallocation of components and hot-swap capability.
- **Serviceability features**: Enable the system to automatically call for support, and provide tools for rapid identification of problems.

The pSeries 670 and pSeries 690 RAS presents features in all these categories, as described in the following sections.

5.3 Predictive functions

In a mission-critical application, any outage caused by system failure will have an impact on users or processes. The extent of the impact depends on the type of the outage and its duration.

Unexpected system outages are the most critical in a system. The disruption caused by these outages not only interrupts the system execution, but can potentially cause data problems of either loss or integrity. Moreover, the recovery procedures after such outages are generally longer than for planned outages, because they involve error recovery mechanisms and log analysis.

Planned system outages are also critical in a mission-critical environment. However, the impact can be minimized by adequately planning the outage time and the procedures to be executed. Applications are properly shut down, and users are aware of the service stop. Therefore, there is less exposure when doing planned outages in a system.

Reliability engineering is the science of understanding, predicting, and eliminating the sources of system failures. Therefore, the ability to detect imminent failures, and the ability to plan system maintenance in advance are fundamental to reducing outages in a system, and to effectively implement a reliable server.

5.3.1 Service processor

The service processor is the interface between the HMC and the pSeries 670 and pSeries 690. It also provides functions for the Service Focal Point application.

Most of the predictive analysis and error recovery actions are directly related to or assisted by the service processor. However, there is no dependency between a server operation and the service processor. This means that even if the service processor fails, the server continues operating, and the running partitions are not affected by the failure.

What the service processor is

A service processor is a separate independent processor that provides hardware initialization during system IPL, operation monitoring of environmental and error events, and maintenance support for the pSeries 670 and pSeries 690. The

service processor communicates with the pSeries 670 and pSeries 690 through attention signals from the hardware and read/write communication through the JTAG ports between the service processor and all the pSeries 670 and pSeries 690 chips. This diagnostic capability is simultaneous, asynchronous, and transparent to any system activity running on the machine.

Figure 5-1 on page 160 illustrates an overview of how the service processor works.

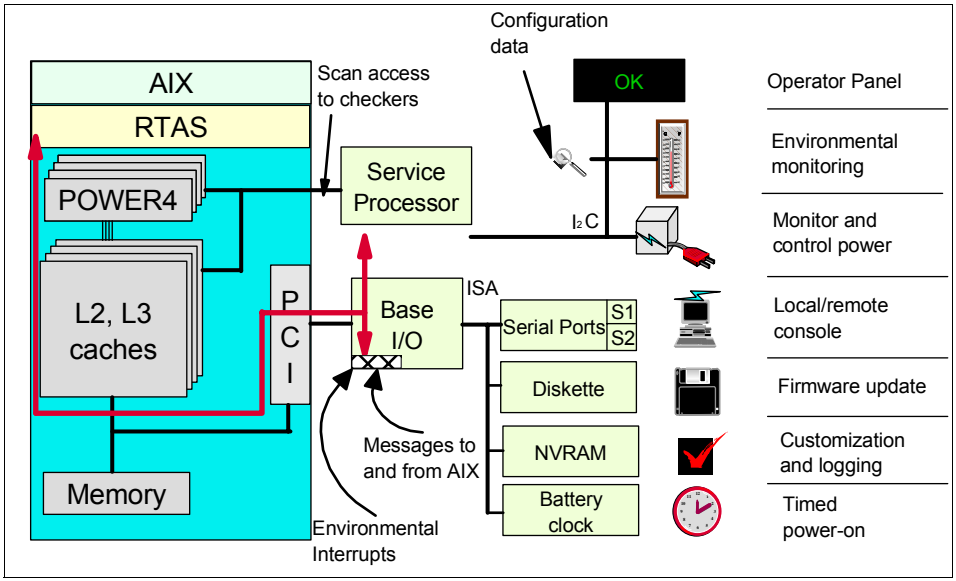


Figure 5-1 Service processor schematic

The NVRAM has a special function for the service processor. It acts like a *mailbox* for the service processor, system firmware, and the AIX running on the pSeries 670 and pSeries 690. An FRU callout, for example, will be removed by the system firmware and passed to AIX, then written into the AIX system error log.

5.3.2 First Failure Data Capture (FFDC)

The pSeries 670 and pSeries 690 have the capability of error detection and logging, using AIX and service processor functions. These functions work in a special way to capture all possible failures, and to store information about them.

First Failure Data Capture is the keystone to building a system problem determination strategy around the capture and analysis of errors as they happen, as opposed to attempting to recreate them later. This not only provides the exact information about the status of the system when the failure happened, but it also provides a way to detect and capture information related to temporary failures, or failures related to server load. Hardware error checkers are strategically placed on the system, operating full time to detect and capture precise error information.

All error checkers have a blocking logic so that for every detected error the error is recorded only by the first checker that encounters it. The error information is stored in Fault Isolation Registers (FIRs) to be later examined.

There are over 5600 fault isolation register bits representing over 15000 internal error checkers. All error check stations have been placed in the data and control paths of pSeries 670 and pSeries 690 systems to deterministically isolate physical faults based on run-time detection of each unique failure that may occur. Run-time error diagnostics are deterministic in that for every check station, the unique error domain for that checker is defined and documented.

Figure 5-2 shows a schematic example of error checker positioning and FIR updating.

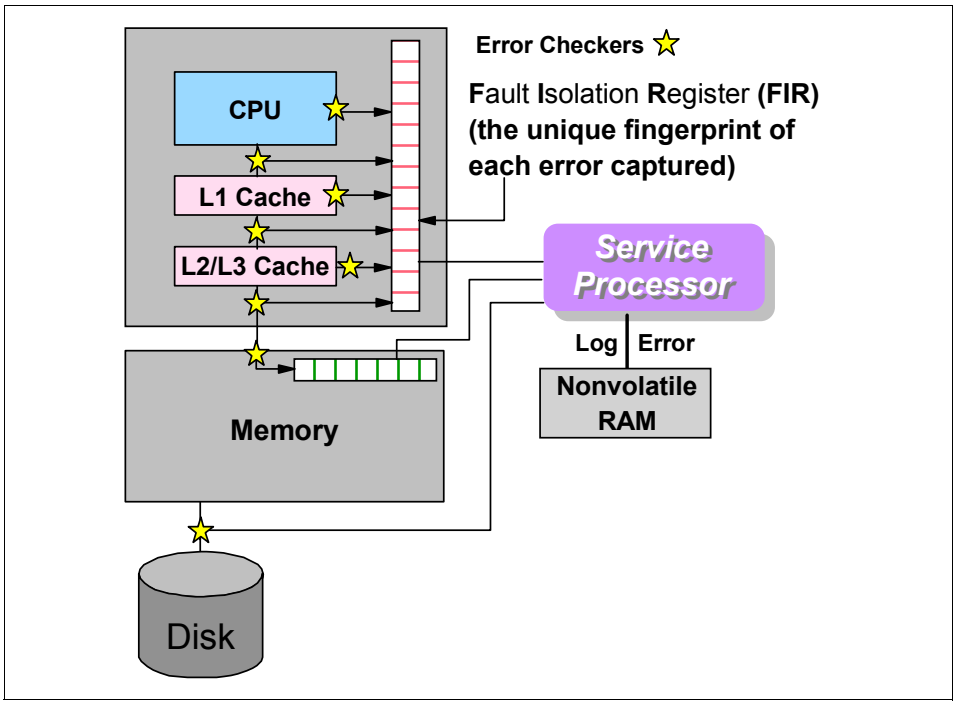


Figure 5-2 FFDC error checkers and fault isolation registers

5.3.3 Predictive failure analysis

Statistically, there are two main situations where a component has a catastrophic failure: Shortly after being manufactured, and when it has reached its useful life period. Between these two regions, the failure rate for a given component is generally low, and normally gradual. A complete failure usually happens after some degradation has happened, be it in the form of temporary errors, degraded performance, or degraded functionality.

The pSeries 670 and pSeries 690 have the ability to monitor critical components such as processors, memory, cache, I/O subsystem, and internal disks, and detect possible indications of failures. By continuously monitoring these components, upon reaching a threshold, the system can isolate and deallocate the failing component without system outage, thereby avoiding a complete failure.

5.3.4 Component reliability

The components used in the CEC provide superior levels of reliability that are available and undergo additional stress testing and screening above and beyond the industry-standard components that are used in several UNIX-based systems today.

Fault avoidance is also served by minimizing the total number of components, and this is inherent in POWER4 chip technology, with two processors per chip. In addition, internal array soft errors throughout the POWER4 chip are systematically masked using internal ECC recovery techniques whenever an error is detected. Going beyond ECC in the memory subsystem, the basic memory DIMM technology has been significantly improved in reliability through the use of more reliable soldered connections to the memory card.

5.3.5 Extended system testing and surveillance

The design of the pSeries 670 and pSeries 690 aids in the recognition of intermittent errors that are either corrected dynamically or reported for further isolation and repair. Parity checking on the system bus, cyclic redundancy checking (CRC) on the remote I/O bus, and the use of error correcting code on memory and processors contribute to outstanding RAS characteristics.

During the boot sequence, built-in self test (BIST) and power-on self test (POST) routines check the processors, cache, and associated hardware required for a successful system start. These tests run every time the system is powered on.

Additional testing can be selected at boot time to fully verify the system memory and check the chip interconnect wiring. When a system reboots after a hard failure, it performs extended mode tests to verify that everything is working properly and that nothing was compromised by the failure. This behavior can be overridden by the systems administrator.

5.4 Redundancy in components

The pSeries 690 and pSeries 670 system design incorporates redundancy in several components to provide fault-tolerant services where failures are not allowed. Power supplies, fans, blowers, boot disks, I/O links, and power cables offer redundancy to eliminate single points of failure. Some of these features are highlighted, as described in the following sections.

5.4.1 Power and cooling

The pSeries 670 and pSeries 690 provide full power and cooling redundancy, with dual power cords and variable-speed fans and blowers, for both the CEC and the I/O drawers.

Within the CEC rack, the N+1 power and cooling subsystem provides complete redundancy in case of failures in the bulk or regulated power supplies, the power controllers, and the cooling units, as well as the power distribution cables. As on the zSeries server, concurrent repair is supported on all of the CEC power and cooling components. See Table 5-1 on page 173 for the hot-pluggable components.

Note: The pSeries 670 supports either single-phase power or three-phase, while pSeries 690 supports only three-phase power.

There is also a redundant feature called internal battery features (IBF) designed to maintain system operation during brown-out conditions. In case of a total power loss, the batteries can be used to execute a complete shutdown of the system in an ordered way. For full power loss protection, the pSeries 670 and pSeries 690 supports optional uninterruptible power supply (UPS) systems in addition to, or in place of, the IBF features. You should see the IBF feature as a redundancy feature only; it will not replace the UPS capabilities.

In case of a fan or blower failure, the remaining fans automatically increase speed to compensate for the air flow from the failed component.

5.4.2 Memory redundancy mechanisms

There are several levels of memory protection. From the internal L1 caches to the main memory, several features are implemented to assure data integrity and data recovery in case of memory failures.

- The L1 caches are protected by parity, and if an error is detected, the data is fetched from the L2 cache, and the cache line with the parity error is invalidated by the hardware. All data stored in the L1 data cache is available in the L2 cache, guaranteeing no data loss.
- The L2 and L3 caches and the main memory are protected by ECC. In case of a single bit failure, it is corrected and the system continues to operate normally. In the case of double bit, errors are detected but not corrected.
- The L2 cache directory has two parity protected copies of data. If a parity error is detected, the directory request is recycled while the entry with the error is refreshed from the good one.
- The L1 and L2 caches and the L2 and L3 cache directories all have spare bits in their arrays that can be used to recover from bit failures.
- Each L3 chip has a spare line to allow for line delete and replacement if the error becomes hard to recover.

Figure 5-3 graphically represents the redundancy and error recovery mechanisms on the main memory.

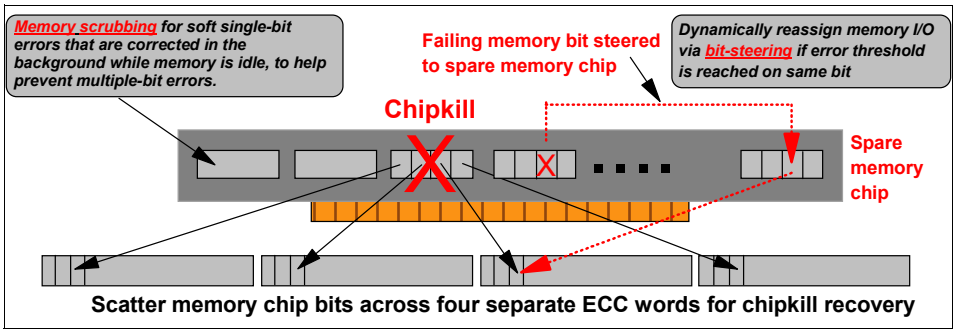


Figure 5-3 Memory error recovery mechanisms

The memory implements the following three techniques, which complement the basic error correction design and significantly improve the robustness.

1. Chipkill™ recovery

   Chipkill recovery is accomplished by bit scattering, so that even an entire DRAM chip failure appears to the ECC logic as correctable single bit errors in several different ECC words. When a whole memory chip fails, it can be recovered and replaced by a spare chip.

   A technique named *bit scattering* scatters memory chip bits across four separate ECC words in order to improve recovery from a memory chip failure. The failure of any specific memory module only affects a single bit within an ECC word.

2. Scrubbing

   Scrubbing is a hardware memory controller function that determines, during normal run time, whether or not a correctable permanent failure has occurred in one or more bit locations. It also reads memory out when it would otherwise be idle and corrects any single bit errors it detects. When the scrubbing threshold is reached, the third technique, called *redundant bit steering*, is invoked, which steers the data lines away from the faulty chip and onto a redundant chip. Memory scrubbing takes place while the memory controller is idle, is completely transparent to software, and uses no CPU cycles.

3. Redundant bit steering

   Redundant bit steering is enabled by having the scrubbing hardware detect a fault, together with steering logic, to provide automatic access to the spare bits. After redundant bit steering is invoked, the memory region is again free of permanent bit errors. In case of another bit failure on the same word, the ECC correction can again be used.

   This is named *bit steering*, and it greatly reduces system outages due to memory failures.

5.4.3 Multiple data paths

The communication between processors in an MCM is done over four local buses. Different MCMs are connected through a distributed switch topology, where the local MCM buses are connected to the buses on the other MCMs, and data is routed between them. Any processor can access resources connected to any MCM.

The I/O subsystem is based on the Remote I/O link technology. This link uses a loop interconnect technology to provide redundant paths to I/O drawers. Each I/O drawer is connected to two RIO ports, and each port can access every component in the I/O drawer. During normal operations the I/O is balanced across the two ports. If an RIO link fails, the hardware is designed to automatically initiate a RIO bus reassignment to route the data through the alternate path to its intended destination.

Any break in the loop is recoverable using alternate routing through the other link path and can be reported to the service provider for a deferred repair.

Power to the drawers is controlled from the power controller in the CEC through the RAS infrastructure (see 2.4.2, “I/O subsystem communication and monitoring” on page 56).

5.5 Fault recovery

The pSeries 670 and pSeries 690 offer new features to recover from several types of failure automatically, without requiring a system reboot. The ability to isolate and deconfigure components while the system is running is of special importance in a partitioned environment, where a global failure can impact different applications running on the same system.

Some faults require special handling. We will discuss these in the following sections.

5.5.1 PCI bus parity error recovery and PCI bus deallocation

PCI bus errors, such as data or address parity errors and timeouts, can occur during either a DMA operation being controlled by a PCI device, or on a load or store operation being controlled by the host processor.

During DMA, a data parity error results in the operation being aborted, which usually results in the device raising an interrupt to the device driver, allowing the driver to attempt recovery of the operation.

However, all the other error scenarios are difficult to handle as gracefully. On previous systems, these errors resulted in a bus critical error, followed by a machine check interrupt and system termination.

In the pSeries 670 and pSeries 690, a new I/O drawer hardware, system firmware, and AIX interaction has been designed to allow transparent recovery of intermittent PCI bus parity errors, and graceful transition to the I/O device unavailable state in the case of a permanent parity error in the PCI bus. This mechanism is known as the *PCI Extended Error Handling (EEH)*.

Standard server PCI implementations can detect errors on the PCI bus, but cannot limit the scope of the damage, so system termination is often the result. The I/O drawer (7040-61D) used for the pSeries 670 and pSeries 690 has enhanced the handling of errors on the PCI bus to limit the scope of the damage to one PCI slot and loss of use of the PCI adapter in that PCI slot rather than system termination.

While the basic enablement of PCI EEH is in the circuitry of the I/O drawer, changes are required in the OS device driver to fully exploit EEH. Many of the current AIX device drivers are fully enabled for EEH, and most of those that are not EEH enabled will be enabled in the next AIX release. Also, IBM is working to have future releases of Linux device drivers EEH enabled.

Use of EEH is particularly important on the pSeries 670 and pSeries 690 running with multiple partitions. If non-EEH enabled PCI adapters are in use, it is possible for multiple partitions to be disrupted by PCI bus errors. If a customer's application requires the use of non-EEH enabled PCI adapters, careful placement of those PCI adapters in the I/O drawer can limit PCI bus error disruptions to a single logical partition.

The ultimate situation is to use only EEH-enabled PCI adapters, to eliminate system and partition disruptions due to PCI bus errors, and merely suffer the loss of a single PCI adapter if that adapter causes a PCI bus error. Most adapters support EEH. It is part of the PCI 2.0 Specification, although its implementation is not required for PCI compliance.

For those adapters that do not support EEH, special care should be taken when configuring partitions for maximum availability. See the *IBM @server* pSeries 690 *Availability Best Practices* white paper for I/O adapter configuration guidelines, which can be found at:

`http://www.ibm.com/servers/eserver/pseries/hardware/whitepapers/p690_avail.html`

The principle of how the error is recovered is described as follows:

1. An operation to access a PCI slot causes an error.
2. Upon detection of the error, the access to the bus is denied.
3. The error recovery mechanism probes the PCI bus for the error code.
4. If it is an unrecoverable error, the bus returns `0xFFFFFFFF`.

5. The device driver performs a call to the firmware requesting a PCI bus reset.
6. After the reset, the device driver retries the operation. If it fails three times, the slot is deconfigured and a permanent error is logged.

A similar error-handling procedure is performed during system startup when the PCI bridges are initialized.

5.5.2 Dynamic CPU deallocation

Dynamic CPU deallocation has been available since AIX Version 4.3.3 on previous RS/6000 and pSeries systems, and is part of the pSeries 670 and pSeries 690 RAS features. The dynamic CPU deallocation is also supported in a partitioned environment.

Processors are continuously monitored for errors, such as L1 and L2 cache errors. When a predefined error threshold is met, an error log with warning severity and threshold exceeded status is returned to AIX. At the same time, the service processor marks the CPU for deconfiguration at the next boot. In the meantime, AIX will attempt to migrate all resources associated with that processor (threads, interrupts, and so on) to another processor, and then stop the failing processor.

The typical flow of events for processor deallocation is as follows:

1. The firmware detects that a recoverable error threshold has been reached by one of the processors.
2. AIX logs the firmware error report in the system error log and, when executing on a machine supporting processor de-allocation, starts the deallocation process.
3. AIX notifies non-kernel processes and threads bound to the last logical CPU.
4. AIX waits for all the bound threads to move away from the last logical CPU. If threads remain bound, AIX eventually times out (after ten minutes) and aborts the deallocation.
5. Otherwise, AIX invokes the previously registered *High Availability Event Handlers* (*HAEHs*). An HAEH may return an error that will abort the deallocation.
6. Otherwise, AIX goes on with the deallocation process and ultimately stops the failing processor.

5.5.3 CPU Guard

It is necessary that periodic diagnostics not run against a processor already found to have an error by a current error log entry. CPU Guard provides this

feature. Otherwise, when an error occurs and the processor cannot be deallocated, we could end up with multiple periodic diagnostic failures for the same processor.

Figure 5-4 illustrates how AIX handles the CPU Guard activities.

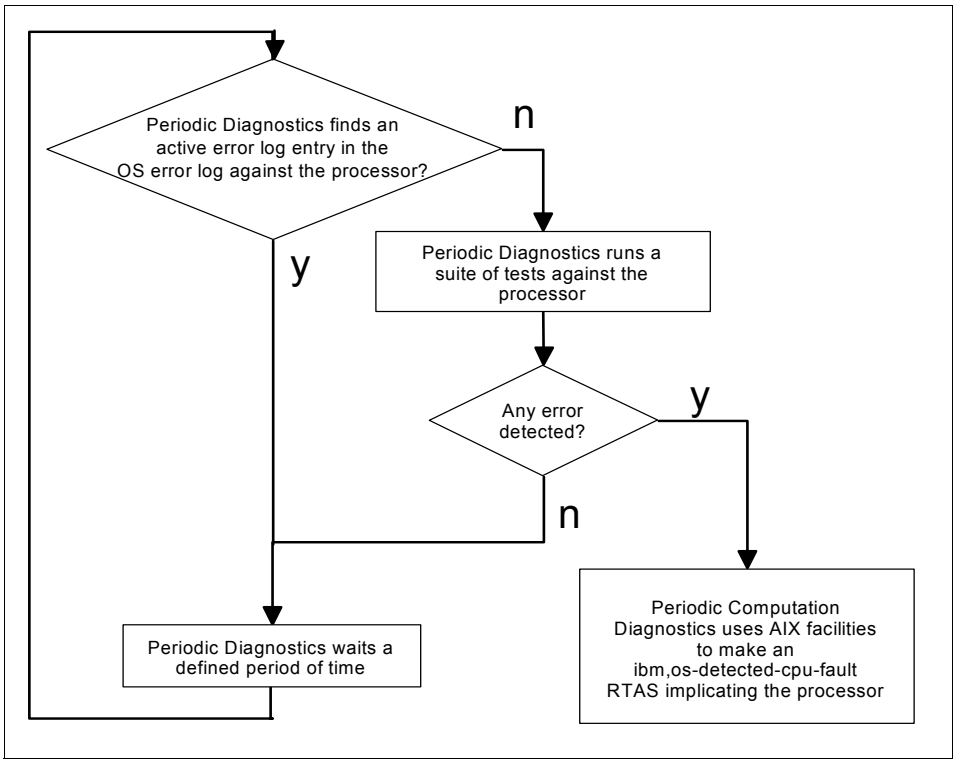


Figure 5-4 CPU Guard activities handled by AIX

Figure 5-5 illustrates how the firmware also handles those errors.

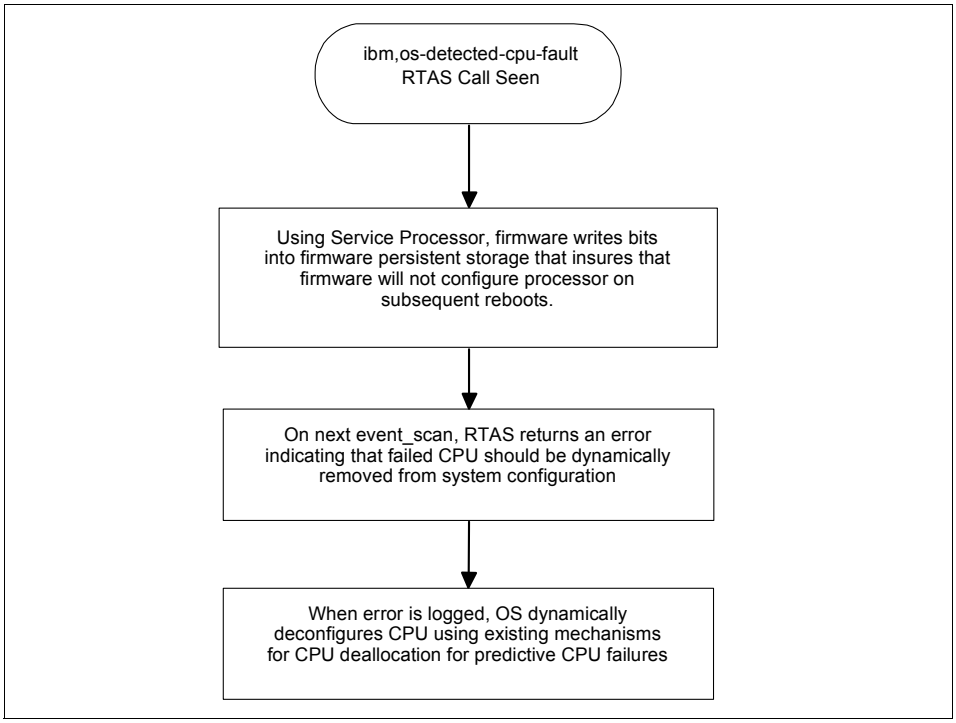


Figure 5-5 CPU Guard activities handled by firmware

Enabling CPU Guard

In case of failure at any point of the deallocation, AIX logs the failure and the reason why the deallocation was aborted. The system administrator can look at the error log, take corrective action (when possible), and restart the deallocation. For example, if the deallocation was aborted because at least one application did not unbind its bound threads, the system administrator could stop the application(s), restart the deallocation (which should go through this time), and restart the application. You can turn off the dynamic CPU deallocation by selecting the following SMIT panels[1]:

```
# smit
    System Environments
        Change / Show Characteristics of Operating System
```

See the SMIT panel shown in Example 5-1.

[1] The SMIT shortcut is smitty chgsys.

Example 5-1 smitty chgsys

```
          Change / Show Characteristics of Operating System

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                  [Entry Fields]
  Maximum number of PROCESSES allowed per user       [128] +#
  Maximum number of pages in block I/O BUFFER CACHE  [1000] +#
  Maximum Kbytes of real memory allowed for MBUFS    [0] +#
  Automatically REBOOT system after a crash           false +
  Continuously maintain DISK I/O history              false +
  HIGH water mark for pending write I/Os per file    [0] +#
  LOW water mark for pending write I/Os per file     [0] +#
  Amount of usable physical memory in Kbytes         1048576
  State of system keylock at boot time               normal
  Enable full CORE dump                              false                  +
  Use pre-430 style CORE dump                        false                  +
  CPU Guard                                           enable +
  ARG/ENV list size in 4K byte blocks                [6] +#
```

The default value for the CPU Guard flag is *enable*[2] in AIX 5L Version 5.2.

Note: Even if the dynamic deallocation is disabled, AIX still reports the error in the AIX system error log and you will see the error indicating that AIX was notified with the problem of a CPU (CPU_FAILURE_PREDICTED).

Sometimes the processor deallocation fails because, for example, an application did not move its bound threads away from the last logical CPU. Once this problem has been fixed, by either unbinding (when it is safe to do so) or stopping the application, the system administrator can restart the processor deallocation process using the **ha_star** command, as shown in the following example:

```
# ha_star -C
ha_star: 1735-008 No previous CPU failure event received.
```

Use the syntax **ha_star -C**, where -C is for a CPU predictive failure event. The processors are represented in the ODM database as *procX*, where X is the number of the processor. With the **lsattr** command you can look at the current status of a processor, as shown below.

```
# lsattr -El proc2
state     enable         Processor state False
type      PowerPC_POWER4 Processor type  False
frequency 1100264056     Processor Speed False
```

[2] In AIX 5L Version 5.1 or lower, the default value is *disable*.

The state of this processor is enable, so there is no problem encountered so far with that processor. There are three possible states:

enable The processor is used by AIX.

disable The processor has been dynamically deallocated by AIX.

faulty The processor was declared defective by the firmware at boot time.

5.5.4 Caches and memory deallocation

The L3 cache is ECC protected, so single bit errors are transparently corrected. When an error occurs, the service processor checks to see if any previous L3 failure occurred and if the failing address matches the previous error. If there is no match, the service processor will then save the failed address and issue a L3 purge (flush the cache line) to attempt to remove the soft error.

If the address is the same as a previous failure, then this will be treated as a hard single bit error (SBE), and the service processor has two recovery mechanisms:

Cache line delete The service processor can use the cache line delete capability to permanently remove this cache line from being used. There are two L3 directory cache lines (one per half) that can be deleted for each L3 module.

L3 cache deconfiguration If the number of errors exceed two, and both cache lines are already deleted, the cache continues to run, but a message calling for deferred repair is issued. If the machine is rebooted without such repair, the L3 cache is placed in bypass mode, and the system comes up with this cache deconfigured.

The L2 cache can also be deconfigured, as part of the dynamic CPU deallocation process. If the L2 caches errors inside a POWER4 chip, the CPU deallocation process is invoked, and the processors and L2 cache are deconfigured from the system.

Memory cards can be deconfigured at boot time when they fail to pass POST, or when marked to be deconfigured by the service processor.

5.5.5 Hot-swappable components

Both the pSeries 670 and pSeries 690 provide many parts as hot-swappable Field Replaceable Units (FRUs). This feature allows you to replace most parts of the pSeries 670 and pSeries 670 concurrently without the need to power off the system. There are some parts like the MCMs. L3 cache, memory, etc., that still

require a scheduled maintenance window to perform the replacement. Table 5-1 provides you with an overview of which components are hot-pluggable.

Table 5-1 Hot-swappable FRUs

Processor subsystem FRUs	**Hot-swappable concurrent maintenance**
Blowers	Yes
DCA	Yes
Bulk power enclosure	Yes
Bulk power controller (BPC)	Yes
Bulk power regulator (BPR)	Yes
Bulk power distributor (BPD)	Yes
Bulk power fan (BPF)	Yes
UEPO switch panel	Yes
Internal battery feature (IBF)	Yes
Capacitor book	No
Processor subsystem chassis	No
MCM	No
Memory cards	No
L3 cache	No
Clock card	No
I/O books	No
I/O subsystem FRUs	**Hot-swappable concurrent maintenance**
I/O backplane and riser card	No
DASD backplane	No
Disk drives	Yes
DCA (power supplies)	Yes
I/O fan assemblies	Yes
PCI adapters	Yes
Media subsystem FRUs	**Hot-swappable concurrent maintenance**
Operator panel	No

Processor subsystem FRUs	Hot-swappable concurrent maintenance
Diskette drive	Yes
CD-ROM/DVD-RAM/DVD-ROM	Yes
Optional media SCSI devices	Yes

5.5.6 Hot-swappable boot disks

The I/O drawer (7040-61D) of the pSeries 670 and pSeries 690 provides up to 16 hot-swappable bays for internal disks, organized in four cages, each one connected to a separate Ultra3 SCSI controller.

Disk mirroring is strongly suggested for the operating system and paging in order, to eliminate system outages because of operating system disk failures. If a mirrored disk fails, AIX automatically reroutes the I/O requests to the other disk, without service interruption. You can replace the failed drive with the system online, and mirror again, to restore disk redundancy.

5.5.7 Hot-Plug PCI adapters

All PCI slots on the pSeries 670 and pSeries 690 are PCI 2.2-compliant and are hot-plug enabled, which allows most PCI adapters to be removed, added, or replaced without powering down the system. This function enhances system availability and serviceability.

The function of Hot-Pluggable PCI adapters is to provide concurrent adding or removal of PCI adapters when the system is running. In the I/O drawer, the installed adapters are protected by plastic separators called *blind swap cassettes*. These are used to prevent grounding and damage when adding or removing adapters.

Figure 5-6 shows a drawing of the blind swap cassette.

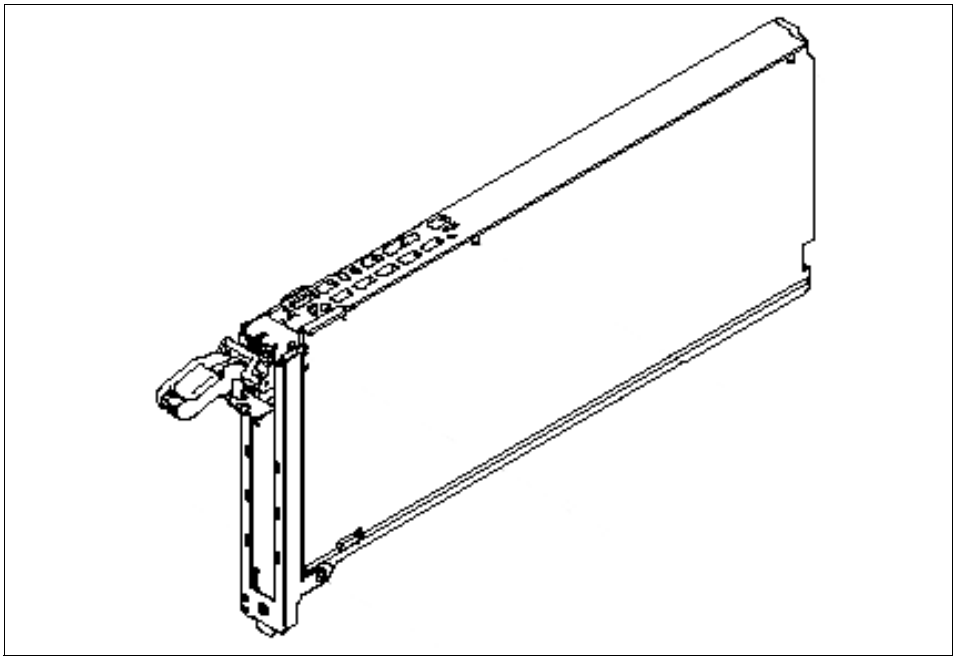


Figure 5-6 Blind swap cassette

The Hot-Plug LEDs outside the I/O drawer indicate whether an adapter can be plugged into or removed from the system. The Hot-Plug PCI adapters are secured with retainer clips on top of the slots; therefore, you do not need a screwdriver to add or remove a card, and there is no screw that can be dropped inside the drawer. Just in case of exchanging an PCI adapter from the blind swap cassette, a screwdriver is needed to remove/replace it from the cassette. To fit several adapters that are different sizes, there are several different blind swap cassettes available.

Figure 5-7 shows a schematic of an I/O drawer with the status LEDs for the RIO ports and PCI slots.

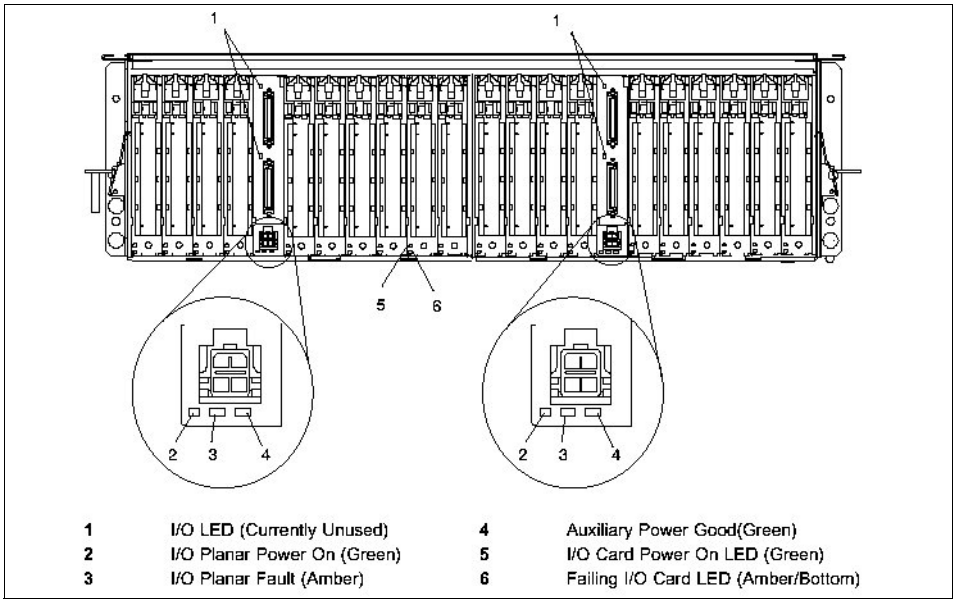


Figure 5-7 Hot-Plug PCI adapters, blind swap cassette, and status LEDs

The function of Hot-Plug is not only provided by the PCI slot, but also by the function of the adapter. Most adapters are Hot-Pluggable, but some are not. Be aware that some adapters must not be removed when the system is running, such as the adapter with the operating system disks connected to it or the adapter that provides the system console.

For further information, refer to *AIX 5L Version 5.2 System Management Guide: Operating System and Devices* (located on the documentation CD-ROM that ships with the AIX operating system) and from *PCI Adapter Placement References*, SA38-0538.

5.5.8 Light Path Diagnostics

The pSeries 670 and pSeries 690 introduce a new concept named *Light Path Diagnostics*, which provides a visual identification of a failed component in order to facilitate the maintenance. This functionality is available for detecting problems in I/O drawers, PCI planers, fans, blowers, and disks.

When a component presents a problem and it is detected by the surveillance mechanisms, the physical location occupied by the failed component flashes a LED, allowing easy identification of the component requiring maintenance.

Disk drives, fans, blowers, and PCI adapters have specific LEDs to indicate the operational state and faults when they occur. Figure 5-8 shows the LED placement on I/O drawers from the front view.

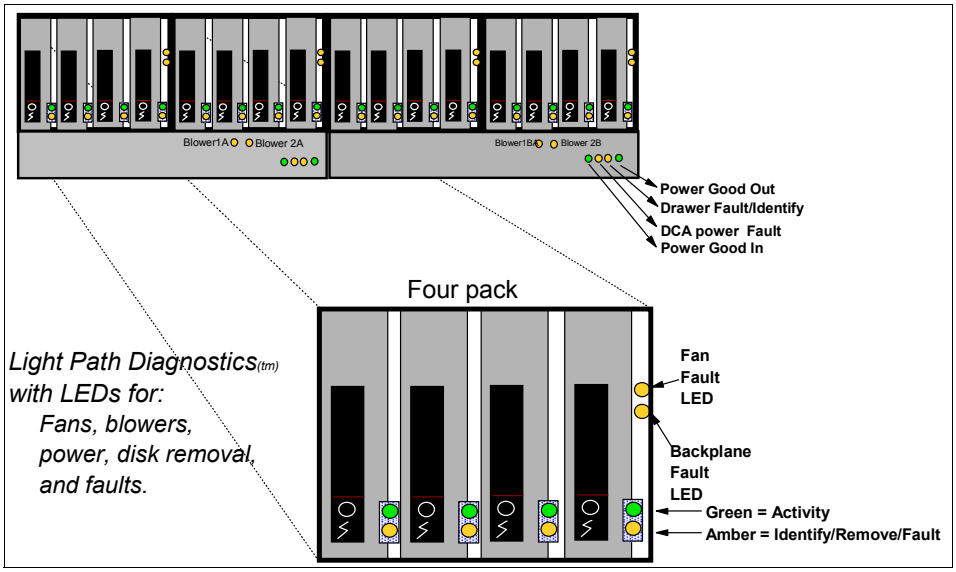


Figure 5-8 Status LEDs for components in I/O drawers

The Hardware Service Functions available in the Service Focal Point application on the HMC provides you with the method to identify the LEDs for the frame, cards, and power supplies (see “How to use Hardware Service Function” on page 199).

5.6 Serviceability features

The serviceability features on the pSeries 670 and pSeries 690 are centralized on the HMC and provided under **Software Maintenance** (see Figure 5-9 on page 178) and **Service Applications** in the Navigation Area (see Figure 5-10 on page 179).

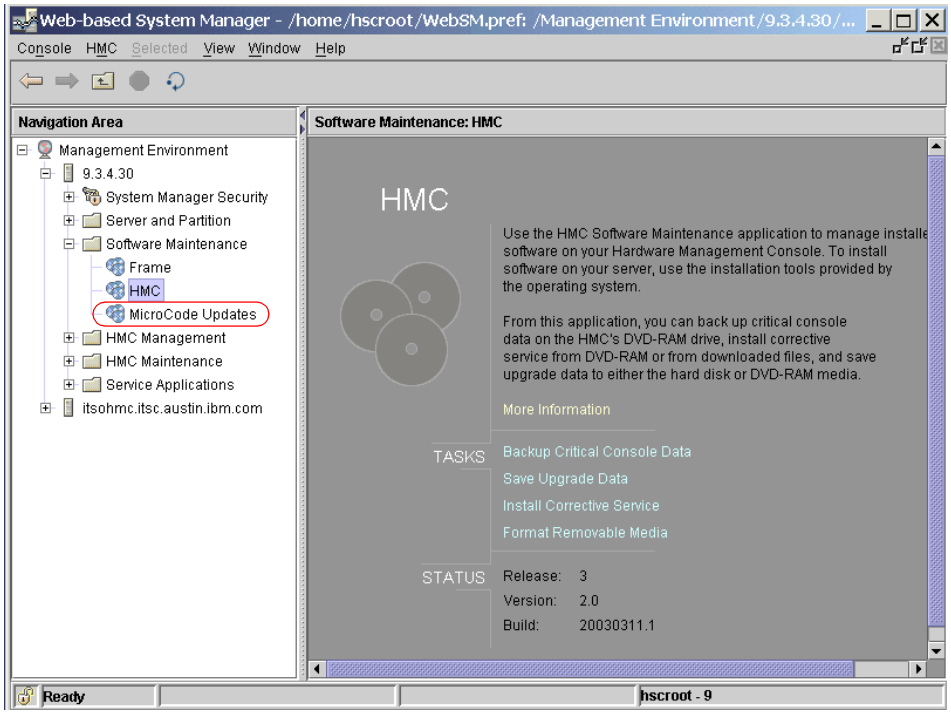


Figure 5-9 Software maintenance

Note: “Microcode Update” will be shown only when HMC software release Version 1.3.2 or later is installed.

Software Maintenance contains the following applications:

- Frames[3]
- HMC
- Microcode Updates

In 5.6.2, “Upgrading HMC” on page 181 we describe the steps involved to upgrade the HMC software via the HMC application, and in 5.6.3, “Microcode Updates function” on page 184 we describe the steps involved to update the microcodes via the Microcode Updates application.

[3] Frame functions will only be available in the future.

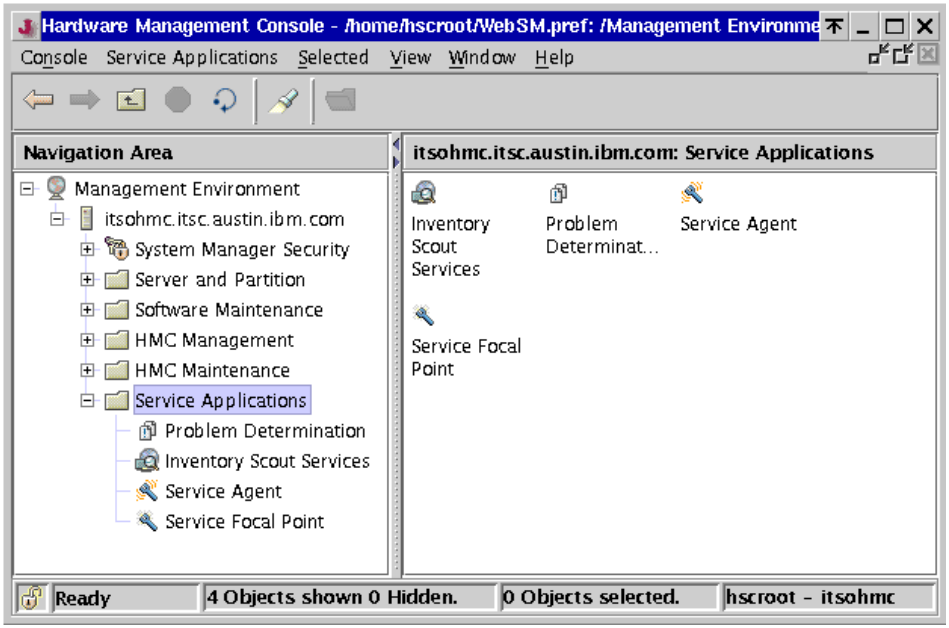


Figure 5-10 Service Applications

Service Applications contain the following applications:

- Problem Determination[4]
- Inventory Scout Services
- Service Agent
- Service Focal Point

Figure 5-11 illustrates the relationship between these components.

[4] The problem determination function is only available for the product support engineers.

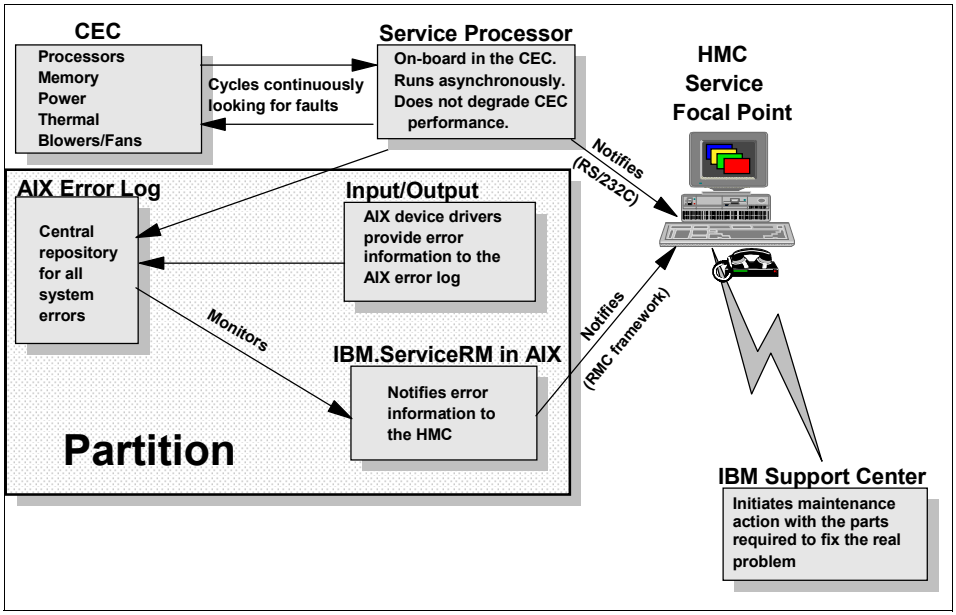


Figure 5-11 Error reporting and consolidation

All the applications' usage is well explained in *IBM Hardware Management Console for pSeries Installation and Operations Guide*, SA38-0590, therefore, we focused on the following points in this section:

- What components comprise these applications
- Problem determination hints

5.6.1 Back up of HMC

There are two types of backup that you need to be aware of: Backup Critical Console Data and Save Upgrade Data.

For Backup Critical Console Data, its function is to provide a means for customers to save the data on the HMC that has changed since the creation/installation of the code on the HMC to the DVD-RAM. All files including those belonging to packages installed after the date the HMC Recovery image was created will be backed up. The typical types of files that are backed up with this function are backup copies of LPAR/Profiles, system information such as network configurations, user configuration file and individual user's home directory. The backup function should be performed each time configuration changes are made to the system or the HMC.

Note: If the HMC software is upgraded from one release to another, do not use the Backup Critical Console Data to restore the configurations. This back up will restore the configurations including the previous HMC release.

For Save Upgrade Data, this function should only be used when customer wants to upgrade the current code on their HMC from one release/version to a higher release/version with a new HMC Recovery CD. This is not required if the code from the new release is downloaded/installed from the Support Web site or from an update CD ordered from IBM. Data saved by this function resides only in a special partition on the HMC disk. After the installation of the new HMC software from the HMC Recovery CD, on the first reboot, a service will examine the partition if there is any data, and then restore the data. For more information, refer to *IBM Hardware Management Console for pSeries Installation and Operations Guide*, SA38-0590.

5.6.2 Upgrading HMC

Note: If you have an HMC with a release earlier than Release 3, please read the chapter about “Installing and Configuring the HMC” in the *IBM Hardware Management Console for pSeries Installation and Operations Guide*, SA38-0590.

Here are the steps for updating HMC software from V1.3.0 to V1.3.2:

1. Before the upgrade, verify the current HMC software version by selecting **Software Maintenance** -> **HMC** as shown in Figure 5-12.

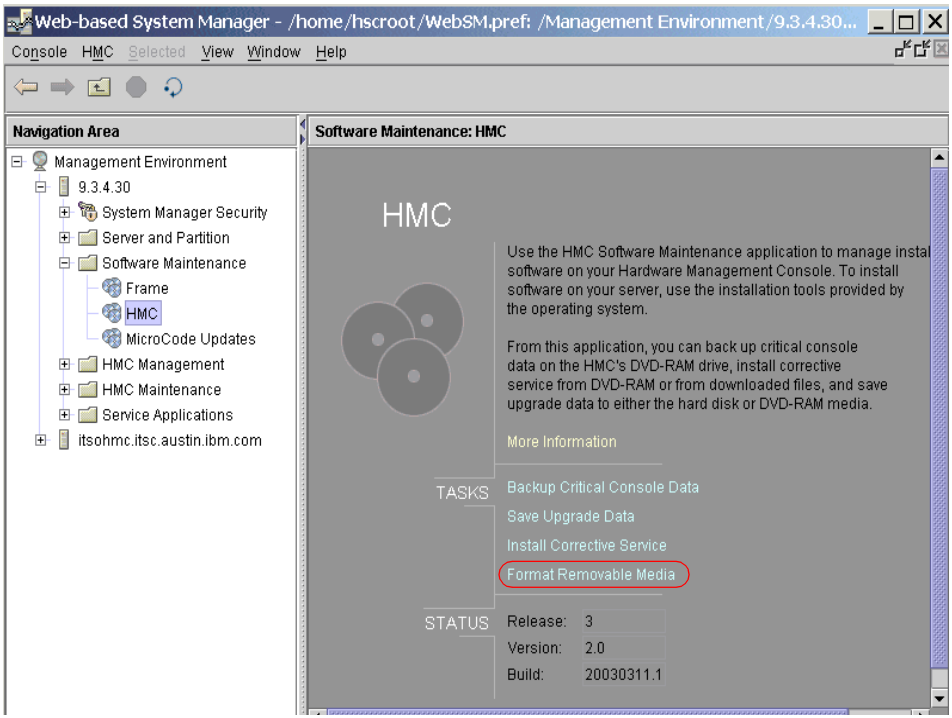


Figure 5-12 Software Maintenance: HMC

2. Select **Format Removable Media** (see circle section in Figure 5-12) -> **DVD RAM** (see Figure 5-13) to format the DVD RAM which will be used for back up.

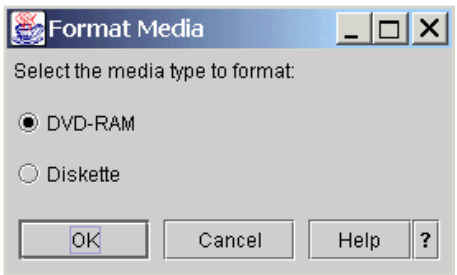


Figure 5-13 Format Media

3. Select **Backup Critical Console Data** to back up the data onto the DVD RAM (see Figure 5-14).

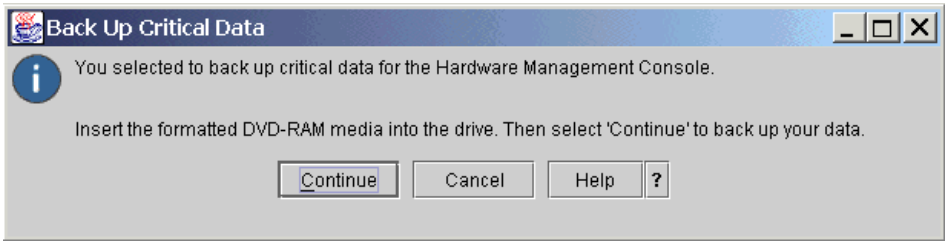


Figure 5-14 Backup Critical Data

4. Select **Install Corrective Service** to upgrade the HMC software. The user needs to select where the corrective service files are located, which can be either on a CD or a remote site. For this example, the files are located at a remote site (see Figure 5-15).

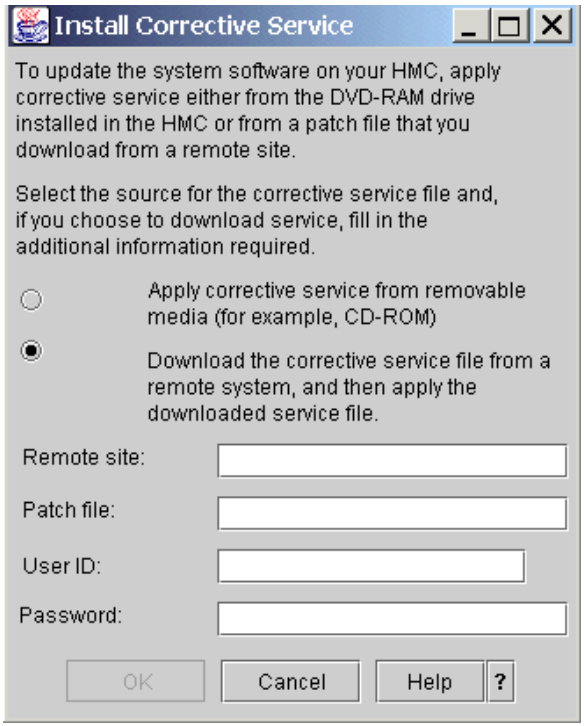


Figure 5-15 Install Corrective Service

5. After the upgrade is completed, if the managed systems are connected through 8-port or 128-port asynchronous adapters, you would need to reconfigure these serial adapters. The details can be found in the chapter about “Using One HMC to Connect to More than One Managed System” in

the *IBM Hardware Management Console for pSeries Installation and Operations Guide*, SA38-0590.

6. Reboot the HMC and verify that the HMC software is upgraded correctly by selecting **Software Maintenance** -> **HMC** (see Figure 5-16).

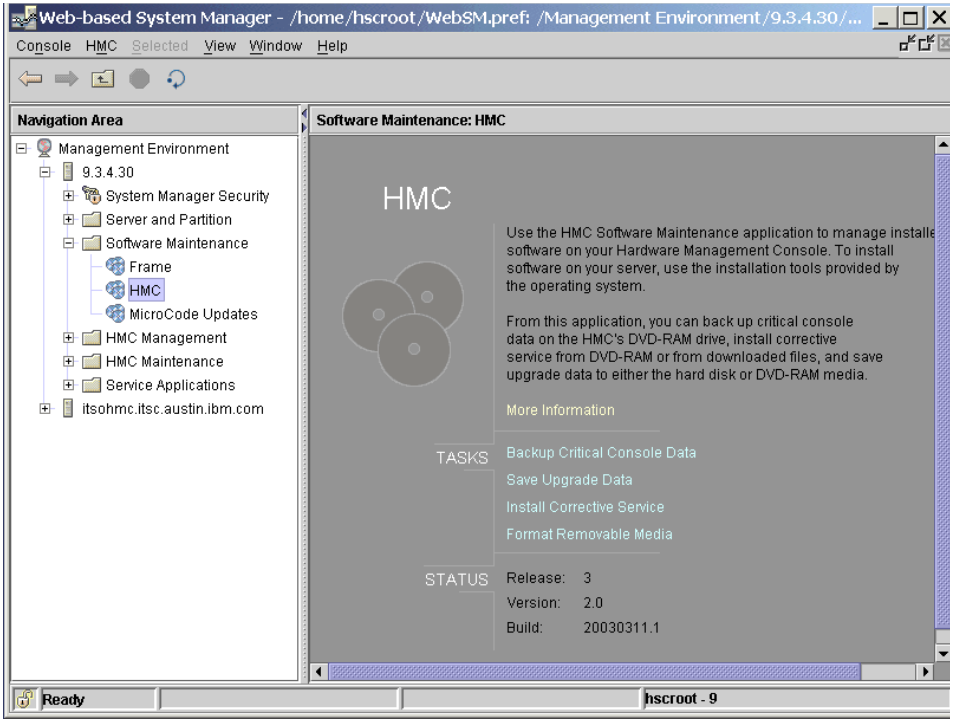


Figure 5-16 Software Maintenance: HMC (2)

5.6.3 Microcode Updates function

There are several aspects to the overall microcode management strategy which include the survey, distribution and installation of microcodes. In the past, it would required an IBM Customer Engineer (CE) to go to the customer's site to survey and update the microcodes. With this Microcode Updates function available in the HMC and AIX, the customer can now manage their microcodes without the need for a CE.

Figure 5-17 on page 185 shows how the mechanism of the microcode update works from a stand-alone AIX server and from an HMC. A single application called Inventory Scout is used to manage the microcodes from both AIX and HMC. The application is able to survey and report the existing and latest microcode levels, and provide suggested actions for each device. It also provides

the capability for installation of microcodes and warns the user before a reboot or installation of back level codes. With this, the customer is able to keep current on the microcode levels at their convenience.

Note: In order to have the Microcode Updates function:

- For a stand-alone AIX server, the operating system must be either be 5100-04 Recommended Maintenance Level, 5200-01 Recommended Maintenance Level, or later.
- For HMC, the HMC software must be V1.3.2 or later.

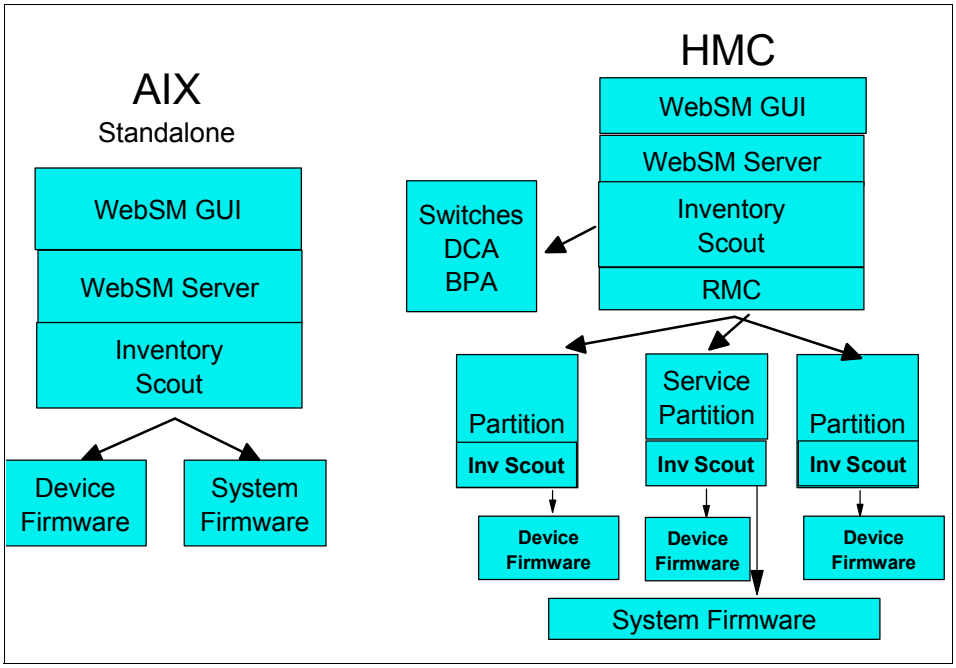


Figure 5-17 Mechanism of the Microcode Updates

Here are the steps to survey and install the latest microcode levels from the HMC:

1. Select **Software Maintenance** -> **MicroCode Updates**, as shown in Figure 5-18.

Figure 5-18 Microcode Updates

2. Select **Change Location** if you want to change from the default Web site Service location, as shown in Figure 5-19.

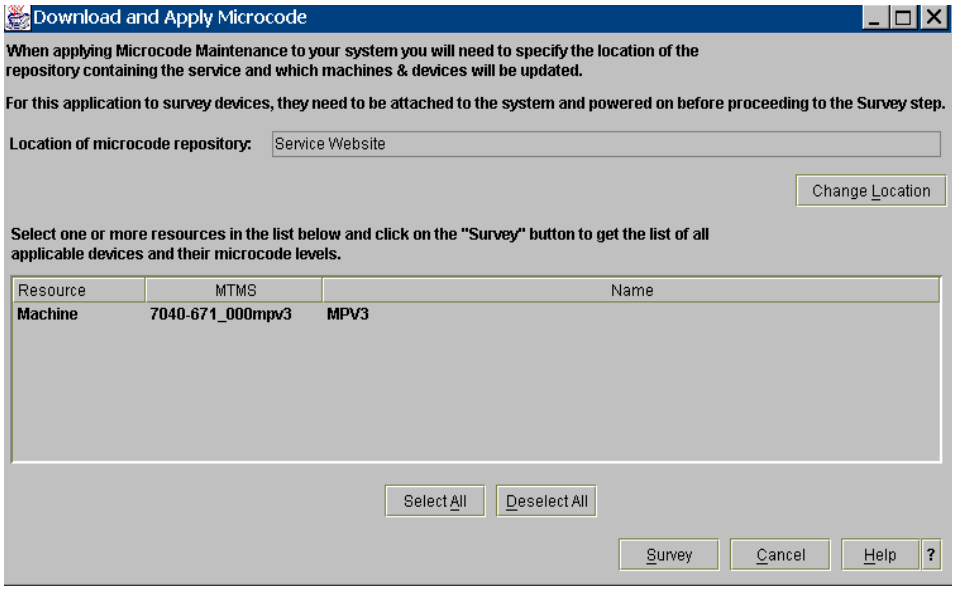


Figure 5-19 Download and Apply Microcode

3. Select the location where the latest microcode levels are stored and select **OK** as shown in Figure 5-20.

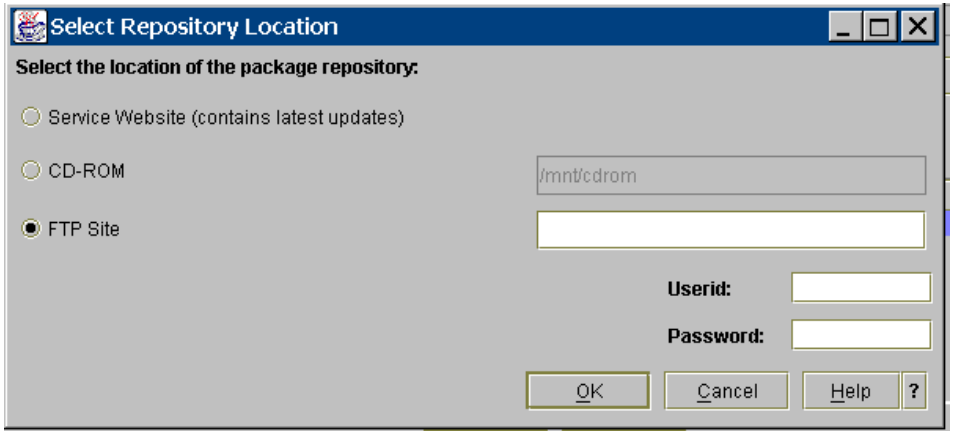


Figure 5-20 Select Repository Location

4. After entering the location, highlight the systems that you want to survey and select **Survey**. When the survey is completed, a summary of all the devices, their current and latest microcode levels, effects of the updates and suggested actions will be displayed as shown in Figure 5-21.

Note: For any selected devices, based on the “Effect” column:

- “Take offline” - The user must take devices offline prior the update of the microcode, otherwise the update will fail.
- “Reboot” - The user must confirm that he understands that the system will reboot as a result of the microcode updates. Applications should be stopped, all users should be notified and the non service partitions should be shut down.

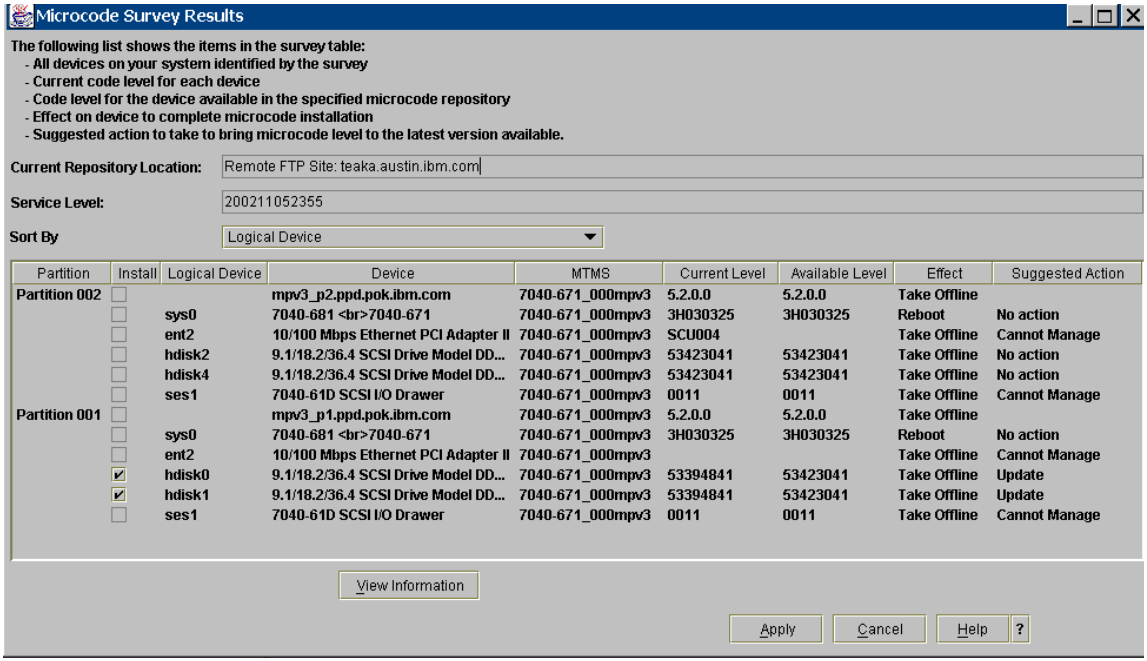


Figure 5-21 Microcode Survey Results

5. It is also possible to view the details for each device. Highlight the device to be viewed and select **View Information**. The details of the device will then be shown as in Figure 5-22.

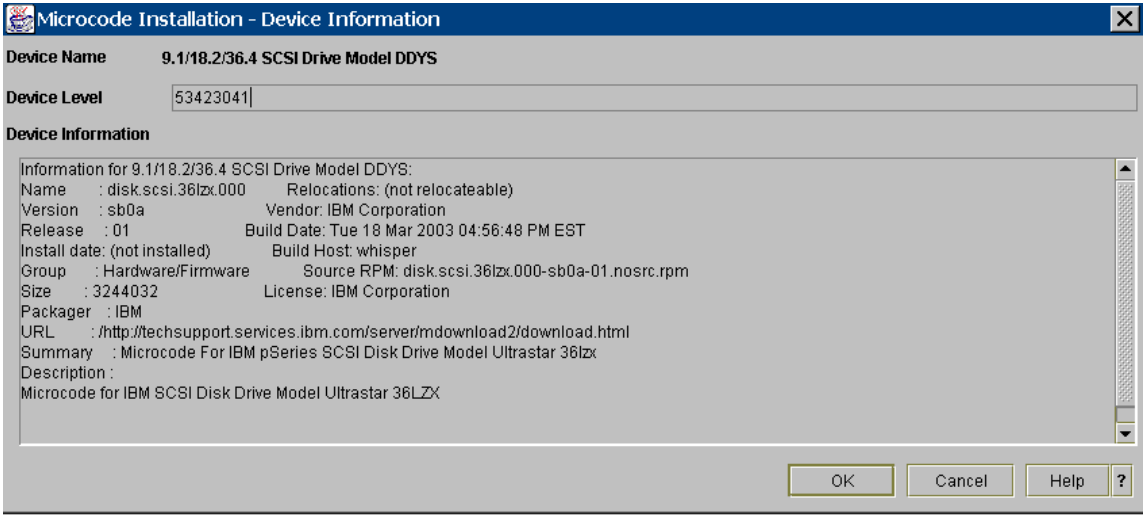


Figure 5-22 Microcode Installation - Device Installation

Note: It is only possible to view one device at a time.

6. The user can either update all or some of the devices by selecting on the “Install” check box on the device line(s) and click **Apply** as shown in Figure 5-21.

   A warning message appears as in Figure 5-23 to warn you about the effects of the updates for system firmware updates. The user can either proceed with the update by selecting **OK** or abort the update by selecting **Cancel**.

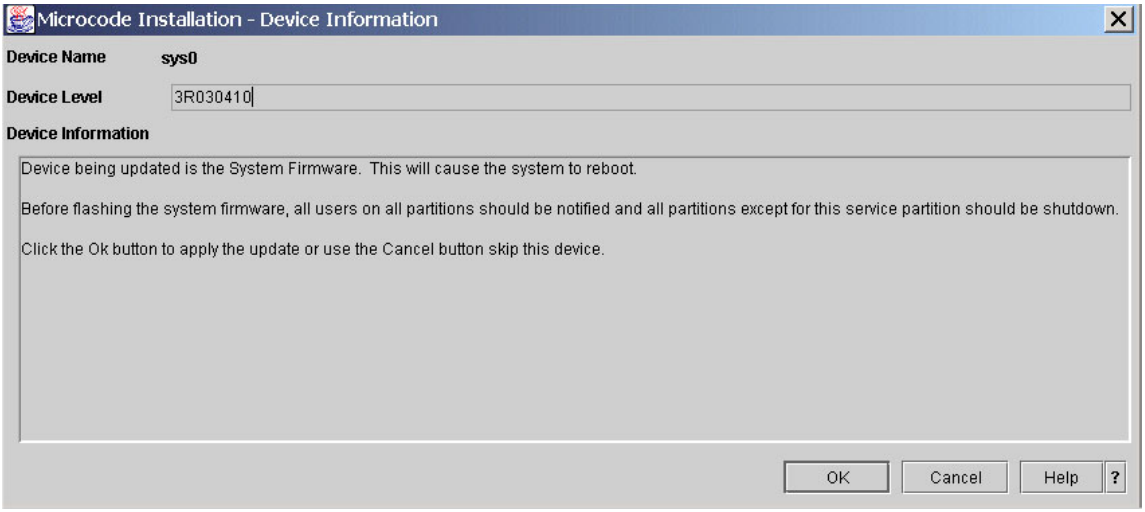


Figure 5-23 Warning message

7. You can monitor the Microcode Updates process from a menu as shown in Figure 5-24.

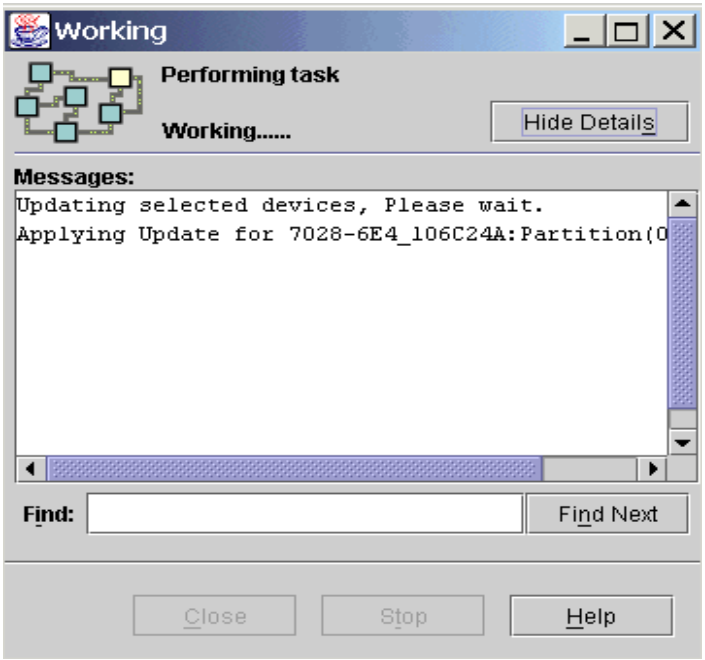


Figure 5-24 Updating microcodes

8. When the microcode update is completed, a window appears as shown in Figure 5-25. The system will automatically reboot if required.

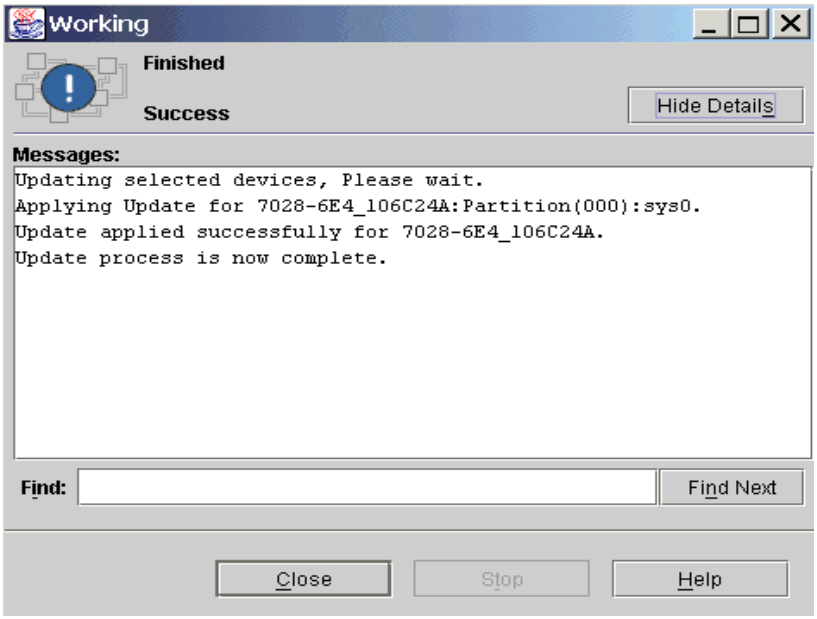


Figure 5-25 Microcode Updates completed

5.6.4 Inventory Scout Services

Inventory Scout Services are pSeries tools that perform the following two functions:

- Microcode Discovery Service

  This generates a real-time comparison report showing subsystems that may need to be upgraded. This microcode survey function is similar to the microcode survey feature discussed in 5.6.3, “Microcode Updates function” on page 184, but without the updates capability. For further information about Microcode Discovery Service, visit the following URL:

  `http://techsupport.services.ibm.com/server/aix.invscoutMDS`

- VPD Capture Service

  This transmits your server’s vital product data (VPD) information to IBM. For further information about VPD Capture Service, visit the following URL:

  `http://techsupport.services.ibm.com/server/aix.invscoutVPD`

You can perform these operations by selecting the following two tasks, as shown in Figure 5-26:

- Conduct Microcode Survey

- Collect VPD Information

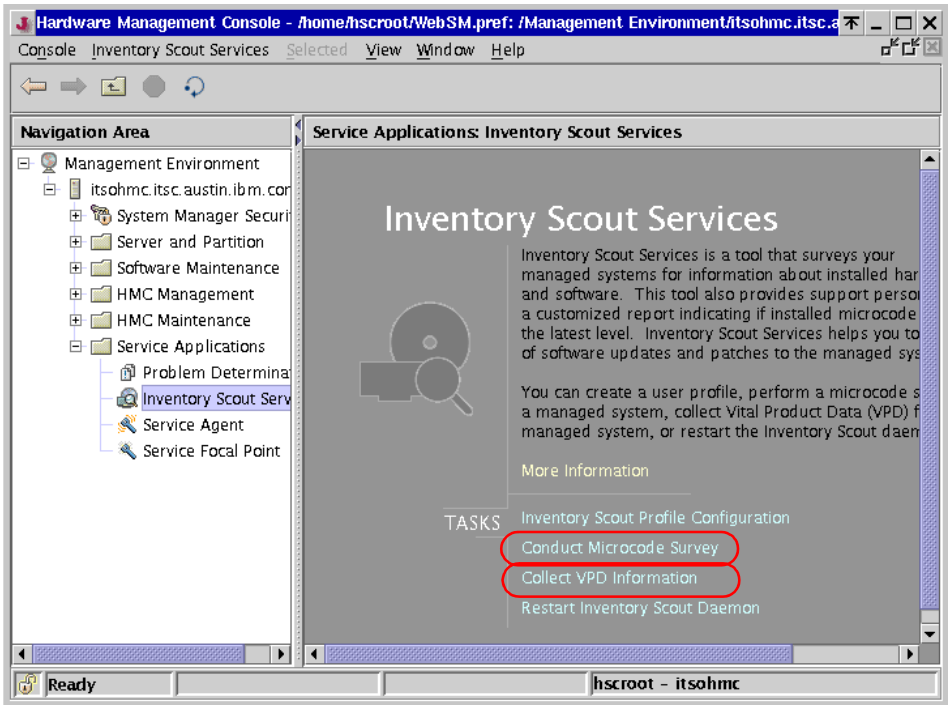


Figure 5-26 Inventory Scout Services

By selecting these two tasks on the HMC, you can check the managed system for needed microcode updates and collect vital product data (VPD) from partitions or the Full System Partition. If Service Agent is configured on the HMC, these files are automatically sent to IBM. You can optionally save files on DOS-formatted diskettes on the HMC diskette drive.

AIX 5L Version 5.2 includes the following Inventory Scout file sets[5]:

```
# lslpp -L invscout.*
  Fileset                      Level  State  Type  Description (Uninstaller)
  ----------------------------------------------------------------------------
  invscout.ldb               1.5.0.0    C     F    Inventory Scout Logic
                                                   Database
  invscout.msg.en_US.rte     1.2.0.0    C     F    Inventory Scout Messages -
                                                   U.S. English
  invscout.rte               1.5.0.0    C     F    Inventory Scout Runtime
```

You can invoke the **invscout** command on AIX using the command line interface to Inventory Scout, as shown below.

[5] AIX 5L Version 5.1 includes a different level of file sets.

```
# /usr/sbin/invscout

******  Command  ----  V1.5.0.0
******  Logic Database V1.5.0.0

Initializing ...
Identifying the system ...
Getting system firmware level(s) ...
Scanning for device firmware level(s) ...

64 devices detected; each dot (.)
represents 10 devices processed:
......

Writing Microcode Survey upload file ...

Microcode Survey complete

The output files can be found at:
Upload file: /var/adm/invscout/lpar04.mup
Report file: /var/adm/invscout/invs.mrp

To transfer the invscout 'Upload file' for microcode
comparison, see your service provider's web page.
```

In fact, this is equivalent to the Conduct Microcode Survey task on the HMC, and the command examines the microcode levels of all hardware components visible in the partition, then accumulates the results in the microcode survey upload file. The brief usage of **invscout** is shown below.

```
# invscout -h

Usage:
invscout [-v] [-c] [-r] [-m] type-model [-s] serial [-g] [-q] [-k] [-h]

-v  -- Change the survey type from
       'Microcode' to 'VPD'
-c  -- Change the main action from 'perform a
       new survey' to 'concatenate existing
       survey upload files'
-r  -- For a new Microcode Survey, sends a
       copy of the formatted text report
       file to the screen from which the
       command was invoked
-m  -- Allows input of the platform machine
       type and model for VPD surveys
-s  -- Allows input of the platform serial
       number for VPD surveys
-g  -- Displays the versions of this command
```

```
       and the logic database currently in
       use
-q  -- Suppresses most run-time messages
-k  -- Keeps temporary files when the
       command is complete
-h  -- Generates this usage statement
```

On the pSeries 670 and pSeries 690 managed by an HMC, the accumulated file can be sent to IBM using the following steps:

1. The file is transferred to the HMC from the partition.
2. The file is sent to IBM from the HMC.

Transferring files to the HMC from partitions

If the following condition is met, files will be automatically transferred to the HMC from the partition after the initial *automatic configuration* is made on the HMC.

- The firmware level of pSeries 670 or pSeries 690 is 10/2002 system microcode update or later.
- The software level of HMC is Release 3 or higher.
- The partition is installed with either of the following:
  - AIX 5L Version 5.1 with 5100-03 Recommended Maintenance Level
  - AIX 5L Version 5.2

Otherwise, Inventory Scout uses its own authentication method between the partition and the HMC in order to talk to the Inventory Scout daemon (invscoutd) on AIX. Therefore, it requires the following additional setup on AIX in a partition:

- A user, invscout, must be defined on the partition.
- A password (for example, invscout) must be set for the invscout user.

To configure Inventory Scout on partitions, do the following:

1. In the Navigation area, click the **Service Applications** icon.
2. In the Contents area, double-click the **Inventory Scout Services** icon.
3. In the Contents area, click **Inventory Scout Profile Configuration**, shown in Figure 5-26. The Inventory Scout Configuration Assistant window will open.
4. Select the managed system you want to configure and click **Next**.

5. A list of partitions, along with each partition's configuration status, will be displayed, as shown in Figure 5-27.

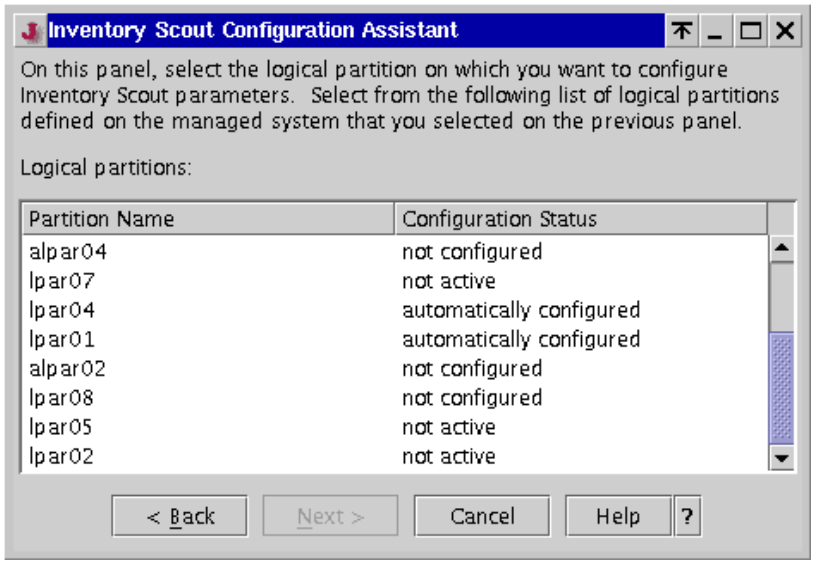


Figure 5-27 Inventory Scout Configuration Assistant

If one of active partitions shows `automatically configured`, then the partition is already configured. If one of the active partitions shows `not configured`, instead of `automatically configured`, you must manually configure the Inventory Scout service for that partition. To configure, select the partition (shown in Figure 5-27) and click **Next**, then enter the following information (the host name or IP address of the partition will be filled automatically):

- The password of the invscout user on that partition
- The listening port of invscoutd (default value is 808)

5.6.5 Service Agent

Electronic Service Agent (also known as Service Agent) is an application program that runs on either AIX or Linux[6] to monitor the system for hardware errors. On pSeries systems managed by the HMC, the primary path for system hardware errors detection and analysis consists of the diagnostics function provided by AIX, the service processor, and the Service Focal Point (see "Service Agent" on page 195). Service Agent provides the transport facility to IBM.

[6] Service Agent supports the Linux operating system on HMC only.

Service Agent can execute several tasks, including:

- Automatic problem analysis
- Problem-definable threshold levels for error reporting
- Automatic problem reporting
- Automatic customer notification
- Visualize hardware error logs

Although there are several scenarios available for the Service Agent network configuration, we only explain the configuration, which is used on the HMC with managed systems, shown in Figure 5-28.

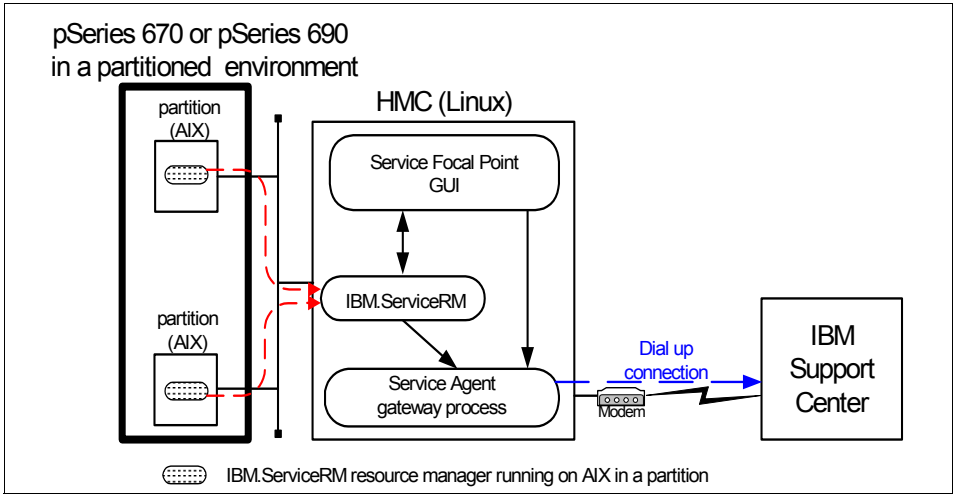


Figure 5-28 Service Agent on the HMC

In this configuration, the Service Agent gateway process running on the HMC places service calls to the IBM support center via a dial-up connection with the attached modem, if necessary.

Note: No human intervention is required for this process.

By utilizing Service Agent, the pSeries 670 and pSeries 690 can reduce the amount of downtime experienced in the event of a system component failure by giving the service provider the ability to view the error report entry and, if needed, order any necessary replacement parts prior to arriving on site. The opportunity for human misinterpretation or miscommunication in problem determination is therefore mitigated.

Figure 5-29 shows the initial configuration panel of Service Agent.

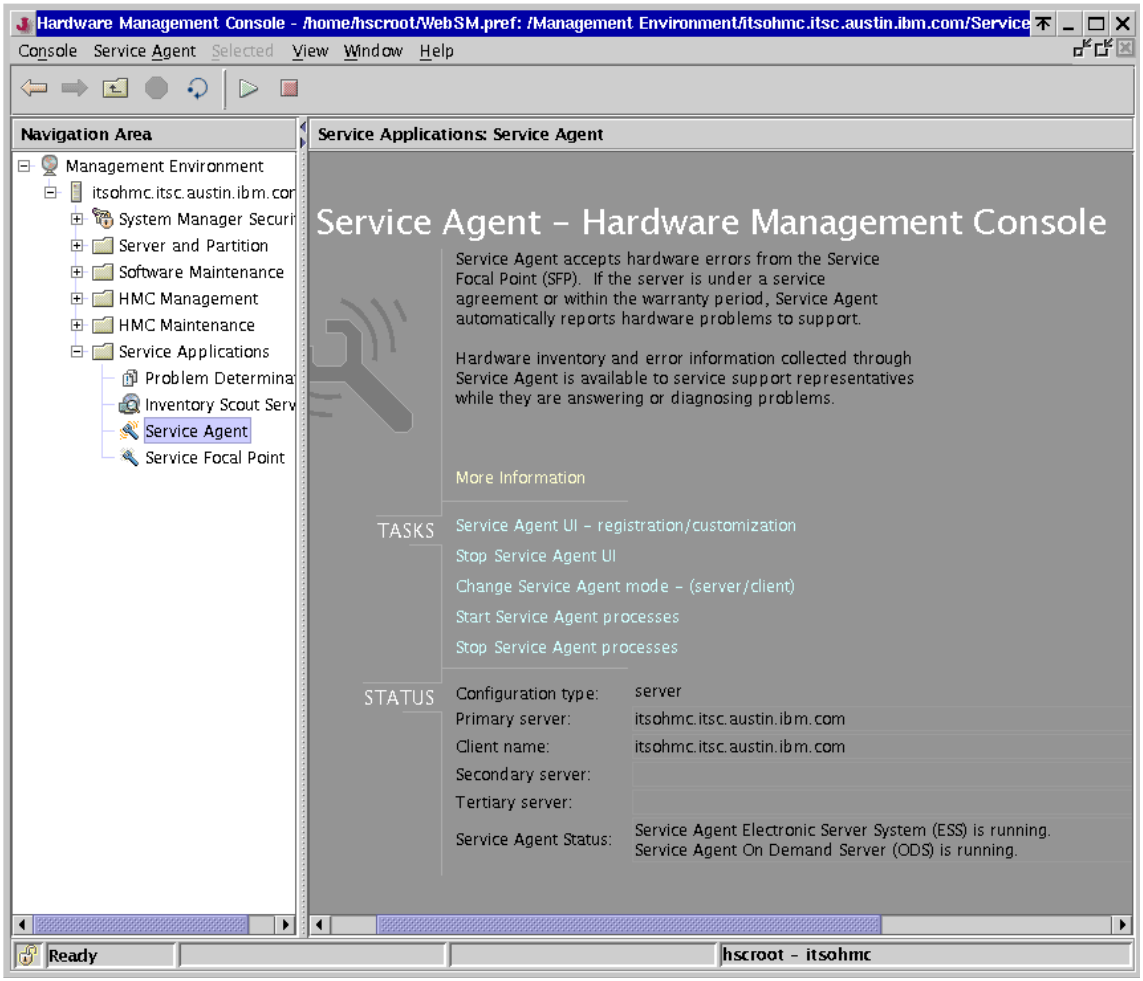


Figure 5-29 Service Agent on the HMC

For more information on the Service Agent, please refer to the following publications:

- *Electronic Service Agent for pSeries and RS/6000 User's Guide*, available at:

  `ftp://ftp.software.ibm.com/aix/service_agent_code/AIX/svcUG.pdf`

- *Electronic Service Agent for pSeries Hardware Management Console User's Guide*, available at:

  `ftp://ftp.software.ibm.com/aix/service_agent_code/HMC/HMCSAUG.pdf`

- *IBM Hardware Management Console for pSeries Installation and Operations Guide*, SA38-0590

5.6.6 Service Focal Point

Traditional service strategies become more complicated in a partitioned environment. Each partition runs on its own, unaware that other partitions exist on the same system. If one partition reports an error for a shared resource, such as a managed system power supply, other active partitions report the same error. To enable service representatives to avoid long lists of repetitive call-home information, the HMC provides the Service Focal Point application. Service Focal Point recognizes that these errors repeat, and filters them into one *serviceable event* for the service representative to review.

The Service Focal Point is a system infrastructure on the HMC that manages serviceable event information for the system building blocks. It includes resource managers that monitor and record information about different objects in the system. It is designed to filter and correlate events from the resource managers and initiate a call to the service provider when appropriate. It also provides a user interface that allows a user to view the events and perform problem determination.

Note: Service Focal Point only collects hardware errors, such as *PERMANENT* errors from AIX (marked as P) and NON BOOT errors from the service processor.

Upon hardware failure events, the corresponding error entry is notified from the partition to the HMC, as shown in Figure 5-11 on page 180. The IBM.ServiceRM subsystem is in charge of this notification. The AIX diagnostic function creates a serviceable event through IBM.ServiceRM when a hardware problem is determined and events will be notified to the HMC using the *Resource Monitoring and Control (RMC)* framework. The IBM.ServiceRM is running as the IBM.ServiceRMd daemon and packaged in the devices.chrp.base.ServiceRM file set in AIX, as shown below.

```
# lssrc -g rsct_rm | head -1; lssrc -g rsct_rm | grep ServiceRM
Subsystem         Group            PID          Status
 IBM.ServiceRM    rsct_rm          307354       active
# ps -ef | head -1; ps -ef | grep ServiceRM | grep -v grep
 UID    PID   PPID   C    STIME    TTY  TIME CMD
 root 307354 122982   0   Sep 11      -  0:31 /usr/sbin/rsct/bin/IBM.ServiceRMd
# lslpp -w /usr/sbin/rsct/bin/IBM.ServiceRMd
  File                                      Fileset               Type
  ----------------------------------------------------------------------------
  /usr/sbin/rsct/bin/IBM.ServiceRMd
                                devices.chrp.base.ServiceRM        File
# lslpp -L devices.chrp.base.ServiceRM
  Fileset                     Level  State  Type  Description (Uninstaller)
  ----------------------------------------------------------------------------
```

```
  devices.chrp.base.ServiceRM
                             1.2.0.0    C     F    RSCT Service Resource
Manager
```

From the Service Focal Point interface, you can execute maintenance procedures such as examining the error log history, checking for components requiring replacement, and performing a Field Replaceable Unit (FRU) replacement. If Service Agent is configured on the HMC, the serviceable events are automatically sent to IBM (call-home support) for automatic generation of a maintenance request.

How to use Hardware Service Function

This function allows you either to just identify a frame when you have several frames connected to your HMC, or to turn off the rack indicator light. You are also able to get a Field Replaceable Unit (FRU) list when the rack indicator light is lit and check which component has problems. When a component is shown here with the LED state ON, it is much easier to identify the failing component.

To use this function, do the following:

1. Click the **+** mark left of the “Service Applications” in the Navigation Area.
2. Select Service Focal Point in the Navigation Area. You will see the Service Focal Point task panel, as shown in Figure 5-30.
3. Select **Hardware Service Functions** (Figure 5-30).

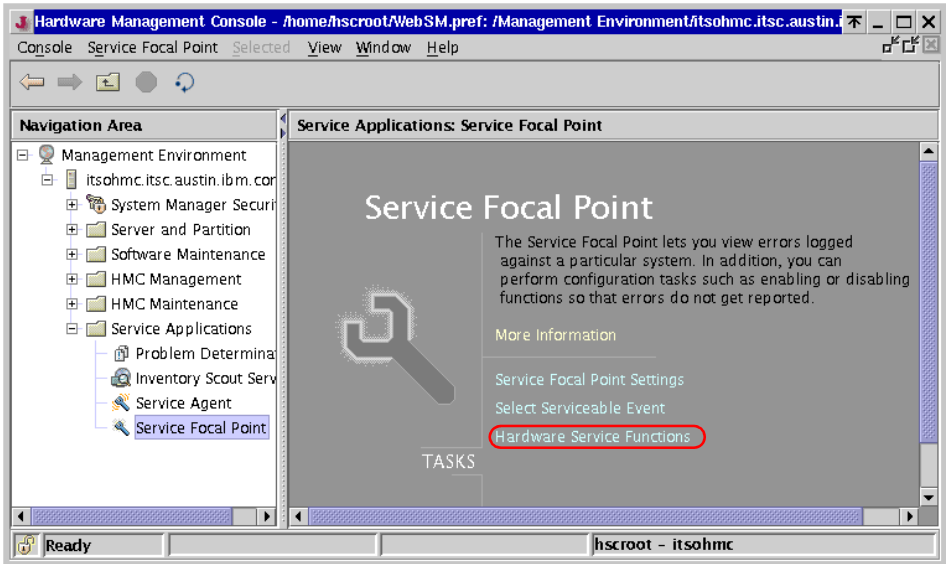


Figure 5-30 Service Focal Point: Hardware Service Functions

4. You will see the Hardware Service Management: Overview window, as shown in Figure 5-31. Select the managed system for which you want to check the LED state, then select the **List FRUs** button.

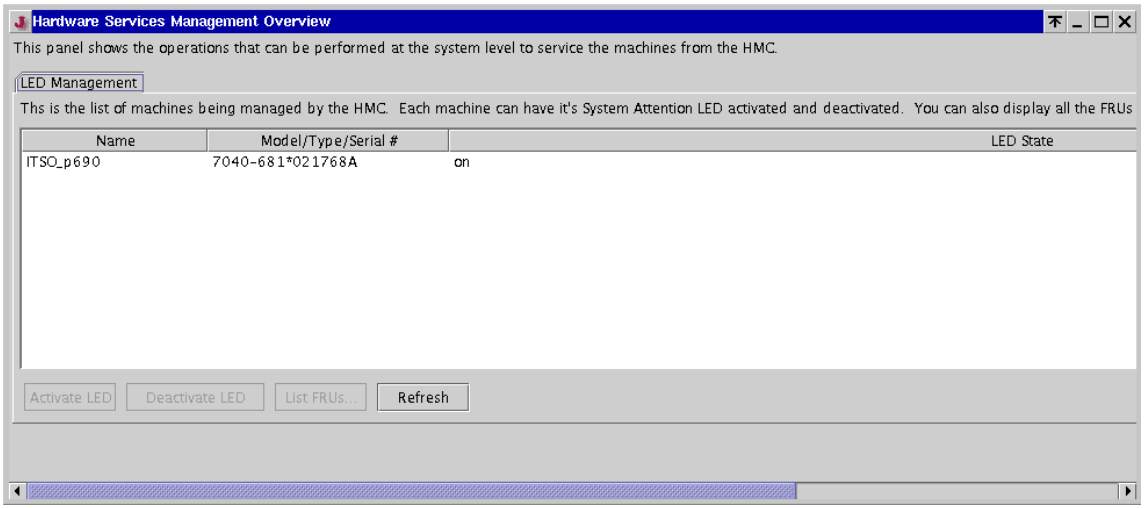


Figure 5-31 Hardware Service Functions overview

5. You will see the FRU LED Management window (Figure 5-32).

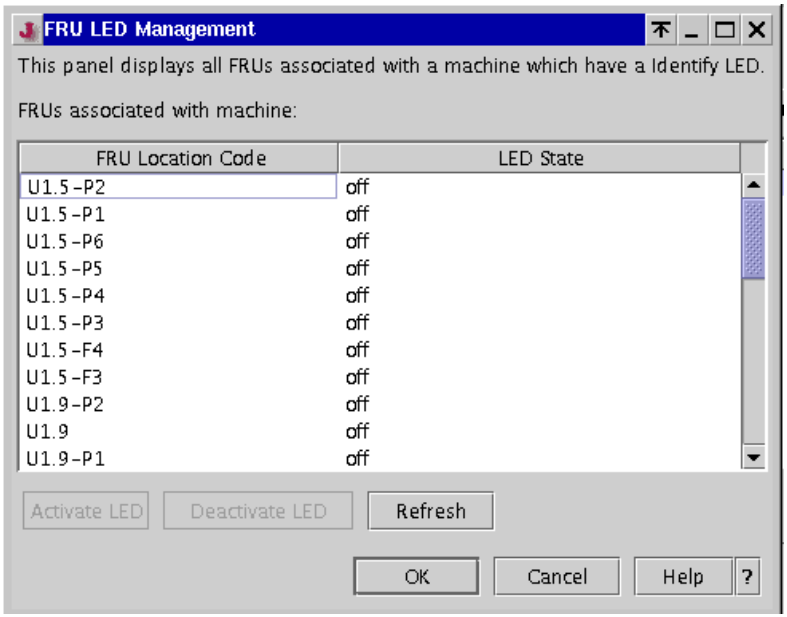


Figure 5-32 FRU LED Management

If any of the LEDs are ON, it would mean that the system has a problem with the indicated component. If the Service Agent is configured to notify IBM of the errors to IBM, then IBM customer service representatives will be informed of the problem.[7]

For more information, please refer to *IBM Hardware Management Console for pSeries Installation and Operations Guide*, SA38-0590.

5.6.7 Problem determination hints of Service Functions

To prevent problems with the configuration of Service Functions on the HMC, such as Inventory Scout, Service Agent, and Service Focal Point, the Ethernet network between the HMC and partitions, there should be a mandatory administrative network used by only these applications.

With careful network planning, you should not have any problems using these applications; however, if an AIX administrator mistakenly changes the TCP/IP configuration on a partition without notifying the HMC administrator, it might result in severe communication problems.

Authentication mechanism

The Service Focal Point and DLPAR functions rely on the RMC framework between the HMC and partitions. The RMC framework performs not only session management, but also authentication between network peers.

The ctcas subsystem, also known as the *cluster authentication* daemon, is in charge of this authentication mechanism. It is running as the ctcasd daemon and is packaged in the rsct.core.sec file set in AIX, as shown below.

```
# lssrc -g rsct
Subsystem         Group            PID          Status
 ctrmc            rsct             299204       active
 ctcas            rsct             188658       active
# ps -ef | head -1; ps -ef | grep ctcas | grep -v grep
     UID    PID   PPID   C    STIME    TTY  TIME CMD
    root 188658 139350   0   Sep 11      -  0:03 /usr/sbin/rsct/bin/ctcasd
# lslpp -w /usr/sbin/rsct/bin/ctcasd
  File                                      Fileset               Type
  ----------------------------------------------------------------------------
  /usr/sbin/rsct/bin/ctcasd                 rsct.core.sec         File
```

The configuration process of authentication between the HMC and partitions can be briefly summarized as shown in Table 5-2.

[7] In order to dispatch IBM customer service representatives, you need the maintenance agreement (MA) for this system.

Table 5-2 Authentication process

Sequence	On the HMC	On an AIX partition
1	The DMSRM resource manager places the secrete key and the HMC host name in the NVRAM of the managed system. Every reboot of the HMC, it places a new secret key.	
2		The IBM.CSMAgentRM resource manager reads the secret key and the HMC host name from NVRAM using an RTAS call. The NVRAM is checked every five minutes to detect new HMC(s) and/or key changes. An existing HMC with a changed key will cause the registration process (the next 2 steps) to be performed again.
3		Once the HMC and LPAR have authenticated each other using the secret key and have exchanged some information about each other, (for example, public keys), IBM.CSMAgentRM grants the HMC permission to access the necessary resource classes on the partition. Without proper permission on AIX, the HMC will be able to establish a session with the partition, but will not be able to query for the operating system information, such as DLPAR capabilities, or execute DLPAR operation commands afterward.
4		The last part of the registration process is the creation of an IBM.ManagedNode resource with a Hostname attribute set to the partition's host name on the HMC. Then, an IBM.ManagementServer resource will be created with a Hostname attribute set to the HMC's host name on the partition.
5	After the ManagedNode resource is created and authenticated, the ServiceRM and LparCmdRM resource managers establish a session with the partition for DLPAR operation and receive serviceable events.	

Note: The current implementation of authentication mechanism used in the RMC framework is called *UNIX hostname authentication*. The RMC, and therefore the HMC, may implement new authentication mechanisms in accordance with the future development plan of RMC.

Trouble-free network planning rules

To avoid unnecessary configuration errors in Service Function applications, you must understand the following rules:

- All the combinations of a host name and an IP address must be unique.
- All the network interfaces on the HMC and partitions must be assigned different host names and, therefore, different IP addresses.
- The assigned IP address must be consistently resolved regardless the location (on the HMC or partitions). If some name services, such as NIS, DNS, and LDAP, are used, they must be reliable and return the consistent results.
- The network interface on the HMC, which is resolved to the node name (the string returned from the **hostname** command), must be reachable from all the partitions.

The following examples show inappropriate network configurations.

- Duplicate IP addresses

  Two partitions have different host names, but the same IP address on their network interface.

- Unresolvable host name

  A partition does not have the valid DNS configuration, while the HMC uses DNS for the name resolution. The partition cannot resolve the HMC's host name to an IP address (unresolvable).

- Inconsistent name resolution

  The HMC is assigned the fully qualified domain name (FQDN) itsohmc.itsc.austin.ibm.com for both node name and the host name for eth0 interface. An AIX partition uses DNS for the name resolution, but there are the following files on the partition:

```
# cat /etc/netsvc.conf
hosts=local,bind
# grep itsohmc /etc/hosts
9.3.4.30       itsohmc    itsohmc.itsc.austin.ibm.com
```

Therefore, the same IP address 9.3.4.30 is resolved as:

On the HMC itsohmc.itsc.austin.ibm.com

On the partition itsohmc

- Unreachable network interface

  The HMC has two network interfaces, eth0 and eth1. Although the FQDN itsohmc.itsc.austin.ibm.com is assigned for both node name and host name for the eth0 interface, all partitions can reach to eth1 interface only.

We strongly suggest that you do the following before doing any recovery activities:

1. Issue the **hostname** command on the HMC and all partitions. To issue the **hostname** command on the HMC, you can use OpenSSH, as shown in the following example:

   ```
   $ whence ssh
   /usr/bin/ssh
   $ ssh -l hscroot itsohmc.itsc.austin.ibm.com hostname
   hscroot@itsohmc.itsc.austin.ibm.com's password: XXXXXX
   itsohmc.itsc.austin.ibm.com
   ```

   For further information about how to use OpenSSH on AIX, please refer to *Managing AIX Server Farms*, SG24-6606.

2. Issue the **host** command against all of the network interfaces on the HMC and all of the partitions.

   a. Confirm how many interfaces are available.

      ```
      $ ssh -l hscroot itsohmc.itsc.austin.ibm.com\
          "/sbin/ifconfig -l | grep Link"
      hscroot@itsohmc.itsc.austin.ibm.com's password: XXXXXX
      eth0      Link encap:Ethernet  HWaddr 00:02:55:13:85:2E
      lo        Link encap:Local Loopback
      ```

   b. Confirm the IP address of eth0.

      ```
      $ ssh -l hscroot itsohmc.itsc.austin.ibm.com /sbin/ifconfig eth0
      hscroot@itsohmc.itsc.austin.ibm.com's password: XXXXXX
      eth0      Link encap:Ethernet  HWaddr 00:02:55:13:85:2E
                inet addr:9.3.4.30  Bcast:9.3.5.255  Mask:255.255.254.0
                UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
                RX packets:1256676 errors:0 dropped:0 overruns:0 frame:7
                TX packets:1381966 errors:0 dropped:0 overruns:0 carrier:13
                collisions:404844 txqueuelen:100
                RX bytes:132305448 (126.1 Mb)  TX bytes:1048151698 (999.5 Mb)
                Interrupt:10 Base address:0x5000
      ```

c. Confirm both the reverse and regular name resolutions.

```
# ssh -l hscroot itsohmc.itsc.austin.ibm.com host 9.3.4.30
hscroot@itsohmc.itsc.austin.ibm.com's password: XXXXXX
30.4.3.9.in-addr.arpa. domain name pointer itsohmc.itsc.austin.ibm.com.
# ssh -l hscroot itsohmc.austin.ibm.com host itsohmc.itsc.austin.ibm.com
hscroot@itsohmc.itsc.austin.ibm.com's password:
itsohmc.itsc.austin.ibm.com. has address 9.3.4.30
```

Changing the host name manually

To change the host name on your partition manually without having any problems with Service Functions applications, do the steps below on the partition.

Note: These steps are not applicable on the partition installed with AIX 5L Version 5.1 with 5100-03 Recommended Maintenance Level or AIX 5L Version 5.2.

The steps are:

1. If you are running AIX 5L Version 5.1 in the partition without applying 5100-03 Recommended Maintenance Level, then do the following:

   ```
   # /usr/sbin/rsct/bin/runact -c IBM.ManagementServer SetRTASPollingInterval\
       Seconds=0
   ```

2. Run the following command:

   ```
   # /usr/sbin/rsct/bin/lsrsrc IBM.ManagementServer Hostname
   ```

   You will receive output similar to the following example (there may be multiple entries if the partitions are managed by multiple HMCs):

   ```
   Resource Persistent Attributes for: IBM.ManagementServer
   resource 1:
           Hostname = "itsohmc.itsc.austin.ibm.com"
   ```

3. The **rmrsrc** command is now used to delete the resources using the host name shown before. Repeat this step for each entry (using its Hostname) and verify none remain with the command shown in step 2.

   ```
   # /usr/sbin/rsct/bin/rmrsrc -s'Hostname = "itsohmc.itsc.austin.ibm.com"'\
       IBM.ManagememementServer
   ```

4. Stop the RMC subsystem using the **rmcctrl** command:

   ```
   # /usr/sbin/rsct/bin/rmcctrl -z
   ```

5. Change the partition host name.

6. Restart the RMC subsystem:

   ```
   # /usr/sbin/rsct/bin/rmcctrl -A
   0513-071 The ctrmc Subsystem has been added.
   ```

```
0513-071 The ctcas Subsystem has been added.
0513-059 The ctrmc Subsystem has been started. Subsystem PID is 19628.
```

You can verify with the command in step 2 whether the host name of the HMC is filed again in the resource configuration database.

For more information on the Inventory Scout Services, please refer to *IBM Hardware Management Console for pSeries Installation and Operations Guide*, SA38-0590.

For further information about the RMC framework and its resource managers, please refer to the following publications:

- *A Practical Guide for Resource Monitoring and Control (RMC)*, SG24-6606
- *IBM Reliable Scalable Cluster Technology for AIX 5L RSCT Guide and Reference*, SA22-7889
- *IBM Reliable Scalable Cluster Technology for AIX 5L Messages*, GA22-7891
- *IBM Reliable Scalable Cluster Technology for AIX 5L Technical Reference*, SA22-7890

5.7 AIX RAS features

The RAS features described in the previous sections are all based on the pSeries 670 and pSeries 690 hardware. There are some additional RAS features that work in conjunction with the AIX operating system, and some that are entirely dependent on AIX. These features are covered in the following sections.

5.7.1 Unrecoverable error analysis

In prior RS/6000 and pSeries systems, data corruption due to parity or cyclic redundancy check (CRC) errors on buses as well as ECC uncorrectable errors in all levels of cache and memory usually resulted in a machine crash. The pSeries 670 and pSeries 690 adds a new functionality so that these errors are captured and sent to the partition affected by the failed resources, and provide the operating system with the error code for analysis.

AIX analyzes the results of the firmware recovery attempt and determines which process is corrupted by the uncorrectable hardware error. If the process is a user process, it will be terminated. If the process is within the AIX kernel, then the operating system will be terminated. Again, terminating AIX in a partition will not impact any of the other partitions.

5.7.2 System hang detection

AIX 5L Version 5.1 and higher offers a feature called *system hang detection* that provides a mechanism to detect system hangs and initiates a pre-configured action. It relies on a new daemon named shdaemon, and a corresponding configuration command named **shconf**.

In a case where applications adjust their processes or thread priorities using system calls, there is a potential problem that their priorities will become so high that regular system shells are not scheduled. In this situation, it is difficult to distinguish a system that really hangs (it is not doing any meaningful work anymore) from a system that is so busy that none of the lower priority tasks, such as user shells, have a chance to run.

The new system hang detection feature uses a shdaemon entry in the /etc/inittab file with an *action* field that specifies what should be done when certain conditions are met. The actions can be, for example, generating an entry in the error log, displaying a warning message, running a predefined command, opening a high-priority login prompt on tty0, or restarting the system.

For more information on how to configure this feature, please refer to *Managing AIX Server Farms*, SG24-6606.

5.7.3 AIX disk mirroring and LVM sparing

Mirroring the operating system boot disks (called *rootvg mirroring*) is a feature available on AIX since Version 4.2.1. It enables continued operation of the system, even in the case of failure on an operating system disk. Two or three copies can be created and accessed in case of failure.

Beginning with AIX 5L Version 5.1, it is possible to designate disks as hot spare disks in a volume group and to specify a policy to be used in the case of failing disks. This enables fast recovery for mirrored disks when a failure occurs, using the automatic sparing option. As soon as one mirrored disk fails, data is copied to the spare disk or to a pool of disks. When the copy finishes, the system is again protected against disk failures.

For detailed information on rootvg mirroring on AIX, please refer to *AIX Logical Volume Manager, From A to Z: Introduction and Concepts*, SG24-5432. For LVM hot spare disk support in a volume group, please refer to *AIX 5L Differences Guide,* SG24-5765.

5.7.4 TCP/IP RAS enhancements

Beginning with AIX 5L Version 5.1, AIX provides several availability features in the TCP/IP network access. It is possible to completely eliminate failures due to network adapters, failed gateways, and incorrect routes, using these functions.

The *Dead Gateway Detection (DGD)* feature in AIX 5L Version 5.1 implements a mechanism for hosts to detect a dysfunctional gateway, adjust its routing table accordingly, and re-route network traffic to an alternate backup route, if available.

With the *multi-path routing* feature in AIX 5L Version 5.1, routes no longer need to have a different destination, netmask, or group ID lists. If there are several routes that equally qualify as a route to a destination, AIX will use a cyclic multiplexing mechanism (round-robin) to choose between them. The benefit of this feature is twofold:

- Enablement of load balancing between two or more gateways.
- Feasibility of load balancing between two or more interfaces on the same network can be realized.

These two features working together enable the system to route TCP/IP traffic through multiple routes and avoid the defective routes. This is as important as the server availability itself, because a network failure also causes an interrupt in server access.

For Ethernet adapters, the *network interface backup* support allows multiple adapters to be grouped together and act as a single interface. In case of a failure in one interface, another interface will automatically take over the TCP/IP traffic and continue operation.

For more information on how to configure this feature, please refer to *Managing AIX Server Farms*, SG24-6606.

A

Minimum and default configurations

This appendix contains the minimum configuration listed in the sales manual as well as the e-config default configuration of pSeries 670 and pSeries 690.

A.1 pSeries 670 configurations

The pSeries 670 least expensive configuration contains the following features listed in Example A-1.

Example: A-1 pSeries 670 minimum configuration

```
One - 7040-671 Central Electronics Complex (17U rack space required)
            One - CD-ROM (#2624)
            One - Diskette Drive Cable (#3155)
            One - Operator Panel Attachment Cable (#3255)
            One - 4 GB Memory Card, Inward Facing (#4196)
            One - 4-Way, POWER4 Processor with 128 MB L3 Cache (#5253)
            One - Processor Clock Card, Programmable (#5251)
            One - Power Cable Group, Bulk Power to CEC and Fans (#6161)
            One - Interface Cable, Power to Service Processor (#6162)
            Two - Power Converter Assembly, Central Electronics Complex (#6170)
            One - Power Cable Group, CEC to Power Controller, First Processor Module (#6181)
            One - Capacitor Book, First and Second Processor Modules (#6198)
            One - Service Processor with Remote I/O Loop Attachment, Two Loops (#6404)
            One - Media Drawer, Operator Panel, Diskette (#8692) - (1U Rack Space Required)
            One - Drawer Placement Indicator, 18U Position, Primary Rack (#4618)
            One - Language Specify (#9xxx)
One - 7040-61D I/O Drawer (4U rack space required)
            One - SCSI Cable - I/O Drawer to Media Drawer (#2122)
            Two - Remote I/O Cables, 2M (#3149)
            Two - 18.2 GB, 10,000 RPM Ultra3 SCSI Disk Drive Assembly (#3157)
            One - I/O Drawer Attachment Cable Group, Drawer Position #4609 (#6121)
            Two - Power Converter Assembly, I/O Drawer (#6172)
            One - Power Cable, I/O Drawer to Media Drawer (#6179)
            One - PCI Single Ended SCSI Adapter (Media Drawer Attach) (#6206)
            One - I/O Drawer PCI Planar, 10 Slot, 2 Integrated Ultra3 SCSI Ports (#6563)
            Two - Ultra3 SCSI 4-Pack Hot Swap Back Plane (#6564)
            One - Drawer Placement Indicator, 9U Position, Primary Rack (#4609)
            One - Language Specify (#9xxx)
One - 7040-61R System Rack (24-inch rack with 42U available space)
            One - Front Door, Black with Copper Accent, Primary Rack (#6070)
            One - Rear Door, Slim Line, Primary or Secondary Rack (#6074)
            Two - Bulk Power Regulators (#6186)
           Two - Power Control, Four Cooling Fans,Three DC Power Converter Connections (#6187)
            Two - Power Distribution Assembly, Ten DC Power Converter Connections (#6188)
            One - Bulk Power Assembly, Redundant (#8690)
            One - Rack Content Specify, 7040/671 - 17U (#0209)
            One - Rack Content Specify, 7040/61D - 4U (#0191)
            One - Rack Content Specify, Media Drawer
            One - Rack Content Specify -7040/61R Feature #8490 - 8U (#0193)
            Two - Line Cords, 60A/240VAC, 6-AWG, 14 feet, IEC309 Plug (#8678)
            One - Language Specify (#9xxx)
```

The default configuration in Example A-2 lacks the mandatory rear door, contains a full RIO planar while the minimum configuration contains a half populated IO planar and contains country (or region) specific features, chosen according to e-config country (or region) settings.

Example: A-2 pSeries 670 default configuration

Product	Description	Qty
7040-671	Rack Server 1:pSeries 670 Model 671 CEC	1
2624	32X (Max) SCSI-2 CD-ROM Drive	1
3155	Diskette Drive Cable	1
3255	Operator Panel Attachment Cable	1
4196	4GB Memory Card, Inward Facing	1
4618	Drawer Placement Indicator, 18U Position, Primary Rack	1
4651	**Rack Indicator, Rack #1**	**1**
5005	**Software Preinstall (RS)**	**1**
5251	Processor Clock Card, Programmable	1
5253	4-Way POWER4 p670 Processor	1
6161	Cable Group, Power Controller to CEC and Fans	1
6162	Interface Cable, Service Processor to Power Subsystem	1
6170	Power Converter Assembly, Central Electronics Complex	2
6181	Power Cable Group, CEC to Power Controller, First Processor Module	1
6198	Capacitor Book, Two Processor Modules	1
6404	Support Processor with Remote I/O Loop Attachment, Two Loops	1
8121	**Attachment Cable, Hardware Management Console to Host, 15-Meter**	**1**
8692	Media Drawer, Operator Panel, Diskette	1
9670	**Manufacturing Optimization Indicator**	**1**
9703	**Language Group: French**	**1**
7040-61D	I/O Drawer 1:Model 61D I/O Drawer	1
2122	SCSI Cable - I/O Drawer to Media Drawer	1
3145	**Remote I/O Cable, 0.5M**	**1**
3149	Remote I/O Cable, 2M	2
3158	36.4 GB DISK DRIVE ASSEMBLY	2
4609	DRAWER PLACEMENT IND, 9U, PRI	1
4651	**Rack Indicator, Rack #1**	**1**
6121	I/O Drawer Attachment Cable Group, Drawer Position #4609	1
6172	Power Converter Assembly, I/O Drawer	2
6179	Power Cable, I/O Drawer to Media Drawer	1

```
    6206  Ultra SCSI Adapter - SE (PCI)                          1
    6563  I/O Drawer PCI Planar, 10 Slot, 2 Integrated           2
          Ultra3 SCSI Ports
    6564  Ultra3 SCSI 4-Pack Hot Swap Back Plane                 4
    9703  Language Group: French                                 1

7040-61R  7040-61R : Rack 1:Model 61R Rack                       1
    0191  Rack Content Specify: 7040/61D - 4U                    1
    0192  Rack Content Specify: Media Drawer Feature             1
          #8692 - 1U
    0193  Rack Content Specify: 7040/61R Feature #8690           1
          - 8U
    0209  Rack Content Specify: 7040/671 - 17U                   1
    4651  Rack Indicator, Rack #1                                1
    6070  Front Door, Black with Copper Accent, Primary          1
          Rack
    6186  Bulk Power Regulator                                   2
    6187  Power Controller, Four Cooling Fans, Three             2
          DC Power Converter Connections
    6188  Power Distribution Assembly                            2
    8677  Line Cord, 8AWG, 14ft, No Plug                         2
    8690  Bulk Power Assembly, Redundant                         1
    9703  Language Group: French                                 1
```

A.2 pSeries 690 configurations

The pSeries 690 least expensive configuration contains the features listed in Example A-3.

Example: A-3 pSeries 690 minimum configuration

```
One - 7040-681 Central Electronics Complex (17U rack space required):
            One - CD-ROM (#2624)
            One - Diskette Drive Cable (#3155)
            One - Operator Panel Attachment Cable (#3255)
            One - 128 MB Level 3 Cache (4 X 32 MB), 400 MHz (#4138)
            One - 8 GB Memory Card, Inward Facing (#4181)
            One - 8-Way, POWER4 Processor (First) (#5242)
            One - Processor Clock Card, Programmable (#5251)
            One - Power Cable Group, Bulk Power to CEC and Fans (#6161)
            One - Interface Cable, Power to Service Processor (#6162)
            Two - Power Converter Assembly, Central Electronics Complex (#6170)
            One - Power Cable Group, CEC to Power Controller, First Processor Module (#6181)
            One - Capacitor Book, First and Second Processor Modules (#6198)
            One - Service Processor with Remote I/O Loop Attachment, Two Loops (#6404)
```

```
        One - Backplane, Central Electronics Complex (#6565)
        One - Media Drawer, Operator Panel, Diskette (#8692) - (1U Rack Space Required)
        One - Drawer Placement Indicator, 18U Position, Primary Rack (#4618)
        One - Language Specify (#9xxx)
    One - 7040-61D I/O Drawer (4U rack space required):
        One - SCSI Cable - I/O Drawer to Media Drawer (#2122)
        One - Remote I/O Cable, 05.M (#3145) (between drawer halves)
        Two - Remote I/O Cables, 2M (#3149)
        Two - 18.2 GB, 10,000 RPM Ultra3 SCSI Disk Drive Assembly (#3157)
        One - I/O Drawer Attachment Cable Group, Drawer Position #4605 (#6121)
        Two - Power Converter Assembly, I/O Drawer (#6172)
        One - Power Cable, I/O Drawer to Media Drawer (#6179)
        One - PCI Single Ended SCSI Adapter (Media Drawer Attach) (#6206)
        Two - I/O Drawer PCI Planar, 10 Slot, 2 Integrated Ultra3 SCSI Ports (#6563)
        Four - Ultra3 SCSI 4-Pack Hot Swap Back Plane (#6564)
        One - Drawer Placement Indicator, 9U Position, Primary Rack (#4609)
        One - Language Specify (#9xxx)
    One - 7040-61R System Rack (24-inch rack with 42U available space):
        One - Front Door, Black with Copper Accent, Primary Rack (#6070)
        One - Rear Door, Slim Line, Primary or Secondary Rack (#6074)
        Two - Bulk Power Regulators (#6186)
        Two - Power Control,Four Cooling Fans, Three DC Power Converter Connections (#6187)
        Two - Power Distribution Assembly, Ten DC Power Converter Connections (#6188)
        One - Bulk Power Assembly, Redundant (#8690)
        One - Rack Content Specify, 7040/681 - 17U (#0190)
        One - Rack Content Specify, 7040/61D - 4U (#0191)
        One - Rack Content Specify, 7040/681 Feature #8692 - 1U (#0192)
        One - Rack Content Specify -7040/61R Feature #8490 - 8U (#0193)
        Two - Line Cords, 60A/240VAC, 6-AWG, 14 feet, IEC309 Plug (#8678)
        One - Language Specify (#9xxx)
```

The default configuration in Example A-4 lacks the mandatory rear door, and contains country (or region) specific features, chosen according to e-config country (or region) settings.

Example: A-4 pSeries 690 default configuration

Product	Description	Qty
7040-681	Rack Server 1:pSeries 690 Model 681 CEC	1
2624	32X (Max) SCSI-2 CD-ROM Drive	1
3155	Diskette Drive Cable	1
3255	Operator Panel Attachment Cable	1
4138	128MB Level 3 Cache (4 X 32MB), 400MHz	1
4181	8GB Memory Card, Inward Facing	1
4618	Drawer Placement Indicator, 18U Position, Primary Rack	1
4651	**Rack Indicator, Rack #1**	**1**
5005	**Software Preinstall (RS)**	**1**

5242	8-Way POWER4 Processor	1
5251	Processor Clock Card, Programmable	1
6161	Cable Group, Power Controller to CEC and Fans	1
6162	Interface Cable, Service Processor to Power Subsystem	1
6170	Power Converter Assembly, Central Electronics Complex	2
6181	Power Cable Group, CEC to Power Controller, First Processor Module	1
6198	Capacitor Book, Two Processor Modules	1
6404	Support Processor with Remote I/O Loop Attachment, Two Loops	1
6565	Backplane, Central Electronics Complex	1
8121	**Attachment Cable, Hardware Management Console to Host, 15-Meter**	**1**
8692	Media Drawer, Operator Panel, Diskette	1
9703	**Language Group: French**	**1**
7040-61D	I/O Drawer 1:Model 61D I/O Drawer	1
2122	SCSI Cable - I/O Drawer to Media Drawer	1
3145	Remote I/O Cable, 0.5M	1
3149	Remote I/O Cable, 2M	2
3158	36.4 GB DISK DRIVE ASSEMBLY	2
4609	Drawer Placement Indicator, 9U Position, Primary Rack	1
4651	**Rack Indicator, Rack #1**	**1**
6121	I/O Drawer Attachment Cable Group, Drawer Position #4609	1
6172	Power Converter Assembly, I/O Drawer	2
6179	Power Cable, I/O Drawer to Media Drawer	1
6206	Ultra SCSI Adapter - SE (PCI)	1
6563	I/O Drawer PCI Planar, 10 Slot, 2 Integrated Ultra3 SCSI Ports	2
6564	Ultra3 SCSI 4-Pack Hot Swap Back Plane	4
9703	**Language Group: French**	**1**
7040-61R	7040-61R : Rack 1:Model 61R Rack	1
0190	Rack Content Specify: 7040/681 - 17U	1
0191	Rack Content Specify: 7040/61D - 4U	1
0192	Rack Content Specify: Media Drawer Feature #8692 - 1U	1
0193	Rack Content Specify: 7040/61R Feature #8690 - 8U	1
4651	Rack Indicator, Rack #1	1
6070	Front Door, Black with Copper Accent, Primary Rack	1
6186	Bulk Power Regulator	2
6187	Power Controller, Four Cooling Fans, Three	2

	DC Power Converter Connections	
6188	Power Distribution Assembly	2
8677	**Line Cord, 8AWG, 14ft, No Plug**	**2**
8690	Bulk Power Assembly, Redundant	1
9703	**Language Group: French**	**1**

B

I/O loop cabling and performance

This section first presents five graphical views of loop cabling, using either RIO components, RIO-2 components, or a mix of the two technologies. The diagrams contains an indication of the data throughput burst and a sustained rate that can be expected in each part of the loops.

The three following figures present examples of the cabling of fully configured pSeries 690 with the maximum possible number of installed IO drawers.

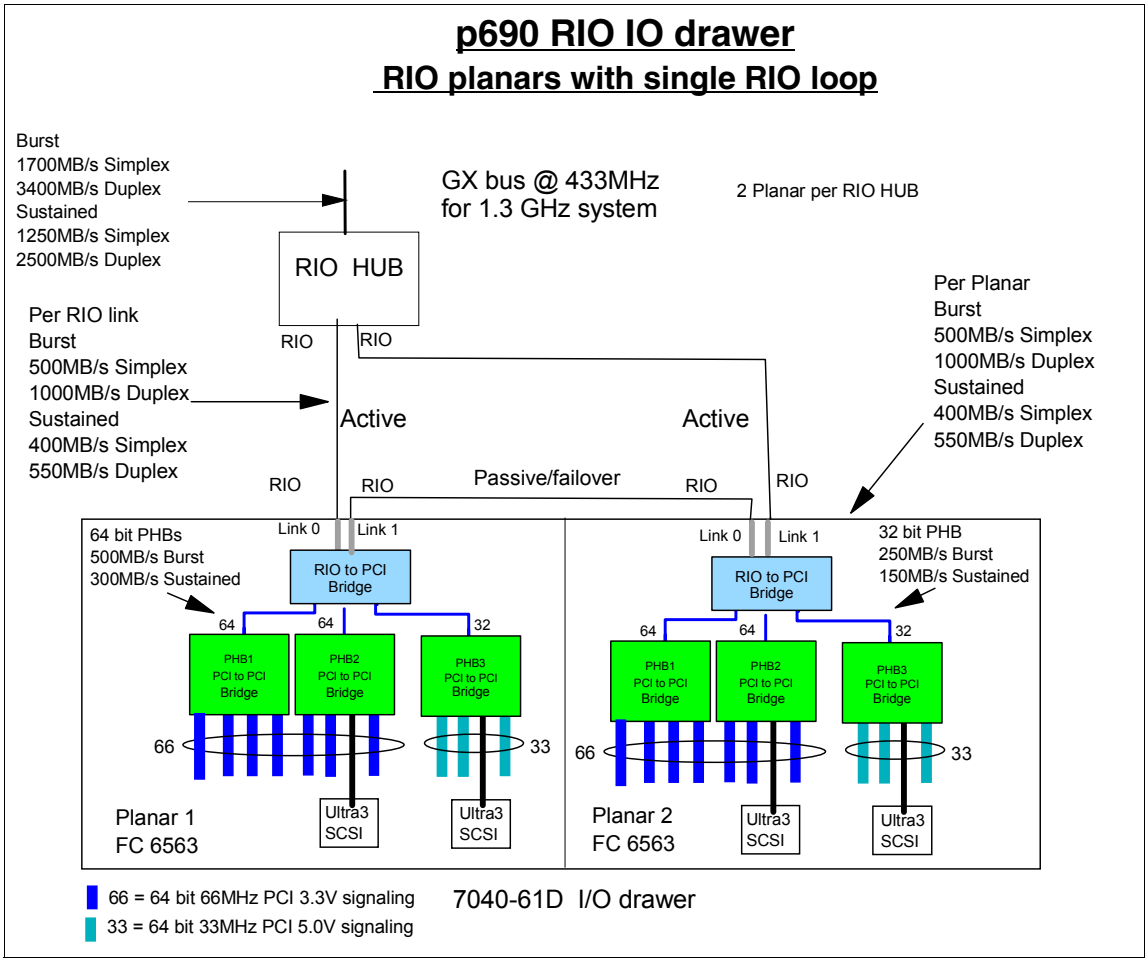


Figure B-1 RIO drawer, in single-loop mode on a 1.3 GHz server

In Figure B-1, all components are using the RIO technology. The two planers are connected to the same loop. Each planar therefore transfers data through only one RIO link.

The transfer rate within the CEC between the RIO book (RIO HUB) and the MCM depends on the processor speed.

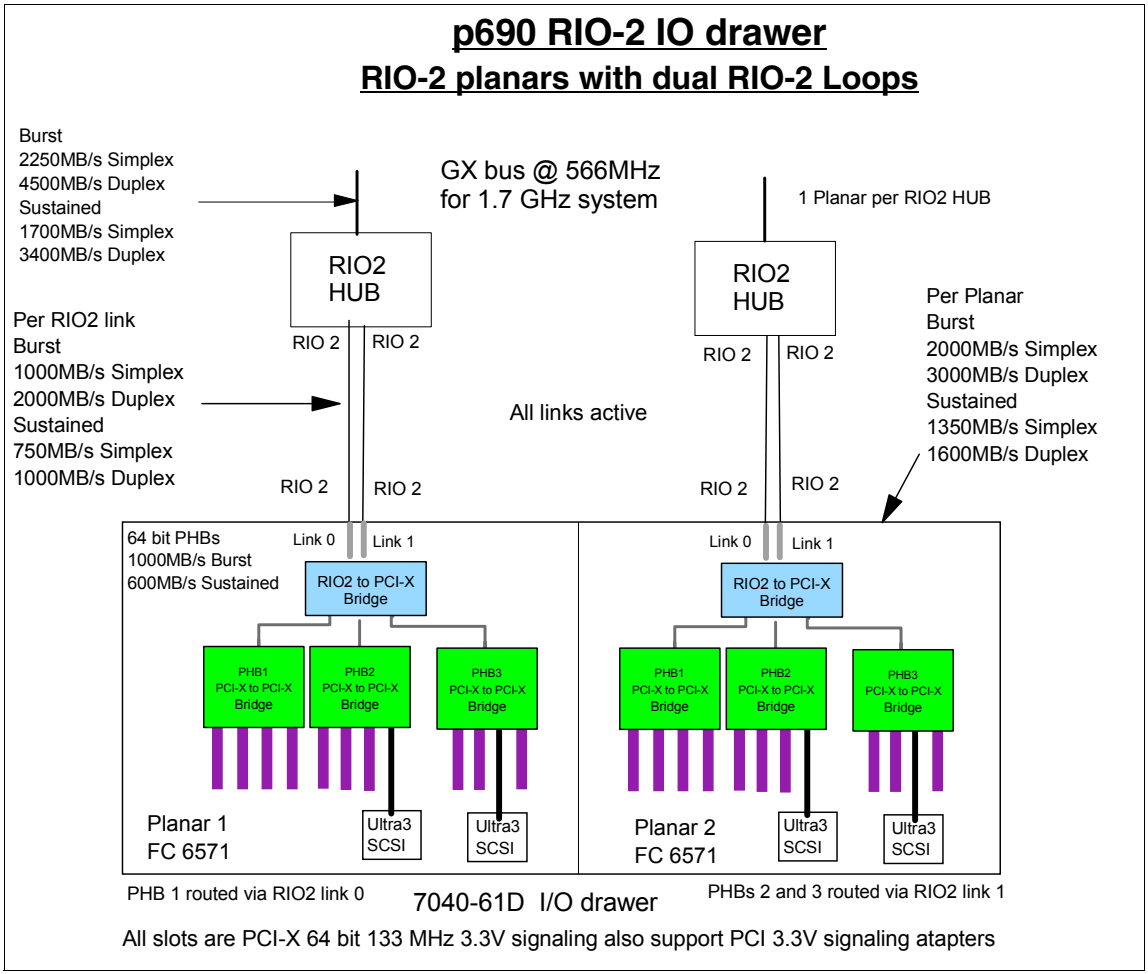


Figure B-2 RIO-2 drawer, in dual-loop mode on a 1.7GHz server

In Figure B-2, all components are using the RIO-2 technology. Each of the two planers are using the two links of a dedicated loop.

This is the recommended and default loop configuration for RIO-2 loops.

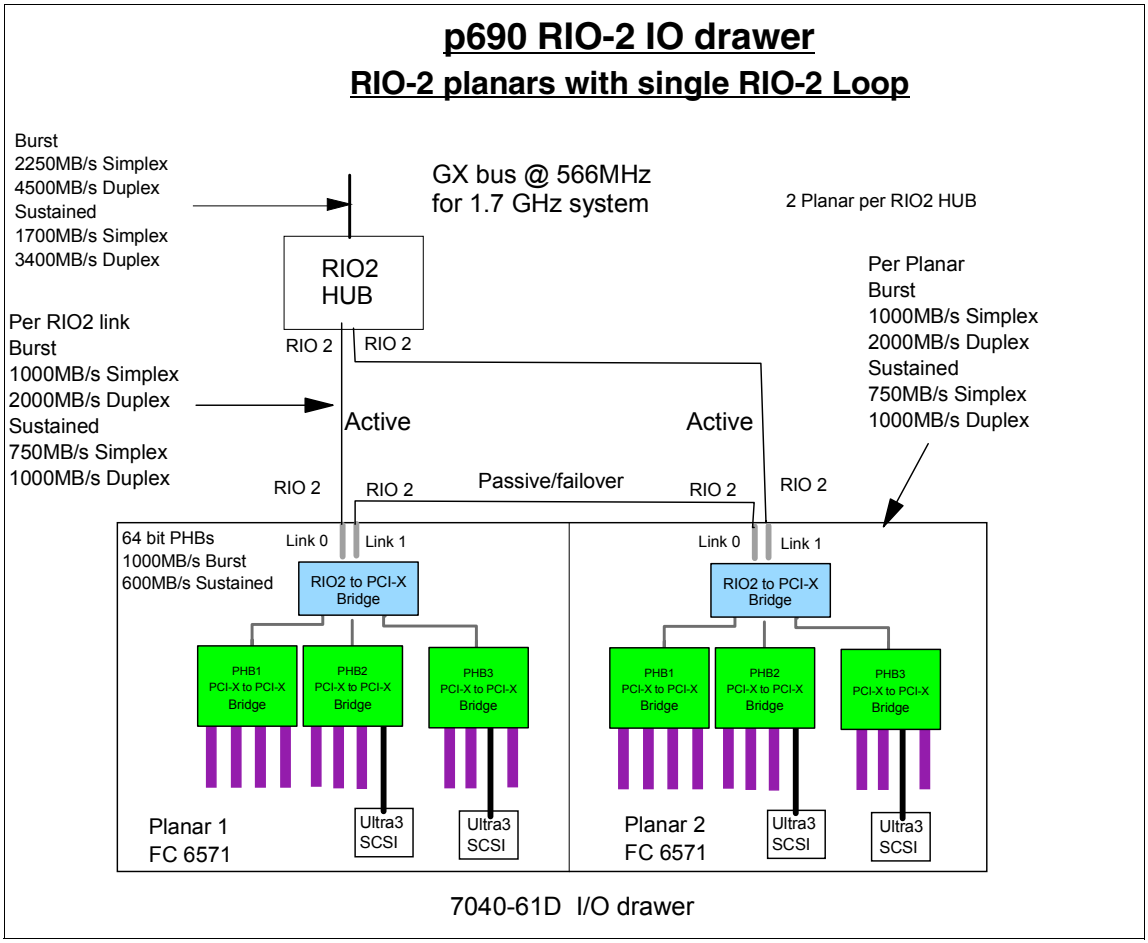


Figure B-3 RIO-2 drawer, in single-loop mode on a 1.7 GHz server

In Figure B-3, all components are using the RIO-2 technology, but the two planers are connected in a single-loop mode, as would be an RIO loop. Each planar therefore transfers data through only one RIO link.

This does not yield the best performance available using these components. However, the throughput of this RIO-2 loop is 32% higher than for the RIO loop of Figure B-1.

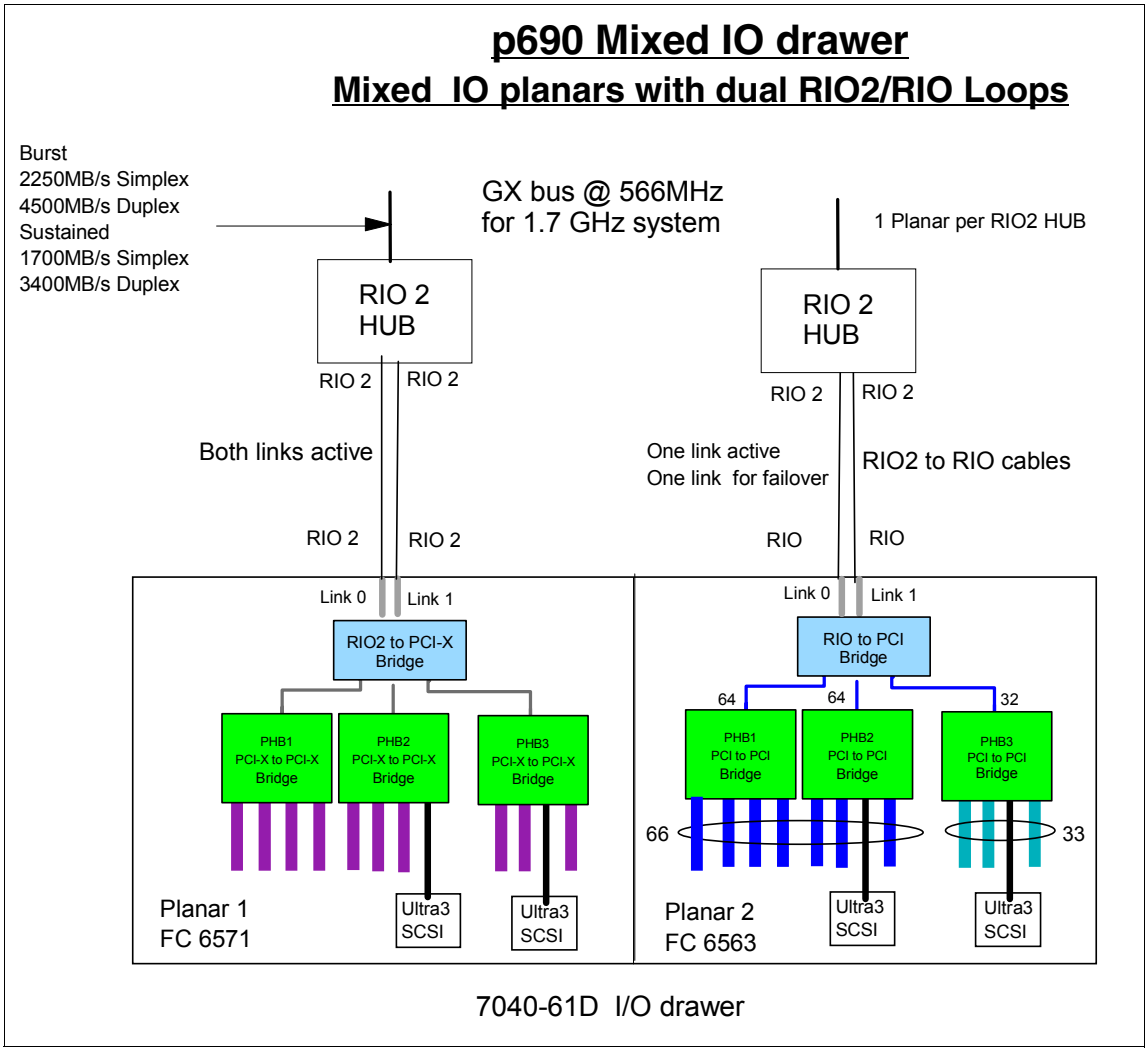


Figure B-4 Mixed RIO/RIO-2 drawer, in dual-loop mode on a 1.7 GHz server

In Figure B-4, one of the I/O planars in the drawer is using the RIO technology, therefore providing three I/O slots for PCI 5V adapters. Each of the two planers are connected in a single-loop mode, as would be an RIO loop. However, only the RIO-2 planar transfers data through to the links, while the RIO planar can fly transmit data through one RIO link.

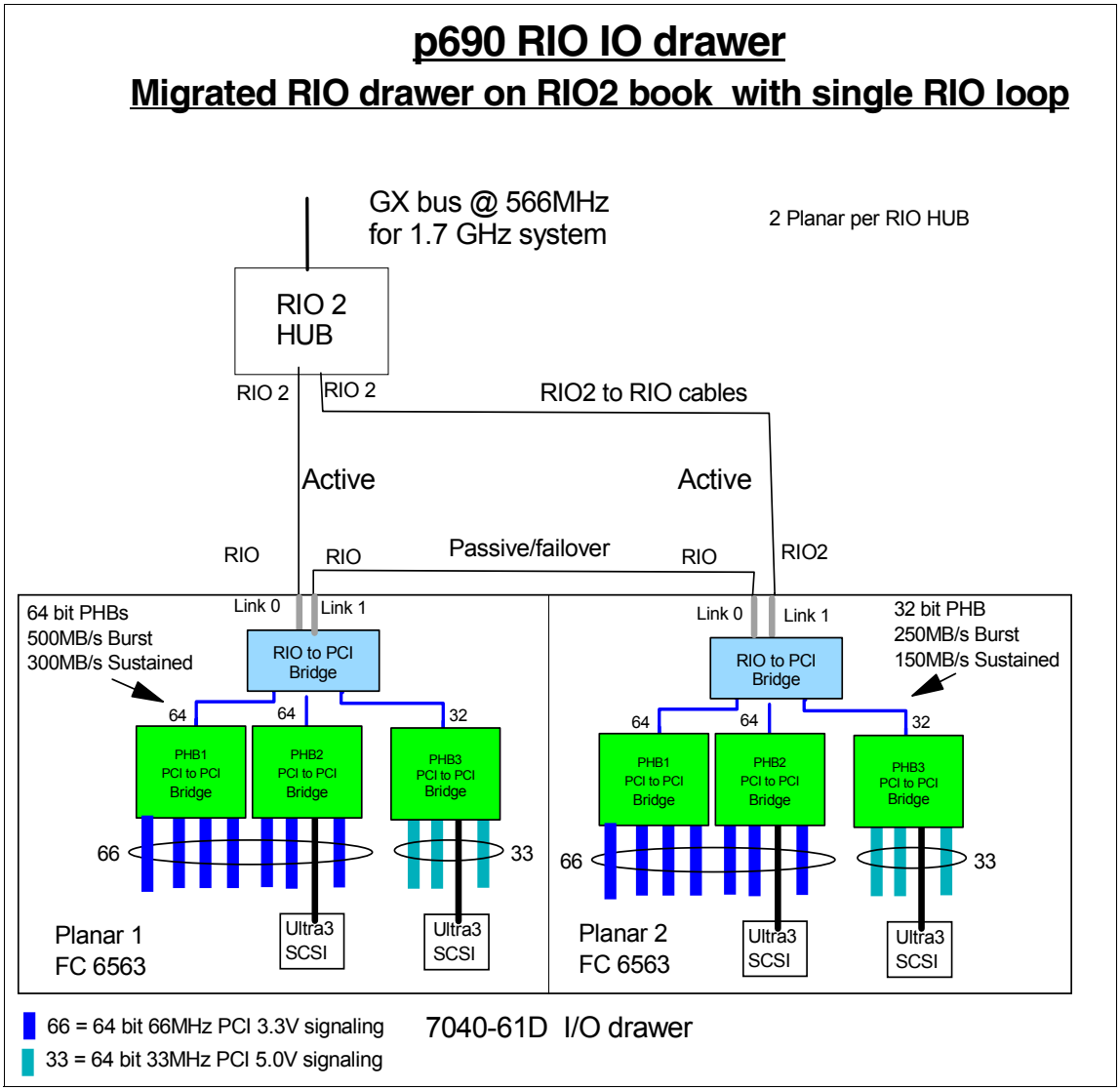


Figure B-5 Migrated single-loop mode on a 1.7 GHz server

Figure B-5, presents a configuration that can be found in a pSeries 690 where an MES has been installed, replacing the processors and RIO books by 1.7 GHz MCM and RIO-2 books.

The configuration of the RIO loop is unchanged, identical to the one presented in Figure B-1. However, the RIO cables have been replaced with RIO to RIO-2 cables.

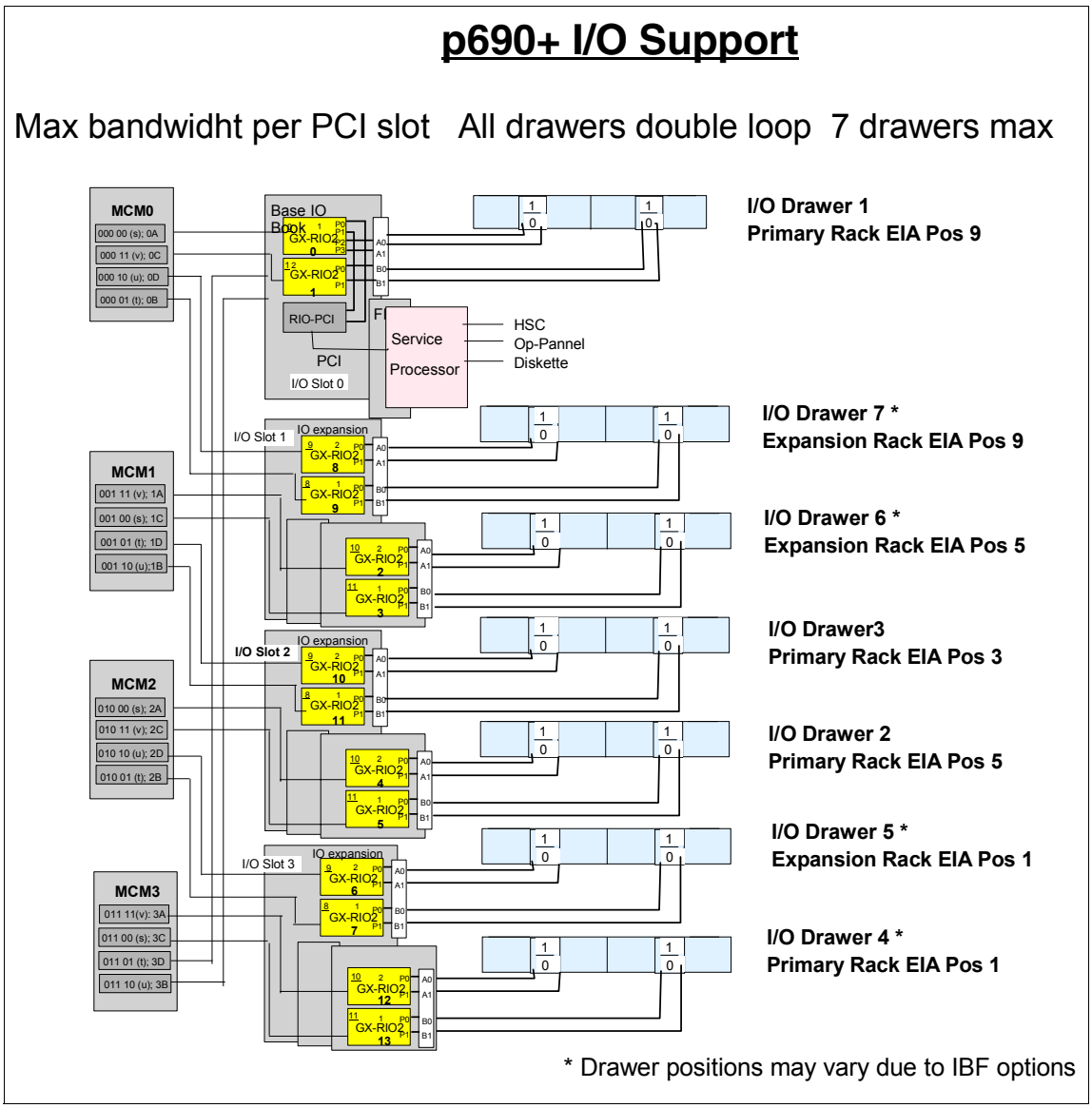


Figure B-6 Configuration for maximum bandwidth

In an I/O bound system where I/O adapters and internal disks access are the most critical bottleneck, you can obtain the maximum bandwidth on each adapter by having as many adapters as possible, and the smallest number of adapters for each uplink between the IO books and the MCM. This is provided by a seven IO drawers configuration, all connected in dual-loop mode (Figure B-6).

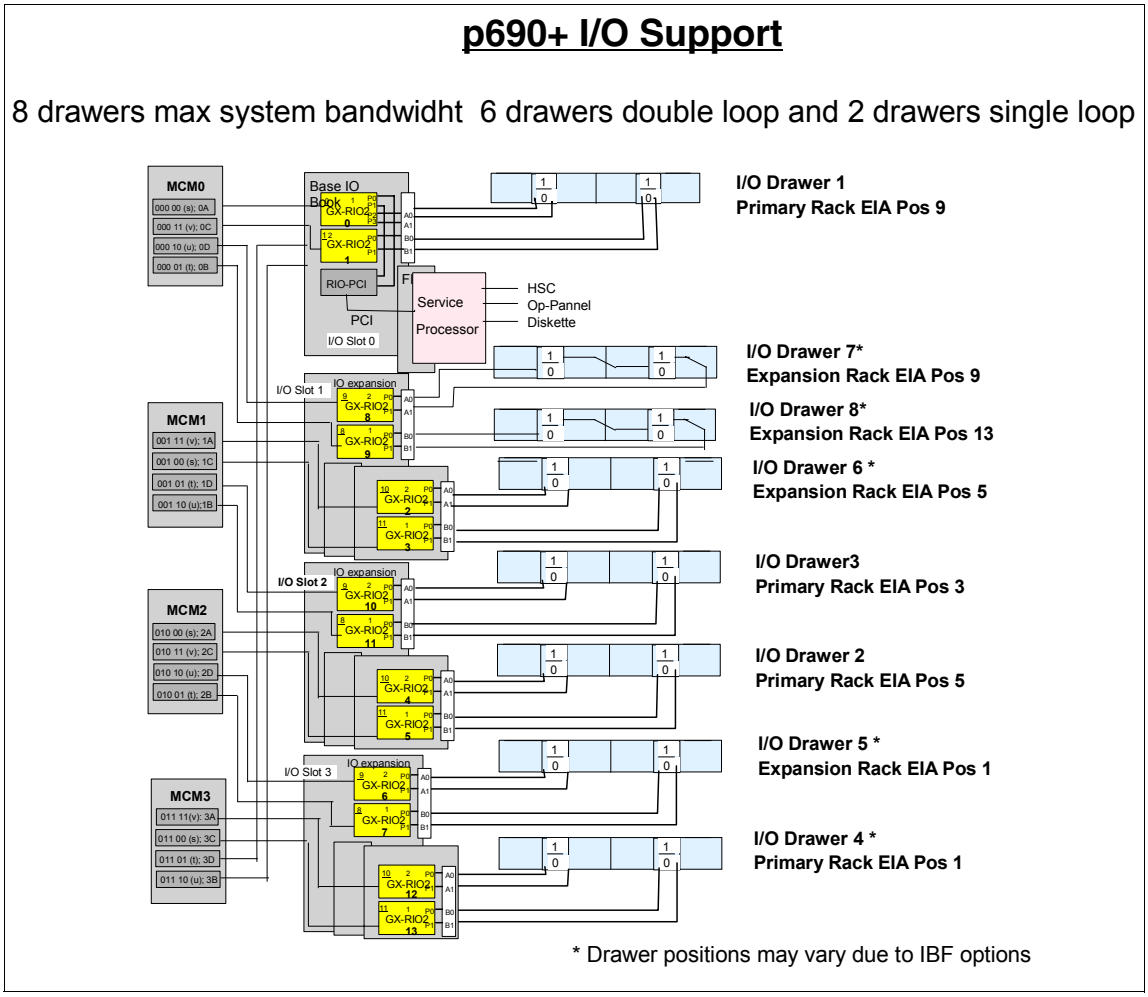


Figure B-7 Configuration for maximum number of adapters and disks

If you are looking for as many adapters and disks as possible in your configuration (for example, in a server consolidation environment with 32 partitions and many adapters per LPAR), and you also want to obtain the best possible I/O throughput, you need 8 I/O drawers, but since there are only 14 pairs of RIO-2 ports in the IO books, you need to configure two drawers in single-loop mode, and the six others in dual-loop mode as in Figure B-7.

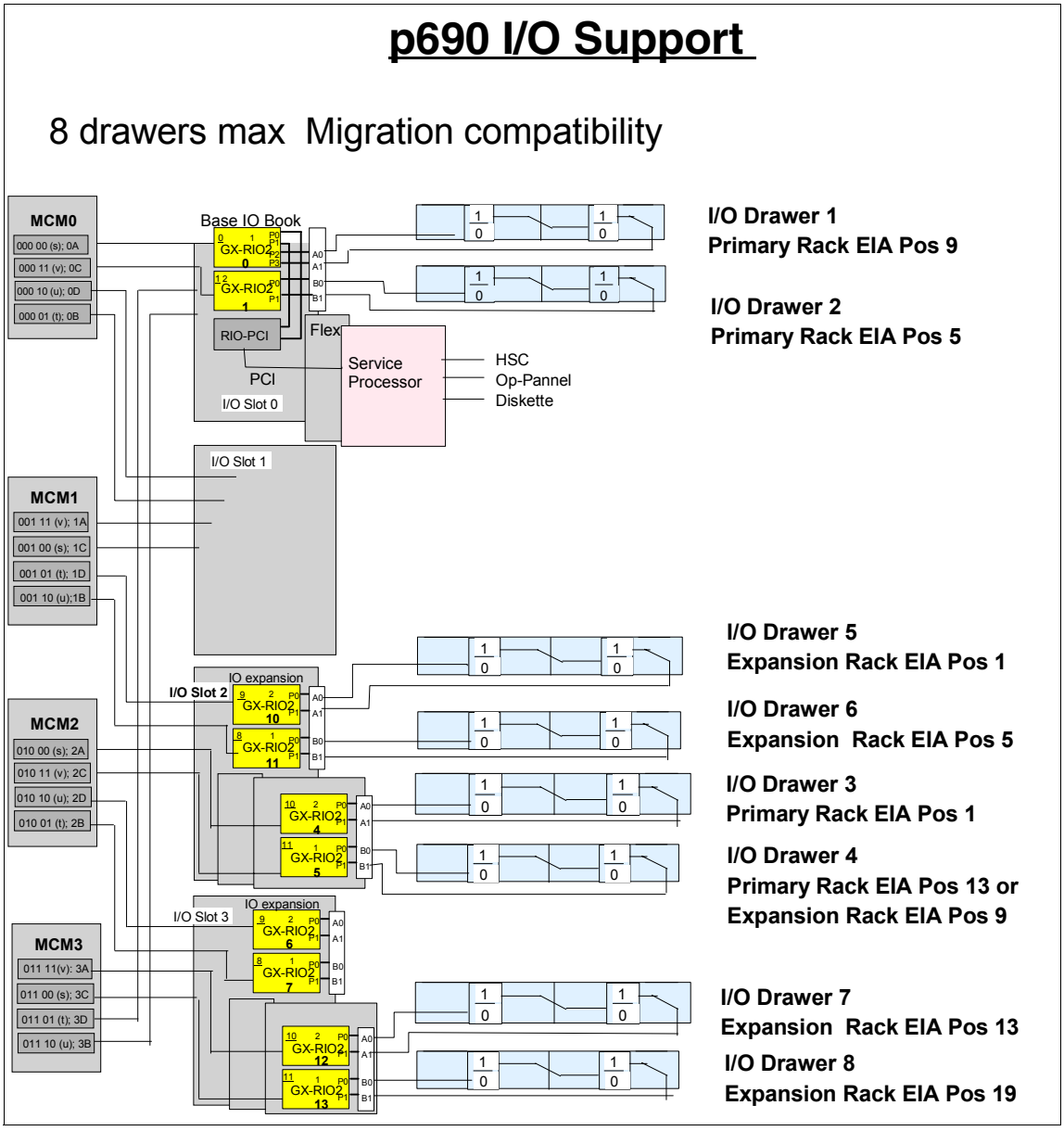


Figure B-8 Configuration after migration to RIO-2 Books without recabling.

Figure B-8 shows a valid configuration after migrating from RIO to RIO-2 books. The previous cabling topology has been kept, to maintain the LPAR configurations. However, the RIO cables have been replaced by RIO to RIO-2 cables.

Abbreviations and acronyms

AC	Alternating Current
AIX	Advanced Interactive eXecutive
APAR	Authorized Program Analysis Report
ATM	Asynchronous Transfer Mode
BI	Business Intelligence
BIST	Built-in Self Test
BPA	Bulk Power Assembly
BPC	Bulk Power Controller
BPF	Bulk Power Fan
BPR	Bulk Power Regulator
bps	Bits per Second
CD	Compact Disk
CD-ROM	Compact Disk - Read Only Media
CEC	Central Electronics Complex
cfm	Cubic feet per minute
CIM	Common Interface Model
CIU	Core Interface Unit
CMOS	Complementary Metal-Oxide-Silicon
CPU	Central Processing Unit
CRAD	Customer Requested Arrival Date
CRC	Cyclic Redundancy Check
CRM	Customer Relationship Management
CTA	Click to Accept
CUoD	Capacity Upgrade on Demand
DAT	Digital Audio Tape
DC	Direct Current
DCA	Distributed Converter Assembly
DDR	Double Data Rate
DGD	Dead Gateway Detection
DIMM	Dual Inline Memory Module
DLPAR	Dynamic logical partitioning
DMA	Direct Memory Access
DMTF	Distributed Management Task Force
DRAM	Dynamic Random Access Memory
DVD	Digital Versatile Disk
DVD-RAM	Digital Versatile Disk - Random Access Media
DVD-ROM	Digital Versatile Disk - Random Only Media
ECC	Error Checking and Correcting
e-config	IBM Configurator for e-business
EEH	Extended Error Handling
EIA	Electronic Industry Alliance
EMF	Electro-motive Force
EPO	Emergency Power On
EPOW	Early Power Off Warning
ERP	Enterprise Resource Planning
ESCON	Enterprise Systems Connection
FC	Feature Code
FDDI	Fibre Distributed Data Interface
FFDC	First Failure Data Capture
FIR	Fault Isolation Register
FLOP	Floating-point Operation

FQDN	Fully Qualified Domain Name
FRU	Field Replaceable Unit
ft	feet
GA	General Announcement
GB	Gigabyte
GHz	Gigahertz
HACMP	High Availability Cluster Multiprocessing
HAEH	High Availability Event Handler
HIPPI	High Performance Parallel Interface
HMC	IBM Hardware Management Console for pSeries
HPC	High Performance Computing
I/O	Input/Output
IBF	internal battery feature
IBM	International Business Machines Corporation
ID	identification
IDE	Integrated Data Equipment
IP	Internet Protocol
IPL	Initial Program Load
ISDN	Integrated Services Digital Network
ISV	Independent Software Vendor
ITSO	International Technical Support Organization
JTAG	Joint Test Action Group
kW	kilowatt
kwh	Kilowatt per Hour
L1	Level 1
L2	Level 2
L3	Level 3
LAN	Local Area Network
LDAP	Lightweight Directory Access Protocol
LED	Light Emitting Diode
LPAR	Logical Partition
LVD	Low Voltage Differential
LVM	Logical Volume Manager
MA	Maintenance Agreement
MB	Megabyte
Mbps	Megabit per Second
MCM	Multichip Module
MDA	Motor Drive Assembly
MES	Miscellaneous Equipment Specification
MHz	Megahertz
mm	millimeter
MPOA	Multiprotocol over ATM
MRPD	Machine ReportedProduct Data
MSA	Motor/Scroll Assembly
NC	Non-cacheable
NIM	Network Installation Management
NIS	Network Information Service
NVRAM	Non-volatile Random Access Memory
ODM	Object Data Manager
OLAP	Online Analytical Processing
PC	Personal Computer
PCI	Peripheral Component Interconnect
PFT	Page Frame Table
PHB	PCI Host Bridge
POR	Power On Reset
POST	Power-On Self-Test
POWER	Performance Optimization With Enhanced RISC
PSSP	Parallel System Support Program
QBB	Quad Building Block
RAM	Random Access Memory

RAN	Remote Asynchronous Node
RAS	Reliability, Availability, and Serviceability
RIO	Remote I/O
RISC	Reduced Instruction Set Computer
RMC	Resource Monitoring and Control
RML	Real Mode Limit
RMO	Real Mode Offset
RPM	Rotation per Minute
RTAS	Run Time Abstraction Services
SBE	Single Bit Error
SCM	Supply Chain Management
SCSI	Small Computer System Interface
SDRAM	Synchronous DRAM
SES	SCSI Enclosure Service
SMI	Synchronous Memory Interface
SMIT	System Management Interface Tool
SMP	Symmetric Multiprocessing
SMS	System Management Services
SOI	Silicon-on-Insulator
SP	IBM RS/6000 Scalable POWERparallel® Systems
SSA	Serial Storage Architecture
TB	Terabyte
TCE	Translation Control Entry
TCP	Transmission Control Protocol
TLB	Translation Look-aside Buffer
TOD	Time-of-Day
UPS	Uninterruptible Power Supply
URL	Uniform Resource Locator
USB	Universal Serial Bus
UTP	Untwisted pair
VHDCI	Very High Density Connector Interface
VM	Virtual Machine
VPD	Vital Product Data
WAN	Wide Area Network

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this redbook.

IBM Redbooks

For information on ordering these publications, see “How to get IBM Redbooks” on page 234. Note that some of the documents referenced here may be available in softcopy only.

- *AIX 5L Differences Guide,* SG24-5765
- *AIX Logical Volume Manager, From A to Z: Introduction and Concepts*, SG24-5432
- *A Practical Guide for Resource Monitoring and Control (RMC)*, SG24-6606
- *Managing AIX Server Farms*, SG24-6606
- *The Complete Partitioning Guide for IBM @server* pSeries Servers, SG24-7039
- *The POWER4 Processor Introduction and Tuning Guide*, SG24-7041

Other publications

These publications are also relevant as further information sources. (The publications marked with an asterisk (*) are located on the documentation CD-ROM that ships with the AIX operating system).

- *AIX 5L Version 5.2 AIX Installation in a Partitioned Environment* *
- *AIX 5L Version 5.2 Installation Guide and Reference* *
- *AIX 5L Version 5.2 Reference Documentation: Commands Reference* *
- *AIX 5L Version 5.2 System Management Guide: AIX 5L Version 5.2 Web-based System Manager Administration Guide* *
- *AIX 5L Version 5.2 System Management Guide: Communications and Networks* *
- *AIX 5L Version 5.2 System Management Guide: Operating System and Devices* *

- *Electronic Service Agent for pSeries and RS/6000 User's Guide*, available at:

  ftp://ftp.software.ibm.com/aix/service_agent_code/AIX/svcUG.pdf
- *Electronic Service Agent for pSeries Hardware Management Console User's Guide*, available at:

  ftp://ftp.software.ibm.com/aix/service_agent_code/HMC/HMCSAUG.pdf
- IBM @server *Cluster 1600 Harware Planning, Installation and Service*, GA22-7863
- *IBM @server pSeries 670 Installation Guide*, SA38-0613-01
- *IBM @server pSeries 670 Service Guide*, SA38-0615
- *IBM @server pSeries 670 User's Guide*, SA38-0614-01
- *IBM @server pSeries 690 Installation Guide*, SA38-0587
- *IBM @server pSeries 690 Service Guide*, SA38-0589
- *IBM @server pSeries 690 User's Guide*, SA38-0588
- *IBM Hardware Management Console for pSeries Maintenance Guide*, SA38-0603
- *IBM Hardware Management Console for pSeries Installation and Operations Guide*, SA38-0590
- *IBM Reliable Scalable Cluster Technology for AIX 5L RSCT Guide and Reference*, SA22-7889
- *IBM Reliable Scalable Cluster Technology for AIX 5L Messages*, GA22-7891
- *IBM Reliable Scalable Cluster Technology for AIX 5L Technical Reference*, SA22-7890
- *Installation Guide 61D I/O drawer 61R Second I/O Rack*, SA23-1281
- *Exploring Expect: A Tcl-based Toolkit for Automating Interactive Programs*, by Don Libes. O'Reilly & Associates, Inc., January 1995. ISBN 1565920902
- *Adapter, Devices, and Cable Information for Multiple Bus Systems*, SA38-0516
- *RS/6000 and eServer pSeries Diagnostics Information for Multiple Bus Systems,* SA38-0509
- *PCI Adapter Placement References*, SA38-0538
- *Site and Hardware Planning Information,* SA38-0508

You can access all of the pSeries hardware related documentation through the Internet at the following URL:

http://publib16.boulder.ibm.com/pseries/en_US/infocenter/base/hardware.htm

You can also access all of the AIX documentation through the Internet at the following URL:

http://publib16.boulder.ibm.com/pseries/en_US/infocenter/base/aix.htm

The following whitepapers can be found on the Internet:

- *IBM @server* pSeries 690 *Availability Best Practices* white paper

  http://www.ibm.com/servers/eserver/pseries/hardware/whitepapers/p690_avail.html
- *IBM @server* pSeries 690 *Configuring for Performance* white paper

  http://www.ibm.com/servers/eserver/pseries/hardware/whitepapers/p690_config.html
- *IBM @server* pSeries 690*: Reliability, Availability, Serviceability* white paper

  http://www.ibm.com/servers/eserver/pseries/hardware/whitepapers/p690_ras.html
- *IBM @server* pSeries 690 *with the HPC feature* white paper

  http://www.ibm.com/servers/eserver/pseries/hardware/whitepapers/p690_hpc.html
- *Linux for IBM @server pSeries: An overview for customers* white paper

  http://www.ibm.com/servers/eserver/pseries/linux/whitepapers/linux_pseries.html
- *Partitioning for the IBM @server* pSeries 690 *System* white paper

  http://www.ibm.com/servers/eserver/pseries/hardware/whitepapers/lpar.html
- *POWER4 System Microarchitecture* white paper

  http://www.ibm.com/servers/eserver/pseries/hardware/whitepapers/power4.html

Online resources

These Web sites are also relevant as further information sources:

- AIX 5L operating system and related IBM products information

  http://www.ibm.com/servers/aix/
- AIX toolkit for Linux applications

  http://www.ibm.com/servers/aix/products/aixos/linux/download.html
- Application availability on the AIX 5L operating system (alphabetical listing and advanced search options for IBM software products and third-party software products)

  http://www.ibm.com/servers/aix/products/
- CuOD process, brief explanation

  http://www.ibm.com/servers/eserver/pseries/cuod/tool.html
- IBM AIX 5L Solution Developer Application Availability Web page

  http://www.ibm.com/servers/aix/isv/availability.html

- IBM AIX: IBM Application Availability Guide Web page

  http://www.ibm.com/servers/aix/products/ibmsw/list
- IBM Configurator for e-business

  http://ftp.ibmlink.ibm.com/econfig/announce/index.htm
- IBM @server pSeries LPAR documentation and references Web site

  http://www.ibm.com/servers/eserver/pseries/lpar/resources.html
- Linux for pSeries system guide

  http://www.ibm.com/servers/eserver/pseries/linux
- Linux on pSeries information

  http://www.ibm.com/servers/eserver/pseries/linux/
- Microcode Discovery Service information

  http://techsupport.services.ibm.com/server/aix.invscoutMDS
- OpenSSH Web site

  http://www.openssh.com
- VPD Capture Service

  http://techsupport.services.ibm.com/server/aix.invscoutVPD

How to get IBM Redbooks

You can search for, view, or download Redbooks, Redpapers, Hints and Tips, draft publications and Additional materials, as well as order hardcopy Redbooks or CD-ROMs at this Web site:

ibm.com/redbooks

Index

D

E

F

G

H

I

J

K

Q

R

S

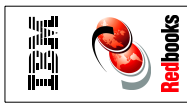
T

U

V

W

Z



IBM @server pSeries 670 and pSeries 690 System Handbook

(0.5” spine)
0.475”<->0.875”
250 <-> 459 pages

IBM @server pSeries 670 and pSeries 690 System Handbook

Redbooks

Component-based description of the hardware architecture

How to use e-config to configure pSeries 670 and pSeries 690

Capacity Upgrade on Demand explained

This IBM Redbook explains the IBM @server pSeries 670 Model 671 and the IBM @server pSeries 690 , a new level of UNIX servers providing world-class performance, availability, and flexibility. Capable of data center, application, and high performance computing, these servers include mainframe-inspired self-management and security to meet your most demanding needs.

This redbook includes the following topics:
-Overview of the pSeries 670 and pSeries 690
-Hardware architecture of the pSeries 670 and pSeries 690
-Using the IBM Configurator for e-business
-Capacity Upgrade on Demand
-Reliability, availability, and serviceability

This book is an ideal desk-side reference for IBM professionals and Business Partners, and technical specialists who support the pSeries 670 and pSeries 690, and for those who want to learn more about this radically new server in a clear, single-source handbook. It provides the necessary information to successfully order a pSeries 670 or pSeries 690 for a production environment and then, upon successful installation, configure service support functions, such as Service Focal Point and Inventory Scout.

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

SG24-7040-02 ISBN 0738499501